Introduction à la théorie des sondages

Echantillonnage

- Utilisé pour décrire une **population** dont on ne connait pas toutes les caractéristiques.
- Dans ce cas, on sélectionne une partie de la population (**échantillon**) pour obtenir plus d'information.
- Si l'échantillon est représentatif et suffisamment grand, on peut extrapoler ses résultats à la population

Estimateurs

- Ce sont des valeurs numériques obtenues sur l'**échantillon** servant à se donner une idée de la valeur d'indicateurs de la **population** (moyenne, variance, corrélation, proportion, etc)

Estimateurs (exemples)

- On estime la moyenne de la population μ par la moyenne de l'échantillon $\bar{x}$
- On estime la variance de la population σ^2 par

$$n/(n-1) * S^2$$

où n et S^2 sont respectivement la taille et la variance de l'échantillon

Estimation d’une proportion

On estime la proportion dans la population p par la proportion dans l’échantillon $\hat{p}$

=> base de la théorie des sondages

Marge de fluctuation

Probleme: la proportion dans l'échantillon $\hat{p}$ varie aléatoirement

Pour un sondage donné, on n'est pas sur qu'elle est proche de la bonne valeur p

Par contre, on peut calculer une **marge de fluctuation** qui va indiquer dans quel intervalle devrait se trouver la valeur de la population

Marge de fluctuation (exemple)

On fait un sondage sur l’opinion des étudiants par rapport aux prix de la cafétéria

Dans le sondage, 45% en sont mécontents.

Si la marge de fluctuation est de 3%, alors la proportion de mécontents dans la population devrait se trouver entre 42% et 48%

Marge de fluctuation (calcul)

Pour une technique d’échantillonnage simple, elle dépend de:

- La taille de l’échantillon;
- Le degré d’erreur du sondage;
- L’homogenéité de la population.

Taille de l’échantillon

- Plus l’échantillon est grand, plus la précision augmente (et donc, plus la marge de fluctuation diminue).
- Le rapport entre la taille de l’échantillon et la marge de fluctuation est de $\sqrt{n_1 / n_2}$
- Par exemple, si on multiplie la taille par 2, la marge de fluctuation sera divisée par $\sqrt{2}$ soit environ 1.41

Degré d’erreur

Chaque estimation a un degré d’erreur associé. En effet, il se peut que l’on ait sélectionné par hasard un échantillon “extrême” complètement atypique.

Il est impossible de réduire complètement cette erreur, sinon les résultats seraient inutilisables. On va donc se fixer un degré d’erreur acceptable (en général 5%)

Plus le degré d’erreur est faible, plus la marge de fluctuation augmente

Homogenéité de la population

Plus la population est homogène par rapport au caractère observé et plus la marge de fluctuation sera faible, car moins on laisse de place au hasard.

En général, le degré d'homogenéité n'est pas connu puisque il dépend justement de ce que l'on veut mesurer.

En pratique, on essaiera de le deviner en se basant sur l'intuition, des résultats partiels ou en supposant que la population est complètement hetérogene ("worst case scenario")

Marge de fluctuation (formule)

Pour un degré d'erreur égal a 5%, et pour une taille de population suffisamment grande (au moins plusieurs centaines) on a la formule suivante pour la marge de fluctuation:

$$1.96\, p^{*}(1-p^{*})/\sqrt{n}$$

où p^{*} représente la proportion choisie pour l'hypothese d'homogenéité (pour le worst case scenario, on a $p^{*}=0.5$)

Choix de la taille d’échantillon

En fait, lorsque l’on effectue un sondage, on pose souvent le problème différemment. On cherche la taille d’échantillon minimum pour obtenir une certaine marge de fluctuation (sachant que le degré d’erreur et le degré d’homogenéité sont fixés).

Dans ce cas (pour un degré d’erreur de 5%), la taille est:

$$p^{*}(1-p^{*})*(1.96/MF)^{2}$$

où *MF* est la marge de fluctuation.

Choix de la taille d'échantillon

Par exemple, si on veut faire un sondage pour l'élection présidentielle francaise avec une marge de fluctuation de 2%, on devrait avoir un échantillon d'au moins 2401 personnes (worst case scenario, échantillonnage simple).

ATTENTION: en général, la taille de l'échantillon ne dépend pas de la taille de la population (pour des populations suffisamment grandes). Un échantillon de 1000 personnes pour une enquete à Geneve est tout aussi correct qu'un échantillon de 1000 personnes pour une enquete sur toute la France !!!

