ENSTA ParisTech 1ère année
Statistiques (MA101), 2017-2018
christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

TD1: Population et échantillon
Eléments de corrigé

Exercice 1
Reconnaître les variables aléatoires parmi la liste suivante :

1. la moyenne de la population
2. la taille de l'échantillon
3. la moyenne de l'échantillon
4. la plus grande valeur de la population
5. la variance empirique de la population
6. la valeur observée de l'estimation de la variance de la population

variables aléatoires: 3, 5. Rem: la valeur observée est un résultat qui est un réel

Exercice 2
On considère un n-échantillon iid $(X_1, \ldots, X_n)$ d'une loi d'espérance μ connue et de variance σ^2 inconnue.

1. Soit $V_n = 1$. V_n est-il un estimateur de σ^2? Si oui, quel est son risque?

   V_n est la fonction de l'échantillon constante égale à 1, c'est donc un estimateur, par exemple de σ^2.

   $$\mathbb{R}(V_n, \sigma^2) = \text{var}(V_n) + Biais(V_n)^2 = 0 + (\mathbb{E}(V_n) - \sigma^2)^2 = (1 - \sigma^2)^2$$

2. Soit $W_n = \frac{1}{n}\sum_i (X_i - \mu)^2$. W_n est-il estimateur de σ^2? Si oui, quel est son risque?

   W_n est une fonction de l'échantillon qui dépend de μ connu à valeurs dans $\mathbb{R}^+$. Elle est calculable, c'est un estimateur. En utilisant la propriété d'un échantillon iid, $\text{var}(W_n) = \sum_i \text{var}((X_i - \mu)^2)/n^2 = \text{var}((X_i - \mu)^2)/n$ *et* $\mathbb{E}(W_n) = \sum_i \text{var}(X_i)/n = \sigma^2$

   $$\mathbb{R}(W_n, \sigma^2) = \text{var}(W_n) + Biais(W_n)^2 = \text{var}((X_i - \mu)^2)/n$$

3. W_n domine-t-il V_n?

   W_n ne domine pas V_n, puisque $\mathbb{R}(V_n, 1) = 0 < \mathbb{R}(W_n, 1)$.

4. On suppose que les X_i sont iid de loi exponentielle d'espérance μ connue. V_n domine-t-il W_n? Proposer un estimateur admissible de σ^2.

   On calcule $\mathrm{var}((X_i-\mu)^2)$. *La fonction génératrice des moments est* $\varphi(t) = (1-\mu t)^{-1}$. *On a* $\mathbb{E}(X^p) = [d^p\varphi(t)/dt^p]_{|t=0}$, *d'où* $\mathbb{E}(X^2) = 2\mu^2$, $\mathbb{E}(X^3) = 6\mu^3$, $\mathbb{E}(X^4) = 24\mu^4$ *soit* $\mathrm{var}((X_i-\mu)^2) = 8\mu^4$. *On a*

   $$\mathbb{R}(V_n, \sigma^2) - \mathbb{R}(W_n, \sigma^2) < 0 \Leftrightarrow n < 8\mu^4/(1-\mu^2)^2$$

   Il existe $\mu_{min}(n) \geq 0$, $\mu_{max} > 1$, *tels que* V_n *domine* W_n *sur* $[\mu_{min}, \mu_{max}]$ *et est dominé sur le complémentaire. Donc* V_n *ne domine pas* W_n *sur* $\mathbb{R}^+$. *Aucun des deux estimateurs n'est uniformément meilleur. Comme* μ *est connu,* μ^2 *est admissible pour* σ^2.

Exercice 3
Soit $(X_1, \ldots, X_n)$ un n-échantillon iid d'une loi d'espérance μ et de variance $\sigma^2 > 0$ inconnues.

