



Année Universitaire 2008-2009

Service Universitaire d'Enseignement à Distance

Licence A E S - Troisième année

Introduction aux sondages

Laurent Rouvière

ENSAI - Campus de Ker Lann
Rue Blaise Pascal - BP 37203
35172 BRUZ cedex
Tel : 02 99 05 32 63

Mel : laurent.rouviere@ensai.fr

Préambule

Résumé : En présence d'une taille de population très élevée, on a souvent recours à un plan de sondage pour évaluer une caractéristique précise de cette population. Dit brutalement, le sondage consiste à mesurer la caractéristique sur une partie de la population (appelée échantillon). Le statisticien doit ensuite étendre les tendances observées sur l'échantillon à la population entière. Une telle procédure soulève plusieurs difficultés telles que le choix des personnes à sonder ou encore leur nombre. Plusieurs plans de sondage sont présentés dans ce cours. La mise en oeuvre pratique ainsi que les propriétés mathématiques de ces différents plans sont étudiés en détail. Les différents concepts sont illustrés par de nombreux exemples et exercices.

Mots clés : plan de sondage aléatoire - estimateur - biais - variance - plan simple - plans stratifiés.

Prérequis Les différents thèmes de la statistique abordés en première et deuxième année de licence sont nécessaires à la compréhension de ce cours. Plus précisément les notions de variables aléatoires, biais et variance d'un estimateur ainsi que d'intervalle de confiance doivent être maîtrisées.

Objectifs d'apprentissage

- Etre capable de choisir un échantillon de manière judicieuse avant de réaliser le plan de sondage
- Savoir présenter les résultats d'un sondage, donner par exemple des marges d'erreurs (ou un niveau de confiance)

Modalités d'apprentissage Ce polycopié est composé de

- Trois chapitres de cours illustrés par des exemples et des exercices en fin de chapitre ;
- Les corrections des exercices se trouvent en Annexe B.
- De propositions de devoirs en Annexe C et D.

Conseils méthodologiques

- Les notations utilisées peuvent paraître complexes. Travailler toujours avec un exemple en tête et relier les notations avec l'exemple que vous avez choisi.
- Refaire chacun des exemples présentés dans le cours avant de passer aux exercices.
- Le fait d'avoir les corrections des exercices peut s'avérer dangereux. Regarder les uniquement pour vérifier vos réponses ou lorsque vous avez passé un temps suffisamment long sur la question.
- Venez aux stages... Il est en effet difficile de faire des mathématiques uniquement sur un polycopié. Lors des stages, j'essaie de résumer chacun des chapitres en une heure et quart environ avant de passer à des exercices "types".

- N'hésitez pas à m'envoyer par courrier les devoirs que vous avez faits. Vous pouvez poser des questions sur la copie, j'y répondrai.. Rédigez proprement.
- Vous pouvez m'envoyer par mail vos questions sur ce cours, j'y réponds assez rapidement en général (à condition que les questions soient bien détaillées...)
- Si vous avez de grandes difficultés de compréhension, vous pouvez passer à mon bureau (contactez moi avant pour être sûr que je sois la!).

Modalités d'évaluation Vous aurez un **examen écrit** de deux heures en fin d'année universitaire. Vous n'aurez droit à **aucun** document, seulement une calculatrice. Un formulaire sera distribué.

Bon courage...

Table des matières

1	Introduction	3
1.1	Qu'est-ce qu'un sondage	3
1.2	Modélisation et notation	4
1.3	Les estimateurs sont des variables aléatoires	5
1.4	Plan de sondage et qualité d'un estimateur	6
2	Sondage aléatoire simple	9
2.1	Définition du plan de sondage aléatoire simple	9
2.1.1	Plans avec ou sans remise	9
2.1.2	Plan aléatoire simple	9
2.1.3	Récapitulatif - Notations	10
2.2	Estimation de la moyenne	11
2.2.1	Estimation ponctuelle	11
2.2.2	Estimation par intervalle de confiance	14
2.3	Estimation d'une proportion	15
2.3.1	Estimation ponctuelle	16
2.3.2	Estimation par intervalle de confiance	16
2.4	Taille d'échantillon	17
2.4.1	Cas de la moyenne	17
2.4.2	Cas de la proportion	18
2.5	Exercices	20
3	Sondages stratifiés	23
3.1	Principe et justification	23
3.2	Plan de sondage stratifié	24
3.3	Estimateur de la moyenne	26
3.3.1	Un exemple	26
3.3.2	Cas général	27
3.4	Répartition de l'échantillon	27
3.4.1	Plan avec allocation proportionnelle	28
3.4.2	Plan avec allocation optimale	32
3.5	Exercices	35
A	Intervalle de confiance pour une moyenne dans un plan de sondage aléatoire simple	39

B Correction des exercices	41
C Sujet Licence AES 3 : juin 2006 (assidus)	53
D Sujet Licence AES 3 : septembre 2006 (assidus)	57
E Sujet Licence AES 3 : mai 2007 (non assidus)	61
F Sujet Licence AES 3 : mai 2008 (non assidus)	65
G Sujet Licence AES 3 : juin 2008 (non assidus)	69
H Un dernier problème...	73

Chapitre 1

Introduction

1.1 Qu'est-ce qu'un sondage

Il existe deux approches pour connaître les caractéristiques statistiques d'un caractère sur une population.

- Le **recensement** est l'approche descriptive. Il consiste à mesurer le caractère sur toute la population.
- Le **sondage** est l'approche inférentielle. Lorsque le recensement n'est pas possible pour des raisons de coût, de temps ou à cause de certaines contraintes (test destructif par exemple), on a recours à un sondage, c'est-à-dire à l'étude statistique sur un sous-ensemble de la population totale, appelé **échantillon**. Si l'échantillon est constitué de manière correcte, les caractéristiques statistiques de l'échantillon seront proches de celles de la population totale.

Exemple 1.1

Je désire connaître l'âge moyen de TOUS les étudiants de Rennes 2.

- Recensement : je demande l'âge à tous les étudiants et je calcule la moyenne... ça risque d'être long!!!
- Sondage : je choisis une partie des étudiants (échantillon), je calcule la moyenne des âges sur cette partie en espérant que cette moyenne soit "proche" de l'âge moyen de tous les étudiants.

Nous voyons sur cet exemple que la mise au point d'un sondage nécessite plusieurs choix pour le statisticien :

- comment choisir les étudiants ?
- combien d'étudiants doit-on choisir ?
- comment doit-on formuler la réponse :
 - sous la forme d'une valeur, c'est à dire que l'on donne une estimation de l'âge moyen sous la forme d'un réel (24.8 ans par exemple) ;
 - sous la forme d'un ensemble de valeurs. On pourra par exemple donner une fourchette ou un intervalle ($[23.4 ; 26.3]$ par exemple).
- est-ce que l'estimation est satisfaisante ? Dit autrement suis-je capable de donner une estimation de l'erreur commise par la prédiction. On pourra par exemple dire "*l'âge moyen des étudiants de Rennes 2 se trouvent dans l'intervalle $[23.4 ; 26.3]$ avec un niveau de confiance de 95%.*".

L'objectif de ce cours consiste à étudier des procédures de sondage pour lesquelles nous pourrions répondre à ces questions. Nous allons dans ce chapitre présenter le contexte, les notations ainsi que les critères permettant d'évaluer la qualité d'un sondage. Nous proposerons dans les chapitres 2 et 3 différentes méthodes de sondage permettant d'estimer des moyennes et proportions.

1.2 Modélisation et notation

Nous présentons dans cette partie le cadre d'étude et introduisons les notations qui seront utilisées tout au long de ce cours.

On s'intéresse à une population U composés d'individus ou **unités** (étudiants de Rennes 2). Chaque unité est représentée par un numéro allant de 1 à N :

$$U = \{U_1, \dots, U_N\} = \text{base de sondage.}$$

On souhaite évaluer une caractéristique de la population (l'âge par exemple). On note X_i la valeur de ce caractère mesuré sur l'individu i (X_i est donc ici l'âge du $i^{\text{ème}}$ individu). On peut utiliser un sondage pour estimer l'âge moyen

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i.$$

Une autre caractéristique souvent étudiée est le total

$$T = \sum_{i=1}^N X_i.$$

On peut également s'intéresser à une proportion d'individus qui vérifie un certain critère. Dans ce cas, X_i prendra deux valeurs :

- 1 si l'individu U_i satisfait le critère ;
- 0 sinon.

La proportion d'individus appartenant à la catégorie qui nous intéresse sera alors :

$$p = \frac{1}{N} \sum_{i=1}^N X_i.$$

Exemple 1.2

Considérons le cas d'un sondage électoral. On s'intéresse à la proportion d'individus votant pour un candidat A. On définit alors X_i la variable qui prend pour valeurs :

- 1 si l'individu U_i vote pour un candidat A ;
- 0 sinon.

Le nombre d'individus qui votent pour A est

$$\sum_{i=1}^n X_i,$$

on en déduit que la proportion d'individus qui votent pour A est

$$p = \frac{1}{N} \sum_{i=1}^N X_i.$$

Pour différentes raisons (coûts, temps...), on ne peut pas mesurer la caractéristique sur tous les individus. Par conséquent les paramètres μ, T ou p sont **inconnus**. On sélectionne alors un sous ensemble de la population U constitué de n unités de la population ($n \leq N$) (voir Figure 1.1). Ce sous-ensemble est appelé *échantillon* et sera noté E .

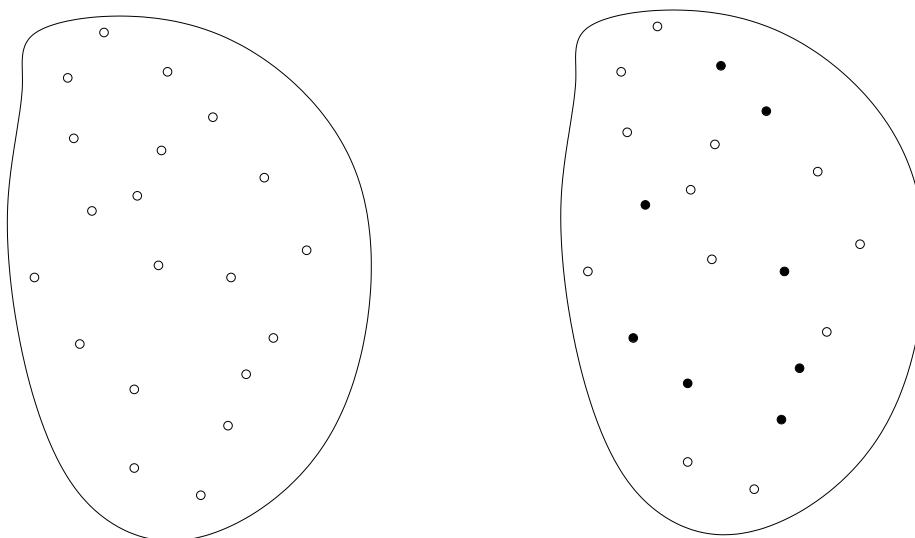


FIGURE 1.1 – Population composée de $N = 20$ individus (gauche) dans laquelle on sélectionne un échantillon de $n = 8$ individus représentés par des ronds noirs (droite).

On désignera par x_1, \dots, x_n les valeurs de la caractéristique (âge) observées sur l'échantillon. Ces valeurs sont **connues**, et tout le problème consiste désormais à estimer les paramètres inconnus à partir des valeurs mesurées sur l'échantillon (qui elles sont connues).

Exemple 1.3

Un moyen naturel d'estimer la moyenne μ consiste à prendre la moyenne observée sur l'échantillon :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Le total T sera quant à lui estimé par

$$t = \sum_{i=1}^n x_i.$$

1.3 Les estimateurs sont des variables aléatoires

Considérons l'exemple suivant.

Exemple 1.4

Nous disposons d'une population composée de $N = 5$ individus. Nous nous posons le problème de connaître l'âge moyen μ de ces individus. Pour certaines raisons, on ne peut demander l'âge qu'à $n = 2$ individus qui constitueront l'échantillon (bien entendu, une telle situation ne se produit jamais en réalité...). Le statisticien propose d'estimer l'âge moyen des 5 étudiants par l'âge moyen $\hat{\mu}$ des deux étudiants de l'échantillon.

Supposons que l'âge des 5 étudiants soit : 15, 25, 18, 14, 20. Si l'échantillon est constitué par les deux premiers individus, l'estimation de μ sera $\frac{15+25}{2} = 20$. Si maintenant l'échantillon est constitué des deux derniers individus alors l'estimation vaudra $\frac{14+20}{2} = 17$. Nous voyons clairement que la valeur de $\hat{\mu}$ va dépendre des individus présents dans l'échantillon. C'est en ce sens que nous affirmons que l'estimateur $\hat{\mu}$ est une **variable aléatoire** (il peut prendre différentes valeurs suivant l'échantillon choisi).

Ce qui est aléatoire dans un sondage est le fait qu'un individu donné appartienne ou non à l'échantillon.

Dans la suite, pour les différents plans de sondage que nous étudierons, nous noterons les estimateurs avec des "chapeaux" (voir la tableau suivant).

	Vraie valeur	Estimateur
Moyenne	μ	$\hat{\mu}$
Total	T	\hat{T}
Proportion	p	\hat{p}

1.4 Plan de sondage et qualité d'un estimateur

Nous nous plaçons dans le cas de l'estimation de la moyenne μ d'une certaine caractéristique sur une population. Tous les concepts étudiés dans cette partie sont également valables pour l'estimation d'un total ou d'une proportion. Nous rappelons que

$$U = (U_1, \dots, U_N)$$

désigne la population ou la base de sondage et nous noterons

$$E = (u_1, \dots, u_n)$$

un sous-ensemble de u de taille $n \leq N$ qui constituera l'échantillon. Le problème consiste à construire un estimateur $\hat{\mu}$ de μ à partir de l'échantillon.

Comment être sûr que $\hat{\mu}$ soit proche de μ .

Eléments de réponse :

- si n est proche de N , alors l'échantillon est proche de la population. n joue donc un rôle dans la réponse.
- E doit "représenter" U . Si par exemple μ est le revenu annuel moyen de la population française et que l'échantillon est constitué d'un groupe d'étudiants, il sera difficile de construire un estimateur $\hat{\mu}$ qui sera proche de μ .

Plusieurs questions peuvent être posées concernant le choix de E :

- Comment s'assurer que E soit représentatif de U ? En contrôlant la façon dont il est sélectionné.
- Mais U est inconnu : comment faire pour que E “ressemble” à U ? Le problème est insoluble. Au mieux, on peut seulement maximiser les chances que E représente U .
- Comment maximiser les chances ? En utilisant un sondage probabiliste.

Définition 1.1

Un **plan de sondage** est une procédure permettant de sélectionner un échantillon E dans une population U . Un plan de sondage est dit **probabiliste** ou **aléatoire** si chaque individu de la population U a une probabilité connue de se retrouver dans l'échantillon E .

Dans les chapitres à venir, nous nous intéresserons à différents plans de sondage aléatoires. Pour un plan donné, un estimateur $\hat{\mu}$ de la moyenne μ sera construit sur l'échantillon. La qualité du sondage est mesurée par la qualité de l'estimateur.

Nous avons vu dans la partie précédente que pour un plan de sondage aléatoire, l'estimateur $\hat{\mu}$ est une variable aléatoire. On va donc pouvoir calculer son espérance et sa variance. Ces deux quantités seront utilisées pour mesurer la qualité de l'estimateur.

Définition 1.2

On définit le **biais** d'un estimateur $\hat{\mu}$ par :

$$B(\hat{\mu}) = \mathbf{E}(\hat{\mu}) - \mu.$$

Ainsi, on dira que $\hat{\mu}$ est un **estimateur sans biais** de μ si

$$B(\hat{\mu}) = 0 \iff \mathbf{E}(\hat{\mu}) = \mu.$$

Dit autrement, $\hat{\mu}$ “tombe” en moyenne sur sa cible μ .

Remarque

- Dire que l'estimateur est sans biais ne veut pas dire que le résultat soit exact. Avant de réaliser l'échantillon, on ne connaît pas la valeur de $\hat{\mu}$, on sait seulement que c'est une variable aléatoire qui en moyenne vaut μ .
- Dire que l'estimateur est sans biais revient à dire que la valeur moyenne de $\hat{\mu}$ sur tous les échantillons possibles est la vraie valeur μ .

Sur la Figure 1.2, nous schématisons cette notion de biais. La vraie valeur de μ est la cible à atteindre (carré). Les points désignent les différentes valeurs de l'estimateur $\hat{\mu}$ suivant l'échantillon.

L'estimateur de gauche est sans biais : la valeur moyenne de toutes les valeurs $\hat{\mu}$ est égale à la cible μ . Ce n'est clairement pas le cas pour l'estimateur associé à la figure de droite.

Pour un estimateur sans biais $\hat{\mu}$, il est aussi utile de savoir comment l'ensemble des valeurs possibles de $\hat{\mu}$ se répartit autour de la cible μ , si elles en sont proches ou s'il y a un risque de tomber sur une combinaison malheureuse (un “mauvais” échantillon).

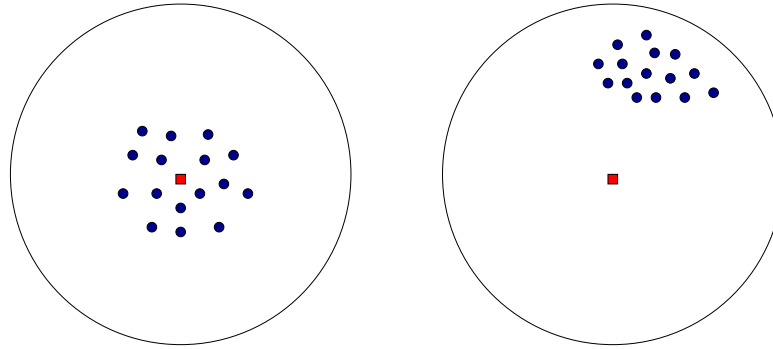


FIGURE 1.2 – Un exemple d'estimateur sans biais (gauche) et biaisé (droite).

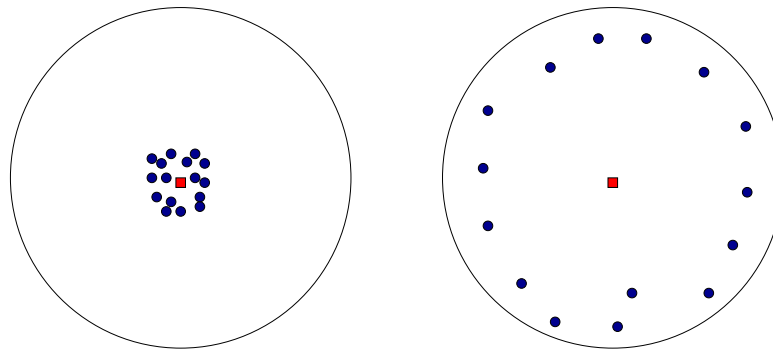


FIGURE 1.3 – Deux exemples d'estimateur sans biais : à gauche la variance est faible, à droite elle est élevée.

