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

"Enquêtes et sondages"

Unité 18323 Statistique B8

TRAVAUX DIRIGES

Sylvie Rousseau

Année scolaire 2004– 2005

SOMMAIRE

Le volume global des travaux dirigés est de 30 heures réparties à raison de 15 séances de 2 heures chacune.

Elles se tiendront normalement le lundi de 18h30 à 20h30 du 11 Octobre 2004 au 07 Février 2005.

RAPPELS DE PROBABILITES
ET DE STATISTIQUE INFERENTIELLE

Exercice 1 — Lecture d'abaques pour la loi normale

Soit une variable aléatoire *Z* distribuée selon la loi normale centrée réduite, notée *N(0,1)*. Utiliser les tables statistiques pour obtenir :

1. *P(Z > 1),*
2. *P(-1,645 < Z < 1,645),*
3. *P(-1,96 < Z < 1,96),*
4. *P(-3,09 < Z < 3,09),*
5. Les quantiles $z_{0,05}$ et $z_{0,95}$ d'ordre respectifs 5% et 95%.

Exercice 2 — Contrôle qualité en usine n°1

Une usine fabrique des canettes de diamètre intérieur moyen de 50 mm avec un écart-type 0,8 mm. Le cahier des charges alloue des tolérances inférieure de 48 mm et supérieure à 52 mm. Dans le cas où ces tolérances ne sont pas respectées, la canette est déclarée non conforme.

1. En admettant que les diamètres sont distribués selon une loi normale, quelle est la proportion de canettes non conformes ?
2. On suppose que le processus de fabrication s'est déréglé et produit désormais des canettes avec un diamètre d'espérance 49 mm. Quelle est dans ce cas la proportion de canettes non conformes ?

Exercice 3 — Distribution de la taille moyenne de joueurs de basket

On considère que la taille des joueurs de basket d'une ville donnée possède une distribution d'espérance 1,85 m et d'écart type 7 cm. On interroge de manière indépendante 35 joueurs choisis aléatoirement et on relève la taille de chacun.

1. Quelle est la loi suivie par la moyenne des tailles des joueurs ?
2. Calculer la probabilité pour que cette moyenne soit
   - supérieure à 1,90 m
   - inférieure à 1,82 m.

Exercice 4 — Précision d'un appareil de mesure

Préalable : soient $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées (*i.i.d*). On note la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Calculer $E(\overline{X})$ et $V(\overline{X})$.

Un appareil de mesure possède une certaine précision définie comme l'écart type σ des mesures prises sur le même objet par le même opérateur. Dans l'optique de s'affranchir au mieux de cette erreur de mesure inhérente à l'appareil, on décide de mesurer n fois le même objet et de prendre la moyenne des résultats comme mesure finale de l'objet.

2. Justifier statistiquement cette procédure.

3. On décide de prendre l'écart type de la mesure finale comme indicateur de précision du procédé. Comment évolue la précision du procédé en fonction de n ?
4. Combien de mesures faut-il prendre successivement pour que cette précision vaille 1 ?
5. Pour choisir n, on décide de minimiser la fonction de coût suivante : $f(n) = a\,n + b\,\sigma / \sqrt{n}$ où
   - a désigne le coût comptable d'une mesure,
   - b le coût de l'imprécision de la mesure finale.

   Justifier cette fonction de coût et trouver la solution optimale.
   Réaliser l'application numérique avec $\sigma = 4$, $a = 1$ € et $b = 10$ €.

Exercice 5 — Contrôle qualité en usine n°2

Le responsable qualité d'une usine contrôle 20 objets dans chaque lot de 1000 objets avant de le laisser partir vers le client. Il accepte seulement les lots pour lesquels il ne trouve aucun objet non conforme dans l'échantillon ; dans le cas contraire, le lot est trié unité par unité.

1. Quelle est la probabilité pour qu'un lot contenant une proportion $p = 0{,}05$ d'objets non conformes soit accepté ?
2. Même question pour $p = 0{,}1$.
3. Le responsable qualité proclame partout qu'il produit du « zéro défaut » parce qu'il n'accepte aucun produit non conforme. Qu'en pensez-vous ?

