Conservatoire National des Arts et Métiers
Pôle Sciences et Techniques de l'Information et de la Communication, Spécialité Mathématiques

Enquêtes et sondages

UE STA 108

MANUEL D'EXERCICES

Sylvie Rousseau

Année scolaire 2020– 2021

Table des matières

Quelques rappels

Variable aléatoire *X*

C'est une grandeur qui peut prendre différentes valeurs avec différentes probabilités. Elle est définie sur l'ensemble des résultats possibles (ou événements) d'une expérience aléatoire (ex : résultat d'un jeu de hasard, durée d'attente,…).

Loi de probabilité

La loi de probabilité, ou distribution, d'une variable aléatoire *X* est définie par l'ensemble des valeurs prises par *X* ainsi que par :

- la probabilité de chaque valeur possible de *X* quand *X* est une v.a. discrète,
- la probabilité que *X* se réalise dans un intervalle donné quand *X* est une v.a. continue. La fonction de densité de *X*, dérivée de la fonction de répartition caractérise la loi de probabilité.

Espérance *E(X)*

C'est la valeur que l'on peut espérer obtenir, en moyenne, en réalisant une v.a. *X*. On l'assimile à la moyenne de *X* par abus de langage.

Pour une variable aléatoire discrète, $E(X) = \sum_k k \times P(X = k)$.

Pour une variable aléatoire continue admettant une densité *f(x)*, $E(X) = \int_{-\infty}^{+\infty} x f(x)$

Propriétés :

- Pour *c* constante réelle, $E(c) = c$
- $E(X+Y) = E(X) + E(Y)$: on dit que l'espérance est un opérateur linéaire
- Si *X* et *Y* sont indépendantes alors $E(XY) = E(X) \times E(Y)$

Variance *Var(X)*

C'est une mesure de la variabilité des valeurs par rapport à la moyenne. Plus les valeurs de X sont « imprévisibles », plus elle est grande. Elle se définit par $Var(X) = \sigma_X^2 = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$ (« moyenne des carrés des écarts à la moyenne »)

Propriétés :

- La variance est toujours positive ou nulle
- $Var(X) = 0 \Leftrightarrow$ *X* constante
- $Var(cX) = c^2 Var(X)$ où *c* est une constante réelle
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
  - $Cov(X,Y) = \sigma_{XY} = E[X - E(X)] \times E[Y - E(Y)]$
  - $Cov(X,Y) = 0$ si *X* et *Y* sont indépendantes

Loi de Bernoulli *B(p)*

C'est la loi de la variable *X* qui indique si le résultat d'une épreuve est un échec ou un succès (par exemple : jouer à pile ou face)

Loi de probabilité : $P(X=1) = p$ et $P(X=0) = 1-p$

Espérance : $E(X) = p$

Variance : $Var(X) = p(1-p)$

Loi binomiale *B(n,p)*

C'est la loi de la variable X qui compte le nombre de boules blanches obtenues à l'issue de *n* tirages, indépendants et avec remise, dans une urne de taille *N* contenant *p* % de boules blanches.

Loi de probabilité : $P(X=k)=C_n^k p^k(1-p)^{n-k}$ avec $k\in\{0,1,...,n\}$

Espérance : $E(X)=np$

Variance : $Var(X)=np(1-p)$

Remarque. : une loi binomiale de paramètres n et p est aussi la somme de n lois de Bernoulli indépendantes et de même paramètre p.

Loi hypergéométrique *H(N, n,p)*

C'est la loi de la variable *X* qui compte le nombre de boules blanches sélectionnées à l'issue de *n* tirages sans remise dans une urne de taille *N* contenant des boules blanches en proportion *p*.

Loi de probabilité : $P(X=k)=\frac{C_{Np}^k C_{N-Np}^{n-k}}{C_N^n}$ avec $\max(0,n-(N-Np))\leq k\leq\min(n,Np)$

Espérance : $E(X)=np$

Variance : $Var(X)=np(1-p)\frac{N-n}{N-1}$

Convergence de la loi hypergéométrique vers la loi binomiale

Si N tend vers l'infini, la loi H(N,n,p) tend vers la loi B(n, p), c'est-à-dire que lorsqu'on effectue un tirage dans une grande population, il importe peu que ce tirage se fasse avec ou sans remise (en pratique, on considèrera que la population est « grande » lorsque l'échantillon représente moins de 10% de cette population : n /N < 0,1).

Loi normale ou loi de Laplace-Gauss *N(m,σ²)*

C'est la loi d'une variable *X* continue, variant de -∞ à +∞ , dont la densité de probabilité vaut :

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$$

Espérance : $E(X)=m$

Variance : $Var(X)=\sigma^2$

Convergence de la loi binomiale vers la loi normale

Si X suit une B(n,p) et que n tend vers l'infini alors $\frac{X-np}{\sqrt{np(1-p)}}\longrightarrow N(0,1)$

En pratique, on considère que l'approximation est correcte dès que $n\,p(1-p)>18$, *d'autant plus que* n *est grand et* p *proche de* $0{,}5$.

Loi uniforme *U(0,1)*

Une variable *X* suit une loi uniforme *U(0,1)* si sa densité de probabilité vaut : $f(x)=1_{]0,1[}(x)$

Espérance : $E(X)=1/2$

Variance : $Var(X)=1/12$

$F(x)=P(X\leq x)=x$ sur $[0,1]$

Loi faible des grands nombres

Si $(X_1,X_2,\ldots,X_n)$ sont des variables indépendantes et identiquement distribuées (i.i.d.) selon une loi quelconque de même moyenne *m*, alors: $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow[n\to\infty]{p} m$

Autrement dit, la moyenne d'une variable sur un échantillon aléatoire simple tend vers la moyenne dans la population, quand la taille de l'échantillon tend vers l'infini. *Par exemple, si l'on pouvait jouer indéfiniment à "pile ou face" avec une pièce bien équilibrée, le pourcentage de "pile" obtenu tendrait vers 50 %.*

Théorème central limite

Si $(X_1,X_2,\ldots,X_n)$ sont des variables i.i.d. selon une loi quelconque de moyenne *m* et de variance σ^2, alors: $\sqrt{n}\frac{\overline{X}_n-m}{\sigma} \xrightarrow[n\to\infty]{Loi} N(0,1)$

Intervalle de confiance

Plutôt que d'estimer ponctuellement la vraie valeur inconnue du paramètre θ, on recherche un intervalle recouvrant «très vraisemblablement » cette vraie valeur.

Définition : On appelle intervalle de confiance de niveau de confiance $1-\alpha$ du paramètre θ tout intervalle *IC* tel que : $P(IC \ni \theta)=1-\alpha$ pour $\alpha \in [0,1]$ fixé.

Les bornes de l'intervalle de confiance *IC* dépendent de l'échantillon, elles sont donc aléatoires.

I. Sondage aléatoire simple

Exercice 1 — Un petit exemple

L'exercice propose de retrouver sur un exemple les résultats de la théorie pour un sondage aléatoire simple sans remise de taille fixe. On considère pour cela tous les échantillons possibles de taille 2 pris dans une population de taille *N* = 5. On connaît par ailleurs les valeurs de la variable d'intérêt *Y* pour chaque unité de la population, à savoir respectivement : 8, 3, 11, 4 et 7.

1. Calculer la moyenne $\overline{Y}$ et la dispersion S_Y^2 du caractère d'intérêt sur la population.
2. Lister tous les échantillons possibles de taille 2.
3. Pour chacun de ces échantillons, calculer l'estimateur $\hat{\overline{Y}}$ de la moyenne de la variable d'intérêt.
4. Vérifier que $\hat{\overline{Y}}$ estime sans biais la vraie moyenne.
5. Calculer la variance de cet estimateur $V\left(\hat{\overline{Y}}\right)$.
6. Vérifier que cette variance $V\left(\hat{\overline{Y}}\right)$ coïncide avec la formule donnée par la théorie.
7. Calculer l'estimateur de variance $\hat{V}\left(\hat{\overline{Y}}\right)$ pour chacun des échantillons possibles.
8. Vérifier que $\hat{V}\left(\hat{\overline{Y}}\right)$ estime sans biais la vraie variance $V\left(\hat{\overline{Y}}\right)$.

Exercice 2 — Rappels de cours

L'exercice propose de démontrer des résultats présentés dans le cours et d'insister sur des techniques de raisonnement usuelles en sondage. Considérons qu'on veuille estimer le total et la moyenne d'une grandeur *Y* dans une population *U* de taille *N*. Pour cela, on procède à un sondage aléatoire simple sans remise de taille *n* et on note *S* l'échantillon aléatoire obtenu.

1. Combien y a-t-il d'échantillons possibles ? Quelle est la probabilité de tirer chacun d'entre eux ?

2. On considère un individu *k* quelconque dans *U*. Combien y a-t-il d'échantillons contenant cet individu ? En déduire la probabilité de tirage de *k*.

3. On note I_k la variable aléatoire valant 1 si *k* appartient à l'échantillon et 0 sinon.
   a. Que vaut $E(I_k)$?
   b. Comment peut-on réécrire $\sum_{k\in S} Y_k$ à partir des I_k ?

4. En déduire que :
   a. $\hat{t}_y = \frac{N}{n}\sum_{k\in S} Y_k$ estime sans biais le vrai total $t_y = \sum_{k\in U} Y_k$
   b. et que $\hat{\overline{Y}} = \frac{1}{n}\sum_{k\in S} Y_k$ estime sans biais la vraie moyenne $\overline{Y} = \frac{1}{N}\sum_{k\in U} Y_k$.

5. Combien y a-t-il d'échantillons comprenant les individus identifiés *k* et *l* ? En déduire la probabilité de tirer ces deux individus conjointement. Que vaut alors $E(I_k I_l)$? En déduire $Cov(I_k, I_l)$.

6. On note $S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(Y_k - \bar{Y}\right)^2$ et $f = \frac{n}{N}$. Montrer que :
   a. $Var\left(\hat{t}_y\right) = N(N-n)\frac{S_y^2}{n}$
   b. $Var\left(\hat{\bar{Y}}\right) = (1-f)\frac{S_y^2}{n}$
7. Quel est l'intérêt du sondage sans remise par rapport au sondage avec remise ?
8. Montrer que $s^2 = \frac{1}{n-1}\sum_{k\in S}\left(Y_k - \hat{\bar{Y}}\right)^2$ estime sans biais S_y^2.
9. En déduire des estimateurs sans biais de $Var\left(\hat{t}_y\right)$ et de $Var\left(\hat{\bar{Y}}\right)$.

Exercice 3 — Estimation de la surface agricole utile d'un canton

(d'après P.Ardilly et Y.Tillé, Exercices corrigés de méthode de sondage, Ellipses, 2003)

On veut estimer la surface moyenne cultivée dans les fermes d'un canton rural. Sur 2010 fermes que comprend ce canton, on en tire 100 par sondage aléatoire simple. On mesure Y_k la surface cultivée par la ferme *k* en hectares et on trouve :

$$\sum_{k\in S} Y_k = 2907\ ha \text{ et } \sum_{k\in S} Y_k^2 = 154\,593\ ha^2$$

1. Donner la valeur de l'estimateur sans biais classique de la moyenne $\bar{Y} = \frac{1}{N}\sum_{k\in U} Y_k$.
2. Donner un intervalle de confiance à 95% pour $\bar{Y}$.

