Chapitre 1
Premiers pas en théorie des sondages

1.1 Qu'est-ce qu'un sondage ?

Le premier chapitre de ce cours va nous permettre de faire nos premiers pas en théorie des sondages. Nous allons y définir la terminologie de base et spécifier les objectifs et le principe général d'un sondage. Nous allons également nous intéresser à ce qui différencie les méthodes de sondage dites *aléatoires* ou *probabilistes* des méthodes de sondage dites *empiriques* ou *à choix raisonné*. Ce chapitre va ainsi nous permettre de préparer le terrain pour les chapitres suivants.

1.1.1 Qu'est-ce qu'un sondage ?

a) Définitions

Avant toute chose, il nous faut nous poser la question suivante : « Qu'est-ce donc qu'un sondage ? »

Les dictionnaires de la langue française définissent le mot « sondage » comme « l'action de sonder », ou encore comme « l'exploration locale et méthodique d'un milieu à l'aide d'une sonde ou de procédés techniques particuliers ». On pense ici, par exemple, au sondage des fonds marins.

Mais le mot « sondage » possède également une signification précise en statistique ! On s'intéresse à un ensemble d'éléments, d'individus, d'objets…. Cet ensemble – généralement de grande taille – constitue ce que l'on appelle la « population ». Réaliser un sondage dans cette population consiste à y prélever un sous-ensemble d'éléments – appelé un « échantillon » – pour extrapoler ensuite ce qu'on observe dans cet échantillon à l'ensemble de la population.

« Sonder une population » revient donc à étudier ce qui se passe dans l'ensemble de la population à partir des observations réalisées auprès d'un échantillon prélevé dans la population.

Le dictionnaire complète généralement la définition du mot « sondage » en précisant ce que signifie le terme d' « enquête par sondage » ou « sondage d'opinion ». Il s'agit d'une enquête visant à déterminer la répartition des opinions sur une question, dans une population donnée, en recueillant, auprès d'un échantillon prélevé dans la population, des réponses individuelles exprimant ces opinions.

Mais les sondages ne servent-ils qu'à mener des enquêtes d'opinion ? La réponse à cette question est clairement « Non » !

b) Domaines d'application

Le terme « sondage » est surtout connu du grand public de par les très nombreux sondages pré-électoraux, sondages d'opinion ou baromètres politiques réalisés ces dernières années. Mais il serait regrettable de croire que les activités des « sondeurs » ne se limitent qu'à ce type particulier de sondages. Les domaines d'application des sondages sont en réalité très nombreux et variés.

De nombreuses enquêtes démographiques ou socio-économiques, comme celles sur le budget et la consommation des ménages par exemple, ne peuvent être menées auprès de l'ensemble de la population : elles doivent donc être réalisées par sondage.

On trouve également une très grande utilisation des sondages dans les études touchant à l'agriculture, dans les pays développés comme en voie de développement.

Plus largement, on trouve des utilisateurs de sondages dans de très nombreux domaines tels que la psychologie, les sciences de l'éducation, les sciences de la santé, l'écologie, le contrôle industriel de qualité, l'audit, les sciences sociales et politiques, sans oublier bien sûr les études de marché et les mesures d'audience.

1.1.2 Objectif et principe général

Sans doute vous demandez-vous : quand a-t-on besoin de réaliser un sondage ? Et pourquoi le réaliser ?

Répondre à ces questions va nous permettre de dégager l'objectif et le principe général de tout sondage.

Nous venons de voir que les domaines d'application des sondages sont particulièrement diversifiés. Toutefois, quel que soit le domaine considéré, on mettra en œuvre un sondage dans le même type de contexte : celui où l'on doit mener une étude sur un *ensemble d'éléments*, généralement relativement nombreux. Cet ensemble constitue ce que nous allons appeler la *population*.

Quand je vous parle de « population », vous imaginez sans doute un ensemble relativement large d'*individus*. Et c'est effectivement le cas pour bon nombre de sondages. Mais, selon la problématique de l'étude, les éléments de la population peuvent être d'une autre nature. On peut en effet être amené à considérer une population de ménages, de logements, d'entreprises, etc. De façon générale, nous dirons que la population à étudier est constituée d'*unités statistiques*.

Etudier la population consiste notamment à vouloir déterminer la valeur d'une ou plusieurs caractéristique(s) quantitative(s) de cette dernière :

- On peut être intéressé par une *proportion* ; par exemple, la proportion de travailleurs à temps partiel dans la population d'individus considérée.
- On peut vouloir déterminer une *moyenne* ; par exemple, le nombre moyen d'enfants par ménage dans la population de ménages considérée.
- On peut aussi désirer connaître un *total* ; par exemple, le chiffre d'affaires total d'une certaine population d'entreprises.
- Etc.

Ces caractéristiques quantitatives de la population sont ce que nous appellerons des *paramètres-population*. Leurs valeurs ne peuvent être déterminées que si l'on peut mener une étude exhaustive — un recensement — des unités statistiques de la population.

Si une telle analyse exhaustive de la population est impossible, nous pouvons procéder en trois temps :

1. Nous pouvons tout d'abord *prélever un échantillon* d'unités statistiques de la population.
2. Nous pouvons ensuite *enquêter* — autrement dit « interroger » — les unités de cet échantillon.
3. Enfin, à partir des réponses recueillies, nous pouvons déterminer une approximation — nous dirons une *estimation* — de la (ou des) caractéristiques quantitatives ou *paramètres* de la population qui nous intéressent.

En résumé, l'objectif d'un sondage est d'obtenir une *estimation* de la valeur de l'un ou l'autre *paramètre* d'une *population* à partir des observations réalisées dans un *échantillon* prélevé dans cette dernière.

Nous reviendrons de manière un peu plus formelle sur ces notions-clés d'estimation, de paramètres, de population et d'échantillon dans la section suivante.

1.2 Concepts de base

Nous avons conclu la section précédente en résumant l'objectif de tout sondage comme suit : il s'agit d'*estimer* la valeur de l'une ou l'autre caractéristique quantitative ou *paramètre* d'une *population* à partir des observations réalisées dans un *échantillon* prélevé dans cette dernière.

Prenons quelques instants pour formaliser proprement cette situation et pour introduire quelques notations que nous utiliserons régulièrement dans la suite. Pour cela, partons d'un exemple.

1.2.1 Exemple illustratif

Supposons que nous soyons chargés de mener une étude sur les montants dépensés au début de cette année scolaire par les parents des 1 245 élèves d'une certaine école pour l'achat de leur matériel scolaire. Nous aimerions notamment déterminer le montant *total* dépensé pour l'ensemble des élèves de l'école, le montant *moyen* par élève de ces dépenses, ainsi que la *proportion* d'élèves pour lesquels le montant des achats a excédé 200 euros. Comment pouvons-nous procéder ?

Décortiquons soigneusement la situation…

La *population* à laquelle nous nous intéressons correspond à l'ensemble des 1 245 élèves inscrits pour cette année scolaire dans une certaine école. Nous dirons que la population est de *taille* égale à 1 245.

Que voulons-nous étudier dans cette population ?

Nous sommes en fait intéressés par la *distribution* d'une certaine *caractéristique* ou *variable* dans la population. En d'autres termes, nous sommes intéressés par l'ensemble des valeurs que prend cette variable auprès des différents élèves de la population.

Cette variable, que nous appellerons la *variable d'intérêt*, correspond dans notre cas au « montant dépensé en début d'année scolaire par les parents d'un élève pour l'achat de ses fournitures scolaires ».

De manière plus précise, nous aimerions déterminer différentes valeurs typiques ou synthétiques de la distribution de la variable d'intérêt dans la population.

Nous souhaiterions notamment connaître :

- premièrement, le *total* de notre variable d'intérêt dans la population, c'est-à-dire la somme des montants dépensés pour tous les élèves de l'école ;
- deuxièmement, sa *moyenne* dans la population, c'est-à-dire la moyenne des montants dépensés pour tous les élèves de l'école ;
- et enfin, la *proportion* d'élèves de l'école pour lesquels le montant dépensé dépasse 200 euros.

Ces trois quantités caractérisent d'une certaine façon la distribution de notre variable d'intérêt dans la population : elles correspondent à ce que l'on appelle des *paramètres*, ou encore des *paramètres-population*.

Pour calculer les valeurs exactes de ces trois caractéristiques ou paramètres de la population, il nous faut impérativement connaître les montants dépensés pour *chaque* élève de l'école. Bien évidemment, nous ne disposons pas de tous ces montants et il nous est impossible d'interroger l'ensemble des élèves de l'école pour obtenir l'information nécessaire.

Que faire ? La solution naturelle à ce problème consiste à réaliser une enquête par sondage.

Dans notre exemple, on peut ainsi sélectionner un *échantillon* d'une centaine d'élèves inscrits dans l'école durant cette année scolaire, interroger ensuite ces élèves — ou plutôt leurs parents — sur les montants qui ont été dépensés en début d'année pour l'achat de leurs fournitures scolaires, et utiliser enfin les montants observés dans l'échantillon pour obtenir une *estimation* des trois paramètres de la population qui nous intéressent.

En particulier, on peut penser à utiliser le montant moyen des dépenses déclarées pour les élèves de l'*échantillon* pour estimer le montant moyen des dépenses pour l'ensemble des élèves de l'école. De la même manière, la proportion des élèves de l'*échantillon* pour lesquels plus de 200 euros ont été dépensés nous fournit une estimation de la proportion équivalente parmi l'ensemble des élèves de l'école.

Présentons à présent de manière plus formelle les principaux acteurs intervenant dans un sondage et avec lesquels nous avons fait connaissance dans le cadre de notre exemple : la population, la variable d'intérêt, les paramètres, l'échantillon et l'estimation. Cette formalisation vous permettra également de découvrir les notations que nous utiliserons tout au long de ce cours pour les désigner.

1.2.2 Population

La **population** est l'ensemble des unités statistiques sur lesquelles se porte notre intérêt. Nous la désignerons de manière générale par la lettre majuscule U (U comme « univers », autre terme que l'on peut utiliser en lieu et place de « population »). Cette population est composée de N unités statistiques : l'unité 1 (u_1), l'unité 2 (u_2), … , l'unité N (u_N). Nous écrirons de manière synthétique :

$$U = \{u_1, u_2, \dots, u_N\}.$$

Nous dirons que la population U est de **taille** égale à N.

1.2.3 Variable d'intérêt

Nous sommes tout particulièrement intéressés par la **distribution** dans la population d'une certaine variable, appelée **variable d'intérêt** et désignée, tout au long de ce cours, par y. Sa distribution dans la population est caractérisée par les valeurs qu'elle prend auprès des différentes unités statistiques de celle-ci. Nous désignerons ces

valeurs par $y_1, y_2, \dots, y_N$; y_1 correspond à la valeur que prend la variable $\mathcal{Y}$ auprès de l'unité 1, y_2 la valeur qu'elle prend auprès de l'unité 2, etc. Dans l'exemple présenté dans la section précédente, $y_1, y_2, \dots, y_N$ correspondent aux montants dépensés pour l'achat des fournitures scolaires des $N = 1\,245$ élèves de l'école.

Cette variable $\mathcal{Y}$ peut être **quantitative**, comme c'est le cas dans notre exemple. Comme son nom l'indique, elle quantifie alors un certain caractère pour chaque unité de la population et ses valeurs sont de nature numérique.

Mais la variable d'intérêt $\mathcal{Y}$ peut aussi être **qualitative** ou **catégorielle**, auquel cas ses « valeurs » correspondent plutôt à des modalités ou catégories. Si $\mathcal{Y}$ n'a que **deux** modalités possibles, comme la variable « sexe » par exemple, nous dirons que $\mathcal{Y}$ est une variable **dichotomique** et nous prendrons l'habitude de coder ses modalités à l'aide des chiffres 0 et 1. Mais $\mathcal{Y}$ peut aussi être une variable à plus de deux modalités : c'est le cas, par exemple, des variables « catégorie socioprofessionnelle », « domaine d'études », « situation matrimoniale », « nationalité », etc.

1.2.4 Paramètres-population

Nous intéresser à la distribution de la variable $\mathcal{Y}$ dans la population U nous conduit de manière naturelle à vouloir déterminer l'une ou l'autre **valeur typique** ou **synthétique** de cette distribution. Ces valeurs particulières constituent ce qu'on appelle des **paramètres** (ou **paramètres-population**). Dans les prochains chapitres de ce cours, nous considérerons préférentiellement les paramètres suivants :

- le **total** de $\mathcal{Y}$ dans la population U. Ce total, que nous désignerons par la lettre grecque τ (le « t » grec ; se prononce « tau ») ou encore par τ_y lorsque plusieurs variables sont impliquées dans l'étude, se définit comme la somme des valeurs que prend $\mathcal{Y}$ chez toutes les unités de la population U :
$$\tau = y_1 + y_2 + \cdots + y_N = \sum_{i \in U} y_i$$
où la notation mathématique $\Sigma_{i \in U}$ désigne la somme sur toutes les unités u_i ($i = 1, \dots, N$) de la population ;
- la **moyenne** de $\mathcal{Y}$ dans la population. Nous la désignerons par la lettre grecque μ (le « m » grec ; se prononce « mu ») ou encore par μ_y. Elle est égale au total de $\mathcal{Y}$ dans la population, divisé par la taille N de cette dernière :
$$\mu = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N} \sum_{i \in U} y_i = \frac{\tau}{N} ;$$
- la **proportion** d'unités de la population U chez lesquelles la variable $\mathcal{Y}$ prend une valeur satisfaisant une certaine condition : nous désignerons cette proportion par la lettre grecque π (le « p » grec ; se prononce « pi »).

De manière générale, tout paramètre-population θ (se prononce « theta ») synthétisant la distribution de la variable d'intérêt $\mathcal{Y}$ dans la population U se définit comme une certaine fonction des valeurs $y_1, y_2, \dots, y_N$ que prend la variable $\mathcal{Y}$ chez toutes les unités de la population :
$$\theta = \theta(y_1, y_2, \dots, y_N).$$

1.2.5 Sondage : échantillonnage et estimation

Si l'on ne connaît pas ou que l'on ne peut pas déterminer les valeurs de y pour *tous* les individus de la population, il nous est impossible de déterminer la valeur exacte des paramètres-population qui nous intéressent. On fera alors appel à un **sondage** afin d'obtenir une *estimation* de ces derniers.

De manière générale, réaliser une enquête par sondage exige tout d'abord le prélèvement dans la population U d'un **échantillon**. Ce dernier, désigné par la lettre s (première lettre du mot anglais « sample »), contient n_s unités statistiques : nous dirons qu'il est de *taille* n_s. Puisque l'échantillon est un *sous-ensemble* de la population, sa taille n_s est nécessairement inférieure à la taille N de la population ($n_s < N$). Notez ici que, pour bien différencier l'échantillon et la population, nous utiliserons systématiquement un « n » minuscule pour la taille de l'échantillon et un « n » majuscule pour la taille de la population.

Une fois l'échantillon sélectionné, on doit relever ou observer les valeurs y_i prises par la variable d'intérêt y sur les différentes unités u_i de l'échantillon s.

Nous pouvons enfin calculer une certaine fonction mathématique de ces valeurs afin d'obtenir une **estimation** $\hat{\tau}$ du total-population τ, ou une estimation $\hat{\mu}$ de la moyenne-population μ, ou enfin une estimation $\hat{\pi}$ de la proportion-population π. En toute généralité, on estimera le paramètre θ en utilisant une certaine fonction $\hat{\theta}$ des valeurs y_i prises par la variable y auprès des unités statistiques u_i de l'échantillon s :

$$\hat{\theta} = \hat{\theta}(y_i\,; u_i \in s).$$

1.2.6 Remarque

Dans le domaine de la théorie des sondages, les notations varient assez fortement d'un ouvrage à l'autre, ou d'un article à l'autre. J'ai donc pris le parti de vous proposer les notations que j'ai personnellement l'habitude d'utiliser, même si elles ne sont pas celles que l'on rencontre le plus fréquemment. C'est ainsi que, tout au long du cours, tout paramètre relatif à la population sera désigné par une lettre grecque, tandis que les lettres latines seront réservées pour les fonctions des observations de l'échantillon.

Ne soyez pas étonné de ne pas toujours retrouver cette pratique dans les ouvrages que vous pourriez consulter. Ainsi, par exemple, le total τ est parfois noté Y majuscule et la moyenne μ est parfois désignée par $\bar{Y}$.

1.3 Qu'est-ce qu'un « bon » échantillon ?

Dans la section précédente, nous avons vu que l'estimation d'un paramètre-population se fonde sur les valeurs que prend la variable d'intérêt y dans un échantillon prélevé dans la population.

Rappelez-vous ! Dans notre exemple, il semblait naturel d'estimer la moyenne des montants dépensés pour l'ensemble des élèves de l'école par la moyenne des montants dépensés pour les élèves de l'échantillon prélevé.

Dans ce contexte, il est naturel de se poser les deux questions suivantes :

- Premièrement, comment faire pour sélectionner un « bon » échantillon dans la population ? Quelle *procédure d'échantillonnage* doit-on mettre en œuvre ?
- Deuxièmement, comment utiliser les observations réalisées dans l'échantillon, comment les associer ou les combiner, pour obtenir une estimation du paramètre-population qui nous intéresse ?

Nous n'aborderons ici que la première question. Nous reviendrons à la deuxième question dès le prochain chapitre.

Qu'est-ce donc qu'un « bon » échantillon ?

Vous avez peut être envie de me répondre qu'un « bon » échantillon est un échantillon « représentatif de la population ». Mais qu'est-ce qu'un échantillon « représentatif » ?

La notion de « représentativité » d'un échantillon n'a en réalité jamais été définie en théorie des sondages et ce terme « représentatif » est trop souvent utilisé à tort et à travers, sans que celui qui l'utilise ne prenne la peine de préciser la signification qu'il lui donne.

Le plus souvent, toutefois, le caractère « représentatif » d'un échantillon sous-entend le fait que l'échantillon constitue une sorte de « modèle réduit » de la population. En d'autres termes, un échantillon est considéré comme représentatif de la population lorsque la répartition de certaines catégories dans l'échantillon est similaire à la répartition de ces catégories dans la population.

FIGURE 1.1 – *L'échantillon, un modèle réduit de la population*

Si la population est constituée de 53% de femmes, par exemple, on souhaite retrouver un pourcentage similaire de femmes dans l'échantillon. Si l'on retrouve nettement plus ou nettement moins de femmes dans l'échantillon, on aura tendance à considérer que celui-ci est de mauvaise qualité !

Il existe cependant des situations où l'on a intérêt à surreprésenter ou, au contraire, à sous-représenter dans l'échantillon l'une ou l'autre catégorie de la population. Nous verrons cela un peu plus tard lorsque nous étudierons le sondage stratifié.

Donc, le fait que l'échantillon soit un modèle réduit de la population n'est pas ce qui importe le plus pour la qualité de l'échantillon. Ce qui est par contre primordial — et vous vous en rendrez compte au fur et à mesure que vous avancerez dans le cours — c'est qu'on retrouve dans l'échantillon toute la diversité des individus ou des unités statistiques de la population. Il faut que l'échantillon reflète aussi bien que possible toute l'étendue ou l'importance de la variabilité que présente la variable d'intérêt dans la population.

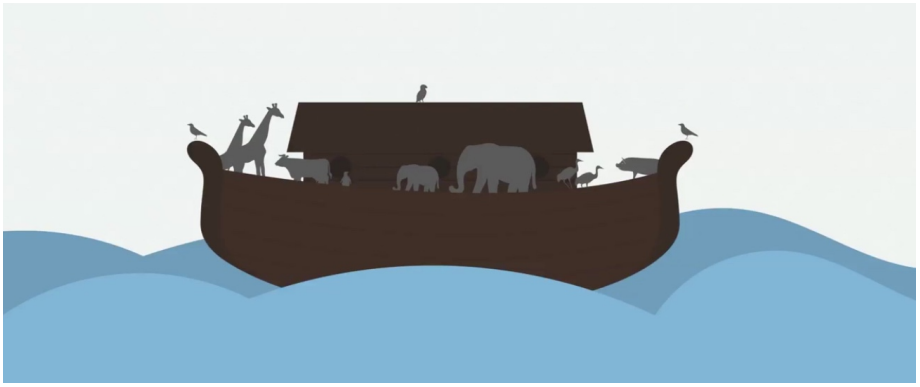


FIGURE 1.2 – *L'échantillon, un sous-ensemble présentant une diversité aussi riche que la population*

Ce n'est que si la diversité des individus de l'échantillon est quasiment aussi riche que celle des individus de la population que l'on pourra valablement extrapoler les résultats obtenus dans l'échantillon au niveau de la population.

Une « bonne » méthode de sondage est donc avant tout une méthode assurant une bonne hétérogénéité de l'échantillon. Il s'agit là d'un principe fondamental à ne jamais perdre de vue lorsqu'on doit réaliser un sondage !

1.4 Les étapes d'un sondage

Réaliser une enquête par sondage ne s'improvise pas. Il faut prévoir son déroulement et faire des choix de stratégies en tenant compte des moyens dont on dispose.

On peut décomposer la démarche à suivre en différentes grandes étapes que je vous propose de passer en revue au travers d'un exemple pratique d'enquête.

1.4.1 Un exemple illustratif

Le service social d'une commune a chargé une équipe de chercheurs en sociologie d'une certaine université de conduire une enquête sur les besoins des personnes âgées résidant dans la commune. Les principales questions soulevées par le service social étaient les suivantes : quel type d'aide à domicile faut-il développer pour les personnes du troisième âge ? Quel lieu d'accueil faut-il privilégier ? Quel type d'activité faut-il proposer ?

Dans un premier temps, les chercheurs ont travaillé à la *conception générale* de l'enquête. Ils ont mené plusieurs réunions de travail avec les responsables du service social de la commune afin de spécifier les grandes lignes du projet, son planning, ainsi que les moyens techniques, humains, financiers… disponibles pour mener ce projet à bien.
Pour rédiger une liste précise des besoins qui allaient faire l'objet de l'enquête, les chercheurs ont organisé des rencontres avec les travailleurs sociaux de la commune en contact avec les personnes âgées ; ils ont également mené des entretiens libres avec 5 personnes âgées de la commune et sont allés consulter un centre documentaire spécialisé en gérontologie.

L'équipe de sociologues s'est ensuite consacrée à la préparation du *plan* ou de la *stratégie d'observation*.
Ils ont spécifié quelle procédure allait être utilisée pour prélever l'échantillon de personnes âgées à interroger. Ils ont également fixé la taille de cet échantillon. La détermination du nombre de personnes à enquêter s'est faite en tenant compte du fait que les chercheurs avaient décidé d'administrer le questionnaire d'enquête via des interviews en face à face, menés par des enquêteurs bien formés devant se rendre au domicile des personnes âgées sélectionnées.
C'est également à cette étape du projet qu'a été préparé le questionnaire d'enquête. Sa rédaction a été relativement délicate et coûteuse en temps de travail. Les chercheurs ont veillé à soumettre le questionnaire qu'ils avaient préparé à un prétest : ils ont administré le projet de questionnaire à une dizaine de personnes âgées, de manière à pouvoir apporter les dernières corrections et modifications indispensables avant la mise en œuvre réelle de l'enquête.

Une fois le plan d'observation finalisé, les chercheurs ont procédé au *prélèvement de l'échantillon* dans la population-cible. Ils ont ensuite lancé la *collecte des données* en envoyant les enquêteurs interroger les personnes âgées sélectionnées. Il va de soi que toutes les personnes sélectionnées pour faire partie de l'échantillon ont d'abord été

contactées personnellement par les chercheurs de l'équipe pour s'assurer de leur accord quant à leur participation à l'enquête et pour fixer le rendez-vous avec l'enquêteur.

L'équipe de recherche s'est ensuite consacrée au *dépouillement* et à l'*encodage* des réponses recueillies. Ils ont pris le soin de contrôler leur fiabilité et leur cohérence, avant de les *analyser* via différentes méthodes statistiques.

Ces analyses ont conduit les chercheurs sociologues à une série de résultats dont ils ont rendu compte au travers d'un *rapport* écrit détaillé. Les résultats ont également été longuement discutés avec le commanditaire de l'enquête — le service social de la commune — au cours d'une journée de séminaire organisée dans le service. Des réponses concrètes ont pu être apportées aux questions que se posait initialement le service social. Les résultats de l'enquête ont permis de dégager les pistes à privilégier dans la politique de service à la population âgée de la commune.

Vous le voyez… cette enquête par sondage a nécessité un travail de longue haleine !

1.4.2 Les étapes d'un sondage

On peut décomposer la démarche d'enquête par sondage en cinq grandes étapes (voir la figure 1.3).

Etape 1

La première étape est celle de la **conception générale** de l'enquête.

Cette première étape consiste, dans un premier temps, à énoncer le problème qui nécessite le recours à l'enquête, à spécifier la question de départ et les grandes lignes du projet, à préciser les moyens techniques, humains, financiers... disponibles pour mener le projet à bien.

Il faut ensuite décomposer l'objectif général de l'enquête en objectifs plus spécifiques formulés sous la forme de questions ou d'hypothèses de recherche bien précises. C'est au cours de cette étape qu'est définie avec soin la population-cible à sonder, que sont précisées les données à recueillir dans l'enquête, que sont définies les variables d'intérêt et les paramètres à estimer.

Ceci impose de rechercher l'information déjà disponible sur le problème considéré et la population visée, que ce soit via une recherche documentaire, via la consultation d'experts et d'acteurs dans le domaine, ou encore via l'analyse d'études similaires. Cette information peut améliorer la compréhension du problème étudié, suggérer des hypothèses à vérifier, aider à la définition de la population-cible.

Etape 2

La deuxième étape est celle de la spécification du **plan** ou de la **stratégie d'observation** : il s'agit de déterminer de quelle façon va se dérouler l'enquête.

Il faut choisir la *méthode d'échantillonnage* qui sera appliquée pour prélever l'échantillon dans la population et déterminer le nombre d'unités à prélever.

Dans le même temps, il faut choisir la *méthode de collecte* de l'information auprès de l'échantillon qui sera sélectionné : de quelle façon va-t-on interroger les unités de l'échantillon ? Par enquêteur en face à face, par voie postale, par téléphone ou par internet ?

C'est également à cette étape qu'il faut élaborer le *questionnaire* d'enquête. Cette tâche est souvent longue, délicate et doit faire l'objet d'un soin tout particulier. Nous en reparlerons plus longuement au dernier chapitre de ce cours, consacré à différentes questions fondamentales de méthodologie d'enquête. Un ou plusieurs essais ou *prétests* du questionnaire doivent impérativement être prévus dans le planning de préparation de l'enquête.

Enfin, il faut garder à l'esprit que le choix de la méthode d'échantillonnage et celui de la méthode de collecte des données ne se font pas indépendamment l'un de l'autre, et doivent tenir compte des diverses contraintes matérielles auxquelles vous êtes soumis (De quel budget disposez-vous ? Quel est le délai de livraison des résultats de l'enquête ? Disposez-vous d'une équipe d'enquêteurs ou êtes-vous pratiquement seul pour mener l'enquête ?)

Etape 3

La troisième étape est celle de la **mise en œuvre de l'enquête** sur le terrain.

On *sélectionne un échantillon* en appliquant sur la population la procédure d'échantillonnage mise au point à l'étape précédente. On *recueille* ensuite *les données* en administrant le questionnaire d'enquête aux unités de cet échantillon.

Etape 4

La quatrième étape est celle du **dépouillement** et de l'**analyse des données**.

Les informations recueillies à l'étape précédente ne peuvent être utilisées telles quelles : il faut commencer par les dépouiller, les présenter sous une forme qui permette l'analyse prévue. Il faut donc réaliser le codage et la saisie des données, en prévision des analyses statistiques qui seront effectuées ensuite. Ce travail doit également être l'occasion d'une vérification de la qualité des données recueillies : celles-ci sont-elles cohérentes ? Certaines d'entre elles apparaissent-elles comme aberrantes ? etc.

Etape 5

La dernière étape est celle de la rédaction du **rapport** final. Ce rapport doit décrire à la fois les objectifs de l'étude, la méthodologie suivie, les résultats obtenus au terme des analyses statistiques des données et leurs interprétations.

En conclusion

La démarche à suivre pour réaliser une enquête par sondage peut donc s'avérer relativement longue et complexe. Il faut garder à l'esprit que l'étape de recueil de l'information est un point de non-retour : une fois la collecte des données terminée, vous ne pourrez que vous mordre les doigts de n'avoir pas pensé à recueillir telle ou telle information.

Par ailleurs, les deux premières étapes de la démarche ne peuvent pas être traitées de façon indépendante : en effet, les objectifs déterminent le plan d'observation mais le plan d'observation peut vous amener à revoir les objectifs.

Un dernier petit mot pour conclure : toutes les étapes que nous avons passées en revue sont indispensables à la bonne conduite d'une enquête par sondage. Leur importance et leur complexité peuvent cependant varier d'une enquête à l'autre, selon la nature de la population-cible et de la problématique abordée dans l'enquête.

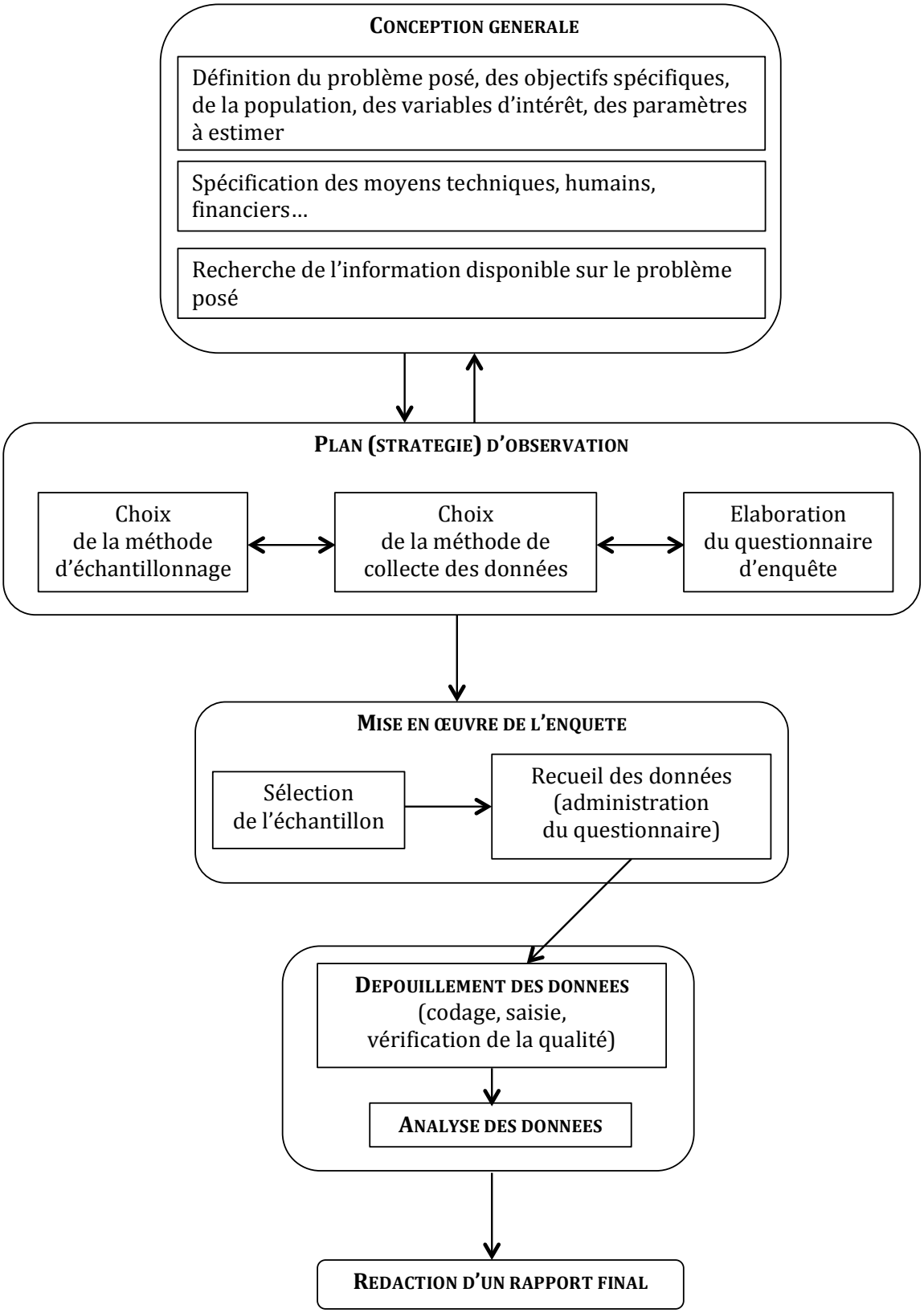


FIGURE 1.3 – *Etapes de la démarche d'enquête par sondage*

1.5 Les méthodes de sondage aléatoires

1.5.1 Catégorisation des méthodes de sondage

Revenons à présent aux méthodes de *sondage* ou d'*échantillonnage* à proprement parler, c'est-à-dire aux méthodes de prélèvement d'un échantillon dans une population. Choisir une méthode de sondage consiste en fait à définir la façon dont on va prélever des unités statistiques dans la population qui nous intéresse, afin de constituer un échantillon.

On distingue deux familles de méthodes de sondage : celle des *méthodes de sondage aléatoires*, aussi appelées méthodes de sondage *probabilistes*, et celle des *méthodes de sondage empiriques*, encore appelées méthodes de sondage *à choix raisonné*.

Cette distinction se fonde essentiellement sur la connaissance que l'on a de la manière dont le *hasard* régit la procédure de sélection de l'échantillon. Comme nous allons le voir dans la suite de cette section, nous pouvons décrire très précisément comment intervient le hasard dans le cas d'une méthode de sondage aléatoire ; une telle description n'est plus possible pour une méthode de sondage empirique.

Nous allons consacrer cette section à l'étude des spécificités des méthodes de sondage dites *aléatoires* qui, pour le statisticien, sont de loin les méthodes de sondage les plus intéressantes à appliquer. Nous ferons connaissance avec les méthodes de sondage empiriques à la prochaine section.

1.5.2 Méthodes aléatoires : la base de sondage

Une méthode de sondage aléatoire nécessite tout d'abord de disposer d'une *base de sondage*.

Celle-ci consiste en une liste exhaustive — c'est-à-dire complète, à jour et sans double compte — des unités statistiques de la population, liste dans laquelle chaque unité est représentée par un identifiant unique : son nom, un numéro de registre, ou, plus simplement, un numéro compris entre 1 et N.

Ainsi, par exemple, si l'on s'intéresse à la population des élèves d'une certaine école pour une certaine année scolaire, le fichier des inscriptions pour cette année peut constituer une bonne base de sondage.

Un autre exemple… Chaque banque possède un fichier reprenant l'ensemble des personnes qui détiennent au moins un compte chez elle ; ce fichier peut être considéré comme une base de sondage pour la population des clients de la banque.

C'est en réalité dans la base de sondage associée à la population que la méthode de sondage *aléatoire* prélèvera l'échantillon.

Attention ! Ne croyez pas que l'on dispose toujours ou presque toujours d'une base de sondage. Imaginez, par exemple, que vous deviez mener une étude sur la population des « sans-abris » présents dans une certaine agglomération au cours d'une certaine période, ou sur la population des clients des restaurants rapides d'une certaine chaîne… Vous aurez beau retourner ciel et terre, vous ne trouverez pas de listes ou de fichiers reprenant de manière exhaustive l'ensemble des unités de ces populations. Puisque vous ne disposez pas de bases de sondage complètes pour ces populations, vous n'allez pas pouvoir les sonder à l'aide d'une méthode *aléatoire*. Vous n'aurez dans ce cas d'autres solutions que de faire appel à une méthode de sondage *empirique*.

Une bonne base de sondage se caractérise par le fait qu'il existe bien une correspondance 1-1 entre elle et la population (voir la figure 1.4).

La population U correspond à l'ensemble des unités $u_1, u_2, \ldots, u_N$. De son côté, la base de sondage, généralement désignée par U_B, peut se concevoir comme l'ensemble des nombres entiers ou numéros de 1 à N. La correspondance entre la base de sondage et la population doit être telle que chaque unité de la population soit représentée par un et un seul numéro de la base de sondage et inversement, chaque numéro de la base de sondage corresponde à une et une seule unité statistique de la population.

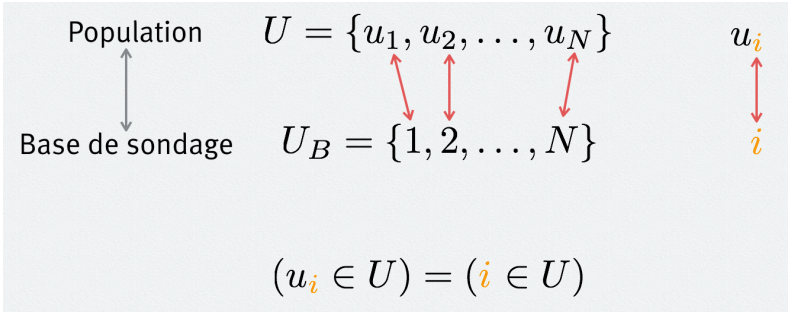


FIGURE 1.4 – *Correspondance 1-1 entre la population et la base de sondage*

Grâce à cette correspondance, on peut se simplifier un peu la vie… On peut considérer que les unités $u_1, u_2, \ldots, u_N$ de la population portent respectivement, dans la base de sondage, les numéros 1,2, … , N. Ceci nous conduit de manière naturelle à représenter chaque unité statistique u_i de la population par son indice i, c'est-à-dire par son numéro dans la base de sondage. Ainsi, plutôt que d'écrire que l'unité u_i appartient à la population U, nous écrirons désormais que l'unité i appartient à la population U, et nous confondrons en pratique la population U et sa base de sondage U_B.

Cette simplification terminologique a pour principal objectif d'alléger les formules mathématiques qui nous permettront de retraduire, sous une forme synthétique, de très nombreux résultats rencontrés dans la suite de ce cours.

Par ailleurs, puisque nous sommes occupés avec des simplifications terminologiques, nous en ferons souvent une autre : dans les explications que je vous donnerai, je considérerai généralement, par facilité, que U est une population d'*individus*. Je parlerai donc très souvent des individus de la population plutôt que des unités statistiques de celle-ci.

1.5.3 Méthodes aléatoires : le plan de sondage

Une méthode de sondage *aléatoire* ne peut donc être appliquée que si l'on dispose d'une base de sondage pour la population ciblée. Par ailleurs, une méthode de sondage *aléatoire* va être définie par l'intermédiaire de ce que nous allons appeler un *plan de sondage*. Qu'est-ce que cela ?

Prenons un petit exemple tout simple pour illustrer cette nouvelle notion.

a) Exemple illustratif

Imaginons que nous ayons une population U — ou une base de sondage — constituée de 5 individus numérotés de 1 à 5 : $U = \{1, 2, 3, 4, 5\}$.

On décide d'y prélever un échantillon en tirant successivement, « au hasard » et sans remise, deux individus dans cette population. En d'autres termes, on place dans une urne 5 papiers ou 5 boules numéroté(e)s de 1 à 5, et on décide de procéder comme suit : on tire à l'aveugle une première boule dans l'urne, puis on tire, de nouveau à l'aveugle, une deuxième boule parmi les 4 boules restant dans l'urne.

Nous avons donc fait le choix d'une procédure de prélèvement de l'échantillon dans laquelle le hasard intervient. Cette procédure peut en réalité être vue comme une sorte de « jeu de hasard ». Mais nous sommes ici dans une situation où nous pouvons aisément décrire, de manière précise et rigoureuse, les différents résultats possibles de notre petit jeu et la probabilité d'obtenir chacun de ces résultats.

En effet, puisque nous disposons de la liste des individus qui composent la population U, nous pouvons aisément énumérer l'ensemble de tous les échantillons qu'il est possible de sélectionner avec la méthode de sondage que nous avons choisie.

Le hasard peut nous faire sélectionner les individus 1 et 2, ou encore les individus 1 et 3, ou les individus 1 et 4, etc. On a vite fait de vérifier que l'on a, en tout et pour tout, 10 échantillons possibles. Pour nous faciliter la vie pour la suite, appelons $\mathbb{S}$ (« grand S ») l'ensemble de ces 10 échantillons possibles :

$$\mathbb{S} = \big\{\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{3,4\},\{3,5\},\{4,5\}\big\}.$$

Non seulement nous pouvons dresser la liste de tous les échantillons qu'il est *a priori* possible d'obtenir, mais nous pouvons aussi déterminer, pour chacun de ces échantillons, la probabilité qu'il soit effectivement sélectionné.

Considérons par exemple l'échantillon contenant les individus 1 et 2. La probabilité d'obtenir cet échantillon particulier est donnée par la probabilité que le hasard nous fasse d'abord tirer le numéro 1 et ensuite le numéro 2, à laquelle se rajoute la probabilité que le hasard nous fasse d'abord tirer le numéro 2 et ensuite le numéro 1 :

$$\begin{aligned} p(\{1,2\}) &= \mathrm{P}[1 \text{ puis } 2] + \mathrm{P}[2 \text{ puis } 1] \\ &= \frac{1}{5}\times\frac{1}{4} + \frac{1}{5}\times\frac{1}{4} = \frac{1}{20} + \frac{1}{20} = \frac{2}{20} = \frac{1}{10}. \end{aligned}$$

Nous obtenons donc une probabilité de sélectionner cet échantillon particulier égale à un dixième. Nous dirons que l'échantillon contenant les individus numéros 1 et 2 a une *probabilité de sélection* égale à un dixième.

Le même raisonnement peut être tenu pour les probabilités de sélection des 9 autres échantillons. Nous avons donc 10 échantillons s possibles ($s \in \mathbb{S}$), ayant tous la même probabilité d'être prélevés : ils ont tous une même probabilité de sélection égale à un dixième :

$$p(s) = \frac{1}{10} \quad \text{pour tout } s \in \mathbb{S}.$$

En dressant la liste de tous les échantillons qu'il est *a priori* possible d'obtenir avec la méthode de prélèvement choisie et en associant à chaque échantillon possible sa probabilité de sélection, nous avons en fait décrit de quelle manière le hasard conduisait la procédure d'échantillonnage. Nous avons spécifié ce que l'on appelle le *plan de sondage* associé à cette procédure.

Le fait d'avoir pu associer un plan de sondage à la méthode de prélèvement choisie fait de celle-ci une méthode *aléatoire* à proprement parler.

b) Formalisation

Une **méthode de sondage** est une procédure spécifique de prélèvement d'un échantillon au sein de la population U. Pour qu'elle puisse être qualifiée d'**aléatoire**, il ne suffit pas que cette méthode consiste à sélectionner « au hasard » un certain nombre d'individus dans la population. Non ! Il faut impérativement qu'on puisse lui associer un **plan de sondage**, décrivant précisément de quelle manière le hasard régit le prélèvement de l'échantillon.

Le **plan de sondage** associé à une méthode de sondage aléatoire spécifie deux choses :

1. l'**ensemble** $\mathbb{S}$ de tous les échantillons s qu'il est *a priori* possible de sélectionner lorsqu'on met en œuvre la méthode de prélèvement choisie ;
2. la probabilité qu'a chaque échantillon possible s d'être sélectionné. Cette probabilité, que nous désignerons par $p(s)$, est appelée la **probabilité de sélection** de l'échantillon s.

Deux remarques s'imposent ici. Premièrement, la spécification de l'ensemble $\mathbb{S}$ de tous les échantillons possibles ne peut se faire que si l'on dispose d'une *base de sondage* pour la population considérée. En effet, sans base de sondage reprenant l'ensemble des unités de la population, il est impossible de dresser la liste précise de tous les échantillons possibles. Ceci explique pourquoi il est indispensable de disposer d'une base de sondage si l'on veut réaliser un sondage *aléatoire* au sens strict du terme.

Deuxièmement, les probabilités de sélection des différents échantillons possibles se caractérisent par deux propriétés fondamentales :

- si s est l'un des échantillons possibles, la probabilité de sélection qui lui est associée est strictement positive :

  $$p(s) > 0 \qquad \text{pour tout } s \in \mathbb{S}.$$

  Ceci traduit simplement le fait qu'il y a « une certaine chance » de se retrouver avec l'échantillon s au terme de la procédure de prélèvement.

- la somme des probabilités de sélection associées à tous les échantillons possibles est égale à 1 :

$$\sum_{s\in\mathbb{S}} p(s) = 1.$$

  Cette propriété traduit le fait que la méthode de sondage nous amènera nécessairement à sélectionner l'un des échantillons possibles, c'est-à-dire l'un des échantillons repris dans l'ensemble $\mathbb{S}$. Dans l'exemple considéré ci-avant, nous avions 10 échantillons possibles, chacun ayant une probabilité de sélection égale à un dixième ; la somme des probabilités de sélection des 10 échantillons possibles est donc bien égale à 1.

En d'autres termes encore, si l'on considère l'échantillon comme un objet (une « variable ») *aléatoire* — nous le désignerons alors à l'aide de la lettre majuscule S —, on peut définir le plan de sondage comme la **distribution de probabilités de S** : l'ensemble $\mathbb{S}$ est simplement l'ensemble de toutes les « valeurs » (ou réalisations) possibles s de S, et les probabilités de sélection $p(s)$ sont les probabilités associées aux différentes réalisations possibles de S.

En pratique, dans le cadre d'un sondage aléatoire, la procédure de prélèvement sera le plus souvent dirigée par un *algorithme informatique* appliqué à la base de sondage et conçu pour respecter le plan de sondage que l'on s'est fixé.

1.5.4 Méthodes aléatoires : les probabilités d'inclusion

a) Exemple illustratif

Nous l'avons vu, le plan de sondage spécifie de quelle manière le hasard régit la procédure d'échantillonnage. Mais nous pouvons aller encore un peu plus loin dans la description de la façon dont travaille le hasard dans notre méthode de sondage ! La connaissance du plan de sondage va en effet nous permettre de déterminer, pour chaque individu de la population, quelle est sa probabilité de faire partie de l'échantillon qui sera sélectionné ; cette probabilité porte le nom de *probabilité d'inclusion*.

Rappelez-vous ce que nous avons vu jusqu'ici !

La population U est constituée de 5 personnes : $U = \{1, 2, 3, 4, 5\}$. Si l'on décide d'y sélectionner un échantillon en y tirant successivement et « au hasard » (à l'aveugle) 2 individus, nous pouvons nous retrouver avec l'un des 10 échantillons possibles repris dans l'ensemble $\mathbb{S}$:

$$\mathbb{S} = \big\{\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{3,4\},\{3,5\},\{4,5\}\big\}.$$

Chacun de ces 10 échantillons possibles a la même probabilité d'être sélectionné, égale à un dixième :

$$p(s) = \frac{1}{10} \quad \text{pour tout } s \in \mathbb{S}.$$

Dans cette situation, que vaut la *probabilité d'inclusion* de l'individu n° 1 ? Autrement dit, quelle est la probabilité p_1 que l'individu 1 fasse partie de l'échantillon sélectionné ?

Cela revient à se demander quelle est la probabilité de prélever un échantillon contenant l'individu 1.

La probabilité recherchée est donc égale à :

$$p_1 = \mathrm{P}(1 \in S) = p(\{1,2\}) + p(\{1,3\}) + p(\{1,4\}) + p(\{1,5\})$$
$$= \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{4}{10} = \frac{2}{5}.$$

Puisque les autres individus de la population sont également contenus dans exactement 4 des 10 échantillons possibles, on obtient cette même probabilité d'inclusion de 2 cinquièmes pour chacun des 5 individus de la population :

$$p_i = \mathrm{P}(i \in S) = \frac{2}{5} \quad \text{pour tout } i \in U.$$

Ainsi, vous savez « à l'avance » qu'avec la méthode de sondage choisie, chaque individu de la population a 2 chances sur 5 de se retrouver dans l'échantillon qui va être prélevé. C'est le *hasard* qui va déterminer si l'individu numéro 1, par exemple, va être sélectionné, mais vous savez maintenant avec quelle probabilité cela risque de se produire. Vous savez exactement comment le hasard « gère » les tirages d'individus dans la population.

b) Formalisation

La détermination du plan de sondage permet d'affecter à chaque individu de la population U une probabilité *non nulle* et *connue* de faire partie de l'échantillon qui sera sélectionné ; cette probabilité porte le nom de **probabilité d'inclusion**.

Comment fait-on pour déterminer ces probabilités à partir du plan de sondage ? Désignons par p_i la probabilité d'inclusion de l'individu n° i, c'est-à-dire la probabilité pour que cet individu i appartienne à l'échantillon aléatoire S qui sera sélectionné. Cette probabilité n'est autre que la probabilité de prélever un échantillon qui contiendra l'individu i. Elle peut dès lors être obtenue en sommant les probabilités de sélection de tous les échantillons possibles s qui contiennent l'individu i :

$$p_i = \mathrm{P}(i \in S) = \sum_{s \in \mathbb{S} \text{ tel que } i \in s} p(s).$$

Notez encore que, dans de nombreux ouvrages consacrés à la théorie des sondages, la probabilité d'inclusion de l'individu i est notée π_i plutôt que p_i. Je ne me suis pas ralliée à cette pratique, puisque j'ai pris le parti de réserver la lettre grecque π pour désigner une proportion dans la population.

Enfin, lorsque tous les individus de la population ont la *même* probabilité d'inclusion, on parle de sondage aléatoire **à probabilités égales** (on dira qu'on a à faire à un sondage **PE**). Nous venons de rencontrer un tel sondage dans notre petit exemple. Dans le cas où les probabilités d'inclusion diffèrent d'un individu à l'autre, on parle de sondage aléatoire **à probabilités inégales** ou encore de sondage **PI**.

1.5.5 Exercices

Objectif – Les exercices proposés ci-dessous visent à vous familiariser avec les notions de plan de sondage et de probabilités d'inclusion caractérisant une méthode de sondage *aléatoire*.

Des correctifs détaillés de ces exercices sont disponibles sur l'UV.

a) Exercice 1.1

Considérons une population U constituée de 6 individus (numérotés de 1 à 6) et intéressons-nous à la méthode de sondage aléatoire consistant à effectuer 3 prélèvements successifs dans U, « au hasard » et sans remise. En d'autres termes, on tire l'échantillon en prélevant « au hasard » (à l'aveugle) un premier individu parmi les 6 individus de la population, puis en sélectionnant « au hasard » un deuxième individu parmi les 5 individus restants dans la population, et enfin en prélevant « au hasard » un troisième individu parmi les 4 personnes encore dans la population.

N.B. : La méthode de sondage considérée ici est appelée *méthode de sondage aléatoire simple sans remise*. Elle sera étudiée en détail dans le deuxième chapitre de ce cours.

Quelles sont les caractéristiques du plan de sondage associé à cette méthode d'échantillonnage et des probabilités d'inclusion qui en découlent ?

1. Quel est le nombre d'échantillons qu'il est possible d'obtenir lorsqu'on applique cette méthode d'échantillonnage ?
Indice – Pour répondre à cette question sans devoir faire appel à une formule de dénombrement particulière, il vous suffit de dresser la liste de tous les échantillons qu'il est *a priori* possible d'obtenir avec la méthode d'échantillonnage considérée ici.

- ○ 120
- ○ 20
- ○ 216
- ○ 40
- ○ 36

2. Que vaut la probabilité de sélection de l'échantillon contenant les individus 2, 4 et 5 ?

- ○ 1/120
- ○ 3/6
- ○ 1/40
- ○ 1/216
- ○ 1/20

3. Parmi les propositions suivantes, cochez celle(s) qui est (sont) correcte(s).

- ○ Certains échantillons possibles ont davantage de chance d'être sélectionnés que d'autres.
- ○ Tous les échantillons possibles ont la même probabilité de sélection.
- ○ La somme des probabilités de sélection des différents échantillons possibles est égale au nombre d'individus sélectionnés, c'est-à-dire 3.
- ○ La somme des probabilités de sélection des différents échantillons possibles vaut 1.
- ○ La somme des probabilités de sélection des différents échantillons possibles coïncide avec la taille de la population, c'est-à-dire 6.

4. Que vaut la probabilité d'inclusion de l'individu n° 3 ?

- ○ 1/2
- ○ 1/216
- ○ 1/20
- ○ 1/6
- ○ 1/3

5. Les individus 2 et 4 possèdent-ils la même probabilité d'inclusion ?

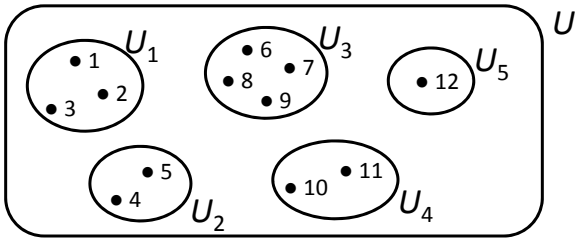
- ○ Oui
- ○ Non

6. Parmi les propositions suivantes, cochez celle(s) qui est (sont) correcte(s).

- ○ Certains individus de la population ont davantage de chance que d'autres de se retrouver dans l'échantillon qui sera tiré.
- ○ Tous les individus de la population possèdent la même chance de se retrouver dans l'échantillon qui sera tiré.
- ○ La somme des probabilités d'inclusion des 6 individus de la population est égale à 6, la taille de la population.
- ○ La somme des probabilités d'inclusion des 6 individus de la population est égale à 3, le nombre d'individus prélevés lors de la procédure d'échantillonnage.
- ○ La somme des probabilités d'inclusion des 6 individus de la population est égale à 1.

b) Exercice 1.2

Considérons la population U représentée ci-dessous. Elle est constituée de 12 individus (numérotés de 1 à 12) qui se répartissent entre 5 ménages : le ménage U_1 est constitué des individus 1, 2 et 3 ; le ménage U_2 est formé des individus 4 et 5 ; le ménage U_3 rassemble les individus 6, 7, 8 et 9 ; les individus 10 et 11 forment le ménage U_4, tandis que l'individu 12 est le seul membre du ménage U_5.



On décide de sélectionner un échantillon *d'individus* en procédant en deux étapes : 1) on prélève « au hasard » et sans remise (c'est-à-dire l'un à la suite de l'autre) 2 ménages parmi les 5 ménages de la population ; 2) on sélectionne ensuite tous les individus des 2 ménages prélevés. Ainsi, par exemple, si le hasard nous fait prélever les ménages U_1 et U_4, notre échantillon final sera constitué des individus 1, 2, 3, 10 et 11.

Cette méthode de sondage particulière porte le nom de *méthode de sondage en grappes*. Nous l'étudierons en détail dans le chapitre 5. Il s'agit d'une méthode de sondage *aléatoire*. Les questions qui suivent portent sur les caractéristiques du plan de sondage qui lui est associé et des probabilités d'inclusion qu'elle affecte aux individus de la population.

1. Quel est le nombre d'échantillons d'individus qu'il est possible d'obtenir lorsqu'on applique cette procédure d'échantillonnage ?

- ○ 66
- ○ 144
- ○ 10
- ○ 25
- ○ 20

2. Que vaut la probabilité de sélection de l'échantillon contenant les individus 1, 2, 3, 10 et 11 ?

- ○ 1/10
- ○ 1/20
- ○ 5/12
- ○ 2/10
- ○ 2/5

3. Que vaut la probabilité de sélection de l'échantillon contenant les individus 1, 2, 3, 4 et 5 ?

- ○ 1/10
- ○ 1/20
- ○ 5/12

- ○ 2/10
- ○ 2/5

4. Parmi les propositions suivantes, cochez celle(s) qui est (sont) correcte(s).

- ○ Tous les échantillons ont la même taille.
- ○ Tous les échantillons possibles ont la même probabilité de sélection et la somme des probabilités de sélection de tous les échantillons possibles vaut 2, le nombre de ménages prélevés dans la population.
- ○ Tous les échantillons possibles ont la même probabilité de sélection et la somme des probabilités de sélection de tous les échantillons possibles vaut 1.
- ○ La somme des probabilités de sélection des différents échantillons possibles coïncide avec la taille de la population, c'est-à-dire 12.
- ○ La probabilité de sélection d'un échantillon est liée au nombre d'individus qu'il contient. Ainsi, par exemple, l'échantillon contenant les individus 1, 2, 3, 6, 7, 8 et 9 a davantage de chance d'être tiré que l'échantillon ne regroupant que les individus 1, 2, 3 et 12.
- ○ Certains échantillons ont davantage de chance d'être sélectionnés que d'autres, mais la probabilité de sélection d'un échantillon n'est pas liée à sa taille, c'est-à-dire au nombre d'individus qu'il contient.

5. Que vaut la probabilité d'inclusion de l'individu 1 ?

- ○ 1/10
- ○ 1/15
- ○ 1/12
- ○ 1/5
- ○ 2/5

6. Parmi les propositions suivantes, cochez celle(s) qui est (sont) correcte(s).

- ○ Les individus d'un même ménage ont la même probabilité d'inclusion, mais les individus du ménage U_3 ont davantage de chance de se retrouver dans l'échantillon final que les autres individus de la population car ils appartiennent au plus grand ménage.
- ○ Tous les individus de la population ont la même probabilité d'inclusion.
- ○ La probabilité d'inclusion d'un individu varie selon la taille du ménage auquel il appartient.
- ○ Tous les individus de la population ont des probabilités d'inclusion différentes.

c) Exercice 1.3

Considérons une population U constituée de 5 individus numérotés de 1 à 5. La méthode de sondage aléatoire choisie pour sonder cette population est caractérisée par le plan de sondage suivant :

s	$p(s)$
{1,2}	0,08
{1,3}	0,06
{1,4}	0,10
{1,5}	0,12
{2,3}	0,12
{2,4}	0,11
{2,5}	0,10
{3,4}	0,10
{3,5}	0,12
{4,5}	0,09
Total	1

Aide à la lecture de ce tableau : ce tableau reprend l'ensemble des échantillons s qu'il est possible de prélever avec la méthode de sondage considérée ici et indique la probabilité de sélection $p(s)$ de chaque échantillon possible. Ainsi, par exemple, on a une probabilité de 0,08 de tirer l'échantillon contenant les individus 1 et 2 ; on a une probabilité de 0,12 de tirer l'échantillon contenant les individus 3 et 5 ; etc.

1. Que vaut la probabilité d'inclusion de l'individu 1 ?

- ○ 0,40
- ○ 0,72
- ○ 0,09
- ○ 0,36
- ○ 1/5

2. Que vaut la probabilité d'inclusion de l'individu 5 ?

- ○ 0,40
- ○ 0,43
- ○ 1/5
- ○ 0,86
- ○ 0,11

1.6 Les méthodes de sondage empiriques

Nous l'avons vu dans la section précédente, les méthodes de sondage aléatoires nécessitent de pouvoir disposer d'une base de sondage. Mais il n'existe pas toujours de base de sondage pour la population que l'on veut étudier ; par ailleurs, même si une telle base de sondage existe, on peut être confronté à des problèmes liés au droit de possession et aux coûts d'accès de ces listes. Les organismes de sondage ont donc été amenés à introduire des méthodes de sondage *non aléatoires*, encore appelées *méthodes de sondage empiriques* ou *à choix raisonné*.

La méthode empirique la plus utilisée est sans nul doute la *méthode des quotas*. Elle repose sur la connaissance de la répartition de la population selon différentes catégories importantes : par exemple, la répartition hommes-femmes de la population, la répartition de la population selon différentes catégories d'âges, ou selon différentes catégories socio-professionnelles... Le principe de la méthode des quotas est alors de prélever un échantillon en choisissant les individus de telle sorte que l'on retrouve dans l'échantillon la même répartition selon les différentes catégories retenues. Si, par exemple, la population comprend 55% d'hommes, 10% d'ouvriers, 27% de personnes âgées de plus de 23 ans, etc., on choisira un échantillon qui respectera ces proportions : il devra lui aussi comprendre 55% d'hommes, 10% d'ouvriers, etc. D'une certaine façon, l'échantillon est construit comme un « modèle réduit » de la population. On retrouve ici un objectif de « représentativité » de l'échantillon déjà évoqué précédemment dans ce chapitre (cf. Section 1.3).

Mais il existe aussi d'autres méthodes de sondage empiriques ! La *méthode des itinéraires*, la *méthode des unités-types*, l'*échantillonnage « sur place »*, etc. Nous en reparlerons plus longuement au chapitre 8.

Ces méthodes empiriques sont particulièrement appréciées des sociétés chargées de réaliser des sondages dans la mesure où le choix des individus n'est pas régi par une procédure nécessitant une base de sondage et un algorithme de tirage. Cette plus grande liberté permet souvent une exécution plus rapide du sondage et une réduction des coûts, comparativement aux méthodes aléatoires.

Il faut toutefois garder à l'esprit que les méthodes de sondage non aléatoires présentent des inconvénients qui sont loin d'être mineurs ! Dans un sondage empirique recourant à l'emploi d'enquêteurs, le fait de laisser à ceux-ci une certaine liberté dans le choix des personnes à interroger peut avoir pour conséquence d'obtenir des estimations des paramètres de la population entachées de biais importants. Les choix des enquêteurs peuvent en effet être influencés par leurs heures de travail, leurs goûts personnels, etc.

En outre, l'application d'une méthode de sondage empirique rend impossible la mesure objective de la précision des résultats obtenus. En effet, comme nous allons le voir dès le prochain chapitre, on ne peut quantifier de manière rigoureuse la précision de la procédure d'estimation des paramètres-population que si l'on dispose d'un plan de sondage pour la méthode d'échantillonnage choisie et que l'on connaît les probabilités

d'inclusion des individus de la population. Ceci explique pourquoi les statisticiens préfèrent les méthodes de sondage aléatoires aux méthodes empiriques.

Il faut donc être prudent dans l'usage des méthodes de sondage empiriques et dans l'interprétation de leurs résultats.

1.7 L'information auxiliaire

Avant de conclure ce chapitre, il nous reste à aborder une dernière notion qui peut jouer un rôle fort important en théorie des sondages : celle d'**information auxiliaire**.

On dispose souvent d'informations diverses sur les unités statistiques de la population, informations qui sont soit contenues dans la base de sondage elle-même, soit fournies par des fichiers administratifs, ou encore des études ou des recensements antérieurs.

Dans certains cas, la base de sondage nous indique, pour chaque unité statistique, la valeur de l'une ou l'autre variable, dite « auxiliaire ». Ainsi, par exemple, s'il s'agit d'une base de sondage d'individus, on y trouve très souvent indiqué le sexe et la date de naissance — donc l'âge — de chaque personne. Mais d'autres variables peuvent également être reprises dans la base de sondage, selon la nature de cette dernière. S'il s'agit d'une base de sondage d'entreprises, on y trouve souvent le secteur d'activité et la taille de chaque entreprise de la population. On dispose dans ce cas d'une information auxiliaire relativement riche qui va pouvoir être exploitée pour concevoir un plan de sondage bien adapté, c'est-à-dire un plan de sondage susceptible de nous permettre d'estimer avec une bonne précision les paramètres-population qui nous intéressent.

Dans d'autres cas, on ne dispose que d'une information auxiliaire *agrégée* : on connaît la moyenne ou le total de telle variable dans la population, on connaît la répartition dans la population de l'un ou l'autre caractère. Cette information auxiliaire agrégée va pouvoir être exploitée non pas au niveau de la conception du plan de sondage, mais plutôt au niveau de l'étape de l'estimation des paramètres-population. Ceci nous conduit aux *méthodes de calage et de redressement* que nous découvrirons au chapitre 7 de ce cours. Notez que ces méthodes peuvent également intervenir dans le traitement du problème de la non-réponse ; nous évoquerons ce sujet au chapitre 9.

1.8 Conclusion

Ce premier chapitre vous a permis de faire vos premiers pas en théorie des sondages.

Vous savez désormais ce qui se cache derrière le terme de *sondage statistique*.

Vous avez fait connaissance avec les acteurs-clés qui interviennent dans une procédure de sondage : la population, la variable d'intérêt, les paramètres-population, l'échantillon, les estimateurs.

Vous avez pris conscience du fait qu'une enquête par sondage était une démarche de recherche d'information constituée de plusieurs étapes.

Vous avez découvert qu'il existait deux familles de méthodes de sondage : les méthodes empiriques et les méthodes aléatoires. Vous savez également que ces dernières, auxquelles nous allons préférentiellement nous intéresser dans le cadre de ce cours, nécessitent de disposer d'une base de sondage, et peuvent être parfaitement caractérisées, pour ce qui est de leur caractère aléatoire, par un plan de sondage et par les probabilités d'inclusion attribuées aux individus de la population avant la mise en œuvre de l'échantillonnage.

Vous voilà fin prêts pour aborder l'étude de la méthode de sondage aléatoire *de base* en théorie des sondages : celle que l'on qualifie de « *simple* ». Ce sera l'objet du chapitre suivant.

Chapitre 2
Le sondage aléatoire simple

2.1 Introduction

Dans le premier chapitre, nous avons vu que l'on pouvait distinguer deux familles de méthodes de sondage : les méthodes *aléatoires* et les méthodes *empiriques*.

Une méthode de sondage aléatoire impose de disposer d'une base de sondage pour la population à sonder ; par ailleurs, elle peut être caractérisée par le plan de sondage qui lui correspond, ainsi que par les probabilités d'inclusion qu'elle affecte *a priori* aux individus de la population.

Du point de vue du statisticien, ce sont sans nul doute les méthodes aléatoires qui sont les plus intéressantes à étudier ; elles permettent en effet, contrairement aux méthodes empiriques, d'analyser de manière objective la qualité — et notamment la précision — de la procédure d'estimation qui sera mise en œuvre au terme de la procédure d'échantillonnage.

Je vous propose de revoir tout cela en détail pour la méthode de sondage aléatoire la plus simple que l'on puisse concevoir : celle qui consiste à tirer un échantillon en effectuant un nombre fixé de prélèvements successifs et « au hasard » (à l'aveugle) dans la population.

Cette méthode de sondage apparaît comme fondamentale pour deux raisons :

- Premièrement, elle sert en quelque sorte d'étalon en théorie des sondages ; c'est par rapport à ses propriétés qu'on a l'habitude de juger les propriétés d'autres méthodes de sondage aléatoires, plus complexes.
- Deuxièmement, elle constitue la « brique » élémentaire de plusieurs autres méthodes de sondage aléatoires fréquemment utilisées, telles que, par exemple, les méthodes de sondage stratifié que nous étudierons dans le chapitre 3. Il est donc important d'en connaître parfaitement toutes les caractéristiques.

Nous avons déjà rencontré cette méthode de sondage très simple dans l'un des exemples présentés dans le chapitre précédent. Nous allons ici l'étudier de manière plus approfondie.

Dans un premier temps, nous allons rechercher les caractéristiques générales du *plan de sondage* et des *probabilités d'inclusion* qui lui sont associées : cela nous permettra de décrire précisément de quelle manière le hasard régit la procédure d'échantillonnage.

Nous nous pencherons ensuite sur la problématique de l'estimation d'une *proportion*, d'une *moyenne* ou d'un *total* de la population à partir d'un échantillon prélevé selon cette méthode. Quelle fonction des observations réalisées dans l'échantillon doit-on utiliser pour estimer le paramètre qui nous intéresse ? Dans quelle situation peut-on s'attendre à obtenir une « bonne » estimation de ce paramètre, autrement dit une estimation proche de sa valeur exacte ? Quels sont les facteurs, propres à la population ou à l'échantillon, qui influencent la précision de l'estimation ? Voici quelques-unes des questions auxquelles nous allons tenter de répondre.

2.2 La procédure d'échantillonnage

Dans le cadre du sondage aléatoire simple, le tirage de l'échantillon s'effectue en prélevant successivement, « au hasard » (c'est-à-dire à l'aveugle), n unités statistiques ou individus parmi les N unités de la population.

Les n prélèvements successifs peuvent se faire soit « sans remise », soit « avec remise ». Voyons cela de plus près !

2.2.1 Sans remise (SR)

Supposons que nous ayons une population constituée de 10 individus et que l'on doive y prélever successivement, « au hasard » et sans remise, 2 individus. Vu la taille très réduite de la population, nous pouvons très facilement organiser la procédure d'échantillonnage comme suit.

On place dans une urne 10 boules ou 10 papiers numérotés de 1 à 10. On tire alors à l'aveugle un premier papier dans l'urne, puis on tire, de nouveau à l'aveugle, un second papier parmi les 9 papiers restant dans l'urne. On effectue ainsi deux prélèvements successifs et « sans remise » dans l'urne, le terme « SANS remise » spécifiant bien qu'une fois qu'un papier est tiré, il n'est pas remis dans l'urne avant que l'on effectue le tirage suivant.

FIGURE 2.1 – *Echantillonnage aléatoire SANS remise*

2.2.2 Avec remise (AR)

Cette procédure de prélèvement « sans remise » que nous venons de décrire est à distinguer de celle que l'on qualifierait de « AVEC remise » et qui correspondrait au cas où, dès qu'on a prélevé un papier dans l'urne et qu'on a lu le numéro qu'il portait, on remet ce papier dans l'urne avant d'effectuer, à l'aveugle, le prélèvement suivant.

Cette méthode de tirage aléatoire « avec remise » est certainement moins naturelle que celle « sans remise ». Par ailleurs, en prélevant l'échantillon « avec remise », on encourt

le risque de tirer à plusieurs reprises le même numéro, autrement dit de sélectionner à plusieurs reprises le même individu. Ce risque est bien sûr négligeable si la population est de grande taille, mais nous montrerons qu'il altère quelque peu l'efficacité de la méthode de sondage si la population est de petite taille.

Dans la section 2.3 de ce chapitre, nous étudierons en détail la procédure de sondage SANS remise (SR) ; nous procèderons à une analyse détaillée de la procédure de sondage AVEC remise (AR) dans la section 2.4.

Notez encore que la solution du tirage de boules ou de papiers dans une urne n'est bien sûr plus praticable lorsque la population est de grande taille. On peut heureusement faire appel dans ce cas à l'un ou l'autre logiciel statistique, tel que le logiciel libre R, par exemple, qui permet de « mimer » cette procédure de prélèvement. Nous verrons également dans la dernière section de ce chapitre (section 2.9) quelques algorithmes excessivement simples permettant de prélever adéquatement l'échantillon et requérant tout au plus l'utilisation d'une bonne machine à calculer ou d'un tableur.

2.3 Le sondage aléatoire simple *sans remise* (PESR)

2.3.1 Le plan de sondage

Dans la section précédente, nous avons vu comment procéder pour le tirage de l'échantillon dans le cadre du sondage aléatoire simple sans remise. Rappelez-vous : on prélève successivement, « au hasard » (à l'aveugle) et sans remise, n unités statistiques ou individus parmi les N unités de la population.

Cette méthode de sondage est « aléatoire » au sens où nous l'avons défini au chapitre précédent : nous pouvons en effet décrire très précisément de quelle manière le hasard régit la procédure d'échantillonnage, en déterminant à la fois le plan de sondage et les probabilités d'inclusion qui lui sont associées.

Commençons donc par le plan de sondage. Nous allons tout d'abord le déterminer dans le cadre d'un exemple illustratif ; nous formaliserons ensuite les résultats.

Exemple

Supposons que notre population soit constituée de 10 personnes et que l'on doive y prélever successivement, « au hasard » et sans remise, 2 individus :

$$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \text{ ; } N = 10 \text{ et } n = 2.$$

Pour déterminer le plan de sondage associé à cette situation, nous pouvons suivre exactement la même démarche que celle que nous avons déjà suivie dans l'exemple de plan de sondage présenté dans le chapitre précédent.

Décrivons tout d'abord l'ensemble $\mathbb{S}$ de tous les échantillons s qu'il est *a priori* possible de sélectionner :

$$\begin{aligned}\mathbb{S} = \big\{&\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{1,7\},\{1,8\},\{1,9\},\{1,10\},\\ &\{2,3\},\{2,4\},\{2,5\},\{2,6\},\{2,7\},\{2,8\},\{2,9\},\{2,10\},\{3,4\},\\ &\{3,5\},\{3,6\},\{3,7\},\{3,8\},\{3,9\},\{3,10\},\{4,5\},\{4,6\},\{4,7\},\\ &\{4,8\},\{4,9\},\{4,10\},\{5,6\},\{5,7\},\{5,8\},\{5,9\},\{5,10\},\{6,7\},\\ &\{6,8\},\{6,9\},\{6,10\},\{7,8\},\{7,9\},\{7,10\},\{8,9\},\{8,10\},\{9,10\}\big\}.\end{aligned}$$

Il y a ainsi 45 échantillons possibles différents. On vérifie aisément que 45 n'est autre que le nombre de combinaisons de 2 éléments parmi 10 :

$$\binom{10}{2} = C_{10}^2 = \frac{10!}{2!\,(10-2)!} = \frac{10\times 9}{2} = 45.$$

Que valent les probabilités de sélection de ces 45 échantillons possibles ?

La probabilité de sélectionner l'échantillon contenant les individus 1 et 2 est égale à :

$$\begin{aligned}p(\{1,2\}) &= \mathrm{P}(\text{prélever 1 puis 2}) + \mathrm{P}(\text{prélever 2 puis 1})\\ &= \frac{1}{10}\times\frac{1}{9} + \frac{1}{10}\times\frac{1}{9} = \frac{2}{90} = \frac{1}{45}.\end{aligned}$$

Vous pouvez facilement vérifier que chaque échantillon possible a cette même probabilité de $1/45$ d'être sélectionné :

$$p(s) = \frac{1}{45} \quad \text{pour tout } s \in \mathbb{S}.$$

Notez également que cette probabilité est égale à 1 divisé par le nombre total d'échantillons possibles.

Enfin, chaque échantillon possible contient 2 des 10 individus de la population. On dira que l'on a un *taux de sondage* de $2/10$, soit 20%.

Quelles sont donc les caractéristiques du **plan de sondage** associé à la procédure d'échantillonnage consistant à prélever successivement, « au hasard » et sans remise, n individus dans une population U de taille N ?

a) L'ensemble des échantillons possibles

L'ensemble $\mathbb{S}$ des échantillons qu'il est possible d'obtenir lorsqu'on applique cette procédure d'échantillonnage est en fait l'ensemble de tous les *sous-ensembles* s de taille n de la population U. En d'autres termes, chaque échantillon possible s contient n individus *distincts* de la population U.

b) Le taux de sondage

Le taux de sondage — habituellement désigné par f — associé à une méthode d'échantillonnage est égal au rapport entre la taille de l'échantillon prélevé et la taille de la population. En particulier, lorsque le sondeur choisit d'effectuer n prélèvements successifs sans remise dans la population, il se fixe *a priori* un taux de sondage

$$f = \frac{n}{N}.$$

c) Le nombre d'échantillons possibles

Le nombre d'échantillons possibles — nous désignerons ce nombre par M — est donné par ce que l'on appelle, en calcul des probabilités, le nombre de **combinaisons** de n éléments parmi N. Ce nombre de combinaisons, noté $\binom{N}{n}$ ou encore C_N^n, est égal à

$$\binom{N}{n} = C_N^n = \frac{N!}{n!\,(N-n)!}$$

où $r!$ désigne la *factorielle* du nombre entier r ($r! = r\times(r-1)\times(r-2)\times \ldots \times 2\times 1$).

d) Les probabilités de sélection

Comme nous l'avons observé dans notre exemple, tous les échantillons possibles ont la même probabilité d'être sélectionnés, égale à $1/M$, c'est-à-dire l'inverse du nombre total d'échantillons possibles :

$$p(s) = \frac{1}{M} \quad \text{pour tout } s \in \mathbb{S}.$$

C'est le fait que tous les échantillons possibles aient la même probabilité de sélection qui motive l'utilisation de l'appellation « méthode de sondage aléatoire **simple** » pour la méthode considérée ici.

2.3.2 Les probabilités d'inclusion

Intéressons-nous à présent aux probabilités d'inclusion affectées aux individus de la population par la méthode de sondage aléatoire simple sans remise.

Revenons une nouvelle fois à notre exemple.