Exercice 6 — Analyse sensorielle

Deux brioches A et B assez semblables possèdent néanmoins des caractéristiques de fabrication distinctes. On souhaite estimer la proportion p de consommateurs capables de distinguer les deux produits. Pour cela, on choisit n personnes aléatoirement et on fait goûter à chacune 3 brioches : 2 de type A et 1 de type B, chacune devant ensuite se prononcer sur celle qui lui semble différente des autres.

1. En appelant π la proportion de bonnes réponses, exprimer π en fonction de p.
2. Donner la loi suivie par le nombre de bonnes réponses.
3. Proposer un estimateur de π et de p. Calculer leur espérance et variance respectives.
4. Si $p = 0$, autrement dit si personne n'est capable de distinguer les deux brioches, donner la valeur limite en dessous de laquelle doit se situer le nombre de bonnes réponses dans 90% des cas. On considèrera $n = 15$.
5. On a obtenu 9 bonnes réponses parmi les 15 personnes interrogées. Qu'en concluez-vous ?

Exercice 7 — Intervalles de confiance pour une moyenne et une variance

On a pesé sur pieds 10 bœufs de trois ans de la même race lors de leur arrivée à l'abattoir et on a obtenu les résultats suivants mesurés en kg : 775 ; 750 ; 755 ; 756 ; 761 ; 765 ; 770 ; 752 ; 760 et 767. On suppose que ces résultats sont issus d'une population infinie distribuée selon une loi normale de moyenne m et de variance σ^2.

1. Construire un intervalle de confiance pour m au niveau de confiance 95%.
2. Construire un intervalle de confiance pour σ^2 au niveau de confiance 95%.

Exercice 8 — Intervalle de confiance pour une proportion

On étudie une population animale dont certains membres sont albinos. On a extrait de cette population un échantillon de 40 animaux parmi lesquels on comptabilise 3 albinos.

1. Construire un intervalle de confiance pour la proportion d'albinos au niveau de confiance 95%.
2. Même question pour un échantillon de taille 400 avec 30 albinos. Commenter.

SONDAGE ALEATOIRE SIMPLE

Exercice 1 **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 $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (Y_k - \overline{Y})^2$ et $f = 1 - \frac{n}{N}$. Montrer que :
   a. $Var(\hat{t}_y) = N(N-n)\frac{S_y^2}{n}$
   b. $Var(\hat{\overline{Y}}) = (1-f)\frac{S_y^2}{n}$

7. Quel est l'intérêt du sondage sans remise par rapport au sondage avec remise ?

8. Montrer que $s^2 = \frac{1}{n-1} \sum_{k \in S} (Y_k - \hat{\overline{Y}})^2$ estime sans biais S_y^2 .

9. En déduire des estimateurs sans biais de $Var(\hat{t}_y)$ et de $Var(\hat{\overline{Y}})$.

Exercice 2 **Application directe du cours**

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 ainsi que l'estimateur de sa variance $\hat{V}\left(\hat{\overline{Y}}\right)$.
4. Calculer la variance $V\left(\hat{\overline{Y}}\right)$.
5. Vérifier que $\hat{\overline{Y}}$ estime sans biais la vraie moyenne.
6. Vérifier que $V\left(\hat{\overline{Y}}\right)$ coïncide avec la formule de la variance donnée par la théorie.
7. Vérifier que $\hat{V}\left(\hat{\overline{Y}}\right)$ estime sans biais la vraie variance $V\left(\hat{\overline{Y}}\right)$.

Exercice 3 **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 4 **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 $\overline{Y} = \frac{1}{N}\sum_{k\in U} Y_k$.
2. Donner un intervalle de confiance à 95% pour $\overline{Y}$.

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, 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 travailleurs. On sait par ailleurs que trois travailleurs sur dix sont ordinairement touchés 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'hommes 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 potentiels 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 000adresses. 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 % (ou 0,5 points) et un degré de confiance de 95%

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

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 **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 10 **Protection de l'anonymat**

Pour préserver l'anonymat dans certaines enquêtes par sondage, le procédé suivant peut être suivi. Admettons que l'on veuille estimer la proportion de personnes qui remplissent leur déclaration fiscale de manière honnête. On demande alors à chaque personne interrogée de se retirer dans une pièce isolée, et de jouer à pile ou face.

