MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE.

UNIVERSITE ABDERRAHMANE MIRA BEJAIA

FACULTE DES SCIENCES EXACTES

DEPARTEMENT DE MATHEMATIQUES

THEME

Initiation aux méthodes de Sondage

POLYCOPIE présenté

Par

GUEBLI Sofia épouse GRITLI

Année 2017/2018

Chapitre 01 : Echantillonnage

Chapitre 02: Sondage

Chapitre 03: Sondage Stratifie

Chapitre 04: Sondage par grappes

Chapitre 5 : Sondage à Plusieurs Degrés

Chapitre 6 : Méthodes à Choix raisonnées

Chapitre 07: Jackknife et bootstrap appliqués aux sondages

Chapitre 01 : Echantillonage

1.1. Notion d'échantillonnage

➢ **Définition :** On considère une population Ω de taille N. On appelle échantillon un sous-ensemble de cette population. Un échantillon de taille n n'est donc qu'une liste de n individus ($\omega_1, \ldots, \omega_n$), extraits de la population mère.

➢ **Définition :** On appelle échantillonnage le prélèvement d'échantillons. Le rapport t de L'effectif n de l'échantillon sur l'effectif N de la population dans laquelle il a été prélevé, est appelé taux d'échantillonnage ou fraction de sondage : $t = \frac{n}{N}$

➢ **Définition :** On appelle échantillonnage aléatoire un prélèvement de n individus dans une population mère tel que toutes les combinaisons possibles de n individus aient la même probabilité d'être prélevés.

On appelle n-échantillon de valeurs de X la liste des valeurs ($x_1, x_2, \ldots x_n$) observées prises par X sur un échantillon ($\omega_1 \ldots \omega_n$) de la population Ω. Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur de variables aléatoires ($X_1, X_2, \ldots, X_n$) appelé n-échantillon de X où les X_i sont de même loi, indépendantes

➢ **Définition :** X_1 est alors la variable aléatoire « valeur du premier élément de l'échantillon », X_2 la variable aléatoire « valeur du second élément de l'échantillon »…etc.

➢ **Définition :** On appelle statistique toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, \ldots, X_n$.

1.2. Distributions d'échantillonnage

1.2.1. Moyenne d'échantillon - variance d'échantillon

➢ **Définitions :**

On considère une population Ω dont les éléments possèdent un caractère quantitatif C qui est la réalisation d'une variable aléatoire X qui suit une loi de probabilité d'espérance μ et d'écart-type σ.

On suppose que la famille est de taille infinie ou que l'échantillonnage se fait avec remise.

On prélève un n-échantillon ($X_1, \ldots, X_n$) de X de valeurs ($x_1,\ldots, x_n$). La moyenne $\overline{X}$ de l'échantillon est donnée par

$$\overline{X} = (x_1, x_2,\ldots, x_n)/n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Il s'agit de la valeur prise par la variable aléatoire.

$$(x_1, x_2,\ldots, x_n)/n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Définition : On définit la variable aléatoire $\overline{X}$, appelée moyenne d'échantillon, par

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

De la même manière, la variance v de l'échantillon ($x_1, x_2, \ldots, x_n$) est donnée par

$$v = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 .$$

Il s'agit de la valeur prise par la variable aléatoire

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Définition : On définit la variable aléatoire S^2, appelée variance d'échantillon, par

$$S^2 = \frac{1}{n-1}\sum_{i=n}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}V$$

1.2.2. Paramètres descriptifs de la distribution

➢ **Proposition**

✓ Quelle que soit la loi de X, on a

$E(\bar{X}) = \mu$ Var $(\bar{X}) = \frac{\sigma}{n}$,

$E(V) = \frac{n-1}{n}\sigma^2$ et $E(S^2) = \sigma^2$.

✓ Si $X \sim N(\mu, \sigma)$, on a :

(i) Si σ est connu : $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

(ii) Si σ n'est pas connu : $\frac{\bar{X}-\mu}{\sqrt{\frac{V}{n}}} \sim T_{n-1}$,

(iii) $\frac{nV}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim X^2_{N-1}$

1.2.3. Proportion d'échantillon

Il arrive que le caractère à estimer ne soit pas quantitatif mais qualitatif. Dans ce cas, on recherche la proportion p des individus présentant ce caractère.
La proportion p sera estimée à l'aide des résultats obtenus sur un n-échantillons.

➢ **Définition**

La proportion f obtenue dans un n-échantillons est la valeur observée d'une variable aléatoire F, fréquence d'apparition de ce caractère dans un échantillon de taille n, appelée proportion d'échantillon ou fréquence statistique, on peut écrire

$$F = \frac{K}{n}$$

Où' K est la variable aléatoire qui compte le nombre d'apparition du caractère dans un échantillon de taille n.

Par définition K sui la loi binomial de paramètre n et p, $k \sim B(n, p)$, soit :

$$E(K) = np \text{ et } Var(k) = npq \text{ avec } q=1-p$$

D'où la proposition suivante :

- **Proposition :**

$$E(\bar{F}) = p \text{ et } Var(F) = \frac{pq}{n}$$

Remarque

Pour $n \geq 30$, $np \geq 15$ on peut approcher F par une loi normale de paramètre $p, \sqrt{\frac{pq}{n}}$

$$N = \left(p, \sqrt{\frac{pq}{n}}\right).$$

1.3. Estimation - Intervalles de confiance

1.3.1. Les estimateurs

Estimer un paramètre c'est de chercher une valeur approchée à partir des résultats obtenus sur un échantillon.

Exemple :

Estimer la taille moyenne d'une population à partir de la moyenne empirique obtenue sur un échantillon de cette population.

1.3.2. Définitions

- **Définition :** Un estimateur $\hat{\theta}$ du paramètre inconnu θ est une fonction qui fait correspondre à une suite d'observations une valeur approchée $\widehat{\theta n}$ de θ, appelée estimation :

$\hat{\theta} : (x_1, x_2, \ldots, x_n) \rightarrow \widehat{\theta n} = f(x_1, x_2, \ldots, x_n)$.

Un estimateur $\hat{\theta}$ est donc une variable aléatoire, on peut en calculer son espérance $E(\hat{\theta})$ et sa

variance $Var(\hat{\theta})$. Ces quantités vont permettre de déterminer la qualité d'un estimateur du paramètre θ à estimer.

Un paramètre peut en effet avoir plusieurs estimateurs. Dans le cas de la taille moyenne d'une population, on peut choisir la moyenne arithmétique, la médiane… etc.

➢ **Définition :** On dit que $\widehat{\theta}$ est un estimateur sans biais si la moyenne de sa distribution d'échantillonnage est égale à la valeur θ du paramètre à estimer :

$$\mathrm{E}(\hat{\theta}) = \theta.$$

Sinon, on parle d'estimateur biaisé. Pour comparer les estimateurs biaisés, on introduit la quantité suivante :

➢ **Définition :** On appelle biais d'un estimateur $\hat{\theta}$ la quantité

$$\mathrm{Biais}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

➢ **Définition :** On dit qu'un estimateur sans biais est efficace si sa variance est la plus petite parmi les variances des estimateurs sans biais. Si $\hat{\theta}_1$ est un estimateur de θ, on dit que $\hat{\theta}_1$ est efficace si pour tout estimateur sans biais $\hat{\theta}_2$:

$$\mathrm{E}(\hat{\theta}_1) = \mathrm{E}(\hat{\theta}_2) = \theta \quad \text{et} \quad \mathrm{Var}(\hat{\theta}_1) < \boldsymbol{Var}(\hat{\theta}_2).$$

➢ **Définition**: Un estimateur $\hat{\theta}$ est convergent si sa distribution tend à se concentrer autour de la valeur θ à estimer, en d'autres termes si sa variance tend vers zéro lorsque la taille de l'échantillon augmente :

$$\lim_{n \to +\infty} Var(\hat{\theta}) = 0.$$

1.3.3. Estimateurs usuels

1.3.3.1.Cas d'un caractère quantitatif

Soit X une variable aléatoire de moyenne μ et d'écart-type σ définie sur une population mère Ω. Soit(X_1, X_2, . . ., X_n) un n-échantillon de X.

➢ **Propriétés :** On a les résultats suivants **:**

$\overline{X}= \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent de μ

$V=\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$est un estimateur biaisé de la variance σ^2

$S^2= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2=\frac{n}{n-1}V$ est un estimateur sans biais et convergent de la Variance σ^2

1.3.3.2.Cas d'un caractère qualitatif

On considère un caractère qualitatif d'une population dont on cherche à estimer la proportion p.

Propriété : La proportion d'échantillon F est un estimateur sans biais et convergent de la proportion p.

1.4. Intervalles de confiance

Plutôt que de déterminer une valeur approchée d'un paramètre θ obtenue à l'aide d'un estimateur $\widehat{\theta}$, on va rechercher un intervalle dans lequel on sait avec une probabilité satisfaisante que la valeur de θ s'y trouve.

1.4.1. Définitions

➢ **Définition :** Soit X une variable aléatoire dont la loi dépend d'un paramètre θ. Les intervalles de confiance de risque α pour le paramètre θ, issus des différents n-échantillons (x_1, . . ., x_n), sont les intervalles [a(x_1,x_2,. . ., x_n) ; b(x_1,x_2,. . ., x_n)] tels qu'une proportion α de ces intervalles contiennent θ.

Remarque

La quantité $1 - \alpha$ est appelée niveau de confiance de l'intervalle [a, b] :

$$P(a \leq \hat{\theta} \leq b) = 1 - \alpha.$$

Dans la pratique, on ne dispose bien souvent que d'un seul échantillon qui fournit un intervalle de confiance [a, b].

Le paramètre à estimer est souvent l'espérance ou la variance dans le cas d'un caractère quantitatif, la proportion dans le cas d'un caractère qualitatif.

Dans la suite on s'attachera à rechercher des intervalles de confiance [a; b] symétriques, c'est-à-dire :

$$p(\hat{\theta}<a) = \frac{\alpha}{2} \quad \text{et} \quad p(\hat{\theta} > a) = \frac{\alpha}{2}$$

On détermine ensuite les variables aléatoires A_n et B_n en fonction de $\widehat{\theta}$ telles que :

$$P(A_n \leq \theta \leq B_n) = 1 - \alpha.$$

Un intervalle de confiance [a, b] de risque α pour θ, issu d'un n-échantillon $(x_1, \ldots, x_n)$ de valeurs de X, s'obtient alors en calculant :

$$a = A_n(x_1, x_2, \ldots, x_n)$$

$$b = B_n(x_1, x_2, \ldots, x_n)$$

1.4.2. Intervalle de confiance pour une moyenne

On se place dans le cas où X suit une loi normale de paramètres μ et σ ou bien dans le cas où l'on ne connait pas forcément la loi de X mais pour laquelle on dispose d'un échantillon de taille $n>30$.

Dans le premier cas $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$, dans le second cas $\overline{X}$ suit approximativement cette même loi.

On considère un n-échantillon $(x_1, x_2, \ldots, x_n)$ de valeurs de X. on note :

$$m = \frac{x_1 + \cdots + x_n}{n} \text{ et } s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

Cas σ connu

$$I = [\, m - t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \,;\, m + t_{1+\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}] \,,$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite

Dessin

- Démonstration: On sait que $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$, soit encore

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Donc

$$P(-t_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Soit

$$p(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

1.4.2.1. Cas σ inconnu

$$I = [\, m - t_{1-\frac{\alpha}{2}n-1} \, , \quad \frac{s}{\sqrt{n}} \; ; m + t_{1-\frac{\alpha}{2}n-1} \frac{s}{\sqrt{n}} \,]$$

Où $t_{1-\frac{\alpha}{2}n-1}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté

- **Remarque**
- si $n > 30$, $t_{1-\frac{\alpha}{2}\,n-1} = t_{1-\frac{\alpha}{2}}$

1.4.3. Intervalle de confiance pour une variance

On se place dans le cas où X suit une loi normale de paramètres μ et σ.

1.4.3.1. Cas μ connu

$$I = [\frac{nv}{x_1^2\,(n)} \; ; \frac{nv}{x^2_{\frac{\alpha}{2}}\,(n)} \,]$$

Où $x^2_{\frac{\alpha}{2}}(n)$ et $x_1^2(n)$ sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à n degrés de liberté et

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

1.4.3.2. cas μ inconnu

$$I = [\,\frac{(n-1)s^2}{x^2_{1-\frac{\alpha}{2}}(n-1)} \; ; \frac{(n-1)s^2}{x^2_{\frac{\alpha}{2}}(n-1)}]$$

Où $x^2_{1-\frac{\alpha}{2}}(n-1)$ et $x^2_{\frac{\alpha}{2}}(n-1)$ sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à n-1 degrés de liberté.

