Cours Méthodes de sondage

© Claire Durand, 2023
Département de sociologie, Université de Montréal

L’échantillon, combien? Échantillon théorique, échantillon de départ, pas, pondération

Synthèse des notions

- *n théorique ou n attendu*: Taille de l'échantillon que l'on voudrait obtenir ***et donc le nombre de répondants souhaité.***
- *Taux de réponse estimé: proportion des unités éligibles -- ou dont on présume l'éligibilité -- avec lesquelles il serait possible de compléter l'entrevue.*
- *Taux d'éligibilité/d'incidence estimé des unités: estimation de la proportion de personnes qui correspondront à la population telle que définie.*
- *Taux de validité* de la base de sondage: estimation de la qualité de la liste
- *n de départ*: Taille de l'échantillon à sélectionner pour obtenir l'échantillon final voulu.
- *Fraction de sélection/pas*: proportion de l'échantillon de départ sur le nombre d'unités de la base de sondage.
- *Rendement du plan échantillonnal*: proportion de questionnaires complétés sur le nombre d'unités de départ.

Combien dois-je avoir de répondants?

- Supposons que je veuille mener une enquête...
- **La première question que je dois me poser est...**
- Combien devrais-je avoir de répondants pour pouvoir faire des analyses ...
  - Avec une marge d'erreur suffisamment petite pour pouvoir détecter les différences significatives.
  - Tout en tenant compte du fait que je veux une taille de l'échantillon "réaliste", que je peux atteindre à un coût acceptable.

Comment déterminer combien de répondants on devrait avoir dans l'échantillon final?

- La marge d'erreur égale:

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

- Si on connaît *la marge d'erreur que l'on est prêt à accepter* -- **e** -- et *la proportion qui nous intéresse* -- **p** --(prop. d'intention de vote pour le parti x, de personnes qui compostent ou qui utilisent le métro, etc.) --, il "suffit d'isoler" ***n*** dans l'équation pour calculer sa valeur.
- La valeur de *n* est ***l'échantillon théorique* ou *échantillon attendu, soit le nombre de répondants que l'on voudrait avoir suite à la collecte*.**
- Si on ne connait pas la proportion de la caractéristique qui nous intéresse dans l'échantillon...
  - Alors on prendra la proportion où la marge d'erreur est maximale (soit p=0,5).

L'échantillon théorique

Et donc...

- Si e égale...

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

- Si on isole n dans l'équation, on se retrouve avec les équations suivantes:

$$\sqrt{n} = \frac{Z_\alpha * \sqrt{p*(1-p)}}{e} \quad et \quad \boxed{n = \frac{Z_\alpha{}^2 * p*(1-p)}{e^2}}$$

Concrètement,...

- Si j'estime qu'une marge d'erreur de 4% dans mon cas est acceptable (attention, en décimales dans l'équation, soit **0,04**),...
- Et si je garde le seuil de confiance habituel de 95% (Z=1,96), avec une proportion maximale de 50% (0,50),...

$$n = \frac{1{,}96^2 * 0{,}5 * (1 - 0{,}5)}{0{,}04^2} = \frac{3{,}84 * 0{,}25}{0{,}0016} = 600$$

- Il faut donc **600 répondants** dans l'échantillon *théorique* pour avoir une marge d'erreur de 4% lorsque la proportion est maximale (à 50%) et le seuil de confiance à 95%.

Le calcul de l'échantillon théorique simplifié...

- Et si je garde le seuil de confiance habituel de 95%, avec une proportion maximale de 50%, il est possible d'avoir recours à une équation "approximative" rapide à utiliser.

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

- Et évidemment, vous pouvez vérifier vos calculs en allant sur le site de Circum (entre autres) : http://callweb.ca/callweb.cgi?fr:echanticalc

Comme pour la marge d'erreur

Il faut comprendre...

- Que la taille de l'échantillon nécessaire augmente ***(vite)*** avec la réduction de la marge d'erreur que l'on est prêt à accepter:
  - Pour une *marge d'erreur de 2%* et une proportion de 50%, la taille de l'échantillon théorique nécessaire serait de 2400 personnes (comparé à 600 pour 4%).

$$n = \frac{1{,}96^2 * 0{,}5 * (1 - 0{,}5)}{0{,}02^2} = \frac{3{,}84 * 0{,}25}{0{,}0004} = 2400$$

Comme pour la marge d'erreur

Il faut bien comprendre...

