Cours en Ligne

Dr. Mohammed Bouznit
Maître de Conférences A
Faculté SEGC- Université de Bejaia

Spécialité	**Master Economie Quantitative**
Semestre	**1er**
Intitulé de l'UE	Unité d'enseignement fondamentale
Intitulé de la matière	**Techniques des Sondages**
Crédits	**05**
Coefficients	**02**
Mode d'évaluation	Contrôle continu + Examen final (en présentiel)

Outils d'interaction, d'échange et de collaboration

- **E-mail:** mohammed.bouznit@univ-bejaia.dz
- **Researchgate:**

https://www.researchgate.net/profile/Mohammed-Bouznit

- **Linkedin**: www.linkedin.com/in/mohamed-bouznit-9bb181a2
- **E-learning -Université de Bejaia:**

https://elearning.univ-bejaia.dz/

Objectifs du cours & Prérequis

- **Objectifs**

Le présent cours vise deux objectifs:

- ✓ Présentation des différentes étapes d'une enquêtes de trains (objectifs de l'enquête, base de sondage, méthodes d'échantillonnage, conception du questionnaire ….etc.)
- ✓ Estimation, à partir des données de l'échantillon, les paramètres de la population (moyenne, total, proportion)

- **Prérequis**
- Avoir des connaissances approfondies en:
  - ✓ Statistique descriptive
  - ✓ Probabilités
  - ✓ Inférence statistique

Plan de cours

Chapitre I : Conception générale d'une enquête

1 Généralités

2. Types d'enquête par sondage

2.1. Enquête qualitative

2.2. Enquête quantitative (enquête par questionnaire)

Chapitre II. Phases d'une enquête quantitative (enquête par questionnaire)

1. Formulation de l'énoncé des objectifs

2. Plan d'enquête

3. Population cible et population d'enquête

4. Base de sondage

6. Méthodes de collecte de données

6. Méthodes d'échantillonnage

6.1. Méthodes non probabilistes (Méthodes empiriques)

a Echantillonnage à l'aveuglette

b. Echantillonnage à participation volontaire

c. Echantillonnage par quotas

d. Echantillonnage probabiliste modifié

6.2.Méthodes d'échantillonnage probabilistes

a. Sondage Aléatoire Simple (SAS)

b. Sondage stratifié

c. Sondage à plusieurs degrés

d. Sondage en grappes

7.Conception du questionnaire

8. Calcul de la taille de l'échantillon

Chapitre III: Estimation

III.1. Cas d'un plan d'échantillonnage aléatoire simple (SAS)

1. Estimation de la moyenne Simple d'une population finie
   a. Variance de la moyenne
   b. Estimation de la variance de la moyenne
2. Estimation par intervalle de confiance de la moyenne simple
3. Estimation du total (T) d'une population finie
   a. Variance de l'estimateur du total
   b. Estimation de la variance de l'estimateur du total
4. Estimation par intervalle de confiance du total
5. Estimation d'une proportion
   a. Variance de l'estimateur de la proportion
   b. Estimation de la variance de $V(\widehat{P})$
6. Intervalle de confiance de la proportion

III.2. Cas d'un plan d'échantillonnage stratifié

1. Notation
2. Estimation de la moyenne simple de la population
   a. Variance de $\hat{\bar{Y}}_{st}$
   b. Estimation de $V\left(\hat{\bar{Y}}_{st}\right)$
3. Estimation par Intervalle de confiance de $\bar{Y}_{st}$
4. Estimation du total (T) d'une population finie
   a. Variance de l'estimateur du total
   b. Estimation de la variance de l'estimateur du total
5. Estimation par intervalle de confiance du total
6. Estimation de la proportion d'une population
   a. Variance de l'estimateur de la proportion
   b. Estimation de la variance de $V\left(\hat{P}_{st}\right)$
7. Estimation par intervalle de confiance de la proportion

Références bibliographiques

- 1. Ardilly P., (1994). Les techniques de sondage, Ed. TCHNIP, Paris France
- 2. Ardilly P., (2006), Les techniques de sondage, ISBN :9782710808473, 2710808471 ; Éditeur: Technip
- 3. Durand C., (2002). Notes de cours, deuxième partie « L'échantillonnage : La gestion du terrain ». Méthodes de sondage - SOL3017. Département de sociologie. Université de Montréal. Canada
- 4. Dussaix A-M. et Grosbras J-M. (1992), Exercices de sondage, Economica
- 5. Haziza D., (2008). Notes de cours : Échantillonnage STT-2000. Département de mathématiques et de statistique Université de Montréal.
- 6. Statistique Canada (2010). Méthodes et pratiques d'enquête, publié en 2010 par Statistique Canada sous le N° 12-587-X au catalogue. https://www150.statcan.gc.ca/n1/fr/pub/12-587-x/12-587-x2003001-fra.pdf?st=2jx8cqh9
- 7. Salès-Wuillemin, E., (2006). Méthodologie de l'enquête, in : M., Bromberg et A., Trognon (Eds.) Psychologie Sociale 1, Presses Universitaires de France, 45-77.

Chapitre I : Conception générale d'une enquête

- **1.1 Généralités :**

Dans ce qui suit, on considère une population bien déterminée et une variable formalisant l'information qui nous intéresse, appelée variable d'intérêt, définie sur chaque individu de cette population.

- Qu'est ce qu'une population ?

La population est l'ensemble des unités, individus, de mêmes natures caractérisées par une variable d'intérêt qu'on peut la mesurer sans ambigüité.

- Qu'est ce qu'un échantillon ?

Un échantillon est un sous ensemble de la population sur lequel on effectue une étude statistique. Selon Durand (2002), « Un échantillon est constitué dès que l'on sélectionne un nombre restreint d'unités à partir d'une population d'unités. Cette population doit être définie de telle manière que l'on peut toujours savoir si une unité fait partie de la population **».**

- Qu'est-ce qu'une enquête par sondage ?

Selon Statistique Canada (2010), « l'enquête est une activité organisée et méthodique de collecte de données sur des caractéristiques d'intérêt d'une partie ou d'une totalité des unités d'une population à l'aide de concepts, de méthodes et de procédures bien définis. Elle est suivie d'un exercice de compilation permettant de présenter les données recueillies sous une forme récapitulative utile ».

- **Exemple** :
- la population: des ménages d'une wilaya, (unité statistique : un ménage)
- Selon Ardilly (1994), une population est définie par la conjonction de 4 facteurs, à savoir:
  - ✓ Sa nature (individu, un logement, une entreprise, un ménage, un hôpital, une commune, ….etc.)
  - ✓ Ses caractéristiques (le sexe, le type de logement, le secteur d'activité, taille de ménage, qualité de soin… etc.)
  - ✓ Sa localisation (Alger, lieu d'implantation, lieu de résidence….etc.)
  - ✓ La date à laquelle on la considère.
- Dans la diapo suivante , un schéma qui retrace les éléments qu'il fallait tenir en compte le moment de la conception d'une enquête de terrain.