Les deux estimateurs schématisés sur la Figure 1.3 sont sans biais. Nous voyons cependant que les valeurs de $\hat{\mu}$ pour l'estimateur de gauche sont plus proches de μ que pour celui de droite. On préférera ainsi l'estimateur de gauche à celui de droite.

La dispersion de $\hat{\mu}$ autour de μ se mesure par la variance de l'estimateur :

- à gauche, la variance est faible \rightarrow les différentes valeurs de $\hat{\mu}$ sont faiblement dispersées autour de μ .
- à droite, la variance est élevée \rightarrow les différentes valeurs de $\hat{\mu}$ sont fortement dispersées autour de μ .

Le tableau ci-dessous résume la mesure de la qualité de l'estimateur en fonction de son biais (espérance) et de sa dispersion (variance).

Qualité	Biais	Dispersion
bonne	faible	faible
mauvaise	élevée	élevée

Pour des plans de sondage aléatoires, la difficulté consiste à rechercher des estimateurs sans biais (éventuellement de biais faible), et de variance minimale.

Chapitre 2

Sondage aléatoire simple

2.1 Définition du plan de sondage aléatoire simple

Le sondage aléatoire simple est le modèle d'échantillonnage en apparence le plus simple que l'on puisse imaginer : il consiste à considérer que, dans une population d'effectif N , tous les échantillons de n unités sont possibles avec la même probabilité.

2.1.1 Plans avec ou sans remise

Définition 2.1

Un plan de sondage est dit **avec remise** si un même individu peut apparaître plusieurs fois dans l'échantillon et si l'ordre dans lequel apparaissent les individus compte.

Exemple 2.1

$\mathcal{P} = \{1, 2, 3, 4, 5\}$, $n = 3$. L'échantillon $\{1, 1, 2\}$ est différent de l'échantillon $\{1, 2, 1\}$.

Dans le cas d'un plan avec remise, il y a N^n échantillons possibles.

Définition 2.2

Un plan de sondage est dit **sans remise** si un même individu ne peut apparaître qu'une seule fois dans l'échantillon.

Dans l'exemple précédent, l'échantillon $\{1, 1, 2\}$ n'est donc pas possible.

Dans le cas d'un plan sans remise, il y a $C_N^n = \frac{N!}{n!(N-n)!}$ échantillons possibles.

La plupart du temps, nous nous intéresserons aux plans sans remise : interroger deux fois le même individu n'apporte pas d'information supplémentaire. Cependant, il n'est pas inintéressant de considérer parfois des plans avec remise, ne serait-ce que pour servir d'élément de comparaison et de référence.

2.1.2 Plan aléatoire simple

Définition 2.3 (Plan simple)

Un plan de sondage aléatoire est dit simple, ou à probabilités égales, si chaque échantillon a la même probabilité qu'un autre d'être tiré au sort.

Exemple 2.2

Dans le cas d'un plan simple sans remise, un échantillon de taille fixe n a donc une probabilité égale à $\frac{1}{C_N^n} = \frac{n!(N-n)!}{N!}$ d'être tiré au sort. Si $N = 5$ et $n = 2$, cette probabilité est donc égale à $\frac{1}{5 \times 4 \times 3 \times 2} = \frac{1}{10}$.

Proposition 2.1 (Probabilité d'inclusion)

Tous les individus ont la même probabilité d'être sélectionnés dans l'échantillon et cette probabilité est égale à $\frac{n}{N}$.

2.1.3 Récapitulatif - Notations**Remarque (très importante)**

- Les données concernant la **population** toute entière (X_i pour tous les i , μ , T , $p...$) sont **inconnues et déterministes** (puisque l'on a pas accès aux informations concernant toute la population);
- En revanche, les valeurs obtenues à partir de l'**échantillon** sont **connues et aléatoires**. Elles dépendent en effet du hasard puisqu'elles varient d'un échantillon aléatoire à un autre, et elles sont connues puisque l'on dispose des informations nécessaires pour les calculer sur l'échantillon.

Le tableau suivant récapitule les notions relatives à la population et à l'échantillon.

	Population U <u>inconnu, déterministe</u>	Échantillon E <u>connu, aléatoire</u>
Taille	N	n
Moyenne	$\mu = \frac{1}{N} \sum_{k=1}^N X_k$	$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$
Total	$T = \sum_{k=1}^N X_k = N\mu$	$t = \sum_{k=1}^n x_k = n\bar{x}$
Variance	$\sigma^2 = \frac{1}{N} \sum_{k=1}^N (X_k - \mu)^2$	
Variance corrigée	$S^2 = \frac{1}{N-1} \sum_{k=1}^N (X_k - \mu)^2$ $= \frac{N}{N-1} \sigma^2$	$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$

Rappels : moyenne et écart-type Pour toute variable aléatoire X , on peut calculer sa moyenne et son écart-type.

$$\text{Moyenne} = \frac{\sum \text{valeur}}{\text{Effectif total}}$$

$$\text{Variance} = \frac{\sum (\text{valeur} - \text{moyenne})^2}{\text{Effectif total}} = \frac{\sum \text{valeur}^2}{\text{Effectif total}} - \text{moyenne}^2$$

$$\text{Ecart-type} = \sqrt{\text{Variance}}$$

On rappelle que l'écart-type donne une idée de la dispersion des données autour de la moyenne.

Remarque (très importante)

La moyenne \bar{x} observée sur l'échantillon est une variable aléatoire qui prend des valeurs différentes d'un échantillon à un autre. On peut donc calculer son espérance et sa variance (à ne surtout pas confondre avec la variance du caractère dans la population notée σ^2 ou dans l'échantillon notée s^2).

2.2 Estimation de la moyenne

2.2.1 Estimation ponctuelle

On va estimer μ par une valeur $\hat{\mu}$.

Problème : Trouver une méthode qui nous permette de donner une estimation de μ à partir de l'échantillon sélectionné par un plan de sondage aléatoire simple ?

Solution : Dans ce chapitre, nous estimons la moyenne μ par la moyenne observée sur l'échantillon. On appelle **estimateur** de μ la "formule" qui nous permet de calculer une estimation du paramètre inconnu (μ). Dans le cas que nous étudions, l'estimateur de μ , que nous noterons $\hat{\mu}$ n'est rien d'autre que \bar{x} :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}. \quad (2.1)$$

Exemple 2.3

On dispose de $N = 5$ jetons portant les valeurs -1, 2, 4, 10, 20.

1. Calculer la moyenne et la variance de la valeur sur toute la population ($\mu = 7$, $\sigma^2 = 55.1$, $\sigma = 7.43$).
2. On souhaite estimer la moyenne μ calculée précédemment par un sondage aléatoire simple (ça n'a aucun sens, juste mieux comprendre le problème). On tire un échantillon de taille $n = 2$ sans remise. Établir la liste de tous les échantillons possibles, et calculer la moyenne pour chacun d'eux.

Ech	$\hat{\mu}$ ou \bar{x}	Ech	$\hat{\mu}$ ou \bar{x}
{-1, 2}	0.5	{2, 10}	6
{-1, 4}	1.5	{2, 20}	11
{-1, 10}	4.5	{4, 10}	7
{-1, 20}	9.5	{4, 20}	12
{2, 4}	3	{10, 20}	15

3. Calculer l'espérance de la variable aléatoire ainsi obtenue.

Soit x_i ($i = 1, 2$) la variable aléatoire correspondant à la valeur du $i^{\text{ème}}$ jeton dans l'échantillon. La moyenne empirique des x_i est l'estimateur $\hat{\mu}$

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2}{2}.$$

Cet estimateur est une variable aléatoire dont la loi est donnée par :

Valeurs de $\hat{\mu}$ ou \bar{x}	0.5	1.5	4.5	9.5	3	6	11	7	12	15
Probabilités	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Les probabilités sont égales car on est dans un plan aléatoire simple (tous les échantillons ont la même probabilité). On déduit ainsi l'espérance et la variance de \bar{X} .

$$\mathbf{E}(\bar{x}) = 7, \quad \mathbf{V}(\bar{x}) = 20.7.$$

Exemple 2.4

Une société bancaire souhaite mener une étude approfondie auprès des particuliers ayant un compte chez elle : il s'agit de préparer le lancement d'un nouveau produit financier. La société dispose d'un fichier de N (N grand) clients et l'étude par sondage doit porter sur n ($n < N$) d'entre eux. Pour illustrer les propriétés du SAS, nous allons simplifier à l'extrême : supposons que le fichier comporte $N = 5$ titulaires de comptes et prélevons un échantillon d'effectif $n = 2$. A la date de l'étude, les dépôts sur ces 5 comptes sont, en millier de francs : 13, 15, 17, 25, 30. La moyenne de ces 5 valeurs est égale à $\mu = 20$. On suppose que l'organisme chargé de l'enquête ignore ces montants et se fixe pour objectif d'évaluer leur moyenne à partir de deux valeurs qu'il constatera sur l'échantillon.

1. Établir la liste de tous les échantillons possibles et calculer la moyenne pour chacun d'eux.

Ech	\bar{x}	Ech	\bar{x}
{13, 15}	14	{15, 25}	20
{13, 17}	15	{15, 30}	22.5
{13, 25}	19	{17, 25}	21
{13, 30}	21.5	{17, 30}	23.5
{15, 17}	16	{25, 30}	27.5

2. Calculer l'espérance et la variance de la variable aléatoire ainsi obtenue.

Soit x_i ($i = 1, 2$) la variable aléatoire correspondant à la valeur du i -ème compte prélevée. La moyenne empirique des x_i

$$\bar{x} = \frac{x_1 + x_2}{2}$$

est une variable aléatoire dont la loi est donnée par :

Valeurs de \bar{x}	14	15	19	21.5	16	20	22.5	21	23.5	27.5
Probabilités	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

On déduit ainsi l'espérance et la variance de \bar{X} .

$$\mathbf{E}(\bar{x}) = 20, \quad \mathbf{V}(\bar{x}) = 15.6.$$

Nous remarquons que pour les exemples 2.3 et 2.4, l'estimateur $\hat{\mu}$ est sans biais. Le théorème suivant montre que ceci est toujours le cas pour un plan de sondage aléatoire simple.

Théorème 2.1

Soit $\hat{\mu}$ l'estimateur d'une moyenne μ pour un plan de sondage aléatoire simple défini par (2.1). On a alors

$$\mathbf{E}(\hat{\mu}) = \mu.$$

Dit autrement, $\hat{\mu}$ est un **estimateur sans biais** de μ , c'est à dire qu'il "tombe" en moyenne sur sa cible μ .

On peut utiliser ce résultat pour calculer directement l'espérance de $\hat{\mu}$ dans les exemples 2.3 et 2.4.

Il est aussi utile de savoir comment l'ensemble des résultats possibles (l'ensemble des moyennes des échantillons) se répartit autour de la cible μ , s'ils en sont proches, ou s'il y a un risque de tomber sur une combinaison malheureuse (sur un mauvais échantillon). Pour cela, nous rappelons que la variance de $\hat{\mu}$ est un indice qui permet de mesurer cette dispersion.

Théorème 2.2

Soit f le taux de sondage $f = n/N$. Alors

$$\mathbf{V}(\hat{\mu}) = (1 - f) \frac{S^2}{n} = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}. \quad (2.2)$$

On peut aussi écrire

$$\mathbf{V}(\hat{\mu}) = \frac{\sigma^2}{n} \frac{N - n}{N - 1}.$$

On peut vérifier à l'aide de ce résultat les calculs de variance de $\hat{\mu}$ pour les exemples 2.3 et 2.4.

Pour l'exemple 2.3, on a $\sigma = 7.43$, $N = 5$, $n = 2$ donc

$$S^2 = \frac{N}{N - 1} \sigma^2 = \frac{5}{4} 7.43^2 = 69.$$

Par conséquent, d'après le Théorème 2.2

$$\mathbf{V}(\hat{\mu}) = (1 - f) \frac{S^2}{n} = \left(1 - \frac{2}{5}\right) \frac{69}{2} = 20.7.$$

Remarque

La formule (2.2) permet de caractériser la précision d'un SAS (plus la variance est faible, plus l'estimateur est précis).

- Plus la taille n de l'échantillon est grande, plus la variance de $\hat{\mu}$ diminue et donc plus l'estimateur est précis. À l'extrême, si $n = N$ la variance est nulle. Ceci est "normal", car dans ce cas on a réalisé un recensement et on connaît de façon certaine la vraie moyenne.

- La précision dépend également de la variance de la variable d'intérêt σ^2 (ou S^2) dans la base de sondage. C'est une condition naturelle : plus une population est homogène (variance faible), plus le sondage y est efficace. A l'extrême, si la variance σ^2 est nulle (tous les individus ont le même âge), la variance de l'estimateur est nulle et nous aurons besoin d'un seul individu pour connaître μ de manière parfaite. A l'inverse, sonder dans une population très hétérogène nécessite des tailles d'échantillons de taille importante, ou un découpage au préalable en sous populations homogènes (c'est le principe des *sondages stratifiés* que nous verrons dans le chapitre 3).

Exemple 2.5

Reprenons l'exemple de la société bancaire. La société dispose d'un fichier de $N = 50\,000$ clients et l'étude par sondage doit porter sur $n = 200$ d'entre eux. On note μ le montant moyen des comptes des 5000 clients. On suppose que la variance σ^2 du montant est connue et vaut 41.6. On a alors

$$V(\hat{\mu}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} = \frac{41.6}{200} \frac{50000-200}{50000-1} \approx 0.21.$$

Pour un échantillon de taille 500, on obtient

$$V(\hat{\mu}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} = \frac{41.6}{500} \frac{50000-500}{50000-1} \approx 0.08.$$

2.2.2 Estimation par intervalle de confiance

On cherche une fourchette de valeurs possibles pour μ à laquelle on puisse associer un certain degré de confiance (par exemple 95%).

Exemple 2.6

Si une enquête montre que l'on peut affirmer avec un niveau de confiance de 95% que le temps moyen passé par jour par les français à regarder la télévision se situe entre 1h30 et 3h00, on dit que $[1, 5; 3]$ est un intervalle de confiance à 95% pour la durée moyenne passée par jour par les français à regarder la télévision.

Notations :

$(1 - \alpha)$: niveau de confiance

α : risque

$z_{1-\frac{\alpha}{2}}$: quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

- Si la variance corrigée S^2 est connue :

$$IC_{1-\alpha}(\mu) \simeq \left[\hat{\mu} \pm z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\mu})} \right] \simeq \left[\hat{\mu} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{(1-f)}{n} S^2} \right]$$

- Si S^2 est inconnue, on la remplace par une estimation :

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{n}{n-1} \left[\frac{\sum_{k=1}^n x_k^2}{n} - \bar{x}^2 \right]$$

Preuve : voir Annexe A pour un rappel sur le Théorème central limite et la construction de cet intervalle de confiance.

Récapitulatif : L'estimation d'une moyenne μ d'un caractère sur une population de taille se réalise de la manière suivante :

- On prélève "au hasard" n individus parmi les N sur lesquels on mesure le caractère. On obtient alors une suite de variables aléatoires x_1, \dots, x_n (échantillon).
- $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ est l'estimateur ponctuel de μ .
- Son espérance vaut μ et sa variance vaut $(1 - \frac{n}{N}) \frac{S^2}{n}$.
- $\left[\hat{\mu} - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}, \hat{\mu} + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \right]$ est un intervalle de confiance de niveau $1 - \alpha$ pour μ .

Exemple 2.7

Reprenons l'exemple de la société bancaire. La société dispose de $N = 50000$ clients et l'organisme chargé de l'enquête recueille les données relatives à $n = 200$ clients. On s'intéresse à nouveau au montant présent sur les comptes des clients. Par conséquent le paramètre à estimer sera μ : le montant moyen présent sur les comptes des 50000 clients.

Les 200 comptes sondés ont un montant moyen $\hat{\mu} = 22.5$ et une variance $s^2 = 42.2$. Calculons l'intervalle de confiance de niveau $1 - \alpha = 0.95$. L'intervalle est donné par :

$$\left[\hat{\mu} - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}, \hat{\mu} + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}} \right].$$

$z_{1-\alpha/2}$ est la quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$. Ici $1 - \alpha = 0.95$ donc $\alpha = 0.05$ et $1 - \alpha/2 = 0.975$. $z_{1-\alpha/2}$ est donc le quantile d'ordre 0.975 de la loi $\mathcal{N}(0, 1)$ que l'on lit dans la table. On trouve $z_{1-\alpha/2} = 1.96$.

Un IC de niveau 0.95 est donc

$$\left[22.5 - 1.96 \sqrt{\frac{42.2}{200}}, 22.5 + 1.96 \sqrt{\frac{42.2}{200}} \right] = [21.6, 23.4].$$

Remarque

Donner une estimation par intervalle de confiance est doublement prudent ; d'une part, on ne fournit pas une valeur ponctuelle, mais une plage de valeur possibles ; d'autre part, on prévient qu'il existe un risque faible que la vraie valeur soit en dehors de la fourchette.

2.3 Estimation d'une proportion

Une proportion peut-être considérée comme un cas particulier de la moyenne.

2.3.1 Estimation ponctuelle

Exemple 2.8

Poursuivons l'exemple de la société bancaire qui souhaite réaliser une enquête pour estimer la proportion p de clients prêts à souscrire à un nouveau produit financier. La société dispose de $N = 50000$ clients et souhaite réaliser son enquête sur $n = 200$ clients.

Construisons la variable aléatoire x_i qui au $i^{\text{ème}}$ client interrogé fait correspondre la valeur suivante :

- $x_i = 1$ si le client i a l'intention de souscrire au produit ;
- $x_i = 0$ sinon.

