Méthodes de sondage – SOL3017 et SOL 6448

Notes de cours

L'échantillon, combien d'unités doit-on prendre?

Département de sociologie
Université de Montréal

Professeur : Claire Durand

1.1 La détermination de la taille de l'échantillon nécessaire :

Pour déterminer la taille de l'échantillon nécessaire pour obtenir la précision voulue (marge d'erreur), au seuil de confiance déterminé, pour une proportion maximale (ou à l'occasion pour une proportion spécifique connue), il suffit d'utiliser l'inverse de la formule de la marge d'erreur.

Pour connaître la taille "n" nécessaire pour avoir une marge d'erreur déterminée pour une proportion déterminée à un seuil de confiance α, l'inverse de l'équation de la marge d'erreur donnera la formule suivante :

$$n = \frac{Z_{\alpha}^2 * p\,(1-p)}{e^2}$$

où Z_α est la surface où l'on retrouve 1-α de la courbe normale (Z_α) et donc 1,96 lorsque le seuil de confiance accepté est de 95%,
p est la proportion de personnes ayant le comportement dont on estime la précision,
et *e* est la marge d'erreur que l'on est prêt à accepter **en décimales** i.e 2% s'inscrit 0,02.

Dans le cas d'une proportion maximale de 50% (0,5), d'un seuil de confiance de 95% et d'une marge d'erreur acceptée de ± 5% (,05), on aura le calcul suivant:

$$n = \frac{1{,}96^2 * 0{,}5(1-0{,}5)}{0{,}05^2} = \frac{3{,}84 * 0{,}5 * 0{,}5}{0{,}0025} = 384$$

Il faut donc un minimum de 384 répondants pour avoir une marge d'erreur maximale de 5% pour une proportion de 50% à un seuil de confiance de 95%. Notez bien l'erreur fréquente qui consiste à confondre 5% (0,05, la marge d'erreur) et 50% (0,50), la proportion dont on estime la marge d'erreur. Attention!

Tout comme pour la marge d'erreur, quand la proportion est maximale (0,50) et que le seuil de confiance est de 95% (Z=1,96), on note que le numérateur de l'équation précédente est approximativement égale à 1. De telle sorte que la taille théorique voulue dans ce cas est approximativement égale à 1 divisé par la marge d'erreur au carré.

$$n = \frac{1{,}96^2 * 0{,}5(1-0{,}5)}{e^2} = \frac{3{,}84 * 0{,}25}{e^2} \approx \frac{1}{e^2}$$

Si la proportion attendue du comportement était de 45% et que l'on acceptait une marge d'erreur assez faible telle 2%, les résultats seraient les suivants :

$$n = \frac{1{,}96^2 * 0{,}45(1-0{,}45)}{0{,}02^2} = \frac{3{,}84 * 0{,}45 * 0{,}55}{0{,}0004} = 2376$$

Il faudrait au moins 2376 répondants pour obtenir une marge d'erreur aussi faible pour une proportion de 45%.

On voit que le nombre d'unités nécessaire augmente très vite lorsque l'on veut réduire substantiellement la marge d'erreur, surtout lorsque la population est répartie relativement également en deux groupes. Plus la proportion est faible, moins il faut d'unités pour obtenir une marge d'erreur acceptable. Ainsi, si la proportion est de 10% et que l'on accepte une marge d'erreur de 2%, l'échantillon nécessaire est de 864.

$$n = \frac{1{,}96^2 * 0{,}1(1-0{,}1)}{0{,}02^2} = \frac{3{,}84 * 0{,}1(0{,}9)}{0{,}0004} \approx 864$$

Tout comme pour la marge d'erreur, il existe également une formule qui permet de corriger pour les populations finies, c'est-à-dire lorsque l'on peut penser que l'échantillon sera plus grand que un vingtième (1/20) de la population. Voici la formule. Elle peut paraître complexe mais il suffit de mettre les bons chiffres au bon endroit et de faire les calculs!

$$n = \frac{p^*(1-p) + \frac{e^2}{Z_\alpha^2}}{\frac{e^2}{Z_\alpha^2} + \frac{p^*(1-p)}{N}}$$

où Z_α est la surface où l'on retrouve 1-α de la courbe normale (Z_α) et donc 1,96 lorsque le seuil de confiance accepté est de 95%,
p est la proportion de personnes ayant le comportement dont on estime la précision,
e est la marge d'erreur que l'on est prêt à accepter **en décimales** i.e 2% s'inscrit ,02,
et *N* est la taille de la base échantillonnale.

