Université Panthéon-Sorbonne (Paris 1)
Unité de formation et de recherche en Économie

Théorie des sondages

Notes du cours de Joseph RYNKIEWICZ

Maîtrise de Modélisation statistique appliquée à l'économie et à la finance (M1)
Année universitaire 2007-2008

Prolégomènes

Jusqu'à présent, dans les cours de Statistique de Licence, nous ne nous sommes jamais souciés de la provenance des données utilisées. Celles-ci ont été considérées comme acquises, et aucune question n'a été posée sur la méthode de construction des échantillons. Des pondérations ont été appliquées et de l'inférence a été faite sur ces données partielles sans remettre en cause la partialité des données.
Dans ce cours de méthodes de sondages, nous allons voir comment construire des échantillons, sur lesquels nous calculerons des indicateurs de statistique descriptive et nous ferons de l'inférence. La théorie des sondages va alors se distinguer de la statistique inférentielle par le fait que l'aléatoire va provenir de la probabilité qu'aura un individu d'être tiré dans l'échantillon. En sondages, l'aléatoire est le fait d'être tiré ou pas, et non pas la variable d'intérêt en elle-même (considération de la statistique inférentielle). De la même manière, en régression, on suppose que les régresseurs (X_i) sont déterministes. Enfin, les méthodes de sondages vont permettre de reconstituer des données. Aujourd'hui, le flux de données est tel qu'il n'est pas possible de toutes les stocker. Seule une partie d'entre elles va être conservée. Il faut alors disposer de méthodes permettant de retrouver les données d'origine à partir de l'échantillon conservé. Les outils de la théorie des sondages seront employés à cet effet.

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

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

Objectif de la matière. À l'issue de ce cours, les étudiants doivent connaître et maîtriser les principales méthodes d'échantillonnage utilisées dans le cas d'une population finie, ainsi que les propriétés des estimateurs associés. Une partie du cours est également consacrée à la présentation de méthodes de redressement, où une information externe est utilisée pour modifier les estimateurs afin de diminuer leur variance.

Sommaire

1. Principes de base des sondages **7**

1.1. Qu'est-ce qu'un sondage ? 7

1.2. Recensement par sondage 10

1.3. Formalisation 11

1.4. Loi d'un estimateur et intervalle de confiance 15

1.5. Bases de sondage 16

1.6. Exemple 16

1.7. Théorie de l'échantillonnage *versus* statistique classique 17

2. Sondage aléatoire simple à probabilités égales **19**

2.1. Probabilité d'inclusion 19

2.2. Expressions des estimateurs du total et de la moyenne 20

2.3. Expression de la variance des estimateurs 22

2.4. Estimateur de S^2 23

2.5. Estimation des intervalles de confiance 23

2.6. Cas particulier des proportions 24

2.7. Algorithme de tirage 24

2.8. Algorithme de tirage 25

2.9. Un exemple (intentions de vote au second tour : Dijon) 25

Exercices 26

Leçon **1**

Principes de base des sondages

La première leçon 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*. Cette leçon va ainsi nous permettre de préparer le terrain pour les leçons suivantes.

1.1. Qu'est-ce qu'un sondage ?

1.1.1. 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 » !

1.1.2. Domaines d'application

Les sondages font partie de ces disciplines qui, tout en étant très mal connues dans leurs fondements par le grand public, n'en demeurent pas moins abondamment mises en œuvre dans les aspects les plus divers de la réalité quotidienne. Ce sont les très nombreux sondages pré-électoraux, sondages d'opinion, baromètres politiques ou sondages sur les modes de vie réalisés ces dernières années qui, bien adaptés à la médiatisation, constituent la forme extérieure la plus « envahissante » des sondages. Cependant, il serait regrettable de croire que les activités des « sondeurs » ne se limitent qu'à ce type particulier de sondages, et de restreindre l'emploi du sondage au cas des enquêtes sur les opinions et les modes de vie. Les domaines d'application des sondages sont en réalité très nombreux et variés. On peut ainsi donner des exemples moins connus d'utilisation des techniques de sondage :

- la recherche de gisements pétroliers ;
- la détermination du volume de certaines productions agricoles ;
- la vérification d'une comptabilité d'entreprise importante ;
- les contrôles fiscaux ;
- les calculs des grands indices « médiatiques », comme l'indice des prix à la consommation, l'indice du coût de la construction ;
- le contrôle de la qualité de fabrication des automobiles sur une chaîne d'usine ;
- les contrôles antidopage.

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.

1.1.3. 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*.

En parlant 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*.

Étudier 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.

Dans une prochaine section, nous allons revenir de manière un peu plus formelle sur ces notions-clés d'estimation, de paramètres, de population et d'échantillon.

Il existe une théorie scientifique fondée des sondages ; ce sera l'objet de ce cours.

On considère une population bien déterminée et une variable formalisant l'information qui nous intéresse, appelée *variable d'intérêt* et que l'on notera Y, définie sur chaque individu de cette population. L'*individu* est l'unité de base à laquelle on s'intéresse et la *population* est l'ensemble des individus. Par la suite, sauf mention contraire, on supposera toujours que la population est de taille finie et connue. Cette taille sera notée N. En général, on ne cherche pas à connaître la valeur Y_i prise par chacun des individus i. L'intérêt porte plutôt sur une fonction de ces Y_i qui constitue l'information que l'on cherche :

- variables quantitatives : total, moyenne, dispersion, quantiles ;
- variables qualitatives : pourcentage.

1.2. Recensement par sondage

La solution au problème est très simple si nous décidons de ne pas regarder à la dépense ; il suffit d'effectuer une enquête par recensement de la population. On a alors une connaissance exhaustive des Y_i. La valeur obtenue est alors rigoureusement égale à la vraie valeur (aux erreurs de réponse près). La plupart des budgets supportent, malheureusement, assez mal les recensements. Il est alors nécessaire de limiter ses ambitions et de collecter l'information Y sur une partie de la population, en constituant un *échantillon* d'individus, réalisant ainsi par définition une enquête par sondage.

Du point de vue du sondeur, le premier temps fort du processus d'enquête par sondage réside dans la sélection de l'échantillon d'individus. Il existe de nombreuses méthodes de tirage concurrentes pour produire un échantillon.

Sa seconde préoccupation est l'agrégation des réponses recueillies auprès des individus :

- Étape des estimateurs, c'est-à-dire des expressions mathématiques qui permettent de proposer une valeur pour la fonction des Y_i.
- Calcul de la précision de cet estimateur, qui conforte le sondeur dans son approche, car il revendique la possibilité de faire aussi bien que le recensement avec un coût beaucoup plus faible.

La méthode de tirage adoptée est la partie déterminante du coût global résultant du processus d'enquête. Déterminer la méthode de sélection de l'échantillon et la formulation de l'estimateur, c'est déterminer par définition le *plan de sondage*.

1.3. Formalisation

La théorie des sondages est un ensemble d'outils statistiques permettant d'extrapoler les résultats obtenus sur une partie de la population à la totalité de celle-ci.