Exercice 4 — Estimation d'une retombée touristique

(d'après A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

145 ménages de touristes séjournant en France dans une région donnée ont dépensé 830 € en moyenne par jour. L'écart type estimé de leurs dépenses s'élève à 210 €. Sachant que 50 000 ménages de touristes ont visité la région où a été effectuée l'enquête, que peut-on dire de la dépense totale journalière de l'ensemble de ces ménages ? On supposera pour cela que l'échantillon est issu d'un plan aléatoire simple à probabilités égales.

Exercice 5 — Taille d'échantillon pour un sondage d'opinion

(d'après A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

Un sondage sur la popularité d'une personnalité politique lui accorde un pourcentage $\hat{p} = 30\%$ 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 regard 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 d'opinions favorables dans la population ne s'écarte pas de $\hat{p}$ de plus de deux points ?

Exercice 6 — Taille d'échantillon pour une proportion

(d'après P.Ardilly et Y.Tillé, Exercices corrigés de méthode de sondage, Ellipses, 2003)

On s'intéresse à l'estimation de la proportion *P* d'individus atteints par une maladie professionnelle dans une entreprise de 1500 salariés. On sait par ailleurs que trois personnes sur dix sont ordinairement touchées par cette maladie dans des entreprises du même type. On se propose de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

1. Quelle taille d'échantillon faut-il sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance 0,95 soit inférieure à 0,02 pour un plan simple :
   a. avec remise ?
   b. sans remise ?

2. Que faire dans le cas du plan sans remise si on ne connaît pas la proportion d'individus habituellement touchés par la maladie ?

Exercice 7 — Nombre d'espaces de stationnement à prévoir

(d'après A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

Une entreprise de promotion immobilière désire estimer le nombre d'espaces de stationnement requis pour une nouvelle tour devant abriter des bureaux. Elle décide de procéder à un sondage aléatoire simple sans remise. Elle sait que le nouveau bâtiment abritera 5 000 personnes et que, dans des entreprises de même type que celles devant emménager dans les futurs locaux, la proportion de personnes se rendant à leur bureau en utilisant les moyens de transport en commun est toujours supérieure à 75%. Quelle doit être la taille de l'échantillon pris au sein des futurs occupants des bureaux pour pourvoir estimer le nombre d'espaces de stationnement à prévoir avec une marge d'erreur symétrique d'au plus 150 places au niveau de confiance 90% ?

Exercice 8 — Application au marketing direct

(d'après A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

Les sondages sont très largement utilisés dans le marketing direct : il arrive souvent que l'on estime par sondage le rendement d'un fichier donné, ou que l'on souhaite comparer les rendements de plusieurs fichiers, ou encore, que disposant de plusieurs fichiers, on souhaite estimer par sondage le rendement global de l'ensemble de ces fichiers. Dans cet exercice, on suppose l'existence d'un fichier de N = 200 000 adresses. On note p le rendement inconnu du fichier à une offre d'abonnement à prix réduit avec calculette offerte en prime ; c'est donc la proportion d'individus qui s'abonneraient si l'offre était offerte à tous les individus du fichier. Selon l'usage $\hat{p}$ est l'estimation de p obtenue à partir d'un test fait sur un échantillon de n adresses choisies à probabilités égales et sans remise sur le fichier.

1. On sait par expérience que les rendements à ce type d'offre sur ce fichier ne dépassent pas généralement 3%. Quelle taille d'échantillon doit-on prendre pour estimer p avec une précision absolue de 0,5 point et un degré de confiance de 95% ?

2. Mêmes questions pour une précision de 0,3 point et 0,1 point.

3. Le test a porté sur 10 000 adresses et on a noté 230 abonnements. En déduire l'intervalle de confiance bilatéral à 95% pour le rendement p ainsi que le pour le nombre total d'abonnements si la même offre était faite sur l'ensemble du fichier.

*Rappel : on appelle **précision absolue** au niveau de confiance 1-α- la quantité* $t_{1-\frac{\alpha}{2}}\sqrt{V(\hat{p})}$ *où* $t_{1-\frac{\alpha}{2}}$ *est le fractile d'ordre* $1-\frac{\alpha}{2}$ *de la loi normale centrée réduite.*

Exercice 9 — Un cas d'enquête répétée

(d'après P.Ardilly et Y.Tillé, Exercices corrigés de méthode de sondage, Ellipses, 2003)

On considère une population de 10 stations-services et on s'intéresse au prix du litre de supercarburant que chacune d'entre elles affiche. Plus exactement, sur deux mois consécutifs, mai et juin, les données de prix figurent dans le tableau ci-dessous :

Prix du litre de supercarburant

Station	1	2	3	4	5	6	7	8	9	10
Mai	5,82	5,33	5,76	5,98	6,20	5,89	5,68	5,55	5,69	5,81
Juin	5,89	5,34	5,92	6,05	6,20	6,00	5,79	5,63	5,78	5,84

On veut estimer l'évolution du prix moyen du litre entre mai et juin. On choisit, comme indicateur de cette évolution la différence des prix moyens On propose deux méthodes concurrentes:

- ***Méthode 1*** : on échantillonne n stations ($n < 10$) en mai et n stations en juin, les deux échantillons étant totalement indépendants ;
- ***Méthode 2*** : on échantillonne n stations en mai, et on interroge de nouveau ces stations en juin (technique de *panel*).

1. Comparer l'efficacité des deux méthodes.
2. Même question si on souhaite cette fois estimer un prix moyen sur la période globale mai-juin.
3. Si on s'intéresse au prix moyen de la question 2, ne vaut-il pas mieux tirer, non pas 2 fois n relevés avec la méthode 1 (n chaque mois) mais directement $2n$ relevés sans se soucier des mois (méthode 3) ? Aucun calcul n'est nécessaire.

Exercice 10 — Echantillonnages successifs

En cours de collecte, la taille d'un échantillon s'avère parfois insuffisante pour assurer la précision attendue. Une solution naturelle est d'enquêter un échantillon complémentaire. Intéressons-nous au plan de sondage final obtenu après :

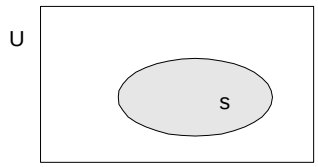
- Un premier échantillonnage simple sans remise de n_1 unités parmi N à probabilités égales,
- Suivi d'un second tirage simple sans remise de n_2 unités parmi $N-n_1$ à probabilités égales

La sélection des $n = n_1 + n_2$ unités ainsi retenues obéit-elle à un plan simple sans remise et à probabilités égales dans la population de taille N?

Rappels sur le sondage aléatoire simple

I/ Définition

Tirage d'un échantillon de *n* unités sans remise et à probabilités égales dans une population finie composée de *N* unités identifiables.



II/ Notations

1. ***Dans la population (ou univers)*** $U = \{1,2,...,k,...,N\}$
   - Variable d'intérêt : Y de caractéristique individuelle Y_k
   - Total : $T_Y = \sum_{k\in U} Y_k$
   - Moyenne: $\bar{Y} = \frac{T_Y}{N} = \frac{1}{N}\sum_{k\in U} Y_k$
   - Variance : $\sigma_y^2 = \frac{1}{N}\sum_{k\in U}\left(Y_k - \bar{Y}\right)^2$
   - Dispersion (variance modifiée) : $S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(Y_k - \bar{Y}\right)^2 = \frac{N}{N-1}\sigma_y^2$

2. ***Dans l'échantillon*** *s* : sous-ensemble de *U* de taille *n(s)*
   - Ensemble des échantillons possibles : $\mathcal{S}$
   - Plan de sondage probabiliste : loi de probabilité sur $\mathcal{S}$

$$p(s) \geq 0, \forall s \in S, \text{ et } \sum_{s\in S} p(s) = 1.$$

   - Moyenne : $\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} Y_k$
   - Dispersion empirique : $\hat{s}_y^2 = \frac{1}{n-1}\sum_{k\in S}\left(Y_k - \hat{\bar{y}}\right)^2$
   - Probabilité d'inclusion d'ordre un de *k* : $\pi_k = P(k \in s) = \sum_{s\in S / k\in s} p(s)$
   - Probabilité d'inclusion ou double de *k* et *l* : $\pi_{kl} = P(k \in s, l \in s) = \sum_{s\in S / k,l\in s} p(s)$
   - $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$

III/ Formulaire du sondage aléatoire simple

Probabilité de sélectionner l'échantillon s : $p(s) = 1/C_N^n$

Probabilité de sélectionner l'individu k : $\forall k \in U,\ \pi_k = P(k \in s) = \frac{n}{N} = f$ (taux de sondage)

Paramètre d'intérêt / Statistique	**Moyenne**	**Proportion** $p = N_0/N$	**Total**
Estimateur du paramètre d'intérêt	$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} Y_k = \hat{\bar{y}}(s)$	$\hat{p} = \frac{1}{n}\sum_{k\in s} y_k = \frac{n_0}{n}$	$\hat{t}y = N \times \hat{\bar{y}} = \frac{N}{n}\sum_{k\in S} Y_k$
Vraie variance d'échantillonnage de cet estimateur	$Var\left(\hat{\bar{y}}\right) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$	$Var(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{N}{N-1}\frac{p(1-p)}{n}$	$Var\left(\hat{t}_y\right) = N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$
Estimateur de la variance d'échantillonnage	$\hat{Var}\left(\hat{\bar{y}}\right) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$	$\hat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$	$\hat{Var}\left(\hat{t}_y\right) = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right) = \left[\hat{\bar{y}} - 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \ \hat{\bar{y}} + 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que *n* est grand $\frac{\hat{\bar{y}} - \bar{Y}}{\sqrt{Var(\hat{\bar{y}})}} \to N(0, 1)$

PLANS A PROBABILITES INEGALES

Exercice 1 — Rappels de cours sur l'estimateur d'Horvitz-Thompson

On considère une population U et on s'intéresse à l'estimation du total d'une variable d'intérêt Y noté $t_y = \sum_{k \in U} Y_k$. Pour cela, on prélève un échantillon *s* avec des probabilités individuelles de sélection notées $(\pi_k)_{k \in U}$.

1. Rappeler l'expression de l'estimateur d'Horvitz-Thompson (ou « π-estimateur » ou encore « estimateur des valeurs dilatées »).
2. Etudier son espérance et sa variance.

Exercice 2 — Application directe du cours

On considère une population $U = \{1,2,3\}$,sur laquelle on définit le plan de sondage suivant :

$$p(\{1,2\}) = \frac{1}{2},\ p(\{1,3\}) = \frac{1}{4},\ p(\{2,3\}) = \frac{1}{4}$$

Y est une variable définie sur *U*, telle que : $Y_1 = Y_2 = 3, Y_3 = 6$ dont on veut estimer le total t_y .

1. Calculer les probabilités d'inclusion simple π_k et double $\pi_{k\ell}$.
2. Donner la distribution de probabilité de l'estimateur de Horvitz-Thompson $\hat{t}_{Y\pi}$ du total. Calculer la variance de cet estimateur.
3. Donner la distribution de probabilité d'un estimateur de variance de $\hat{t}_{Y\pi}$ (il est conseillé de choisir l'estimateur le plus simple à calculer). On pourra vérifier que cet estimateur est sans biais.

Exercice 3 — Volume d'archives

On désire estimer à l'échelle d'un canton le nombre de kilomètres linéaires d'archives stockées dans les mairies. Pour cela, on procède à un tirage de 4 communes parmi les 9 du canton, proportionnellement à leur population.