Exemple (suite)

La probabilité d'inclusion de l'individu numéro 1 est la probabilité que cet individu appartienne à l'échantillon aléatoire S qui sera prélevé. En d'autres termes, il s'agit de la probabilité de prélever un échantillon qui contient l'individu numéro 1 : cette probabilité est égale à la somme des probabilités de sélection des échantillons possibles s qui contiennent l'individu 1. Nous avons donc :

$$p_1 = \mathrm{P}(1 \in S) = \sum_{s \in \mathbb{S} \text{ tel que } 1 \in s} p(s).$$

Or, 9 échantillons sur les 45 échantillons possibles contiennent l'individu numéro 1 et — nous l'avons vu précédemment — chacun de ces 9 échantillons a une chance sur 45 d'être sélectionné. Dès lors :

$$p_1 = p(\{1,2\}) + p(\{1,3\}) + \cdots + p(\{1,10\}) = \frac{9}{45} = \frac{1}{5}.$$

On a vite fait de vérifier que chacun des 9 autres individus de la population est contenu lui aussi dans exactement 9 échantillons possibles. On en déduit que tous les individus de la population ont la même probabilité d'inclusion, égale à $1/5$: pour tout individu i appartenant à la population U, p_i est égal à $1/5$.

$$p_i = \frac{1}{5} \quad \text{pour tout } i \in U.$$

Remarquons que $1/5$ est aussi égal à $2/10$, c'est-à-dire au rapport entre la taille des échantillons possibles et la taille de la population. Ainsi, la probabilité d'inclusion d'un individu de la population coïncide avec le taux de sondage f que l'on s'est fixé.

Par ailleurs, si l'on somme les probabilités d'inclusion des 10 individus de notre population, on obtient :

$$\sum_{i \in U} p_i = p_1 + p_2 + \cdots + p_{10} = 10 \times \frac{1}{5} = 2 = n.$$

Ces caractéristiques des probabilités d'inclusion ne sont pas spécifiques à notre exemple. Il s'agit là de caractéristiques générales des probabilités d'inclusion pour le sondage aléatoire simple sans remise.

a) Les probabilités d'inclusion

La probabilité d'inclusion p_i de l'individu i est donnée par :

$$p_i = \mathrm{P}(i \in S) = \mathrm{P}(\text{prélever un échantillon contenant } i)$$
$$= \sum_{s \in \mathbb{S} | i \in s} p(s) = \sum_{s \in \mathbb{S} | i \in s} \frac{1}{M} = \sum_{s \in \mathbb{S} | i \in s} \frac{1}{C_N^n}$$

$$= \left(\text{nombre d'échantillons possibles contenant } i\right) \times \frac{1}{C_N^n}.$$

Or, le nombre d'échantillons possibles (de taille n) contenant i est égal au nombre d'échantillons qu'il est possible d'obtenir lorsqu'on prélève successivement, « au hasard » et sans remise, $(n-1)$ individus parmi les $(N-1)$ individus de la population autres que i ; ce nombre est donc égal à C_{N-1}^{n-1}. Nous avons ainsi, pour tout $i \in U$:

$$p_i = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{\dfrac{(N-1)!}{(n-1)!\,\big((N-1)-(n-1)\big)!}}{\dfrac{N!}{n!\,(N-n)!}}$$
$$= \frac{(N-1)!\,n!\,(N-n)!}{(n-1)!\,(N-n)!\,N!} = \frac{(N-1)!\,n\,(n-1)!}{(n-1)!\,N\,(N-1)!} = \frac{n}{N} = f\,.$$

Tout individu i de la population possède donc une *même* probabilité d'inclusion, égale au taux de sondage n/N que l'on s'est fixé. Ceci explique pourquoi cette méthode de sondage est aussi appelée « méthode de sondage **à probabilités égales sans remise** » : nous la désignerons d'ailleurs désormais de manière synthétique par « méthode **PESR** » (les lettres PE faisant référence à « probabilités égales » et les lettres SR à « sans remise »).

Par ailleurs :

$$\sum_{i\in U} p_i = \sum_{i\in U} \frac{n}{N} = N\,\frac{n}{N} = n\,.$$

b) Les variables indicatrices d'inclusion

On peut également associer aux individus de la population les *variables indicatrices d'inclusion* suivantes : pour $i \in U$,

$$I_i = \begin{cases} 1 & \text{si } i \in S \\ 0 & \text{sinon.} \end{cases}$$

Ces variables indicatrices vont s'avérer bien pratiques à considérer dans toute une série de développements mathématiques, grâce à leurs propriétés particulières :

- I_i admet une loi de Bernoulli $\mathrm{Bin}(1, p_i)$:
  $$\mathrm{P}(I_i = 1) = \mathrm{P}(i \in S) = p_i \quad \text{et} \quad \mathrm{P}(I_i = 0) = 1 - p_i\,.$$
  Cela implique en particulier que
  $$\mathrm{E}(I_i) = p_i \quad \text{et} \quad \mathrm{V}(I_i) = p_i(1-p_i).$$
  Dans le cas particulier du sondage PESR, $p_i = n/N$ pour tout $i \in U$ et donc
  $$I_i \sim \mathrm{Bin}\left(1, \frac{n}{N}\right) \quad , \quad \mathrm{E}(I_i) = \frac{n}{N} \quad \text{et} \quad \mathrm{V}(I_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n(N-n)}{N^2}\,.$$
- La somme des variables indicatrices d'inclusion associées à tous les individus de la population correspond au nombre d'individus sélectionnés pour faire partie de l'échantillon. Dans le cas du sondage PESR, on a donc :
  $$\sum_{i\in U} I_i = n\,.$$

2.3.3 L'estimation d'une proportion π

a) Paramètre π à estimer et variable d'intérêt y

La population U qui nous intéresse est de taille N. Le **paramètre-population** que l'on souhaite estimer est la proportion π d'individus de la population U possédant une certaine caractéristique.

Dans ce contexte, il est naturel de définir la **variable d'intérêt** y comme la variable *indicatrice* de la présence de cette caractéristique chez un individu : pour l'individu i, cette variable prend la valeur y_i telle que

$$y_i = \begin{cases} 1 & \text{si l'individu } i \text{ possède la caractéristique} \\ 0 & \text{si l'individu } i \text{ ne possède pas la caractéristique.} \end{cases}$$

En définissant la variable d'intérêt y de cette façon, la proportion π coïncide avec la *moyenne* de y dans la population U :

$$\pi = \frac{1}{N}\sum_{i \in U} y_i \,,$$

puisque la somme (ou le total) des valeurs y_i que prend la variable y chez tous les individus i de la population U correspond au *nombre* d'individus de la population qui possèdent la caractéristique prise en considération.

b) L'estimateur $\hat{\pi}$ de π

Lorsqu'on dispose d'un échantillon aléatoire S de taille n prélevé par sondage PESR, on prend pour **estimateur** de π la proportion $\hat{\pi}$ d'individus de cet échantillon S qui présentent la caractéristique en question.

Cet estimateur est une fonction des valeurs observées pour la variable d'intérêt y sur les individus de l'échantillon ; $\hat{\pi}$ correspond en fait à la *moyenne* de la variable y dans l'échantillon S :

$$\hat{\pi} = \frac{1}{n}\sum_{i \in S} y_i \,,$$

puisque la somme des valeurs y_i que prend la variable y chez tous les individus i de l'échantillon S correspond au *nombre* d'individus de l'échantillon qui possèdent la caractéristique considérée.

c) La fluctuation et l'erreur d'échantillonnage

Introduisons ces deux notions fondamentales que sont la fluctuation et l'erreur d'échantillonnage en repartant de notre exemple.

Exemple (suite)

Notre population U est en fait constituée de 10 salariés et nous voulons y estimer la proportion π de travailleurs à temps partiel, autrement dit la proportion d'individus de la population chez qui la variable d'intérêt y, indicatrice du fait de travailler à temps partiel, prend la valeur 1.

Supposons que les valeurs de y dans la population soient en réalité les suivantes :

i	1	2	3	4	5	6	7	8	9	10	Total
y_i	0	0	1	0	1	1	0	0	1	0	4

La proportion π de travailleurs à temps partiel dans cette population est dans ce cas égale à $4/10 = 0{,}4$.

Plaçons-nous à présent dans la situation où l'on ne connaît pas les valeurs que prend y dans la population U et où l'on souhaite estimer la proportion inconnue π en réalisant un sondage PESR de taille $n = 2$. Ceci revient à partir des observations de la variable d'intérêt y dans un échantillon S de 2 individus prélevés par sondage PESR.

Supposons que le hasard nous fasse tomber sur l'échantillon contenant les individus 2 et 5. On interroge alors ces deux individus sur leur temps de travail : l'individu 2 répond qu'il travaille à temps plein (y_2 est égal à 0), alors que l'individu 5 dit qu'il travaille à temps partiel (y_5 est égal à 1). On obtient ainsi, dans l'échantillon, une proportion de travailleurs à temps partiel égale à un demi. Cette proportion observée nous fournit une *estimation* de π.

Et si le hasard nous avait fait sélectionner les individus 4 et 8, nous aurions obtenu une estimation de π égale à 0, puisque les salariés 4 et 8 travaillent tous deux à temps plein.

Si le hasard nous avait conduit à l'échantillon contenant les individus 3 et 6, deux individus travaillant tous deux à temps partiel, nous aurions estimé π par la valeur 1 (ou encore par 100%).

Ces constatations illustrent les concepts d'erreur et de fluctuation d'échantillonnage. En effet, la proportion de travailleurs à temps partiel dans l'*échantillon* de taille 2 ne peut être égale qu'à 0, à 1/2 ou à 1 ; elle ne va dès lors jamais coïncider avec la proportion réelle de travailleurs à temps partiel dans la *population*, égale à 0,4.

L'estimation de π que nous fournit l'échantillon prélevé n'est donc jamais qu'une *approximation* de π ; on commet une certaine erreur — l'*erreur d'échantillonnage* — en remplaçant la valeur exacte de π par l'estimation qu'on a pu en faire dans un échantillon.

Par ailleurs, la proportion d'individus travaillant à temps partiel varie — fluctue — d'un échantillon à l'autre ; par conséquent, l'estimation de la proportion-population π varie selon l'échantillon prélevé. C'est ce qu'on appelle la *fluctuation d'échantillonnage*.

Ainsi, puisque le hasard intervient dans le prélèvement de l'échantillon, il intervient également dans la détermination de la valeur que va prendre notre estimateur $\hat{\pi}$: l'**estimateur** $\hat{\pi} = \hat{\pi}(S)$ est une **variable aléatoire**[1].

[1] Intuitivement, le terme « variable » témoigne du fait que la proportion-échantillon *varie* d'un échantillon à l'autre.
Le terme « aléatoire » rend compte du fait que le hasard intervient dans la détermination de la valeur de la proportion-échantillon. Vous ne pouvez pas déterminer à l'avance quelle valeur particulière prendra la proportion dans l'échantillon, puisque vous ne pouvez pas connaître à l'avance quels seront les individus qui se retrouveront dans l'échantillon prélevé. Le caractère aléatoire de la procédure d'échantillonnage induit le caractère aléatoire de l'estimateur de π.

La **valeur** $\hat{\pi}(s)$ que prend l'estimateur $\hat{\pi}$ dans un échantillon s particulier est appelée une **estimation** de la proportion-population π.

La valeur de $\hat{\pi}$ — autrement dit l'estimation qu'il nous donne de la proportion-population π — varie selon l'échantillon prélevé : ce phénomène porte le nom de **fluctuation d'échantillonnage**.

Par ailleurs, l'estimation de π fournie par $\hat{\pi}$ n'est en général pas égale à la valeur exacte de π : l'estimation de π s'accompagne donc d'une *erreur*, appelée **erreur d'échantillonnage**. Cette erreur n'est pas due à une incompétence du sondeur. Elle est en réalité inévitable et intrinsèquement liée au processus d'estimation dans lequel on cherche à déterminer la valeur d'un paramètre propre à la population tout entière à partir des observations réalisées dans un échantillon, c'est-à-dire un sous-ensemble de cette population. Nous verrons toutefois, dans la suite de ce chapitre, sur quelles caractéristiques de l'échantillon il est possible de jouer pour limiter l'ampleur de cette erreur d'échantillonnage.

d) La distribution d'échantillonnage de $\hat{\pi}$

L'estimateur $\hat{\pi}$ de la proportion-population π est une *variable aléatoire* : *variable* car sa valeur varie d'un échantillon possible à l'autre ; et *aléatoire* car, du fait que le prélèvement de l'échantillon est aléatoire, on ne peut prédire à l'avance quels individus seront effectivement sélectionnés et donc quelle valeur particulière sera obtenue pour $\hat{\pi}$.

Nous pouvons toutefois appréhender les caractéristiques du comportement aléatoire de $\hat{\pi}$ en étudiant ce que l'on appelle sa *distribution d'échantillonnage*.

Cette distribution d'échantillonnage décrit en fait le comportement de $\hat{\pi}$ dans l'ensemble des échantillons possibles. Déterminer cette distribution d'échantillonnage consiste :

- premièrement, à dresser la liste des valeurs que peut prendre $\hat{\pi}$, autrement dit la liste des estimations de la proportion π auxquelles peut conduire $\hat{\pi}$;
- et deuxièmement, à déterminer avec quelle probabilité chacune de ces valeurs possibles de $\hat{\pi}$ peut être obtenue.

Comme nous allons le voir tout de suite dans notre exemple, ces probabilités ne peuvent être déterminées que si l'on connaît le plan de sondage ; il est indispensable de connaître les probabilités de sélection des différents échantillons possibles.

Exemple (suite)

Repartons du plan de sondage relatif à notre exemple (sondage PESR de taille $n = 2$ dans la population U de taille $N = 10$) : nous avons 45 échantillons possibles ayant tous une même probabilité de sélection égale à $1/45$.

Le tableau suivant présente la distribution d'échantillonnage de l'estimateur $\hat{\pi}$ de la proportion π de travailleurs à temps partiel dans la population. Il nous indique, pour chaque échantillon possible $s \in \mathbb{S}$, la valeur particulière $\hat{\pi}(s)$ qu'y prend l'estimateur, et la probabilité associée $\mathrm{P}\big(\hat{\pi} = \hat{\pi}(s)\big) = \mathrm{P}(S = s) = p(s)$, qui n'est autre que la

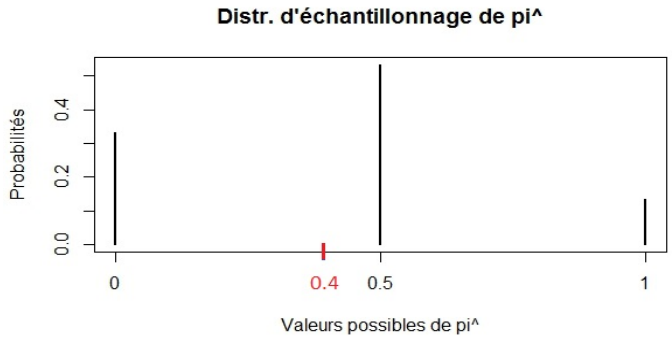
probabilité de sélection de l'échantillon particulier s.

$s \in \mathbb{S}$	$p(s)$	Valeurs observées pour $\mathcal{Y}$ dans s		$\hat{\pi}(s)$
{1,2}	1/45	0	0	0
{1,3}	1/45	0	1	0,5
{1,4}	1/45	0	0	0
{1,5}	1/45	0	1	0,5
{1,6}	1/45	0	1	0,5
{1,7}	1/45	0	0	0
{1,8}	1/45	0	0	0
{1,9}	1/45	0	1	0,5
{1,10}	1/45	0	0	0
{2,3}	1/45	0	1	0,5
{2,4}	1/45	0	0	0
{2,5}	1/45	0	1	0,5
{2,6}	1/45	0	1	0,5
{2,7}	1/45	0	0	0
{2,8}	1/45	0	0	0
{2,9}	1/45	0	1	0,5
{2,10}	1/45	0	0	0
{3,4}	1/45	1	0	0,5
{3,5}	1/45	1	1	1
{3,6}	1/45	1	1	1
{3,7}	1/45	1	0	0,5
{3,8}	1/45	1	0	0,5
{3,9}	1/45	1	1	1
{3,10}	1/45	1	0	0,5
{4,5}	1/45	0	1	0,5
{4,6}	1/45	0	1	0,5
{4,7}	1/45	0	0	0
{4,8}	1/45	0	0	0
{4,9}	1/45	0	1	0,5
{4,10}	1/45	0	0	0
{5,6}	1/45	1	1	1
{5,7}	1/45	1	0	0,5
{5,8}	1/45	1	0	0,5
{5,9}	1/45	1	1	1
{5,10}	1/45	1	0	0,5
{6,7}	1/45	1	0	0,5
{6,8}	1/45	1	0	0,5
{6,9}	1/45	1	1	1
{6,10}	1/45	1	0	0,5
{7,8}	1/45	0	0	0
{7,9}	1/45	0	1	0,5
{7,10}	1/45	0	0	0
{8,9}	1/45	0	1	0,5
{8,10}	1/45	0	0	0
{9,10}	1/45	1	0	0,5

En passant ainsi en revue les 45 échantillons possibles, on se rend compte que $\hat{\pi}$ vaut 0 dans 15 des 45 échantillons possibles, vaut un demi dans 24 échantillons et vaut 1 dans 6 échantillons. La distribution d'échantillonnage de $\hat{\pi}$ peut donc être synthétisée dans le tableau suivant :

Valeurs possibles de $\hat{\pi}$	Nombre d'échantillons	Probabilités associées
0	15	15/45 = 0,33
0,5	24	24/45 = 0,53
1	6	6/45 = 0,13
Somme	45	1

Elle peut également être représentée à l'aide du diagramme en bâtons ci-dessous :



De manière générale

Dans le cadre d'un sondage aléatoire caractérisé par le plan de sondage $\{(s, p(s));\ s \in \mathbb{S}\}$, la **distribution d'échantillonnage** de l'estimateur $\hat{\theta} = \hat{\theta}(S)$ du paramètre-population θ correspond à

$$\left\{\left(\hat{\theta}(s), p(s)\right);\ s \in \mathbb{S}\right\},$$

où $p(s) = \mathrm{P}(S = s) = \mathrm{P}\left(\hat{\theta} = \hat{\theta}(s)\right)$.

La distribution d'échantillonnage de $\hat{\theta}$ peut être synthétisée à l'aide d'une première valeur typique particulièrement importante : sa **moyenne**, aussi appelée l'**espérance** de $\hat{\theta}$.

L'*espérance* de $\hat{\theta}$, notée $\mathrm{E}(\hat{\theta})$, n'est autre que la *moyenne* de la distribution d'échantillonnage de $\hat{\theta}$: c'est autour de cette valeur *centrale* que varient les valeurs possibles de $\hat{\theta}$. Elle se définit par :

$$\mathrm{E}(\hat{\theta}) = \sum_{s \in \mathbb{S}} p(s)\hat{\theta}(s).$$

L'estimateur $\hat{\theta}$ est dit *sans biais* si

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

Ainsi, l'estimation de θ fournie par $\hat{\theta}$ ne coïncide généralement pas avec la valeur exacte de θ. Mais, si $\hat{\theta}$ est sans biais — non biaisé —, il estime correctement θ *en moyenne* ; un estimateur $\hat{\theta}$ sans biais est un estimateur qui, en moyenne, vise juste !

Si $\hat{\theta}$ est un estimateur *biaisé* de θ, on peut définir son biais par

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

Une autre caractéristique de la distribution d'échantillonnage de $\hat{\theta}$ est sa dispersion autour de la valeur de θ. Une faible dispersion implique qu'il y a une grande probabilité que $\hat{\theta}$ prenne une valeur peu éloignée de la valeur exacte de θ, c'est-à-dire, de manière équivalente, qu'il n'y a qu'une faible probabilité — un faible risque — que $\hat{\theta}$ fournisse une « mauvaise » estimation de θ (une estimation de θ fort éloignée de la valeur exacte de θ). A l'inverse, un estimateur $\hat{\theta}$ dont la distribution d'échantillonnage présente une forte dispersion autour de θ est un estimateur ayant une probabilité non négligeable de nous donner une estimation de mauvaise qualité du paramètre θ.

En d'autres termes, plus la dispersion de la distribution d'échantillonnage de $\hat{\theta}$ autour de θ est faible, plus l'estimateur $\hat{\theta}$ est *précis*.

La dispersion de la distribution d'échantillonnage de $\hat{\theta}$ autour de θ peut être quantifiée à l'aide de l'**erreur quadratique moyenne** de $\hat{\theta}$:

$$\begin{aligned}\mathrm{EQM}(\hat{\theta}) &= \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] \\ &= \sum_{s\in\mathbb{S}} p(s)\left(\hat{\theta}(s) - \theta\right)^2.\end{aligned}$$

Cette erreur quadratique moyenne mesure l'amplitude moyenne de l'erreur d'échantillonnage associée à l'estimateur $\hat{\theta}$.

On vérifie aisément que

$$\mathrm{EQM}(\hat{\theta}) = \mathrm{V}(\hat{\theta}) + \left(\mathrm{B}(\hat{\theta})\right)^2$$

où $\mathrm{B}(\hat{\theta})$ est le biais et $\mathrm{V}(\hat{\theta})$ est la variance[2] de $\hat{\theta}$:

$$\begin{aligned}\mathrm{V}(\hat{\theta}) &= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2\right] \\ &= \textstyle\sum_{s\in\mathbb{S}} p(s)\left(\hat{\theta}(s) - \mathrm{E}(\hat{\theta})\right)^2.\end{aligned}$$

Il est clair que si $\hat{\theta}$ est un estimateur non biaisé de θ — un estimateur qui, en moyenne, estime correctement le paramètre θ —,

$$\mathrm{EQM}(\hat{\theta}) = \mathrm{V}(\hat{\theta}) ;$$

la précision d'un estimateur sans biais peut donc être directement quantifiée par l'intermédiaire de sa variance.

e) L'espérance de $\hat{\pi}$

On vérifie aisément que l'espérance de $\hat{\pi}$ coïncide toujours avec la valeur exacte de la proportion-population π :

---

[2] La variance de $\hat{\theta}$ mesure la dispersion de la distribution d'échantillonnage de $\hat{\theta}$ autour de la moyenne de cette distribution d'échantilonnage, c'est-à-dire autour de $\mathrm{E}(\hat{\theta})$.

$$\mathrm{E}(\hat{\pi}) = \pi$$

(voir la démonstration toute simple de ce résultat dans l'annexe 2.1). Dans le cas du sondage PESR, la proportion-échantillon $\hat{\pi}$ est donc un estimateur *sans biais* de la proportion-population π.

Exemple (suite)

En repartant de la distribution d'échantillonnage de $\hat{\pi}$ déterminée précédemment, on obtient que sa moyenne — autrement dit l'espérance de $\hat{\pi}$ — est égale à

$$\frac{15}{45}\times 0 + \frac{24}{45}\times(0{,}5) + \frac{6}{45}\times 1 = \frac{18}{45} = \frac{2}{5} = 0{,}4 = \pi.$$

Le caractère sans biais de $\hat{\pi}$ est donc bien vérifié.

f) La variance et les facteurs de précision de $\hat{\pi}$

Moyennant quelques calculs (relativement simples mais un peu longs), on vérifie que la variance de $\hat{\pi}$ est égale à :

$$\mathrm{V}(\hat{\pi}) = \frac{N-n}{N-1}\,\frac{\pi(1-\pi)}{n}$$

$$\simeq \frac{N-n}{N}\,\frac{\pi(1-\pi)}{n} = (1-f)\,\frac{\pi(1-\pi)}{n} \qquad \text{si } N \text{ est grand,}$$

où $f = n/N$ est le taux de sondage.

Exemple (suite)

Dans notre exemple, $N = 10$, $n = 2$ et $\pi = 0{,}4$, ce qui nous conduit à :

$$\frac{N-n}{N-1}\,\frac{\pi(1-\pi)}{n} = \frac{8}{9}\,\frac{(0{,}4)(0{,}6)}{2} = 0{,}1067.$$

C'est bien la valeur que l'on obtient pour la variance de l'estimateur $\hat{\pi}$ lorsqu'on la calcule directement à partir de la distribution d'échantillonnage de $\hat{\pi}$:

$$\mathrm{V}(\hat{\pi}) = \frac{15}{45}\times(0-0{,}4)^2 + \frac{24}{45}\times(0{,}5-0{,}4)^2 + \frac{6}{45}\times(1-0{,}4)^2$$

$$= \frac{15\times(0{,}16) + 24\times(0{,}01) + 6\times(0{,}36)}{45} = 0{,}1067.$$

Deux remarques fondamentales peuvent être formulées ici.

1. La variance de $\hat{\pi}$, et donc sa précision, dépendent de la taille n de l'échantillon et du taux de sondage appliqué : plus la taille de l'échantillon est grande, plus le taux de sondage se rapproche de 1, plus la variance de $\hat{\pi}$ est faible et plus l'estimateur $\hat{\pi}$ est précis.
   Ce résultat n'est guère étonnant ! Intuitivement, plus grand est l'échantillon, plus élevée est la probabilité qu'il nous donne une image fidèle de la population et nous fournisse ainsi une estimation de π proche de la valeur exacte de cette proportion-population. A la limite, si le taux de sondage f était égal à 1, c'est-à-dire si n était égal à N — ce qui correspondrait au cas où l'on effectue un recensement de tous les individus de la population —, la variance de $\hat{\pi}$ serait nulle : l'échantillon correspondrait à la population tout entière et $\hat{\pi}$ coïnciderait avec π ; $\hat{\pi}$ estimerait alors parfaitement la proportion-population π et il n'y aurait plus d'erreur d'échantillonnage.

2. L'expression de la variance de $\hat{\pi}$ nous montre aussi que la précision de l'estimateur $\hat{\pi}$ dépend de la valeur même de la proportion-population qu'il cherche à estimer ; en effet, la variance de $\hat{\pi}$ est proportionnelle au produit $\pi(1-\pi)$.
   La valeur de ce produit varie en fonction de la valeur de π comme le montre la figure ci-dessous : $\pi(1-\pi)$ atteint sa valeur maximale, un quart (0,25), lorsque π vaut un demi (0,5) ; lorsque π s'éloigne de un demi et prend une valeur qui se rapproche de 0 ou une valeur qui se rapproche de 1, le produit $\pi(1-\pi)$ décroît et se rapproche de plus en plus de la valeur zéro.

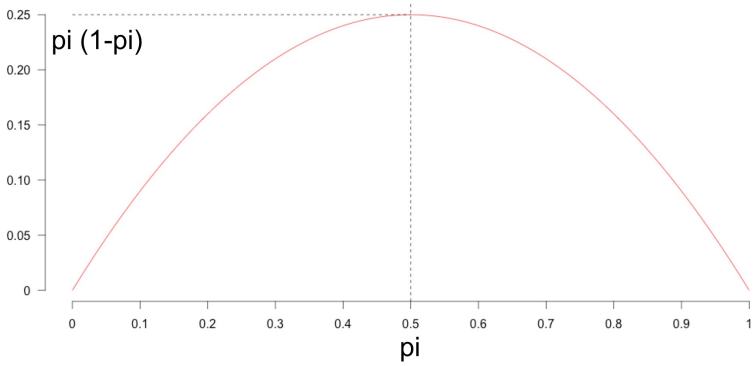


FIGURE 2.2 – *Le produit $\pi(1-\pi)$ en fonction de la valeur de π*

   Ainsi, à taille n d'échantillon fixée, l'estimateur $\hat{\pi}$ a une plus grande variance, donc une moins bonne précision, lorsqu'il doit estimer une proportion-population π proche de un demi que lorsqu'il doit estimer une proportion-population plus faible (plus proche de zéro) ou plus élevée (c'est-à-dire plus proche de 1).

Ces remarques conduisent à formuler le conseil suivant : ne vous lancez pas tête baissée dans l'estimation d'une proportion-population. Prenez d'abord quelques secondes pour réfléchir à l'ordre de grandeur de π. Si vous pensez que π possède une valeur proche de un demi, prévoyez une taille n d'échantillon suffisamment grande pour limiter la variance de l'estimateur $\hat{\pi}$ et lui assurer ainsi un niveau de précision acceptable. En revanche, si vous savez que la proportion π est relativement faible ou, au contraire, fort élevée, vous savez à présent qu'un échantillon de taille plus restreinte vous permettra d'estimer π avec une bonne précision.

Nous reviendrons à ce problème du choix de la taille n de l'échantillon dans la section 7 de ce chapitre.

g) L'estimation de la variance de $\hat{\pi}$

Nous venons d'analyser l'expression mathématique de la variance de la proportion-échantillon $\hat{\pi}$ dans le cas du sondage aléatoire PESR. Cela nous a permis de mettre en évidence de quelle manière la taille n de l'échantillon et la valeur même de la proportion-population π à estimer intervenaient dans cette variance et influençaient dès lors la précision de $\hat{\pi}$.

Mais, si nous disposons d'une expression mathématique pour la variance de $\hat{\pi}$, nous restons toutefois confrontés à un problème pratique non négligeable ! Dans les situations réelles de mise en œuvre d'un sondage, nous sommes bien incapables de

calculer la *valeur* de la variance de $\hat{\pi}$, puisque l'expression de cette variance fait intervenir la proportion-population π qui nous est inconnue.

Comment pouvons-nous alors, dans la pratique, évaluer la variance, et donc la précision, de $\hat{\pi}$?

La réponse est toute simple. Nous allons *estimer* la variance de $\hat{\pi}$ à partir de l'échantillon que nous avons prélevé pour estimer la proportion-population π.

On peut montrer que l'on peut estimer *sans biais* la variance de $\hat{\pi}$ en faisant appel à l'estimateur défini par l'expression :

$$\widehat{\mathrm{V}}(\hat{\pi}) = (1-f)\frac{\hat{\pi}(1-\hat{\pi})}{n-1}.$$

La démonstration du caractère non biaisé de $\widehat{\mathrm{V}}(\hat{\pi})$ est donnée dans l'annexe 2.2 (à lecture facultative, pour ceux d'entre vous qui sont curieux et n'ont pas peur des calculs !).

Nous verrons ultérieurement dans ce chapitre tout l'intérêt de pouvoir estimer la variance de l'estimateur d'une proportion-population, notamment lorsque nous chercherons à construire un intervalle de confiance pour cette proportion.

h) Exercice 2.1

Remarque – Le correctif détaillé de l'exercice proposé ci-dessous est disponible sur l'UV.

Le service social d'une université est intéressé par la proportion π_{job} d'étudiants exerçant de manière régulière (au moins une fois par semaine) un job d'étudiant parmi les 8 750 étudiants de premier ou deuxième cycle.

Une enquête est menée auprès d'un échantillon de 350 étudiants sélectionnés par tirage PESR. Les réponses fournies par ces étudiants à la question « *Exercez-vous un job d'étudiant de manière régulière* ? » sont reprises dans le fichier Excel « Data_ex_2_1 ». Dans ce fichier, la réponse « Oui » a été codée par le chiffre 1 et la réponse « Non » par le chiffre 0. Ce fichier vous indique également, pour chaque étudiant, si celui-ci vit en semaine chez ses parents (valeur 1 pour la variable y_{parents}) ou non (valeur 0).

1. A quelle estimation de π_{job} vous conduisent les réponses obtenues dans l'échantillon ?

- ○ 0,52
- ○ 0,34
- ○ 0,66
- ○ 0,0136
- ○ 119

2. A quelle estimation de π_{parents} — la proportion d'étudiants de premier ou deuxième cycle de l'université qui vivent en semaine chez leurs parents — vous conduisent les réponses obtenues dans l'échantillon ?

- ○ 0,52
- ○ 0,34
- ○ 0,48
- ○ 0,0208
- ○ 182

3. Par quelle valeur (arrondie à 5 décimales) peut-on estimer la variance de l'estimateur $\hat{\pi}_{\text{job}}$ de la proportion π_{job} ?

- ○ 0,02484
- ○ 0,00062
- ○ 0,02620
- ○ 0,00064
- ○ 0,00069

4. Par quelle valeur (arrondie à 5 décimales) peut-on estimer la variance de l'estimateur $\hat{\pi}_{\text{parents}}$ de la proportion π_{parents} ?

- ○ 0,02484
- ○ 0,00062
- ○ 0,02620
- ○ 0,00072
- ○ 0,00069

5. Entre $\hat{\pi}_{\text{job}}$ et $\hat{\pi}_{\text{parents}}$, quel est l'estimateur qui se montre le plus précis ?

- ○ Impossible à dire.
- ○ $\hat{\pi}_{\text{job}}$ semble plus précis que $\hat{\pi}_{\text{parents}}$.
- ○ $\hat{\pi}_{\text{parents}}$ semble plus précis que $\hat{\pi}_{\text{job}}$.
- ○ $\hat{\pi}_{\text{job}}$ et $\hat{\pi}_{\text{parents}}$ sont aussi précis l'un que l'autre.

2.3.4 L'estimation d'une moyenne μ et d'un total τ

a) Les paramètres-population

Considérons ici la situation où la variable d'intérêt y est **quantitative** : autrement dit, y est une variable qui *quantifie* une certaine propriété des individus de la population. Dans ce cas, les valeurs $y_1, y_2, \ldots, y_N$ que prend cette variable auprès des N individus de la population sont, par nature même, *numériques*.

Nous allons plus spécifiquement nous intéresser à l'estimation de deux valeurs typiques (ou synthétiques) de la distribution (ou ensemble des valeurs) de y dans la population U :

- le **total** de y dans U : nous allons le désigner par τ (le « t » grec ; se prononce « tau ») ; il correspond à la somme — au total donc — des valeurs que prend y auprès de tous les individus de la population :
$$\tau = \sum_{i \in U} y_i \ ;$$
- la **moyenne** de y dans U : cette moyenne, que nous allons désigner par μ (le « m » grec ; se prononce « mu »), n'est autre que le total de y dans U, divisé par le nombre N d'individus dans cette population U :
$$\mu = \frac{\tau}{N} = \frac{1}{N} \sum_{i \in U} y_i \ .$$

Il découle de l'égalité $\mu = \tau / N$ que $\tau = N\mu$. Dès lors, puisque la taille N de la population est connue — il s'agit du nombre d'enregistrements dans la base de sondage —, nous pourrons directement déduire l'estimation du total τ de celle de la moyenne μ.

La distribution de la variable d'intérêt y dans la population U n'est pas seulement caractérisée par sa moyenne ou son total ; elle l'est également par sa *dispersion*. On peut appréhender celle-ci par l'intermédiaire de la **variance** de y dans U. Cette variance, désignée par σ^2 (σ correspond au « s » grec et se prononce « sigma »), mesure la dispersion des valeurs que prend y dans la population autour de la moyenne μ de ces valeurs. Elle se définit comme suit :
$$\sigma^2 = \frac{1}{N} \sum_{i \in U} (y_i - \mu)^2 .$$

Si la population est *homogène*, autrement dit si les valeurs de y diffèrent relativement peu d'un individu à l'autre, et donc diffèrent relativement peu de la moyenne μ, la variance σ^2 est faible ; si, au contraire, la population est fort *hétérogène*, c'est-à-dire que les valeurs de y diffèrent fortement d'un individu à l'autre et présentent ainsi une forte dispersion autour de la valeur moyenne μ, la variance σ^2 est élevée. Plus les individus de la population sont « différents » les uns des autres pour ce qui est de la variable y, plus la variance σ^2 est grande.

Nous allons le voir très rapidement, ce n'est pas tant cette variance σ^2, que l'on pourrait qualifier de « classique », qui va jouer un rôle important dans la suite, mais plutôt la **variance** dite « **corrigée** », qui se définit en reprenant la même somme que celle qui intervient dans σ^2, mais en divisant cette somme non pas par N, mais par $(N-1)$:
$$\sigma^2_{\text{corr}} = \frac{1}{N-1} \sum_{i \in U} (y_i - \mu)^2 = \frac{N}{N-1} \sigma^2 .$$

En d'autres termes, la variance corrigée de la variable d'intérêt y dans la population U est égale à la variance « classique », multipliée par le rapport entre N et $(N-1)$. Il est clair que si la taille N de la population est grande — ce qui est généralement le cas — ce rapport $N/(N-1)$ est pratiquement égal à 1 et la variance corrigée coïncide approximativement avec la variance « classique » ; l'intérêt de distinguer la variance corrigée de la variance « classique » est alors purement théorique.

Du point de vue mathématique, un des avantages de la variance corrigée sur la variance « classique » est que σ^2_{corr} n'est définie que si la taille N de la population est supérieure ou égale à 2. Quoi de plus naturel ! Pour pouvoir commencer à parler de « dispersion », il faut qu'il y ait au moins deux individus en jeu !

b) Les estimateurs $\hat{\mu}$ et $\hat{\tau}$ de μ et τ

Prenons un petit exemple pour illustrer ces paramètres que sont le *total*, la *moyenne* et la *variance* d'une variable quantitative y dans la population. Nous verrons également comment estimer les deux premiers de ces paramètres.

Exemple

Considérons une population constituée de 10 personnes et intéressons-nous à la variable y correspondant à la somme d'argent contenue, en billets, dans le porte-monnaie ou le portefeuille de ces personnes : désignons par y_i le montant (en euros) détenu par l'individu i.

La *distribution* de la variable d'intérêt y dans notre population est en réalité la suivante :

i	y_i
1	100
2	80
3	100
4	55
5	40
6	10
7	50
8	25
9	60
10	120

L'individu numéro 1 détient 100 euros, l'individu numéro 2 possède 80 euros, etc. Les 10 individus de la population détiennent au total 640 euros, ce qui nous fait un montant moyen par individu égal à 640/10, c'est-à-dire 64 euros. En d'autres termes, le *total* τ et la *moyenne* μ de la variable y dans la population s'élèvent respectivement à 640 euros et à 64 euros.

Quelle est la dispersion des montants détenus par les 10 personnes de la population autour du montant moyen de 64 euros ? Pour répondre à cette question, calculons la *variance classique*, puis la *variance corrigée*, de ces 10 montants.

La variance classique s'obtient en prenant :

$$\sigma^2 = \frac{(100-64)^2 + (80-64)^2 + \cdots + (120-64)^2}{10} = 1129.$$

La variance corrigée vaut dès lors :

$$\sigma^2_{\text{corr}} = \frac{10}{9} 1129 = 1254{,}44.$$

Imaginons à présent que l'on nous cache l'ensemble de la population et qu'on nous demande d'estimer le montant total τ détenu par l'ensemble des 10 personnes, ainsi que le montant μ détenu en moyenne par chaque individu. Pour ce faire, on nous donne le droit d'interroger 2 individus sélectionnés « au hasard » et sans remise parmi les 10 personnes qui composent la population. En d'autres termes, on nous donne le droit d'utiliser un échantillon s de taille $n = 2$, prélevé par sondage PESR.

Nous allons, comme pour l'estimation d'une proportion, faire appel aux estimateurs les plus naturels qui soient : nous allons estimer la moyenne μ de la variable y dans la population par la moyenne de la variable y dans l'échantillon s. Cette moyenne-échantillon, que l'on désigne habituellement par $\bar{y}$, s'obtient en faisant la somme des valeurs que prend y auprès des individus de l'échantillon, puis en divisant cette somme par la taille n de l'échantillon.

Et pour estimer le total τ de la variable y dans la population ? Il suffit de se rappeler que τ est égal à la taille N de la population multipliée par la moyenne-population μ. Dès lors, puisque N est connu — il vaut 10 dans notre exemple —, nous pouvons prendre pour estimateur de τ, N fois l'estimateur de μ, c'est-à-dire N fois $\bar{y}$.

En pratique, supposons que le hasard nous fasse sélectionner les individus 2 et 9. Ceux-ci nous indiquent qu'ils possèdent respectivement 80 et 60 euros. Nous pouvons alors estimer la moyenne-population μ par la moyenne de ces deux montants, soit $(80 + 60)/2 = 70$ euros ; quant au total-population τ, nous pouvons l'estimer par la valeur de 10 fois 70 euros, c'est-à-dire 700 euros.

De manière générale

Comment estimer la moyenne μ et le total τ de la variable y dans la population dans le cadre d'un sondage PESR?

On tire un échantillon aléatoire S de taille n en prélevant successivement, à l'aveugle et sans remise, n individus dans la population U. Puis, on observe les valeurs y_i que prend la variable d'intérêt y chez les individus i qui appartiennent à l'échantillon. On fait ensuite appel aux estimateurs les plus naturels qui soient :

- on estime la moyenne μ de la variable y dans la population par la moyenne de la variable y dans l'échantillon S :

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i \in S} y_i \,.$$

- puisque $\tau = N\mu$ et que la taille N de la population est connue, on prend pour estimateur de τ :

$$\hat{\tau} = N\hat{\mu} = N\bar{y} \,.$$

Nous pouvons encore réécrire $\hat{\tau}$ comme suit :

$$\hat{\tau} = \frac{N}{n}\sum_{i \in S} y_i = \sum_{i \in S} \frac{N}{n} y_i \,.$$

La dernière formulation de $\hat{\tau}$ nous montre que l'estimateur du total-population peut s'exprimer comme la somme des valeurs que prend la variable y auprès des individus de l'échantillon, chaque valeur étant répliquée N/n fois dans cette somme. Cette

reformulation de $\hat{\tau}$ met clairement en évidence l'idée intuitive suivie par cet estimateur. Dans le cas du sondage aléatoire simple, chaque individu de la population a la même probabilité d'être sélectionné pour faire partie de l'échantillon : cette probabilité, on l'a vu précédemment, est égale au taux de sondage n/N. On peut dès lors considérer que chaque individu de l'échantillon représente ainsi N/n individus de la population. Pour reconstituer le total de y dans la population, il est alors naturel de faire la somme des valeurs que prend cette variable y dans l'échantillon S, mais de comptabiliser la valeur associée à l'individu i de l'échantillon autant de fois qu'il y a d'individus de la population représentés par i, soit N/n fois.

Dans l'exemple ci-dessus, on tire un échantillon de taille $n = 2$ dans une population de taille $N = 10$; le taux de sondage est donc égal à 2/10, soit un cinquième. L'échantillon prélevé est constitué d'un cinquième des individus de la population, et tout se passe comme si chacun des 2 individus de l'échantillon représentait 10/2 = 5 individus de la population. Il est alors naturel d'estimer le total de y dans la population en additionnant 5 fois la valeur prise par y chez le premier individu de l'échantillon et 5 fois la valeur prise par y chez le second individu de l'échantillon.

c) L'espérance de $\hat{\mu}$ et de $\hat{\tau}$

On vérifie aisément que $\hat{\mu}$ et $\hat{\tau}$ sont des estimateurs sans biais de μ et τ, respectivement :

$$\mathrm{E}(\hat{\mu}) = \mu \quad \text{et} \quad \mathrm{E}(\hat{\tau}) = \tau$$

(voir l'annexe 2.3 pour la démonstration toute simple de ce résultat).

d) La variance et les facteurs de précision de $\hat{\mu}$ et de $\hat{\tau}$

On vérifie (moyennant quelques calculs…) que la variance de $\hat{\mu}$ est donnée par :

$$\mathrm{V}(\hat{\mu}) = (1 - f)\frac{\sigma^2_{\mathrm{corr}}}{n}.$$

Il s'ensuit que :

$$\mathrm{V}(\hat{\tau}) = \mathrm{V}(N\hat{\mu}) = N^2\mathrm{V}(\hat{\mu}) = N^2(1 - f)\frac{\sigma^2_{\mathrm{corr}}}{n}.$$

Puisque la variance de $\hat{\tau}$ est proportionnelle à N^2, elle peut être excessivement élevée dès que la taille N de la population est grande.

Analysons de plus près la variance de $\hat{\mu}$. Cette variance dépend de la taille n de l'échantillon et du taux de sondage $f = n/N$: plus n est grand et se rapproche de la taille N de la population, plus la variance de $\hat{\mu}$ diminue et, par conséquent, plus la précision de $\hat{\mu}$ est grande. Cette propriété est similaire à celle que nous avions vue pour l'estimateur d'une proportion-population : elle exprime simplement le fait que plus l'échantillon est grand, plus on peut avoir confiance dans l'estimation qui en résulte, qu'il s'agisse de l'estimation d'une proportion ou de la moyenne d'une variable quantitative.

Notons encore que si la taille N de la population est très grande, le taux de sondage f que l'on peut atteindre reste toujours excessivement faible et le facteur $(1 - f)$ est alors assimilable à 1 ; dans ce cas, on peut considérer que la précision de l'estimateur de μ dépend de la taille de l'échantillon, mais pratiquement plus de celle de la population. Il s'agit là d'un point difficile à faire admettre au public qui ne connaît pas la théorie des sondages. On pourrait penser *a priori* qu'une enquête sur 2 000 individus dans une ville

de 500 000 habitants est bien plus précise qu'une enquête auprès de 2 000 individus prélevés dans une région comptant 10 millions d'habitants, mais il n'en est rien, du moins si les deux populations-cibles présentent la même hétérogénéité (nous allons y revenir tout de suite) : pour ces deux enquêtes, le taux de sondage est tellement faible — il vaut 0,4% dans le premier cas et 0,02% dans le second — que le facteur $(1-f)$ n'intervient pratiquement pas dans la variance de $\hat{\mu}$.

Une autre caractéristique influençant la précision de l'estimateur d'une moyenne ou d'un total-population est le caractère homogène ou hétérogène de la population vis-à-vis de la variable d'intérêt y. Plus précisément, à taille n d'échantillon fixée, plus la variance de y dans la population est élevée, plus les variances des estimateurs de μ et de τ sont grandes, moins ces estimateurs sont précis.

Que se passe-t-il en réalité ? Intuitivement, si y possède une variance élevée, c'est que les valeurs de y varient fortement d'un individu à l'autre de la population (on dira que la population est *hétérogène*) ; dans ce cas, les valeurs observées pour y peuvent varier assez fortement d'un échantillon à l'autre, et donc l'estimation obtenue pour la moyenne ou le total-population varie elle aussi fortement d'un échantillon à l'autre. Ainsi, l'hétérogénéité de la population induit une plus grande dispersion de la distribution d'échantillonnage de l'estimateur et a donc un effet négatif sur la précision de celui-ci.

Que faire face à une population fort hétérogène, si l'on veut que la précision de l'estimateur de la moyenne ou du total-population soit acceptable ? On peut bien sûr prévoir une taille n d'échantillon suffisamment grande pour contrer l'effet négatif de la valeur élevée de σ^2_{corr}. Mais cette solution n'est pas toujours praticable : de par le budget, les moyens techniques et le temps dont on dispose, on peut être limité à une taille d'échantillon qui ne permet pas d'atteindre la précision souhaitée. Il existe heureusement une solution efficace au manque de précision des estimateurs induit par l'hétérogénéité de la population : cette solution consiste à mettre en œuvre un sondage *stratifié* plutôt que le sondage aléatoire simple. Nous étudierons cette méthode de sondage particulière dans le chapitre 3 de ce cours.

e) L'estimation de la variance de $\hat{\mu}$ et de $\hat{\tau}$

Les expressions des variances de $\hat{\mu}$ et de $\hat{\tau}$ font intervenir σ^2_{corr}, la variance corrigée de la variable d'intérêt y dans la population. Or, nous ne connaissons généralement pas la valeur exacte de cette variance ! Comment faire alors, dans la pratique, pour évaluer la variance de $\hat{\mu}$ ou de $\hat{\tau}$?

La solution à ce problème consiste tout simplement à estimer σ^2_{corr} à partir de l'échantillon prélevé pour estimer μ ou τ.

On montre (ceux d'entre vous qui le souhaitent peuvent consulter l'annexe 2.4) que l'on peut estimer *sans biais* σ^2_{corr} via la variance *corrigée* de la variable y dans l'échantillon : cette variance-échantillon corrigée, désignée par s^2_{corr}, correspond à la variance classique de y dans l'échantillon, si ce n'est que la somme des carrés des écarts entre les valeurs de y dans l'échantillon et la moyenne $\bar{y}$ de ces valeurs est divisée par $(n-1)$ plutôt que par n :

$$s^2_{\text{corr}} = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2.$$

On obtient donc des estimateurs *sans biais* des variances de $\hat{\mu}$ et de $\hat{\tau}$ en remplaçant, dans les expressions de ces variances, la variance-population σ^2_{corr} inconnue par son estimateur sans biais, la variance-échantillon s^2_{corr} :

$$\widehat{\mathrm{V}}(\hat{\mu}) = (1-f)\frac{s^2_{\text{corr}}}{n} \quad \text{et} \quad \widehat{\mathrm{V}}(\hat{\tau}) = N^2\widehat{\mathrm{V}}(\hat{\mu}) = N^2(1-f)\frac{s^2_{\text{corr}}}{n}.$$

f) Exercice 2.2

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Remarque – Le correctif détaillé de l'exercice proposé ci-dessous est disponible sur l'UV dans le fichier « Corr_ex_2_2.xlsx.

On s'intéresse à deux populations différentes de médecins : la population 1 compte 3 593 médecins, tandis que la population 2 n'en compte que 752.

On désire estimer les moyennes des variables y_{pa} et y_{exp} dans chacune des deux populations, où y_{pa} est la variable correspondant au nombre moyen de patients que voit un médecin pendant sa journée de travail et y_{exp} est la variable correspondant au nombre d'années d'expérience professionnelle d'un médecin.
Pour obtenir les estimations de μ_{pa} et μ_{exp}, on prélève un échantillon aléatoire simple de taille $n_1 = 360$ dans la population 1 et un échantillon aléatoire simple de taille $n_2 = 75$ dans la population 2.

Les valeurs observées pour les deux variables d'intérêt, y_{pa} et y_{exp}, auprès des médecins des deux échantillons prélevés sont reprises dans le fichier Excel « Data_ex_2_2 ».

Attention ! Pour les **questions 1 à 4**, introduisez vos réponses dans les cases prévues à cet effet, avec une précision de **1 décimale**

1. Qu'obtient-on comme estimation de la moyenne μ_{pa} dans la population 1 (*précision de votre réponse : 1 décimale*) ?
2. Qu'obtient-on comme estimation de la moyenne μ_{pa} dans la population 2 (*précision de votre réponse : 1 décimale*) ?
3. Qu'obtient-on comme estimation de la moyenne μ_{exp} dans la population 1 (*précision de votre réponse : 1 décimale*) ?
4. Qu'obtient-on comme estimation de la moyenne μ_{exp} dans la population 2 (*précision de votre réponse : 1 décimale*) ?
5. Qu'obtient-on comme estimation du nombre total τ_{pa} de patients que voient l'ensemble des médecins de la population 1 sur une journée de travail ?

6. Qu'obtient-on comme estimation du nombre total τ_{pa} de patients que voient l'ensemble des médecins de la population 2 sur une journée de travail ?

7. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\mu}_{pa}$ pour le sondage réalisé dans la population 1 (*précision de votre réponse : 2 décimales*) ?

8. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\mu}_{pa}$ pour le sondage réalisé dans la population 2 (*précision de votre réponse : 2 décimales*) ?

9. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\mu}_{exp}$ pour le sondage réalisé dans la population 1 (*précision de votre réponse : 2 décimales*) ?

10. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\mu}_{exp}$ pour le sondage réalisé dans la population 2 (*précision de votre réponse : 2 décimales*) ?

11. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\tau}_{pa}$ pour le sondage réalisé dans la population 1 (*précision de votre réponse : 2 décimales*) ?

12. Qu'obtient-on comme estimation de la variance de l'estimateur $\hat{\tau}_{pa}$ pour le sondage réalisé dans la population 2 (*précision de votre réponse : 2 décimales*) ?

g) Moyenne et proportion

Au point 2.3.3, nous nous sommes intéressés à l'estimation d'une *proportion*-population dans le cadre du sondage aléatoire simple PESR. Nous l'avons vu, la proportion d'unités de la population possédant une certaine caractéristique peut être vue comme la *moyenne*, dans la population, d'une variable d'intérêt Y dite « dichotomique », indiquant la présence ou l'absence de cette caractéristique : ainsi, par exemple, la proportion de femmes de moins de 25 ans dans la population n'est autre que la moyenne, dans la population, de la variable qui prend la valeur 1 chez les individus qui sont des femmes de moins de 25 ans, et qui prend la valeur 0 sinon.

L'annexe 2.5 (que je vous invite vivement à lire) fait explicitement le lien entre les résultats que nous avons obtenus au point 2.3.3, et ceux que nous venons d'obtenir pour l'estimation d'une moyenne-population. Il suffit de retraduire les résultats obtenus pour $\hat{\mu}$ dans le cas particulier où la variable d'intérêt Y n'a que deux valeurs possibles : les valeurs 1 et 0.

Ainsi, estimer une proportion, c'est estimer la moyenne d'une variable un peu particulière. Par conséquent, nous ne distinguerons plus toujours explicitement, désormais, l'estimation d'une proportion de celle d'une moyenne !

2.4 Le sondage aléatoire simple *avec remise* (PEAR)

Intéressons-nous à présent à la procédure d'échantillonnage qui consiste à effectuer n prélèvements successifs, « au hasard » et AVEC remise, dans la population : on tire « à l'aveugle » un premier individu dans la population et on lui administre l'enquête ; on le remet ensuite dans la population avant d'effectuer « à l'aveugle » un deuxième prélèvement d'individu dans la population, etc. En d'autres termes, chacun des n prélèvements est effectué dans l'ensemble de la base de sondage, indépendamment des prélèvements précédents ; les n prélèvements se font donc indépendamment l'un de l'autre[1].

2.4.1 Le plan de sondage

Partons d'un petit exemple qui va nous permettre de découvrir facilement les caractéristiques du plan de sondage associé au sondage PEAR et ce qui le différencie du plan de sondage associé au tirage PESR.

Exemple

Considérons une population constituée de 5 individus : $U = \{1, 2, 3, 4, 5\}$. Quel est le plan de sondage associé à la procédure d'échantillonnage consistant à effectuer 2 prélèvements successifs, « au hasard » et avec remise, dans cette population ?

L'ensemble $\mathbb{S}$ de tous les échantillons possibles est le suivant :
$\{\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{2,2\},\{2,3\},\{2,4\},\{2,5\},\{3,3\},\{3,4\},\{3,5\},\{4,4\},\{4,5\},\{5,5\}\}$,
où $\{i, i\}$ $(i = 1, \dots, 5)$ correspond à l'échantillon obtenu lorsqu'on tire à deux reprises l'individu n° i.

Comparativement au cas PESR, on a 5 échantillons possibles supplémentaires correspondant aux cas où l'on sélectionne à deux reprises le même individu.

Quelles sont les probabilités de sélection des échantillons possibles ?

- $p(\{1,1\}) = \mathrm{P}(\text{tirer l'individu 1 au 1er et au 2e prélèvements}) = \frac{1}{5} \times \frac{1}{5} = \frac{1}{25}$.
- $p(\{1,2\}) =$
  P(tirer l'individu 1 au 1er prélèvement et l'individu 2 au 2e prélèvement)
  +P(tirer l'individu 2 au 1er prélèvement et l'individu 1 au 2e prélèvement)
  $= \frac{1}{5} \times \frac{1}{5} + \frac{1}{5} \times \frac{1}{5} = \frac{1}{25} + \frac{1}{25} = \frac{2}{25}$.

Il est clair que les 5 échantillons $\{i, i\}$ $(i = 1, \dots, 5)$ ont une probabilité de sélection égale

[1] Notez que dans le tirage PESR, les prélèvements ne se font pas indépendamment l'un de l'autre, puisque chaque prélèvement modifie l'ensemble des unités statistiques dans lequel va s'effectuer le prélèvement suivant.

à $1/25$, alors que les 10 autres échantillons $\{i, j\}$ constitués de deux individus i et j distincts ont une probabilité de sélection égale à $2/25$.

De manière générale

- Pour la procédure d'échantillonnage consistant à effectuer n prélèvements successifs, « au hasard » et AVEC remise, dans une population de taille N, les échantillons possibles sont de la forme :

$$s = \{i_1, i_2, \dots, i_n\},$$

avec $i_1, i_2, \dots, i_n \in U$ et non nécessairement distincts (puisqu'un même individu peut être sélectionné à plusieurs reprises).

- L'expression donnant le nombre M d'échantillons possibles est fort complexe. En effet, M est égal au nombre d'échantillons dont les n individus sont distincts, plus le nombre d'échantillons dans lesquels un individu est sélectionné deux fois et les $(n-2)$ autres individus sont distincts, plus le nombre d'échantillons dans lesquels deux individus sont sélectionnés à deux reprises et les $(n-4)$ autres individus sont distincts, plus…

- Les échantillons possibles n'ont pas tous la même probabilité de sélection.

Il peut apparaître « gênant » de ne pas avoir la même probabilité de sélection pour tous les échantillons possibles. Mais ce « problème » n'en est pas réellement un ! Il nous suffit de définir les échantillons possibles en tenant compte de l'ordre des prélèvements pour se retrouver avec un plan de sondage *simple*, dans lequel tous les échantillons possibles ont la même probabilité d'être tirés.

- Si l'on tient compte de l'ordre dans lequel les individus sont prélevés, les échantillons possibles sont de la forme :

$$s_{\text{ord}} = (i_1, i_2, \dots, i_n)$$

où $i_1, i_2, \dots, i_n \in U$ et i_k est l'individu sélectionné lors du k-ème prélèvement $(k = 1, \dots, n)$.
- Le nombre M d'échantillons possibles est alors simplement égal à N^n.
- Les N^n échantillons possibles ont tous la même probabilité d'être sélectionnés, égale à :

$$p(s_{\text{ord}}) = \frac{1}{N^n}.$$

Dans notre exemple, nous aurions ainsi $5^2 = 25$ échantillons possibles, ayant tous 1 chance sur 25 d'être sélectionnés.

2.4.2 Les probabilités d'inclusion

a) Les probabilités d'inclusion

Dans le cas du sondage PESR de taille n dans une population de taille N, tous les individus de la population possèdent la même probabilité d'inclusion, égale au taux de sondage $f = n/N$.

Que valent les probabilités d'inclusion des individus de la population dans le cas du tirage AVEC remise de l'échantillon ?

Quel que soit l'individu i considéré, sa probabilité d'inclusion p_i est égale à :

$$\begin{aligned} p_i &= \mathrm{P}(i \text{ est sélectionné lors d'au moins l'un des } n \text{ prélèvements}) \\ &= 1 - \mathrm{P}(i \text{ n'est sélectionné à aucun des } n \text{ prélèvements}) \\ &= 1 - \left(1 - \frac{1}{N}\right) \times \left(1 - \frac{1}{N}\right) \times \ldots \times \left(1 - \frac{1}{N}\right) \\ &= 1 - \left(1 - \frac{1}{N}\right)^n = 1 - \left(\frac{N-1}{N}\right)^n . \end{aligned}$$

Tous les individus de la population ont donc la même probabilité d'inclusion, ce qui motive l'appellation PEAR – *à probabilités égales* avec remise – de la procédure d'échantillonnage considérée ici.

Exemple (suite)

Dans l'exemple considéré précédemment, $N = 5$ et $n = 2$:

- pour l'échantillonnage PESR : $p_i = 2/5 = 0{,}4$ pour tout $i \in U$;
- pour l'échantillonnage PEAR : $p_i = 1 - \left(\frac{4}{5}\right)^2 = 0{,}36$.

b) Comparaison des probabilités d'inclusion pour PESR et PEAR

Si le nombre n de prélèvements à effectuer est très petit par rapport à N, c'est-à-dire si le taux de sondage $f = n/N$ est très faible,

$$1 - \left(1 - \frac{1}{N}\right)^n \simeq \frac{n}{N} ;$$

les probabilités d'inclusion pour le sondage PEAR sont pratiquement identiques à celles pour le sondage PESR.

Intuitivement, ceci s'explique par le fait que, si $n \ll N$, la probabilité de sélectionner à deux ou plusieurs reprises un même individu est pratiquement nulle. Dans ce cas, le plan de sondage PEAR est très ressemblant au plan de sondage PESR, car les probabilités de sélection associées aux échantillons possibles qui contiennent à deux ou plusieurs reprises un même individu sont pratiquement nulles.

Ainsi, par exemple, pour $N = 1000$ et $n = 10$: $f = \frac{n}{N} = 1\%$. Dès lors :

- PESR : $p_i = f = 1\%$;
- PEAR : $p_i = 1 - \left(1 - \frac{1}{1000}\right)^{10} = 0{,}00995 \simeq 0{,}01 = 1\%$.

La probabilité que, au cours des 10 prélèvements, on « tombe » à plus d'une reprise sur un même individu est extrêmement faible.

2.4.3 Tirage PEAR de l'échantillon et échantillon aléatoire simple en statistique « classique »

Dans les cours de statistique « classique », la notion d'*échantillon aléatoire simple* de taille n renvoie à un ensemble de n observations (ou variables aléatoires) $Y_1, Y_2, \ldots, Y_n$

indépendantes et identiquement distribuées (i.i.d.). Un tel échantillon aléatoire simple d'observations est en réalité étroitement lié au prélèvement par tirage PEAR d'un échantillon de n individus (ou unités statistiques) dans une population.

Nous pouvons en effet formaliser la situation comme suit :

- Supposons que nous soyons intéressés par la *variable d'intérêt* $\mathcal{Y}$ dans une certaine population U d'individus.

- Tirer « au hasard » (à l'aveugle) un individu dans la population U est une *expérience aléatoire* dont l'ensemble Ω de tous les résultats possibles est la population U elle-même (puisque n'importe lequel des individus de U peut être sélectionné, avec une probabilité égale à $1/N$).

  On peut associer à cette expérience aléatoire la *variable aléatoire* Y correspondant à la variable d'intérêt $\mathcal{Y}$ pour l'individu sélectionné : en d'autres termes, Y prend la valeur y_i (la valeur de la variable d'intérêt $\mathcal{Y}$ pour l'individu i de la population) si le tirage « au hasard » dans la population nous fait sélectionner l'individu i.

  La *distribution de probabilité* de l'observation aléatoire Y coïncide parfaitement avec la *distribution de fréquences* de la variable d'intérêt $\mathcal{Y}$ dans la population U. Ceci a pour conséquence que :
  - $\mathrm{E}(Y) = \mu$ où $\mu = \frac{1}{N}\sum_{i\in U} y_i$ est la moyenne de la variable d'intérêt $\mathcal{Y}$ dans la population ;
  - $\mathrm{V}(Y) = \sigma^2$ où $\sigma^2 = \frac{1}{N}\sum_{i\in U}(y_i - \mu)^2$ est la variance de la variable d'intérêt $\mathcal{Y}$ dans la population.

- Effectuer n prélèvements PEAR dans la population revient à répéter l'expérience aléatoire décrite ci-dessus à n reprises, sous des conditions expérimentales identiques et de manière indépendante. Au k-ème prélèvement $(k = 1, \dots, n)$ peut être associée la variable aléatoire Y_k correspondant à la variable d'intérêt $\mathcal{Y}$ observée auprès de l'individu sélectionné au cours de ce k-ème prélèvement. Aux n prélèvements PEAR sont ainsi associées les n observations aléatoires $Y_1, \dots, Y_n$ qui ont pour caractéristiques d'être indépendantes les unes des autres et identiquement distribuées : elles ont toutes la même distribution de probabilité, qui coïncide avec la distribution de fréquences de la variable d'intérêt $\mathcal{Y}$ dans la population U. Les variables aléatoires Y_k $(k = 1, \dots, n)$ constituent ainsi un *échantillon aléatoire simple* tel que considéré dans les cours de statistique « classique » et sont telles que, pour tout $k = 1, \dots, n$:
  $$\mathrm{E}(Y_k) = \mu \text{ et } \mathrm{V}(Y_k) = \sigma^2$$
  où μ et σ^2 sont la moyenne et la variance de la variable d'intérêt $\mathcal{Y}$ dans la population U.

La situation est plus complexe dans le cas où les n prélèvements se font SANS remise. On peut toujours associer au k-ème prélèvement $(k = 1, \dots, n)$ la variable aléatoire Y_k correspondant à la variable d'intérêt $\mathcal{Y}$ pour l'individu sélectionné au cours de ce k-ème

prélèvement. Cependant, les observations aléatoires $Y_1, \dots, Y_n$ ainsi définies ne sont plus ici ni indépendantes, ni identiquement distribuées !

En conclusion, appliquer une méthode statistique « classique » construite à partir d'observations aléatoires $Y_1, \dots, Y_n$ i.i.d. n'est théoriquement valide que si l'échantillon d'individus sur lesquels on va observer la variable d'intérêt y est prélevé dans la population par tirage PEAR.

Cela veut-il dire que vous ne pouvez pas appliquer ce qu'on vous a enseigné dans vos autres cours de statistique (tests d'hypothèses, modèles de régression…) dès que l'échantillon d'individus est prélevé par tirage PESR dans la population… autrement dit dans la plupart des cas ? La réponse à cette question est nuancée.

Nous avons vu que si le nombre n de prélèvements effectués est fort petit par rapport à la taille N de la population – c'est le cas dans bon nombre de sondages appliqués dans des populations de grande taille – le plan de sondage PESR est très proche du plan de sondage PEAR ; dans ce cas, les outils de la statistique « classique » peuvent être appliqués sans problème. En revanche, si le taux de sondage n/N est relativement élevé, le sondage PESR ne peut plus être assimilé à un sondage PEAR et les méthodes de la statistique classique ne sont plus valides en cas de tirage SANS remise de l'échantillon des individus. Il est important de garder cela à l'esprit lorsqu'on désire mener une analyse statistique inférentielle sur les résultats d'un sondage aléatoire !

2.4.4 L'estimation d'une proportion π

a) L'estimateur $\hat{\pi}$ de π

Soit la proportion (inconnue) π d'individus de la *population* qui possèdent une certaine caractéristique A. L'**estimateur** de π dans le cas du sondage PEAR est identique à celui utilisé dans le cas du sondage PESR : il s'agit de la proportion $\hat{\pi}$ d'individus de l'*échantillon* possédant la caractéristique A.

b) Les propriétés de $\hat{\pi}$

Quelles sont les **propriétés** de $\hat{\pi}$ dans le cas du sondage PEAR ?

La variable d'intérêt y considérée ici est la variable indicatrice de la présence de la caractéristique A chez un individu : pour l'individu i, cette variable prend la valeur

$$y_i = \begin{cases} 1 & \text{si l'individu } i \text{ possède la caractéristique A} \\ 0 & \text{sinon.} \end{cases}$$

La distribution de fréquences de y dans la population est toute simple :

- y prend la valeur 1 auprès d'une proportion π d'individus de la population ;
- y prend la valeur 0 auprès d'une proportion $1 - \pi$ d'individus de la population.

Si l'on se rappelle les explications données au point 2.4.3 de ce document, on peut alors associer à la procédure de tirage PEAR de n individus les observations aléatoires $Y_1, \dots, Y_n$ indépendantes et identiquement distribuées : pour tout $k = 1, \dots, n$,

$$Y_k \sim \text{Bin}(1, \pi),$$

puisque Y_k ne peut prendre que les valeurs 1 et 0, avec les probabilités

$\mathrm{P}(Y_k = 1)$
$= \mathrm{P}$(l'individu sélectionné au k-ème prélèvement possède la caractéristique A) $= \pi$
et
$\mathrm{P}(Y_k = 0)$
$= \mathrm{P}$(l'individu sélectionné au k-ème prélèvement ne possède pas la caractéristique A)
$= 1 - \pi$.
Dès lors, pour tout $k = 1, \dots, n$:

$$\mathrm{E}(Y_k) = \pi \quad \text{et} \quad \mathrm{V}(Y_k) = \pi(1-\pi).$$

Il est clair que la proportion-échantillon $\hat{\pi}$ peut se voir comme la moyenne arithmétique de ces observations aléatoires i.i.d. $Y_1, \dots, Y_n$:

$$\hat{\pi} = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

Dès lors, dans le cas du sondage PEAR :

$$\mathrm{E}(\hat{\pi}) = \mathrm{E}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}(Y_k) = \frac{1}{n}\sum_{k=1}^{n} \pi = \frac{1}{n} n\pi = \pi$$

et

$$\mathrm{V}(\hat{\pi}) = \mathrm{V}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{V}(Y_k) = \frac{1}{n^2}\sum_{k=1}^{n} \pi(1-\pi) = \frac{1}{n^2} n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}.$$

Vous retrouvez là des résultats que vous avez certainement rencontrés dans vos autres cours de statistique !

Que pouvons-nous conclure de ces résultats ?

La proportion-échantillon $\hat{\pi}$ est un estimateur *sans biais* de la proportion-population π, que l'échantillon d'individus ait été prélevé par tirage PEAR ou PESR.

La variance de $\hat{\pi}$ est, comme c'était d'ailleurs le cas aussi pour le sondage PESR, proportionnelle au produit $\pi(1-\pi)$ et inversement proportionnelle à n. Il s'ensuit que $\hat{\pi}$ est d'autant plus précis que le nombre de prélèvements effectués est grand et que la proportion π à estimer est petite (proche de 0) ou, au contraire, élevée (proche de 1).

Notez encore que la taille N de la population n'intervient pas dans l'expression de la variance de $\hat{\pi}$. Contrairement à ce qui se passe pour le sondage PESR, le taux de sondage appliqué n'a donc aucune influence sur la précision de $\hat{\pi}$ dans le cas du sondage PEAR. Ceci peut paraître quelque peu paradoxal… mais est simplement dû au fait que les prélèvements dans la population se font SANS remise.

c) Comparaison avec le sondage PESR

Comparons à présent la variance de $\hat{\pi}$ dans le cas du tirage PEAR et dans celui du tirage PESR :

$$\mathrm{V}_{\mathrm{PESR}}(\hat{\pi}) = \frac{N-n}{N-1}\frac{\pi(1-\pi)}{n} \quad \text{et} \quad \mathrm{V}_{\mathrm{PEAR}}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}.$$

Puisque le rapport de $(N-n)$ sur $(N-1)$ est nécessairement inférieur ou égal à 1 (car $n \geq 1$), nous avons :

$$V_{PESR}(\hat{\pi}) \leq V_{PEAR}(\hat{\pi})$$

(cette inégalité est même « stricte » dès que $n \geq 2$). En terme de précision de l'estimateur $\hat{\pi}$, il est donc plus intéressant d'effectuer les n prélèvements dans la population SANS remise plutôt qu'AVEC remise !

2.4.5 L'estimation d'une moyenne-population

a) L'estimateur $\hat{\mu}$ de μ

Considérons à présent une variable d'intérêt quantitative y, de moyenne (inconnue) μ et de variance (également inconnue) σ^2 dans la *population* U.

Tout comme dans le cas du sondage PESR, l'estimateur utilisé pour estimer la moyenne-population μ dans le cas du sondage PEAR est la moyenne de la variable d'intérêt y dans l'échantillon :

$$\hat{\mu} = \bar{y}.$$

b) Les propriétés de $\hat{\mu}$

Nous l'avons vu au point 2.4.3 de ce document, on peut associer à la procédure de tirage PEAR de n individus les observations aléatoires $Y_1, \ldots, Y_n$, indépendantes et identiquement distribuées, telles que, pour tout $k = 1, \ldots, n$:

$$E(Y_k) = \mu \quad \text{et} \quad V(Y_k) = \sigma^2.$$

Dans ce contexte, l'estimateur $\hat{\mu}$ n'est autre que la moyenne arithmétique de ces observations aléatoires i.i.d. $Y_1, \ldots, Y_n$:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

On en déduit très facilement que :

$$E(\hat{\mu}) = \frac{1}{n}\sum_{k=1}^{n} E(Y_k) = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

et

$$V(\hat{\mu}) = \frac{1}{n^2}\sum_{k=1}^{n} V(Y_k) = \frac{1}{n^2}\sum_{k=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Vous retrouvez à nouveau des résultats bien connus !

La moyenne-échantillon $\hat{\mu}$ est donc un estimateur sans biais de la moyenne-population μ dans le cas du sondage PEAR comme dans celui du sondage PESR.

De manière similaire à ce que nous avions découvert dans le cas du sondage PESR, la précision de $\hat{\mu}$ dans le cas du sondage PEAR est d'autant plus grande que le nombre n de prélèvements effectués dans la population est élevé et que la population est bien homogène (c'est-à-dire que la dispersion des valeurs prises par la variable d'intérêt y dans la population est faible).

c) Comparaison avec le sondage PESR

Comparons à présent la variance de $\hat{\mu}$ dans le cas du tirage PESR et dans celui du tirage PEAR :

$$\mathrm{V}_{\mathrm{PESR}}(\hat{\mu}) = (1-f)\,\frac{\sigma^2_{\mathrm{corr}}}{n} = \Big(1-\frac{n}{N}\Big)\frac{\frac{N}{N-1}\sigma^2}{n} = \frac{N-n}{N-1}\,\frac{\sigma^2}{n}$$

et

$$\mathrm{V}_{\mathrm{PEAR}}(\hat{\mu}) = \frac{\sigma^2}{n}.$$

Puisque $N-n \leq N-1$, il est clair que

$$\mathrm{V}_{\mathrm{PESR}}(\hat{\mu}) \leq \mathrm{V}_{\mathrm{PEAR}}(\hat{\mu}).$$

Ainsi, en terme de précision de l'estimateur $\hat{\mu}$, il est plus intéressant d'effectuer les n prélèvements dans la population SANS remise plutôt qu'AVEC remise !

2.4.6 Remarque finale : PEAR versus PESR

Nous avons vu que

$$\frac{\mathrm{V}_{\mathrm{PESR}}(\hat{\pi})}{\mathrm{V}_{\mathrm{PEAR}}(\hat{\pi})} = \frac{\mathrm{V}_{\mathrm{PESR}}(\hat{\mu})}{\mathrm{V}_{\mathrm{PEAR}}(\hat{\mu})} = \frac{N-n}{N-1} \leq 1.$$

Ainsi l'échantillonnage SANS remise de n unités statistiques, outre le fait d'être plus naturel, permet aux estimateurs d'une proportion ou d'une moyenne d'être plus précis que l'échantillonnage AVEC remise. Ceci explique pourquoi, dans la pratique, les prélèvements des unités statistiques dans la population se font (presque) toujours SANS remise.

Il nous faut toutefois relativiser quelque peu l'avantage du tirage PESR de l'échantillon sur le tirage PEAR. Si la taille N de la population est fort grande et si n est négligeable par rapport à N — cette situation se rencontre dans bon nombre de sondages d'opinion — le sondage PEAR se montre quasiment aussi efficace que le sondage PESR. En effet, nous avons dans ce cas :

$$\frac{N-n}{N-1} \simeq \frac{N-n}{N} = 1-\frac{n}{N} \simeq 1$$

et donc

$$\mathrm{V}_{\mathrm{PEAR}}(\hat{\pi}) \simeq \mathrm{V}_{\mathrm{PESR}}(\hat{\pi}) \quad \text{et} \quad \mathrm{V}_{\mathrm{PEAR}}(\hat{\mu}) \simeq \mathrm{V}_{\mathrm{PESR}}(\hat{\mu}).$$

Le tableau 2.1 vous donne la valeur du rapport $\mathrm{V}_{\mathrm{PESR}}(\hat{\pi})/\mathrm{V}_{\mathrm{PEAR}}(\hat{\pi}) = (N-n)/(N-1)$ pour différentes valeurs de N et de n. Il illustre clairement les résultats que nous venons d'énoncer.

TABLEAU 2.1 – *Valeurs de* $\frac{N-n}{N-1}$

$N \rightarrow$	100	10 000	1 000 000
$n \downarrow$			
10	0,909	0,999	0,99999
100	0	0,990	0,99990
1 000	—	0,900	0,99900
10 000	—	0	0,99000

Remarque complémentaire

Jusqu'à présent, pour la procédure d'estimation de π ou de μ dans le cadre du sondage PEAR, nous avons toujours pris en compte l'ensemble des n observations associées aux n prélèvements effectués dans la population. Cela signifie que si, par exemple, un individu est sélectionné à deux reprises au cours de la procédure de tirage avec remise, la valeur que prend y chez cet individu intervient à deux reprises dans le calcul de l'estimation ; tout se passe comme si on ne remarquait pas que l'individu en question apparaissait à deux reprises dans l'échantillon.

Dans cette situation, il est assez naturel de s'interroger sur les propriétés statistiques de l'estimateur qui ne prendrait en compte qu'une et une seule fois chaque individu *distinct* sélectionné.