1.3.1. Le paramètre

La fonction des Y_i qui nous intéresse est un paramètre, c'est-à-dire une grandeur fixée mais inconnue. Notons-la :

$$\theta = f(Y_1, Y_2, \ldots, Y_N)$$

où N est la taille connue de la population. Pour le sondeur, elle représente la « vraie valeur » qu'il faut estimer. Dans le cas d'un total, par exemple :

$$f(Y_1, Y_2, \ldots, Y_N) = \sum_{i=1}^{N} Y_i = T.$$

Dans le cas d'une moyenne :

$$f(Y_1, Y_2, \ldots, Y_N) = \frac{1}{N} \sum_{i=1}^{N} Y_i = \overline{Y}.$$

Une proportion est une moyenne particulière. Il faut alors coder $Y_i \in \{0, 1\}$.

1.3.2. L'échantillon

L'information est collectée sur un échantillon de taille n tiré par une méthode appropriée. L'échantillon, traditionnellement noté s (pour *sample*), doit être décrit par la totalité des identifiants des individus qu'il contient. Il est nécessaire d'utiliser un système de double indice pour être clair : le j-ème individu de l'échantillon est i_j où j correspond au numéro du tirage ; i_j est donc l'identifiant de l'individu tiré en j-ème position (j varie séquentiellement de 1 à n, mais i_j peut prendre toute valeur entre 1 et N, avec de nombreux « trous »).

Pour estimer θ, on tire un échantillon $s = \{Y_{i_1}, \ldots, Y_{i_n}\}$ confondu avec $\{i_1, \ldots, i_n\}$, des éléments de Ω, par abus de notation (on n'a besoin de ne savoir que les indices choisis, mais on confond cela avec la population). Les indices sont tirés au hasard dans Ω.

Exemple. Soit $N = 5$ pour une population composée de « Messieurs » 1, 2, 3, 4 et 5 (identifiants). Si $n = 3$:

1er tirage → « Monsieur » 2 est tiré ;

2e tirage → « Monsieur » 5 est tiré ;

3e tirage → « Monsieur » 1 est tiré.

Alors : $i_1 = 2$, $i_2 = 5$, $i_3 = 1$ et $s = \{i_1, i_2, i_3\} = \{2, 5, 1\}$. □

Une fois l'échantillon tiré, nous disposons de l'information $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}$. Pour trouver la quantité θ qui nous intéresse, il faut combiner ces n valeurs pour obtenir une expression dont la valeur numérique soit proche de celle de θ. La formule « agrégeant » les n valeurs s'appelle l'*estimateur de* θ. Il s'agit d'une fonction $g(.)$ calculée seulement à partir des données de l'échantillon :

$$\hat{\theta}(s) = g(Y_{i_1}, \ldots, Y_{i_n}).$$

Par exemple, pour estimer $\theta = \overline{Y}$, on pose : $g = \dfrac{1}{N} \sum_{i \in s} Y_i$.

Exemple. Soit la population de salaires mensuels en euros :

$$Y_1(1000) \qquad Y_2(500) \qquad Y_3(1200) \qquad Y_4(300) \qquad Y_5(2000).$$

$\theta =$ salaire moyen $= \dfrac{1}{N} \sum_{i=1}^{N} Y_i$, où N est la taille de la population.

On a $\{Y_i : i = 1, \ldots, N\}$, où Y_i est la variable. L'indice de la population est fixé, et le hasard porte ici sur l'indice qui va être choisi dans la population.

Si on tire deux individus au hasard dans cette population, l'ensemble des couples (non ordonnés) possibles est le suivant :

$s_1 = \{Y_1, Y_2\}$	$s_6 = \{Y_2, Y_4\}$	
$s_2 = \{Y_1, Y_3\}$	$s_7 = \{Y_2, Y_5\}$	$i \in s$ (tous les indices de s)
$s_3 = \{Y_1, Y_4\}$	$s_8 = \{Y_3, Y_4\}$	Pour $s_1 : i \in s_1 \iff i \in \{1, 2\}$
$s_4 = \{Y_1, Y_5\}$	$s_9 = \{Y_3, Y_5\}$	Pour $s_5 : i \in s_5 \iff i \in \{2, 3\}$
$s_5 = \{Y_2, Y_3\}$	$s_{10} = \{Y_4, Y_5\}$	

$\hat{\theta}(s) = \dfrac{1}{n} \sum_{i \in s} Y_i$ où n est la taille de l'échantillon (ici, $n = 2$).

$$\hat{\theta}(s_1) = \frac{1}{2} \sum_{i \in s_1} Y_i = \frac{1}{2}(Y_1 + Y_2) = \frac{1}{2}(1000 + 500) = 750\,€.$$

$$\hat{\theta}(s_5) = \frac{1}{2} \sum_{i \in s_5} Y_i = \frac{1}{2}(Y_2 + Y_3) = \frac{1}{2}(500 + 1200) = 850\,€.$$

À chaque fois que l'on « construit » une méthode d'estimation de θ, on s'arrange pour avoir un estimateur sans biais. □

1.3.3. Les mesures des erreurs d'échantillonnage

Un point important est la nature de l'aléa introduit dans notre problème. L'aléa se situe exclusivement au niveau des identifiants des individus de l'échantillon. Ce qui est aléatoire, ce sont $i_1, i_2, \ldots, i_n$, et non $Y_{i_1}, \ldots, Y_{i_n}$. Notre estimateur $\hat{\theta}(.) = g(.)$ est donc aléatoire, fonction de l'échantillon s. Notre étude porte essentiellement sur les « dégâts » occasionnés par l'aléa sur $g(.)$. Elle fait appel à trois notions :

- le biais,
- la variance,
- l'erreur quadratique moyenne.

Le biais

La valeur prise par l'estimateur $\hat{\theta}(.)$ est fonction de l'échantillon s. Soit $K = \mathrm{C}_N^n$ le nombre d'échantillons distincts de taille n parmi une population de taille finie N. Notons $(s_1, \ldots, s_K)$ la liste des K échantillons envisageables. La valeur $g(Y_{i_1}, \ldots, Y_{i_n})$ est la valeur prise par $g(.)$ si on tire l'échantillon $\{i_1, \ldots, i_n\}$ qui est quelque part dans la liste $(s_1, \ldots, s_K)$. Notons $\mathbb{P}[s_k]$ la probabilité de tirer s_k. On a la contrainte :

$$\sum_{k=1}^{K} \mathbb{P}[s_k] = 1,$$

relation que doit vérifier toute famille de probabilité. Les probabilités $\mathbb{P}[s_k]$ sont contrôlées par le sondeur (par la méthode de tirage) et la valeur moyenne de l'estimateur $g(.)$ est donc :

$$\sum_{k=1}^{K} \mathbb{P}(s_k) g(s_k).$$

Pour être plus formel, si on cherche à estimer le paramètre θ, au lieu d'utiliser $g(.)$, on note l'estimateur correspondant $\hat{\theta}$ et, si s désigne l'échantillon courant, on a :

$$E\Big(\hat{\theta}(s)\Big) = \sum_{\text{“}s\text{” possibles}} \hat{\theta}(s) \cdot \mathbb{P}[s].$$

On comprend qu'une des préoccupations des sondeurs soit de réaliser un tirage pour lequel $\mathbb{E}\Big[\hat{\theta}\Big]$ soit proche de θ. On cherche alors à réduire la quantité

$$\mathbb{E}\big[\hat{\theta}(s) - \theta\big] = \mathbb{E}\Big[\hat{\theta}(s)\Big] - \theta,$$

que l'on appelle *biais de* $\hat{\theta}(s)$, et qui constitue une première mesure de l'erreur d'échantillonnage que l'on commet.