➢ **Remarque**

➢ si n>30,x_{α}^{2} (n-1) ≈ $\frac{1}{2}$ ($t_{\alpha} + \sqrt{2n-3}$)2 , Si bien que l'on choisit :

$$I = [\frac{(n-1)s^2}{(t_{1-\frac{\alpha}{2}} + \sqrt{2n-3})^2} ; \frac{(n-1)s^2}{(t_{1+\frac{\alpha}{2}} + \sqrt{2n-3})^2}]$$

D'autre part, la symétrie de la loi normale centrée réduite assure que $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$

1.4.4. intervalle de confiance pour une proportion

On a vu dans le chapitre précédent que la proportion d'échantillon F peut être approchée par une loi normale N (p, $\sqrt{\frac{pq}{n}}$)

$$I = [f - t_{1-\frac{\alpha}{2}}\sqrt{\frac{f(f-1)}{n}} ; f + t_{1-\frac{\alpha}{2}}\sqrt{\frac{f(f-1)}{n}}]$$

Ou' f est la proportion de l'échantillon analysé

1.5. Tests d'hypothèse

1.5.1. Introduction aux tests d'hypothèse

On étudie un caractère quantitatif ou qualitatif X d'une certaine population Ω dont une au moins des valeurs des paramètres décrivant X est inconnue.

On formule une hypothèse sur la valeur de ce paramètre.
On s'interroge sur la pertinence de cette hypothèse en la confrontant aux résultats obtenus sur un échantillon.

Les distributions d'échantillonnage d'une moyenne, d'une variance et d'une proportion vues dans le chapitre précédent nous permettrons d'élaborer des tests d'hypothèse**.**

1.5.2. Définitions

➢ **Définition:** Un test d'hypothèse est une procédure basée sur l'observation d'un ou plu- sieurs échantillons permettant de faire un choix entre deux hypothèses formulées.

➢ **Définition:** L'hypothèse mise en avant dans le cadre d'un test d'hypothèse est notée (H_0), appelée hypothèse nulle. Toute autre hypothèse à laquelle on peut la confronter s'appelle hypothèse alternative, notée (H_1).

Remarque :

C'est l'hypothèse (H_0) qui est soumise au test et que l'on suppose comme vraie.

La décision d'accepter ou rejeter le test repose sur la confrontation aux valeurs observées sur un échantillon. L'information contenue dans cet échantillon étant incomplète, toute décision est associée à prise de risque.

➢ **Définition :** On appelle erreur de première espèce l'erreur commise lorsqu'on rejette

L'hypothèse nulle (H_0) alors que cette dernière est vraie. La probabilité d'une telle erreur s'appelle risque de première espèce et se note α :

$$\alpha = \mathrm{P}\ (\text{rejeter}(H_0)/(\ H_0\ )\ \text{vraie}).$$

Remarque :

On choisit souvent en pratique $\alpha = 0.05$ ou $\alpha = 0.01$.

➢ **Définition**: On appelle erreur de seconde espèce l'erreur commise lorsqu'on accepte l'hypothèse (H_0) alors que cette dernière est fausse. La probabilité d'une telle erreur s'appelle risque de seconde espèce et se note β :

$$\beta = \mathrm{P}\ (\text{accepter}(H_0)/(\ H_0\ )\ \text{fausse}).$$

La valeur $1 - \beta$ est appelée puissance du test.

La puissance du test correspond à la probabilité de rejeter l'hypothèse (H_0) sachant que cette dernière

est fausse. Plus β est petit et plus le test sera puissant.

Le test est concluant lorsqu'on qu'on ne commet aucune erreur, ni de première ni de seconde espèce**.**

P	(H_0) acceptée	(H_0) refusée
(H_0) vraie	A	1 – α
(H_0) fausse	– β	B

1.5.3. Fonctionnement d'un test

Pour chaque test, on fera appel à une variable aléatoire de décision T qui suit une certaine loi théorique. En supposant que l'hypothèse (H_0) est vraie, on cherche la valeur idéale pour T. Le risque de première espèce α étant choisi, on détermine une zone de probabilité $1-\alpha$, généralement un intervalle, contenant cette valeur idéale.

Si la valeur idéale constitue l'une des bornes de cette zone, on parle de test unilatéral.

Sinon on parle de test bilatéral. Dans la suite, on s'intéresse surtout aux tests bilatéraux**.**

Si la valeur de T obtenue en se basant sur les résultats d'un échantillon appartient à la zone critique de probabilité α, alors l'hypothèse (H_0) est rejetée, sinon elle est acceptée.

1.5.4. Tests de conformité

Soit X un caractère dont la loi dépend d'un paramètre θ inconnu. Soit θ_0 une valeur donnée.

On se donne un niveau de risque **α**

On va alors tester (H_0) contre (H_1) ou $\begin{cases}(H_1): \theta \neq \theta_0 \\ (H_2): \theta = \theta_0\end{cases}$

Au risque de première espèce α

On détermine la variable aléatoire de décision : dans le cas de la moyenne $\bar{X}$ Semble appropriée.

1.5.4.1. σ est connu

Dans ce cas,

$$U = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

D'où la zone d'acceptation avec un risque de niveau α pour $\overline{X}$:

$$I = [\, \mu_0 - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \, , \mu_0 + t_{1+\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \,]$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On sait en effet que :

$$P(\overline{X} \in I) = 1 - \alpha$$

Si la valeur $\bar{x}$ obtenue sur le n-échantillon considéré appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

1.5.4.2. σ est inconnu

Dans ce cas,

$$U = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim T_{n-1},$$

D'où la zone d'acceptation avec un risque de niveau α pour $\overline{X}$:

$$I = [\ \mu_0 - \ t_{1-\frac{\sigma}{2,n-1}} \frac{s}{\sqrt{n}}\ ,\ \mu_0 + \ t_{1+\frac{\sigma}{2,n-1}} \frac{s}{\sqrt{n}}\ \ ]$$

Où $t_{1-\frac{\alpha}{2}}$, n-1 est le quantile d'ordre 1- $\frac{\alpha}{2}$ de la loi de student à n-1 degrés de liberté. On sait en effet que **:**

$$P(\overline{X}\ \in I) = 1 - \ \alpha$$

Si la valeur obtenue sur le n-échantillon considéré appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

Remarque

L'intervalle que l'on obtient est centrée sur la valeur supposée μ_0, alors que pour les intervalles de confiance il était centré sur la valeur de la moyenne de l'échantillon

$\bar{x}$. Dans le cas des tests, on conclut alors en regardant si cette valeur de x appartient ou non à notre intervalle centré sur la valeur supposée μ_0.

1.6. Test de conformité d'une variance :

On dispose d'un n-échantillon (x_1, x_2, . . ., x_n) de valeurs de X. On suppose que X suit une loi normale de paramètres μ et σ. Soit $\sigma 0$ une valeur plausible de la variance fixée a priori. On teste (H_0) contre (H_1) où

On va alors tester (H_0) contre (H_1) ou$\begin{cases}(H_0) : \ \sigma = \ \sigma_0 \\ (H_1): \ \ \sigma \neq \sigma_0\end{cases}$

Au risque de première espèce α

On détermine la variable aléatoire de décision : dans le cas de la variance S^2 semble appropriée. On ne traite ici que le cas où μ est inconnu, le cas où μ est connu étant relativement peu fréquent.

Dans ce cas

$$U = \frac{n-1}{\sigma_0^2} S^2 \sim X^2_{n-1}$$

D'où la zone d'acceptation avec un risque de niveau α pour S^2 :

$$I = [\frac{\sigma_0^2}{n-1} x^2_{1-\frac{\alpha}{2}}(n-1); \frac{\sigma_0^2}{n-1} x^2_{\frac{\alpha}{2}}(n-1)$$

Ou $x^2_{1-\frac{\alpha}{2}}(n-1)$ et $x^2_{\frac{\alpha}{2}}(n-1)$; sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à $n-1$ Degrés de liberté. On sait en effet que :

$$P(S^2) \in I) = 1 - \alpha.$$

Si la valeur S^2 obtenue sur le n-échantillon considéré appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

Remarque

Contrairement à l'intervalle de confiance sur la variance quand la moyenne est inconnue, pour l'intervalle d'acceptation, les quantiles sont aux numérateurs.

1.7. Test de conformité d'une proportion :

Pausible de p, fixée a priori. On teste (H_0) contre (H_1) où

$$\begin{cases} (H_0) :: p = p_0 \\ (H_1): \ p \neq p_0 \end{cases}$$

Au risque de première espèce α.

On détermine la variable aléatoire de décision : la fréquence F semble appropriée.

On suppose que F suit approximativement une loi $N\left(p, \sqrt{\frac{pq}{n}}\right)$

La zone d'acceptation avec un risque de niveau $\boldsymbol{\alpha}$ pour F est

$$I = \left[p_0 - t_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0}{n}} \; ; \; p_0 + t_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0}{n}} \right]$$

Ou' $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. On sait en effet que :

$$P(F \in I) = 1 - \alpha$$

Si la valeur de la proportion f obtenue sur le n-échantillon considéré appartient à I, alors on accepte (H_0), sinon on rejette (H_0),

1.8. Teste de comparaison

Il est fréquent de comparer des échantillons issus de deux groupes d'individus. On considère deux variables aléatoires X_1 et X_2 définies sur chacun de ces groupes et on souhaite tester si les caractères étudiés suivent la même loi. Les variables aléatoires que nous emploierons pour ces tests sont la différence des moyennes d'échantillon, le quotient des variances d'échantillon ou la différence des fréquences d'échantillon.

En général, on travaille dans le cas particulier où les tailles d'échantillons sont supérieures à 30.

1.9. Test de comparaison de deux moyennes

On considère que les variables aléatoires X_1 et X_2 suivent respectivement des lois normales $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$

On teste (H_0) contre (H_1) où

$$\begin{cases} (H_0) : \mu_1 = \mu_2 \\ (H_1) : \mu_1 \neq \mu_2 \end{cases}$$

Au risque de première espèce α

On dispose d'un échantillon de taille n_1 pour X_1 et de taille n_2 pour X_2.

1.9.1. $\sigma_1$et σ_2 connus

On pose

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

La variable aléatoire de décision du test. L'hypothèse (H_0) étant supposée vraie, T suit une loi normale centrée réduite. La zone d'acceptation avec un risque de niveau α pour T est :

$$I = [-t_{1-\frac{\alpha}{2}} ; t_{1-\frac{\alpha}{2}}]$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. On sait en effet que :

$$P(T \in I) = 1 - \alpha$$

Si la valeur

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Obtenue sur les échantillons considérés appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

1.9.2. σ_1 et σ_2 inconnus

On remplace alors les variances théoriques par les estimateurs sans biais S_1 et S_2 de ces variances. La variable de décision s'écrit :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Les échantillons étant de tailles supérieures à 30, on peut montrer que T peut être approché par une loi normale centrée réduite.

La zone d'acceptation avec un risque de niveau α pour T est à nouveau :

$$I = \left[-t_{1-\frac{\alpha}{2}} \; ; \; t_{1-\frac{\alpha}{2}}\right],$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On sait en effet que :

$$P(T \in I) = 1 - \alpha$$

Si la valeur

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Obtenue sur les échantillons considérés appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

1.10.Test de comparaison de deux variances

On conserve les même notations que précédemment et on teste (H_0) contre (H_1) où

$$\begin{cases} (H_0) \; : \sigma_1 =: \sigma_2 \\ (H_1) \; : \sigma_1 \neq: \sigma_2 \end{cases}.$$

Au risque de première espèce α.

On pose

$$T = \frac{S_1^2}{S_2^2}.$$

La variable aléatoire de décision du test. Sous l'hypothèse (H_0), T suit une loi de Fischer-Snédécor à $(n_1 - 1; n_2 - 1)$ degrés de liberté. La zone d'acceptation avec un risque de niveau α pour T est :

$$I = [F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) ; F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)],$$

Où $F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Fischer-Snédécor. On sait en effet que :

$$P(T \in I) = 1 - \alpha.$$

Remarque on a :

$$F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = \frac{1}{F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)}.$$

Si la valeur $t = {s_1^2}/{s_2^2}$ obtenue sur les échantillons considérés appartient à I, alors on accepte (H_0), sinon on rejette (H_0).

1.11. Test de comparaison de deux proportions

On considère un caractère qualitatif et on note p_1 la proportion d'individus présentant ce caractère dans une première population-mère, p_2 la proportion dans une seconde population-mère. On souhaite déterminer s'il s'agit d'une même population en ce qui concerne ce caractère.