Note: Le tableau I en appendice donne les résultats approximatifs de ces calculs pour diverses proportions et marges d'erreur avec un seuil de confiance de 95%. On peut également consulter

différents sites sur Internet qui donne le résultat de ces calculs. Il suffit alors de savoir quel chiffre mettre à quel endroit!

A l'aide de ces formules, on aura déterminé la taille de l'échantillon final attendu – que l'on appelle aussi *l'échantillon théorique* – que l'on veut obtenir. Cet échantillon serait équivalent à l'échantillon de départ dans un monde parfait, c'est-à-dire un monde où les listes ne comportent aucune erreur, tout le monde peut être rejoint, est disponible, en bonne santé et intéressé à ma fabuleuse enquête. Mais... la vie n'est pas parfaite, d'où la section suivante portant sur l'estimation de l'échantillon de départ nécessaire étant donné les informations possédées ou estimées sur le déroulement de la collecte.

1.2 Population, base de sondage et collecte des données

Rappel : La population est constituée de l'ensemble des unités auxquelles les résultats de l'enquête s'appliqueront. Elle doit être définie de manière précise de telle sorte que l'on puisse savoir très facilement si une personne fait ou non partie de l'échantillon.

La **base d'échantillonnage ou base de sondage** est constituée par la liste des unités d'échantillonnage (liste matérielle ou conceptuelle), c'est-à-dire liste des unités à partir de laquelle se fera la sélection. Cette liste doit constituer la meilleure approximation possible de la population : *Chaque membre de la population doit y apparaître une fois et une seule fois*.

Lorsque des membres de la population n'apparaissent pas dans la base de sondage, on parle de **biais de la base de sondage**. Ainsi, dans un sondage téléphonique auprès de la population, les personnes qui n'ont pas le téléphone ne seront pas rejointes. Il s'agit d'un biais, qui pourra avoir des conséquences plus ou moins grandes selon l'ampleur du biais. On tente de choisir une base de sondage le moins biaisée possible. Ainsi, pour un sondage téléphonique, une base produite par génération aléatoire de numéros de téléphones (GANT ou RDD pour Random Digit Dialing) sera moins biaisée que les bottins téléphoniques puisque ceux-ci ne listent pas les numéros confidentiels ni les numéros des personnes abonnées depuis la dernière édition du bottin.

Partons maintenant de notre base de sondage, que l'on tente de choisir la moins biaisée possible. Après avoir déterminé l'échantillon théorique nécessaire, il faut déterminer combien d'unités de départ il est nécessaire de tirer dans la base de sondage pour obtenir le nombre théorique voulu. Pour cela, il faut tenir compte de la **qualité de la liste (validité) de même que de la qualité des** unités inscrites sur la liste (**éligibilité**) et du taux de réponse.

1.2.1 La validité des unités sélectionnées ou la *qualité de la liste*.

Il s'agit ici de tenir compte des erreurs et des imperfections de la liste utilisée comme base d'échantillonnage. *Les unités non valides sont celles qui ne devraient pas figurer sur la liste si celle-ci était à jour et sans erreur et pouvait correspondre parfaitement à la population.*

Lorsque l'on utilise le GANT , c'est-à-dire la génération aléatoire de numéros de téléphones (RDD pour random digit dialing en anglais) pour générer des numéros de téléphone au hasard (en utilisant un programme informatique), une certaine proportion des numéros ainsi générés ne sont pas attribués (Pas de service), sont des lignes FAX, DATA, commerciales (Non résidentiel). **L'ensemble des numéros non-attribués et non-résidentiels constituent des numéros non-valides**.