Remarquons que x_i suit une loi de Bernoulli de paramètre p . La proportion p de clients favorables est naturellement estimée par la proportion \hat{p} de clients interrogés (sondés) favorable. On remarque que

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Ainsi en utilisant les Théorèmes 2.1 et 2.2, on montre que :

$$\mathbf{E}(\hat{p}) = p$$

et

$$\mathbf{V}(\hat{p}) = (1 - f) \frac{S^2}{n} = (1 - f) \frac{Np(1 - p)}{n(N - 1)}.$$

2.3.2 Estimation par intervalle de confiance

En suivant un raisonnement analogue au cas de la moyenne, on montre qu'un IC de niveau $1 - \alpha$ pour une proportion p est donné par :

$$\left[\hat{p} - z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{p})}, \hat{p} + z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{p})} \right],$$

avec

$$\mathbf{V}(\hat{p}) = (1 - f) \frac{S^2}{n} = (1 - f) \frac{Np(1 - p)}{n(N - 1)}.$$

D'où l'IC

$$\left[\hat{p} - z_{1-\alpha/2} \sqrt{(1 - f) \frac{S^2}{n}}; \hat{p} + z_{1-\alpha/2} \sqrt{(1 - f) \frac{S^2}{n}} \right]. \quad (2.3)$$

Remarque

$\mathbf{V}(\hat{p})$ dépend de la proportion p qui est inconnue. En pratique dans la formule (2.3), on remplace $\mathbf{V}(\hat{p})$ par son estimateur

$$\hat{\mathbf{V}}(\hat{p}) = (1 - f) \frac{s^2}{n} = (1 - f) \frac{\hat{p}(1 - \hat{p})}{n - 1},$$

ce qui donne l'intervalle

$$\left[\hat{p} - z_{1-\alpha/2} \sqrt{(1 - f) \frac{\hat{p}(1 - \hat{p})}{n - 1}}, \hat{p} + z_{1-\alpha/2} \sqrt{(1 - f) \frac{\hat{p}(1 - \hat{p})}{n - 1}} \right]. \quad (2.4)$$

Exemple 2.9 (Calcul d'un IC pour une proportion)

La banque possède $N = 1\,000$ clients. Sur $n = 200$ clients interrogés, 30 se déclarent favorable à souscrire au nouveau produit financier. Déterminer un IC de niveau 0.95 pour p .

$1 - \alpha = 0.95$ donc $z_{1-\alpha/2} = 1.96$. Sur les 200 clients interrogés, 30 sont favorables donc la proportion de personnes favorable sur l'échantillon est $\hat{p} = \frac{30}{200} = 0.15$. Un IC de niveau 0.95 est :

$$\left[0.15 - 1.96 \sqrt{\left(1 - \frac{200}{1\,000}\right) \frac{0.15(1 - 0.15)}{200}}, 0.15 + 1.96 \sqrt{\left(1 - \frac{200}{1\,000}\right) \frac{0.15(1 - 0.15)}{200}} \right]$$

$$\approx [0.106, 0.194]$$

2.4 Taille d'échantillon

Jusqu'à présent la taille d'échantillon n était fixée. Cependant, on pose souvent la question au statisticien : "A partir de combien d'élément un échantillon est-il valable ?". Bien entendu, il faut définir ce qu'on entend par valable. Dans le contexte qui est le nôtre, nous conviendrons d'un écart maximum toléré de l'intervalle de confiance. C'est à dire que nous chercherons la taille d'échantillon minimum n_0 de manière à ce que l'intervalle de confiance ne soit *pas trop grand*. Plus précisément, nous fixons une demi-longueur h_0 pour l'intervalle de confiance et nous cherchons la taille d'échantillon n_0 pour laquelle la demi-longueur de l'intervalle de confiance vaut h_0 .

2.4.1 Cas de la moyenne

Dans le cadre de l'estimation d'une moyenne, on rappelle que l'intervalle de confiance de niveau $1 - \alpha$ est donné par :

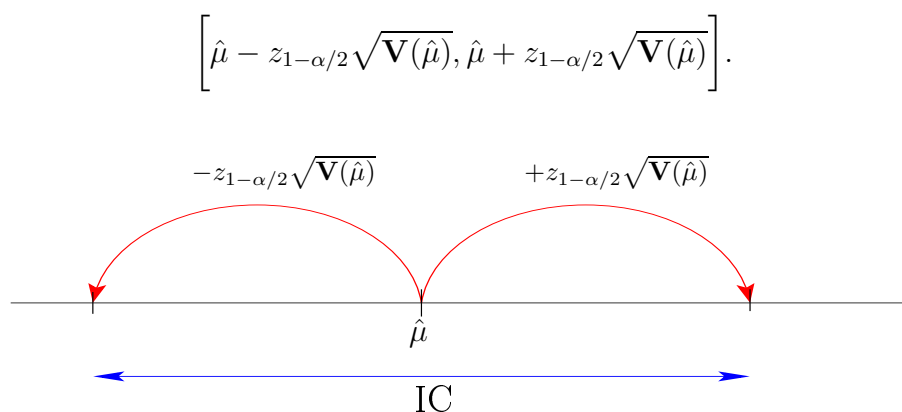


FIGURE 2.1 – Intervalle de confiance.

La demi longueur de l'IC vaut donc (voir Figure 2.1)

$$z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{\mu})},$$

ou encore

$$\begin{aligned} z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{\mu})} &= z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \\ &\approx z_{1-\alpha/2} \sqrt{\frac{S^2}{n}} \end{aligned}$$

on considère que le taux de sondage n/N est proche de 0.

$$\approx z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}}$$

on approche la variance corrigée par la variance.

Problème : cette demi longueur dépend de la variance de tous les individus qui est inconnue. Une solution consiste à utiliser un majorant σ_{max}^2 de cette variance σ^2 (ce majorant sera en général déterminé sur la base d'une enquête précédente). La demi longueur de l'IC sera alors au plus égale à

$$z_{1-\alpha/2} \sqrt{\frac{\sigma_{max}^2}{n}}$$

(on se place dans le pire des cas, c'est à dire celui où la variance vaut σ_{max}^2). Par conséquent la taille d'échantillon minimum n_0 telle que la demi longueur de l'IC ne dépasse pas h_0 sera la solution de l'équation

$$z_{1-\alpha/2} \sqrt{\frac{\sigma_{max}^2}{n_0}} = h_0,$$

c'est-à-dire

$$n_0 = \frac{z_{1-\alpha/2}^2 \sigma_{max}^2}{h_0^2}.$$

2.4.2 Cas de la proportion

Pour la proportion, on négligera le taux de sondage et on approchera la demi-longueur de l'IC par :

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

Ici le problème est que cette demi longueur dépend de la proportion p qui est inconnue. Cependant une simple étude de fonction montre que

$$\forall p \in [0, 1], \quad p(1-p) \leq 1/4.$$

Par conséquent, la demi longueur de l'IC est au plus égale à

$$z_{1-\alpha/2} \sqrt{\frac{1}{4n}}$$

(on se place dans le pire des cas où $p(1-p) = 1/4$). La taille d'échantillon minimum n_0 telle que la demi longueur de l'IC ne dépasse pas h_0 est la solution de l'équation

$$z_{1-\alpha/2} \sqrt{\frac{1}{4n_0}} = h_0$$

c'est-à-dire

$$n_0 = \frac{z_{1-\alpha/2}^2}{4h_0^2}.$$

2.5 Exercices

Exercice 2.1

Soit une caractéristique X définie sur une population de $N = 4$ unités.

Individu	1	2	3	4
Valeur de X	11	10	8	11

1. Calculer la valeur des paramètres suivants de la population : la moyenne, la variance, et la variance corrigée, notées respectivement μ , σ^2 , et S^2 .
2. On tire un échantillon sans remise de taille $n = 2$ à probabilités égales.
 - (a) Combien d'échantillons peut-on tirer ?
 - (b) Pour chaque échantillon possible, calculer la moyenne \bar{x} et la variance corrigée s^2 obtenues sur l'échantillon.
 - (c) Calculer $\mathbf{E}(\bar{x})$, $\mathbf{V}(\bar{x})$, et $\mathbf{E}(s^2)$.

Exercice 2.2

Sur la population $\{1, 2, 3\}$, on considère le plan de sondage suivant :

$$n = 2$$

$$\mathbf{P}(\{1, 2\}) = \frac{1}{2} \text{ (c'est-à-dire que l'échantillon } \{1, 2\} \text{ a une probabilité } \frac{1}{2} \text{ d'apparaître)}$$

$$\mathbf{P}(\{1, 3\}) = \frac{1}{4}$$

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

1. Est-ce un sondage aléatoire simple ?
2. Calculer la probabilité pour que l'individu 1 fasse partie de l'échantillon. Même question pour les individus 2 et 3.
3. Calculer la valeur de l'estimateur de la moyenne pour chaque échantillon possible.
4. Vérifier que cet estimateur est biaisé.

Exercice 2.3

On veut estimer la superficie moyenne cultivée dans les fermes d'un canton rural. Sur les 2010 fermes que comprend le canton, on en tire 100 par sondage aléatoire simple. On mesure (en hectares) la surface cultivée x_k par la ferme numéro k de l'échantillon et on trouve :

$$\sum_{k=1}^{100} x_k = 2907 \quad \text{et} \quad \sum_{k=1}^{100} x_k^2 = 154593.$$

1. Donner la valeur de l'estimateur de la moyenne $\hat{\mu} = \bar{x}$.
2. Donner un intervalle de confiance à 95% pour $\hat{\mu}$.

Exercice 2.4

Un pépiniériste souhaite estimer la taille moyenne de ses arbustes d'une même variété. Sur les 10000 plantes de la serre, on en sélectionne 200 par sondage aléatoire simple, puis on mesure la hauteur de chacune de ces plantes. Les résultats sont les suivants (en m) :

$$\sum_{k=1}^{200} x_k = 248, \quad \sum_{k=1}^{200} x_k^2 = 331.$$

1. Donner un intervalle de confiance à 95% pour la taille moyenne des arbustes.
2. Le pépiniériste a de bonnes raisons de penser que l'écart-type calculé sur la population de tous les arbustes se situe entre 0.25 et 0.45 m . En négligeant le taux de sondage, quelle taille d'échantillon doit-on retenir pour donner un intervalle de confiance à 95% ayant une demi-longueur d'au plus 2 cm ?

Exercice 2.5

On souhaite estimer la quantité d'eau moyenne (exprimée en m^3) consommée annuellement par les habitants d'une ville donnée de 100 000 habitants. On sélectionne par un plan simple un échantillon de 250 habitants. Les résultats obtenus sont les suivants :

$$\sum_{i=1}^n x_i = 15\,125 \quad \sum_{i=1}^n x_i^2 = 921\,310.$$

1. Traduire en quelques mots l'information contenue dans la formule : $\sum_{i=1}^n x_i = 15\,125$.
2. Donner un intervalle de confiance à 95% pour la quantité d'eau moyenne consommée annuellement par les habitants de cette ville.
3. On s'intéresse maintenant à la quantité totale consommée annuellement par l'ensemble des habitants de la ville. Donner une estimation, puis un intervalle de confiance à 95% pour cette quantité totale.

Exercice 2.6

Dans une région qui possède 250 hôtels, on souhaite estimer la proportion d'hôtels deux étoiles qui ont un parking. On sélectionne par plan simple 50 hôtels deux étoiles de la région. Parmi les 50 hôtels de l'échantillon, 34 possèdent un parking. Donner une estimation par intervalle de confiance à 95% de la proportion d'hôtels deux étoiles de la région possédant un parking. Même question avec un intervalle de confiance à 90%.

Exercice 2.7

Quelle taille d'échantillon doit-on retenir, si on choisit un sondage aléatoire simple, pour donner un intervalle de confiance à 95% ayant une demi-longueur d'au plus 2% pour la proportion de parisiens qui portent des lunettes ?

Indications

1. La taille de la population de la ville de Paris étant très grande, on suppose que le taux de sondage est négligeable.

2. N'ayant manifestement aucune indication *a priori* sur la proportion recherchée, on se place dans le cas le plus défavorable qui conduit à une taille d'échantillon maximale (taille "de précaution"). Montrer que cette taille maximale correspond au cas où la vraie proportion dans population p est égale à 50% (indication : étudier les variations de la fonction $f(p) = p(1-p)$ sur l'intervalle $[0, 1]$. Montrer qu'elle prend son maximum pour $p = 50\%$)
3. Trouver la taille d'échantillon recherchée.

Exercice 2.8

On souhaite réaliser un sondage d'opinion dans le but d'estimer la proportion p d'individus qui ont une opinion favorable d'une certaine personnalité politique. On suppose que la taille de la population est très grande, ce qui nous conduit à négliger le taux de sondage. En admettant que l'on utilise un sondage aléatoire simple, combien de personnes doit-on interroger pour que l'on puisse donner un intervalle de confiance à 95% pour p ayant une demi-longueur d'au plus 0.02 ?

Indication : en l'absence d'informations complémentaires, on peut utiliser "l'intervalle de précaution" consistant à considérer la plus grande demi-longueur possible (c'est-à-dire le pire des cas).

Chapitre 3

Sondages stratifiés

3.1 Principe et justification

Dans un sondage aléatoire simple, tous les échantillons d'une population de taille N sont possibles avec la même probabilité. On imagine que certains d'entre eux puissent s'avérer *a priori* indésirables. Dans le cas de l'exemple 2.3, nous disposons de 5 jetons : -1, 2, 4, 10 et 20 dont nous souhaitons évaluer la moyenne ($\mu = 7$) à l'aide d'un échantillon de taille 2. Parmi les échantillons à deux unités, on trouve les cas extrêmes $\{-1, 2\}$ et $\{10, 20\}$, qui sont particulièrement "mauvais".

Plus concrètement, dans l'étude du lancement d'un nouveau produit financier, on peut supposer des différences de comportement entre les "petits" et les "gros" clients de la banque. Il serait malencontreux que les hasards de l'échantillonnage conduisent à n'interroger que les clients appartenant à une seule de ces catégories, ou simplement que l'échantillon soit trop déséquilibré en faveur de l'une d'elles. S'il existe dans la base de sondage une information auxiliaire permettant de distinguer, *a priori*, les catégories de petits et gros clients, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque sous-population. C'est le principe de la **stratification** : découper la population en sous-ensembles appelés *strates* et réaliser un sondage dans chacune d'elles.

L'intérêt de cette méthode, en comparaison des plans simples, est qu'elle permet d'améliorer la précision des estimateurs. Elle nécessite l'utilisation d'une information auxiliaire connue pour l'ensemble de la population.

Exemple 3.1

Reprenons l'exemple initial où nous souhaitons estimer l'âge moyen de toutes les personnes évoluant sur le site de Rennes 2. La base de sondage est composée de l'ensemble des personnes de Rennes 2. Supposons que nous disposions de la répartition des éléments de la base suivant les catégories :

- étudiants ;
- enseignants ;
- IATOS.

Dit autrement nous connaissons la répartition des personnes de Rennes 2 suivant ces 3 catégories (voir Figure 3.1). Il y a fort à parier que la variable âge ne se comporte pas de la même manière dans ces trois classes ("en moyenne", on peut en effet penser que la

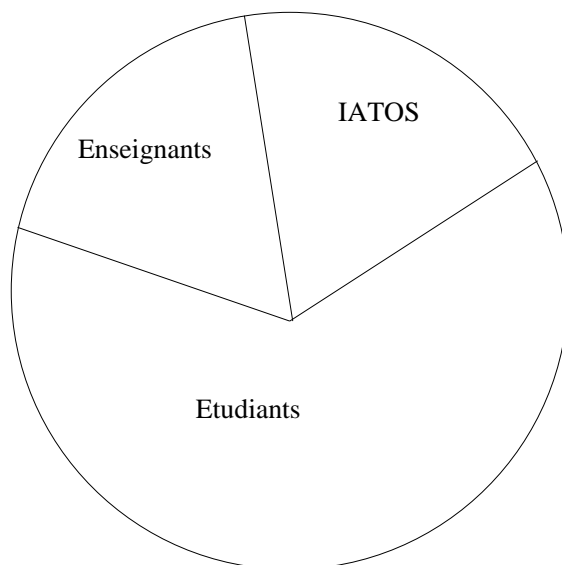


FIGURE 3.1 – Exemple de répartition des personnels de Rennes 2.

population enseignant ou IATOS est plus âgée que la population étudiante). Il paraît dès lors pertinent d'essayer de prendre en compte cette information dans le plan de sondage.

La répartition des personnes de Rennes 2 fournit une information auxiliaire à notre problématique. L'objectif principal consiste donc à mettre à profit cette information pour obtenir des résultats précis. L'information auxiliaire peut être utilisée à deux moments :

- à l'étape de la conception du plan de sondage ;
- à l'étape de l'estimation des paramètres.

Dans ce chapitre, nous utiliserons cette information uniquement pour bâtir le plan de sondage.

3.2 Plan de sondage stratifié

Nous précisons maintenant quelques notations utiles à la définition d'un plan stratifié.

Rappel du contexte : on note N le nombre d'individus dans la population. On souhaite évaluer une caractéristique de la population. On note X_i la valeur de ce caractère mesurée sur le $i^{\text{ème}}$ individu. On cherche estimer la moyenne du caractère sur la population

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i.$$

Dans ce chapitre, nous nous restreindrons à l'estimation de la moyenne. Cependant tout comme dans le chapitre précédent, tous les concepts s'étendent facilement à l'estimation d'un total ou d'une proportion.

On suppose que la population \mathcal{P} est partagée en H sous-ensembles ou strates notées \mathcal{P}_h , $h = 1, \dots, H$. On définit :

- taille de la strate h : N_h ;
- moyenne de la strate h : $\mu_h = \frac{1}{N_h} \sum_{i \in \mathcal{P}_h} X_i$.
- variance de la strate : $\sigma_h^2 = \frac{1}{N_h} \sum_{i \in \mathcal{P}_h} (X_i - \mu_h)^2$;
- variance corrigée de la strate h : $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in \mathcal{P}_h} (X_i - \mu_h)^2 = \frac{N_h}{N_h - 1} \sigma_h^2$.

Proposition 3.1

1. Réécriture de μ :

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \frac{1}{N} \sum_{i=1}^N N_h \mu_h.$$

2. Réécriture de σ^2 :

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{N} \sum_{h=1}^H N_h \sigma_h^2 + \frac{1}{N} \sum_{h=1}^H N_h (\mu_h - \mu)^2 \\ &= \text{Variance intra-strate} + \text{Variance inter-strate}. \end{aligned}$$

Le premier terme représente la moyenne des variances des strates. Le second est dû aux différences entre strates : si par exemple l'échantillon est stratifié entre étudiant, enseignant, IATOS, ce terme représente le contraste d'âge entre ces différentes catégories.

Nous sommes maintenant en mesure de définir un plan stratifié.

Définition 3.1

Un plan de sondage est dit **stratifié** si dans chaque strate on sélectionne un échantillon aléatoire de taille fixe n_h et que les sélections sont réalisées **indépendamment** d'une strate à une autre. On suppose en outre dans ce cours qu'au sein de chaque strate les plans sont simples et sans remise.

Les n_h doivent vérifier $\sum_{h=1}^H n_h = n$.

Exemple 3.2

Reprenons l'exemple de la stratification de la "population" Rennes 2 suivant : étudiant, enseignant, IATOS. Pour simplifier à l'extrême, supposons que la population est composée de $N = 20$ individus :

- 10 étudiants (strate 1, $N_1 = 10$) ;
- 6 enseignants (strate 2, $N_2 = 6$) ;
- 4 IATOS (strate 3, $N_3 = 4$) ;

La population est donc composée de $N = N_1 + N_2 + N_3 = 20$ individus. On effectue un plan de sondage stratifié : on sélectionne un échantillon aléatoire de taille $n = 10$ de la manière suivante (voir Figure 3.2) :

- $n_1 = 5$ dans la strate 1 ;
- $n_2 = 3$ dans la strate 2 ;
- $n_3 = 2$ dans la strate 3.