Souvent, $\mathbb{P}[s] = \dfrac{1}{\mathrm{C}_N^n}$ (quand les échantillons sont équiprobables).

Remarque. Ceci n'est pas vérifié pour les sondages d'opinion. ⊛

La variance

La notion de moyenne ne suffit pas à mesurer la qualité d'un échantillonnage. Il faut une autre grandeur liée à la dispersion des valeurs des $\hat{\theta}$. On décide donc de calculer la variance des estimateurs $\hat{\theta}(s)$. La *variance de* $\hat{\theta}(s)$ est par définition la moyenne des carrés des écarts à la moyenne :

$$\mathrm{Var}\Big[\hat{\theta}(s)\Big] = \mathbb{E}\Big[\hat{\theta}(s) - \mathbb{E}\Big[\hat{\theta}(s)\Big]\Big] = \sum_{\text{“}s\text{”possibles}} \mathbb{P}[s] \cdot \Big[\hat{\theta}(s) - \mathbb{E}\Big[\hat{\theta}(s)\Big]\Big]^2.$$

De même, *l'écart-type* de l'estimateur $\hat{\theta}(s)$ sera :

$$\sigma\Big[\hat{\theta}(s)\Big] = \sqrt{\mathrm{Var}\Big[\hat{\theta}(s)\Big]}.$$

En terme de sondages, $\sigma\Big[\hat{\theta}\Big]$ et $\mathrm{Var}\Big[\hat{\theta}\Big]$ mesurent la précision et réalisent, après le biais, une seconde mesure de l'erreur d'échantillonnage que l'on commet.

Cas particulier : Si $\mathbb{E}\Big[\hat{\theta}(s)\Big] - \theta = 0$ (pas de biais), $\mathrm{Var}\Big[\hat{\theta}(s)\Big] = \sum_s \mathbb{P}[s] \cdot \Big(\hat{\theta}(s) - \theta\Big)^2$.

L'erreur quadratique moyenne

On peut construire un indicateur de précision qui englobe les notions de biais et de variance. Il suffit de calculer la moyenne des carrés des écarts des estimateurs à la vraie valeur selon :

$$\mathbb{E}\Big[\hat{\theta} - \theta\Big]^2.$$

Cette grandeur s'appelle l'*erreur quadratique moyenne* (ou *écart quadratique moyen*) (EQM). On a la relation :

$$\mathrm{EQM} = \mathrm{Variance} + (\mathrm{Biais})^2.$$

L'écart quadratique moyen mesure la distance entre θ et $\hat{\theta}(s)$:

$$\begin{aligned}\mathrm{EQM} &= \mathrm{Var}\Big[\hat{\theta}(s)\Big] + \mathbb{E}^2\Big[\hat{\theta}(s) - \theta\Big] \\ &= \sum_s \mathbb{P}[s] \cdot \Big(\hat{\theta}(s) - \theta\Big)^2.\end{aligned}$$

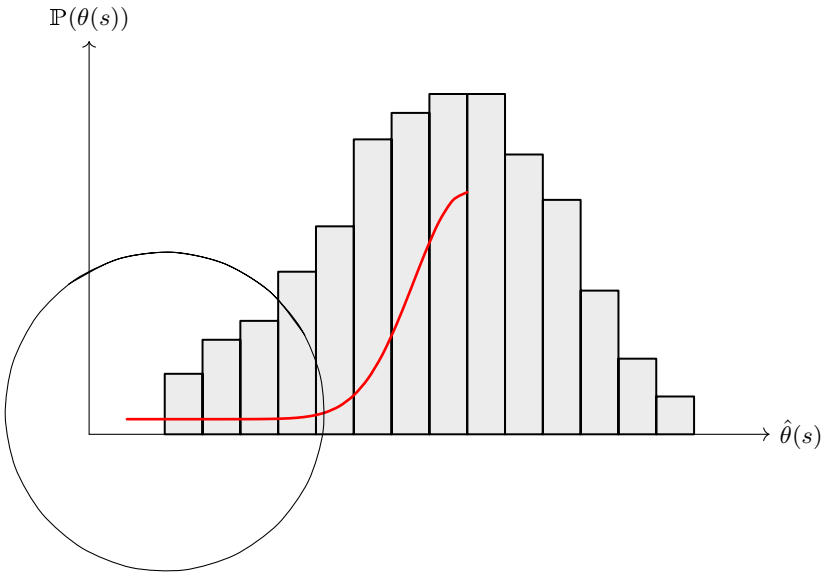
Remarque. Le nombre de s (d'échantillons) possibles est : $\mathrm{C}_N^n = \dfrac{N!}{n!(N-n)!}$ ✳

1.4. Loi d'un estimateur et intervalle de confiance

L'idéal est d'arriver à connaître non seulement le biais et la variance mais aussi la loi de l'estimateur, c'est-à-dire connaître les couples :

$$\Big(\mathbb{P}\left[s\right], \hat{\theta}(s)\Big).$$

La connaissance de la loi de $\hat{\theta}(s)$ amènerait la connaissance de la qualité de l'approximation de θ.



On a toujours affaire à une variable discrète.

Il est évidemment impossible de connaître la loi de $\hat{\theta}(s)$, cependant, on fera une approximation par une loi gaussienne (difficile à justifier théoriquement) :

$$\mathcal{N}\Big(\mathbb{E}\Big[\hat{\theta}(s)\Big], \mathrm{Var}\Big[\hat{\theta}(s)\Big]\Big).$$

Cela va nous donner un intervalle de confiance à 95 % ($\mathrm{IC}_\theta[95\,\%]$) de θ :

$$\mathrm{IC}_\theta[95\,\%] = \left[\hat{\theta}(s) - 1{,}96\sqrt{\mathrm{Var}\Big[\hat{\theta}(s)\Big]}\,;\hat{\theta}(s) + 1{,}96\sqrt{\mathrm{Var}\Big[\hat{\theta}(s)\Big]}\right],$$

soit

$$\mathrm{IC}_\theta[95\,\%] = \left[\hat{\theta}(s) - 1{,}96 \cdot \sigma\Big[\hat{\theta}(s)\Big]\,;\hat{\theta}(s) + 1{,}96 \cdot \sigma\Big[\hat{\theta}(s)\Big]\right]$$

pour les estimateurs sans biais, par approximation gaussienne (n est « grand » ; $n \geqslant 30$).

1.5. Bases de sondage

Pour pouvoir réaliser un tirage probabiliste proprement, c'est-à-dire où chaque individu a une probabilité connue et fixée de faire partie de l'échantillon, il est nécessaire de disposer d'une liste de toutes les unités d'échantillonnage faisant partie du champ de l'enquête. Cette liste doit avoir trois qualités principales :

1. Elle doit permettre de repérer l'unité sans ambigüité. Exemple : nom, prénom, adresse, date de naissance.
2. Elle doit être exhaustive.
3. Elle doit être sans double compte.

1.6. Exemple

On a une population P composée de cinq individus ayant chacun un salaire mensuel. On veut connaître le salaire moyen de cette population.

$$\theta = \frac{1\,000 + 500 + 1\,200 + 300 + 2\,000}{5} = 1\,000\,€.$$

On a le droit de tirer deux individus au hasard pour estimer ce salaire moyen.