Supposons que les n prélèvements successifs avec remise donnent lieu à la sélection de m individus distincts, avec $m \leq n$. Désignons par S_m l'échantillon aléatoire constitué de ces m individus distincts (S_m est un sous-ensemble de taille *aléatoire* m de la population U) et définissons l'estimateur

$$\hat{\mu}_m = \frac{1}{m} \sum_{i \in S_m} y_i$$

de la moyenne-population μ. On peut montrer que :

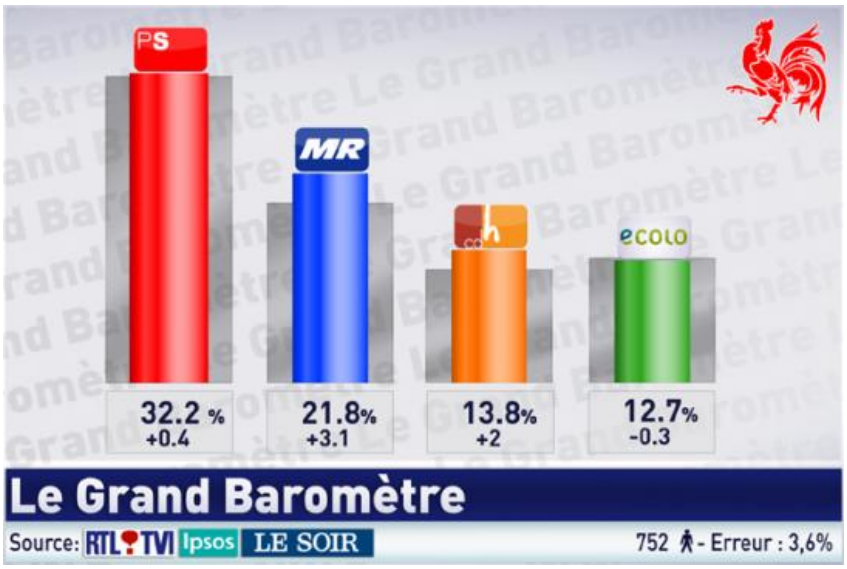
- $\hat{\mu}_m$ est un estimateur sans biais de μ ;
- $\mathrm{V}(\hat{\mu}_m) \simeq \left(\frac{1}{n} - \frac{1}{2N} + \frac{n-1}{12N^2}\right) \sigma^2_{\text{corr}}$;
- $\mathrm{V}(\hat{\mu}_m) \leq \mathrm{V}(\hat{\mu})$: dans le cas du prélèvement de n individus par sondage PEAR, il est donc toujours plus intéressant (en terme de précision de l'estimateur) de ne conserver que les unités statistiques distinctes !

2.5 L'estimation par intervalle de confiance

2.5.1 Introduction

Nous sommes régulièrement abreuvés de résultats de sondages de toutes sortes. Pensez, par exemple, à tous ces sondages d'opinion politique ou à tous ces sondages pré-électoraux qui foisonnent à l'approche d'élections. Mais comment aborder les résultats de ces sondages ? Comprenez-vous toujours l'ensemble des résultats présentés au grand public ?

Pour débuter cette section, je voudrais vous soumettre un graphique présentant quelques-uns des résultats du sondage d'opinion politique « Le Grand Baromètre » réalisé en Belgique en septembre 2012.



Ce graphique présente les pourcentages d'intention de vote pour les 4 partis politiques les plus importants en Wallonie — une des trois régions de la Belgique fédérale — et la hausse ou la baisse de ces pourcentages d'intention de vote par rapport au sondage précédent.

Dans le coin inférieur droit, sont indiqués le nombre de répondants — 752 — ainsi que la valeur d'une « erreur » — 3,6 %. Mais à quoi correspond donc cette *erreur* ?

Voici à présent la dernière page du rapport technique rédigé pour un autre sondage d'opinion politique réalisé du 30 mai au 4 juin 2012 en Belgique. Qu'y lisons-nous ?

Cette vague de 2343 répondants, formant des échantillons représentatifs des Belges de 18 ans et plus à raison de 918 en Wallonie, 925 en Flandre et 500 dans les 19 communes de la Région Bruxelles-Capitale, a été réalisée du 30 mai au 4 juin 2012.
Les interviews ont eu lieu via l'Ipsos On Line Panel.
La marge d'erreur maximale, pour un pourcentage de 50% et un taux de confiance de 95% est de 3,2% en Wallonie et en Flandre, de 4,4% à Bruxelles.
Affiliations: ESOMAR, FEBELMAR.
Le rapport technique complet a été publié sur www.febelmar.be

Mais en quoi consistent ces *marges d'erreur* ? Et ce *taux de confiance* ? Comment devons-nous les interpréter ? Comment devons-nous tenir compte des erreurs mentionnées dans notre lecture des résultats du sondage ?

Ces marges d'erreur et ce taux de confiance sont en réalité liés à l'estimation de proportions-population par ce que l'on appelle des *intervalles de confiance*.

Je vous propose de revoir dans cette section cette notion d'intervalle de confiance (dans le cadre du sondage PESR). Nous verrons comment construire un tel intervalle de confiance pour un paramètre de la population et comment interpréter correctement l'intervalle de confiance déterminé à partir de l'échantillon particulier que l'on a prélevé.

2.5.2 Objectifs

Jusqu'à présent, nous nous sommes intéressés à l'estimation dite « ponctuelle » d'une proportion, d'une moyenne ou d'un total dans une population. Pourquoi « ponctuelle » ?

Parce qu'on estime le paramètre inconnu par UNE valeur particulière, calculée à partir des observations effectuées sur les individus de l'échantillon que l'on a prélevé : si la moyenne des valeurs de la variable d'intérêt y dans l'échantillon vaut 50, par exemple, c'est par cette valeur de 50 que l'on estime la moyenne μ de y dans la population.

Dans la pratique, donc, on prélève UN échantillon particulier dans la population et cet échantillon particulier nous conduit à UNE valeur particulière pour estimer le paramètre-population.

Mais comment rendre compte du fait que, si le hasard nous avait fourni un autre échantillon, nous aurions très probablement obtenu une autre estimation du paramètre ? Comment tenir compte de ce que nous avons appelé la fluctuation d'échantillonnage ? Et comment rendre compte, de manière claire et facilement interprétable, du niveau de précision attaché au processus d'estimation ?

Une manière de faire consiste à estimer le paramètre-population non plus par UNE valeur particulière, mais plutôt par une FOURCHETTE ou un INTERVALLE de valeurs, intervalle construit sur la base de notre connaissance (au moins théorique) de la distribution d'échantillonnage de l'estimateur du paramètre.

2.5.3 Définition et construction

a) Définition

Désignons de manière générale par la lettre grecque θ le paramètre-population que nous voulons estimer : selon le cas, θ correspondra à une proportion-population π, à une moyenne-population μ ou à un total-population τ.

Un **intervalle de confiance** pour le paramètre θ est un intervalle dont les bornes L^- et L^+ sont déterminées à partir des valeurs observées pour la variable d'intérêt y dans l'échantillon prélevé, et qui a une probabilité très élevée de contenir la valeur inconnue du paramètre θ. Cette probabilité, appelée « *niveau de confiance* », est le plus souvent choisie égale à 95% ; les autres valeurs usuelles du niveau de confiance sont 90% ou 99%.

Ainsi, l'intervalle $[L^-, L^+]$ est un intervalle de confiance pour le paramètre θ au niveau de confiance de 95% si la probabilité que cet intervalle contienne la valeur exacte de θ s'élève à 95% :

$$\mathrm{P}(L^- \leq \theta \leq L^+) = 95\%.$$

b) Construction

Pour construire un intervalle de confiance pour le paramètre-population θ, nous allons partir de la *distribution d'échantillonnage* d'un estimateur *sans biais* $\hat{\theta}$ de θ.

En pratique, il nous est généralement impossible de déterminer la distribution d'échantillonnage *exacte* de $\hat{\theta}$: on en connaît juste certaines caractéristiques. On sait que $\hat{\theta}$ est sans biais, autrement dit que la moyenne de sa distribution d'échantillonnage coïncide avec la valeur exacte (mais inconnue) de θ ; on dispose également d'une expression générale pour la variance de $\hat{\theta}$.

Mais la théorie statistique[1] nous fournit en réalité des informations supplémentaires sur la forme de la distribution d'échantillonnage de $\hat{\theta}$. On peut en effet montrer que, s'il n'y a « pas trop » d'individus atypiques dans la population, et si les tailles N de la population et n de l'échantillon sont « grandes », la distribution d'échantillonnage de $\hat{\theta}$ correspond en bonne approximation à une distribution dite « normale » (ou loi de Gauss) de moyenne égale à l'espérance de $\hat{\theta}$, donc θ, et de variance égale à la variance de $\hat{\theta}$:

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \mathrm{V}(\hat{\theta})\right).$$

Nous restons volontairement très vague ici sur ce que recouvre l'expression « pas trop d'individus atypiques », car la formulation rigoureuse de cette condition est très complexe. Quant à la taille de l'échantillon requise, disons que si n est supérieur ou égal à 50, on peut abandonner tout scrupule à assimiler la distribution d'échantillonnage de $\hat{\theta}$ à une distribution normale.

Ainsi, si n est supérieur ou égal à 50, on peut considérer que l'estimateur $\hat{\theta}$ du paramètre θ est approximativement distribué selon une loi normale de moyenne égale à θ et de variance égale à $\mathrm{V}(\hat{\theta})$. Par ailleurs, on sait comment estimer $\mathrm{V}(\hat{\theta})$, et, si n est

[1] On fait en réalité appel ici à un théorème central limite pour population *finie*.

raisonnablement grand, on peut s'attendre à ce que l'échantillon prélevé nous fournisse une estimation proche de la valeur exacte de cette variance. Ceci nous permet alors de considérer qu'en bonne approximation toujours, l'estimateur $\hat{\theta}$ est distribué selon une loi normale de moyenne égale à θ et de variance égale à $\hat{\mathbb{V}}(\hat{\theta})$:

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \hat{\mathbb{V}}(\hat{\theta})\right).$$

En d'autres termes encore, $\hat{\theta}$ moins sa moyenne θ, et divisé par la racine carrée de $\hat{\mathbb{V}}(\hat{\theta})$, est, en bonne approximation, distribué selon une loi normale de moyenne nulle et de variance égale à 1 :

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\mathbb{V}}(\hat{\theta})}} \approx \mathcal{N}(0,1).$$

Or, une variable distribuée selon la loi normale $\mathcal{N}(0,1)$ a 95 chances sur 100 de prendre une valeur entre $-1{,}96$ et $+1{,}96$ (voir la figure 2.3). Nous pouvons donc écrire :

$$\mathrm{P}\left(-1{,}96 \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{\mathbb{V}}(\hat{\theta})}} \leq 1{,}96\right) = 0{,}95 = 95\%.$$

Il découle de ce dernier résultat que :

$$\mathrm{P}\left(\hat{\theta} - 1{,}96\sqrt{\hat{\mathbb{V}}(\hat{\theta})} \leq \theta \leq \hat{\theta} + 1{,}96\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\right) = 0{,}95 = 95\%.$$

Ceci nous indique que l'intervalle

$$\left[\hat{\theta} - 1{,}96\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\,;\hat{\theta} + 1{,}96\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\right]$$

est un intervalle de confiance pour le paramètre θ au niveau de confiance de 95%.

Si l'on préfère travailler à un niveau de confiance de 90% seulement, la valeur de 1,96 doit être remplacée par 1,64 ; pour un niveau de confiance de 99%, la valeur de 1,96 doit être remplacée par 2,58 (voir la figure 2.3) :

- intervalle de confiance pour θ au niveau de confiance de 90% :

$$\left[\hat{\theta} - 1{,}64\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\,;\hat{\theta} + 1{,}64\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\right];$$

- intervalle de confiance pour θ au niveau de confiance de 99% :

$$\left[\hat{\theta} - 2{,}58\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\,;\hat{\theta} + 2{,}58\sqrt{\hat{\mathbb{V}}(\hat{\theta})}\right].$$

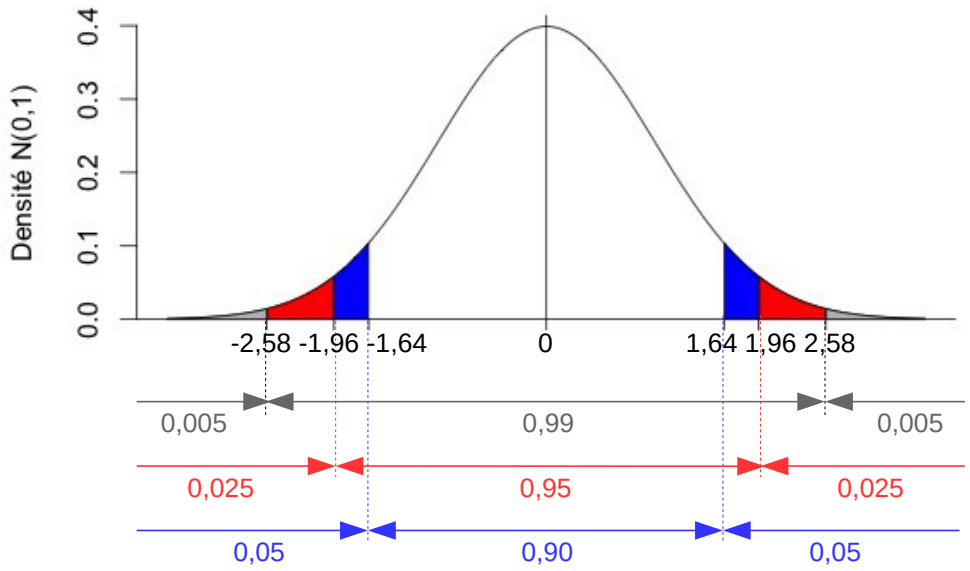


FIGURE 2.3 – *Distribution* $\mathcal{N}(0,1)$

Il suffit à présent de remplacer le paramètre général θ par π, μ ou τ pour obtenir les expressions des intervalles de confiance pour une proportion, une moyenne ou un total de la population :

- intervalle de confiance pour π au niveau de confiance de 95% :
$$\left[\hat{\pi} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\pi})}\right] \quad \text{où} \quad \widehat{\mathrm{V}}(\hat{\pi}) = (1-f)\frac{\hat{\pi}(1-\hat{\pi})}{n-1} \; ;$$
- intervalle de confiance pour μ au niveau de confiance de 95% :
$$\left[\hat{\mu} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\mu})}\right] \quad \text{où} \quad \widehat{\mathrm{V}}(\hat{\mu}) = (1-f)\frac{s^2_{\mathrm{corr}}}{n} \; ;$$
- intervalle de confiance pour τ au niveau de confiance de 95% :
$$\left[\hat{\tau} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\tau})}\right] \quad \text{où} \quad \widehat{\mathrm{V}}(\hat{\tau}) = N^2(1-f)\frac{s^2_{\mathrm{corr}}}{n}.$$

L'intervalle de confiance est centré en l'estimateur du paramètre considéré et ses bornes sont définies en retranchant et en rajoutant à cet estimateur 1,96 fois la racine carrée de la variance estimée de l'estimateur (voir la section 2.3 pour les expressions de $\widehat{\mathrm{V}}(\hat{\pi})$, $\widehat{\mathrm{V}}(\hat{\mu})$ et $\widehat{\mathrm{V}}(\hat{\tau})$ dans le cadre du sondage PESR).

2.5.4 Interprétation

Reprenons l'expression des bornes de l'intervalle de confiance pour un paramètre θ, au niveau de confiance de 95% : elles sont obtenues en retranchant et en ajoutant à $\hat{\theta}$ 1,96 fois la racine carrée de la variance estimée de $\hat{\theta}$.

Puisque les valeurs de $\hat{\theta}$ et de $\widehat{\mathrm{V}}(\hat{\theta})$ varient d'un échantillon possible à l'autre, les bornes de l'intervalle de confiance voient elles aussi leurs valeurs varier d'un échantillon à l'autre.

Ainsi, à chaque échantillon possible correspond une réalisation différente de l'intervalle de confiance. Le fait de considérer un niveau de confiance de 95% nous assure que, parmi tous les échantillons aléatoires simples possibles de taille n, 95% d'entre eux fournissent des réalisations de l'intervalle qui contiennent bien la valeur exacte du paramètre θ considéré et seulement 5% d'entre eux donnent des réalisations de l'intervalle qui ne recouvrent pas la valeur exacte de θ.

Dès lors, en pratique, lorsqu'on tire UN échantillon pour estimer θ, nous pouvons nous dire qu'il y a 95 chances sur 100 que cet échantillon particulier que le hasard nous a fourni, fasse partie des « bons » échantillons, c'est-à-dire donne lieu à une réalisation de l'intervalle de confiance qui contient bien la valeur exacte de θ.

2.5.5 L'effet du niveau de confiance

Le niveau de confiance correspond donc au *degré de certitude* que l'on peut avoir dans le fait que l'intervalle que nous a fourni l'échantillon particulier que l'on a prélevé recouvre bien la valeur exacte du paramètre.

Comme déjà mentionné, 95% est la valeur la plus fréquemment considérée pour ce niveau de confiance. On travaille donc le plus souvent avec l'intervalle de confiance que voici pour le paramètre θ :

$$\left[\hat{\theta} \pm 1{,}96\sqrt{\widehat{\mathbb{V}}(\hat{\theta})}\right].$$

Mais si vous êtes plus exigeant quant au degré de certitude à atteindre, et que vous décidez ainsi de fixer le niveau de confiance à 99%, cela aura pour effet d'élargir l'intervalle, puisque la valeur de 1,96 apparaissant dans l'expression de ses bornes est remplacée par la valeur de 2,58.

Si vous acceptez un niveau de confiance plus faible, de 90% par exemple, vous obtiendrez un intervalle plus étroit, puisque cette fois c'est la valeur de 1,64 qui doit être prise en compte dans l'expression de ses bornes.

Nous pouvons en conclure que plus le niveau de confiance que l'on se fixe est élevé, plus l'intervalle de confiance est large ; et plus le niveau de confiance considéré est faible, plus l'intervalle de confiance est étroit.

Et oui ! Il n'y a pas de miracle en statistique ! Pour être pratiquement certain — à 99%, disons — que l'intervalle de confiance contienne bien la valeur exacte du paramètre θ, on est obligé de considérer un intervalle couvrant une fourchette relativement large de valeurs *a priori* possibles pour θ.

A l'inverse, si vous vous fixez un niveau de confiance plus faible — disons de 90% — et que vous acceptez donc un risque non négligeable — de 10% — d'avoir un intervalle qui ne contienne pas la valeur exacte de θ, vous pouvez considérer un intervalle couvrant une fourchette plus étroite de valeurs *a priori* possibles pour θ.

En conclusion, donner une estimation par intervalle de confiance d'un paramètre-population est doublement prudent.

D'une part, on ne fournit pas une valeur PONCTUELLE comme valeur plausible pour le paramètre, mais bien une PLAGE de valeurs plausibles, ce qui rend bien compte de la fluctuation d'échantillonnage liée au processus d'estimation.

D'autre part, on prévient qu'il existe un risque faible que la vraie valeur du paramètre soit en dehors de la fourchette donnée, ce qui rend honnêtement compte des limitations du processus d'estimation.

2.5.6 Exercice 2.3

Objectif – Cet exercice a pour objectif de vous permettre de vérifier que vous êtes capable de calculer des intervalles de confiance pour une proportion, pour une moyenne ou pour un total, à différents niveaux de confiance, et que vous êtes à même d'en proposer des interprétations correctes.

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé (dans un fichier Excel ainsi que dans un fichier pdf) de cet exercice sur l'UV.

La population qui nous intéresse est constituée de 2 390 individus âgés de 18 ans ou plus et exerçant tous une certaine activité professionnelle.

On souhaite estimer :

- la proportion d'individus de la population qui ont changé d'employeur à au moins une reprise au cours des trois dernières années ;
- le nombre d'individus de la population ayant 10 ans ou plus d'expérience professionnelle ;
- le nombre hebdomadaire moyen d'heures de travail des individus de la population ;
- le revenu hebdomadaire moyen des individus de la population ;
- le revenu horaire moyen des individus de la population.

Pour ce faire, on a sélectionné par sondage aléatoire simple un échantillon de 200 personnes. Le fichier Excel « Data_ex_2_3 » disponible sur l'UV reprend les valeurs observées auprès de ces 200 personnes pour les variables suivantes :

- Chgt : variable prenant la valeur 1 chez un individu ayant changé d'employeur à au moins une reprise au cours des trois dernières années, et prenant la valeur 0 sinon ;
- Exp : variable indiquant le nombre d'années d'expérience professionnelle d'un individu ;
- NbrHebd : variable indiquant le nombre hebdomadaire d'heures de travail d'un individu ;
- RevHebd : variable indiquant le revenu hebdomadaire (en euros) d'un individu.

Partie 1

1. Donnez une estimation (ponctuelle) de la proportion d'individus de la population qui ont changé d'employeur à au moins une reprise au cours des trois dernières années.

2. Déterminez l'intervalle de confiance pour cette proportion, au niveau de confiance de 95%.

3. Cochez la (les) proposition(s) qui exprime(nt) correctement la conclusion que l'on peut tirer des valeurs des bornes de l'intervalle de confiance que vous venez de calculer.

- ○ On peut affirmer que la proportion π_{Chgt} d'individus de la population qui ont changé d'employeur à au moins une reprise au cours des trois dernières années a une valeur qui appartient à l'intervalle calculé.
- ○ Il y a 95 chances sur 100 pour que l'intervalle calculé contienne bien la valeur de la proportion π_{Chgt} d'individus de la population qui ont changé d'employeur à au moins une reprise au cours des trois dernières années.
- ○ Si l'on sélectionne par tirage PESR un échantillon de taille 200 dans la population étudiée, on peut être sûr à 95% d'obtenir une estimation de la proportion π_{Chgt} qui appartienne à l'intervalle de confiance calculé.
- ○ 95% des valeurs possibles pour l'estimateur de π_{Chgt}, dans le cadre du sondage aléatoire simple considéré, appartiennent à l'intervalle calculé.

4. Parmi les propositions suivantes, cochez celle(s) qui est (sont) correcte(s).

- ○ Si nous avions calculé l'intervalle de confiance au niveau de confiance de 99%, nous aurions obtenu un intervalle plus large que celui correspondant au niveau de confiance de 95%.
- ○ L'intervalle de confiance au niveau de confiance de 90% est plus étroit que celui au niveau de confiance de 95%, lui-même plus étroit que celui au niveau de confiance de 99%. Il est donc vivement conseillé de travailler au niveau de confiance de 90%, puisque c'est ce niveau qui donne lieu à la plus grande précision pour l'estimation.
- ○ Plus le niveau de confiance est élevé, plus le niveau de précision de l'intervalle de confiance est élevé et donc plus l'intervalle de confiance est étroit.
- ○ Plus le niveau de confiance est élevé, plus il y a de chance que l'intervalle de confiance calculé recouvre bien la valeur exacte du paramètre à estimer.

Partie 2

1. Donnez une estimation (ponctuelle) du nombre d'individus de la population qui ont 10 ans ou plus d'expérience professionnelle.

2. Déterminez l'intervalle de confiance pour ce nombre d'individus, au niveau de confiance de 90%.

Partie 3

1. Donnez une estimation (ponctuelle) du nombre hebdomadaire moyen d'heures de travail des individus de la population.

2. Déterminez l'intervalle de confiance pour cette moyenne, au niveau de confiance de 95%.

Partie 4

1. Donnez une estimation (ponctuelle) du revenu hebdomadaire moyen des individus de la population.

2. Déterminez l'intervalle de confiance pour cette moyenne, au niveau de confiance de 95%.

Partie 5

1. Donnez une estimation (ponctuelle) du revenu horaire moyen des individus de la population.

2. Déterminez l'intervalle de confiance pour cette moyenne, au niveau de confiance de 95%.

2.6 Les incertitudes absolue et relative

2.6.1 Définitions

Deux notions fondamentales en théorie des sondages sont directement liées à l'intervalle de confiance pour un paramètre au niveau de confiance de 95% : il s'agit des notions d'*incertitude absolue* et d'*incertitude relative*. Comme nous allons le voir, ces notions jouent un rôle particulièrement important car elles permettent de quantifier la précision de la procédure d'estimation du paramètre θ de manière beaucoup plus intuitive que ne le permet la valeur estimée de la variance de l'estimateur $\hat{\theta}$ de θ.

L'intervalle de confiance pour θ, au niveau de confiance de 95%, est donné par :

$$\left[\hat{\theta} \pm 1{,}96\sqrt{\hat{\mathrm{V}}(\hat{\theta})}\right].$$

Cet intervalle est centré en $\hat{\theta}$ et ses bornes inférieure et supérieure sont obtenues en retranchant et en ajoutant à $\hat{\theta}$ la quantité $1{,}96\sqrt{\hat{\mathrm{V}}(\hat{\theta})}$. Cette dernière quantité n'est autre que la *demi-largeur* de l'intervalle de confiance : elle constitue ce qu'on appelle l'« **incertitude (ou erreur) absolue** » ou, plus communément, la « **marge d'erreur** ». Nous la désignerons de façon générale par la lettre d :

$$d = 1{,}96\sqrt{\hat{\mathrm{V}}(\hat{\theta})}.$$

On peut aussi définir l'**incertitude (ou erreur) relative** comme étant le rapport

$$d/\hat{\theta}\ ,$$

c'est-à-dire la demi-largeur de l'intervalle de confiance rapportée à la valeur centrale de cet intervalle. L'incertitude relative s'exprime généralement en pourcents.

L'incertitude absolue dépend de la variance de l'estimateur du paramètre considéré, au travers de l'estimation qu'on a pu faire de cette variance : plus la variance de l'estimateur est élevée, autrement dit, moins l'estimateur est précis, plus l'incertitude absolue est grande, plus l'intervalle de confiance est large. L'incertitude absolue peut donc se voir comme une nouvelle façon de quantifier la précision liée à l'estimation du paramètre.

L'incertitude absolue étant directement liée à la largeur de l'intervalle de confiance, elle constitue une mesure de précision aisément interprétable, ce qui n'était pas le cas de la valeur estimée de la variance de l'estimateur.

2.6.2 Pour l'estimation d'une proportion

a) Incertitude absolue pour l'estimation d'une proportion

L'incertitude absolue (ou marge d'erreur) pour l'estimation d'une proportion π est donnée par :

$$d = 1{,}96\sqrt{\hat{\mathrm{V}}(\hat{\pi})} = 1{,}96\sqrt{(1-f)\frac{\hat{\pi}(1-\hat{\pi})}{n-1}}.$$

Elle dépend de la valeur même de la proportion-population, au travers de son estimation $\hat{\pi}$; elle dépend également de la taille n de l'échantillon, ainsi que du taux de sondage $f = n/N$ appliqué.

Dans bien des cas, n est grand, mais malgré tout négligeable devant la taille N de la population, ce qui fait que le taux de sondage f est extrêmement faible : dans ce cas, $(n-1)$ est à peu de chose près égal à n et $(1-f)$ à 1. L'incertitude absolue est alors approximativement égale à :

$$d \simeq 1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Si on arrondit la valeur 1,96 à 2 et puisque le produit $\hat{\pi}(1-\hat{\pi})$ atteint sa valeur maximale de un quart lorsque $\hat{\pi}$ vaut un demi, nous pouvons écrire :

$$d \simeq 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$$\leq 2\sqrt{\frac{1}{4n}} = \frac{1}{\sqrt{n}}.$$

L'incertitude absolue (ou marge d'erreur) pour l'estimation d'une proportion-population vaut donc au maximum $1/\sqrt{n}$.

TABLEAU 2.2 – *Valeur de la marge d'erreur maximale $1/\sqrt{n}$ pour l'estimation de π au niveau de confiance de 95%*

n	$1/\sqrt{n}$
100	0,10 = 10%
400	0,05 = 5%
1 000	0,03 = 3%
1 600	0,025 = 2,5%
10 000	0,01 = 1%

Le tableau 2.2 ci-dessus vous donne cette **incertitude absolue maximale** de $1/\sqrt{n}$ pour différentes valeurs de n. Pour une taille d'échantillon égale à 100, l'incertitude absolue peut atteindre 0,10, soit 10 points (de pourcents) si la proportion estimée est proche de un demi ! Pour être assuré que l'incertitude absolue ne dépasse pas 0,03 ou 3 points de pourcents, il faut travailler avec un échantillon de taille supérieure ou égale à 1 000.

Remarque

Vous remarquerez qu'on exprime l'incertitude absolue associée à l'estimation d'une proportion en **points** ou **points de pourcents**, plutôt que simplement en pourcents. Pourquoi cela ? Tout simplement parce que l'incertitude absolue d n'est pas de la même nature que l'estimation de π ou que l'incertitude relative. Si $\hat{\pi}$ et $d/\hat{\pi}$ sont par définition des *taux* ($\hat{\pi}$ est une *proportion* et $d/\hat{\pi}$ est un *rapport* de deux nombres) et peuvent dès lors s'exprimer de manière naturelle sous la forme d'un pourcentage, la marge d'erreur

d est quant à elle un *écart* entre des proportions. Le fait d'exprimer l'incertitude absolue en points de pourcents et l'incertitude relative en pourcents permet d'éviter toute confusion entre ces deux types d'erreurs.

Pour conclure

Vous avez à présent toutes les clés pour comprendre les deux extraits de résultats de sondage politique présentés en introduction à la section 5 pour entamer l'étude des intervalles de confiance.

Dans le premier extrait, on vous indique le nombre de répondants à l'enquête : 752. Le rapport de 1 sur la racine carrée de 752 est égal aux 3,6% indiqués comme valeur de l'erreur.

Si l'on peut raisonnablement considérer que l'ensemble des 752 répondants constitue un sous-ensemble du corps électoral assimilable à un échantillon qui aurait été sélectionné par sondage PESR, ces 3,6% correspondent à la marge d'erreur maximale associée à l'estimation d'une proportion par intervalle de confiance et au niveau de confiance de 95%.

Dans le second extrait, on vous parle de 918 répondants en Wallonie, 925 en Flandre et 500 en Région de Bruxelles-Capitale : avec de telles tailles d'échantillons de répondants, on obtient des marges d'erreur maximales pour une proportion égales :

- à $1/\sqrt{918}$, c'est-à-dire approximativement 3,2 points de pourcents en Wallonie,
- à $1/\sqrt{925}$, soit également à peu près 3,2 points de pourcents en Flandre,
- et à $1/\sqrt{500}$, soit approximativement 4,4 points de pourcents en Région de Bruxelles-Capitale.

Notez que l'on fait à nouveau ici l'hypothèse que l'on peut assimiler les ensembles de répondants à des échantillons aléatoires simples. Nous reparlerons de cette problématique lorsque nous étudierons la méthode de sondage par quotas dans le chapitre 8 de ce cours. C'est en effet cette méthode de sondage à choix raisonné (donc non aléatoire) qui est presque toujours utilisée pour les sondages d'opinion politique.

b) Incertitude relative pour l'estimation d'une proportion

L'incertitude relative associée à l'estimation d'une proportion-population π est donnée par :

$$\frac{d}{\hat{\pi}} = \frac{1{,}96\sqrt{(1-f)\dfrac{\hat{\pi}(1-\hat{\pi})}{n-1}}}{\hat{\pi}}.$$

Si n est grand et f très petit, cette incertitude relative est approximativement égale à :

$$\frac{2\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}}{\hat{\pi}} = 2\sqrt{\frac{1-\hat{\pi}}{n\hat{\pi}}}.$$

Le tableau 2.3 vous montre comment évolue cette incertitude relative en fonction de la taille n de l'échantillon et de la valeur estimée de π. Si la proportion à estimer est faible, il faut une taille d'échantillon très grande pour que l'incertitude relative ne dépasse pas 10%, par exemple.

TABLEAU 2.3 – *Incertitude relative* $2\sqrt{\frac{1-\hat{\pi}}{n\hat{\pi}}}$ *pour l'estimation de* π *au niveau de confiance de 95%* (n *grand mais* f *négligeable*)

$\hat{\pi} \rightarrow$ $n \downarrow$	0,10	0,20	0,30	0,40	0,50
100	60 %	40 %	31 %	24 %	20 %
200	42 %	28 %	21 %	18 %	14 %
300	34 %	23 %	17 %	14 %	12 %
500	26 %	19 %	15 %	12 %	10 %
1 000	18 %	13 %	9 %	8 %	6 %
2 000	14 %	9 %	7 %	6 %	4 %
5 000	8 %	6 %	4 %	4 %	3 %
10 000	6 %	4 %	3 %	3 %	2 %

2.7 Le choix de la taille de l'échantillon

2.7.1 La problématique

Lorsqu'on doit mener une enquête par sondage, une des premières questions que l'on se pose est : « Quelle doit être la taille de l'échantillon pour que l'on puisse considérer notre étude par sondage comme valable ? ».

Pour pouvoir répondre à cette question, il nous faut tout d'abord préciser ce que l'on entend par « valable ». Etant donné que l'objectif d'une enquête par sondage est avant tout d'estimer différents paramètres de la population, nous pouvons nous mettre d'accord sur le fait qu'une étude par sondage est « valable » dès le moment où elle nous garantit une précision acceptable pour nos estimations.

Au vu des formules de variances des estimateurs et des expressions de l'incertitude absolue, il est évident que plus la taille n de l'échantillon est grande, plus le sondage est précis. Il nous faut cependant garder à l'esprit que nous sommes toujours soumis à certaines contraintes budgétaires et matérielles, susceptibles de limiter le nombre d'individus ou d'unités statistiques que l'on va pouvoir enquêter. En pratique, la question de la taille de l'échantillon doit souvent être reformulée de la manière suivante : « Quel budget faudrait-il consacrer à l'enquête pour garantir une précision acceptable ? »

Même formulée dans ces termes, la question posée n'a pas de réponse toute faite. Il faut d'abord définir ce qu'on entend par « précision acceptable ».

On peut convenir d'une largeur maximale tolérée pour l'intervalle de confiance au niveau de confiance de 95%, c'est-à-dire fixer une borne supérieure pour l'incertitude absolue. On peut aussi, dans certains cas, préférer se fixer une borne supérieure pour l'incertitude relative.

2.7.2 Pour l'estimation d'une moyenne

a) Contrôle de l'incertitude absolue

Considérons la situation où l'on doit estimer une moyenne-population μ et où l'on désire que l'incertitude absolue (ou marge d'erreur) d ne dépasse pas la valeur d_0 fixée *a priori*.

Il nous faut rechercher quelle valeur n donner à la taille de l'échantillon pour que l'inégalité suivante soit satisfaite :

$$d = 1{,}96\sqrt{(1-f)\frac{s_{\text{corr}}^2}{n}} = 1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{s_{\text{corr}}^2}{n}} \leq d_0\,.$$

Si on arrondit 1,96 à 2, la résolution de cette inéquation nous conduit à prendre

$$n \geq n_{\min}$$

où

$$n_{\min} = \frac{N}{1 + (Nd_0^2)/(4s_{\text{corr}}^2)}.$$

$n_{\min}$ correspond à la taille *minimale* que doit avoir l'échantillon si l'on souhaite que l'incertitude absolue n'excède pas d_0.

Le problème avec ce résultat est que l'on ne peut pas calculer $n_{\min}$ puisque l'on ne connaît pas la valeur de s_{corr}^2 étant donné que l'on n'a pas encore prélevé l'échantillon ! (Nous sommes seulement en train de réfléchir à la taille qu'il devrait avoir !) Il n'y a qu'une seule solution à ce problème : donner une valeur *a priori* à s_{corr}^2, soit à partir de conseils d'« experts », soit à partir d'une autre enquête réalisée dans un passé pas trop éloigné et portant sur la variable d'intérêt $\mathcal{Y}$ ou sur une autre variable $\mathcal{Z}$ bien corrélée à $\mathcal{Y}$ (on recherche alors la taille d'échantillon permettant d'estimer la moyenne-population de $\mathcal{Z}$ avec la précision souhaitée, en faisant l'hypothèse que, grâce à la corrélation entre les variables $\mathcal{Z}$ et $\mathcal{Y}$, ce qui est bon pour l'une doit l'être pour l'autre). On peut également envisager d'enquêter quelques dizaines ou centaines d'individus afin d'obtenir une première valeur pour s_{corr}^2, et de compléter ensuite l'échantillon en fonction de la taille $n_{\min}$ à atteindre.

Illustrons nos propos à l'aide d'un exemple.

Exemple

On s'intéresse à l'ensemble des 2 010 exploitations d'une certaine région rurale et on souhaite y estimer la surface μ cultivée en moyenne par chaque exploitation.

Un premier échantillon de 100 exploitations agricoles prélevé par sondage PESR nous fournit comme estimation de μ la valeur $\hat{\mu}$ égale à 30,09 hectares. Dans ce même échantillon, la variance corrigée des surfaces cultivées vaut 93,29 (hectares^2).

Nous obtenons ainsi comme intervalle de confiance pour μ, au niveau de confiance de 95%, l'intervalle

$$\left[30{,}09 \pm 1{,}96\sqrt{\left(1 - \frac{100}{2\,010}\right)\frac{93{,}29}{100}}\right] = [30{,}09 \pm 1{,}85] = [28{,}24\,;\,31{,}94] \text{ (hectares)}.$$

L'incertitude absolue, soit la demi-largeur de cet intervalle de confiance, s'élève donc à 1,85 hectares.

Supposons à présent que nous voulions limiter l'incertitude absolue à 1 hectare. Quelle taille d'échantillon devons-nous considérer pour atteindre cet objectif de précision ?

Nous devons prendre une taille n d'échantillon supérieure ou égale à $n_{\min}$, où $n_{\min}$ est donné par :

$$n_{\min} = \frac{N}{1 + (Nd_0^2/4s_{\text{corr}}^2)} = \frac{2\,010}{1 + (2\,010\times 1)/(4\times 93{,}29)}$$
$$= 314{,}73 \simeq 315.$$

Pour que la marge d'erreur ne dépasse pas 1 hectare, il nous faut donc prendre une taille d'échantillon au moins égale à 315.

Vous le voyez, augmenter la précision n'est pas sans coûts ! Pour réduire la marge d'erreur de 1,85 à 1 hectare, il nous faut passer d'un échantillon de 100 exploitations agricoles à un échantillon de plus de 300 exploitations !

b) Exercice 2.4

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de déterminer la taille d'échantillon nécessaire pour estimer une moyenne-population avec un niveau de précision fixé.

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

On souhaite estimer le salaire mensuel net moyen dans une entreprise constituée de 1 000 employés.

Pour ce faire, on prélève un échantillon de 50 employés par tirage PESR (à probabilités égales et sans remise). La moyenne des salaires mensuels nets de ces 50 employés vaut 1 100 euros ; la variance corrigée de leurs salaires mensuels nets s'élève à 62 500.

1. Déterminez l'intervalle de confiance pour le salaire mensuel net moyen des employés de l'entreprise, au niveau de confiance de 95%.

2. Déterminez la marge d'erreur (ou incertitude absolue) associée à cette estimation.

3. La marge d'erreur obtenue à la question 2 est jugée comme étant trop élevée par les responsables du sondage. Ils disposent heureusement d'un certain budget pour compléter l'échantillon déjà tiré par de nouvelles observations de manière à réduire la marge d'erreur. Quelle taille devrait avoir l'échantillon final si l'on souhaite que la marge d'erreur n'excède pas 25 euros ?

2.7.3 Pour l'estimation d'une proportion

a) Contrôle de l'incertitude absolue

On peut suivre la même démarche que celle présentée dans le cadre de l'estimation d'une moyenne pour déterminer la taille d'échantillon minimale nous permettant d'estimer une proportion-population π avec une marge d'erreur d'au plus d_0. Il nous faut résoudre l'inégalité suivante :

$$d = 1{,}96\sqrt{(1-f)\frac{\hat{\pi}(1-\hat{\pi})}{n-1}} = 1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{\pi}(1-\hat{\pi})}{n-1}} \leq d_0\,.$$

En approximant l'incertitude absolue d par $2\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, on trouve qu'il faut prendre

$$n \geq n_{\min},$$

avec

$$n_{\min} = \frac{N}{1 + (N d_0^2)/\left(4\hat{\pi}(1-\hat{\pi})\right)}.$$

Notons que si la taille N de la population est grande,

$$n_{\min} \simeq \frac{4\hat{\pi}(1-\hat{\pi})}{d_0^2}.$$

A nouveau, la détermination de $n_{\min}$ nécessite de donner *a priori* une valeur à $\hat{\pi}$, en fonction de l'ordre de grandeur de la proportion-population π que l'on a en tête (on a en effet généralement une idée assez précise de l'ordre de grandeur de π dès que l'on a une certaine connaissance sur le sujet ou que l'on consulte un expert de la problématique considérée).

Dans le tableau 2.4, on peut lire la taille d'échantillon nécessaire pour assurer une incertitude absolue égale à d_0 sur la proportion π que l'on cherche à estimer : d_0 se lit en ligne, $\hat{\pi}$ en colonne et $n_{\min}$ à l'intersection ligne-colonne. Ainsi, par exemple, si vous souhaitez une incertitude absolue de l'ordre de 1 point de pourcent alors que $\hat{\pi}$ devrait tourner aux alentours de 30%, vous devez prévoir de prélever un échantillon d'au moins 8 400 individus ou unités statistiques.

TABLEAU 2.4 – *Taille d'échantillon* $n_{\min} \simeq \frac{4\hat{\pi}(1-\hat{\pi})}{d_0^2}$ *nécessaire pour que la marge d'erreur pour l'estimation de* π *ne dépasse pas* d_0 *(N grand)*

$\hat{\pi} \rightarrow$ $d_0 \downarrow$	0,05 ou 0,95	0,10 ou 0,90	0,20 ou 0,80	0,30 ou 0,70	0,40 ou 0,60	0,50
0,005	7 600	14 400	25 600	33 600	38 400	40 000
0,01	1 900	3 600	6 400	8 400	9 600	10 000
0,02	475	900	1 600	2 100	2 400	2 500
0,03	211	400	711	933	1 066	1 111
0,04	119	225	400	525	600	625
0,05	76	144	256	336	384	400

Dans le cas où l'on n'a vraiment aucune idée préalable sur la valeur de π et donc de $\hat{\pi}$, on peut tenir le raisonnement suivant :

$$n_{\min} \simeq \frac{4\hat{\pi}(1-\hat{\pi})}{d_0^2} \leq \frac{4(1/4)}{d_0^2} = \frac{1}{d_0^2}.$$

Dès lors, si l'on ne sait vraiment pas par quelle valeur remplacer $\hat{\pi}$ dans l'expression de $n_{\min}$, on peut faire preuve de prudence en décidant d'adopter $1/d_0^2$ comme taille d'échantillon. Nous désignerons cette dernière taille par $n^*_{\min}$.

Le tableau 2.5 vous donne la taille $n^*_{\min}$ de l'échantillon qu'il est nécessaire de considérer si vous voulez pouvoir estimer la proportion-population π avec une marge d'erreur d'au plus d_0, et cela quelle que soit l'ordre de grandeur de la valeur exacte de π. On y voit

clairement que si l'on veut réduire la marge d'erreur de moitié, il faut être prêt à multiplier par 4 la taille de l'échantillon. La précision a donc un prix non négligeable !

TABLEAU 2.5 – *Taille d'échantillon $n^*_{min} = 1/d_0^2$ nécessaire pour que la marge d'erreur pour l'estimation de π ne dépasse pas d_0, quelle que soit la valeur réelle de π (N grand)*

d_0	n^*_{min}
0,10	100
0,08	156
0,06	278
0,05	400
0,04	625
0,03	1 111
0,02	2 500
0,01	10 000
0,005	40 000

Illustrons ces résultats à l'aide d'un nouvel exemple.

Exemple

Supposons que la population-cible compte $N = 10\,000\,000$ d'individus et qu'on souhaite y estimer la proportion π de personnes qui soutiennent un certain programme de réformes sociales.

On prélève par sondage PESR un échantillon de taille n égale à 1 000 ; 20% des personnes de cet échantillon se déclarent en faveur du programme. On a donc $\hat{\pi}$ égal à 0,20.

La variance estimée de $\hat{\pi}$ est égale à :

$$(1-f)\frac{\hat{\pi}(1-\hat{\pi})}{n-1} = \left(1 - \frac{1\,000}{10\,000\,000}\right)\frac{(0{,}20)(1-0{,}20)}{999} = 0{,}00016\,.$$

L'intervalle de confiance pour π au niveau de confiance de 95% a donc pour bornes :

$$\left[0{,}20 \pm 1{,}96\sqrt{0{,}00016}\right] = [0{,}20 \pm 0{,}025] = [0{,}175\,;\,0{,}225].$$

Il lui correspond une incertitude absolue de 0,025, soit 2,5 points de pourcents, et une incertitude relative de 0,025 divisé par 0,20, soit 12,5%.

Quelle taille d'échantillon aurait-il fallu prendre pour que l'incertitude absolue ne dépasse pas 1 point de pourcent, c'est-à-dire 0,01 ?

Reprenons le tableau 2.4. En considérant $\hat{\pi}$ égal à 0,20 comme estimation préliminaire de la proportion-population π, le tableau 2.4 nous indique qu'il nous faut interroger au moins 6 400 personnes !

Si nous ne disposions pas d'estimation préliminaire de π et n'avions aucune idée de l'ordre de grandeur de cette proportion, nous aurions dû nous référer au tableau 2.5 qui nous aurait conseillé de prélever au moins 10 000 individus.

A moins de disposer d'un budget conséquent et d'une armée d'enquêteurs, notre objectif de précision semble donc ici peu réaliste.

b) Exercice 2.5

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de déterminer la taille d'échantillon nécessaire pour estimer une proportion-population avec un niveau de précision fixé.

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

1. Vous êtes chargé de réaliser un sondage dans une population constituée de 5 000 unités statistiques. Vous devez estimer diverses proportions-population : vous savez que certaines d'entre elles sont relativement faibles, que d'autres sont relativement élevées et qu'il y a aussi quelques proportions à estimer qui sont de l'ordre de un demi.
Quelle taille d'échantillon devez-vous choisir si vous souhaitez que les marges d'erreur (ou incertitudes absolues) associées à l'estimation des différentes proportions-population par intervalle de confiance, au niveau de confiance de 95%, soient toutes inférieures ou égales à 5 points de pourcents ?

2. Quelle taille d'échantillon devez-vous choisir si vous souhaitez que les marges d'erreur soient toutes inférieures ou égales à 2,5 points de pourcents ?

3. Répondez à la question 2 en supposant cette fois que toutes les proportions-population à estimer sont inférieures à 30% ou supérieures à 70% ?

4. Répondez à la question 3 en vous plaçant cette fois dans la situation où l'on ne connaît pas la taille de la population mais où l'on sait que cette taille est fort grande.

c) Contrôle de l'incertitude relative

Il faut être conscient du fait qu'un raisonnement en terme d'incertitude *relative* peut radicalement changer le choix de la taille de l'échantillon : si l'on se réfère à nouveau au tableau 2.3, on s'aperçoit que si l'on souhaite maintenir l'incertitude relative pour l'estimation d'une proportion-population π à une valeur assez faible, il faut, si cette proportion est très faible, prévoir une taille d'échantillon très grande (alors que si l'on ne souhaite contrôler que l'incertitude absolue, on peut au contraire se contenter d'une taille d'échantillon assez petite).

Illustrons cette remarque à l'aide d'un exemple.

Exemple

Un recensement passé a montré que les professions scientifiques constituaient 1% de la population active d'un certain pays. On désire aujourd'hui estimer la part des professions scientifiques dans la population active actuelle. Quelle taille d'échantillon faut-il considérer si l'on veut s'assurer que l'incertitude *relative* n'excède pas 5% ?

Nous avons vu que l'incertitude relative était approximativement égale, pour n grand et f petit, à $2\sqrt{\frac{1-\hat{\pi}}{n\hat{\pi}}}$. Pour que cette quantité n'excède pas les 5% que nous nous sommes fixés, nous devons prendre n supérieur à

$$\left(\frac{2}{0{,}05}\right)^2 \frac{1-\hat{\pi}}{\hat{\pi}}.$$

En remplaçant dans cette borne $\hat{\pi}$ par la valeur de 1% obtenue au dernier recensement, on obtient que n doit être supérieur à 158 400… Cette taille d'échantillon est bien trop grande que pour pouvoir être respectée en pratique. Pour se ramener à une taille d'échantillon plus raisonnable, il nous faut obligatoirement revoir à la baisse nos exigences de précision en terme d'incertitude relative !

Et si l'on reformulait nos souhaits de précision en terme d'incertitude *absolue*… Une incertitude relative de 5%, pour une proportion de l'ordre de 1%, correspond à une incertitude absolue de (0,05)(0,01), c'est-à-dire de 0,0005, ou encore de 0,05 points de pourcents. Il s'agit là d'une incertitude absolue très faible, et l'on comprend mieux pourquoi ce niveau de précision coûte si cher en terme de nombre d'observations à réaliser. Si l'on se fixe plutôt comme objectif de faire en sorte que l'incertitude absolue ne dépasse pas 0,5 points de pourcents, nous pouvons réduire la taille d'échantillon à

$$\frac{4(0{,}01)(1-0{,}01)}{(0{,}05)^2} = 1\,584\,,$$

ce qui est déjà nettement plus praticable !

2.7.4 A garder à l'esprit

Que pouvons-nous tirer comme conclusions pour la problématique du choix de la taille de l'échantillon ?

Lorsqu'il y a de nombreuses variables étudiées dans l'enquête, on peut chercher à déterminer la taille de l'échantillon en se basant sur le cas — bien pratique — des proportions.

Mais il faut garder à l'esprit qu'une taille n déterminée de cette façon permettra d'atteindre des précisions qui seront différentes, voire très différentes, selon les variables. Autrement dit, chaque variable détermine « sa » propre taille d'échantillon optimale. Il est donc nécessaire d'adopter une solution de compromis dès lors que l'on ne peut pas choisir d'emblée la taille optimale la plus importante.

Enfin, il ne faut pas oublier qu'en cas d'existence de non-réponse — autrement dit lorsque certains individus de l'échantillon ne participent pas à l'enquête pour une raison ou l'autre —, c'est en fait la taille de l'ensemble des REPONDANTS à l'enquête qui

doit être considérée dans les calculs de précision, et non la taille originelle de l'échantillon prélevé.

Il est donc vivement conseillé d'anticiper le taux de non-réponse pour gonfler la taille de l'échantillon d'origine de façon à atteindre *in fine* le nombre d'observations souhaité.

2.8 La comparaison de deux proportions

2.8.1 Introduction

Les enquêtes pré-électorales n'ont pas pour seul objectif de fournir des estimations brutes des intentions de vote. Elles cherchent également à répondre à diverses questions que se posent les citoyens ou les médias. Les questions les plus fréquentes peuvent être formulées de la manière suivante :

1. Quelle est aujourd'hui l'intention de vote pour le candidat ou le parti A ?
2. Le candidat (ou parti) A devance-t-il aujourd'hui le candidat (ou parti) B ?
3. Le candidat (ou parti) A a-t-il progressé depuis l'enquête précédente ?
4. L'enquête finale fournit-elle des résultats statistiquement compatibles avec le résultat du vote ?

En produisant les deux bornes de l'intervalle de confiance pour la proportion du corps électoral ayant l'intention de voter pour le candidat ou le parti A, on répond de manière pertinente aux questions 1 et 4.

Mais comment tester l'avantage d'un candidat ou d'un parti sur ses adversaires ? Ou comment tester s'il y a une évolution significative du score d'un candidat ou d'un parti, d'une enquête à l'autre ? Un raisonnement probabiliste s'impose pour répondre à ces questions. C'est ce que nous allons étudier dans cette section.

2.8.2 Premier problème de comparaison de deux proportions

a) Le problème

Deux candidats A et B s'affrontent au dernier tour des élections. Gagnera les élections le candidat qui obtiendra plus de 50% des voix.

Un sondage est mené auprès du corps électoral, cinq jours avant les élections. On interroge 2 000 électeurs sélectionnés par une méthode assimilable au sondage aléatoire simple : 52,5% des personnes interrogées disent avoir l'intention de voter pour le candidat A.

Ce résultat nous permet-il de conclure à la prochaine victoire du candidat A aux élections, du moins sous l'hypothèse que l'opinion de la population reste inchangée entre le moment du sondage pré-électoral et le jour des élections ?

En d'autres termes, pouvons-nous conclure du score de 52,5% en faveur du candidat A observé dans l'*échantillon* que la proportion inconnue π_A d'électeurs en faveur de ce candidat dans l'*ensemble du corps électoral* est supérieure à 50% et permet ainsi la victoire de A ?

Répondre à cette question revient à résoudre le problème de test dont l'hypothèse nulle H_0 et l'hypothèse alternative H_1 sont les suivantes :

$$\begin{cases} H_0 : \pi_A \leq 50\% \\ H_1 : \pi_A > 50\% \end{cases}$$

Pour déterminer la règle de décision à suivre face à ce problème de test, je vous propose de tenir un raisonnement similaire à celui qui nous a permis de construire un intervalle de confiance pour une proportion-population. Mais nous allons déterminer cette fois ce que nous pourrions appeler une limite INFERIEURE de confiance pour π_A, au niveau de confiance de 95%.

b) La règle de décision

La question : le candidat A va-t-il gagner le dernier tour des élections?

Rappelons brièvement le problème à résoudre. Deux candidats A et B s'affrontent au dernier tour des élections : le gagnant sera celui qui obtient plus de 50% des suffrages.

Soit π_A, la proportion — inconnue — d'électeurs en faveur du candidat A dans l'ensemble du corps électoral. On dispose d'une estimation $\hat{\pi}_A$ de π_A, correspondant à la proportion d'électeurs en faveur du candidat A dans un échantillon de taille n prélevé par sondage PESR (ou par une méthode assimilable au sondage PESR) dans le corps électoral. Il s'avère que $\hat{\pi}_A$ est supérieure à 50%. Peut-on, sur la base de cette estimation, prédire la victoire de A aux prochaines élections ? En d'autres termes, peut-on déduire du résultat du sondage que la proportion inconnue π_A est strictement supérieure à 0,50 ?

Le raisonnement probabiliste

Repartons de ce que nous savons de la distribution d'échantillonnage de l'estimateur $\hat{\pi}_A$ de la proportion-population π_A : si la taille n de l'échantillon et la taille N de la population sont grandes,

$$\frac{\hat{\pi}_A - \pi_A}{\sqrt{\hat{\mathrm{V}}(\hat{\pi}_A)}} \approx \mathcal{N}(0,1).$$

Il y a dès lors (en bonne approximation) une probabilité de 95% que cette variable, comme toute variable de loi $\mathcal{N}(0,1)$, prenne une valeur inférieure ou égale à 1,64 (le quantile d'ordre 95% de la loi normale centrée réduite) :

$$\mathrm{P}\left[\frac{\hat{\pi}_A - \pi_A}{\sqrt{\hat{\mathrm{V}}(\hat{\pi}_A)}} \leq 1{,}64\right] \simeq 95\%.$$

Nous pouvons déduire de ce dernier résultat que :

$$\mathrm{P}\left[\hat{\pi}_A - 1{,}64\sqrt{\hat{\mathrm{V}}(\hat{\pi}_A)} \leq \pi_A\right] \simeq 95\%.$$

Il y a donc une probabilité de 95% que la quantité $\hat{\pi}_A - 1{,}64\sqrt{\hat{\mathrm{V}}(\hat{\pi}_A)}$ soit inférieure à la valeur exacte de la proportion-population π_A. Cette quantité constitue une **limite inférieure de confiance pour π_A, au niveau de confiance de 95%**. Nous avons la quasi-certitude que la proportion d'électeurs en faveur du candidat A dans le corps électoral excède cette limite inférieure.

Si cette limite inférieure est elle-même supérieure à 50% (voir la figure 2.4), nous pouvons conclure, avec moins de 5% de risque de nous tromper, que la proportion-population π_A est elle aussi supérieure à 50%, autrement dit que le candidat A devrait gagner les élections.

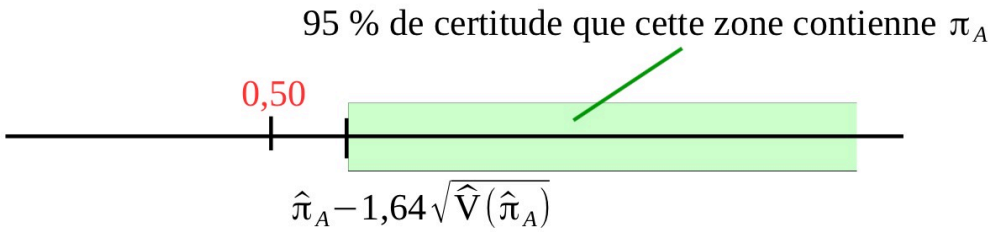


FIGURE 2.4 – *Limite inférieure de confiance pour π_A, au niveau de confiance de 95%*

La règle de décision

En conclusion, les résultats du sondage pré-électoral nous permettent de conclure à la victoire du candidat A ($\pi_A > 0{,}50$), avec moins de 5% de probabilité de nous tromper, si

$$\hat{\pi}_A - 1{,}64\sqrt{\hat{V}(\hat{\pi}_A)} > 0{,}50,$$

c'est-à-dire si

$$\hat{\pi}_A > 0{,}50 + 1{,}64\sqrt{\hat{V}(\hat{\pi}_A)}\,.$$

Puisque, dans un sondage pré-électoral, le taux de sondage f est négligeable et la taille n de l'échantillon est relativement grande, nous avons

$$\hat{V}(\hat{\pi}_A) \simeq \frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n}.$$

L'estimation $\hat{\pi}_A$ de π_A nous permet par conséquent de prédire la victoire du candidat A avec un niveau de certitude d'au moins 95% si

$$\boxed{\hat{\pi}_A > 0{,}50 + 1{,}64\sqrt{\frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n}}.}$$

c) Un exemple

Revenons au cas de figure que nous avons envisagé pour introduire le problème : 52,5% des 2 000 personnes interrogées ont déclaré avoir l'intention de voter pour le candidat A. Peut-on en déduire que π_A est strictement supérieur à 50% ?

Nous avons donc $\hat{\pi}_A$ égal à 52,5% et n égal à 2 000. La borne à laquelle nous devons comparer $\hat{\pi}_A$ vaut

$$0{,}50 + 1{,}64\sqrt{\frac{(0{,}525)(1-0{,}525)}{2\,000}} = 0{,}518.$$

Puisque $\hat{\pi}_A$ dépasse 0,518, nous pouvons prédire la victoire du candidat A, avec une probabilité d'au plus 5 % que cette prédiction soit erronée.

En revanche, si l'on avait observé une proportion de 52,5% dans un échantillon de taille n égal à 1 000, la borne à laquelle comparer $\hat{\pi}_A$ aurait pris pour valeur

$$0{,}50 + 1{,}64\sqrt{\frac{(0{,}525)(1-0{,}525)}{1\,000}} = 0{,}526\,;$$

nous n'aurions plus été dans les conditions pour conclure à la prochaine victoire du candidat A au dernier tour des élections, puisque $\hat{\pi}_A$ est cette fois inférieur à cette borne.

Ainsi, pour que l'on puisse déduire de l'estimation $\hat{\pi}_A$ que la proportion-population π_A est supérieure à 50%, avec un niveau de certitude d'au moins 95%, il ne suffit pas que $\hat{\pi}_A$ ait une valeur supérieure à 50% ! Non ! Il faut non seulement que la proportion $\hat{\pi}_A$ ait une valeur qui excède 50% de manière suffisamment marquée, mais aussi que $\hat{\pi}_A$ ait été observée dans un échantillon de taille suffisamment grande. Une même valeur de $\hat{\pi}_A$, supérieure à 50%, peut nous permettre de conclure à la prochaine victoire du candidat A si l'on a disposé d'un échantillon de grande taille, mais ne pas autoriser cette même prédiction si l'on a travaillé avec un échantillon plus restreint.

d) Exercice 2.6

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de tester statistiquement, sur la base des résultats d'un sondage aléatoire simple, la validité de l'hypothèse selon laquelle la proportion-population inconnue qui vous intéresse est supérieure à 0,5.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Une chaîne de centres de fitness compte 10 245 adhérents. On s'intéresse à leur opinion quant au projet de modification de l'horaire d'ouverture des centres.

1. On interroge 200 adhérents sélectionnés par tirage PESR ; 109 d'entre eux (soit 54,5%) se déclarent en faveur du projet.
Si l'on se fixe un niveau de confiance de 95%, peut-on conclure des résultats de ce sondage que le projet de modification de l'horaire d'ouverture des centres jouit du soutien de la majorité des adhérents des centres ? (Cochez la réponse correcte.)

- ○ Oui. La proportion observée de 54,5% est suffisamment supérieure à 50% pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet.
- ○ Oui. La proportion observée de 54,5% est suffisamment supérieure à 50% pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet, avec toutefois une probabilité d'au plus 5% que cette conclusion ne soit pas conforme à la réalité.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% n'est pas suffisamment grand pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% n'est pas suffisamment grand. On peut en conclure, avec un degré de certitude de 95%, qu'il y a moins de la moitié des adhérents des centres qui sont favorables au projet.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% est tellement faible qu'il n'y a qu'une probabilité d'au plus 5% que la majorité des adhérents des centres soient favorables au projet.

2. Répondez à nouveau à la question 1 mais en considérant cette fois la situation où l'on a observé 54,5% de personnes en faveur du projet dans un échantillon aléatoire simple de taille 500. (Cochez la réponse correcte.)

- ○ Oui. La proportion observée de 54,5% est suffisamment supérieure à 50% pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet.
- ○ Oui. La proportion observée de 54,5% est suffisamment supérieure à 50% pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet, avec toutefois une probabilité d'au plus 5% que cette conclusion ne soit pas conforme à la réalité.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% n'est pas suffisamment grand pour que l'on puisse en conclure que la majorité des adhérents des centres soutiennent le projet.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% n'est pas suffisamment grand. On peut en conclure, avec un degré de certitude de 95%, qu'il y a moins de la moitié des adhérents des centres qui sont favorables au projet.
- ○ Non. L'écart entre la proportion observée (54,5%) et 50% est tellement faible qu'il n'y a qu'une probabilité d'au plus 5% que la majorité des adhérents des centres soient favorables au projet.

2.8.3 Deuxième problème de comparaison de deux proportions

a) Le problème

Considérons à présent un deuxième problème de comparaison de proportions fréquemment rencontré dans le contexte de sondages pré-électoraux.

Imaginons que plusieurs candidats s'opposent les uns aux autres. Au cours d'un sondage d'opinion mené auprès de n personnes sélectionnées selon une méthode que nous pouvons assimiler à un sondage aléatoire simple, les candidats A et B ont récolté les faveurs d'une proportion $\hat{\pi}_A$ et $\hat{\pi}_B$ de l'échantillon.

Il s'avère que $\hat{\pi}_A$ est supérieure à $\hat{\pi}_B$: peut-on en déduire que le candidat A devance aujourd'hui le candidat B dans l'ensemble du corps électoral ?

En d'autres termes, peut-on en conclure que la proportion inconnue π_A du corps électoral en faveur du candidat A est supérieure à la proportion inconnue π_B du corps électoral en faveur du candidat B ?

Le problème de test à résoudre est donc le suivant :

$$\begin{cases} H_0 : \pi_A \leq \pi_B \\ H_1 : \pi_A > \pi_B \end{cases}$$

Résoudre ce problème nécessite à nouveau de faire appel à un raisonnement probabiliste. Ce dernier permet en réalité de déterminer, en fonction de la taille n de l'échantillon, de combien de points de pourcents la proportion observée $\hat{\pi}_A$ doit excéder

la proportion observée $\hat{\pi}_B$ pour que l'on puisse en déduire que π_A est supérieure à π_B, avec un niveau de certitude d'au moins 95%.

b) La règle de décision

La question : le candidat (parti) A devance-t-il aujourd'hui le candidat (parti) B ?

Soient π_A et π_B, les proportions — inconnues — d'électeurs en faveur du candidat A et en faveur du candidat B, respectivement, dans l'ensemble du corps électoral. On dispose d'une estimation $\hat{\pi}_A$ de π_A et d'une estimation $\hat{\pi}_B$ de π_B, correspondant aux proportions d'électeurs ayant l'intention de voter pour le candidat A et pour le candidat B, respectivement, dans un échantillon de taille n prélevé dans le corps électoral par sondage PESR (ou par une méthode assimilable au sondage PESR).

Il s'avère que $\hat{\pi}_A > \hat{\pi}_B$: peut-on en déduire que $\pi_A > \pi_B$?

Le raisonnement probabiliste

Pour répondre à cette question, il ne suffit pas de vérifier si les intervalles de confiance pour π_A et pour π_B se recoupent ou non. En effet, le raisonnement probabiliste peut nous amener à conclure que π_A est effectivement supérieure à π_B même lorsque les intervalles de confiance se recoupent… Voyons cela de plus près !

Se demander si $\pi_A > \pi_B$ revient à vouloir déterminer si $\pi_A - \pi_B > 0$. Nous souhaitons donc tirer des résultats du sondage une conclusion quant au signe de la différence $\pi_A - \pi_B$ des proportions inconnues π_A et π_B. Il est clair que cette différence peut être estimée par $\hat{\pi}_A - \hat{\pi}_B$. Pour répondre à la question posée ci-dessus, il faut s'intéresser à la distribution d'échantillonnage de $\hat{\pi}_A - \hat{\pi}_B$ et rechercher une limite inférieure de confiance pour la différence inconnue $\pi_A - \pi_B$, au niveau de confiance de 95%.

L'annexe technique 2.6 présente les détails du raisonnement probabiliste menant à la règle de conduite à suivre pour répondre à la question considérée ici.

La règle de décision

Les résultats du sondage pré-électoral nous permettent de conclure que la proportion du corps électoral en faveur du candidat A est supérieure à celle en faveur du candidat B (c'est-à-dire que $\pi_A > \pi_B$, ou encore que $\pi_A - \pi_B > 0$), avec moins de 5% de probabilité de nous tromper, si

$$\hat{\pi}_A - \hat{\pi}_B > 1{,}64\sqrt{\frac{1}{n}[\hat{\pi}_A(1-\hat{\pi}_A) + \hat{\pi}_B(1-\hat{\pi}_B) + 2\hat{\pi}_A\hat{\pi}_B]}\,.$$

c) Un exemple

Illustrons ce raisonnement statistique en faisant appel à un célèbre cafouillage médiatique alimenté par des sondages apparemment contradictoires lors des élections présidentielles françaises de 2012 opposant Nicolas Sarkozy à François Hollande.

Le lundi 12 mars 2012, l'IFOP annonce : « Croisement des courbes ! Le candidat Sarkozy en tête au premier tour avec 28,5% des intentions de vote (...) ».

Le lendemain, mardi 13 mars, la SOFRES annonce : « Le candidat Hollande confirme son avance avec 30% des intentions de vote (...) ».

A la lecture de ces résultats, les citoyens ont des raisons de se montrer perplexes. Mais analysons d'un peu plus près les résultats de ces deux sondages et vérifions s'ils permettent réellement d'affirmer qu'un candidat est en avance sur l'autre. Désignons par π_H et π_S les proportions du corps électoral ayant l'intention de voter pour Hollande et pour Sarkozy, respectivement, au moment du sondage.

Penchons-nous tout d'abord sur les résultats du sondage IFOP. La fiche technique du sondage (cf. http://www.ifop.com/media/poll/1793-1-study_file.pdf) nous indique que les estimations de π_S et π_H ont été obtenues auprès d'un échantillon de $n = 1\,638$ personnes inscrites sur les listes électorales : $\hat{\pi}_H = 27\%$ et $\hat{\pi}_S = 28{,}5\%$, ce qui nous conduit à un écart observé de 1,5 point de pourcents en faveur de Sarkozy. Peut-on déduire de ces résultats, à un degré de certitude de 95%, que Sarkozy devance désormais Hollande ? Nous avons $\hat{\pi}_S - \hat{\pi}_H = 1{,}5\%$ et

$$\frac{1}{n}[\hat{\pi}_S(1-\hat{\pi}_S) + \hat{\pi}_H(1-\hat{\pi}_H) + 2\hat{\pi}_S\hat{\pi}_H]$$
$$= \frac{1}{1\,638}[(0{,}285)(1-0{,}285) + (0{,}27)(1-0{,}27) + 2(0{,}285)(0{,}27)] = 0{,}00034.$$

La borne à laquelle il nous faut comparer $\hat{\pi}_S - \hat{\pi}_H$ vaut donc

$$1{,}64\sqrt{0{,}00034} = 0{,}03 = 3\%.$$

Puisque la différence de 1,5% entre les scores observés de Sarkozy et de Hollande est inférieure à cette borne de 3%, nous ne pouvons pas conclure des résultats du sondage IFOP que Sarkozy est passé en tête, du moins si nous désirons raisonner à un degré de certitude de 95%.

Et qu'en est-il du sondage de la TNS SOFRES ? La consultation de la fiche de ce sondage (cf. http://www.tns-sofres.com/sites/default/files/2012.03.13-iv8.pdf) nous indique que 1 000 personnes ont été interrogées, mais que 20% d'entre elles n'ont pas exprimé d'intention de vote, ce qui nous ramène à $n = 800$. Ce sondage a fourni les estimations suivantes : $\hat{\pi}_H = 30\%$ et $\hat{\pi}_S = 26\%$, soit un écart de 4 points de pourcents en faveur de Hollande ! Cependant,

$$\frac{1}{n}[\hat{\pi}_S(1-\hat{\pi}_S) + \hat{\pi}_H(1-\hat{\pi}_H) + 2\hat{\pi}_S\hat{\pi}_H]$$
$$= \frac{1}{800}[(0{,}26)(1-0{,}26) + (0{,}30)(1-0{,}30) + 2(0{,}26)(0{,}30)] = 0{,}0007.$$

La borne à laquelle il nous faut comparer la différence observée $\hat{\pi}_H - \hat{\pi}_S$ s'élève donc à

$$1{,}64\sqrt{0{,}0007} = 0{,}043 = 4{,}3\%.$$

Cette borne excède $\hat{\pi}_H - \hat{\pi}_S$. En conclusion, l'institut de sondage TNS SOFRES a été quelque peu imprudent en affirmant que Hollande confirmait son avance sur Sarkozy. En effet, si l'on se fixe un degré de certitude de 95%, les résultats du sondage ne permettaient pas d'affirmer que l'avance de Hollande sur Sarkozy observée dans l'échantillon était statistiquement significative.

d) Exercice 2.7

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de tester statistiquement, sur la base des résultats d'un sondage aléatoire simple, la validité de l'hypothèse selon laquelle une première proportion-population inconnue qui vous intéresse est supérieure à une seconde proportion-population inconnue.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Les élections pour la présidence du parti Union doivent avoir lieu dans un mois. Quatre candidats sont en lice.

Un sondage est mené auprès d'un échantillon aléatoire simple de 350 personnes affiliées au parti, afin de tenter de prédire les résultats de ces élections. Les deux candidats favoris — appelons-les A et B — ont reçu respectivement les faveurs de 37,2% et 34,5% des personnes de l'échantillon.

Peut-on, sur la base de ces résultats et si l'on se fixe un niveau de confiance de 95%, affirmer que le candidat A est préféré au candidat B dans l'ensemble des membres du parti Union ? (Cochez la réponse correcte.)

- ○ Oui. Puisque 37,2% est clairement supérieur à 34,5%, on peut effectivement affirmer que le candidat A est préféré au candidat B dans l'ensemble des membres du parti Union.
- ○ Oui. La supériorité du score du candidat A sur celui du candidat B dans l'échantillon est suffisamment forte pour que l'on puisse affirmer que le candidat A est préféré au candidat B dans l'ensemble des membres du parti Union, avec toutefois une probabilité de 5% (au maximum) que cette affirmation ne soit pas conforme à la réalité.
- ○ Non. La supériorité du score du candidat A sur celui du candidat B dans l'échantillon est trop faible pour pouvoir affirmer que le candidat A est préféré au candidat B dans l'ensemble des membres du parti Union.
- ○ Non. La supériorité du score du candidat A sur celui du candidat B dans l'échantillon est tellement faible qu'on peut au contraire affirmer, avec un degré de certitude de 95%, que le candidat A ne devance pas le candidat B dans le cœur des membres du parti Union.
- ○ Non. La supériorité du score du candidat A sur celui du candidat B dans l'échantillon est tellement faible qu'il n'y a qu'une probabilité d'au plus 5% que le candidat A devance réellement le candidat B dans le cœur des membres du parti Union.

2.8.4 Troisième problème de comparaison de deux proportions

a) Le problème

Une autre problématique fréquemment soulevée lors de l'analyse des résultats de sondages pré-électoraux est celle de l'évolution du pourcentage d'électeurs en faveur d'un certain parti ou d'un certain candidat au cours de la période pré-électorale.

Imaginons la situation suivante.

Un premier sondage a été mené auprès d'un premier échantillon aléatoire simple d'électeurs : une proportion $\hat{\pi}_1$ de personnes s'y sont déclarées en faveur du candidat en question. Un nouveau sondage a été réalisé 10 jours plus tard auprès d'un autre

échantillon aléatoire simple d'électeurs et une proportion $\hat{\pi}_2$ d'entre eux y ont affirmé avoir l'intention de voter pour le candidat.

Il s'avère que $\hat{\pi}_2$ est supérieure à $\hat{\pi}_1$: peut-on en déduire que la proportion du corps électoral en faveur du candidat a effectivement augmenté entre les dates de la première et de la seconde enquêtes ?

En d'autres termes, le fait que $\hat{\pi}_2$ soit supérieure à $\hat{\pi}_1$ permet-il d'affirmer que π_2 est supérieure à π_1, où π_2 et π_1 sont les proportions inconnues du corps électoral en faveur du candidat au moment de la deuxième enquête et de la première enquête, respectivement ?

Le problème de test à résoudre est donc le suivant :

$$\begin{cases} H_0 : \pi_2 \leq \pi_1 \\ H_1 : \pi_2 > \pi_1 \end{cases}$$

Comme pour les deux autres problèmes de test déjà considérés, il nous faut recourir à un raisonnement probabiliste. Ce dernier va nous indiquer, en fonction des tailles n_1 et n_2 des deux échantillons, de combien de points de pourcents la proportion observée $\hat{\pi}_2$ doit excéder la proportion observée $\hat{\pi}_1$ pour que l'on puisse conclure des résultats des deux sondages que la proportion du corps électoral en faveur du candidat est effectivement à la hausse, avec un niveau de certitude d'au moins 95%.

b) La règle de décision

La question : le candidat (parti) A a-t-il progressé depuis l'enquête précédente ?

Soient

- π_1, la proportion — inconnue — d'électeurs en faveur du candidat A dans l'ensemble du corps électoral au moment de la première enquête, et $\hat{\pi}_1$, son estimation obtenue dans un échantillon aléatoire simple de taille n_1 ;
- π_2, la proportion — inconnue — d'électeurs en faveur du candidat A dans l'ensemble du corps électoral au moment de la deuxième enquête, et $\hat{\pi}_2$, son estimation obtenue dans un échantillon aléatoire simple de taille n_2.

Il s'avère que $\hat{\pi}_2 > \hat{\pi}_1$: peut-on en déduire que $\pi_2 > \pi_1$, c'est-à-dire que $\pi_2 - \pi_1 > 0$?

Le raisonnement probabiliste

Pour répondre à cette question, nous pouvons suivre une démarche similaire à celle présentée pour résoudre le deuxième problème de test considéré précédemment : nous allons partir de la distribution d'échantillonnage de l'estimateur $(\hat{\pi}_2 - \hat{\pi}_1)$ de la différence inconnue $(\pi_2 - \pi_1)$ afin de déterminer une limite inférieure de confiance pour cette différence, au niveau de confiance de 95%. La procédure va cependant être un peu simplifiée grâce au fait que, les échantillons associés aux deux sondages étant généralement prélevés indépendamment l'un de l'autre, les estimateurs $\hat{\pi}_1$ et $\hat{\pi}_2$ sont des variables aléatoires indépendantes l'une de l'autre. Cette indépendance implique que la variance de la différence $(\hat{\pi}_2 - \hat{\pi}_1)$ des deux estimateurs est simplement égale à la somme de leurs variances respectives.

