Méthodologie
de l'enquête de terrain

Amor BELHEDI

2011

SOMMAIRE

Introduction

La collecte des données constitue la première étape de toute analyse après avoir précisé le problème, défini la problématique et fixé les hypothèses de la recherche.

L'enquête constitue souvent la seule méthode pour disposer de données appropriées mais l'analyse des différentes sources en est un préalable pour pouvoir définir les besoins de l'enquête. Ce n'est qu'après avoir pris connaissance des données disponibles, potentielles (données non accessibles directement et nécessitant un traitement) ou manquantes qu'il faut procéder à la collecte de l'information sur le terrain. Ce tour d'horizon des différentes sources est incontournable pour pouvoir définir convenablement la problématique, les méthodes d'analyse et la nature de l'information requise à demander.

On examinera dans un premier point les sources d'information avant d'aborder l'enquêtre proprement dit, les sondages et l'estimation.

On oublie souvent que l'enquête a pour finalité de pouvoir généraliser le résulat ontenu à partir d'un échantillon représentatif faute de pouvoir analyser l'ensemble de la population concernée. Pour cela, la représentativité de l'échantillon constitue la clef. C'est pour cette raison qu'on consacrera tout un chapitre aux sondages, aux types de sondages et aux méthodes de choix des unités à enquêter.

La généralisation des données obtenues au niveau de l'échantillon à l'ensemble de la population concernée est

incontournable, pour cela la conaissance des lois de distribution d’échantillonnage est nécessaire.

Enfin, la taille de l’échantillon à enquêter constitue un autre volet central que le chercheur ne doit pas négliger pour garantir la représentativité minimale des données recueillies, on lui a consacré le dernier chapitre.

Chapitre 1

LES SOURCES DE L'INFORMATION

On peut distinguer plusieurs sources d'information dont on peut citer: les études et les recherches, les documents administratifs, les cartes et les photographies, les recensements et les enquêtes. Dans une analyse donnée, le recoupement des sources est incontournable et on ne doit jamais se fier à une seule source quelque soit son degré de fiabilité.

1 - Les documents administratifs

Les sources administratives couvrent tous les documents produits ou commandés par l'administration dans son sens le plus large (Etat, Ministères, offices, banques, Sociétés, Collectivités, Associations...). Les documents sont établis pour des fins purement administratives suivant des dispositions réglementaires et dans des buts précis de contrôle, de suivi ou de gestion.

Ces sources sont souvent faciles à en disposer mais elles ne répondent pas toujours aux besoins spécifiques de l'analyse escomptée, leur utilisation doit être très prudente et dans tous les cas recoupée avec d'autres sources dans la mesure où on peut relever des erreurs.

Les sources administratives sont très variées et on peut distinguer : les comptabilités, les rapports d'activité, les fichiers, les archives et les annuaires...

a- Les comptabilités

Les comptabilités sont des documents établis régulièrement par divers services dans un objectif de contrôle, de suivi et de gestion. Les

services de contrôle, de promotion et de gestion tiennent une comptabilité permanente pour pouvoir suivre l'activité, dresser un bilan ou contrôler les bénéficiaires tel est le cas du service des mines (permis, carte grise), de la construction (permis de bâti, lotissement), l'Office de l'Emploi (demande, offre et placement), les services fiscaux (patente, impôts...) et douaniers (importation, exportation), c'est aussi le cas de l'API, APIA, FOSDA, FOSEP, FONAPRAM, FNAH,

b- Les rapports d'activité

La plupart des sociétés, agences, offices, Ministères et banques établissent des rapports d'activité annuelle ou occasionnelle qui contiennent d'importantes informations sur leur activité, les réalisations, la stratégie suivie et les moyens mis en oeuvre.

c- Les fichiers

Certains organismes détiennent des fichiers permettant le contrôle et le suivi de l'activité, c'est le cas de l'Agence de Promotion de l'Industrie (API), celle de l'Agriculture (APIA), du service des mines... L'analyse de ces fichiers lorsqu'ils sont accessibles permet de disposer de données très précieuses qu'on ne trouve nulle part. C'est le cas par exemple du fichier des salariés lorsqu'on veut analyser le lieu de résidence, la taille du ménage, la qualification, l'ancienneté, l'âge...

On peut citer à ce niveau le Fichier des Entreprises Non Agricoles (FENA) établi par l'Office puis de l'Agence de l4Emploi qui fournit des indications sur l'activité, l'emploi, l'emploi des établissements non agricoles de plus de 10 salariés

d- Les archives

Ce sont tous les documents administratifs anciens qui ont été gardés et permettant de suivre l'évolution de certains faits dans le temps.

e- Les annuaires

Ce sont des documents qui contiennent d'importantes informations, on peut citer les annuaires téléphoniques, les pages jaunes, l'annuaire agricole, industriel ou économique, l'annuaire statistique...

f- Les rapports et les études

Ce sont des études élaborées ou commandées par l'administration pour comprendre un phénomène, permettre le choix ou la prise de décision dans un domaine donné. On peut citer les plans d'aménagement, les études stratégiques, les études d'impact ou les études de projet...

Les sources d'erreur sont nombreuses et l'utilisation doit être très prudente. Ces sources sont indispensables mais pas suffisantes et doivent être confrontées à d'autres sources.

2 - Cartes, photographies et images

Pour l'étude de l'espace, il est important de localiser et de suivre l'évolution spatiale des faits géographiques si bien qu'une des principales sources est représentée par la carte sous toutes ses formes et échelle: carte topographique, géologique, pédologique, carte d'occupation du sol, carte d'érosion.

La photographie constitue une importante source d'information pour représenter un témoin de l'observation, elle peut être latérale, oblique ou aérienne.

Enfin l'image satellitaire conquiert, de nos jours, du terrain et est devenue un instrument indispensable pour l'analyse spatiale: occupation du sol, dynamique spatiale, pollution.

3 - Les recensements

Le recensement est une enquête exhaustive portant sur toutes les unités1, c'est une opération très lourde surtout que la population concernée est très nombreuse. Il mobilise des moyens considérables tant financiers que humains, exigent de longs délais de préparation, d'exécution et de dépouillement. Il faut presque une année pour le préparer et une autre pour sortir les premiers résultats malgré les progrès considérables dus à l'informatisation et aux techniques de sondage, le recensement de 1975 a coûté 1.8 MD (l'enquête de 1980 300.000D).

C'est ce qui explique que le recensement se fait périodiquement tous les dix ans 1956, 1966, 1975, 1984, 1994, 2004[1]. L'importance des coûts fait qu'on souvent recours à des enquêtes intermédiaires (1980 et 1989).

C'est dans le domaine démographique que les recensements on été les plus précoces[2] et les plus fréquents répondant ainsi à des impératifs stratégiques, fiscaux ou de planification. Le premier recensement français date de 1801, en Tunisie les dénombrements datent de 1921 tandis que les recensements datent de 19562 : c'est le recensement général de la population et des logements qui a pour base le ménage, il touche les caractères démographiques (taille, structure, émigration, âge, sexe, état matrimonial...), éducationnels (niveau scolaire, fréquentation,..) et économiques (activité, statut, branche, secteur, chômage..) et le logement (nature, pièces, état, confort...).

Sur le plan des activités, on peut citer le recensement industriel : porte sur les établissements industriels et leurs activités, le premier fait en France date de 1963, en Tunisie c'est en 1967 que fut élaboré le premier recensement industriel par l'INS. Ce recensement donnait lieu au début à une publication détaillée qui fut abrégée par la suite pour disparaître totalement.

En 1977-78, l'INS a procédé à un recensement des établissements privés non agricoles, en 1981-83 un recensement des activités, plus exhaustif cette fois-ci, a été élaboré mais n'a pas donné lieu à une publication.

Le Fichier des Entreprises Non Agricoles (FENA), établi par l'office de l'emploi constitue un recensement depuis 1977 des entreprises non agricoles de plus de 50 salariés. Sur un autre niveau, l'Enquête Suivi de l'API depuis 1973, constitue un recensement des projets agréés par l'API mais la libéralisation de l'agrément depuis quelques temps a limité cette enquête seulement aux projets demandant des avantages.

---

[1] Avant 1946, il s'agissait plutôt de dénombrements, le recensement ne concernait que la population européenne entre 1921 – 1936

[2] Le premier recensement dans l'histoire est celui de Jules César en l'an 0.

Ces recensements constituent aussi une base de sondage pour des enquêtes plus limitées.

5 – Les enquêtes

Ce sont des mini recensements, des études sur échantillon surune population donnée souvent menés par des organismes officiels comme l'INS, l'API, l'Agence de l'Emploi permettant d'analyser l'état des lieux dans un domaine particulier. L'exemple des enquêtes consommation des ménages menées par l'INS tous les cinq ans est significatif, depuis 1967, l'INS procédait périodiquement à une enquête de consommation : 1975, 1980, 1985, 2000, 2005. L'INS procédait aussi à une enquête population–emploi tous les cinq ans au début, puis tous les deux ans depuis les années 2000 : l'enquête 1980, 1989 en constituent des exemples.

Des enquêtes non périodiques existent aussi, c'est le cas par exemple de l'enquête migration-emploi à Tunis menée en 1972 par l'INS, c'est le cas aussi des enquêtes sur le secteur non structuré dans les années 1980.

Un autre exemple est donné par l'enquête suivi de l'Agence de Promotion des Investissements Industriels (API) depuis 1973, c'est une enquête annuelle que menait l'API pour assurer le suivi des agréments industriels. Depuis la libération des investissements, l'enquête ne concerne que les projets qui demandent des avantages.

6 – Les études et les publications académiques

Elles renferment les articles publiés dans les revues spécialisées, les mémoires et les thèses présentées dans les différentes universités ainsi que toute la documentation académique.

7 – Le document objet d'étude

On peut avoir affaire aussi à l'étude d'un document donnée comme le cas d'un manuel scolaire, le programme d'enseignement, des archives ou tout autre document. Dans ce cas, l'approche doit être externe et le référentiel indépendant du document pour pouvoir apporter des éclairages adéquats.

Le nécessaire recoupement des données

Chaque source statistique reflète les objectifs qui ont régis sa mise en oeuvre et il n'est pas rare non plus qu'elle se trouve entachée d'erreurs provenant souvent de simples oublis ou faute de suivi et de contrôle dans les différentes phases de l'élaboration.

Pour cela, il faut veiller toujours à utiliser plusieurs sources à la fois et ne jamais se fier à une seule, lorsque cela est possible.

Deux types de recoupements sont nécessaires :

- Le recoupement interne : à ce niveau, c'est l'analyse logique du texte, des conclusions et des chiffres qui permet de détecter les erreurs imputables parfois à des fautes de frappe mais pas toujours, à une fausse interprétation ou à un manque de suivi.