Le paramètre θ est *déterministe,* fixé ; et on veut l'estimer. On va avoir plusieurs estimations possibles selon l'échantillon.

$s_1 = \{a, b\}$ $\quad$ $s_6 = \{b, d\}$
$s_2 = \{a, c\}$ $\quad$ $s_7 = \{b, e\}$
$s_3 = \{a, d\}$ $\quad$ $s_8 = \{c, d\}$
$s_4 = \{a, e\}$ $\quad$ $s_9 = \{c, e\}$
$s_5 = \{b, c\}$ $\quad$ $s_{10} = \{d, e\}$

$\mathrm{C}_5^2 = \dfrac{5!}{3!2!} = 10$ choix possibles.

L'estimation de θ sera la moyenne de l'échantillon, $\hat{\theta}(s)$.

$\hat{\theta}(s_1) = 750$ $\quad$ $\hat{\theta}(s_6) = 400$
$\hat{\theta}(s_2) = 1\,100$ $\quad$ $\hat{\theta}(s_7) = 1\,250$
$\hat{\theta}(s_3) = 650$ $\quad$ $\hat{\theta}(s_8) = 750$
$\hat{\theta}(s_4) = 1\,500$ $\quad$ $\hat{\theta}(s_9) = 1\,600$
$\hat{\theta}(s_5) = 850$ $\quad$ $\hat{\theta}(s_{10}) = 1\,150$

$\hat{\theta}$ est une *variable aléatoire*. Une variable aléatoire est une fonction de Ω dans $\mathbb{R}$, plus précisément ici dans $\{400, 650, 750, 850, 1\,100, 1\,150, 1\,250, 1\,500, 1\,600\}$.

La valeur idéale pour $\hat{\theta}$ serait $1\,000$; mais elle ne se trouve pas dans l'espace d'arrivée. *Une variable aléatoire ne donne donc jamais une estimation exacte.*

Ω est le tirage dans la population, c'est-à-dire l'ensemble des couples s_i, $i = \{1, \ldots, 10\}$.

L'*espérance de* $\hat{\theta}(s)$ est donnée par :

$$\mathbb{E}\Big[\hat{\theta}(s)\Big] = \sum_{i=1}^{10} \hat{\theta}(s_i) \cdot \mathbb{P}\left[s_i\right]$$

En supposant que tous les tirages sont équiprobables, on a $\mathbb{P}\left[s_i\right] = \dfrac{1}{10}$. Alors :

$$\mathbb{E}\Big[\hat{\theta}(s)\Big] = \frac{750 + 1\,100 + 650 + 1\,500 + 850 + 400 + 1\,250 + 750 + 1\,600 + 1\,150}{10} = 1\,000.$$

L'estimateur est sans biais.

La loi de probabilité est spécifique au problème ; on ne peut pas la donner, mais on peut en faire un graphique.

Il faut sortir la variable aléatoire afin de connaître sa loi et pouvoir l'utiliser.

1.7. Théorie de l'échantillonnage *versus* statistique classique

La théorie des sondages offre une approche d'inférence dite "*sous le plan de sondage*" (approche par randomisation). C'est une approche qui peut être qualifiée de *non-paramétrique*, dans la mesure où on ne postule pas un modèle paramétrique pour décrire la distribution des caractéristiques socio-économiques étudiées (c'est-à-dire les y_p et x_q).

Ce qui est aléatoire, c'est la sélection de l'échantillon. Et on connaît la loi de probabilité de cette variable aléatoire.

Il existe toutefois un certain nombre d'hypothèses de régularité de la distribution empirique des variables d'intérêt pour obtenir certains résultats asymptotiques, comme le théorème Central Limite.

Leçon 2

Sondage aléatoire simple à probabilités égales

Il s'agit d'une méthode de tirage. Le sondage aléatoire simple consiste à tirer, dans une population de taille N, un échantillon de taille n sans remise, de façon à ce que chaque individu ait la même probabilité d'inclusion. Le sondage aléatoire simple attribue également, à chaque échantillon s qui peut être formé, la même probabilité de sortie $\mathbb{P}[s]$.

On a une population $P = \{Y_1, \ldots, Y_N\}$ avec N donné.
$s = \{Y_{i_1}, \ldots, Y_{i_n}\} := \{i_1, \ldots, i_n\}$ de taille n.
Le tirage des individus est équiprobable ; tous les échantillons ont la même probabilité : $\mathbb{P}[s] = \dfrac{1}{\mathrm{C}_N^n}$.

2.1. Probabilité d'inclusion

On note $\mathbb{P}_i$ la probabilité qu'a l'individu i d'être présent dans l'échantillon (probabilité d'inclusion) :

$$\mathbb{P}_i = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$$

en sachant qu'un cas est favorable quand l'individu se trouve dans l'échantillon.

On a alors : $\displaystyle\sum_{i=1}^{N} \mathbb{P}_i = n.$

PREUVE. Il y a C_N^n échantillons possibles et équiprobables. La probabilité que l'individu i apparaisse une fois au cours des n tirages est

$$\mathbb{P}_i = \frac{\mathrm{C}_{N-1}^{n-1}}{\mathrm{C}_N^n} = \frac{n}{N}.$$

∎

On peut constater que si $\mathbb{P}[s]$ est la probabilité de tirer l'échantillon s, alors on obtient $\mathbb{P}_i$ par :

$$\mathbb{P}_i = \sum_{s \ni i} \mathbb{P}[s].$$

La grandeur $\mathbb{P}_i = \dfrac{n}{N}$ est aussi appelée le *taux de sondage*. Plus ce taux est grand, plus l'individu a de chances d'être inclus dans l'échantillon.

2.2. Expressions des estimateurs du total et de la moyenne

Quelque soit la famille de probabilités $\mathbb{P}_i$, sous la seule condition qu'elles soient toutes strictement positives, un estimateur sans biais du total $T = \sum_{i=1}^{N} Y_i$ est

$$\widehat{T} = \sum_{i \in s} \frac{Y_i}{\mathbb{P}_i}.$$

Cet estimateur est appelé estimateur de *Horvitz-Thompson*.

Ici, $\mathbb{P}_i = \dfrac{n}{N}$, donc $\widehat{T} = \displaystyle\sum_{i \in s} \frac{Y_i}{n/N} = N \times \frac{1}{n} \sum_{i \in s} Y_i$.