En fait, quelque soit la base échantillonnale, il y a habituellement une proportion de la base qui est non-valide. On doit estimer cette proportion de façon à déterminer le plus précisément possible l'échantillon de départ requis.

De même, dans un sondage auprès des étudiants, les noms des étudiants qui ont abandonné leurs études mais qui sont encore inscrits sur la liste des étudiants seront considérés non-valides : si la liste était à jour, leurs noms n'y apparaîtraient pas.

1.2.2 L'éligibilité des unités sélectionnées ou *la qualité des personnes*

En fonction de la définition de la population et de la base échantillonnale choisie, certaines unités peuvent être considérées comme non-éligibles. Ainsi, si la population d'intérêt est définie comme l'ensemble des personnes de 18 ans et plus pouvant conduire une entrevue de 10 minutes en français ou en anglais, les personnes ne pouvant converser en français ou en anglais (problème de langue), les personnes malades ou confuses (Age, maladie) seront considérées comme étant des **unités valides mais non-éligibles.** De même, les personnes n'ayant pas droit de vote sont non-éligibles dans un sondage sur l'intention de vote.

De même, si la population est définie comme celle des jeunes de 18 à 34 ans, les personnes de moins de 18 ans et de plus de 34 ans seront considérées comme non-éligibles. Dans ce cas, on devra estimer la proportion des ménages comprenant un jeune de 18 à 34 ans, c'est-à-dire l'*incidence* dans la population.

L'éligibilité, c'est-à-dire la proportion des éligibles sur l'ensemble des éligibles et des non-éligibles et donc sur l'ensemble des unités valides, doit tenir compte de l'incidence lorsque l'on échantillonne un sous-ensemble d'une population.

Le *taux d'éligibilité* dans un sondage auprès de l'ensemble de la population est habituellement d'environ 95%. Il est un peu plus faible à Montréal (il y a plus de personnes ne pouvant parler ni le français, ni l'anglais) et dans les grandes villes en général.

Le taux *d'incidence* varie en fonction de la population à l'étude. Ainsi, si l'on voulait faire un sondage uniquement auprès des jeunes de 18 à 34 ans par exemple, il faudrait que j'estime la proportion de ménages où on est susceptible de trouver au moins un jeune de 18 à 34 ans. Cette proportion est le taux d'incidence.

1.2.3 Le taux de réponse

Parmi les personnes éligibles, certaines ne pourront pas être rejointes pour diverses raisons: Dans le cas des sondages auprès des ménages, on classe habituellement les non-réponses de la façon suivante:

refus du ménage: La personne qui répond au téléphone ou à la porte refuse que l'on fasse la sélection ou refuse de nous permettre de parler à la personne sélectionnée.

refus de la personne: La personne sélectionnée refuse de répondre au questionnaire.

pas de réponse: Personne ne répond après plusieurs appels téléphoniques ou visites effectués à différentes heures.

absence prolongée: La personne sélectionnée est absente pour la durée de la collecte des données (partie en voyage, en vacances, etc.).

incomplet: La personne sélectionnée a été rejointe et l'entrevue commencée mais l'entrevue n'a pas pu être complétée avant la fin de la période de cueillette.

Dans les sondages téléphoniques, on vise un taux de réponse (nombre de questionnaires complétés divisé par le nombre d'unités éligibles) de plus de 50%.

Dans le cas des sondages postaux ou dans les sondages internet faits auprès de populations pour lesquelles il y a une liste d'adresses courriel (membres d'une organisation, étudiants d'une université, etc.), l'absence de retour de questionnaire et le retour de questionnaires non remplis sont considérés comme des non-réponses. Les mauvaises adresses sont habituellement considérées comme non-valides.