- Que la taille de l'échantillon nécessaire diminue si le comportement qui nous intéresse est peu fréquent:
  - Pour une marge d'erreur similaire de 2% mais une proportion de 20% -- une situation où on saurait à l'avance que la proportion risque de se situer autour de 20% --, la taille de l'échantillon théorique nécessaire serait de 1536 personnes.

$$n = \frac{1{,}96^2 * 0{,}2 * (1 - 0{,}2)}{0{,}02^2} = \frac{3{,}84 * 0{,}16}{0{,}0004} = 1536$$

Pour une proportion de 0,20 et une marge d'erreur de 4%, on en arrive à 384.

Se rappeler au besoin

- Pour une proportion de 0,50,
  - Pour une marge d'erreur de
    - 3%, il faut 1067 répondants.
    - 4%, il faut 600 répondants.
    - 5%, il faut 384 répondants.

Comme pour la marge d'erreur

Lorsque la population est finie, soit plus petite que 20 fois l'échantillon...

- Il y a une correction pour population finie qui permet de tenir compte de cette situation.
- Et donc, si la population est finie, l'échantillon théorique nécessaire sera plus petit pour une marge d'erreur donnée (à garder "au cas où" mais ne fait pas partie de la matière à retenir pour l'examen).

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

Pour p de 20% et une marge d'erreur de 2%, si N=20,000, n=1333 (au lieu de 1536).

En pratique...

Lorsque la population est finie, soit plus petite que 20 fois l'échantillon...

- En pratique, on procède différemment, c'est-à-dire que,
  - Plutôt que de se demander combien on doit avoir de répondants pour avoir une certaine marge d'erreur,
  - On procède à l'inverse: on se demande combien on aurait de répondants en utilisant une fraction de sélection f *(on peut faire de l'essai-erreur avec différentes fractions),* et quelle serait alors la marge d'erreur.
- Nous reviendrons là-dessus plus loin.

À faire...

- Il y a des notes de cours sur le site web.
- Un test pour cette semaine.
- ET des exercices sur le site web avec les solutions, pour vous pratiquer.

Mais comment détermine-t-on la marge d'erreur que l'on est prêt à accepter?

- D'abord, à partir des connaissances que nous avons sur notre sujet.
  - Ex: Je veux être capable de prédire de façon très précise les intentions de vote
    - et je sais que la différence entre les deux principaux partis est habituellement de ***5% (et donc la marge d'erreur doit être au maximum à 2,5%)***
    - et que l'intention de vote pour ces partis se situe autour de 30%.
  - Je devrais être en mesure de dire avec assez de certitude qu'un parti est en avance sur l'autre. Si je fais le calcul, j'arrive à 1290 répondants.

$$n = \frac{1{,}96^2 * 0{,}3 * (1 - 0{,}3)}{0{,}025^2} = \frac{3{,}84 * 0{,}21}{0{,}000625} = 1290$$

Les critères pour décider de la marge d'erreur acceptable et du nombre de répondants

- 1. Plus la population est divisée moitié-moitié sur un enjeu, plus la taille de l'échantillon doit être grande.
- 2. On doit se demander quelle sera la marge d'erreur pour une sous-population d'intérêt, le cas échéant (les jeunes, les allophones, etc.), ce qui entraîne que...
  - Plus la population est hétérogène, plus la taille de l'échantillon doit être grande.
  - À l'inverse, plus la population est homogène, plus la taille peut être petite.

Si tout était parfait...

Mais

Si tout était parfait,...

- Si toutes les personnes de mon échantillon étaient rejoignables et acceptaient de répondre à mon sondage,
- Si toutes les personnes de mon échantillon correspondaient à ma définition de la population,
- Si ma base de sondage (liste) correspondait parfaitement à ma population telle que je la définis,
- Il suffirait de tirer de la base de sondage le nombre de répondants que je veux à la fin dans l'échantillon, pour obtenir ce nombre,
- MAIS tout n'est pas parfait.
- **D'où la deuxième étape qui suit...**

Les questions auxquelles on doit répondre

- 1. Jusqu'à quel point la base de sondage à laquelle j'ai accès est-elle biaisée?
- 2. Jusqu'à quel point les informations de la base de sondage sont-elles valides, fiables?
- 3. Jusqu'à quel point toutes les unités de ma base de sondage sont-elles éligibles?
- 4. Quel est le taux de réponse auquel je pourrais m'attendre?