- Le recoupement externe : par la comparaison de diverses sources, il n'est pas rare de trouver des données voire des conclusions contradictoires si non différentes. A ce niveau le choix s'impose et sa justification est plus que nécessaire.

Il ne faut jamais se fier à une source d'information.

Chapitre 2

L'ENQUETE

L’enquête est l'étude d'une population à partir d'un sous-ensemble représentatif, appelé échantillon, en vue d’une généralisation des résultats obtenus sur l’ensemble de la population d’origine ou population-mère. On l'appelle aussi sondage[3] ou enquête par sondage.

On peut distinguer deux types d’enquête : l’enquête par sondage et l’enquête partielle

- L'enquête par sondage: c'est une enquête sur un échantillon représentatif de la population concernée, tiré selon des règles définies permettant la généralisation des résultats recueillies à l'ensemble de la population. Elle est appelée aussi sondage

- L'enquête partielle: C'est une enquête sur un sous-ensemble limité en nombre et représentant l'essentiel de la population concernée, c'est le cas d'une enquête industrielle qui touche un nombre réduit d'unités regroupant la majorité de la main d'oeuvre. On utilise ce type d'enquête partielle lorsque la taille ne détermine pas le comportement et la population considérée est homogène.

On l’utilise aussi quand la population concernée est infinie comme est le cas par exemple de l’étude de la pollution de l’air ou de la mer.

[3] Le terme sondage exprime aussi la manière dont ont fait le choix des individus à enquêter, l’échantillon retenu pour l’analyse.

L'enquête, sous réserve d'être représentative, permet d'avoir des données rapidement à des coûts réduits, une richesse de l'information demandée et une souplesse élevée dans les concepts et les champs d'étude.

L'enquête s'impose dans les cas suivants lorsque :

- il y' a un grand risque de détruire l'unité: c'est le cas des tests de résistance ou de durée de vie en particulier dans l'industrie.
- le phénomène étudié est limité et invisible ou inconnu, c'est le cas des maladies
- la population est infinie ou nombreuse comme est le cas de la pollution atmosphérique ou de la consommation.
- l'étude s'intéresse à la structure, c'est le cas des enquêtes agricoles ou de consommation.

L'intérêt de l'enquête réside dans la possibilité de la généralisation des résultats, la richesse de l'information et le coût bas. C'est là aussi que résident ses limites:

- le tirage de l'échantillon
- la mesure de l'erreur d'échantillonnage et de là, précision des résultats obtenus.

Quelque soit le type d'enquête, il y a lieu de définir la problématique générale, les objectifs et le champs de l'enquête avec précision. Les résultats en dépendent fortement.

2.1 - Problématique et objectifs

La problématique générale de l'étude à mener fixe en quelque sorte les objectifs poursuivis à travers l'enquête et son champ d'investigation. Ces objectifs doivent être définis avec précision et dans les moindres détails, ils déterminent le type d'information à demander et de là, la formulation des questions à poser.

L'unité de base doit être définie avec précision pour éviter les ambiguïtés au même titre que le champ et les concepts utilisés. Si on veut demander le revenu, il faut définir quel type de revenu s'agit-il ?: revenu salarial, brut, net, incluant les allocations familiales et le prime de rendement que les enquêtés ont toujours tendance à exclure... Si on

veut étudier la population urbaine, encore faut-il définir l'urbain et ses limites...

Le champ spatial et temporel doit être précisé et défini. Pour les données de mouvement comme les naissances, la migration ou les entrées, la période doit être limitée et précise: un jour donné, une année... Pour les données d'état qui consistent à établir des bilans, la date doit être précise : c'est le cas du recensement par exemple... Lorsque les délais sont longs, il ne faut pas que les variations soient importantes.

Au même titre, le champ spatial de l'enquête doit être défini avec une très grande précision.

2.2 - Le support de l'enquête

On peut distinguer deux types de supports qui peuvent être complémentaires: le questionnaire et l'interview.

a- Le questionnaire :

Le questionnaire est la série de questions posées directement ou indirectement à l'enquêté qui constitue l'unité de base.

Le questionnaire doit être à la fois, clair et aéré, facile à lire et sans ambiguïté, peu long et riche tout en s'attaquant à des problèmes complexes, c'est de ce compromis difficile que dépend le résultat.

Ce paradoxe pose le problème du choix de questions à poser, de leur priorité ce qui exige que les objectifs soient définis et la problématique précisée.

L'exploitation doit être facile si bien qu'il faut toujours penser à la manière dont les réponses vont être exploitées pour pouvoir à la fois bien formuler les questions et saisir leur opportunité.

Aussi bien les questions que les réponses doivent être sans équivoque et sans ambiguïté, la formulation simple, adaptée aux

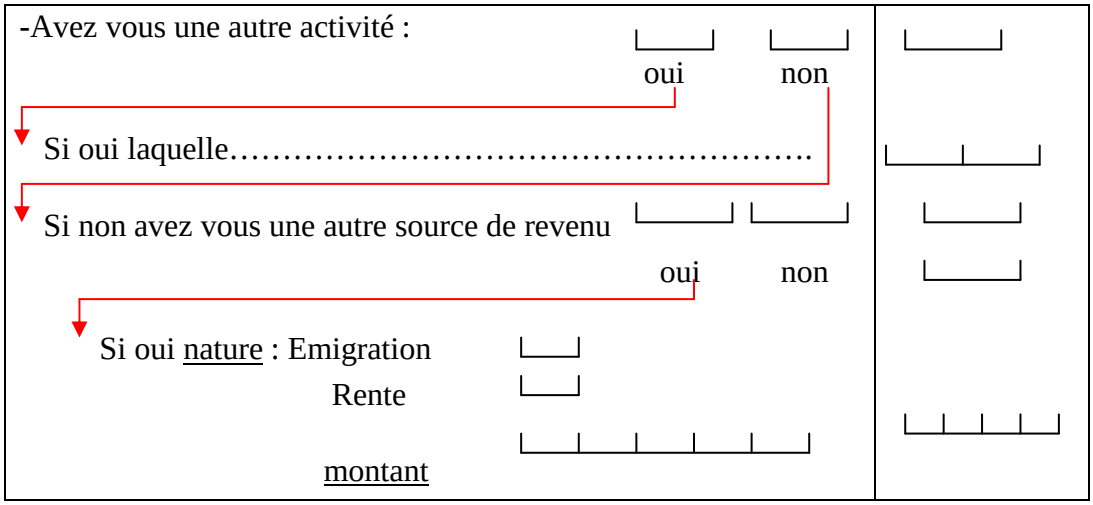
concernés tout en évitant les termes savants, pédant, recherchés, vagues ou polysémiques.

Il faut éviter la double alternative comme est le cas de la question suivante: votre logement est-il pire ou meilleur qu'il y a un an?: Oui/ non. C'est ainsi que des termes comme *sac*, *ouiba* ou *saa'*, souvent ou collectif peuvent avoir des sens différents selon les régions et les enquêtés. Les termes abstraits sont à éviter ou à clarifier, c'est le cas par exemple du *ménage*, *sous-location*....

Les questions longues et complexes sont à éviter dans la mesure où l'enquêté peut facilement perdre le fil de ses idées et répondre à aveuglement.

S'il est intéressant d'avoir des réponses, il importe encore plus de tirer des résultats, le souci de la facilité d'exploitation doit être présent dès le début dans l'élaboration des questions. C'est souvent ce qu'on oublie pour se rendre compte après enquête que telle ou telle question ne peut pas être exploitée convenablement...

Le questionnaire doit avoir un enchaînement logique et progressif: du général au particulier, du commun au privé, du facile au complexe... Il faut éviter les sauts ou les va et viens entre les rubriques.



*** Les types de questions**

On peut distinguer plusieurs types de questions selon le niveau de réponse, la forme et l'objectif :

Selon le niveau des réponses on a trois formes :

- Les questions à réponses numérique : la réponse est une valeur simple et concrète , c'est le cas de l'âge ou du revenu..
- Les questions alternatives: ce sont des questions dont la réponse est une alternative comme le sexe, l'occupation, la maladie.
- Les questions à voies (à choix) multiples : QVM ou QCM: la réponse est plus complexe c'est le cas de la profession, du lieu, d'une attitude ...

Selon la forme des questions, on peut distinguer trois types de questions :

- Les questions ouvertes: l'enquêté n'est pas orienté et on lui donne la totale liberté de réponse. L'éventail de réponse est large ce qui présente uns richesse d'information mais difficile quant à l'exploitation. Ce type de questions est à utiliser là où l'éventail de réponse est élevé comme la profession ou lorsqu'on veut avoir le maximum d'information ou on cherche des données qualitatives. C'est le cas par exemple lorsqu'on demande les raisons du choix de localisation (à préciser) .....

- Les questions fermées: les possibilités de réponse sont fixées d'avance, l'enquêté est orienté dans ses réponses ce qui constitue une perte d'information mais offre en contre partie, une plus grande facilité d'exploitation. C'est le type de question qui convient lorsque la gamme de réponses est limitée ou quand on privilégie la facilité d'exploitation dans le cas de l'exemple précédent on a : héritage, prix acceptable, site, possibilité d'extension.

- Les questions mixtes : ce sont des questions qui combinent les deux types précédents, elles orientent l'enquêté dans le maximum de réponses qu'on juge les plus importantes ou les plus représentatives tout en lui laissant la possibilité de s'exprimer librement si jamais il a

une réponse à la quelle on n'a pas pensé. Il s'agit d'ajouter souvent une case : autre .....

Ces questions constituent un compromis entre la finesse de l'information et la facilité de l'exploitation et sont souvent conseillées. Si on reprend le même exemple, on a : héritage, prix acceptable, site, possibilité d'extension, autre (à préciser): .....

Selon l'objectif de la question on a deux types de questions:

- Les questions d'information : destinées à récolter les données.
- Les questions test (ou piège): ce sont des questions dont l'objectif est de tester la véracité des informations livrées et de les recouper avec d'autres réponses ou de pouvoir détecter ceux qui répondent au hasard. C'est le cas par exemple de la dépense pour recouper le revenu. En demandant le revenu, on a tendance à s'écarter un peu de la réalité pour diverses raisons mais lorsqu'on demande la dépense, l'enquête tente toujours de se reprendre pour faire coïncider les deux réponses.

b- L'interview

L'interview est une enquête semi-directive qui laisse à l'enquêté une grande liberté tout en le guidant au cours de l'entretien. Il s'agit plutôt d'axes d'entretien beaucoup plus que de questions proprement dites.

Ces axes doivent être définis dans un guide d'entretien pour rendre la comparaison possible et éviter de sortir trop du sujet.

L'interview est très riche en informations si bien qu'il est très lourd en exploitation, c'est pourquoi la taille de l'échantillon doit être limitée.