On teste (H_0) contre (H_1) où

$$\begin{cases} (H_0) : p_1 =: p_2 \\ (H_1) : p_1 \neq: p_2 \end{cases},$$

au risque de première espèce α On suppose que F_1 et F_2 suivent approximativement des lois normales. Soient f_1 et f_2 les fréquences observées sur les deux échantillons. On note

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}.$$

Et on pose

$$T = \frac{F_1 - F_2}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1}+\frac{1}{n_2})}},$$

La variable aléatoire de décision du test. Sous l'hypothèse (H_0), T suit approximativement une loi normale centrée réduite.

La zone d'acceptation avec un risque de niveau $\boldsymbol{\alpha}$ pour T est à nouveau :

$$I = \left[-t_{1-\frac{\alpha}{2}} \; ; \; t_{1-\frac{\alpha}{2}}\right],$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On sait en effet que :

$$P(T \in I) = 1 - \alpha,$$

Si la valeur

$$t = \frac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}},$$

Obtenue sur les échantillons considérés appartient à I, alors on accepte (H_0), sinon on rejette(H_0).

1.12. Test du Chi-deux d'ajustement

Soit X une variable aléatoire définie sur une population-mère. On dispose d'un échantillon de valeurs (x_1, x_2,…, x_n) et on souhaite tester si la distribution expérimentale observée correspond à une distribution théorique donnée.

On définit sur Ω, k événements $E1, \dots, Ek$ formant un système complet d'événements, c'est à-dire :

(i) $\Omega = \bigcup_{i=1}^{k} E_i$

(ii) $\forall\, i \neq j, E_i \cap E_j = \emptyset$

Dans le modèle théorique, on note p_1, p_2,…, p_k les probabilités de ces événements.

Sur l'échantillon (x_1, x_2,…, x_n) on note n_1, n_2,…, n_k les effectifs observés de ces événements. On va les comparer aux effectifs théoriques obtenus pour un échantillon de taille n : ces effectifs

Valent np_1,np_2,…, np_k .

On teste :

(H_0) : la distribution observée est conforme à la distribution théorique choisie, contre

(H_1) : la distribution observée n'est pas conforme à la distribution théorique choisie,

Au risque de première espèce α.

- **Décision concernant le test :**

On considère la variable aléatoire de décision :

$$x_c^2 = \sum_{i=1}^{k} \left(\frac{N_i - np_i}{np_i}\right)^2$$

Où Ni est la variable aléatoire qui compte l'effectif observé de l'événement E_i sur un échantillon donné **:**

$$N_i\,(x_1, x_2, \ldots, x_n) = n_i\,.$$

Sous l'hypothèse (H_0), la variable aléatoire X_C^2

Suit une loi du X^2à $k - 1 - r$ degrés de liberté où r est le nombre de paramètres à estimer éventuellement pour connaître la distribution théorique.

On impose à la zone d'acceptation d'être un intervalle ayant 0 pour borne inférieure, donc

On pose :

$$I = [0, X_{1-\alpha}^2(k - 1 - r)],$$

Où $X_{1-\alpha}^2(k - 1 - r)$ est le quantile d'ordre $1 - \alpha$ de la loi du X^2à $k - 1 - r$ degrés de liberté

On note $\mathbf{X_0^2}$ la valeur prise par $\mathbf{X_c^2}$ sur l'échantillon :

$$x_0^2 = \sum_{i=1}^{k} \left(\frac{n_i - np_i}{np_i}\right)^2$$

Si $X_0^2 \in I$, alors l'hypothèse (H_0) est acceptée, sinon elle est rejetée.

Remarque

On exige que les effectifs observés ni pour chaque événement soient ≥ 5. Si ce n'est pas le cas, on fusionne deux ou plusieurs événements.

Exemple1

Dans un centre avicole, des études antérieures ont montré que la masse d'un œuf choisi au hasard peut être considéré comme la réalisation d'une variable aléatoire normale X, de moyenne m et de variance

$\boldsymbol{\sigma}^2$. On admet que les masses des œufs sont indépendantes les unes des autres. On prend un échantillon de taille n=36 œufs que l'on pèse. Les mesures sont données par ordre croissant :

50,34	52,62	53,79	54,99	55,82	57,67
51,41	53,13	53,89	55,04	55,91	57,99
51,51	53,28	54,63	55,12	55,95	58,10
52,07	53,30	54,76	55,24	57,05	63,15
52,22	53,32	54,78	55 ,28	57,18	50,58
52,38	53,39	54,93	55,56	57,31	63,15.

a) Calculer la moyenne empirique et l'écart- type empirique de cette série statistique.
b) Donner une estimation des paramètres m et σ.
c) Donner un intervalle de confiance au niveau 95%, puis 98%, de la masse moyenne m d'un œuf.
d) Tester si la moyenne de cette variable est égale à 56.

Solution

a) $\bar{x}$=55,083 s= 2,683 Q_1=53,29 Med= 54,96 Q_3=56,5.

b) $\bar{x}$ est une estimation de m, et s est une estimation de σ.

c) L'intervalle de confiance 1-α =95% pour m :

$$[\bar{x}\text{-}z_{\alpha/2}\frac{s}{\sqrt{36}}\ ,\ \bar{x}\text{+}z_{\alpha/2}\frac{s}{\sqrt{36}}\ ] = [54{,}207 \quad 55{,}96],$$

Car $z_{\alpha/2}=z_{0.001}$, P[Z≤2.3263]=0.99 quand Z de loi N(0,1), et donc $z_{\alpha/2}$=2.3263.

Exemple2

On suppose que le poids d'un nouveau né est une variable d'écart-type égale à 0.5 kg.

Le poids moyen des 49 enfants nés au mois de janvier 2004 dans l'hôpital de Terga Ouzemour a été de 3.6 kg.

a) Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau né dans cet hôpital.
b) Quel serait le niveau de confiance d'un intervalle de longueur 0.1 kg centré en 3.6 kg pour ce poids moyen ?

Solution

a) IC de niveau de confiance 95% pour le poids moyen :

$$[\bar{x}\text{-}1.9\frac{\sigma}{7}\ ,\ \bar{x}\text{+}z_{\alpha/2}\frac{\sigma}{7}\ ] = [3.46 \quad 3.74]$$

b) $[\bar{x}\text{-}0.05 \le m \le \bar{x}\text{+}0.05] = P[\frac{-0.05}{\frac{\sigma}{7}} \le \frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} \le \frac{0.05}{\frac{\sigma}{7}}]$

$$= 2F(\frac{0.05}{0.5/7}) = 2F(0.7)\text{-}1 = 2*0.758\text{-}1 = 0.516$$

Le niveau de confiance est donc 0.516

Chapitre 02 : Sondage

L'objectif dans ce chapitre consiste à étudier des procédures de sondage pour lesquelles nous pourrons répondre à ces questions, présenter le contexte, les notations ainsi que les critères permettant d'évaluer la qualité d'un sondage.

Nous proposerons différents plans de sondage permettant d'estimer des moyennes et proportions.

2.1. Définitions

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

- 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, 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.

2.2. Notation

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

On souhaite évaluer une caractéristique de la population. On note Xi la valeur de ce caractère mesuré sur l'individu i. On peut utiliser un sondage pour estimer la moyen, le total, la variance.....etc.

La moyenne :

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

Le total

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

Une proportion d'individus : qui vérifie un certain critère. Dans ce cas, Xi prendra deux valeurs :

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

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

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

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 ($n \leq N$). Ce sous-ensemble est appelé échantillon, noté par E.

On désignera par $x_1, x_2, \ldots, x_n$ les valeurs de la caractéristique observées sur l'échantillon. Ces valeurs sont connues, et tout le problème consiste à estimer les paramètres inconnus à partir des valeurs mesurées sur l'échantillon (valeur connues).

2.3. 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. Les mêmes concepts étudiés sont également valables pour l'estimation d'un total ou d'une proportion.

Soient $U = (U_1, U_2, \ldots, U_N)$

Désigne la population ou la base de sondage et

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

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

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

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 μˆ, 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 μˆ sur tous les échantillons possibles est la vraie valeur μ.

➢ **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.

2.3.1. Plans avec ou sans remise

➢ **Définition**

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

Exemple

P = {1, 2, 3, 4, 5}, n = 2. L'échantillon {1, 2} est différent de l'échantillon {1, 3}. Dans le cas d'unplan avec remise, il y a N^n échantillons possibles.

- **Définition**

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

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

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

2.4. Plan aléatoire simple

- **Définition (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.

Remarque

Dans le cas d'un plan simple sans remise, un échantillon de taille fixe n a donc une probabilité égal 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 $\frac{2\times3\times2}{5\times4\times3\times2}=\frac{1}{10}$

- **Proposition (Probabilité d'inclusion)**

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

2.5. Estimation de la moyenne

2.5.1. Estimation ponctuelle

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

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 $\overline{X}$:

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

- **Théorème**

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

On a alors

$$E(\bar{\mu}) = \mu.$$

Dit autrement, $\bar{\mu}$ est un estimateur sans biais de μ.

- **Théorème**

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

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

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. A 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, 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 qui suive).

2.5.2. 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%).

➢ **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 est connue

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

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

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

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 possible ; d'autre part, on prévient qu'il existe un risque faible que la vraie valeur soit en dehors de la fourchette

2.6. Estimation d'une proportion

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

2.7.1. Estimation ponctuelle

Construisons la variable aléatoire x_i qui à l'ième individu 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 l'individu 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 :

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

Et $V(\hat{p}) = (1-f)\frac{s^2}{n} = (1-f)\frac{Np(1-p)}{n(N-1)}$

2.7.2. 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 :

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

Avec

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

D'où L'IC

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

Remarque

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

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

Ce qui donne l'intervalle

$$\hat{p} - z_{1-\alpha/2}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}\ ;\ z_{1-\alpha/2}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}} \qquad (2.4)$$

2.8. Taille d'échantillon

La taille d'échantillon n était fixée. Cependant, on pose souvent la question au : "A partir de combien d'élément un échantillon est-il représentons?". Bien entendu, il faut définir ce qu'on entend par représentons, 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

2.8.1. 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 :

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

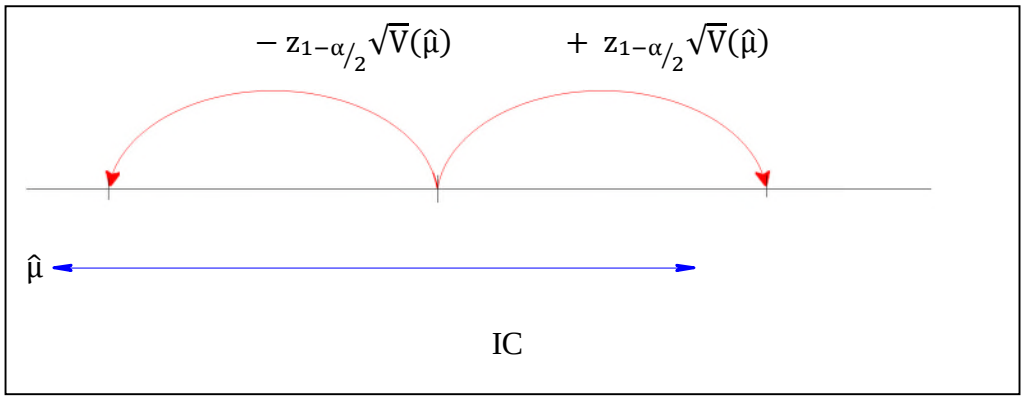

Fig. 2.1 – Intervalle de confiance.

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

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

Ou encore

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

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

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^2_{max} 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

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

(On se place dans le pire des cas, c'est à dire celui où la variance vaut $\boldsymbol{\sigma^2_{max}}$). Par conséquentla taille d'échantillon minimum n_0 telle que la demi-longueur de l'IC ne dépasse pas h0 sera la solution de l'équation

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

C’est-à-dire

$$n_0 = \frac{z_{1-\alpha/_2}^2 \sigma_{max}^2}{h_0^2}$$

2.8.2. 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

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

Donc, 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

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

Chapitre 3 Sondage Stratifié

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.

- ✓ 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.

La stratification consistent à :

1. partitionner la population U en H groupes (les strates), notés u_1, …,u_H telle que à l'intérieure de chaque strate h, la dispersion S_h^2 de Y est faible ;
2. à l'intérieure de chaque strate h, tirer des échantillons indépendants selon un plan p_h.

Grace à la faible dispersion dans chaque strate, les estimateurs devaient être plus précis, ce qui donnera une variance globale plus faible.