L'annexe technique 2.7 présente les détails du raisonnement probabiliste menant à la règle de conduite à suivre pour répondre à la question considérée ici.

La règle de décision

L'accroissement du score du candidat observé d'un sondage à l'autre nous permet de conclure que la proportion du corps électoral en faveur du candidat est effectivement à la hausse ($\pi_2 > \pi_1$, ou encore $\pi_2 - \pi_1 > 0$), avec moins de 5% de probabilité de nous tromper, si

$$\hat{\pi}_2 - \hat{\pi}_1 > 1{,}64\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

c) Un exemple

Un premier sondage, réalisé auprès d'un échantillon aléatoire simple de $n_1 = 1\,750$ électeurs, a donné lieu à une proportion estimée $\hat{\pi}_1 = 43{,}2\%$ de personnes en faveur d'un certain candidat. Dix jours après, un deuxième sondage mené auprès d'un autre échantillon aléatoire simple de $n_2 = 1\,810$ électeurs a conduit à une proportion estimée $\hat{\pi}_2 = 45{,}1\%$ de personnes en faveur de ce même candidat. Peut-on conclure de ces estimations que la proportion de l'ensemble du corps électoral en faveur de ce candidat est réellement à la hausse ?

La différence observée $\hat{\pi}_2 - \hat{\pi}_1$ vaut $45{,}1\% - 43{,}2\% = 1{,}9\% = 0{,}019$. La quantité $1{,}64\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$ vaut quant à elle 0,027 et est ainsi supérieure à $\hat{\pi}_2 - \hat{\pi}_1$. La hausse observée de 1,9% n'est donc pas statistiquement significative ! Elle n'est pas suffisamment marquée pour que l'on puisse en conclure qu'il y a bien hausse du score du candidat dans l'ensemble du corps électoral, du moins si l'on fixe à 5% au maximum la probabilité de nous tromper en affirmant l'existence d'une telle hausse.

d) Remarque : l'importance du choix du niveau de confiance

Comme le montrent les exemples présentés pour les deuxième et troisième problèmes de test, les échantillons habituels sont très souvent de tailles insuffisantes pour différencier les intentions entre candidats au coude à coude, et pour suivre à court terme les évolutions d'intentions de vote. Ainsi, à moins de considérer un (des) échantillon(s) de taille(s) économiquement inaccessible(s), un institut de sondage ne pourrait que rarement conclure « scientifiquement » sur la dynamique électorale. Mais il serait naïf de penser qu'il faille attendre la confirmation d'une évaluation avec une certitude de 95% pour qu'un candidat juge si sa stratégie est positive ou s'il recule. S'il nous fallait toujours ce degré de certitude dans la vie, nous ne prendrions pas beaucoup de décisions !

Reprenons donc l'exemple que nous venons de considérer pour le troisième problème de test et baissons quelque peu le degré de certitude exigé. Si nous voulons un degré de certitude de 90%, il nous faut remplacer, dans la borne à laquelle la différence $(\hat{\pi}_2 - \hat{\pi}_1)$ doit être comparée, la valeur 1,64 par le quantile d'ordre 90% de la loi $\mathcal{N}(0,1)$. Ce quantile correspond à la valeur laissant à sa gauche 90% des valeurs possibles d'une variable de loi $\mathcal{N}(0,1)$: il vaut 1,28. La quantité $1{,}28\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$ est égale à 0,021 et est encore supérieure à $(\hat{\pi}_2 - \hat{\pi}_1)$: à ce degré de certitude, on ne peut donc toujours pas conclure à une augmentation du score du candidat dans le corps électoral.

Si l'on baisse le degré de certitude à 85%, la valeur 1,64 doit être remplacée par le quantile d'ordre 85% de la loi $\mathcal{N}(0,1)$, soit 1,04. La quantité $1{,}04\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}+\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$ est alors égale à 0,017 : puisque sa valeur est inférieure à $(\hat{\pi}_2-\hat{\pi}_1)$, on peut conclure cette fois à une progression statistiquement significative du score du candidat dans le corps électoral, mais en gardant à l'esprit qu'il y a une probabilité de 15% (au maximum) de nous tromper en tirant cette conclusion.

e) Exercice 2.8

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de tester statistiquement, sur la base des résultats de deux sondages aléatoires simples successifs réalisés dans la même population, la validité de l'hypothèse selon laquelle une certaine proportion-population inconnue qui vous intéresse a augmenté (ou diminué).

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Le service médical d'une grande université s'intéresse depuis de nombreuses années au problème du tabagisme chez les étudiants. Il réalise tous les deux ans un premier sondage auprès de la population étudiante masculine et un second sondage auprès de la population étudiante féminine, afin d'étudier de quelle manière cette question de santé évolue au cours du temps.

En 2013, 1225 garçons et 1440 filles ont été interrogés ; 25% des garçons et 16% des filles ont indiqué qu'ils fumaient quotidiennement.

En 2015, 1520 garçons et 1455 filles ont participé à l'enquête (les personnes chargées de l'enquête se sont préalablement assurées que les étudiants et étudiantes interrogées en 2015 n'avaient pas fait partie des échantillons sélectionnés en 2013) ; 23% des garçons et 22% des filles ont déclaré qu'ils fumaient quotidiennement.

1. Peut-on conclure de ces résultats de sondages qu'il y a eu, entre 2013 et 2015, une augmentation du pourcentage de fumeurs quotidiens dans la population des étudiantes de l'université ? Raisonnez en vous fixant un niveau de confiance de 95%.

- ○ Oui
- ○ Non

2. Peut-on conclure de ces résultats de sondages qu'il y a eu, entre 2013 et 2015, une diminution du pourcentage de fumeurs quotidiens dans la population des étudiants masculins de l'université ? Raisonnez en vous fixant un niveau de confiance de 95%.

- ○ Oui
- ○ Non

2.9 Le tirage de l'échantillon

Dans cette section, nous allons revenir de manière plus pratique sur le tirage à proprement parler de l'échantillon selon le plan de sondage PESR. Comment procéder, en pratique, pour obtenir un échantillon selon ce plan de sondage en vue d'estimer ensuite l'un ou l'autre paramètre de la population ?

Vous pourriez bien évidemment penser à tirer l'échantillon selon la procédure présentée dans la section 2 de ce chapitre : vous mettez dans une **urne** N boules ou papiers numérotés de 1 à N, puis vous prélever successivement, « à l'aveugle » et sans remise, n boules ou papiers dans cette urne. Les numéros apparaissant sur les boules ou papiers ainsi sélectionnés sont ceux des individus de la population à contacter pour leur soumettre l'enquête.

Mais vous imaginez bien que cette façon de faire s'avère très peu pratique dès que la population est de grande taille. On a donc cherché à retraduire cette procédure de tirage de l'échantillon sous la forme d'un **algorithme informatique** directement applicable sur la **base de sondage**. Celle-ci est d'ailleurs très souvent disponible sous la forme d'un fichier Excel ou d'un fichier texte de format txt ou csv, par exemple. Cet algorithme informatique est censé *mimer* la procédure aléatoire de tirage de l'échantillon et doit être conçu de telle sorte à respecter les caractéristiques du sondage aléatoire simple. En particulier, il faut qu'avec cet algorithme d'échantillonnage, toute combinaison de n éléments parmi les N éléments de la base de sondage soit sélectionnée avec la même probabilité ; mais il faut aussi qu'avec cet algorithme, toute unité statistique de la population ait la même « chance » que les autres d'être sélectionnée (autrement dit que tous les éléments de la population aient la même probabilité d'inclusion).

Différents algorithmes ont été proposés. Citons par exemple l'algorithme de la **méthode du tri aléatoire** (voir ci-dessous), celui de la **méthode de sélection-rejet**, celui de la **méthode de mise à jour de l'échantillon**. Ces algorithmes diffèrent de par leur complexité, leur rapidité, leur consommation en espace mémoire.

Certains algorithmes présentent aussi l'avantage de pouvoir être mis en œuvre sans devoir connaître au préalable la taille de la population : l'algorithme parcourt progressivement la base de sondage et s'arrête dès qu'il arrive à la dernière unité du fichier. Ce type d'algorithme est bien pratique à utiliser lorsqu'il faut sélectionner un échantillon par tirage à probabilités égales sans remise dans une base de sondage qui se construit progressivement : c'est le cas, par exemple, lorsqu'on veut tirer un échantillon dans la population des clients qui vont se présenter aux guichets d'une certaine banque au cours d'une journée donnée. Certains de ces algorithmes sont disponibles dans des logiciels statistiques tels que Stata, SAS, ou le logiciel libre R, par exemple.

2.9.1 La méthode du tri aléatoire

a) La procédure de tirage

On commence par générer N nombres (pseudo-)aléatoires $u_1, u_2, \dots, u_N$ suivant la loi $\mathcal{U}(0,1)$ (loi uniforme sur l'intervalle (0,1)) et, pour $i = 1, \dots, N$, on affecte le nombre u_i à l'individu i de la base de sondage.

N.B. : On peut très facilement générer un nombre aléatoire u suivant la loi $\mathcal{U}(0,1)$ à l'aide de la fonction ALEA() de Excel ou du tableur de LibreOffice ou d'OpenOffice (ou la fonction RAND() si on dispose d'une version anglophone du tableur).

On trie ensuite la base de sondage par ordre croissant (ou décroissant) de ces nombres aléatoires u_i : cette procédure conduit en fait à effectuer un tri aléatoire de la base de sondage.

On sélectionne enfin pour l'échantillon les n premiers (ou les n derniers) individus de la base de sondage ainsi ordonnée.

b) Remarques

On montre que cette méthode de tirage respecte bien le plan de sondage PESR ; il donne lieu à des échantillons de taille n fixée *a priori.*

Son principal avantage est sans nul doute sa grande facilité de mise en œuvre. Cet avantage est cependant contrebalancé par deux défauts majeurs :

(i) l'application de cette méthode d'échantillonnage nécessite de connaître au préalable la taille N de la population (il est donc nécessaire de disposer dès le départ de l'ensemble de la base de sondage) ;

(ii) cette méthode requiert de trier toute la base de sondage : cette opération peut s'avérer très longue quand le fichier est grand.

2.9.2 Le tirage systématique

Parmi les nombreux algorithmes permettant de prélever un échantillon dans une base de sondage selon le plan de sondage PESR, l'algorithme du *tirage systématique* s'avère particulièrement simple à appliquer et est dès lors très largement utilisé.

Tout comme pour la méthode du tri aléatoire, il n'est pas nécessaire de faire appel à un logiciel statistique pour le mettre en œuvre. Il suffit que la base de sondage soit constituée d'un fichier dans lequel les unités de la population sont numérotées de 1 à N, et de disposer d'un tableur ou d'une machine à calculer pouvant générer des nombres aléatoires entre 0 et 1 selon une loi uniforme sur l'intervalle (0,1).

a) La procédure de tirage

Supposons que l'on veuille sélectionner un échantillon de taille n.

On commence par calculer ce que l'on appelle le « PAS » du tirage, qui correspond au rapport entre la taille N de la population et la taille n de l'échantillon :

$$\text{PAS} = N/n.$$

Le PAS n'est pas nécessairement un nombre entier : il ne le sera que si N est un multiple de n. Quoi qu'il en soit, ce PAS va représenter la distance, comptabilisable en nombre d'individus après arrondi, qu'il faut parcourir dans le fichier de la base de sondage entre deux sélections successives.

Une fois que l'on dispose du PAS du tirage, on détermine *aléatoirement* le numéro du premier individu à sélectionner : ce numéro, que nous pouvons désigner par INIT, est le résultat de l'opération :

$$1 + \text{ENT}(\text{ALEA}*\text{PAS}),$$

où ALEA est un nombre compris entre 0 et 1, généré aléatoirement (selon la loi $\mathcal{U}(0,1)$) par la fonction *ad hoc* du tableur ou de la machine à calculer que vous utilisez, et ENT représente la partie entière d'un nombre positif, c'est-à-dire l'entier que l'on obtient si l'on supprime la partie décimale du nombre.

On peut ensuite déterminer, de manière *automatique*, les numéros des $(n-1)$ autres individus à sélectionner : il s'agira, pour le compteur j allant de 1 à $(n-1)$, des numéros donnés par :

$$1 + \text{ENT}((\text{ALEA} + j)*\text{PAS}),$$

ou encore, si vous préférez,

$$1 + \text{ENT}(\text{ALEA}*\text{PAS} + j*\text{PAS}).$$

Le tirage systématique a donc pour particularité que la sélection *aléatoire* de la première unité détermine *automatiquement* l'échantillon dans son ensemble.

Notez encore que si l'on prend j égal à 0 dans la dernière formule, on retrouve le numéro INIT du premier individu à sélectionner. Ainsi, les numéros des n individus à sélectionner sont donnés par :

$$1 + \text{ENT}(\text{ALEA}*\text{PAS} + j*\text{PAS}) \quad \text{pour } j = 0,1,\dots,n-1.$$

Illustrons cette procédure sur un petit exemple.

Exemple

Considérons une population de 32 dossiers, numérotés de 1 à 32 selon leur ordre d'arrivée sur notre bureau. Nous aimerions y sélectionner un échantillon de taille 5 par tirage systématique.

On peut appliquer ce tirage en nous aidant d'Excel ou du tableur de LibreOffice ou OpenOffice.

Nous avons $N = 32$ et $n = 5$. Le PAS du tirage est égal à 32/5, c'est-à-dire 6,4.

On demande alors à Excel de nous générer aléatoirement un nombre entre 0 et 1 à l'aide de la fonction ALEA() si vous utilisez la version francophone d'Excel et à l'aide de la fonction RAND() si vous utilisez la version anglophone d'Excel : nous obtenons par exemple 0,0932.

Pour déterminer les numéros des 5 dossiers à sélectionner, il nous faut appliquer la formule 1+ ENT(ALEA*PAS+j*PAS) pour le compteur j allant de 0 à 4.

Ainsi, pour le premier dossier à sélectionner ($j = 0$), nous avons :

1+ ENT(ALEA*PAS+j*PAS) = 1+ ENT(0,5964) = 1+0 = 1.

Pour le deuxième dossier à sélectionner ($j = 1$), nous avons :

1+ ENT(ALEA*PAS+j*PAS) = 1+ ENT(6,9964) = 1+6 = 7.

Et ainsi de suite pour les trois autres dossiers qu'il nous reste à sélectionner :

$j = 2 \Rightarrow$ 1+ ENT(ALEA*PAS+j*PAS) = 1+ ENT(13,3965) = 1+13 = 14.

$j = 3 \Rightarrow$ 1+ ENT(ALEA*PAS+j*PAS) = 1+ ENT(19,7965) = 1+19 = 20.

$j = 4 \Rightarrow$ 1+ ENT(ALEA*PAS+j*PAS) = 1+ ENT(26,1965) = 1+26 = 27.

Au terme de la procédure de tirage, nous voyons qu'il nous faut donc prélever les dossiers n° 1, 7, 14, 20 et 27.

b) Remarques

La simplicité du tirage systématique fait son succès. Il faut néanmoins être vigilant sur ses propriétés, comme expliqué ci-dessous.

On peut montrer que, si l'ordre selon lequel les unités statistiques sont listées dans la base de sondage est « aléatoire » ou, à tout le moins, est tout à fait indépendant de la variable d'intérêt, le tirage systématique est rigoureusement équivalent au sondage aléatoire simple.

En revanche, s'il y a une périodicité dans la base de sondage et que le « pas » du tirage systématique est égal à la période (ou à un multiple de celle-ci), on risque de sélectionner des unités très particulières. Imaginons par exemple que, pour réaliser une enquête sur la consommation de chauffage, on tire des logements avec un pas de 4 alors que le tirage se réalise dans une tour où chaque étage comprend 4 logements. Dans ce cas, si le premier logement sélectionné est orienté au nord, tous les logements de l'échantillon seront orientés au nord, et on surestimera certainement la dépense moyenne de chauffage par logement ! Dans une telle situation, nous serons confrontés à un *biais* dit *d'échantillonnage*.

Une dernière petite remarque : dans certaines circonstances, on peut être amené à adopter un tirage systématique en se fixant *a priori*, non pas la taille n de l'échantillon, mais plutôt le « pas » du tirage. Cette manière de faire s'avère bien pratique lorsqu'on ne connaît pas la taille de la population au moment où l'on doit lancer la procédure de tirage, c'est-à-dire que l'on ne dispose pas d'emblée de l'ensemble de la base de sondage. Supposons que l'on ait, par exemple, à échantillonner des dossiers qui arrivent par lots successifs dans un bureau, mais que l'on ne sache pas par avance combien on va recevoir de dossiers au total. On peut alors sélectionner très facilement un échantillon en tirant un dossier sur dix, par exemple, de manière systématique au fur et à mesure que les lots de dossiers se présentent. Si cette procédure de tirage de l'échantillon est très pratique et facile à mettre en œuvre, elle ne peut malheureusement pas être assimilée au sondage aléatoire simple. En effet, avec une telle procédure de tirage, non seulement la composition de l'échantillon est aléatoire, mais sa taille l'est aussi. Nous ne pouvons plus, alors, en toute rigueur, utiliser les estimateurs que nous avons étudiés pour le sondage PESR. Ainsi, par exemple, la taille de l'échantillon étant variable, on n'estime plus sans biais la moyenne-population μ par la moyenne-échantillon $\bar{y}$, mais

plutôt par $(n/N) \times \text{PAS} \times \bar{y}$, où PAS correspond au « pas » du tirage que l'on s'est choisi arbitrairement au début de la procédure.

c) Exercice 2.9

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de tirer un échantillon dans une population par tirage systématique.

Consignes – Cet exercice doit être réalisé à l'aide de votre tableur (Excel, LibreOffice ou OpenOffice).

Correctif – Vous pouvez télécharger sur l'UV un fichier Excel (Corr_ex_2_9.xlsx) vous permettant de vérifier si vous avez effectué correctement les échantillonnages par tirage systématique qui vous sont demandés dans cet exercice.

1. Sélectionnez par tirage systématique un échantillon de taille 50 dans une population de taille 1 242 (où les unités statistiques sont numérotées de 1 à 1 242).

2. Sélectionnez par tirage systématique un échantillon de taille 200 dans une population de taille 250 000 (où les unités statistiques sont numérotées de 1 à 250 000).

2.9.3 Le tirage de Bernoulli

Nous venons de le voir, le *tirage systématique* est une procédure de tirage de l'échantillon aux multiples avantages : il est extrêmement facile à mettre en œuvre, ne nécessite pas l'utilisation d'un logiciel statistique particulier (une simple machine à calculer ou un tableur suffisent), ne coûte pratiquement rien en temps de calcul… et, si la base de sondage est triée aléatoirement ou selon un ordre qui n'a *a priori* aucun lien avec les variables d'intérêt, respecte le plan de sondage aléatoire simple (les probabilités de sélection des échantillons possibles et les probabilités d'inclusion des unités de la population).

Toutefois, on ne peut prélever un échantillon par tirage systématique que si l'on connaît dès le départ la taille N de la population. Comment faire alors si la base de sondage se construit simultanément à l'échantillonnage[1] et que l'on ne dispose qu'*a posteriori* de la connaissance de N ? Le tirage de Bernoulli est une solution possible à ce problème.

Le *tirage de Bernoulli* est une procédure de tirage de l'échantillon excessivement simple à mettre en œuvre, ne nécessitant pas la connaissance *a priori* de la taille N de la population à sonder, et respectant un plan de sondage qui ne coïncide pas exactement avec le plan de sondage PESR, mais qui présente toutefois de fortes similarités avec ce dernier.

[1] Il en est ainsi, par exemple, lorsqu'on veut sélectionner « sur place » un échantillon parmi l'ensemble des personnes qui se présentent aux guichets d'une certaine agence bancaire, un jour donné ; on ne connaît la taille de la population à sonder qu'en fin de journée, à la fermeture des guichets.

a) La procédure de tirage

On commence par se choisir une quantité $p \in (0,1)$. Nous verrons ultérieurement à quoi correspond exactement cette quantité et, dès lors, comment la choisir de manière raisonnée.

Ensuite, de manière séquentielle[2], pour chaque unité statistique $i = 1, \dots, N$ de la population :

- on génère un nombre (pseudo-)aléatoire u_i selon une loi $\mathcal{U}(0,1)$ (loi continue uniforme sur l'intervalle (0,1)), via la fonction ALEA() ou RAND() du tableur ;
- si $u_i \leq p$, alors l'unité statistique i est sélectionnée pour faire partie de l'échantillon ;
  si, au contraire, $u_i > p$, alors l'unité statistique i n'est pas sélectionnée.

La procédure de tirage s'arrête une fois que l'on a passé en revue, une à une, toutes les unités de la population.

b) Les caractéristiques du plan de sondage associé au tirage de Bernoulli

- ***Indépendance des prélèvements successifs***

Le tirage de Bernoulli se caractérise tout d'abord par le fait que la procédure de sélection est indépendante d'une unité statistique de la population à l'autre : chaque unité statistique est sélectionnée (ou non) indépendamment de ce qui se passe pour les autres unités.

- ***Les probabilités d'inclusion***

Que vaut la probabilité d'inclusion de l'unité i ?

La probabilité que l'individu i soit sélectionné pour faire partie de l'échantillon est égale à la probabilité que le nombre u_i qu'on lui a associé soit inférieur ou égal au nombre p que l'on s'est fixé :

$$p_i = \mathrm{P}(i \in S) = \mathrm{P}(u_i \leq p).$$

Or, u_i a été généré selon une loi uniforme $\mathcal{U}(0,1)$. La probabilité d'inclusion de l'individu i est donc égale à la probabilité qu'une variable aléatoire de loi $\mathcal{U}(0,1)$ prenne une valeur inférieure ou égale à p et, par conséquent, est égale à p :

$$p_i = \mathrm{P}(\mathcal{U}(0,1) \leq p) = p.$$

Ainsi, le tirage de Bernoulli associe la *même* probabilité d'inclusion à toutes les unités statistiques de la population : il s'agit donc d'une méthode d'échantillonnage « à probabilités égales » (PE).

Par ailleurs, les prélèvements dans la population se font *sans remise* (SR), puisque chaque unité statistique ne peut jamais être prélevée à plusieurs reprises.

Mais le tirage de Bernoulli respecte-t-il pour autant le plan de sondage PESR que nous avons étudié dans ce chapitre ? La réponse à cette question est négative. En effet, si le

[2] Au fur et à mesure que les clients se présentent aux guichets, par exemple.

sondage aléatoire simple (PESR) donne lieu à des échantillons de taille fixe, le tirage de Bernoulli en revanche correspond à une méthode d'échantillonnage de taille aléatoire !

- ***Echantillonnage de taille aléatoire***

Désignons par n_S la taille de l'échantillon S qui sera sélectionné par tirage de Bernoulli.

Le tirage de Bernoulli n'est autre qu'une procédure aléatoire respectant le schéma de Bernoulli. En effet :

- on répète à N reprises, sous des conditions expérimentales identiques et de manière indépendante, l'expérience aléatoire $\mathcal{E}$ consistant à générer une valeur selon la loi uniforme $\mathcal{U}(0,1)$.
- Cette expérience $\mathcal{E}$ donne lieu à un « succès » si le nombre généré est inférieur ou égal à p (auquel cas l'individu correspondant est sélectionné pour l'échantillon), à un « échec » dans le cas contraire.
- La probabilité de succès est égale à p.

Dès lors, puisque la taille n_S de l'échantillon correspond au « nombre de succès au cours des N répétitions de l'expérience aléatoire $\mathcal{E}$, il est clair que n_S est *aléatoire* et suit la loi binomiale d'exposant N et de probabilité p :

$$n_S \sim \mathrm{Bin}(N, p).$$

Il en découle que

$$\mathrm{E}(n_S) = Np \quad \text{et} \quad \mathrm{V}(n_S) = Np(1-p).$$

Le tirage de Bernoulli fournit donc des échantillons de tailles *variables*, variant de 0 à N selon la loi $\mathrm{Bin}(N, p)$. Les échantillons possibles ont, en moyenne, une taille égale à Np, ce qui revient encore à dire que le nombre p que l'on s'est choisi au début de la procédure correspond au *taux de sondage moyen* appliqué dans la population. Il est bon d'avoir cela à l'esprit lorsqu'on se fixe la valeur de p en début de tirage !

c) Les estimateurs utilisés

Dans le cas du sondage aléatoire simple (PESR), on estime la moyenne μ de la variable d'intérêt y dans la population par la moyenne $\bar{y}$ de la variable y dans l'échantillon S prélevé.

Nous verrons plus tard dans le cours (lorsque nous étudierons l'estimateur de Horvitz-Thompson introduit dans le cadre du sondage à probabilités inégales) que, dans le cas du tirage de Bernoulli, $\bar{y}$ est un estimateur *biaisé* de μ. Dès lors, pour estimer μ sur la base d'un échantillon obtenu par tirage de Bernoulli, nous utiliserons plutôt l'estimateur *non biaisé* défini comme suit :

$$\hat{\mu}_{\mathrm{B}} = \frac{1}{\mathrm{E}(n_S)} \sum_{i \in S} y_i = \frac{1}{Np} \sum_{i \in S} y_i \,.$$

L'espérance et la variance de $\hat{\mu}_{\mathrm{B}}$ se déterminent assez facilement grâce au fait que le tirage de Bernoulli donne lieu à n prélèvements à probabilités égales et sans remise, *indépendants* l'un de l'autre. On vérifie ainsi que :

$$\mathrm{E}(\hat{\mu}_{\mathrm{B}}) = \mu \,,$$

ce qui prouve le caractère non biaisé de $\hat{\mu}_B$, et que :

$$V(\hat{\mu}_B) = \frac{1}{N^2}\left(\frac{1}{p} - 1\right)\sum_{i \in U} y_i^2.$$

Nous verrons ultérieurement (lorsque nous étudierons l'estimateur de Horvitz-Thompson) que la variance de $\hat{\mu}_B$ peut être estimée sans biais par

$$\hat{V}(\hat{\mu}_B) = \frac{1}{N^2 p}\left(\frac{1}{p} - 1\right)\sum_{i \in S} y_i^2.$$

d) Exercice 2.10

Objectif – Cet exercice doit vous permettre de vérifier que vous êtes à même de tirer un échantillon selon la procédure de Bernoulli et d'appliquer ensuite la procédure d'estimation adéquatement.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Un Commissaire aux Comptes doit contrôler un stock composé de $N = 100$ références, d'une valeur totale τ inconnue. Afin d'estimer τ, il souhaite prélever un échantillon de taille moyenne égale à 25, par tirage de Bernoulli, dans la base de sondage.

Le tableau ci-dessous vous indique, pour chaque référence i du stock ($i = 1, \ldots, 100$) la valeur u_i du nombre aléatoire compris entre 0 et 1 généré par Excel (selon une loi uniforme sur l'intervalle [0,1]).

i	u_i
1	0,0625
2	0,1507
3	0,5633
4	0,8957
5	0,4397
6	0,8990
7	0,5178
8	0,0548
9	0,1462
10	0,9193
11	0,3194
12	0,2236
13	0,6972
14	0,5821
15	0,9896
16	0,8346
17	0,7807
18	0,3848
19	0,2222
20	0,7386
21	0,5238
22	0,3288

i	u_i
26	0,1680
27	0,2398
28	0,3136
29	0,8105
30	0,5064
31	0,3306
32	0,5756
33	0,3255
34	0,6641
35	0,1181
36	0,7293
37	0,2468
38	0,1839
39	0,7200
40	0,1377
41	0,8518
42	0,0817
43	0,4428
44	0,4659
45	0,8793
46	0,0597
47	0,6401

i	u_i
51	0,3219
52	0,1744
53	0,9174
54	0,1368
55	0,3446
56	0,5849
57	0,8288
58	0,9659
59	0,2685
60	0,1380
61	0,2544
62	0,8432
63	0,1697
64	0,7498
65	0,0688
66	0,2830
67	0,7953
68	0,9964
69	0,1285
70	0,2166
71	0,6953
72	0,9223

i	u_i
76	0,0827
77	0,6857
78	0,9445
79	0,4547
80	0,9375
81	0,2313
82	0,9097
83	0,3483
84	0,2508
85	0,6211
86	0,9468
87	0,8653
88	0,5391
89	0,6193
90	0,8983
91	0,8245
92	0,0661
93	0,8899
94	0,2400
95	0,8179
96	0,2859
97	0,8323

23	0,8791
24	0,0828
25	0,2773

48	0,1513
49	0,9822
50	0,6105

73	0,0770
74	0,8761
75	0,3759

98	0,6712
99	0,5678
100	0,9668

1. Sur la base de ce tableau, indiquez les numéros des références du stock retenues pour faire partie de l'échantillon.

2. Voici 40 valeurs de la variable d'intérêt y (cette variable associe à chaque référence sa valeur comptable, en euros) :

273 487 610 544 625 260 729 563 261 619

612 430 689 568 372 573 538 274 252 655

614 542 528 675 472 556 610 453 338 445

727 502 512 675 531 570 716 583 335 252

Soit n_s la taille de l'échantillon s que vous avez prélevé au point 1 ; considérez que les n_s premières valeurs y_i parmi les 40 valeurs listées juste avant correspondent aux valeurs comptables des n_s références qui constituent votre échantillon s.

Quelle estimation de la valeur comptable totale du stock (τ) obtenez-vous alors ?

3. En utilisant les mêmes valeurs observées pour la variable d'intérêt y qu'au point 2, quelle estimation de la variance de l'estimateur de τ obtenez-vous ?

4. Déterminez l'intervalle de confiance, au niveau de confiance de 95%, pour la valeur moyenne μ de la variable d'intérêt y dans l'ensemble du stock.

Chapitre 3
Le sondage stratifié

3.1 Pourquoi recourir au sondage stratifié ?

Dans le chapitre précédent (chapitre 2), nous avons vu que, lorsqu'on appliquait un sondage aléatoire simple, on estimait la moyenne μ de la variable d'intérêt y dans la population via la moyenne $\bar{y}$ de cette variable dans l'échantillon.

Nous avons également mis en avant le fait que la variance de $\bar{y}$ était proportionnelle à la variance de la variable y dans la population, et inversement proportionnelle à la taille n de l'échantillon. A taille d'échantillon fixée, plus la variable y prend des valeurs fort différentes les unes des autres dans la population, autrement dit plus la population est hétérogène, plus la précision de $\bar{y}$ se détériore. Dès lors, face à une population fort hétérogène, il faut absolument prévoir une taille n d'échantillon suffisamment grande si l'on veut contrôler la variance de $\bar{y}$ et lui assurer ainsi une précision acceptable.

Malheureusement, le temps, les moyens matériels et le budget que l'on peut consacrer à l'enquête, ne sont pas illimités et ne sont parfois pas suffisants pour pouvoir se permettre un accroissement important du nombre de personnes à interroger. Que pouvons-nous faire dans ce cas ? Comment lutter contre l'effet néfaste de l'hétérogénéité de la population sur la précision de l'estimateur de μ, sans pour autant devoir augmenter la taille de l'échantillon ?

La solution à ce problème est fournie par le recours au SONDAGE STRATIFIE. Comme nous allons le voir au cours de ce chapitre, le sondage stratifié tient compte du découpage — souvent naturel — de la population en différentes sous-populations ou catégories, et de la façon dont l'hétérogénéité de la population s'organise AU SEIN de ces catégories et ENTRE celles-ci.

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

C'est le principe de la stratification : découper la population en sous-populations appelées *strates* et réaliser un sondage aléatoire simple dans chacune d'elles.

On a vu qu'à taille égale un sondage aléatoire simple était plus efficace dans une population homogène que dans une population hétérogène. On va dès lors avoir clairement intérêt à découper la population en strates les plus homogènes possible : chaque sondage partiel, dans une strate particulière, s'effectuera alors de façon efficace. Et l'assemblage des sondages partiels — tous relativement précis — donnera des

résultats plus fiables qu'un sondage aléatoire simple de même taille effectué « en vrac » dans l'ensemble de la population, sans découpage préalable de celle-ci.

Mais la stratification dans un sondage répond souvent aussi à un objectif de réduction des coûts d'enquête ou d'optimisation de sa gestion. C'est en particulier le cas lorsqu'on utilise un critère géographique, comme la région par exemple, pour le découpage de la population : cela permet d'organiser l'administration de l'enquête région par région et de diminuer ainsi les frais de déplacement des enquêteurs.

La section suivante sera consacrée à une formalisation plus rigoureuse du principe général du sondage stratifié. Nous verrons que tout commence par ce qu'on appelle la *stratification de la population*. Nous nous pencherons également sur la manière d'obtenir un échantillon dit *stratifié*, puis verrons comment définir les estimateurs d'une moyenne, d'un total ou d'une proportion dans le cadre d'un sondage stratifié.

3.2 Principe général

3.2.1 La stratification

a) La stratification de la population

Pour le sondage stratifié, tout commence par ce qu'on appelle la stratification de la population. Mais quel est donc le principe général de cette stratification ?

Considérons une population U constituée de N individus ou unités statistiques.

On parle de stratification de cette population lorsque les informations contenues dans la base de sondage nous permettent de partitionner la population en H sous-populations ou *strates*, $U_1, U_2, \ldots, U_H$ de tailles respectives $N_1, N_2, \ldots, N_H$ (voir la figure 3.1). Nous parlerons de façon générale des strates U_h de taille N_h, l'indice h allant de 1 à H. Nous avons ainsi :

$$U = \bigcup_{h=1}^{H} U_h \quad \text{avec } U_h \cap U_k = \emptyset \text{ pour tout } h \neq k \text{ ;}$$

$$N = \sum_{h=1}^{H} N_h \,.$$

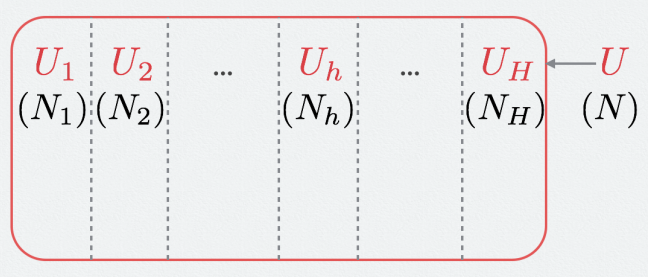


FIGURE 3.1 – *Population stratifiée*

Par exemple, la base de sondage nous indique le sexe de chaque individu de la population : on peut alors découper la population en la strate des hommes et celle des femmes.

Ou la base de sondage nous indique la date de naissance de chaque individu de la population : on peut alors diviser la population en différentes classes d'âges.

Ou encore la base de sondage d'entreprises dont nous disposons indique le secteur d'activité de chacune des entreprises de la population étudiée : on peut alors stratifier la population d'entreprises par secteur d'activité.

L'important est donc que la base de sondage contienne une information auxiliaire nous permettant réellement de *partitionner* la population, c'est-à-dire de classer chaque unité de la population dans une et une seule strate, et cela avant même de mettre en route la procédure d'échantillonnage.

b) L'échantillonnage stratifié

Comment prélève-t-on l'échantillon dans la population stratifiée ?

On a notre population U de taille N partitionnée en H strates, $U_1, U_2, \ldots, U_H$, de tailles respectives $N_1, N_2, \ldots, N_H$.

On commence par se fixer la taille n de l'échantillon S que l'on veut prélever dans cette population U.

On décide alors d'une certaine répartition — on parlera aussi d'*allocation* — entre les différentes strates, de ce nombre total n de prélèvements à effectuer : on choisit d'effectuer n_1 prélèvements du type PESR dans la première strate U_1, n_2 prélèvements PESR dans la deuxième strate U_2, et ainsi de suite, avec

$$n_1 + n_2 + \cdots + n_H = n\,.$$

En d'autres termes, on prélève dans chaque strate U_h un échantillon aléatoire simple S_h de taille n_h.

L'échantillon global S est alors obtenu en réunissant les H sous-échantillons $S_1, S_2, \ldots, S_H$ (voir la figure 3.2) :

$$S = \bigcup_{h=1}^{H} S_h\,.$$

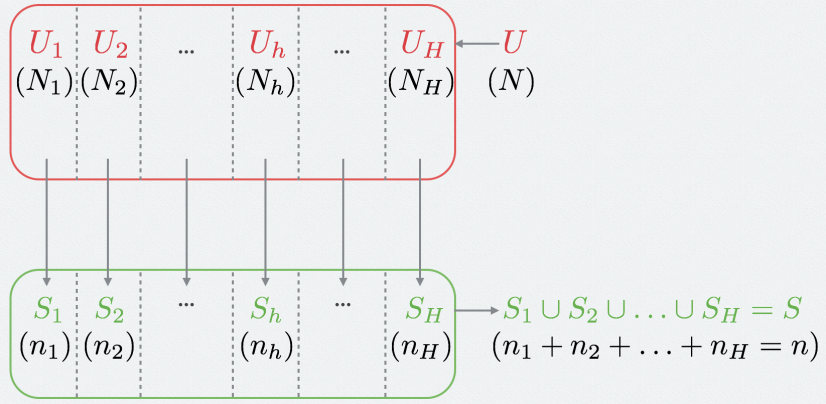


FIGURE 3.2 – *Echantillonnage stratifié*

Le sondage stratifié fait donc en sorte que chaque strate de la population soit représentée dans l'échantillon final par un nombre d'unités statistiques que nous aurons nous-mêmes fixé.

Il est important de noter ici que les H sous-échantillons sont prélevés dans les différentes strates *indépendamment* les uns des autres. Nous verrons dans la suite pourquoi cette indépendance est importante.

Par ailleurs, nous désignerons le taux de sondage global appliqué dans la population par

$$f = n/N$$

et le taux de sondage appliqué dans la strate U_h par

$$f_h = n_h/N_h\,.$$

c) Le plan de sondage

Quelles sont les caractéristiques probabilistes de la procédure de prélèvement de l'échantillon dans le cadre du sondage stratifié ? Répondons à cette question en décrivant le *plan de sondage* correspondant.

L'échantillon stratifié aléatoire S de taille n est obtenu en regroupant les sous-échantillons aléatoires simples S_h de tailles n_h prélevés dans les différentes strates U_h de la population $(h = 1, \dots, H)$:

$$S = S_1 \cup S_2 \cup \dots \cup S_H = \bigcup_{h=1}^{H} S_h \, .$$

Dès lors, le nombre M d'échantillons stratifiés possibles de taille n est tout simplement égal au nombre d'échantillons qu'il est possible d'obtenir en effectuant n_1 prélèvements PESR dans la 1re strate de taille N_1, multiplié par le nombre d'échantillons qu'il est possible d'obtenir en effectuant n_2 prélèvements PESR dans la 2e strate de taille N_2, etc. Autrement dit, si vous vous rappelez ce que nous avons vu dans le chapitre 2 de ce cours, M est égal au nombre de combinaisons de n_1 unités parmi N_1, multiplié par le nombre de combinaisons de n_2 unités parmi N_2, etc. :

$$M = C_{N_1}^{n_1} \times C_{N_2}^{n_2} \times \dots \times C_{N_H}^{n_H} \, .$$

On vérifie par ailleurs que ces M échantillons stratifiés s possibles ont tous la même probabilité d'être sélectionnés, égale à $1/M$.

d) Les probabilités d'inclusion

Que valent les probabilités d'inclusion affectées aux individus de la population U par le plan de sondage stratifié ?

L'échantillon stratifié de taille n est obtenu en effectuant n_1 prélèvements PESR dans la strate n° 1 de taille N_1, n_2 prélèvements PESR dans la strate n° 2 de taille N_2, … , n_H prélèvements PESR dans la strate n° H de taille N_H. Autrement dit, on réalise un sondage aléatoire simple dans la strate n° h $(h = 1, \dots, H)$ avec un taux de sondage $f_h = n_h/N_h$.

Nous pouvons directement en déduire la probabilité d'inclusion de chaque individu de la population. Si i est un individu appartenant à la strate n° h, la probabilité qu'il se retrouve dans l'échantillon final S est égale à la probabilité qu'il appartienne au sous-échantillon aléatoire simple S_h prélevé dans cette strate ; sa probabilité d'inclusion p_i vaut donc le taux de sondage f_h appliqué dans cette strate n° h : pour $i \in U_h$,

$$p_i = \mathrm{P}(i \in S) = \mathrm{P}(i \in S_h) = f_h = \frac{n_h}{N_h} \, .$$

Ainsi, tous les individus d'une même strate ont la même probabilité d'inclusion. Mais si les taux de sondage que l'on s'est fixés varient d'une strate à l'autre, les individus de deux strates différentes n'ont pas la même probabilité d'inclusion.

3.2.2 L'estimateur stratifié d'un paramètre

a) L'estimateur stratifié et ses propriétés

- ***L'estimateur stratifié***

Intéressons-nous à un paramètre-population θ qui peut s'écrire sous la forme d'une combinaison linéaire des paramètres (de même nature) θ_h relatifs aux strates U_h $(h = 1, \dots, H)$ qui partitionnent la population U :

$$\theta = \sum_{h=1}^{H} a_h \theta_h \qquad \text{où } a_h \text{ est une constante réelle de valeur } \textit{connue}.$$

Supposons que nous disposions d'estimateurs $\hat{\theta}_h$ pour les paramètres θ_h $(h = 1, \dots, H)$. Dans ce cas, il est naturel d'estimer θ à l'aide de l'estimateur suivant, généralement appelé **estimateur stratifié** de θ :

$$\hat{\theta}_{\mathrm{ST}} = \sum_{h=1}^{H} a_h \hat{\theta}_h .$$

- ***L'espérance de $\hat{\theta}_{\mathrm{ST}}$***

Nous avons :

$$\mathrm{E}\left(\hat{\theta}_{\mathrm{ST}}\right) = \mathrm{E}\left(\sum_{h=1}^{H} a_h \hat{\theta}_h\right) = \sum_{h=1}^{H} a_h \mathrm{E}\left(\hat{\theta}_h\right).$$

En particulier, si les estimateurs $\hat{\theta}_h$ $(h = 1, \dots, H)$ sont des estimateurs sans biais des paramètres θ_h, nous obtenons :

$$\mathrm{E}\left(\hat{\theta}_{\mathrm{ST}}\right) = \sum_{h=1}^{H} a_h \mathrm{E}\left(\hat{\theta}_h\right) = \sum_{h=1}^{H} a_h \theta_h = \theta \ ;$$

l'estimateur stratifié $\hat{\theta}_{\mathrm{ST}}$ du paramètre-population θ est lui aussi sans biais.

- ***La variance de $\hat{\theta}_{\mathrm{ST}}$***

Puisque les sous-échantillons sont prélevés indépendamment d'une strate à l'autre, les estimateurs $\hat{\theta}_h$ et $\hat{\theta}_k$, avec $h \neq k$, sont des variables aléatoires indépendantes. Dès lors :

$$\mathrm{V}\left(\hat{\theta}_{\mathrm{ST}}\right) = \mathrm{V}\left(\sum_{h=1}^{H} a_h \hat{\theta}_h\right) = \sum_{h=1}^{H} \mathrm{V}\left(a_h \hat{\theta}_h\right) = \sum_{h=1}^{H} a_h^2 \mathrm{V}\left(\hat{\theta}_h\right).$$

La précision de l'estimateur stratifié $\hat{\theta}_{\mathrm{ST}}$ est donc étroitement liée à celle des estimateurs $\hat{\theta}_h$ $(h = 1, \dots, H)$. De petites variances pour les estimateurs $\hat{\theta}_h$ conduisent à une faible variance — donc une bonne précision — pour l'estimateur stratifié $\hat{\theta}_{\mathrm{ST}}$; en revanche, de grandes variances pour les estimateurs $\hat{\theta}_h$ donnent lieu à une variance élevée — donc une mauvaise précision — pour l'estimateur stratifié $\hat{\theta}_{\mathrm{ST}}$. L'efficacité de l'estimateur stratifié $\hat{\theta}_{\mathrm{ST}}$ est ainsi directement liée à l'efficacité des estimateurs dans chacune des strates de la population.

- ***Un estimateur sans biais de la variance de $\hat{\theta}_{\mathrm{ST}}$***

Si l'on dispose d'estimateurs sans biais $\hat{\mathrm{V}}\left(\hat{\theta}_h\right)$ des variances $\mathrm{V}\left(\hat{\theta}_h\right)$ (pour $h = 1, \dots, H$), alors

$$\widehat{\mathrm{V}}(\hat{\theta}_{\mathrm{ST}}) = \sum_{h=1}^{H} a_h^2 \widehat{\mathrm{V}}(\hat{\theta}_h)$$

est un estimateur sans biais de la variance $\mathrm{V}(\hat{\theta}_{\mathrm{ST}})$ de l'estimateur stratifié de θ.

b) L'estimateur stratifié d'une moyenne

• *L'estimateur stratifié de μ*

Comment estimer la moyenne μ de la variable d'intérêt y dans la population ? L'estimateur de μ que nous allons utiliser va être construit à partir des estimateurs des moyennes de la variable y dans les différentes strates de la population.

On vérifie aisément que la moyenne μ de y dans l'ensemble de la population est en fait une moyenne *pondérée* — donc une combinaison linéaire — des moyennes de y dans les différentes strates. Plus précisément, si l'on désigne par μ_h la moyenne de y dans la strate n° h, on a la relation suivante :

$$\mu = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \,.$$

Dans cette relation, le poids affecté à la moyenne μ_h est égal au poids de la strate n° h dans la population ($h = 1, \ldots, H$) ; ce poids a une valeur connue.

Or, nous avons prélevé un échantillon aléatoire simple dans chaque strate de la population. Nous pouvons donc estimer la moyenne μ_h de y dans la strate n° h par la moyenne $\bar{y}_h$ des valeurs que prend la variable y dans le sous-échantillon aléatoire simple S_h :

$$\hat{\mu}_{h;\mathrm{PESR}} = \bar{y}_h = \frac{1}{n_h} \sum_{i \in S_h} y_i \,.$$

Il est dès lors naturel, dans le cadre du sondage stratifié, d'estimer la moyenne globale μ par :

$$\hat{\mu}_{\mathrm{ST}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\mu}_{h;\mathrm{PESR}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \,.$$

• *Remarque*

Il est important de noter ici que $\hat{\mu}_{\mathrm{ST}}$ ne coïncide généralement pas avec la moyenne $\bar{y}$ de la variable y dans l'échantillon global S. En effet, $\hat{\mu}_{\mathrm{ST}}$ est la somme des moyennes-échantillons $\bar{y}_h$ pondérées par les rapports de N_h sur N, alors que $\bar{y}$ peut s'exprimer comme la somme des moyennes-échantillons $\bar{y}_h$ pondérées par les rapports de n_h sur n :

$$\hat{\mu}_{\mathrm{ST}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \quad \text{alors que} \quad \bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h \,.$$

Ainsi, le poids affecté à la moyenne arithmétique $\bar{y}_h$ correspond, dans $\hat{\mu}_{\mathrm{ST}}$, à l'importance relative de la strate U_h dans la population U, alors qu'il correspond, dans $\bar{y}$, à l'importance relative du sous-échantillon S_h dans l'échantillon global S. En général, les rapports N_h/N ne sont pas (tous) égaux aux rapports n_h/n, et $\hat{\mu}_{\mathrm{ST}}$ diffère donc de $\bar{y}$.

Nous reviendrons sur cette remarque dans la section consacrée au sondage stratifié proportionnel.

Afin d'illustrer la procédure de sondage stratifié et d'estimation d'une moyenne-population, prenons un exemple inspiré du livre de 2006 de Pascal Ardilly.

Exemple 1

On dispose d'une population de 1 200 entreprises et on s'intéresse au *nombre moyen d'employés* par entreprise.

Il se fait que l'utilisation de documents fiscaux dans lesquels sont spécifiées différentes caractéristiques des entreprises permet de diviser la population étudiée en 3 strates, correspondant à 3 classes de tailles :

- la première strate U_1 contient les petites entreprises, comptant au plus 20 employés ;
- la deuxième strate U_2 est constituée des entreprises comptant de 21 à 100 employés ;
- la troisième strate U_3 est celle des entreprises comptant plus de 100 employés.

La feuille « U » du fichier Excel « Chap3_Section2_exemple1.xlsx » mis à disposition sur l'UV contient la population que nous voulons étudier, avec :

- dans la colonne A, les numéros de 1 à 1 200 des entreprises ;
- dans la colonne B, le nombre exact d'employés pour chacune des entreprises, autrement dit les valeurs de notre variable d'intérêt y ;
- et, dans la colonne C, la strate à laquelle appartient chaque entreprise.

Les données enregistrées dans cette feuille Excel nous ont permis de déterminer quelques premières caractéristiques de la population et des strates, reprises dans le tableau que voici :

Strates	Tailles	Moyennes
$U_1 : 1-20$	$N_1 = 400$	$\mu_1 = 12$
$U_2 : 21-100$	$N_2 = 500$	$\mu_2 = 50$
$U_3 : 101$ et plus	$N_3 = 300$	$\mu_3 = 185$
Population U	$N = 1\ 200$	$\mu = 71,08$

- la première strate U_1, contenant les petites entreprises de 1 à 20 employés, regroupe au total 400 entreprises, comptant chacune 12 employés en moyenne ;
- la deuxième strate U_2, constituée des entreprises de 21 à 100 employés, rassemble 500 entreprises de 50 employés en moyenne ;
- la troisième et dernière strate U_3 contient 300 entreprises de 185 employés en moyenne.

La moyenne globale μ, qui vaut 71,08, est bien égale à :

$$\frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 + \frac{N_3}{N}\mu_3 = \frac{400}{1\ 200}(12) + \frac{500}{1\ 200}(50) + \frac{300}{1\ 200}(185) = 71,08.$$

Imaginons à présent que l'on ne mette plus à notre disposition qu'une base de sondage dressant la liste des 1 200 entreprises, numérotées de 1 à 1 200, et indiquant également, pour chacune des entreprises, à quelle strate elle appartient : strate 1, 2 ou 3. Tentons alors d'estimer la taille moyenne μ des entreprises à l'aide d'un échantillon stratifié.

Pour commencer, considérons un échantillon stratifié de taille $n = 300$, obtenu en prélevant par sondage PESR 100 entreprises dans chacune des 3 strates : nous avons donc $n_1 = n_2 = n_3 = 100$.

La feuille « ST » du fichier « Chap3_Section2_exemple1.xlsx » présente l'échantillon d'entreprises auquel nous a conduit le hasard et les nombres exacts d'employés déclarés par chaque entreprise de cet échantillon.

Le tableau que voici présente les premières caractéristiques de cet échantillon stratifié :

Echantillons	Tailles	Moyennes
S_1	$n_1 = 100$	$\bar{y}_1 = 13,14$
S_2	$n_2 = 100$	$\bar{y}_2 = 50,27$
S_3	$n_3 = 100$	$\bar{y}_3 = 190,63$
S	$n = 300$	

On observe un nombre moyen d'employés égal à 13,14 dans l'échantillon prélevé dans la première strate, égal à 50,27 dans l'échantillon prélevé dans la deuxième strate et égal à 190,63 dans l'échantillon prélevé dans la troisième strate.

Notez que ces trois moyennes-échantillons nous fournissent des estimations ponctuelles des nombres moyens d'employés par entreprise dans chacune des 3 strates de la population.

L'estimation $\hat{\mu}_{ST}$ du nombre moyen μ d'employés par entreprise, dans la population globale, est alors obtenue en calculant :

$$\hat{\mu}_{ST} = \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + \frac{N_3}{N}\bar{y}_3$$
$$= \frac{400}{1\,200}(13,14) + \frac{500}{1\,200}(50,27) + \frac{300}{1\,200}(190,63) = 72,98\,.$$

Et que vaut la moyenne arithmétique $\bar{y}$ de l'échantillon global S ? Elle vaut 84,68, et a donc une valeur différente de $\hat{\mu}_{ST}$. On vérifie aisément que $\bar{y}$ est bien égale à

$$\frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2 + \frac{n_3}{n} = \frac{100}{300}(13,14) + \frac{100}{300}(50,27) + \frac{100}{300}(190,63)\,.$$

$\bar{y}$ diffère de $\hat{\mu}_{ST}$ tout simplement parce que les rapports n_h/n intervenant dans $\bar{y}$ ne sont pas tous égaux aux rapports N_h/N intervenant dans $\hat{\mu}_{ST}$: ainsi, par exemple, le rapport n_3/n est égal à $100/300 = 0,33$, alors que le rapport N_3/N est égal à $300/1\,200 = 0,25$.

- ***L'espérance de $\hat{\mu}_{\text{ST}}$***

Puisque $\hat{\mu}_{h;\text{PESR}} = \bar{y}_h$ est un estimateur sans biais de μ_h, pour tout $h = 1, \dots, H$,

$$\text{E}(\hat{\mu}_{\text{ST}}) = \text{E}\left(\sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h\right) = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h = \mu \, ;$$

$\hat{\mu}_{\text{ST}}$ est donc un estimateur sans biais de la moyenne-population μ.

En revanche, la moyenne $\bar{y}$ de l'échantillon global S est un estimateur *biaisé* de μ dès que les rapports n_h/n ne sont pas égaux aux rapports N_h/N. Attention donc à ne pas utiliser $\bar{y}$ pour estimer μ dans ce cas !

- ***La variance de $\hat{\mu}_{\text{ST}}$***

Puisque $\hat{\mu}_{\text{ST}}$ est une moyenne pondérée des estimateurs *indépendants* $\bar{y}_h$ des moyennes μ_h $(h = 1, \dots, H)$, nous avons :

$$\text{V}(\hat{\mu}_{\text{ST}}) = \text{V}\left(\sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \text{V}(\bar{y}_h).$$

Or, puisqu'on applique un sondage PESR de taille n_h dans la strate U_h de taille N_h, pour tout $h = 1, \dots, H$, on a :

$$\text{V}(\bar{y}_h) = (1 - f_h) \frac{\sigma^2_{h;\text{corr}}}{n_h}$$

où

$$\sigma^2_{h;\text{corr}} = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \mu_h)^2$$

est la variance corrigée de la variable d'intérêt y dans la strate n° h, et $f_h = n_h/N_h$ est le taux de sondage appliqué dans cette même strate.

Dès lors :

$$\text{V}(\hat{\mu}_{\text{ST}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h) \frac{\sigma^2_{h;\text{corr}}}{n_h} \, .$$

Que nous indique cette expression de la variance de l'estimateur stratifié de μ ? Elle met en évidence le fait que la variance de $\hat{\mu}_{\text{ST}}$ est d'autant plus petite — donc $\hat{\mu}_{\text{ST}}$ est d'autant plus précis — que les variances $\sigma^2_{h;\text{corr}}$ sont petites. Pour que le sondage stratifié soit efficace, il faut que les sondages aléatoires simples réalisés dans les différentes strates de la population soient eux-mêmes efficaces, ce qui est d'autant plus facile que les strates sont bien homogènes.

Cette remarque est vraiment fondamentale et est réellement la clé de la stratification. Si l'on stratifie la population dans le but d'obtenir une bonne précision pour l'estimation de μ sans avoir à prévoir une taille d'échantillon excessive, il faut veiller à ne pas découper la population « n'importe comment ». De *bonnes* strates sont avant tout des strates aussi homogènes que possible, c'est-à-dire des strates au sein desquelles les valeurs de la variable y sont aussi peu dispersées que possible. En pratique donc, il faut chercher à définir les strates de telle sorte à ce qu'elles soient constituées d'individus ou

d'unités statistiques qui ont tendance à se ressembler pour ce qui est de la variable d'intérêt.

C'est ainsi que les populations de ménages ou d'individus, pour les enquêtes usuelles, sont classiquement stratifiées par région croisée par type d'habitat (ou taille des communes) ; les populations d'entreprises sont stratifiées par secteur et par taille (exprimée en nombre de salariés ou en importance du chiffre d'affaires) ; les populations d'exploitations agricoles sont stratifiées par classes de surface cultivée ; les populations de jeunes sortis de l'enseignement supérieur sont stratifiées par discipline. Pour l'étude du lancement d'un nouveau produit financier, il sera pertinent de stratifier la population selon des critères liés aux revenus, à l'âge, éventuellement au sexe, c'est-à-dire de découper la population selon des facteurs susceptibles d'expliquer les différences de comportement financier.

Nous reviendrons sur cette idée de « bonne » stratification un peu plus loin dans ce chapitre, lorsque nous étudierons le sondage stratifié dit « proportionnel ».

Exemple 1 (suite)

Revenons à notre exemple consacré à l'estimation du nombre moyen d'employés par entreprise.

Le tableau que voici reprend les tailles de la population et de ses strates, ainsi que les moyennes du nombre d'employés par entreprise dans chacune des strates et dans la population globale. Nous avions déjà vu ces caractéristiques auparavant.

h	1	2	3	Population
N_h	400	500	300	$N = 1\ 200$
μ_h	12	50	185	$\mu = 71,08$
σ_h^2	19,54	114,56	5 552,84	$\sigma^2 = 6\ 035,53$
$\sigma_{h;\text{corr}}^2$	19,58	114,79	5 571,41	$\sigma_{\text{corr}}^2 = 6\ 040,56$
n_h	100	100	100	$n = 300$

Ce tableau présente également les valeurs des variances classiques et des variances corrigées du nombre d'employés dans les 3 strates et dans l'ensemble de la population. Si vous avez téléchargé le fichier Excel relatif à cet exemple, vous trouverez le calcul de ces variances dans la première feuille de ce fichier.

Il ressort clairement des valeurs de ces variances que c'est au sein de la première strate, constituée des plus petites entreprises, que les nombres d'employés sont les moins dispersés ; en revanche, la troisième strate — celle des plus grandes entreprises — est beaucoup plus hétérogène pour ce qui est du nombre d'employés.

La dernière ligne du tableau rappelle les tailles des sous-échantillons prélevés dans les différentes strates, ainsi que la taille de l'échantillon global.

Les éléments repris dans ce tableau nous permettent de déterminer la valeur de la variance de $\hat{\mu}_{\text{ST}}$:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{\mathrm{ST}}) &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(1-f_h)\frac{\sigma^2_{h;\mathrm{corr}}}{n_h}\\ &= \left(\frac{400}{1\,200}\right)^2\left(1-\frac{100}{400}\right)\frac{19{,}58}{100}+\left(\frac{500}{1\,200}\right)^2\left(1-\frac{100}{500}\right)\frac{114{,}79}{100}\\ &\quad+\left(\frac{300}{1\,200}\right)^2\left(1-\frac{100}{300}\right)\frac{5\,571{,}41}{100}\\ &= 2{,}50\,.\end{aligned}$$

Mais qu'aurait valu la variance de l'estimateur de μ dans le cas d'un sondage aléatoire simple de taille n égale à 300 ?

Si l'on avait mis en œuvre un sondage PESR de taille 300, on aurait estimé la moyenne-population μ via la simple moyenne-échantillon $\bar{y}$, et celle-ci aurait eu pour variance :

$$(1-f)\frac{\sigma^2_{\mathrm{corr}}}{n} = \left(1-\frac{300}{1\,200}\right)\frac{6\,040{,}56}{300} = 15{,}10\,.$$

L'efficacité de la stratification est ici flagrante ! Le simple fait d'avoir stratifié notre population d'entreprises, très hétérogène du point de vue des tailles de celles-ci, en 3 strates plus homogènes a permis de réduire drastiquement la variance de l'estimateur de μ.

Ainsi, le sondage stratifié permet d'estimer la moyenne-population avec une bien meilleure précision que le sondage aléatoire simple, et cela sans exiger la moindre augmentation de la taille de l'échantillon !

- ***Un estimateur sans biais de la variance de $\hat{\mu}_{\mathrm{ST}}$***

Comme d'habitude, nous ne nous contentons pas de l'expression théorique de la variance de l'estimateur stratifié de μ. Nous allons chercher à l'estimer, sans biais, pour pouvoir ensuite utiliser cette estimation pour déterminer un intervalle de confiance pour μ.

Il suffit de se rappeler les résultats vus dans le chapitre 2 pour le sondage aléatoire simple. Nous avons vu que nous pouvions estimer sans biais les variances corrigées $\sigma^2_{h;\mathrm{corr}}$ dans les strates via les variances corrigées $s^2_{h;\mathrm{corr}}$ des différents sous-échantillons S_h. Nous pouvons directement en déduire l'expression de $\widehat{\mathrm{V}}(\hat{\mu}_{\mathrm{ST}})$:

$$\widehat{\mathrm{V}}(\hat{\mu}_{\mathrm{ST}}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(1-f_h)\frac{s^2_{h;\mathrm{corr}}}{n_h}\,.$$

L'intervalle de confiance pour la moyenne-population μ, au niveau de confiance de 95%, est alors l'intervalle

$$\left[\hat{\mu}_{\mathrm{ST}} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\mu}_{\mathrm{ST}})}\right].$$

Illustrons les résultats relatifs à l'estimation de la variance de l'estimateur stratifié de la moyenne-population sur notre exemple.

Exemple 1 (suite)

Dans la feuille « ST » du fichier Excel présentant l'échantillon stratifié d'entreprises prélevé dans notre exemple, nous avons calculé les variances corrigées du nombre d'employés dans les différents sous-échantillons.

h	1	2	3	Population
N_h	400	500	300	$N = 1\ 200$
n_h	100	100	100	$n = 300$
$s^2_{h;\text{corr}}$	18, 51	108, 24	5 727, 12	

On obtient ainsi comme estimation de la variance de $\hat{\mu}_{\text{ST}}$:

$$\begin{aligned}\widehat{\mathrm{V}}(\hat{\mu}_{\text{ST}}) &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(1-f_h)\frac{s^2_{h;\text{corr}}}{n_h}\\ &= \left(\frac{400}{1\ 200}\right)^2\left(1-\frac{100}{400}\right)\frac{18{,}51}{100}+\left(\frac{500}{1\ 200}\right)^2\left(1-\frac{100}{500}\right)\frac{108{,}24}{100}\\ &\quad+\left(\frac{300}{1\ 200}\right)^2\left(1-\frac{100}{300}\right)\frac{5\ 727{,}12}{100}\\ &= 2{,}55\,.\end{aligned}$$

L'intervalle de confiance pour μ, au niveau de confiance de 95%, a donc pour bornes :

$$\begin{aligned}\left[\hat{\mu}_{\text{ST}} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\mu}_{\text{ST}})}\right] &= \left[72{,}98 \pm 1{,}96\sqrt{2{,}55}\right]\\ &= [72{,}98 \pm 3{,}13] = [69{,}85\,;\,76{,}11]\,.\end{aligned}$$

En conclusion, on peut être sûr à 95% que le nombre moyen d'employés par entreprise est compris entre, disons, 70 et 76.

c) L'estimateur stratifié d'un total

• *L'estimateur stratifié de τ*

L'expression de l'estimateur stratifié du total τ de la variable d'intérêt y dans la population se déduit directement de celle de l'estimateur stratifié de sa moyenne μ. En effet, puisque $\tau = N\mu$, il est naturel d'estimer τ à l'aide de l'estimateur

$$\hat{\tau}_{\text{ST}} = N\hat{\mu}_{\text{ST}} = N\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h = \sum_{h=1}^{H}N_h\bar{y}_h\,.$$

Nous aurions pu arriver à cette expression de $\hat{\tau}_{\text{ST}}$ par une autre voie. Le total de la variable y dans l'ensemble de la population U est bien évidemment égal à la somme des totaux de y dans les différentes strates de la population :

$$\tau = \sum_{h=1}^{H}\tau_h\,.$$

Or, le total τ_h de y dans la strate n° h est estimé par $\hat{\tau}_{h;\text{PESR}} = N_h \bar{y}_h$, puisque le sondage effectué dans la strate est un sondage aléatoire simple. On retrouve ainsi l'expression de $\hat{\tau}_{\text{ST}}$ présentée ci-dessus :

$$\hat{\tau}_{\text{ST}} = \sum_{h=1}^{H} \hat{\tau}_{h;\text{PESR}} = \sum_{h=1}^{H} N_h \bar{y}_h \, .$$

Exemple 1 (suite)

Revenons à notre exemple portant sur une population d'entreprises et estimons maintenant le nombre total τ d'employés dans l'ensemble des 1 200 entreprises de cette population.

Nous avions estimé le nombre moyen d'employés par entreprise par $\hat{\mu}_{\text{ST}}$, égal à 72,98. Nous en déduisons directement l'estimation de τ : $\hat{\tau}_{\text{ST}}$ est égal à 1 200 fois 72,98, c'est-à-dire à 87 580, aux arrondis près.

En d'autres termes, on estime que la population d'entreprises considérée compte au total près de 87 580 employés.

- ***L'espérance de $\hat{\tau}_{\text{ST}}$***

On a :

$$\text{E}(\hat{\tau}_{\text{ST}}) = \sum_{h=1}^{H} \text{E}\left(\hat{\tau}_{h;\text{PESR}}\right) = \sum_{h=1}^{H} \tau_h = \tau \, .$$

$\hat{\tau}_{\text{ST}}$ est donc un estimateur non biaisé du total-population τ.

- ***La variance de $\hat{\tau}_{\text{ST}}$***

On a :

$$\begin{aligned} \text{V}(\hat{\tau}_{\text{ST}}) &= \text{V}(N\hat{\mu}_{\text{ST}}) = N^2 \text{V}(\hat{\mu}_{\text{ST}}) \\ &= \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{\sigma_{h;\text{corr}}^2}{n_h} = \sum_{h=1}^{H} \text{V}\left(\hat{\tau}_{h;\text{PESR}}\right) . \end{aligned}$$

- ***Un estimateur sans biais de la variance de $\hat{\tau}_{\text{ST}}$***

On peut estimer sans biais la variance de $\hat{\tau}_{\text{ST}}$ à l'aide de

$$\widehat{\text{V}}(\hat{\tau}_{\text{ST}}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{h;\text{corr}}^2}{n_h} \, .$$

d) L'estimateur stratifié d'une proportion

- ***L'estimateur stratifié de π***

Comment estimer une proportion π dans le cadre d'un sondage stratifié ? Il suffit de se rappeler qu'une proportion s'estime exactement de la même façon qu'une moyenne, puisque toute proportion peut être considérée comme la moyenne d'une variable y dichotomique (n'ayant que deux valeurs possibles : 1 ou 0) ; rappelez-vous ce que nous avons vu au chapitre 2 de ce cours !

La proportion π d'individus de la population qui possèdent une certaine caractéristique est liée aux proportions π_h d'individus possédant cette même caractéristique dans les strates U_h $(h = 1, \dots, H)$ par la relation suivante :

$$\pi = \sum_{h=1}^{H} \frac{N_h}{N} \pi_h \, .$$

Il est dès lors naturel d'estimer π à l'aide de l'estimateur stratifié

$$\hat{\pi}_{\text{ST}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\pi}_{h;\text{PESR}}$$

où $\hat{\pi}_{h;\text{PESR}} = \hat{\pi}_h$ est la proportion de personnes présentant la caractéristique considérée dans le sous-échantillon S_h.

- ***Remarque***

Notez que $\hat{\pi}_{\text{ST}}$ n'est généralement pas égal à la proportion $\hat{\pi}$ d'individus de l'échantillon global S ayant la caractéristique en question. En effet,

$$\hat{\pi}_{\text{ST}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\pi}_h \quad \text{alors que} \quad \hat{\pi} = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\pi}_h \; ;$$

$\hat{\pi}_{\text{ST}}$ ne coïncidera donc avec $\hat{\pi}$ que si

$$\frac{N_h}{N} = \frac{n_h}{n} \qquad \text{pour tout } h = 1, \dots, H.$$

Illustrons l'estimation d'une proportion dans le cadre du sondage stratifié à l'aide d'un autre exemple.

Exemple 2

On souhaite estimer la proportion π d'employés qui possèdent au moins un véhicule, parmi les 7 500 employés d'une grosse société.

Pour chaque individu de la base de sondage, on dispose de la valeur de son revenu. Cette information auxiliaire nous permet de constituer 3 strates dans la population : la première strate est celle des 3 500 employés de revenus faibles, la deuxième strate contient les 2 000 employés de revenus moyens et la troisième strate rassemble les 2 000 employés de revenus élevés.

U_h	N_h
$h = 1$ (revenus faibles)	3 500
$h = 2$ (revenus moyens)	2 000
$h = 3$ (revenus élevés)	2 000
Total	7 500

On décide de prélever un échantillon stratifié de taille 1 000, en effectuant 500 tirages PESR dans la première strate, 300 tirages PESR dans la deuxième strate et 200 tirages PESR dans la troisième strate.

Il se fait que, dans l'échantillon aléatoire simple S_1 prélevé dans la première strate, 13% des employés déclarent posséder au moins un véhicule ;

dans l'échantillon S_2 prélevé dans la deuxième strate, on observe une proportion de 45% d'employés ayant au moins un véhicule,
et cette proportion s'élève à 50% dans l'échantillon S_3 prélevé dans la troisième strate.

U_h	N_h	n_h	$\widehat{\pi}_h$
$h = 1$ (revenus faibles)	3 500	500	0,13
$h = 2$ (revenus moyens)	2 000	300	0,45
$h = 3$ (revenus élevés)	2 000	200	0,50
Total	7 500	1 000	

L'estimation $\hat{\pi}_{ST}$ de π vaut ainsi :

$$\hat{\pi}_{ST} = \frac{3\,500}{7\,500}(0{,}13) + \frac{2\,000}{7\,500}(0{,}45) + \frac{2\,000}{7\,500}(0{,}50) = 0{,}314 = 31{,}4\%.$$

En conclusion, notre échantillon stratifié selon les trois niveaux de revenus nous permet d'estimer qu'un peu moins d'un tiers seulement des employés de la société considérée possèdent au moins un véhicule.

- ***L'espérance de $\widehat{\pi}_{ST}$***

On a :

$$\mathrm{E}(\hat{\pi}_{ST}) = \sum_{h=1}^{H} \frac{N_h}{N} \mathrm{E}(\hat{\pi}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \pi_h = \pi\,.$$

$\hat{\pi}_{ST}$ est donc un estimateur non biaisé de la proportion-population π.

- ***La variance de $\widehat{\pi}_{ST}$***

On a :

$$\mathrm{V}(\hat{\pi}_{ST}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \mathrm{V}(\hat{\pi}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h - 1} \frac{\pi_h(1-\pi_h)}{n_h}.$$

- ***Un estimateur sans biais de la variance de $\widehat{\pi}_{ST}$***

On peut estimer sans biais la variance de $\hat{\pi}_{ST}$ à l'aide de

$$\widehat{\mathrm{V}}(\hat{\pi}_{ST}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \widehat{\mathrm{V}}(\hat{\pi}_h)$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h) \frac{\hat{\pi}_h(1-\hat{\pi}_h)}{n_h - 1}.$$

e) Exercice 3.1

Objectif – Cet exercice doit vous permettre de vérifier que vous pouvez : (i) estimer correctement une proportion, une moyenne ou un total ; (ii) estimer correctement les variances des estimateurs d'une proportion, d'une moyenne ou d'un total, dans le cadre d'un sondage stratifié.

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très

rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Le personnel d'une entreprise comporte 4 225 employés, 1 300 cadres moyens et 975 cadres supérieurs.

Par sondage auprès de 1 000 membres du personnel, la direction de l'entreprise désire estimer un niveau de satisfaction moyen de son personnel, celui-ci correspondant à la moyenne d'un certain indice de satisfaction, assimilable à une variable quantitative positive y dont on peut déterminer la valeur (sur une échelle de 0 à 20) chez un individu à partir d'un ensemble de questions. La direction désire également estimer la part de son personnel désireuse de pouvoir suivre à court terme l'une ou l'autre formation courte en lien étroit avec son activité professionnelle.

L'échantillon de taille 1 000 est sélectionné par sondage stratifié : on prélève 500 personnes dans la strate des employés, 250 personnes dans la strate des cadres moyens et 250 personnes dans la strate des cadres supérieurs. Les données récoltées auprès des personnes interrogées ont permis de compléter le tableau suivant :

Echantillon	Employés	Cadres moyens	Cadres supérieurs
Taille	500	250	250
Proportion de personnes désireuses de suivre une formation	0,22	0,15	0,10
Indice de satisfaction moyen	11	12	15
Variance corrigée des indices de satisfaction	9	25	36

Partie 1

1.1) Estimez, à partir de cet échantillon stratifié, l'indice de satisfaction moyen pour l'ensemble du personnel de l'entreprise.

1.2) Si vous aviez estimé l'indice de satisfaction moyen pour l'ensemble du personnel de l'entreprise en oubliant qu'il y a eu stratification et en considérant que l'échantillon a été prélevé par sondage aléatoire simple, quelle estimation auriez-vous obtenue ?

1.3) Estimez la variance de l'estimateur stratifié de l'indice de satisfaction moyen du personnel de l'entreprise.

1.4) Que vaut, dans le sondage considéré ici, la marge d'erreur (ou incertitude absolue) associée à l'estimation de l'indice de satisfaction moyen du personnel de l'entreprise ?

Partie 2

2.1) Estimez, à partir de cet échantillon stratifié, la part du personnel de l'entreprise désireuse de suivre une formation courte dans le cadre de son activité professionnelle.

2.2) Si vous aviez estimé cette proportion en oubliant qu'il y a eu stratification et en considérant que l'échantillon a été prélevé par sondage aléatoire simple, quelle estimation auriez-vous obtenue ?

2.3) Estimez la variance de l'estimateur stratifié de la proportion des membres du personnel qui sont désireux de suivre une formation courte dans le cadre de leur activité professionnelle.

2.4) Que vaut, dans le sondage considéré ici, la marge d'erreur (ou incertitude absolue) associée à l'estimation de la part du personnel désireuse de suivre une formation courte dans le cadre de son activité professionnelle ?

Partie 3

3.1) Estimez, à partir de cet échantillon stratifié, le nombre de membres du personnel de l'entreprise désireux de suivre une formation courte dans le cadre de leur activité professionnelle.

3.2) Estimez la variance de l'estimateur stratifié du nombre de membres du personnel qui sont désireux de suivre une formation courte dans le cadre de leur activité professionnelle.

3.3) Que vaut, dans le sondage considéré ici, la marge d'erreur (ou incertitude absolue) associée à l'estimation du nombre de membres du personnel qui sont désireux de suivre une formation courte dans le cadre de leur activité professionnelle ?

3.3 Le sondage stratifié proportionnel (STP)

3.3.1 Le principe de base

a) L'allocation proportionnelle

Une question fondamentale se pose lorsqu'on décide de réaliser un sondage stratifié. Comment faut-il répartir les n prélèvements entre les différentes strates de la population ? En d'autres termes, comment choisit-on les tailles n_h des sous-échantillons sélectionnés dans les strates ?

La solution la plus simple, et de loin la plus utilisée, consiste à appliquer le même taux de sondage dans chacune des strates de la population.

Si l'on a décidé de prélever au total n individus dans la population, autrement dit si l'on a décidé d'appliquer dans la population U un taux de sondage $f = n/N$, on appliquera ce même taux de sondage f dans chaque strate.

Si l'on a décidé, par exemple, de prélever 5 % des unités de la population, on tirera 5 % des unités de chacune des strates.

On effectuera de la sorte $n_h = fN_h$ prélèvements PESR dans la strate numéro h. Bien évidemment, en pratique, si fN_h n'est pas un nombre entier, on l'arrondit à l'entier le plus proche pour obtenir la valeur de n_h.

Le fait que, pour tout $h = 1, \dots, H$,

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

implique que, pour tout $h = 1, \dots, H$,

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

Cette dernière égalité nous indique que les strates ont, dans l'échantillon S, des poids n_h/n égaux à leurs poids N_h/N dans la population. En d'autres termes, les différentes catégories définissant la stratification présentent la même répartition dans l'échantillon et dans la population.

C'est ainsi que, dans un sondage stratifié par sexe, obtenu en appliquant le même taux de sondage dans la strate des hommes et dans celle des femmes, on retrouve la même répartition hommes-femmes dans l'échantillon et dans la population.

Si un échantillon d'entreprises est stratifié par secteur d'activité, les proportions d'entreprises par secteur sont identiques dans l'échantillon et dans la base de sondage.