Le questionnaire et l’interview peuvent être complémentaire l’un à l’autre dans une même étude en joignant la dimension quantitative (représentative) et qualitative (limitée mais approfondie) en concernant la même population cible ou des populations différentes.

2.3 - Les sources d'erreur

Aussi bien pour le questionnaire que pour l'interview, les sources d'erreur sont multiples et diverses qu'on peut résumer comme suit :

- L'oubli : l'oubli pur et simple constitue une importante source d'erreur qui prend de l'importance avec le recul et la longueur de la période. Interrogé à 6 mois d'intervalle, un ménage ne peut dater à moins de 3 mois près l'achat d'une voiture effectué dans les 18 derniers mois. Il y' a donc tout intérêt à limiter la période d'étude.

- La confusion : Un ménage confond toujours ses enfants et tendance à donner une moyenne de la dépense de ses enfants. Cette confusion augmente avec le temps.

- La non sincérité : Par crainte, l'enquêté peut mentir. Ce manque de sincérité peut provenir de plusieurs sources : la crainte des ennuis, la réticence à l'égard des statistiques, de la fiscalité, de l'intrusion dans la vie privée et du dévoilement. Elle provient aussi du désir d'apparat et d'impressionner l'enquêteur, lui faire plaisir ou l'intéresser à son cas.

L'enquêté a tendance à faire taire tout ce qui est spécifique, ce qu'il considère comme anormal par rapport aux valeurs dominantes et intériorisées que ce soit au niveau des attitudes ou des paramètres socio-économiques : revenu, dépense, logement...

- La mauvaise période d'enquête : en choisissant mal la période d'enquête on peut passer totalement à côté, c'est le cas par exemple d'une enquête sur les déplacements scolaires en été, d'une étude sur les migrants à une période où ils sont à l'étranger.

- La mauvaise formulation des questions: la mauvaise formulation des questions, ambiguïté des termes utilisés ou leur caractère abstrait conduisent à des erreurs au niveau de l'interprétation des enquêtés et du chercheur même.

2.4 - La pré-enquête ou le test

Pour améliorer le questionnaire, on est amené à le tester auprès d'un nombre réduit d'unités avant de lancer définitivement l'enquête. Deux types de procédés peuvent être utilisés :

- Les études de motivations: consistent à faire parler l'enquêté longuement ou poser des questions indirectes.
- l'interview en profondeur : consiste à mener des interviews auprès d'un petit échantillon.
- l'animation de groupe: elle est intéressante mais gênante.
- les études indirectes: l'enquêté ignore l'objet de l'enquête

- Le test : c'est la technique la plus utilisée, elle consiste à tester le questionnaire auprès d'un nombre réduit d'unités avant de lancer définitivement l'enquête. Le test a pour objectif d'améliorer la formulation, l'enchaînement des questions, détecter les défauts, éviter les sources d'erreur et réduire le nombre de non-réponses.

Avant d'entamer définitivement l'enquête, il y a lieu d'effectuer le test sur un nombre limité d'enquêtés (10 à 50 selon les cas). Ce nombre peut être récupéré par la suite moyennant quelques retouches.

Le test a pour objectif d'améliorer le questionnaire vise à :

- Vérifier l'opportunité des questions.
- Améliorer la formulation.
- Voir le contenu des réponses dans une perspective d'exploitation.
- Voir le bon emplacement des questions et leur enchaînement.
- Voir les raisons des nons-réponses à certaines questions.
- Définir les échelles de variation de certaines variables et de voir les réponses-types…

2.5 - Chiffrement et codage

Avant même d'effectuer l'enquête, il convient de penser au codage et à la manière dont le questionnaire va être exploité : manuelle ou informatique. Selon les possibilités d'exploitation dépend la quantité de questions et le type de traitement.

Le codage consiste à donner un code à l'information recueillie selon un référentiel typologique prédéfini permettant de faciliter l'exploitation des résultats. Le code doit être simple et d'interprétation souple et facile.

On peut distinguer plusieurs types de codes :

- le codes numérique: le code est un chiffre, le codage le plus utilisé est le code décimal, c'est le cas du code des branches d'activité économique de l' INS, le code des professions...

On a aussi le code de groupe qui consiste à regrouper les données et leur donner un code, c'est le cas par exemple du code de l'âge : 1 : 0-5, 2: 5-10...

- le code alphabétique : le code est un caractère, ce qui permet d'augmenter la capacité de codage: un alphabet de 26 ou 28 lettres au lieu de 10 chiffres: AA, AB, AF, ... AX, AABCD
- le code alphanumérique : il combine les deux systèmes de codes: AB10, BX60...

2.6 - Le plan d'exploitation

Le plan d'exploitation est l'ensemble des différents types de traitements à faire et des combinaisons de questions. Ce plan d'exploitation constitue le tableau de bord de traitement et il convient de l'établir avant même de lancer l'enquête ce qui permet d'améliorer les performances et éliminer les questions inutiles.

Le plan doit comporter les types de traitement et d'analyse de données avec les questions correspondantes. En gros, on peut distinguer les traitements suivants :

- Le listing

- La somme, les moyennes, les paramètres statistiques caractéristiques
- Le calcul de fréquences ou des pourcentages (en lignes ou en colonnes)
- Le croisement de données sous forme de tableaux avec ou sans filtre
- L'analyse de corrélation et de régression
- L'analyse multivariée des données (régression multiple, analyse factorielle…)
- L'analyse classificatoire et typologique

Des préalables

Quelque soit l'outil choisi, trois préalables sont indispensables : la présentation de l'enquête, la mise en confiance et le test.

- La présentation de l'enquête

L'enquêté doit être informée de l'objectif réel de l'enquête, du statut de l'enquêteur et de son utilisation future.

- La mise en confiance

La qualité de l'information recueillie dépend du degré de confiance ou de méfiance instauré entre enquêté-enquêteur. A défaut de cette confiance, l'information se trouve biaisée.

Rien ne sert à induire l'enquêté en erreur ou lui cacher le but réel de l'étude. En expliquant clairement l'objectif, tout en garantissant le secret personnel (utilisation fiscale, ou policière…). L'enquêté ne peut guère cacher sa propre vérité ou essayer d'en cacher une partie.

- Le test ou la pré-enquête

Une pré-enquête est nécessaire pour tester le questionnaire, l'améliorer et le débarrasser des questions non indispensables, floues, ambigues ou inutiles.

Il s’agit d’administrer le questionnaire à un nombre réduit d’individus dans le but d’éviter les mauvaises formulations, les contresens et de limiter les sources d’erreur. Le test est de nature à permettre de voir pourquoi les enquêtés ne répondent pas à certaines questions, quitte à les éliminer ou en améliorer la formulation, déterminer le bon emplacement des questions, leur enchainement logique et progressif, éviter le va et vient entre les questions, de mieux ordonner les rubriques et les questions posées.

Le nombre de questions doit être limité et suffisant à la fois pour parfaire cette finalité. On admet souvent qu’une dizaine est suffisante pour répondre à la finalité assignée au test à moins qu’il s’agisse d’un champ d’investigation totalement nouveau.

De préférence, les unités choisies doivent faire partie de l’échantillon pour ne pas gaspiller l’énergie déployée. Dans tous les cas, les unités doivent être choisies au hasard pour éviter de biaiser la représentativité.

Chapitre 3

LES SONDAGES
CONCEPT, CATEGORIES ET TYPES DE SONDAGES

I - CONCEPT ET FONDEMENT DES SONDAGES

Un sondage est une enquête portant sur un sous-ensemble représentatif et dont les résultats sont généralisables à l'ensemble de la population. Il signifie aussi la méthode de choisir l'échantillon. Il s'appuie sur la notion de représentativité.

1- Le concept de sondage

Le sondage est une enquête sur un échantillon représentatif de la population mère, il est défini par un taux, un plan et une base de sondage. C'est aussi le procédé qui consiste à tirer l'échantillon, on dit faire un sondage ou une enquête par sondage.

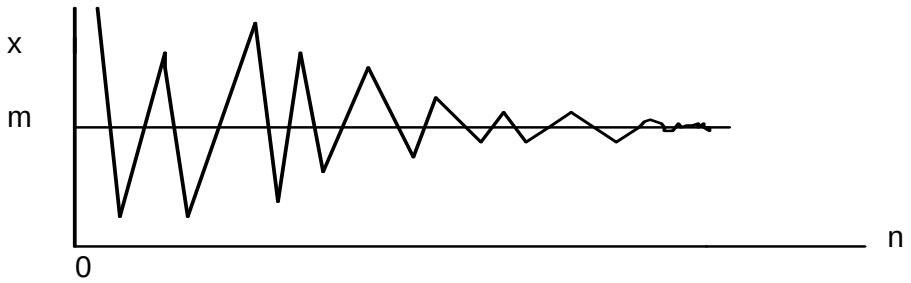
L'échantillon est un sous-ensemble représentatif d'une population, tiré selon des règles précises conformément à un plan de sondage précis donné et une base de sondage. *L'échantillonnage* est la méthodologie suivie pour déterminer l'échantillon: taille, base de sondage, plan de tirage...

La base de sondage est l'ensemble des critères ou variables de base servant à définir et à tirer l'échantillon. C'est aussi la liste exhaustive des unités à partir de la quelle se fait le tirage.

Le taux de sondage est le rapport entre la taille de l'échantillon (n) et celle de la population (N) $t = n/N$

2 - Le fondement

Le sondage trouve son fondement théorique dans ***la loi des grands nombres*** ou *la loi de convergence*. Lorsque la taille de l'échantillon n augmente, les valeurs observées dans l'échantillon tendent à converger vers les valeurs réelles de la population avec un certain risque déterminé. Ce risque d'erreur diminue lorsque la taille de l'échantillon augmente et tend vers zéro dans le cas d'un recensement.



Convergence de la valeur du paramètre échantillon vers la valeur réelle dans la population

Les lois de distribution d'échantillonnage permettent ainsi de fixer la taille optimale d'un échantillon pour un risque d'erreur fixé à l'avance.

3 - Les catégories de sondages : Sondage raisonné et sondage aléatoire

On peut distinguer deux grandes catégories de sondages: les sondages raisonnées et les sondages aléatoires.

a- Les sondages raisonnés

Un sondage raisonné est un sondage où les critères de choix et les unités sont raisonnablement choisis par le chercheur. Il présente la même structure que la population d'origine, l'échantillon en est *la miniature*. Le fondement de ce type d'échantillonnage est *la*

représentativité structurelle qui fait qu'en enquêtant un échantillon de même structure on peut généraliser les résultats.

La base de sondage est représentée par les variables, liés au phénomène étudié, de préférence observables pour pouvoir contrôler la structure de l'échantillon. Ces variables ont la même distribution que le phénomène étudié si bien qu'elles constituent des variables de contrôle.