- si elle obtient « pile » alors elle doit répondre honnêtement par « oui » ou « non » à la question « Votre déclaration fiscale est-elle honnête ? »
- si elle obtient « face », elle devra lancer la pièce une nouvelle fois et répondre par « oui » ou « non » à la question « Avez-vous obtenu « face » au deuxième tirage ? ».

Grâce à ce procédé, il est impossible à l'enquêteur de savoir à quelle question se rapporte la réponse de la personne interrogée, celle-ci peut donc fournir sans crainte une réponse sincère.

1. On note p la proportion inconnue de déclarations fiscales remplies honnêtement dans la population et π la proportion de réponses « oui ». Montrer que $\pi = p/2 + 1/4$.

2. Soit X la variable aléatoire désignant le nombre de réponses « oui » dans une enquête auprès de n personnes. Quelle est la loi de X ? Donner un estimateur de π et un estimateur de p. Calculer leur espérance et variance respectives.

3. En déduire un intervalle de confiance de niveau $1-\alpha$ pour p. On utilisera l'approximation normale de la loi binomiale.

4. Application numérique avec $n = 1000$ et 600 réponses affirmatives. Donner une estimation de p et un intervalle de confiance pour p au niveau 95%. Quel est le prix payé pour la confidentialité ?

PLANS A PROBABILITES INEGALES

Exercice 1 **Rappels de cours sur l'estimateur d'Horvitz-Thomson**

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-Thomson (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 **Tirage systématique de ménages**

Une population est composée de 6 ménages. Le tableau suivant donne la taille des ménages (i.e. le nombre de personnes du ménage), et le nombre de visites reçues par chaque ménage une semaine donnée. On tire un échantillon de 3 ménages sans remise, avec des probabilités proportionnelles à la taille du ménage.

Ménage	A	B	C	D	E	F
Taille	2	4	3	1	2	9
Visites	1	2	3	4	5	6

1. Donner, sous forme fractionnaire, les probabilités d'inclusion des 6 ménages (on pourra être amené à recalculer certaines probabilités).
2. Réaliser le tirage par une méthode systématique, en conservant l'ordre des ménages ci-dessus (en utilisant n'importe quel générateur de nombres aléatoires : calculette, logiciel, cerveau humain...).

3. A partir de l'échantillon obtenu, donner une estimation de la taille moyenne des ménages. Le résultat était-il prévisible ?. Donner une estimation du nombre total de visites dans la population.

4. Lister les 4 échantillons possibles que l'on peut obtenir avec un tirage systématique, et indiquer les probabilités de tirage de chacun d'eux. Vérifier que l'estimateur du nombre total de visites est sans biais.

5. Calculer la matrice des probabilités d'inclusion d'ordre 2 ? Commenter.

Exercice 3 **Tirage systématique d'entreprises**

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

On a répertorié dans une petite municipalité 6 entreprises dont les chiffres d'affaires(variable X_k) valent respectivement 40, 10, 8, 1, 0.5 et 0.5 millions d'euros. Dans le but d'estimer l'emploi salarié total, sélectionner trois entreprises au hasard et sans remise à probabilités inégales selon le chiffre d'affaires par la méthode du tirage systématique en justifiant votre démarche. Pour ce faire, on utilisera la réalisation suivante d'une variable aléatoire uniforme sur [0,1] : 0,83021. Que se passe-t-il si l'ordre du fichier est modifié ?

Exercice 4 **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 ?