L'échantillon apparaît ainsi comme un modèle réduit de la population. On parle alors pour S d'échantillon stratifié *proportionnel,* ou encore d'échantillon stratifié

représentatif. Rappelez-vous ici de ce que nous avions dit de la notion de représentativité dans le premier chapitre de ce cours (cf. Section 1.3).

Le sondage stratifié proportionnel — que nous désignerons brièvement par le sigle STP — possède des propriétés ou caractéristiques particulièrement intéressantes pour le sondeur. Nous allons regarder cela de plus près dans la suite de cette section !

b) Les probabilités d'inclusion

Nous avons vu que, de manière générale, dans le cadre du sondage stratifié, tout individu i appartenant à la strate U_h avait une probabilité d'inclusion égale au taux de sondage f_h appliqué dans cette strate. Ainsi, tous les individus d'une même strate ont la même probabilité d'être sélectionnés, mais les individus de deux strates différentes ont des probabilités différentes de se retrouver dans l'échantillon dès le moment où l'on applique des taux de sondage différents dans ces deux strates.

Dans le cas du sondage stratifié proportionnel, puisqu'on applique le *même* taux de sondage dans toutes les strates, les probabilités d'inclusion sont égales pour tous les individus de la base de sondage : ceux-ci ont tous la même probabilité d'être sélectionnés pour faire partie de l'échantillon, égale au taux de sondage unique f. Pour tout $i \in U$:

$$p_i = \mathrm{P}(i \in S) = f = \frac{n}{N}.$$

c) Les estimateurs STP

Le sondage stratifié proportionnel donne lieu à une seconde propriété, intéressante car simplificatrice. Le fait que

$$\frac{N_h}{N} = \frac{n_h}{n} \qquad \text{pour tout } h = 1, \ldots, H,$$

implique que l'estimateur $\hat{\mu}_{\mathrm{STP}}$ de la moyenne-population dans le cadre du sondage stratifié proportionnel coïncide avec la moyenne $\bar{y}$ de la variable d'intérêt dans l'échantillon global S : en effet,

$$\hat{\mu}_{\mathrm{STP}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} = \frac{1}{n} \sum_{i \in S} y_i \,.$$

Il s'ensuit que $\hat{t}_{\mathrm{STP}} = N\bar{y}$.

De la même façon, l'estimateur $\hat{\pi}_{\mathrm{STP}}$ de la proportion-population π dans le cas du sondage stratifié proportionnel coïncide avec la proportion $\hat{\pi}$ dans l'échantillon global.

On dit que l'échantillon est *autopondéré* puisque, dans l'expression de l'estimateur STP, tous les individus de l'échantillon ont un poids identique. Parfois même, on dit que le sondage *se dépouille comme un recensement*, parce que, comme dans l'exploitation d'un recensement, il n'y a pas besoin de se préoccuper de l'appartenance des individus à telle ou telle catégorie ou strate.

3.3.2 La décomposition de la variance

Mon objectif, dans la suite de cette section, est de vous montrer que l'on peut aisément quantifier le gain en précision que permet le sondage stratifié proportionnel par

rapport au sondage aléatoire simple de même taille n. L'analyse de ce gain nous permettra par ailleurs d'affiner notre définition de ce qu'est une « bonne » stratification.

Que ce soit dans le cas du sondage stratifié proportionnel ou dans celui du sondage aléatoire simple, on estime la moyenne-population μ à l'aide de la moyenne $\bar{y}$ des valeurs que prend la variable d'intérêt y dans l'échantillon S.

Mais, si cette moyenne-échantillon est bien sans biais sous ces deux plans de sondage, elle ne jouit pas de la même précision. Elle est en réalité toujours plus précise sous le plan de sondage stratifié proportionnel que sous le plan de sondage aléatoire simple. Nous allons établir ce résultat de manière rigoureuse ci-dessous.

Mais pour cela, il nous faut avant tout nous pencher sur une caractéristique fondamentale de la variance de la variable d'intérêt y dans la population stratifiée U : cette caractéristique se traduit sous la forme d'une formule dite « de décomposition » de la variance.

a) La formule de décomposition de la variance

La variance σ^2 de la variable d'intérêt y dans la population stratifiée U peut se décomposer en une somme de deux termes, comme suit :

$$\sigma^2 = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu)^2$$
$$= (\mathrm{I}) + (\mathrm{II}) .$$

Le terme (I) est une somme pondérée des variances de y dans les différentes strates de la population, la variance σ_h^2 dans la strate n° h ayant comme poids le rapport N_h/N, c'est-à-dire l'importance relative de la strate n° h dans la population. Ce premier terme apparaît ainsi comme une mesure globale de la dispersion de la variable y *à l'intérieur* même des strates : ceci explique pourquoi ce premier terme porte le nom de **variance intra-strates** et est désigné par σ^2_{intra}.

Le terme (II) quantifie la dispersion des moyennes μ_h de y dans les différentes strates autour de la moyenne globale μ de y dans la population. Ce terme a une valeur d'autant plus grande que les moyennes de y varient fortement d'une strate à l'autre, ou encore sont nettement contrastées les unes par rapport aux autres. Il mesure ainsi la dispersion de la variable y *entre* les strates, et porte d'ailleurs le nom de **variance inter-strates**. On désigne ce second terme par σ^2_{inter}.

Découper la population en strates bien homogènes signifie que l'on fait en sorte de rendre la variance *intra*-strates aussi petite que possible ; dans ce cas, ce seront les contrastes entre strates qui expliqueront, pour la plus large part, la variance de y dans la population.

Mais faire en sorte que ce soit la variance *inter*-strates qui ait le plus de poids dans la décomposition de σ^2 signifie aussi que l'on stratifie la population selon une variable

auxiliaire X fortement liée ou associée à la variable d'intérêt Y : dans ce cas là, en effet, le niveau moyen de Y varie fortement d'une strate à l'autre, autrement dit d'une modalité à l'autre de la variable auxiliaire X. Ainsi, par exemple, si la moyenne de Y dans la strate des hommes est fort différente de la moyenne de Y dans la strate des femmes, c'est bien qu'il y a un lien, une association étroite entre la variable Y et la variable auxiliaire « sexe » qui a donné lieu à la stratification.

Par conséquent, découper la population en strates homogènes, c'est la découper selon les modalités d'une variable auxiliaire X fortement liée à la variable d'intérêt Y.

b) Le rapport de corrélation

On peut quantifier l'intensité de l'association entre la variable auxiliaire X définissant la stratification — X est donc une variable qualitative à H modalités, auxquelles correspondent les H strates de la population U — et la variable d'intérêt Y à l'aide du **rapport de corrélation** η dont le carré se définit comme suit :

$$\eta^2 = \frac{\sigma^2_{\text{inter}}}{\sigma^2} = \frac{\sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu)^2}{\frac{1}{N} \sum_{h=1}^{H} \sum_{i \in U_h} (y_i - \mu)^2}.$$

De par la formule de décomposition de la variance σ^2, il est clair que

$$0 \leq \eta^2 \leq 1.$$

L'égalité $\eta^2 = 0$ signifie que $\mu_h = \mu$ pour tout $h = 1, \ldots, H$: toutes les strates ont la même moyenne, $\sigma^2_{\text{inter}} = 0$ et il n'y a aucune association entre la variable de stratification X et la variable d'intérêt Y.

L'égalité $\eta^2 = 1$ signifie quant à elle que $\sigma^2_{\text{inter}} = \sigma^2$ et donc que $\sigma^2_{\text{intra}} = 0$: la variable d'intérêt Y est constante au sein de chaque strate et ce sont les contrastes entre strates qui expliquent la totalité de la variance de Y dans la population. Dans ce cas, il existe un lien déterministe entre la variable de stratification X et la variable d'intérêt Y.

En conclusion, le coefficient η^2 prendra une valeur d'autant plus proche de 1 que la variable auxiliaire de stratification X est fortement liée à la variable d'intérêt Y, et donc, de manière équivalente, que les H strates qui partitionnent la population sont bien homogènes (c'est-à-dire que les individus d'une même strate ont des comportements similaires pour ce qui est de la variable Y).

Exemple 1 (suite)

Illustrons la formule de décomposition de la variance à l'aide de notre exemple relatif à la population d'entreprises et voyons ce qu'elle nous donne comme information sur la façon dont s'organise l'hétérogénéité de la population DANS et ENTRE les trois strates.

Reprenons les caractéristiques de la population déjà présentées précédemment :

h	1	2	3	Population
N_h	400	500	300	$N = 1\ 200$
μ_h	12	50	185	$\mu = 71,08$
σ_h^2	19,54	114,56	5 552,84	$\sigma^2 = 6\ 035,53$

La variance *intra*-strates est égale à :

$$\begin{aligned}\sigma^2_{\text{intra}} &= \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 + \frac{N_3}{N}\sigma_3^2 \\ &= \frac{400}{1\ 200}(19{,}54) + \frac{500}{1\ 200}(114{,}56) + \frac{300}{1\ 200}(5\ 552{,}84) \\ &= 1\ 442{,}45\ .\end{aligned}$$

Quant à la variance *inter*-strates, elle s'élève à :

$$\begin{aligned}\sigma^2_{\text{inter}} &= \frac{N_1}{N}(\mu_1 - \mu)^2 + \frac{N_2}{N}(\mu_2 - \mu)^2 + \frac{N_3}{N}(\mu_3 - \mu)^2 \\ &= \frac{400}{1\ 200}(12 - 71{,}08)^2 + \frac{500}{1\ 200}(50 - 71{,}08)^2 + \frac{300}{1\ 200}(185 - 71{,}08)^2 \\ &= 4\ 593{,}08\ .\end{aligned}$$

On vérifie que la somme de la variance intra-strates et de la variance inter-strates redonne bien, aux erreurs d'arrondis près, la variance globale σ^2.

On peut également noter que la dispersion de y dans la population s'explique principalement par la dispersion des moyennes de y dans les différentes strates autour de la moyenne μ de y dans la population globale : en effet, le rapport η^2 de la variance inter-strates sur la variance globale σ^2 est égal à 0,76, ce qui revient à dire que la variance inter-strates constitue 76% de la variance globale.

D'autre part, les strates sont plus homogènes que la population dans son ensemble : la variance intra-strates ne représente que 24% de la variance globale.

Ces caractéristiques de la décomposition de la variance globale sont liées au fait que la variable selon laquelle nous avons stratifié notre population d'entreprises est très fortement liée à la variable d'intérêt. Effectivement, la variable d'intérêt est le nombre d'employés par entreprise et nous avons stratifié la population en trois classes de tailles d'entreprise, selon les informations fournies par divers documents fiscaux.

3.3.3 STP versus PESR

Déterminons à présent le gain en précision qu'offre le sondage stratifié proportionnel (ou STP) par rapport au sondage aléatoire simple (ou PESR) de même taille n. Nous allons pour cela comparer la variance de l'estimateur de μ sous le plan de sondage STP avec sa variance sous le plan de sondage PESR ; cette comparaison peut se faire en déterminant le rapport de ces deux variances. Ce rapport particulier porte le nom d'*effet de sondage* du STP par rapport au PESR.

Que ce soit dans le cadre du sondage stratifié proportionnel ou celui du sondage aléatoire simple, on estime μ par le même estimateur : la moyenne-échantillon $\bar{y}$. Mais pour éviter toute confusion, nous désignerons cet estimateur par $\hat{\mu}_{\text{STP}}$ lorsqu'il est considéré sous le plan de sondage stratifié proportionnel et par $\hat{\mu}_{\text{PESR}}$ lorsqu'il est étudié sous le plan de sondage aléatoire simple.

Voyons en premier lieu ce que nous pouvons dire de la variance de l'estimateur de μ dans le cas du sondage stratifié proportionnel. Repartons de l'expression générale de la variance de l'estimateur de μ dans le cas du sondage stratifié :

$$\mathrm{V}(\hat{\mu}_{\text{STP}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h) \frac{\sigma^2_{h;\text{corr}}}{n_h} = \sum_{h=1}^{H} \left(\frac{N_h}{N} \times \frac{N_h}{N}\right) (1-f_h) \frac{\sigma^2_{h;\text{corr}}}{n_h}.$$

Puisque, pour le sondage stratifié proportionnel, les rapports N_h/N sont égaux aux rapports n_h/n et les taux de sondage f_h sont tous égaux au taux de sondage global f, nous obtenons :

$$\mathrm{V}(\hat{\mu}_{\text{STP}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \times \frac{n_h}{n}\right) (1-f) \frac{\sigma^2_{h;\text{corr}}}{n_h} = \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{h;\text{corr}}.$$

En posant

$$\sigma^2_{\text{intra;corr}} = \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{h;\text{corr}},$$

nous obtenons :

$$\mathrm{V}(\hat{\mu}_{\text{STP}}) = (1-f) \frac{\sigma^2_{\text{intra;corr}}}{n}.$$

D'autre part, nous avons vu au chapitre 2 que la variance de l'estimateur de μ dans le cas du sondage aléatoire simple de taille n était donnée par :

$$\mathrm{V}(\hat{\mu}_{\text{PESR}}) = (1-f) \frac{\sigma^2_{\text{corr}}}{n}.$$

Nous pouvons à présent facilement comparer la précision de l'estimateur de μ dans le cas du sondage stratifié proportionnel et dans le cas du sondage aléatoire simple en calculant le rapport de ses variances :

$$\frac{\mathrm{V}(\hat{\mu}_{\text{STP}})}{\mathrm{V}(\hat{\mu}_{\text{PESR}})} = \frac{\sigma^2_{\text{intra;corr}}}{\sigma^2_{\text{corr}}}.$$

Notons que :

$$\begin{aligned}
\sigma^2_{\text{intra;corr}} &= \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{h;\text{corr}} \\
&= \frac{1}{N} \sum_{h=1}^{H} (N_h - 1) \sigma^2_{h;\text{corr}} + \frac{1}{N} \sum_{h=1}^{H} \sigma^2_{h;\text{corr}} \\
&= \frac{1}{N} \sum_{h=1}^{H} N_h \sigma^2_h + \frac{1}{N} \sum_{h=1}^{H} \sigma^2_{h;\text{corr}}
\end{aligned}$$

$$= \sigma^2_{\text{intra}} + \frac{1}{N}\sum_{h=1}^{H} \sigma^2_{h;\text{corr}} .$$

Dès lors, si N est grand et si les variances $\sigma^2_{h;\text{corr}}$ $(h = 1, \dots, H)$ ne sont pas trop grandes, le terme $\frac{1}{N}\sum_{h=1}^{H} \sigma^2_{h;\text{corr}}$ est pratiquement nul et

$$\sigma^2_{\text{intra;corr}} \simeq \sigma^2_{\text{intra}} .$$

Par ailleurs, si N est grand, on a aussi

$$\sigma^2_{\text{corr}} \simeq \sigma^2 .$$

Dans ce cas :

$$\frac{\text{V}(\hat{\mu}_{\text{STP}})}{\text{V}(\hat{\mu}_{\text{PESR}})} \simeq \frac{\sigma^2_{\text{intra}}}{\sigma^2} = 1 - \frac{\sigma^2_{\text{inter}}}{\sigma^2} = 1 - \eta^2 .$$

De par la formule de décomposition de la variance, il est clair que, si N est grand, le rapport de la variance de $\hat{\mu}_{\text{STP}}$ et de la variance de $\hat{\mu}_{\text{PESR}}$ est toujours inférieur à 1 :

$$\frac{\text{V}(\hat{\mu}_{\text{STP}})}{\text{V}(\hat{\mu}_{\text{PESR}})} \simeq 1 - \eta^2 \leq 1 \Rightarrow \text{V}(\hat{\mu}_{\text{STP}}) \leq \text{V}(\hat{\mu}_{\text{PESR}}) .$$

Ainsi, l'estimateur de μ, soit la moyenne-échantillon $\bar{y}$, a une variance plus petite et est donc plus précis sous le plan de sondage stratifié proportionnel que sous le plan de sondage aléatoire simple. On a donc toujours intérêt à stratifier la population et à y appliquer un sondage stratifié proportionnel plutôt que de se contenter d'un sondage PESR !

Mais le gain en précision attaché à la stratification proportionnelle est d'autant plus important que la variance *intra-strates* est faible par rapport à la variance globale, c'est-à-dire que les strates sont bien homogènes. Cela signifie dans le même temps que la stratification proportionnelle est d'autant plus efficace que la variance *inter-strates* constitue une part importante de la variance globale (η^2 proche de 1), autrement dit que la variable auxiliaire de stratification est fortement liée à la variable d'intérêt y.

Stratifier la population est donc toujours positif, mais on a tout intérêt à bien réfléchir à la stratification que l'on va réaliser ! La règle fondamentale de construction des strates peut être résumée ainsi : une bonne stratification donne lieu à la constitution de sous-populations d'individus telles que, vis-à-vis de la variable d'intérêt y, *les comportements moyens au sein des différentes sous-populations soient les plus différents possibles*

ou encore, de manière équivalente,

telles que, *au sein de chacune des sous-populations, les comportements des individus soient les plus semblables possible*.

Exemple 1 (suite)

Revoici les résultats auxquels nous étions arrivés pour la décomposition de la variance dans notre exemple :

$$\sigma^2_{\text{intra}} + \sigma^2_{\text{inter}} = 1\,442{,}45 + 4\,593{,}08 = 6\,035{,}53 = \sigma^2 ;$$

$$\frac{\sigma^2_{\text{intra}}}{\sigma^2} = \frac{1\,442{,}45}{6\,035{,}53} = 0{,}24 = 1 - \eta^2 ;$$

$$\frac{\sigma^2_{\text{inter}}}{\sigma^2} = \frac{4\,593{,}08}{6\,035{,}53} = 0{,}76 = \eta^2 \,.$$

La variance INTRA-strates est relativement faible — elle ne constitue que 24% de la variance globale de y dans notre population d'entreprises — alors que la variance INTER-strates représente 76% de la variance globale.

On peut donc s'attendre à une bien meilleure précision de l'estimateur de μ sous le plan de sondage stratifié proportionnel que sous le plan de sondage aléatoire simple.

En effet, supposons que nous choisissions n égal à 300, soit un taux de sondage global f égal à 300 sur 1 200, c'est-à-dire égal à 25%.

Si l'on met en œuvre un sondage stratifié proportionnel, nous allons prélever un quart des entreprises dans chacune des 3 strates. Nous allons ainsi tirer $n_1 = 100$ entreprises dans la première strate, $n_2 = 125$ entreprises dans la deuxième strate et $n_3 = 75$ entreprises dans la troisième strate.

Dans le cas de ce sondage stratifié proportionnel, la variance de $\bar{y}$ est à peu de choses près égale à :

$$\frac{1-f}{n}\sigma^2_{\text{intra}} = \frac{1-0{,}25}{300}(1\,442{,}45) = 3{,}61 \,.$$

Nous avons déjà eu l'occasion de calculer la variance de $\bar{y}$ sous le plan de sondage PESR de taille n égale à 300 : elle s'élève à 15,10.

Ainsi, pour une même taille d'échantillon, la stratification proportionnelle a permis de réduire de 76% la variance de l'estimateur de μ ! C'est un gain en précision plus qu'appréciable !

3.3.4 Exercice 3.2

Objectif – Cet exercice doit vous permettre de vérifier que vous avez bien compris le principe du sondage stratifié proportionnel et que vous êtes capable d'identifier dans quelle situation ce dernier offre une meilleure précision que le sondage aléatoire simple (PESR).

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

La population étudiée ici est celle des 8 000 médecins (généralistes) dont le cabinet médical se situe dans une certaine région géographique relativement vaste. On s'intéresse à la distribution — notamment à la moyenne μ — de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail ». Pour

estimer cette moyenne, on projette de prélever, par sondage aléatoire simple ou par sondage stratifié, un échantillon de 600 médecins.

On hésite entre deux stratifications possibles de la population de médecins étudiée.

La première stratification se fonde sur un découpage de la région ciblée en 4 zones géographiques distinctes. Le tableau suivant indique le nombre de médecins dans chacune de ces zones :

	Zone 1	Zone 2	Zone 3	Zone 4	Population
Taille	3 800	2 000	1 200	1 000	8 000

La seconde stratification partitionne la population de médecins en 3 groupes selon leur niveau d'expérience professionnelle : les « débutants » (strate 1), les « confirmés » (strate 2) et les « très expérimentés » (strate 3). Les tailles de ces 3 strates sont présentées dans le tableau ci-dessous :

	Débutants	Confirmés	Très expérimentés	Population
Taille	1 600	4 000	2 400	8 000

Par ailleurs, un sondage réalisé il y a 5 ans auprès des médecins de la même région a conduit aux estimations suivantes, pour chaque strate et pour la population entière, de la moyenne et de l'écart-type corrigé de la variable d'intérêt « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail ».

	Zone 1	Zone 2	Zone 3	Zone 4	Population
Moyenne	16,6	16,1	16,3	16,4	16,4
Ecart-type corrigé	6,1	5,8	5,9	6,0	6,0

	Débutants	Confirmés	Très expérimentés	Population
Moyenne	10	15	23	16,4
Ecart-type corrigé	3,4	2,0	5,6	6,0

Le sondage à mener aujourd'hui a pour but principal d'actualiser les estimations réalisées il y a 5 ans, mais il est raisonnable de penser que les écarts-types ont gardé le même ordre de grandeur. Nous pouvons dès lors utiliser ces écarts-types corrigés obtenus il y a 5 ans pour évaluer les niveaux de précision que l'on devrait pouvoir atteindre avec les différents plans de sondage envisagés aujourd'hui.

Partie 1

Si l'on fait le choix de prélever les 600 médecins par tirage aléatoire simple (PESR), que vaudrait la variance de l'estimateur de la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » ?

Partie 2

Au vu des informations fournies par le sondage d'il y a 5 ans, pourriez-vous — sans effectuer le moindre calcul supplémentaire — établir un classement, selon le niveau de précision qui leur est associé, entre le sondage PESR, le sondage stratifié proportionnel avec stratification selon les 4 zones géographiques (STP1) et le sondage stratifié proportionnel avec stratification selon les 3 niveaux d'expérience (STP2), ces trois sondages étant tous de même taille 600 ?

Désignons par μ la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » et par $\hat{\mu}_{\text{PESR}}$, $\hat{\mu}_{\text{STP1}}$ et $\hat{\mu}_{\text{STP2}}$ les estimateurs de μ pour le sondage PESR, STP1 et STP2, respectivement. Cochez la proposition correcte parmi les propositions suivantes (« $\simeq$ » signifie « est à peu près égal à »).

- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP1}}) \leq \mathrm{V}(\hat{\mu}_{\text{STP2}})$
- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \leq \mathrm{V}(\hat{\mu}_{\text{STP1}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP2}})$
- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP1}}) \geq \mathrm{V}(\hat{\mu}_{\text{STP2}})$
- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \geq \mathrm{V}(\hat{\mu}_{\text{STP1}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP2}})$
- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP2}}) \geq \mathrm{V}(\hat{\mu}_{\text{STP1}})$
- ○ $\mathrm{V}(\hat{\mu}_{\text{PESR}}) \simeq \mathrm{V}(\hat{\mu}_{\text{STP2}}) \leq \mathrm{V}(\hat{\mu}_{\text{STP1}})$
- ○ Aucune des propositions précédentes n'est correcte.

Partie 3

Si l'on fait le choix de prélever les 600 médecins par sondage stratifié proportionnel, en considérant la première stratification selon les 4 zones géographiques, ...

1. combien de médecins devrait-on prélever dans chaque zone ?

$$n_1 = ? \qquad n_2 = ? \qquad n_3 = ? \qquad n_4 = ?$$

2. que vaudrait la variance de l'estimateur de la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » ?

Partie 4

Si l'on fait le choix de prélever les 600 médecins par sondage stratifié proportionnel, en considérant la seconde stratification selon les 3 niveaux d'expérience, ...

1. combien de médecins devrait-on prélever dans chaque strate ?

$$n_1 = ? \qquad n_2 = ? \qquad n_3 = ?$$

2. que vaudrait la variance de l'estimateur de la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » ?

3.4 Le sondage stratifié optimal (STO)

3.4.1 Le principe de base

Nous avons vu dans la section précédente que le sondage stratifié proportionnel, qui consiste à appliquer le même taux de sondage dans toutes les strates de la population, permettait toujours d'obtenir une meilleure précision pour l'estimation que le sondage aléatoire simple de même taille n.

Mais peut-on faire mieux que le sondage stratifié proportionnel ? Autrement dit, y a-t-il moyen d'atteindre une précision encore meilleure en choisissant les tailles des sous-échantillons autrement que selon l'allocation proportionnelle ?

La réponse à cette question est… oui, si l'on sait *a priori* que certaines strates sont beaucoup plus hétérogènes que d'autres.

Comment faire ? Nous savons que le sondage aléatoire simple est plus efficace lorsqu'il est appliqué dans un ensemble d'unités homogène. Pour l'échantillonnage stratifié, nous pouvons dès lors diminuer le nombre de prélèvements dans les strates plutôt homogènes pour, au contraire, augmenter le nombre de tirages dans les strates plus hétérogènes.

En d'autres termes, si les strates de la population présentent des niveaux d'hétérogénéité bien différents les uns des autres, on a intérêt à ne pas se contenter du sondage stratifié proportionnel ; dans une telle situation, on a plutôt intérêt à sur-échantillonner quelque peu les strates les plus hétérogènes et à sous-échantillonner en contrepartie les strates les plus homogènes.

Pour vous en convaincre, reprenons rapidement notre exemple relatif aux tailles des entreprises.

Exemple 1 (suite)

En nous fixant $n = 300$, nous avons décidé d'appliquer un taux de sondage global $f = 300/1\,200 = 25\%$.

L'allocation proportionnelle nous indique d'appliquer ce même taux de sondage de 25% dans les trois strates, c'est-à-dire de prendre $n_1 = 100$, $n_2 = 125$ et $n_3 = 75$. Avec cette allocation proportionnelle, la variance de l'estimateur stratifié de μ s'élève, nous l'avons vu, à 3,61.

Mais rappelez-vous la première allocation que nous avions considérée. Nous avions choisi de tirer 100 entreprises dans chacune des 3 strates de la population et nous avions alors obtenu une variance pour $\hat{\mu}_{\mathrm{ST}}$ égale à 2,50. Cette variance est plus faible que la variance de 3,61 associée à l'allocation proportionnelle. Pourquoi ?

Parce que, comparativement à l'allocation proportionnelle, la première allocation que nous avons considérée nous fait prélever davantage d'entreprises dans la troisième

strate – plus petite mais beaucoup plus hétérogène que les deux autres strates – et, en contrepartie, nous indique de sélectionner moins d'entreprises dans la deuxième strate, la plus grande mais d'hétérogénéité non excessive.

Dans la suite de cette section, nous allons étudier l'allocation dite « optimale » nous indiquant comment répartir les n prélèvements PESR entre les différentes strates de telle sorte à minimiser la variance – donc maximiser la précision – de l'estimateur stratifié d'une moyenne ou d'un total-population.

3.4.2 L'allocation optimale (de Neyman)

Supposons à nouveau que l'on se soit fixé comme objectif de sélectionner au total n individus dans la population. Si l'on veut minimiser la variance de $\hat{\mu}_{\text{ST}}$, autrement dit rendre l'estimateur stratifié de μ le plus précis possible, un calcul d'optimisation nous indique qu'il nous faut répartir les n prélèvements entre les strates de la population de la manière suivante[1] : dans la strate U_h, il faut réaliser un nombre n_h de tirages PESR avec

$$n_h = \left(\frac{\frac{N_h}{N}\sigma_{h;\text{corr}}}{\bar{\sigma}_{\text{corr}}} \right) n \,,$$

où

$$\bar{\sigma}_{\text{corr}} = \sum_{h=1}^{H} \frac{N_h}{N}\sigma_{h;\text{corr}} \,;$$

$\bar{\sigma}_{\text{corr}}$ est une moyenne pondérée des écarts-types corrigés dans les différentes strates de la population. Cette répartition particulière des n prélèvements entre les différentes strates porte le nom d'**allocation optimale (de Neyman)**.

Il va de soi que si les nombres n_h donnés par l'allocation optimale de Neyman ne sont pas des nombres entiers, on les arrondit aux entiers les plus proches (puisque nous ne pouvons pas prélever des « portions » d'individus !).

L'allocation optimale de Neyman nous indique clairement que le nombre d'individus à sélectionner dans la strate n° h est fonction, non seulement de l'importance relative de cette strate dans la population – au travers du rapport N_h/N –, mais également du caractère homogène ou hétérogène de cette strate, au travers de l'écart-type corrigé $\sigma_{h;\text{corr}}$.

Comparativement à l'allocation proportionnelle qui ne prenait en compte que l'importance relative de la strate dans la population, l'allocation de Neyman nous conduit à augmenter le nombre d'individus à prélever dans les strates hétérogènes (c'est-à-dire à écart-type élevé), et à réduire en contrepartie le nombre d'individus à prélever dans les strates homogènes (c'est-à-dire à écart-type faible).

---

[1] Les détails du calcul d'optimisation sont présentés dans l'annexe technique 3.1. La lecture de cette annexe technique est tout à fait facultative. Par ailleurs, seuls les étudiants d'ECON et d'INGE ont la formation mathématique nécessaire pour pouvoir comprendre cette annexe ; ils y trouveront une belle illustration de l'utilisation d'une fonction lagrangienne pour la résolution du problème d'optimisation sous contrainte.

Remarquez que, pour déterminer les tailles n_h selon l'allocation optimale de Neyman, il nous faut connaître les valeurs des écarts-types corrigés $\sigma_{h;\text{corr}}$ $(h = 1, \ldots, H)$. Or, il est rare que l'on dispose de cette connaissance. En pratique, on se basera sur des estimations de ces écarts-types obtenues au cours d'une enquête précédente portant sur le même sujet, ou obtenues dans des échantillons préliminaires ; dans certains cas, on se basera plutôt sur notre connaissance des écarts-types dans les strates d'une variable auxiliaire étroitement associée à la variable d'intérêt y.

Exemple 1 (suite)

Reprenons une nouvelle fois notre exemple. Fixons-nous à nouveau $n = 300$, soit un taux de sondage global f de 25%.

Que nous avait donné le sondage stratifié proportionnel ? L'allocation proportionnelle nous avait conduit à $n_1 = 100$, $n_2 = 125$ et $n_3 = 75$.

Que nous donne à présent l'allocation optimale de Neyman ? Revoici le tableau spécifiant les tailles, les variances corrigées et les écarts-types corrigés dans les trois strates et dans la population d'entreprises :

h	1	2	3	Population
N_h	400	500	300	$N = 1\ 200$
$\sigma^2_{h;\text{corr}}$	19,58	114,79	5 571,41	$\sigma^2_{\text{corr}} = 6\ 040,56$
$\sigma_{h;\text{corr}}$	4,43	10,71	74,64	$\sigma_{\text{corr}} = 77,72$

Nous avons :

$$\begin{aligned}\bar{\sigma}_{\text{corr}} &= \frac{N_1}{N}\sigma_{1;\text{corr}} + \frac{N_2}{N}\sigma_{2;\text{corr}} + \frac{N_3}{N}\sigma_{3;\text{corr}} \\ &= \frac{400}{1\,200}(4{,}43) + \frac{500}{1\,200}(10{,}71) + \frac{300}{1\,200}(74{,}64) \\ &= 24{,}60\,.\end{aligned}$$

Dès lors, en appliquant la formule de l'allocation optimale, nous obtenons :

$$n_1 = \left(\frac{\frac{400}{1\,200}(4{,}43)}{24{,}60}\right)300 = 17{,}99 \simeq 18\,;$$

$$n_2 = \left(\frac{\frac{500}{1\,200}(10{,}71)}{24{,}60}\right)300 = 54{,}42 \simeq 54\,;$$

$$n_3 = \left(\frac{\frac{300}{1\,200}(74{,}64)}{24{,}60}\right)300 = 227{,}56 \simeq 228\,.$$

Comparons l'allocation proportionnelle et l'allocation optimale de Neyman.

On voit clairement qu'avec l'allocation optimale, les deux premières strates, plus

homogènes, sont fortement sous-représentées dans l'échantillon par rapport au sondage stratifié proportionnel : avec l'allocation optimale, on prélève 18 et 54 unités dans la première et la deuxième strate, respectivement, alors qu'on en prélevait 100 et 125 avec l'allocation proportionnelle.

En revanche, la troisième strate, beaucoup plus hétérogène, est très nettement surreprésentée dans l'échantillon avec l'allocation optimale : on doit prélever dans cette strate 228 unités statistiques, alors que l'allocation proportionnelle nous indiquait de n'en prélever que 75.

Avec l'allocation proportionnelle, les taux de sondage étaient de 25% dans chacune des 3 strates. Par contre, l'allocation optimale de Neyman nous indique d'appliquer des taux de sondage d'à peine 5% et 11% dans les deux premières strates, mais de 76% dans la troisième strate beaucoup plus hétérogène.

3.4.3 STO versus STP

Pour une stratification donnée et un nombre total n de prélèvements fixé, l'allocation optimale (de Neyman) est l'allocation donnant lieu à la meilleure précision – autrement dit, la plus petite variance – pour l'estimateur stratifié de μ. Il est donc certain que :

$$\mathrm{V}(\hat{\mu}_{\mathrm{STO}}) \leq \mathrm{V}(\hat{\mu}_{\mathrm{STP}}).$$

Mais que vaut précisément le gain en précision apporté par l'allocation optimale de Neyman par rapport à l'allocation proportionnelle ?

On montre (voir l'annexe 3.2 pour les détails de calcul) que l'allocation optimale de Neyman conduit à la variance

$$\mathrm{V}(\hat{\mu}_{\mathrm{STO}}) = \frac{\bar{\sigma}^2_{\mathrm{corr}}}{n} - \frac{\sigma^2_{\mathrm{intra;corr}}}{N}$$

où

$$\bar{\sigma}^2_{\mathrm{corr}} = (\bar{\sigma}_{\mathrm{corr}})^2 = \left(\sum_{h=1}^{H} \frac{N_h}{N} \sigma_{h;\mathrm{corr}} \right)^2 \qquad \text{et} \qquad \sigma^2_{\mathrm{intra;corr}} = \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{h;\mathrm{corr}} .$$

Par ailleurs, nous avons vu que

$$\mathrm{V}(\hat{\mu}_{\mathrm{STP}}) = (1-f) \frac{\sigma^2_{\mathrm{intra;corr}}}{n}.$$

Dès lors :

$$\begin{aligned}
\mathrm{V}(\hat{\mu}_{\mathrm{STP}}) - \mathrm{V}(\hat{\mu}_{\mathrm{STO}}) &= (1-f) \frac{\sigma^2_{\mathrm{intra;corr}}}{n} - \frac{\bar{\sigma}^2_{\mathrm{corr}}}{n} + \frac{\sigma^2_{\mathrm{intra;corr}}}{N} \\
&= \frac{\sigma^2_{\mathrm{intra;corr}}}{n} - \frac{n}{N} \frac{\sigma^2_{\mathrm{intra;corr}}}{n} - \frac{\bar{\sigma}^2_{\mathrm{corr}}}{n} + \frac{\sigma^2_{\mathrm{intra;corr}}}{N} \\
&= \frac{1}{n} \left(\sigma^2_{\mathrm{intra;corr}} - \bar{\sigma}^2_{\mathrm{corr}} \right) = \cdots = \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} \left(\sigma_{h;\mathrm{corr}} - \bar{\sigma}_{\mathrm{corr}} \right)^2 \\
&= \frac{1}{n} \left(\text{variance des } \sigma_{h;\mathrm{corr}}, \text{ pour } h = 1, \ldots, H \right) \geq 0 .
\end{aligned}$$

Ainsi, le gain en précision apporté par l'allocation optimale de Neyman par rapport à l'allocation proportionnelle n'est substantiel que si les dispersions $\sigma_{h;\mathrm{corr}}$ varient

nettement d'une strate à l'autre ; si les strates ont des écarts-types assez similaires, les deux allocations sont très voisines et donnent dès lors lieu à des variances très similaires... et autant se contenter alors de la simplicité de l'allocation proportionnelle !

En pratique donc, si l'on sait que la variable d'intérêt présente une dispersion fort différente d'une strate à l'autre, on a tout intérêt à ne pas se contenter de l'allocation proportionnelle ; on améliore significativement la précision de la procédure d'estimation en appliquant un taux de sondage plus élevé dans les strates les plus hétérogènes et en réduisant, en contrepartie, le taux de sondage pour les strates les plus homogènes.

Exemple 1 (suite)

Que nous avait donné le sondage stratifié proportionnel ? L'allocation proportionnelle nous avait conduit à $n_1 = 100, n_2 = 125$ et $n_3 = 75$. La variance de l'estimateur de μ valait 3,61.

Que vaut la variance de l'estimateur de μ si l'on applique l'allocation optimale de Neyman ($n_1 = 18, n_2 = 54$ et $n_3 = 228$) ?

Puisque $\bar{\sigma}_{\text{corr}} = 24{,}60$ et

$$\sigma^2_{\text{intra;corr}} = \frac{400}{1\,200}(19{,}58) + \frac{500}{1\,200}(114{,}79) + \frac{300}{1\,200}(5\,571{,}41) = 1\,447{,}21,$$

on a :

$$\text{V}(\hat{\mu}_{\text{STO}}) = \frac{\bar{\sigma}^2_{\text{corr}}}{n} - \frac{\sigma^2_{\text{intra;corr}}}{N} = \frac{(24{,}60)^2}{300} - \frac{1\,447{,}21}{1\,200} = 0{,}81\,.$$

Ainsi, l'allocation optimale permet ici une forte réduction de la variance de l'estimateur de μ par rapport à l'allocation proportionnelle, grâce au fait qu'elle tient compte de la nette différence d'hétérogénéité d'une strate à l'autre.

3.4.4 Exercice 3.3

Objectif – Cet exercice doit vous permettre de vérifier que vous avez bien compris le principe du sondage stratifié optimal et que vous êtes capable d'identifier dans quelle situation ce dernier offre une bien meilleure précision que le sondage stratifié proportionnel.

Consignes – Effectuez les calculs qui vous sont demandés dans votre tableur (Excel, LibreOffice ou OpenOffice). Utiliser une machine à calculer vous confronterait très rapidement à des erreurs d'arrondis qui vous empêcheraient d'obtenir les réponses attendues.

Correctif – Vous trouverez un correctif détaillé de cet exercice sur l'UV.

Replacez-vous dans la situation considérée dans l'exercice 3.2 de la Section 3.3 de ce chapitre. On désire estimer, par sondage stratifié de taille 600, la moyenne μ de la variable « nombre de patients que voit (en moyenne) un médecin par journée de travail » dans une population constituée de 8 000 médecins généralistes.

La base de sondage et un sondage réalisé il y a 5 ans nous fournissent les informations suivantes sur les strates et la population.

- Première stratification :

	Zone 1	Zone 2	Zone 3	Zone 4	Population
Taille	3 800	2 000	1 200	1 000	8 000
Ecart-type corrigé	6,1	5,8	5,9	6,0	6,0

- Seconde stratification :

	Débutants	Confirmés	Très expérimentés	Population
Taille	1 600	4 000	2 400	8 000
Ecart-type corrigé	3,4	2,0	5,6	6,0

Dans l'exercice 3.2, nous avons étudié l'efficacité du sondage stratifié proportionnel par rapport au sondage aléatoire simple, pour chacune des deux stratifications considérées. Nous allons à présent nous intéresser au niveau de précision que l'on peut atteindre si on prélève l'échantillon par un sondage stratifié optimal.

Partie 1

Si l'on fait le choix de prélever les 600 médecins par sondage stratifié optimal, en considérant la première stratification selon les 4 zones géographiques, ...

1. combien de médecins devrait-on prélever dans chaque zone ?

$$n_1 = ? \qquad n_2 = ? \qquad n_3 = ? \qquad n_4 = ?$$

2. que vaudrait la variance de l'estimateur de la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » ?

Partie 2

Si l'on fait le choix de prélever les 600 médecins par sondage stratifié optimal, en considérant la seconde stratification selon les 3 niveaux d'expérience, ...

1. combien de médecins devrait-on prélever dans chaque zone ?

$$n_1 = ? \qquad n_2 = ? \qquad n_3 = ?$$

2. que vaudrait la variance de l'estimateur de la moyenne de la variable « nombre de patients que voit (en moyenne) un médecin pendant une journée de travail » ?

3.5 Le sondage stratifié optimal en termes de coûts (STOC)

[!! Section réservée aux étudiants d'ECON et INGE !!]

3.5.1 Le principe de base

Rappelons-nous l'objectif poursuivi dans le cadre du *sondage stratifié optimal*. On commence par se fixer le nombre total n de prélèvements à réaliser dans la population, c'est-à-dire la taille de l'échantillon global S. L'allocation optimale de Neyman nous indique alors quelles sont les tailles $n_1, n_2, \dots, n_H$ des sous-échantillons PESR à prélever dans les différentes strates de la population pour que la variance de l'estimateur stratifié de μ soit minimale, sous la contrainte que $\sum_{h=1}^{H} n_h = n$. Autrement dit, l'allocation optimale de Neyman nous indique comment répartir les n prélèvements entre les H strates de la population de manière à assurer la meilleure précision possible à l'estimateur stratifié de μ.

Il existe toutefois des situations où il est difficile de se fixer *a priori* la taille n de l'échantillon final S, tout simplement parce que ce choix de n est essentiellement assujetti à des contraintes de nature budgétaire.

Considérons par exemple la situation suivante :

- On a pu évaluer que chaque « observation » (incluant le déplacement de l'enquêteur jusqu'au domicile d'un individu, l'interview de ce dernier, …) réalisée dans la strate U_h avait un coût égal à C_h — on dit que C_h est le *coût unitaire* d'une observation dans la strate n° h — et que ce coût unitaire diffère d'une strate à l'autre[1].
- On dispose par ailleurs d'un budget global égal à C_0 pour réaliser les observations sur le terrain et il est hors de question de dépasser ce budget.

La question naturelle que l'on peut se poser dans ce contexte est alors la suivante : combien de prélèvements $n_1, n_2, \dots, n_H$ doit-on effectuer dans les strates $U_1, U_2, \dots, U_H$ de la population si l'on veut minimiser la variance — donc maximiser la précision — de l'estimateur stratifié de μ, tout en respectant la contrainte que le coût total du recueil des données, c'est-à-dire $\sum_{h=1}^{H} n_h C_h$, soit inférieur ou égal au budget C_0 à notre disposition ?

La réponse à cette question est donnée par l'*allocation optimale en termes de coûts*[2].

---

[1] Les coûts peuvent en effet s'avérer très différents d'une strate à l'autre. Par exemple, une interview peut être plus coûteuse en milieu rural qu'en milieu urbain à cause des plus grandes distances à parcourir par l'enquêteur.

[2] L'allocation optimale en termes de coûts est, tout comme l'allocation optimale de Neyman, solution d'un problème de minimisation sous contrainte, à résoudre à l'aide d'une fonction lagrangienne.

3.5.2 L'allocation optimale en termes de coûts

Le premier résultat à garder à l'esprit est le suivant : on ne pourra maximiser la précision de l'estimateur stratifié de μ que si l'on accepte de dépenser la totalité du budget C_0 qui nous a été alloué ! Il est donc impossible d'atteindre la précision maximale tout en faisant des économies sur le budget consacré au recueil des données…

Sachant que le recueil des données va nous coûter l'ensemble de notre budget C_0 — autrement dit que $\sum_{h=1}^{H} n_h C_h = C_0$ — on atteindra la précision maximale pour l'estimateur stratifié de μ en prélevant dans la strate U_h $(h = 1, \dots, H)$ un nombre n_h d'unités statistiques égal à

$$\frac{\frac{C_0}{\sqrt{C_h}} \frac{N_h}{N} \sigma_{h;\text{corr}}}{\sum_{\ell=1}^{H} \frac{N_\ell}{N} \sigma_{\ell;\text{corr}} \sqrt{C_\ell}} . \qquad (1)$$

En pratique, n_h est choisi comme étant le nombre *entier* le plus proche du rapport (1) défini ci-dessus, avec $\sigma_{h;\text{corr}}$ $(h = 1, \dots, H)$ éventuellement remplacé par une estimation obtenue au cours d'une étude préliminaire ou d'une étude antérieure du même type.

Que nous indique cette allocation optimale en termes de coûts ?

Au vu du numérateur du rapport (1), il est clair que cette allocation particulière établit un savant compromis entre trois éléments. En effet, le nombre de prélèvements à effectuer dans une strate est fonction :

(i) de l'*importance relative* de la strate dans la population : on aura tendance à prélever davantage d'unités statistiques dans les plus grandes strates, et moins d'unités dans les strates de plus petites tailles ;

(ii) du *niveau d'hétérogénéité* de la strate : on aura tendance à prélever davantage d'unités statistiques dans les strates plus hétérogènes et, au contraire, à sous-représenter quelque peu les strates plus homogènes ;

(iii) du *coût unitaire* d'une observation dans la strate : afin de respecter le budget dont on dispose, on aura tendance à restreindre le nombre de prélèvements dans les strates « coûteuses » (celles où le coût unitaire d'une observation est élevé) pour favoriser au contraire les prélèvements dans les strates « moins chères » (celles où le coût unitaire d'une observation est plus faible).

3.5.3 Si le coût unitaire d'une observation est identique dans toutes les strates

Plaçons-nous dans le cas particulier où une observation « coûte » le même montant dans toutes les strates de la population. Dans ce cas, se fixer *a priori* le budget global C_0 que l'on ne peut pas dépasser pour le recueil des données revient à se fixer le nombre total n de prélèvements que l'on peut se permettre d'effectuer dans la population, et l'allocation optimale en termes de coûts coïncide avec l'allocation optimale de Neyman.

En effet :

- Si $C_h = C$ pour tout $h = 1, \dots, H$, la contrainte budgétaire

$$\sum_{h=1}^{H} n_h C_h = C_0$$

revient à

$$C \sum_{h=1}^{H} n_h = Cn = C_0,$$

c'est-à-dire encore à

$$n = C_0/C.$$

Se fixer C_0 revient à se fixer n. Ainsi, par exemple, si l'on sait qu'on dispose d'un budget global de $C_0 = 1000$ euros pour le recueil des données et que, dans toutes les strates, chaque observation coûte 5 euros, il est clair que nous pouvons sélectionner au total $1000/5 = 200$ unités statistiques dans la population.

- Si $C_h = C$ pour tout $h = 1, \dots, H$, le nombre n_h de prélèvements à effectuer dans la strate n° h selon l'allocation optimale en termes de coûts est donné par :

$$\frac{\frac{C_0}{\sqrt{C_h}} \frac{N_h}{N} \sigma_{h;\text{corr}}}{\sum_{\ell=1}^{H} \frac{N_\ell}{N} \sigma_{\ell;\text{corr}} \sqrt{C_\ell}} = \frac{\frac{Cn}{\sqrt{C}} \frac{N_h}{N} \sigma_{h;\text{corr}}}{\sum_{\ell=1}^{H} \frac{N_\ell}{N} \sigma_{\ell;\text{corr}} \sqrt{C}} = \left(\frac{\sqrt{C} \frac{N_h}{N} \sigma_{h;\text{corr}}}{\sqrt{C} \sum_{\ell=1}^{H} \frac{N_\ell}{N} \sigma_{\ell;\text{corr}}} \right) n$$

$$= \left(\frac{\frac{N_h}{N} \sigma_{h;\text{corr}}}{\sum_{\ell=1}^{H} \frac{N_\ell}{N} \sigma_{\ell;\text{corr}}} \right) n = \left(\frac{\frac{N_h}{N} \sigma_{h;\text{corr}}}{\bar{\sigma}_{\text{corr}}} \right) n ;$$

il est exactement identique au nombre n_h de prélèvements à effectuer dans la strate n° h selon l'allocation optimale de Neyman (avec $n = C_0/C$).

3.5.4 Exercices

a) Exercice 3.4

Supposons que nous voulions sonder une population constituée de $N = 800$ unités statistiques et partitionnée en 4 strates. Le tableau suivant vous présente différentes caractéristiques de ces strates U_h $(h = 1, \dots, 4)$:

- N_h désigne la taille de la strate U_h ;
- $\sigma_{h;\text{corr}}$ désigne la valeur (évaluée au cours d'un sondage antérieur) de l'écart-type corrigé (la racine carrée de la variance corrigée) de la variable d'intérêt y dans la strate U_h ;
- C_h désigne le coût unitaire (évalué au cours d'un sondage antérieur et exprimé dans une certaine unité monétaire) d'une observation dans la strate U_h.

Strates	U_1	U_2	U_3	U_4
N_h	380	200	120	100
$\sigma_{h;\text{corr}}$	9,6	6,2	3,1	1,8
C_h	6,25	4,00	2,25	1,44

Nous allons considérer dans cet exercice différents sondages stratifiés et rechercher pour chacun d'eux : (i) les répartitions (allocations) des prélèvements qu'ils préconisent

dans les différentes strates ; (ii) la précision qu'ils permettent d'atteindre pour l'estimation de la moyenne μ de la variable d'intérêt y dans la population.

Attention ! Prenez le temps d'analyser de manière critique les résultats de vos calculs et de comparer les résultats obtenus pour les différents sondages stratifiés envisagés : vos résultats sont-ils conformes à ce à quoi vous pouviez vous attendre ? Vous semblent-ils logiques ?

PARTIE 1 : le sondage stratifié proportionnel ($n = 40$)

On décide de tirer un échantillon de taille $n = 40$ par sondage stratifié proportionnel.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des quatre strates de la population :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :
   - nombre de prélèvements dans la 4e strate (n_4) :

b) Que vaut la variance de l'estimateur stratifié de la moyenne μ de la variable d'intérêt y dans la population ? (Donnez votre réponse avec une précision de 2 décimales.)

PARTIE 2 : le sondage stratifié optimal ($n = 40$)

On décide de tirer un échantillon de taille $n = 40$ par sondage stratifié optimal.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des quatre strates de la population, selon l'allocation optimale de Neyman :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :
   - nombre de prélèvements dans la 4e strate (n_4) :

b) Que vaut la variance de l'estimateur stratifié de la moyenne μ de la variable d'intérêt y dans la population ? (Donnez votre réponse avec une précision de 2 décimales.)

PARTIE 3 : le sondage stratifié optimal en termes de coûts ($C_0 = 350$)

On dispose d'un budget global $C_0 = 350$ pour la collecte des données.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des quatre strates de la population, selon l'allocation optimale en termes de coûts :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :
   - nombre de prélèvements dans la 4e strate (n_4) :

b) Que vaut la variance de l'estimateur stratifié de la moyenne μ de la variable d'intérêt y dans la population ? (Donnez votre réponse avec une précision de 2 décimales.)

PARTIE 4 : le sondage stratifié optimal en termes de coûts ($C_0 = 200$)

On dispose d'un budget global $C_0 = 200$ pour la collecte des données.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des quatre strates de la population, selon l'allocation optimale en termes de coûts :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :
   - nombre de prélèvements dans la 4e strate (n_4) :

b) Que vaut la variance de l'estimateur stratifié de la moyenne μ de la variable d'intérêt y dans la population ? (Donnez votre réponse avec une précision de 2 décimales.)

b) Exercice 3.5

Rappels utiles

Rappelons-nous les résultats suivants rencontrés dans le chapitre 2 de ce cours :

- la proportion π d'individus de la population qui possèdent une certaine caractéristique coïncide avec la moyenne μ dans la population de la variable y indicatrice de la présence de cette caractéristique chez un individu ($y_i = 1$ si l'individu i possède la caractéristique ; $y_i = 0$ sinon) ;
- la variance σ^2 de cette variable indicatrice y dans la population est égale au produit $\pi(1-\pi)$; la variance corrigée σ^2_{corr} de y dans la population est donc égale à $\frac{N}{N-1}\pi(1-\pi)$.

Ces résultats vous seront utiles pour trouver les réponses aux questions posées ci-dessous.

Considérons une population de $N = 10\,000$ individus, partitionnée en 3 catégories socio-professionnelles, U_1, U_2 et U_3, de tailles respectives $N_1 = 4\,500$, $N_2 = 3\,000$ et $N_3 = 2\,500$.
On souhaite estimer, via un sondage stratifié, la proportion π_{TP} d'individus de la population ayant déjà travaillé à temps partiel durant leur carrière professionnelle. Des experts ont par ailleurs établi au cours d'une étude antérieure que, dans chacune des strates, la proportion de personnes ayant déjà travaillé à temps partiel est du même ordre de grandeur que la proportion de femmes. Or, la base de sondage nous permet de déterminer de manière exacte la proportion $\pi_{\text{F};h}$ de femmes dans chaque strate U_h de la population : elle est de 75% dans U_1, de 52% dans U_2 et d'à peine 15% dans U_3.
Enfin, l'institut de sondage chargé de réaliser le sondage a évalué que le coût unitaire d'une observation s'élevait à 8 € dans U_1, à 5 € dans U_2 et à 10 € dans U_3.

Nous allons considérer dans cet exercice différents sondages stratifiés et rechercher pour chacun d'eux : (i) les répartitions (allocations) des prélèvements qu'ils préconisent

dans les différentes strates ; (ii) la précision qu'ils permettent d'atteindre pour l'estimation de la proportion π_{TP} dans la population.

Attention ! Prenez le temps d'analyser de manière raisonnée les résultats de vos calculs et de comparer les résultats obtenus pour les différents sondages stratifiés envisagés : vos résultats sont-ils conformes à ce à quoi vous pouviez vous attendre ? Vous semblent-ils logiques ?

PARTIE 1 : le sondage stratifié proportionnel

On décide de tirer un échantillon de taille $n = 500$ par sondage stratifié proportionnel.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des trois strates de la population :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :

b) Que vaut, en bonne approximation, l'écart-type de l'estimateur stratifié de la proportion π_{TP} dans la population ? (Donnez votre réponse avec une précision de 4 décimales.)

PARTIE 2 : le sondage stratifié optimal

On décide de tirer un échantillon de taille $n = 500$ par sondage stratifié optimal, en faisant comme si la procédure qui serait optimale pour l'estimation de la proportion de femmes π_F était également optimale pour l'estimation de la proportion π_{TP} de personnes ayant déjà travaillé à temps partiel.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des trois strates de la population, selon l'allocation optimale de Neyman :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :

b) Que vaut, en bonne approximation, l'écart-type de l'estimateur stratifié de la proportion π_{TP} dans la population ? (Donnez votre réponse avec une précision de 4 décimales.)

PARTIE 3 : le sondage stratifié optimal en termes de coûts

On dispose d'un budget global $C_0 = 2000$ € pour la collecte des données.

a) Indiquez le nombre de prélèvements à effectuer dans chacune des trois strates de la population, selon l'allocation optimale en termes de coûts, en faisant comme si la procédure qui serait optimale pour l'estimation de la proportion de femmes π_F était également optimale pour l'estimation de la proportion π_{TP} de personnes ayant déjà travaillé à temps partiel. :
   - nombre de prélèvements dans la 1re strate (n_1) :
   - nombre de prélèvements dans la 2e strate (n_2) :
   - nombre de prélèvements dans la 3e strate (n_3) :

b) Que vaut, en bonne approximation, l'écart-type de l'estimateur stratifié de la proportion π_{TP} dans la population ? (Donnez votre réponse avec une précision de 4 décimales.)

Chapitre 4
Le sondage à probabilités inégales et sans remise (PISR)

4.1 Introduction

4.1.1 Préambule

Ce chapitre va nous faire découvrir un type de sondage aléatoire très général, englobant en particulier le sondage aléatoire simple et le sondage aléatoire stratifié que nous avons étudiés dans les chapitres 2 et 3.

Cela nous permettra d'introduire une démarche spécifique de construction d'un estimateur *sans biais* d'un total ou d'une moyenne-population, applicable pour toute méthode de sondage aléatoire. C'est d'ailleurs cette démarche que nous suivrons pour déterminer les estimateurs à utiliser dans le cadre des sondages en grappes ou à plusieurs degrés que nous étudierons dans la suite du cours.

Par ailleurs, nous verrons comment il est possible de définir ce plan de sondage « à probabilités inégales » de manière à le rendre particulièrement efficace, c'est-à-dire de manière à ce qu'il nous permette d'estimer τ ou μ avec une excellente précision.

4.1.2 Pourquoi un sondage à probabilités inégales ?

Sélectionner l'échantillon par tirage à probabilités *égales* sans remise est certainement la manière la plus naturelle de procéder lorsqu'on ne connaît rien *a priori* sur la distribution des variables étudiées dans la population. Mais on est rarement aussi démuni dans la pratique. Souvent, des données ou des hypothèses sur la population préexistent à l'enquête, et il est normal d'en tenir compte pour organiser le sondage. On peut alors être amené à accorder une probabilité plus grande d'être sélectionné à certaines unités de la population qu'à d'autres.

C'est d'ailleurs ce que nous sommes contraints de faire si l'on décide de mettre en œuvre le sondage stratifié optimal. Rappelez-vous : si les informations dont nous disposons sur la distribution de la variable d'intérêt dans la population nous indiquent que certaines strates sont nettement plus hétérogènes que d'autres, nous avons tout intérêt à appliquer des taux de sondage plus élevés dans ces strates fort hétérogènes et plus faibles dans les strates plus homogènes. Cela revient à attribuer des probabilités d'inclusion plus élevées aux individus des strates hétérogènes qu'à ceux appartenant à des strates plus homogènes. Notre plan de sondage sera dès lors à probabilités inégales.

Mais il y a d'autres contextes que celui du sondage stratifié où l'on peut juger opportun d'attribuer des probabilités d'inclusion qui diffèrent d'un individu à l'autre de la population. Ainsi, par exemple, si l'on étudie le volume d'une production, on veillera à ce que les entreprises aient une probabilité d'être sélectionnées proportionnelle à leur taille (nombre de salariés) ou à leur chiffre d'affaires, ou l'on fera en sorte que les exploitations agricoles aient une probabilité d'inclusion proportionnelle à leur surface. Ceci nous assurera que les entreprises ou les exploitations agricoles les plus importantes pour ce qui est du volume de la production soient bien représentées dans l'échantillon.

Intuitivement, on imagine bien que la méthode n'a d'intérêt que si les probabilités d'inclusion sont corrélées avec la variable d'intérêt. Il apparaît également naturel de devoir tenir compte des différences entre les probabilités d'inclusion dans la construction de l'estimateur : on peut raisonnablement penser que, dans les calculs d'estimation, les données recueillies sur les individus de l'échantillon devront être pondérées par l'inverse des probabilités d'inclusion de ces individus, si on veut en quelque sorte reconstituer une image fidèle (une « maquette ») de la population. Ainsi, si on donne plus d'importance à l'individu i lors du tirage de l'échantillon, on fera en sorte de lui en donner moins dans le calcul de l'estimation de telle façon, en un sens, à « rétablir l'équilibre ». Nous allons préciser ces idées dans la suite de ce chapitre.

4.2 Le tirage PISR

On parle de méthode de **sondage aléatoire *à probabilités inégales* sans remise** — nous dirons PISR — dès que les probabilités d'inclusion p_i affectées par la méthode aux individus i de la population ne sont pas toutes égales ; la méthode d'échantillonnage est donc telle que certains individus de la population ont une plus grande probabilité de se retrouver dans l'échantillon que d'autres.

En pratique, sélectionner un échantillon selon une méthode de sondage PISR ne peut se faire qu'en appliquant à la base de sondage un certain algorithme informatique conçu de telle sorte que les probabilités d'inclusion associées à la méthode choisie soient bien respectées.

Attention ! Les **probabilités d'inclusion** affectées aux individus de la population ne peuvent pas être quelconques ! Elles doivent satisfaire les deux conditions suivantes :

1. $p_i > 0$ pour tout $i \in U$.
   Tout individu de la population doit avoir une probabilité strictement positive d'être sélectionné ; la procédure d'échantillonnage ne peut exclure *a priori* aucune unité statistique de la population.
2. Si la procédure d'échantillonnage choisie conduit à un échantillon aléatoire S de taille aléatoire n_S : $\sum_{i\in U} p_i = \mathrm{E}(n_S)$;
   si la procédure d'échantillonnage est de taille fixe n : $\sum_{i\in U} p_i = n$.

D'où vient la seconde condition ?

- Désignons par I_i $(i \in U)$ la variable aléatoire (v.a.) indicatrice de l'inclusion de l'individu i dans l'échantillon S qui sera prélevé :
  $$I_i = \begin{cases} 1 & \text{si } i \in S \\ 0 & \text{sinon} \end{cases} ;$$
  $$\mathrm{P}(I_i = 1) = \mathrm{P}(i \in S) = p_i \text{ et } \mathrm{P}(I_i = 0) = \mathrm{P}(i \notin S) = 1 - p_i.$$
  La v.a. I_i suit donc une loi de Bernoulli $\mathrm{Bin}(1, p_i)$ et, par conséquent, $\mathrm{E}(I_i) = p_i$ et $\mathrm{V}(I_i) = p_i(1 - p_i)$.

- La v.a. $\sum_{i\in U} I_i$ comptabilise le nombre d'individus sélectionnés au cours de la procédure d'échantillonnage et correspond donc à la taille n_S de l'échantillon S prélevé :
  $$\sum_{i\in U} I_i = n_S.$$
  Dès lors :
  $$\mathrm{E}(n_S) = \mathrm{E}\left(\sum_{i\in U} I_i\right) = \sum_{i\in U} \mathrm{E}(I_i) = \sum_{i\in U} p_i.$$

Si la procédure d'échantillonnage est de taille fixe n (tous les échantillons possibles s sont de taille n), on a $\mathrm{E}(n_S) = n$ et l'on retrouve la seconde partie de la condition 2.

4.3 L'estimateur de Horvitz-Thompson

4.3.1 Définition : l'estimateur de Horvitz-Thompson d'un total

Intéressons-nous dans un premier temps à l'estimation du total τ d'une variable d'intérêt $\mathcal{Y}$ dans la population :

$$\tau = \sum_{i \in U} y_i \, .$$

Dans le contexte général d'un plan de sondage PISR, il est classique d'estimer τ à l'aide de l'**estimateur *de Horvitz-Thompson*** : cet estimateur se définit comme la somme, sur les individus de l'échantillon prélevé S, des valeurs observées pour la variable $\mathcal{Y}$ sur ces individus, divisées par les probabilités d'inclusion de ces individus. Nous prendrons l'habitude de désigner cet estimateur par $\hat{\tau}_{\mathrm{HT}}$:

$$\hat{\tau}_{\mathrm{HT}} = \sum_{i \in S} \frac{y_i}{p_i} \, .$$

L'estimateur de Horvitz-Thompson est aussi appelé « **estimateur *par valeurs dilatées*** ». La raison en est simple ! p_i est une probabilité et a donc une valeur inférieure (ou égale) à 1 ; dès lors, diviser la valeur y_i de la variable $\mathcal{Y}$ par cette probabilité p_i revient à « gonfler » ou « dilater » la valeur y_i, et cela d'autant plus que la probabilité d'inclusion p_i est faible. L'estimateur de Horvitz-Thompson tente donc d'estimer le total-population en donnant à chaque individu de l'échantillon un poids d'autant plus important que cet individu avait peu de chance d'être sélectionné.

Illustrons le calcul de l'estimateur de Horvitz-Thompson à l'aide d'un exemple. Nous allons considérer ici une taille de population et une taille d'échantillon très petites, de manière à ce que les calculs « à la main » ne soient pas trop fastidieux.

Exemple

Notre population U est constituée de 5 dépenses, numérotées de 1 à 5 dans l'ordre chronologique selon lequel elles sont survenues. Notre variable d'intérêt $\mathcal{Y}$ dans cette population est le montant (en euros) des dépenses.

Il se fait que la première dépense s'élève à 24€, la deuxième à 20€, la troisième à 50€, la quatrième à 32€ et la dernière à 54€ ; le montant total de ces dépenses vaut donc 180€ :

$$y_1 = 24, y_2 = 20, y_3 = 50, y_4 = 32 \text{ et } y_5 = 54 \Longrightarrow \tau = 180.$$

Nous allons nous intéresser à l'estimation de ce total τ via un échantillon de taille 2 prélevé par tirage PISR.

Fixons *a priori* les probabilités d'inclusion suivantes (nous verrons un peu plus loin dans ce chapitre comment l'on peut judicieusement choisir les valeurs de ces probabilités d'inclusion) :

$$p_1 = 0{,}2 \, , p_2 = 0{,}3 \, , p_3 = 0{,}4 \, , p_4 = 0{,}5 \text{ et } p_5 = 0{,}6.$$

Remarquez que nous avons bien choisi des probabilités d'inclusion telles que leur

somme est égale à 2, la taille souhaitée pour l'échantillon.

Supposons qu'au terme de la procédure d'échantillonnage, nous nous retrouvions avec l'échantillon $\{1,3\}$. Dans ce cas, nous observons les valeurs $y_1 = 24$ et $y_3 = 50$. Les valeurs dilatées correspondantes sont

$$\frac{y_1}{p_1} = \frac{24}{0{,}2} = 120 \quad \text{et} \quad \frac{y_3}{p_3} = \frac{50}{0{,}4} = 125.$$

Nous obtenons alors comme estimation du montant total τ des dépenses :

$$\hat{\tau}_{\text{HT}} = 120 + 125 = 245\ €.$$

Si nous tirons l'échantillon $\{3,5\}$, nous observons les valeurs $y_3 = 50$ et $y_5 = 54$. Les valeurs dilatées correspondantes sont

$$\frac{y_3}{p_3} = \frac{50}{0{,}4} = 125 \quad \text{et} \quad \frac{y_5}{p_5} = \frac{54}{0{,}6} = 90,$$

ce qui nous donne :

$$\hat{\tau}_{\text{HT}} = 125 + 90 = 215\ €.$$

4.3.2 Propriété intéressante de l'estimateur de Horvitz-Thompson

L'estimateur de Horvitz-Thompson est très simple à calculer. Mais… est-ce un « bon » estimateur ?

Nous étudierons sa variance dans la sous-section 4.3.5. Regardons d'abord ce qu'il en est de son espérance.

L'estimateur de Horvitz-Thompson du total-population τ jouit d'une propriété bien intéressante : il est *sans biais*, quel que soit le plan de sondage PISR considéré. En d'autres termes, quelles que soient les probabilités d'inclusion affectées aux individus de la population (et respectant les deux conditions spécifiées dans la Section 4.2), $\hat{\tau}_{\text{HT}}$ est un estimateur de τ qui, en moyenne, « vise juste ».

Ce résultat se vérifie aisément. En effet :

$$\hat{\tau}_{\text{HT}} = \sum_{i \in S} \frac{y_i}{p_i} = \sum_{i \in U} \frac{y_i}{p_i} I_i\ ,$$

où I_i est la variable indicatrice de la sélection de l'individu i (cf. Section 4.2). Dès lors :

$$\text{E}(\hat{\tau}_{\text{HT}}) = \text{E}\left(\sum_{i \in U} \frac{y_i}{p_i} I_i\right) = \sum_{i \in U} \frac{y_i}{p_i} \text{E}(I_i).$$

Puisque $\text{E}(I_i) = p_i$ pour tout $i \in U$ (cf. Section 4.2), nous avons :

$$\text{E}(\hat{\tau}_{\text{HT}}) = \sum_{i \in U} \frac{y_i}{p_i} p_i = \sum_{i \in U} y_i = \tau,$$

ce que nous voulions démontrer.

Remarque

La condition selon laquelle $p_i > 0$ pour tout $i \in U$ est indispensable pour que l'on puisse réécrire $\hat{\tau}_{\text{HT}}$ sous la forme $\sum_{i \in U} \frac{y_i}{p_i} I_i$.

Si p_i était nul pour certains individus, c'est-à-dire si on excluait d'emblée volontairement des individus de la procédure d'échantillonnage pour être certain de ne

pas les retrouver dans l'échantillon, on introduirait un phénomène assimilable à celui de défaut de couverture de la base de sondage (cf. Chapitre 9 de ce cours : l'erreur de couverture). Dans ce cas, $\hat{\tau}_{\mathrm{HT}}$ serait un estimateur biaisé de τ, et n'estimerait sans biais que le total des valeurs y_i sur l'ensemble des individus i de la population pour lesquels les probabilités p_i sont strictement positives :

$$E(\hat{\tau}_{\mathrm{HT}}) = \sum_{i \in U \text{ tel que } p_i > 0} y_i$$
$$= \tau - \sum_{i \in U \text{ tel que } p_i = 0} y_i.$$

4.3.3 L'estimateur de Horvitz-Thompson d'une moyenne

$\hat{\tau}_{\mathrm{HT}}$ est un estimateur du total τ de la variable y dans la population. Mais comment fait-on pour estimer la moyenne μ de y ? La réponse est évidente : puisque $\mu = \tau/N$, l'estimateur de Horvitz-Thompson de μ n'est autre que :

$$\hat{\mu}_{\mathrm{HT}} = \frac{\hat{\tau}_{\mathrm{HT}}}{N}.$$

4.3.4 Remarque

Vous avez peut-être l'impression que l'estimateur de Horvitz-Thompson est un estimateur bien étrange… sortant de nulle part ! Et pourtant, cet estimateur nous accompagne depuis le début du chapitre 2 de ce cours. En effet, les estimateurs de τ et de μ que nous avons considérés dans le cadre du sondage aléatoire simple, du tirage de Bernoulli, ou encore du sondage stratifié sont en réalité des estimateurs de Horvitz-Thompson du total et de la moyenne-population. Vérifions cela…

- Dans le cadre du sondage aléatoire simple (ou sondage PESR), tous les individus de la population ont la même probabilité d'inclusion[1], égale au taux de sondage n/N. Dès lors, l'estimateur de Horvitz-Thompson de τ est égal à :

$$\hat{\tau}_{\mathrm{HT}} = \sum_{i \in S} \frac{y_i}{n/N} = \frac{N}{n} \sum_{i \in S} y_i = N\bar{y} = \hat{\tau}_{\mathrm{PESR}}.$$

  L'estimateur de Horvitz-Thompson de μ est dès lors :

$$\hat{\mu}_{\mathrm{HT}} = \frac{\hat{\tau}_{\mathrm{HT}}}{N} = \bar{y} = \hat{\mu}_{\mathrm{PESR}}.$$

- Dans le cadre du tirage de Bernoulli, tous les individus de la population ont la même probabilité d'inclusion, égale à la valeur p choisie au début du tirage. Dès lors :

$$\hat{\tau}_{\mathrm{HT}} = \sum_{i \in S} \frac{y_i}{p} = \frac{1}{p} \sum_{i \in S} y_i$$

  et

$$\hat{\mu}_{\mathrm{HT}} = \frac{\hat{\tau}_{\mathrm{HT}}}{N} = \frac{1}{Np} \sum_{i \in S} y_i = \hat{\mu}_{\mathrm{B}}.$$

[1] Le sondage PESR donne lieu à des probabilités d'inclusion *égales*, mais qui respectent bien les conditions imposées aux probabilités d'inclusion d'un sondage PISR; le sondage PESR peut dès lors être vu comme un cas particulier du sondage PISR.

- Et pour le sondage aléatoire stratifié ? Rappelons-nous que le taux de sondage peut varier d'une strate à l'autre. Ainsi, de manière générale, la probabilité d'inclusion d'un individu i de la strate U_h est égale au taux de sondage f_h appliqué dans cette strate : pour $i \in U_h$,
$$p_i = f_h = \frac{n_h}{N_h}.$$
Il s'ensuit que l'estimateur de Horvitz-Thompson de τ prend la forme :
$$\hat{\tau}_{\mathrm{HT}} = \sum_{i \in S} \frac{y_i}{p_i} = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_i}{n_h/N_h} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_h} y_i = \sum_{h=1}^{H} N_h \bar{y}_h = \hat{\tau}_{\mathrm{ST}}.$$
L'estimateur de Horvitz-Thompson de μ est égal à :
$$\hat{\mu}_{\mathrm{HT}} = \frac{\hat{\tau}_{\mathrm{HT}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \hat{\mu}_{\mathrm{ST}}.$$

Vous voyez donc que l'estimateur de Horvitz-Thompson ne vous est pas si étranger que cela !

4.3.5 La variance de l'estimateur de Horvitz-Thompson

Nous avons vu ci-dessus que l'estimateur de Horvitz-Thompson permettait d'estimer **sans biais** un total ou une moyenne-population, quel que soit le plan de sondage PISR considéré (quelles que soient les probabilités d'inclusion affectées aux individus de la population, pour autant que celles-ci respectent bien les deux conditions énoncées à la Section 4.2).

Intéressons-nous à présent à la précision de cet estimateur, au travers de sa variance.

On trouve dans la littérature consacrée à la théorie des sondages différentes expressions de la variance de $\hat{\tau}_{\mathrm{HT}}$, toutes relativement complexes et lourdes à calculer. Voyons-en quelques-unes ci-dessous. Nous nous placerons systématiquement dans le cas où $p_i > 0$ pour tout $i \in U$.

a) 1re expression de la variance de $\hat{\tau}_{\mathrm{HT}}$

Observons que :
$$\mathrm{V}(\hat{\tau}_{\mathrm{HT}}) = \mathrm{V}\left(\sum_{i \in S} \frac{y_i}{p_i}\right) = \mathrm{V}\left(\sum_{i \in U} \frac{y_i}{p_i} I_i\right)$$
$$= \sum_{i \in U} \frac{y_i^2}{p_i^2} \mathrm{V}(I_i) + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{y_i}{p_i} \frac{y_j}{p_j} \mathrm{Cov}(I_i, I_j).$$
Ceci nous conduit, moyennant quelques calculs, à l'expression suivante :
$$\mathrm{V}(\hat{\tau}_{\mathrm{HT}}) = \sum_{i \in U} \frac{y_i^2}{p_i}(1 - p_i) + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{y_i}{p_i} \frac{y_j}{p_j}(p_{ij} - p_i p_j) := \mathrm{V}_1(\hat{\tau}_{\mathrm{HT}}),$$
où $p_{ij} = \mathrm{P}\big((i \in S) \text{ et } (j \in S)\big)$ est la *probabilité d'inclusion* dite « *d'ordre 2* » associée aux individus i et j $(i \neq j)$.

b) 2e expression de la variance de $\hat{\tau}_{\mathrm{HT}}$ (Sen, Yates et Grundy, 1953)

Sen, Yates et Grundy ont montré en 1953 que, lorsque l'échantillonnage est de taille n *fixe* :

$$\mathrm{V}(\hat{\tau}_{\mathrm{HT}}) = -\frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\left(p_{ij}-p_ip_j\right) \coloneqq \mathrm{V}_2(\hat{\tau}_{\mathrm{HT}})\,.$$

Insistons sur le fait que $\mathrm{V}_1(\hat{\tau}_{\mathrm{HT}})$ et $\mathrm{V}_2(\hat{\tau}_{\mathrm{HT}})$ font toutes deux intervenir les probabilités d'inclusion p_{ij} *d'ordre 2*, généralement relativement difficiles à déterminer !

c) Estimation de la variance de $\hat{\tau}_{\mathrm{HT}}$

Le théorème suivant va nous permettre de proposer un estimateur sans biais de la variance de $\hat{\tau}_{\mathrm{HT}}$.

Théorème

Soit $\varphi(y_i, y_j)$ une fonction à valeurs réelles définie sur $U\times U$: la somme

$$\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\varphi(y_i,y_j)$$

est estimée sans biais par

$$\sum_{i\in S}\sum_{\substack{j\in S\\ j\neq i}}\frac{\varphi(y_i,y_j)}{p_{ij}}$$

si et seulement si $p_{ij}>0$ pour tout $i,j\in U, i\neq j$.

Ainsi, l'expression de l'estimateur sans biais s'obtient tout simplement en remplaçant, dans l'expression à estimer, la double somme sur les individus distincts de la *population* par la double somme sur les individus distincts de l'*échantillon*, et en dilatant les termes de la somme en les divisant par les probabilités d'inclusion d'ordre 2.