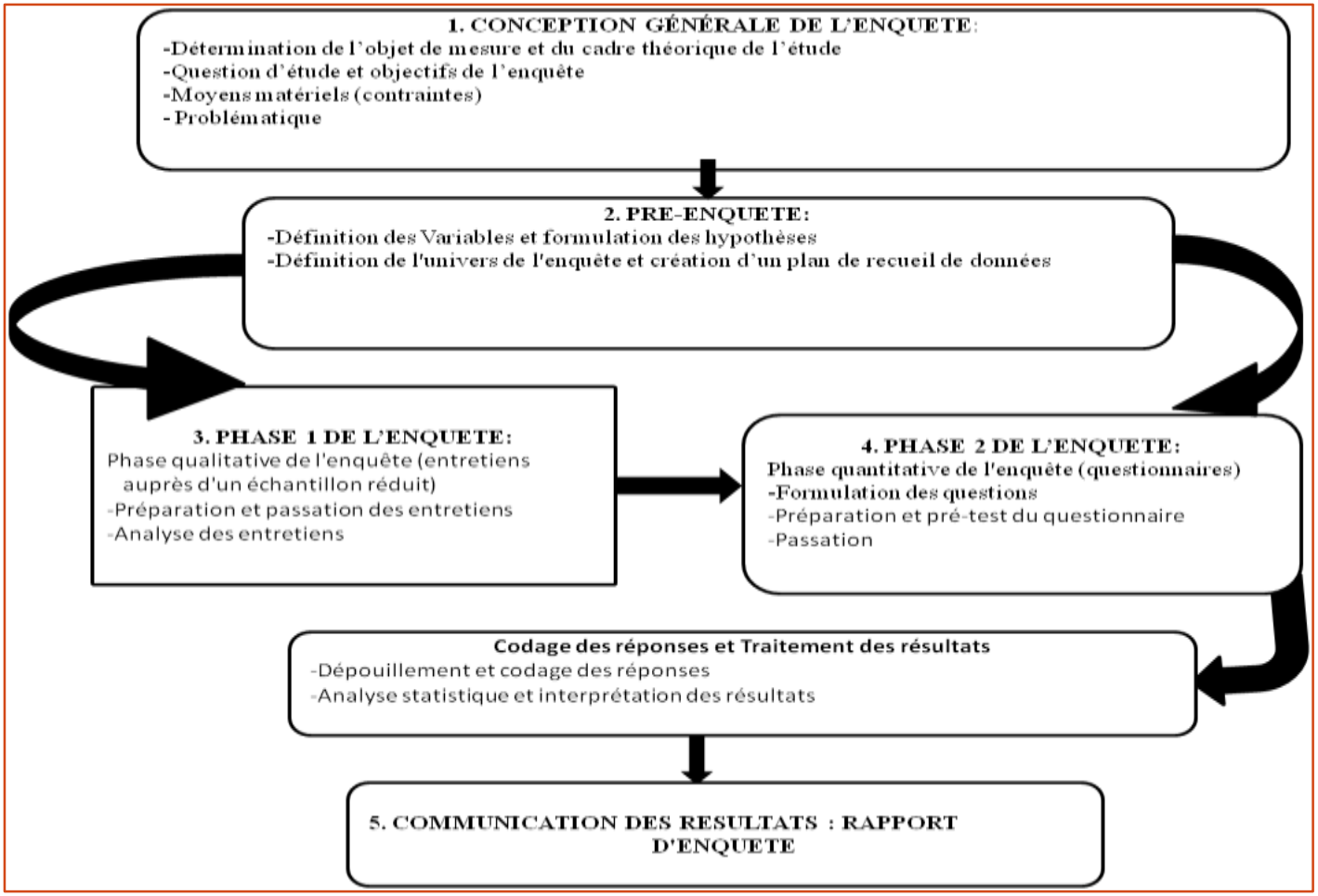


Figure 1. Démarche générale pour la réalisation d'une enquête par entretien et questionnaire

Source : Salès-Wuillemin, E. (2006). Méthodologie de l'enquête, in : M., Bromberg et A., Trognon (Eds.) *Psychologie Sociale 1*, Presses Universitaires de France, 45-77.

1.2. Types d'enquête par sondage

- **Enquête qualitative :** elle vise à recueillir des informations qualitatives à partir d'un échantillon de taille réduite. Il s'agit d'avoir des informations pour comprendre soit l'attitude, le comportement, la motivation, ou la perception d'une population. Ce type d'enquête requiert un guide d'entretien pour effectuer des entretiens individuels ou collectifs.
- **Enquête quantitative (enquête par questionnaire)**: c'est une enquête par questionnaire qui permet de recueillir des données quantifiables sur un échantillon représentatif d'une population donnée. Ce type d'enquêtes est largement utilisé pour étudier, mesurer et analyser des questions-problématiques- socioéconomiques. Les étapes requises pour mener une enquête quantitative (enquête par questionnaire) sont présentées dans le chapitre 2 et chapitre III

Chapitre II. Phases d'une enquête quantitative (enquête par questionnaire)

- On distingue deux phases; la phase préliminaire et la phase opérationnelle. La phase préliminaire regroupe notamment, la formulation de l'énoncé des objectifs, la sélection d'une base de sondage, et le choix d'un plan d'échantillonnage. La phase opérationnelle est relative à la conception du questionnaire, la collecte des données, la saisie et codage des données, la vérification et imputation, estimation, l'analyse de données, la diffusion des données et la documentation de l'enquête. Par conséquent, la fiabilité des données, l'exactitude et la significativité des résultats sont étroitement liées à la pertinence de ces étapes suivies. Ces dernières sont expliquées brièvement dans les diapos ci-dessous :

1.Formulation de l'énoncé des objectifs

- Cette étape concerne les tâches qui sont dédiées à établir les besoins d'information de l'enquête, les principaux utilisateurs et les principales utilisations des données, les définitions opérationnelles qui seront utilisées, les principaux concepts et les définitions opérationnelles, les sujets particuliers à considérer et le plan d'analyse. Il est vivement recommandé que chaque enquête doit être bien planifiée. Il s'agit de préciser de manière très claire les objectifs de l'enquête, ce qui permettra une meilleure orientation des étapes ultérieures en garantissant que les résultats dégagés correspondent aux objectifs originaux. Les définitions opérationnelles particulières concernent la définition précise de la population cible et la délimitation de la période de référence. Cela permet d'identifier facilement les unités statistiques, et de mesurer de façon précise la variable d'intérêt.

2. Plan d'enquête

- Il s'agit de mettre en place un mécanisme opérationnel pour mener l'enquête. On distingue deux types de mécanismes : Enquête-échantillon Enquête par recensement
- Plusieurs facteurs seront pris en considération pour choisir entre enquête échantillon et enquête par recensement à savoir :
  - ✓ le budget et les ressources disponibles,
  - ✓ la taille de la population et des sous-populations (petite population/ recensement ; grande population/ enquête échantillon).
  - ✓ l'échéancier des résultats de l'enquête.
  - ✓ L'erreur d'échantillonnage (seulement dans l'enquête-échantillon).
  - ✓ L'erreur non due à l'échantillonnage (dans l'enquête-échantillon et le recensement).
  - ✓ Besoins spécialisés: dans certaines études, l'enquête-échantillon s'impose parce que la collecte de données nécessite un personnel hautement qualifié, un matériel de mesure qui coûte cher. Par exemple, le cas des études sur la tension artérielle, le groupe sanguin, ou la condition physique des répondants, ne peuvent être réalisées sans qu'il y ait un professionnel de la santé (enquête échantillon).
  - ✓ La création d'une base de sondage exhaustive (dans ce cas le recours au recensement est une obligation. A titre d'exemple en Algérie, le Recensement Général de la Population et de l'habitat réalisé tous les 10 ans par l'Office National des Statistiques.

3.Population cible et population d'enquête

- la population cible comporte l'ensemble des individus sur lesquels devrait porter l'étude. Cependant, la population d'enquête est celle qui couvre l'enquête, donc elle regroupe les individus qui seront accessibles au moment de l'enquête. Les résultats de l'enquête seront généralisés uniquement sur la population d'enquête. pour que les résultats soient plus fiables, il faut que la population cible soit identique à la population d'enquête.
- Exemple 1 : enquête sur le revenu des ménages
- Population cible : toute la population résidante en Algérie au 30 Mars 2020.
- Population d'enquête: la population résidante en Algérie au 30 Mars 2020 en excluant les personnes qui ont une résidence permanente à l'étranger.
- Exemple 2 : enquête sur la performance des entreprises industrielles
- Population cible : toutes les PME industrielles de la Wilaya de Bejaia au 1 Avril 2020.
- Population d'enquête : toutes les PME industrielles de la Wilaya de Bejaia au 1 Avril 2020 à l'exception celles qui emploient moins de 5 travailleurs.

4.Base de sondage

- La base de sondage est une liste exhaustive qui permet d'identifier et donner accès aux unités de la population d'enquête. De ce fait, une base de sondage doit comporter des données d'indentification, des données de communication, des données de classification, et des données de mise à jour,
- **Remarque 1** : On distingue trois genres d'unités, mais dans certaines enquêtes elles sont les mêmes:
- Unité d'échantillonnage : elle fait partie de la base de sondage, mais aussi elle a une possibilité qu'elle soit sectionnée dans l'échantillon.
- Unité d'analyse: l'unité à laquelle s'intéresse l'étude.
- Unité déclarante : l'unité avec qui communique l'enquêteur pour avoir l'information recherchée.
- **Remarque 2** : trois types de bases de sondage peuvent être distingués :
- la nomenclature : elle comporte toutes les informations nécessaires pour accéder à la population d'enquête
- la base aléatoire : est une liste des régions géographiques qui donnent indirectement accès à des unités (comme les quartiers d'une localité)
- Une base de sondage multiple : est comporte les nomenclatures et les bases aléatoires.

5. Méthodes de collecte de données

- Il s'agit des mécanismes qui permettent de recueillir l'information auprès des unités de l'échantillon**.** Le taux de réponse, le coût et la précision des données collectées sont les principaux facteurs pour choir telle ou telle méthode de collecte de données. Les méthodes les plus utilisées sont :
- Interview sur place
- Interview téléphonique
- Auto-dénombrement
- Observation directe

6. Méthodes d'échantillonnage

- Selon Statistique Canada (2010), « L'échantillonnage est un moyen de sélectionner un sous-ensemble d'unités dans une population aux fins de la collecte de l'information sur ces unités pour formuler des inférences sur l'ensemble de la population ». On distingue deux catégories de plans d'échantillonnage :
- méthodes d'échantillonnage non probabilistes (méthodes empiriques)
- méthodes d'échantillonnage probabilistes

6.1. Méthodes non probabilistes (Méthodes empiriques)

- Les méthodes d'échantillonnage empiriques consistent à sélectionner un échantillon dans une population à l'aide une méthode subjective (non-aléatoire). Ce plan d'échantillonnage est largement utilisé dans les études de marché, car il donne des résultats rapides à prix raisonnable. On peut l'appliquer également dans des études qui servent :
  - ✓ d'outil pour donner des idées,
  - ✓ d'étape préliminaire à l'élaboration d'une enquête par échantillon probabiliste
  - ✓ étape de suivi pour aider à comprendre les résultats d'une enquête par échantillonnage probabiliste

Quatre méthodes d'échantillonnage non probabilistes peuvent être distinguées:

a. Echantillonnage à l'aveuglette : On fait appel à cet échantillonnage lorsque la population est homogène et les unités de l'échantillon sont sélectionnées de façon arbitraire. En effet, les unités de la population sont semblables, alors n'importe quelle unité peut être choisie dans l'échantillon.