Exercice 5 **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. Les communes sont listées ci-dessous avec leur nombre d'habitants :

Numéro 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

1. Donner les probabilités d'inclusion de chacune des communes.

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

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

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 :

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

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) :

	Effectifs N_h	Moyennes $\overline{Y}_h$	Variances S_h^2
mâles	60	6	4
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 **Calcul de la surface cultivée en maïs**

On cherche à estimer la surface moyenne en maïs des exploitations agricoles d'une région donnée. On a stratifié cette région avec la variable surface agricole utile des exploitations découpée en 7 classes. Le tableau ci-dessous précise pour chaque strate l'effectif, la surface moyenne cultivée en maïs et l'écart type dans la strate. On veut un échantillon total de taille n = 100 exploitations et on note :

- $Y_{i,h}$ la surface en maïs de la *i* ème exploitation de la strate *h*,
- $N = \sum_{h=1}^{7} N_h$ le nombre total d'exploitations,
- $\overline{Y} = \frac{1}{N}\sum_{h=1}^{7}\sum_{i=1}^{N_h} Y_{i,h}$ la surface moyenne inconnue en maïs de la région.
- $S_y^2 = \frac{1}{N-1}\sum_{h=1}^{7}\sum_{i=1}^{N_h}\left(Y_{i,h} - \overline{Y}\right)^2$ la dispersion totale de la surface en maïs.

Taille des exploitations en ha Strate h	Nombre d'exploitations N_h	Surface moyenne en maïs $\overline{Y}_h$	Ecart type S_{yh}
De 0 à 20 ha	394	2.7	4.1
De 21 à 40 ha	461	8.3	6.6
De 41 à 60 ha	391	12.1	7.5
De 61 à 80 ha	334	17.2	8.9
De 81 à 100 ha	169	21.1	12.2
De 101 à 120 ha	113	25.1	13.0
Plus de 121 ha	148	31.9	17.6
Total ou moyenne	2 010	13.1	

1. Définir comment est réparti l'effectif de l'échantillon dans les strates
   a. si on utilise un échantillon stratifié proportionnel,
   b. si on utilise une allocation optimale.

2. Démontrer que :

$$(N-1)S_y^2 = \sum_{h=1}^{7}(N_h - 1)S_{yh}^2 + \sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2$$

et en déduire la valeur de S_y^2.

3. Comparer les précisions des deux méthodes d'échantillonnage stratifié avec celle de l'échantillon aléatoire simple de même taille.

On calculera pour cela le design effect défini comme le rapport des précisions $\frac{V\left(\hat{\overline{Y}}_{stratifié}\right)}{V\left(\hat{\overline{Y}}_{sas}\right)}$.

4. Démontrer la formule de l'échantillon optimal $n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$.

Exercice 4 **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	Poids dans l'ensemble du personnel	Ecart type des âges
1	20%	18
2	30%	12
3	50%	3.6
Ensemble	100%	16

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 5 **Estimation d'une proportion**

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

Sur les 7500 employés de l'INSEE, 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 :

	$h = 1$	$h = 2$	$h = 3$
N_h	3500	2000	2000
n_h	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 ?

4. Que pensez-vous de l'allocation (n_1, n_2, n_3) choisie ? Quelle est la perte de variance par rapport à une allocation proportionnelle ? Par rapport à une allocation optimale ?

Exercice 6 **Optimalité 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.

Exercice 7 **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.

1. 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.

2. On réalise finalement l'enquête selon le plan de sondage permettant d'atteindre l'objectif 2. On obtient: pour les exécutants : $\bar{y}_e = 13\ s_e^2 = 9$ et pour les cadres : $\bar{y}_c = 15\ s_c^2 = 36$.

   a. Donner des intervalles de confiance à 95% pour la moyenne de l'indice de satisfaction dans chaque catégorie.
   b. La différence entre les deux valeurs moyenne de l'indice est-elle significativement différente de 0 ?

Exercice 8 **Estimation d'une différence**

On considère une population *U* de taille *N* partitionnée en *H* strates notées $U_1 \dots U_h \dots U_H$, de tailles respectives $N_1 \dots N_h \dots N_H$.On note $\bar{Y}_1 \dots \bar{Y}_h \dots \bar{Y}_H$ les moyennes d'une variable d'intérêt *Y* au sein de chaque strate, et $S_1^2 \dots S_h^2 \dots S_H^2$ les dispersions.

La moyenne de *Y* dans la population vaut bien sûr : $\bar{Y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h = \sum_{h=1}^{H} w_h \bar{Y}_h$.

