Le choix de l'échantillon

Statistiques, logiciels et enquête

Benoît Le Maux

- Produire et préparer les variables -

Pourquoi un échantillon ?

Pourquoi un échantillon ?

Pour des raisons de coûts ou de délais. L'objectif est alors de construire un échantillon tel que les observations pourront être généralisées à l'ensemble de la population.

Deux méthodes pour constituer un échantillon.

- *Méthode probabiliste* : sélection de l'échantillon par tirage aléatoire dans la population-mère. Chaque individu statistique doit avoir exactement la même chance que les autres de participer à l'enquête.
- *Méthode non-probabiliste* : identifier dans la population-mère, quelques critères de répartition significatifs puis d'essayer de respecter cette répartition dans l'échantillon d'individus interrogés.

La bonne question

1. Est-ce qu'un échantillon de taille 500 suffit pour une population de 10 000 ?

2. Quelle est la taille de l'échantillon qui assure tel degré précision ?

Il est **impossible de répondre** par oui ou par non à la **première question**. Un échantillon doit fournir une estimation aussi précise que possible d'une variable, et la précision s'améliore indéfiniment lorsque la taille de l'échantillon augmente.

La notion de précision

La notion de précision (ou fiabilité d'échantillonnage) est matérialisée par **un seuil de confiance** et une **marge d'erreur.**

Dans le cas de l'estimation d'une proportion :
Un échantillon défini à un seuil de confiance de 95% et avec une marge d'erreur de $e = 3\%$ vous permettra d'extrapoler chaque résultat issu de votre enquête, avec 5% de risques de vous tromper de + ou - 3%.

Dans le cas de l'estimation d'une proportion (1)

Pour calculer la taille de l'échantillon dans le cas de l'estimation d'une proportion, vous pouvez appliquer la formule suivante :

$$n = \frac{z^2 p(1-p)}{e^2}$$

où n est la taille de l'échantillon, t est une constante issue de la loi normale selon un certain seuil de confiance (en général 95% et z=1,96), p : est le pourcentage de gens qui présentent le caractère observé, e est la marge d'erreur d'échantillonnage choisie.

Dans le cas de l'estimation d'une proportion (2)

Répartition des réponses		**Erreur d'échantillonnage**				
p	$1-p$	**1 %**	**5 %**	**10 %**	**15 %**	**20 %**
50 %	**50 %**	9604	385	97	43	25
60 %	**40 %**	9220	369	93	41	24
70 %	**30 %**	8068	323	81	36	21
80 %	**20 %**	6147	246	62	28	16
90 %	**10 %**	3458	139	35	16	9

Table: Taille de l'échantillon dans le cas d'une proportion

Quelques remarques (1)

La généralisation aux méthodes non-probabilistes :

En théorie, les méthodes de calcul scientifique de la taille d'échantillon ne s'applique que sur les échantillons obtenus par la méthode probabiliste. En pratique, ces méthodes de calcul sont quand même utilisées.

Un échantillon de grande taille mais pas trop

Plus votre échantillon est important, plus la généralisation sera fiable. Mais, les gains de fiabilité ne sont pas proportionnels à l'augmentation de la taille de l'échantillon.

Quelques remarques (2)

La taille de la population-mère

c'est un facteur seulement lorsqu'elle est en deçà de 100 000 ou à peu près. En deçà, il faut utiliser, pour déterminer la taille de l'échantillon, un élément appelé « facteur de correction » :

- *Dans le cas d'une proportion :* $n' = \frac{n}{1+\frac{n+1}{N}} \approx \frac{n}{1+\frac{n}{N}}$

N représente la population mère.

Quelques remarques (3)

Population	Echantillon
N	n'
50	45
100	80
200	132
1 000	278
2 000	323
5 000	357
10 000	371
100 000	384
200 000	385

Table: Population et échantillon dans le cas d'une proportion

Quelques remarques (4)

Un faux problème

En général, pour déterminer la taille d'un l'échantillon, une étude quantitative se base généralement sur les chiffres obtenus dans le tableau ci-dessus. Un échantillon de 200-300 individus fournit donc de bon résultats. On choisit ensuite un seuil de confiance (en général 95%). La marge d'erreur, et par conséquent l'intervalle de confiance, sont ensuite déduits via la formule suivante :

- *Pour une proportion :* $e = z \times \sqrt{\frac{p \times (1-p)}{n}}$.

Résumé

La précision dépend de deux éléments principaux :

1. **La taille n de l'échantillon :** plus l'échantillon est grand plus l'estimation est précise.
2. **Le seuil de confiance** : plus celui-ci sera grand, plus z grandira et plus l'intervalle de confiance sera large.

Méthode d'échantillonnage probabiliste

Définition

Les échantillons probabilistes *ou aléatoires sont constitués par tirage au sort dans la population mère pour laquelle on dispose de la liste complète de toutes les unités de sondage qui la composent (individus, familles, entreprises, etc.). On distingue 4 méthodes :*

- *Echantillonnage aléatoire simple*
- *Echantillonnage aléatoire systématique*
- *Echantillonnage stratifié*
- *Echantillonnage en grappes et à plusieurs degrés*

Echantillonnage aléatoire simple

Principe

- *Chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon.*
- *Chaque combinaison de membres de la population a aussi une chance égale de composer l'échantillon.*

Mode d'administration

Vous devez dresser une liste de toutes les unités incluses dans la population observée pour sélectionner un échantillon aléatoire simple. Un échantillonnage aléatoire simple peut s'effectuer avec ou sans remise.