Les biais de la base de sondage

Relation population - base de sondage - les biais

Population

Base de sondage

Biais: Personnes dans la population non présentes dans la base de sondage

Validité de la liste: noms sur la liste qui ne devraient pas y être.

Propriétés de la base de sondage

Le biais

- *La population* est constituée de l'ensemble des unités auxquelles on veut que les résultats obtenus dans l'échantillon puissent être généralisés.
- *La base échantillonnale (ou base de sondage)* est constituée de la **liste des unités à partir de laquelle on fera la sélection** de l'échantillon.
- Lorsque des membres de la population telle que définie n'apparaissent pas dans la base de sondage -- la liste des unités --, on parle de *biais de la base de sondage*.
- On cherche une base de sondage la moins biaisée possible.

Propriétés de la base de sondage

Le biais, concrètement...

- Si ma base de sondage comprend uniquement les numéros de téléphone fixe publiés,
  - Le biais est le fait que les personnes qui ont seulement un cellulaire, celles qui n'ont pas fait publier leur numéro de téléphone fixe (sur liste rouge) et celles qui n'ont pas de téléphone ne sont pas sur ma liste.
- Si ma base comprend des pâtés de maison, dans lesquels on sélectionne des appartements,
  - Les personnes itinérantes, en transit, dans des résidences pour personnes âgées ou pour étudiants ne sont pas sur ma liste.
- Si je fais un sondage web, les personnes qui n'ont pas accès à internet ne peuvent pas être rejointes. C'est un biais de départ.

La validité

Relation population - base de sondage - La validité

Population

Base de sondage

Biais: Personnes dans la population non présente dans la base de sondage

Validité de la liste: noms sur la liste qui ne devraient pas y être.

Propriétés de la base de sondage

La validité

- La base peut contenir des erreurs et des imperfections. On parle alors de *qualité de la liste* et donc de *validité des unités*.
- Les unités non valides sont celles qui ne devraient pas figurer dans la base de sondage -- sur la liste --si celle-ci était à jour et sans erreur et correspondait parfaitement à la population.
- Concrètement:
  - Les listes les plus valides sont habituellement celles qui sont utilisées pour des questions essentielles d'une organisation: liste des employés recevant un chèque de paie, liste des étudiants (registraire), liste des personnes ayant une carte d'assurance-maladie, liste des payeurs d'impôt, etc.

Validité

Exemples

- Dans une liste de numéros de téléphone *générée aléatoirement*, les unités non valides sont:
  - Les numéros de téléphone non attribués et les numéros qui ne correspondent pas à une résidence principale (commerces, résidences secondaires, entrepôts, etc.)
- Dans un sondage auprès des étudiants,
  - Les unités pour lesquelles on a une mauvaise adresse postale ou courriel, un mauvais numéro de téléphone sont non-valides.
- Dans un sondage web auprès de la population, les unités dont l'adresse courriel n'est plus fonctionnelle sont non-valides.

Validité

Définitions

- Le *taux de validité* est la proportion d'unités valides sur l'ensemble des unités sélectionnées.
- Dans un sondage téléphonique auprès de la population utilisant la **génération aléatoire de numéros de téléphone* de lignes fixes,* on pouvait au départ l'estimer au Québec à environ 70%. Il était moins élevé en dehors des grandes villes (parce qu'il y a moins de numéros de téléphones attribués). Avec le développement des technologies, on peut maintenant repérer d'avance une bonne partie des numéros non valides.
- Lorsque l'on travaille avec des listes dont la validité apparaît presque parfaite, on se garde une marge de manoeuvre et on l'estime à 95%.

L’éligibilité

La qualité des unités/individus/ménages

L'éligibilité (ou admissibilité) et l'incidence

- L'*éligibilité* a trait au fait qu'il y a des personnes qui ne sont pas membres de la population telle que nous l'avons définie.
- L'*incidence* est un terme utilisé lorsque l'on fait une enquête sur des sous-populations. Elle réfère à la proportion d'unités valides où l'on est susceptible de trouver au moins une personne membre de la sous-population telle que définie.