Le sondage raisonné est simple et constitue le seul procédé à défaut d'une liste exhaustive des unités, c'est ce qui explique son utilisation fréquente. Utilisé avec *prudence*, il donne des résultats acceptables: > 10% et une bonne spartiale.

Les limites découlent de l'absence de fondement théorique solide au niveau de *la représentativité et l'absence de règles de choix* des unités. De là, la difficulté de mesurer la précision du résultat et de la généralisation. Il suppose une connaissance préalable de la structure

.

b- Les sondages aléatoires

Le sondage aléatoire est un sondage où le tirage se fait selon des règles précises à partir d'une liste exhaustive appelée *base de sondage*, il se fonde sur la loi des grands nombres. Les unités sont choisies aléatoirement en utilisant *la Table des Nombres au Hasard* (TNH).

La TNH est une table où les chiffres ont la même probabilité d'apparition qui se présente sous forme de 2 à 7 chiffres qu'on peut lire dans tous les sens (vertical, horizontal, diagonal), les ordres (premiers, derniers chiffres, chiffres alternés...) et avec un nmbre variable de chiffres (2 à 7). Il est conseillé cependant de choisir, pour la rapidité de l'opération, le même nombre de chiffres que la valeur extême de la population (N). L'utilisation de la TNH passe par les stades suivants :

- Numéroter la population à étudier de 1 à N.
- Etablir la taille de l'échantillon n
- Lire la TNH selon un ordre et un sens donnés en utilisant autant de chiffres qu'il y a dans la population.
- Relever les chiffres <=N qui apparaissent par ordre jusqu'à obtenir n (n: taille de l'échantillon).

- Pour disposer des unités de remplacement, on continue le processus de tirage de 25 à 33% unités supplémentaires. Le remplacement des défaillants se fait dans l'ordre des remplaçants.

Lorsque la taille de l'échantillon est suffisamment élevée, on peut voir apparaître une grande proportion constituée des mêmes unités qui se trouvent choisies par des méthodes différentes à l'aide de la TNH (premiers ou derniers chiffres, deux, trois ou quatre chiffres, lecture horizontale ou verticale).

Exemple :

Soit une population de 360 éleveurs, on veut enquêter le dixième, soit t = 0,1 = Dn = 36. N = 360, on peut utiliser trois chiffres seulement. En lisant la TNH dans n'importe quel sens et ordre, on retient les 36 premiers nombres rencontrés qui sont inférieurs ou égaux à 360.

Trois premiers chiffres, sens horizontal

159, 198, 189, 216, 7, 66, 71, 227, 208, 209, 127, 284, 250, 217, 275, 184, 97, 322, 236, 115, 336, 176, 181, 129, 177, 162, 47, 334, 2, 348, 158, 357, 179, 322, 331, 75.

Trois derniers chiffres, sens horizontal

230, 10, 106, 350, 98, 292, 259, 88, 6, 5, 44, 121, 343, 226, 357, 211, 129, 162, 194, 51, 238, 270, 314, 308, 163, 68, 102, 197, 331, 217, 312, 202, 262, 254, 250, 209.

Trois premiers chiffres, sens vertical

159, 217, 162, 158, 275, 271, 43, 146, 133, 30, 93, 313, 53, 80, 308, 211, 25, 208, 173, 178, 51, 1, 264, 151, 70, 249, 118, 136, 60, 191, 248, 66, 176, 357, 152, 304.

Trois derniers chiffres, sens vertical

259, 343, 129, 238, 207, 188, 17, 316, 200, 20, 349, 75, 65, 41, 10, 18, 140, 230, 88, 331, 297, 208, 258, 87, 1, 102, 318, 146, 289, 170, 328, 97, 155, 355, 96, 226.

Trois chiffres du milieu, sens vertical

259, 125, 172, 834, 312, 299, 270, 88, 336, 60, 31, 301, 75, 292, 137, 351, 320, 2, 244, 50, 138, 343, 307, 114, 308, 67, 33, 254, 259, 42, 101, 283, 10, 37, 128, 297.

On voit que quelque soit les sens ou l'ordre choisi, il y a un certain nombre de chiffres qui se répètent.

Le sondage aléatoire assure une très grande précision et permet de connaître le risque ce qui permette la généralisation des résultats. On peut estimer la valeur réelle des paramètres en définissant un intervalle de variation avec une probabilité connue.

II - LES TYPES DE SONDAGES

On peut distinguer plusieurs types de sondage indépendamment de la catégorie (raisonné ou aléatoire), le sondage élémentaire, systématique, parquota, stratifié, en grappes..

1 - Le sondage élémentaire

C'est un sondage sans contrainte majeure, il s'agit simplement de choisir n unités parmi la population N sans critère précis sinon la représentativité.

Dans un sondage aléatoire (S.A), chaque unité a la même probabilité d'être choisie, c'est un sondage sans remise. La méthode présentée ci-dessus permet de choisir les unités. Dans un sondage raisonné, il suffit de choisir n unités.

Exemple: Choisir 10 unités dans une popolation de 100. Pour un sondage aléatoire, il s'agit d'abord de numéroter la population de 1 à 100, fixer le sens et l'ordre de la leceture de la TNH, par exemple: les trois premiers horizontalement. Relever les éléments qui apparaîssent inéférieurs à 100 jusqu'à obtenir 10. Continuer le processus et relever 3 unités supplémentaires pour un éventuel remplacement des défaillants. Pour un sondage raisonné, on n'a pas de règle du choix, l'essentiel est de choisir 10.

2 - Le sondage systématique

Il s'agit de choisir les unités à intervalle régulier de manière à couvrir toute la population. La première unité (ou la base b) est choisie dans l'intervalle [1 - N/n] tandis que le pas de la progression arithmétique (ou raison r) est de r = N/n. Dans un S.A, la base est choisie dans la TNH selon le procédé indiqué ci-dessus. C'est un sondage plus facile, mais pose le problème de remplacement en cas de défaillance et ne convient pas aux phénomènes périodiques

Exemple: Choisir un échantillon systématique avec un taux de sondage de 1/10 pour une population de 100. Pour un SAS: On numérote la population de 1 à 100, on fixe le sens et l'ordre de lecture de la TNH: 2 derniers verticalement puisque la base est comprise entre 1 et 10 (N/n) qui est en même temps la raison: les unités choisies sont par exemple: 5, 15, 25, 35, 45,..95 . Pour un SRS: On choisit un chiffre entre 1 et 10 qui constitue la base puis on ajoute 10 pour les autres.

3 - Le sondage par quota

Lorsqu'on a une population hétérogène où on a plusieurs modalités ou classes estimées discriminantes (taille, forme) on peut assurer une répartition équitable en fonction de l'effectif de la strate (quota proportionnel) ou privilégier les petites strates en leur donnant plus de chance d'être enquêtées contrairement aux grandes strates où on peut se contenter d'un nombre réduit d'unités mais suffisamment grand pour permettre la généralisation (quota non proportionnel). Le choix des unités dans un SA se fait toujours à l'aide de la TNH selon les mêmes règles à l'intérieur de chaque strate séparément.

Dans *le quota proportionnel*, on a: Qi = Si .t avec t: taux de sondage global, Si: effectif de la strate i, Qi: quota de la strate i. Chaque strate et chaque individu a la même probabilité d'être choisis.

Dans *le quota non proportionnel*, on a: Qi = Si.ti avec t = (Σti.Si)/Pi et Pi: population totale, t: taux global de sondage.

Exemple: On a 10 gros propriétaires qui accaparent 60% du sol à côté de 500 petits exploitants qui n'ont que 15% des terres et on a fixé le taux de sondage à 1/10°.

Dans un quota proportionnel, on a 1 et 50 respectivement. Dans un quota non proportionnel, on peut enquêter tous les gros propriétaires (10) mais seulement 31 petits ce qui respecte le taux global de 1/10.

4 - Le sondage stratifié

Le tirage se fait à plusieurs niveaux ou degrés avec le choix au tirage des *unités primaires* (UP) dans un premier degré puis *les unités secondaires* (US) choisies dans les UP... Les unités échantillons sont *en cascade*. C'est le cas lorsque le tirage se fait selon plusieurs critères, la combinaison des critères définit les strates. On peut distinguer plusieurs types de sondages selon que les probabilités de tirage aux différents degrés sont égales ou inégales:

U. Primaires	U. Secondaires	
	P. Egales	P. Inégales
P. Egales	EE	EI
P. Inégales	IE	II

On peut citer l'enquête population-emploi et l'enquête consommation de l'INS. Dans ce cas, les UP sont définies par la combinaison Région (5) et le milieu (3). Dans ces UP, la combinaison taille du ménage (5) et activité de son chef (5) constitue les US.

5 - Le sondage par grappe ou aréolaire

Il constitue un cas particulier des sondages stratifiés notamment dans les sondages où l'espace constitue un paramètre important (région, quartier...), il contribue à réduire les déplacements sur le terrain. Ce type de sondage nécessite *un nombre réduit de types représentatifs*. Dans une enquête portant sur une ville, il vaut mieux enquêter le minimum de quartier représentatifs avec le maximum de ménages dans chaque quartier qu'un nombre réduit de ménages

répartis dans tous les quartiers ce qui réduit fortement les déplacements; ceci est encore plus pertinent lorsque le terrain d'enquête est trop vaste comme l'ensemble du pays: au lieu d'enquêter touts les villes, on dresse une typologie de 4 à 5 groupes où on choisit une ville.

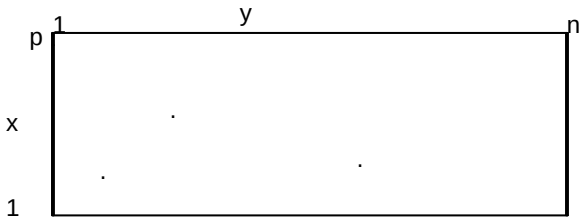
III - LES SONDAGES SPATIAUX

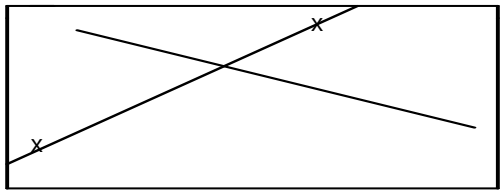
Dans l'espace, l'unité est bidimentionnelle et peut être de trois formes selon les besoins de l'analyse: ponctuelle, linéaire ou aréale. Sur une carte, une photographie aérienne ou un plan, on peut tirer un certain nombre de points ou de lieux à étudier. Il s'agit de graduer la carte horizontalement (X) et verticalement (Y) pour pouvoir tirer les points ou *les transects* (coupe) aléatoirement selon la TNH (Cf. P Haggett, 1973: L'analyse spatiale en géographie, AC).