Exemple : dans une interview de l'homme de la rue, l'enquêteur peut interroger n'importe quel passant.

b. Echantillonnage à participation volontaire: cette méthode fait appel à des répondants volontaires. Ces derniers doivent faire l'objet d'un examen pour obtenir un ensemble de caractéristiques qui conviennent aux objectifs de l'enquête

Exemple 1: les personnes atteintes d'une maladie en particulier (expériences médicales)

Exemple 2 : au cours d'une émission radio, une question fait l'objet d'une discussion et les auditeurs sont invités à téléphoner pour exprimer leurs opinions. Certainement, seuls les auditeurs qui sont intéressés par le sujet vont répondre.

c. Echantillonnage par quotas : C'est l'une des méthodes échantillonnages non probabilistes la plus utilisée. Elle consiste à sélectionner un nombre déterminé d'unités (quotas) dans diverses sous populations. Les quotas seront établis selon des proportions de la population. C'est à dire, si nous avons 60 hommes et 40 femmes dans une population, par exemple s'il faut tirer un échantillon de 10 personnes, alors 6 hommes et 4 femmes doivent être interviewés.

l'échantillonnage par quotas peut être considéré meilleur par rapport à d'autres fromes d'échantillonnage puisque la structure de l'échantillon est identique à celle de la population

l'échantillonnage par quotas ressemble à l'échantillonnage stratifié parce que des unités semblables sont regroupées, cependant la méthode de sélection des unités est différente (une sélection aléatoire des unités pour un sondage stratifié sera utilisée).

d. Echantillonnage probabiliste modifié: C'est une combinaison d'échantillonnage probabiliste et non probabiliste. Les premières étapes sont habituellement axées sur l'échantillonnage probabiliste, tandis que la dernière étape est un échantillonnage non probabiliste (souvent on fait appel à une méthode par quotas)

Exemple : à l'aide d'un plan d'échantillonnage probabiliste, on sélectionne des secteurs géographiques. Ensuite, dans chaque région sectionnée, un échantillon de personnes peut être choisi en utilisant la méthode par quotas.

6.2.Méthodes d'échantillonnage probabilistes

- Un sondage est dit probabiliste (aléatoire) si toutes les unités de la population ont une probabilité non nulle et connue d'être dans l'échantillon. Le choix de l'échantillon est déterminé par le hasard ce qui permet de formuler des inférences sur la population. En outre, le choix de l'échantillon nécessite une base de sondage exhaustive.
- Les principaux critères de l'échantillonnage probabiliste :
- La sélection des unités de la population d'enquête est aléatoire
- Toutes les unités de la population d'enquête ont une probabilité d'inclusion non nulle et connue dans l'échantillon.
- Il n'est pas nécessaire que toutes les unités aient la même probabilité d'inclusion
- Les méthodes d'échantillonnage probabilistes les plus utilisées sont présentées ci-dessous:

a. Sondage Aléatoire Simple (SAS)

- Le SAS est une méthode d'échantillonnage probabiliste qui consiste à trier un échantillon à partir d'une population finie (de taille N). On sélectionne un échantillon de taille n où chaque unité statistique a la même probabilité d'inclusion (c'est à dire tous les individus de la population ont la même probabilité non nulle d'être dans l'échantillon), et sans aucune manipulation préalable dans la population. Ce type de tirage ne nécessite pas l'existence de l'information auxiliaire pour qu'il soit appliqué. En outre, l'échantillon peut être tiré soit avec remise (tirage non exhaustif) ou sans remise (triage exhaustif)
- **Remarque** : Dans un SAS, la portabilité d'inclusion $\pi_i = \frac{n}{N} = \frac{1}{C_N^n}$,. tq. C_N^n : le nombre de d'échantillons possibles de taille n parmi une population de tille finie N
- $\sum_{i=1}^{N} \pi_i = N.\frac{n}{N} = n$
- $\frac{n}{N}$: Le taux de sondage
- Pour réaliser un tirage aléatoire simple, il faut :
  - Se procurer d'une base de sondage complète et à jour (exhaustive)
  - Numéroter les individus de 1 à N
  - Se donner la taille de l'échantillon n
  - Tirer n nombre d'individus compris entre 1 et N en accordant à chaque individu la même probabilité d'être choisi

b. Sondage stratifié

- Le sondage stratifié consiste à diviser la population d'enquête en groupes homogènes (appelés strates), qui doivent être mutuellement exclusifs, et à tirer indépendamment un échantillon aléatoire dans chaque strate. La stratification permet d'améliorer la précision des estimateurs.

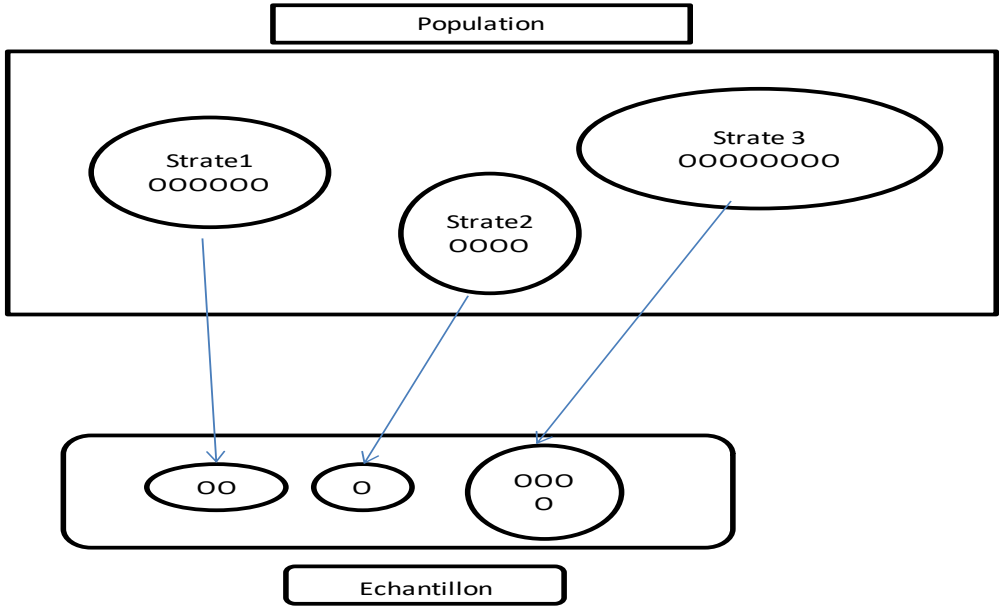


Figure1: Processus de sondage stratifié

c. Sondage à plusieurs degrés

La sélection de l'échantillon se fait en plusieurs étapes. Il consiste à diviser la population en groupes d'individus qui soient disjoints, puis on utilise un plan d'échantillonnage, par exemple le SAS, pour tirer un certain nombre de groupes (appelés unités primaires, notée UP). Ensuite, on sélectionne, par un SAS par exemple, certains individus de chaque unité primaire pour constituer un échantillon représentatif de la population d'enquête (les individus tirés sont appelés des unités secondaires, notées US).

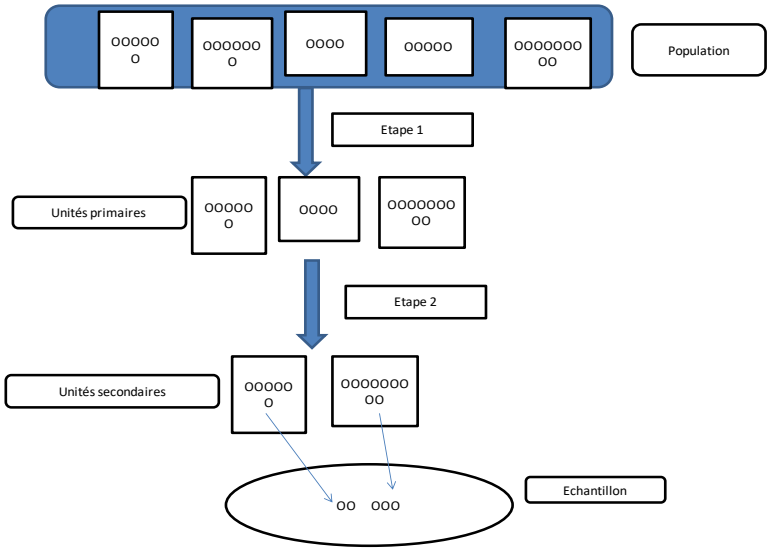


Figure 3: Processus du sondage à plusieurs degrés

d. Sondage en grappes

C'est un cas particulier du sondage à plusieurs degrés où le processus de sélection de l'échantillon se réalise en deux étapes. Il consiste à diviser la population en groupes d'individus qui soient disjoints, puis à tirer tous les individus des unités primaires, c'est-à-dire l'échantillon comporte tous les individus des groupes qui sont déjà tirés à la première étape.