Preuve. On fabrique la variable aléatoire δ_i telle que :

$$\delta_i = \begin{cases} 1 & \text{si } i \in s \text{ (l'individu d'identifiant } i \text{ est dans } s) \\ 0 & \text{sinon} \end{cases}.$$

Alors on a :

$$\widehat{T} = \sum_{i \in s} \frac{Y_i}{\mathbb{P}_i} = \sum_{i=1}^{N} \frac{Y_i}{\mathbb{P}_i} \times \delta_i.$$

Y_i est non aléatoire, et : $\mathbb{E}[\delta_i] = \mathbb{P}[\delta_i = 1] = \mathbb{P}[i \in s] = \mathbb{P}_i$, d'où :

$$\begin{aligned} \mathbb{E}\left[\widehat{T}\right] &= \mathbb{E}\left[\sum_{i=1}^{N} \frac{Y_i}{\mathbb{P}_i} \times \delta_i\right] \\ &= \sum_{i=1}^{N} \frac{Y_i}{\mathbb{P}_i} \times \mathbb{E}[\delta_i] \\ &= \sum_{i=1}^{N} \frac{Y_i}{\mathbb{P}_i} \times \mathbb{P}_i \\ &= \sum_{i=1}^{N} Y_i \\ \mathbb{E}\left[\widehat{T}\right] &= T. \end{aligned}$$

■

On notera $\overline{y}$ la moyenne simple des Y_i calculée sur l'échantillon tiré s. L'estimateur sans biais du total T, dans le cas du sondage aléatoire simple, est :

$$\widehat{T} = \sum_{i \in s} \frac{Y_i}{\frac{n}{N}} = N \times \frac{\sum_{i \in s} Y_i}{n} = N \times \overline{y}.$$

Comme $\overline{Y} = \dfrac{T}{N}$, on en déduit qu'un estimateur sans biais de la moyenne $\overline{Y}$ est :

$$\widehat{\overline{Y}} = \frac{\widehat{T}}{N} = \overline{y}.$$

Cet estimateur est sans biais puisque :

$$\mathbb{E}[\overline{y}] = \mathbb{E}\left[\frac{\hat{T}}{N}\right] = \frac{\mathbb{E}\left[\hat{T}\right]}{N} = \frac{T}{N} = \overline{Y}.$$

$$\frac{\widehat{T}}{N} = \frac{1}{N}\sum_{i \in s} \frac{Y_i}{\mathbb{P}_i} = \frac{1}{n}\sum_{i \in s} Y_i = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (moyenne sur l'échantillon).}$$

La preuve directe que $\overline{y}$ est sans biais n'est pas immédiate, car les indices dans s ne sont pas *iid* (indépendants et identiquement distribués) du fait qu'il s'agit d'un tirage sans remise (on se trouve plutôt dans le cas d'une distribution hypergéométrique que d'une distribution binomiale).

Tout ça a lieu dans le cadre « sondage ».

L'estimation d'une proportion (pourcentage) est un cas particulier de moyenne. En effet, la population des individus dont on cherche la proportion constitue ce qu'on appelle un domaine. Pour chaque individu i de la population, fabriquons une variable Y_i qui vaut 1 si i est dans le domaine d'intérêt noté D (un étudiant, une femme de plus de 40 ans…), et 0 sinon (Y est la variable indicatrice du domaine D). La moyenne vaut :

$$\overline{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} = \frac{N_D}{N} = P.$$

où N_D est l'effectif vrai (inconnu) d'individus appartenant au domaine D, et P est le pourcentage recherché. Par conséquent, P apparaît comme une moyenne, et on l'estime sans biais par :

$$\widehat{P} = \overline{y}.$$

2.3. Expression de la variance des estimateurs

Introduisons $f = \dfrac{n}{N}$ le taux de sondage. On considère l'expression :

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \Big(Y_i - \overline{Y}\Big)^2.$$

Il s'agit de la variance « corrigée » ou « dispersion » de la variable Y (des Y_i) dans la population $\mathcal{P}$.

2.3.1. Cas de la variance de l'estimateur d'une moyenne

On a alors la variance de $\overline{y}$:

$$\mathrm{Var}[\overline{y}] = (1-f) \times \frac{S^2}{n}.$$

Cette formule prouve que, pour réaliser un sondage aléatoire simple qui fournisse des résultats précis, il faut :

- une taille d'échantillon n grande ;
- un taux de sondage f grand (proche de 1) ;
- une dispersion S^2 faible.

On peut montrer que la moyenne simple $\overline{y}$ estime sans biais $\overline{Y}$ si le tirage est simple *avec remise*, et que sa précision vaut

$$\mathrm{Var}[\overline{y}] = \frac{1}{n} \times \frac{N-1}{N} \times S^2.$$

Par conséquent, le rapport de la variance du sondage aléatoire simple à la variance d'un tirage avec remise de même taille est égal à $(1-f)$. On retrouve le principe qui dit que si f tend vers 0, c'est-à-dire si n est très petit devant N, alors le tirage avec et sans remise finissent par se confondre. Les tirages avec remise sont donc systématiquement moins précis que les tirages sans remises, et donc sans intérêts. En effet, se donner la possibilité de tirer plusieurs fois le même individu, c'est risquer de collecter plusieurs fois la même information, et donc gaspiller des tirages.

2.3.2. Cas de la variance de l'estimateur d'une proportion

On note ici $\overline{y} = p = \widehat{P}$. La dispersion S^2 se simplifie dans le cas de la variable indicatrice Y. En effet,

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N} {Y_i}^2 - 2 \times \sum_{i=1}^{N} Y_i \times \overline{Y} + N(\overline{Y})^2\right) \\ &= \frac{N}{N-1} \times \left(\frac{1}{N} \times \sum_{i=1}^{N} {Y_i}^2 - \overline{Y}^2\right) \end{aligned}$$

d'où

$$S^2 = \frac{N}{N-1} \times P \cdot (1-P).$$

On suppose N grand, soit $S^2 \simeq P \cdot (1-P)$, où P est toujours la vraie proportion. D'où

$$\mathrm{Var}[p] = (1-f) \times \frac{P \times (1-P)}{n}. \qquad [2.1]$$

2.3.3. Cas de la variance de l'estimateur d'un total

La variance de l'estimateur du total vaut :

$$\mathrm{Var}\left[\widehat{T}\right] = N^2 \times \mathrm{Var}[\overline{y}],$$

soit

$$\mathrm{Var}\left[\widehat{T}\right] = N^2 \times (1-f) \times \frac{S^2}{n}.$$

2.4. Estimateur de S^2

2.5. Estimation des intervalles de confiance

Les intervalles de confiance ne peuvent être formés que si on connaît la loi de l'estimateur $\overline{y}$. Les résultats traitant de la loi de $\overline{y}$, lorsqu'il n'y a pas de remise, sont assez pauvres. On peut montrer que s'il n'y a « pas trop » d'individus atypiques, et si n est très grand, on peut considérer que $\overline{y}$ suit une loi de Gauss. Dans ces conditions favorables, l'intervalle de confiance vrai à 95 % pour $\overline{y}$ est

$$I = \left[\overline{y} - 1{,}96 \times \sqrt{(1-f)\frac{S^2}{n}}\ ; \overline{y} + 1{,}96 \times \sqrt{(1-f)\frac{S^2}{n}}\right].$$

Ce véritable intervalle de confiance fait intervenir S^2, incalculable. Il doit donc, en pratique, être remplacé par un intervalle estimé

$$\hat{I} = \left[\overline{y} - 1{,}96 \times \sqrt{(1-f)\frac{s^2}{n}} \; ; \overline{y} + 1{,}96 \times \sqrt{(1-f)\frac{s^2}{n}} \right].$$

Dans le cas des proportions, la formule approchée donne l'intervalle de confiance

$$P \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \; ; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right].$$

2.6. Cas particulier des proportions

2.7. Algorithme de tirage

Le problème concret est de savoir comment extraire n unités d'un ensemble de taille N à probabilités égales sans remise. On supposera que les unités de la population constituent les enregistrements d'un fichier, et qu'on dispose d'un générateur de nombres aléatoires selon une loi uniforme $\mathcal{U}_{[0\,;\,1]}$.