Dans les sondages postaux ou internet auprès d'une population pour laquelle une liste existe, un taux de réponse d'environ 50% (nombre de questionnaires retournés complétés divisé par le nombre d'unités éligibles) devrait être visé.

Si l'on peut difficilement contrôler le taux de validité et le taux d'éligibilité, le contrôle du taux de réponse constitue le centre même de tous les efforts dans le déroulement d'un sondage. Si le taux de réponse est trop bas, on pourrait penser que ceux qui ne répondent pas ont des caractéristiques particulières susceptibles de biaiser les résultats. Plus la non-réponse est élevée, plus les différences de caractéristiques entre les répondants et les non-répondants peuvent avoir des conséquences importantes sur l'estimation. On parle alors de biais de non-réponse.

Pour estimer les différents taux (de validité, d'éligibilité, de réponse) en vue de déterminer l'échantillon de départ, on recueille les informations auprès des sources disponibles. Un fournisseur d'échantillon de numéros de téléphone pourra habituellement nous indiquer la validité des numéros fournis. D'autres sondages réalisés par des moyens similaires auprès de populations similaires nous donneront des indications sur la validité, l'éligibilité, le taux de réponse qu'il est possible d'atteindre. Les données statistiques (StatCan, ISQ, etc) nous permettront d'estimer l'incidence.

1. 3 Comment calculer l'échantillon de départ et le rendement du plan échantillonnal

L'échantillon de départ nécessaire se calcule en prenant l'échantillon théorique (*c'est-à-dire la taille d'échantillon que l'on vise à obtenir lorsque l'enquête sera terminée)* **que l'on multiplie par l'inverse des taux de validité, d'éligibilité – et d'incidence lorsque pertinent – et de réponse estimés.**

La formule est la suivante:

$$n_{départ} = n_{théorique} * \frac{1}{tx\ réponse} * \frac{1}{tx\ éligib. * [tx\ incidence]} * \frac{1}{tx\ validité}$$

où $n_{départ}$ est le nombre d'unités que je devrai tirer dans la base échantillonnale,
$n_{théorique}$ est l'échantillon théorique, le nombre d'unités sur lesquelles je veux que l'analyse porte,
et les différents taux sont en décimales (i.e. 0,6 pour 60%).

Exemple:

Si le taux de réponse prévu est de 60% (0,6),
le taux d'éligibilité est de 95% (0,95),
et le taux de validité de 80% (0,80)

et si je désire avoir un échantillon théorique de 384 répondants (marge d'erreur de 5% pour une proportion de 0,50 avec un seuil de confiance de 95%),

Je ferai le calcul suivant:

$$n_{départ} = 384 * \frac{1}{.6} * \frac{1}{.95} * \frac{1}{.80} = 842$$

Je peux donc évaluer que je dois tirer **842 unités** de la base échantillonnale pour obtenir un **échantillon théorique de 384 répondants** dans les conditions citées plus haut.

L'ensemble des facteurs mentionnés plus haut (validité, éligibilité/incidence et taux de réponse) amène à parler du **rendement du plan échantillonnal, c'est-à-dire la proportion attendue de questionnaires complétés étant donné l'échantillon de départ. Ce nombre peut être trouvé en divisant l'échantillon théorique (celui que l'on veut obtenir) par l'échantillon de départ.**

On dira que le rendement prévu du plan échantillonnal est de 384/842, c'est-à-dire 45,6%.
On obtiendrait le même nombre en multipliant le taux de réponse par le taux d'éligibilité (et d'incidence lorsque pertinent) et par le taux de validité (0,6*0,95*0,8=0,456).

1.4 La fraction de sélection, le pas ou l'intervalle

Une fois que nous avons toutes les informations en main, soit la base échantillonnale et une estimation de l'échantillon de départ nécessaire, il ne reste plus qu'à déterminer la fraction de sélection ou, à l'inverse, le pas ou l'intervalle de sélection.