On réalise un sondage stratifié, avec sondage aléatoire simple sans remise dans chaque strate, de taux de sondage $f_h = n_h / N_h$. La taille de l'échantillon total est $n = \sum_{h=1}^{H} n_h$.

L'objectif est de comparer une strate particulière U_i à la population totale : on veut estimer $D_i = \bar{Y}_i - \bar{Y}$

1. Donner l'expression de l'estimateur de Horvitz-Thompson de D_i, noté $\hat{D}_i$, ainsi que l'expression de sa variance.

2. Pour une taille d'échantillon fixée n, trouver l'allocation optimale $n_1 \dots n_h \dots n_H$, qui minimise la variance de $\hat{\mathrm{D}}_{\mathrm{i}}$. Comparer avec l'allocation optimale de Neyman.

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

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

  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 : $\eta^2 = \frac{\sum_{i=1}^{M}\left(\bar{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{M}\sum_{k=1}^{N_0}\left(Y_k - \bar{Y}\right)^2} = \frac{\sigma^2_y inter}{\sigma^2_y}$

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 :

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

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 $S^2_{\bar{y}}$. Les grappes sont de même taille.

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

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

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

Exercice 2 — 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-Thomson.

2. Estimez sans biais la variance de l'estimateur d'Horvitz-Thomson 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 3 **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 4 **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.

PLANS à PLUSIEURS DEGRES

Exercice 1 **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 2 **Estimation d'un effectif**

Un camion transportant des vis contient 500 caisses, chacune d'elles contenant 40 boîtes de vis. L'industriel réceptionnant ces caisses souhaite estimer le nombre moyen de vis par boîte. Pour cela, il tire un échantillon de 100 caisses, selon un sondage aléatoire simple sans remise, puis il tire dans chacune de ces 100 caisses 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 caisse *i* de son échantillon le nombre moyen de vis par boîte, et la variance du nombre de vis par boîte (ces deux quantités sont calculées à partir des 5 boîtes échantillonnées dans la caisse).

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

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

Il calcule aussi la variance des 100 estimations du nombre total de vis par caisse 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 3 **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 4 **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 $V(\hat{p})=\left(\frac{M}{N}\right)^2(1-f_1)\frac{S_I^2}{m}+\frac{1-f_2}{Nf_1f_2}\sum_{i=1}^{M}\frac{N_i}{N}p_i(1-p_i)$
3. Montrer que $\hat{V}(\hat{p})=\left(\frac{M}{N}\right)^2(1-f_1)\frac{s_I^2}{m}+\frac{1-f_2}{Nf_1f_2}\sum_{i=1}^{M}\frac{N_i}{N}\hat{p}_i(1-\hat{p}_i)$
4. Application numérique : donner un intervalle de confiance à 95% pour *p* avec les résultats d'enquête suivants : $\sum_{k\in S}y_k=102$, $s_I^2=1200$, $\sum_{i\in sI}\hat{p}i(1-\hat{p}i)=0{,}01$

Exercice 5 **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 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 :

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

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

REDRESSEMENTS

Exercice 1 **Post-stratification**

Un institut de sondage est chargé de mesure 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-réponses. 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 urbaine	Habitant en zone rurale	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 **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) : $n_1=70$ $n_2=30$ $\hat{\bar{y}}_1=19{,}40$ $\hat{\bar{y}}_2=51{,}63$ $s^2_{y1}=312$ $s^2_{y2}=922$.

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 : $\hat{\bar{x}}=132{,}25$ $s^2_x=9173$ $s^2_y=708$ $\hat{\rho}=0{,}57$ 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 3 **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 (on le note $\overline{Y}$)? 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 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 : $s_y=2, s_x=10, s_{xy}=15, \hat{\overline{x}}=169, \hat{\overline{y}}=24$ .Sachant que 400 000 hommes adultes vivent dans cette ville,

1. Calculez l'estimateur d'Horvitz-Thomson, 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 $\hat{\overline{x}}$, $\hat{\overline{y}}$ 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 $\overline{Y}$ 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 $\overline{X}=15$ d'un caractère auxiliaire *x*.