Le plan stratifié va permettre de poser a priori une exigence de précision minimale par strate, en choisissant judicieusement les tailles d'échantillon dans chaque strate.

3.1. Plan de sondage stratifié

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

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 l'IIème individu. On cherche d'estimer la moyenne du caractère sur la population

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

On suppose que la population p est partagée en H sous-ensembles ou strates notées p_h

h = 1,. . . ,H. On définit:

- ✓ Taille de la strate h **:** N_h **:**
- ✓ Moyenne de la strate h : $\mu_h = \mu = \frac{1}{N}\sum_{i=1}^{n} X_i$

✓ Variance de la strate : $\sigma_h^2 = \frac{1}{N_h}\sum_{i\in P_h}(X_i - \mu_h)^2$

Variance corrigée de la strate h : $S_h^2 = \frac{1}{N_h-1}\sum_{i\in h}(X_i - \mu_h)^2 = \frac{N_h}{N_h-1}\sigma_h^2$

➢ **Proposition**

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

✓ $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 = \frac{1}{N}\sum_{h=1}^{H} N_h\sigma_h^2 + \frac{1}{N}\sum_{h=1}^{H} N_h(\mu_h - \mu)^2$

= Variance intra-strate + Variance inter-strate

Le premier terme représente la moyenne des variances des strates. Le second est dû aux différences entre strates.

Nous sommes maintenant en mesure de définir un plan stratifié.

➢ **Définition**

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

$$\sum_{h=1}^{H} n_h = n$$

3.2. Estimateur de la moyenne

Une fois l'échantillonnage effectué, en passent à l'estimateur de la moyenne μ.

$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{n} N_i \bar{x}_i = \frac{1}{N}(N_1\bar{x}_1 + \ldots + N_n\bar{x}_n)$

3.2.1. 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 **:**

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

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

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

- **Théorème**

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

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

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

3.3. Répartition de l'échantillon

Dans le plan stratifié, on suppose que les tailles d'échantillons $n_h$étaient fixées pour chaque strate. En pratique, au coure du sondage, on doit se poser la question suivante : combien de personnes nous sondons par strate pour que notre estimateur soit le plus précis possible ?

Dit autrement, comment choisir les n_h ?

Pour la répartition du plan stratifié on a besoin d'un plan avec allocation proportionnelle ou un plan avec allocation optimal.

		Population P inconnu, déterministe	Echantillon E connu, aléatoire
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	$\overline{X\,h}$
	Variance	σ_h	
	Variance Corrigée	S_h^2	S_h^2

Tab. 3.2 – Notations pour le plan stratifié

3.3.1. 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 :
- ✓ $f_h = n_h /N=n/N=f$

➢ **Définition**

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

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

Remarque

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 $\sum_{i=1}^{h} ni = n$)

➢ **Proposition**

Soit μˆ l'estimateur construit pour un plan avec allocation proportionnelle. On a :

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

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) :

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

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

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

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

$\sigma^2_{intra} < \sigma^2$

Remarque

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

On a par exemple la partition suivante :

strate	N_h
Etudiant	6000
Enseignant	2500
ATS	1500

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

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

3.3.2. Plan avec allocation optimal

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**

Dans un plan stratifié avec allocation optimale, on choisit les tailles d'échantillons $n_1, \ldots, n_H$

Telles que $\sum_h^H n_h = n$ et telles que la variance de l'estimateur $V(\hat{\mu})$ 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

- ✓ 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.
- ✓ La formule précédente nécessite de connaître les variances corrigées de chaque strate Sh (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 α est donné par

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

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

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 souhaitent 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 stratifié selon ces catégories. Pour simplifier, on considère 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,…,n l'âge de la $i^{éme}$ personne interrogée. L'estimateur de μ est

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

   La variance d'un tel estimateur est donnée par

   $V(\hat{\mu})=(1-n/N)S^2/n$.

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

$S^2 = (N/N-1)\,\sigma^2 = (10000/999)16^2 = 256.03$

On déduit

$V(\hat{\mu}) = (1-100/10000)256.03/100 = 2.53$

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

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

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 suivants :

$n_h = n\,(N_h/N)$.

Par conséquent,

$n_1 = 100\times2000/10000=20$, $n_2 = 100\times3000/10000=30$, $n_3 = 100\times5000/10000=50$.

Calculons les variances corrigées par strate $S^2_h = (N/N-1)\,\sigma^2_h$

$S^2_1 = 324.03$ $S^2_2 = 124.01$ $S^2_3 = 12.96$

La variance de l'estimateur est donner par :

$V(\hat{\mu}) = 1/n(1-n/N)1/N\sum_{h=1}^{H} S_h^2 N_h = 1.10$

3. Pour un plan avec allocation optimal, je laisse à l'étudiant d'inspirer la solution.

Chapitre 4 : Sondage par grappes

4.1.Principe et justification

Quel que soit l'estimateur utilisé, le sondage aléatoire simple présente deux inconvénients

Il faut disposer d'une base de sondage complète

Lorsqu'à chaque individu est associé un coût de déplacement pour pouvoir réaliser l'interview, le sondage aléatoire simple peut disperser suffisamment l'échantillon pour que le coût total soit positif.

Pour contourner les deux difficultés, on peut utiliser la technique du sondage par grappes. On commence par construire des groupes d'individus qui soient disjoints et dont la réunion est la population toute entière. On tire alors (généralement par sondage aléatoire simple) un certain nombre de groupes dans la base de sondage de groupes qui a été constituée. Chaque groupe est un individu ou unité d'échantillonnage (On parle de grappes). On réalise ensuite l'enquête auprès de tous les individus de la grappe (recensement de la grappe).

Exemples

- ✓ Enquêtes passagers (AIR Algerie) : Les passagers sont regroupés naturelle- ment par vols. L'échantillonnage consiste alors à choisir un certain nombre de vols. Tous les passager d'un vol seront interroger (grappes de passagers).

- ✓ Estimation à 20 heures : Dans les bureaux de vote les bulletins sont comptés en paquet de 100. L'échantillonnage consiste donc à choisir un certain nombre de bureaux de vote (fermant à 18 h) et retenir les résultats des premiers paquets de 100 bulletins (grappes de votes)

- ✓ Contrôle de qualité : Les produits alimentaires sont livrés par lots (ou caisses). L'Échantillonnage consiste à prélever un certain nombre de lots. Ensuite on contrôle le contenu des lots (Grappes de marchandises).

4.2. Estimation et calcul de précision

On note :

M : nombre total d'unité primaire constituée

Ni : taille de l'unité primaire i.

N : Taille de la population totale

$$N = \sum_{i=1}^{M} N_i$$

$\overline{N}$ **:** La taille moyenne des grappes

$$\overline{N} = \frac{1}{M} \sum_{i=1}^{M} N_i$$

$\boldsymbol{Y_{ij}}$: la valeur de la variable d'intérét pour l'individu j de l'unité primaire i.

$\boldsymbol{T_i}$: $\sum_{j=1}^{N_i} \boldsymbol{Y_{ij}}$le total de la grappe i

$\overline{\overline{\boldsymbol{Y}}}_{\boldsymbol{i}}$: $\frac{\mathbf{1}}{\boldsymbol{N_i}} \boldsymbol{T_i}$la moyenne de la grappe i. (double barre car moyenne d'unité secondaire)

$\overline{\boldsymbol{T}}$: $\frac{\mathbf{1}}{\boldsymbol{M}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} \boldsymbol{T_i}$le total moyen

$\boldsymbol{T}$: $\sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} \boldsymbol{T_i}$le total global

$\overline{\overline{\boldsymbol{Y}}}$: $\frac{\boldsymbol{T}}{\boldsymbol{M}}$la moyenne générale

$\boldsymbol{m}$: taille de l'échantillon d'unité primaire.

L'estimateur du total est

$$\hat{T} = \sum_{i \in s} \frac{T_i}{\frac{m}{M}} = M \times \frac{\sum_{i \in s} T_i}{m}$$

Contrairement au as du sondage aléatoire simple, il n'y a donc pas besoin de connaître la taille totale N de la population pour estimer un total T . Le sondage par grappe permet don d'estimer N (qui est un total particulier). Cependant l'estimation sera de mauvaise qualité dès que les tailles Ni seront sensiblement différente d'une U P à l'autre. La connaissance de N est par contre nécessaire pour estimer la moyenne $\overline{\boldsymbol{Y}}$:

$$\hat{\overline{\overline{\boldsymbol{Y}}}} = \frac{\mathbf{1}}{\boldsymbol{N}} \hat{\boldsymbol{T}}$$

4.3. Calcul de précision

Ces formules étant des cas particuliers des formules des tirages à plusieurs degrés.
Soit

$$\boldsymbol{S_1^2} = \frac{\mathbf{1}}{\boldsymbol{M-1}} \times \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} (\boldsymbol{T_i} - \overline{\boldsymbol{T}})^{\mathbf{2}}$$

La dispersion des$(\boldsymbol{T_i})_{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{M}}$. La variance de l'estimateur est alors

$$\boldsymbol{V(\widehat{T}) = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m}}$$

De plus si on note

$$\boldsymbol{S_1^2 = \frac{1}{m-1} \sum_{i \in s} \left(\widehat{T}_i - \frac{\widehat{T}}{M}\right)^2}$$

Alors $\boldsymbol{S_1^2}$ est un estimateur sans biais de la variance. On en déduit la variance de l'estimateur sans biais de la moyenne :

$$V\left(\hat{\bar{\bar{Y}}}\right) = \left(\frac{\boldsymbol{M}}{\boldsymbol{N}}\right)^2 \left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{M}}\right) \times \frac{\boldsymbol{S_1^2}}{\boldsymbol{m}} = \left(\frac{\boldsymbol{1}}{\boldsymbol{N}}\right)\left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{M}}\right)^2 \times \frac{\boldsymbol{S_1^2}}{\boldsymbol{m}}$$

Dont un estimateur sans biais est

$$V\left(\hat{\bar{\bar{Y}}}\right) = \left(\frac{M}{N}\right)^2 \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m} = \left(\frac{1}{N}\right)\left(1 - \frac{m}{M}\right)^2 \times \frac{S_1^2}{m}$$

4.4.Cas particulier des grappes uniformes

4.4.1. Précision :

Dans ce cas on a $\boldsymbol{N_i}$=$\boldsymbol{\widehat{N}}$:=$\boldsymbol{N_0}$, n=m× $\overline{\boldsymbol{N}}$ est une constante non aléatoire et

$$\hat{\bar{\bar{Y}}} = \frac{M}{N} \sum_{i \in s} \frac{T_i}{m} = \frac{1}{m} \sum_{i \in s} \frac{T_i}{N} = \frac{1}{m} \sum_{i \in s} \bar{\bar{Y}}_i$$

Et

$$V\left(\hat{\bar{\bar{Y}}}\right) = \left(1 - \frac{m}{M}\right)\frac{1}{M-1}\frac{\sum_{i=1}^{M}\left(\bar{\bar{Y}}_i - \bar{\bar{Y}}\right)^2}{m}$$

4.5. Comparaison avec le sondage simple

Supposons qu'on puisse réaliser un échantillon de même taille n, sans tenir compte du groupement en grappes.

On a

$$\mathrm{n} = \sum_{\mathrm{i}\in\mathrm{s}} \mathrm{N_i} = \mathrm{mN_0}$$

La moyenne $\bar{\bar{\mathrm{Y}}}$ sera estimée par

$$\bar{\bar{\mathrm{Y}}} := \frac{1}{n}\sum_{i\in s} y_i$$

Dont la précision sera

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

Où $S^2 = \frac{1}{\mathrm{N}-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2$

Comme $\mathrm{N} = \mathrm{MN_0}$ et $\mathrm{n} = \mathrm{mN_0}$ on a :

$$V(\bar{\bar{y}}) = \frac{1}{N_0}\left(1 - \frac{m}{M}\right)\frac{S^2}{m}$$

Comparons$\bar{\bar{y}}$ et $\hat{\bar{\bar{Y}}}$ on a

$$V\left(\hat{\bar{\bar{Y}}}\right) \simeq \left(1 - \frac{m}{M}\right)\frac{1}{M-1}\frac{\sum_{i=1}^{M}\left(\bar{\bar{Y}}_i - \bar{\bar{Y}}\right)^2}{m} = \left(1 - \frac{m}{M}\right)n^2 S^2$$

où n^2est le rapport de corrélation inter-grappes

$$n^2 = \frac{\sum \frac{N_0}{N}\left(\bar{\bar{Y}}_i - \bar{\bar{Y}}\right)^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2}$$

Par conséquent

$V\left(\hat{\bar{\bar{Y}}}\right) < V(\bar{\bar{y}}) \leftrightarrow n^2 < \frac{1}{N_0}$. Le sondage en grappe est donc meilleur que le sondage simple si le rapport de corrélation inter grappes est inférieur à $\frac{1}{N_0}$

En conclusion

- ✓ Il est souhaitable que les moyennes de grappes $\bar{\bar{Y}}_i$ Soient le plus semblable possible
- ✓ Il est souhaitable que la taille N_0Soit petite.