.

1 - Le sondage élémentaire

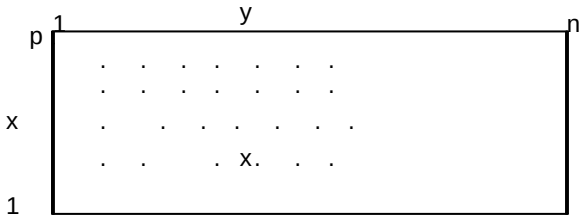
Dans un sondage élémentaire on tire les coordonnées des lieux (x, y) selon la TNH. Pour les phénomènes linéaires, on choisit au hasard deux points (x, y) pour chaque traverse. C'est le cas par exemple lorsqu'on veut étudier les prix fonciers dans un espace urbain.

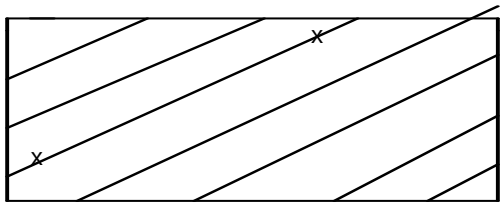




2 - Le sondage systématique

Dans un sondage systématique, seul le premier point ou la première traverse sont tirés au hasard, les autres sont choisis selon le même schéma pour couvrir l'ensemble de l'espace avec des points ou des traverses à des distances régulières.

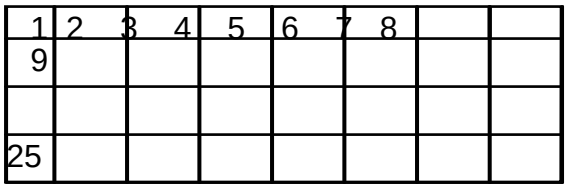


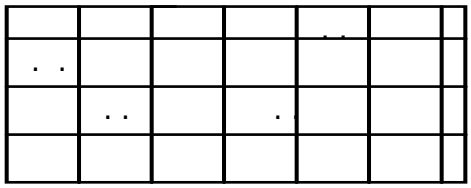


3 - Le sondage hiérarchique

C'est un sondage stratifié à deux degrés qui consiste à tracer des carrés et les numéroter, choisir les UP représentées par *les carrés de base* et à l'intérieur de ces carrés, fixer *les points* selon le même procédé de graduation et de tirage (avec un nombre égal ou variable par carré).

La grille peut être établie par les coordonnées d'un point tiré au hasard, les autres points sont choisis de telle manière qu'il y 'ait une grille de carrés où on tire un nombre égal de points selon la TNH.





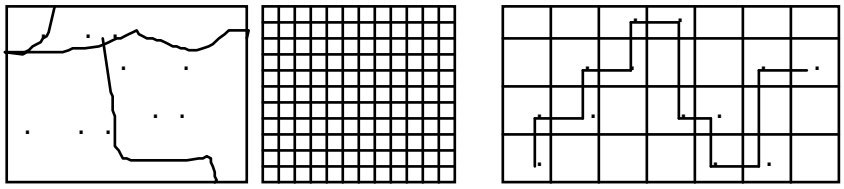
4 - Le sondage aréolaire

Il consiste à définir les strates constituées par des unités spatiales homogènes comme le type de sol ou les types de culture et tirer dans chaque strate un certain nombre de points selon la TNH. Ce nombre peut être égal ou variable. Ces strates sont des unités homogènes.

5 - Le sondage systématique stratifié sans alignement

C'est un sondage stratifié, systématique mais qui évite l'alignement. Il a été élaboré par Berry et comprend les étapes suivantes:

- Définir les différentes strates
- Déterminer une grille graduée de 0 à 9 (pour chaque strate)qu'on on place au dessus de la strate
- A l'aide de la TNH, on choisit un point P à partir du coin inférieur gauche et on y place la grille
- On fait varier y à l'aide de la TNH (alors que x est fixe) ce qui nous ramène à un second point P1 de même abscisse mais d'ordonné différent.
- On refait le même processus précédent pour faire varier non pas y mais x, ce qui donne le troisième point P2
- On répète les étapes jusqu'à remplir la grille.



Selon l'effet de la distance, on peut choisir le type adéquat de sondage, si la distance est aléatoire on peut utiliser le sondage systématique, pour un effet linéaire on utilise le sondage systématique ou stratifié. Lorsqu'on a des processus périodiques, on utilise le sondage stratifié mais quand l'effet de la distance est inconnu, il vaut mieux utiliser le sondage systématique stratifié non aligné.

Chapitre 4

LES DISTRIBUTIONS D'ECHANTILLONNAGE

Dans un échantillon, la valeur d'un paramètre statistique quelconque, comme la moyenne ou la variance, tend à suivre une certaine distribution bien déterminée et converge vers la valeur réelle du paramètre correspondant dans la popultion-mère d'où on a tiré l'échantillon lorsque la taille augmente, c'est ce qu'on appelle la *distribution d'échantillonnage*.

Sur cette propriété fondamentale, s'appuient les techniques de sondage pour fixer la taille minimale d'un échantillon à enquêter si on veut qu'il soit représentatif ou estimer la valeur réelle d'un paramètre de la population en fonction du résultat obtenu sur l'échantillon avec un risque connu d'erreur.

Les valeurs observées dans un échantillon tendent à converger vers les paramètres d'origine de la population mère avec une probabilité bien déterminée et un risque donné d'erreur ($1/t^2$). Certains paramètres tendent à être distribués selon une loi donnée comme la loi normale ou la loi de Student, on dit qu'ils ont une distribution *asymptotiquement* normale ou de Student...

1 - L'inégalité de Tchebicheff

L'inégalité de Bienaymé Tchebicheff (Cf. Chap.3) se trouve à la base de cette règle. Elle stipule que la probabilité qu'une valeur x de l'échantillon s'écarte de la moyenne réelle observée dans la population-

mère (m) de plus de (t) fois son écart type est au plus égale à $1/t^2$. Elle s'écrit comme suit: $P (| x - m | \geq t\sigma) \leq 1/t^2$
Si on pose $\varepsilon = t\sigma/(n)^{1/2}$ on a la relation suivante : $P (| x - m | \geq \sigma^2/n\varepsilon^2$

On constate dans cette inégalité que lorsque n tend vers l'infini, la différence entre les valeurs de l'échantillon et celle de la population tend vers zéro : $P (| x - m | \geq \varepsilon)$ tend vers 0.

Cette inégalité peut être généralisée aux différents paramètres statistiques ce qui permet de mieux définir la distribution d'échantillonnage et fixer le degré de précision.

Exemple: Une population renferme 40% d'analphabètes, on observe un taux de 35% dans un échantillon. Quelle est la probabilité d'observer un tel écart dans un intervalle de deux écart-types?. On a: $P | f - p | < | t (pq/n)^{1/2} | > 1 - 1/t^2$ ce qui nous donne $P |0.05 | < 0.48/n^{1/2}| > 1 - ¼ = 0.75$.

2 - La moyenne

La moyenne d'un échantillon suit la loi normale ou de Student selon deux facteurs : la nature de la population mère et la taille de l'échantillon:

- La moyenne x suit une loi normale de moyenne m et d'écart-type $\sigma/(n)^{1/2}$ lorsque la population mère est normale ou lorsque la taille de l'échantillon (n) dépasse 30.

- Elle suit une loi de Student à (n-1) degrés de liberté, de moyenne m et d'écart-type: $s^2/(n -1)$, lorsque n est inférieur à 30 ou σ est inconnu .

Loi de distribution de la moyenne selon la taille de l'échantillon et la nature de la popultion-mère

Conditions	Moyenne	Variance	Loi de Distribution	
$n \geq 30$ Population normale	m	σ^2/n	$N (m, \sigma/n^{1/2})$	$(x - m)/s/n^{1/2}$ suit N(0, 1)
$n < 30$	m	$\sigma^2/(n -1)$	T_{n-1}	$(x - m)/(\sigma/(n -1)_{1/2})$ suit T_{n-1}
s inconnu $s^2_x = \sigma^2/n = s^2/(n-1)$	m	$\sigma^2/(n -1)$	T_{n-1}	$(x - m)/(\sigma/(n -1)_{1/2})$ suit T_{n-1}
Population infinie	m	$\sigma(N - n)/ n(N - 1)$		

s^2_x: Variance observée dans l'échantillon. σ^2: Variance de la population, s^2: variance estimée de la population. N: Effectif de la population infinie.

Exemple: Une enquête sur 100 individus donne un âge moyen de 35 et un écart type de 12. Quelle est la distribution d'échantillonnage? Puisque n > 30, la moyenne suit la loi Normale N(m, $\sigma/n^{1/2}$). puisque l'écart type est inconnu on procède à son estimation: $s^2 = ns^2/(n - 1)$, d'où $S^2x = \sigma^2/n = s^2/(n - 1)$. La moyenne suit la N(m, $s/(n(n - 1)^{1/2})$, soit N(35, 0.1206). Puisque l'écart type est inconnu, la moyenne suit la loi de Student Tn-1.

Exemple: On a enquêté 10000 sur un total de 700000 ménages, on a obtenu une moyenne de 950 D/an de consommation et un écart type de 700. Quelle est la loi de distribution?. Il s'agit là d'un tirage exhaustif: n/N = 10/700 = 0.014 mais (N - n)/(n - 1) = 0.9857 ce qui est très proche de 1 et on l'assimile à un tirage indépendant. Puisque n > 30, on a x suit N(m, $\sigma/n^{1/2}$) avec Sx = 700/100 = 7, d'où on écrit: $(x - m)/(s/n^{1/2})$ suit N(0, 1).

Exemple: Un comptage routier a donné 1.33 véhicule/mn. Quelle est la distribution d'échantillonnage? Il s'agit d'une loi de Poisson de paramètre m = 1.33.

2 - La variance

La variance d'un échantillon suit une loi normale de moyenne : $(n-1)\sigma^2/n$ et d'écart-type égal à: $(2(n - 1))^{1/2}\ \sigma^2/n$.

