STT-2000, Échantillonnage

Pierre Duchesne

August 30, 2019

Coordonnées

- Professeur: Pierre Duchesne
- Courriel: duchesne@dms.umontreal.ca
- Téléphone: 514-343-7267
- Bureau: 4251 du Pavillon André-Aisenstadt
- Site web: http://www.dms.umontreal.ca/∼duchesne

Plan de cours

1. Structure d'une enquête par sondage
2. Méthodes d'échantillonnage
3. Utilisation de l'information auxiliaire
4. Sondages empiriques
5. Types d'erreur dans les sondages
6. Méthodes de Monte Carlo

Barême

Le barême proposé est le suivant:

- Examen intra: 30%;
- Examen final: 40%;
- Devoirs: 22.5%($7.5 \times 3 = 22.5$) (donc au nombre de trois);
- Projet: 7.5%.

Ouvrages de référence (recommandés):

- Lohr (2010). *Sampling: Design and Analysis*. Seconde édition. Duxbury Press, New York.
- Särndal, Swensson et Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Qu'est-ce qu'un sondage?

La vie en société implique une collection d'individus, la population, et nous avons souvent besoin d'information sur cette dernière.

- Préférences, choix (choix politiques, préférences en matière de consommation).
- Besoins (qu'est-ce que le consommateur est prêt à payer).
- Comportement des individus (études de marché).

Sondage d'opinion

Certainement la partie la plus visible du grand public.
Cependant, les sondages ne se résument pas aux sondages d'opinion.
De manière générale, on peut considérer que ceux qui ont besoin des sondages sont:

- Gouvernements;
- Entreprises;
- Institutions sociales.

Sondage d'opinion

- Secteur gouvernemental: Statistique Canada (www.statcan.gc.ca); Institut de la Statistique du Québec (www.stat.gouv.qc.ca/); U.S. Bureau of Census (www.census.gov);
- Secteur privé: Gallup (www.gallup.com); Harris (harrispollonline.com); Augus Reid (www.angusreidforum.com);
- Recherche: sondages maison dans les universités, hôpitaux.
- Gestion, affaires: études de marché, marketing.

Définition d'un sondage

Méthode de collecte de l'information sur un échantillon d'individus. On parlera en général d'unités. Ces unités pourraient être des humains, des animaux, des maisons ou encore des entreprises, pour ne citer que ces exemples.

L'échantillon n'est donc qu'une partie (une fraction) de la population. Ceci est en opposition avec le recensement, où tous les membres de la population sont étudiés (on dira également sondés).

Exemples de sondages

- Exemple 1. Un échantillon de personnes aptes à voter est questionné à l'avance sur une élection. Parmi les questions, intentions de vote, perception des différents candidats, résultats anticipés de l'élection.
- Exemple 2. Directeur de la santé publique de Montréal veut dresser un portrait de la sexualité chez les jeunes (www.dsp.santemontreal.qc.ca).

Portrait de la sexualité chez les jeunes

Exemples de questions:

- En matière de sexualité, vers qui les jeunes se tournent d'abord comme principale source de renseignements sur la sexualité?
- D'accord ou pas d'accord avec l'éducation de la sexualité dans les écoles secondaires?
- Chez Tel-Jeunes, quels sont les deux sujets les plus abordés?
- Age moyen et/ou médian de la première relation sexuelle?

Pour des raisons administratives, on voudrait les résultats d'ici un mois

- Option 1. On réunit le personnel nécessaire et on va voir chaque adolescent qui fréquente une école secondaire: Frais de personnel? Temps? Frais de déplacement? Contraintes de coûts?
- Option 2. Mise en oeuvre d'un sondage. On choisit un échantillon représentatif (possiblement échantillon d'écoles et on tente de rejoindre tous les adolescents d'une école sélectionnée). On aura donc un sous-ensemble de tous les adolescents (la population). Pour une fraction du coût, permet de gagner du temps.

Erreurs lorsque l'on dispose d'un échantillon (SSW, p.14; Lohr, p. 15)

- Erreurs dues à l'échantillonnage: l'échantillon n'est pas la population.
- Erreurs non dues à l'échantillonnage: erreurs de mesures; biais de sélection.

Biais de sélection

- Représentation exagérée d'une partie de la population;
- Sous-couverture de la population;
- Doubles dans la base de sondage;
- Interviewer néglige certaines personnes;
- Population cible versus population échantillonnée;
- Choix délibérée des personnes à interviewer;
- Non-réponse.

Avec un recensement, qu'en est-il?

Échantillon versus recensement

- L'échantillon est souvent plus fiable que le recensement.
- Il faut aller au-delà des considérations mathématiques.
- Avec un recensement, en théorie, par d'erreurs dues à l'échantillonnage.
- Les ressources nécessaires à la mise en oeuvre du recensement sont telles que la précision n'est pas toujours au rendez-vous (besoin de personnel qualifié en quantité suffisante; si le travail est trop imposant, risque d'être bâclé; jamais à l'abri de la non-réponse pouvant fausser les résultats).