Estimateurs sans biais de $V(\hat{\tau}_{HT})$

En appliquant la démarche de Horvitz-Thompson et le théorème ci-dessus, on obtient — pour autant que toutes les probabilités d'inclusion p_{ij} d'ordre 2 soient strictement positives, autrement dit que tous les couples d'individus (i,j) que l'on peut former dans la population soient susceptibles d'être tirés — qu'un estimateur sans biais de

$$\mathrm{V}_1(\hat{\tau}_{\mathrm{HT}}) = \sum_{i\in U}\frac{y_i^2}{p_i}(1-p_i)+\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i}{p_i}\frac{y_j}{p_j}\left(p_{ij}-p_ip_j\right)$$

est

$$\widehat{\mathrm{V}}_1(\hat{\tau}_{\mathrm{HT}}) = \sum_{i\in S}\frac{y_i^2}{p_i^2}(1-p_i)+\sum_{i\in S}\sum_{\substack{j\in S\\ j\neq i}}\frac{y_i}{p_i}\frac{y_j}{p_j}\frac{\left(p_{ij}-p_ip_j\right)}{p_{ij}}\,;$$

un estimateur sans biais de

$$\mathrm{V}_2(\hat{\tau}_{\mathrm{HT}}) = -\frac{1}{2}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\left(p_{ij}-p_ip_j\right)$$

est

$$\widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}}) = -\frac{1}{2}\sum_{i\in S}\sum_{\substack{j\in S\\ j\neq i}}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\frac{(p_{ij}-p_ip_j)}{p_{ij}}.$$

Remarques

1. On peut avoir $\widehat{\mathrm{V}}_1(\hat{\tau}_{\mathrm{HT}}) \neq \widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}})$.

   Notez cependant que l'égalité entre $\widehat{\mathrm{V}}_1(\hat{\tau}_{\mathrm{HT}})$ et $\widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}})$ est assurée dans le cas du sondage PESR ou ST (stratifié).

2. $\widehat{\mathrm{V}}_1(\hat{\tau}_{\mathrm{HT}})$ et $\widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}})$ peuvent être ≤ 0.

   Notez cependant que $\widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}}) \geq 0$ si
   $$p_{ij} - p_ip_j \leq 0 \qquad \text{pour tout } i,j \in U, i \neq j,$$
   c'est-à-dire si
   $$p_ip_j - p_{ij} \geq 0 \qquad \text{pour tout } i,j \in U, i \neq j\ .$$
   Ces conditions portent le nom de **conditions de Yates-Grundy**.

Conclusion

On reviendra ultérieurement dans ce chapitre aux difficultés liées au calcul de la valeur de $\widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{HT}})$, difficultés dues en particulier à la présence des probabilités d'inclusion d'ordre 2 dans son expression. Même si l'on dispose des probabilités p_{ij}, le calcul de $\widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{HT}})$ est plus complexe qu'il n'y paraît à première vue : la difficulté à laquelle on reste confronté est d'ordre numérique, car les sommes à calculer mettent en jeu tous les couples d'individus que l'on peut former dans l'échantillon de taille n. Sachant qu'il y a $n(n-1)/2$ couples différents possibles, on vérifie que, pour un échantillon de taille 500 par exemple, le calcul de l'estimation de la variance de $\hat{\tau}_{\mathrm{HT}}$ nous impose de faire la somme de 124 750 termes !

Aussi complexe cela soit-il, il est important de chercher à déterminer $\widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{HT}})$! En effet, si l'on arrive à estimer la variance de $\hat{\tau}_{\mathrm{HT}}$, on peut dans la foulée calculer l'intervalle de confiance pour le total-population τ, au niveau de confiance de 95%. Cet intervalle est centré en $\hat{\tau}_{\mathrm{HT}}$ et ses bornes inférieure et supérieure sont obtenues en retranchant et en ajoutant à $\hat{\tau}_{\mathrm{HT}}$ 1,96 fois la racine carrée de la variance estimée de $\hat{\tau}_{\mathrm{HT}}$:

$$\left[\hat{\tau}_{\mathrm{HT}} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{HT}})}\right].$$

d) Cas particulier : le tirage de Bernoulli

Revenons au tirage de Bernoulli. Les probabilités d'inclusion d'ordre 1 et 2 associées à cette méthode de sondage aléatoire présentent les caractéristiques suivantes :

- pour tout $i \in U$:
  $$p_i = p\ ;$$
- pour tout $i,j \in U$, avec $i \neq j$: grâce à l'indépendance des prélèvements successifs dans la population,
  $$p_{ij} = \mathrm{P}\big((i \in S) \text{ et } (j \in S)\big) = \mathrm{P}(i \in S)\,\mathrm{P}(j \in S) = p^2\ .$$

Par conséquent, pour tout $i,j \in U$, avec $i \neq j$:

$$p_{ij} - p_i p_j = p^2 - p^2 = 0.$$

On obtient ainsi que[2]

$$\mathrm{V}(\hat{\tau}_{\mathrm{B}}) = \mathrm{V}_1(\hat{\tau}_{\mathrm{B}}) = \sum_{i \in U} \frac{y_i^2}{p_i}(1 - p_i) = \frac{1-p}{p} \sum_{i \in U} y_i^2$$

et

$$\widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{B}}) = \widehat{\mathrm{V}}_1(\hat{\tau}_{\mathrm{B}}) = \sum_{i \in S} \frac{y_i^2}{p_i^2}(1 - p_i) = \frac{1-p}{p^2} \sum_{i \in S} y_i^2 \,.$$

Notez encore que, puisque $\hat{\mu}_{\mathrm{B}} = \hat{\tau}_{\mathrm{B}}/N$:

$$\mathrm{V}(\hat{\mu}_{\mathrm{B}}) = \mathrm{V}(\hat{\tau}_{\mathrm{B}})/N^2 \qquad et \qquad \widehat{\mathrm{V}}(\hat{\mu}_{\mathrm{B}}) = \widehat{\mathrm{V}}(\hat{\tau}_{\mathrm{B}})/N^2 \,.$$

On retrouve bien les expressions indiquées dans le point c) de la section 2.9.3 (cf. Chapitre 2).

[2] Le tirage de Bernoulli donnant lieu à un échantillonnage de taille *aléatoire*, on ne peut faire appel à l'expression de Sen, Yates et Grundy pour la variance de $\hat{\tau}_{\mathrm{B}}$. Il faut donc partir de l'expression $\mathrm{V}_1(\hat{\tau}_{\mathrm{B}})$.

4.4 Le sondage à probabilités proportionnelles aux valeurs d'une variable auxiliaire $\mathcal{X}$ (sondage PPS)

4.4.1 Principe de base

Revenons un moment sur l'expression $V_2(\hat{\tau}_{HT})$ (de Sen, Yates et Grundy) de la variance de $\hat{\tau}_{HT}$. Que nous indique-t-elle ? Dans quelle situation cette variance est-elle petite ? Comment faut-il choisir les probabilités d'inclusion associées au sondage PISR pour que ce dernier soit efficace, autrement dit pour que l'estimateur de Horvitz-Thompson jouisse d'une bonne précision ?

Plaçons-nous dans le cas — très souvent rencontré — où la variable d'intérêt $\mathcal{Y}$ ne prend que des valeurs *strictement positives*, et imaginons que l'on puisse définir les probabilités d'inclusion de telle sorte qu'elles soient approximativement proportionnelles aux valeurs de $\mathcal{Y}$. Autrement dit, supposons que, pour chaque individu i de la population, sa probabilité d'inclusion p_i soit approximativement égale, à une certaine constante de proportionnalité λ près, à la valeur y_i que prend la variable d'intérêt $\mathcal{Y}$ chez cet individu :

$$p_i \simeq \lambda y_i \qquad \text{pour tout } i \in U.$$

Dans ce cas, quels que soient les individus distincts i et j que l'on considère :

$$\frac{y_i}{p_i} - \frac{y_j}{p_j} \simeq \frac{y_i}{\lambda y_i} - \frac{y_j}{\lambda y_j} = \frac{1}{\lambda} - \frac{1}{\lambda} = 0.$$

Dès lors, tous les termes de la double somme définissant la variance de $\hat{\tau}_{HT}$ sont, à peu de chose près, égaux à zéro, ce qui va nous donner une variance pratiquement nulle pour $\hat{\tau}_{HT}$.

Ainsi donc, si l'on pouvait affecter aux individus de la population des probabilités d'inclusion p_i approximativement proportionnelles aux valeurs y_i de la variable d'intérêt, l'estimateur de Horvitz-Thompson estimerait de manière très précise le total-population τ, sans presque plus aucune erreur d'échantillonnage.

Mais comment faire, en pratique, pour s'assurer que la probabilité d'inclusion p_i de l'individu i soit approximativement proportionnelle à y_i, alors que nous ne connaissons pas *a priori* les valeurs que prend la variable $\mathcal{Y}$ dans la population ? La solution à ce problème est relativement simple dès le moment où l'on dispose — dans la base de sondage ou dans un fichier annexe — des valeurs, pour *tous* les individus de la population, d'une variable auxiliaire $\mathcal{X}$ strictement positive et approximativement proportionnelle à la variable d'intérêt $\mathcal{Y}$. Dans ce cas, en effet, il suffit de définir les probabilités d'inclusion p_i comme *exactement* proportionnelles aux valeurs (connues !) x_i de la variable $\mathcal{X}$ pour se retrouver — comme souhaité — avec des probabilités d'inclusion *approximativement* proportionnelles aux valeurs y_i de la variable $\mathcal{Y}$: si, pour tout $i \in U$, il existe une constante a telle que

$$x_i \simeq a y_i$$

et que l'on définit

$$p_i = c x_i$$

où c est une certaine constante, alors

$$p_i \simeq c a y_i = \lambda y_i$$

avec $\lambda = ca$.

Cette manière de déterminer les probabilités d'inclusion conduit au sondage PISR du type PPS (*Probability Proportional to Size*). Cette appellation de « sondage à probabilité proportionnelle à la *taille* » provient du fait que, dans de très nombreuses situations, la variable auxiliaire X utilisée pour définir les probabilités d'inclusion est une variable rendant compte de la « taille » de chaque unité statistique de la population.

4.4.2 Détermination des probabilités d'inclusion p_i ($i \in U$)

Nous venons de voir que si l'on connaissait les valeurs d'une variable auxiliaire X — strictement positive et approximativement proportionnelle à la variable d'intérêt Y — pour tous les individus de la population, il était particulièrement judicieux d'affecter à tout individu $i \in U$ une probabilité d'inclusion

$$p_i = c x_i \,,$$

où c est une certaine constante. Mais quelle valeur doit-on donner à c ?

Pour répondre à cette question, il suffit de se rappeler que, dans le cas d'un échantillonnage de taille fixe n, la somme des probabilités d'inclusion de tous les individus de la population doit être égale à n (cf. Section 4.2)[1] :

$$\sum_{i \in U} p_i = n.$$

Dès lors, si l'on souhaite un sondage PPS de taille fixe n, il faut que

$$\sum_{i \in U} c x_i = n \,,$$

c'est-à-dire que

$$c = \frac{n}{\sum_{i \in U} x_i} = \frac{n}{\tau_x}$$

où $\tau_x = \sum_{i \in U} x_i$ est le total (connu !) de la variable auxiliaire X dans la population. La probabilité d'inclusion p_i que l'on doit attribuer à l'individu i doit donc être égale à

$$p_i = c x_i = \frac{n x_i}{\tau_x}.$$

Remarque

Pour que la quantité p_i définie de la sorte soit bien une *probabilité*, il faut nécessairement que

$$\frac{n x_i}{\tau_x} \leq 1 \,,$$

autrement dit que

[1] Il en est de même si l'échantillonnage est de taille aléatoire, mais donne lieu à des échantillons de taille *moyenne* égale à n.

$$x_i \leq \frac{\tau_x}{n},$$

ce qui n'est pas automatiquement assuré ! S'il s 'avère que, pour certains individus i de la population, la valeur x_i est strictement supérieure au rapport τ_x/n, on procède de la manière suivante :

- Désignons par A l'ensemble de ces individus i dont la définition de p_i pose problème et par n_A le nombre de ces individus :
$$A = \left\{ i \in U \text{ tels que } x_i > \frac{\tau_x}{n} \right\} \text{ et } n_A = \text{card}(A).$$
- A chaque individu de cet ensemble A, on affecte d'autorité une probabilité d'inclusion égale à 1 : ils se retrouveront donc obligatoirement dans l'échantillon qui sera sélectionné.
- Quant aux individus i qui n'appartiennent pas à cet ensemble A, on recalcule leurs probabilités d'inclusion comme suit : puisqu'il ne nous faut plus prélever, par sondage PPS, que $(n - n_A)$ individus dans la population $U \setminus A$ (la population U dont on a extrait le sous-ensemble A), on prend
$$p_i = \frac{(n - n_A)x_i}{\sum_{j \in U \setminus A} x_j}.$$
- On réitère cette procédure si certains des nouveaux p_i calculés de la sorte posent à nouveau problème.

Exemple

Considérons la situation où $N = 6, n = 3$ et $x_1 = 1, x_2 = 9, x_3 = 10, x_4 = 70, x_5 = 90$ et $x_6 = 120$.

Nous avons :

$$\tau_x = 1 + 9 + 10 + 70 + 90 + 120 = 300\,,$$

ce qui nous conduit à :

$$\frac{nx_1}{\tau_x} = \frac{1}{100}, \quad \frac{nx_2}{\tau_x} = \frac{9}{100}, \quad \frac{nx_3}{\tau_x} = \frac{1}{10}, \quad \frac{nx_4}{\tau_x} = \frac{7}{10},$$
$$\frac{nx_5}{\tau_x} = \frac{9}{10} \text{ et } \frac{nx_6}{\tau_x} = \frac{6}{5} > 1\,.$$

On sélectionne d'office l'unité 6 dans l'échantillon (car on se fixe $p_6 = 1$).

On répète ensuite la même procédure en considérant cette fois que l'on doit sélectionner $n - 1 = 2$ unités dans $U \setminus \{6\}$:

$$\sum_{i \in U \setminus \{6\}} x_i = 180$$

et donc :

$$\frac{(n-1)x_1}{\sum_{i \in U \setminus \{6\}} x_i} = \frac{1}{90}, \quad \frac{(n-1)x_2}{\sum_{i \in U \setminus \{6\}} x_i} = \frac{1}{10}, \quad \frac{(n-1)x_3}{\sum_{i \in U \setminus \{6\}} x_i} = \frac{1}{9},$$
$$\frac{(n-1)x_4}{\sum_{i \in U \setminus \{6\}} x_i} = \frac{7}{9}, \quad \frac{(n-1)x_5}{\sum_{i \in U \setminus \{6\}} x_i} = 1.$$

Les probabilités d'inclusion sont donc :

$$p_1 = \frac{1}{90}, \quad p_2 = \frac{1}{10}, \quad p_3 = \frac{1}{9}, \quad p_4 = \frac{7}{9}, \quad p_5 = 1, \quad p_6 = 1\,.$$

Les unités 5 et 6 sont sélectionnées d'office et le problème d'échantillonnage se réduit à la sélection d'une unité dans $U \setminus \{5,6\} = \{1,2,3,4\}$ tout en respectant les probabilités d'inclusion p_1, p_2, p_3 et p_4 spécifiées ci-dessus.

4.4.3 Un exemple de sondage PPS

L'exemple numérique suivant, tiré de l'ouvrage de Pascal Ardilly de 2006 (p. 146-147), illustre l'avantage qu'il y a à choisir un tirage à probabilités proportionnelles à une variable auxiliaire $\mathcal{X}$, elle-même approximativement proportionnelle à la variable d'intérêt $\mathcal{Y}$.

Exemple

Dans cet exemple, la population U est constituée de 5 communes. La variable d'intérêt $\mathcal{Y}$ est la variable « nombre d'habitants dans la commune » ; nous désirons en estimer la moyenne dans la population, soit le nombre moyen d'habitants par commune.

Si les valeurs de $\mathcal{Y}$ dans la population nous sont *a priori* inconnues, nous disposons en revanche dans la base de sondage du nombre x_i de logements dans chaque commune i de la population. Cette information auxiliaire fort riche va nous permettre de définir un sondage PPS.

Les distributions de $\mathcal{X}$ et de $\mathcal{Y}$ dans la population sont en réalité les suivantes (d'après le Recensement français de 1990) :

TABLEAU 4.1 – *La population U*

Communes	Nombre de logements ($\mathcal{X}$)	Nombre d'habitants ($\mathcal{Y}$)
(1) Antibes	48 812	70 688
(2) Cagnes	23 227	41 303
(3) St Laurent du Var	12 383	24 475
(4) Vence	9 341	15 364
(5) Villefranche/Mer	4 915	8 123
Total	$\tau_x = 98\,678$	$\tau_y = 159\,953$
Moyenne	$\mu_x = 19\,735{,}6$	$\mu_y = 31\,990{,}6$

Supposons que, pour des raisons budgétaires, on doive se limiter à un échantillonnage de taille 2. On a alors 10 échantillons possibles.

Considérons d'abord le cas où l'on échantillonne à probabilités égales sans remise. On estime alors la moyenne-population μ_y par $\hat{\mu}_{y;\mathrm{PESR}} = \bar{y}$, la moyenne des valeurs observées pour la variable $\mathcal{Y}$ dans l'échantillon prélevé. Les valeurs (arrondies) de cet estimateur dans les 10 échantillons possibles sont présentées dans la deuxième colonne du tableau 4.4.

Si l'on fait appel à l'échantillonnage PPS, on affecte aux 5 communes de la population les probabilités d'inclusion calculées dans la dernière colonne du tableau 4.2, selon la

formule

$$p_i = \frac{nx_i}{\tau_x} = \frac{2x_i}{98\,678}.$$

On estime alors le total τ_y par l'estimateur de Horvitz-Thompson

$$\hat{\tau}_{y;\mathrm{HT}} = \sum_{i \in S} \frac{y_i}{p_i}$$

et la moyenne μ_y par

$$\hat{\mu}_{y;\mathrm{HT}} = \frac{\hat{\tau}_{y;\mathrm{HT}}}{N} = \frac{\hat{\tau}_{y;\mathrm{HT}}}{5}.$$

Les valeurs dilatées y_i/p_i de la variable y sont calculées dans la dernière colonne du tableau 4.3. Les valeurs de l'estimateur de Horvitz-Thompson de μ_y dans les 10 échantillons possibles sont présentées dans la troisième colonne du tableau 4.4.

TABLEAU 4.2 – *Les probabilités d'inclusion du sondage PPS*

i	x_i	p_i
(1) Antibes	48 812	0,99
(2) Cagnes	23 227	0,47
(3) St Laurent du Var	12 383	0,25
(4) Vence	9 341	0,19
(5) Villefranche/Mer	4 915	0,10
Total	$\tau_x = 98\,678$	$n = 2$

TABLEAU 4.3 – *Les valeurs dilatées de la variable* y

i	y_i	p_i	y_i/p_i
(1) Antibes	70 688	0,99	71 402,02
(2) Cagnes	41 303	0,47	87 878,72
(3) St Laurent du Var	24 475	0,25	97 900,00
(4) Vence	15 364	0,19	80 863,16
(5) Villefranche/Mer	8 123	0,10	81 230,00

TABLEAU 4.4 – *Les valeurs de* $\hat{\mu}_{y;\mathrm{PESR}}$ *et de* $\hat{\mu}_{y;\mathrm{HT}}$ *dans les 10 échantillons possibles de 2 communes*

Echantillon s	$\hat{\mu}_{y;\mathrm{PESR}}$	$\hat{\mu}_{y;\mathrm{HT}}$
{1,2}	55 996	31 856
{1,3}	47 582	33 860
{1,4}	43 026	30 453
{1,5}	39 406	30 526
{2,3}	32 889	37 156
{2,4}	28 334	33 748
{2,5}	24 713	33 822
{3,4}	19 920	35 753
{3,5}	16 299	35 826
{4,5}	11 744	32 419

Pour disposer de la distribution d'échantillonnage de $\hat{\mu}_{y;\mathrm{HT}}$ et pouvoir ensuite calculer la variance de cet estimateur dans le cadre du sondage PPS, il nous faudrait compléter

le tableau 4 par une colonne supplémentaire contenant les probabilités de sélection $p(s)$ de chacun des 10 échantillons s possibles. Ceci est malheureusement impossible, car la valeur des probabilités de sélection dépend de l'algorithme utilisé pour effectuer le tirage. Il faudrait non seulement que l'on nous indique précisément les modalités du tirage, mais que l'on nous fournisse aussi une formule pour $p(s)$, ce qui est en réalité infaisable à de très rares exceptions près car les calculs deviennent rapidement inextricables.

Mais dans notre exemple, nous n'avons pas besoin de déterminer la valeur de la variance de l'estimateur de Horvitz-Thompson de μ_y pour nous rendre compte du gain important en efficacité du sondage PPS par rapport au sondage PESR. Le tableau 4.4 nous montre très clairement que les estimations possibles du nombre moyen d'habitants par commune sont bien moins dispersées autour de la valeur exacte de μ_y — égale, rappelons-le, à 31 991 — dans le cas du tirage à probabilités proportionnelles au nombre X de logements dans la commune que dans le cas du tirage à probabilités égales. La raison de cette bien meilleure efficacité du sondage PPS réside dans le lien de proportionnalité fort qui existe entre le nombre de logements et le nombre d'habitants des communes de la population.

Notez que si l'on n'avait pas pu disposer du nombre de logements fourni par le Recensement Général de la Population de 1990, on aurait pu penser à utiliser comme autre variable auxiliaire X le nombre d'habitants des communes de la population fourni par le Recensement Général de la Population de 1982.

4.4.4 Sélection des individus selon le plan de sondage PPS

Déterminer les probabilités d'inclusion des individus de la population pour un plan de sondage PPS n'est donc pas très compliqué. Mais comment fait-on ensuite pour tirer un échantillon tout en respectant ces probabilités d'inclusion particulières ?

Différents algorithmes de sélection de l'échantillon ont été proposés et sont disponibles dans certains logiciels spécifiques. Les procédures de tirage correspondantes constituent autant de façons différentes d'attribuer des valeurs aux probabilités de sélection $p(s)$ des échantillons possibles s, tout en satisfaisant la contrainte selon laquelle, pour chaque individu i de la population, sa probabilité d'inclusion p_i est bien égale à la somme des probabilités de sélection des échantillons qui contiennent cet individu i : pour tout $i \in U$, il faut que

$$p_i = \sum_{s \in \mathbb{S} \text{ tel que } i \in s} p(s) \, .$$

De façon générale, un des obstacles majeurs à l'utilisation du tirage à probabilités inégales réside dans la difficulté rencontrée pour le calcul des probabilités d'inclusion p_{ij} d'ordre 2, indispensables pour le calcul de la variance des estimateurs de Horvitz-Thompson, mais malheureusement fonctions de l'algorithme de tirage utilisé.

Par ailleurs, il peut arriver — nous l'avons déjà mentionné — que l'on obtienne des estimations *négatives* de la variance de $\hat{\tau}_{\mathrm{HT}}$. En effet, si l'on regarde l'expression de $\widehat{\mathrm{V}}_2(\hat{\tau}_{\mathrm{HT}})$ donnée dans la Section 4.3.5, on se rend compte que l'on risque d'obtenir une

estimation négative de la variance de $\hat{\tau}_{\mathrm{HT}}$ lorsque suffisamment de facteurs $p_i p_j - p_{ij}$ sont eux-mêmes négatifs. Pour éviter de se retrouver dans une telle situation, le sondeur apprécie les algorithmes satisfaisant les **conditions de Yates-Grundy** , c'est-à-dire assurant, pour tout $i, j \in U$ (avec $j \neq i$), que

$$p_i p_j - p_{ij} \geq 0.$$

Parmi les procédures d'échantillonnage permettant d'assurer à chaque individu i une probabilité d'inclusion p_i fixée d'avance, nous pouvons relever deux algorithmes particulièrement simples et aux caractéristiques fort intéressantes : le *tirage de Poisson* et le *tirage systématique sur un fichier de probabilités cumulées*.

a) Le tirage de Poisson

Procédure

Le tirage de Poisson est la généralisation naturelle du tirage de Bernoulli au cas des probabilités inégales : chaque unité i de la population (ou base de sondage) U est sélectionnée de manière indépendante avec une probabilité p_i.

Pour chaque unité $i = 1, \ldots, N$ de la population :

- on génère un nombre aléatoire u_i selon une loi $\mathcal{U}(0,1)$ (loi continue uniforme sur l'intervalle $(0,1)$), via la fonction ALEA() ou RAND() du tableur ;
- si $u_i \leq p_i$, alors l'unité statistique i est sélectionnée pour faire partie de l'échantillon ;
  si, au contraire, $u_i > p_i$, alors l'unité statistique i n'est pas sélectionnée.

Illustrons cette procédure d'échantillonnage sur un petit exemple.

Exemple

Nous souhaitons tirer un échantillon de taille 5 dans une population constituée de 30 opérations comptables ; pour ce faire, nous décidons de suivre un plan de sondage à probabilités proportionnelles aux montants (en centaines d'euros) de ces opérations comptables, montants que l'on trouve dans la base de sondage.

Nous sommes ici dans une situation bien connue dans le contexte des vérifications comptables, où l'on peut assez régulièrement faire l'hypothèse que les montants des *erreurs* comptables sont approximativement proportionnels aux montants des opérations comptables.

Vous trouverez dans le tableau 4.5 le montant x_i de chaque opération comptable i de la population (colonne 2), ainsi que la probabilité d'inclusion $p_i = 5x_i/\tau_x$ qui en découle (colonne 3). Remarquez que la somme des probabilités d'inclusion associées aux 30 opérations comptables de la population est bien égale à 5, la taille que nous nous sommes fixée pour l'échantillon.

Appliquons le tirage de Poisson. Les nombres u_i obtenus via la fonction ALEA() pour chaque opération i sont repris dans la colonne 4 du tableau 4.5. Leur comparaison avec les probabilités d'inclusion p_i nous conduit à sélectionner les opérations comptables n° 6, 8, 22, 27 et 30.

TABLEAU 4.5 – *Tirage de Poisson*

i	x_i	p_i	u_i
1	48	0,1485	0,7941
2	54	0,1671	0,2564
3	98	0,3032	0,5929
4	42	0,1300	0,4655
5	8	0,0248	0,0533
6	99	0,3063	0,0677
7	37	0,1145	0,5602
8	71	0,2197	0,1797
9	99	0,3063	0,4316
10	12	0,0371	0,5675
11	39	0,1207	0,3139
12	1	0,0031	0,7999
13	66	0,2042	0,5202
14	16	0,0495	0,8534
15	72	0,2228	0,6308
16	37	0,1145	0,3625
17	50	0,1547	0,2876
18	70	0,2166	0,8041
19	61	0,1887	0,8778
20	89	0,2754	0,7841
21	38	0,1176	0,6275
22	61	0,1887	0,0124
23	90	0,2785	0,3025
24	41	0,1269	0,5845
25	34	0,1052	0,1436
26	8	0,0248	0,4487
27	43	0,1330	0,0046
28	82	0,2537	0,8852
29	64	0,1980	0,4689
30	86	0,2661	0,1959
Total	1616	5	

Propriétés

Passons en revue quelques propriétés marquantes du tirage de Poisson.

1. La taille de l'échantillon est en réalité *aléatoire*, et il y a une probabilité non nulle de sélectionner un échantillon vide ou de sélectionner toute la population ! La valeur n considérée dans le calcul des probabilités d'inclusion p_i correspond ici à la taille *moyenne* des échantillons possibles.

2. Comme les unités sont sélectionnées indépendamment les unes des autres, on a
$$p_{ij} = p_i p_j \quad \text{pour tout } i \neq j.$$
Par conséquent, $p_i p_j - p_{ij} = 0$ pour tout $i \neq j$.

3. Le plan de sondage est le suivant : l'ensemble $\mathbb{S}$ de tous les échantillons possibles n'est autre que l'ensemble de toutes les parties (ou sous-ensembles) de U et, pour tout échantillon possible s :
$$p(s) = \left\{\prod_{i \in s} p_i\right\} \times \left\{\prod_{i \in U \setminus s} (1 - p_i)\right\}$$
où le premier terme entre accolades correspond au produit des probabilités d'inclusion des individus qui composent l'échantillon s et le second terme entre accolades est le produit des probabilités de NON inclusion des individus qui ne font pas partie de l'échantillon s.

4. On montre que, si l'échantillon est sélectionné par tirage de Poisson[2] :

[2] Le tirage de Poisson donnant lieu à un échantillonnage de taille *aléatoire*, on ne peut faire appel à l'expression de Sen, Yates et Grundy (cf. Section 4.3.5) pour la variance de $\hat{\tau}_{\text{HT}}$. Il faut donc partir de l'expression $\text{V}_1(\hat{\tau}_{\text{HT}})$.

$$V(\hat{\tau}_{HT}) = V_1(\hat{\tau}_{HT}) = \sum_{i \in U} \frac{y_i^2}{p_i}(1 - p_i) .$$

Cette variance peut être estimée sans biais en appliquant une nouvelle fois la démarche de Horvitz-Thompson :

$$\hat{V}(\hat{\tau}_{HT}) = \hat{V}_1(\hat{\tau}_{HT}) = \sum_{i \in S} \frac{y_i^2}{p_i^2}(1 - p_i) .$$

L'intérêt du plan de Poisson est sans nul doute son extrême simplicité et le fait que la seule connaissance des probabilités d'inclusion p_i d'ordre un nous permette d'évaluer la précision associée à l'estimateur de Horvitz-Thompson.

b) Le tirage systématique sur un fichier de probabilités cumulées

[!! Section réservée aux étudiants d'ECON et INGE !!]

Procédure

Contrairement au tirage de Poisson qui donnait lieu à un échantillonnage de taille aléatoire, le tirage systématique considéré ici permet un échantillonnage de taille fixe n.

Les individus de la population ayant une probabilité d'inclusion égale à 1 sont automatiquement sélectionnés pour faire partie de l'échantillon. S'ils sont au nombre de n^*, il nous reste alors à sélectionner $(n - n^*)$ individus parmi les $(N - n^*)$ individus restants dans la population.

Notons N^* ce nombre d'individus qui restent dans la population et considérons le fichier dans lequel sont repris, dans un ordre que l'on peut considérer comme aléatoire, ces N^* individus et leurs probabilités d'inclusion dont la somme, rappelons-le, est nécessairement égale à $(n - n^*)$. Par facilité, renumérotons ces individus de 1 à N^*, selon leur ordre de présentation dans le fichier.

On construit alors la série des probabilités d'inclusion *cumulées* C_i, pour $i = 1, \dots, N^*$:

- à l'individu n° 1 est associée la probabilité cumulée $C_1 = p_1$;
- à l'individu n° 2 est associée la probabilité cumulée $C_2 = p_1 + p_2$;
- à l'individu n° 3 est associée la probabilité cumulée $C_3 = p_1 + p_2 + p_3$;
- etc.

De manière générale, la probabilité cumulée C_i associée à l'individu n° i est obtenue en faisant la somme de sa probabilité d'inclusion et de celles des individus qui le précèdent dans le fichier :

$$C_i = \sum_{j=1}^{i} p_j .$$

Notons aussi que la probabilité cumulée associée au dernier individu du fichier n'est autre que la somme des probabilités d'inclusion des N^* individus restants dans la population et est donc nécessairement égale au nombre $(n - n^*)$ d'individus qu'il nous faut encore sélectionner.

On génère ensuite un nombre ALEA entre 0 et 1, selon une loi uniforme sur l'intervalle (0,1). Le *premier* individu sélectionné est alors le premier individu du fichier dont la

probabilité cumulée dépasse ce nombre ALEA ; le *deuxième* individu sélectionné est le premier du fichier dont la probabilité cumulée dépasse le nombre (ALEA + 1) ; le *troisième* individu sélectionné est le premier du fichier dont la probabilité cumulée dépasse le nombre (ALEA + 2) ; et ainsi de suite jusqu'au dernier individu sélectionné, qui est le premier du fichier dont la probabilité cumulée dépasse (ALEA $+ (n - n^*) - 1$).

On peut résumer cette règle de sélection des individus de la manière suivante : pour k allant de 1 à $(n - n^*)$, le k-ème individu sélectionné porte le numéro i_k si

$$C_{i_k-1} \leq \text{ALEA} + (k-1) < C_{i_k} .$$

Cette procédure d'échantillonnage revient à progresser de manière systématique dans le fichier des probabilités cumulées avec un PAS de tirage valant 1.

Exemple

Reprenons l'exemple considéré pour le tirage de Poisson mais appliquons cette fois le tirage systématique pour sélectionner notre échantillon de taille 5. Notez que les experts en audit ont coutume de désigner cette procédure de tirage particulière appliquée sur une base de sondage d'opérations comptables sous le nom de *sondage MUS* (*Monetary Sampling Unit*).

Le tableau 4.6 reprend, pour chaque opération comptable i, son montant x_i, la probabilité d'inclusion $p_i = 5x_i/\tau_x$ et la probabilité cumulée C_i qui lui sont associées. Supposons que notre tableur attribue la valeur 0,8320 au nombre ALEA. Nous avons alors :

- ALEA = 0,8320 ;
- ALEA + 1 = 1,8320 ;
- ALEA + 2 = 2,8320 ;
- ALEA + 3 = 3,8320 ;
- ALEA + 4 = 4,8320.

Ceci nous conduit à sélectionner en premier lieu l'opération comptable n° 6, puisque cette opération est la première du fichier dont la probabilité cumulée, égale à 1,0798, dépasse le nombre ALEA. Nous sélectionnons ensuite les opérations comptables n° 11, 18, 23 et 30.

TABLEAU 6 – *Tirage MUS*

i	x_i	p_i	C_i	i	x_i	p_i	C_i
1	48	0,1485	0,1485	16	37	0,1145	2,4722
2	54	0,1671	0,3156	17	50	0,1547	2,6269
3	98	0,3032	0,6188	**18**	70	0,2166	**2,8434**
4	42	0,1300	0,7488	19	61	0,1887	3,0322
5	8	0,0248	0,7735	20	89	0,2754	3,3075
6	99	0,3063	**1,0798**	21	38	0,1176	3,4251
7	37	0,1145	1,1943	22	61	0,1887	3,6139
8	71	0,2197	1,4140	**23**	90	0,2785	**3,8923**
9	99	0,3063	1,7203	24	41	0,1269	4,0192
10	12	0,0371	1,7574	25	34	0,1052	4,1244
11	39	0,1207	**1,8781**	26	8	0,0248	4,1491
12	1	0,0031	1,8812	27	43	0,1330	4,2822
13	66	0,2042	2,0854	28	82	0,2537	4,5359
14	16	0,0495	2,1349	29	64	0,1980	4,7339
15	72	0,2228	2,3577	**30**	86	0,2661	**5,0000**

Total	1616	5

Propriétés

Outre sa facilité de mise en œuvre, le tirage considéré ici présente l'avantage de donner lieu à des probabilités d'inclusion p_{ij} d'ordre 2 dont on a pu déterminer l'expression de manière exacte (il s'agit d'une expression relativement complexe… faisant intervenir la taille n de l'échantillon et les probabilités d'inclusion p_i et p_j d'ordre 1). Sur la base de cette expression, on a pu montrer que, si la taille N de la population est grande et que la taille n de l'échantillon est négligeable par rapport à N, la variance de $\hat{\tau}_{\mathrm{HT}}$ pouvait être approximée par l'expression que voici :

$$\sum_{i\in U} p_i \left(1 - \frac{n-1}{n} p_i\right) \left(\frac{y_i}{p_i} - \frac{\tau}{n}\right)^2,$$

et estimée par l'expression suivante :

$$\frac{1}{2(n-1)} \sum_{i\in S} \sum_{\substack{j\in S \\ j\neq i}} \left[1 - (p_i + p_j) + \frac{1}{n} \sum_{i\in U} p_i^2\right] \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

avec un très léger biais si N est grand. Remarquez que cet estimateur de la variance de $\hat{\tau}_{\mathrm{HT}}$ ne fait plus intervenir les probabilités d'inclusion d'ordre 2, ce qui nous permet d'éviter de devoir les calculer… à notre grand soulagement !

Enfin, on peut montrer que la procédure de tirage systématique vérifie la propriété de Yates-Grundy mentionnée précédemment, qui assure que l'estimateur de la variance de l'estimateur de Horvitz-Thompson ne peut prendre que des valeurs positives. Voilà encore un point positif en faveur de cette procédure d'échantillonnage.

4.5 Conclusion : quelques réflexions sur l'opportunité des tirages à probabilités inégales et de la pondération associée

Quand faut-il choisir un sondage PPS plutôt qu'un sondage aléatoire simple ? Idéalement, il faudrait répondre à cette question en passant par la comparaison théorique des variances des estimateurs de τ ou de μ pour le sondage PPS d'une part et le sondage PESR d'autre part. Malheureusement, cette comparaison est extrêmement complexe. Il faut donc souvent s'en remettre à un raisonnement de type qualitatif.

Supposons que la base de sondage contienne les valeurs d'une variable auxiliaire strictement positive X pour tous les individus de la population, et que l'on puisse raisonnablement penser que cette variable X est approximativement proportionnelle à la variable d'intérêt Y. Dans ce cas, si les valeurs de X varient sensiblement d'un individu à l'autre, il est préférable d'utiliser un sondage à probabilités d'inclusion proportionnelles à ces valeurs de X. En revanche, si l'on est dans une configuration où les valeurs de X varient relativement peu d'un individu de la population à l'autre, le sondage aléatoire à probabilités égales sans remise sera sans doute préférable, ne serait-ce que parce qu'il est plus simple tout en donnant lieu à une précision voisine de celle que l'on obtiendrait avec un sondage PPS.

Il me faut également vous mettre en garde contre les effets parfois très néfastes des probabilités inégales. Si la variable d'intérêt Y et la variable auxiliaire X utilisée pour définir les probabilités d'inclusion ne sont pas vraiment liées entre elles par un lien de proportionnalité, faire appel à l'échantillonnage à probabilités inégales risque de causer une augmentation notable de la variance des estimateurs par rapport à l'échantillonnage PESR.

Attention donc à l'usage du sondage PPS lorsque plusieurs variables d'intérêt entrent en jeu : il se peut qu'il soit bien adapté pour certaines de ces variables d'intérêt, mais pas du tout pour d'autres.

En tout état de cause, lorsqu'on doit effectuer une enquête dans laquelle le questionnaire comprend de nombreuses questions, il peut être sage de préférer le simple sondage PESR au sondage PPS. En faisant appel au sondage aléatoire simple plutôt qu'au sondage PPS, il nous faut accepter de faire peut-être « moins bien » sur certaines variables bien corrélées à la variable « taille » X, mais on évite en retour des catastrophes sur d'autres variables qui ne seraient pas du tout en rapport avec la « taille ».

4.6 Exercice 4.1

On désire estimer la production de céréales des 300 exploitations agricoles d'une certaine région. Pour cela, on veut effectuer un sondage PPS de taille (moyenne) $n = 30$ en utilisant comme variable auxiliaire X la variable « Surface Agricole Utile » (SAU ; exprimée en ha).

La base de sondage, contenant les valeurs de la variable SAU pour toutes les exploitations agricoles de la population, est disponible dans la feuille « U » du fichier Data_ex_4_1.xlsx.

a) Déterminez (dans le fichier Excel que vous venez de télécharger) les probabilités d'inclusion des différentes exploitations agricoles de la population. Que vaut la somme de toutes ces probabilités d'inclusion ?

b) Réalisez un tirage de Poisson en utilisant les nombres aléatoires u_i $(i = 1, \dots, 300)$ listés dans la feuille « Poisson » du fichier Data_ex_4_1.xlsx.
- **(i)** Quelle est la taille de l'échantillon sélectionné ?
- **(ii)** Quelles sont les exploitations agricoles sélectionnées au cours de ce tirage de Poisson ?

c) [!! Question réservée aux étudiants d'ECON et INGE !!]
Réalisez un tirage systématique sur le fichier de probabilités cumulées en prenant le nombre ALEA égal à 0,70086.
Quelles sont les exploitations agricoles sélectionnées au cours de ce tirage systématique ?

d) Les productions de céréales (en tonnes) des 30 exploitations agricoles sélectionnées par le tirage systématique que vous venez d'appliquer s'élèvent à (à lire ligne par ligne) :

438	370	389	233	484	419	507	609	309	362
373	588	573	487	458	431	306	186	519	512
524	428	515	262	604	595	558	186	299	594

- **(i)** Vérifiez à partir de ces observations si l'hypothèse sous-jacente à la validité du choix du sondage PPS considéré dans cet exercice semble satisfaite ou non.
- **(ii)** A quelle estimation de la production de céréales *totale* dans la population vous conduisent ces observations ?
- **(iii)** A quelle estimation de la production de céréales *moyenne* par exploitation agricole vous conduisent ces observations ?

Chapitre 5
Le sondage en grappes et le sondage à plusieurs degrés

5.1 Introduction

Nous allons, dans ce chapitre, nous intéresser à deux autres méthodes de sondage aléatoires dont la pratique est très répandue : le *sondage en grappes* et le *sondage à deux degrés* ou, de manière plus générale, *à plusieurs degrés*. Comme nous allons le voir, l'utilisation de ces méthodes particulières est essentiellement motivée par la nature des données à recueillir, des considérations de coût ou de faisabilité, et la mauvaise qualité, voire l'inexistence, d'une base de sondage pour l'ensemble de la population à sonder.

La mise en œuvre de ces méthodes est très fréquente dans la pratique des sondages, alors qu'elles risquent de donner lieu à une moins bonne précision que le sondage aléatoire simple. Il est dès lors nécessaire de prendre conscience des facteurs susceptibles de limiter l'efficacité de ces méthodes. Cette prise de conscience nous permettra d'identifier les critères à prendre en considération lorsqu'on souhaite évaluer s'il est pertinent ou non de faire appel à ce type de méthodes pour effectuer notre sondage.

5.1.1 Le principe général du sondage à deux ou plusieurs degrés et du sondage en grappes

Les méthodes de sondage aléatoires étudiées dans ce chapitre tirent parti du découpage — souvent naturel — de la population-cible en un certain nombre de sous-ensembles.

Supposons que la population U soit partitionnée en un certain nombre de sous-ensembles d'individus. Ces sous-ensembles ou sous-populations sont généralement appelé(e)s les *unités primaires* de la population. Pour le **sondage à deux degrés**, on prélève aléatoirement, par tirage PESR ou PISR, un échantillon d'unités primaires ; puis, dans chaque unité primaire sélectionnée, on prélève, par tirage PESR ou PISR, un échantillon d'individus (voir la figure 5.1). La construction de l'échantillon d'individus de U se fait donc en deux étapes, ce qui explique la terminologie de « sondage à deux degrés ». Le premier degré du sondage s'applique dans l'ensemble des unités primaires, tandis que le second degré du sondage s'applique au niveau des unités primaires qui ont été prélevées au premier degré du sondage.

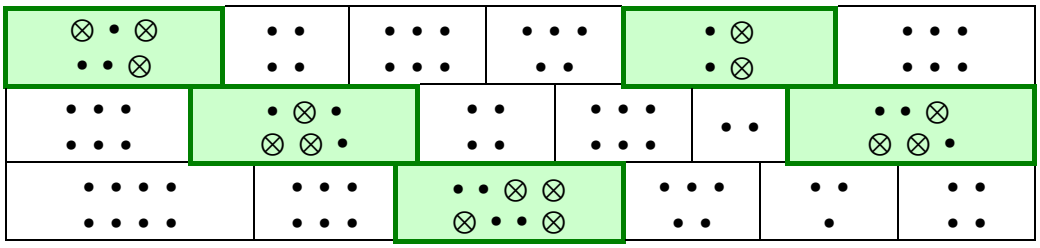


FIGURE 5.1 – *Le principe général du sondage à deux degrés*

Le principe du sondage à deux degrés peut bien évidemment être généralisé pour donner lieu à un **sondage à plusieurs degrés**. Si, par exemple, la population U est découpée en *unités primaires*, elles-mêmes constituées d'*unités secondaires* contenant chacune un certain nombre d'individus, on peut prélever l'échantillon d'individus par

un **sondage à trois degrés** : on sélectionne aléatoirement un échantillon d'unités primaires ; puis, dans chaque unité primaire sélectionnée, on prélève aléatoirement un échantillon d'unités secondaires ; enfin, dans chaque unité secondaire tirée au deuxième degré du sondage, on prélève aléatoirement un échantillon d'individus (voir la figure 5.2). Le sondage à trois degrés est donc un sondage à deux degrés dont le deuxième degré est lui-même un sondage à deux degrés.

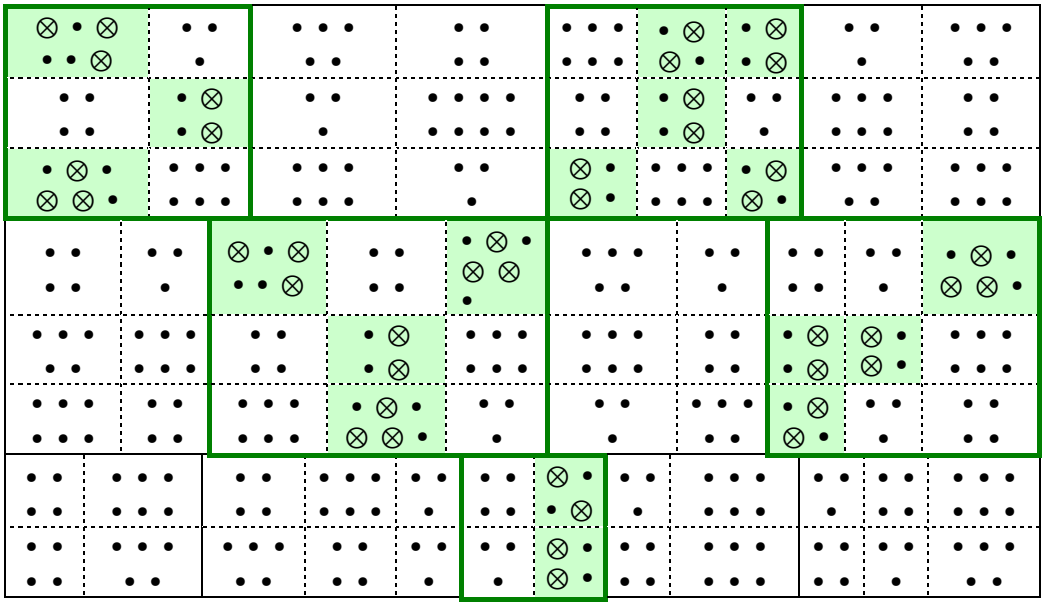


FIGURE 5.2 – *Le principe général du sondage à trois degrés*

Si les unités secondaires sont à leur tour partitionnées en un certain nombre d'unités tertiaires, on peut penser à mettre en place un sondage à quatre degrés. Etc.

Dans certaines situations — essentiellement lorsque les unités primaires qui partitionnent la population-cible *U* sont fort nombreuses et ne regroupent chacune qu'un nombre relativement limité d'individus — on peut décider de tirer un échantillon d'unités primaires, puis de réaliser un *recensement* des unités primaires sélectionnées. En d'autres termes, on fait appel à un sondage à deux degrés où l'on sélectionne, au deuxième degré du sondage, l'ensemble des individus de chaque unité primaire prélevée au premier degré (voir la figure 5.3). Ce type de sondage porte généralement le nom de **sondage en grappes**, car les unités primaires sont réellement considérées comme des « grappes » d'individus. On a d'ailleurs l'habitude, dans ce cas, de parler des « grappes » de la population plutôt que des unités primaires de celle-ci.

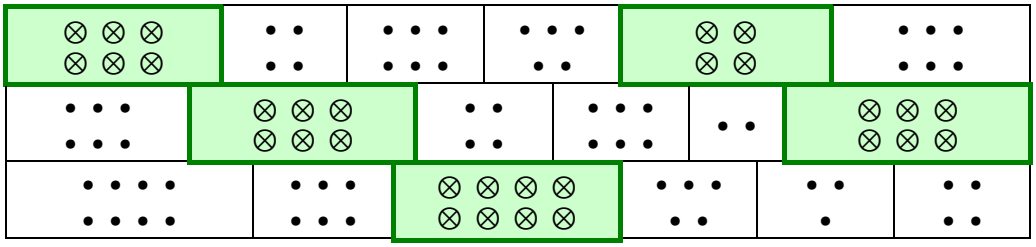


FIGURE 5.3 – *Le principe général du sondage en grappes*

5.1.2 Exemples de situation où l'on fait appel à ce type de sondages aléatoires « complexes »

L'utilisation de ce type de sondages aléatoires « complexes » est essentiellement motivée par des considérations de coût et de faisabilité. Passons en revue quelques exemples de situations où le choix du sondage à plusieurs degrés ou en grappes apparaît comme naturel et bien pratique.

Il peut arriver que l'on ne dispose pas d'une base de sondage complète reprenant l'ensemble des unités de la population à sonder. Il nous est alors impossible de prélever un échantillon par sondage aléatoire simple. Le sondage à deux degrés, par exemple, peut alors être une solution intéressante, pour autant que la population soit découpée en unités primaires dont il est aisé de dresser la liste et qu'il soit possible de constituer ensuite une liste exhaustive des individus ou unités statistiques qui composent chaque unité primaire sélectionnée au premier degré du sondage.
Imaginez, par exemple, que vous deviez tirer aléatoirement un échantillon dans la population des élèves scolarisés dans une certaine région de votre pays. Vous n'avez pas accès à la base de sondage reprenant l'ensemble de ces élèves. En revanche, il ne vous est pas difficile d'obtenir la liste de toutes les écoles de la région considérée. Vous pouvez alors tirer un échantillon d'écoles, puis contacter les directions des écoles sélectionnées afin d'organiser un sondage aléatoire parmi leurs élèves. Si, au sein de chacune de ces écoles, vous prélevez un échantillon d'élèves par un tirage PESR, par exemple, vous mettez en œuvre un sondage à deux degrés. Si vous optez plutôt pour le prélèvement d'un échantillon de classes dans chacune des écoles prélevées au premier degré du sondage, puis pour le recensement des classes sélectionnées, vous réalisez un sondage à deux degrés dont le deuxième degré est en réalité un sondage en grappes. Si vous préférez tirer un échantillon aléatoire d'élèves dans chacune des classes sélectionnées, c'est un sondage à trois degrés que vous appliquez.

L'utilisation d'un sondage à plusieurs degrés ou en grappes permet aussi, dans certains cas, de réduire drastiquement les coûts liés aux déplacements que doivent faire les enquêteurs pour aller à la rencontre des personnes enquêtées.
Imaginez par exemple que la population à sonder soit celle d'une région relativement vaste. Le sondage aléatoire simple risque fort de nous fournir un échantillon dispersé géographiquement sur l'ensemble du territoire de la région, ce qui va induire des coûts de déplacement des enquêteurs extrêmement importants. Si l'on opte plutôt pour un sondage à deux ou plusieurs degrés pour lequel les unités primaires correspondent à des zones géographiques de dimensions relativement faibles, on va très fortement limiter la zone de travail de chaque enquêteur et réduire ainsi significativement les coûts — et les pertes de temps — liés à ses déplacements.

Dans certains cas, l'utilisation du sondage en grappes ou à deux degrés s'impose de par la nature même des unités à sélectionner : il arrive en effet que les unités de sondage soient naturellement groupées « en paquets » et qu'il soit dès lors plus économique d'y accéder par paquets, c'est-à-dire par grappe ou par unité primaire.

- Il en est ainsi, par exemple, des *contrôles par lot*. Il s'agit de contrôler des livraisons de produits fabriqués en grande série, ou de produits alimentaires (par exemple des fruits). Physiquement, les objets ou produits à contrôler sont

conditionnés par caisses. Quoi de plus naturel alors que d'organiser l'échantillonnage par caisses !

- Des enquêtes sont régulièrement menées auprès des passagers des compagnies aériennes pour connaître leur opinion et évaluer leur degré de satisfaction pour les services offerts : moyens d'accès aux aéroports, qualité de l'accueil, confort, repas, etc. On a un regroupement naturel des usagers par vol ou par avion. Il est alors bien pratique de constituer l'échantillon d'usagers en sélectionnant un échantillon de vols.
- Certaines études médicales sont réalisées en prélevant un d'échantillon de médecins considérés, pour l'enquête, comme des grappes ou des unités primaires de patients ou de prescriptions.

Citons encore d'autres exemples de situations où l'utilisation des sondages en grappes ou à plusieurs degrés est fréquente.

- Le sondage en grappes se prête bien à certaines études écologiques, par exemple celles ayant pour objectif d'étudier la santé des arbres de massifs forestiers. A partir de photographies aériennes et de relevés topographiques, on définit sur la carte un quadrillage du territoire, puis on sélectionne aléatoirement un échantillon de carrés dans lesquels les enquêteurs spécialisés iront effectuer leurs analyses. On parle plutôt, dans ce contexte, de *sondage aréolaire*, car les grappes — les grappes d'arbres dans notre exemple — correspondent à des aires géographiques.
- Le sondage en grappes ou à deux degrés peut s'avérer bien utile pour certaines études de marché. Le marché d'un produit de grande distribution, comme la petite confiserie par exemple, est constitué d'une multitude de points de vente dont, souvent, l'activité est autre que la vente de ce produit. Il n'existe bien sûr pas de liste exhaustive de ces points de vente. On découpe alors le territoire concerné par l'étude en quartiers ou îlots géographiques, et on prélève ensuite aléatoirement un échantillon de ces quartiers ou îlots. Dans cet échantillon de « grappes de points de vente », les enquêteurs iront recenser systématiquement tous les points de vente distribuant le produit.
- Les instituts de sondage qui mesurent l'audience des médias (tels que la télévision ou la radio, par exemple) utilisent des méthodes d'échantillonnage à deux degrés : ils sélectionnent des communes selon un plan de sondage à probabilités inégales proportionnelles aux tailles de ces dernières ; puis, dans chaque commune sélectionnée, ils prélèvent un échantillon de ménages ou de foyers en contrôlant un certain nombre de variables de quotas (de manière à mimer un sondage stratifié proportionnel) ; les foyers sélectionnés sont alors considérés comme des grappes d'auditeurs.
- Enfin, les échantillonnages de ménages ou d'individus réalisés par l'INSEE (l'Institut National de la Statistique et des Etudes Economiques) en France sont des exemples de sondages à trois ou quatre degrés. En général, les unités finales de sondage sont des logements (c'est le cas, par exemple, pour les études des logements, des équipements, des dépenses liées à la consommation alimentaire, etc.). Le tirage de ces logements est organisé comme suit :
  - au premier degré du sondage, on procède à un tirage de cantons ou d'agglomérations, avec des probabilités proportionnelles à leur population, et stratification préalable par région et type d'habitat (on distingue les zones rurales et différentes tailles de zones urbaines) ;

- à l'intérieur des unités primaires sélectionnées, on effectue un tirage de communes ou de quartiers — ce sont nos unités secondaires — toujours à probabilités inégales proportionnelles à leur population ;
- enfin, dans les unités secondaires sélectionnées, on tire au sort des logements ; on prélève généralement, par tirage PESR, le même nombre de logements dans toutes les unités secondaires considérées.

Dans certains cas, le sondage a trait non pas aux logements ou aux ménages qui les occupent, mais plutôt aux individus eux-mêmes (il en est ainsi, par exemple, pour les études sur les dépenses de loisirs ou de santé). On peut alors rajouter un quatrième degré au sondage, en prélevant aléatoirement *un* individu dans chaque ménage prélevé au degré précédent du sondage.

Avec ces quelques exemples, vous commencez fort probablement à comprendre pourquoi l'utilisation des sondages aléatoires à plusieurs degrés ou en grappes est si répandue. Mais ces méthodes de sondage particulières ont-elles d'autres qualités que celle de faciliter l'organisation de l'échantillonnage ? Que peut-on dire de leur efficacité ? Donnent-elles lieu à une meilleure ou, au contraire, à une moins bonne précision que le sondage aléatoire simple ? C'est ce que nous allons étudier à présent.

Même s'il peut être considéré comme un cas particulier du sondage à deux degrés, je vous propose de commencer par analyser le sondage en grappes. D'une part, ses propriétés statistiques sont plus faciles à établir que celles du sondage à deux degrés. D'autre part, l'étude du sondage en grappes nous permettra de mettre plus facilement en lumière les caractéristiques des unités primaires qui favorisent l'efficacité du sondage à deux ou plusieurs degrés.

5.2 Le sondage en grappes

5.2.1 Caractéristiques générales de la population et de l'échantillon

Voyons tout d'abord quelques caractéristiques générales de la population et de l'échantillon dans le contexte du sondage en grappes.

La population-cible U est constituée de N individus ou unités statistiques :

$$U = \{1,2, \dots, N\} = \{i; i = 1, \dots, N\}.$$

Cette population U est par ailleurs partitionnée en M sous-ensembles qui sont nos *grappes* d'individus (voir la figure 5.4). Nous les désignerons par $U_1, U_2, \cdots, U_M$. Notez que, dans les expressions mathématiques que nous allons être amenés à écrire, nous désignerons parfois la grappe U_g simplement par son numéro g.

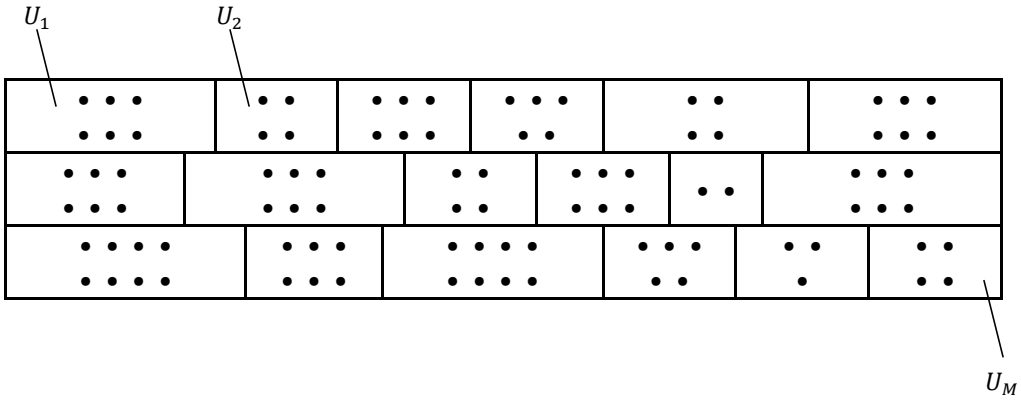


FIGURE 5.4 – *Partition de la population U en M grappes*

Nous aurons besoin, dans la suite, de bien distinguer la population U des individus, de taille N, de la population dont les unités sont les grappes, de taille M : nous désignerons cette dernière par U_{GR}. Ainsi,

$$U_{\text{GR}} = \{U_1, U_2, \cdots, U_M\} = \{U_g; g = 1, \cdots, M\}$$
$$= \{1,2, \dots, M\} = \{g; g = 1, \dots, M\}.$$

Chaque grappe contient un certain nombre d'individus de la population U : nous noterons N_g la taille de la grappe U_g. Il est clair que la taille N de la population U est égale à la somme des tailles des M grappes qui partitionnent U :

$$N = \sum_{g=1}^{M} N_g.$$

Les tailles des grappes ne sont pas toujours connues *a priori*. Dans certaines situations, on ne pourra déterminer que les tailles des grappes sélectionnées au cours du sondage.

La procédure d'échantillonnage consiste à prélever aléatoirement, par tirage PESR ou PISR, m grappes parmi les M grappes qui partitionnent la population U. On tire donc

aléatoirement un échantillon de grappes, S_{GR}, de taille fixe m, dans la population U_{GR}. L'échantillon final S est alors constitué de l'ensemble des individus appartenant aux grappes sélectionnées (voir la figure 5.5) :

$$S = \bigcup_{g \in S_{GR}} U_g.$$

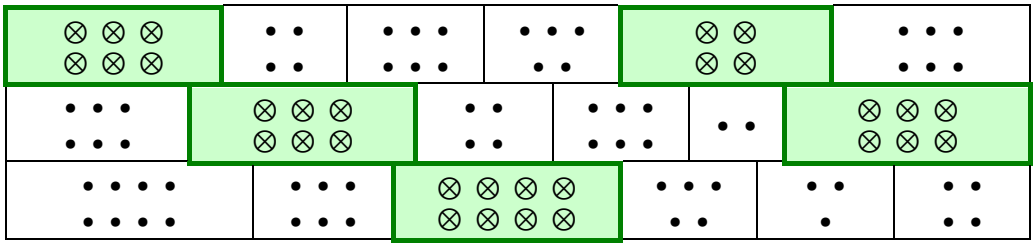


FIGURE 5.5 – *Echantillonnage en grappes*

La taille n_S de l'échantillon S des individus est la somme des tailles des m grappes sélectionnées :

$$n_S = \sum_{g \in S_{GR}} N_g.$$

Si les grappes qui partitionnent la population n'ont pas toutes la même taille (ce qui est souvent le cas), on ne peut prédire à l'avance avec certitude combien d'individus se retrouveront dans l'échantillon S ; autrement dit, la taille de cet échantillon S est *aléatoire*.

5.2.2 Paramètres à estimer : τ, μ et μ_τ

Quels paramètres allons-nous chercher à estimer ?

On peut être intéressé par l'estimation du **total** τ de la variable d'intérêt y dans la population U : τ se définit comme la somme des valeurs que prend la variable y sur tous les individus de U. Mais, si l'on définit τ_g comme le total de y dans la grappe n° g, autrement dit comme la somme des valeurs que prend y sur tous les individus de la grappe U_g, on peut réécrire τ sous la forme de la somme des totaux τ_g associés à toutes les grappes de la population :

$$\tau = \sum_{i \in U} y_i = \sum_{g=1}^{M} \left(\sum_{i \in U_g} y_i \right) = \sum_{g=1}^{M} \tau_g.$$

Exemple

Considérons, par exemple, une population U d'individus partitionnée en ménages (chaque ménage est ainsi une grappe d'individus) et intéressons-nous à la variable y associant à chaque individu le montant de ses frais médicaux au cours des 6 derniers mois. Le paramètre τ correspond au montant total des frais médicaux associés à l'ensemble des individus de la population ; ce montant total est aussi la somme des montants totaux des frais médicaux encourus par les différents ménages qui constituent la population.

On peut donc considérer le paramètre τ en adoptant deux points de vue différents :

1. Il peut se voir comme le total de la variable $\mathcal{Y}$ dans la population U des individus : $\tau = \sum_{i \in U} y_i$.
2. Si l'on se place au niveau de la population U_{GR} des grappes et qu'on y définit la variable $\mathcal{T}$ qui associe à chaque grappe U_g le total τ_g de $\mathcal{Y}$ dans cette grappe, on peut alors considérer le paramètre τ comme le total de cette variable $\mathcal{T}$ dans U_{GR} : $\tau = \sum_{g=1}^{M} \tau_g = \sum_{g \in U_{\text{GR}}} \tau_g$.

Le rapport de τ sur N correspond à la **moyenne** μ de la variable d'intérêt $\mathcal{Y}$ dans la population U. Mais on peut aussi être intéressé par le rapport de τ sur le nombre M de grappes, correspondant au **total moyen** de $\mathcal{Y}$ par grappe, c'est-à-dire à la moyenne des totaux τ_g associés à toutes les grappes de la population (ou encore la moyenne de la variable $\mathcal{T}$ dans la population des grappes U_{GR}) ; nous désignerons ce nouveau paramètre par μ_τ :

$$\mu = \frac{\tau}{N} = \frac{1}{N} \sum_{i \in U} y_i \qquad \text{et} \qquad \mu_\tau = \frac{\tau}{M} = \frac{1}{M} \sum_{g=1}^{M} \tau_g \, .$$

Exemple (suite)

A quoi correspondent ces deux paramètres, μ et μ_τ, dans notre exemple ? La moyenne μ est le montant des frais médicaux encourus, *en moyenne*, *par* chaque *individu* de la population au cours des 6 derniers mois. Le total moyen μ_τ correspond quant à lui au montant des frais médicaux encourus, *en moyenne*, par chaque *ménage* de la population au cours des 6 derniers mois.

5.2.3 Estimateurs de Horvitz-Thompson de τ, μ et μ_τ

Nous allons suivre la démarche de Horvitz-Thompson étudiée dans le chapitre précédent pour construire l'estimateur du total τ : rappelons-nous que, de manière générale, l'estimateur de Horvitz-Thompson du *total* d'une variable dans une population se définit comme la somme, sur les unités de l'échantillon prélevé dans cette population, des valeurs que prend la variable sur ces unités, valeurs divisées par les probabilités d'inclusion de ces unités.
Ensuite, puisque la moyenne μ est égale à τ/N et que le total moyen μ_τ est égal à τ/M, il suffira de diviser par N ou par M l'estimateur de τ pour obtenir les estimateurs de μ et de μ_τ, respectivement.

Le paramètre τ est la somme des valeurs que prend $\mathcal{Y}$ sur tous les individus de U :

$$\tau = \sum_{i \in U} y_i \, .$$

Son estimateur de Horvitz-Thompson — que nous désignerons par $\hat{\tau}_{\text{GR}}$ pour nous rappeler que nous considérons le plan de sondage en grappes — se définit donc comme suit :

$$\hat{\tau}_{\text{GR}} = \sum_{i \in S} \frac{y_i}{p_i}$$

où $p_i = \text{P}(i \in S)$ est la probabilité d'inclusion de l'individu i dans l'échantillon S tiré dans la population U.

Puisque l'échantillon S est constitué de l'ensemble des individus des grappes qui font partie de l'échantillon de grappes S_{GR}, on peut encore reformuler l'estimateur de τ de la manière suivante :

$$\hat{\tau}_{\text{GR}} = \sum_{g \in S_{\text{GR}}} \sum_{i \in U_g} \frac{y_i}{p_i}.$$

Mais que valent les probabilités d'inclusion p_i ? Pour qu'un individu se retrouve dans l'échantillon final S, il faut et il suffit que la grappe à laquelle il appartient fasse partie des m grappes prélevées dans la population des grappes. En d'autres termes, si l'individu i appartient à la grappe n° g, sa probabilité d'inclusion dans l'échantillon S est égale à la probabilité d'inclusion de la grappe n° g dans l'échantillon de grappes S_{GR} : pour tout $i \in U_g$,

$$p_i = \text{P}(i \in S) = \text{P}(g \in S_{\text{GR}}).$$

Cette dernière probabilité d'inclusion dépend du plan de sondage appliqué dans la population U_{GR} des grappes. Nous allons nous intéresser à deux cas particuliers : celui où l'échantillon de grappes est prélevé par tirage PESR, et celui où il est prélevé par tirage à probabilités inégales, proportionnelles aux tailles des grappes.

De manière générale, si l'on remplace dans l'expression de l'estimateur de τ la probabilité d'inclusion de l'individu i par la probabilité d'inclusion de la grappe à laquelle il appartient, on obtient :

$$\begin{aligned}
\hat{\tau}_{\text{GR}} &= \sum_{g \in S_{\text{GR}}} \sum_{i \in U_g} \frac{y_i}{p_i} = \sum_{g \in S_{\text{GR}}} \sum_{i \in U_g} \frac{y_i}{\text{P}(g \in S_{\text{GR}})} \\
&= \sum_{g \in S_{\text{GR}}} \frac{1}{\text{P}(g \in S_{\text{GR}})} \sum_{i \in U_g} y_i \\
&= \sum_{g \in S_{\text{GR}}} \frac{\tau_g}{\text{P}(g \in S_{\text{GR}})}.
\end{aligned}$$

Ainsi, $\hat{\tau}_{\text{GR}}$ se reformule comme la somme, sur toutes les grappes g sélectionnées, des totaux τ_g de ces grappes, divisés par les probabilités d'inclusion de celles-ci. On reconnaît là l'expression de l'estimateur de Horvitz-Thompson du paramètre τ vu cette fois comme le total de la variable $\mathcal{T}$ dans la population U_{GR} des grappes (rappelez-vous : cette variable $\mathcal{T}$ associe à chaque grappe U_g son total τ_g).

Attention ! Ne soyez pas étonnés de voir apparaître les paramètres τ_g dans cette dernière expression de l'estimateur de τ. Ces totaux, relatifs aux grappes sélectionnées lors du sondage, ont des valeurs que l'on peut déterminer de manière exacte puisque, une fois qu'une grappe est sélectionnée, on relève les valeurs de la variable d'intérêt $\mathcal{Y}$ auprès de l'ensemble des individus de cette grappe.

5.2.4 Tirage PESR des grappes

a) Les estimateurs de τ, μ et μ_τ

Considérons à présent la situation particulière où l'on prélève **par tirage PESR** m grappes parmi les M grappes qui partitionnent la population. Dans ce cas, chaque grappe a la même probabilité de se retrouver dans l'échantillon S_{GR}, égale à m/M, le taux de sondage appliqué dans l'ensemble U_{GR}. Par conséquent, tout individu i de la population U se voit attribuer une probabilité d'inclusion p_i égale à ce rapport de m sur M : si $i \in U_g$,

$$p_i = \mathrm{P}(i \in S) = \mathrm{P}(g \in S_{\text{GR}}) = m/M.$$

L'estimateur du total-population τ s'écrit alors :

$$\hat{\tau}_{\text{GR}} = \sum_{i \in S} \frac{y_i}{p_i} = \frac{M}{m} \sum_{i \in S} y_i$$

ou encore, de manière équivalente :

$$\hat{\tau}_{\text{GR}} = \sum_{g \in S_{\text{GR}}} \frac{\tau_g}{\mathrm{P}(g \in S_{\text{GR}})} = \frac{M}{m} \sum_{g \in S_{\text{GR}}} \tau_g.$$

L'estimateur de μ n'est autre que celui de τ divisé par N :

$$\hat{\mu}_{\text{GR}} = \frac{\hat{\tau}_{\text{GR}}}{N} = \frac{M}{mN} \sum_{i \in S} y_i.$$

Quant à l'estimateur du total-moyen μ_τ, il est égal à l'estimateur de τ divisé par M et coïncide donc avec la moyenne arithmétique des totaux τ_g des grappes qui ont été sélectionnées :

$$\hat{\mu}_{\tau;\text{GR}} = \frac{\hat{\tau}_{\text{GR}}}{M} = \frac{1}{m} \sum_{g \in S_{\text{GR}}} \tau_g \, .$$

Exercice 5.1

La population étudiée dans le cadre de cet exercice est constituée de 300 exploitations agricoles d'une certaine région rurale. Cette dernière est partitionnée en 60 zones géographiques bien délimitées, contenant toutes entre 1 et 8 exploitations agricoles.

On désire estimer :

- la production de céréales *totale* (en tonnes) des 300 exploitations agricoles de la population pour l'année passée ;
- la production de céréales *moyenne*, par exploitation agricole, dans la population étudiée ;
- la production de céréales *totale moyenne*, par zone géographique, dans la population étudiée.

Pour ce faire, on décide de procéder au prélèvement, par tirage PESR, de 10 zones géographiques, puis d'interroger les propriétaires de toutes les exploitations agricoles des zones sélectionnées afin de déterminer leurs productions de céréales au cours de l'année passée. Les données ainsi récoltées sont présentées dans la feuille « GR-PESR » du fichier Grappes_ex.xlsx téléchargeable sur l'UV.

a) Que vaut la taille N de la population étudiée ?

b) Quel est le nombre M de « grappes » partitionnant la population étudiée ?

c) Quelle est la taille moyenne $\overline{N}$ des grappes de la population étudiée ?

d) Quel est le nombre m de grappes sélectionnées au cours du sondage ?

e) Quel est le nombre n_s d'exploitations agricoles sélectionnées au cours du sondage ?

f) Quelle estimation obtenez-vous pour la production de céréales *totale* (en tonnes) des 300 exploitations agricoles de la population ?

g) Quelle estimation obtenez-vous pour la production de céréales *moyenne*, par exploitation agricole, dans la population étudiée ?

h) Quelle estimation obtenez-vous pour la production de céréales *totale moyenne*, par zone géographique, dans la population étudiée ?

b) La taille de l'échantillon S

Avant d'étudier la précision des estimateurs que nous venons de définir, penchons-nous un tout petit moment sur la taille n_S de l'échantillon S des individus. Cette taille est *aléatoire*.

On peut cependant facilement déterminer $\mathrm{E}(n_S)$:

$$\mathrm{E}(n_S) = \mathrm{E}\left(\sum_{g\in S_{\mathrm{GR}}} N_g\right) = \mathrm{E}\left(\sum_{g\in U_{\mathrm{GR}}} N_g I_g\right)$$

où

$$I_g = \begin{cases} 1 & \text{si } g \in S_{\mathrm{GR}} \\ 0 & \text{sinon.} \end{cases}$$

Dès lors,

$$\mathrm{E}(n_S) = \sum_{g\in U_{\mathrm{GR}}} N_g \mathrm{E}(I_g) = \sum_{g\in U_{\mathrm{GR}}} N_g \mathrm{P}(g \in S_{\mathrm{GR}})$$
$$= \sum_{g\in U_{\mathrm{GR}}} N_g \frac{m}{M} = m\frac{N}{M} = m\overline{N}\,,$$

où $\overline{N} = N/M$ est la taille *moyenne* des grappes de la population.

Autrement dit, la *taille moyenne* des échantillons d'individus qu'il est possible d'obtenir avec la procédure de sondage que nous sommes en train d'étudier, est égale au nombre m de grappes prélevées, multiplié par $\overline{N}$, la taille *moyenne* des grappes de la population.

c) La variance des estimateurs

Les estimateurs que nous venons de construire sont nécessairement *sans biais*, puisqu'il s'agit d'estimateurs de Horvitz-Thompson.

Que peut-on dire de leur précision, autrement dit de leur variance ? Malgré qu'il s'agisse d'estimateurs de Horvitz-Thompson, on obtient l'expression générale de leur variance très facilement, dès le moment où l'on se rappelle que ce sondage en grappes n'est autre

qu'un sondage aléatoire simple sans remise de m grappes dans la population U_{GR} constituée de M grappes.

Rappelons-nous les résultats que nous avons établis au chapitre 2 de ce cours. Dans le cadre d'un sondage aléatoire simple, la variance de l'estimateur du total-population est égale à la taille de la population au carré — soit ici M^2 —, multipliée par (1 – le taux de sondage appliqué) — soit ici $(1 - m/M)$ —, fois la variance corrigée de la variable d'intérêt dans la population — soit ici la variance corrigée de la variable $\mathcal{T}$ dans la population U_{GR}, notée $\sigma^2_{\tau;\text{corr}}$ —, divisée par la taille de l'échantillon — soit ici m :

$$\text{V}(\hat{\tau}_{\text{GR}}) = M^2 \left(1 - \frac{m}{M}\right) \frac{\sigma^2_{\tau;\text{corr}}}{m},$$

où

$$\sigma^2_{\tau;\text{corr}} = \frac{1}{M-1} \sum_{g=1}^{M} (\tau_g - \mu_\tau)^2.$$

La variance $\sigma^2_{\tau;\text{corr}}$ est tout simplement la variance, corrigée, des totaux τ_g de la variable d'intérêt $\mathcal{Y}$ dans les différentes grappes. Si les grappes ont des totaux très similaires, cette variance $\sigma^2_{\tau;\text{corr}}$ est faible ; si, au contraire, les grappes de la population ont des totaux qui varient assez fortement d'une grappe à l'autre, cette variance $\sigma^2_{\tau;\text{corr}}$ est élevée.

Les variances de $\hat{\mu}_{\text{GR}}$ et de $\hat{\mu}_{\tau;\text{GR}}$ se déduisent directement de celle de $\hat{\tau}_{\text{GR}}$:

$$V(\hat{\mu}_{\text{GR}}) = \text{V}\left(\frac{\hat{\tau}_{\text{GR}}}{N}\right) = \frac{1}{N^2}\text{V}(\hat{\tau}_{\text{GR}}) = \frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{\sigma^2_{\tau;\text{corr}}}{m} ;$$

$$V(\hat{\mu}_{\tau;\text{GR}}) = \text{V}\left(\frac{\hat{\tau}_{\text{GR}}}{M}\right) = \frac{1}{M^2}\text{V}(\hat{\tau}_{\text{GR}}) = \left(1 - \frac{m}{M}\right)\frac{\sigma^2_{\tau;\text{corr}}}{m}.$$

Que pouvons-nous conclure de ces expressions ? Premièrement, que le sondage en grappes, avec tirage PESR des grappes, est d'autant plus efficace que l'on sélectionne un grand nombre de grappes parmi les M grappes existantes. Et deuxièmement, que la précision du sondage sera d'autant meilleure que la valeur de $\sigma^2_{\tau;\text{corr}}$ est faible, c'est-à-dire que les grappes donnent lieu à des totaux τ_g fort similaires.