1. Calculer les probabilités d'inclusion de chaque communes, à partir des données suivantes :

N°de commune	Nom de la commune	Population
1	Val le Grand	1100
2	Les Gries	650
3	Les Combres	500
4	Flins	2300
5	Villers le Lac	4000
6	Fortin	5500
7	Montlebon	1900
8	Sanzeau	200
9	Aumont	150

2. Estimer le métrage total des archives du canton à partir des résultats suivants :

N° de commune	Nom de la commune	Mètres d'archives
2	Les Gries	17
4	Flins	38
5	Villers le Lac	55
6	Fortin	70

Exercice 4 — Tirage systématique d'entreprises

On veut sélectionner un échantillon de taille 4 dans une population de 8 entreprises dont on connaît la taille, mesurée en termes d'effectif salarié. L'échantillon est tiré à probabilités proportionnelles à la taille.

Entreprise	1	2	3	4	5	6	7	8
Taille	300	300	150	100	50	50	25	25

1. Donner les probabilités d'inclusion d'ordre 1 des entreprises.
2. Sélectionner l'échantillon selon un tirage systématique en utilisant 0,27 comme nombre aléatoire ;
3. Lister les échantillons possibles que l'on peut obtenir avec un tirage systématique, et indiquer les probabilités de tirage de chacun d'eux.
4. A partir des échantillons obtenus, donner une estimation du total de l'effectif salarié des entreprises. Le résultat était-il prévisible ?
5. Calculer la matrice des probabilités d'inclusion d'ordre 2 ? Commenter.

Exercice 5 — Tirage de Poisson

(d'après P.Ardilly et Y.Tillé, Exercices corrigés de méthode de sondage, Ellipses, 2003)

Lorsqu'on effectue des tirages à probabilités inégales, on utilise en général des méthodes d'échantillonnage de taille fixe. Il existe cependant des algorithmes très simples permettant des tirages à probabilités inégales mais conférant à l'échantillon une taille variable. On s'intéresse ici au tirage de Poisson dont le principe consiste à effectuer une loterie sur chaque individu de la population indépendamment d'un individu à l'autre. Ainsi, pour une population de taille N où les probabilités d'inclusion individuelles π_k sont connues pour tout *k*, on simule *N* aléas indépendants dans la loi uniforme sur [0,1] et on retient l'individu *k* si et seulement si $u_k \leq \pi_k$

1. Vérifier que l'algorithme de tirage respecte les probabilités d'inclusion d'ordre 1 en calculant la probabilité pour que l'individu k soit sélectionné.
2. La taille de l'échantillon est une variable aléatoire notée n_S .
   a. Ecrire n_S en fonction des variables indicatrices de Cornfield.
   b. Que vaut l'espérance et la variance de n_S ?
   c. Quelle est la probabilité pour que l'échantillon ait une taille au moins égale à 1 ?

   On supposera dans la suite que l'échantillon a une taille au moins égale à 1.
3. On utilise l'estimateur du total $\hat{Y} = \sum_{k \in S} \frac{Y_k}{\pi_k}$ où *S* désigne l'échantillon aléatoire obtenu à l'issue des *N* loteries.
   a. Vérifier que $\hat{Y}$ estime le vrai total sans biais.
   b. Quelle est la variance de $\hat{Y}$? Comment peut-on l'estimer sans biais ?
   c. Que valent les probabilités d'inclusion d'ordre 2 ?
4. Comparer à un plan général de taille fixe *n* de mêmes probabilités d'inclusion. Quelles sont les inconvénients d'un plan de taille non-fixe ?

Rappels sur les plans à probabilités inégales

I/ Intérêt

Retenir de préférence les unités les plus porteuses d'information afin d'accroître la précision.

III/ Formulaire

Probabilité de sélectionner l'individu k :

- Pour un plan à probabilités proportionnelles à une variable X de taille (corrélée positivement à Y)

$$\forall k \in U,\ \pi_k = P(k \in S) = n\frac{X_k}{\sum_{k \in U} X_k}$$

- Pour un plan de taille fixe, $\sum_{k \in U} \pi_k = n$

Paramètre d'intérêt / Statistique	**Moyenne**	**Total**
Estimateur d'Horvitz-Thompson du paramètre d'intérêt (π-estimateur)	Si la taille N est connue : $\hat{\mu}_{y\pi} = \frac{1}{N}\sum_{k \in s}\frac{Y_k}{\pi_k} = \frac{\hat{t}_{y\pi}}{N}$ Sinon, estimateur de Hájek : $\hat{\mu}_{yH} = \frac{1}{\hat{N}_\pi}\sum_{k \in s}\frac{Y_k}{\pi_k} = \frac{\sum_{k \in s}\frac{Y_k}{\pi_k}}{\sum_{k \in s}\frac{1}{\pi_k}} = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}$	$\hat{t}_{y\pi} = \sum_{k \in s}\frac{Y_k}{\pi_k}$ En particulier : $\hat{N}_\pi = \sum_{k \in s}\frac{1}{\pi_k}$
Vraie variance d'échantillonnage de cet estimateur	Cas général $Var(\hat{\mu}_{y\pi}) = \frac{1}{N^2}\sum_{k \in U}\sum_{l \in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$ Si la taille de l'échantillon est fixe $Var(\hat{\mu}_{y\pi}) = -\frac{1}{2N^2}\sum_{k \in U}\sum_{l \in U}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$	Cas général : $Var(\hat{t}_{y\pi}) = \sum_{k \in U}\sum_{l \in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$ Si la taille de l'échantillon est fixe $Var(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k \in U}\sum_{l \in U}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$
Estimateur de la variance d'échantillonnage	Cas général $\hat{Var}_1(\hat{\mu}_{y\pi}) = \frac{1}{N^2}\sum_{k \in s}\sum_{l \in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$ Si la taille de l'échantillon est fixe $\hat{Var}_2(\hat{\mu}_{y\pi}) = -\frac{1}{2N^2}\sum_{k \in s}\sum_{l \in s}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$	Cas général $\hat{Var}_1(\hat{t}_{y\pi}) = \sum_{k \in s}\sum_{l \in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$ Si la taille de l'échantillon est fixe $\hat{Var}_2(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k \in s}\sum_{l \in s}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$

Si *n* est grand, **l'intervalle de confiance pour la moyenne au niveau de confiance 1- α** est :

$$IC_{1-\alpha}(\mu_y) = \left[\hat{\mu}_{y\pi} - u_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\mu}_{y\pi})};\hat{\mu}_{y\pi} + u_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\mu}_{y\pi})}\right]$$

où $u_{1-\frac{\alpha}{2}}$ désigne le fractile d'ordre 1- α/2 de la loi N(0,1)

TP1 : SIMULATIONS DE TIRAGE D'ECHANTILLONS

Objectifs de la séance

- Sélectionner des échantillons en utilisant différents algorithmes de tirages ;
- Évaluer le paramètre d'intérêt et la précision de cette estimation ;
- Valider de manière empirique certaines propriétés de la théorie des sondages ;
- Comparer les méthodes d'échantillonnage.

Le logiciel employé de préférence est SAS. On utilisera alors principalement les procédures SURVEYSELECT et SURVEYMEANS.

A défaut, des macros Excel plus sommaires seront également proposées.

Données utilisées

Les données sont extraites de l'historique des populations des communes métropolitaines, disponible sur le site de l'Insee à cette adresse : http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/populations-legales/doc.asp?page=historique-populations-legales.htm

On souhaite estimer la population au RP2010 de l'ensemble de ces communes ainsi que le nombre moyen d'habitants par commune.

Partie I : Tirage d'un échantillon

Echantillonner 1000 communes en raisonnant successivement à probabilités égales puis à probabilités inégales (proportionnellement à la population recensée en 1999) au moyen des algorithmes suivants :

1) Tirage de Bernoulli ;
2) Méthode du tri aléatoire ;
3) Méthode de sélection-rejet ;
4) Tirage de Poisson ;
5) Tirage systématique ;
6) Algorithme de Sunter.

Partie II : Emploi des procédures SAS d'échantillonnage et simulations de tirage

1) Application aux plans simples sans remise :
   a. Sélectionner un échantillon de 100 communes en utilisant la PROC SURVEYSELECT ;
   b. Avec l'échantillon obtenu, estimer la population au RP2010 de l'ensemble des communes de métropole et en fournir un intervalle de confiance (on utilisera la PROC SURVEYMEANS) ;
   c. Simuler le tirage de 100 échantillons. Commenter les résultats.
2) Mêmes questions pour un plan à probabilités inégales, avec des probabilités proportionnelles à la taille des communes au RP1999.
3) Mêmes questions pour un plan stratifié avec des allocations proportionnelles et en utilisant la tranche d'unité urbaine au RP1999 comme critère de stratification.

PLANS STRATIFIES

Exercice 1 — Rappels de cours

Dans une population de taille N partitionnée en H strates, on sélectionne un échantillon de taille n suivant un plan stratifié. Dans chaque strate h, on tire n_h individus parmi N_h selon un sondage aléatoire simple sans remise de taille fixe.

Préalable : montrer la formule de décomposition de la variance :

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in U}\left(Y_k - \bar{Y}\right)^2 = \frac{1}{N}\sum_{h=1}^{H} N_h \sigma_{yh}^2 + \frac{1}{N}\sum_{h=1}^{H} N_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

1. Pour une variable d'intérêt Y, donner les estimateurs du total t_Y et de la moyenne.
2. Montrer que ces deux estimateurs sont sans biais et calculer leur variance.
3. On considère l'allocation proportionnelle de l'échantillon : on décide de tirer dans chaque strate h un nombre d'individus n_h tel que :

   $\frac{n_h}{N_h} = \frac{n}{N}$ (en supposant que $N_h \frac{n}{N}$ soit entier).

   a. Comment s'écrivent alors les estimateurs du total et de la moyenne ?
   b. Que vaut leur variance ?
   c. Montrer alors, que si on suppose : $\sigma_y^2 \approx S_y^2$ et $\sigma_{yh}^2 \approx S_{yh}^2$ pour tout h, l'allocation proportionnelle est toujours meilleure qu'un sondage aléatoire simple.

4. Le point de vue envisagé maintenant est celui d'une allocation optimale afin de satisfaire un souci de précision. Sous la contrainte que $\sum_{h=1}^{H} n_h = n$,
   a. Quelle est l'allocation des n_h qui minimise la variance de l'estimateur du total ?
   b. Que vaut alors la variance ?
   c. Comment peut-on interpréter le choix des allocations optimales ?

Exercice 2 — Estimation du poids des éléphants d'un cirque

(d'après P.Ardilly et Y.Tillé, Exercices corrigés de méthode de sondage, Ellipses, 2003)

Un directeur de cirque possède 100 éléphants classés en deux catégories : "mâles" et "femelles". Le directeur veut estimer le poids total de son troupeau, car il veut traverser un fleuve en bateau. Il a la possibilité de faire peser seulement 10 éléphants de son troupeau. Cependant, en 1998, 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
Femelles	40	4	2,25

1. Calculer la variance dans la population de la variable "poids de l'éléphant" en 1998.

2. Si, en 1998, le directeur avait procédé à un sondage aléatoire simple sans remise de 10 éléphants, quelle aurait été la variance de l'estimateur du poids total du troupeau ?

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

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

Exercice 3 — L'âge du personnel

Une grande entreprise veut réaliser une enquête auprès de son personnel qui comprend 10 000 personnes. Des études préliminaires ont montré :

- que les variables que l'on cherche à analyser dans l'enquête sont très contrastées selon les catégories de personnel et qu'il y a donc intérêt à stratifier selon ces catégories. Pour simplifier, on considérera qu'il y a 3 grandes catégories qui formeront les strates,

- que ces variables sont également très fortement liées à l'âge des individus.