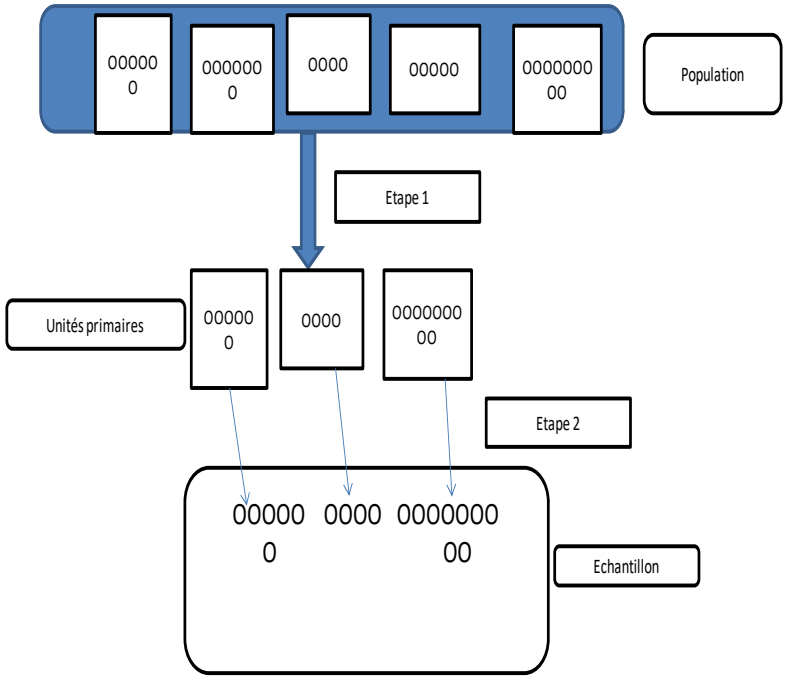


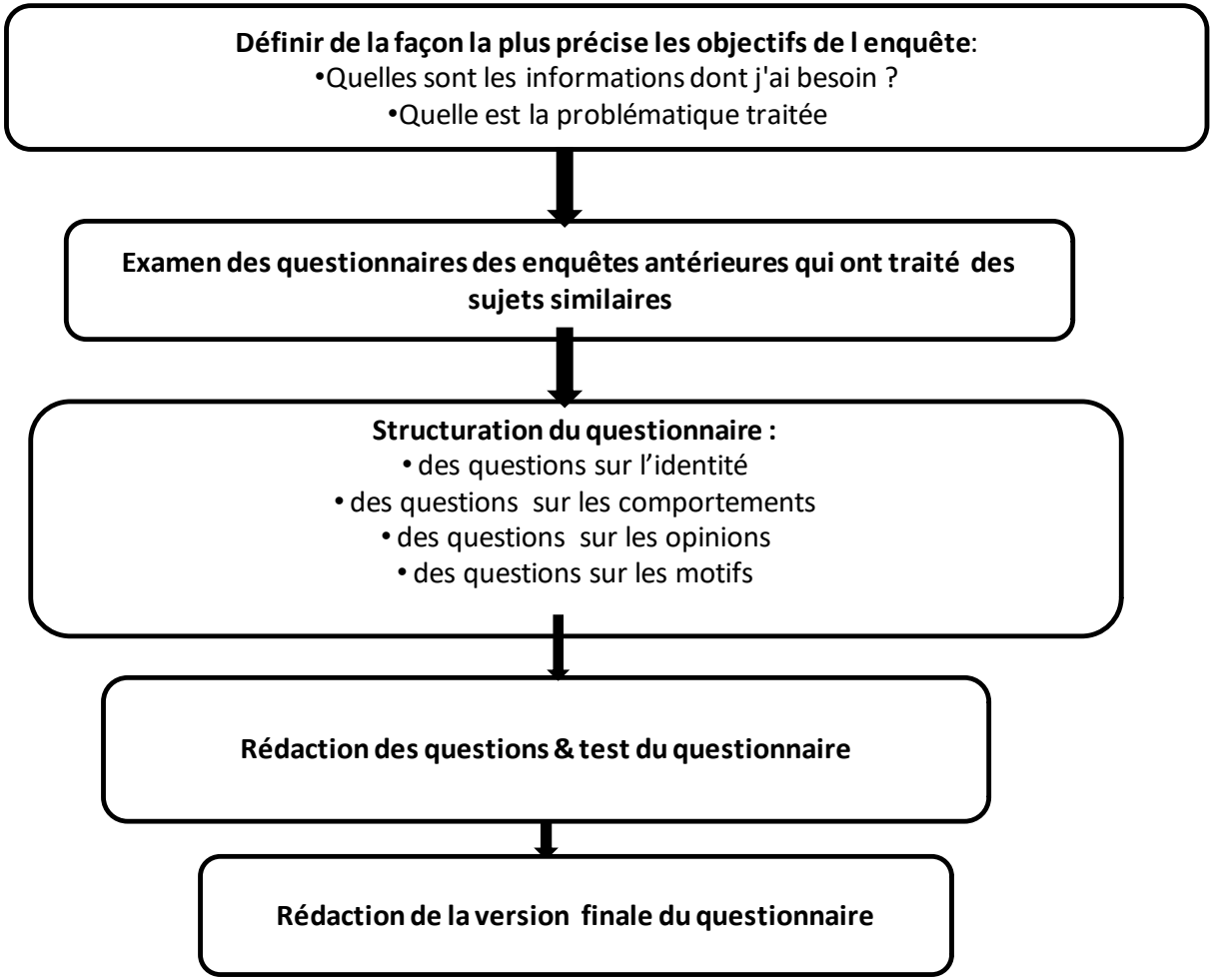
Figure 4: Processus de sondage en grappes

7. Conception du questionnaire

- Selon Statistique Canada (2010), « Un questionnaire est un groupe ou une séquence de questions conçues pour obtenir d'un répondant de l'information sur un sujet ». De ce fait, les questions doivent être rédigées de façon à collecter des informations qui répondent entièrement aux objectifs de l'enquête. Pour ce faire, trois types de questions sont employés:
  - ✓ Questions fermées
  - ✓ Questions ouvertes
  - ✓ Question semi fermées /semi ouvertes
- **Remarque**

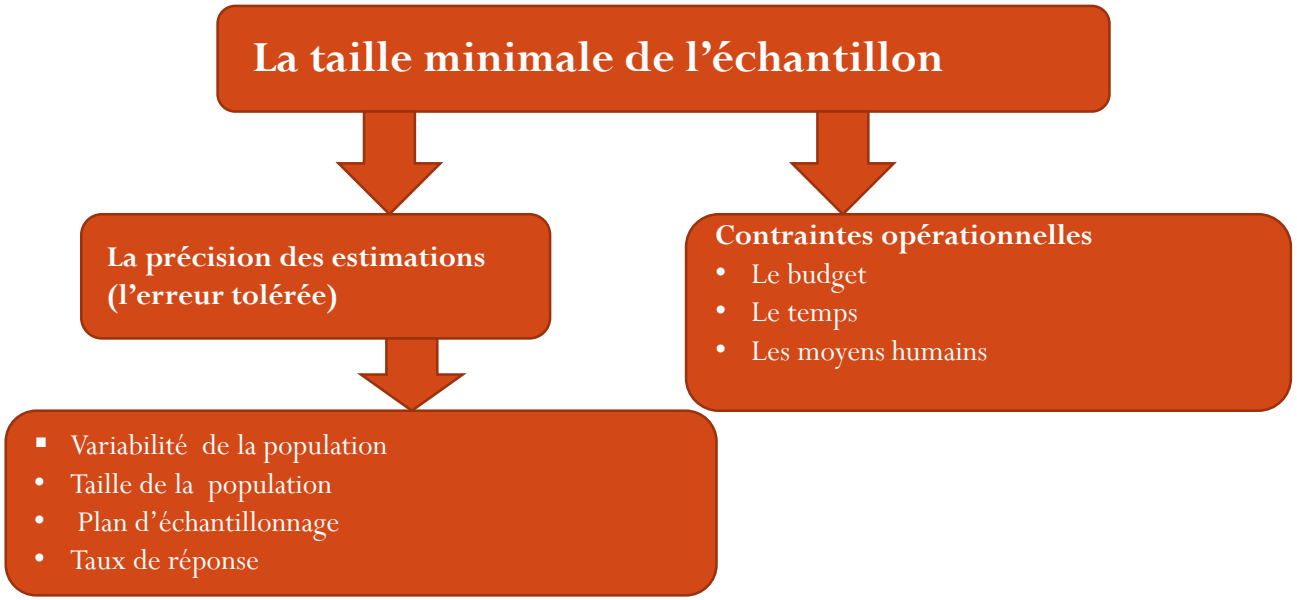
Le lien ci-dessous vous permet de consulter un fichier qui montre de façon détaillée, avec des exemples, comment concevoir un questionnaire. https://central.bac-lac.gc.ca/.item?id=12-587-x2003001-fra&op=pdf&app=Library