Le sondage en grappe est intéressant s'il y a beaucoup de petites grappes le plus ressemblantes possibles. Ces deux exigences sont contradictoires puisque si N_0 est faible il y a beaucoup de grappe et la dissemblance des moyennes $\bar{Y}_i$ augmente. On a donc recours à la stratification des grappes, qui auront des moyennes similaires dans une même strate.

4.6. Tirage stratifie des grappes

On dispose de K strates de tailles $M_1, M_2, \dots, M_k$ on tire alors des échantillons de tailles $m_1, m_2, \dots, m_k$ telles que

$$m = \sum_{k=1}^{k} m_k$$

Pour estimer un total T on aura

$$\widehat{T} = \sum_{k=1}^{K} M_k \, \widehat{T}_k$$

où

$$\widehat{T}_k = \frac{\sum_{i \in s_k} T_i}{m_k}$$

Et

$$V(\widehat{T}) = \sum_{k=1}^{K} M_K^2 \, V(\widehat{T}_k) = \sum_{k=1}^{K} M_K^2 \left(1 - \frac{m_k}{M_k}\right) \times \frac{S_{k1}^2}{m_k}$$

où

$$S_{k1}^2 = \frac{1}{M_k - 1} \times \sum_{i=1}^{M_k} \left(T_i^k - \overline{T}_k\right)^2$$

où $\overline{T}_k$est le total moyen de la strate k et T_i^k le total de la grappe i de la strate k

D'où

$$\hat{V}(\hat{T}) = \sum_{k=1}^{K} M_k^2 \, \hat{V}\left(\hat{\bar{T}}_k\right) = \sum_{k=1}^{K} M_k^2 \left(1 - \frac{m_k}{M_k}\right) \times \frac{S_{k1}^2}{m_k}$$

Avec

$$S_{k1}^2 = \frac{1}{m_k - 1} \times \sum_{i \in s_k} (T_i^k - \hat{\bar{T}}_k)^2$$

L'estimateur de la variance dans la strate k.

Chapitre 5 : Sondage à Plusieurs Degrés

Il s'agit d'une généralisation du sondage par grappes. Au lieu de recenser tous les individus des unités primaires on procède à un deuxième tirage à l'intérieur de

Ces unités. Les individus tirés au sein des UP sont appelés unités secondaires (US). Comme pour le sondage par grappes, le sondage à plusieurs degrés est généralement moins précis que le sondage aléatoire simple. Intuitivement, lorsqu'on forme des uni- tés primaires, les individus qui forment une unité primaire donnée sont semblables du point de vue de la variable d'intérêt Y. Le tirage au second degré perd alors de son information.

On parle alors d'effet de grappe pour exprimer cette idée de redondance d'information au sein de l'unité constituée à chaque degré.

5.1. Estimation et calcul de précision

On se place dans le cadre d'un sondage à deux degré et on suppose qu'on tire des unités par sondage aléatoire simple à chaque degré. On rappelle que l'on note

M : nombre total d'unité primaire constituée
Ni : taille de l'unité primaire i.
N : Taille de la population totale

$$\overline{N} = \sum_{i=1}^{M} N_i$$

$$\overline{N} = \frac{1}{M} \sum_{i=1}^{M} N_i$$

$\boldsymbol{Y_{ij}}$:La valeur de la variable d'intérêt pour l'individu j de l'unité primaire i**.**

$\boldsymbol{T_i}$ **:**$\sum_{j=1}^{N_i} \boldsymbol{Y_{ij}}$la total de la grappe i.

$\overline{\overline{\boldsymbol{Y}}}_{\boldsymbol{i}}$: $\frac{\mathbf{1}}{\boldsymbol{N_i}} \boldsymbol{Y_i}$la moyenne de la grappe i.(double barre car moyenne d'unité secondaire)

$\overline{\boldsymbol{T}}$: $\frac{\mathbf{1}}{\boldsymbol{M}} \sum_{i=1}^{M} \boldsymbol{T_I}$le total moyen

T : $\sum_{i=1}^{M} \boldsymbol{T_1}$ le total global

$\overline{\overline{\boldsymbol{y}}}$: $\frac{\boldsymbol{T}}{\boldsymbol{N}}$La moyenne générale

$\boldsymbol{m}$: taille de l'échantillon d'unité primaire.

$\boldsymbol{n_i}$: la taille de l'échantillon d'unités secondaire dans l'unité i.

5.1.1. Estimation

On sait que le vrai total de l'unité primaire i, est estimé sans biais par

$$\widehat{\boldsymbol{T}}_{\boldsymbol{i}} = \sum_{j \in S_i} \frac{\boldsymbol{Y_{ij}}}{\frac{\boldsymbol{n_i}}{\boldsymbol{N_i}}} = \boldsymbol{N_i} \times \sum_{j \in S_i} \frac{\boldsymbol{Y_{ij}}}{\boldsymbol{n_i}}$$

Où $\boldsymbol{S_i}$ désigne la liste des unités secondaires échantillonnées dans l'unité primaire i.

$\overline{\boldsymbol{N}}$ **;** la taille moyenne des grappes

On a donc

$$\widehat{T}_i = N_i \times \bar{\bar{y}}_i$$

Où $\bar{\bar{y}}_i$désigne la moyenne

$$\bar{\bar{y}}_i = \frac{\sum_{j \in S_i} Y_{ij}}{n_i}$$

L'estimateur sans biais du total est alors

$$\widehat{\mathbf{T}} = \sum_{\mathbf{i} \in \mathbf{s}} \frac{\widehat{\mathbf{T}}_\mathbf{i}}{\frac{\mathbf{m}}{\mathbf{M}}} = \mathbf{M} \times \frac{\sum_{\mathbf{i} \in \mathbf{s}} \widehat{\mathbf{T}}_\mathbf{i}}{\mathbf{m}}$$

Où s désigne la liste des unités primaires échantillonnées. Là n'y a pas besoin de connaitre N pour estimer le total.

On remarquera que dans le cadre de e plan de sondage, la probabilité d'inclusion d'un individu de l'UP i dépend du rapport $\frac{N_i}{n_i}$ Comme ces poids varient généralement selon l'UP, le tirage à plusieurs degré est en général un tirage à probabilités inégales.

5.1.2. Calcul de précision

On a la variance de $\widehat{\mathbf{T}}$:

$$\mathbf{V}(\widehat{\mathbf{T}}) = \mathbf{M}^2 \times \left(1 - \frac{\mathbf{m}}{\mathbf{M}}\right) \times \frac{\mathbf{S}_1^2}{\mathbf{m}} + \frac{\mathbf{M}}{\mathbf{m}} \times \sum_{\mathbf{i}=1}^{\mathbf{M}} \mathbf{N}_\mathbf{i}^2 \times \left(1 - \frac{\mathbf{n}_\mathbf{i}}{\mathbf{N}_\mathbf{i}}\right) \times \frac{\mathbf{S}_{2\mathbf{i}}^2}{\mathbf{n}_\mathbf{i}}$$

Avec

$$\mathbf{S}_1^2 = \frac{1}{\mathbf{M} - 1} \times \sum_{\mathbf{i}=1}^{\mathbf{M}} (\mathbf{T}_\mathbf{i} - \overline{\mathbf{T}})^2$$

Et

$$\mathbf{S_{2i}^2}=\frac{\mathbf{1}}{\mathbf{N_i-1}}\times\sum_{\mathbf{J=1}}^{\mathbf{N_i}}(\mathbf{Y_{ij}}-\bar{\bar{\bar{\mathbf{Y}}}}_{\mathbf{I}})^{\mathbf{2}}$$

Le premier terme $\mathbf{A}\coloneqq\mathbf{M^2}\times(\mathbf{1}-\frac{\mathbf{m}}{\mathbf{M}})\times\frac{\mathbf{S_1^2}}{\mathbf{m}}$ ne fait intervenir que les grandeurs liées au premier degré de tirage (tirage des UP) est la variance INTER (sous-entendu « inter UP » . Le second terme $\mathbf{B}\coloneqq\frac{\mathbf{m}}{\mathbf{M}}\times\sum_{\mathbf{i=1}}^{\mathbf{M}}\mathbf{N_i^2}\times\left(\mathbf{1}-\frac{\mathbf{n_i}}{\mathbf{N_i}}\right)\times\frac{\mathbf{S_{2i}^2}}{\mathbf{n_i}}$ est la variance INTRA (sous-entendu « intra UP »).

Preuve on aq

$$\widehat{\mathbf{T}}=\mathbf{M}\times\sum_{\mathbf{i}\in\mathbf{s}}\frac{\widehat{\mathbf{T}}}{\mathbf{m}}$$

Considérons UP la variable aléatoire des unités primaires échantillonnées, par la formule de la variance conditionnelle.

$$\mathbf{V}\left(\widehat{\mathbf{T}}\right)=\ \mathbf{E_{UP}}\left[\mathbf{V}\left(\widehat{\mathbf{T}}\middle|\mathbf{UP}\right)\right]+\ \mathbf{V_{UP}}\left[\mathbf{E}\left(\widehat{\mathbf{T}}\middle|\mathbf{UP}\right)\right]$$

où

$$\widehat{\mathrm{T}}=\frac{1}{\mathrm{m}}\sum_{\mathrm{i}\in\mathrm{m}}\widehat{\mathrm{T}}_{\mathrm{i}}$$

5.2. Estimation et calcul de précision

$$\mathbf{S_{2i}^2} = \frac{\mathbf{1}}{\mathbf{n_i - 1}} \sum_{\mathbf{j \in s_i}} \left(\mathbf{Y_{ij}} - \bar{\bar{\mathbf{y}}}_\mathbf{i}\right)^\mathbf{2},$$

Où $\bar{\bar{\boldsymbol{y}}}_{\boldsymbol{i}} = \frac{\mathbf{1}}{\boldsymbol{n_i}} \sum_{\boldsymbol{j \in s_i}} \mathbf{Y_{ij}}$ est l'estimateur sans biais de $\bar{\bar{\mathbf{Y}}}_\mathbf{i}$

Preuve on remarque d'abord que $\sum_{i=1}^{M} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{\boldsymbol{S_{2i}^2}}{\boldsymbol{n_i}}$ est estimé au premier degré par

$$\frac{M}{m} \sum_{i \in s} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \times \frac{S_{2i}^2}{n_i}$$

De plus $\boldsymbol{S_{2i}^2}$ est estimé par $\boldsymbol{S_{2i}^2}$(cas aléatoire simple classique pour le deuxième degré). Donc

$$\frac{M}{m} \sum_{i=1}^{M} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{S_{2i}^2}{n_i}$$

Estimé par

$$\frac{M^2}{m^2} \sum_{i \in s} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{S_{2i}^2}{n_i}$$

D'autre part, pour estimer le terme

$$S_1^2 = \frac{1}{M - 1} \times \sum_{i=1}^{M} (T_i - \bar{T})^2$$

On pense naturellement à son homologue sur l'échantillon :

$$S_1^2 = \frac{1}{m-1}\sum_{i\in s}\left(\hat{T}_i - \hat{\bar{T}}\right)^2$$

On va donc calculer l'espérance de

$$(m-1)S_1^2 = \sum_{i\in s}\left(\hat{T}_i - \hat{\bar{T}}\right)^2$$

$$=\sum_{i\in s}\left(\hat{T}_i - \hat{\bar{T}}\right)^2 - m\left(\hat{\bar{T}} - \bar{t}\right)^2$$

D'où

$$E[(m-1)s_1^2] = \sum_{i\in s} V\left(\hat{T}_i\right) - mV(\hat{\bar{T}})$$

Avec

$$V(\hat{T}_i) = V_{UP}\left(E\left[\hat{T}_i \middle| UP\right]\right) + E_{UP}[V(\hat{T}_i|UP)]$$

$$= V_{UP}(T_i) + \frac{1}{M}\sum_{i=1}^{M} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{S_{2i}^2}{n_i}$$

$$= \frac{1}{M}\sum_{i=1}^{M}(T_i - \bar{T})^2 + \frac{1}{M}\sum_{i=1}^{M} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{S_{2i}^2}{n_i}$$

Donc

$$\sum_{i\in s}\left(\hat{T}_i\right) = \frac{m}{M}\sum_{i=1}^{M}(T_i - \bar{T})^2 + \frac{m}{M}\sum_{i=1}^{M} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{S_{2i}^2}{n_i}$$

De plus $\hat{\bar{T}} = \frac{\bar{T}}{M}$ donc

$$V\left(\hat{\bar{T}}\right)=\frac{V(\hat{T})}{M^2}$$

Et par la formule

$$mV\left(\hat{\bar{T}}\right) = \left(1 - \frac{m}{M}\right) S_1^2 + \frac{1}{M}.\sum_{i=1}^{M} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

Finalement

$$E[(m-1)S_1^2] = \frac{1}{M} S_1^2\big(m(M-1) - (M-m)\big) + \frac{m-1}{M} \sum_{i=1}^{M} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

$$= (m-1)(S_1^2 + \frac{1}{M}.\sum_{i=1}^{M} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

Donc $\boldsymbol{S_1^2}$n'est pas l'estimateur de S_1^2, ce dernier est

$$S_1^2 = S_1^2 - \sum_{i \in s} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

Enfin en rassemblant les deux termes de $V(\hat{T})$ on a :

$$\hat{V}(\hat{T}) = M^2 \frac{M-m}{Mm} \left(S_1^2 - \frac{1}{M}. \sum_{i=1}^{M} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i} \right) + \frac{M^2}{m^2} \sum_{i \in s} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

$$= M^2 \left(1 - \frac{m}{M}\right) \frac{S_1^2}{m} + \frac{M}{m} \sum_{i \in s} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2i}^2}{n_i}$$

On remarque donc que la similarité des formes entre l'estimateur et la valeur à estimer n'est

qu'une coïncidence.