1. Donner une condition nécessaire et suffisante sur les constantes réelles $a_1, \ldots, a_n$ pour que $\sum_i a_i X_i$ soit un estimateur sans biais de μ.

   on a : $\mathbb{E}(\sum_i a_i X_i) = \mu \Leftrightarrow \sum_i a_i = 1$

2. Parmi les estimateur sans biais de μ de la forme $\sum_i a_i X_i$, quel est celui de risque quadratique minimal?

   Le biais étant nul, le risque est égal à la variance $\sum_i a_i^2\sigma^2$, *sous la contrainte* $\sum_i a_i = 1$. *Soit on utilise un multiplicateur de Lagrange, soit on écrit*

   $$C(a) = \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n-1} a_i^2 + a_n^2 = \sum_{i=1}^{n-1} a_i^2 + (1 - \sum_{i=1}^{n-1} a_i)^2$$

   puis on dérive par rapport à a_i, $i = 1, \ldots, n-1$:

   $$\partial C(a)/\partial a_i = 2a_i - 2(1 - \sum_{i=1}^{n-1} a_i) = 2(a_i - a_n) = 0$$

   d'où $\tilde{a}_i = \tilde{a}_n$ *pour tout* $i = 1, \ldots, n$ *et* $1 = \sum_i \tilde{a}_i = na_n$ *d'où l'optimum est* $\tilde{a}_i = 1/n$. *L'estimateur de risque quadratique minimal parmi les estimateurs de l'espérance sans biais et de la forme* $\sum_i X_i/n$ *est la moyenne empirique* $\bar{X}$, *qui est donc admissible dans cette classe d'estimateur.*

Exercice 4
On souhaite estimer par échantillonnage la proportion de ménages de plus de 75 ans possédant un micro-ordinateur.

1. Modéliser cette situation.

   La population $\mathcal{P}$ des ménages de plus de 75 ans étant très grande, on peut la considérer comme infinie. On effectue un tirage de n ménages. Sur chaque ménage de l'échantillon, on observe X_i qui vaut 1 si le ménage est équipé d'un ordinateur et 0 sinon. Ainsi, $(X_1, \ldots, X_n)$ est un n-échantillon iid de loi mère de Bernoulli $\mathcal{B}(1, p)$ où la proportion $p = \mathbb{E}(X_i)$ est à estimer.

2. Définir un estimateur non biaisé de la proportion et calculer son risque.

   On peut considérer la variable aléatoire $\bar{X}$ fonction de l'échantillon pour estimer p. On a $\mathbb{E}(\bar{X}) = p$ (estimateur non biaisé) et $\mathrm{var}(\bar{X}) = p(1-p)/n$, égale au risque ici.

3. La proportion à estimer étant inconnue, proposer une taille d'échantillon pour que l'erreur standard de l'estimation soit inférieure à 0.02.

   L'erreur standard d'estimation est $s = \sqrt{\mathrm{var}(\bar{X})} = \sqrt{p(1-p)/n}$. La proportion p étant inconnue, on peut la majorer par $s \leq (2\sqrt{n})^{-1}$ d'où $n = 1/(4 \times 0.02^2) = 625$

4. Sachant par ailleurs que cette proportion est entre 5% et 15%, quelle taille d'échantillon préconiser?

   De part la forme de la fonction $p(1-p)$, on majore la variance en la calculant pour $p = 15\%$, $n = .15 \times .85/0.02^2 = 319$. La connaissance de l'intervalle d'appartenance de la proportion permet une majoration plus fine, et donc une valeur de n moins importante.

Exercice 5

Dans un bassin, on a observé la présence de quatre espèces de poissons. On souhaite estimer les proportions $p_1, \ldots, p_4$ de chacune de ces espèces.

L'expérience suivante a été réalisée: on capture un poisson, on note son espèce, puis on le relâche dans le bassin. On recommence n fois l'expérience.

On observe donc $(n_1, \ldots, n_4)$ où n_i est le nombre de poissons de la i-ème espèce parmi les n poissons capturés. On interprète $(n_1, \ldots, n_4)$ comme une réalisation d'un vecteur aléatoire $(N_1, \ldots, N_4)$.

1. Donner une hypothèse plausible en ce qui concerne la loi du vecteur $(N_1, \ldots, N_4)$. Nous allons travailler sous cette hypothèse pendant le reste de l'exercice.