On va donc proposer des plans d'échantillonnage comme si on voulait étudier l'âge des individus : si une stratégie est meilleure que d'autres pour estimer l'âge moyen, alors on a de bonnes raisons de penser qu'elle le sera aussi pour les variables d'intérêt. Comme on connaît l'âge des membres du personnel, on peut raisonner en faisant les comparaisons exactes.

On dispose des renseignements suivants :

Catégorie de personnel	Poids dans l'ensemble du personnel	Ecart type des âges
1	20%	18,0
2	30%	12,0
3	50%	3,6
Ensemble	100%	16,0

1. Soit $\overline{Y}$ l'âge moyen et $\hat{\overline{Y}}$ l'estimateur issu d'un échantillon aléatoire simple sans remise à probabilités égales de n = 100 individus. Quelle est l'erreur type de $\hat{\overline{Y}}$?

2. On décide que l'échantillon de 100 individus doit être stratifié selon les catégories de personnel. Quelle est la répartition « représentative » ? Quelle est l'erreur type de l'estimateur de $\overline{Y}$ qui en découle ? Comparer avec les résultats de la question 1.

3. Quelle serait la répartition optimale de l'échantillon ? Quelle est l'erreur type de l'estimateur de $\overline{Y}$ qui en découle ? Comparer avec les résultats de la question 2.

Exercice 4 — Estimation d'une proportion

Sur les 7500 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 : individus de faibles revenus (strate 1), de revenus moyens (strate 2) et de revenus élevés (strate 3).

On note :

- N_h la taille de la strate *h,*
- n_h la taille de l'échantillon dans la strate *h,*
- $\hat{p}_h$ l'estimateur de la proportion d'individus possédant au moins un véhicule dans la strate *h.*

On obtient le résultat suivant :

	h = 1	*h* = 2	*h* = 3
Nh	3500	2000	2000
nh	500	300	200
$\hat{p}_h$	0,13	0,45	0,50

1. Quel estimateur $\hat{p}$ de *p* proposez-vous ? Que peut-on dire de son biais ?
2. Calculez la précision de $\hat{p}$, et donnez un intervalle de confiance à 95% pour *p*.
3. Estimez-vous que le critère de stratification est adéquat ?

Exercice 5 — Choix des allocations

(d'après A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

Cet exercice est une application du principe : "à chaque objectif son échantillon". Une entreprise comporte 400 exécutants et 100 cadres. La direction de l'entreprise désire évaluer un indice de satisfaction, assimilable à une variable numérique positive *Y*, mesurable pour chaque individu à partir d'un ensemble de questions : elle décide pour cela de faire réaliser une enquête auprès de 100 personnes employées dans l'entreprise, à l'aide d'un plan de sondage stratifié, avec un sondage aléatoire simple dans chaque strate. Le coût d'une interview est le même dans les deux strates.

On pense *a priori* que la dispersion de la variable *Y* doit être la même au sein de chacun des deux groupes. Comment répartir l'échantillon entre les deux groupes, selon que l'on vise l'un des objectifs suivants :

a. obtenir la meilleure précision possible sur la valeur moyenne de l'indice de satisfaction dans l'entreprise ;

b. obtenir la même précision sur la valeur moyenne de l'indice de satisfaction dans chacune des deux catégories ;

c. obtenir la meilleure précision possible sur la différence entre les valeurs moyennes de l'indice de satisfaction dans les deux catégories.

Exercice 6 — 0ptimalité pour une différence

(d'après J-M. Grosbras, Méthodes statistiques des sondages, Economica, 1987)

Le but de l'exercice est de montrer que si une stratégie est optimale pour estimer précisément une quantité dans l'ensemble d'une population stratifiée, elle peut ne plus l'être tout à fait si l'objectif du sondage est justement de comparer les strates entre elles. La bonne définition des objectifs à atteindre est donc essentielle au choix de la technique à employer. Considérons une population de taille *N* formée de deux strates, de taille N_1 et N_2 et intéressons-nous à la moyenne $\overline{X}$ d'une variable *X*. Les moyennes de *X* dans les strates 1 et 2 sont notées $\overline{X}_1$ et $\overline{X}_2$ et leurs estimateurs $\hat{\overline{X}}_1$ et $\hat{\overline{X}}_2$.

On dispose d'un budget *C* et on suppose que :

- le tirage effectué est un sondage aléatoire simple sans remise de n_h unités parmi N_h dans la strate *h* (*h =1* ou *2*),
- la fonction de coût s'écrit $C_1n_1 + C_2n_2$ où C_h désigne le coût unitaire dans la strate *h.*

1. Si on cherche à estimer précisément la moyenne $\overline{X}$,
   a. Donner l'expression de $\hat{\overline{X}}$, estimateur sans biais de $\overline{X}$ en fonction de $\hat{\overline{X}}_1$ et $\hat{\overline{X}}_2$.
   b. Calculer sa variance.
   c. Quelle répartition *(n_1, n_2)* de l'échantillon donne une variance $V\left(\hat{\overline{X}}\right)$ minimale ? Que vaut alors $V\left(\hat{\overline{X}}\right)$?
   d. Application numérique : calculer n_1, n_2, n et $V\left(\hat{\overline{X}}\right)$ avec :

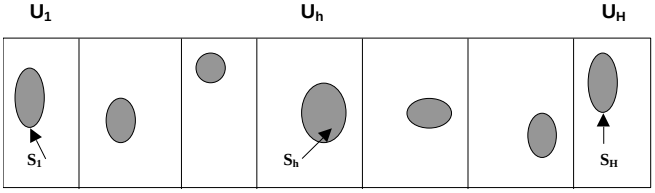
      $N_1 = 10\,000$ $N_2 = 20\,000$
      $S_1 = 2$ $S_2 = 1$
      $C_1 = 4$ $C_2 = 9$ $C = 1\,000$

2. Si on avait appliqué une allocation proportionnelle, c'est-à-dire : $n_h / N_h = n / N$,
   a. Qu'aurait-on trouvé pour n_1, n_2 *et* n *?*
   b. Que vaudrait alors $V\left(\hat{\overline{X}}\right)$?
   c. Avec les mêmes données numériques, évaluer la perte relative de précision par rapport à l'échantillon optimal.
3. En fait, on cherche à évaluer l'écart entre les moyennes des deux groupes : $\overline{X}_1 - \overline{X}_2$.
   a. Montrer que $\hat{\overline{X}}_1 - \hat{\overline{X}}_2$ est un estimateur sans biais de $\overline{X}_1 - \overline{X}_2$.
   b. Calculer sa variance.
   c. Déterminer la répartition *(n_1, n_2)* de l'échantillon pour que $V\left(\hat{\overline{X}}_1 - \hat{\overline{X}}_2\right)$ soit minimale, toujours avec la même contrainte de budget. (on pourra éventuellement utiliser, en les adaptant, certains résultats de la question 1).
   d. Calculer dans ces conditions $V\left(\hat{\overline{X}}\right)$. Comparer ce résultat avec celui de la 1ère question en écrivant la différence Δ des variances de ces deux estimateurs.
   e. Reprendre l'application numérique pour trouver les nouvelles valeurs de n_1, n_2, n, $V\left(\hat{\overline{X}}\right)$ et la perte relative de précision par rapport à l'échantillon optimal.

Rappels sur les plans stratifiés

I/ Définition

Partition de la population en sous-groupes appelés strates selon un critère lié au paramètre d'intérêt puis tirage d'autant d'échantillons indépendants qu'il y a de strates.



Constituer des strates homogènes en intra au regard de la variable d'intérêt permet de gagner en précision.

II/ Notations

1. Dans la population

- $U=\bigcup_{h=1}^{H}U_h$ et $N=\sum_{h=1}^{H}N_h$
- Total : $t_y=\sum_{h=1}^{H}t_{yh}=\sum_{h=1}^{H}N_h\bar{y}_h$
- Moyenne: $\bar{y}=\frac{t_y}{N}=\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h$ avec $\bar{y}_h=\frac{1}{N_h}\sum_{k\in U_h}y_k$
- Variance : $\sigma_y^2=\frac{1}{N}\sum_{k\in U}(y_k-\bar{y})^2=\sum_{h=1}^{H}\frac{N_h}{N}\sigma_{yh}^2+\sum_{h=1}^{H}\frac{N_h}{N}(\bar{y}_h-\bar{y})^2=\sigma_{y\text{intra}}^2+\sigma_{y\text{inter}}^2=\frac{N-1}{N}S_y^2$

  avec $\sigma_{yh}^2=\frac{1}{N_h}\sum_{k\in U_h}(y_k-\bar{y}_h)^2$

2. Dans l'échantillon

- $S=\bigcup_{h=1}^{H}S_h$ et $n=\sum_{h=1}^{H}n_h$
- Moyenne dans S_h : $\hat{\bar{y}}_h=\frac{1}{n_h}\sum_{k\in S_h}y_k$
- Dispersion dans S_h : $\hat{S}_{yh}^2=\frac{1}{n_h-1}\sum_{k\in S_h}(y_k-\hat{\bar{y}}_h)^2$

III/ Formulaire du sondage stratifié

Paramètre d'intérêt / Statistique	Moyenne	Proportion	Total
Estimateur du paramètre d'intérêt	$\hat{\bar{y}}=\sum_{h=1}^{H}\frac{N_h}{N}\hat{\bar{y}}_h$	$\hat{p}=\sum_{h=1}^{H}\frac{N_h}{N}\hat{p}_h$	$\hat{t}_y=N\hat{\bar{y}}=\sum_{h=1}^{H}\hat{t}_{yh}=\sum_{h=1}^{H}N_h\hat{\bar{y}}_h$
Vraie variance d'échantillonnage de cet estimateur	$Var\left[\hat{\bar{y}}\right]=Var\left[\sum_{h=1}^{H}\frac{N_h}{N}\hat{\bar{y}}_h\right]=\sum_{h=1}^{H}Var\left[\frac{N_h}{N}\hat{\bar{y}}_h\right]$ Si plan simple dans chaque strate : $Var\left[\hat{\bar{y}}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{yh}^2}{n_h}$	Si plan simple dans chaque strate $Var\left[\hat{p}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{N_h}{N_h-1}\frac{p_h(1-p_h)}{n_h}$	$Var\left[\hat{t}y\right]=Var\left[N\hat{\bar{y}}\right]=N^2Var\left[\hat{\bar{y}}\right]$ Si plan simple dans chaque strate : $Var\left[\hat{t}y\right]=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{yh}^2}{n_h}$
Estimateur de la variance d'échantillonnage	Si plan simple dans chaque strate $\hat{Var}\left[\hat{\bar{y}}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_{yh}^2}{n_h}$	Si plan simple dans chaque strate $\hat{Var}\left[\hat{p}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$	Si plan simple dans chaque strate $\hat{Var}\left[\hat{t}_y\right]=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_{yh}^2}{n_h}$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right)=\left[\hat{\bar{y}}-1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \hat{\bar{y}}+1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que *n* est grand $\frac{\hat{\bar{y}}-\bar{Y}}{\sqrt{Var(\hat{\bar{y}})}}\rightarrow N(0,1)$

Choix des allocations

- Allocations proportionnelles : $\frac{n_h}{n}=\frac{N_h}{N}\quad\forall h\in\{1,...H\}$
- Allocations optimales de Neyman (sans contrainte de budget) **:** $n_h=n\frac{N_hS_{yh}}{\sum_{l=1}^{H}N_lS_{yl}}$
- Allocations optimales sous contrainte budgétaires **:** $n_h=C\frac{N_hS_{yh}}{\sqrt{C_h}\sum_{l=1}^{H}N_lS_{yl}\sqrt{C_l}}$

PLANS PAR GRAPPES

Exercice 1 — Problématique d'un plan par grappes

L'objet de cet exercice est de rappeler le formulaire établi en cours et de revenir sur les notions d'effet de sondage et d'effet de grappe.
Un sondage en grappes se pratique sur une population partitionnée en groupes d'individus appelés « grappes » : il consiste à sélectionner certaines grappes, selon un plan quelconque, et à retenir tous les individus des grappes désignées dans l'échantillon final. Procéder de la sorte permet de réduire les coûts d'enquête. On s'intéresse ici au cas particulier où *m* grappes sont choisies par sondage aléatoire simple sans remise parmi les *M* grappes de taille N_i d'une population de taille *N*.
On cherche à estimer le total t_y et la moyenne $\bar{y}$ sur la population d'un caractère d'intérêt *Y*.