Éligibilité

Exemples

- Dans un sondage auprès de la population, on définit habituellement celle-ci comme l'*ensemble des personnes de 18 ans et plus, capable de soutenir une conversation en français ou en anglais pendant x minutes (le temps requis pour compléter l'entrevue).*
  - Et donc, les personnes qui ne peuvent pas soutenir une conversation en français ou en anglais (problème de langue), celles qui sont confuses ou malades au point de ne pas pouvoir répondre (âge, maladie) sont considérées comme non-éligibles.
- Et les personnes qui n'ont pas le droit de vote sont inéligibles dans un sondage sur l'intention de vote.

Incidence

Exemples relatifs à l'incidence

- Si la population est définie comme l'ensemble des personnes âgées de 18 à 34 ans, les *ménages* où il n'y a personne de ce groupe d'âge seront considérés comme *inéligibles*.
- Si la population est définie comme l'ensemble des personnes ayant le français comme langue maternelle, les ménages où il n'y a personne de langue maternelle française seront considérés *inéligibles*.
- L'incidence est la proportion des ménages où il y a au moins une personne répondant à la définition de ma population.

Éligibilité

Définitions

- Le *taux d'éligibilité* est la proportion des unités **éligibles** sur les unités **valides**.
  - Il est habituellement de 95% dans un sondage auprès de la population, un peu moins dans les grandes villes (92%-93%).
  - Lorsque l'on travaille avec des listes où l'éligibilité apparaît presque parfaite (étudiants inscrits, employés d'une organisation, liste d'adresses courriel d'un panel Web), on se garde une marge de manoeuvre et on l'estime à 95%.

Éligibilité

Définitions

- Le *taux d'incidence* est la proportion des unités sélectionnées où des personnes correspondant à la définition de la sous-population sont présentes.
  - Il est pertinent seulement lorsque l'on enquête une sous-population.
  - Il varie en fonction de la définition de cette population.
- Dans un sondage auprès des jeunes de 18 à 34 ans, la proportion des ménages comportant au moins un jeune est *l'incidence* de la présence de jeunes dans les ménages.

Notez que...

- On peut difficilement contrôler les biais, les taux de validité et d'éligibilité, sinon en choisissant la base de sondage la plus appropriée possible.
- On va donc se concentrer sur le contrôle du taux de réponse.

Le taux de réponse

Le taux de réponse

Les éléments constitutifs

- Les éléments relatifs à la coopération – volontaire ou involontaire -- des personnes sélectionnées constituent le taux de réponse.
- Pour un sondage téléphonique ou face à face:
  - Si l'unité collective (par exemple, un ménage) sélectionnée refuse de collaborer au sondage, on parle de *refus du ménage*.
  - Si la personne sélectionnée refuse de collaborer, on parle de *refus de la personne*.
  - Si, après de multiples tentatives à diverses heures sur semaine et en fin de semaine, personne ne répond aux tentatives de contact, on parle de *non-contact* ou de "*pas de réponse après plusieurs tentatives".*
  - Si la personne sélectionnée est absente pour la durée de la période de collecte, on parle d'*absence prolongée.*
  - Si le questionnaire a été commencé mais non terminé, on parle de *questionnaire incomplet*.

Le taux de réponse

Les éléments constitutifs

- Pour un sondage auto-administré (par web, par ex.):
  - Si la personne sélectionnée indique clairement qu'elle refuse de collaborer, on parle de *refus*.
  - Si, après de multiples rappels, la personne sélectionnée ne répond pas aux tentatives de contact, on parle de *non-réponse après plusieurs rappels.*
  - Si la personne sélectionnée est absente pour la durée de la période de collecte – on reçoit un message qui dit que la personne est absente pour une période qui dépasse la durée de l'enquête --, on parle d'*absence prolongée.*
  - Si le questionnaire a été commencé mais non terminé, on parle de *questionnaire incomplet*.
- Pour un sondage administré à un groupe (une classe par exemple):
  - L'étudiant absent du cours le jour de la passation sera considéré comme non-répondant - absence.
  - Le refus de coopérer de l'enseignant de la classe sera considéré comme un refus.