2.7.1. Algorithme du tirage systématique

Le principe consiste à calculer d'abord un nombre appelé « PAS de tirage », à tirer un individu au hasard en début de fichier, et, partant de celui-ci, à descendre le long du fichier en retenant un individu tous les PAS individus, aux arrondis près.

Le PAS est égal au rapport de la taille de la population à la taille de l'échantillon

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

On remarquera qu'il n'est pas forcément entier.

L'individu tiré en premier a un rang (c'est-à-dire un numéro d'ordre dans le fichier) égal à

$$1 + \mathrm{INT}(X \times PAS)$$

où X est un nombre au hasard entre 0 et 1, et INT représente la partie entière d'un nombre positif. Les rangs des $n-1$ individus sélectionnés par la suite sont de la forme

$$1 + \mathrm{INT}((X + I) \times PAS)$$

où I est un indice de boucle variant de 1 à $n-1$.

Cette méthode assure l'équiprobabilité de tirage de chacun des individus de la population de départ. De plus, si les individus du fichier se succèdent dans un ordre aléatoire,

on montre que le tirage systématique est rigoureusement équivalent au sondage aléatoire simple. Cependant, si le fichier est effectivement dans un ordre aléatoire, il est plus simple de retenir d'emblée les n premiers individus. Ses performances (la qualité de l'estimateur obtenu) dépendent d'ailleurs du fichier.

- Si le fichier est trié selon une variable auxiliaire, il peut être assimilé à un tirage stratifié à allocation proportionnelle, et la variance de l'estimateur est plus petite que celle obtenue avec un sondage aléatoire simple.
- Si le fichier présente des périodicités, ce tirage s'apparente à un tirage en grappe avec sélection au hasard d'une seule grappe, et l'estimateur de la variance est biaisé.

2.8. Algorithme de tirage

2.8.1. Algorithme du tirage systématique

2.9. Un exemple (intentions de vote au second tour : Dijon)

Les chiffres proviennent de `www.liberation.fr`.

2.9.1. Le protocole du sondage, les résultats

2.9.2. Calcul de précision, interprétation

Exercices

Exercice 2.1 (Application directe du cours)

145 ménages de touristes séjournant en France dans une région donnée ont dépensé chacun, en moyenne journalière, 830 €. L'écart-type estimé de ces 145 dépenses journalières s'élève à 210 €. Sachant que, dans la région où a été effectuée l'enquête, il est venu 50 000 ménages de touristes, que peut-on dire de la dépense totale journalière de l'ensemble de ces ménages (on suppose que l'échantillon est issu d'un plan de sondage simple à probabilités égales) ?

Exercice 2.2

Un sondage portant sur l'opinion relative à une personnalité politique donne un pourcentage d'opinions favorables égale à $\hat{P} = 30\,\%$. En admettant qu'il s'agisse d'un sondage aléatoire simple, combien de personnes ont-elles été interrogées pour que l'on puisse dire, avec un degré de confiance de 95 %, que le vrai pourcentage d'opinions favorables dans la population ne s'écarte pas de $\hat{P}$ de plus de deux points ?

Exercice 2.3

Une entreprise de promotion immobilière désire estimer, en effectuant un sondage aléatoire simple sans remise, le nombre d'espaces de stationnement requis lors de la construction d'une nouvelle tour devant abriter des bureaux. Elle sait que la nouvelle tour abritera 5 000 personnes et que, dans les entreprises de même type que celles devant emménager dans la tour, la proportion de personnes se rendant à leur bureau en utilisant les transports en commun est toujours supérieure à 75 %.

Quelle doit être la taille de l'échantillon (réalisé au sein des futurs occupants potentiels des bureaux) pour pouvoir estimer le nombre de places de stationnement, avec une marge d'erreur symétrique d'au plus 150 places au niveau de confiance 90 % ?

Exercice 2.4

Des signatures pour une pétition ont été collectées sur 676 feuillets. Chacun de ces feuillets peut contenir au maximum 42 signatures, mais les nombres de signatures réellement recueillies sur ces feuillets sont plus faibles. Ces nombres ont été comptés sur un échantillon aléatoire de 50 feuillets (environ 7 % du total) dont les résultats sont donnés dans la table suivante :

Résultat d'un sondage sur 50 feuillets
(y_i = nombre de signatures sur 1 feuillet ; f_i = effectif de ce nombre)

y_i	42	41	36	32	29	27	23	19	16	15	14	11	10	9	7	6	5	4	3
f_i	23	4	1	1	1	2	1	1	2	2	1	1	1	1	1	3	2	1	1

On a

$$n = \sum f_i = 50 \quad ; \quad y = \sum f_i y_i = 1\,471 \quad ; \quad \sum f_i {y_i}^2 = 54\,497.$$

Estimer le nombre total de signatures recueillies et donner un intervalle de confiance au niveau 80 %.

Exercice 2.5 (Application au marketing)

Les sondages sont très largement utilisés dans le marketing direct : il arrive souvent que l'on estime par sondage le rendement d'un fichier donné, que l'on souhaite comparer les rendements de plusieurs fichiers, ou encore que, disposant de plusieurs fichiers, l'on souhaite estimer par sondage leur rendement global.

Dans cet exercice, on suppose l'existence d'un fichier de $N = 200\,000$ adresses. On note P le rendement (inconnu) du fichier à une offre d'abonnement à prix réduit : c'est donc la proportion d'individus qui s'abonneraient si l'offre était faite à tous les individus du fichier. Selon l'usage, $\hat{P}$ est l'estimation de P obtenue à partir d'un test fait sur un échantillon de n adresses, prélevées aux probabilités égales et sans remise dans le fichier.

On sait, par expérience, que les rendements à ce type d'offre sur ce fichier ne dépassent généralement pas les 3 %.

1° Quelle taille d'échantillon doit-on prendre pour estimer P avec une précision absolue de 0,5 % (ou 0,5 points) et un degré de confiance de 95 % ?

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

3° Le test a porté sur 10 000 adresses, et on a noté 230 abonnements. En déduire un intervalle de confiance bilatéral à 95 % pour le rendement P, ainsi que pour le nombre total d'abonnements, si la même offre était faite sur l'ensemble du fichier.

Exercice 2.6

Soit U une population de taille N, et $s \subset U$ un échantillon tiré par sondage aléatoire simple, et soit $s_1 \subseteq U - s$ un échantillon du restant de la population tiré par un sondage aléatoire simple. Soit

$$\hat{\overline{Y}} = \frac{1}{n} \sum_{i \in s} Y_i \quad \text{et} \quad \hat{\overline{Y}}_1 == \frac{1}{n_1} \sum_{i \in s_1} Y_i,$$

où n et n_1 sont les tailles respectives de s et s_1.

1° Déterminer l'espérance de $\hat{\overline{Y}}$ et de $\hat{\overline{Y}}_1$.

2° Montrer que la covariance entre $\hat{\overline{Y}}$ et $\hat{\overline{Y}}_1$ est égale à $-\frac{1}{N} S^2$, avec $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$ et $\hat{\overline{Y}} = \frac{1}{N} \sum_{i=1}^{N} Y_i$.