1. Partie 1 : généralités

1.1. Quelle est la probabilité pour qu'un individu appartienne à l'échantillon ?
1.2. Que pouvez-vous dire de la taille finale de l'échantillon ? Même question si toutes les grappes sont de même taille N_0.
1.3. Quels estimateurs sans biais $\hat{t}_y$ et $\hat{\bar{y}}$ proposez-vous ?
1.3.1. Quelle est la précision de ces estimateurs ?
1.3.2. Montrez que dans le cas où les grappes sont de même taille alors on obtient

$$Var(\hat{\bar{y}}) = \frac{M-m}{M-1}\frac{\sigma^2_{y\,inter}}{m}.$$

1.3.3. En déduire comment constituer les grappes pour obtenir des résultats précis.
1.4. Comment estimez-vous sans biais la précision des estimateurs du total et de la moyenne ?
1.5. Dans le cas où *N* est inconnue, quel estimateur de $\bar{y}$ proposez-vous ? Cet estimateur est-il sans biais ? Approcher son espérance et son erreur quadratique moyenne.

2. Partie 2 : effet de sondage

On souhaite caractériser la précision de l'échantillonnage par grappes par rapport au sondage aléatoire simple de même taille dans le cas où les grappes sont d'effectifs égaux N_0.

2.1. Montrez que l'effet de sondage défini par $Deff = \frac{Var(\hat{\bar{y}})}{Var_{sas}(\hat{\bar{y}})}$ vaut $N_0\eta^2$ où η^2 désigne le rapport de corrélation « inter-grappes » : $\eta^2 = \frac{\sum_{i=1}^{M} N_0\left(\bar{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{M}\sum_{k=1}^{N_0}\left(Y_k - \bar{Y}\right)^2} = \frac{\sigma_y^2 inter}{\sigma_y^2}$

2.2. En déduire quand le plan par grappes est plus précis que le sondage aléatoire simple.

3. Partie 3 : effet de grappe

On définit le coefficient de corrélation « intra-grappes » par :

$$\rho = \frac{\sum_{i=1}^{M}\sum_{k=1}^{N_0}\sum_{l=1,l\neq k}^{N_0}(Y_k - \bar{Y})(Y_l - \bar{Y})}{(N_0 - 1)(N-1)S_Y^2}.$$

Ce coefficient mesure l'effet de grappe. Il se rapproche de 1 si à l'intérieur de chaque grappe, il n'y a pas de différence entre les individus ; au contraire, il est négatif si les individus sont très disparates à l'intérieur de leurs grappes.

3.1. Montrez que l'effet de grappe vaut :

$$\rho = \frac{1}{N_0 - 1}\left[N_0 \frac{\sigma^2_{y\,inter}}{\sigma^2_y} - 1 \right]$$

3.2. En déduire que $Deff = 1 + \rho(N_0 - 1)$ et que $Var(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{1}{n} S_y^2 \left[1 + \rho(N_0 - 1)\right]$.

4. ***Partie 4 subsidiaire: estimation de l'effet de sondage et de l'effet de grappe***

On cherche à estimer l'effet de sondage et l'effet de grappe et donc à estimer sans biais $Var_{sas}(\hat{\bar{y}})$ autrement dit la dispersion S_y^2. Les grappes sont de même taille.

4.1. Montrez que la dispersion empirique observée sur l'échantillon $s_y^2 = \frac{1}{n-1}\sum_{k \in S}(y_k - \hat{\bar{y}})^2$ possède un biais sous un plan complexe de taille fixe et à probabilités égales (comme ici avec des grappes de même taille) donné par :

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

4.2. En déduire que l'expression $\hat{D}eff = \frac{Var(\hat{\bar{y}})}{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}}$ est justifiée si *n* est assez grand.

Exercice 2 — Nombre de signataires d'une pétition
(Extrait de Cochran, Sampling Technics)

On a collecté des signatures pour une pétition sur 676 feuilles. Sur chacune d'entre elles, il y a la place pour 42 signatures, mais beaucoup ne sont pas très remplies. Le nombre de signatures par feuille a été étudié sur un échantillon de 50 feuilles (à peu près 7% de l'ensemble donc). A partir des résultats sont consignés dans le tableau ci–contre, estimer le nombre total de signatures et donner un intervalle de confiance pour ce nombre à 95% et à 80% .

Nombre de signatures	Fréquence
42	23
41	4
36	1
32	1
29	1
27	2
23	1
19	1
16	2
15	2
14	1
11	1
10	1
9	1
7	1
6	3
5	2
4	1
3	1

Exercice 3 — Sélection d'îlots

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

L'objectif est d'estimer le revenu moyen des ménages dans un arrondissement d'une ville composée de 60 îlots de maisons (un îlot est « un pâté de maison », de taille variable). Pour cela, on sélectionne 3 îlots par sondage aléatoire simple sans remise et on interroge tous les ménages qui y résident. On sait en outre que 5 000 ménages résident dans cet arrondissement. Le résultat est donné dans le tableau ci-dessous.

1. Estimez le revenu moyen et le revenu total des ménages de l'arrondissement par l'estimateur d'Horvitz-Thompson.

2. Estimez sans biais la variance de l'estimateur d'Horvitz-Thompson de la moyenne.

3. Estimez le revenu moyen des ménages de l'arrondissement par le ratio de Hajek, et comparez à l'estimation issue de 1. Le sens de variation était-il prévisible ?

Numéro de l'îlot	Nombre de ménages dans l'îlot	Revenu total des ménages de l'îlot
1	120	2100
2	100	2000
3	80	1500

Exercice 4 — Emprunts bancaires

Une société bancaire structurée en 3 980 succursales gère 39 800 clients, à raison de 10 clients par agence. On choisit 40 succursales par sondage aléatoire simple sans remise pour lesquelles on compte le nombre de clients ayant obtenu un prêt durant une période donnée.

On note t_{yi} le nombre obtenu dans la succursale *i* et on observe : $\sum_{i=1}^{40} t_{yi} = 185$ et $\sum_{i=1}^{40} t_{yi}^2 = 1263$.

1. Estimer le nombre total de clients de la banque qui ont obtenu un prêt durant la période de référence ainsi que leur proportion dans l'ensemble de la clientèle. On notera ces estimateurs $\hat{t}_y$ et $\hat{p}$.

2. Calculer la variance des estimateurs $\hat{t}_y$ et $\hat{p}$.

3. Estimer ces variances et fournir un intervalle de confiance approché à 95% pour chacune des quantités estimées.

4. Calculer l'effet de sondage défini comme le ratio mesurant la perte de variance estimée par rapport à un sondage aléatoire simple sans remise de même taille (indication : on commencera par estimer la dispersion S_y^2). On pourra commenter le résultat en comparant les amplitudes des intervalles de confiance à 95% obtenus pour la proportion d'intérêt entre les deux plans de sondage.

5. Calculer le coefficient de corrélation intra-grappe.

Exercice 5 — Influence de la taille et du nombre de grappes échantillonnées

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

Un statisticien souhaite réaliser une enquête sur la qualité des soins assurés dans les services de cardiologie des hôpitaux. Pour cela, il tire par sondage aléatoire simple 100 hôpitaux parmi les 1 000 hôpitaux répertoriés , puis, dans chacun des hôpitaux tirés, il recueille l'avis de tous les malades du service de cardiologie.

1. Comment se nomme ce plan de sondage et quelle est sa raison d'être ?

2. On considère que chaque service de cardiologie comprend exactement 50 lits et que l'intervalle de confiance à 95% sur la vraie proportion *P* de malades insatisfaits est : $P \in [0{,}10 \pm 0{,}018]$, (cela signifie en particulier que, dans l 'échantillon, 10 % des malades sont insatisfaits de la qualité des soins). Comment estimez-vous l'effet de grappe (commencer par estimer S_y^2, dispersion du caractère d'intérêt sur toute la population) ?

3. Le statisticien se demande comment évoluerait la précision de son enquête de satisfaction si, d'un seul coup, il échantillonnait deux fois plus d'hôpitaux mais que dans chaque hôpital tiré, il ne collectait ses données que sur la moitié du service de cardiologie (mettons que les services soient systématiquement partagés par un couloir et que notre statisticien ne s'intéresse exclusivement qu'aux 25 lits qui se situent à droite du couloir) ?

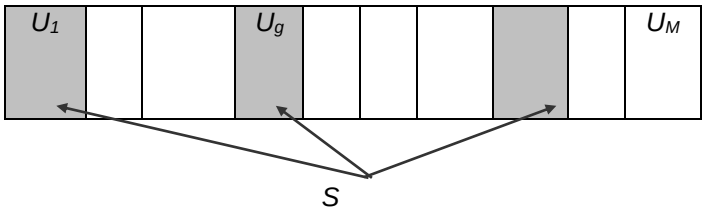
4. Commentez ce résultat par rapport à ce que donnait le premier plan de sondage.

Rappels sur les plans par grappes

I/ Définition

Objectif principal : réduire les coûts d'enquête et/ou de pallier le manque d'une base de sondage.

Principe : partition de la population en sous-groupes appelés grappes, puis tirage de grappes et enfin recensement de toutes les unités qui les composent.



Règle : constituer des grappes hétérogènes en intra au regard de la variable d'intérêt.

II/ Notations

1. Dans la population U constituée de M grappes et N individus

- $U=\bigcup_{g=1}^{M} U_g$ et $N=\sum_{g=1}^{M} N_g$
- $t_y=\sum_{g=1}^{M} t_{yg}=\sum_{g=1}^{M} N_g \bar{y}_g$
- $\bar{y}=\frac{t_y}{N}=\sum_{g=1}^{M} \frac{N_g}{N}\bar{y}_g$ avec $\bar{y}_g=\frac{1}{N_g}\sum_{k\in Ug} y_k$
- $S_G^2=\frac{1}{M-1}\sum_{g=1}^{M}\left(t_{yg}-\frac{t_y}{M}\right)^2$

2. Dans l'échantillon S constitué de m grappes et n_s individus

- $S=\bigcup_{g\in S_G} U_g$ et $n_s=\sum_{g\in S_G} N_g$

III/ Formulaire du plan par grappe dans le cas d'un plan simple de grappes

Paramètre d'intérêt / Statistique	Total	Moyenne
Estimateur du paramètre d'intérêt	$\hat{t}_y = \frac{M}{m}\sum_{g\in S_G} t_{yg}$	$\hat{\bar{y}} = \frac{1}{N}\hat{t}_y = \frac{M}{Nm}\sum_{g\in S_G} N_g \bar{y}_g$
Vraie variance d'échantillonnage de cet estimateur	$Var\left[\hat{t}y\right] = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{g=1}^{M}\left(t_{yg}-\frac{t_y}{M}\right)^2$	$Var\left[\hat{\bar{y}}\right] = \frac{1}{N^2}Var\left[\hat{t}y\right]$
Estimateur de la variance d'échantillonnage	$\hat{Var}\left[\hat{t}y\right] = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{m-1}\sum_{g\in S_G}\left(t_{yg}-\frac{\hat{t}_y}{M}\right)^2$	$\hat{Var}\left[\hat{\bar{y}}\right] = \frac{1}{N^2}\hat{Var}\left[\hat{t}y\right]$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right) = \left[\hat{\bar{y}} - 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \ \hat{\bar{y}} + 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que la taille de l'échantillon est assez grande.