Le taux de réponse

Définitions

- Il y a plusieurs manières de calculer le taux de réponse (voir site AAPOR) selon la manière dont on tient compte des unités pour lesquelles on ne peut pas établir l'éligibilité.
- La manière la plus simple – et une des plus sévères – est de considérer qu'il s'agit du **nombre de questionnaires complétés sur le nombre d'unités éligibles (voir ci-après)** en présumant que toutes les unités -- numéros de téléphones, adresses, courriel -- non-rejointes sont éligibles. C'est celle que nous utilisons dans ce cours.
- Notez que les grandes agences statistiques utilisent souvent une norme encore plus sévère soit le nombre de questionnaires complétés sur le nombre d'unités valides.

Le taux de réponse

Exemples

- Dans un sondage téléphonique auprès de la population fait par une firme privée au Québec, en utilisant une base de sondage probabiliste, les taux de réponse peuvent atteindre au mieux 40% (si on est prêt à payer). Ils sont plus proche de 10% en ce moment.
- Dans les sondages faits par les grands organismes gouvernementaux, téléphoniques ou au domicile, les taux de réponse approchent 60% à 70% et plus.
- Dans les sondages web, on peut aussi atteindre 40% mais c'est souvent beaucoup plus bas.
- Dans les sondages faits auprès de classes d'élèves, en classe, le taux approche 90% (Il dépend des absences, les refus étant très rares).

Que faire?

- Le **taux de réponse** constitue le coeur de tous les efforts dans la gestion des opérations.
- Si le taux de réponse est trop bas, on peut penser que les non-répondants ont des caractéristiques différentes des répondants et
  - Plus il y a de non-répondants
  - ET si leurs caractéristiques diffèrent de celles des répondants
  - Plus la possibilité de biais de non réponse est importante.

Comment faire un plan de sondage?

Comment faire le plan de sondage?

- On sait comment estimer la taille de l'échantillon théorique voulu (le nombre de répondants) étant donné la marge d'erreur que l'on accepte MAIS,
- Combien faut-il sélectionner d'unités au départ pour obtenir le nombre de répondants voulus? Voilà la question!
- On fait le processus expliqué dans les diapos suivantes dans un sens ou l'autre selon:
  - A) Que l'on a une population finie: On sait combien on peut sélectionner d'unités au départ et on veut savoir combien on aurait de répondants.
  - B) Que l'on a une population de très grande taille: On sait combien on veut de répondants et on cherche la taille de l'échantillon de départ .
  - C) que l'on a une stratégie aréolaire.

Taux typiques selon diverses situations

population	mode d'administration	validité	éligibilité	tx de réponse	incidence
étudiants universitaires	web	95%	95%	40% (MAX)	
étudiants universitaires	tétéphonique	95%	95%	75%	
population générale	web (panel opt-in)	95%	95%	40%	
population 18-34 ans	web (panel opt-in)	95%	95%	40%	25%
population générale	téléphonique	80%	92%	40% (MAX)	
population 18-34 ans	téléphonique	80%	95%	40% (MAX)	30%
échantillon de classes	distribué sur place	95%	95%	80%	
échantillon immigrant arrivé 1 an + tôt	web	80%	90%	60%	

Calcul de la fraction de sélection

Population finie

B) Population finie: Grille de calcul excel - estimer le nombre de répondants et la marge d'erreur

Calcul du n de répondants					
Population		**Base échantillonale**		**Échantillon**	
		N de la liste:		n départ:	#DIV/0!
		pas:			
N valide:		*Qualité de la liste*		*taux de validité:*	
		pas:	#DIV/0!	n valide:	#DIV/0!
		Qualité des individus		*taux d'éligibilité:*	
				et taux d'incidence:	
N éligible 1:	#DIV/0!	pondération1(base):	#DIV/0!	n éligible:	#DIV/0!
N éligible 2:	#DIV/0!	pondération2(pop):	#DIV/0!		
		Disponibilité et coopération des individus		*taux de réponse :*	
				n répondants:	#DIV/0!
				(n théorique, n attendu)	
				MARGE D'ERREUR:	#DIV/0!
				(pour popul. Finie)	
				SEUIL DE CONFIANCE	95%
Informations proviennent de StatCan, etc.		Information empirique Bottin, liste de membres, etc.			