- Etapes à suivre pour concevoir un questionnaire sont illustrées dans le schéma ci-dessous :



8. Calcul de la taille de l'échantillon

- Dans une enquête par sondage, le choix de la taille de l'échantillon est principalement tributaire de la précision des estimations de l'enquête. La précision des estimations, à son tour, est liée à la variabilité de la population, la taille de la population, le plan d'échantillonnage et le taux de réponse à l'enquête. Pour calculer la taille minimale de l'échantillon, nous tenons compte les facteurs qui sont illustrés dans le schéma ci-dessous :



Formule de Calcul de la taille minimale de l'échantillon

- La taille minimale de l'échantillon se calcule à partir de la formule l'écart type de l'estimation de la moyenne estimé pour un plan d'échantillonnage aléatoire simple (EAS).
- L'estimation de la moyenne pour un (EAS) est notée $\hat{\bar{Y}}$
- $\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Pour un tirage exhaustive (sans remise), l'écart type de $\hat{\bar{Y}}$: $ET(\hat{\bar{Y}}) = \sqrt{(1-\frac{n}{N})}\frac{\hat{s}}{\sqrt{n}}$
- N : taille de la population ; n : la taille de l'échantillon ; $\hat{s}$: la racine carrée de l'estimation de la variance de la variable d'intérêt Y (estimation de l'Ecart type de la population) .
- L'intervalle de confiance de la moyenne de la population est égale à : $\hat{\bar{Y}} \mp Z_{1-\frac{\alpha}{2}}\sqrt{(1-\frac{n}{N})}\frac{\hat{s}}{\sqrt{n}}$.

où $Z_{1-\frac{\alpha}{2}}$ est le fractile d'une loi normale centrée réduite au seuil de confiance de $1-\frac{\alpha}{2}$

donc, $Z_{1-\frac{\alpha}{2}}\sqrt{(1-\frac{n}{N})}\frac{\hat{s}}{\sqrt{n}}$ est la marge d'erreur (e) de $\hat{\bar{Y}}$.

- On pose $e = Z_{1-\frac{\alpha}{2}}\sqrt{(1-\frac{n}{N})}\frac{\hat{s}}{\sqrt{n}} \Rightarrow e^2 = Z^2_{1-\frac{\alpha}{2}}(1-\frac{n}{N})\frac{\hat{s}^2}{n} \Rightarrow n = \frac{\hat{s}^2 Z^2_{1-\frac{\alpha}{2}}}{e^2 + \frac{\hat{s}^2 Z^2_{1-\frac{\alpha}{2}}}{N}}$

D'où, pour un taux de réponse de 100% et une marge d'erreur tolérée (e), la taille minimale de

l'échantillon doit être supérieure ou égale à $\frac{\hat{s}^2 Z^2_{1-\frac{\alpha}{2}}}{e^2 + \frac{\hat{s}^2 Z^2_{1-\frac{\alpha}{2}}}{N}}$

- $\hat{s}^2$: est une estimation de la variabilité de la population.

Remarque

- Si la variable d'intérêt est la proportion de la population, la variabilité de la population $\hat{s}^2 = \hat{P}(1-\hat{P})$, alors la taille minimale de la population pour un taux de réponse de 100% :
- $n \geq \frac{\hat{P}(1-\hat{P}) Z^2_{1-\frac{\alpha}{2}}}{e^2 + \frac{\hat{P}(1-\hat{P}) Z^2_{1-\frac{\alpha}{2}}}{N}}$
- Si $\hat{s}^2$ est inconnue, l'estimation de la variabilité de la population sera calculée en prenant $\hat{P} = 0{,}50$ (variabilité maximale de la population)

Exemple (Extrait du Livre « Méthodes et pratiques d'enquête », publié en 2010 par Statistique Canada sous le N° 12-587-X au catalogue, ISBN 978-1-100-95206-2)

- L'éditeur d'une revue veut obtenir une estimation de la satisfaction des lecteurs en général. Il serait possible de communiquer avec les 2500 abonnés à l'aide d'un questionnaire envoyé par la poste, mais l'éditeur a décidé d'interviewer un échantillon aléatoire simple par téléphone à cause des contraintes de temps. Combien de lecteurs faudrait-il interviewer?
- Voici certaines hypothèses:
- l'éditeur sera satisfait si la proportion de la population réelle est à $\pm 0{,}10$ de la proportion de la population estimée, compte tenu des résultats de l'échantillon, c.-à-d. que la marge d'erreur nécessaire, $e = 0{,}10$;
- l'éditeur veut obtenir un niveau de confiance de 95 % dans les estimations de l'enquête (c.-à-d. qu'il y aurait seulement une chance sur 20 d'obtenir un échantillon qui donne une estimation hors de l'étendue $\hat{P} \pm 0{,}10$, donc $Z_{1-\frac{\alpha}{2}} = 1{,}96$);
- un EAS sera utilisé;
- un taux de réponse de 65 % environ est prévu, c.-à-d. que $r = 0{,}65$;
- étant donné qu'il n'y a pas d'estimation de $\hat{P}$ disponible, le degré de satisfaction de la clientèle est donc supposé être $\hat{P} = 0{,}5$.

- **Solution**
- Calcul de la taille initiale de l'échantillon, n_1 :
- $n_1 = \frac{\hat{P}(1-\hat{P})\, Z^2_{1-\frac{\alpha}{2}}}{e^2 + \frac{\hat{P}(1-\hat{P})\, Z^2_{1-\frac{\alpha}{2}}}{N}} = \frac{05(1-0{,}5)1{,}96}{0{,}10^2 + \frac{05(1-0{,}5)1{,}96}{2500}} \approx 96$
- Ajustement de la taille de l'échantillon pour tenir compte :
- de la taille de la population : $n_2 = n_1 \frac{N}{N+n_1} = 96 \frac{2500}{(2500+96)} = 92$
- de l'effet de plan d'échantillonnage(EPE : $n_3 = EPE * n_2 = 1 * 92 = 92$ (pour un plan EAS :EPE=1, Stratifié. EPE< 1, en grappe ou plusieurs degrés :EPE> 1)
- du taux de réponse : $n = \frac{n_3}{r} = \frac{92}{0{,}65} = 142$
- D'où, la taille minimale de l'échantillon doit être supérieure ou égale à 142 abonnés

Chapitre III: Estimation

- L'estimation consiste à estimer les paramètres inconnus d'une population finie à partir des données de l'échantillon . De ce fait, ce chapitre comporte des techniques statistiques permettant d'estimer la moyenne, le total et la proportion et l'estimation de leurs précisions , et ce en tenant compte deux types de plans d'échantillonnage:
  - Cas d'un plan d'échantillonnage aléatoire simple ;
  - Cas d'un plan d'échantillonnage stratifié.

III. 1. Cas d'un plan d'échantillonnage aléatoire simple (SAS)

1. Estimation de la moyenne Simple d'une population finie

- La moyenne théorique d'une population finie, de taille N, est donnée par :
- $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ (Inconnue)

Tel que : Y: Variable d'intérêt (le caractère étudié)

i: $1, \ldots \ldots, N$: Individus de la population

Y_i: la valeur de la variable d'intérêt associé à l'individu i

- On note $\bar{y}$ la moyenne simple de la variable d'intérêt calculée à partir des données de l'échantillon, telle que:
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

y_i: Valeur de la variable étudiée (variable aléatoire) associée à l'individu (i) dans l'échantillon. Donc, y_i peut prendre les valeurs $Y_1, Y_2, \ldots \ldots ou\ Y_N$