(Indication : On pourra utiliser les fonctions caractéristiques : $I_k = 1$ si $k \in s$, 0 sinon, et $I_k^1 = 1$ si $k \in s_1$, 0 sinon)

Correction des exercices

Solution de l'exercice 2.1

On déduit des données de $N = 50\,000$ et $n = 145$ que

$$\overline{y} = 830\,€ \quad \text{et} \quad \sqrt{s^2} = s = 210\,€.$$

On dispose de $\overline{y} = \dfrac{1}{n}\sum_{i\in s} Y_i = 830\,€$ pour un échantillon de $n = 145$ individus (des ménages). Donc :

$$\sum_{i\in s} Y_i = n\overline{y} = 145 \times 830 = 120\,350\,€.$$

On peut estimer le total des dépenses des 50 000 ménages, $\hat{T} = \sum_{i\in s} \dfrac{Y_i}{\mathbb{P}_i}$ avec $\mathbb{P}_i = \dfrac{n}{N} = 0{,}002\,9$. Ainsi :

$$\hat{T} = \frac{120\,350}{0{,}002\,9} = 41\,500\,000\,€.$$

On peut maintenant construire l'intervalle de confiance de la dépense totale journalière de l'ensemble des ménages.

$$\text{IC}_{(95\,\%)}(T) = \left[\hat{T} \pm 1{,}96N\sqrt{(1-f)\frac{s^2}{n}}\right]$$

$f = \mathbb{P}_i = \dfrac{145}{50\,000} = 0{,}002\,9$

$s = 210 \implies s^2 = 44\,100$

D'où :

$$\begin{aligned}\text{IC}_{(95\,\%)}(T) &= \left[41\,500\,000 \pm 1{,}96 \times 50\,000\sqrt{(1-0{,}002\,9)\frac{44\,100}{145}}\right] \\ &= [39\,793\,404\,;\,43\,206\,596]\end{aligned}$$

Autre méthode

$\overline{y}$ est un estimateur de $\overline{Y}$.

$\hat{T} = N\overline{Y}$ est un estimateur de T, donc on a ici $\hat{T} = N\overline{y}$

$\widehat{\text{Var}[\hat{T}]} = \widehat{\text{Var}[N\overline{y}]} = N^2\widehat{\text{Var}[\overline{y}]} \implies \sqrt{\widehat{\text{Var}[\hat{T}]}} = N\sqrt{\widehat{\text{Var}[\overline{y}]}}$

Et donc : $\text{IC}_{(95\,\%)}(T) = \left[\hat{T} \pm 1{,}96N\sqrt{\widehat{\text{Var}[\hat{T}]}}\right]$.

Solution de l'exercice 2.2

On veut connaître n tel que :

$$\mathrm{IC}_{[95\,\%]}(P) = \left[\hat{P} \pm 1{,}96\sqrt{(1-f)\frac{s^2}{n}}\right] = [0{,}3 \pm 0{,}02],$$

avec : $f = \dfrac{n}{N}$ et $s^2 = \hat{P}(1-\hat{P}) = 0{,}3 \times 0{,}7 = 0{,}21$.

On cherche donc n tel que :

$$1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{P}(1-\hat{P})}{n}} \leqslant 2\,\%.$$

Première solution
En supposant que $f = \dfrac{n}{N} \to 0$, car c'est un sondage dans la population entière, on a :

$$\begin{aligned}
1{,}96\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leqslant 2\,\% &\implies 1{,}96\sqrt{\frac{0{,}21}{n}} \leqslant \frac{2}{100} \\
&\iff 1{,}96^2 \times \frac{0{,}21}{n} \leqslant \frac{4}{10\,000} \\
&\iff 4n \geqslant 2\,100 \times 1{,}962 \\
&\iff n \geqslant \frac{2\,100 \times 1{,}96^2}{4} \\
&\iff \boxed{n \geqslant 2\,017 \text{ personnes}}
\end{aligned}$$

Deuxième solution
En supposant que $f = \dfrac{n}{N}$, on a :

$$\begin{aligned}
1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{P}(1-\hat{P})}{n}} \leqslant 2\,\% &\implies 1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{0{,}21}{n}} \leqslant \frac{2}{100} \\
&\iff 1{,}96^2\left(1-\frac{n}{N}\right)\frac{0{,}21}{n} \leqslant \frac{4}{10\,000} \\
&\iff 1{,}96^2\left(\frac{0{,}21}{n} - \frac{0{,}21}{N}\right) \leqslant 0{,}000\,4 \\
&\iff \frac{0{,}21}{n} - \frac{0{,}21}{N} \leqslant \frac{0{,}000\,4}{1{,}96^2} \\
&\iff \frac{1}{n} - \frac{1}{N} \leqslant \frac{0{,}000\,4}{1{,}96^2 \times 0{,}21} \\
&\iff \frac{1}{n} \leqslant \frac{1}{1{,}96^2 \times 525} + \frac{1}{N} \\
&\iff \boxed{n \geqslant \left(\frac{1}{1{,}96^2 \times 525} + \frac{1}{N}\right)^{-1}}
\end{aligned}$$

Quand N est petit, la valeur de n varie grandement, mais, quand N est grand, il y a peu de variations.

Solution de l'exercice 2.3

Solution de l'exercice 2.4

$N = 676$ feuillets
$n = 50$ feuillets
$Y_i =$ nombre de signatures $\in \{0, \ldots, 42\}$
$T = \sum_{i=1}^{N} Y_i =$ nombre total de signatures

$$\hat{T} = N \times \overline{y} = 676 \times \frac{1\,471}{50} \implies \boxed{\hat{T} = 19\,888}$$

Le quantile à 80 % est 1,28, d'où : $\mathrm{IC}_{[80\,\%]}(T) = \left[\hat{T} \pm 1{,}28N\sqrt{(1-f)\frac{s^2}{n}}\right]$.

$$\widehat{\mathrm{Var}[\hat{T}]} = N^2 \times \mathrm{Var}[\overline{y}] = N^2 \cdot (1-f)\frac{s^2}{n}$$

Or,

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i\in s}(y_i - \overline{y})^2 \\ &= \frac{1}{n}\sum_{i\in s} {Y_i}^2 - \overline{y}^2 (\text{car } n \text{ grand}) \\ &= \frac{1}{50}\left(54\,497 - \frac{1\,471}{50}\right) \\ s^2 &= 1\,099. \end{aligned}$$

Donc :

$$\begin{aligned} \mathrm{IC}_{[80\,\%]}(T) &= \left[19\,888 \pm 1{,}28 \times 676\sqrt{(1-0{,}07) \times \frac{1\,099}{50}}\right] \\ &= [19\,888 \pm 3\,053] \\ \mathrm{IC}_{[80\,\%]}(T) &= [16\,835\,;\,22\,941]. \end{aligned}$$

Solution de l'exercice 2.5

1° $N = 200\,000$

$$Y_i = \begin{cases} 1 & \text{si l'individu s'abonne} \\ 0 & \text{sinon} \end{cases}$$

$$P = \overline{Y}$$

$$\hat{P} = \sum_{i \in s} y_i \times \frac{1}{n}$$

$$s^2 = P \cdot (1 - P)$$

On sait que $P \leqslant 0{,}03 \implies s^2 \leqslant 0{,}03 \times 0{,}97 \approx 0{,}03$.

$$\mathrm{Var}\left[\hat{P}\right] = (1-f)\frac{S^2}{n} \leqslant (1-f)\frac{0{,}03}{n}.$$