On connaît déjà le N.
On commence par le pas.

Comment calculer?

La grille est équivalente aux calculs suivants

- Si
  - J'ai sélectionné 3289 unités dans l'échantillon de départ (soit un pas de un sur 10 pour une base comprenant 32,890 unités)
  - Que le taux de validité de la liste est de 80% (0,8)
  - Que le taux d'éligibilité estimé est de 95% (0,95)
  - Et le taux de réponse prévu est de 40% (0,4)
  - Je ferai le calcul suivant pour estimer combien j'aurai de répondants:

$$n_{répondants} = 3289*0,8*0,95*0,40=1000$$

- On obtiendra donc 1000 répondants si on sélectionne 3289 unités au départ dans ces conditions.

- Attention, marge d'erreur pour population finie (multipliée par RC ((3290-1000)/(3290-1))=,834): 2,6% au lieu de 3,1%

Calcul de la fraction de sélection

Population infinie

A) Population de très grande taille: Grille de calcul excel - estimer l'échantillon de départ

Calcul du n de départ					
Population		**Base échantillonale**		**Échantillon**	
		N de la liste:		n départ:	#DIV/0!
		pas:	#DIV/0!		
		Qualité de la liste		*taux de validité:*	
N ménages:	3395000	pas:	#DIV/0!	n valide:	#DIV/0!
		Qualité des individus		*taux d'éligibilité:*	
				et taux d'incidence:	
N éligible 1:	#DIV/0!	pondération1(base):	#DIV/0!	n éligible:	#DIV/0!
N éligible 2:	#DIV/0!	pondération2(pop):	#DIV/0!		
		Disponibilité et coopération des individus		*taux de réponse :*	
				n répondants:	
				(n théorique, n attendu)	
				MARGE D'ERREUR:	#DIV/0!
				SEUIL DE CONFIANCE	95%
Informations proviennent de StatCan, etc.		Information empirique Bottin, liste de membres, etc.			

On commence par la marge d'erreur.

Comment calculer?

La grille est équivalente aux calculs suivants

- Si
  - Je désire avoir 1000 répondants dans l'échantillon final (marge d'erreur de 3,1%),
  - Que le taux de réponse prévu est de 40% (0,4)
  - Que le taux d'éligibilité estimé est de 95% (0,95)
  - Et le taux de validité de la liste est de 80% (0,8)
  - Je ferais le calcul suivant:

$$n_{départ} = 1000 * \frac{1}{.4} * \frac{1}{.95} * \frac{1}{.80} = 3289$$

- Il faut donc sélectionner 3289 unités dans la liste de départ pour espérer obtenir 1000 répondants dans ces conditions.

Pourquoi avoir calculé tout ça?

Pour obtenir la fraction de sélection, soit comment procéder...

- C'est la dernière étape, soit j'en prends un sur combien?
- La fraction de sélection égale:

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

- Et le pas égale l'inverse de la fraction de sélection, soit:

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

Remarquez que...

- Si on n'a pas l'information sur la base échantillonnale mais seulement sur la population,
  - Il faut comprendre que le n valide (et non le n de départ) correspond à la population
  - Et donc ... il faudra calculer le pas comme suit:

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

Un sur combien?

Exemple pour une population infinie

- Par exemple, si j'ai besoin d'avoir 3289 personnes dans *l'échantillon de départ* et que j'ai 6,578,000 unités (numéros de téléphone de ménages, par exemple) dans la base de sondage,
  - J'aurai besoin de prendre une unité sur 2000, soit une fraction de sélection de 0,0005 (3,289/6,578,000)
  - Et mon pas sera de 2000 (inverse de la fraction de sélection, soit 6,578,000/3289).
- S'il y avait 275,520 unités dans la base,
  - La fraction de sélection serait de 0,0112 (3289/275,520).
  - Il est plus facile et plus parlant de calculer l'inverse (le pas) en divisant 275,520 par 3289, ce qui est égal à 83,8. On arrondira à 84 dans la pratique s'il faut faire la sélection à la main.