On a alors : $(ns^2 - (n - 1)\sigma^2)/(\sigma^2\ (2(n - 1))^{1/2}\ /n)$ suit une loi N(0,1)

a- Cas général

Variance $(s^2) = (n - 1)((n - 1)U_4 - (n - 3)\sigma^4)/n^3$

Lorsque n est élevé, la variance (s^2) tend vers $(U_4 - \sigma^4)/n$

$(s^2 - (n - 1)/n)/(Vs^2)^{1/2}$ suit la loi normale N(0,1)

$(s^2 - \sigma^2)n^{1/2}/(U_4 - \sigma^4)^{1/2}$ suit la loi normale N(0,1)

b- Relation entre les moyennes (x, m) et les variances (σ^2 et s^2)

Le rapport entre une variable normale réduite : $(x - m)/s/n^{1/2}$ et la racine carrée d'une variable χ^2 indépendante: $(ns^2/(n - 1)s^2)^{1/2}$ suit une loi de Student à (n - 1) ddl. On obtient alors : $(x - m)(n - 1)^{1/2}/s$ suit Tn-1:

$(x - m)/s/n^{1/2}$ et $(ns^2/(n - 1)s^2)^{1/2}$

$(s^2 - \sigma^2)n^{1/2}/(U_4 - \sigma^4)^{1/2}$ suit la loi normale N(0, 1)

Dans le cas d'une distribution normale, on a : $n.s^2/\sigma^2$ suit une loi de χ^2 n-1

3 - La médiane

La distribution de la médiane est asymptotiquement normale pour toute distribution connue.

La médiane Me' suit une loi Normale de moyenne Me et d'écart-type égal à $(\pi s^2/2n)^{1/2}$ en cas d'une population normale, et $1/(4n.f^2(Me^*)$ en cas d'une autre distribution avec $f^2(Me^*)$: la valeur de la fréquence (densité de probabilité ddp) de la médiane de la population. Me: Médiane de la population. Me': Médiane de l'échantillon.

On peut écrire alors dans le premier cas de la loi normale que : $(Me' - Me)2n^{1/2}/\sigma\pi^{1/2}$ suit la loi N(0, 1)

Rapport entre médiane et moyenne

Lorsque n est élevé (il tend vers l'infini), le rapport entre la variance de la médiane-échantillon et la variance de la moyenne tend vers la valeur 1.57

Var Me'/Varx = $\pi\sigma^2$. $n/2n\sigma^2 = \pi/2 = 1.57$

4 - La fréquence

Dans un tirage sans remise on a la fréquence f = x/n avec une moyenne (p) et une variance : $(pq/n)^{1/2}$ avec f: fréquence observée dans l'échantillon, p: la proportion dans la population mère, q: 1 - p. On peut écrire alors selon l'inégalité de Tchebycheff la relation suivante: $P(| f - p | \leq t\ (pq/n)^{1/2} | \geq 1 - 1/t^2$

Lorsque n tend vers l'infini, la fréquence f tend vars p avec un risque d'erreur de $1/t^2$ et f suit la loi N(0, 1): $(f - p)/(pq/n)^{1/2}$ suit N(0, 1)

5 - Les valeurs extrêmes

a- Sur la base d'une distribution polynomiale (Cf. chap. 3), on peut déterminer les probabilités suivantes:

- la probabilité que (n -1) valeurs soient < x,
- la probabilité qu'une valeur x < xi
- la probabilité qu'il n'y a pas de valeurs > x + Δx

On peut alors écrire : $g(x_n) = n[f^{x_n}_{inf}\ px.dx]^{n-1}.f_{(x_n)}$

b- Dans le cas d'une distribution uniforme, on a la relation suivante :

$P(x < x \leq nx + \Delta x) = P(n - 1, 1, 0)$

$= n![P(x \leq x)^{n-1} P(x < X \leq X + \Delta x\ ]/((n - 1)!\ 1!\ 0!) = nx^{n-1} \Delta x$

$g(xn) = f_{nx^{n-1}}{}_{\Delta x\ 0}$

c- La plus grande valeur

La densité de probabilité (ddp) s'écrit : $W(u) = \alpha.Exp\ (-Exp^{-\alpha(U-Mo)}) - \alpha(U - Mo)$

$F(x) = W(u) = Exp\ (-Exp^{-\alpha(U-Mo)})$

u : Plus grande valeur, α : N(f(Mo)

$Mo = Me - 0.36651/\alpha = Me - 0.233(Q_3 - Q_1)$

$\alpha = 1.5725\ U/((Q_3 - Q_1) \quad \alpha = \pi/\sigma^4\sigma^{1/2}$

Cette fonction de répartition donne les effectifs théoriques: W(u) xn

$Uo = U - 0.45005\ \sigma^4$

$y = \alpha(U - Mo) \qquad Mo = U - 0.45005\ \sigma^4$

$\alpha = \pi/\sigma^4\sigma^{1/2}$

Test χ^2_{n-1} si $\chi^2_{obs} > \chi^2_{a,n-1}$: Ajustement retenu.

d- Le critère d'équiprobabilité

On découpe en intervalles de 0.5 = p

$p = W(u) = Exp\ (-e\ -y)$ d'où $y = -ln\ (-lnp)$

Np = p.n = Nombre d'observations pour lesquelles la plus grande valeur U est $\leq$ u

Les points y et u sont alignés sur une droite $y = a(U - Mo)$

U est déterminé par interpolation linéaire selon la méthode des moindres carrés

$y = \alpha(U\ -u) + y$

Exemple: Les températures maximales absolues annuelles ont été les suivantes. Ajuster les données au modèle de la plus grande valeur.

28	28.5	29	29.5	30	30.5	31	31.5	32	32.5	33	33.5	34	34.5	35	35.5	36	36.5	37	37.5	38
28.5	29	29.5	30	30.5	31	31.5	32	32.5	33	33.5	34	34.5	35	35.5	36	36.5	37	37.5	38	38.5
1	0	2	1	3	2	6	9	6	7	4	6	5	4	4	3	2	2	0	2	2

71

On a Q1 = 31.65, Q2 = 32.89 et Q3 = 34.66 ce qui nous donne α = 1.57254/(Q3 - Q1) = 0.522, Mo = Me - 0.3651/α = 32.19. On écrit y = α(u - Mo)= 0.522(u - 32.19) et y = 0.522u - 16.803.
On a Wu = $e^{-0.522u + 16.803}$. Cette fonction de répartition donne les fréquences théoriques et en multipliant Wu par 71, on obtient les effectifs théoriques. Le test de Khi-deux est significatif.

La méthode des moments donne y = 0.583u + 18.77. La méthode des moindres carrés nous donne Wu = e-0.547u + 17.624.

6 - La corrélation

Lorsque les deux variables x et y sont normales ou n est suffisamment élevé, le coefficient de corrélation linéaire suit une loi normale de moyenne r et d'écart-type égal à $(1-r^2)/n^{1/2}$. On a alors : $rn^{1/2}/(1 - r^2)^{1/2}$ suit N(0, 1)

Si n est faible, r suit une loi de Student de moyenne r et d'écart-type $(1 - r^2)/(n - 2)^{1/2}$. On a alors : $r(n - 2)^{1/2}/(1 - r^2)^{1/2}$ suit Tn-2.

Dans ce cas, la transformée de Fisher (z) suit la loi Normale de moyenne : 1/2ln((1 + r)/(1 - r) et d'écart-type égal à: $1/(n - 1)^{1/2}$. Elle s'écrit ainsi : z =1/2ln((1 + r)/(1 - r).

(z. 1/2ln((1 + r)/(1 - r))$(n - 3)^{1/2}$ suit N(0, 1)

$r^2(n - 2)^{1/2}/(1- r^2)^{1/2}$ suit T_{n-2} or $T_{n - 2} = F_{(1, n -2)}$

$r^2(n - 2)/(1 - r^2)$ suit $F_{(1, n - 2)}$

Exemple: On a relevé dans un échantillon de 400 individus une corrélation de 0.8, quelle est la loi de distribution de r? La corrélation suit la loi normale de moyenne r et d'écart type égal à: $(1 - r^2)/n^{1/2}$), soit r suit N(0.8, 0.018).

7 - Les coefficients de corrélation

Lorsque x et y sont des variables normales et n est élevé le coefficient de régression (a) suit une loi normale de moyenne (a) et d'écart-type : $\sigma y((1 - r^2)/n)^{1/2})/\sigma x$.

Lorsque n est faible, il suit la loi de Student à (n - 2) ddl de moyenne (a) et d'écart-type : $\sigma y((1 - r^2)/(n - 2))^{1/2})/\sigma x$.

Le coefficient b suit une loi Normale de moyenne (b) et d'écart-type : $\sigma y((1 - r^2)/n)^{1/2}$.

Souvent on est amené à faire des enquêtes pour collecter les données sur le terrain. Selon quelles règles collecter cette information, comment fixer la taille de l'échantillon et comment procéder à l'estimation des données à partir de l'information recueillie? Autrement, comment généraliser les résultats obtenus dans un échantillon sur l'ensemble de la population. Ces distributions d'échantillonnage sont de nature à nous permettre de fixer la taille des échantillons pour un risque d'erreur donné et l'estimation des valeurs réelles de la population, c'est à dire la généralisation des résultats obtenus sur l'échantillon.

Chapitre 4

L'ESTIMATION

De l'échantillon à la population la généralisation

Nous avons vu que chacun des paramètres de l'échantillon a une distribution d'échantillonnage bien définie et tend, lorsque n est grand à converger vers la valeur réelle de la population.

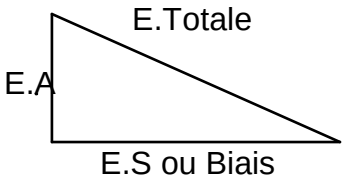
L'inégalité de Bienaymé Tchébicheff permet de déterminer avec précision la valeur réelle du paramètre : $P (I x - Ex I => t\sigma x) <= 1/t^2$

Si on pose $t \sigma/n^{1/2} = \varepsilon$ on a : $P (I x - Ex I => \sigma^2/n\varepsilon^2$

Lorsque n tend vers l'infini: $P (I x - Ex I => \varepsilon$ tend vers 0.

L'objectif de tout sondage est de pouvoir estimer les valeurs réelles des paramètres de la population, cette estimation est d'autant plus significative que l'intervalle de confiance est réduit et la probabilité est élevée.

L'erreur totale $(E (e - E)^2)$ commise se décompose en deux types d'erreur: une erreur systématique ou *biais* (σ_e^2) et *une erreur aléatoire* (β^2_e). Tout le problème consiste à minimiser le biais : $(E (e - E)^2) = (\beta^2_e) + (\sigma_e^2)$.



1 - La moyenne

La moyenne x suit la loi normale N(m, $\sigma/n^{1/2}$) lorsque n > 30 ou quand la loi d'origine est normale, autrement ou quand σ est inconnu, elle suit la loi de Student T(m, $s/(n-1)^{1/2}$ à n-1 ddl.

M = x +/- $L_{\alpha/2}$ σx P(x - Lσx <= m <= x + Lσx) = 1-α

m = x +/- $U_{\alpha/2}$ σx P(x - Uσx <= m <= x + Uσx) = 1- α

(x - U$\sigma/n^{1/2}$ <= m <= x + U$\sigma/n^{1/2}$)

m = x +/- T$\sigma x/(n - 1)^{1/2}$

Sur cette base, on peur estimer la somme en multipliant la moyenne par N.