- Soit $\hat{\bar{Y}}$ l'estimateur de la moyenne simple d'une population, alors :
- $\hat{\bar{Y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: est un estimateur sans biais de $\bar{Y}$

a. Variance de la moyenne

- Soient $\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ la variance de la variable d'intérêt et $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ la variance corrigée
- $V\left(\hat{\bar{Y}}\right) = V(\bar{y}) = \frac{\sigma_Y^2}{n}$ cas d'un tirage avec remise (non-exhaustif) (PEAR)
- $V(\hat{\bar{Y}}) = V(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma_Y^2}{n} = \frac{N-n}{N}\frac{S_Y^2}{n} = (1\text{-}t_x)\frac{S_Y^2}{n}$ cas d'un tirage sans remise (exhaustif) (PESR)

b. Estimation de la variance de la moyenne

- $\hat{V}(\hat{\bar{Y}}) = \hat{V}(\bar{y}) = \frac{\hat{S}_y^2}{n}$ cas d'un tirage avec remise (non-exhaustif) (PEAR)
- $\hat{V}(\hat{\bar{Y}}) = \hat{V}(\bar{y}) = \frac{N-n}{N}\frac{\hat{S}_y^2}{n} = (1\text{-}t_x)\frac{\hat{S}_y^2}{n}$ cas d'un tirage sans remise (exhaustif) (PESR)
- tq : $\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ est la variance corrigée-échantillon estime sans biais σ_Y^2 et S_Y^2 pour PEAR et PESR respectivement.
- **Remarque** : Si $n > 30$ alors, PEAR $\Leftrightarrow$PESR

2. Estimation par intervalle de confiance de la moyenne simple

- $IC_{\bar{Y}} = [\hat{\bar{Y}} - Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\bar{Y}})} \;\; ; \;\; \hat{\bar{Y}} + Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\bar{Y}})}]$

$= [\bar{y} - Z_{1-\frac{\alpha}{2}}\sqrt{V(\bar{y})} \;\; ; \;\; \bar{y} + Z_{1-\frac{\alpha}{2}}\sqrt{V(\bar{y})}]$

$= [\bar{y} - Z_{1-\frac{\alpha}{2}}\frac{\hat{S}_y}{\sqrt{n}} \;\; ; \;\; \bar{y} + Z_{1-\frac{\alpha}{2}}\frac{\hat{S}_y}{\sqrt{n}}]$ pour un PEAR

$=[\bar{y} - Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}} \;\; ; \;\; \bar{y} + Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}}]$ pour un PESR

- $Z_{1-\frac{\alpha}{2}}$: représente le fractile d'ordre $(1-\frac{\alpha}{2})$ de la loi normale centrée réduite.

Exemple 1.

- Soit une population de 100 entreprises, et on veut estimer le nombre moyen d'employés par entreprise. Pour ce faire, un échantillon de 7 entreprises a été sélectionné en utilisant un plan de sondage aléatoire simple sans remise, les résultats sont consignés dans le tableau ci-dessous:

N° Entp.	1	2	3	4	5	6	7
Nombre d'employés	10	5	15	20	12	8	10

- Donner une estimation ponctuelle de la moyenne des employés de 100 entreprises
- Déterminer un intervalle de confiance au niveau de 95% pour la moyenne de la population

- **Réponse**

1. Estimation ponctuelle de la moyenne des employés de 100 entreprises:

Nous avons:

$\bar{Y}$: la vraie moyenne de la population

$\hat{\bar{Y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: est un estimateur sans biais de $\bar{Y}$

Donc; $\hat{\bar{Y}} = \bar{y} = \frac{1}{7}\sum_{i=1}^{7} y_i = \frac{1}{7}[10 + 5 + 15 + 20 + 12 + 8 + 10] = 11{,}42 \approx 11$ employés/ entreprise

2. Intervalle de confiance de $\bar{Y}$

- $IC_{\bar{Y}} = [\bar{y} - Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}} \quad ; \quad \bar{y} + Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}}]$ pour un PESR
- $t_x = \frac{n}{N} = \frac{7}{100}$; $Z_{1-\frac{\alpha}{2}} = 1{,}96$; $\hat{S}_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{7}(y_i - \bar{y})^2} = \sqrt{23{,}952}$

D'où

$IC_{\bar{Y}} = [11{,}42 - 1{,}96\sqrt{(1-0{,}07)}\left(\sqrt{\frac{23{,}952}{7}}\right) ; 11{,}42 + 1{,}96\sqrt{(1-0{,}07)}(\sqrt{\frac{23{,}952}{7}})]$

3. Estimation du total (T) d'une population finie

- Soit $T = \sum_{\alpha=1}^{N} Y_\alpha = N\frac{1}{N}\sum_{\alpha=1}^{N} Y_\alpha = N\bar{Y}$ le total de Y dans la population, alors son estimateur est égal à : $\hat{T} = N\bar{y}$
- **Exemples** : Le total de Y dans la population est le revenu total, le nombre total des chômeurs, la consommation totale….etc.

a. Variance de l'estimateur du total

- $V(\hat{T}) = V(N\bar{y}) = N^2 V(\bar{y})$

$= N^2 \frac{\sigma_Y^2}{n}$ pour un PEAR

$= N^2 \frac{N-n}{N-1}\frac{\sigma_Y^2}{n}$ $= N^2 \frac{N-n}{N}\frac{S_Y^2}{n} = N^2\,(1\text{-}t_x)\frac{S_Y^2}{n}$ pour un PESR

b. Estimation de la variance de l'estimateur du total

- $\hat{V}(\hat{T}) = N^2 \frac{\hat{S}_y^2}{n}$ pour un PEAR

$= N^2\,(1\text{-}t_x)\frac{\hat{S}_y^2}{n}$ pour un PESR

- avec $\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

4. Estimation par intervalle de confiance du total

- $IC_T = [\hat{T} - Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{T})}\ ;\quad \hat{T} + Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{T})}\ ]$

  $= [\hat{T} - Z_{1-\frac{\alpha}{2}}\, N\frac{\hat{S}_y}{\sqrt{n}}\ ;\quad \hat{T} + Z_{1-\frac{\alpha}{2}}\, N\frac{\hat{S}_y}{\sqrt{n}}]$ pour un PEAR

- $=[\hat{T} - Z_{1-\frac{\alpha}{2}}\, N\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}};\quad \hat{T} + Z_{1-\frac{\alpha}{2}}\, N\sqrt{(1-t_x)}\frac{\hat{S}_y}{\sqrt{n}}]$ pour un PESR

5. Estimation de la proportion d'une population finie

- Soit une caractéristique **C** et soit la variable dichotomique Y définie comme ci-dessous :
- $$\boldsymbol{Y_\alpha} = \begin{cases} 1 : si\ l'individu\ \alpha\ possède\ la\ caractéristique\ C \\ \\ 0 :\ Sinon \end{cases}$$
- La proportion (P) d'individus dans la population ayant la caractéristique C est égale à :
- $P = \frac{1}{N}\sum_{i=1}^{N} Y_i$
- Par ailleurs, les valeurs y_i observées sur l'échantillon ne peuvent être que 0 ou 1. Par conséquent, la proportion d'individus ayant la caractéristique C sur l'échantillon est égal à : $p = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Donc, l'estimateur de la proportion d'individus dans la population ayant la caractéristique C est:
- $\hat{P} = p = \frac{1}{n}\sum_{i=1}^{n} y_i$

a. Variance de l'estimateur de la proportion

- $V(\hat{P}) = V(p) = V(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(y_i)$

$= \frac{P(1-P)}{n}$ pour un PEAR

$= (\frac{N-n}{N})\frac{P(1-P)}{n}$ pour un PESR

b. Estimation de la variance de $V(\hat{P})$

- Nous avons : $\hat{P} = p = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$
- Par ailleurs, la variance corrigée sur l'échantillon se calcule comme suit :
- $\widehat{S_y^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)$

$= \frac{1}{n-1}(\frac{n}{n}\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)$ ..... (1)

- Comme les valeurs que peut prendre y_i sont 0 ou 1, alors :
- $p = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ et $p = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \bar{y}$
- D'où l'équation (1) peut être récrite comme suit :
- $\frac{1}{n-1}\left(\frac{n}{n}\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{n-1}(np - np^2) = \left(\frac{n}{n-1}\right)p(1-p)$
- donc :
- $\widehat{S_y^2} = \left(\frac{n}{n-1}\right)p(1-p)$
- $\hat{V}(\hat{P}) = \hat{V}(p) = \frac{1}{n^2}\sum_{i=1}^{n} V(y_i) = \frac{1}{n^2}nV(y_i) = \frac{1}{n^2}n\widehat{S_y^2} = \frac{1}{n^2}n\left(\frac{n}{n-1}\right)p(1-p)$

$= \frac{1}{n-1}p(1-p)$

6. Intervalle de confiance de la proportion d'une population finie

- $IC_P = [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{P})}\ ;\quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{P})}\ ]$

$$= [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{P(1-P)}{n}}\ ;\quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{P(1-P)}{n}}]$$

$$= [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n-1}}\ ;\quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n-1}}]$$ Pour un PEAR

- En outre,
- $IC_P = [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{P(1-P)}{n}}\ ;\quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{P(1-P)}{n}}]$

$$= [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{p(1-p)}{n-1}}\ ;\quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{p(1-p)}{n-1}}]$$

Pour un PESR

Exemple

Afin de contrôler la conformité de la production d'une pièce détachée à une norme établie, on décide de faire un sondage aléatoire systématique. On tire un échantillon au taux de sondage de 1/50 dans un lot de 500 pièces.