   Soit $(X_1, \ldots, X_n)$ un n-échantillon tel que X_1 prend les valeurs

$$\begin{gathered} x_1 \text{ avec probabilité } p_1 \\ x_2 \text{ avec probabilité } p_2 \\ \vdots \\ x_m \text{ avec probabilité } p_m, \end{gathered}$$

christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

avec $p_1 + \cdots + p_m = 1$. *Pour tout* $k \in \{1, \ldots, m\}$, *posons* $N_k = \sum_{i=1}^{n} \mathbb{1}\{X_i = x_k\}$, *le nombre de fois que la valeur* x_k *est prise par les* X_i, $1 \leq i \leq n$.
Nous avons $\forall (n_1, \ldots, n_m) \in \mathbb{N}^m$

$$\mathbb{P}(N_1 = n_1, N_2 = n_2 \ldots, N_m = n_m) = \begin{cases} n! \prod_{\ell=1}^{m} \dfrac{p_\ell^{n_\ell}}{n_\ell!} & si \ \sum_{\ell=1}^{m} n_\ell = n, \\ 0 & sinon. \end{cases}$$

On dit que $(N_1, \ldots, N_m)$ *suit la loi multinomiale de paramètres* n *et* $(p_1, \ldots, p_m)$.
Nous faisons l'hypothèse que chaque expérience est indépendante et de même loi. Dans ce cas le vecteur (N_1, N_2, N_3, N_4) *suit une loi multinomiale de paramètres* n *et* (p_1, p_2, p_3, p_4).

2. Calculer $\mathbb{E}[N_i]$, $\mathrm{var}(N_i)$ et $\mathrm{Cov}(N_i, N_j)$, $i \neq j$ en fonction des paramètres.

pour tout $k \in \{1, \ldots, m\}$

$$\mathbb{E}[N_k] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{X_i = x_k\}] = \sum_{i=1}^{n} \mathbb{P}(X_i = x_k) = np_k,$$

$$\mathrm{var}(N_k) = \sum_{i=1}^{n} \mathrm{var}(\mathbb{1}\{X_i = x_k\}) = np_k(1 - p_k),$$

puisque $X_1, \ldots, X_n$ *sont i.i.d. Enfin, pour tout* $1 \leq k \neq l \leq m$,

$$\begin{aligned} \mathrm{cov}(N_k, N_l) =& \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i = x_k\} \sum_{j=1}^{n} \mathbb{1}\{X_j = x_l\}\right) \\ & - \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i = x_k\}\right) \mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}\{X_j = x_\ell\}\right) \\ =& \underbrace{\sum_{i,j} \mathbb{E}(\mathbb{1}\{X_i = x_k\} \mathbb{1}\{X_j = x_\ell\})}_{=0 \ si \ i=j} - n^2 p_k p_l \\ =& - np_k p_\ell \end{aligned}$$

Remarquons que, si $m = 2$ *et* $p_1 = 1 - p_2 = p$ *alors nous retombons sur la loi binomiale de paramètre* n *et* p.

3. Proposer des estimateurs sans biais des proportions $p_1, \ldots, p_4$.

L'estimateur $\widehat{p}_i = \frac{N_i}{n}$ *est un estimateur sans biais de* p_i, *pour tout* $1 \leq i \leq 4$.

On fait une hypothèse paramétrique sur la forme des probabilités $p_1, \ldots, p_4$: on suppose qu'il existe des réels positifs (θ, ν) tels que

$$p_1 = \theta + \nu, \ p_2 = \theta - \nu, \ p_3 = \nu, \ p_4 = 1 - 2\theta - \nu.$$

On veut maintenant estimer les paramètres θ et ν.

christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

4. Construire un estimateur sans biais de ν.

   L'estimateur $\widehat{\nu} = \widehat{p}_3 = \frac{N_3}{n}$ *est non biaisé.*

5. Construire un estimateur T_1 sans biais de θ à partir des variables (N_1, N_3), et un estimateur T_2 sans biais de θ à partir des variables (N_2, N_3).

   Comparer ces deux estimateurs.