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

COMPLEMENTS

Exercice 1 **Algorithme de tirage bernoullien**

On considère une population *U* de 1000 individus composée de trois sous-populations disjointes U_1,U_2,U_3 de tailles respectives $N_1 = 600, N_2 = 300, N_3 = 100$. On va échantillonner dans cette population au moyen de tirage bernoullien : cette méthode consiste à choisir une probabilité d'inclusion commune π, puis à simuler sur la population une variable aléatoire distribuée selon une loi uniforme sur [0,1[et à sélectionner les individus pour lesquels la réalisation de cette variable est inférieure à π.

1. On décide dans un premier temps de tirer un échantillon dans *U* en utilisant le plan de sondage suivant :
   - dans la sous-population U_1, on réalise un tirage bernoullien, tel que chaque élément *k* a la probabilité $\pi_k = 0.1$ d'être sélectionné,
   - dans la sous-population U_2, on réalise un tirage bernoullien, tel que chaque élément *k* a la probabilité $\pi_k = 0.2$ d'être sélectionné,
   - dans la sous-population U_3, on réalise un tirage bernoullien, tel que chaque élément *k* a la probabilité $\pi_k = 0.8$ d'être sélectionné,
   - l'échantillon complet est constitué de la réunion des trois sous-échantillons ainsi obtenus.

   Calculer l'espérance et la variance de la taille n_s de l'échantillon.

2. On réalise maintenant un tirage bernoullien directement dans *U*, tel que chaque élément a la probabilité π d'être sélectionné.
   a. Déterminer π pour que l'espérance de la taille de l'échantillon, sous ce plan de sondage, soit égale à l'espérance de la taille de l'échantillon calculée à la question précédente.
   b. Calculer alors la variance de la taille de l'échantillon, et comparer cette variance à celle de la question précédente.

Exercice 2 **Tendance linéaire et tirage systématique**
(d'après J-M. Grosbras, Méthodes statistiques des sondages, Economica, 1987)

On considère une population de taille *N* avec *N = n k* où *n* est la taille souhaitée de l'échantillon et *k* un nombre entier. On suppose que pour tout individu *k* de la population, on a $Y_k = k$ pour *k = 1* à *N*.

1. On note respectivement $\overline{Y}$ et S_Y^2 la moyenne et la dispersion du caractère d'intérêt sur la population. Vérifier que $\overline{Y} = \frac{N+1}{2}$ et $S_Y^2 = \frac{N(N+1)}{12}$.

2. On réalise un sondage aléatoire simple sans remise de taille *n*.
   a. Quel est l'estimateur classique $\hat{\overline{Y}}$ de la moyenne ?
   b. Montrer que sa variance vaut : $V\left(\hat{\overline{Y}}\right) = \frac{(k-1)(N+1)}{12}$.

3. On réalise à présent un tirage systématique de taille *n* : on tire un nombre *a* au hasard entre 1 et *k* et on forme un échantillon de taille voulue avec les unités a, *a* + *k*, *a* + 2*k*, …, *a* + *(n-1)k*. Soit $\hat{\bar{Y}}_{sys}$ la moyenne des unités sélectionnées dans l'échantillon. Montrer que : $E\left(\hat{\bar{Y}}_{sys}\right) = \bar{Y}$ et que $V\left(\hat{\bar{Y}}_{sys}\right) = \frac{k^2 - 1}{12}$

4. Comparer $V\left(\hat{\bar{Y}}\right)$ et $V\left(\hat{\bar{Y}}_{sys}\right)$ et commenter

Exercice 3 **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 :

Numéro 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 4 **Algorithme de « sélection-rejet »**

La méthode de sélection-rejet permet d'obtenir un échantillon de taille *n* en une seule lecture du fichier. L'algorithme est le suivant :

- On initialise à 0 les compteurs *k* et *j* renseignant respectivement le nombre d'unités du fichier déjà examinées et le nombre d'unités déjà sélectionnées dans l'échantillon. On se positionne sur le premier individu du fichier.