PLANS A PLUSIEURS DEGRES

Exercice 1 — Probabilités d'inclusion et plans de sondage

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

On considère une population $U = \{1,2,3,4,5,6,7,8,9\}$, sur laquelle on définit le plan de sondage suivant :

$$p(\{1,2\}) = \frac{1}{6},\ p(\{1,3\}) = \frac{1}{6},\ p(\{2,3\}) = \frac{1}{6}$$
$$p(\{4,5\}) = \frac{1}{12},\ p(\{4,6\}) = \frac{1}{12},\ p(\{5,6\}) = \frac{1}{12}$$
$$p(\{7,8\}) = \frac{1}{12},\ p(\{7,9\}) = \frac{1}{12},\ p(\{8,9\}) = \frac{1}{12}$$

1. Calculer les probabilités d'inclusion simple π_k.

2. Ce plan de sondage est-il simple, stratifié, en grappes, à deux degrés, ou aucun de ces plans particuliers?

Exercice 2 — Rappels de cours

Considérons une population de taille *N* répartie en *M* unités primaires elles-mêmes quadrillées en N_i unités secondaires. Le premier degré de tirage consiste à extraire un échantillon d'unités primaires parmi lesquelles, dans un second degré de tirage, sont sélectionnées des unités secondaires. Les individus des unités secondaires désignées composent l'échantillon final. Par exemple, si les UP quadrillent le territoire selon un découpage en communes, elles-mêmes composées d'US définies à partir des îlots (ou « pâtés de maisons »), alors l'enquête sera limitée géographiquement aux communes et îlots sélectionnés.

Dans la suite, on considérera le cas où les UP sont choisies selon un sondage aléatoire simple sans remise de taille *m* et où les US sont tirées dans les UP retenues au 1er degré selon un plan simple sans remise de taille n_i parmi N_i. On s'intéresse au total t_y d'un caractère d'intérêt *Y*.

1. Quelle est l'expression de $\hat{t}_y$ estimateur sans biais de t_y ?

2. Donner l'expression de la variance de $\hat{t}_y$ et interpréter les différents termes de ce calcul.

3. Comment estime-t-on cette variance ?

4. Que pouvez-vous dire de la taille finale de l'échantillon ?

Exercice 3 — Estimation d'un effectif

Un camion transporte des vis sur 500 palettes, chacune d'elles contenant 40 boîtes de vis. L'industriel réceptionnant ces palettes souhaite estimer le nombre moyen de vis par boîte. Pour cela, il tire un échantillon de 100 palettes, selon un sondage aléatoire simple sans remise, puis il tire dans chacune de ces 100 palettes un échantillon de 5 boîtes, selon un sondage aléatoire simple sans remise également, et enfin il compte le nombre de vis dans les boîtes ainsi tirées.

L'industriel, et néanmoins statisticien, calcule pour chaque palette *i* de son échantillon le nombre moyen de vis par boîte, et la dispersion du nombre de vis par boîte (ces deux quantités sont calculées à partir des 5 boîtes échantillonnées dans la palette).

Il calcule ensuite les moyennes, sur les 100 palettes, de ces deux quantités :

- moyenne du nombre moyen de vis par boîte = 50
- moyenne de la dispersion du nombre de vis par boîte = 455.

Il calcule aussi la dispersion des 100 estimations du nombre de vis par palette et obtient 375 000.

1. Donner un estimateur sans biais du nombre moyen de vis par boîte.
2. Donner la précision de cet estimateur.
3. Donnez un intervalle de confiance à 95% pour le nombre moyen de vis par boîte.

Exercice 4 — Nombre de caractères par enregistrement

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

Sur un disque dur de micro-ordinateur, on compte 400 fichiers, chacun comprenant exactement 50 enregistrements. Pour estimer le nombre moyen de caractères par enregistrement, on décide de tirer par sondage aléatoire simple 80 fichiers, puis 5 enregistrements dans chaque fichier. On note *m* = 80 et $\overline{n} = 5$.

On mesure après tirage :

- la dispersion des estimateurs du nombre total de caractères par fichier, soit $s_I^2 = 905000$,
- la moyenne des *m* dispersions s_i^2 est égale à 805 où s_i^2 représente la dispersion du nombre de caractères par enregistrement dans le fichier *i.*

1. Comment estimez-vous le nombre moyen $\overline{Y}$ de caractères par enregistrement ?
2. Comment estimez-vous sans biais la précision de l'estimateur précédent ?
3. Donnez un intervalle de confiance à 95% pour $\overline{Y}$.

Exercice 5 — Etude d'impact préalable au lancement d'un produit financier

En vue de préparer le lancement d'un nouveau produit financier, une société bancaire ayant un réseau de *M* succursales souhaite mener une étude approfondie auprès de particuliers possesseurs de comptes chez elle . Les variables d'intérêt de l'enquête ont trait aux caractéristiques de la clientèle et à ses motivations éventuelles. On cherche à estimer la proportion *p* de personnes potentiellement intéressées par la nouvelle offre. L'enquête opère selon un plan à 2 degrés : dans un premier temps, on choisit *m* succursales pour participer à l'opération parmi lesquelles, au second temps, on désigne des échantillons de titulaires de comptes à interroger. Le plan de sondage est le suivant :

- Au premier degré, on réalise un sondage aléatoire simple sans remise de *m = 10* succursales parmi *M = 100*. Le taux de sondage f_1 vaut 0,10. La société bancaire gère *N = 100 000* titulaires de compte.
- Au second degré de tirage, le taux de sondage f_2 est uniforme à 10%.

1. Donner un estimateur sans biais de *p* qu'on notera $\hat{p}$.

2. Montrer que $V(\hat{p}) \cong \left(\frac{M}{N}\right)^2 (1 - f_1) \frac{S_T^2}{m} + \frac{1 - f_2}{N f_1 f_2} \sum_{i=1}^{M} \frac{N_i}{N} p_i (1 - p_i)$

3. Montrer que $\hat{V}(\hat{p}) \cong \left(\frac{M}{N}\right)^2 (1 - f_1) \frac{\hat{s}_T^2}{m} + \frac{1 - f_2}{N f_1 f_2} \sum_{i \in S_1} \frac{N_i}{N} \hat{p}_i (1 - \hat{p}_i)$

4. Application numérique : donner un intervalle de confiance à 95% pour *p* avec les résultats d'enquête suivants : $\sum_{k \in S} y_k = 102$, $s_T^2 = 1200$, $\sum_{i \in sI} \frac{N_i}{N} \hat{p}i(1 - \hat{p}i) = 0{,}01$

Exercice 6 — Choix entre méthodes concurrentes

Une population de 1010 saucisses est partitionnée en deux unités primaires, de tailles respectives 1000 et 10. Pour estimer le nombre moyen de bouts de saucisses dans cette population, on emploie le plan de sondage suivant :

- on sélectionne une UP selon un sondage aléatoire simple,
- on sélectionne deux saucisses dans l'UP tirée selon un sondage aléatoire simple sans remise.

La première UP est sélectionnée. On observe que chacune des deux saucisses tirées dans l'UP possède deux bouts.

Le statisticien A calcule le nombre moyen de bouts sur son échantillon de deux saucisses et trouve 2. Il affirme que cette valeur est une estimation sans biais du nombre moyen de bouts dans la population.

Le statisticien B propose comme estimation sans biais de ce nombre moyen de bouts la valeur :

$$\frac{1000}{1010} \times 4 = 3.96$$

Discuter les deux méthodes d'estimation, en précisant les logiques qui les sous-tendent.

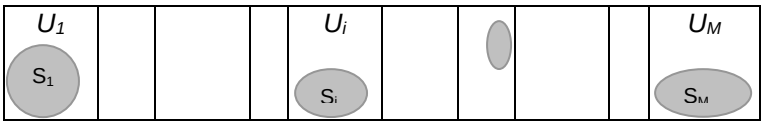
Rappels sur les plans à deux degrés

I/ Définition

Objectif principal : réduire les coûts d'enquête et/ou de pallier le manque d'une base de sondage.

Principe : dans une population partitionnée en sous-groupes appelés unités primaires, eux-mêmes composés d'unités secondaires :

- au 1er degré, tirage d'unités primaires
- au 2nd degré, tirage d'unités secondaires dans les unités primaires retenues au 1er degré (indépendamment d'une unité primaire à l'autre)



Règle : constituer des unités primaires hétérogènes en intra au regard de la variable d'intérêt.

II/ Notations

1. *Dans la population U constituée de M unités primaires et N individus*

- $U=\bigcup_{i=1}^{M} U_i$ et $N=\sum_{i=1}^{M} N_i$
- $t_y=\sum_{i=1}^{M} t_{yi}=\sum_{i=1}^{M} N_i \bar{y}_i$
- $\bar{y}=\frac{t_y}{N}=\sum_{i=1}^{M} \frac{N_i}{N}\bar{y}_i$ avec $\bar{y}_i=\frac{1}{N_i}\sum_{k\in Ui} y_k$
- $S_I^2=\frac{1}{M-1}\sum_{i=1}^{M}\left(t_{yi}-\frac{t_y}{M}\right)^2$ et $S_i^2=\frac{1}{N_i-1}\sum_{k=1}^{N_i}\left(y_k-\bar{y}_i\right)^2$

2. *Dans l'échantillon S constitué de m unités primaires et n_s individus*

- $S=\bigcup_{i\in SUP} S_i$ et $n_s=\sum_{i\in Si} n_i$
- $\hat{s}_I^2=\frac{1}{m-1}\sum_{i\in SUP}\left(\hat{t}_{yi}-\frac{\hat{t}_y}{M}\right)^2$ et $\hat{s}_i^2=\frac{1}{n_i-1}\sum_{k\in Si}^{i}\left(y_k-\hat{\bar{y}}_i\right)^2$

III/ Formulaire du plan à deux degrés dans le cas d'un plan simple des unités primaires et des unités secondaires

Paramètre d'intérêt / Statistique	Total	Moyenne
Estimateur du paramètre d'intérêt	$\hat{t}_y=\sum_{k\in S}\frac{y_k}{\pi_k}=\frac{M}{m}\sum_{i\in S_{UP}}\hat{t}_{yg}=\frac{M}{m}\sum_{i\in S_{UP}}\frac{N_i}{n_i}\sum_{k\in Si}y_k$	$\hat{\bar{y}}=\frac{1}{N}\hat{t}_y=\frac{M}{Nm}\sum_{g\in S_G}N_g\hat{\bar{y}}_g$
Vraie variance d'échantillonnage de cet estimateur	$Var\left[\hat{t}y\right]=M^2\left(1-\frac{m}{M}\right)\frac{S_t^2}{m}+\frac{M}{m}\sum_{i=1}^{M}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{S_i^2}{n_i}$	$Var\left[\hat{\bar{y}}\right]=\frac{1}{N^2}Var\left[\hat{t}y\right]$
Estimateur de la variance d'échantillonnage	$\hat{Var}\left[\hat{t}y\right]=M^2\left(1-\frac{m}{M}\right)\frac{\hat{s}_t^2}{m}+\frac{M}{m}\sum_{i\in S_{UP}}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\hat{s}_i^2}{n_i}$	$\hat{Var}\left[\hat{\bar{y}}\right]=\frac{1}{N^2}\hat{Var}\left[\hat{t}y\right]$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right)=\left[\hat{\bar{y}}-1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \hat{\bar{y}}+1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que la taille de l'échantillon est assez grande

TP2 : CORRECTION DE LA NON-REPONSE

Exercice

Le but de l' étude de cas est de corriger la non-réponse (totale et partielle) pour une enquête conduite auprès de 2 389 personnes interrogées sur leur perception de leur état de santé.