5.3. Simplification

Si on suppose que les taux de sondage au second degré sont constants, 'est-à- dire si on choisit, pour toute UP i, $\boldsymbol{n_i}$ Proportionnel à N_i selon $n_i = f_{2N_i} N_i$ et si on s'intéresse çà un total sur la populaion, on a la formulation simplifiée de l'estimateur du total :

$$\hat{T} = \frac{1}{f_1 \times f_2} \sum_{i \in s} \sum_{j \in s_j} Y_{ij}$$

Ou f_1 est le taux de sondage des UP ($f_1 = \frac{m}{M}$). Avec pour formule de précision :

$$V(\hat{T}) = M(1 - f_1) \times \frac{S_1^2}{f_1} + \frac{1 - f_2}{f_1 \times f_2} \times \left(\sum_{i=1}^{M} N_i \times S_{2i}^2 \right)$$

Si de plus, toute les UP ont la même taille $N_i = \bar{N}$ alors

$$\hat{T} = N \times \bar{\bar{y}}$$

Et la variance précédente s'écrit

$$\mathrm{V}(\hat{\mathrm{T}}) = N^2 \times \left((1 - f_1) \times \frac{\mathrm{S}_1^2}{m} + \frac{1 - f_2}{m \times \bar{n}} \times \left(\frac{\sum_{i=1}^{M} \mathrm{S}_{2\mathrm{i}}^2}{M} \right) \right)$$

Avec $\acute{}\mathrm{S}_1^2 = \frac{1}{\mathrm{M}-1} \sum_{\mathrm{i}=1}^{\mathrm{M}} \left(\bar{\bar{\mathrm{Y}}}_{\mathrm{i}} - \bar{\bar{\mathrm{Y}}} \right)^2$

La dispersion vraie des vrais moyenne $\bar{\bar{\mathrm{Y}}}_{\mathrm{i}}$, des UP i , L'estimateur sans biais de $\mathrm{V}(\hat{\mathrm{T}})$ dans ce cas simplificateur est

$$\hat{V}(\hat{\mathrm{T}}) = N^2 \times \left(\frac{1-f_1}{m} \times \acute{S}_1^2 + f_1 \times \frac{1-f_2}{m \times n} \times \left(\frac{1}{m} \sum_{i \in s} s_{2i}^2 \right) \right)$$

Avec

$$\acute{S}_1^2 = \frac{1}{\mathrm{m}-1} \sum_{\mathrm{i} \in \mathrm{s}} (\bar{\bar{\mathrm{y}}}_\mathrm{i} - \bar{\bar{\mathrm{y}}})^2 \text{ et } \mathrm{s}_{2\mathrm{i}}^2 = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{j} \in \mathrm{s}_\mathrm{i}} (\mathrm{Y}_{\mathrm{ij}} - \bar{\bar{\mathrm{y}}}_\mathrm{i})^2$$

On remarque que le terme contenant les dispersions s_{2i}^2 est multiplié par le taux de sondage au premier degré f_1 qui est généralement très faible. Cela permet de négliger le second terme de $\hat{V}(\hat{\mathrm{T}})$devant le premier terme. D'ou

$$\hat{V}(\hat{\mathrm{T}}) \simeq N^2 \times \frac{1-f_1}{m} \times \acute{S}_1^2$$

La dispersion $\acute{\mathbf{S}}_\mathbf{1}^\mathbf{2}$ exprime à elle seul pratiquement toute la dispersion.

5.4. Sondage à trois degré ou plus

Lorsqu'on introduit des degrés supplémentaires, la précision s'en trouve presque toujours diminué. On peut ainsi voir apparaître dans les variances un terme propre à chaque degré, généralisant ainsi e que l'on avait constaté pour le tirage à deux degrés. Ainsi, dans un tirage à trois degrés ave sondage aléatoire simple à chaque degré, lorsque la population comprend M unité primaires (UP) ayant ha une R unité secondaires (US) et chaque US possède elle-même U unité tertiaires, on montre que si on échantillon m UP, puis r US dans chaque UP, puis u UT dans chaque US, alors un estimateur $\hat{\mathbf{T}}$ sans biais de T est

$$\hat{\mathrm{T}} = \frac{\mathrm{N}}{\mathrm{m} \times \mathrm{r} \times \mathrm{u}} \sum_{\mathrm{i} \in \mathrm{s}} \sum_{\mathrm{j} \in \mathrm{s_i}} \sum_{\mathrm{k} \in \mathrm{s_{ij}}} Y_{ijk} = N \times \bar{\bar{\bar{Y}}}$$

Ou s, $s_i$et s_{ij} sont les échantillons respectifs d'UP, d'US et d'UT, Y_{ijk}la valeur de Y pour L'UT k appartenant à l'US j appartenant ç l'UP i et $\bar{\bar{\bar{\mathbf{Y}}}}$ la moyenne globale de la population. La variance vraie de cet estimateur sera

$V(\hat{T}) = N^2 \times \frac{1-f_1}{m} \times \acute{S}_1^2 +\rightarrow$ Contribution du premier degré

$\frac{1-\mathrm{f_2}}{\mathrm{m \times r \times u}} \times \left(\frac{1}{\mathrm{M}} \sum_{\mathrm{i=1}}^{\mathrm{M}} \acute{\mathrm{S}}_1^2\right) +\rightarrow$Contribution du second degré

$\frac{1-\mathrm{f_3}}{\mathrm{m \times r \times u}} \times \left(\frac{1}{\mathrm{M \times R}} \sum_{\mathrm{i=1}}^{\mathrm{M}} \sum_{\mathrm{j=1}}^{\mathrm{R}} \acute{\mathrm{S}}_{13\mathrm{ij}}^2\right) \rightarrow$Contribution du troisième degré

Avec

$$\acute{\mathrm{S}}_1^2 = \frac{1}{\mathrm{M}-1} \sum_{\mathrm{i=1}}^{\mathrm{M}} \left(\bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{i}} - \bar{\bar{\bar{\mathrm{Y}}}}\right)^2$$

$$\acute{\mathrm{S}}_{2\mathrm{i}}^2 = \frac{1}{\mathrm{R}-1} \sum_{\mathrm{j=1}}^{\mathrm{R}} \left(\bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{i}} - \bar{\bar{\bar{\mathrm{Y}}}}\right)^2$$

Et

$$\acute{\mathrm{S}}_{3\mathrm{ij}}^2 = \frac{1}{\mathrm{U}-1} \sum_{\mathrm{K=1}}^{\mathrm{U}} \left(\bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{ijk}} - \bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{ij}}\right)^2$$

$\bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{i}} = \frac{1}{\mathrm{R-U}} \sum_{\mathrm{j=1}}^{\mathrm{R}} \sum_{\mathrm{K=1}}^{\mathrm{U}} \mathrm{Y_{ijk}}$= vraie moyenne dans L 'UP i.

$\bar{\bar{\bar{\mathrm{Y}}}}_{\mathrm{ij}} = \frac{1}{\mathrm{U}} \sum_{\mathrm{k=1}}^{\mathrm{U}} \mathrm{Y_{ijk}}$= vraie moyenne dans L'us j de L'UP i

Les termes f1, f2 et f3 représentent respectivement les taux de sondage aux degrés un, deux, trois. Ainsi V $\hat{\mathrm{T}}$ est la somme apportée par chaque degré de tirage, le degré x à une contribution du type :

$$\frac{1-\text{Tauxdesondageaudegré x}}{\text{Produitdestaillesd'échantillons}} \times \text{moyennededispersioncalculéeaudegré x} - 1$$

A chaque degré, on introduit la moyenne vraie de toutes les unités finales contenues dans l'unité considérée. On a intérêt à avoir, a chaque degré, une dispersion minimale de es moyennes, 'est-à-dire, en vertu de la formule de dé composition de la variance, une hétérogénéité maximale des individus au sein de chaque unité.

Prolongeant le as du tirage à deux degrés, l'estimation de variance fait apparaître une su cession de termes numériquement de plus en plus faibles, puisque :

$$\hat{V}(\hat{T}) = N^2 \times \left[\frac{1-f_1}{m} \times \acute{S}_1^2 + f_1 \times \frac{1-f_2}{m \times r} \times \left(\frac{1}{m}\sum_{i \in s} \acute{S}_{2i}^2\right) + f_1 \times f_2 \frac{1-f_3}{m \times r \times u} \times \left(\frac{1}{m \times r}\sum_{i \in s}\sum_{j \in s} \acute{S}_{3ij}^2\right)\right]$$

Avec

$$\boldsymbol{\acute{S}_1^2} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\bar{\bar{\bar{y}}}_i - \bar{\bar{\bar{y}}}\right)^2$$

$\acute{S}_{2i}^2 = \frac{1}{r-1}\sum_{j \in s_i}^{M}\left(\bar{\bar{\bar{y}}}_{ij} - \bar{\bar{\bar{y}}}_i\right)^2$ Et $\quad \acute{S}_{3ij}^2 = \frac{1}{u-1}\sum_{k \in s_{ij}}\left(Y_{ijk} - \bar{\bar{\bar{y}}}_{ij}\right)^2$

Où

$\bar{\bar{\bar{y}}}_{\boldsymbol{i}} = \frac{\mathbf{1}}{\boldsymbol{r-1}}\sum_{\boldsymbol{j \in s_i}}\sum_{\boldsymbol{k \in s_{ij}}} \boldsymbol{y_{ijk}}$ = Moyenne de l'échantillon dans l'UPi.

$\bar{\bar{\bar{\bar{y}}}}_{\boldsymbol{ij}} = \frac{\mathbf{1}}{\boldsymbol{u}}\sum_{\boldsymbol{k \in s_{ij}}} \boldsymbol{y_{ijk}}$ = moyenne de l'échantillon dans l'US j de L'UP i

Si le taux de sondage au premier degré f1 est faible, on pourra se contenter du terme faisant intervenir$\boldsymbol{\acute{S}_1^2}$, qui intègre une partie importante de la variance :

$$\hat{V}(\hat{T}) \simeq N^2 \times \frac{1-f_1}{m} \times \acute{S}_1^2$$

5.5.L’effet de grappe

5.4.1. Définition, interprétation

L'e effet de grappe survient dans les tirages à plusieurs degrés. Dans la très grande majorité des as, il traduit un phénomène de perte de précision due à l'existence d'une similarité entre les individus d'une même UP. Chaque degré de tirage amène son e et de grappe. Cet effet peut être mesuré par un coefficient appelé « coefficient de corrélation intra-grappe : ρ. Lorsqu'on manipule deux degrés de sondage, coefficient vaut :

$$\rho = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k\neq 1}^{N_i} \left(Y_{ij} - \overline{Y}\right)(Y_{ik} - \overline{Y})}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(Y_{ij} - \overline{Y}\right)^2} \times \frac{1}{N - 1}$$

Où $\overline{N}$ est la taille moyenne des UP $\left(\overline{N} = \frac{N}{M}\right)$. Il ressemble à un coefficient de corrélation linéaire mais il met en jeu qu'une seule variable Y.