Exemple: Dans un échantillon de 50 éléments, on a une moyenne de 12.5 et un écart-type de 3.5. Quelle serait la moyenne réelle à 95%. Dans ce cas, on n = 50 ce qui fait qu'elle suit la loi normale. Au seuil de 95% on a Uα = 1.96 ou 2. On peut écrire alors : x - U$\sigma x/(n-1)^{1/2}$ =< m =< x + U$\sigma x/(n - 1)^{1/2}$

12.5 - 2.3.5/7 <= m <= x + 2.3.5/7, soit une moyenne comprise entre: 11.5 <= m <=13.5.

Exemple: Dans un échantillon de 10 unités, de moyenne 14.5 et d'écart-type égal à 2.1, on veut estimer la moyenne réelle m au seuil de 95%. Pour a = 0.05, on a T_{10-1} = 2.26.

m = 14.5 +/- 2.6 2.1/(10 - 1) = 14.5 +/- 1,6, d'où la moyenne: 12.9 <= m <= 16.1.

Lorsque la population est finie, on a un tirage avec remise ou *un tirage exhaustif*, il convient de pondérer la variance par le coefficient : (N - n)/(N - 1).

n >30 ou population normale : $\sigma^2x = \sigma^2(N - n)/ n(N - 1)$

σ inconnu ou n > 30 : $\sigma^2x = s^2(N - n)/ N(n - 1)$

Lorsque n est faible, on a (N - n)/ (N - 1) proche de 1 et les tirages sont équivalents.

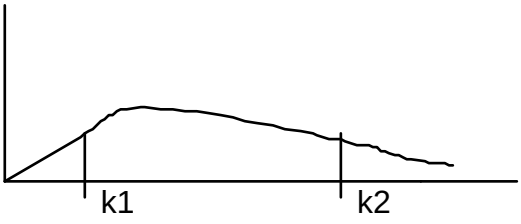
Exemple: Une enquête sur 100 individus a donné un âge moyen de 35 ans et un écart-type de 12. Quelle est la distribution d''échantillonnage de la moyenne x ?. Quel est l'intervalle de confiance?.
n dépasse 30, donc x suit une loi normale de moyenne m et d'écart-type $\sigma/n^{1/2}$. Dans ce cas, σ est inconnu ce qui fait qu'on doit estimer la variance : $\sigma^2 = ns^2/(n - 1)$ ce qui donne $sx^2 = \sigma^2/n = s^2/(n - 1)$. x suit $N(m, s/(n(n-1)^{1/2})$, donc x suit N(35, 0.1206).
x suit Tn-1: 35 - T 12/9.949 < m < 35 + T.12/9.949. Pour P = 95%, on a T= 2.26 et 32.274 < m < 54.725

2 - La variance

Elle représente un estimateur biaisé et il convient de le corriger par: (n - 1)/n. On a: E(s)2 = V - Vx = $\sigma^2 - \sigma^2/n = (n - 1)\sigma^2/n$. s^2 est biaisé, on le remplace par s'^2 qui est sans biais : $s'^2 = ns^2/(n - 1)$ ou $s'^2 = n(N - 1)s^2/N(n - 1)$ dans le cas d'un tirage avec remise.

Lorsque n est élevé, $s^2 = s'^2$ puisque (N - n)/(N - 1) =1 - n/N. Quant n est petit on a:: (N - n)/(N - 1) proche de 1. P ($-ns^2/k2 < \sigma^2 < ns^2/k1$) = 1- α, k1 et k2 sont les bornes de probabilité χ2 n–1. Si v > 30, $\chi 2 = U + (2v - 1)^{1/2}/2$
ns^2/σ^2 suit χ^2_{n-1}
$S^2_1 = n\sigma^2/\chi^2_{1-\alpha/2, n-1} = n\sigma^2/(2n - 3)^{1/2} + U_{1-\alpha/2}$

Si m est inconnu, on a $s'^2 = \Sigma (xi - m)^2/n$ constitue le meilleur estimateur de σ^2. ns'^2/σ^2 suit χ^2_n
P ($-ns'^2/l2 < \sigma^2 < ns'^2/l1$) = 1 - α P ($l1 < ns'^2/\sigma^2 < l2$) = 1 - α



Exemple: Dans un échantillon de 30 individus, on a une variance de 12. Quelle serait la variance réelle au seuil de 90%?. On a la relation: $ns^2/\chi^2_{\alpha/2, n} < \sigma^2 < ns^2/\chi^2_{\alpha/2, n}$

30.144 < σ^2 < 30.144 et la variance serait comprisentre 8.22 et 19.48: 8.22 < σ^2 < 19.48

3 - La somme

On peut estimer la somme (S) sur la base de la moyenne m puisque S = Σ N.m. La variance de la somme est égale à N^2 multiplié par Vx: V'.|S- = N^2 V'x. On a alors les relations selon le cas: $S = N_X$ +/- U_α. Ns_X = 1- α et $S = N_X$ +/- $T_{1-\alpha/2}$. Ns_X = 1- α

Exemple: On a relevé un revenu moyen de 95 D et un écart-type de 75 dans une enquête de 10.000 individus sur un total de 70.000 personnes. Quelle est la masse totale des revenus au seuil de 95%?. On a ainsi la masse comprise entre 56.7 et 76.3 MD: 56.7 <= S <= 76.3 MD

Exemple: On a enquêté 10.000 ménages sur un total de 700.000, la consommation moyenne est de 950 D/an, l'écart-type est de 700. Quelle est la loi de distribution ?. Quel est l'intervalle de m?. Quelle est la masse globale de dépenses au seuil de 95%?.
N = 700.000, x = 950, s = 700 et n = 0.000: c'est un tirage exhaustif puisque n/N = 0.014 mais comme (N -n)/(N - 1) = 0.9857 proche de 1, on peut l'assimiler à un tirage indépendant.
n > 30, d'où x suit N(m, $\sigma/n^{1/2}$). sx = 700/$(10.000)^{1/2}$ = 7 et (x - m)/s/$n^{1/2}$ suit N(0, 1). L'intervalle de m: P(x -Usx <= m<= x +Usx) = 1 - α. P(x -2sx <= m <= x + 2sx) = 0.95. D'où: 950 - 2.7 <= m <= 950 +2.7, soit 936 <= m <= 964 D. La dépense totale est: P(Nx - 2Nsx <= S <= Nx + 2Nsx)= 0.95, soit 9.486 <= S <= 9.514 MD

4 - La proportion

La fréquence f suit une loi normale N(p, $(pq/n)^{1/2}$) lorsque np >20. On peut écrire alors que: p = f +/- $_{\alpha/2}(pq/n)^{1/2}$ ce qui nous donne l'intervalle de variation suivant:
P(f - $U_\alpha(pq/n)^{1/2}$ <= p <= f - $U_\alpha(pq/n)^{1/2}$) = 1- α.

Lorsque le tirage est exhaustif, il faut pondérer par le coefficient (N - n)/(N - 1), la variance devient alors : pq(N - n)/n(N -

1) au lieu de pq/n. Dans le cas où p est inconnu et np > 10, on peut utiliser : f(1- f)/(n - 1) ce qui nous donne: $p = f \pm U_{\alpha/2} (f(1 - f)/(n - 1))^{1/2}$

Exemple: On a constaté dans la population masculine un taux de mortalité de 15% entre 20 et 45 ans, dans une usine on trouve un taux de 20% sur un échantillon de 400. Ce taux diffère-t-il significativement du taux de la population totale au seuil de 95%?.

On a p = 0.2, q = 0.8, f = 0.15 et n = 400. On écrit: $p - 2 (pq/n)^{1/2} <= f <= p + 2 (pq/n)^{1/2}$.
On obtient le résultat suivant : 0.115 <= f <= 0.185. Or, on a f = 0.20 qui se situe à l'extérieur de l'intervalle au seuil de 95%. On conclut que la mortalité dans cette usine est significativement plus élevée que dans la population d'origine.

Exemple: Dans une population, le taux de chômage est de 25%, dans quel intervalle au risque de 0.05 se situent les intervalles de chômage pour un échantillon de 50, 100 et 10.000 individus ?.

Au seuil de 95%, on a $f = p \pm 2 (pq/n)^{1/2}$ d'où $f = 0.25 \pm 0.866/(n)^{1/2}$. Pour n = 50, on a: 0.127 < f < 0.373. pour n = 100, on a: 0.163 < f < 0.337 et pour n = 1000, on a: 0.241 < f < 0.259
Exemple: Dans un échantillon de 100 personnes, on a relevé 36 occupés. Quelle est la proportion des actifs dans la population d'origine au seuil de 95% ?.
On a: $p = f \pm 2 (f(1-f)/n)^{1/2}$ ce qui nous donne: p = 0.36 +/- 0.096, soit 0.226 < p < 0.456

5 - L'effectif

On peut calculer l'effectif en multipliant p par N d'où on a : Np1 <= Np <= Np2. On peut écrire la relation: $Np = Nf \pm U_{\alpha/2} (npq)^{1/2}$
$V'.Nf = N^2.V'f$

Dans le cas d'un tirage exhaustif, il faut pondérer par le coefficient (N - n)/(N - 1) et on obtient: $Np = Nf \pm U_{\alpha/2} (npq(N - n)/(N - 1))^{1/2}$

6 - Le coefficient de variation

$$Cv^2 = \sigma^2_{(\ Nf\ )} / Np^2$$

7 - Le coefficient de corrélation

Le coefficient de corrélation linéaire suit une loi normale: N(r, 1- $r^2/n^{1/2}$) ou la loi de Student T (r, 1 - $r^2/(n - 2)^{1/2}$) à n - 2 ddl. On peut alors écrire pour la loi normale: $\rho = r +/- U\alpha.\sigma r$.

Exemple: L'analyse a donné une corrélation de -0.532 pour un échantillon de 33 individus, le coefficient est-il significatif?. La table de Pearson montre que la relation est significative, on a ici affaire à la distribution de Student $T = (n - 2)^{1/2}/(1 - r2)^{1/2} = -2.508$ or $T_{0,05\ n-2} = 1.697$ d'où la signification de la corrélation.

Au seuil de 95% on a $r = r1 +/- [((1 + r) \exp(-2Ua/(n-3)^{1/2})/(1 - r) - 1]/ [((1 + r) \exp(-2Ua/(n-3)^{1/2})/(1 - r) + 1]$ ce qui nous donne avec Ua = 2 un intervalle [-0.74 <r < - 0.22].