La fraction de sélection est la relation entre l'échantillon de départ et la base échantillonnale, lorsque nous savons combien il y a d'unité exactement dans la base échantillonnale.

$$f = \frac{n}{N} = \frac{\textit{échantillon de départ}}{\textit{base échantillonnale}}$$

Une fraction d'échantillonnage de 1/10 indique que l'on doit tirer une unité de la base échantillonnale à toutes les 10 unités. La fraction de sélection indique aussi le rapport entre l'échantillon éligible et valide et la population (voir pondération).

On utilise aussi la notion de "pas" ou d'intervalle. Il s'agit de l'inverse de la fraction d'échantillonnage:

$$pas = \frac{1}{f} = \frac{N}{n} = \frac{base\ échantillonnale}{échantillon\ de\ départ}$$

Donc, une fraction de 1 sur 10 correspond à un pas de 10, ce qui veut dire que je dois échantillonner une unité de la base échantillonnale à toutes les 10 unités. On utilise habituellement le pas parce qu'on cherche une fraction "entière". On ne veut pas prendre une personne sur 2,7 personnes, ce qui est très difficile à faire, mais une personne sur 3 (on arrondit habituellement au nombre entier suivant. Toutefois lorsque l'échantillon est fait au moyen d'une procédure informatique, les fractions avec décimale peuvent être utilisées la plupart du temps.

Dans le processus pour préciser l'ensemble de ces opérations, certains allers-retours sont parfois nécessaires. Ainsi, on estimera dans un premier temps que l'intervalle requis est de 9,8. Comme on veut arrondir pour être en mesure de tirer l'échantillon, on décidera d'un intervalle de 10. On recalculera alors l'échantillon de départ (N de la base divisé par 10) et donc, par la suite, l'échantillon théorique et la marge d'erreur. Il ne faut pas oublier également que dans le cas d'un échantillon stratifié, on fera plusieurs sous-échantillons avec des fractions de sélection différentes et que l'on devra donc faire les calculs pour chacune des strates.

Il arrive que l'on n'ait pas l'information sur le nombre d'unités dans la base échantillonnale mais uniquement l'information sur la population. Ainsi, on sait qu'il y a 2M de ménages au Québec par exemple et mais on ne sait pas combien exactement il y a de numéros de téléphone dans la base de numéros de téléphone générée par GANT (on sait seulement que son taux de validité est d'environ 70%). Dans ce cas, on estime la fraction de sélection ou le pas au niveau des unités valides – qui correspondent au nombre de ménages dans la population – plutôt qu'au niveau des unités de départ. Ainsi, le pas se calculera:

$$pas = \frac{N}{n} = \frac{nb\ ménages\ population}{échantillon\ valide}$$

Voici ci-bas un tableau récapitulant l'ensemble des informations présentées.

Population	Base échantillonale	Échantillon
	N de la liste _____← ← ← **(f- pas)**	**n départ _____**
	qualité de la liste	***taux de validité: ____%***
N valide	**N valide_____← ← ←** **(f- pas)**	**n valide _____**
	Qualité des individus	***taux d'éligibilité : ____%*** ***et taux d'incidence____%***
N éligible ←	**← ← ← ← ← ← ← ← ← ← ← ←← ← ←** **(pondération)**	**n éligible _____**
	disponibilité et coopération des individus	***taux de réponse :____%***
		n répondants _____ **(n théorique, n attendu)**
		MARGE D'ERREUR ↗ **SEUIL DE CONFIANCE**
Informations proviennent de StatCan, etc.	**Information empirique** **Bottin, liste de membres, etc.**	

Taux de réponse = n répondants effectifs
n éligibles

Taux d'éligibilité = n unités éligibles
n unités valides

Taux de validité = n unités valides
n de départ

Rendement du plan d'échantillonnage = Taux de réponse x Taux d'éligibilité x Taux de validité

$$n_{départ} = n_{théorique} \times 1/\%rep \times 1/\%élig \times 1/\%valid$$

$$f = \frac{n\ départ}{N\ liste} = \frac{n\ valide}{N\ population} \qquad pas = \frac{N\ liste}{n\ départ} = \frac{N\ population}{n\ valide}$$

1.5 La pondération

L'étape de la pondération arrive **après** la cueillette de données. Elle est nécessaire...