Idéalement, il est espéré que l'on pourra mobiliser les ressources afin que les données obtenues sur l'échantillon soient de qualité; meilleur contrôle de la qualité.

Comment choisir l'échantillon?

- Pas n'importe comment: on veut une méthode objective.
- Pour éviter les biais: pas parmi les volontaires.
- On ne veut pas de SLOPS (*self-selected opinion pools*): sondages télé, lignes ouvertes, réseaux sociaux (on veut éviter les *trolls*). Les gens avec des préjugés, fortes opinions (souvent négatives) qui s'expriment souvent sur des questions sensibles (avortement, racisme, etc.)

Comment choisir l'échantillon? (suite)

- On veut une méthode scientifique telle que chaque personne dans la population possède une chance *mesurable* (que l'on peut quantifier) de sélection.
- Commenter l'affirmation suivante: Pour obtenir un échantillon, il faut que chaque unité possède une chance égale de faire partie de l'échantillon (Attention il y a un piège).

Avantages de la méthode scientifique

- Avec la méthode scientifique, tout le monde se voit poser les mêmes questions dans le même ordre.
- On peut projeter les résultats de l'échantillon sur l'ensemble de la population.
- Le but d'un sondage n'est pas de décrire un individu en particulier. On veut une image, un profil, de la population.

Ingrédients pour faire un sondage

- Choix du plan de sondage. Comment choisir les unités dans la population de taille N.
- Combien d'unités choisir? Taille de l'échantillon, notée n.

Pour choisir le n, il faudra introduire des critères. Fonction de la précision souhaitée, la taille peut être plus ou moins grande. On verra que le N n'est pas un facteur majeur. Ainsi, il est possible que 1000 unités soient suffisantes pour refléter des caractéristiques de populations comportant des millions d'individus; autrement dit, un échantillon de taille $n = 1000$ fournira souvent une précision comparable dans une population de 100000 unités ou de 1000000 d'unités.

Les étapes d'un sondage

1. Sélection d'un échantillon.
2. Collecte des données.
3. Vérification et imputation.
4. Estimation et analyse.
5. Publication des résultats.

Population finie

On doit procéder à l'identification de la population cible.
Quelle est la population visée?
On doit construire une base de sondage dans une population finie.

- Idéalement on tente de trouver une base de sondage existante. Sinon on doit en construire une.
- Une base de sondage est une liste des éléments dans la population.
- Tous les éléments de la population sont identifiés: elle contient les coordonnées des unités.

Base de sondage versus population visée

Il faut souvent être conscient des limites de l'étude.

- Est-ce que la base de sondage correspond à la population visée?
- Si l'on s'intéresse à la population québécoise et que l'on utilise les listes de téléphone, est-ce que la population cible = base de sondage?
- C'est quoi aujourd'hui un sondage téléphonique? Utilise les lignes fixes? Les numéros de portable?
- Comment contourner les difficultés?
- Conséquences sinon?

1. Sélection d'un échantillon

Une fois que l'on dispose d'une base de sondage, on cherche à choisir un échantillon, noté s, dans la population U:

$$U = \{1, 2, \ldots, k, \ldots, N\}.$$

On obtiendra alors $s \subseteq U$.
Pour obtenir l'échantillon, on procède à un échantillonnage, c'est-à-dire que l'on procède à une sélection dans U selon un plan de sondage (plan d'échantillonnage) que l'on notera p.

2. Collecte des données

Logiciels spécialisés:

Méthode *CATI* (*Computer Assisted Telephone Interview*): logiciel spécialisé de gestion des sondages par téléphone;

Méthode *Web-CATI*: avec interface web;

Méthode *CAPI* (*Computer Assisted Personal Interview*): interview personnelle; maintenant avec téléphone portable et tablette.

https://www.surveysystem.com/interviewing-cati.htm

https://www.surveysystem.com/CAPI-software.htm

Formation

Parmi les avantages de ces méthodes: le déroulement des entrevues est rigoureusement contrôlé.
Inconvénients: planification des enquêtes, interfaces avec les logiciels.
Exemple de sondages qui utilise ces méthodes: le sondage CPS (*Current Population Survey*)
On se doute que l'implantation de ces méthodes nécessite que les interviewers soient formés.

Questionnaire

Le questionnaire nécessite la présence des experts du sujet. Les questions doivent être validées et la façon de poser les questions peut avoir un impact sur la réponse:
Exemple: aux États-Unis, dans un sondage *NBC/Wall Street Journal*, deux groupes furent sélectionnés, et un groupe donné se voyait poser une des deux questions suivantes:

1. Êtes-vous en faveur de couper dans les programmes tels la sécurité du revenu, les soins médicaux, les subventions agricoles afin de réduire le déficit?
2. Êtes-vous en faveur de coupures gouvernementales afin de réduire de déficit?