   On a: $\theta = p_1 - p_3$ *d'où l'estimateur sans biais* $T_1 = \widehat{p}_1 - \widehat{p}_3 = \frac{N_1 - N_3}{n}$.
   On a: $\theta = p_2 + p_3$ *d'où l'estimateur sans biais* $T_2 = \widehat{p}_2 + \widehat{p}_3 = \frac{N_2 + N_3}{n}$.
   Les deux estimateurs sont sans biais, on compare leur variance :

$$\begin{aligned}
\mathrm{var}(T_1) &= \frac{1}{n^2}\left(\mathrm{var}(N_1) + \mathrm{var}(N_3) - 2\,\mathrm{cov}(N_1, N_3)\right) \\
&= \frac{1}{n}\left(p_1(1-p_1) + p_3(1-p_3) + 2p_1p_3\right) \\
&= \frac{1}{n}\left(\theta + \nu - (\theta+\nu)^2 + \nu - \nu^2 + 2(\theta+\nu)\nu\right) \\
&= \frac{1}{n}\left(\theta + \nu - \theta^2 - \nu^2 - 2\theta\nu + \nu - \nu^2 + 2\theta\nu + 2\nu^2\right) = \frac{1}{n}\left(\theta - \theta^2 + 2\nu\right),
\end{aligned}$$

$$\begin{aligned}
\mathrm{var}(T_2) &= \frac{1}{n^2}\left(\mathrm{var}(N_2) + \mathrm{var}(N_3) + 2\,\mathrm{cov}(N_2, N_3)\right) \\
&= \frac{1}{n}\left(p_2(1-p_2) + p_3(1-p_3) - 2p_2p_3\right) \\
&= \frac{1}{n}\left(\theta - \nu - (\theta-\nu)^2 + \nu - \nu^2 - 2(\theta-\nu)\nu\right) \\
&= \frac{1}{n}\left(\theta - \nu - \theta^2 - \nu^2 + 2\theta\nu + \nu - \nu^2 - 2\theta\nu + 2\nu^2\right) = \frac{1}{n}\left(\theta - \theta^2\right)
\end{aligned}$$

   et donc $\mathbb{R}(T_1, \theta) = \mathrm{var}(T_1) > \mathrm{var}(T_2) = \mathbb{R}(T_2, \theta)$ *: l'estimateur* T_2 *est meilleur que* T_1.

Exercice 6

On considère $\mathcal{P} = \{u_1, \ldots, u_N\}$ une population de taille finie N, où les u_j représentent les différentes unités de la population, dont les valeurs peuvent être identiques. On se place dans le cadre du tirage simple sans remise d'un n-échantillon de $\mathcal{P}$.

1. On considère une population $\mathcal{P}$ constituée des $N = 5$ individus de valeurs suivantes: 1, 2, 2, 4, 8. Calculer la moyenne μ, la variance σ^2, et l'écart type corrigé σ^* de la population.

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j = \frac{1+2+2+4+8}{5} = 3.4$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j-\mu)^2 = \frac{1}{N}\sum_{j=1}^{N}u_j^2 - \mu^2 = \frac{1^2+2\times 2^2+4^2+8^2}{5} - 3.4^2 = 6.24$$

μ et σ^2 représente l'espérance et la variance de la loi qui affecte le poids $1/5$ à chacun des individus de la population. L'écart-type corrigé $\sigma^* = \sqrt{\frac{N}{N-1}\,\mathrm{var}} = \sqrt{6.24\times 5/4} = 2.79$ n'a pas ici de signification particulière.

2. Calculer la distribution d'échantillonnage de la moyenne $\bar{X}$ d'un échantillon simple de taille $n=2$, en générant tous les échantillons possibles. Calculer l'espérance et la variance de la distribution d'échantillonnage.
   Il y a $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$ échantillons non ordonnés

$\sum x_i$	1	2	2	4	8
1	/	3	3	5	9
2		/	4	6	10
2			/	6	10
4				/	12
8					/

$\bar{x}$	1	2	2	4	8
1	/	1.5	1.5	2.5	4.5
2		/	2	3	5
2			/	3	5
4				/	6
8					/

On en déduit la loi de $\bar{X}$

$\bar{x}$	1.5	2	2.5	3	4.5	5	6
proba	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

L'espérance de la loi d'échantillonnage de la moyenne est

$$\mathbb{E}(\bar{X}) = \frac{1.5\times 2+2\times 1+2.5\times 1+3\times 2+4.5\times 1+5\times 2+6\times 1}{10} = 3.4$$

et la variance est

$$\mathrm{var}(\bar{X}) = \sum_j p_j(\bar{x}_j-\mu)^2 = \sum_j p_j(\bar{x}_j)^2 - \mu^2 = 2.34$$