Calcul de la fraction de sélection

Stratégie aréolaire

C) Stratégie aréolaire: Grille de calcul excel - estimer le nombre de groupes et de répondants et la marge d'erreur

Calcul du n de répondants (devis avec choix de groupes - aréolaire)					
Population		**Base échantillonale**		**Échantillon**	
		N pers. Total de la liste:			
		nb pers. par groupe			
		nb groupes total	#DIV/0!		
		nb groupe choisis		n départ:	0
		pas:	#DIV/0!		
N valide:		*Qualité de la liste*		*taux de validité:*	
		pas:	#DIV/0!	n valide:	0
		Qualité des individus		*taux d'éligibilité:*	
				et taux d'incidence:	
N éligible 1:	#DIV/0!	pondération1(base):	#DIV/0!	n éligible:	0
	0				
		Disponibilité et coopération des		*taux de réponse :*	
				n répondants:	0
				(n théorique, n attendu)	
				MARGE D'ERREUR:	#DIV/0!
				(pour popul. Finie)	
				SEUIL DE CONFIANCE	95%
Informations proviennent		Information empirique			

On connaît déjà le N.

On commence par le pas sur le nombre de groupes - et le cas échéant, le pas à l'intérieur des groupes.

Comment calculer?

La grille est équivalente aux calculs suivants

- Si
  - J'ai sélectionné 33 unités dans l'échantillon de départ (soit un pas de un sur 3 pour une base comprenant 100 unités). Et il y a 100 personnes par unité - N=10,000).
  - Que le taux de validité de la liste est de 95% (0,95)
  - Que le taux d'éligibilité estimé est de 95% (0,95)
  - Et le taux de réponse prévu est de 50% (0,5)
  - Je ferai le calcul suivant pour estimer combien j'aurai de répondants:

$$n_{répondants} = 3300*0,95*0,95*0,50=1489$$

- On obtiendra donc 1489 répondants si on sélectionne 33 unités au départ dans ces conditions et que l'on sélectionne toutes les personnes membres de l'unité.

- Attention, marge d'erreur pour population finie (multipliée par RC ((10000-1489)/(10000-1))=,922): 2,3% au lieu de 2,5%

Rendement du plan de sondage

Rendement du plan de sondage

- L'ensemble des facteurs mentionnés plus haut (validité, éligibilité/incidence et taux de réponse) amène à parler du **rendement du plan de sondage,**
  - c'est-à-dire la proportion attendue de répondants étant donné l'échantillon de départ.
- Ce nombre peut être trouvé
  - Soit en divisant l'échantillon théorique (le nombre de répondants attendu) par l'échantillon de départ,
  - Soit en multipliant les divers taux.
- On dira que le rendement attendu du plan est de
  - 1000/3289, soit 30,4%
  - équivalent à 0,4*0,95*0,8=0,304 (soit le taux de réponse multiplié par le taux d'éligibilité multiplié par le taux de validité.

Synthèse

Synthèse des notions

- *n théorique ou n attendu*: Nombre de répondants que l'on voudrait obtenir à la fin du processus de cueillette
- *Taux de réponse estimé : proportion des unités éligibles -- ou dont on présume l'éligibilité -- avec lesquelles on estime qu'il sera possible de compléter l'entrevue.*
- *Taux d'éligibilité/d'incidence* estimé des unités: estimation de la proportion de personnes qui correspondront à la population telle que définie.
- *Taux de validité* estimé de la base: estimation de la qualité de la liste disponible.
- *n de départ*: Taille de l'échantillon à sélectionner pour obtenir le nombre de répondants voulu (le n théorique).
- *Rendement du plan échantillonnal* : proportion estimée de répondants sur le nombre d'unités de départ (ou produit des taux.
- *Fraction de sélection/pas*: proportion de l'échantillon de départ sur le nombre d'unités de la base de sondage.

Votre tâche:

- Il faut estimer au mieux possible
  - La taille de la population à laquelle vous voulez généraliser vos résultats
- Déterminer la stratégie la plus appropriée
  - Comment avoir accès à la population: quelle est...
    - La base de sondage qui serait la plus appropriée
    - Ou la manière dont vous allez procéder pour contacter les répondants
- En fonction de la stratégie, estimer quels seraient
  - Le biais, la validité, l'éligibilité, le taux de réponse et donc le rendement.
- Déterminer l'échantillon attendu et sa marge d'erreur, l'échantillon de départ et la fraction de sélection.