Résultats du sondage *NBC/Wall Street Journal*

Groupe d'individus qui ont répondu à la première question:

- Pour: 23%;
- Contre: 66%;
- Sans opinion: 11%.

Groupe d'individus qui ont répondu à la seconde question:

- Pour: 61%;
- Contre: 25%;
- Sans opinion: 14%.

Vérification et analyse

Si pas déjà dans un fichier informatique, alors il faut procéder à une transcription des données.

- Codification. Souvent des logiciels comme SAS ou SPSS préfèrent une information chiffrée. Exemple: variable Sexe, $1 = H$, $2 = F$.
- Vérification (dans la mesure du possible) si l'information est cohérente. Exemple: né en 2010 et possède un permis de conduire.
- Traitement des valeurs manquantes.

Imputation (SSW, Chap. 15)

L'information peut être manquante.
Le questionnaire pourrait être non-rendu.
Il pourrait également y avoir des trous dans le questionnaire.

- Non-réponse par item: au moins une question est répondue mais pas des réponses à toutes les questions.
- Non-réponse par unité: la personne ou l'unité ne donne aucune réponse.

En présence de non-réponse, il peut être envisagé de procéder à de l'imputation. Imputation: ensemble de méthodes pour boucher les trous.

Méthodologie dans un recensement

Il est intéressant de noter que si un recensement est entrepris, alors beaucoup de la méthodologie des sondages doit être mise en oeuvre.

- Base de sondage à créer, valider;
- Personnel à former pour les entrevues;
- Soucis (comme la non-réponse) qui peuvent survenir.

En fait, les sources d'erreurs non-dues à l'échantillonnage dans les sondages peuvent survenir lors d'un recensement.

On doit choisir un estimateur pour chaque paramètre à estimer de la population finie.
Exemples:

- Moyenne: $\bar{y}_U = \frac{1}{N}\sum_U y_k$;
- Total: $t_{yU} = \sum_U y_k$;
- Variance: $S^2_{yU} = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2$.

On note qu'il fait du sens de parler d'un total car la population est de taille N. On note que N est également un paramètre.

Estimations ponctuelles

Une estimation ponctuelle consiste en un seul nombre dont l'objectif est d'estimer un paramètre.
Pour la moyenne de la population U notée $\bar{y}_U$, un estimateur pourrait s'écrire:

$$\hat{\bar{y}}_U$$

Une seule valeur n'est habituellement pas suffisante pour apprécier la qualité de l'estimation. On a souvent besoin de la variance de l'estimateur, notée:

$$V(\hat{\bar{y}}_U)$$

La variance théorique est rarement calculable sur un échantillon. Il faudra estimer la variance:

$$\hat{V}(\hat{\bar{y}}_U)$$

Estimateurs

En fait, le nombre qui sert à estimer le paramètre inconnu est un **estimateur**.
Un estimateur est une **variable aléatoire**.
Pour chaque échantillon possible, l'estimateur prend une certaine valeur.
La distribution de l'estimateur est obtenue sur l'ensemble de tous les échantillons possibles.
Même s'il est habituellement grand, le nombre d'échantillons possible est **fini**.
On rappelle que la population est finie de taille N.

Variance versus estimateur de variance: V versus $\hat{V}$

Un estimateur a une variabilité qui est quantifiée en calculant sa variance sur l'ensemble des échantillons possibles.
Idéalement c'est la variance de l'estimateur

$$V(\hat{\bar{y}}_U)$$

que l'on aimerait utiliser.
Comme elle est calculée sur l'ensemble des échantillons, elle est habituellement inconnue et doit être estimée.

Estimation par intervalles de confiance

Un intervalle de confiance de niveau 95% pour le paramètre moyenne est donné selon la formule:

$$\hat{\bar{y}}_U \pm 1.96\{\hat{V}(\hat{\bar{y}}_U)\}^{1/2}$$

Quelques questions:

- Dans quel contexte cet intervalle a été vu?
- Quelles étaient les conditions sur l'échantillonnage?
- Y avait-il des conditions sur les distributions?
- Était-il toujours exactement de niveau 95%?

5. Publication des résultats

Dernière étage du sondage qui consiste à publier les résultats. On devrait retrouver les éléments suivants:

- Conditions de la réalisation du sondage: base de sondage, population visée, plan d'échantillonnage;
- Grandes lignes de conduite: précision visée, taille de l'échantillon;
- Discussion des différentes sources d'erreurs: erreurs non dues à l'échantillonnage, dues à l'échantillonnage; décisions prises pour en tenir compte.
- Comment la non-réponse a été traitée.

Présentation sous forme de rapport.