3. Soit un n-échantillon $E=(X_1,\dots,X_n)$ de $\mathcal{P}$. On note ξ_j la variable aléatoire qui vaut 1 si u_j appartient au n-échantillon et 0 sinon. Calculer $\mathrm{var}(\xi_j)$ et $\mathrm{cov}(\xi_j,\xi_{j'})$.
   On remarque que $\xi_j = \mathbb{1}_{u_j\in E}$ est une variable de Bernoulli $\mathcal{B}(1,f)$ de paramètre le taux de sondage

$$\mathbb{P}(\xi_j=1) = \mathbb{P}(u_j\in E) = \frac{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{n}{N} = f$$

christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

Son espérance vaut donc f et sa variance $f(1-f)$. De plus, pour $j \neq j'$,

$$\mathbb{P}(\xi_j = 1, \xi_{j'} = 1) = \mathbb{P}(u_j \in E \cap u_{j'} \in E) = \frac{\begin{pmatrix} N-2 \\ n-2 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{n(n-1)}{N(N-1)}.$$

On en déduit pour $j \neq j'$,

$$cov(\xi_j, \xi_{j'}) = \mathbb{E}(\xi_j \xi_{j'}) - \mathbb{E}(\xi_j)\mathbb{E}(\xi_{j'}) = 1 \times \frac{n(n-1)}{N(N-1)} - f^2 = -\frac{f(1-f)}{N-1}$$

4. Exprimer $\bar{X}$ en fonction des ξ_j. En déduire que $\mathbb{E}(\bar{X}) = \mu$ et

$$\operatorname{var} \bar{X} = \frac{\sigma^{*2}}{n}\left(1 - \frac{n}{N}\right) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Vérifier avec les données de la question 1.
On écrit $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\sum_{j=1}^N u_j \xi_j$, ce qui met en lumière que ce qui est aléatoire, c'est ξ_j (ie, ce qui est lié au tirage), et pas u_j. D'où par linéarité de l'espérance

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^N u_j \xi_j\right) = \frac{1}{n}\sum_{j=1}^N u_j \mathbb{E}(\xi_j) = \frac{1}{n}\sum_{j=1}^N u_j \frac{n}{N} = \frac{1}{N}\sum_{j=1}^N u_j = \mu$$

$$\begin{aligned}
\operatorname{var}(\bar{X}) &= \operatorname{var}\left(\frac{1}{n}\sum_{j=1}^N u_j \xi_j\right) \\
&= \frac{1}{n^2}\left(\sum_{j=1}^N (u_j^2 \operatorname{var}(\xi_j)) + \sum_{j \neq j'} \operatorname{cov}(\xi_j, \xi_{j'})\right) \\
&= \frac{1}{n^2}\left(\sum_j u_j^2\right)\frac{n}{N}(1-f) + \frac{1}{n^2}\sum_{j \neq j'} u_j u_{j'}\left(-\frac{n}{N(N-1)}(1-f)\right)
\end{aligned}$$

Or, $(N\mu)^2 = \left(\sum_j u_j\right)^2 = \sum_j u_j^2 + \sum_{j\neq j'} u_j u_{j'}$ *d'où*

$$\begin{aligned}
\mathrm{var}(\bar{X}) &= \frac{1-f}{nN}\sum_j u_j^2 - \frac{1-f}{nN(N-1)}\left(N^2\mu^2 - \sum_j u_j^2\right) \\
&= \frac{1-f}{nN}\sum_j u_j^2\left(1+\frac{1}{N-1}\right) - \frac{1-f}{n(N-1)}N\mu^2 \\
&= \frac{1-f}{n(N-1)}\left(\frac{N}{N}\sum_j u_j^2 - N\mu^2\right) \\
&= \frac{1-f}{n(N-1)}\sum_j (u_j-\mu)^2 \\
&= \frac{1-f}{n}\sigma^{*2} = \frac{1-f}{n}\frac{N}{N-1}\sigma^2 = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)
\end{aligned}$$

On vérifie que l'espérance de $\bar{X}$ *de la question* 2 *est égale à* μ. *De plus* $\mathrm{var}(\bar{X}) = 2.34 = \frac{6.24\times 5/4}{2}(1-\frac{2}{5}) = \frac{6.24}{2}(1-\frac{1}{4})$

Exercice 7

En échantillonnage stratifié, la population est partitionnée en sous-populations ou strates, qui sont échantillonnées indépendamment. Les résultats sur les strates sont ensuite combinés pour estimer le paramètre de la population totale. On considère K strates de taille $N_1, \ldots, N_K$ d'une population de taille N. La moyenne et variance de la k-ième strate sont notées μ_k et σ_k^2.