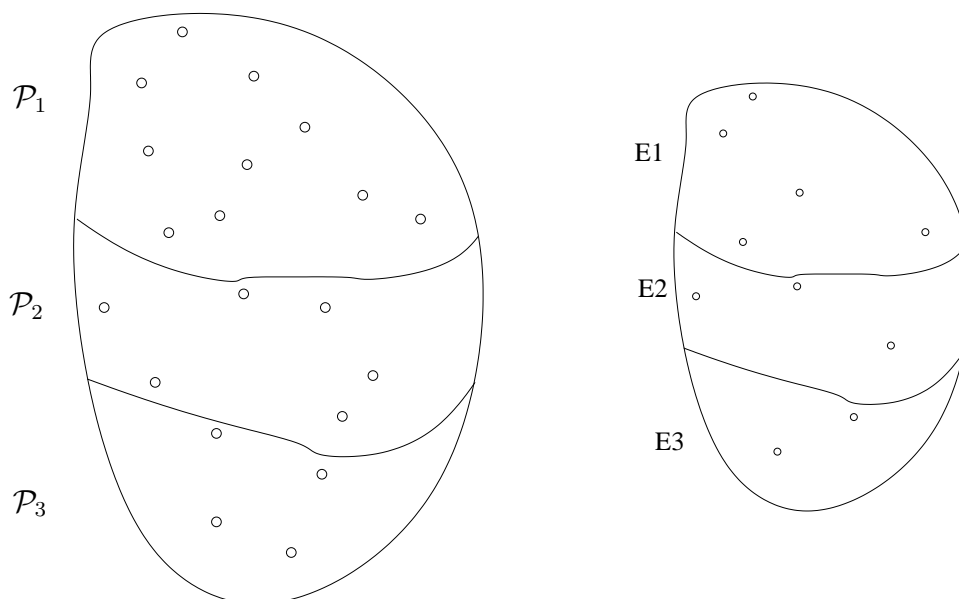


FIGURE 3.2 – Echantillonnage stratifié : à gauche la population, à droite l'échantillon.

3.3 Estimateur de la moyenne

Une fois l'échantillonnage effectué, il se pose bien entendu la question de l'estimateur de la moyenne μ .

3.3.1 Un exemple

Reprenons l'exemple précédent. Pour $i = 1, \dots, n$, on note x_i l'âge du $i^{\text{ème}}$ individu présent dans l'échantillon E . Cet échantillon E est divisé en trois sous-ensembles :

- E_1 contient les étudiants de l'échantillon ;
- E_2 contient les enseignants de l'échantillon ;
- E_3 contient les IATOS de l'échantillon.

On calcule ensuite l'âge moyen des individus de l'échantillon strate par strate :

- $\bar{x}_1 = \sum_{i \in E_1} x_i$: âge moyen des individus de la strate 1 ;
- $\bar{x}_2 = \sum_{i \in E_2} x_i$: âge moyen des individus de la strate 2 ;
- $\bar{x}_3 = \sum_{i \in E_3} x_i$: âge moyen des individus de la strate 3 ;

On rappelle que N_1 est le nombre d'individus présents dans la strate 1 (dans la population entière), par conséquent $N_1\bar{x}_1$ est un estimateur de l'âge total de la population étudiante (strate 1). De même $N_2\bar{x}_2$ est un estimateur de l'âge total de la population enseignante (strate 2) et $N_3\bar{x}_3$ est un estimateur de l'âge total de la population IATOS (strate 3). Par conséquent :

$$\sum_{i=1}^3 N_i \bar{x}_i = N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3$$

est un estimateur de l'âge total de la population. Pour obtenir un estimateur de l'âge moyen μ il suffit donc de diviser par le nombre d'individus dans la population. L'estimateur $\hat{\mu}$ est

donc

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^3 N_i \bar{x}_i = \frac{1}{N} (N_1 \bar{x}_1 + N_2 \bar{x}_2 + N_3 \bar{x}_3).$$

Application numérique : les résultats du sondage sont donnés dans le tableau suivant :

Strate	1	2	1	3	1	1	2	3	2	1
Age	20	50	25	42	23	22	35	44	38	26

TABLE 3.1 – Age des individus sondés.

On calcule la moyenne des âges des individus de l'échantillon par strate :

$$\bar{x}_1 = 23.2, \quad \bar{x}_2 = 42, \quad \bar{x}_3 = 44.$$

Une estimation de μ est donc :

$$\hat{\mu} = \frac{1}{20} (10 \times 23.2 + 6 \times 42 + 4 \times 44) = 33.$$

3.3.2 Cas général

Nous pouvons maintenant définir l'estimateur $\hat{\mu}$ dans un contexte général pour un plan stratifié. Pour chaque strate h , on note \bar{x}_h la moyenne calculée sur l'échantillon issu de la strate h :

$$\bar{x}_h = \frac{1}{n_h} \sum_{i \in E_h} x_i.$$

L'estimateur $\hat{\mu}$ s'écrit alors :

$$\hat{\mu} = \frac{1}{N} \sum_{h=1}^H N_h \bar{x}_h. \quad (3.1)$$

Le tableau 3.3.2 récapitule les notations relatives à la population et à l'échantillon.

Comme pour le plan simple, on étudie la précision de l'estimateur (et donc du sondage) en étudiant son biais et sa variance. On a le résultat suivant.

Théorème 3.1

Soit $\hat{\mu}$ l'estimateur de la moyenne pour un plan stratifié (défini par (3.1)). On a :

- $\mathbf{E}(\hat{\mu}) = \mu$: $\hat{\mu}$ est un estimateur sans biais de μ ;
- La variance de $\hat{\mu}$ est donnée par :

$$\mathbf{V}(\hat{\mu}) = \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2. \quad (3.2)$$

3.4 Répartition de l'échantillon

Jusqu'à présent, dans le plan stratifié, nous avons supposé que les tailles d'échantillons n_h étaient fixés pour chaque strate. En pratique, lors de la planification du sondage, le statisticien doit se poser la question suivante : combien de personnes dois-je sonder par strate pour que mon estimateur soit le plus précis possible ? Dit autrement, comment choisir les n_h ?

		Population \mathcal{P} <u>inconnu, déterministe</u>	Echantillon E <u>connu, aléatoire</u>
Totale	Taille	N	n
	Moyenne	μ	\bar{x}
	Variance	σ^2	
	Variance Corrigée	S^2	s^2
Strate	Taille	N_h	n_h
	Moyenne	μ_h	\bar{x}_h
	Variance	σ_h^2	
	Variance Corrigée	S_h^2	s_h^2

TABLE 3.2 – Notations pour le plan stratifié.

3.4.1 Plan avec allocation proportionnelle

Pour décider des effectifs d'échantillon n_h , la solution la plus simple, et de très loin la plus utilisée, est de les établir au prorata des tailles N_h , ce qui peut s'exprimer de deux façons équivalentes :

- les strates ont dans l'échantillon des poids n_h/n égaux à leurs poids N_h/N dans la population ;
- on applique le même taux de sondage dans toutes les strates : $f_h = n_h/N = n/N = f$.

Pour l'exemple de l'âge moyen de la population "Rennes 2", un tel plan signifie que les proportions de chaque strate dans la population sont les mêmes que dans l'échantillon. Si on a par exemple la répartition suivante :

Strate	N_h
Etudiant	6000
Enseignant	2500
IATOS	1500

Alors un plan stratifié avec allocation proportionnelle de taille $n = 100$ consistera à sonder :

- $n_1 = 60$ étudiants ;
- $n_2 = 25$ enseignants ;
- $n_3 = 15$ IATOS.

Définition 3.2

Dans un plan stratifié avec allocation proportionnelle, on choisit les n_h de telle sorte que la proportion d'individus provenant de la strate h dans l'échantillon soit la même que dans la population, c'est-à-dire :

$$\frac{n_h}{n} = \frac{N_h}{N},$$

d'où

$$n_h = n \frac{N_h}{N}.$$

Attention : Cette procédure ne donne généralement pas de résultat entier. Il faut alors recourir à une procédure d'arrondi (et vérifier que l'on a toujours $\sum_{h=1}^H n_h = n$).

Proposition 3.2

Soit $\hat{\mu}$ l'estimateur construit pour un plan avec allocation proportionnelle. On a :

$$\mathbf{V}(\hat{\mu}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^H N_h S_h^2. \quad (3.3)$$

Remarque

Dans le cas d'un plan avec allocation proportionnelle on aura le choix entre cette formule et (3.2) pour calculer la variance de l'estimateur $\hat{\mu}$.

Si les tailles N_h de chaque strate h sont grandes, on a $S_h^2 \simeq \sigma_h^2$. On peut donc écrire d'après (3.3) :

$$\mathbf{V}(\hat{\mu}) \simeq \frac{1}{n} \left(1 - \frac{n}{N}\right) \sigma_{\text{intra}}^2.$$

Dans le cas d'un plan simple (chapitre précédent), si N est grand, on rappelle que :

$$\mathbf{V}(\hat{\mu}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \sigma^2.$$

On a donc remplacé, grâce à la stratification le terme σ^2 intervenant dans la variance de l'estimateur par le terme σ_{intra}^2 . Comme

$$\sigma_{\text{intra}}^2 \leq \sigma^2,$$

on en déduit que la stratification avec allocation proportionnelle donne presque toujours de meilleurs résultats qu'un plan simple puisque l'on supprime la variance inter-strate dans l'expression de la variance de l'estimateur. Les résultats seront d'autant plus satisfaisants lorsque la variance inter-strate est grande. Celle ci est grande quand la variable de stratification est fortement liée à la variable d'intérêt. C'est pourquoi il faut toujours stratifier avec une variable très dépendante de la variable d'intérêt.

Exemple 3.3

On donne dans le tableau pour chaque individu de Rennes 2 :

- son âge ;
- sa catégorie : 1 si étudiant, 2 si enseignant, 3 si IATOS ;
- sa couleur de cheveux : a si brun, b si blond, c si châtain.

Pour simplifier les calculs, on considère une population de 20 individus.

Age	Cat	Che	Age	Cat	Che
24	1	c	22	1	c
52	2	a	48	2	a
42	3	b	24	1	a
19	1	c	38	3	a
38	3	a	26	1	b
26	1	b	36	3	b
45	2	c	46	2	b
23	1	a	23	1	c
39	2	a	39	2	a
24	1	b	18	1	c

1. On souhaite estimer la moyenne μ à l'aide d'un plan simple. Quel est la variance de l'estimateur ?

D'après le chapitre précédent

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = \left(1 - \frac{10}{20}\right) \frac{115.305}{n} = 5.77.$$

2. On désire stratifier la population suivant la catégorie. Quelle est la variance de l'estimateur $\hat{\mu}$ pour un tel plan ?

La population est divisée selon la Figure 3.3.

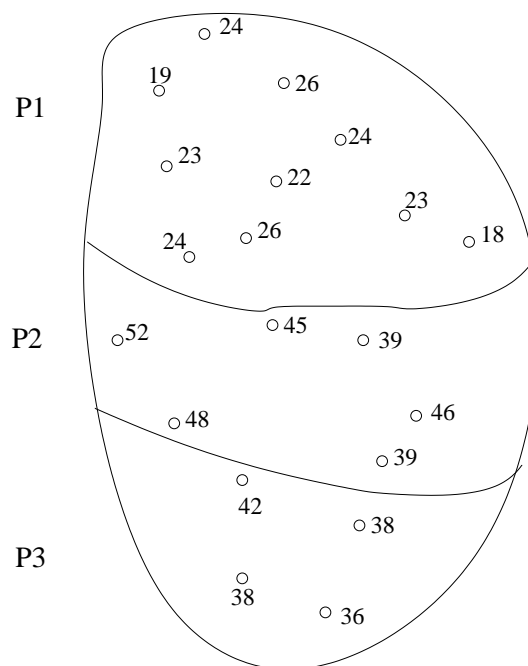


FIGURE 3.3 – Population divisée suivant la catégorie.

Calculons les moyennes et variances corrigées par strate :

- $\mu_1 = 22.9$, $S_1^2 = 6.99$;
- $\mu_2 = 44.83$, $S_2^2 = 26.17$;

- $\mu_3 = 38.5$, $S_3^2 = 6.33$.

On en déduit la variance de l'estimateur à l'aide de la formule (3.3) :

$$\begin{aligned} \mathbf{V}(\hat{\mu}) &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^H N_h S_h^2 \\ &= \frac{1}{10} \left(1 - \frac{10}{20}\right) \frac{1}{20} [10 * 6.99 + 6 * 26.17 + 4 * 6.33] = 0.63. \end{aligned}$$

On peut également retrouver ce résultat avec la formule (3.2).

3. On choisit maintenant de stratifier suivant la couleur des cheveux. Quelle est la variance de l'estimateur pour un tel plan ?

Dans ce cas, la population est divisée selon la Figure 3.4.

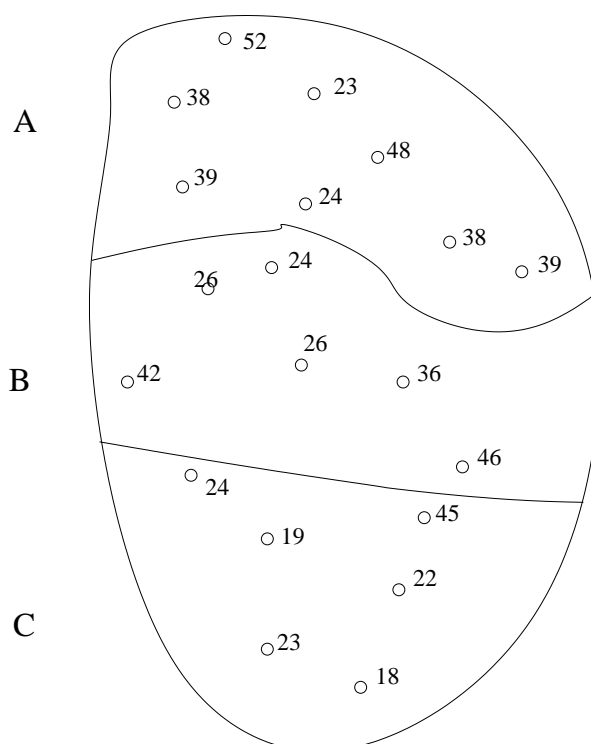


FIGURE 3.4 – Population stratifiée suivant la couleur des cheveux.

Par un raisonnement similaire à celui de la question précédente on peut montrer que la variance de l'estimateur vaut 4.86 pour ce plan de sondage.

Le tableau suivant récapitule les résultats :

Plan	$\mathbf{V}(\hat{\mu})$
simple	5.77
Strat Cat	0.63
Strat Che	4.86

On voit que les deux plans stratifiés possèdent des variances inférieures au plan simple. Le gain de la stratification par la catégorie est significatif comparé à celui de la couleur des cheveux. Ceci vient du fait que la variable d'intérêt (âge) dépend plus de la catégorie que de la couleur de cheveux. Il sera donc beaucoup plus pertinent de stratifier par rapport à la catégorie que par rapport à la couleur de cheveux (on pouvait s'y attendre...)

Nous avons vu qu'en terme de variance de l'estimateur, le plan avec allocation proportionnelle est plus précis que le plan simple. *Peut-on faire encore mieux?*

3.4.2 Plan avec allocation optimale

La réponse à la question précédente est : oui, si l'on sait *a priori* que certaines classes sont beaucoup plus homogènes que d'autres. Intuitivement, on a intérêt à sous-échantillonner les premières pour consacrer plus de moyens aux secondes.

Définition 3.3

Dans un plan stratifié avec **allocation optimale**, on choisit les tailles d'échantillons n_1, \dots, n_H telles que $\sum_{h=1}^H n_h = n$ et telles que la variance de l'estimateur $\mathbf{V}(\hat{\mu})$ soit minimale. La solution de ce problème est

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}.$$

Par définition, l'estimateur construit avec un plan d'allocation optimale possède la plus petite variance possible (parmi tous les plans stratifiés). Le prix à payer est que pour construire un tel estimateur (pour choisir les tailles d'échantillons dans chaque strate), il nous faut connaître la variance corrigée du caractère dans chaque strate de la population.

La variance de l'estimateur associé à ce plan est toujours donnée par (3.2). On ne peut par contre pas utiliser la formule (3.3) qui est valable **uniquement** pour un plan avec allocation proportionnelle.

Remarque

1. Là encore, les n_h ne sont pas nécessairement entiers, il faut recourir à une procédure d'arrondi. De plus la formule précédente peut parfois conduire à des choix de n_h tels que $n_h > N_h$. Dans ce cas, on fait un recensement dans les strates où le problème se pose et on recalcule les valeurs de n_h pour les strates restantes.
2. La formule précédente nécessite de connaître les variances corrigées de chaque strate S_h (ou plutôt leurs racines carrées). En pratique, il faut donc les estimer. En sondage, on utilise souvent les résultats d'enquêtes précédentes.

Pour les estimateurs construits par plans stratifiés, on peut calculer des intervalles de confiance comme pour les plans simples. Un intervalle de confiance de niveau $1 - \alpha$ est donné par

$$IC = \left[\hat{\mu} - z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{\mu})}; \hat{\mu} + z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{\mu})} \right],$$

où $z_{1-\alpha/2}$ désigne le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite. Nous terminons par un exemple sur les plans stratifiés, nous rappelons que tout ce qui a été vu dans ce chapitre peut s'adapter à l'estimation d'un total ou d'une proportion.

Exemple 3.4

Une grande entreprise veut réaliser une enquête auprès de son personnel qui comprend 10000 personnes. Elle s'intéresse à l'évolution de l'âge de ses employés et souhaite commencer par estimer l'âge moyen. Des études préliminaires ont montré que la variable que l'on cherche à analyser est très contrastée 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 trois grandes catégories qui formeront les strates. On va donc proposer des plans d'échantillonnage, on dispose des renseignements suivants :

Catégories	Effectifs	Ecart-type des âges
1	2000	18
2	3000	12
3	5000	3.6
Ensemble	10000	16

On désire estimer l'âge moyen noté μ à partir d'un échantillon de $n = 100$ personnes.

1. On réalise d'abord un plan simple, proposer un estimateur de μ et calculer sa variance.
2. Un sondage stratifié est ensuite envisagé. Proposer un estimateur pour μ . Quels effectifs doit-on sélectionner dans chaque strate si on réalise un plan avec allocation proportionnelle. Calculer la variance de l'estimateur construit.
3. Reprendre la question précédente pour un plan avec allocation optimale.