Enfin, pour estimer sans biais les variances de $\hat{\tau}_{\text{GR}}$, de $\hat{\mu}_{\text{GR}}$ et de $\hat{\mu}_{\tau;\text{GR}}$, il suffit de remplacer la variance $\sigma^2_{\tau;\text{corr}}$ dont la valeur est inconnue par la variance corrigée des totaux τ_g des grappes qui ont été sélectionnées :

$$\widehat{\text{V}}(\hat{\tau}_{\text{GR}}) = M^2 \left(1 - \frac{m}{M}\right) \frac{s^2_{\tau;\text{corr}}}{m}$$

où

$$s^2_{\tau;\text{corr}} = \frac{1}{m-1} \sum_{g \in S_{\text{GR}}} (\tau_g - \hat{\mu}_{\tau;\text{GR}})^2.$$

Exercice 5.2 (suite de l'exercice 5.1)

Reprenez l'exercice 5.1 là où vous l'avez laissé.

i) Donnez une estimation de la variance de l'estimateur de la production de céréales *totale* des 300 exploitations agricoles de la population.

j) Donnez une estimation de la variance de l'estimateur de la production de céréales *moyenne*, par exploitation agricole, dans la population étudiée.

k) Donnez une estimation de la variance de l'estimateur de la production de céréales *totale moyenne*, par zone géographique, dans la population étudiée.

d) Le cas particulier des grappes de tailles égales (GRTE)

Dans certains cas, comme celui du contrôle par lots évoqué dans la Section 5.1, les grappes ont toutes la même taille : elles contiennent toutes le même nombre d'unités de la population U. Désignons par N_0 cette taille commune.

La situation s'avère alors plus simple à étudier du point de vue statistique, notamment grâce au fait que la taille de l'échantillon S d'unités statistiques est alors *fixée*, et non plus aléatoire. En effet, si l'on prélève m grappes qui ont toutes la même taille N_0, l'échantillon S contient nécessairement mN_0 unités. Désignons par n cette taille fixe de l'échantillon S.

Dans ce cas, l'estimateur de la moyenne-population μ coïncide tout simplement avec la moyenne $\bar{y}$ des valeurs que prend la variable y chez les individus de l'échantillon S. En effet :

$$\hat{\mu}_{\text{GRTE}} = \frac{\hat{\tau}_{\text{GR}}}{N} = \frac{M}{mN}\sum_{i\in S} y_i \,.$$

Or, la taille N de la population U est égale ici au nombre M de grappes multiplié par la taille N_0 de ces grappes. Nous avons donc :

$$\hat{\mu}_{\text{GRTE}} = \frac{M}{mMN_0}\sum_{i\in S} y_i = \frac{1}{mN_0}\sum_{i\in S} y_i$$

ou encore, puisque le produit mN_0 correspond à la taille fixe n de l'échantillon S :

$$\hat{\mu}_{\text{GRTE}} = \frac{1}{n}\sum_{i\in S} y_i = \bar{y}.$$

Voilà déjà une simplification bien agréable !

Par ailleurs, le fait que les grappes soient de même taille va également nous permettre de réécrire l'expression de la variance de $\hat{\mu}_{\text{GR}}$ sous une autre forme et d'analyser ainsi plus finement les caractéristiques des grappes qui jouent sur l'efficacité du sondage en grappes.

Rappelons-nous tout d'abord la formule de décomposition de la variance que nous avons déjà rencontrée dans le chapitre 3, lorsque nous étudiions le sondage stratifié proportionnel. Cette décomposition de la variance σ^2 de la variable y dans la population U est d'application dès que cette population U est partitionnée en un certain nombre de sous-ensembles, qu'il s'agisse de strates ou de grappes. On peut donc exprimer ici σ^2 comme la somme de deux termes. Le premier terme, appelé ici la *variance intra-grappes* et désigné par σ^2_{intra}, mesure globalement la dispersion des valeurs de la variable y à l'intérieur même des grappes ; il est une moyenne pondérée des variances σ_g^2 de y au sein des grappes :

$$\sigma^2_{\text{intra}} = \sum_{g=1}^{M} \frac{N_g}{N} \sigma_g^2 \,.$$

Le second terme, appelé la *variance inter-grappes* et désigné par σ^2_{inter}, mesure la dispersion des moyennes μ_g de $\mathcal{Y}$ dans les différentes grappes autour de la moyenne globale μ de $\mathcal{Y}$ dans la population U :

$$\sigma^2_{\text{inter}} = \sum_{g=1}^{M} \frac{N_g}{N} (\mu_g - \mu)^2.$$

Repartons à présent de l'expression générale de la variance de l'estimateur de μ dans le cas du sondage en grappes et voyons comment nous pouvons la réécrire dans le cas particulier du sondage GRTE :

$$\mathrm{V}(\hat{\mu}_{\text{GRTE}}) = \frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{\sigma^2_{\tau;\text{corr}}}{m}$$

avec

$$\begin{aligned}
\sigma^2_{\tau;\text{corr}} &= \frac{1}{M-1}\sum_{g=1}^{M}(\tau_g - \mu_\tau)^2 \\
&= \frac{1}{M-1}\sum_{g=1}^{M}\left(N_0\mu_g - \frac{\tau}{M}\right)^2 = \frac{1}{M-1}\sum_{g=1}^{M}\left(N_0\mu_g - \frac{N\mu}{M}\right)^2 \\
&= \frac{1}{M-1}\sum_{g=1}^{M}\left(N_0\mu_g - \frac{MN_0\mu}{M}\right)^2 = \frac{N_0^2}{M-1}\sum_{g=1}^{M}(\mu_g - \mu)^2 \\
&= \frac{N_0 N}{M-1}\sum_{g=1}^{M}\frac{N_0}{N}(\mu_g - \mu)^2 = \frac{N_0 N}{M-1}\sigma^2_{\text{inter}} = \frac{N_0^2 M}{M-1}\sigma^2_{\text{inter}}
\end{aligned}$$

et

$$\frac{m}{M} = \frac{mN_0}{MN_0} = \frac{n}{N} = f.$$

On obtient ainsi :

$$\begin{aligned}
\mathrm{V}(\hat{\mu}_{\text{GRTE}}) &= \frac{M^2}{N^2}(1-f)\frac{N_0^2 M}{M-1}\frac{\sigma^2_{\text{inter}}}{m} = \frac{M^2}{N_0^2 M^2}(1-f)\frac{N_0^2 M}{M-1}\frac{\sigma^2_{\text{inter}}}{m} \\
&= (1-f)\frac{M}{M-1}\frac{\sigma^2_{\text{inter}}}{m}
\end{aligned}$$

où f est le taux de sondage global n/N dans la population U.

Que faut-il donc pour que la variance de $\hat{\mu}_{\text{GRTE}}$ soit petite, autrement dit pour que le sondage en grappes soit efficace ? Il faut notamment que la variance inter-grappes soit faible, c'est-à-dire que les grappes aient des moyennes μ_g aussi semblables que possible.

Ceci, en réalité, nous le savions déjà ! En effet, nous avons vu que le sondage en grappes, avec tirage PESR des grappes, donnait lieu à une bonne précision si les grappes avaient des totaux τ_g aussi peu dispersés que possible. Lorsque les grappes ont toutes la même

taille N_0, cela revient au même de dire que les grappes doivent avoir des moyennes μ_g aussi peu différentes que possible (puisque $\mu_g = \tau_g / N_0$ pour tout $g = 1, \dots, M$).

Mais, puisque $\sigma^2 = \sigma^2_{\text{intra}} + \sigma^2_{\text{inter}}$, demander que la variance *inter*-grappes soit faible, revient à dire qu'il faut considérer des grappes donnant lieu à une variance *intra*-grappes constituant une large part de σ^2, c'est-à-dire des grappes au sein desquelles on retrouve une large part de la dispersion de la variable y. Autrement dit, il faut que la dispersion totale de y dans la population U soit essentiellement expliquée par l'hétérogénéité des individus à l'intérieur des grappes.

On le voit clairement ici, un découpage de la population en grappes ne doit pas du tout présenter les mêmes caractéristiques qu'un découpage en strates. Pour que le sondage *stratifié* soit efficace, il nous fallait des strates au sein desquelles il y avait peu de dispersion des valeurs de y, et pour lesquelles les moyennes de y étaient bien différentes les unes des autres. Pour que le sondage *en grappes* soit efficace, il nous faut au contraire des grappes au sein desquelles on retrouve l'essentiel de la dispersion de y, et donnant lieu à des moyennes de y aussi similaires que possible. Ainsi, si les strates doivent regrouper des individus aussi semblables que possible, les grappes doivent au contraire contenir des individus plutôt bien différents les uns des autres. En d'autres termes, de bonnes strates sont de mauvaises grappes et de bonnes grappes sont de mauvaises strates !

Nous pouvons enfin nous intéresser à une autre question importante. Supposons que la population U soit partitionnée en M grappes de tailles égales à N_0 : nous pouvons y prélever un échantillon S soit en prélevant par tirage PESR m grappes parmi les M grappes existantes — l'échantillon S comptera alors $n = mN_0$ unités —, soit en prélevant par sondage aléatoire simple n unités directement dans la population U, sans tenir compte du découpage en grappes de U. Quand le sondage en grappes va-t-il être plus efficace que le sondage aléatoire simple ?

Pour répondre à cette question, déterminons l'*effet* du sondage en grappes par rapport au sondage aléatoire simple. Autrement dit, déterminons le rapport de la variance de l'estimateur de μ sous le plan de sondage en grappes de tailles égales sur la variance de l'estimateur de μ sous le plan de sondage aléatoire simple.

Sous le plan de sondage aléatoire simple, l'estimateur de μ est $\bar{y}$ et sa variance est égale à :

$$\mathrm{V}_{\text{PESR}}(\bar{y}) = (1 - f)\frac{\sigma^2_{\text{corr}}}{n} = (1 - f)\frac{\sigma^2_{\text{corr}}}{mN_0}.$$

Si la taille N de la population U est grande, σ^2_{corr} est à peu de chose près égale à σ^2, ce qui nous permet d'écrire :

$$\mathrm{V}_{\text{PESR}}(\bar{y}) \simeq (1 - f)\frac{\sigma^2}{mN_0}.$$

Sous le plan de sondage en grappes de tailles égales, l'estimateur de μ est aussi $\bar{y}$ et sa variance est égale à :

$$\mathrm{V}_{\mathrm{GRTE}}(\bar{y}) = (1-f)\frac{M}{M-1}\frac{\sigma^2_{\mathrm{inter}}}{m}.$$

Si le nombre M de grappes est élevé, alors $M/(M-1) \simeq 1$ et, dès lors :

$$\mathrm{V}_{\mathrm{GRTE}}(\bar{y}) \simeq (1-f)\frac{\sigma^2_{\mathrm{inter}}}{m}.$$

Par conséquent, si N et M sont « grands » :

$$\frac{\mathrm{V}_{\mathrm{GRTE}}(\bar{y})}{\mathrm{V}_{\mathrm{PESR}}(\bar{y})} \simeq \frac{(1-f)\frac{\sigma^2_{\mathrm{inter}}}{m}}{(1-f)\frac{\sigma^2}{mN_0}} = N_0\frac{\sigma^2_{\mathrm{inter}}}{\sigma^2}.$$

Le sondage en grappes de tailles égales est plus efficace que le sondage aléatoire simple si ce rapport $\mathrm{V}_{\mathrm{GRTE}}(\bar{y})/\mathrm{V}_{\mathrm{PESR}}(\bar{y})$ que nous venons de déterminer est inférieur à 1, c'est-à-dire si

$$\frac{\sigma^2_{\mathrm{inter}}}{\sigma^2} < \frac{1}{N_0}.$$

Cette condition est d'autant plus facilement vérifiée que $\sigma^2_{\mathrm{inter}}$ est faible et que l'inverse de N_0 est grand, c'est-à-dire que la taille N_0 des grappes est petite.

D'un point de vue pratique, si la population U est partitionnée en grappes de tailles égales, le sondage en grappes y sera plus efficace que le sondage aléatoire simple si la population est constituée d'un *grand nombre* de grappes de *petite taille* N_0 et de moyennes μ_g fort similaires, de telle sorte que la variance de y dans la population soit essentiellement expliquée par l'hétérogénéité des individus au sein même des grappes.

Exercice 5.3

La population U considérée est constituée de 100 ménages comptant chacun 5 individus. On désire y étudier la variable y associant à chaque individu de U le montant dépensé au cours des trois derniers mois pour l'achat de médicaments qui lui sont destinés : on souhaite en particulier estimer le montant *total* des achats de médicaments pour l'ensemble de la population, le montant *moyen par individu* de ces achats, ainsi que le montant *total moyen* de ces achats, *par ménage*.

Pour estimer ces trois paramètres, on peut faire appel à un sondage aléatoire simple ou à un sondage en grappes (les ménages de la population sont ici des grappes de tailles égales).

La feuille « GRTE » du fichier Grappes_ex.xlsx (téléchargeable sur l'UV) vous présente l'ensemble de la population U et son découpage en 100 grappes (ménages) de taille 5. Les cellules C2 à C501 contiennent les valeurs y_i de la variable d'intérêt y pour chaque individu i de la population U ; les cellules C502, C503 et C504 contiennent les valeurs des paramètres τ, μ et μ_τ de la population, respectivement ; la colonne D contient les totaux τ_g des grappes (ménages) U_g qui partitionnent la population ; la colonne E contient les moyennes μ_g des grappes (ménages) U_g de la population ; la colonne F contient les quantités $(\mu_g - \mu)^2$ associées aux grappes (ménages) U_g de la population ; la colonne G contient les variances σ_g^2 des grappes (ménages) U_g de la population.

PARTIE 1 : SONDAGE ALEATOIRE SIMPLE DE 100 INDIVIDUS

Supposons que l'on choisisse de prélever 100 individus de la population par tirage aléatoire PESR.

(i) Que vaut la variance de l'estimateur du montant *moyen*, par individu de la population, des achats en médicaments ?

PARTIE 2 : SONDAGE PESR DE 20 GRAPPES (MENAGES)

Supposons que l'on choisisse de prélever 20 ménages de la population par tirage aléatoire PESR et de relever ensuite les valeurs de la variable d'intérêt y pour l'ensemble des individus de ces ménages.

(i) Que vaut la variance σ^2 de la variable d'intérêt y dans la population U ?

(ii) Que vaut la variance σ^2_{intra} (la variance intra-grappes) ?

(iii) Que vaut la variance σ^2_{inter} (la variance inter-grappes) ?

(iv) Sur la base des résultats que vous venez d'établir, on peut dire que (cochez la proposition correcte) :

- ○ La dispersion de la variable y dans la population est pour la plus large part expliquée par l'hétérogénéité des montants d'achats de médicaments au sein des ménages.
- ○ La dispersion de la variable y dans la population est principalement expliquée par la variabilité des montants moyens d'achats de médicaments d'un ménage à l'autre.
- ○ La dispersion de la variable y dans la population est expliquée dans la même mesure par l'hétérogénéité des montants d'achats de médicaments au sein des ménages que par la variabilité des montants moyens d'achats de médicaments d'un ménage à l'autre.

(v) Sur la base des résultats que vous venez d'établir, on peut affirmer que (cochez la réponse correcte) :

- ○ Le *sondage en grappes* avec tirage PESR de 20 ménages de la population permettra d'estimer les paramètres qui nous intéressent avec une meilleure précision que le *sondage aléatoire simple* avec tirage PESR de 100 individus de la population.
- ○ Le *sondage en grappes* avec tirage PESR de 20 ménages de la population permettra d'estimer les paramètres qui nous intéressent avec une moins bonne précision que le *sondage aléatoire simple* avec tirage PESR de 100 individus de la population.
- ○ Le *sondage en grappes* avec tirage PESR de 20 ménages de la population permettra d'estimer les paramètres qui nous intéressent avec exactement la même précision que le *sondage aléatoire simple* avec tirage PESR de 100 individus de la population.

(vi) Dans le cadre du *sondage en grappes* avec tirage PESR de 20 ménages, que vaut la variance de l'estimateur du montant *moyen*, par individu de la population, des

achats en médicaments ?

(vii) Dans le cadre du *sondage en grappes* avec tirage PESR de 20 ménages, que vaut la variance de l'estimateur du montant *total* des achats en médicaments effectués pour l'ensemble des individus de la population ?

(viii) Dans le cadre du *sondage en grappes* avec tirage PESR de 20 ménages, que vaut la variance de l'estimateur du montant *total moyen*, par ménage de la population, des achats en médicaments ?

e) En conclusion : les conditions favorables pour effectuer un tirage PESR de grappes

Essayons de refaire le point sur les principales leçons à tirer de notre étude du sondage en grappes, avec tirage PESR des grappes.

1. C'est dans le cas où les grappes ont toutes la même taille que le tirage PESR des grappes donnera lieu à la meilleure précision. On peut en effet montrer que, si les grappes sont de tailles *inégales*, la variance de $\hat{\mu}_{\text{GR}}$ s'accroît d'un facteur positif représentant la dispersion des tailles N_g. On a donc intérêt, pour que le tirage PESR des grappes s'avère efficace, à ce que la population soit partitionnée en un grand nombre de petites grappes, aux tailles aussi semblables que possible.

2. Par ailleurs, même lorsque les grappes sont de même taille, le tirage PESR des grappes ne s'avère plus efficace que le sondage aléatoire simple dans U que si le rapport de la variance inter-grappes sur la variance globale de la variable y dans la population U est suffisamment faible. Cette condition n'est satisfaite que si les grappes donnent lieu à des moyennes μ_g très similaires. Il s'agit en réalité d'une condition assez forte, qui n'est pas toujours respectée dans les situations concrètes. En revanche, cette condition peut être plus facilement respectée pour des sous-ensembles de grappes. Il est dès lors essentiel de réfléchir à l'opportunité de stratifier la population U_{GR} des grappes, c'est-à-dire de découper U_{GR} en strates en faisant en sorte que chaque strate regroupe des grappes qui soient les plus ressemblantes possibles, aussi bien au niveau de leurs tailles que de leurs moyennes.

3. Enfin — nous l'avons déjà souligné aussi — dire que les grappes doivent conduire à une variance inter-grappes aussi faible que possible revient à dire que l'essentiel de la variance de y dans la population doit être expliqué par l'hétérogénéité des valeurs de y à l'intérieur même des grappes. Cela signifie, en pratique, que chaque grappe doit idéalement pouvoir bien refléter toute la diversité des individus de la population tout entière. J'aurais envie de dire que chaque grappe doit pouvoir nous donner une image fidèle de l'hétérogénéité de la population dans son ensemble.

   Pour reprendre les exemples donnés dans la Section 5.1 pour illustrer le sondage en grappes, cela signifie que chaque caisse de fruits devrait être « représentative » de l'ensemble des caisses, que chaque médecin devrait avoir une clientèle semblable aux autres, que chaque quartier devrait être à l'image de l'ensemble des quartiers, que chaque classe d'une école devrait se montrer aussi hétérogène que l'ensemble de l'école.

Vous imaginez bien que, dans la pratique, cette condition pose un problème, justement parce que, très souvent, les individus ou unités statistiques qui forment une grappe ont tendance à se ressembler : ne dit-on pas « qui se ressemble s'assemble ! » ? C'est ainsi que les médecins ont des clientèles typées selon leur lieu d'exercice ou que chaque classe d'une école rassemble des élèves présentant de nombreuses similarités, par exemple. Ce phénomène porte le nom *d'effet de grappe*. Il contribue à réduire l'efficacité du sondage en grappes. La stratification de l'ensemble des grappes évoquée ci-dessus devrait avoir pour effet de réduire son impact négatif sur la précision du sondage. En regroupant dans une même strate des grappes qui se ressemblent, nous pouvons espérer que chaque grappe de cette strate nous donne une image plus fidèle de la diversité de l'ensemble des individus des grappes de cette strate.

5.2.5 Tirage à probabilités proportionnelles aux tailles des grappes

Nous venons de le voir, le sondage en grappes avec tirage PESR des grappes n'a de chance de se montrer efficace que si les grappes qui composent la population ont toutes des tailles très similaires. S'il s'avère que les grappes ont des tailles qui diffèrent assez fortement les unes des autres, on peut éventuellement penser à stratifier (découper) la population des grappes en différentes classes « de tailles » : on peut ainsi penser à définir, par exemple, une première strate qui regroupe les grappes de « petites » tailles, une deuxième strate qui regroupe les grappes de tailles « moyennes » et une troisième strate rassemblant les plus grandes grappes. Le prélèvement des grappes se fera alors par sondage stratifié. Cette solution s'avèrera efficace pour autant que les grappes d'une même strate aient non seulement des tailles similaires, mais aussi des moyennes μ_g fort semblables.

Une autre solution que l'on peut adopter lorsque les grappes ont des tailles fort différentes les unes des autres consiste à sélectionner m grappes par tirage à probabilités *inégales* sans remise, en attribuant à chaque grappe de la population une probabilité d'inclusion proportionnelle à sa taille. Au vu de ce que nous avons étudié dans le chapitre précédent sur le sondage PPS, on se doute qu'il s'agit là d'une manière d'assurer une bonne efficacité au sondage en grappes, du moins s'il est raisonnable de penser que le total τ_g associé à une grappe est approximativement proportionnel à sa taille N_g. Cette solution ne peut cependant être mise en œuvre que si la base de sondage listant l'ensemble des grappes de la population nous indique également quelle est la taille N_g de chaque grappe U_g.

a) Principales caractéristiques du plan de sondage

Voyons rapidement quelles sont les principales caractéristiques de ce plan de sondage particulier.

Que valent les probabilités d'inclusion des différentes grappes de la population ? Nous voulons que chaque grappe U_g ait une probabilité d'inclusion proportionnelle à sa taille N_g, autrement dit que, pour tout $g = 1, \dots, M$,

$$\mathrm{P}(g \in S_{\mathrm{GR}}) = cN_g$$

où c est une certaine constante.

On détermine la valeur de c en se rappelant que, dans tout échantillonnage aléatoire de taille fixe, la somme des probabilités d'inclusion attribuées aux unités statistiques de la population que l'on doit sonder est égale à la taille de l'échantillon à prélever. Dans notre contexte, il faut donc que la somme des probabilités d'inclusion des M grappes qui partitionnent la population soit égale à m :

$$\sum_{g=1}^{M} \mathrm{P}(g \in S_{\mathrm{GR}}) = m\,.$$

Cela revient à dire que :

$$\sum_{g=1}^{M} cN_g = m\,,$$

autrement dit que :

$$c\sum_{g=1}^{M} N_g = cN = m\,.$$

On obtient ainsi que $c = m/N$.
Il s'ensuit que, pour tout $g = 1, \ldots, M$:

$$\mathrm{P}(g \in S_{\mathrm{GR}}) = m\frac{N_g}{N}\,.$$

S'il existe l'une ou l'autre grappe U_g pour laquelle la quantité $m\frac{N_g}{N}$ est strictement supérieure à 1, on attribue automatiquement à ces grappes une probabilité d'inclusion égale à 1 et on recalcule les probabilités d'inclusion des autres grappes de la population en suivant la même procédure que celle décrite pour le sondage PPS dans le chapitre précédent.

Notez que, dans ce sondage en grappes particulier, la taille n_S de l'échantillon final S des individus est toujours aléatoire. On peut cependant montrer que la *taille moyenne* des échantillons S qu'il est possible d'obtenir avec cette procédure de prélèvement PISR de m grappes, est supérieure ou égale à la taille moyenne des échantillons S que l'on peut obtenir lorsqu'on effectue un tirage PESR de m grappes. En effet, la taille de l'échantillon S des individus peut s'écrire sous la forme :

$$n_S = \sum_{g \in S_{\mathrm{GR}}} N_g = \sum_{g \in U_{\mathrm{GR}}} N_g I_g$$

où I_g est la variable aléatoire indiquant si la grappe n° g fait partie ($I_g = 1$) ou non ($I_g = 0$) de l'échantillon de grappes S_{GR} sélectionné. Dès lors :

$$\mathrm{E}(n_S) = \sum_{g \in U_{\mathrm{GR}}} N_g \mathrm{E}(I_g) = \sum_{g \in U_{\mathrm{GR}}} N_g \mathrm{P}(g \in S_{\mathrm{GR}}).$$

Puisque, dans le cas du tirage PPS de m grappes, $\mathrm{P}(g \in S_{\mathrm{GR}}) = m\,N_g/N$, nous obtenons :

$$\mathrm{E}(n_S) = \sum_{g \in U_{\mathrm{GR}}} N_g m\frac{N_g}{N} = \frac{m}{N}\sum_{g \in U_{\mathrm{GR}}} N_g^2 = \frac{m}{M\overline{N}}\sum_{g \in U_{\mathrm{GR}}} N_g^2.$$

Or, la variance des tailles des grappes de la population est nécessairement positive. Par conséquent :

$$\frac{1}{M}\sum_{g \in U_{\mathrm{GR}}} \left(N_g - \overline{N}\right)^2 = \frac{1}{M}\sum_{g \in U_{\mathrm{GR}}} N_g^2 - \overline{N}^2 \geq 0$$

et donc

$$\sum_{g\in U_{\mathrm{GR}}} N_g^2 \geq M\bar{N}^2.$$

Il s'ensuit que :

$$\mathrm{E}(n_S) = \frac{m}{M\bar{N}} \sum_{g\in U_{\mathrm{GR}}} N_g^2 \geq \frac{m}{M\bar{N}} M\bar{N}^2 = m\bar{N}\,,$$

ce que nous voulions établir.

b) Définition des estimateurs

Puisque nous connaissons à présent les probabilités d'inclusion des grappes de la population, on obtient directement l'expression de l'estimateur de Horvitz-Thompson du total τ. Si $m\frac{N_g}{N}$ est inférieur ou égal à 1 pour toutes les grappes de la population, on a :

$$\hat{\tau}_{\mathrm{GR}} = \sum_{g\in S_{\mathrm{GR}}} \frac{\tau_g}{\mathrm{P}(g\in S_{\mathrm{GR}})} = \sum_{g\in S_{\mathrm{GR}}} \frac{\tau_g}{m\frac{N_g}{N}} = \frac{N}{m}\sum_{g\in S_{\mathrm{GR}}} \frac{\tau_g}{N_g} = \frac{N}{m}\sum_{g\in S_{\mathrm{GR}}} \mu_g.$$

Dès lors :

$$\hat{\mu}_{\mathrm{GR}} = \frac{\hat{\tau}_{\mathrm{GR}}}{N} = \frac{1}{m}\sum_{g\in S_{\mathrm{GR}}} \mu_g\,;$$

l'estimateur de la moyenne de la variable y dans la population U n'est autre que la moyenne arithmétique des moyennes de y dans les grappes qui ont été sélectionnées. Quant à l'estimateur du total moyen μ_τ, il s'obtient en divisant $\hat{\tau}_{\mathrm{GR}}$ par M.

Exercice 5.4

On s'intéresse à une certaine population d'élèves de 1re année du secondaire qui ont présenté, en début d'année scolaire, un test standardisé (sur 20 points) destiné à évaluer leur niveau en mathématiques. La variable d'intérêt y que l'on désire étudier est la note obtenue à ce test.

La population considérée est en réalité constituée de 100 classes comptant entre 8 et 30 élèves. La feuille « GR-PPS(1) » du fichier Grappes_ex.xlsx (téléchargeable sur l'UV) vous indique la taille de chacune des 100 classes.

Afin d'estimer la note moyenne obtenue au test standardisé, on décide de procéder à un sondage en grappes en prélevant 20 classes par tirage à probabilités proportionnelles aux tailles des classes (tirage PPS) .

a) Déterminez les probabilités d'inclusion des 100 classes de la population.

Réalisez ensuite le prélèvement de 20 classes en appliquant la procédure de tirage systématique sur le fichier des probabilités cumulées avec le nombre ALEA égal à 0,2075. [!! Uniquement pour les étudiants d'ECON et INGE !!]

b) A quelle estimation de la note moyenne obtenue au test vous conduit l'échantillon prélevé ?
Attention ! Vous trouverez la note moyenne de chaque classe dans la feuille « GR-PPS(2) » du fichier Grappes_ex.xlsx.

c) Variance des estimateurs

Comme bien souvent dans le cas d'un sondage PISR, l'expression de la variance de $\hat{\tau}_{\mathrm{GR}}$ ou $\hat{\mu}_{\mathrm{GR}}$ lorsqu'on effectue un tirage PPS des grappes est très complexe. On peut cependant avoir recours au stratagème suivant pour obtenir une expression approchée fort simple de cette variance : on considère que les grappes sont sélectionnées à probabilités inégales (proportionnelles à leurs tailles) **avec remise**. On peut montrer que cette approximation fournit toujours un résultat numérique de variance supérieure à la variance que l'on aurait si on utilisait les formules exactes du tirage sans remise.

On obtient ainsi que

$$\mathrm{V}(\hat{\mu}_{\mathrm{GR}}) = \mathrm{V}\left(\frac{1}{m}\sum_{g\in S_{\mathrm{GR}}}\mu_g\right) \simeq \frac{\sigma^2_{\mathrm{inter}}}{m}$$

où $\sigma^2_{\mathrm{inter}}$ est la variance des moyennes des grappes de la population :

$$\sigma^2_{\mathrm{inter}} = \sum_{g=1}^{M}\frac{N_g}{N}(\mu_g - \mu)^2.$$

On montre également que l'on peut prendre

$$\widehat{\mathrm{V}}(\hat{\mu}_{\mathrm{GR}}) = \frac{1}{m(m-1)}\sum_{g\in S_{\mathrm{GR}}}(\mu_g - \hat{\mu}_{\mathrm{GR}})^2.$$

5.3 Le sondage à deux degrés

Passons à présent à l'étude du sondage à deux degrés. Si vous en comprenez bien le principe et les caractéristiques, vous n'aurez aucune peine à généraliser ensuite les résultats que nous allons voir pour le sondage à deux degrés au cas du sondage à trois, quatre ou même davantage de degrés.

5.3.1 Caractéristiques générales de la population et de l'échantillon

Comme auparavant, la population-cible U est constituée de N individus ou unités statistiques. Elle est par ailleurs partitionnée en M unités primaires : $U_1, U_2, \cdots, U_M$. Chaque unité primaire U_g compte N_g individus. La taille N de la population U est la somme des tailles N_g des unités primaires qui la composent :

$$N = \sum_{g=1}^{M} N_g \, .$$

L'échantillon d'individus S est obtenu en procédant à un sondage aléatoire en deux étapes. On tire d'abord un échantillon aléatoire de m unités dans l'ensemble (ou population) $U_{\text{PRIM}} = \{U_1, \ldots, U_M\} = \{1, \ldots, M\}$ des unités primaires ; nous pouvons désigner cet échantillon d'unités primaires par S_{PRIM}. Dans un second temps, on tire, dans chaque unité primaire U_g sélectionnée au premier degré du sondage, un échantillon aléatoire S_g de n_g individus. L'échantillon S est construit en réunissant les différents sous-échantillons S_g prélevés au deuxième degré du sondage :

$$S = \bigcup_{g \in S_{\text{PRIM}}} S_g \, .$$

La taille de l'échantillon final S correspond à la somme des tailles de ces différents sous-échantillons S_g :

$$n_S = \sum_{g \in S_{\text{PRIM}}} n_g \, .$$

Le plus souvent, cette taille n_S est aléatoire ; c'est le cas, par exemple, si on décide d'appliquer le même taux de sondage dans chacune des unités primaires préalablement sélectionnées.

Notez encore que le tirage des individus s'effectue de manière indépendante d'une unité primaire à l'autre.

5.3.2 Estimateur de Horvitz-Thompson du total-population τ

Comment estimer le paramètre τ ? Rappelons une nouvelle fois que τ est le total de la variable d'intérêt y dans la population U. Mais, comme dans le contexte du sondage en grappes, τ peut aussi se voir comme le total de la variable $\mathcal{T}$ dans la population U_{PRIM},

autrement dit comme la somme des totaux de y dans toutes les unités primaires qui partitionnent la population :

$$\tau = \sum_{i \in U} y_i = \sum_{g \in U_{\text{PRIM}}} \tau_g \, .$$

Adoptons la première vision du paramètre τ. Puisque τ est la somme des valeurs que prend la variable y sur tous les individus de U, son estimateur de Horvitz-Thompson se définit comme la somme des valeurs observées pour la variable y sur les individus i de l'échantillon S, chaque valeur y_i étant divisée par la probabilité d'inclusion p_i de l'individu i auquel elle se rapporte :

$$\hat{\tau} = \sum_{i \in S} \frac{y_i}{p_i}$$

où $p_i = \mathrm{P}(i \in S)$.

Comment déterminer ces probabilités d'inclusion p_i ? Leurs valeurs dépendent bien évidemment des plans de sondage choisis pour les deux degrés du sondage. Regardons cela d'un peu plus près.

Pour qu'un individu i se retrouve dans l'échantillon final S, il faut d'abord que l'unité primaire à laquelle il appartient soit prélevée au premier degré du sondage et fasse ainsi partie de l'échantillon S_{PRIM}, puis que l'individu i soit sélectionné au cours du tirage effectué dans son unité primaire. Supposons que l'individu i appartienne à l'unité primaire n° g. La théorie des probabilités nous indique en fait que la probabilité que cet individu i appartienne à l'échantillon final S est donnée par la probabilité que l'unité primaire U_g soit sélectionnée pour faire partie de l'échantillon S_{PRIM}, multipliée par la probabilité que l'individu i se retrouve dans le sous-échantillon S_g prélevé, au deuxième degré du sondage, dans l'unité primaire U_g (sachant que celle-ci a été prélevée au premier degré du sondage) : pour $i \in U_g$,

$$p_i = \mathrm{P}(i \in S) = \mathrm{P}\big(U_g \in S_{\text{PRIM}}\big)\mathrm{P}\big(i \in S_g | U_g \in S_{\text{PRIM}}\big).$$

Les probabilités p_i dépendent donc des probabilités d'inclusion dans l'échantillon S_{PRIM} des unités primaires de la population et des probabilités d'inclusion dans l'échantillon S_g des individus de l'unité primaire n° g ; elles varient donc selon le type d'échantillonnage aléatoire choisi pour chacun des deux degrés du sondage.

Nous allons nous contenter ici d'étudier les deux situations particulières les plus fréquemment rencontrées dans la pratique : la première situation est celle où l'on effectue un échantillonnage du type PESR aux deux degrés du sondage ; la seconde situation est celle où l'on réalise un sondage PPS des unités primaires et un sondage PESR de taille fixe des individus dans les unités primaires sélectionnées au premier degré du sondage.

Avant de nous pencher sur ces deux cas particuliers de sondage à deux degrés, revenons encore un petit moment sur l'estimateur de τ. Nous avons vu qu'il s'exprimait comme la somme, sur tous les individus i de l'échantillon S, des valeurs y_i divisées par les probabilités d'inclusion p_i :

$$\hat{\tau} = \sum_{i \in S} \frac{y_i}{p_i}.$$

Mais l'échantillon final S est la réunion des sous-échantillons S_g tirés dans les unités primaires U_g qui appartiennent à l'échantillon S_{PRIM} prélevé au premier degré du sondage. Par conséquent :

$$\hat{\tau} = \sum_{g \in S_{\text{PRIM}}} \sum_{i \in S_g} \frac{y_i}{p_i}.$$

Nous pouvons remplacer la probabilité d'inclusion p_i par l'expression que nous lui avons trouvée. On obtient alors :

$$\begin{aligned}\hat{\tau} &= \sum_{g \in S_{\text{PRIM}}} \sum_{i \in S_g} \frac{y_i}{\mathrm{P}(U_g \in S_{\text{PRIM}})\mathrm{P}(i \in S_g | U_g \in S_{\text{PRIM}})} \\ &= \sum_{g \in S_{\text{PRIM}}} \frac{1}{\mathrm{P}(U_g \in S_{\text{PRIM}})} \sum_{i \in S_g} \frac{y_i}{\mathrm{P}(i \in S_g | U_g \in S_{\text{PRIM}})} \\ &= \sum_{g \in S_{\text{PRIM}}} \frac{\hat{\tau}_g}{\mathrm{P}(U_g \in S_{\text{PRIM}})}.\end{aligned}$$

Ainsi, $\hat{\tau}$ peut être réécrit comme la somme, sur toutes les unités primaires g sélectionnées au premier degré, des estimateurs de Horvitz-Thompson de leurs totaux τ_g, divisés par les probabilités d'inclusion de ces unités primaires g. Vous voyez l'analogie avec l'estimateur de τ dans le cas du sondage en grappes. Dans le cas du sondage en grappes, c'étaient les valeurs exactes des totaux τ_g des grappes ou unités primaires sélectionnées qui apparaissaient dans l'expression de $\hat{\tau}$. Dans le cas du sondage à deux degrés, on ne peut plus déterminer les valeurs exactes de ces totaux τ_g, car les unités primaires sélectionnées ne sont plus *recensées* ; il nous faut donc estimer ces totaux à l'aide des sous-échantillons S_g prélevés dans ces unités primaires.

De manière générale, la variance de $\hat{\tau}$ va pouvoir s'écrire comme une somme de deux termes : le premier terme est induit par la fluctuation d'échantillonnage engendrée par le premier degré du sondage, tandis que le deuxième terme est lié à la fluctuation d'échantillonnage engendrée par le second degré du sondage. Pour rendre le sondage à deux degrés efficace, il faut donc choisir un plan de sondage efficace pour le premier degré du sondage, et un plan de sondage efficace pour le deuxième degré du sondage.

5.3.3 Premier cas particulier : tirage PESR aux deux degrés du sondage

a) Estimateur de τ

Si l'on décide de réaliser un échantillonnage PESR de taille m au premier degré du sondage, chaque unité primaire U_g possède alors une probabilité de se retrouver dans l'échantillon S_{PRIM} égale au taux de sondage m/M : pour tout $g = 1, \cdots, M$,

$$\mathrm{P}(g \in S_{\text{PRIM}}) = \frac{m}{M}.$$

Et si l'on décide de réaliser un échantillonnage PESR de taille n_g dans l'unité primaire U_g prélevée au premier degré du sondage, chaque individu i de U_g possède alors une

probabilité d'appartenir au sous-échantillon S_g égale au taux de sondage n_g/N_g : pour tout $i \in U_g$ (avec $g \in S_{\text{PRIM}}$),

$$\mathrm{P}(i \in S_g) = \frac{n_g}{N_g}.$$

Dès lors, pour tout individu i appartenant à U_g :

$$p_i = \mathrm{P}(i \in S) = \frac{m}{M}\frac{n_g}{N_g}.$$

Nous obtenons alors :

$$\hat{\tau} = \sum_{g \in S_{\text{PRIM}}} \sum_{i \in S_g} \frac{y_i}{\frac{m}{M}\frac{n_g}{N_g}} = \frac{M}{m} \sum_{g \in S_{\text{PRIM}}} \frac{N_g}{n_g} \sum_{i \in S_g} y_i$$
$$= \frac{M}{m} \sum_{g \in S_{\text{PRIM}}} N_g \bar{y}_g ,$$

où $\bar{y}_g$ est la moyenne des valeurs observées pour la variable d'intérêt y dans le sous-échantillon S_g.

Dans le cas particulier où, au second degré du sondage, on applique le même taux de sondage f_2 dans toutes les unités primaires prélevées, c'est-à-dire dans le cas où $\frac{n_g}{N_g} = f_2$ pour tout g appartenant à S_{PRIM}, $\hat{\tau}$ peut se réécrire sous la forme suivante :

$$\hat{\tau} = \frac{M}{m f_2} \sum_{g \in S_{\text{PRIM}}} \sum_{i \in S_g} y_i = \frac{M}{m f_2} \sum_{i \in S} y_i .$$

On est en réalité dans le cas où tous les individus de la population U ont la *même* probabilité d'inclusion, égale à mf_2/M. Ceci peut être vu comme un atout pour ce sondage à deux degrés particulier. Cependant, cette allocation très particulière peut être rejetée dans certaines circonstances car elle impose des tailles d'échantillons variables selon l'unité primaire considérée, ce qui pose en pratique des problèmes de gestion des charges des enquêteurs, notamment lorsque chaque enquêteur est affecté à une unité primaire particulière.

Pour faciliter la gestion du travail des enquêteurs, on préférera dans certains cas tirer le même nombre d'individus, fixé *a priori*, dans chacune des unités primaires sélectionnées au premier degré du sondage : dans ce cas, la taille de l'échantillon final S sera elle aussi fixée *a priori*, et non plus aléatoire comme dans les autres cas.

b) La variance de l'estimateur de τ

On peut montrer que la variance de $\hat{\tau}$ vaut :

$$\mathrm{V}(\hat{\tau}) = M^2 \left(1 - \frac{m}{M}\right) \frac{\sigma^2_{\tau;\text{corr}}}{m} + \frac{M}{m} \sum_{g=1}^{M} \mathrm{V}(\hat{\tau}_g) \qquad (= A + B)$$

avec

$$\sigma^2_{\tau;\text{corr}} = \frac{1}{M-1} \sum_{g=1}^{M} (\tau_g - \mu_\tau)^2 = \frac{1}{M-1} \sum_{g=1}^{M} \left(\tau_g - \frac{\tau}{M}\right)^2$$

et, pour $g = 1, \ldots, M$:

$$\mathrm{V}(\hat{\tau}_g) = N_g^2 \left(1 - \frac{n_g}{N_g}\right) \frac{\sigma^2_{g;\text{corr}}}{n_g},$$

où $\sigma^2_{g;\text{corr}}$ n'est autre que la variance corrigée de la variable d'intérêt y dans l'unité primaire n° g.

On voit donc que la variance de $\hat{\tau}$ se décompose comme une somme de deux termes que nous appellerons A et B :

- Le terme A est lié au tirage PESR des m unités primaires dans U_{PRIM}. Il correspond à la variance que nous avions pour l'estimateur de τ dans le cas du sondage en grappes avec tirage PESR de m grappes. Il est directement lié, au travers de la variance $\sigma^2_{\tau;\text{corr}}$, à la dispersion des totaux τ_g des différentes unités primaires qui composent la population. Il est par ailleurs inversement proportionnel au nombre m d'unités primaires prélevées.
- Le second terme — le terme B — est lié au second degré du sondage : il fait intervenir la somme, sur toutes les unités primaires de la population, de la variance de l'estimateur du total τ_g de chaque unité primaire g, dans le cadre d'un sondage PESR dans cette unité primaire.

Que pouvons-nous en conclure quant à l'efficacité du sondage à deux degrés ?

1) Tout d'abord, on peut déduire du terme A que, comme dans le cas du sondage en grappes, il est intéressant de disposer d'un grand nombre d'unités primaires de tailles voisines et aussi faibles que possible, et de « comportement moyen » semblable. Pour parvenir à ce résultat, il faut que soient regroupés au sein d'une même unité primaire, des individus bien différents les uns des autres : la dispersion de la variable d'intérêt y doit être concentrée *à l'intérieur* des unités primaires, et doit donc être faible *entre* les unités primaires.

2) Le terme B fait intervenir la dispersion de y au sein des unités primaires de la population. On pourrait penser que la construction des unités primaires selon les règles que nous venons d'énoncer va rendre les variances $\sigma^2_{g;\text{corr}}$ « grandes », et va dès lors venir gonfler la valeur du terme B. Heureusement, l'effet « négatif » des valeurs élevées des variances $\sigma^2_{g;\text{corr}}$ est contrebalancé par la présence du facteur $1/m$ dans le terme B, ainsi que par la présence, pour chaque unité primaire, de la taille d'échantillon n_g au dénominateur de la variance de $\hat{\tau}_g$.

   En réalité, le terme A est d'ordre de grandeur $1/m$, alors que le terme B est d'ordre de grandeur $1/(m\bar{n})$ où $\bar{n}$ est la taille moyenne des échantillons d'individus tirés au sein des unités primaires ($\bar{n} = \frac{1}{M}\sum_{g=1}^{M} n_g$). L'existence de variances $\sigma^2_{g;\text{corr}}$ rendues aussi grandes que possible par le découpage en unités primaires est donc un moindre mal, le coefficient $1/(m\bar{n})$ étant, en général, très inférieur à $1/m$.

3) Si le budget dont on dispose nous permet d'augmenter la taille de l'échantillon final S, il vaut mieux le faire en augmentant le nombre m d'unités primaires prélevées au premier degré du sondage plutôt qu'en augmentant uniquement les tailles n_g des échantillons tirés au second degré du sondage. Si on augmente m seulement, sans toucher aux tailles n_g, on va diminuer à la fois le terme A et le terme B ; si on augmente les tailles n_g sans toucher à m, on ne va diminuer que le

terme B. La seconde opération permettra donc un moins grand accroissement de la précision que la première.

4) Enfin, il faut être bien conscient du fait que, dans la plupart des populations courantes et pour la plupart des variables traitées, la variance $\sigma^2_{\tau;\text{corr}}$ des totaux est un terme d'une importance réellement majeure dans la valeur de la variance de $\hat{\tau}$: $\sigma^2_{\tau;\text{corr}}$ peut facilement prendre une valeur telle que la variance de $\hat{\tau}$ s'avère excessivement grande. Il est donc impératif de contrôler $\sigma^2_{\tau;\text{corr}}$ *en priorité*. A l'opposé, les dispersions $\sigma^2_{g;\text{corr}}$ à l'intérieur même des unités primaires conservent en général des valeurs raisonnables.
Pour illustrer cette remarque, prenons l'exemple de l'estimation, par un sondage à deux degrés, du nombre total d'hommes dans une population découpée en communes. La variable d'intérêt y est la variable dichotomique prenant la valeur 1 chez tout individu de sexe masculin, la valeur 0 chez tout individu de sexe féminin. Dans chaque commune U_g, la variance $\sigma^2_{g;\text{corr}}$ de y vaut $\pi_g(1-\pi_g)$ où π_g est la proportion exacte d'hommes dans la commune ; les variances $\sigma^2_{g;\text{corr}}$ sont dès lors toutes inférieures ou égales à 0,25. Par contre, le total τ_g attaché à la commune n° g correspond au nombre exact d'hommes que compte cette commune ; la variance $\sigma^2_{\tau;\text{corr}}$ mesure donc la dispersion des effectifs d'hommes dans les différentes communes. Si on n'y prend pas garde, et que les communes qui partitionnent la population sont de tailles sensiblement différentes, les effectifs d'hommes seront eux-mêmes très dispersés : ainsi, par exemple, même si on s'en tient à des communes rurales, les tailles de ces communes peuvent varier de 20 à 2 000 personnes, et l'on peut alors se retrouver avec des nombres d'hommes par commune allant d'une dizaine à un millier. Dans ces conditions, le terme $\sigma^2_{\tau;\text{corr}}$ sera explosif, rendant la précision de $\hat{\tau}$ catastrophique !

c) Estimation de la variance de l'estimateur de τ

Notez que l'on peut estimer de manière naturelle la variance de $\hat{\tau}$ en remplaçant, dans le premier terme de son expression, la variance inconnue $\sigma^2_{\tau;\text{corr}}$ par son estimation $\hat{\sigma}^2_{\tau;\text{corr}}$ définie ci-dessous, et en remplaçant, dans le second terme, la somme sur toutes les unités primaires de la population par la somme sur les unités primaires sélectionnées, ainsi que les variances inconnues $\sigma^2_{g;\text{corr}}$ par $s^2_{g;\text{corr}}$, les variances corrigées de y dans les échantillons S_g tirés au deuxième degré du sondage :

$$\widehat{\mathrm{V}}(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\frac{\hat{\sigma}^2_{\tau;\text{corr}}}{m} + \frac{M}{m}\sum_{g\in S_{\text{PRIM}}}\widehat{\mathrm{V}}(\hat{\tau}_g) \qquad = (a+b)$$

avec

$$\hat{\tau}_g = N_g\bar{y}_g\,,$$

$$\hat{\sigma}^2_{\tau;\text{corr}} = \frac{1}{m-1}\sum_{g\in S_{\text{PRIM}}}\left(\hat{\tau}_g - \frac{\hat{\tau}}{M}\right)^2$$

et, pour $g = 1, \ldots, M$:

$$\widehat{\mathrm{V}}(\hat{\tau}_g) = N_g^2\left(1-\frac{n_g}{N_g}\right)\frac{s^2_{g;\text{corr}}}{n_g}$$

où

$$s^2_{g;\text{corr}} = \frac{1}{n_g - 1} \sum_{i \in S_g} \left(y_i - \bar{y}_g\right)^2.$$

Attention ! Le terme a n'est pas un estimateur sans biais du terme A de l'expression de $\mathrm{V}(\hat{\tau})$ et le terme b n'est pas un estimateur sans biais du terme B. Toutefois, par un petit miracle algébrique, $\widehat{\mathrm{V}}(\hat{\tau}) = a + b$ s'avère être un estimateur *sans biais* de $\mathrm{V}(\hat{\tau}) = A + B$!

Exercice 5.5

Remarque : cet exercice porte sur la même population que celle étudiée dans l'exercice 5.4.

On s'intéresse à une certaine population constituée de 1 934 élèves de 1re année du secondaire qui ont présenté, en début d'année scolaire, un test standardisé (sur 20 points) destiné à évaluer leur niveau en mathématiques. On désire y estimer :

(i) la note moyenne (μ_{note}) obtenue à ce test ;
(ii) le nombre ($N_{\geq 10}$) et la proportion ($\pi_{\geq 10}$) d'élèves qui ont obtenu une note supérieure ou égale à 10 à ce test.

La population considérée est en réalité constituée de 100 classes comptant entre 8 et 30 élèves. Pour obtenir les estimations demandées, on décide de prélever un échantillon par un sondage à deux degrés : on réalise un tirage PESR de 40 classes parmi les 100 classes qui partitionnent la population, puis de sélectionner par tirage PESR 5 élèves dans chacune des 40 classes prélevées au premier degré du sondage.

Vous trouverez dans la feuille « 2D-PESR » du fichier Degrés_ex.xlsx (téléchargeable sur l'UV) les numéros des classes sélectionnées, leurs tailles, ainsi que les notes obtenues au test d'évaluation en mathématiques par les 200 élèves de l'échantillon sélectionné.

a) Estimez la note moyenne (μ_{note}) obtenue au test par les élèves de la population.

b) Estimez la variance de l'estimateur de μ_{note}.

c) Déterminez l'intervalle de confiance pour μ_{note}, au niveau de confiance de 95%.
c.1) Que vaut la borne inférieure de cet intervalle de confiance ?
c.2) Que vaut la borne supérieure de cet intervalle de confiance ?

d) Estimez le nombre ($N_{\geq 10}$) d'élèves qui ont obtenu une note supérieure ou égale à 10 au test d'évaluation en mathématiques.

e) Estimez la proportion ($\pi_{\geq 10}$) d'élèves qui ont obtenu une note supérieure ou égale à 10 au test d'évaluation en mathématiques.

5.3.4 Second cas particulier : sondage à deux degrés autopondéré (sondage PPS des unités primaires et sondage PESR de taille fixe des unités secondaires)

a) Estimateur de τ

Considérons à présent le second cas particulier de sondage à deux degrés. Au premier degré du sondage, on choisit de réaliser un tirage PPS de m unités primaires : ceci revient à affecter aux unités primaires des probabilités d'inclusion proportionnelles à leurs tailles. Ainsi, la probabilité que l'unité primaire n° g se retrouve dans l'échantillon S_{PRIM} est égale à $m\frac{N_g}{N}$.

Pour le second degré du sondage, on décide de faire appel à un sondage aléatoire simple de taille fixe n_0 (autrement dit, on décide de prélever par tirage PESR n_0 individus ou unités dans chacune des unités primaires qui ont été sélectionnées au premier degré du sondage). Dans ce cas, si l'unité primaire U_g a été prélevée au premier degré, tous les individus qu'elle contient se voient affecter une probabilité d'inclusion dans S_g égale à $\frac{n_0}{N_g}$, le taux de sondage appliqué dans l'unité primaire.

La taille n_S de l'échantillon final S est fixe, égale à mn_0. Par ailleurs, tout individu i appartenant à U_g possède une probabilité p_i de se retrouver dans l'échantillon final S égale à

$$m\frac{N_g}{N}\frac{n_0}{N_g} = \frac{mn_0}{N} = \frac{n_S}{N}.$$

On le voit, tous les individus de la population se retrouvent avec la *même* probabilité d'appartenir à l'échantillon final S, probabilité égale au taux de sondage global appliqué dans la population-cible U. Ceci explique pourquoi ce sondage à deux degrés particulier est dit *autopondéré*.

Enfin, l'estimateur $\hat{\tau}$ de τ prend la forme particulièrement simple de $N\bar{y}$, où $\bar{y}$ est la moyenne des valeurs observées pour la variable y sur l'ensemble des individus de l'échantillon final S. En effet :

$$\hat{\tau} = \sum_{i\in S}\frac{y_i}{n_S/N} = N\frac{1}{n_S}\sum_{i\in S} y_i = N\bar{y}.$$

Ainsi, le sondage à deux degrés autopondéré nous conduit, malgré sa complexité apparente, aux mêmes probabilités d'inclusion et au même estimateur de τ que le sondage aléatoire simple (PESR) !

Exercice 5.6

Replaçons-nous dans le même contexte de travail que dans l'exercice 5.5, mais supposons à présent que l'échantillon d'élèves soit prélevé de la manière suivante : on prélève 40 classes par sondage PPS parmi les 100 classes qui partitionnent la population, puis on sélectionne par tirage PESR 5 élèves dans chacune des 40 classes prélevées au premier degré du sondage.

Vous trouverez dans la feuille « 2D-auto » du fichier Degrés_ex.xlsx les numéros des classes sélectionnées, leurs tailles, ainsi que les notes obtenues au test d'évaluation en mathématiques par les 200 élèves de l'échantillon sélectionné.

a) Que vaut la probabilité d'inclusion d'un élève de la population ?

b) Estimez la note moyenne (μ_{note}) obtenue au test par les élèves de la population.

c) Estimez le nombre ($N_{\geq 10}$) d'élèves qui ont obtenu une note supérieure ou égale à 10 au test d'évaluation en mathématiques.

d) Estimez la proportion ($\pi_{\geq 10}$) d'élèves qui ont obtenu une note supérieure ou égale à 10 au test d'évaluation en mathématiques.

5.4 Conclusion

Les sondages à deux degrés ou en grappes peuvent se montrer efficaces pour certains domaines d'études, et pas du tout pour d'autres. Leur efficacité dépendra essentiellement de la capacité à construire des unités primaires hétérogènes, regroupant des individus bien différents les uns des autres pour ce qui est de la variable d'intérêt. Ce n'est qu'en assurant l'hétérogénéité des unités primaires que l'on peut être assuré que chaque grappe sélectionnée ou chaque échantillon tiré dans une unité primaire participe bien à rendre compte au mieux de la diversité des individus de la population tout entière.

Ainsi, la pertinence du plan de sondage et du découpage en unités primaires dépendra du sujet étudié. Prenons le cas où les unités primaires sont des ménages. Un ménage est une grappe d'individus composée en général de personnes de sexes et d'âges différents et comprenant souvent des actifs, des inactifs, des écoliers, etc. Dans une enquête socio-économique, le ménage sera une grappe très efficace pour estimer la proportion d'hommes dans la population, la proportion de personnes actives dans la population, la part d'une certaine classe d'âges dans la population… Mais ce sera une grappe moins efficace pour étudier le niveau d'instruction, l'origine ethnique, etc., des individus de la population.

Par ailleurs, on aura intérêt à forcer sur le nombre d'unités primaires prélevées plus que sur le nombre d'individus enquêtés au deuxième degré du sondage : il vaut mieux sélectionner beaucoup d'unités primaires avec relativement peu de questionnaires administrés dans chacune d'elles, plutôt que tirer peu d'unités primaires, même si on y administre ensuite un grand nombre de questionnaires. Ceci doit bien sûr tenir compte aussi des différents facteurs de coût.

Chapitre 6
Diverses problématiques

6.1 Estimation de la taille de la population

6.1.1 Introduction

Nous avons toujours fait jusqu'ici comme si la taille N de la population U était connue. C'est effectivement le cas lorsqu'on dispose d'une base de sondage complète pour cette population. Il arrive cependant que l'on ne dispose pas d'une telle base de sondage et qu'il nous soit très difficile de la constituer par nous-même : c'est le genre de situation qui peut nous pousser d'ailleurs à mettre en œuvre un sondage en grappes ou à deux degrés, avec tirage PESR des unités primaires, pour lequel il nous faut juste disposer d'une liste exhaustive des unités primaires qui composent la population et construire une base de sondage des individus pour les unités primaires sélectionnées au cours du sondage. Vous pouvez vérifier, en retournant aux résultats que nous avons étudiés dans le chapitre précédent, que pour ce type particulier de sondage aléatoire, l'estimation du total-population τ ne nécessite pas de connaître N. Ce n'est que si nous désirons estimer la moyenne-population μ que nous aurons besoin de N.

Que faire si N est inconnu ? Il n'y a qu'une solution possible : estimer N.

6.1.2 Comment estimer N ?

Pour estimer N, nous pouvons une nouvelle fois suivre la démarche de Horvitz-Thompson.

Définissons la variable $\mathcal{Z}$ comme la variable indicatrice de l'appartenance à la population U : $\mathcal{Z}$ prend la valeur 1 sur tout individu de la population U et la valeur 0 sur tout individu qui n'appartient pas à cette population. Le total de cette variable $\mathcal{Z}$ dans U correspond au nombre d'individus qui appartiennent à U, et est donc égal à la taille N de U :

$$\tau_z = \sum_{i \in U} z_i = \sum_{i \in U} 1 = N.$$

Puisque N correspond au total de $\mathcal{Z}$ dans la population U, on peut l'estimer via l'estimateur de Horvitz-Thompson de τ_z. On a donc :

$$\widehat{N} = \sum_{i \in S} \frac{z_i}{p_i},$$

c'est-à-dire, puisque z_i vaut 1 pour tout individu i de l'échantillon S,

$$\widehat{N} = \sum_{i \in S} \frac{1}{p_i}.$$

6.1.3 Remarques sur $\widehat{N}$

$\widehat{N}$ est un estimateur de Horvitz-Thompson : il estime donc sans biais la taille N de la population.

Que peut-on dire de sa précision ? Rappelez-vous ce que nous avons dit de la variance de l'estimateur de Horvitz-Thompson d'un total dans le chapitre 4 consacré au sondage PISR, et qui nous avait d'ailleurs conduit au sondage PPS : dans le cas d'un sondage aléatoire de taille fixe n, l'estimateur de Horvitz-Thompson du total d'une variable dans la population a une variance nulle si les probabilités d'inclusion affectées aux individus de la population sont proportionnelles aux valeurs de la variable considérée. Ainsi, $\widehat{N}$ aura une variance nulle — $\widehat{N}$ estimera parfaitement la taille N de la population, sans aucune erreur d'échantillonnage — si les probabilités d'inclusion p_i affectées aux individus i de la population sont proportionnelles aux valeurs z_i de la variable Z. Mais puisque ces valeurs z_i sont toutes égales à 1, cela revient à dire que les probabilités d'inclusion p_i des individus de la population doivent toutes avoir la même valeur si l'on veut que $\widehat{N}$ soit « parfait ».

En conclusion, avec un plan de sondage à probabilités *égales* et donnant lieu à des échantillons de *taille fixe*, $\widehat{N}$ estime N parfaitement quel que soit l'échantillon prélevé. Plus on s'éloigne du plan de sondage à probabilités égales — plus les probabilités d'inclusion associées au plan de sondage diffèrent fortement d'un individu à l'autre de la population — plus la variance de $\widehat{N}$ aura tendance à être grande, moins la précision de $\widehat{N}$ sera bonne.

6.1.4 Exercice 6.1

Replaçons-nous dans le contexte de l'étude de la population considérée dans les exercices 5.4, 5.5 et 5.6 proposés dans le chapitre 5 consacré au sondage en grappes et au sondage à deux degrés.

Dans ces exercices, on s'intéresse à une population constituée de $N = 1\,934$ élèves de 1re année du secondaire. Cette population est partitionnée en 100 classes comptant entre 8 et 30 élèves.

Dans l'exercice 5.5, on met en œuvre un premier sondage à deux degrés : on sélectionne par tirage PESR 40 classes parmi les 100 classes qui composent la population, puis on prélève par tirage PESR 5 élèves dans chacune des 40 classes prélevées au premier degré du sondage. Notez que la détermination de ce plan de sondage et des probabilités d'inclusion qui lui sont associées ne nécessite pas la connaissance de la taille N de la population d'élèves.

Nous aurions pu imaginer un autre plan de sondage à deux degrés ne nécessitant pas la connaissance de N : celui consistant à sélectionner par tirage PESR 40 classes parmi les 100 classes qui composent la population, puis à prélever par tirage PESR un quart des élèves dans chacune des 40 classes prélevées au premier degré du sondage.

Dans le présent exercice, nous allons tenter d'estimer N à partir des échantillons prélevés selon ces deux plans de sondage particuliers.

PARTIE 1 : estimation de N à partir de l'échantillon prélevé dans l'exercice 5.5 (premier plan de sondage à deux degrés : tirages PESR aux 2 degrés ; $m = 40$; $n_g = 5$ pour tout $g \in S_{\text{PRIM}}$)

L'échantillon prélevé dans le cadre de l'exercice 5.5 est repris dans la feuille « 2D-PESR-1 » du fichier Estimation_N_ex.xlsx.

Quelle estimation de *N* vous fournit cet échantillon ?
(Arrondissez votre réponse à l'entier le plus proche.)

PARTIE 2 : estimation de N à partir de l'échantillon prélevé selon le deuxième plan de sondage à deux degrés (tirages PESR aux 2 degrés ; $m = 40$; $n_g = N_g/4$ pour tout $g \in S_{\text{PRIM}}$)

L'échantillon prélevé selon ce plan de sondage particulier est repris dans la feuille « 2D-PESR-2 » du fichier Estimation_N_ex.xlsx.

Quelle estimation de *N* vous fournit cet échantillon ?
(Arrondissez votre réponse à l'entier le plus proche.)

6.2 Estimation d'un ratio

6.2.1 Introduction

Il arrive que l'on veuille estimer le rapport (ou ratio) des totaux de deux variables différentes dans la population. Considérons, par exemple, les variables $\mathcal{Y}$ et $\mathcal{Z}$ dans la population U et supposons que nous cherchions à estimer le rapport

$$\Gamma = \frac{\tau_y}{\tau_z}$$

où τ_y et τ_z sont les totaux, inconnus, de $\mathcal{Y}$ et de $\mathcal{Z}$ dans U.

Γ peut correspondre, par exemple, à un ratio comptable, à un taux de chômage, à un certain coefficient budgétaire, etc. Si $\mathcal{Z}$ est la variable indicatrice de l'appartenance d'un individu à la population U, autrement dit si le total τ_Z correspond à la taille N inconnue de la population, estimer Γ revient à estimer la moyenne μ_y de la variable $\mathcal{Y}$ dans la population.

6.2.2 Comment estimer Γ ?

Puisque Γ est le rapport de τ_y sur τ_z, il est naturel de prendre pour estimateur de Γ le rapport de l'estimateur de τ_y sur l'estimateur de τ_z :

$$\hat{\Gamma} = \frac{\hat{\tau}_y}{\hat{\tau}_z}.$$

Ainsi, en particulier, si l'on a estimé le total τ_y de la variable $\mathcal{Y}$ et que l'on désire ensuite en estimer la moyenne μ_y alors qu'on ne connaît pas la taille N de la population, on fera appel à l'estimateur défini par le rapport de $\hat{\tau}_y$ sur $\widehat{N}$:

$$\hat{\mu}_y = \frac{\hat{\tau}_y}{\widehat{N}}.$$

6.2.3 Remarques sur $\hat{\Gamma}$

Il faut être conscient du fait que, même si $\hat{\tau}_y$ et $\hat{\tau}_z$ sont des estimateurs sans biais de τ_y et τ_z, respectivement, $\hat{\Gamma}$ en revanche est généralement un estimateur *biaisé* du rapport Γ. Ceci est tout simplement dû au fait que l'espérance d'un rapport de variables aléatoires n'est généralement pas égal au rapport des espérances de ces variables.

Dans le cas du sondage aléatoire simple (PESR), on montre que le biais de $\hat{\Gamma}$ est proportionnel au facteur $(1-f)/n$: ceci nous indique que, plus la taille n de l'échantillon se rapproche de la taille N de la population, plus le biais de $\hat{\Gamma}$ est faible. Par ailleurs, le biais de $\hat{\Gamma}$ apparaît également négligeable lorsque la variable $\mathcal{Y}$ est approximativement proportionnelle à la variable $\mathcal{Z}$; plus précisément, quelle que soit la taille n de l'échantillon, le biais de $\hat{\Gamma}$ est nul lorsque la droite de régression des moindres

carrés de y en z *dans la population* est une droite passant par l'origine et de pente égale au rapport Γ.

On trouve dans la littérature bien d'autres résultats sur le biais, mais aussi sur la précision de l'estimateur d'un ratio, que ce soit dans le cas d'un sondage aléatoire simple, d'un sondage stratifié ou d'un sondage PISR en général.

6.3 Introduction à la problématique de l'étude de domaines

6.3.1 Introduction

a) Objectif et domaine d'étude

Il arrive que, dans le cadre d'un sondage mené dans une population U, on veuille estimer l'un ou l'autre paramètre caractérisant la distribution d'une variable y non plus dans la population U tout entière, mais plutôt dans une certaine sous-population D de U. Cette sous-population à laquelle on s'intéresse tout particulièrement constitue ce que l'on appelle un *domaine* d'étude. On peut ainsi, par exemple, vouloir estimer le total ou la moyenne de la variable y dans le domaine D. Ce total, que nous désignerons par $\tau_{y|D}$ n'est autre que la somme des valeurs y_i que prend la variable y sur les individus i qui appartiennent à D. Quant à la moyenne de y dans D, que nous noterons $\mu_{y|D}$, elle est égale au total de y dans D divisé par la taille N_D du domaine :

$$\tau_{y|D} = \sum_{i \in D} y_i \quad \text{et} \quad \mu_{y|D} = \frac{\tau_{y|D}}{N_D}.$$

La difficulté rencontrée dans ce contexte provient du fait que la base de sondage associée à la population globale U ne nous permet pas d'identifier *a priori* quels sont les individus de la population qui appartiennent au domaine D ; il nous est donc impossible de réaliser un échantillonnage spécifique dans le domaine. Par ailleurs, il faut être conscient du fait que la taille N_D du domaine D est également très souvent inconnue ; cela aussi aura un impact sur la construction des estimateurs utilisés.

On se retrouve dans une telle situation lorsque, par exemple, on mène une enquête par sondage auprès de la population étudiante d'une grande université et que l'on souhaite estimer le montant moyen des dépenses mensuelles en alimentation des étudiants qui vivent en semaine en logement-étudiant. Le domaine considéré est celui constitué par ces étudiants vivant en logement-étudiant ; les étudiants de ce domaine d'études ne sont pas identifiables dans la base de sondage que constitue, par exemple, le fichier des inscriptions de l'université en question et leur nombre nous est inconnu.

Un autre exemple d'étude de domaine est celui où l'on mène une enquête par sondage dans une certaine population relativement vaste d'individus et que l'on veut estimer le montant total des frais médicaux encourus au cours des six derniers mois par les individus de la population atteints d'un type particulier de maladie. Ou enfin lorsqu'on veut estimer la part des personnes actives dans le domaine constitué des individus de la population qui habitent une région géographique bien spécifique.

b) L'importance relative du domaine

Une caractéristique du domaine d'étude va jouer un rôle fondamental pour la suite de l'analyse : il s'agit de *l'importance relative du domaine* dans la population. Cette

importance relative n'est autre que la part des individus de la population U qui appartiennent au domaine D. Il s'agit donc de la proportion, que nous noterons π_D, définie par le rapport de la taille N_D du domaine sur la taille N de la population :

$$\pi_D = \frac{N_D}{N}.$$

Cette proportion est inconnue dès le moment où la taille N_D de D est inconnue.

Précisons que nous allons supposer ici que la taille N de la population U est, quant à elle, bien connue.

c) Variables supplémentaires utilisées

A côté de la variable d'intérêt y dont on veut estimer le total ou la moyenne dans le domaine d'étude, nous pouvons définir deux autres variables qui vont nous permettre de formaliser la situation de façon simple.

Nous pouvons tout d'abord introduire la variable z, indicatrice de l'appartenance d'un individu au domaine D : la valeur z_i que prend la variable z chez l'individu i est égale à 1 si cet individu i appartient au domaine D, et est égale à 0 sinon. Pour tout $i \in U$:

$$z_i = \begin{cases} 1 & \text{si } i \in D \\ 0 & \text{si } i \notin D \end{cases}.$$

Le total de cette variable z dans la population U correspond au nombre d'individus de la population qui appartiennent au domaine D, c'est-à-dire à la taille N_D de ce dernier. Quant à la moyenne de la variable z dans la population U, elle est égale au total de z divisé par N, soit à N_D divisé par N ; elle coïncide donc avec l'importance relative π_D du domaine D dans la population. Ainsi :

$$\tau_z = \sum_{i \in U} z_i = N_D \quad \text{et} \quad \mu_z = \frac{\tau_z}{N} = \frac{N_D}{N} = \pi_D .$$

Nous pouvons également définir la variable v comme étant le produit de la variable y et de la variable z : pour tout $i \in U$,

$$v_i = y_i z_i = \begin{cases} y_i & \text{si } i \in D \\ 0 & \text{si } i \notin D \end{cases}.$$

Cette variable v nous permet en fait de restreindre au seul domaine D notre champ d'analyse de la variable y.

Le total de la variable v dans la population U n'est autre que le total de la variable y dans le domaine D :

$$\tau_v = \sum_{i \in U} v_i = \sum_{i \in D} y_i = \tau_{y|D} .$$

Par ailleurs, la moyenne de y dans le domaine D correspond au rapport du total de la variable v et du total de la variable z dans la population U : en effet,

$$\mu_{y|D} = \frac{\tau_{y|D}}{N_D} = \frac{\tau_v}{\tau_z}.$$

d) L'échantillon

Quelles sont à présent les caractéristiques de l'échantillon ?

On prélève aléatoirement un échantillon S de taille n dans la population U (voir la figure 6.1). En interrogeant les individus de cet échantillon sur leur appartenance ou non au domaine d'étude, on fait apparaître le sous-échantillon S_D constitué des individus de l'échantillon S qui appartiennent au domaine D :

$$S_D = S \cap D.$$

Non seulement la composition de cet échantillon S_D est aléatoire, mais sa taille n_D l'est aussi. On vérifie cependant facilement que l'espérance de n_D, autrement dit la taille moyenne des échantillons S_D qu'il est possible d'obtenir, est égale à $n\pi_D$ où, rappelons-le, π_D est l'importance relative du domaine D dans la population U :

$$\mathrm{E}(n_D) = n\pi_D = n\frac{N_D}{N}.$$

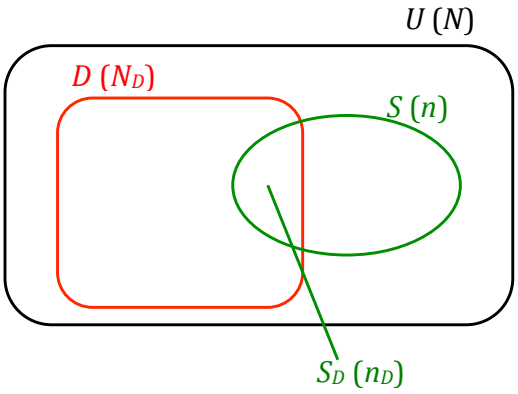


Figure 6.1 – *Echantillonnage pour l'étude du domaine D*

Il est clair que si le domaine D est de faible taille, on risque de se retrouver avec un sous-échantillon S_D ne contenant que très peu d'individus. Dans le cas de tout petits domaines d'étude, on risque même de n'avoir aucun individu de S qui appartient à D ! C'est de là que vient toute la difficulté d'estimer des paramètres dans ce qu'on appelle des « petits domaines », c'est-à-dire des domaines de faible importance relative. Nous en reparlerons de manière plus spécifique à la fin de ce chapitre.

6.3.2 Estimation directe

a) Estimateur de $\tau_{y|D}$

Commençons par rechercher l'estimateur du total de y dans le domaine D.

Puisque $\tau_{y|D}$ correspond au total de la variable v dans la population U, nous pouvons l'estimer via l'estimateur de Horvitz-Thompson de τ_v :

$$\hat{\tau}_{y|D} = \hat{\tau}_v = \sum_{i \in S} \frac{v_i}{p_i},$$

où p_i désigne, comme d'habitude, la probabilité d'inclusion affectée à l'individu i par le plan de sondage choisi pour prélever l'échantillon S. Puisque v_i est égal à y_i si l'individu i appartient à D et est nul sinon, on obtient :

$$\hat{\tau}_{y|D} = \sum_{i \in S_D} \frac{y_i}{p_i}.$$

Ainsi, l'estimateur du total de la variable y dans le domaine D a la même forme que l'estimateur du total de y dans l'ensemble de la population U, si ce n'est qu'au lieu de prendre en considération les n individus de l'échantillon S, on ne prend en compte que les n_D individus du sous-échantillon S_D.

b) Estimateur de $\mu_{y|D}$

Qu'en est-il de l'estimateur de la moyenne de y dans le domaine D ? Cette moyenne est égale au total de y dans D divisé par la taille N_D de D :

$$\mu_{y|D} = \frac{\tau_{y|D}}{N_D}.$$

b.1) Si la taille N_D du domaine D est connue

Si la taille N_D est connue, on peut tout simplement estimer $\mu_{y|D}$ via

$$\hat{\mu}_{y|D} = \frac{\hat{\tau}_{y|D}}{N_D}.$$

b.2) Si la taille N_D du domaine D est inconnue

Si la taille N_D du domaine D est inconnue, il va nous falloir l'estimer par son estimateur de Horvitz-Thompson $\widehat{N}_D$ et l'on estimera alors $\mu_{y|D}$ par le rapport de $\hat{\tau}_{y|D}$ sur $\widehat{N}_D$. Pour distinguer cet estimateur de celui utilisé dans le cas où N_D est connu, nous le noterons $\hat{\hat{\mu}}_{y|D}$:

$$\hat{\hat{\mu}}_{y|D} = \frac{\hat{\tau}_{y|D}}{\widehat{N}_D}.$$

N_D étant le total de la variable z dans la population U, on estime la taille du domaine D via l'estimateur de Horvitz-Thompson de τ_z :

$$\widehat{N}_D = \hat{\tau}_z = \sum_{i \in S} \frac{z_i}{p_i},$$

c'est-à-dire, puisque z_i est égal à 1 pour tout individu i appartenant au domaine D et est égal à 0 pour tout individu i en dehors de D :

$$\widehat{N}_D = \sum_{i \in S_D} \frac{1}{p_i}.$$

Il en résulte que

$$\hat{\hat{\mu}}_{y|D} = \frac{\sum_{i \in S_D} \frac{y_i}{p_i}}{\sum_{i \in S_D} \frac{1}{p_i}}.$$

Exercice 6.2

Considérons une dernière fois la population des 1 934 élèves de 1re année du secondaire qui ont présenté, en début d'année scolaire, un test standardisé (sur 20 points) destiné à évaluer leur niveau en mathématiques. Cette population est constituée de 100 classes comptant entre 8 et 30 élèves.