1. Calculer la moyenne μ et la variance σ^2 de la population totale en fonction de celles des strates.
   Pour $k = 1, \ldots, K$ *et* $i = 1, \ldots, N_k$, *soit* x_{ik} *la valeur de la* i*-ème unité de la* k*-ième strate. On note* $p_k = N_k/N$ *le poids de la strate* k. *On a:*

$$\mu = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} x_{ik} = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k = \sum_{k=1}^{K} p_k\mu_k$$

   La variance se décompose en un terme de variance inter- strates et un terme *de variance intra-strate:*

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ik}-\mu)^2 \\
&= \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ik}-\mu_k+\mu_k-\mu)^2 \\
&= \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{N_k}(x_{ik}-\mu_k)^2 - 2\underbrace{\sum_{i=1}^{N_k}(x_{ik}-\mu_k)}_{=0}(\mu_k-\mu)+N_k(\mu_k-\mu)^2\right) \\
&= \underbrace{\sum_k p_k\sigma_k^2}_{\textit{variance intra}} + \underbrace{\sum_k p_k(\mu_k-\mu)^2}_{\textit{variance inter}}
\end{aligned}
$$

2. Pour chaque strate k, un échantillonnage simple de taille n_k est effectué. Soit $\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ik}$ la moyenne calculée de l'échantillon de la strate k et soit $\bar{X}_s = \sum_{k=1}^{K} p_k\bar{X}_k$ la moyenne calculée par stratification. Montrer que $\mathbb{E}(\bar{X}_s) = \mu$ et
$$\mathrm{var}(\bar{X}_s) = \sum_{k=1}^{K} p_k^2\frac{\sigma_k^2}{n_k}\left(1-\frac{n_k-1}{N_k-1}\right)$$
En déduire la variance de la moyenne stratifiée dans le cas où toutes les sous populations sont infinies.
On se placera dans ce cas pour la suite.
On note d'abord que $\mathbb{E}(\bar{X}_k) = \frac{1}{n_k}\sum_{i=1}^{N_k}\mathbb{E}(X_{ik} = \frac{1}{n_k}n_k\mu_k = \mu_k$, *d'où*
$$\mathbb{E}(\bar{X}_s) = \sum_k p_k\mathbb{E}(\bar{X}_k) = \sum_k p_k\mu_k = \mu$$
Les strates étant indépendantes,
$$\mathrm{var}(\bar{X}_s) = \sum_k p_k^2\,\mathrm{var}(\bar{X}_k) = \sum_k p_k^2\frac{\sigma_k^2}{n_k}\left(1-\frac{n_k-1}{N_k-1}\right) \to_{N_k\to\infty} \sum_k p_k^2\frac{\sigma_k^2}{n_k}$$

3. Déterminer $n_1,\dots,n_k$ minimisant $\mathrm{var}(\bar{X}_s)$. En déduire la variance de $\bar{X}_{so}$, l'estimateur stratifié utilisant l'allocation optimale.
On utilise un multiplicateur de Lagrange λ. Le lagrangien à minimiser est
$$\mathcal{L} = \sum_k p_k^2\frac{\sigma_k^2}{n_k} + \lambda\left(\sum_k n_k - n\right)$$

christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

On le dérive par rapport à n_k, $k = 1, \dots K$:

$$\frac{\partial \mathcal{L}}{\partial n_k} = -\sum_k \frac{p_k^2 \sigma_k^2}{n_k^2} + \lambda, \ \textit{soit}\ n_k = \frac{p_k \sigma_k}{\sqrt{\lambda}}$$

De plus, la contrainte doit être satisfaite, donc

$$\sum_k n_k = n, \ \textit{soit}\ \frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_k p_k \sigma_k}$$