1. Quelle est la taille de l'échantillon ?

2. Dans l'échantillon, on constate que 10 pièce ne répondent à la norme. Estimer ponctuellement et par intervalle de confiance la proportion des pièces défectueuses dans le lot (on admet un risque d'erreur de 5%) (pour un tirage sans remise)

✓ **Réponse**

1. $t_x = \frac{n}{N} \Rightarrow n = t_x N = \frac{1}{50}(500) = 50$

2. **Estimation de (P)**

$\hat{P} = p = \frac{10}{50} = \frac{1}{5} = 0{,}2$ est un estimateur de la vraie proportion des pièces défectueuses (P)

Intervalle de confiance de (P)

$IC_P = [\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{p(1-p)}{n-1}} \;;\; \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{(1-t_x)\frac{p(1-p)}{n-1}}]$ Pour un PESR

$IC_P = [0{,}2 - 1{,}96\sqrt{\left(1-\frac{1}{50}\right)\frac{0{,}2(1-0{,}2)}{50-1}} \;; 0{,}2 + 1{,}96\sqrt{(1-1/50)\frac{0{,}2(1-0{,}2)}{50-1}}]$

Exercice 1

Une compagnie aérienne a demandé des statistiques afin d'améliorer la sûreté au décollage et définir un poids limite de bagages. Pour l'estimation du poids des voyageurs, un échantillon tiré en utilisant un SAS est constitué de 300 passagers qui ont accepté d'être pesés : on a obtenu une moyenne de 68 Kg, avec un écart-type de 7 Kg.

Définir un intervalle de confiance pour la moyenne des passagers. (On admet que le poids des passagers suit une loi normale de moyenne μ, d'écart-type σ.)

Solution

$\bar{Y}$: la moyenne de tous les passagers (la vraie moyenne de la population)

Nous avons : $\bar{y}$: la moyenne des 300 passagers de l'échantillon.

$\bar{y} = 68$ Kg

$\hat{\bar{Y}} = \bar{y} = 68$ Kg: estimateur sans biais de $\bar{Y}$

Intervalle de confiance de $\bar{Y}$

$$IC_{\bar{Y}} = [\bar{y} - Z_{1-\frac{\alpha}{2}} \frac{\hat{S}_y}{\sqrt{n}} \quad ; \quad \bar{y} + Z_{1-\frac{\alpha}{2}} \frac{\hat{S}_y}{\sqrt{n}}]$$ pour un PESR

$$IC_{\bar{Y}} = [68 - 1{,}96 \frac{7}{\sqrt{300}} \quad ; \quad 68 + 1{,}96 \frac{7}{\sqrt{300}}]$$

Exercice 2. (extrait du A-M. Dussaix et J-M. Grosbras, Exercices de sondage, Economica, 1992)

145 ménages de touristes séjournant en France dans une région donnée ont dépensé 830 € en moyenne par jour. L'écart type estimé de leurs dépenses s'élève à 210 €. Sachant que 50000 ménages de touristes ont visité la région où a été effectuée l'enquête, que peut-on dire de la dépense totale journalière de l'ensemble de ces ménages ? On supposera pour cela que l'échantillon est issu d'un plan aléatoire simple à probabilités égales.

Solution

La dépense moyenne/jour pour 145 ménages : $\bar{y}$=830 €/jour

$\hat{T} = N\bar{y}$ est un estimateur sans biais de la vraie totale de la population pour le cas d'un SAS

$\hat{T} = N\bar{y}$: Estimation de la dépense totale journalière des 50000 ménages

Donc ; $\hat{T} = 50000(830)$ =41500000 €/jour

Exercice 3 (Extrait de la série de TD de Pr. Myriam Maumy-Bertrand- https://irma.math.unistra.fr/~mmaumy/enseignement/M2StatsM2Actu/td2.pdf)

- Une entreprise possède cinq succursales. Un inspecteur ne peut en examiner que deux par tournées. Dans chaque succursale, nous mesurons une variable d'intérêt Y (Nombre de nouveaux clients dans l'année). La situation réelle des cinq succursales est la suivante :

$Y_1 = 100; Y_2 = 80; Y_3 = 100; Y_4 = 120; Y_5 = 90.$

1. Calculer la moyenne $\bar{Y}$
2. Enumérer tous les échantillons possibles (e_i) correspondant à une tournée et pour chaque échantillon calculer $\bar{y}_{e_i}$
3. Calculer l'estimateur de la moyenne $\bar{Y}$,
4. Vérifier que $\hat{\bar{Y}}$ est un estimateur sans biais de la vraie moyenne de la variable d'intérêt.
3. Calculer la variance de l'estimateur de la moyenne $\bar{Y}$ (i.e. $V(\hat{\bar{Y}})$)
4. Pour chaque échantillon possible, calculer sa variance corrigée $s^2_{e_i,c}$
5. Soit s_c^2 la variance corrigée-échantillon, Vérifier que celle-ci estime sans biais la vraie valeur de la variance corrigée de la population

III.2. Cas d'un plan d'échantillonnage stratifié :

- **1. Notation**
- **1.1. Paramètres de la population**

	Strates						
	1	2	.....	h	.....	k	
Effectifs	N_1	N_2	.....	N_h	.....	N_k	N
Moyenne	$\bar{Y}_1$	$\bar{Y}_2$	.....	$\bar{Y}_h$	.....	$\bar{Y}_k$	$\bar{Y}$
Variance	σ_1^2	σ_2^2	.....	σ_h^2	.....	σ_k^2	σ^2
Variance corrigée	S_1^2	S_2^2		S_h^2		S_k^2	S^2

- **Dans chaque strate (h)**:
- $\bar{Y}_h = \frac{1}{N_h}\sum_{i,h}^{N_h} Y_{i,h}$, $\sigma_h^2 = \frac{1}{N_h}\sum_{i,h}^{N_h}(Y_{i,h} - \bar{Y}_h)^2$

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i,h}^{N_h}(Y_{i,h} - \bar{Y}_h)^2$$

- **Variance de la population totale (Variance de la variable d'intérêt Y)**:
- $\sigma_Y^2 = \frac{1}{N}\sum_{h=1}^{k}\sigma_h^2 + \frac{1}{N}\sum_{h=1}^{K} N_h(\hat{\bar{Y}}_{st} - \bar{Y}_h)^2$

  $\sigma_Y^2 = \sigma_{intra}^2 + \sigma_{inter}^2$
- Où
- σ_{intra}^2 est la variance intra-strates ($\sigma_{intra}^2 = \frac{1}{N}\sum_{h=1}^{k}\sigma_h^2$)
- σ_{inter}^2 est la variance inter-strates ($\frac{1}{N}\sum_{h=1}^{K} N_h(\hat{\bar{Y}}_{st} - \bar{Y}_h)^2$)

- **1.2. Échantillon**

	Echantillon tiré de la Strates						Échantillon
	1	2	.....	h	.....	k	
Effectifs	n_1	n_2	.....	n_h	.....	n_k	n
Moyenne	$\bar{y}_1$	$\bar{y}_2$	.....	$\bar{y}_h$	.....	$\bar{y}_k$	$\bar{y}$
Variance Corrigée	$\hat{s}_1^2$	$\hat{s}_2^2$	.....	$\hat{s}_h^2$	.....	$\hat{s}_k^2$	$\hat{s}^2$

- Tels que : $\bar{y}_h = \frac{1}{n_h}\sum_{i,h}^{n_h} y_{i,h}$ la moyenne calculée à partir de l'échantillon tiré de la strate h, et $\hat{s}_h^2 = \frac{1}{n_h - 1}\sum_{i,h}^{n_h}(y_{i,h} - \bar{y}_h)^2$ la variance modifiée de l'échantillon tiré de la strate h

Choix de la taille de l'échantillon tiré de chaque strate

- **Plan stratifié avec allocation proportionnelle:** dans ce cas, le taux de sondage ($t_{xh} = \frac{n_h}{N_h}$) est le même pour toutes les strate donc nous avons: $\frac{n}{N} = \frac{n_h}{N_h} \Longrightarrow n_h = \frac{N_h}{N} n$
- **Plan stratifié avec allocation optimale**

- Dans ce cas, nous utilisons la répartition de Neyman telle que
- $\frac{n_h}{n} = \frac{N_h \sigma_h^2}{\sum_{h=1}^{k} N_h \sigma_h^2} \Longrightarrow n_h = \frac{N_h \sigma_h^2}{\sum_{h=1}^{k} N_h \sigma_h^2} n$

2. Estimation de la moyenne simple de la population ($\bar{Y}_{st}$)

- Soit $\bar{Y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{Y}_h$ la vraie moyenne de la population, alors;