L'échantillon a été choisi par sondage aléatoire simple sans remise dans une population de 2 millions d'individus. La base de données est fournie au format SAS et contient les informations suivantes :

- l'identifiant de l'enquêté (variable « ident ») ;
- son poids de sondage initial (« poids ») ;
- son âge (« âge ») ;
- son sexe (« sexe ») ;
- son niveau de revenu (« revenu ») ;
- sa région d'habitat (« region ») ;
- son nombre de consultations chez un professionnel de la santé en un an (« visites ») ;
- sa consommation de tabac (« tabac ») ;
- sa perception de son état de santé (« sante ») ;
- une indicatrice de la non-réponse totale (« nrt ») ;
- une indicatrice de la non-réponse partielle (« nrp »).

Les modalités des caractéristiques qualitatives sont définies de la sorte :

Variable	Modalités
AGE	3-4 : Junior 5-6 : Jeune adulte 7-8 : Adulte 9-11: Senior
SEXE	1 : Homme 2 : Femme
REVENU	1-4 : Bas revenus 5-6 : Moyens revenus 7-8 : Bons revenus 9-11: Hauts revenus
TABAC	Tabac 1 : Fume quotidiennement 2 : Ayant fumé quotidiennement 3 : Fume occasionnellement 4 : Ayant fumé occasionnellement 5 : Jamais fumé
SANTE	1-2 : Excellente 3 : Bonne 4-5 : Passable

1. Dresser un état des lieux sur l'importance des non-réponses.
2. Corriger la non-réponse totale. On commencera par décrire le comportement de non-réponse totale en fonction des caractéristiques disponibles pour tous les individus.
3. Corriger la non-réponse partielle pour la variable d'intérêt en envisageant diverses méthodes (imputation par la moyenne, imputation par la moyenne par classe, imputation par déduction, imputation par hot-deck, imputation par hot-deck par classe, etc.).

REDRESSEMENTS

Exercice 1 — Post-stratification

Un institut de sondage est chargé de mesurer l'audience d'un nouveau magazine. Il interroge pour cela un échantillon de taille *n* selon un procédé que l'on assimilera à un plan simple à probabilités égales et sans remise au sein de la population française des individus âgés de 15 ans et plus. On supposera de plus qu'il n'y a pas de non-réponse. Pour satisfaire à la demande de l'éditeur, les résultats sont ventilés selon le critère « habitant en zone urbaine » ou « habitant en zone rurale ». Les données recueillies se présentent ainsi :

	Habitant en zone rurale	Habitant en zone urbaine	Total
Lecteurs	64	476	540
Non lecteurs	576	884	1 460
Total	640	1 360	2 000

1. Estimez la proportion du lectorat du magazine dans l'ensemble de la population et proposez un intervalle de confiance à 95% de ce taux de lecture.
2. Sachant que la proportion réelle d'habitants en zone urbaine vaut 75%, proposez un nouvel estimateur de la proportion de lecteurs et donnez–en un intervalle de confiance à 95%. Quel gain de précision obtient-on ?

Exercice 2 — Chiffre d'affaires et effectif salarié

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

Dans une population de 10 000 entreprises, on veut estimer le chiffre d'affaires moyen $\overline{Y}$. Pour cela, on échantillonne n=100 entreprises par sondage aléatoire simple. On dispose par ailleurs de l'information auxiliaire « nombre de salariés » notée *x* par entreprise. Les données issues du sondage sont :

- $\overline{X}=50$ salariés (vraie moyenne sur les x_k),
- $\hat{\overline{y}}=5.2\times10^6$ euros (chiffre d'affaires moyen dans l'échantillon),
- $\hat{\overline{x}}=45$ salariés (effectif moyen dans l'échantillon),
- $s_y^2=25\times10^{10}$ (dispersion corrigée des y_k calculée dans l'échantillon),
- $s_x^2=15$ (dispersion corrigée des x_k calculée dans l'échantillon),
- $\hat{\rho}=0.8$ (coefficient de corrélation linéaire entre *x* et *y* calculé dans l'échantillon).

1. Que vaut l'estimateur par le ratio ? Cet estimateur est-il biaisé ?
2. Rappelez la formule de variance « vraie » de cet estimateur.
3. Calculez une estimation de la variance vraie. L'estimateur de variance utilisé est-il biaisé ?
4. Donnez un intervalle de confiance à 95% pour $\overline{Y}$.

Exercice 3 — Estimation d'une surface cultivée

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

On considère une région agricole comprenant *N = 2010* fermes où on cherche à estimer la moyenne de la surface cultivée en céréales (variable *Y* mesurée en hectares). On possède l'information auxiliaire sur la surface agricole totale cultivée de chaque ferme. En particulier, on sait qu'il y a 1 580 fermes de moins de 160 hectares (post-strate 1) et 430 fermes d'au moins 160 hectares (post-strate 2). On réalise un sondage aléatoire simple de *n = 100* exploitations et on obtient (avec les indices 1 et 2 pour les deux post-strates définies) : $n_1=70$ $n_2=30$ $\hat{\overline{y}}_1=19{,}40$ $\hat{\overline{y}}_2=51{,}63$ $s_{y1}^2=312$ $s_{y2}^2=922$.

1. a. Quel est l'estimateur post-stratifié $\hat{\bar{Y}}_{post}$? Est-il différent de la moyenne simple?
   b. Quelle est la loi de n_1 ? Que valent son espérance et sa variance?
   c. Calculer l'estimateur sans biais de la variance de $\hat{\bar{Y}}_{post}$ et donner un intervalle de confiance à 95% pour la surface moyenne cultivée en céréales.

2. On exploite désormais la variable auxiliaire *X* mesurant la surface agricole totale cultivée pour construire un estimateur par le ratio. On connaît la moyenne $\bar{X}=118{,}32$ ha et on obtient sur l'échantillon : $\hat{\bar{x}}=132{,}25$ $s_x^2=9173$ $s_y^2=708$ $\hat{\rho}=0{,}57$ où $\hat{\rho}$ est l'estimateur du vrai coefficient de corrélation linéaire inconnu ρ.
   a. Rappeler l'expression de ρ.
   b. Comment définissez-vous $\hat{\rho}$? S'agit-il d'une estimation sans biais de ρ ?
   c. Montrez que l'estimateur par le ratio de $\bar{Y}$ apparaît préférable à la moyenne simple si et seulement si $\hat{\rho}>\frac{1}{2}\frac{\hat{CV}(x)}{\hat{CV}(y)}$ où les $\hat{CV}$ estiment les coefficients de variation. Qu'obtient-on dans le cas présent ?
   d. Calculez l'estimateur par le ratio $\hat{\bar{y}}_q$ de $\bar{Y}$.
   e. Estimez sa précision et donnez un intervalle de confiance à 95% pour $\bar{Y}$.

Exercice 4 — Taille des pieds

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

Le directeur d'une entreprise de confection de chaussures veut estimer la longueur moyenne des pieds droits des hommes adultes d'une ville. Soient *y* le caractère « longueur du pied droit » (en centimètre) et *x* la taille de l'individu (en centimètres). Le directeur sait en outre par les résultats d'un recensement que la taille moyenne des hommes adultes de cette ville est de 168 cm. Pour estimer la longueur des pieds, le directeur effectue un sondage aléatoire simple sans remise de 100 hommes adultes. Les résultats sont les suivants : $s_y=2, s_x=10, s_{xy}=15, \hat{\bar{x}}=169, \hat{\bar{y}}=24$ .Sachant que 400 000 hommes adultes vivent dans cette ville,

1. Calculez l'estimateur d'Horvitz-Thompson, l'estimateur par le quotient, l'estimateur par différence et l'estimateur par la régression.
2. Estimez les variances de ces 4 estimateurs
3. Quel estimateur conseilleriez-vous au directeur ?
4. Exprimez la différence littérale entre la variance de l'estimateur par le quotient et la variance de l'estimateur par la régression en fonction de $\hat{\bar{x}}$, $\hat{\bar{y}}$ et de la pente $\hat{b}$ de la droite de régression de *y* sur *x* dans l'échantillon. Commentez.

Exercice 5 — Comparaison d'estimateurs

(Ardilly P, Tillé Y, 2003, Exercices corrigés de méthodes de sondage, Ellipses)

On se propose d'estimer la moyenne $\bar{Y}$ d'un caractère d'intérêt au moyen d'un échantillon sélectionné selon un plan aléatoire simple sans remise de taille 1 000 dans une population de taille 1 000 000. On connaît la moyenne $\bar{X}=15$ d'un caractère auxiliaire *x* et on donne, avec les notations usuelles,

$$\hat{\bar{y}}=10 \; ; \; \hat{\bar{x}}=14 \; ; \; s_x^2=25 \; ; \; s_y^2=20 \text{ et } s_{xy}=15$$

1. Estimez $\bar{Y}$ au moyen des estimateurs d'Horvitz-Thompson, par différence, par le quotient et par la régression. Estimez les variances de ces estimateurs.
2. Quel estimateur choisiriez-vous pour estimer $\bar{Y}$?

Rappels sur les redressements

I/ Intérêt

Accroître la précision en tirant parti d'information auxiliaire liée au caractère d'intérêt.

Selon la nature de l'information auxiliaire, techniques de post-stratification, d'estimation par différence, par le ratio, par la régression, par calage généralisé.