Chapitre 6 ; Méthodes à Choix raisonnées

6.1. Méthodes à Choix raisonnées

Les méthodes à Choix raisonnées s'oppose aux sondages probabiliste en e sens où il n'est plus possible de déterminer à priori quelle est la probabilité qu'à chaque individu de la population d'appartenir à l'échantillon. En général, on les utilise lorsqu'il y a absence de base de sondage. Le but de l'opération est de se rapprocher au maximum d'un tirage rigoureusement aléatoire.

6.1.1. Méthodes des quotas

Principe

La méthode empirique la plus fréquemment rencontrée est la méthode des quotas : on fait en sorte que la structure de l'échantillon soit exactement elle de la population toute entière selon certain critère que l'on a préalablement choisis. Par exemple, si on sait que la population est constituée de 45% d'hommes et de 55% de femmes,

On cherchera à avoir 45% d'hommes et 55% de femmes (d'où le terme de quotas). L'estimateur d'une moyenne $\bar{Y}$ est alors égal à la moyenne $\bar{y}$ annulée sur l'échantillon. En e et si il y a Q catégories de populations sur lesquelles on s'impose un quota (par
Exemple les Q modalités d'une variable explicative) on a

$$\hat{\bar{Y}} = \sum_{q=1}^{Q} \frac{N_q}{N} \times \hat{\bar{Y}} = \sum_{q=1}^{Q} \frac{n_q}{n} \times \bar{y}_q = \bar{y}$$

où

- ✓ $\bar{Y}_k$ est la moyenne des Y dans la catégorie q, annulée sur les individus de l'échantillon$N_q$est la taille vraie (connue) de la population dans la catégorie q

- ✓ n_q est la taille de l'échantillon dans la catégorie q

- ✓ Puisque le propre de la méthode est d'imposer

$$\frac{\mathbf{n_q}}{\mathbf{n}} = \frac{\mathbf{N_q}}{\mathbf{N}}$$

Ainsi, si l'enquêteur essuie un refus ou trouve un logement vide, il ne sera pas tenu, Contrairement à e qui se fait avec le sondage probabiliste, à insister ou à renouveler plusieurs fois sa tentative de prise de conta t

6.1.1. Biais

Si on veut éviter un biais substantiel, il faut supposer que les valeurs des variables d'intérêt ne sont expliquée que par nos critères de quotas et qu'il n'y a plus de Critère caché non pris en compte qui pourrait les influencer. Si on réalise par exemple un sondage d'opinion politique, sachant grâce à des études préalables que l'opinion est bien expliquée par le sexe et l'âge, on peut baser nos quotas sur le croisement sexe-âge. Si l'opinion ne dépendait rigoureusement que du croisement sexe-âge, on pourrait laisser l'enquêteur complètement libre de sélectionner les individus dans chaque croisement, puisque son choix n'aurait finalement plus d'importance. Si par contre l'opinion dépend, en plus, du revenu et si e critère n'est pas in lut dans les quotas, l'enquêteur peut sélectionner par mal han e les individus d'un croisement ayant plutôt des revenus élevés. Cela entraînera un biais dans l'estimateur que l'on dénomme biais de sélection.

Les biais de sélection peuvent être diminués :

- ✓ Etalement des interviews sur l'ensemble de la journée et de la soirée.
- ✓ Etalement sur l'ensemble du territoire concerné …etc.

6.1.2. Précision

La précision des estimateurs par quotas n'est pas calculable, puisque au une probabilité p(s) n'est connue. Les contrôles qui sont imposés par les quotas limitent la marge de manœuvre laissé à l'aléa. Si on tient absolument à fournir un résultat numérique de précision, on peut toujours utiliser la formule de la variance d'un sondage stratifié à allocation proportionnelle.

Dans le cas de petits échantillons, le sondage probabiliste peut être de biais nul et de variance assez forte. Dans les mêmes conditions, le sondage empirique est biaisé

Mais de variance assez faible. Dans la pratique on utilise les quotas pour les petits échantillons (de taille 1000 et moins) et les sondages probabilistes pour les gros.

6.1.3. Quotas marginaux et quotas croisés

Bien que les méthodes par quota soient fa iles et relativement rapides à mettre en œuvre, il subsiste une difficulté liée à l'absence fréquente d'information concernant les effectif correspondant aux croisements des variables de quota. En e et si on dispose d'une part de la structure par sexe et d'autre part de la structure par âge, rien ne dit que l'on dispose de la structure croisée sexe-âge. Le problème consiste alors à remplir les cases d'un tableau rectangulaire pour obtenir les effectifs échantillonnés connaissant les deux marges. On parle dans e as de quotas marginaux. Si on connaît la structure croisée, on parle de quotas croisés et tout se passe comme si les quotas étaient imposés sur une seule variable.
Il peut être di ile de remplir les ases du tableau si on ne dispose que des seuls quotas marginaux :

Exemple

On dispose de deux critères de quotas qui sont le sexe et l'indicateur d'activité. La taille d'échantillon est 1000. On sait :

Moins de 30 ans : 20%
30-55 ans : 50%

Plus de 55 ans 30%

Ainsi que

Actifs 50%

Chômeurs 10%

Retraités 20%

Inactifs 20%

On peut alors réaliser

Actifs	**500**	**100**	**380**	**20**
Chômeurs	100	25	25	50
Retraités	200	0	20	180
Inactifs	200	75	75	50
		200	500	300
		-30ans	30-55 ans	+55ans

OU

Actifs	**500**	**100**	**380**	**20**
Chômeurs	100	25	25	50
Retraités	200	0	20	180
Inactifs	200	75	75	50
		200	500	300
		-30ans	30-55 ans	+55ans

Ces deux tableaux sont très différents mais tous deux acceptables.

Bien entendu, il ne s'agit pas pour l'enquêteur de pré remplir d'avance et à son gré les effectifs croisés du tableau à partir des marges mais de laisser faire le hasard des rencontres qui devrait le conduire à une structure croisé proche de la véritable structure croisée inconnue.

Après une enquête par quotas marginaux, on utilise en général comme estimateur brut une moyenne simple $\bar{y}$. En e et si on repère par (i, j) la case du tableau qui correspond au

croisement des modalités i (lignes) et j (colonne), on appelle n_{ij} la taille de l'échantillon dans la case (i, j), N_{ij} la taille (inconnue) de la population totale dans (i, j) et en n $\overline{Y}_{ij}$la vraie moyenne (inconnue) dans (i, j). La vraie moyenne globale

$$\overline{\mathrm{Y}} = \sum_{\mathrm{i,j}} \frac{\mathrm{N_{ij}}}{\mathrm{N}} \times \overline{\mathrm{Y}}_{\mathrm{ij}}$$

Or

$$\overline{\boldsymbol{y}} = \sum_{\boldsymbol{i,j}} \frac{\boldsymbol{n_{ij}}}{\boldsymbol{Nn}} \times \overline{\boldsymbol{y}}_{\boldsymbol{ij}}$$

L'argument consiste à dire que $\overline{\boldsymbol{y}}_{\boldsymbol{ij}}$ estime bien en moyenne $\overline{\boldsymbol{Y}}_{\boldsymbol{ij}}$ et que la porpotion $\frac{\boldsymbol{n_{ij}}}{\boldsymbol{n}}$ est non seulement égale en moyenne

à $\frac{\boldsymbol{N_{ij}}}{\boldsymbol{N}}$mais doit etre peu variable. La raison intuitive de cette faible variabilité est le contrôle des marges, qui impose les effectifs marginaux :

- ✓ $\boldsymbol{n_{.j}} = \sum_{\boldsymbol{i}} \boldsymbol{n_{ij}}$ pour tout j
- ✓ $\boldsymbol{n_{i.}} = \sum_{\boldsymbol{j}} \boldsymbol{n_{ij}}$ pour tout i

Cela va limiter la variabilité des effectif croisés $\boldsymbol{n_{ij}}$ (en retournant l'argument, on peut dire que si les effectifs par case $\boldsymbol{n_{ij}}$ peuvent beaucoup varier, alors les effectifs des marges $\boldsymbol{n_{i.}}$et $\boldsymbol{n_{.j}}$peuvent aussi beaucoup varier). Aussi $\frac{\boldsymbol{n_{ij}}}{\boldsymbol{n}}$ sera un bon estimateur de $\frac{\boldsymbol{N_{ij}}}{\boldsymbol{N}}$ si n est suffisamment grand. Ainsi

$$\overline{\boldsymbol{y}} \simeq \frac{\boldsymbol{N_{ij}}}{\boldsymbol{N}} \times \overline{\boldsymbol{y}}_{\boldsymbol{ij}}$$

Cette équation approximative traduit la for e de la méthode qui conduit, à partir d'une moyenne simple $\bar{y}$, à une précision qui pourrait être assimilée en première approximation à elle d'un estimateur

Post-stratifié. Il subsiste cependant une différence de fond : Le système des quotas permet d'agir au stade de l'échantillonnage et d'utiliser par la suite un estimateur simple, alors que la post-stratification agit au stade de l'estimation car il s'agit d'une méthode de redressement.

Si on considère la variable auxiliaire X_q^i qui vaut 1 si l'individu i vérifie la modalité q de la variable qualitative sur laquelle on fonde les quotas, et 0 sinon. Alors on voit que le propre de la méthode des quotas est d'imposer

$$\bar{X}_q = \bar{x}_q$$

Semblables à l'individu moyen pour la variable d'intérêt de l'enquête. Cela revient à admettre une relation stable reliant la variable d'intérêt aux variables servant à définis l'individu représentatif.

En ore plus que la méthode des quotas, les résultats seront sensibles au tirage au travers du choix que l'on a fait de l'individu

Moyen. Historiquement, elle a été appliquée pendant la période de guerre au sondage agricoles, alors que l'échantillon- nage aléatoire en était à ses débuts et que la situation exigeait un technique rapide et peu coûteuse. Des statisticiens italiens (Gini et Galvani) ont conçu un échantillon de districts tirés du recensement italien de 1921 à partir de districts type qui produisait des estimations très satisfaisante de sept moyenne sociodémographique importantes (taux de natalité, mortalité, et ...). Des études ultérieures ont montré que des moyennes de variables autres que elles qui étaient contrôlées s'avéraient être mal estimées, ainsi que d'autres fonction que les moyennes des variables ayant étaient contrôlées

6.2. Echantillonnage de volontaires

Si on pousse la logique de l'économie à l'extrême, on aboutit à l'échantillonnage de volontaire, très en vogue sur internet. Cette méthode très peu onéreuse n'est pas « statistiquement » valide puisque :

- ✓ On ne connait pas la probabilité d'inclusion des personnes et elle- i doit être extrêmement variable d'un individu à l'autre sur la population.
- ✓ On ne peut pas mesurer le taux de non-réponse
- ✓ Il y a, a priori, une très forte corrélation entre les valeurs Yi des volontaires et leur probabilité de sélection Pi.

Ce type de risque existe à peu près systématiquement avec les enquêtes par courrier n'ayant pas un caractère obligatoire, par exemple lorsqu'un magazine demande à ses lecteurs leur sentiment sur tel ou tel aspects : Ceux qui répondent ont quelque choses à dire et sont plutôt très contents ou très mécontents de et aspe t.

En 1936 aux Etats-Unis, un grand journal organisa un sondage préélectoral auprès de ses lecteurs volontaires. Plus de 2 millions lecteurs répondirent et le journal prédit la victoire de Landon et ce fut

6.2.1. Sondage probabiliste ou méthode des quotas ?

La Banque ne pratique pas la méthode des quotas, car ils manipulent de très gros échantillons. Des études comparatives ont été menées en vraie grandeur sur des é han- tillons de petite taille. On a constaté l'équivalence des performances des deux méthodes. Il semble que l'on puisse accorder une assez grande on an e aux enquêtes par quotas sérieusement menées.

6.3. Méthode des itinéraires

C'est une variante de la méthode des quotas, au lieu de laisser l'enquêteur dé- terminer librement les individus à interroger, on lui impose de réaliser ses interviews dans certains endroits

définis par avance sur une carte. L'avantage, par rapport à la méthode des quotas classiques est de limiter la liberté de choix de l'enquêteur.

6.4. Méthode des unités types

Elle consiste à choisir un individu « moyen » que l'on dé lare représentatif d'un groupe d'individus possédant les mêmes caractéristiques. Sa validité repose sur le pari que les individus semblables à l'individu moyen selon les variables qui ont servi précisément à définis et individu seront également Roosevelt qui fut élu.