  $\hat{\bar{Y}}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$ est l'estimateur de $\bar{Y}_{st}$

a. Variance de $\hat{\bar{Y}}_{st}$

- $V\left(\hat{\bar{Y}}_{st}\right) = V\left(\sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h\right)$
- Les tirages sont indépendants d'une strate à l'autre $\Longrightarrow Cov(\bar{y}_h, \bar{y}_k) = 0$ donc :
- $V\left(\hat{\bar{Y}}_{st}\right) = V\left(\sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} V(\bar{y}_h)$
- d'où $V\left(\hat{\bar{Y}}_{st}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{\sigma_h^2}{n_h}$ Tirage avec remise dans chaque strate
- $V\left(\hat{\bar{Y}}_{st}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h} = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$ Tirage sans remise dans chaque strate

b. Estimation de $V\left(\hat{\bar{Y}}_{st}\right)$

- Si σ_h^2 et S_h^2 sont inconnues alors :
- $\hat{V}\left(\hat{\bar{Y}}_{st}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{\hat{s}_h^2}{n_h}$ Tirage avec remise dans chaque strate
- $\hat{V}\left(\hat{\bar{Y}}_{st}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{\hat{s}_h^2}{n_h}$ Tirage sans remise dans chaque strate

c. Estimation par Intervalle de confiance de $\bar{Y}_{st}$

- $IC_{\bar{Y}_{st}} = [\hat{\bar{Y}}_{st} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}\left(\hat{\bar{Y}}_{st}\right)}]$

$= [\hat{\bar{Y}}_{st} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{\hat{s}_h^2}{n_h}}]$ Tirage avec remise dans chaque strate

$= [\hat{\bar{Y}}_{st} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{\hat{s}_h^2}{n_h}}]$ Tirage sans remise dans chaque strate

- **Exemple (Extrait du livre « Les techniques de sondage » publié par Pascal Ardilly, 2006, ISBN :9782710808473, 2710808471 ; Éditeur:Technip)**
- On dispose une population de 1060entreprises, et on s'intéresse au nombre moyen ($\bar{Y}_{st}$) d'employés par entreprise. La population est constituée de 5 strates définies par des tranches de tailles en nombre d'employés. Réalisant un S.A.S. dans chaque strate pour sélectionner un échantillon de 300 entreprises. On mesure la moyenne ($\bar{Y}_h$) et la dispersion de la variable « nombre d'employés » dans l'échantillon des entreprises tirées. Les données relatives à cette répartition sont résumées dans le tableau suivant :

Tranche de taille	N_h	$\bar{y}_h$	s_h^2	n_h
0-9	500	5	1,5	130
10-19	300	12	4	80
20- 49	150	30	8	60
50-499	100	150	100	25
500 et plus	10	600	2500	5
TOTAL	1060			300

- 1. Quel est l'estimateur de $\bar{Y}_{st}$ et sa précision
- Donner l'intervalle de confiance de $\bar{Y}_{st}$ au seuil de confiance de 95%

- **Réponse:**

1. l'estimateur de $\bar{Y}_{st}$

- **Nous avons:** $\bar{Y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{Y}_h$, la vraie moyenne de la population
- $\hat{\bar{Y}}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$ est l'estimateur de $\bar{Y}_{st}$
- $\hat{\bar{Y}}_{st} = \sum_{h=1}^{5} \frac{N_h}{N} \bar{y}_h = \frac{1}{1060}[500(5) + \cdots + 10(600)] = 29{,}51$
- **Précision de $\hat{\bar{Y}}_{st}$**
- $V(\hat{\bar{Y}}_{st}) = \hat{V}\left(\hat{\bar{Y}}_{st}\right) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{\hat{s}_h^2}{n_h}$ (lorsque n$> 30, PESR \approx PEAR$)
- $\hat{V}\left(\hat{\bar{Y}}_{st}\right) = \left(\frac{1}{1060^2}\right)\left[500^2 \frac{1{,}5}{130} + \cdots + 10^2 \frac{2500}{5}\right] + 0{,}089$

2. Intervalle de confiance

$IC_{\bar{Y}_{st}} = [29{,}51 \pm 1{,}96\sqrt{0{,}089}]$

3. Estimation du total (T) de la variable d'intérêt (Y) d'une population finie

- Soit $T_h = N_h \bar{Y}_h$ le total dans la strate h , alors : $\hat{T}_h = N_h \bar{y}_h$ est l'estimateur du T_h
- Par ailleurs, le total T pour les k strates est égale à : $T = \sum_{h=1}^{k} T_h = \sum_{h=1}^{k} N_h \bar{Y}_h$
- D'où $\hat{T} = \sum_{h=1}^{k} N_h \bar{y}_h$ est l'estimateur de T

a. Variance de l'estimateur du total

- $\boldsymbol{V(\hat{T})} = V\left(\sum_{h=1}^{k} N_h \bar{y}_h\right) = \sum_{h=1}^{k} N_h^2 V(\bar{y}_h)$ Car les tirages sont indépendants d'une strate à l'autre $\Longrightarrow Cov(\bar{y}_h, \bar{y}_k) = 0$
- $\boldsymbol{V(\hat{T})} = \sum_{h=1}^{k} N_h^2 \frac{\sigma_h^2}{n_h}$ Tirage avec remise dans chaque strate
- $\boldsymbol{V(\hat{T})} = \sum_{h=1}^{k} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$ Tirage sans remise dans chaque strate

b. Estimation de la variance de l'estimateur du total

- $\widehat{\boldsymbol{V}}(\hat{T}) = \sum_{h=1}^{k} N_h^2 \frac{\hat{s}_h^2}{n_h}$ Tirage avec remise dans chaque strate
- $\widehat{\boldsymbol{V}}(\hat{T}) = \sum_{h=1}^{k} N_h^2 \frac{N_h - n_h}{N_h} \frac{\hat{s}_h^2}{n_h}$ Tirage sans remise dans chaque strate

5. Estimation par intervalle de confiance du total

- $\boldsymbol{IC_T} = [\hat{T} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\boldsymbol{V}}(\hat{T})}]$

$= [\hat{T} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{k} N_h^2 \frac{\hat{s}_h^2}{n_h}}]$ Tirage avec remise dans chaque strate

$= [\hat{T} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{k} N_h^2 \frac{N_h - n_h}{N_h} \frac{\hat{s}_h^2}{n_h}}]$ Tirage sans remise dans chaque strate

6. Estimation de la proportion

- Soit P_{st} : la vraie proportion de la population (inconnue)
- $P_{st} = \sum_{h=1}^{k} \frac{N_h}{N} P_h$

Si, P_h la proportion de la strate h est inconnue, donc on va l'estimer par la proportion calculée à partir de l'échantillon tiré de la strate h (notée p_h) , alors: $\hat{P}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} p_h$ est un estimateur sans biais de P_{st}

a. Variance de $\widehat{P}_{st}$

$V(\hat{P}_{st}) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} (1 - t_{x,h}) (\frac{p_h(1-p_h)}{n_h - 1})$ pour un tirage sans remise

b. Intervalle de confiance de P_{st}

$$IC_{P_{st}} = [\hat{P}_{st} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{k} \frac{N_h^2}{N^2} (1 - t_{x,h}) (\frac{p_h(1-p_h)}{n_h - 1})}$$ **(tirage sans remise)**

Exercice 1 (extrait du polycopié du Christophe Chesneau (2022). Eléments de théorie des sondages. Université de Caen-Normandie. https://chesneau.users.lmno.cnrs.fr/sondage-cours.pdf)

Une population U est partagée en 3 strates U_1, U_2 et U_3 de tailles respectives : N1 = 12, N2 = 28 et N3 = 50. On prélève un échantillon de n = 20 individus suivant un plan de sondage aléatoire de type Sondage stratifié (ST) avec : n1 = 2 individus pour U_1, n2 = 6 individus pour U_2, n3 = 12 individus pour U_3.
On mesure un caractère quantitatif Y sur chacun d'entre eux. Les résultats obtenus sont :

Pour U_1	1450	1598				
Pour U_2	718	626	922	823	901	823
Pour U_3	201	268	225	231	453	387
	401	368	325	331	253	197

1. Donner une estimation ponctuelle de la moyenne-population Y_U .
2. Donner une estimation ponctuelle de l'écart-type de l'estimateur de Y_U .
3. Déterminer un intervalle de confiance pour Y_U au niveau 95%.

Exercice 2

- Sur les 7500 employés de l'entreprise A, on souhaite connaitre 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 des strates dans la population : individus de revenu faible (strate 1), de revenu moyen (strate 2), et de revenu élevé (strate 3). On note :N_h : La taille de la strate h ; n_h : La taille de l'échantillon dans la strate h ; $\hat{P}_h$: L'estimateur de la proportion d'individus possédant au moins un véhicule dans la strate h. Les résultats de l'enquête sont consignés dans le tableau ci-dessous :

	Strate1	Strate 2	Strate 3
N_h	3500	2000	2000
n_h	500	300	300
$\hat{P}_h$	0,13	0,45	0,45

- Quel estimateur $\hat{P}$ de P proposez-vous ? que peut-on dire de son biais ?
- Calculer la précision de $\hat{P}$ et donner un intervalle de confiance à 95% pour P ?
- Estimez-vous que le caractère de stratification est adéquat ?, justifiez votre réponse ?