Synthèse des étapes

Pour une population finie

- Pour une population finie (base=moins de 20 fois l'échantillon théorique voulu), il faut procéder à l'inverse, soit:
- 1) Combien d'unités y a-t-il dans la base?
- 2) Quelle pourrait être la fraction de sélection? Ça donnerait combien d'unités de départ? Essais-erreurs...(1 sur 2? 1 sur 3? 1 sur 10?)
- 3) Estimer le taux de validité, le taux d'éligibilité (et d'incidence, le cas échéant) et le taux de réponse.
- 4) Étant donné les taux de validité, d'éligibilité et de réponse estimés, ça donnerait combien d'unités attendues (de répondants)?
- 5) Quelle serait alors la marge d'erreur? Serait-elle acceptable?

Synthèse des étapes

Pour une population de grande taille

- 1) Recueillir les informations sur la population et/ou sur la base de sondage et déterminer comment procéder. Est-ce une population de grande taille ou finie ?
- 2) Déterminer la taille de l'échantillon (ou des sous-échantillons, le cas échéant) théorique souhaitée (nombre de répondants) en fonction de la question de recherche et des informations disponibles.
- 3) Estimer le taux de validité, le taux d'éligibilité (et **d'incidence, le cas échéant**) et le taux de réponse.
- 4) Estimer l'échantillon de départ.
- 5) Estimer la fraction de sélection.

Synthèse des étapes

Pour une stratégie aréolaire

- Pour une stratégie aréolaire, il faut prendre la fraction sur le nombre d'unités collectives, par exemple des classes, des pâtés de maison (ilots), soit:
- 1) Combien d'unités collectives y a-t-il dans la base? Et combien de personnes en moyenne dans chaque unité?
- 2) Quelle pourrait être la fraction de sélection des unités collectives? Ça donnerait combien d'unités de départ? Essais-erreurs...(1 sur 2? 1 sur 3? 1 sur 10?). Il peut aussi y avoir une fraction de sélection à l'intérieur des unités.
- 3) Estimer le taux de validité, le taux d'éligibilité (et d'incidence, le cas échéant) et le taux de réponse.
- 4) Étant donné les taux estimés, ça donnerait combien de répondants?
- 5) Quelle serait alors la marge d'erreur? Serait-elle acceptable?

Pondération - redressement

Étape postérieure à la collecte des données

- Pour généraliser à la population totale, après avoir collecté les données, on se demande combien de personnes “vaut” chaque répondant.
- On pondère les données en fonction de ce qui s’est passé pendant la collecte et de la (ou des) fractions de sélection appliquées.
  - Le poids d’échantillonnage est égal au pas.
  - Le poids de non-réponse est égal à l’inverse du taux de réponse.
  - Le poids total = $poids = pas * \frac{1}{tx\,de\,réponse}$

Pondération -redressement

Étape postérieure à la collecte des données

- Si on a deux phases de sélection, par exemple un échantillon à deux degrés où on a d'abord sélectionné des ménages et ensuite des personnes dans les ménages.
- Alors chaque personne vaut le nombre de personnes éligibles dans le ménage.
- On aura le poids ménage = $$poids = pas * \frac{1}{tx\,de\,réponse}$$
  - Et le poids individu: $$poids = pas * \frac{1}{tx\,de\,réponse} * Nbpersondsménage$$

Pondération - redressement

- Si, en appliquant le poids d'échantillonnage, on se rend compte que la distribution de l'échantillon s'écarte significativement de celle de la population (ex: pas assez de jeunes dans l'échantillon, pas assez d'hommes, etc.), on corrige en appliquant un poids de redressement.
- En pratique, la plupart des firmes et organismes appliquent uniquement un poids de redressement.
- Concrètement, si j'étais supposé avoir 20% de 18-24 ans (selon le recensement) et que j'en ai 15%, j'applique un poids égal à 20/15=1,33 à tous les répondants jeunes. Chaque répondant jeune en vaut maintenant 1,33...et il y aura 20% de répondants jeunes dans l'échantillon pondéré.