On en déduit

$$n_k = n \frac{p_k \sigma_k}{\sum_k p_k \sigma_k}$$

Il reste à reporter les n_k optimaux dans l'expression de $\operatorname{var}(\bar{X}_s)$ et noter que les $\bar{X}_k$ sont indépendants

$$\operatorname{var}(\bar{X}_{so}) = \sum_k p_k^2 \operatorname{var}(\bar{X}_k) = \sum_k p_k \sigma_k \left(\frac{p_k \sigma_k}{n_k} \right) = \sum_k \sigma_k p_k \sum_\ell \frac{p_\ell \sigma_\ell}{n} = \frac{1}{n} \left(\sum_k \sigma_k p_k \right)^2$$

Remarque 1: on peut aussi remplacer n_K par $n - \sum_{k=1}^{K-1}$ dans l'expression de la variance, dériver par rapport à n_k, $k = 1, \dots, K-1$ et obtenir le même résultat.
Remarque 2: le résultat obtenu nécessite de connaître les σ_k ou à défaut de les estimer.

4. La méthode d'allocation proportionnelle utilise la même fraction d'échantillonnage dans chaque strate, ie $n_1/N_1 = \ldots = n_K/N_K$. Calculer $\bar{X}_{sp}$ la moyenne stratifiée par cette méthode et sa variance.
Soit $\rho = n_1/N_1 = \ldots = n_K/N_K$. On a $\rho N = n$, d'où $n_k = \frac{n}{N} N_k = n p_k$. $\bar{X}_{sp}$ est la moyenne non pondérée

$$\bar{X}_{sp} = \sum_k p_k \bar{X}_k = \sum_k \frac{n_k}{n} \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik} = \frac{1}{n} \sum_k \sum_i X_{ik}$$

La variance s'exprime comme suit

$$\operatorname{var}(\bar{X}_{sp}) = \sum_k p_k^2 \operatorname{var}(\bar{X}_k) = \sum_k p_k^2 \frac{\sigma_k^2}{n_k} = \frac{1}{n} \sum_k p_k \sigma_k^2$$

5. Montrer que

$$\operatorname{var}(\bar{X}_{sp}) - \operatorname{var}(\bar{X}_{so}) = \frac{1}{n} \sum_{k=1}^K p_k (\sigma_k - \bar{\sigma})^2 \text{ où } \bar{\sigma} = \sum_{k=1}^K p_k \sigma_k$$

et

$$\text{var}(\bar{X}) - \text{var}(\bar{X}_{sp}) = \frac{1}{n}\sum_{k=1}^{K} p_k(\mu_k - \mu)^2$$

Commenter.

Comparaison des variances des allocations proportionnelle et optimale:

$$\begin{aligned}
\frac{1}{n}\sum_k p_k(\sigma_k - \bar{\sigma})^2 &= \frac{1}{n}\sum_k p_k\sigma_k^2 - \frac{2}{n}\underbrace{\sum_k p_k\sigma_k}_{=\bar{\sigma}} + \frac{1}{n}\underbrace{\sum_k p_k}_{=1}\bar{\sigma}^2 \\
&= \frac{1}{n}\left(\sum_k p_k\sigma_k^2 - \bar{\sigma}^2\right) \\
&= \text{var}(\bar{X}_{sp}) - \text{var}(\bar{X}_{so}) \geq 0
\end{aligned}$$

Si les variances des strates sont les mêmes, l'allocation proportionnelle a la même variance que l'allocation optimale: plus les variances sont hétérogènes, mieux il faut utiliser l'allocation optimale. Par ailleurs, $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$, *d'où*

$$\begin{aligned}
\text{var}(\bar{X}) - \text{var}(\bar{X}_{sp}) &= \frac{1}{n}\sum_k p_k\sigma_k^2 + \frac{1}{n}\sum_k p_k(\mu_k - \mu)^2 - \frac{1}{n}\sum_k \sigma_k^2 \\
&= \frac{1}{n}\sum_k p_k(\mu_k - \mu)^2 \geq 0
\end{aligned}$$

L'allocation proportionnelle donne une variance inférieure à celle de $\bar{X}$ *et est d'autant meilleure si les moyennes des strates sont différentes.*