Eléments de correction :

1. $n = 100$, on note x_i , $i = 1, \dots, n$ l'âge de la $i^{\text{ème}}$ personne interrogée. L'estimateur de μ est

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

La variance d'un tel estimateur est donnée par

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}.$$

Ici S^2 est inconnu mais on connaît σ^2 , donc

$$S^2 = \frac{N}{N-1} \sigma^2 = \frac{10000}{9999} 16^2 = 256.03.$$

On déduit

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{100}{10000}\right) \frac{256.03}{100} = 2.53.$$

2. Plan stratifié : soit n_h , $h = 1, 2, 3$ le nombre de personnes interrogées dans chaque strate. L'estimateur est donné par :

$$\hat{\mu} = \frac{1}{N} \sum_{h=1}^H N_h \bar{x}_h,$$

où \bar{x}_h est l'âge moyen des personnes interrogées dans la strate h . Pour un plan avec allocation proportionnelle, les effectifs sont choisis suivant :

$$n_h = n \frac{N_h}{N}.$$

Par conséquent,

$$n_1 = 100 \times \frac{2000}{10000} = 20, \quad n_2 = 100 \times \frac{3000}{10000} = 30, \quad n_3 = 100 \times \frac{5000}{10000} = 50.$$

Calculons les variances corrigées par strate $S_h^2 = \frac{N}{N-1} \sigma_h^2$:

$$S_1^2 = \frac{10000}{9999} 18^2 = 324.03, \quad S_2^2 = \frac{10000}{9999} 12^2 = 124.01, \quad S_3^2 = \frac{10000}{9999} 3.6^2 = 12.96.$$

La variance de l'estimateur est donnée par (3.2) ou (3.3) :

$$\begin{aligned} \mathbf{V}(\hat{\mu}) &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^H N_h S_h^2 \\ &= \frac{1}{100} \left(1 - \frac{100}{10000}\right) \frac{1}{10000} \left[2000 \times 324.03 + 3000 \times 124.01 + 5000 \times 12.96\right] \\ &= 1.10 \end{aligned}$$

3. Pour un plan avec allocation optimale, les effectifs sont choisis suivant :

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}.$$

On calcule

$$\sum_{h=1}^H N_h S_h = 2000 \times \sqrt{324.03} + 3000 \times \sqrt{124.01} + 5000 \times \sqrt{12.96} = 87409.6$$

On déduit

$$\begin{aligned} n_1 &= 100 \times \frac{2000 \times \sqrt{324.03}}{87409.6} = 41.18, \quad n_2 = 100 \times \frac{3000 \times \sqrt{124.01}}{87409.6} = 38.22, \\ n_3 &= 100 \times \frac{5000 \times \sqrt{12.96}}{87409.6} = 20.59. \end{aligned}$$

On arrondit

$$n_1 = 41, \quad n_2 = 38, \quad n_3 = 21$$

en vérifiant que la somme fait bien 100. On peut maintenant calculer la variance à l'aide de la formule (3.2)

$$\begin{aligned} \mathbf{V}(\hat{\mu}) &= \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2 \\ &= \frac{1}{10000^2} \left[2000 \frac{2000 - 41}{41} 324.03 + 3000 \frac{3000 - 38}{38} 124.01 + 5000 \frac{5000 - 21}{21} 12.96\right] \\ &= 0.75. \end{aligned}$$

3.5 Exercices

Exercice 3.1

Soit une population $\mathcal{P} = \{1, 2, 3, 4\}$ et $X_1 = X_2 = 0$, $X_3 = 1$, $X_4 = -1$ les valeurs prises par la variable à laquelle on s'intéresse.

1. Calculer la variance de l'estimateur de la moyenne pour un plan aléatoire simple sans remise de taille $n = 2$.
2. Calculer la variance de l'estimateur de la moyenne pour un plan aléatoire stratifié pour lequel une seule unité est prélevée par strate, les strates étant données par :

$$E_1 = \{1, 2\}, \quad E_2 = \{3, 4\}.$$

Exercice 3.2

Dans une grande ville, on s'intéresse au nombre moyen de clients que peut avoir un médecin pendant une journée de travail. On part de l'idée a priori que plus le médecin a d'expérience, plus il a de clients. On classe donc la population de médecins en trois groupes : les "débutants" (classe 1), les "confirmés" (classe 2), et les "très expérimentés" (classe 3). Par ailleurs, on suppose que l'on connaît, dans la base de sondage des médecins, la classe de chacun d'entre eux. On tire par sondage aléatoire simple 200 médecins dans chaque classe. On obtient les résultats suivants :

	h=1	h=2	h=3
\bar{x}_h	10	15	20
s_h^2	4	7	10
N_h	500	1000	2500

1. Comment s'appelle ce plan de sondage ?
2. Comment estimez vous le nombre moyen de clients soignés par jour et par médecin ?
3. Donner un intervalle de confiance à 95% pour le vrai nombre moyen de clients soignés par jour et par médecin.
4. Si vous n'aviez comme contrainte que le nombre total de médecin à enquêter (soit 600), procéderiez-vous comme ci-dessus ?

Exercice 3.3

Un directeur de cirque possède 100 éléphants classés en deux catégories : les mâles et les femelles. Le directeur veut estimer le poids total de son troupeau car il veut traverser un fleuve en bateau. Cependant, l'année précédente, le directeur de cirque avait fait peser tous les éléphants de son troupeau et avait obtenu les résultats suivants (les moyennes sont exprimées en tonnes) :

	Effectif N_h	Moyenne μ_h	S_h^2
Mâles	60	6	4
Femelles	40	4	2.25

1. Calculer σ^2 et S^2 pour l'année précédente.

2. Le directeur suppose désormais que les dispersions de poids n'évoluent pas sensiblement d'une année sur l'autre (ce type d'hypothèse reste ici très raisonnable et se rencontre couramment en pratique quand on répète des enquêtes dans le temps). Si le directeur procède à un tirage aléatoire simple de 10 éléphants, quelle est la variance de l'estimateur du poids total du troupeau ?
3. Si le directeur procède à un tirage stratifié avec allocation proportionnelle de 10 éléphants, quelles tailles d'échantillon doit-on retenir dans chaque strate ? Quelle est alors la variance de l'estimateur du poids total du troupeau ?
4. Si le directeur procède à un tirage stratifié optimal de 10 éléphants, quelles tailles d'échantillon doit-on retenir dans chaque strate ? Quelle est alors la variance de l'estimateur du poids total du troupeau ?

Exercice 3.4

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 faible revenu (strate 1), individus de revenu moyen (strate 2), individus de revenu élevé (strate 3). On note \bar{p}_h la proportion d'individus possédant au moins un véhicule dans l'échantillon issu de la strate h . Les résultats obtenus sont les suivants :

	h=1	h=2	h=3
N_h	3500	2000	2000
n_h	500	300	200
\bar{p}_h	0.13	0.45	0.50

1. Quel estimateur \hat{p} de p proposez-vous ?
2. Donner un intervalle de confiance à 95% pour p .

indications : dans le cas d'une proportion, on peut estimer la variance corrigée S_h^2 par $s_h^2 = \frac{N}{N-1} \bar{p}_h (1 - \bar{p}_h)$.

Exercice 3.5

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

Strate	N_h	\bar{x}_h	S_h^2	n_h
Moins de 40 ans	5000	25	16	40
De 40 à 50 ans	3000	45	10	20
Plus de 50 ans	2000	58	20	40

1. Quelle est la valeur de l'estimateur stratifié de l'âge moyen μ ?
2. Calculer la variance de cet estimateur.
3. Quelles tailles d'échantillons n_h doit-on choisir pour chaque strate si on souhaite réaliser une allocation proportionnelle afin de constituer un échantillon de $n = 100$ individus ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
4. On souhaite maintenant réaliser une allocation optimale (toujours avec $n = 100$). Calculer alors la valeur des n_h ainsi que la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
5. Parmi les trois plans de sondage proposés, lequel vous semble le plus approprié ?

Exercice 3.6

La variable d'intérêt est ici le chiffre d'affaire moyen réalisé par un ensemble de 1060 entreprises. Celles-ci étant de tailles très différents, on a constitué cinq strates en fonction du nombre de salariés dans chaque entreprise. De plus, grâce à une enquête précédente, on

Nombre de salariés	0 à 9	10 à 19	20 à 29	50 à 499	500 et plus
Nombre d'entreprises	500	300	150	100	10

dispose d'estimations pour les variances corrigées S_h^2 de chaque strate. On considère donc que :

$$S_1^1 = 1.5, \quad S_2^2 = 4, \quad S_3^3 = 8, \quad S_4^4 = 100, \quad S_5^5 = 2500.$$

1. A l'intérieur de chaque strate, on réalise un sondage aléatoire simple avec les tailles d'échantillon suivantes :

$$n_1 = 130, \quad n_2 = 80, \quad n_3 = 60, \quad n_4 = 25, \quad n_5 = 5.$$

Les résultats sont les suivants :

$$\bar{x}_1 = 5, \quad \bar{x}_2 = 12, \quad \bar{x}_3 = 30, \quad \bar{x}_4 = 150, \quad \bar{x}_5 = 600.$$

Donner un intervalle de confiance à 90% pour le chiffre d'affaire moyen.

2. En conservant toujours la même taille globale d'échantillon, quels effectifs d'échantillon faut-il prendre dans chaque strate
 - (a) pour une allocation proportionnelle ?
 - (b) pour une allocation optimale ?
3. Calculer les variances de l'estimateur pour le plan avec allocation proportionnelle puis pour le plan avec allocation optimale.

Annexe A

Intervalle de confiance pour une moyenne dans un plan de sondage aléatoire simple

Théorème A.1 (Théorème central limite)

Soit x_1, \dots, x_n une suite de n variables aléatoires i.i.d telles que $\mathbf{E}(x_i) = \mu$. Soit $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ la moyenne empirique des x_i . Alors on peut approcher la loi de \bar{x} par la loi normale $\mathcal{N}(\mu, \mathbf{V}(\bar{x}))$. Ou encore, on peut approcher la loi de la variable aléatoire

$$Z = \frac{\bar{x} - \mu}{\sqrt{\mathbf{V}(\bar{x})}}$$

par la loi $\mathcal{N}(0, 1)$.

On appellera *intervalle de confiance* pour μ de niveau $1 - \alpha$ un intervalle aléatoire $[\bar{x} - h, \bar{x} + h]$ tel que

$$\mathbf{P}([\bar{x} - h, \bar{x} + h] \ni \mu) = 1 - \alpha.$$

Calculons un intervalle de confiance pour μ de niveau $1 - \alpha$. On remarque que :

$$\begin{aligned} \mathbf{P}([\bar{x} - h, \bar{x} + h] \ni \mu) &= \mathbf{P}(\bar{x} - h \leq \mu \leq \bar{x} + h) \\ &= \mathbf{P}(-h \leq \mu - \bar{x} \leq h) \\ &= \mathbf{P}(-h \leq \bar{x} - \mu \leq h) \\ &= \mathbf{P}\left(-\frac{h}{\sqrt{\mathbf{V}(\bar{x})}} \leq \frac{\bar{x} - \mu}{\sqrt{\mathbf{V}(\bar{x})}} \leq \frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right). \end{aligned}$$

Il suffit donc de trouver h tel que

$$\mathbf{P}\left(-\frac{h}{\sqrt{\mathbf{V}(\bar{x})}} \leq \frac{\bar{x} - \mu}{\sqrt{\mathbf{V}(\bar{x})}} \leq \frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right) = 1 - \alpha.$$

En notant F la fonction de répartition de la loi $\mathcal{N}(0, 1)$, on a donc

$$F\left(\frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right) - F\left(-\frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right) = 1 - \alpha$$

$$2F\left(\frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right) - 1 = 1 - \alpha$$

$$F\left(\frac{h}{\sqrt{\mathbf{V}(\bar{x})}}\right) = 1 - \frac{\alpha}{2}.$$

Avec $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, on obtient $\frac{h}{\sqrt{\mathbf{V}(\bar{x})}} = z_{1-\alpha/2}$ et donc

$$h = z_{1-\alpha/2} \sqrt{\mathbf{V}(\bar{x})}.$$

Un intervalle de confiance de niveau $1 - \alpha$ est donc donnée par

$$\left[\bar{x} - z_{1-\alpha/2} \sqrt{\mathbf{V}(\bar{x})}, \bar{x} + z_{1-\alpha/2} \sqrt{\mathbf{V}(\bar{x})} \right]$$

avec

$$\mathbf{V}(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

pour un plan de sondage aléatoire simple. L'IC de niveau $1 - \alpha$ s'écrit alors

$$\left[\bar{x} - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}, \bar{x} + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \right].$$

Annexe B

Correction des exercices

Exercice B.1

La population est composée de $N = 4$ individus.

1. **Moyenne :**

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \frac{1}{4}(11 + 10 + 8 + 11) = 10.$$

Variance :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - \mu^2 = \frac{1}{4}(11^2 + 10^2 + 8^2 + 11^2) - 10^2 = 1.5.$$

Variance corrigée :

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{3}((11 - 10)^2 + (10 - 10)^2 + (8 - 10)^2 + (11 - 10)^2) = 2.$$

On peut également utiliser la formule

$$S^2 = \frac{N}{N-1} \sigma^2.$$

2. (a) On effectue un sondage aléatoire simple sans remise, il y a donc $C_N^n = C_4^2$ échantillons possibles, soit :

$$C_4^2 = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = 6.$$

(b)

Ech	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
\bar{x}	10.5	9.5	11	9	10.5	9.5
s^2	0.5	4.5	0	2	0.5	4.5

3. $\mathbf{E}(\bar{x})$ est la moyenne des valeurs de \bar{x} sur tous les échantillons possibles :

$$\mathbf{E}(\bar{x}) = \frac{1}{6}(10.5 + 9.5 + 11 + 9 + 10.5 + 9.5) = 10,$$

on retrouve ici que \bar{x} est un estimateur sans biais de μ (Théorème 1 du cours) : $\mathbf{E}(\bar{x}) = \mu = 10$.

$$\mathbf{V}(\bar{x}) = \frac{1}{6}((10.5-10)^2 + (9.5-10)^2 + (11-10)^2 + (9-10)^2 + (10.5-10)^2 + (9.5-10)^2) = \frac{1}{2},$$

on peut aussi calculer $\mathbf{V}(\bar{x})$ à l'aide du théorème 2 :

$$\mathbf{V}(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}.$$

$$\mathbf{E}(s^2) = \frac{1}{6}(0.5 + 4.5 + 0 + 2 + 0.5 + 4.5) = 2.$$

Exercice B.2

1. On n'est pas dans le cas d'un plan de sondage aléatoire simple puisque l'échantillon $\{1, 2\}$ a ici plus de chances d'apparaître que les autres.
2. On note $\mathbf{P}(\{j\})$ la probabilité que l'individu j fasse partie de l'échantillon. L'individu 1 fait partie de l'échantillon si on tire l'échantillon $\{1, 2\}$ ou l'échantillon $\{1, 3\}$, donc

$$\mathbf{P}(\{1\}) = \mathbf{P}(\{1, 2\}) + \mathbf{P}(\{1, 3\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

De même,

$$\mathbf{P}(\{2\}) = \mathbf{P}(\{1, 2\}) + \mathbf{P}(\{2, 3\}) = \frac{3}{4}$$

et

$$\mathbf{P}(\{3\}) = \mathbf{P}(\{1, 3\}) + \mathbf{P}(\{2, 3\}) = \frac{1}{2}.$$

3. Soit \bar{x} la moyenne de l'échantillon prélevé :

Ech	$\{1, 2\}$	$\{1, 3\}$	$\{2, 3\}$
\bar{x}	1.5	2	2.5
Proba	1/2	1/4	1/4

4. \bar{x} est un estimateur de la moyenne $\mu = \frac{1}{3}(1 + 2 + 3) = 2$. On a

$$\mathbf{E}(\bar{x}) = \frac{1}{2}1.5 + \frac{1}{4}2 + \frac{1}{4}2.5 = 1.875 \neq 2.$$

$\mathbf{E}(\bar{x}) \neq \mu$, donc \bar{x} n'est pas un estimateur sans biais de μ ici. Ceci vient du fait qu'on ne réalise pas un plan de sondage aléatoire simple (tous les échantillons n'ont pas la même probabilité d'être tirés).

Exercice B.3

1. On estime la moyenne inconnue μ (moyenne des surfaces cultivées par les 2010 fermes) par la moyenne des surfaces cultivées de l'échantillon :

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{100} \sum_{k=1}^{100} x_k = \frac{2907}{100} = 29.07.$$

2. Un intervalle de confiance de niveau 95% pour μ est donné par :

$$\left[\hat{\mu} - z_{0.975} \sqrt{(1-f) \frac{S^2}{n}}; \hat{\mu} + z_{0.975} \sqrt{(1-f) \frac{S^2}{n}} \right]$$

où

- $z_{0.975}$ est le quantile d'ordre 0.975 de la loi normale $\mathcal{N}(0,1)$, on lit sur la table $z_{0.975} = 1.96$;
- $f = n/N = 100/2010 = 0.05$ est le taux de sondage ;
- S^2 est la variance corrigée des superficies sur toute la population, elle est inconnue ici. On l'estime par la variance corrigée sur l'échantillon :

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{n}{n-1} \left[\frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2 \right] = \frac{100}{99} \left[\frac{154593}{100} - 29.07^2 \right] \simeq 707.94.$$

On obtient après calcul l'intervalle de confiance :

$$[23.99; 34.15].$$

Exercice B.4

1. Par un raisonnement analogue à celui de l'exercice précédent on trouve l'intervalle de confiance :

$$[1.19; 1.29].$$

2. La demi longueur de l'intervalle de confiance de niveau α est (voir poly page 17-18) :

$$z_{1-\alpha/2} \sqrt{(1-f) \frac{S^2}{n}} \simeq z_{1-\alpha/2} \sqrt{\frac{S^2}{n}}$$

car on néglige ici le taux de sondage f . On cherche la taille d'échantillon n de manière à ce que cette demi-longueur ne dépasse pas 2 cm, dit autrement, on cherche n tel que :

$$z_{1-\alpha/2} \sqrt{\frac{S^2}{n}} \leq 0.02 \iff z_{1-\alpha/2}^2 \frac{S^2}{n} \leq 0.02^2 \iff z_{1-\alpha/2}^2 \frac{S^2}{0.02^2} \leq n.$$

Ici $z_{1-\alpha/2}^2 = 1.96^2$ mais la variance corrigée de la population S^2 est inconnue. On sait cependant d'après l'énoncé que la variance de la population σ^2 est comprise entre 0.25^2 et 0.45^2 , comme

$$S^2 = \frac{N}{N-1} \sigma^2$$

on déduit :

$$\frac{N}{N-1}0.25^2 \leq S^2 \leq \frac{N}{N-1}0.45^2 \iff 0.06 \leq S^2 \leq 0.20.$$

Rappel : on cherche n tel que :

$$n \geq 1.96^2 \frac{S^2}{0.02^2} \quad (\text{B.1})$$

et $S^2 \leq 0.20$. Ce qui signifie que dans le pire des cas la variance corrigée vaut 0.20. Si on trouve une taille d'échantillon qui satisfait (B.1) dans le pire des cas, alors cette taille d'échantillon vérifiera toujours (B.1). On cherche donc n qui vérifie (B.1) dans le cas le plus défavorable, c'est-à-dire :

$$n \geq 1920.8.$$

A partir de $n = 1921$, la demi longueur de l'intervalle de confiance est au plus égale à 2cm.