Sélection de l’échantillon

- Sélection aléatoire (échantillonnage simple): on prend des personnes completement au hasard (souvent en se basant sur des registres téléphoniques)
- Sélection systematique: on se base sur une liste et on choisit par exemple la 100-ème, 200-ème, 300-ème, etc personne. Probleme: peut introduire un biais selon comment la liste a été organisée.

Sélection par strates

Si la population est partagée en groupes, on construit un échantillon final à partir d'échantillons simples tirés dans chaque groupe. Le résultat final devra etre pondéré par l'importance relative de chaque groupe.

- Exemple: je veux prendre un échantillon de 100 étudiants. Or je sais que les étudiants étrangers représentent 1% de la population. Si je procède par sélection simple, j'en aurais seulement (en moyenne) 1 dans l'échantillon (pas assez). Je peux donc décider de procéder par strates en sélectionnant 2 sous-échantillons et en "sur-représentant" les étrangers. Dans ce cas, je sélectionnerais (par exemple) 80 suisses et 20 étrangers. Mon résultat final sera une moyenne des proportions de chaque groupe pondérés par 99/100 (Suisses) et 1/100 (étrangers)
- Avantage: plus précis que la sélection simple
- Inconvenient: suppose que l'on connaisse la distribution de la population

Sélection par quotas

Cas particulier de la sélection par strates. On suppose que l'échantillon doit avoir la même structure par groupe que la population (ex: 50% de femmes, 10% d'ivrognes, etc). C'est en général la méthode utilisee pour les enquêtes d'opinion.

- Avantage: plus précis que la sélection simple, peu couteux
- Inconvénient: peut introduire un biais selon la structure imposée

Sélection par grappes

On suppose que la population est composée de “grappes” (unités) semblables entre elles. Au lieu de faire une sélection simple sur la population, on fera juste une sélection sur certaines grappes uniquement.

- Exemple: je fais un sondage dans la rue sur l’opinion politique des habitants du canton de Genève. Pour réduire les couts, je décide d’interroger uniquement des gens à Carouge et à Meyrin.
- Avantage: moins couteux que la sélection simple, plus précis
- Inconvenient: biais si les grappes ne sont pas semblables

Sondages politiques

En général, effectués selon la méthode des quotas et par téléphone.

Problèmes: échantillons pas forcement representatifs (ex: téléphone ?). De plus, certaines personnes ne donnent pas nécessairement la “vraie” réponse

Les instituts procèdent alors à des **redressements.**

Redressements

Exemple du vote Le Pen en France.

En général, sous-evalué dans les réponses “brutes”.

Différentes techniques appliquées pour “corriger” les données brutes.

Redressement (sondages passes)

(Exemple fictif).

Les réponses brutes donnaient Le Pen à 8% en 2002. Il a fait 17% aux elections.

Les réponses brutes donnent Le Pen à 7% en 2007. Si l'on suppose que les gens ont le meme comportement par rapport aux sondages qu'en 2002, Le Pen devrait avoir 17% * (7/8) en 2007 (règle de trois).

Problème: le comportement des gens (ainsi que leur positionnement politique) peut évoluer dans le temps…

Redressement (votes passés)

(Exemple fictif).

Les réponses brutes donnent Le Pen a 7% en 2007. Dans le meme sondage, si l'on demande aux gens pour qui ils ont voté en 2002, 10% répondent "Le Pen".

Sachant que Le Pen a fait 17% aux élections en 2002, il devrait faire 7% * (17/10) en 2007 (regle de trois).

Problème: les gens oublient souvent ce qu'ils ont voté dans le passé…

Redressement (autres facteurs)

Utilisation de proxies: si les électeurs de Le Pen ont un certain profil, on peut les “détecter” en utilisant des questions non-directement liées au vote (par exemple: attitude par rapport à l’Europe, aux immigrés, à l’Economie, etc).

Probleme: assez incertain, des électeurs avec un profil similaire peuvent décider de voter pour des candidats différents, notamment si les deux candidats développent des themes similaires

Redressements

Problème général: chaque institut a ses propres techniques de redressement et les garde secrètes (secret professionnel)

Théoriquement, il est donc difficile de vérifier s'il n'y a pas "manipulation" des résultats.