On désire estimer la note moyenne obtenue par les élèves *qui ont réussi le test* (autrement dit qui ont obtenu une note supérieure ou égale à 10). Pour ce faire, on dispose de l'échantillon présenté dans la feuille « Dom1 » du fichier « Domaines_ex.xlsx » et obtenu en prélevant par tirage PESR 40 classes parmi les 100

classes qui composent la population, puis en prélevant par tirage PESR 5 élèves dans chacune des classes sélectionnées au premier degré du tirage.

a) Estimez la *taille* du domaine considéré, autrement dit le nombre d'élèves de la population qui ont réussi le test d'évaluation en mathématiques.
(Arrondissez votre réponse à l'entier le plus proche.)

b) Estimez le *total* des notes obtenues par les élèves qui ont réussi le test d'évaluation en mathématiques.
(Arrondissez votre réponse à l'entier le plus proche.)

c) Estimez la *moyenne* des notes obtenues par les élèves qui ont réussi le test d'évaluation en mathématiques.
(Précision : 2 décimales)

c) Dans le cas particulier du sondage PESR de taille n

Plaçons-nous à présent dans le cas particulier du sondage aléatoire simple de taille n et voyons à quoi correspondent alors les estimateurs du total et de la moyenne de y dans D que nous venons de construire dans un cadre tout à fait général.

Si l'on décide de prélever dans la population U un échantillon S de taille n par tirage PESR, les probabilités d'inclusion p_i affectées aux individus de la population sont toutes égales au taux de sondage n/N. Nous avons vu au chapitre 2 que, dans ce cas, la moyenne de la variable y dans la population U est estimée par $\bar{y}$, la moyenne des valeurs prises par y dans l'échantillon S ; quant au total de y dans la population U, il est estimé par $N\bar{y}$:

$$\hat{\mu}_y = \bar{y} \quad \text{et} \quad \hat{\tau}_y = N\bar{y}\,.$$

Voyons comment se modifient ces estimateurs quand on restreint notre étude de la variable y au domaine D.

c.1) Si la taille N_D du domaine D est inconnue

Considérons tout d'abord le cas le plus complexe : celui où la taille N_D du domaine D est inconnue.

L'estimateur de N_D prend la forme suivante :

$$\widehat{N}_D = \sum_{i\in S_D} \frac{1}{p_i} = \sum_{i\in S_D} \frac{1}{n/N} = N\frac{n_D}{n},$$

où le rapport $\frac{n_D}{n}$ correspond à la proportion d'individus de l'échantillon S qui appartiennent au domaine D. Notez que, puisque $N_D = N\pi_D$, ce rapport $\frac{n_D}{n}$ est un estimateur de l'importance relative du domaine D dans la population U :

$$\hat{\pi}_D = \frac{n_D}{n}.$$

L'estimateur du total de y dans D est, quant à lui, égal à :

$$\hat{\tau}_{y|D} = \sum_{i\in S_D} \frac{y_i}{p_i} = \sum_{i\in S_D} \frac{y_i}{n/N} = \frac{N}{n}\sum_{i\in S_D} y_i = N\frac{n_D}{n}\cdot\frac{1}{n_D}\sum_{i\in S_D} y_i = N\frac{n_D}{n}\bar{y}_D$$

où

$$\bar{y}_D = \frac{1}{n_D}\sum_{i\in S_D} y_i \,.$$

Dès lors, si N_D est inconnu, on estime la moyenne de y dans D via

$$\hat{\bar{\mu}}_{y|D} = \frac{\hat{\tau}_{y|D}}{\widehat{N}_D} = \frac{N\frac{n_D}{n}\bar{y}_D}{N\frac{n_D}{n}} = \bar{y}_D \,;$$

l'estimateur de la moyenne de la variable y dans le domaine D coïncide tout simplement avec la moyenne des valeurs de la variable y dans le sous-échantillon S_D. Il s'agit là de l'estimateur le plus naturel auquel on pouvait penser !

En conclusion, si N_D est inconnu, les estimateurs de la moyenne et du total de la variable y dans le domaine D sont, respectivement :

$$\hat{\bar{\mu}}_{y|D} = \bar{y}_D \quad \text{et} \quad \hat{\tau}_{y|D} = \widehat{N}_D\bar{y}_D \,.$$

c.2) Si la taille N_D du domaine D est connue

Si la taille N_D du domaine D est connue, on pourrait très bien décider d'estimer le total de y dans D via l'estimateur de Horvitz-Thompson que nous avons déjà construit, $\hat{\tau}_{y|D} = N\frac{n_D}{n}\bar{y}_D = \widehat{N}_D\bar{y}_D$, puis estimer la moyenne de y dans D via $\hat{\mu}_{y|D} = \frac{\hat{\tau}_{y|D}}{N_D} = \frac{\widehat{N}_D}{N_D}\bar{y}_D$.

Mais ces estimateurs n'apparaissent pas très naturels. Par analogie aux estimateurs considérés dans le cas où N_D est inconnu, on préférera plutôt estimer la moyenne et le total de y dans D via les estimateurs

$$\tilde{\mu}_{y|D} = \bar{y}_D \quad \text{et} \quad \tilde{\tau}_{y|D} = N_D\bar{y}_D \,.$$

c.3) Propriétés de ces estimateurs

Quelles sont les propriétés de ces estimateurs ? On trouve dans la littérature de nombreux résultats relatifs aux propriétés des estimateurs de la moyenne et du total de la variable y dans le domaine D, même dans le cadre général où l'échantillon S est prélevé dans la population par tirage PISR. Ces résultats découlent du fait que les estimateurs de $\tau_{y|D}$ et N_D sont les estimateurs de Horvitz-Thompson des totaux de la variable v et de la variable z, respectivement, dans la population U, et que l'estimateur de $\mu_{y|D}$ est obtenu en faisant le rapport de l'estimateur de $\tau_{y|D}$ sur N_D ou sur l'estimateur de N_D.

Nous allons nous contenter ici de citer quelques propriétés importantes des estimateurs du total et de la moyenne de y dans D, dans le cas particulier où l'on réalise un sondage PESR de taille n dans la population U.

On montre tout d'abord que ces estimateurs sont tous *sans biais*, et cela que la taille N_D du domaine D soit connue ou non.

Quant à leur précision, on retiendra essentiellement que la précision de l'estimateur de la moyenne ou d'un total sur un domaine D varie en première approximation comme l'inverse de la taille *attendue* de l'échantillon dans le domaine, autrement dit comme

l'inverse de l'espérance de la taille n_D de l'échantillon S_D. Or, cette espérance de n_D est égale, nous l'avons vu, à n fois l'importance relative π_D du domaine D (et si π_D est une proportion très faible, son inverse $1/\pi_D$ a une valeur fort élevée !). Ceci explique la difficulté qu'il y a à estimer avec une précision acceptable une moyenne ou un total sur un domaine D de faible taille par rapport à la taille de la population U tout entière. Si D est ce que l'on appelle un « petit domaine », la précision des estimateurs que nous avons introduits jusqu'à présent peut réellement s'avérer exécrable !

Exercice 6.3

La population qui nous intéresse est constituée des 825 prescriptions rédigées par un médecin au cours des quatre derniers mois. On désire estimer la proportion de prescriptions contenant l'indication de la prise d'un antibiotique parmi les prescriptions destinées à des enfants âgés d'au plus 15 ans.

Pour ce faire, on dispose d'un échantillon constitué de 200 prescriptions prélevées par sondage PESR dans la population des prescriptions considérée. La feuille « Dom2 » du fichier « Domaines_ex.xlsx » vous indique, pour chaque prescription de l'échantillon, si celle-ci était destinée ($z_i = 1$) ou non ($z_i = 0$) à un enfant âgé d'au plus 15 ans, et si celle-ci comportait l'indication d'un antibiotique ($y_i = 1$) ou non ($y_i = 0$).

a) Estimez le *nombre* de prescriptions destinées à des enfants âgés d'au plus 15 ans dans l'ensemble des 825 prescriptions considérées.
(Arrondissez votre réponse à l'entier le plus proche.)

b) Estimez la *proportion* de prescriptions comportant l'indication d'un antibiotique parmi celles destinées à des enfants âgés d'au plus 15 ans.
(Précision : 2 décimales.)

6.3.3 Le cas des petits domaines

Et que faire lorsqu'on s'intéresse à un « petit domaine », c'est-à-dire à une sous-population rare ?

Imaginons par exemple que, dans le cadre d'une enquête dont l'échantillon provient d'un sondage aléatoire simple réalisé dans la population d'une certaine région d'un pays, on veuille estimer le taux d'activité dans une toute petite zone rurale donnée : la variable d'intérêt y considérée ici est la variable qui prend la valeur 1 ou la valeur 0 selon que l'individu a ou n'a pas une activité professionnelle ; le domaine D est l'ensemble des habitants de la zone rurale particulière à laquelle on s'intéresse ; le taux d'activité que l'on souhaite estimer coïncide alors avec la moyenne de y dans D. Si la taille N_D du domaine est fort petite par rapport à la taille N de la population ciblée par le sondage, autrement dit si l'importance relative π_D du domaine est très faible, le risque est grand que le nombre n_D d'habitants de la zone rurale qui se trouvent dans l'échantillon S soit particulièrement limité (si pas nul !). Dans ce contexte, l'estimateur du taux d'activité des habitants de la zone rurale est le taux d'activité $\bar{y}_D$ dans le tout petit sous-échantillon S_D constitué des individus de l'échantillon S qui viennent de la zone en question ; puisque sa variance est inversement proportionnelle à la valeur

attendue pour n_D , c'est-à-dire à $n\pi_D$, $\bar{y}_D$ va souffrir d'une précision vraiment catastrophique !

Comment peut-on contourner cette difficulté ? On peut penser à faire appel à un *modèle de comportement* : on fait l'hypothèse que le taux d'activité $\mu_{y|D}$ dans la petite zone rurale considérée est identique au taux d'activité μ_y dans l'ensemble de la région sondée (ou, du moins, est identique au taux d'activité dans l'ensemble, d'importance relative non négligeable, D^* constitué de toutes les zones géographiques de la région dont la population est comparable à celle de la zone rurale qui nous intéresse). Selon l'hypothèse formulée, on débouche sur une estimation de $\mu_{y|D}$ — appelée *estimation synthétique* — du type : $\hat{\mu}_{y|D;\text{synth}} = \bar{y}$, où $\bar{y}$ est le taux d'activité dans l'ensemble de l'échantillon régional S de taille n ; ou encore, $\hat{\mu}_{y|D;\text{synth}} = \bar{y}_{D^*}$, où $\bar{y}_{D^*}$ est le taux d'activité dans le sous-échantillon S_{D^*} constitué des individus de l'échantillon S qui proviennent du domaine D^*.

Qu'y gagne-t-on ? La précision est bien meilleure, puisque la variance de $\bar{y}$ varie en $1/n$ ou celle de $\bar{y}_{D^*}$ varie en $1/\mathrm{E}(n_{D^*})$, c'est-à-dire en 1 sur n fois l'importance relative du domaine D^*, alors que la variance de $\bar{y}_D$ varie en 1 sur n fois l'importance relative du domaine D et est donc nettement plus élevée ! En effet, n ou $n\pi_{D^*}$ sont bien plus élevés que $n\pi_D$.

Qu'y perd-t-on ? L'estimateur synthétique risque d'être biaisé si on se « trompe » dans l'hypothèse de comportement formulée : le biais de $\bar{y}$ est en fait égal à la différence entre μ_y et $\mu_{y|D}$; celui de $\bar{y}_{D^*}$ correspond à la différence entre $\mu_{y|D^*}$ et $\mu_{y|D}$:

$$\mathrm{B}(\bar{y}) = \mathrm{E}(\bar{y}) - \mu_{y|D} = \mu_y - \mu_{y|D} \quad et \quad \mathrm{B}(\bar{y}_{D^*}) = \mathrm{E}(\bar{y}_{D^*}) - \mu_{y|D} = \mu_{y|D^*} - \mu_{y|D} \,.$$

Faire appel à un estimateur synthétique revient donc à déplacer le risque d'une variance élevée vers le risque d'un biais dû à une « mauvaise » hypothèse de comportement faite sur la population du domaine étudié. On espère toutefois que cette hypothèse de comportement est suffisamment valide pour que le biais éventuellement introduit soit de faible amplitude et ne contrecarre pas la nette amélioration globale de la précision de l'estimateur.

On peut construire d'autres estimateurs synthétiques ou encore d'autres types d'estimateurs censés estimer les paramètres d'un petit domaine de manière plus précise que les estimateurs directs. Ces dernières années ont été le témoin de très nombreuses avancées théoriques dans l'étude des « petits » domaines !

Chapitre 7
Méthodes classiques de redressement et de calage

7.1 Introduction générale au redressement et au calage

Les redressements d'estimations sont très fréquemment réalisés par les praticiens des sondages. Ce chapitre a pour objectif d'en expliquer l'objectif, le principe de base et de présenter les méthodes de redressement les plus simples ou les plus fréquemment utilisées.

7.1.1 L'utilisation de l'information auxiliaire

Avant de réaliser un sondage en vue d'estimer l'un ou l'autre paramètre relatif à la variable d'intérêt y, il est bon de faire le point sur toute l'information auxiliaire dont on dispose sur la population ciblée.

Dans certains cas, la base de sondage elle-même nous fournit la valeur ou la modalité d'une *variable auxiliaire* x, quantitative ou qualitative, pour chaque individu de la population. Ainsi, par exemple, une base de sondage de logements indiquera très souvent le nombre de pièces de chaque logement (ce nombre de pièces ayant, par exemple, été relevé au cours du dernier recensement des logements de la population) ; une base de sondage d'individus contient fréquemment à la fois l'âge et le sexe de chaque individu ; une base de sondage d'entreprises indique dans certains cas le secteur de l'activité principale de chaque entreprise. Si la variable auxiliaire x est étroitement liée à la variable d'intérêt y, nous pouvons judicieusement tirer parti de notre connaissance précise de la distribution de x dans la population pour mettre au point un plan de sondage efficace. Lorsque la variable x est qualitative, elle peut, par exemple, nous permettre de stratifier la population ; lorsque la variable x est quantitative, on peut penser à faire appel à ses valeurs pour déterminer les probabilités d'inclusion des individus dans le cadre d'un sondage PPS.

Mais on rencontre aussi des situations où notre connaissance de la ou des variables auxiliaires est moins fine. On ne connaît pas la valeur que prend la variable quantitative x sur chaque individu de la population, mais on connaît le total de cette variable dans la population ; si la variable x est qualitative, autrement dit que ses modalités définissent différentes catégories dans la population, on ne sait pas *a priori* à quelle catégorie particulière appartient chaque individu de la population, mais on connaît en revanche le nombre ou la proportion d'individus de la population dans chacune des catégories. Si ce type d'information auxiliaire, plus agrégée, n'est pas suffisamment riche que pour pouvoir être prise en compte dans l'élaboration du plan de sondage, elle peut par contre être prise en considération *a posteriori* et participer à une correction — on dira un *redressement* — de l'estimateur classique de Horvitz-Thompson en vue d'obtenir une « meilleure » estimation de la moyenne ou du total de la variable y dans la population.

De manière générale, si l'on dispose de plusieurs variables auxiliaires, on peut, selon la connaissance que l'on en a, utiliser certaines de ces variables pour améliorer le plan de

sondage et en utiliser d'autres pour améliorer l'estimation fournie par l'estimateur de Horvitz-Thompson.

7.1.2 Le principe général du redressement et du calage

Rappelons-nous tout d'abord l'expression générale de l'estimateur de Horvitz-Thompson du total de la variable y dans la population :

$$\hat{\tau}_y = \sum_{i \in S} \frac{y_i}{p_i} = \sum_{i \in S} d_i y_i \quad \text{où} \quad d_i = \frac{1}{p_i}.$$

On le voit, cet estimateur peut s'exprimer sous la forme d'une somme pondérée des valeurs observées pour la variable y sur les individus i de l'échantillon S, le poids affecté à la valeur y_i correspondant à l'inverse de la probabilité d'inclusion p_i de l'individu i. Ce poids, que nous désignerons par d_i, est appelé le *poids de sondage* associé à l'unité i de l'échantillon.

L'idée est assez naturelle. Pour passer de l'échantillon à la population, et plus spécifiquement pour estimer le total de y dans la population à partir des valeurs observées pour y dans l'échantillon, on fait en quelque sorte « comme si » l'individu i de l'échantillon représentait d_i individus de la population. Le poids de sondage de chaque individu i de l'échantillon est directement lié à la probabilité d'inclusion de cet individu et est donc déterminé par le plan de sondage que l'on s'est fixé.

Supposons à présent qu'il existe une variable x dont on connaît le total τ_x dans la population et qui soit étroitement liée à la variable d'intérêt y. Comment pourrions-nous utiliser cette information auxiliaire pour améliorer l'estimation du total de y fournie par $\hat{\tau}_y$? La solution proposée est assez naturelle : nous pourrions exploiter cette information auxiliaire pour venir corriger (ou « redresser ») légèrement les poids de sondage d_i associés aux individus i de l'échantillon S et définir ainsi des poids *redressés* d_i^* de manière à ce que l'estimateur du type Horvitz-Thompson de τ_x défini à partir de ces poids redressés estime parfaitement, sans aucune erreur d'échantillonnage, la valeur connue de τ_x. En d'autres termes, on va transformer légèrement les poids de sondage initiaux d_i en poids redressés d_i^* tels que

$$\hat{\tau}_x^* = \sum_{i \in S} d_i^* x_i = \tau_x$$

quel que soit l'échantillon S sélectionné. Les poids redressés sont donc définis de telle sorte à assurer ce que l'on appelle un « calage » sur le total connu de la variable x dans la population.

Cette idée se fonde en réalité sur le pari suivant : puisque les variables x et y sont étroitement liées l'une à l'autre, si les poids redressés d_i^* affectés aux individus de l'échantillon permettent d'estimer parfaitement le total de x dans la population, ils devraient aussi permettre de très bien estimer le total de y dans la population ; $\hat{\tau}_y^* = \sum_{i \in S} d_i^* y_i$ devrait nous fournir une estimation de τ_y fort proche de la valeur exacte de τ_y.

Outre le fait que le redressement (par calage) des poids de sondage vise à améliorer notre estimation de τ_y, il peut également avoir pour objectif de réduire les différences entre diverses sources d'estimations, de corriger le problème de la sur-représentation ou de la sous-représentation de certaines catégories d'individus dans l'échantillon, ou encore, de participer à la correction de la non-réponse totale (nous évoquerons ce point au cours du chapitre 9 de ce cours).

La manière avec laquelle on vient redresser les poids de sondage diffère selon le nombre de variables auxiliaires considérées, et selon que la ou les variables auxiliaires considérées sont quantitatives ou qualitatives. Lorsque les variables auxiliaires qui interviennent sont quantitatives, il faudra également tenir compte de la nature de la relation qui lie ces variables auxiliaires à la variable d'intérêt y. On distingue ainsi :

- l'*estimation par le quotient* (ou *par le ratio*) ou l'*estimation par la régression*, par exemple, lorsqu'on considère *une* variable auxiliaire quantitative dont on connaît le total dans la population ;
- l'*estimation par régression généralisée* lorsqu'on considère plusieurs variables auxiliaires quantitatives dont on connaît les totaux dans la population ;
- la *stratification a posteriori* (ou *post-stratification*) lorsqu'on dispose d'une variable auxiliaire qualitative dont on connaît la distribution dans la population ;
- le *redressement sur critères multiples* dans le cas de plusieurs variables auxiliaires qualitatives dont on ne connaît que les distributions marginales dans la population.

Mon objectif ici n'est pas de vous présenter ces méthodes de façon exhaustive mais plutôt de vous montrer le principe de leur fonctionnement et leur intérêt. Je m'arrêterai essentiellement aux deux méthodes de redressement les plus simples : l'estimation par ratio et la post-stratification. Ceux d'entre vous qui le souhaitent trouveront quelques éléments d'information supplémentaires sur les méthodes de redressement dans les annexes techniques proposées à la fin de ce chapitre.

7.2 L'estimation par le ratio (ou par le quotient)

Considérons le problème de l'estimation de τ_y, le total de la variable $\mathcal{Y}$ dans la population, à partir d'un échantillon aléatoire S prélevé selon un certain plan de sondage. Comme nous l'avons déjà rappelé dans l'introduction de ce chapitre, la démarche classique d'estimation veut que l'on fasse appel à l'*estimateur de Horvitz-Thompson* de τ_y, autrement dit à l'estimateur

$$\hat{\tau}_y = \sum_{i \in S} d_i y_i$$

où les *poids de sondage* d_i se définissent comme suit :

$$d_i = \frac{1}{p_i} = \frac{1}{\mathrm{P}(i \in S)}.$$

Supposons à présent qu'il existe une variable auxiliaire quantitative $\mathcal{X}$ dont on connaît le total τ_x dans la population et qui soit, en bonne approximation, liée à la variable d'intérêt $\mathcal{Y}$ par un lien de proportionnalité. Il existe donc une constante β telle que

$$\mathcal{Y} \simeq \beta \mathcal{X},$$

autrement dit telle que, pour tout individu i de la population U,

$$y_i \simeq \beta x_i \,.$$

Ceci implique que le total de $\mathcal{Y}$ dans la population est lui aussi approximativement égal à β fois le total de $\mathcal{X}$ dans la population :

$$\tau_y \simeq \beta \tau_x \,.$$

Mais alors, puisque le total τ_x de la variable $\mathcal{X}$ dans la population est connu, il est naturel de penser à estimer τ_y en prenant un estimateur $\hat{\beta}$ de la constante de proportionnalité β, multiplié par τ_x. C'est ce nouvel estimateur de τ_y que nous appellerons l'*estimateur par le quotient* (ou *par le ratio*) de τ_y:

$$\hat{\tau}_{y;\mathrm{quot}} = \hat{\beta} \tau_x \,.$$

Que prendre comme estimateur $\hat{\beta}$ de β ? Puisque τ_y est approximativement égal à β fois τ_x, β est approximativement égal au rapport du total de $\mathcal{Y}$ sur le total de $\mathcal{X}$. On peut dès lors décider de prendre $\hat{\beta}$ égal au rapport des estimateurs de Horvitz-Thompson de τ_y et de τ_x :

$$\hat{\beta} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$$

où

$$\hat{\tau}_y = \sum_{i \in S} d_i y_i \quad \text{et} \quad \hat{\tau}_x = \sum_{i \in S} d_i x_i \,.$$

L'estimateur par le quotient de τ_y peut donc se réécrire sous la forme suivante :

$$\hat{\tau}_{y;\mathrm{quot}} = \frac{\hat{\tau}_y}{\hat{\tau}_x} \tau_x$$

ou encore

$$\hat{\tau}_{y;\text{quot}} = \hat{\tau}_y \frac{\tau_x}{\hat{\tau}_x}.$$

La dénomination de cet estimateur se justifie aisément : on parle d'estimateur *par le quotient ou par le ratio* puisque sa construction se base sur l'estimation de la constante de proportionnalité β, autrement dit sur l'estimation du ratio de totaux τ_y/τ_x.

L'estimateur par le quotient de τ_y est obtenu en « redressant » l'estimateur de Horvitz-Thompson $\hat{\tau}_y$ de τ_y : en effet, on vient légèrement corriger cet estimateur de Horvitz-Thompson en le multipliant par le rapport entre la valeur exacte et connue du total de $\mathcal{X}$ dans la population et l'estimation (de Horvitz-Thompson) qu'on a pu en faire dans l'échantillon S. Il y a bien redressement des poids de sondage :

$$\hat{\tau}_y = \sum_{i \in S} d_i y_i ,$$

tandis que

$$\hat{\tau}_{y;\text{quot}} = \sum_{i \in S} d_i^* y_i \quad \text{avec} \quad d_i^* = d_i \frac{\tau_x}{\hat{\tau}_x}.$$

Notez que la correction apportée à l'estimateur de Horvitz-Thompson $\hat{\tau}_y$ est généralement assez légère, car le rapport de τ_x sur $\hat{\tau}_x$ est généralement assez peu différent de la valeur 1. Tout dépend bien sûr de la qualité de l'estimation $\hat{\tau}_x$ obtenue dans l'échantillon S pour le total τ_x. Par ailleurs, l'idée de cette correction apparaît assez naturelle dès lors que la variable $\mathcal{Y}$ est effectivement approximativement proportionnelle à $\mathcal{X}$: si l'échantillon S fournit une estimation $\hat{\tau}_x$ inférieure à la valeur exacte τ_x du total de $\mathcal{X}$, il est intéressant de venir gonfler un peu l'estimation $\hat{\tau}_y$ obtenue pour le total de $\mathcal{Y}$; inversement, si l'échantillon S fournit une estimation $\hat{\tau}_x$ supérieure à la valeur exacte τ_x du total de $\mathcal{X}$, il est intéressant de réduire quelque peu l'estimation $\hat{\tau}_y$ obtenue pour le total de $\mathcal{Y}$.

Par ailleurs, on vérifie très facilement que ce redressement particulier des poids de sondage assure bien le *calage* sur le total connu de $\mathcal{X}$ dans la population. En effet,

$$\hat{\tau}_{x;\text{quot}} = \sum_{i \in S} d_i^* x_i = \frac{\tau_x}{\hat{\tau}_x} \sum_{i \in S} d_i x_i = \frac{\tau_x}{\hat{\tau}_x} \hat{\tau}_x = \tau_x ;$$

si l'on estime le total de $\mathcal{X}$ en utilisant les poids de sondage redressés d_i^*, on retrouve la valeur exacte de τ_x, quel que soit l'échantillon S prélevé. Le redressement effectué permet ainsi d'assurer la cohérence des données par rapport à la variable $\mathcal{X}$.

Illustrons ces résultats à l'aide d'un petit exemple.

Exemple

On désire estimer le revenu annuel total τ_y des 100 propriétaires d'un certain quartier. On connaît par ailleurs la valeur totale τ_x des propriétés du quartier : τ_x est égal à 35 000 000 euros.

On prélève, par tirage PESR, un échantillon de 8 propriétaires du quartier et on mesure le revenu annuel y_i et la valeur de la propriété x_i de chaque propriétaire i.

Propriétaire i	Revenu y_i	Valeur x_i de la propriété
1	45 000	200 000
2	40 000	250 000
3	42 000	350 000
4	44 000	400 000
5	45 000	300 000
6	62 000	200 000
7	44 000	450 000
8	47 000	300 000

Puisqu'on est dans le cadre d'un sondage PESR, les 100 individus de la population ont tous la même probabilité d'inclusion égale au taux de sondage 8/100. Les poids de sondage d_i sont dès lors tous égaux à 100/8, soit à 12,5. Selon la démarche de Horvitz-Thompson, chacun des 8 propriétaires sélectionnés représente 12,5 propriétaires de la population.

Avec ces poids de sondage, on obtient comme revenu total estimé :

$$\hat{\tau}_y = (12{,}5)\sum_{i=1}^{8} y_i = 4\,612\,500,$$

et comme valeur totale estimée des propriétés :

$$\hat{\tau}_x = (12{,}5)\sum_{i=1}^{8} x_i = 30\,625\,000.$$

La valeur estimée de τ_x est quelque peu inférieure à la valeur exacte de τ_x. On va dès lors venir « gonfler » légèrement $\hat{\tau}_y$ en le multipliant par le rapport de τ_x sur $\hat{\tau}_x$. On obtient ainsi l'estimation par le quotient du revenu total :

$$\hat{\tau}_{y;\text{quot}} = \hat{\tau}_y \frac{\tau_x}{\hat{\tau}_x} = 4\,612\,500\,\frac{35\,000\,000}{30\,625\,000} = 4\,612\,500\,(1{,}14) = 5\,258\,250.$$

Les poids de sondage redressés d_i^* valent en réalité $(1{,}14)d_i = (1{,}14)(12{,}5) = 14{,}25$.

Exercice 7.1

On veut estimer le chiffre d'affaires moyen μ_y (en millions d'euros) dans une population de 10 000 entreprises. Pour cela, on décide de prélever 100 entreprises par tirage PESR.

Par ailleurs, la base de sondage nous indique, pour chaque entreprise de la population, la valeur de la variable « nombre de salariés » (variable auxiliaire $\mathcal{X}$). On en déduit que le nombre moyen μ_x de salariés par entreprise est égal à 50 dans la population d'entreprises considérée.

Vous trouverez dans la feuille « Ex1 » du fichier Redressement_ex.xlsx les valeurs du chiffre d'affaires (variable d'intérêt $\mathcal{Y}$) et du nombre de salariés (variable auxiliaire $\mathcal{X}$) relevées pour chacune des 100 entreprises de l'échantillon prélevé.

1. A quelles estimations du chiffre d'affaires moyen μ_y et du chiffre d'affaires total τ_y vous conduisent ces données, si vous n'utilisez pas l'information auxiliaire disponible ?
   $\hat{\mu}_y$ (avec une précision de 2 décimales) : ……..
   $\hat{\tau}_y$ (arrondi à l'entier le plus proche) : ……..
2. Quelles estimations du chiffre d'affaires moyen μ_y et du chiffre d'affaires total τ_y vous fournissent les estimateurs par le quotient $\hat{\mu}_{y;\text{quot}}$ et $\hat{\tau}_{y;\text{quot}}$?
   $\hat{\mu}_{y;\text{quot}}$ (avec une précision de 2 décimales) : ……..
   $\hat{\tau}_{y;\text{quot}}$ (arrondi à l'entier le plus proche) : ……..

N.B. : Vous êtes également invité à vérifier de manière empirique – par une analyse exploratoire des données relatives aux entreprises de l'échantillon – si le recours à l'estimateur par le quotient s'avère judicieux ou non.

Les propriétés statistiques de cet estimateur par le quotient de τ_y peuvent être résumées comme suit.
Premièrement, cet estimateur est *biaisé* (sa moyenne sur tous les échantillons possibles pour le plan de sondage considéré n'est pas égale à τ_y). Mais, heureusement, son biais tend vers zéro lorsque la taille *n* de l'échantillon croît. On considère qu'il est négligeable pour des échantillons de taille courante (pour *n* supérieur à 100, par exemple).

Deuxièmement, dans le cas où l'on fait appel à un sondage aléatoire simple, la variance de l'estimateur par le quotient est inférieure à la variance de l'estimateur classique (de Horvitz-Thompson) de τ_y, c'est-à-dire $N\bar{y}$, pour autant que les variables X et Y soient bien approximativement liées l'une à l'autre dans la population par un lien de proportionnalité.

Si vous le désirez, vous trouverez davantage de précisions sur ces propriétés de l'estimateur par le quotient dans l'annexe technique n° 1.

Quant aux annexes techniques 2 et 3, elles abordent d'autres problématiques liées à l'estimation par le quotient : l'estimation par le quotient dans un sondage stratifié et l'intérêt de l'estimation par le quotient avec calage sur la taille *N* de la population. Enfin, l'annexe technique 4 introduit une alternative à l'estimation par le quotient : l'estimation par le produit.

7.3 L'estimation par régression

Abordons à présent de manière très brève le principe de l'estimation par la régression.

Supposons à nouveau qu'il existe une variable auxiliaire quantitative $\mathcal{X}$ dont on connaît le total τ_x et donc — de manière équivalente, si N est connu — la moyenne μ_x dans la population. Pour l'estimation *par le quotient*, on supposait qu'il existait une relation de simple proportionnalité entre la variable d'intérêt $\mathcal{Y}$ et la variable auxiliaire $\mathcal{X}$, c'est-à-dire une relation du type $\mathcal{Y} \simeq \beta\mathcal{X}$. Pour l'estimation *par la régression*, nous allons plutôt supposer qu'il existe entre les variables $\mathcal{Y}$ et $\mathcal{X}$ une relation dite *affine*, du type $\mathcal{Y} \simeq \alpha + \beta\mathcal{X}$.

Nous avons donc, pour tout individu i de la population U,

$$y_i \simeq \alpha + \beta x_i \,,$$

ce qui implique que

$$\tau_y \simeq N\alpha + \beta\tau_x \,,$$

ou encore que

$$\mu_y \simeq \alpha + \beta\mu_x \,.$$

Plaçons-nous dans le cas du sondage aléatoire simple. Puisque μ_x est connu, on peut penser à estimer μ_y par

$$\hat{\mu}_{y;\text{reg}} = \hat{\alpha} + \hat{\beta}\mu_x \,,$$

où $\hat{\alpha}$ et $\hat{\beta}$ sont les estimateurs des coefficients de la droite de régression des moindres carrés de $\mathcal{Y}$ en $\mathcal{X}$ dans la population. Et on estimera τ_y par $\hat{\tau}_{y;\text{reg}}$, égal à N fois $\hat{\mu}_{y;\text{reg}}$.

Dans le cas particulier du sondage aléatoire simple, $\hat{\alpha}$ et $\hat{\beta}$ sont simplement les coefficients de la droite de régression des moindres carrés de $\mathcal{Y}$ en $\mathcal{X}$ dans l'échantillon S que l'on a prélevé, et $\hat{\mu}_{y;\text{reg}}$ correspond alors à l'ordonnée du point de cette droite dont l'abscisse est μ_x. On a donc :

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i\in S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in S}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \,.$$

Ceci nous permet de réécrire l'estimateur — appelé l' « estimateur par la régression » — de μ_y sous la forme suivante :

$$\hat{\mu}_{y;\text{reg}} = \bar{y} + \hat{\beta}(\mu_x - \bar{x}).$$

Cette dernière expression met clairement en évidence le fait que l'estimateur par la régression de μ_y corrige (redresse) l'estimateur classique $\bar{y}$: on rajoute à $\bar{y}$ un terme proportionnel à l'écart qui existe entre la valeur exacte de μ_x et l'estimation $\bar{x}$ qu'on a pu faire de cette valeur à partir de l'échantillon. Ce redressement assure un calage sur la moyenne connue μ_x de la variable $\mathcal{X}$. En effet, si on estime μ_x via son estimateur par la régression, on obtient, en remplaçant simplement y par x dans l'expression de $\hat{\mu}_{y;\text{reg}}$, que

$$\hat{\mu}_{x;\text{reg}} = \bar{x} + \frac{s_x^2}{s_x^2}(\mu_x - \bar{x}) = \bar{x} + \mu_x - \bar{x} = \mu_x \,.$$

L'estimateur par la régression de μ_y est biaisé, mais son biais est d'ordre $1/n$ et devient donc négligeable lorsque n est grand. Quant à la variance d'échantillonnage de $\hat{\mu}_{y;\text{reg}}$, on montre qu'elle est approximativement égale à la variance de $\bar{y}$, multipliée par $(1-\rho^2)$, où ρ est le coefficient de corrélation linéaire entre les variables Y et X dans la population :

$$\mathrm{V}\left(\hat{\mu}_{y;\text{reg}}\right) \simeq \mathrm{V}(\bar{y})(1-\rho^2)\,.$$

Puisque ρ est nécessairement compris entre -1 et $+1$, ρ^2 est nécessairement compris entre 0 et 1, et l'estimateur par la régression de μ_y a toujours une variance plus faible que l'estimateur classique $\bar{y}$.

Citons enfin, pour mémoire, l'*estimateur par la différence* (aussi appelé en audit *estimateur de la méthode des écarts*)

$$\hat{\mu}_{y;\text{diff}} = \bar{y} + (\mu_x - \bar{x}).$$

Cet estimateur est obtenu en ajoutant à $\bar{y}$, l'estimateur classique de μ_y, la différence constatée dans l'échantillon entre μ_x et $\bar{x}$. Il correspond en fait à l'estimateur par la régression pour lequel la valeur de β est choisie *a priori* égale à 1 dans la relation supposée lier Y à X dans la population.

Exercice 7.2

Le directeur d'une entreprise de confection de chaussures souhaite estimer la longueur moyenne des pieds droits des hommes adultes d'une ville. Soient Y la variable « longueur du pied droit » (en centimètres) et X la variable « taille de l'individu » (également en centimètres). Le directeur sait en outre, par les résultats d'une vaste enquête de santé menée récemment, que la taille moyenne des hommes adultes de cette ville est de l'ordre de 168 cm.

Pour estimer la moyenne qui l'intéresse, le directeur prélève un échantillon de 100 hommes adultes ; nous allons faire comme si cet échantillon avait été prélevé par sondage aléatoire simple. Les valeurs de X et de Y relevées sur les individus de cet échantillon sont disponibles dans la feuille « Ex2 » du fichier Redressement_ex.xlsx.

a) Calculez l'estimateur de Horvitz-Thompson, l'estimateur par le quotient, l'estimateur par la différence et l'estimateur par la régression de μ_y.

$\hat{\mu}_y$ (avec une précision de 2 décimales) : ……

$\hat{\mu}_{y;\text{quot}}$ (avec une précision de 2 décimales) : ……

$\hat{\mu}_{y;\text{diff}}$ (avec une précision de 2 décimales) : ……

$\hat{\mu}_{y;\text{reg}}$ (avec une précision de 2 décimales) : ……

b) Parmi les 4 estimations calculées, laquelle conseilleriez-vous au directeur ? (Pour répondre à cette question, vous pouvez vous appuyer sur une analyse exploratoire des données relatives aux individus de l'échantillon.)

- l'estimation de Horvitz-Thompson
- l'estimation par le quotient
- l'estimation par la différence
- l'estimation par la régression

Les estimateurs par le quotient, par la régression et par la différence se généralisent au cas de plusieurs variables auxiliaires et de plans de sondage plus complexes que le plan PESR. Si ces méthodes de redressement vous intéressent, vous pouvez consulter les ouvrages de Ardilly (2006) et Särndal *et al.* (1992)[1].

[1] Särndal, C.E., B. Swensson, and J. Wretman (1992), *Model assisted survey sampling*, Springer-Verlag.

7.4 La post-stratification ou stratification *a posteriori*

Les méthodes de redressement que nous avons étudiées jusqu'à présent tiraient parti de la connaissance que nous avions du total ou de la moyenne, dans la population, d'une ou plusieurs variables *quantitatives*. Nous allons à présent nous intéresser à une méthode de redressement, très simple et très fréquemment mise en œuvre, qui se fonde sur notre connaissance de la distribution d'une certaine variable *qualitative* dans la population.

Considérons une variable auxiliaire qualitative $\mathcal{X}$ dont les H modalités définissent H strates $U_1, U_2, \cdots, U_H$ dans la population. Nous connaissons les effectifs $N_1, N_2, \cdots, N_H$ de ces strates ou leurs importances relatives $\frac{N_1}{N}, \frac{N_2}{N}, \cdots, \frac{N_H}{N}$, grâce, par exemple, aux données du dernier recensement, à des statistiques professionnelles ou administratives, etc. Mais la base de sondage dont nous disposons pour la population U ne nous indique pas à quelle strate appartient chaque individu de la population : il nous est donc impossible de mettre en place un sondage stratifié.

Supposons que l'on décide de prélever un échantillon S de taille n par sondage aléatoire simple : les estimateurs classiques de la moyenne et du total de la variable $\mathcal{Y}$ dans la population sont alors

$$\hat{\mu}_y = \bar{y} \quad \text{et} \quad \hat{\tau}_y = N\bar{y}\,,$$

où $\bar{y}$ est la moyenne des valeurs observées pour $\mathcal{Y}$ sur les individus qui appartiennent à S. Mais ne pourrait-on pas améliorer l'estimation de μ_y et de τ_y en utilisant *a posteriori* la stratification de la population et notre connaissance de N_h ou de N_h/N pour tout $h = 1, \cdots, H$?

En demandant à chaque individu de l'échantillon S à quelle strate U_h il appartient, on peut stratifier S *a posteriori*, autrement dit partitionner S en sous-échantillons $S_h = S \cap U_h$ $(h = 1, \ldots, H)$. Attention ! Chaque sous-échantillon S_h ainsi défini est de taille n_h *aléatoire*, puisqu'on ne peut pas déterminer à l'avance avec certitude combien d'individus de l'échantillon S proviendront de la strate n° h. La situation envisagée ici est donc différente de celle rencontrée dans le cadre du sondage stratifié où chaque sous-échantillon S_h, prélevé par tirage PESR dans la strate U_h, est de taille n_h fixée *a priori*.

On peut ensuite déterminer la moyenne $\bar{y}_h$ des valeurs prises par la variable $\mathcal{Y}$ dans chaque sous-échantillon S_h :

$$\bar{y}_h = \frac{1}{n_h} \sum_{i \in S_h} y_i \ ;$$

$\bar{y}_h$ est un estimateur de $\mu_{y|h}$, la moyenne de $\mathcal{Y}$ dans la strate U_h.
Mais alors, puisque

$$\mu_y = \sum_{h=1}^{H} \frac{N_h}{N} \mu_{y|h} ,$$

on peut définir l'estimateur dit « post-stratifié » de μ_y comme suit :

$$\hat{\mu}_{y;\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h .$$

Dans ce cadre, l'estimateur post-stratifié du total τ_y est égal à N fois l'estimateur post-stratifié de la moyenne μ_y :

$$\hat{\tau}_{y;\text{post}} = N\hat{\mu}_{y;\text{post}} = \sum_{h=1}^{H} N_h \bar{y}_h .$$

Illustrons cette procédure de stratification *a posteriori* à l'aide d'un petit exemple.

Exemple

Considérons une enquête sur le revenu mensuel des individus d'une certaine population, où l'on décide de post-stratifier sur une variable « tranche d'âges ». Le choix d'une telle variable auxiliaire est guidé par le fait qu'il existe bien évidemment une forte association entre l'âge et le revenu.

La base de sondage dont on dispose ne nous indique pas l'âge de chaque individu de la population. Cependant, les résultats du dernier recensement nous indiquent que les tranches d'âges que nous désirons considérer se répartissent dans la population comme suit :

	< 20 ans	21-35 ans	36-50 ans	> 50 ans
N_h/N	20%	35%	30%	15%

Les moins de 20 ans constituent 20% de la population étudiée. Les individus de 21 à 35 ans représentent 35% de la population, les 36-50 ans constituent 30% de la population, et les plus de 50 ans, 15% de la population.

On tire l'échantillon par sondage PESR et on y interroge les individus sur leur âge ; on obtient les parts n_h/n suivantes pour les différentes tranches d'âges :

	< 20 ans	21-35 ans	36-50 ans	> 50 ans
n_h/n	15%	30%	30%	25%

On le voit, les caprices du hasard ont fait que les individus des deux premières tranches d'âges de la population sont légèrement sous-représentés dans l'échantillon, contrairement aux individus les plus âgés (de plus de 50 ans) qui sont quelque peu surreprésentés dans l'échantillon.

Les individus de l'échantillon sont également interrogés sur leur revenu mensuel. Les réponses obtenues nous conduisent aux revenus mensuels moyens par tranche d'âges suivants :

	< 20 ans	21-35 ans	36-50 ans	> 50 ans
N_h/N	20%	35%	30%	15%
n_h/n	15%	30%	30%	25%
$\bar{y}_h$	900	1 350	2 250	1 800

Si on ne redresse pas sur l'âge, on estime le revenu mensuel moyen dans la population par :

$$\bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = (0.15)900 + (0.30)1\,350 + (0.30)2\,250 + (0.25)1\,800 = 1\,665\,.$$

Si on redresse selon la tranche d'âges, on estime alors le revenu mensuel moyen dans la population par :

$$\hat{\mu}_{y;\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = (0.20)900 + (0.35)1\,350 + (0.30)2\,250 + (0.15)1\,800 = 1\,597{,}5.$$

En l'absence de post-stratification, le revenu mensuel moyen est donc estimé à un montant plus élevé. Ceci est dû au fait que, sous l'effet du « hasard », l'échantillon comprend « trop » de personnes de plus de 50 ans et « pas assez » de personnes d'au plus 35 ans. Or, les personnes de la dernière tranche d'âges ont un revenu mensuel moyen plus élevé que celles des deux premières tranches d'âges. Dès lors, le petit déséquilibre de l'échantillon par rapport à la population tire la moyenne générale des revenus $\bar{y}$ vers une valeur trop élevée.

Intuitivement, on peut penser que l'estimation par post-stratification sera meilleure que celle fournie par $\bar{y}$ puisque, dans la post-stratification, on tient compte de la vraie répartition des tranches d'âges dans la population, et non de la répartition à laquelle le hasard nous a conduits dans l'échantillon.

Les estimateurs *post-stratifiés* de μ_y et τ_y ont exactement la même expression que les estimateurs *stratifiés* de ces deux paramètres. Ils n'ont cependant pas les mêmes propriétés car, ici, on ne prend en compte la stratification de la population qu'*a posteriori*, une fois l'échantillon aléatoire simple S prélevé. Mais quelles sont ces propriétés ?

(i) L'estimateur classique $\hat{\tau}_y = N\bar{y}$ est l'estimateur de Horvitz-Thompson de τ_y, construit à partir des poids de sondage $d_i = \frac{1}{p_i} = \frac{N}{n}$, identiques pour tous les individus de l'échantillon S. L'estimateur post-stratifié de τ_y fait quant à lui intervenir des **poids de sondage redressés** : il donne à tout individu i du sous-échantillon S_h un poids

$$d^*_{i,h} = d_i \frac{(n/N)}{(n_h/N_h)} = \frac{N_h}{n_h}.$$

Vous trouverez la petite démonstration de ce résultat dans l'annexe technique à la fin de ce chapitre.

Le poids de sondage d_i d'un individu i de l'échantillon S qui appartient à la strate n° h est donc redressé, corrigé par le rapport entre le taux de sondage n/N appliqué dans la population U et le taux de sondage n_h/N_h qui en résulte dans la strate n° h.

Le facteur de correction $\frac{(n/N)}{(n_h/N_h)}$ est supérieur à 1 si le rapport $\frac{n_h}{N_h}$ est inférieur au rapport $\frac{n}{N}$, c'est-à-dire si $\frac{n_h}{n} < \frac{N_h}{N}$: la post-stratification vient donc gonfler le poids de sondage d'un individu de la strate h lorsque cette strate apparaît quelque peu sous-représentée dans l'échantillon. Inversement, le facteur de correction $\frac{(n/N)}{(n_h/N_h)}$ est inférieur à 1 si le rapport $\frac{n_h}{N_h}$ est supérieur au rapport $\frac{n}{N}$, c'est-à-dire si $\frac{n_h}{n} > \frac{N_h}{N}$: la post-stratification diminue le poids de sondage d'un individu de la strate h lorsque cette strate est surreprésentée dans l'échantillon.

Notez que si la taille n de l'échantillon aléatoire simple prélevé est relativement grande et si le hasard ne fait pas de « caprices » particuliers au cours du sondage, les rapports n_h/N_h ne devraient pas être très différents du taux de sondage n/N appliqué dans la population ; les poids de sondage redressés intervenant dans l'estimateur post-stratifié de τ_y ne devraient donc pas être fort différents des poids de sondage initiaux.

Enfin, le redressement effectué dans le cadre de la post-stratification assure un **calage sur les tailles connues des différentes strates**. La vérification de ce résultat est faite dans l'annexe technique 5 en fin de chapitre.

(ii) Que peut-on dire du **biais** des estimateurs post-stratifiés ? On montre que, si $n_h > 0$ pour tout h (autrement dit si chaque strate est représentée par au moins un individu dans l'échantillon S), l'estimateur post-stratifié de μ_y ou de τ_y est exactement sans biais :

$$\mathrm{E}\left(\hat{\mu}_{y;\mathrm{post}}\right) = \mu_y \quad \text{et} \quad \mathrm{E}\left(\hat{\tau}_{y;\mathrm{post}}\right) = \tau_y \,.$$

Mais attention ! Cette propriété de non-biais ne vaut que si les effectifs N_h ou les rapports N_h/N intervenant dans l'expression de $\hat{\tau}_{y;\mathrm{post}}$ et de $\hat{\mu}_{y;\mathrm{post}}$ correspondent bien aux valeurs exactes des tailles et des importances relatives des différentes strates dans la population. Si, par exemple, les vraies valeurs des fréquences associées aux strates sont N_h^*/N^* au lieu de N_h/N, $\hat{\mu}_{y;\mathrm{post}}$ est alors biaisé et son biais vaut :

$$\mathrm{B}\left(\hat{\mu}_{y;\mathrm{post}}\right) = \mathrm{E}\left(\hat{\mu}_{y;\mathrm{post}}\right) - \mu_y = \sum_{h=1}^{H} \left(\frac{N_h}{N} - \frac{N_h^*}{N^*}\right) \mu_{y|h} \,.$$

Il est donc essentiel que les effectifs ou les fréquences des strates soient connus de manière précise et surtout récente : une stratification *a posteriori* ajustant un échantillon sur une distribution ancienne (et susceptible de s'être déformée) est à déconseiller !

(iii) Passons à présent à la **variance** de $\hat{\mu}_{y;\mathrm{post}}$. On montre qu'elle est approximativement égale à

$$\frac{(1-f)}{n} \sigma^2_{y\ \mathrm{intra;corr}} + \frac{(1-f)}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \sigma^2_{y|h;\mathrm{corr}}$$

où $\sigma^2_{y|h;\mathrm{corr}}$ est la variance corrigée de la variable y dans la strate n° h ($\sigma^2_{y|h;\mathrm{corr}} = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \mu_{y|h})^2$) et

$$\sigma^2_{y\,\text{intra;corr}} = \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{y|h;\text{corr}} .$$

L'estimateur post-stratifié ne se montrera donc efficace que si la variable auxiliaire X de stratification de la population définit de « bonnes » strates, c'est-à-dire des strates bien homogènes, à l'intérieur desquelles les valeurs de Y sont relativement peu dispersées.

(iv) Le redressement par stratification *a posteriori* est-il toujours intéressant à réaliser ? $\hat{\mu}_{y;\text{post}}$ est-il toujours plus précis que $\bar{y}$, l'estimateur classique de μ_y ?

Pour répondre à cette question, il faut comparer la variance de $\hat{\mu}_{y;\text{post}}$ à celle de $\bar{y}$. Cette comparaison est faite dans l'annexe technique 6. Que nous apprend-elle ?

La variance de $\hat{\mu}_{y;\text{post}}$ est généralement inférieure à celle de $\bar{y}$ quand les moyennes de la variable Y dans les différentes strates sont bien différentes les unes des autres et que les variances de Y dans les différentes strates sont faibles, indiquant par là que la variable de stratification X est fortement liée à la variable d'intérêt Y. Ainsi, l'estimateur post-stratifié peut être plus précis que l'estimateur classique lorsque les strates U_h ont toutes les caractéristiques pour être de « bonnes » strates, c'est-à-dire des strates avec lesquelles le sondage stratifié se serait montré très efficace.

Mais si la variable de stratification n'est pas associée à la variable d'intérêt, l'estimateur post-stratifié peut avoir une variance plus grande, donc une précision plus faible, que l'estimateur classique.

Notons toutefois que la différence entre les variances de l'estimateur post-stratifié et de l'estimateur classique est proportionnelle à $(1-f)/n$; cette différence est donc petite lorsque l'échantillon est de grande taille. Ceci s'explique par le fait que, lorsque l'échantillon aléatoire simple prélevé est de taille n relativement grande, on peut s'attendre, comme nous l'avons déjà vu, à ce que le redressement des poids de sondage soit très léger et donc à ce que l'estimateur post-stratifié et l'estimateur classique fournissent des estimations très proches l'une de l'autre.

(v) On peut enfin très facilement comparer la précision de l'estimateur post-stratifié avec celle de l'estimateur stratifié proportionnel.

Rappelons-nous que les estimateurs de μ_y et τ_y dans le cadre du sondage stratifié proportionnel coïncident avec les estimateurs classiques $\bar{y}$ et $N\bar{y}$. Par ailleurs, la variance de $\hat{\mu}_{y;\text{STP}}$ est égale à $\frac{(1-f)}{n}\sigma^2_{y,\text{intra;corr}}$. Il s'ensuit que

$$\text{V}\left(\hat{\mu}_{y;\text{post}}\right) - \text{V}\left(\hat{\mu}_{y;\text{STP}}\right) \simeq \frac{(1-f)}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \sigma^2_{y|h;\text{corr}} .$$

Cette différence est toujours supérieure ou égale à 0, nous indiquant par là que, si l'on dispose de l'information auxiliaire nécessaire, il vaut mieux utiliser la stratification au niveau du plan de sondage lui-même et mettre en œuvre un sondage stratifié proportionnel plutôt que de ne tirer parti de la stratification qu'*a posteriori*.

Notez cependant que cette différence est en $1/n^2$ et devient donc quasiment négligeable lorsque n est « relativement grand ».

Exercice 7.3

Considérons une région agricole dans laquelle on comptabilise 2 010 fermes. Nous aimerions estimer la surface que consacre en moyenne une ferme de la région à la culture céréalière (en d'autres termes, nous aimerions estimer la moyenne-population de la variable y associant à chaque ferme la surface sur laquelle elle cultive des céréales).

On sait par ailleurs que, parmi les 2 010 fermes de la région, 1 080 fermes disposent d'une surface cultivée totale de moins de 150 hectares (strate 1), 620 fermes ont une surface cultivée totale comprise entre 150 et 200 hectares (strate 2), et 310 fermes cultivent plus de 200 hectares (strate 3).

On prélève par tirage PESR 200 fermes de la région et on relève, pour chaque ferme sélectionnée, sa surface cultivée totale ainsi que sa surface cultivée en céréales. Ces relevés conduisent aux résultats suivants :

- 102 fermes de l'échantillon appartiennent à la première strate ; la moyenne des surfaces cultivées en céréales par ces 102 fermes vaut 19,40 (hectares) ;
- 64 fermes de l'échantillon appartiennent à la deuxième strate ; la moyenne des surfaces cultivées en céréales par ces 64 fermes vaut 61,63 (hectares) ;
- 34 fermes de l'échantillon appartiennent à la troisième strate ; la moyenne des surfaces cultivées en céréales par ces 34 fermes s'élève à 96,45 (hectares).

a) Quelle est l'estimation PESR de μ_y ?
(*Indiquez votre réponse avec une précision de 2 décimales.*)

b) Que vaut l'estimation post-stratifiée de μ_y ?
(*Indiquez votre réponse avec une précision de 2 décimales.*)

Exercice 7.4

L'institut de sondage SOFOP est chargé d'étudier l'audience de différents journaux et magazines. Il interroge pour cela un échantillon de taille 2 000, sélectionné selon un procédé qu'on assimilera à un plan simple à probabilités égales et sans remise, au sein de la population française des individus âgés de 15 ans et plus. On supposera qu'il n'y a pas de non-réponses.

Pour satisfaire à la demande du magazine « PARIS-STAR », les résultats concernant cette publication sont ventilés selon le critère « habitant en zone rurale – habitant en zone urbaine ». Les données recueillies se présentent donc selon le tableau suivant :

Réponses	Zone rurale	Zone urbaine	Ensemble
Ont lu au moins une fois PARIS-STAR	64	476	540
N'ont pas lu PARIS-STAR	576	884	1 460
Total	640	1 360	2 000

a) Donnez la valeur de $\hat{\pi}$, l'estimateur de la proportion de lecteurs du magazine PARIS-STAR dans la population.
(*Indiquez votre réponse avec une précision de 2 décimales.*)

b) On sait que la proportion réelle d'habitants en zone urbaine s'élève à 75%. Proposez alors une nouvelle estimation de la proportion π de lecteurs du magazine PARIS-STAR dans la population.
(*Indiquez votre réponse avec une précision de 2 décimales.*)

Bien évidemment, la méthode de redressement par post-stratification peut être étendue à d'autres plans de sondage que le sondage aléatoire simple (voir, par exemple, Ardilly (2006)).

7.5 Le redressement sur plusieurs variables auxiliaires qualitatives

Abordons un dernier point relatif aux méthodes de redressement sur variables qualitatives.

Dans bon nombre d'enquêtes, on dispose de plusieurs variables auxiliaires permettant de réaliser une stratification *a posteriori*. Par exemple, dans les enquêtes sur les individus ou sur les ménages d'un pays ou d'une région, on connaît la distribution dans la population de différentes caractéristiques socio-démographiques grâce aux résultats du recensement.

Utiliser simultanément les différentes variables auxiliaires ne pose pas de problème méthodologique pour autant que l'on connaisse la distribution des unités de la population selon les croisements des modalités de ces différentes variables : on se retrouve alors dans le cadre classique de la post-stratification.
Supposons ainsi, par exemple, que l'on veuille redresser un échantillon d'individus selon la variable « sexe » (à deux modalités : homme, femme) et selon la variable « âge » (à 4 modalités correspondant à 4 tranches d'âges particulières) : si l'on connaît les poids dans la population des 8 classes obtenues en croisant ces deux variables, on peut tout simplement faire appel à l'estimateur post-stratifié.

Certaines difficultés peuvent toutefois survenir dans la pratique. Tout d'abord, si les variables auxiliaires considérées sont nombreuses, on risque de ne trouver dans l'échantillon qu'un très faible nombre d'individus, voire même aucun individu, au croisement de certaines modalités des variables. Cela pose bien sûr un problème pour estimer la moyenne ou le total de y pour ces croisements, et cela risque d'introduire un biais dans la procédure d'estimation.
Par ailleurs, on ne connaît bien souvent que les distributions *marginales* des variables ou, au mieux les distributions des variables croisées deux à deux. Dans ce cas, on ne peut pas appliquer la stratification *a posteriori* car on ne dispose pas de toute l'information auxiliaire nécessaire pour cette post-stratification.

On se retrouve alors confronté au problème dit de « l'équilibrage d'un tableau de contingence dont on connaît les marges », ou encore, de manière équivalente, au problème de la définition de poids de sondage redressés assurant un calage sur les effectifs des distributions *marginales* des variables auxiliaires. Ainsi, par exemple, face à une population de ménages, on peut vouloir réaliser un calage d'une part sur la distribution des ménages selon leur taille, et d'autre part sur leur distribution selon le niveau d'instruction du chef de ménage.

La méthode la plus utilisée pour réaliser ce type de calage sur marges est la *méthode RAS* (Raking Adjusted Statistics), aussi appelée *méthode du raking-ratio*. L'idée de base de cette méthode est la suivante. Supposons que l'on considère deux variables auxiliaires : on commence par modifier les poids de sondage de telle sorte à assurer le calage sur la distribution marginale d'une des deux variables. Dans un deuxième temps,

on remodifie les poids pour se caler sur la distribution marginale de l'autre variable. Puis on recommence avec la première distribution pour repasser ensuite à la deuxième, etc. Après un certain nombre d'itérations, on obtient des poids redressés assurant simultanément un calage sur les deux distributions marginales.

Cette méthode peut être adoptée dans le cas d'un sondage aléatoire simple, ou dans celui d'un plan de sondage plus complexe à probabilités d'inclusion égales et donnant lieu à des échantillons de taille fixe.

7.6 Conclusion

Quelles leçons tirer de ce chapitre ?

Les méthodes de redressement sont très couramment utilisées dans la pratique. Il en existe une large variété : les méthodes varient selon le nombre de variables auxiliaires prises en considération, selon leur nature (quantitative ou qualitative), selon la richesse de l'information auxiliaire disponible, selon la nature de la relation liant, dans la population, la ou les variables auxiliaires à la variable d'intérêt. Elles tentent toutes de tirer parti, de manière naturelle, de l'information auxiliaire à disposition du sondeur.

Il faut toutefois garder à l'esprit que les méthodes de redressement ne sont justifiées et efficaces qu'à certaines conditions.

Premièrement, la ou les variables auxiliaires ou de contrôle doivent être bien associées ou corrélées avec les variables qui font l'objet de l'enquête (les variables d'intérêt).

Deuxièmement, les valeurs des moyennes, des totaux ou des effectifs sur lesquelles on réalise le calage doivent être fiables : elles doivent concerner la même population, provenir d'un recensement ou d'une très grosse enquête, ne doivent pas être périmées, etc.

Troisièmement, l'échantillon prélevé doit être de taille suffisante. Il est peu sensé de vouloir corriger un échantillon d'à peine 100 individus.

Quatrièmement, les redressements doivent être conçus comme un « lissage » des estimations ; ils ne doivent normalement modifier qu'assez légèrement les poids de sondage initiaux des individus de l'échantillon. Une modification marquée des poids de sondage peut être vue comme le signe que l'échantillon prélevé est probablement de mauvaise qualité (c'est effectivement un risque auquel nous sommes soumis à cause des éventuels caprices du hasard).

Chapitre 8
Les sondages empiriques

8.1 Avantages et inconvénients

Dans le chapitre 1 au cours duquel vous avez fait vos tout premiers pas en théorie des sondages, nous avons distingué deux grandes catégories de méthodes de sondage. Vous avez d'une part les méthodes de sondage dites *aléatoires*, dont le sondage PESR, le sondage stratifié, le sondage PISR, le sondage en grappes et le sondage à plusieurs degrés que nous avons étudiés dans les chapitres précédents sont les exemples les plus courants. Et vous avez d'autre part les méthodes de sondage *non aléatoires*, aussi appelées méthodes de sondage *empiriques* ou *à choix raisonné*.

C'est à ce deuxième type de méthodes qu'est consacré le chapitre 8.

Comparativement aux méthodes de sondage aléatoires, les méthodes empiriques présentent des avantages certains, mais également des inconvénients majeurs qu'il ne faut pas perdre de vue.

Les méthodes de sondage empiriques ont le grand avantage de pouvoir être utilisées alors qu'aucune base de sondage n'est disponible. Contrairement au sondage *aléatoire*, la sélection des individus à interroger dans un sondage empirique n'exige pas l'application d'un algorithme informatique sur une base de sondage : cette plus grande liberté dans la procédure d'échantillonnage permet souvent une exécution plus rapide du sondage et une réduction significative des coûts, comparativement aux méthodes aléatoires.

Ceci explique pourquoi les méthodes empiriques sont très fréquemment utilisées par les instituts de sondage privés pour les sondages d'opinion et les études de marché, par exemple.

Cependant, les méthodes de sondage empiriques présentent également des défauts majeurs. Tout d'abord, dans un sondage empirique recourant à l'emploi d'enquêteurs, le fait de laisser à ceux-ci une certaine liberté dans le choix des personnes à interroger peut avoir pour conséquence que l'échantillon sélectionné soit quelque peu *biaisé*. En effet, il se peut que certains enquêteurs aient tendance, même inconsciemment, à sélectionner plus facilement certaines personnes que d'autres, pour des raisons d'affinités socio-culturelles par exemple ; l'échantillon interrogé risque alors de ne pas refléter correctement toute la diversité ou l'hétérogénéité de la population ciblée.

En outre, l'application d'une méthode de sondage empirique rend impossible la mesure objective de la précision de la procédure d'estimation et le calcul rigoureux de marges d'erreur. Pourquoi donc ? Parce que la mesure objective de la précision d'un estimateur ne peut se faire que via le calcul de son espérance et de sa variance, calcul qui requiert la connaissance, au moins théorique, du plan de sondage caractérisant la méthode de sondage ou la connaissance des probabilités d'inclusion affectées aux individus de la population. Or, sans base de sondage pour la population ciblée et sans procédure parfaitement contrôlée de tirage de l'échantillon, il nous est impossible de déterminer un plan de sondage et les probabilités qu'ont *a priori* les individus de la population d'être sélectionnés.

Il faut donc être particulièrement prudent dans l'usage des méthodes de sondage empiriques. Il faut avant tout essayer de contrôler au mieux la phase de sélection de l'échantillon. Ce contrôle doit notamment veiller à limiter autant que possible le biais d'échantillonnage qui risque de découler de l'implication personnelle des enquêteurs dans le choix des personnes à interroger. Il faut en réalité tenter de « mimer » au mieux une procédure de prélèvement aléatoire au sens strict du terme, afin de pouvoir raisonnablement considérer que l'échantillon prélevé présente toutes les caractéristiques d'un échantillon qui aurait été prélevé par sondage PESR ou par sondage stratifié, par exemple. Ce n'est qu'alors qu'on pourra évaluer le niveau de précision des résultats du sondage empirique en faisant appel aux calculs de précision valables dans le cadre d'un sondage aléatoire. En pratique, dans la publication des résultats des sondages d'opinion — sondages le plus souvent empiriques — par exemple, les marges d'erreur qui sont mentionnées sont généralement celles relatives au simple sondage PESR. Ces marges d'erreur ne sont que très approximatives… et il faut garder cela à l'esprit lorsqu'on tire des conclusions des résultats de tels sondages. La prudence s'impose !

8.2 Quelques méthodes empiriques

8.2.1 La méthode des quotas

a) Le principe

La méthode de sondage empirique la plus fréquemment mise en œuvre est certainement la méthode du sondage *par quotas*. Elle consiste à faire en sorte que l'échantillon apparaisse comme un modèle réduit de la population étudiée, selon des critères dont on connaît la répartition dans la population. Autrement dit, on choisit quelques caractéristiques dont on connaît la distribution dans la population étudiée (le sexe, l'âge, la catégorie socio-professionnelle du chef de ménage, par exemple) et on sélectionne librement les personnes à interroger pourvu qu'au final, la distribution de ces caractéristiques dans l'échantillon soit similaire à leur distribution dans la population. La constitution de l'échantillon est ainsi soumise au respect de certains quotas, d'où le nom de cette méthode de sondage empirique.

Le sondage par quotas apparaît en quelque sorte comme le compétiteur, *non aléatoire*, du sondage stratifié proportionnel.

b) Le choix des quotas

Quels quotas se fixer ? Tout comme dans le cas du sondage stratifié proportionnel, les variables permettant de définir les quotas doivent être liées aux variables d'intérêt de l'enquête ; elles doivent exercer une influence sur les phénomènes étudiés.

Par ailleurs, les quotas ne doivent pas être trop nombreux et doivent être facilement identifiables par les enquêteurs en début d'interview.

Ils doivent enfin être déterminés sur la base de statistiques récentes et fiables.

Si l'on cherche à prélever un échantillon d'*individus*, les variables de quotas les plus utilisées sont :

- la région et la taille de la commune de résidence ;
- le sexe ;
- l'âge ;
- la catégorie socio-professionnelle du chef de ménage complétée par un quota sur l'activité de la personne interrogée.

Si c'est un échantillon de *ménages* qui doit être sélectionné, on prendra généralement en compte les variables de quotas suivantes :

- la région et la taille de l'agglomération ;
- l'activité socio-professionnelle du chef de ménage ;
- la taille du ménage.

Notez qu'on donne généralement à l'enquêteur non pas des quotas *croisés* (relatifs au croisement de deux ou plusieurs variables de quotas), mais plutôt des quotas *marginaux* (relatifs à chaque variable de quotas considérée séparément) : avec des enquêteurs

expérimentés, cette manière de définir les quotas diminue le temps de recherche des interviewés. Illustrons cette remarque en reprenant l'exemple présenté p. 61-62 dans l'ouvrage *Les sondages : principes et méthodes* de la collection « Que sais-je ? » des Presses Universitaires de France (A.-M. Dussaix et J.-M. Grosbras, 1993) :
Imaginons qu'un enquêteur doive réaliser 16 interviews dans une ville de la région Ouest (taille de la commune : 2 000 à 20 000 habitants), le champ de l'enquête étant les individus de 15 ans et plus. Si les caractéristiques ou *quotas* retenus sont le sexe, l'âge et la catégorie socio-professionnelle du chef de ménage, le plan de travail de l'enquêteur sera résumé dans la feuille de quotas suivante :

Feuille de quotas

Région	Ouest	
Habitat	2 000 à 20 000 habitants	
16 interviews à réaliser		
Sexe de l'interviewé		
• Homme	8	** ** ** ** ** ** ** **
• Femme	8	** ** ** ** ** ** ** **
Age de l'interviewé		
• 15 à 24 ans	3	** ** **
• 25 à 34 ans	3	** ** **
• 35 à 49 ans	4	** ** ** **
• 50 à 64 ans	3	** ** **
• 65 ans et plus	3	** ** **
CSP du chef de ménage		
• Agriculteur et salarié agricole	2	** **
• Artisan et petit commerçant	2	** **
• Industriel + Gros comm. + Prof. lib. + Cadre supérieur	1	**
• Cadre moyen+ Employé + Divers	2	** **
• Ouvriers + Personnel service	5	** ** ** ** **
• Inactifs + Retraités	4	** ** ** **

Selon cette feuille, l'enquêteur doit interviewer 8 hommes et 8 femmes ; 3 interviewés doivent avoir entre 15 et 24 ans, 3 doivent avoir entre 25 et 34 ans, etc. (l'enquêteur biffera les étoiles dans la partie droite de la feuille de quotas au fur et à mesure du déroulement de l'enquête). La tâche de l'enquêteur sera assez facile pour les premiers interviews, plus difficile pour les derniers : le bon enquêteur doit être à même de réaliser correctement ses fins de quotas et doit éviter de se retrouver dans la situation où son dernier interviewé doit être un homme de 35 à 49 ans, inactif ou retraité !

c) Les consignes de travail

Dans la méthode des quotas, l'enquêteur est libre d'interroger qui il veut, pourvu qu'il respecte les quotas qu'on lui a fixés. Mais il va de soi qu'une série de consignes de travail lui sont normalement communiquées afin d'assurer autant que possible la validité de la procédure d'échantillonnage en faisant en sorte qu'elle reproduise le mieux possible un tirage équiprobable (c'est-à-dire un tirage dans lequel tous les individus de la population-cible ont la même « chance » d'être sélectionnés).

L'enquêteur doit veiller à disperser ses contacts autant que possible car la dispersion, par la diversité qu'elle implique, est un facteur de qualité de l'échantillon. Il lui est ainsi demandé, par exemple, de :

- ne pas interroger plus d'une personne sur dix dans un immeuble ;
- varier les étages dans lesquels sont réalisés les interviews ;
- ne pas interroger plusieurs personnes dans une même famille ;
- ne pas abuser de contacts dans une même rue. Si plusieurs contacts y sont effectués, éviter des maisons proches dont les occupants pourraient se connaître ;
- respecter la répartition indiquée des interviews dans la journée et en soirée.

Il faut également éviter d'interroger des personnes qui l'ont déjà été plus ou moins récemment dans le cadre d'une autre enquête, afin de ne pas provoquer le phénomène de « fatigue » du répondant. II faut aussi assurer que l'attitude de l'enquêteur soit toujours parfaitement neutre au cours de l'administration de l'enquête et qu'aucun lien n'unisse l'enquêteur à la personne interrogée, au risque de voir les réponses de cette dernière influencées par ce lien. Il est ainsi demandé à chaque enquêteur :

- d'éviter d'interroger des personnes qui leur sont proches, et notamment des membres de leur famille ;
- de changer de quartier après chaque enquête ;
- de ne pas interroger une personne interrogée depuis moins d'un an.

Au-delà des consignes strictes données aux enquêteurs avant le début de l'enquête, il faut aussi idéalement contrôler leur travail *a posteriori*, une fois le sondage réalisé. Ceci peut se faire, par exemple, en recontactant par voie postale ou par téléphone, certaines personnes interrogées dans le cadre de l'enquête afin de vérifier que les interviews ont été correctement réalisées. Mais cela passe aussi par l'analyse des réponses recueillies par les enquêteurs. L'analyse des correspondances multiples, par exemple, est une méthode statistique qui peut s'avérer efficace pour faire apparaître les anomalies dans les réponses.

8.2.2 La méthode des unités-types

Une autre méthode de sondage empirique est celle des « unités-types ». Elle consiste à partager la population en groupes homogènes — en groupes au sein desquels les unités statistiques ou individus « se ressemblent » — et bien différents les uns des autres, puis à choisir dans chaque groupe *une* unité statistique « représentative » du groupe. Celle-ci constitue l'*unité-type* ; elle est censée se situer dans la moyenne du groupe eu égard à un certain nombre de caractères.

On postule donc que cette unité-type a le même comportement vis-à-vis des variables étudiées que la moyenne des unités du groupe dont elle est censée être représentative. On pourrait ainsi, par exemple, décider d'interroger les parents d'élèves d'*un* groupe scolaire dont la répartition socioprofessionnelle est conforme à celle de l'ensemble des parents d'une ville ou d'une région.

L'utilisation de cette méthode d'échantillonnage se fonde sur des motivations non pas statistiques, mais plutôt sociologiques.

8.2.3 La méthode des itinéraires

La méthode dite « des itinéraires » est une méthode de sondage empirique utilisée pour obtenir des échantillons de ménages ou de logements. Elle consiste à imposer à l'enquêteur un itinéraire en lui indiquant exactement les points du circuit où il doit procéder à une interview.

L'identification de ces points d'enquête résulte de la combinaison de tirages aléatoires parmi des coordonnées géographiques et de la consultation d'une carte détaillée.

Pour les immeubles collectifs, cette identification est réalisée par la désignation, en principe aléatoire, de l'étage et de la porte sur le palier du logement à enquêter.

L'itinéraire peut également être repris sur un schéma. L'enquêteur reçoit une carte qui mentionne le nom de la commune où les enquêtes auront lieu, l'adresse exacte à laquelle la première enquête doit être effectuée. L'enquêteur poursuit ensuite son chemin en suivant le plan qu'on lui a donné, en ayant soin de respecter les consignes qui lui indiquent, par exemple, quelles sonnettes utiliser dans les immeubles, que faire en cas d'absence, quelle personne interroger dans le ménage, etc.

Cette méthode a l'avantage de laisser très peu d'initiative aux enquêteurs. Il est donc plus facile de les contrôler. La qualité de l'échantillon sélectionné dépend en fait ici des « connaissances géographiques » du quartier ou de la région ciblée, et du discernement de la personne définissant l'itinéraire. Cette méthode de sondage reste cependant *non aléatoire* au sens strict du terme.

8.2.4 Le volontariat

Il s'agit d'enquêtes réalisées auprès de « volontaires », c'est-à-dire, par exemple, de lecteurs d'un journal ou d'adhérents d'une association acceptant de répondre à un questionnaire, de téléspectateurs acceptant de répondre par téléphone ou par internet à une ou plusieurs questions posées au cours d'une émission, ou encore d'internautes qui répondent à un questionnaire proposé sur un site.

Les résultats de tels « sondages » (peut-on vraiment appeler cela des sondages, étant donné qu'il n'y a pas à proprement parler de *tirage* d'un échantillon dans la population cible ?) doivent être interprétés avec beaucoup de précautions et ne doivent en aucun cas être extrapolés à la population tout entière !

On peut aussi classer dans cette catégorie les échantillons obtenus sur « Access Panels ». Les Access Panels sont des échantillons, souvent de très grande taille, constitués d'individus pré-recrutés par un institut de sondage et acceptant le principe d'être interrogés s'ils sont sollicités. Le questionnaire peut leur être administré par voie postale, par téléphone ou par internet. Leur grande taille permet d'identifier un nombre suffisant de consommateurs d'un produit ou d'un service à faible taux de pénétration ou acheté rarement.
Le développement des Access Panels est dû aux difficultés croissantes de recrutement d'échantillons *ad hoc* pour les études de marché et à la multiplication des produits qui rendent nécessaires les études sur cible fine.

Il va de soi que, dans le cas des Access Panels, la « représentativité » de l'ensemble des répondants doit être étudiée au cas par cas, en fonction du thème de l'enquête et de la population-cible.

8.2.5 L'échantillonnage sur place

On parle d'échantillonnage sur place lorsque l'échantillonnage se fait directement sur le lieu d'achat ou le lieu d'activité (lorsque la population étudiée est définie par son activité) : on échantillonne par exemple des clients d'un centre commercial, des clients d'une certaine chaîne de restaurants, des visiteurs d'un musée ou d'une manifestation, des voyageurs dans un aéroport ou dans une gare, des automobilistes sur les routes ou les stations-service, etc.

Si l'on désire échantillonner des clients d'un centre commercial, par exemple, il faut veiller à assurer la plus grande diversité possible de clients dans l'échantillon. Ceci nécessite de bien réfléchir aux endroits où enquêter dans le centre commercial, aux périodes d'enquête, etc.

Cette méthode de sondage est surtout utilisée pour des études de marché.

8.2.6 La méthode « boule de neige »

Cette technique est adaptée à des enquêtes auprès de personnes possédant certaines caractéristiques rares. On commence par identifier quelques personnes appartenant à la population-cible, puis on leur demande d'indiquer d'autres personnes de leur connaissance possédant la même caractéristique.

Cette technique convient aussi à l'étude de réseaux relationnels (réseaux d'amis, réseaux de connaissances…).

Si cette méthode « boule de neige » n'est pas défendable d'un point de vue statistique, elle a pour elle des arguments pratiques et quelquefois sociologiques.