Exercice B.5

1. Le total de la consommation d'eau des 250 habitants de l'échantillon est $15\,125m^3$.
2. Pour calculer l'intervalle de confiance, on procède comme dans les exercices 3 et 4 et on obtient :

$$[59.88; 61.12].$$

3. On note T la somme totale dépensée par tous les habitants de la ville. T est inconnu, on l'estime à l'aide d'un plan de sondage aléatoire simple. On note x_k la somme dépensée par l'habitant numéro k de l'échantillon et $\hat{\mu}$ l'estimateur de la somme moyenne dépensée par les habitants :

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k = \frac{15125}{250} = 60.5.$$

Pour obtenir un estimateur de la somme totale dépensée T il suffit de multiplier la somme moyenne dépensée par les habitants de l'échantillon par le nombre d'habitants, on estime donc T par

$$\hat{T} = N\hat{\mu} = 100\,000 * 60.5 = 6\,050\,000.$$

Par analogie avec la moyenne, un intervalle de confiance de niveau $1 - \alpha$ pour le total est donné par :

$$\left[\hat{T} - z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{T})}; \hat{T} + z_{1-\alpha/2} \sqrt{\mathbf{V}(\hat{T})} \right].$$

Ici $1 - \alpha = 0.95$, donc $z_{1-\alpha/2} = z_{0.975} = 1.96$. Il reste à calculer $\mathbf{V}(\hat{T})$:

$$\mathbf{V}(\hat{T}) = \mathbf{V}(N\hat{\mu}) = N^2\mathbf{V}(\hat{\mu}) = (100\,000)^2\mathbf{V}(\hat{\mu}) = 1\,000\,000\,000 = 10^9,$$

car $\mathbf{V}(\hat{\mu}) = 0.1$ a été calculé à la question précédente. On obtient donc l'intervalle de confiance :

$$\left[6\,050\,000 - 1.96 * \sqrt{10^9}; 6\,050\,000 + 1.96 * \sqrt{10^9} \right] = [5\,988\,019; 6\,111\,981]$$

Exercice B.6

Soit p la proportion inconnue d'hôtels deux étoiles admettant un parking et \hat{p} la proportion d'hôtels deux étoiles **de l'échantillon** admettant un parking. Un intervalle de confiance de niveau $1 - \alpha$ pour p est donné par :

$$\left[\hat{p} - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}} \right]$$

(voir poly page 16). Ici, $n = 50$, $N = 250$ et $\hat{p} = 34/50 = 0.68$.

- Si le niveau de l'intervalle de confiance est 0.95, $\alpha = 0.05$ et $z_{1-\alpha/2} = z_{0.975} = 1.96$, ce qui donne :

$$\left[0.68 - 1.96 \sqrt{(1-0.2) \frac{0.68(1-0.68)}{49}}; 0.68 + 1.96 \sqrt{(1-0.2) \frac{0.68(1-0.68)}{49}} \right] \\ = [0.563; 0.797];$$

- Si le niveau de l'intervalle de confiance est 0.90, $\alpha = 0.1$ et $z_{1-\alpha/2} = z_{0.95} = 1.64$, ce qui donne :

$$\left[0.68 - 1.64 \sqrt{(1-0.2) \frac{0.68(1-0.68)}{49}}; 0.68 + 1.64 \sqrt{(1-0.2) \frac{0.68(1-0.68)}{49}} \right] \\ = [0.582; 0.778];$$

Exercice B.7

Soit p la proportion (inconnue) de parisiens qui portent des lunettes et \hat{p} la proportion de parisiens **de l'échantillon** qui portent des lunettes. En négligeant le taux de sondage, la demi longueur d'un intervalle de confiance de niveau $1 - \alpha$ est donnée par :

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

(voir poly page 18). Ici $z_{1-\alpha/2} = z_{0.975} = 1.96$, on cherche donc une taille d'échantillon n telle que

$$z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq 0.02 \iff n \geq 1.96^2 \frac{p(1-p)}{0.02^2}.$$

Le problème vient bien entendu du fait que p est ici inconnue. Etudions comme l'indique l'énoncé la fonction $f(p) = p(1-p)$ sur $[0, 1]$. $f'(p) = 1 - 2p$, donc f est croissante sur $[0, 1/2]$ et décroissante sur $[1/2, 1]$, elle atteint donc son maximum en $p = 1/2$, ce qui implique $f(p) \leq f(1/2) = 1/4$.

Rappel : on cherche n tel que

$$n \geq 1.96^2 \frac{p(1-p)}{0.02^2} \tag{B.2}$$

et $p(1-p) \leq 1/4$, ce qui signifie que dans le pire des cas $p(1-p) = 1/4$. On se place donc dans ce cas le plus défavorable qui va conduire à une taille d'échantillon maximale (si (B.2) est vraie dans le pire des cas, elle sera vraie dans tous les autres cas). On cherche donc n tel que :

$$n \geq 1.96^2 \frac{1}{4 \times 0.02^2} = 2401.$$

Il faut interroger 2401 personnes pour être sûr que l'intervalle de confiance de niveau 95% pour la proportion de parisiens qui portent des lunettes ait une demi longueur d'au plus 0.02.

Exercice B.8

Même raisonnement et même réponse que pour l'exercice 7.

Exercice B.9

1. Pour un plan simple, la variance de $\hat{\mu}$ vaut :

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{N}.$$

Il faut donc calculer S^2 la variance corrigée sur la population :

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{3}((0-0)^2 + (0-0)^2 + (1-0)^2 + (-1-0)^2) = \frac{2}{3}.$$

On déduit :

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{2}{4}\right) \frac{\frac{2}{3}}{4} = \frac{1}{6}.$$

2. Pour un plan stratifié la variance est donnée par :

$$\mathbf{V}(\hat{\mu}) = \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2. \quad (\text{B.3})$$

Il faut cette fois calculer la variance corrigée dans les deux strates :

$$S_1^2 = 0, \text{ et } S_2^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N_2} (X_i - \mu_2)^2 = 2.$$

On déduit :

$$\mathbf{V}(\hat{\mu}) = \frac{1}{4^2} \left(2 \times \frac{2-1}{1} \times 0 + 2 \times \frac{2-1}{1} \times 2\right) = \frac{1}{4}.$$

Exercice B.10

1. La population (ensemble des médecins) est ici divisée en trois catégories dans lesquelles on réalise un plan simple, il s'agit donc d'un plan de sondage stratifié.
2. L'estimateur du nombre moyen de clients soignés par jour par médecin pour un tel plan est donné par

$$\hat{\mu} = \frac{1}{N} \sum_{h=1}^H N_h \bar{x}_h = \frac{1}{4000} (500 * 10 + 1000 * 15 + 2500 * 20) = 17.5.$$

3. Il faut d'abord calculer la variance de $\hat{\mu}$, en utilisant la formule (B.3), on trouve

$$\mathbf{V}(\hat{\mu}) = 0.0199.$$

On calcule l'intervalle de confiance de niveau 0.95 à partir de la formule :

$$IC = \left[\hat{\mu} - z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})}; \hat{\mu} + z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})} \right] = [17.22; 17.78].$$

4. Si la variance corrigée S_h^2 de chaque strate est inconnue, on effectue un plan stratifié avec allocation proportionnelle. On choisit alors comme taille d'échantillon dans chaque strate :

$$n_1 = 75, \quad n_2 = 150, \quad n_3 = 375.$$

Si S_h^2 est connu pour chaque strate, on fait alors un plan stratifié avec allocation optimale, *i.e.*, on choisit les tailles d'échantillon suivant :

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}.$$

Exercice B.11

1. Pour calculer σ_h^2 , on utilise la formule :

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^H N_h \sigma_h^2 + \frac{1}{N} \sum_{h=1}^H N_h (\mu_h - \mu)^2 = \sigma_{\text{intra}}^2 + \sigma_{\text{inter}}^2.$$

$\sigma_h^2 = \frac{N_h - 1}{N_h} S_h^2$, donc

$$\sigma_1^2 = \frac{59}{60} 4 = 3.93, \quad \sigma_2^2 = \frac{39}{40} 2.25 = 2.19.$$

D'où

$$\sigma_{\text{intra}}^2 = \frac{1}{100} [60 * 3.93 + 40 * 2.19] = 3.24$$

et

$$\sigma_{\text{inter}}^2 = \frac{1}{100} [60 * (6 - 5.2)^2 + 40 * (4 - 5.2)^2] = 0.96.$$

On déduit

$$\sigma^2 = 4.2, \quad S^2 = \frac{N-1}{N} \sigma^2 = \frac{100}{99} 4.2 = 4.24.$$

2. On réalise ici un plan simple. Soit $\hat{\mu}$ l'estimateur de μ pour ce plan. Pour avoir une estimation du total T , il suffit de multiplier le poids moyen de l'échantillon par le nombre total d'éléphants, ce qui donne $\hat{T} = N\hat{\mu}$. On a donc

$$\mathbf{V}(\hat{T}) = \mathbf{V}(N\hat{\mu}) = N^2 \mathbf{V}(\hat{\mu}) = 100^2 * \left(1 - \frac{10}{100}\right) \frac{4.24}{10} = 3816.$$

3. On note n_M (resp n_F) le nombre de mâles (resp femelles) dans l'échantillon. Pour un plan avec allocation proportionnelle, on a :

$$n_H = n \frac{N_H}{N} = 10 \frac{60}{100} = 6$$

et

$$n_F = n \frac{N_F}{N} = 10 \frac{40}{100} = 4$$

On calcule la variance en utilisant la formule (B.3) et on trouve :

$$\mathbf{V}(\hat{\mu}) = 2970.$$

4. Pour un plan avec allocation optimale, les tailles d'échantillons sont données par :

$$n_H = 10 \times \frac{60 \times 2}{60 \times 2 + 40 \times \sqrt{2.25}} = 6.66$$

et

$$n_F = 10 \times \frac{40 \times \sqrt{2.25}}{60 \times 2 + 40 \times \sqrt{2.25}} = 3.33.$$

Cela donne $n_H = 7$ et $n_F = 3$ après arrondi. On utilise toujours (B.3) pour obtenir la variance

$$\mathbf{V}(\hat{\mu}) = 2927.$$

Parmi les trois plans de sondage étudiés dans cet exercice, la variance de l'estimateur $\hat{\mu}$ est la plus faible pour le plan stratifié avec allocation optimale. Ce plan est donc le plus précis.

Exercice B.12

On cherche à estimer la proportion p (inconnue) d'employés qui possèdent un véhicule.

1. On interroge n_h personnes dans chaque strate E_h , \bar{p}_h désigne la proportion de personnes **interrogées** (de l'échantillon) dans la strate E_h qui possèdent un véhicule. On estime p par

$$\hat{p} = \frac{1}{N} \sum_{h=1}^H N_h \bar{p}_h,$$

la moyenne des proportions par strate pondérée par le nombre d'individus dans chaque strate N_h . Compte tenu des résultats du sondage on a

$$\hat{p} = \frac{1}{7500} (3500 \times 0.13 + 2000 \times 0.45 + 2000 \times 0.5) = 0.314.$$

2. Comme pour l'estimation de la moyenne, un intervalle de confiance de niveau 0.95 est donné par :

$$IC_{0.95} = \left[\hat{p} - z_{0.975} \sqrt{\mathbf{V}(\hat{p})}; \hat{p} + z_{0.975} \sqrt{\mathbf{V}(\hat{p})} \right].$$

On lit sur la table $z_{0.975} = 1.96$. Une proportion étant une moyenne, on a

$$\mathbf{V}(\hat{p}) = \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2.$$

La variance corrigée de chaque strate S_h^2 est ici inconnue, on va l'estimer à partir des résultats de l'enquête par la variance corrigée prise sur l'échantillon s_h^2 . La difficulté consiste ici à déterminer cette variance. On utilise la formule donnée dans l'énoncé :

$$s_h^2 = \frac{N}{N-1} \bar{p}_h (1 - \bar{p}_h).$$

On déduit

$$s_1^2 = \frac{7500}{7449} 0.13(1 - 0.13) = 0.114, \quad s_2^2 = 0.248, \quad s_3^2 = 0.251,$$

et on obtient

$$\begin{aligned} \mathbf{V}(\hat{p}) &= \frac{1}{7500^2} \left(3500 \frac{3500 - 500}{500} 0.114 + 2000 \frac{2000 - 300}{300} 0.248 + 2000 \frac{2000 - 200}{200} 0.251 \right) \\ &= 0.0001724. \end{aligned}$$

On trouve donc l'intervalle de confiance

$$IC_{0.95} = [0.288; 0.339].$$

Exercice B.13

1. L'estimateur stratifié de l'âge moyen μ est donné par

$$\hat{\mu} = \frac{1}{N} \sum_{h=1}^H N_h \bar{x}_h = \frac{1}{10000} (5000 \times 25 + 3000 \times 45 + 2000 \times 58) = 37.6.$$

2. La variance de cet estimateur se calcule à l'aide de la formule (B.3), on trouve après calcul

$$\mathbf{V}(\hat{\mu}) = 0.16.$$

3. L'allocation proportionnelle propose de choisir les tailles d'échantillon de sorte que les proportions d'individus dans les strates de l'échantillon soient les mêmes que dans les strates de la population :

$$\frac{n_h}{n} = \frac{N_h}{N} \iff n_h = n \frac{N_h}{N}.$$

On obtient

$$n_1 = 50, \quad n_2 = 30, \quad n_3 = 20.$$

Toujours par la formule (B.3), on obtient

$$\mathbf{V}(\hat{\mu}) = 0.1485.$$

4. Pour un plan stratifié avec allocation optimale, on choisit les tailles d'échantillon de manière à minimiser la variance de l'estimateur $\hat{\mu}$,

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}.$$

On obtient après calcul

$$n_1 = 52.04, \quad n_2 = 24.68, \quad n_3 = 23.27,$$

en arrondissant

$$n_1 = 52, \quad n_2 = 25, \quad n_3 = 23.$$

On calcule toujours la variance à l'aide de (B.3) :

$$\mathbf{V}(\hat{\mu}) = 0.1462.$$

Exercice B.14

1. Un intervalle de confiance de niveau 0.90 est donné par

$$IC_{0.90} = \left[\hat{\mu} - z_{0.95} \sqrt{\mathbf{V}(\hat{\mu})}, \hat{\mu} + z_{0.95} \sqrt{\mathbf{V}(\hat{\mu})} \right],$$

avec $z_{0.95} \simeq 1.64$. On calcule $\mathbf{V}(\hat{\mu})$ grâce à (B.3) et on obtient

$$\mathbf{V}(\hat{\mu}) = 0.055.$$

On calcule $\hat{\mu} = 29.81$ et on déduit

$$IC_{0.90} = [29.43; 30.19].$$

2. (a) Pour une allocation proportionnelle $n_h = n \frac{N_h}{N}$, donc

$$n_1 = 141.51, \quad n_2 = 84.91, \quad n_3 = 42.45, \quad n_4 = 28.30, \quad n_5 = 2.83,$$

en arrondissant

$$n_1 = 142, \quad n_2 = 85, \quad n_3 = 42, \quad n_4 = 28, \quad n_5 = 3.$$

- (b) (**plus difficile**) Pour une allocation optimale

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h},$$

ce qui donne

$$n_1 = 58.57, \quad n_2 = 57.39, \quad n_3 = 40.58, \quad n_4 = 95.64, \quad n_5 = 47.82,$$

en arrondissant

$$n_1 = 59, \quad n_2 = 57, \quad n_3 = 40, \quad n_4 = 96, \quad n_5 = 48.$$

On doit interroger 48 personnes dans la strate 5 alors qu'elle n'en contient que 10 !!! C'est bien entendu impossible, on choisit donc d'interroger les 10 personnes de la strate 5 ($n_5 = 10$) et on recalcule les tailles d'échantillons pour les quatre autres strates avec $n = 300 - 10 = 290$. On a par exemple pour n_1

$$n_1 = 290 \frac{500\sqrt{1.5}}{500\sqrt{1.5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100}} = 67.35,$$

de même

$$n_2 = 65.99, \quad n_3 = 46.66, \quad n_4 = 109.98.$$

Encore une fois, on doit interroger $n_4 = 110$ individus dans la strate 4 qui en contient 100. On les interroge donc toutes ($n_4 = 100$) et on recalcule n_1, n_2 et n_3 avec $n = 290 - 100 = 190$. On obtient après arrondi

$$n_1 = 71, \quad n_2 = 70, \quad n_3 = 49.$$

Pour résumer

$$n_1 = 71, \quad n_2 = 70, \quad n_3 = 49, \quad n_4 = 100, \quad n_5 = 10.$$

3. Pour l'allocation proportionnelle on obtient grâce à (B.3)

$$\mathbf{V}(\hat{\mu}) = 0.0819.$$

Pour l'allocation optimale, on obtient :

$$\mathbf{V}(\hat{\mu}) = 0.00974.$$

Annexe C

Sujet Licence AES 3 : juin 2006 (assidus)

NB : Ce devoir vous sera corrigé si vous me le remettez à l'occasion d'un stage ou me l'expédiez par courrier (n'oubliez pas de joindre une enveloppe à votre adresse) :

Laurent Rouvière
Département MASS
Université Rennes 2-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France

e-mail : laurent.rouviere@uhb.fr
tel : 02 99 14 18 21

Exercice C.1

Expliquer en quoi consiste un plan de sondage aléatoire simple ainsi qu'un plan stratifié. Dans le cas de la stratification, quel est le principe de l'allocation proportionnelle ? Et de l'allocation optimale ? Quel est l'intérêt de la stratification ?

Exercice C.2

On souhaite estimer la quantité d'eau moyenne (exprimée en m^3) consommée annuellement par les habitants d'une ville donnée de 100 000 habitants. On sélectionne par un plan simple un échantillon de 250 habitants. Les résultats obtenus sont les suivants :

$$\sum_{i=1}^n x_i = 15\,125 \quad \sum_{i=1}^n x_i^2 = 921\,310.$$

1. Traduire en quelques mots l'information contenue dans la formule : $\sum_{i=1}^n x_i = 15\,125$.
2. Donner un intervalle de confiance à 95% pour la quantité d'eau moyenne consommée annuellement par les habitants de cette ville.

	Effectif N_h	S_h^2
Mâles	60	4
Femelles	40	2.25

	Effectif N	S^2
Mâles et femelles confondus	100	4.24

3. On s'intéresse maintenant à la quantité totale consommée annuellement par l'ensemble des habitants de la ville. Donner une estimation, puis un intervalle de confiance à 95% pour cette quantité totale.