- Tant que *j* est strictement inférieur à la taille d'échantillon voulue, on a généré un nombre aléatoire u selon une loi uniforme sur [0,1[pour l'individu de rang *k+1* sur lequel on est positionné et on décide :
  - Si on obtient $u < \frac{n-j}{N-k}$, alors on sélectionne l'unité de rang *k+1* . On incrémente donc *j* d'une unité, puis on passe à l'individu suivant en incrémentant *k*.
  - Sinon, l'unité *k+1* n'est pas tirée et on passe à l'individu suivant en incrémentant *k.*

1. Montrer qu'il suffit effectivement donc d'au plus *N* opérations pour sélectionner *n* unités et que les probabilités d'inclusion individuelles sont bien respectées : $\pi_k = \frac{n}{N}, \forall k \in U$.

2. Application : sélectionner un échantillon de taille 4 dans une population de taille 10 selon cette méthode en utilisant les réalisations suivantes d'une variable aléatoire U uniforme sur [0,1[:

Individu k	1	2	3	4	5	6	7	8	9	10
u_k	0.375489	0.62004	0.517951	0.0454450	0.632912	0.246090	0.927398	0.32595	0.645951	0.178048

Exercice 5 **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 10 relevés avec la méthode 1. (10 chaque mois) mais directement 20 relevés sans se soucier des mois (méthode 3.) ? Aucun calcul n'est nécessaire.

Exercice 6 **Non-réponse dans les enquêtes par quotas**

(A-M. Dussaix, J-M. Grosbras, 1992, Exercices de sondage, Economica)

L'objet de cet exercice est de montrer l'existence de biais pouvant découler de non-réponses dans les enquêtes par quotas. On considère une enquête où sont imposés des quotas relatifs à une variable qualitative donnée. Pour fixer les idées, on supposera, par exemple, qu'il y a dans la population, *H* variables d'âge ou de poids en proportion N_h/N pour *h = 1 à H*. On demande aux enquêteurs de compléter un échantillon représentatif, c'est-à-dire tel que $n_h/n = N_h/N$. A la fin de l'enquête, la moyenne $\overline{Y}$ de la variable d'intérêt est estimée par la moyenne simple sur l'échantillon $\hat{\overline{y}}$, ce qui peut encore s'écrire :

$$\hat{\overline{y}} = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\overline{y}}_h = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\overline{y}}_h \quad \text{où} \quad \hat{\overline{y}}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k \ .$$

Pour étudier l'influence de la non-réponse, on fait l'hypothèse qu'il existe dans la population un partage en 2 catégories :

- La 1ère est celle des personnes accessibles et répondant volontiers à l'enquête caractérisée par les effectifs N_1 et N_{h1} dans les tranches d'âge *h*, et les moyennes $\overline{Y}_1$ et $\overline{Y}_{h1}$.

- La 2ème est celle des personnes inaccessibles ou refusant de répondre à l'enquête caractérisée par les effectifs N_0 et N_{h0} dans les tranches d'âge *h*, et les moyennes $\overline{Y}_{\wedge}$ et $\overline{Y}_{h\wedge}$.

Naturellement, les quantités N_1, N_{h1}, N_0, N_{h0}, $\overline{Y}_1$, $\overline{Y}_{h1}$, $\overline{Y}_{\wedge}$ et $\overline{Y}_{h\wedge}$ sont inconnues.

1. Si on fait l'hypothèse que les n_h réponses constituent un échantillon d'un plan aléatoire simple sans remise prélevé dans un ensemble d'effectif N_{h1}, montrer que $\hat{\overline{y}}$ est un estimateur biaisé pour $\overline{Y}$. On écrira l'expression du biais en fonction de N, N_h, N_{h0} et $\Delta_h = \overline{Y}_{h1} - \overline{Y}_{h0}$.

2. Commentez brièvement cette expression. Construire un exemple numérique illustrant une situation où le biais est élevé (on prendra *H = 3*).