On veut que :

$$\begin{aligned}
1{,}96\sqrt{\mathrm{Var}\left[\hat{P}\right]} \leqslant 0{,}005 &\implies 1{,}96\sqrt{(1-f)\frac{s^2}{n}} \leqslant 0{,}005 \\
&\implies 1{,}96\sqrt{(1-f)\frac{0{,}03}{n}} \leqslant 0{,}005 \\
&(1-f)\frac{0{,}03}{n} \leqslant \left(\frac{0{,}005}{1{,}96}\right)^2 \\
&\iff \left(1 - \frac{n}{N}\right) \times \frac{0{,}03}{n} \leqslant \left(\frac{0{,}005}{1{,}96}\right)^2 \\
&\iff \frac{0{,}03}{n} \leqslant \left(\frac{0{,}005}{1{,}96}\right)^2 + \frac{0{,}03}{N} \\
&\frac{1}{n} \leqslant \left(\frac{0{,}005}{1{,}96}\right)^2 \times \frac{1}{0{,}03} + \frac{1}{N} \\
&n \geqslant \left(\frac{1}{0{,}03} \times \left(\frac{0{,}005}{1{,}96}\right)^2 + \frac{1}{N}\right)^{-1} \\
&n \geqslant (6{,}66 \times 10^{-6})^{-1} \times 0{,}03 \\
&\boxed{n \geqslant 4\,505}
\end{aligned}$$

2° On veut que :

$$\begin{aligned}
1{,}96\sqrt{\mathrm{Var}\left[\hat{P}\right]} \leqslant 0{,}003 &\implies 1{,}96\sqrt{(1-f)\frac{0{,}03}{n}} \leqslant 0{,}003 \\
&\iff n \geqslant \left(\frac{1}{0{,}03} \times \left(\frac{0{,}003}{1{,}96}\right)^2 + \frac{1}{N}\right)^{-1} \\
&\iff n \geqslant \left(\frac{1}{0{,}03} \times \left(\frac{0{,}003}{1{,}96}\right)^2 + \frac{1}{200\,000}\right)^{-1} \\
&\iff \boxed{n \geqslant 12\,035}
\end{aligned}$$

On veut que :

$$\begin{aligned}
1{,}96\sqrt{\mathrm{Var}\left[\hat{P}\right]} \leqslant 0{,}001 &\implies 1{,}96\sqrt{(1-f)\frac{0{,}03}{n}} \leqslant 0{,}001 \\
&\iff n \geqslant \left(\frac{1}{0{,}03} \times \left(\frac{0{,}001}{1{,}96}\right)^2 + \frac{1}{N}\right)^{-1} \\
&\iff n \geqslant \left(\frac{1}{0{,}03} \times \left(\frac{0{,}001}{1{,}96}\right)^2 + \frac{1}{200\,000}\right)^{-1} \\
&\iff \boxed{n \geqslant 73\,116}
\end{aligned}$$

3° $n = 10\,000$

$\displaystyle\sum_{i \in s} Y_i = 230$

$P = \overline{Y}$

$\hat{P} = \dfrac{230}{10\,000} = 0{,}023 = 2{,}3\,\%$

$\widehat{\mathrm{Var}[\hat{P}]} = (1-f)\dfrac{s^2}{n}$ avec $s^2 = \hat{P} \cdot (1-\hat{P})$

$$\begin{aligned}
\mathrm{IC}_{[95\,\%]}(P) &= \left[\hat{P} \pm 1{,}96\sqrt{(1-f)\frac{s^2}{n}}\right] \\
&= \left[0{,}023 \pm 1{,}96\sqrt{\left(1 - \frac{10\,000}{200\,000}\right) \times \frac{0{,}023}{10\,000}}\right] \\
&= \left[0{,}023 \pm 1{,}96\sqrt{\frac{19}{20} \times \frac{0{,}023}{10\,000}}\right]
\end{aligned}$$

D'où :

$$\boxed{\mathrm{IC}_{[95\,\%]}(P) = [0{,}023 \pm 0{,}003]} \quad \text{et} \quad \boxed{\mathrm{IC}_{[95\,\%]}(T) = [4\,600 \pm 600]}$$

Solution de l'exercice 2.6

1° Le nombre d'échantillons contenant l'individu i dans s est égal à : C_{N-1}^{n-1}. Le nombre de cas possibles est égal à : C_N^n. D'où :

$$\mathbb{P}_i = \frac{\mathrm{C}_{N-1}^{n-1}}{\mathrm{C}_N^n} = \frac{n}{N}$$

$\mathbb{E}\left[\hat{\overline{Y}}\right] = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} Y_i = \overline{Y}$ (vu en cours)

Le nombre d'échantillons qui contiennent i dans s_1 est égal à : C_{N-1}^{n-1}. Le nombre de cas possibles est égal à : $\mathrm{C}_N^{n_1}$. D'où :

$$\mathbb{P}_i^1 = \frac{n_1}{N}.$$

Soit $\delta_i^1 = \begin{cases} 1 & \text{si } i \in s_1 \\ 0 & \text{sinon} \end{cases}$

$$\mathbb{E}\left[\hat{\overline{Y}}_1\right] = \frac{1}{n_1}\sum_{i\in s_1} Y_i = \frac{1}{n_1}\sum_{i=1}^{N} Y_i\delta_i^1$$

$$\mathbb{E}[\delta_i^1] = \mathbb{P}_i^1 = \frac{n_1}{N}$$

$$\Longrightarrow \mathbb{E}\left[\hat{\overline{Y}}_1\right] = \frac{1}{n_1}\sum_{i=1}^{N} Y_i\,\mathbb{E}[\delta_i^1] = \frac{1}{N}\sum_{i=1}^{N} Y_i = \overline{Y}$$

2° Trouve la solution toi-même

Table des matières

1. Principes de base des sondages **7**
1.1. Qu'est-ce qu'un sondage ? 7
1.1.1. Définitions 7
1.1.2. Domaines d'application 8
1.1.3. Objectif et principe général 9
1.2. Recensement par sondage 10
1.3. Formalisation 11
1.3.1. Le paramètre 11
1.3.2. L'échantillon 11
1.3.3. Les mesures des erreurs d'échantillonnage 13
1.4. Loi d'un estimateur et intervalle de confiance 15
1.5. Bases de sondage 16
1.6. Exemple 16
1.7. Théorie de l'échantillonnage *versus* statistique classique 17

2. Sondage aléatoire simple à probabilités égales **19**
2.1. Probabilité d'inclusion 19
2.2. Expressions des estimateurs du total et de la moyenne 20
2.3. Expression de la variance des estimateurs 22
2.3.1. Cas de la variance de l'estimateur d'une moyenne 22
2.3.2. Cas de la variance de l'estimateur d'une proportion 23
2.3.3. Cas de la variance de l'estimateur d'un total 23
2.4. Estimateur de S^2 23
2.5. Estimation des intervalles de confiance 23
2.6. Cas particulier des proportions 24
2.7. Algorithme de tirage 24
2.7.1. Algorithme du tirage systématique 24
2.8. Algorithme de tirage 25
2.8.1. Algorithme du tirage systématique 25

2.9. Un exemple (intentions de vote au second tour : Dijon) 25
2.9.1. Le protocole du sondage, les résultats 25
2.9.2. Calcul de précision, interprétation 25
Exercices 26