Exercice C.3

Un directeur de cirque possède un troupeau de 100 éléphants et souhaite estimer le poids moyen de ses éléphants. Cependant, l'année précédente, le directeur de cirque les avait classés en deux catégories, les mâles et les femelles, puis avait fait peser tous les éléphants de son troupeau. Il avait obtenu les résultats suivants (les moyennes sont exprimées en tonnes) :

1. Le directeur suppose désormais que les dispersions de poids n'évoluent pas sensiblement d'une année sur l'autre, c'est-à-dire que les valeurs des S_h^2 restent inchangées (ce type d'hypothèse reste ici très raisonnable et se rencontre couramment en pratique quand on répète des enquêtes dans le temps). Si le directeur procède à un tirage aléatoire simple de 10 éléphants, quelle est la variance de l'estimateur du poids moyen du troupeau ?
2. Le directeur procède à un tirage stratifié et sélectionne cinq femelles et cinq mâles. Il obtient pour l'échantillon des mâles une moyenne de $\bar{x}_1 = 6.5$ et de $\bar{x}_2 = 3.9$ pour celui des femelles. Donner une estimation du poids moyen du troupeau. Calculer la variance de l'estimateur de ce poids moyen.
3. Si le directeur procède à un tirage stratifié avec allocation proportionnelle de 10 éléphants, quelles tailles d'échantillon doit-on retenir dans chaque strate ? Quelle est alors la variance de l'estimateur du poids moyen du troupeau ?
4. Si le directeur procède à un tirage stratifié optimal de 10 éléphants, quelles tailles d'échantillon doit-on retenir dans chaque strate ? Quelle est alors la variance de l'estimateur du poids moyen du troupeau ?
5. Parmi les quatre plans de sondage proposés, lequel vous semble le plus approprié ?

Exercice C.4

Une équipe est chargée de réaliser une enquête dans le but d'estimer la proportion de restaurants disposant d'une salle entièrement non fumeur en France. On sélectionne par plan simple un échantillon de 120 restaurants. Parmi ces 120 restaurants sélectionnés, 51 disposent d'une salle entièrement non fumeur.

Dans cet exercice on négligera le taux de sondage f .

-
1. On souhaite donner un intervalle de confiance à 90% puis à 95% pour la proportion p de restaurants disposant d'une salle entièrement non-fumeur.
 - (a) Avant d'effectuer les calculs, pouvez-vous dire, en justifiant votre réponse, quel sera l'intervalle le plus large ?
 - (b) Donner ces intervalles de confiance.
 2. Quelle taille d'échantillon doit-on retenir pour que l'on puisse donner un intervalle de confiance à 95% pour p ayant une demi-longueur d'au plus 3%, en utilisant "l'intervalle de précaution" ?

Indications :

- (a) Montrer que la fonction $f(p) = p(1 - p) = p - p^2$ définie pour $0 \leq p \leq 1$ atteint son maximum en $p = \frac{1}{2}$ et que ce maximum est égal à $\frac{1}{4}$.
- (b) En déduire que le "pire des cas", c'est-à-dire le cas où la demi-longueur de l'intervalle de confiance est la plus grande, correspond au cas où $\hat{p} = \frac{1}{2}$.
- (c) Trouver la taille d'échantillon n recherchée.

Annexe D

Sujet Licence AES 3 : septembre 2006 (assidus)

NB : Ce devoir vous sera corrigé si vous me le remettez à l'occasion d'un stage ou me l'expédiez par courrier (n'oubliez pas de joindre une enveloppe à votre adresse) :

*Laurent Rouvière
Département MASS
Université Rennes 2-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France*

*e-mail : laurent.rouviere@uhb.fr
tel : 02 99 14 18 21*

Exercice D.1

- Qu'est-ce qu'un plan de sondage aléatoire ? Donner un exemple de plan non aléatoire.
- Expliquer en quoi consiste un plan de sondage aléatoire simple ainsi qu'un plan stratifié. Dans le cas de la stratification avec allocation optimale, de quelle(s) information(s) supplémentaire(s) par rapport à l'allocation proportionnelle a-t-on besoin pour calculer les tailles des échantillons issus des différentes strates ? Comment obtient-on en pratique ces informations ?

Exercice D.2

On souhaite estimer la quantité moyenne de fruits (exprimée en kg) consommée annuellement par les habitants d'une ville de 100 000 habitants. On sélectionne par un plan simple un échantillon de 200 habitants. Les résultats obtenus sont les suivants :

$$\sum_{i=1}^n x_i = 18\,700 \quad \sum_{i=1}^n x_i^2 = 1\,766\,500.$$

1. Donner un intervalle de confiance à 95% pour la quantité de fruits moyenne consommée annuellement par les habitants de cette ville.

2. On s'intéresse maintenant à la quantité totale consommée annuellement par l'ensemble des habitants de la ville. Donner une estimation, puis un intervalle de confiance à 95% pour cette quantité totale.
3. On souhaite dans cette question donner un intervalle de confiance à 95% pour la quantité de fruits moyenne consommée annuellement par les habitants de cette ville ayant une demi-longueur d'au plus 1 kg. On cherche une taille d'échantillon n qui permette de construire un tel intervalle.
 - (a) Pour trouver cette taille n , on néglige le taux de sondage f . Pouvez-vous donner une interprétation "concrète" de cette hypothèse et expliquer pourquoi elle est raisonnable ? A l'inverse, que signifie un taux de sondage égal à 1 ?
 - (b) Un premier expert estime en se basant sur des enquêtes précédentes que l'on peut considérer que la variance corrigée S^2 calculée sur l'ensemble de la population est égale à 100. Un autre expert estime que la variance corrigée S^2 est un peu plus élevée, et est égale à 125.
 - i. Si vous souhaitez être prudent et vous placer dans le pire des cas possibles, de quel expert allez-vous suivre l'avis ?
 - ii. Calculer n (dans ce pire des cas).

Exercice D.3

Une grande entreprise qui comprend 10 000 personnes souhaite estimer l'âge moyen de son personnel. Des études préliminaires ont montré que l'âge est fortement lié aux différentes catégories de personnels. Pour simplifier, on considérera qu'il y a 3 grandes catégories qui formeront les strates. Cinq années auparavant, le directeur avait recensé l'âge de tous ses employés, il avait obtenu les résultats suivants : Le directeur souhaite estimer l'âge moyen

Catégories	Effectif N_h	S_h^2
1	2000	324
2	3000	144
3	5000	100
Ensemble	10000	256

des employés noté μ à partir d'un échantillon de 100 personnes. Il suppose désormais que les dispersions des âges n'ont pas évolué sensiblement au cours des 5 dernières années (ce type d'hypothèse reste ici très raisonnable et se rencontre couramment en pratique quand on répète des enquêtes dans le temps).

1. Si le directeur procède à un tirage aléatoire simple de 100 employés, quelle est la variance de l'estimateur de l'âge moyen des employés ?
2. Le directeur procède à un tirage stratifié avec allocation proportionnelle de 100 employés.
 - (a) Quelles tailles d'échantillon doit-on retenir dans chaque strate ? Quelle est alors la variance de l'estimateur du l'âge moyen des employés ?

- (b) Il obtient pour l'échantillon de la catégorie 1, une moyenne de $\bar{x}_1 = 34$, pour l'échantillon de la catégorie 2, une moyenne de $\bar{x}_2 = 38$ et pour l'échantillon de la catégorie 3, une moyenne de $\bar{x}_3 = 50$. Donner un intervalle de confiance de à 90% pour l'âge moyen des employés.
3. Dans cette question, nous négligerons le taux de sondage f . Nous sommes toujours dans le cas d'un tirage stratifié avec allocation proportionnelle. Le directeur souhaite connaître la taille d'échantillon n qu'il doit retenir pour qu'un intervalle de confiance de niveau 90% pour μ ait une demi-longueur d'au plus 1 an.
- (a) Avant d'effectuer les calculs, pouvez vous dire, en justifiant votre réponse, si cette taille d'échantillon sera supérieure ou inférieure à 100 ?
- (b) Calculer cette taille d'échantillon.

Annexe E

Sujet Licence AES 3 : mai 2007 (non assidus)

NB : Ce devoir vous sera corrigé si vous me le remettez à l'occasion d'un stage ou me l'expédiez par courrier (n'oubliez pas de joindre une enveloppe à votre adresse) :

*Laurent Rouvière
Département MASS
Université Rennes 2-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France*

*e-mail : laurent.rouviere@uhb.fr
tel : 02 99 14 18 21*

Exercice E.1 (Vrai ou Faux : +0.5 bonne réponse, -0.5 mauvaise réponse.)

On souhaite estimer l'âge moyen μ dans une population de taille N . La population est découpée suivant trois strates. On estime μ à l'aide des trois plans de sondage suivant :

- \mathcal{P}_1 : un plan de sondage aléatoire simple. On note $\hat{\mu}_1$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_2 : un plan stratifié avec allocation proportionnelle. On note $\hat{\mu}_2$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_3 : un plan stratifié avec allocation optimale. On note $\hat{\mu}_3$ l'estimateur de μ pour un tel plan.

Pour les trois plans de sondage ci-dessus, les échantillons sont de même tailles n . Dire sans justifier si les assertions suivantes sont vraie ou fausses.

1. μ est une variable aléatoire (il peut prendre plusieurs valeurs suivant l'échantillon choisi).
2. $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont des variables aléatoires (ils peuvent prendre plusieurs valeurs suivant l'échantillon choisi).
3. Les estimateurs $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont tous sans biais.

4. Les intervalles de confiance de niveau 0.95 construits à partir de ces trois plans ont tous la même longueur.
5. Les intervalles de confiance de niveau 0.95 construits à partir de ces trois plans ont tous le même centre.
6. Pour la plan \mathcal{P}_2 , le centre de l'intervalle de confiance de niveau 0.95 est $\hat{\mu}_2$.
7. La variance de $\hat{\mu}_3$ est toujours inférieure ou égale à la variance de $\hat{\mu}_2$.
8. Si le taux de sondage $f = n/N$ est égal à 1, on a forcément $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \mu$.

Exercice E.2 (7.5 points)

On souhaite estimer μ la distance moyenne (exprimée en kilomètres) parcourue en vélo par les habitants d'une ville de $N = 50\,000$ habitants en mai 2005. On sélectionne par un plan de sondage aléatoire simple un échantillon de taille $n = 250$. On note x_i la distance (exprimée en kilomètres) parcourue en mai 2005 par le $i^{\text{ème}}$ individu de l'échantillon. Les résultats sont :

$$\sum_{i=1}^{250} x_i = 15\,150, \quad \sum_{i=1}^{250} x_i^2 = 1\,155\,400.$$

1. Traduire en quelques mots l'information contenue dans la formule $\sum_{i=1}^{250} x_i = 15\,150$.
2. Avec les notations du cours, on rappelle que la variance corrigée s^2 de l'échantillon peut se calculer de la manière suivante :

$$s^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right).$$

Calculer cette variance corrigée s^2 .

3. On souhaite donner un intervalle de confiance de niveau 90%, puis 95% pour μ .
 - (a) Avant d'effectuer les calculs, pouvez vous dire, en justifiant votre réponse, quel sera l'intervalle le plus large ?
 - (b) Donner ces intervalles de confiance (pour la loi normale centrée réduite, on rappelle que le quantile d'ordre 0.95 vaut 1.64, celui d'ordre 0.975 vaut 1.96).
4. On souhaite dans cette question donner un intervalle de confiance de niveau 95% pour μ ayant une demi-longueur d'au plus 2 kilomètres. On considère que la variance corrigée S^2 calculée sur l'ensemble de la population est la même que la variance corrigée s^2 calculée sur l'échantillon (elle a été calculée à la question 2).
 - (a) Avant d'effectuer les calculs, pouvez vous dire, en justifiant votre réponse, si la taille d'échantillon cherchée sera supérieure ou inférieure à 250.
 - (b) Calculer cette taille d'échantillon (on négligera le taux de sondage $f = n/N$ pour simplifier les calculs).

Exercice E.3 (7.5 points)

Le chef d'une entreprise de $N = 10\,000$ employés souhaite estimer μ l'âge moyen de ses employés. Pour chaque individu de son l'entreprise, l'entrepreneur connaît la répartition de ses employés suivant deux variables :

- le salaire net partagée en 3 catégories :
 - inférieur à 1 400 euros ;
 - entre 1 400 et 2 500 euros ;
 - supérieur à 2 500 euros ;
- l'ancienneté (mesurée en nombre d'années dans l'entreprise) :
 - moins de 8 ans ;
 - entre 8 et 18 ans ;
 - plus de 18 ans.

Les répartitions des individus suivant ces deux variables sont données dans les tableaux suivants :

Salaires	Effectifs N_h	S_h^2
$[0; 1\,400[$	2 000	100
$[1\,400; 2\,500[$	6 500	64
plus de 2 500	1 500	81

TABLE E.1 – Répartition selon les salaires.

Ancienneté	Effectifs N_h	S_h^2
moins de 8 ans	1 500	16
entre 8 et 18 ans	4 500	25
plus de 18 ans	4 000	9

TABLE E.2 – Répartition selon l'ancienneté.

La colonne S_h^2 désigne la variance corrigée de la variable âge mesurée sur la population qui compose la strate h .

Le patron de l'entreprise décide de faire réaliser l'étude par deux instituts de sondage. Le premier institut I_1 décide de réaliser un plan stratifié en découpant la population suivant les classes de salaires proposées dans le tableau E.1. Le second institut propose de stratifier la population suivant les classes d'ancienneté du tableau E.2.

1. Avant d'effectuer les calculs, pouvez vous dire quel est le plan qui vous semble le plus pertinent parmi les deux plans proposés par I_1 et I_2 ? Justifier votre réponse.
2. Les deux instituts de sondage décide de constituer un échantillon de taille $n = 100$.
 - (a) Quelles tailles d'échantillon doit retenir l'institut I_1 dans chaque strate s'il réalise un plan avec allocation proportionnelle? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
 - (b) Quelles tailles d'échantillon doit retenir l'institut I_2 dans chaque strate s'il réalise un plan avec allocation optimale? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
3. Pour le plan réalisé par l'institut I_2 dans la question 2-b), on a les résultats suivants :

$$\bar{x}_1 = 28, \quad \bar{x}_2 = 40, \quad \bar{x}_3 = 52,$$

où \bar{x}_h désigne l'âge moyen des individus de l'échantillon dans la strate h .

- (a) Donner $\hat{\mu}$ l'estimateur ponctuel de μ pour ce plan de sondage.
- (b) Donner un intervalle de confiance de niveau 0.95 pour μ .

Annexe F

Sujet Licence AES 3 : mai 2008 (non assidus)

NB : Ce devoir vous sera corrigé si vous me le remettez à l'occasion d'un stage ou me l'expédiez par courrier (n'oubliez pas de joindre une enveloppe à votre adresse) :

*Laurent Rouvière
Département MASS
Université Rennes 2-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France*

*e-mail : laurent.rouviere@uhb.fr
tel : 02 99 14 18 21*

Exercice F.1 (Vrai ou Faux : +0.5 bonne réponse, -0.5 mauvaise réponse.)

On souhaite estimer l'âge moyen μ dans une population de taille N . La population est découpée suivant trois strates. On estime μ à l'aide des trois plans de sondage suivant :

- \mathcal{P}_1 : un plan de sondage aléatoire simple. On note $\hat{\mu}_1$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_2 : un plan stratifié avec allocation proportionnelle. On note $\hat{\mu}_2$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_3 : un plan stratifié avec allocation optimale. On note $\hat{\mu}_3$ l'estimateur de μ pour un tel plan.

Pour les trois plans de sondage ci-dessus, les échantillons sont de même tailles n . Dire sans justifier si les assertions suivantes sont vraie ou fausses.

1. μ est une variable aléatoire (il peut prendre plusieurs valeurs suivant l'échantillon choisi).
2. Plus la taille n de l'échantillon est grande, plus la variance de $\hat{\mu}_1$ est petite.
3. $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont des variables aléatoires (ils peuvent prendre plusieurs valeurs suivant l'échantillon choisi).
4. Les estimateurs $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont tous sans biais.
5. Si $n = N$ alors la variance de $\hat{\mu}_2$ est nulle.

6. Les intervalles de confiance de niveau 0.95 construits à partir de ces trois plans ont tous le même centre.
7. La demi-longueur d'un intervalle de confiance de niveau 0.90 est toujours plus grande que celle d'un intervalle de confiance de niveau 0.95
8. Si le taux de sondage $f = n/N$ est égal à 1, on a forcément $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \mu$.

Exercice F.2 (7.5 points)

On souhaite estimer μ le poids moyen (exprimé en kilogrammes) des habitants d'une ville de $N = 50\,000$ habitants. On sélectionne par un plan de sondage aléatoire simple (sans remise) un échantillon de taille $n = 500$. On note x_i le poids (exprimée en kilogrammes) du $i^{\text{ème}}$ individu de l'échantillon. Les résultats sont :

$$\sum_{i=1}^{500} x_i = 40\,200, \quad \sum_{i=1}^{500} x_i^2 = 3\,300\,000.$$

1. Traduire en quelques mots l'information contenue dans la formule $\sum_{i=1}^{500} x_i = 40\,200$.
2. Donner une estimation ponctuelle du poids moyen ainsi que du poids total des habitants de la ville.
3. Avec les notations du cours, on rappelle que la variance corrigée s^2 de l'échantillon peut se calculer de la manière suivante :

$$s^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right).$$

Calculer cette variance corrigée s^2 .

4. On souhaite donner un intervalle de confiance de niveau 90%, puis 95% pour μ .
 - (a) Avant d'effectuer les calculs, pouvez vous dire, en justifiant votre réponse, quel sera l'intervalle le plus large ?
 - (b) Donner ces intervalles de confiance (pour la loi normale centrée réduite, on rappelle que le quantile d'ordre 0.95 vaut 1.64, celui d'ordre 0.975 vaut 1.96).
5. On souhaite dans cette question donner un intervalle de confiance de niveau 95% pour μ ayant une demi-longueur d'au plus 1 kilogramme. On considère que la variance corrigée S^2 calculée sur l'ensemble de la population est la même que la variance corrigée s^2 calculée sur l'échantillon (elle a été calculée à la question 2).
 - (a) Avant d'effectuer les calculs, pouvez-vous dire, en justifiant votre réponse, si la taille d'échantillon cherchée sera supérieure ou inférieure à 500.
 - (b) Calculer cette taille d'échantillon (on négligera le taux de sondage $f = n/N$ pour simplifier les calculs).

Exercice F.3 (7.5 points)

Le ministère de l'industrie souhaite estimer μ le chiffre d'affaire moyen en millions d'euros des $N = 10\,000$ entreprises d'un département. Pour chaque entreprise du département, la

personne chargée de l'étude connaît la répartition des entreprises du département suivant deux variables :

- le nombre d'employés :
 - inférieur à 15 employés ;
 - entre 15 et 50 employés ;
 - supérieur à 50 employé ;
- l'âge moyen des employés
 - moins de 35 ans ;
 - entre 35 et 48 ans ;
 - plus de 48 ans.