- **Avantages :** facile à mettre en œuvre.
- **Inconvénients :** La non-représentativité, le coût.

Echantillonnage systématique

Principe

Il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon.

Mode d'administration

1. *Numéroter de 1 à N les unités incluses dans votre base de sondage (où N est la taille de la population totale).*
2. *Déterminer l'intervalle d'échantillonnage ou pas de sondage (K) en divisant la population N par la taille de l'échantillon que vous désirez obtenir.*
3. *Sélectionner au hasard un nombre entre 1 et K. Ce nombre s'appelle l'**origine choisie au hasard**.*
4. *Sélectionner chaque K ème unité après ce premier nombre.*

- **Avantages :** La probabilité d'être sélectionnée = celle d'un EAS.
- **Inconvénients :** Le coût, problème si la population est ordonnée.

Echantillonnage stratifié

Principe

Découper la population en sous ensembles appelés strates et réaliser un sondage dans chacune d'elles.

Mode d'administration

1. *On divise la population en groupes homogènes (appelés strates), qui sont mutuellement exclusifs (selon l'âge, le sexe, la province de résidence, le revenu, etc.)*
2. *On sélectionne à partir de chaque strate des échantillons indépendants. On peut utiliser n'importe quelle des méthodes d'échantillonnage*
3. *La méthode d'échantillonnage peut varier d'une strate à une autre.*

- **Avantages :** La probabilité d'être sélectionnée = celle d'un EAS. Echantillon plus représentatif.
- **Inconvénients :** Le coût.

Echantillonnage en grappes et à plusieurs degrés

Principe

Limiter les zones géographiques qui font l'objet de l'enquête

Mode d'administration

Si la population est répartie sur M grappes (usines, établissements d'enseignement, subdivisions électorales) :

1. *1er degré : choisir un échantillon de m grappes.*
2. *2ème degré : réaliser une enquête dans chacune des m grappes :*
   - *soit auprès de tous les éléments (dits aussi unités secondaires) : sondage par grappes.*
   - *soit en désignant des échantillons d'unités secondaires : sondage à deux degrés.*

- **Avantages :** réduire les coûts
- **Inconvénients :** Effet de grappe (variance intra qui est faible) dû à l'existence de similarite entre individus d'une même grappe.

Méthode d'échantillonnage non-probabiliste

Définition

La méthode d'échantillonnage non-probabiliste *est utilisée lorsqu'il n'est pas possible de constituer une liste exhaustive de toutes les unités du sondage.*

- *Dans le cas de l'échantillonnage probabiliste, chaque unité a une chance d'être sélectionnée.*
- *Ce n'est plus vrai dans le cas de l'échantillonnage probabiliste. On se fixe alors comme règle que l'échantillon retenu doit avoir la même composition que la population mère par rapport à une ou plusieurs caractéristiques.*

Echantillonnage par quotas

Principe

Il s'effectue jusqu'à ce qu'un nombre précis d'unités (de quotas) pour diverses sous-populations ait été sélectionné.

Mode d'administration

- *Les quotas peuvent être fondés sur des proportions de la population. (par exemple 50% d'hommes et 50% de femmes)*
- *Ne retenir qu'un nombre restreint de quotas. Au delà de 2 ou 3 quotas, on complique la tâche des enquêteurs.*

- **Avantages :** L'échantillonnage par quotas est généralement moins coûteux que l'échantillonnage aléatoire. Il est également facile à administrer.
- **Inconvénients :** Certaines unités peuvent n'avoir aucune chance d'être sélectionnées.

Les autres méthodes non-probabilistes (1)

Le volontariat

On prélève l'échantillon à partir d'un groupe de volontaires.
⇒ Inconvénients : échantillon biaisé

La méthode des itinéraires

On impose à l'enquêteur :

- Un point de départ dans une commune.
- Un itinéraire à suivre avec tirage systématique des logements dans lesquels effectuer les interviews
  ⇒ **Objectif**: reproduire un certain tirage aléatoire des enquêtés, sans donner explicitement des noms et adresses à l'enquêteur.

Les autres méthodes non-probabilistes (2)

Technique de « boule de neige »

Utilisation de personnes comme source d'identification d'unités additionnelles.

Échantillonnage de convenance ou au jugé

On prélève un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population.

Échantillonnage sur place

L'échantillon étudié est définie par un lieu. Cette méthode est utilisé dans l'échantillonnage de populations mobiles, rares ou spécifiques. Avec cette méthode, il faut faire attention à :

- ne pas sur-représenter les individus passant + de temps sur place
- les périodes d'enquête
- les pondérations *a posteriori* pour tenir compte de la probabilité de présence

Conclusion

- **Pourquoi un échantillon ?** $\Rightarrow$ La population cible est généralement trop nombreuse et pour des raisons de coûts, de délais, il est pratiquement impossible d'étudier tous les individus d'une population c'est-à-dire d'effectuer un recensement.
- **Quelle taille d'échantillon ?** En général, on utilise la formule $n' = \frac{385}{1+\frac{385}{N}}$ pour trouver la taille nécessaire (pour que la marge d'erreur dans l'estimation de la proportion soit inférieur à 5 % et ce, pour un seuil de confiance de 95%).
- **Objectif :** Construire un échantillon tel que les observations pourront être généralisées à l'ensemble de la population (méthode probabiliste ou non-probabiliste).
- **Condition :** Il faut que l'échantillon présente les mêmes caractéristiques que la population cible. En d'autres termes, qu'il soit **représentatif**. Si ce n'est pas le cas, l'échantillon est biaisé.