II/ Formulaire pour le total

En notant *X* l'information auxiliaire,

Méthode	Estimateur du total	Vraie erreur quadratique moyenne de cet estimateur
Estimateur d'Horvitz-Thompson	$\hat{t}_{y_\pi}=\sum_{k\in S}\frac{y_k}{\pi_k}=\frac{N}{n}\sum_{k\in S}y_k$	$Var\left(\hat{t}_{y_\pi}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$
Estimateur par la différence	$\hat{t}_{yD}=\hat{t}_{y_\pi}+t_x-\hat{t}_{x\pi}$	$Var\left(\hat{t}_{y_D}\right)=N^2\left(1-\frac{n}{N}\right)\frac{\left(S_y^2+S_x^2-2S_{xy}\right)}{n}$
Estimateur par le ratio (ou le quotient)	$\hat{t}_{yQ}=\hat{t}_{y_\pi}\frac{t_x}{\hat{t}_{x\pi}}$	$Var\left(\hat{t}_{yQ}\right)=N^2\left(1-\frac{n}{N}\right)\frac{\left(S_y^2+R^2S_x^2-2RS_{xy}\right)}{n}$ avec $R=\frac{t_y}{t_x}=\frac{\bar{Y}}{\bar{X}}$
Estimateur par la régression	$\hat{t}_{yD}=\hat{t}_{y_\pi}+\hat{b}\left(t_x-\hat{t}_{x\pi}\right)$ avec $\hat{b}=\frac{\hat{s}_{xy}}{\hat{s}_x^2}$	$Var\left(\hat{t}_{yQ}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}\left(1-\rho^2\right)$ avec $\rho=\frac{s_{xy}}{s_x s_y}$
Estimateur post-stratifié	$\hat{t}_{y\,post}=\sum_{h=1}^{H}N_h\hat{\bar{y}}_h=\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}y_k$	$Var\left(\hat{t}_{y\,post}\right)=\frac{N-n}{n}\sum_{h=1}^{H}N_hS_{yh}^2+\frac{N-n}{N-1}\frac{N^2}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}S_{yh}^2$

Estimateur par substitution de l'erreur quadratique moyenne

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right)=\left[\hat{\bar{y}}-1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \ \hat{\bar{y}}+1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que la taille de l'échantillon est assez grande.

TP3 : CALAGE SUR MARGES

Cette séance utilise la macro SAS %CALMAR développée par l'Insee. Elle est disponible sur le site insee.fr, accompagnée de sa documentation. Les données sont fournies au format SAS :

Exercice 1

Un institut spécialisé a réalisé une enquête auprès des salariés d'une entreprise, qui compte 230 salariés répartis sur deux établissements A (70 salariés) et B (160 salariés). L'institut a effectué un sondage aléatoire simple dans chaque établissement, de taux de sondage respectifs 1/10 (A) et 1/20 (B). Le but est d'estimer la proportion de salariés prêts à monétariser une partie de leurs congés.

Pour chaque salarié enquêté, on dispose de :

- son identifiant (variable ID), à 3 caractères : le premier indique l'établissement, les deux suivants constituent un numéro d'ordre dans l'établissement ;
- la variable SERVICE indiquant si le salarié travaille dans un service productif (1) ou administratif (2) ;
- la variable CATEG qui indique la catégorie de personnel à laquelle appartient le salarié : employés (1), ouvriers (2), autres (9) ;
- la variable SEXE ;
- la variable SALAIRE annuel brut ;
- la variable Y indiquant si l'employé est intéressé par le paiement de jours de son compte-épargne temps (oui = 1, non = 0).

Par ailleurs, la direction de l'entreprise a aimablement fourni les informations suivantes sur ses salariés : l'entreprise compte 80 employés, 90 ouvriers, 140 hommes, 100 personnes travaillent dans le secteur productif, et le salaire total annuel vaut 47 000.

On vous demande d'estimer le paramètre d'intérêt en utilisant cette information auxiliaire via différents calages :

- par la méthode linéaire ;
- par la méthode raking ratio :
- par la méthode logit LO=0.5 UP=2.2.

Exercice 2

Vous disposez d'une base de sondage de 11 600 individus décrits par la région, l'âge, le niveau scolaire, la catégorie socio-professionnelle, etc. (cf. tableau ci-dessous).

Le but de l'exercice est d'en sélectionner un échantillon, puis de procéder à des estimations et des redressements, en faisant comme si l'information d'intérêt avait été collectée sur l'échantillon seulement. Les variables d'intérêt mesurent l'importance consacrée aux activités sportives et culturelles.

1 / Donner la répartition de la population par tranche d'âge et niveau scolaire.

2 / Sélectionner un échantillon de taille 1 160 selon un sondage aléatoire simple.

Pour rappel, la syntaxe de la procédure SURVEYSELECT de SAS est la suivante :

```
PROC SURVEYSELECT DATA = nom de la base de sondage lue en entrée
                  STATS
                  METHOD   = SRS pour un sondage aléatoire simple sans remise
                  SEED     = germe
                  SAMPSIZE = taille de l'échantillon souhaitée
                  OUT       = nom de la table de sortie (l'échantillon);
RUN;
```

3 / A partir de l'échantillon, estimer la répartition de la population par tranche d'âge et niveau scolaire. Évaluer également le nombre moyen d'heure par semaine consacrées à la lecture, au sport, passées devant la télévision ainsi que le nombre moyen d'expositions visitées en une année et le nombre moyen de séances de cinéma en un an.

Pour rappel, la syntaxe de la procédure SURVEYMEANS de SAS est la suivante :

```
PROC SURVEYMEANS DATA = nom de la table-échantillon
                 N = Effectif de la population
                 MEAN STDERR CLM CV  = Statistiques éditées en sortie;
VAR listes de variable d'intérêt;
WEIGHT variable de pondération;
RUN;
```

4 / Caler l'échantillon sur la vraie structure par tranche d'âge et niveau scolaire. Pourquoi ces variables de calage sont-elles pertinentes ?

5/ Ré-estimer les grandeurs citées à la question 3.

Ci-dessous le contenu de la base de données :

Nom	Type	Libellé et modalités
IDENTIND	C	Identifiant
TRAGE	C	Tranche d'âge 1 : de 15 à 25 ans 2 : de 25 à 29 ans 3 : de 30 à 39 ans 4 : de 40 à 49 ans 5 : de 50 à 64 ans 6 : de 65 à 69 ans 7 : plus de 70 ans
NIVSCO2	C	Niveau scolaire 1 : inférieur au baccalauréat 2 : supérieur au baccalauréat
CS	C	Catégorie socio-professionnelle 1 : agriculteurs 2 : artisans, commerçants, chefs d'entreprises, professions libérales 3 : cadres 4 : professions intermédiaires 5 : employés 6 : ouvriers 7 : retraités
REGION	C	Région
ZEAT	C	Zone d'emploi et d'aménagement du territoire 1 : Région parisienne 2 : Bassin parisien 3 : Nord 4 : Est 5 : Ouest 6 : Sud-ouest 7 : Centre-est 8 : Méditerranée
CINEMA	N	Nombre de fois où l'individu est allé au cinéma au cours des 12 derniers mois
EXPO	N	Nombre d'expositions visitées au cours des 12 derniers mois
SPORT	N	Nombre d'heures de sport pratiquées au cours de la dernière semaine
LECTURE	N	Nombre d'heures de lecture au cours de la dernière semaine
TELE	N	Nombre d'heures passées devant la télévision au cours de la dernière semaine

REVISIONS

Exercice 1 — Algorithme du tri aléatoire

On veut estimer le poids moyen de 10 éléphants d'un cirque. Pour cela, on réalise un sondage aléatoire simple sans remise de taille 5 à l'aide d'un tri aléatoire. On simule donc une variable aléatoire uniforme *U ~ U[0,1]* sur la population des éléphants, puis on trie les réalisations obtenues par ordre croissant (ou décroissant) et on retient l'échantillon correspondant aux 5 plus grandes valeurs (ou plus petites). La simulation a été effectuée à partir de la fonction *ALEA()* sous Excel et a donné les réalisations ci-dessous :

N° de l'éléphant	Valeur générée
1	0,84
2	0,12
3	0,36
4	0,60
5	0,68
6	0,11
7	0,87
8	0,44
9	0,21
10	0,77

1. Quel est l'échantillon tiré ?
2. On pèse les éléphants retenus et on obtient en tonnes les poids respectifs suivants : 3,65 ; 3,17 ; 4,18 ; 3,55 et 4,26.
3. Donnez un estimateur du poids moyen des éléphants puis un intervalle de confiance à 95% de ce poids moyen.
4. Finalement, on réalise une pesée exhaustive des éléphants. On obtient un poids moyen de 3,45 tonnes. Que dire de l'intervalle de confiance précédent ? D'où peut venir le problème ?

Exercice 2 — Cas pratique dans une pisciculture

Un éleveur de poissons souhaite connaître le poids moyen de ses poissons. Il dispose de 3 bassins selon l'âge des animaux : n°1 pour ceux « de petite taille », n°2 « de taille moyenne » et n°3 « de grande taille ». Le nombre total de poissons par bassin est respectivement de 1000, 900 et 950.

Notre pisciculteur appelle un statisticien à sa rescousse pour estimer le poids moyen des poissons. Armé de son épuisette, le statisticien attrape 20 poissons dans le bassin n°1, 15 dans le n°2 et 10 dans le n°3. Ensuite, il calcule le poids moyen sur les 3 échantillons relatifs aux 3 bassins. Il trouve : 0.152 Kilo pour le bassin N°1, 0.255 Kilo pour le n°2 et 0.305 Kilo pour le n°3. Il calcule également la dispersion corrigée des poids des poissons sur les 3 échantillons et trouve respectivement: $(0.05)^2$ $Kilo^2$, $(0.02)^2$ $Kilo^2$ et $(0.01)^2$ $Kilo^2$ pour les bassins N°1, 2 et 3.

On admettra que le mode de tirage des échantillons de poissons dans chacun des trois bassins est assimilable à un sondage aléatoire simple de taille fixe.

1)

a) Proposer un estimateur sans biais du poids moyen des poissons relativement à un bassin.

b) Donner les 3 estimations des poids moyens relatifs aux 3 bassins puis les 3 intervalles de confiance à 95% correspondants.

c) Pour estimer le poids moyen relatif à l'ensemble des 3 bassins, le statisticien a mis en œuvre l'estimateur stratifié. Après avoir rappelé la forme générale de cet estimateur et précisé les strates adoptées par le statisticien, donner l'estimation recherchée et l'intervalle de confiance à 95% correspondant.

2)

a) Est-ce que l'allocation définie par le statisticien correspond à l'allocation proportionnelle?
b) Compte tenu des mesures effectuées sur les échantillons, expliquer (qualitativement) pourquoi l'allocation du statisticien semble être légitime.
c) A partir des résultats obtenus sur les trois échantillons, calculer l'allocation de Neyman pour une taille totale de l'échantillon de poissons de 45.

3) Le pisciculteur propose d'estimer le poids moyen des poissons sur l'ensemble des 3 bassins en faisant la moyenne arithmétique des poids des poissons sur l'ensemble des 3 échantillons.

a) Calculer l'estimation fournie par le pisciculteur.
b) Montrer que cet estimateur est en réalité biaisé (on exprimera ce biais théorique en fonction des vrais poids moyens des poissons relatifs aux bassins, des vrais effectifs de poissons et des tailles des échantillons de poissons relatifs aux bassins).
c) Donner une estimation de ce biais.

4) Le statisticien apprend par hasard, en discutant avec l'un des employés, qu'un contrôle de la taille des poissons a été réalisé récemment. Ce contrôle a été effectué dans chacun des bassins et de façon quasi-exhaustive. Il révèle que la taille moyenne des poissons par bassin est de : 25 cm pour le bassin n°1, 40 cm pour n°2 et 50 cm pour le n°3.

a) Expliquer pourquoi la connaissance de cette nouvelle information est intéressante par rapport au phénomène étudié.
b) A partir de cette nouvelle information, proposer un nouvel estimateur du poids moyen des poissons pour un bassin fixé . Donner les 3 nouvelles estimations du poids moyen relatives à chacun des bassins. On donne pour cela les tailles moyennes des poissons mesurées sur les échantillons : 23 cm (bassin n°1), 42 cm (n°2), 51 cm (n°3).
c) Proposer une nouvelle estimation du poids moyen pour l'ensemble des 3 bassins.