... lorsque les fractions de sélection varient selon certaines caractéristiques de départ (échantillon stratifié),

... lorsque les taux de réponse varient en fonction de certaines caractéristiques importantes que l'on peut et veut contrôler,

... lorsque l'on veut faire une estimation s'appliquant à la population de référence,

Les résultats sont pondérés en fonction des informations que l'on a de façon à ce que les résultats reflètent le poids réel de chaque unité dans la population.

Le rapport entre le nombre d'unités éligibles et la population définie est le même que le rapport entre l'échantillon de départ et la base échantillonnale.

En pondérant, on cherche à estimer la population des unités auxquelles je veux généraliser mes résultats.

Les poids que l'on doit appliquer sont les suivants:

- *Poids d'échantillonnage* : inverse de la fraction de sélection ou "pas".
- *Poids de non-réponse* : inverse du taux de réponse

Ainsi la pondération sera égale à : *pas/taux de réponse*

Ménages et individus

Lorsque l'on sélectionne des individus en utilisant un échantillon correspondant à des ménages (unités de sélection), soit un échantillon à deux degrés, il faut procéder à une sélection à l'intérieur des ménages. Des grilles de sélection basées soit sur l'âge seul (grille de Kish), soit sur l'âge et le sexe (Troldahl-Carter), soit sur la date de naissance (non recommandée), sont utilisées.

Par la suite, on appliquera automatiquement un poids égal au nombre de personnes dans le ménage où l'individu a été sélectionné; il s'agit du poids de sélection dans les ménages.

Lorsque l'on effectue une sélection dans les ménages, on doit normalement utiliser ***un facteur de correction de la marge d'erreur*** pour tenir compte de ce mode de sélection. Ce facteur s'applique autant à la formule de détermination de la taille qu'à la formule de détermination de la marge d'erreur. Au cours des dernières années, des méthodes statistiques très raffinées se sont développées pour tenir compte des biais entraînés par les divers modes de sélection.

Redressement

On effectue un redressement lorsque, après pondération, on s'aperçoit que la répartition selon certaines caractéristiques importantes dans l'échantillon final pondéré s'écarte encore significativement de la répartition dans la population.

poids de redressement: poids qu'il faut appliquer pour que chaque catégorie d'intérêt soit représentée en fonction de son poids réel dans la population

1. 6 Synthèse des étapes à suivre:

1. Trouver les informations sur la base échantillonnale ou la population selon la disponibilité.

2. Déterminer le mode d'échantillonnage qui sera utilisé (mode de tirage, de collecte, stratification, etc.)

3. Déterminer la taille de l'échantillon théorique nécessaire à partir des informations recueillies ainsi que de la précision et du seuil de confiance voulus. Il peut s'agir de plusieurs sous-échantillons dans le cas d'un échantillon stratifié.

4. Estimer le taux de réponse attendu, le taux d'éligibilité (incluant le taux d'incidence le cas échéant) et le taux de validité de la liste. Justifier votre estimation.

5. Déterminer la taille de l'échantillon de départ requis (ou des échantillons de départ s'il y a des sous-groupes) en fonction de ces informations.

6. Déterminer la fraction de sélection (échantillon de départ requis divisé par N de la base échantillonnale) ou le pas (N de la base échantillonnale divisé par l'échantillon de départ requis) ou alternativement (échantillon valide divisé par N de la population ou l'inverse dans les cas où l'information sur la population est disponible mais non celle sur la base échantillonnage.

7. Vous pouvez également estimer le rendement attendu et la pondération prévue.

A partir de ces informations et des résultats finaux, on pourra déterminer les poids à appliquer pour la pondération de base.

A partir d'informations complémentaires sur la population, on déterminera les poids de redressement.