Exercice 7 **Nombre de titulaires de comptes CODEVI à interroger**

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

Une banque désire étudier par sondage (interviews par enquêteur) les caractéristiques socio-démographiques (âge, catégorie sociale,…) et les comportements financiers des titulaires de comptes CODEVI. Leur répartition en fonction des montants moyens annuels des comptes est la suivante :

Solde moyen annuel	Nombre de comptes
De 0 à 100 €	15 000
De 100 à 900 €	15 000
Plus de 900 €	30 000
Ensemble	60 000

Pour chacun des trois groupes, on veut étudier la répartition des titulaires par classe d'âge, catégorie sociale, etc. Par exemple, on s'intéresse à la proportion de titulaires ayant entre 25 et 35 ans. Quelle taille d'échantillon doit-on prévoir dans chaque groupe s'il s'agit de déterminer les différentes proportions avec une précision de $\pm$ 2,5% au niveau de confiance 95% ?

Exercice 8 **Tirage des UP avec remise – Taille de ménages**

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

Pour estimer le nombre moyen $\overline{Y}$ de personnes par ménage dans un pays donné, on réalise un tirage à 2 degrés :

- 1er degré : tirage aléatoire avec remise de *m = 4* villages parmi *M = 400* proportionnellement à leur taille. La taille d'un village est le nombre de ménages qu'il contient. Ainsi, à chacun des 4 tirages indépendants, un village est sélectionné avec une probabilité proportionnelle à sa taille.

- 2ème degré : tirage aléatoire simple de n_i ménages parmi les N_i si le village *i* est tiré.

Le nombre total de ménages dans le pays est *N = 10 000*. Les données sont représentées dans le tableau ci-dessous ; $\hat{\overline{Y}}_i$ est le nombre moyen de personnes par ménage dans le village *i* d'après l'échantillon.

i	1	2	3	4
N_i	20	23	25	18
$\hat{\overline{Y}}_i$	5.25	5.50	4.50	5

1. a. Quelle est la probabilité de tirage p_i de chacun des 4 villages sélectionnés ? (on appelle probabilité de tirage la probabilité qu'a le village d'être choisi lors de chacun des 4 tirages indépendants réalisés successivement dans les mêmes conditions).
   b. Calculer $\Pr(i \notin S)$ en fonction de (1- p_i). En déduire la probabilité d'inclusion $\pi_i = \Pr(i \in S)$ en fonction de p_i . Examiner le cas où p_i est petit.

2. Quelle est l'expression de $\overline{Y}$ (vraie valeur) et quel est son estimateur sans biais ?

3. Estimer la variance de cet estimateur. Quel intérêt a-t-on à utiliser un tirage avec remise au 1er degré ?

Exercice 9 **Raking-ratio**

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

On s'intéresse à la population des 10 000 étudiants inscrits en première année dans une université. On connaît le nombre total d'étudiants dont les parents ont un diplôme d'études primaires, secondaires et supérieures. On effectue un sondage selon un plan aléatoire simple sans remise de 150 étudiants. On ventile ces 150 étudiants selon le diplôme des parents et leurs résultats (échec ou réussite) à l'examen de première année et on obtient le résultat ci-dessous. Les nombres d'étudiants dont les parents ont un diplôme d'études primaires, secondaires et supérieures sont respectivement de 5000, 3000 et 2000.

	Echec	Réussite
Primaire	45	15
Secondaire	25	25
Supérieure	10	30

1. Estimez le taux de réussite des étudiants en utilisant l'estimateur de Horvitz-Thomson et donnez un estimateur de variance et un intervalle de confiance à 95% de ce taux.
2. Expliquez pourquoi il est a priori intéressant d'effectuer un redressement, et pourquoi le redressement doit diminuer la valeur de l'estimation issue de 1.
3. Estimez le taux de réussite des étudiants par l'estimateur post-stratifié et donnez un estimateur de variance et un intervalle de confiance à 95% de ce taux.
4. Estimez le taux de réussite par niveau d'études des parents en utilisant une technique de raking-ratio et sachant que dans la population totale étudiante, le taux de réussite est en réalité de 40%.

Exercice 10 **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 oeuvre 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) L'allocation définie par le statisticien correspond elle 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.

b) Calculer l'estimation fournie par le pisciculteur.
c) 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).
d) 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).

Proposer une nouvelle estimation du poids moyen pour l'ensemble des 3 bassins.