6.5. En conclusion

Il existe encore d'autres méthodes empiriques (unités-type, échantillonnage de volontaires) mais les résultats de ces méthodes sont très peu fiables

Chapitre 07: Jackknife et bootstrap appliqués aux sondages

Les méthodes bootstrap et jackknife peuvent être utilisées pour estimer le biais et l'erreur d'une estimation, et les mécanismes des deux méthodes d'échantillonnage ne sont pas très différents: échantillonnage avec remplacement ou exclusion d'une observation à la fois. Cependant, le jackknife n'est pas aussi populaire que le bootstrap dans la recherche et la pratique, donc que donnent-elles comme résultats une foi appliquées aux sondages ?

7.1. Introduction au Jackknife

7.1.1. Jackknife classique

Estimation du biais Cette technique a été introduite par Quenouille en 1949 pour estimer le biais d'un estimateur. Soit $T_n = T(X_1, X_2, \cdots, X_n)$ un estimateur d'un paramètre inconnu θ. Le biais de $T_n$est

$$\text{Biais}(Tn) = E[Tn] - \theta$$

Soit

$$T_{n-1,i} = T_{n-1}(X_1, , X_2, \cdots, X_{i-1}, X_{1+1}, \cdots, X_n)$$

La statistique basée sur toutes les observations sauf l'IIème. L'estimateur Jackknife du biais est

$$b_{jack} = (n-1)(\bar{T}_n - T_n)$$

$$\widehat{T}_{n,i} \coloneqq nT_n - (n-1)T_{n-1,i}$$

Et il estima $\mathrm{Var}(T_n)$ par l'estimateur jackknife de la variance :

$$v_{jack} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{\bar{T}}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \hat{\bar{T}}_{n,j} \right)^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{n-1,i} - \frac{1}{n} \sum_{j=1}^{n} (T_{n-1,j}) \right)^2$$

Qui est généralement asymptotiquement équivalent à l'estimateur obtenu par le TCL

7.1.2. Delete-dJackknife

On calcule ici

$$T_{r,s} := T_r (X_i, i \in s^c)$$

Où S est un sous-ensemble de $\{1, \cdots, n\}$ de taille d et $\boldsymbol{s^c}$est le complémentaire de s de taille $r = n - d$. Pour une statistique T_n donnée, $T_{r,s}$est la même statistique mais basée sur les r observations obtenues en enlevant $\{X_i, , i \in s\}$ de l'échantillon

$\{X_i, X_2, \cdots, X_1, \}$. Il s'agit don de la généralisation du ja kknife pré dent qui est le cas particulier $d = 1$.

Pour une statistique T_n, l'estimateur de la variance du delete-d Jackknife est

$v_{jack-d} = \frac{r}{dN} \sum_{s \in S} \left(T_{r,s} - \frac{1}{N} \sum_{s \in S} T_{r,s} \right)^2$ Ou

$$\overline{T}_n = \frac{\sum_{i=1}^{n} T_{n-1,i}}{n}$$

On obtient alors l'estimateur débiaisé par jackknife

$$T_{jack} = T_n - b_{jack} = nT_n - (n-1)\overline{T}_n$$

Estimation de la variance Tuckey en 1958 prouva que jackknife pouvait être utilisé pour estimer la variance.

Il définit

$$\widehat{T}_{n,i} := nT_n - (n-1)T_{n-1,i}$$

Et il estima $Var(T_n)$ par l'estimateur jackknife de la variance :

$$v_{jack} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\widehat{T}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\widehat{T}_{n,j}\right)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}(T_{n-1,j})\right)^2$$

Qui est généralement asymptotiquement équivalent à l'estimateur obtenu par le TCL

7.1.3. Delete-dJackknife

On calcule ici

$$T_{r,s} := T_r\,(X_i, i \in s^c)$$

Où S est un sous-ensemble de $\{1,\cdots,n\}$ de taille d et s^cest le complémentaire de s de taille $r = n - d$. Pour une statistique T_n donnée, $T_{r,s}$ est la même statistique mais basée sur les r observations obtenues en enlevant $\{X_i,, i \in s\}$ de l'échantillon

$\{X_i, X_2,\cdots, X_1,\}$. Il s'agit don de la généralisation du ja kknife pré dent qui est le cas particulier $d = 1$.

Pour une statistique T_n, l'estimateur de la variance du delete-d Jackknife est

$$\mathrm{v} = \frac{r}{dN}\sum_{s\in S}\left(T_{r,s} - \frac{1}{N}\sum_{s\in S}T_{r,s}\right)^2$$

où S est l'ensemble des sous-ensembles de taille d de $\{1,\cdots,n\}$ et $N = C_n^d$ son cardianal

7.2. Introduction au bootstrap

Cette technique a été introduite par E fron en 1979, supposons que $\mathrm{X_1},,\cdots,\mathrm{X_n}$ sont des variables i.i.d. de fonction de distribution F estimée par $\widehat{\mathrm{F}}$. L'estimateur de la variance bootstrap est :

$$\mathrm{v_{boot}} = \int\left[\mathrm{T_n(x)} - \int \mathrm{T_n\,(y)d}\prod_{\mathrm{i=1}}^{\mathrm{n}}\hat{\mathrm{F}}\,(\mathrm{x_i})\right]^2 \mathrm{d}\prod_{\mathrm{i=1}}^{\mathrm{n}}\hat{\mathrm{F}}\,(\mathrm{x_i})$$

Comme cette expression est généralement in calculable on utilise l'approximation par la Méthode de Monte-Carlo :

$$\mathrm{v}_{boot}^{(B)} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum\nolimits_{l=1}^{B}T_{n,l}^*\right)^2$$

où $(\mathrm{X_{1b}^*}\ \cdots,\mathrm{X_{nb}^*}),\mathrm{b} = 1,\cdots,\mathrm{B}$ sont tirés au hazard (avec remise) dans $\mathrm{X_1},,\cdots,\mathrm{X_n}$
et $\mathrm{X_{1b}^*} = \mathrm{T_n}\,(\mathrm{X_{1b}^*}\ \cdots,\mathrm{X_{nb}^*})$. On obtient de manière similaire l'estimateur du biais :

$$b_{boot}^{(B)} = \frac{1}{B}\sum_{b=1}^{B}T(X_{1b}^*\ \cdots,X_{nb}^*) - T_n$$

7.3. Application du Jackknife et du bootstrap aux sondages

7.3.1. Notation générale

On considère la forme générale d'un estimateur du total :

$$\widehat{\mathrm{T}} = \sum_{\mathrm{h}=1}^{\mathrm{H}} \sum_{\mathrm{i}\in \mathrm{S_h}} \sum_{\mathrm{j}\in \mathrm{S_{h,i}}} \mathrm{w_{hij}}\,\mathrm{Y_{hij}}$$

où w_{hij}est le poids de sondage de l'individu Y_{hij}, par exemple :

- ✓ si H = 1 et que $\mathrm{S_{h,i}}$est réduit à un individu (cas du sondage aléatoire simple) on aura $w_i = \frac{N}{n}$.
- ✓ Si le sondage est stratifié et que $\mathrm{S_{h,i}}$est réduit ç un individu, on aura $w_{\mathrm{h}i} = \frac{\mathrm{N}_h}{n_h}$
- ✓ Si le sondage est stratifie à deux degrés on aura $w_{\mathrm{h}ij} = \frac{\mathrm{M}_h \mathrm{N}_{hi}}{\mathrm{m}_h n_{hi}}$

Dans la plupart des cas l'estimateur du paramètre θ peuvent s'écrire g(t) ou g est une fonction connue. On remarquera que θ peut être une ration, un coefficient de corrélation, des coefficients de régression.

7.3.2. Application Jackknife

- ➢ **Jackknife classique**

On suppose que nombre total d'unités primaires $\mathrm{m} = \sum_{\mathrm{h}=1}^{\mathrm{H}} \mathrm{m_h}$est grand, pour des entier $\mathrm{h}', \mathrm{i}', 1 \leq \mathrm{H}$ et $1 \leq \mathrm{i}' \leq \mathrm{m}'_{\mathrm{h}}$ fixés soit

$$\widehat{\mathrm{T}}_{\mathrm{h}'\mathrm{i}'} = \sum_{\mathrm{h}\neq \mathrm{h}'} \sum_{\mathrm{i}\in s_{\mathrm{h}}} \sum_{\mathrm{j}\in s_{\mathrm{hi}}} \mathrm{w_{hij}}\,\mathrm{Y_{hij}} + \frac{\mathrm{m_{h}}'}{\mathrm{m_{h}}' - 1} \sum_{\mathrm{i}\in s_{\mathrm{h}}, \mathrm{i}\neq \mathrm{i}'} \sum_{\mathrm{j}\in s_{\mathrm{h}'\mathrm{i}}} \mathrm{w_{h'ij}}\,\mathrm{Y_{h'ij}}$$

L'analogue de $\widehat{\mathrm{T}}$, après que la grappe i' de la strate h' soit détruite. Soit

$\hat{\theta}_{\mathrm{h}',\mathrm{i}'} = \mathrm{g}\left(\widehat{\mathrm{T}}_{\mathrm{h}'\mathrm{i}'}\right)$ et $\widehat{\theta}_{\mathrm{h}'} = \frac{1}{\mathrm{m_h}} \sum_{\mathrm{i}'\in s_{\mathrm{h}}} \hat{\theta}_{\mathrm{h}',\mathrm{i}'}$

Un estimateur jackknife de la variable est ;

$$v_{jack} = \sum_{h=1}^{H} \frac{(1-f_h)(n_h-1)}{n_h} \sum_{i \epsilon s_h} \left(\hat{\theta}_{hi} - \hat{\theta}_h\right)^2$$

De même, pour $\hat{\theta} = g(\hat{T})$, on aura

$$b_{jack} = \sum_{h=1}^{H} (n_h - 1)(\bar{\theta}_h - \bar{\theta})$$

L'utilisation de l'estimateur $\bar{\theta} - b_{jack}$élimine en générale biais d'ordre $\frac{1}{n}$

- Méthode des répliques répétées équilibrés

Cette méthode (Balanced Repeated Replication Method, BRR) est applicable pour les sondages stratifiés quad H est grand et n_h pas trop grand. Supposons que chaque strate h de l'échantillon soit découpée en deux grappes, G_{h1} et G_{h2} dont les totaux sont $\hat{T}_{h1}$et $\hat{T}_{h2}$, on définit les estimateurs des demi-échantillons suivant

$$\hat{T}_k = \sum_{h=1}^{H} \left(\delta_{kh} \hat{T}_{h1} + (1 - \delta_{kh} \hat{T}_{h2})\right)$$

Ou $(\delta_{k1}, \ldots \delta_{kH})$ est un vecteur de R^H avec des 1 et des 0 (2^H vecteur possible). On aura finalement un estimateur du total

$$\hat{T} = \frac{1}{2^H} \sum_{h=1}^{H} \overline{T}_{(k)}$$

Et un estimateur de sa variance :

$$\hat{V}(\overline{T}) = \frac{1}{2^H} \sum_{h=1}^{2^L} \left(\overline{T}_{(k)} - \overline{T}\right)^2$$

Conclusion

De manière peut-être paradoxale, la plupart des composantes des plans de sondage ne font pas appel à des mathématiques très élaborées. Quant à l'aspect probabiliste et statistique, il s'accommode d'un niveau de premier cycle, une connaissance de base en calcul de probabilité et en statistique descriptive suffit, car les sondages constituent une discipline qui ne fait pas appel à des « modèles », du moins dans la perception traditionnelle qu'en ont les statisticiens. Le bon vieux modèle avec des variables aléatoires indépendantes et identiquement distribuées n'est pas vraiment adapté au contexte des sondages. En effet, on travaille en population finie et la problématique est, de ce fait, « descriptive ». On pourrait d'ailleurs offrir à des étudiants un cours de sondage avant même qu'ils n'aient abordé le domaine de la statistique mathématique.

[1] Ardilly P. Les techniques de sondage. Paris : Editions Technip, 1994.

[2] Desabie J. Théorie et pratique des sondages. Paris : Dunod, 1966.

[3] Dussaix AM, Grosbras JM. Exercice de sondages. Collection « Economie etstatistique avancées ». Paris :Economicaédition, 1992.

[4] Levy PS, Lemeshow S. Sampling of populations: methods andapplications. 3rd edition. New-York : J. Wiley and Sons, Inc., 1999.

[5] Tillé Y. Théorie des sondages : échantillonnage et estimation en populations finies. Paris : Dunod, 2001.

[6] Warszawski J, Lellouch J. Méthodes d'estimation dans une enquête parsondage. RevEpidemiol Sante Publique 1997 ; 45 : 150-68.

[6] Rouvière L. Enquête et Sondage, universite RENNE 2, 2008-2009.