Les répartitions des individus suivant ces deux variables est donnée dans les tableaux suivants :

Age moyen \ Nb employé	Nb employé			Total
	[0; 15[[15; 50[plus de 50	
[0; 35[1 500	500	500	2 500
[35; 48[2 000	1 500	1 000	4 500
plus de 48	500	1 500	1 000	3 000
Total	4 000	3 500	2 500	10 000

TABLE F.1 – Répartition des entreprises selon l'âge moyen et le nombre d'employés.

L'écart type corrigé de la variable chiffre d'affaire suivant les variables nombre d'employés et âge moyen des employés est connu. Il est donné dans les tableaux suivants :

Nombre d'employés	S_h
[0; 15[10
[15; 50[6
plus de 50	12

TABLE F.2 – Ecart-type corrigé selon le nombre d'employés.

Age moyen	S_h
[0; 35[17
[35; 48[14
plus de 48	28

TABLE F.3 – Ecart type corrigé selon l'âge moyen.

La personne chargée de l'étude décide de faire appel à deux instituts de sondage. Le premier institut I_1 décide de réaliser un plan stratifié en découpant la population suivant l'âge moyen des salariés de l'entreprise. Le second institut I_2 propose de stratifier la population suivant le nombre d'employés des entreprises.

1. Avant d'effectuer les calculs, pouvez vous dire quel est le plan qui vous semble le plus pertinent parmi les deux plans proposés par I_1 et I_2 ? Justifier votre réponse.
2. Les deux instituts de sondage décide de constituer un échantillon de taille $n = 100$.
 - (a) Quelles tailles d'échantillon doit retenir l'institut I_1 dans chaque strate s'il réalise un plan avec allocation proportionnelle ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
 - (b) Quelles tailles d'échantillon doit retenir l'institut I_2 dans chaque strate s'il réalise un plan avec allocation optimale ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.

3. Pour le plan réalisé par l'institut I_2 dans la question 2-b), on a les résultats suivants :

$$\bar{x}_1 = 18.4, \quad \bar{x}_2 = 31.8, \quad \bar{x}_3 = 90.2,$$

où \bar{x}_h désigne le chiffre d'affaire moyen des individus de l'échantillon dans la strate h .

- (a) Donner $\hat{\mu}$ l'estimateur ponctuel de μ pour ce plan de sondage.
- (b) Donner un intervalle de confiance de niveau 0.95 pour μ .

Annexe G

Sujet Licence AES 3 : juin 2008 (non assidus)

NB : Ce devoir vous sera corrigé si vous me le remettez à l'occasion d'un stage ou me l'expédiez par courrier (n'oubliez pas de joindre une enveloppe à votre adresse) :

*Laurent Rouvière
Département MASS
Université Rennes 2-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France*

*e-mail : laurent.rouviere@uhb.fr
tel : 02 99 14 18 21*

Exercice G.1 (Vrai ou Faux : +0.5 bonne réponse, -0.5 mauvaise réponse.)

On souhaite estimer l'âge moyen μ dans une population de taille N . La population est découpée suivant trois strates. On estime μ à l'aide des trois plans de sondage suivant :

- \mathcal{P}_1 : un plan de sondage aléatoire simple. On note $\hat{\mu}_1$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_2 : un plan stratifié avec allocation proportionnelle. On note $\hat{\mu}_2$ l'estimateur de μ pour un tel plan.
- \mathcal{P}_3 : un plan stratifié avec allocation optimale. On note $\hat{\mu}_3$ l'estimateur de μ pour un tel plan.

Pour les trois plans de sondage ci-dessus, les échantillons sont de même tailles n . Dire sans justifier si les assertions suivantes sont vraie ou fausses.

1. μ est une variable aléatoire (il peut prendre plusieurs valeurs suivant l'échantillon choisi).
2. Plus la taille n de l'échantillon est grande, plus la variance de $\hat{\mu}_1$ est petite.
3. $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont des variables aléatoires (ils peuvent prendre plusieurs valeurs suivant l'échantillon choisi).
4. Les estimateurs $\hat{\mu}_1, \hat{\mu}_2$ et $\hat{\mu}_3$ sont tous sans biais.
5. Si $n = N$ alors la variance de $\hat{\mu}_2$ est nulle.

6. Les intervalles de confiance de niveau 0.95 construits à partir de ces trois plans ont tous le même centre.
7. La demi-longueur d'un intervalle de confiance de niveau 0.90 est toujours plus grande que celle d'un intervalle de confiance de niveau 0.95
8. Si le taux de sondage $f = n/N$ est égal à 1, on a forcément $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \mu$.

Exercice G.2 (7.5 points)

On souhaite estimer μ le poids moyen (exprimé en kilogrammes) des habitants d'une ville de $N = 50\,000$ habitants. On sélectionne par un plan de sondage aléatoire simple (sans remise) un échantillon de taille $n = 500$. On note x_i le poids (exprimée en kilogrammes) du $i^{\text{ème}}$ individu de l'échantillon. Les résultats sont :

$$\sum_{i=1}^{500} x_i = 40\,200, \quad \sum_{i=1}^{500} x_i^2 = 3\,300\,000.$$

1. Traduire en quelques mots l'information contenue dans la formule $\sum_{i=1}^{500} x_i = 40\,200$.
2. Donner une estimation ponctuelle du poids moyen ainsi que du poids total des habitants de la ville.
3. Avec les notations du cours, on rappelle que la variance corrigée s^2 de l'échantillon peut se calculer de la manière suivante :

$$s^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right).$$

Calculer cette variance corrigée s^2 .

4. On souhaite donner un intervalle de confiance de niveau 90%, puis 95% pour μ .
 - (a) Avant d'effectuer les calculs, pouvez vous dire, en justifiant votre réponse, quel sera l'intervalle le plus large ?
 - (b) Donner ces intervalles de confiance (pour la loi normale centrée réduite, on rappelle que le quantile d'ordre 0.95 vaut 1.64, celui d'ordre 0.975 vaut 1.96).
5. On souhaite dans cette question donner un intervalle de confiance de niveau 95% pour μ ayant une demi-longueur d'au plus 1 kilogramme. On considère que la variance corrigée S^2 calculée sur l'ensemble de la population est la même que la variance corrigée s^2 calculée sur l'échantillon (elle a été calculée à la question 2).
 - (a) Avant d'effectuer les calculs, pouvez-vous dire, en justifiant votre réponse, si la taille d'échantillon cherchée sera supérieure ou inférieure à 500.
 - (b) Calculer cette taille d'échantillon (on négligera le taux de sondage $f = n/N$ pour simplifier les calculs).

Exercice G.3 (7.5 points)

Le ministère de l'industrie souhaite estimer μ le chiffre d'affaire moyen en millions d'euros des $N = 10\,000$ entreprises d'un département. Pour chaque entreprise du département, la

personne chargée de l'étude connaît la répartition des entreprises du département suivant deux variables :

- le nombre d'employés :
 - inférieur à 15 employés ;
 - entre 15 et 50 employés ;
 - supérieur à 50 employé ;
- l'âge moyen des employés
 - moins de 35 ans ;
 - entre 35 et 48 ans ;
 - plus de 48 ans.

Les répartitions des individus suivant ces deux variables est donnée dans les tableaux suivants :

Age moyen \ Nb employé	Nb employé			Total
	[0; 15[[15; 50[plus de 50	
[0; 35[1 500	500	500	2 500
[35; 48[2 000	1 500	1 000	4 500
plus de 48	500	1 500	1 000	3 000
Total	4 000	3 500	2 500	10 000

TABLE G.1 – Répartition des entreprises selon l'âge moyen et le nombre d'employés.

L'écart type corrigé de la variable chiffre d'affaire suivant les variables nombre d'employés et âge moyen des employés est connu. Il est donné dans les tableaux suivants :

Nombre d'employés	S_h
[0; 15[10
[15; 50[6
plus de 50	12

TABLE G.2 – Ecart-type corrigé selon le nombre d'employés.

Age moyen	S_h
[0; 35[17
[35; 48[14
plus de 48	28

TABLE G.3 – Ecart type corrigé selon l'âge moyen.

La personne chargée de l'étude décide de faire appel à deux instituts de sondage. Le premier institut I_1 décide de réaliser un plan stratifié en découpant la population suivant l'âge moyen des salariés de l'entreprise. Le second institut I_2 propose de stratifier la population suivant le nombre d'employés des entreprises.

1. Avant d'effectuer les calculs, pouvez vous dire quel est le plan qui vous semble le plus pertinent parmi les deux plans proposés par I_1 et I_2 ? Justifier votre réponse.
2. Les deux instituts de sondage décide de constituer un échantillon de taille $n = 100$.
 - (a) Quelles tailles d'échantillon doit retenir l'institut I_1 dans chaque strate s'il réalise un plan avec allocation proportionnelle ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.
 - (b) Quelles tailles d'échantillon doit retenir l'institut I_2 dans chaque strate s'il réalise un plan avec allocation optimale ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.

3. Pour le plan réalisé par l'institut I_2 dans la question 2-b), on a les résultats suivants :

$$\bar{x}_1 = 18.4, \quad \bar{x}_2 = 31.8, \quad \bar{x}_3 = 90.2,$$

où \bar{x}_h désigne le chiffre d'affaire moyen des individus de l'échantillon dans la strate h .

- (a) Donner $\hat{\mu}$ l'estimateur ponctuel de μ pour ce plan de sondage.
- (b) Donner un intervalle de confiance de niveau 0.95 pour μ .

Annexe H

Un dernier problème...

On réalise une enquête pour évaluer le salaire moyen des employés d'une entreprise. L'entreprise est composée de 20 salariés, on connaît la répartition des salariés suivant deux catégories : ouvrier (O) ou cadre (C). Les salaires ainsi que les catégories se trouvent dans le tableau H.1.

Employés	Catégories	salaire mensuel
1	C	2225
2	C	1616
3	C	2456
4	C	3350
5	C	2600
6	C	2028
7	C	3025
8	C	2756
9	C	1965
10	C	2618
11	O	1415
12	O	1415
13	O	1469
14	O	1335
15	O	1554
16	O	1465
17	O	1498
18	O	1325
19	O	1598
20	O	1484

TABLE H.1 – Salaires et catégories des employés.

1. Calculer le salaire moyen μ (que l'on va ensuite chercher à estimer!!!) et la variance corrigée S^2 ?
2. Un employé parmi les ouvrier souhaitent estimer le salaire moyen des employés en effectuant un plan de sondage aléatoire simple (avec un échantillon de taille $n = 8$).

- (a) Rappeler la formule qui permet de calculer l'estimateur de μ pour ce plan de sondage.
 - (b) Quelle est la variance de cet estimateur ?
3. Les cadres se trouvant dans des locaux éloignés du sien, il décide d'interroger uniquement des ouvriers de l'entreprise. Dans le cas où il interroge les 8 premiers ouvriers du tableau H.1, donner la valeur de l'estimateur de la moyenne $\hat{\mu}$.

Un ouvrier (un peu plus malin) se dit que l'estimation du salaire moyen serait "meilleure" en interrogeant des ouvriers et des cadres. Il décide de réaliser un plan de sondage stratifié (la taille de l'échantillon est toujours égale à 8).

4. Décrire l'enquête permettant de réaliser un tel plan de sondage ainsi que la manière de calculer l'estimateur $\hat{\mu}$ du salaire moyen. Quel est l'intérêt d'une telle procédure en comparaison avec les plans simples ?
5. On note n_C le nombre de personnes interrogées parmi les cadres et n_O parmi les ouvriers. Dans le cas d'un plan stratifié avec allocation proportionnelle :
 - (a) Calculer n_C et n_O .
 - (b) Calculer la variance de $\hat{\mu}$.
 - (c) On a interrogé les n_C premiers cadres et les n_O premiers ouvriers du tableau H.1, quelle est la valeur de $\hat{\mu}$?
 - (d) En déduire un intervalle de confiance de niveau 0.95 pour μ .
6. Reprendre la question 6 dans le cas d'un plan avec allocation optimale.
7. Comparer et commenter les différences entre les variances des estimateurs pour les trois plans de sondage proposés dans cet exercice.

CORRECTION

1. **Moyenne :**

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = 1959.4.$$

Variance corrigée :

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \mu)^2 = 399\,906.7.$$

2. (a) Pour $i = 1, \dots, 8$, on note x_i le salaire de la $i^{\text{ème}}$ personne interrogée, l'estimateur de μ est donné par :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

(b) La variance de cet estimateur est donnée par :

$$\mathbf{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{N} = \left(1 - \frac{8}{20}\right) \frac{399\,906.7}{20} = 11\,997.2$$

3. **Valeur de $\hat{\mu}$ sur l'échantillon :**

$$\hat{\mu} = \frac{1415 + 1469 + \dots + 1325}{8} = 1434.5.$$

4. Un plan de sondage stratifié consiste à découper la population suivant les deux catégories (ouvriers et cadres) et à réaliser un plan de sondage aléatoire simple dans chacune de ces deux populations (strates). Plus précisément, on interroge n_C salariés parmi les cadres et n_O parmi les ouvriers. On note \bar{x}_C (resp \bar{x}_O) le salaire moyen des cadres (resp ouvriers) interrogés. L'estimateur du salaire moyen de **tous** les salariés est obtenu grâce à la formule :

$$\hat{\mu} = \frac{N_O \bar{x}_O + N_C \bar{x}_C}{N} = \frac{10\bar{x}_O + 10\bar{x}_C}{20}. \quad (\text{H.1})$$

L'intérêt d'une telle procédure est de fournir des estimateurs plus précis (ayant une variance plus faible). Pour augmenter la précision, il est nécessaire d'utiliser une variable de stratification fortement liée à la variable d'intérêt. C'est le cas ici puisque intuitivement, on sent bien que les salaires des cadres sont plus élevés que ceux des ouvriers.

5. Pour réaliser le plan stratifié, il reste maintenant à choisir les tailles d'échantillon n_C et n_O , c'est à dire le nombre de cadres et d'ouvriers que l'on va interroger.

- (a) L'allocation proportionnelle propose de choisir les tailles d'échantillon dans les strates de manière à ce que la proportion d'individus dans les strates de l'échantillon soit la même que dans les strates de la population. On choisit donc n_C tel que

$$\frac{n_C}{n} = \frac{N_C}{N} \iff n_C = n \frac{N_C}{N} = 8 * \frac{10}{20} = 4.$$

De même

$$n_O = n \frac{N_O}{N} = 8 * \frac{10}{20} = 4.$$

- (b) Calculons d'abord la variance corrigée pour les deux strates :

$$\begin{aligned} S_C^2 &= \frac{1}{N_C - 1} \sum_{i=1}^{N_C} (X_i - \mu_C)^2 \\ &= \frac{(2225 - 2463.9)^2 + (1616 - 2463.9)^2 + \dots + (2618 - 2463.9)^2}{10 - 9} = 271\,397.7, \end{aligned}$$

et

$$\begin{aligned} S_O^2 &= \frac{1}{N_O - 1} \sum_{i=1}^{N_O} (X_i - \mu_O)^2 \\ &= \frac{(1415 - 1454.9)^2 + (1415 - 1454.9)^2 + \dots + (1484 - 1454.9)^2}{10 - 1} = 7\,249.211. \end{aligned}$$

La variance de $\hat{\mu}$ pour un plan stratifié avec allocation proportionnelle est donnée par :

$$\begin{aligned} \mathbf{V}(\hat{\mu}) &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^H N_h S_h^2 \\ &= \frac{1}{8} \left(1 - \frac{8}{20}\right) \frac{1}{20} (10 * 271\,397.7 + 10 * 7\,249.211) = 10\,449.26. \end{aligned}$$

- (c) Le salaire moyen des cadres et ouvriers interrogés est

$$\bar{x}_O = \frac{1415 + 1415 + 1465 + 1335}{4} = 1\,408.5$$

et

$$\bar{x}_C = \frac{2225 + 1616 + 2456 + 3350}{4} = 2\,411.75.$$

On déduit de (H.1)

$$\hat{\mu} = \frac{10 * 1408.5 + 10 * 2411.75}{20} = 1\,910.125.$$

- (d) Un intervalle de confiance à 95% est donné par

$$\left[\hat{\mu} - z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})}; \hat{\mu} + z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})} \right] = [1\,709.771; 2\,110.479].$$

6. Pour un sondage avec allocation optimale, on choisit les tailles d'échantillon de manière à minimiser la variance de l'estimateur $\hat{\mu}$.

(a) Les tailles d'échantillon sont données par :

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}.$$

Par conséquent

$$n_C = 8 \times \frac{10 * 520.9584}{10 * 520.9584 + 10 * 85.1423} = 6.87,$$

$$n_O = 8 \times \frac{10 * 85.1423}{10 * 520.9584 + 10 * 85.1423} = 1.13.$$

Il faut arrondir $n_C = 7$ et $n_O = 1$.

(b) La variance de $\hat{\mu}$ se calcule à partir de

$$\begin{aligned} \mathbf{V}(\hat{\mu}) &= \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2 \\ &= \frac{1}{20^2} \left(10 \times \frac{10 - 7}{7} 271\,397.7 + 10 \times \frac{10 - 1}{1} 7\,249.211 \right) = 4\,538.905. \end{aligned}$$

(c) Le salaire moyen des cadres et ouvriers interrogés est

$$\bar{x}_O = \frac{1415}{1} = 1415$$

et

$$\bar{x}_C = \frac{2225 + 1616 + 2456 + 3350 + 2600 + 2028 + 3025}{7} = 2\,471.429.$$

On déduit de (H.1)

$$\hat{\mu} = \frac{10 * 1415 + 10 * 2471.429}{20} = 1\,943.215$$

(d) Un intervalle de confiance à 95% est donné par

$$\left[\hat{\mu} - z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})}; \hat{\mu} + z_{0.975} \sqrt{\mathbf{V}(\hat{\mu})} \right] = [1\,811.167; 2\,075.263].$$

7. Le tableau H.2 récapitule les variance de l'estimateur $\hat{\mu}$ en fonction du plan de sondage :

plans	$\mathbf{V}(\hat{\mu})$
Simple	11 997.2
Alloc. prop	10 449.26
Alloc opti	4 538.9

TABLE H.2 – Variances de $\hat{\mu}$ pour les trois plans de sondage étudiés.

Les plans simple et stratifié avec allocation proportionnelle conduisent à des estimateurs possédant des variances similaires. Le plan stratifié avec allocation optimale permet de réduire la variance de manière significative. En regardant les données, on s'aperçoit que ceci vient du fait que les disparités sont beaucoup plus importantes chez les cadres que chez les ouvriers ($S_C^2 = 271\,397.7$ et $S_O^2 = 7\,249.211$), il est donc nécessaire d'interroger plus de cadres que d'ouvriers pour estimer au mieux le salaire moyen dans chacune des catégories. C'est ce que propose l'allocation optimale puisque qu'on interroge 7 cadres et un seul ouvrier.