8 - Les coefficients de régression

Le coefficient de régression (a) suit la loi Normale: N(a, $\sigma y/\sigma x((1 - r^2)/n)^{1/2}$) ou de Student: (a, $\sigma y/\sigma x\ ((1 - r^2)/(n - 2)^{1/2}$) à (n - 2) ddl selon la taille de l'échantillon n. On a ainsi:

$a^* = a +/- U\alpha\ \sigma y/\sigma x\ ((1 - r^2)/n)^{1/2}$) ou $a^* = a +/- T\alpha\ \sigma y/\sigma x\ ((1 - r^2)/(n - 2)^{1/2}$)

Le coefficient de régression (b) suit la loi Normale N(b, $\sigma y\ ((1 - r^2)/n)^{1/2}$) ou de Student (b, $\sigma y\ ((1 - r^2)/(n-2)^{1/2}$) à (n-2) ddl selon la taille de l'échantillon n:

$b^* = b+/- U\alpha\ \sigma y\ ((1 - r^2)/n)^{1/2}$) ou $b^* = b +/- T\alpha\ \sigma y\ ((1 - r^2)/(n - 2)^{1/2}$)

L'estimation de la valeur réelle des paramètres de la population à partir de l'échantillon n'a de sens que lorsque les échantillons sont strictement aléatoires et la taille de l'échantillon est représentative. Dans les sondages raisonnés, l'erreur n'est pas connu interdisant toute

généralisation et les résultats ne sont valables qu'ua niveau de l'échantillon enquêté.

Intervalle de confiance d'un pourcentage au seuil de 95%

	5	10	15	20	25	30	35	40	45	50
10	-	0 45	1 50	3 56	5 60	7 65	9 70	12 74	15 78	19 81
20	0 25	1 32	3 38	6 44	9 49	12 54	15 59	19 64	23 68	27 73
30	0 20	2 27	5 33	8 39	11 44	15 49	19 54	23 59	27 64	31 69
40	0 17	3 24	6 30	9 36	13 41	17 47	21 52	25 57	29 62	34 66
40	1 15	3 22	6 28	10 34	14 39	18 45	22 50	26 55	31 60	36 64
60	1 14	4 21	7 27	11 32	15 38	19 43	23 48	28 53	32 58	37 63
70	1 13	4 21	8 26	11 31	15 37	20 42	24 47	28 52	33 57	38 62
80	1 12	4 19	8 25	12 30	16 36	20 41	25 46	29 52	34 57	39 61
90	2 12	5 18	8 24	12 30	16 35	21 41	25 46	30 51	34 56	39 61
100	2 11	5 18	9 24	13 29	17 35	21 40	26 45	30 50	35 55	40 60
150	2 10	6 16	10 22	14 27	18 33	23 38	27 43	32 48	37 53	42 58
200	2 9	6 15	10 21	15 26	19 32	24 37	28 42	33 47	38 52	43 57
500	3 7	8 13	12 18	17 24	21 29	26 34	31 39	36 44	41 49	46 54
1000	4 7	8 12	13 17	18 23	22 28	27 33	32 38	37 43	42 48	47 53
2000	4 6	9 11	13 17	18 22	23 27	28 32	33 37	38 42	43 47	48 52

Lecture da la Table

Sur 100 individus, on a enregistré un pourcentage de 10% (une proportion p = 0,1). La table montre que le pourcentage théorique dans la population (p) est compris dans un intervalle (5% – 18%) au seuil de 95% (au risque de 5%).

Quand le pourcentage observé (f) dépasse 50%, on utilise le pourcentage complémentaire (1- f) : lorsque f est de 65%, (1 – f) est de 35%.

Les intervalles qui correspondent à des résultats impossibles (par exemple 15% de cas positifs sur 10 cas) ne figurent dans la Table que pour permettre des interpolations entre les lignes ou les colonnes.

Chapitre 5

LA TAILLE DE L'ECHANTILLON

Sur la base de la loi des grands nombres, les distributions d'échantillonnage et en connaissant certains paramètres de la population, on peut déterminer la taille minimale de l'échantillon à enquêter avec un risque d'erreur connu à l'avance et un seuil de probabilité bien déterminé qui nous permettent de généraliser à la population les résultats obtenus sur les échantillons.

1 - La moyenne

Pour la moyenne, on a la relation: $m = x +/- U_{\alpha/2}\, \sigma x$ et $P(x - U\sigma x <= m <= x + U\sigma x) = 1 - \alpha$ on peut écrire ainsi que: $(x - U\sigma/n^{1/2} <= m <= x + \sigma/ n^{1/2})$, d'où: $|x - m| <= U\sigma/n^{1/2}$
$U\sigma/ n^{1/2} <= km$ avec k : précision de l'estimation en % de la moyenne m.

On écrit ainsi: $n^{1/2} => U\alpha.\sigma/k.m$, d'où on détermine la taille de l'échantillon requis pour ce risque d'erreur et ce niveau de précision: $n => U\alpha^2.\ \sigma^2 /k^2\, m^2$

Si on connaît le coefficient de variation, on a: $n => (V.\, U\alpha /k)^2$. Si le tirage est exhaustif, on a : $n => U\alpha^2\, \sigma^2\, N/U\alpha^2\sigma^2 + \alpha\,(N - 1)$ avec a: précision fixée en valeur absolue.

Exemple: Dans une population de moyenne 35 et d'écart type 10, on veut enquêter un échantillon avec une précision de 0.1 de la moyenne et un seuil de signification de 95%, quelle serait la taille minimale à tirer?. On a $n >= (U\alpha.\ s/km)^2 = (1{,}96.10/35.0{,}1)^2 = 31.36$, soit 32 individus. Si on veut que la précision soit au 1/100, on doit enquêter 56.

2 - La proportion

Si la taille dépasse 30, la fréquence suit la loi normale N(p, $pq/n)^{1/2}$), on a alors la relation: | f - p | <= Uα. $(pq/n)^{1/2}$, d'où on obtient: Uα. $(pq/n)^{1/2}$ <= k.p avec k : précision en % de p. on détermine ainsi la taille n => $U\alpha^2.(1 - p) / p.k^2$

n maximum = $1/p^2$ nous donne: n <= $(U_{\alpha/2}/2\Delta P)^2$. Le tableau ci-dessous représente la taille de l'échantillon en fonction de k et p.

ΔP	1-α		
	0.9	0.95	0.98
0.01	6760	9600	13530
0.02	1700	2400	3380
0.05	270	380	540

Taille n de l'échantillon en fonction de k et p

p / k	0.9	0.8	0.7	0.5	0.3	0.2	0.1		0.01
0.1	45	100	172	400	934	1600	3600		39600
0.05				1600					

Exemple: Une population contient 40% d'alphabètes, on désire que le % dans l'échantillon se trouve dans l'intervalle: p +/-0.01 avec une probabilité de 99%. Quelle est la taille de l'échantillon à enquêter?.
P(| f - p | <= 0.01 I => 0.99, $1 - 1/t^2 = 0.99$ donne t = 10 et t $(pq/n)^{1/2}$ <= 0.01 donne n => 240.000

Si on utilise la loi normale pour l'approximation de la loi binomiale, on a f suit β(n, p) avec Ex = p et Vx = pq/n, d'où P|p - U(pq/n)1/2) <= f <= | >= 0.99. Avec t = 2.58 on a $t(pq/n)^{1/2}$ <= 0.01, soit n >= 15975.

3 - La variance

Comme les autres paramètres, on peut écrire la relation: σ* +/- $U_{1-\alpha/2}$ $\sigma^*/(2n)^{1/2}$ avec $\delta = U_{1-\alpha/2}\sigma^*/(2n)^{1/2}$ et dr =

$100.U_{1-\alpha/2}/(2n)^{1/2}$ où dr : la différence relative par rapport à σ en %. On obtient alors n: n = $5000\ U_{1-\alpha/2}\ /\ dr^2$

Sur la base de la loi des grands nombres, les distributions d'échantillonnage et la connaissance de certains paramètres de la population (x, σ, f), on peut déterminer la taille minimale de l'échantillon à enquêter avec un risque d'erreur connu à l'avance (α) qui nous permet de généraliser à la population les résultats obtenus sur l'échantillon.

Lorsque la distribution est *normale*, ou quand la taille de l'échantillon (n) *dépasse 30 unités* (choisis aléatoirement) et on *connaît certains paramètres* (à travers d'autres études ou enquêtes similaires récentes) comme la moyenne et l'écart type, le coefficient de variation ou la proportion, on peut déterminer la taille minimale à enquêter.

1- la moyenne et l'écart type ou le coefficient de variation

La formule est : $\mathbf{n \geq (U_\alpha . \sigma/k.m)^2}$ ou $\mathbf{n \geq (U_\alpha\ v/k)^2}$

avec m : la moyenne observée d'une variable significative, σ: son écart type, v : le coefficient de variation v = σ/m. n : la taille de l'échantillon à enquêter. U_α : valeur de la loi normale au risque d'erreur α et au seuil de (1 - α)% (au risque de 5%, correspond un seuil de signification de 95%)...). Au seuil de signification de 95%, la valeur de U_α est 1,96 qu'on peut arrondir à 2 (au seuil de 99%, la valeur de U_α est 2,32).

Le tableau suivant donne la taille minimale de l'échantillon pour un risque d'erreur de 5% (seuil de signification de 95%, la valeur de U_α = 1,96). Dans le cas où on veut être plus sûr (seuil de 99%), il faut multiplier les effectifs suivants par 1,4 $(2,32/1,96)^2$.

Taille de l'échantillon au seuil de 95% selon les valeurs de k et v

k / v	**0.5**	**1**	**1.5**	**2**	**5**	**10**
0.500	***4***	***16***	***35***	***256***	***385***	**1537**
0.200	***25***	***97***	***217***	***385***	**2401**	9605
0.150	***43***	***171***	***385***	***683***	**4269**	17074
0.100	***97***	***385***	***865***	**1537**	9605	38417
0.050	***385***	**1537**	**3458**	6147	38417	153665
0.010	9605	38417	86437	153665	960401	3841601
0.005	38417	153665	345745	614657	3841601	15366401
0.001	960401	3841601	8643600	15366401	96040001	384160001

Les chiffres en gras indiquent un échantillon inférieur à 5000. Les chiffres en gras et italique indiquent une taille inférieure à un millier. v : coefficient de variation, v = σ/m. k : degré de précision de la moyenne en % (0,1 = à 10% près)

Si le tirage est exhaustif, on a : n ≥ Uα² σ² N/Uα²σ² + α (N – 1) avec a: précision fixée en valeur absolue.

Exemple: Dans une population de moyenne 35 et d'écart type 10, on veut enquêter un échantillon avec une précision de de 10% près et 1% (k = 0.1 et 0,01) de la moyenne et un seuil de signification de 95% et 99%, quelle serait la taille minimale à tirer?
On a $n \geq (U_\alpha \sigma/k.m)^2 = (1{,}96.10/35.0{,}1)^2 = 31.36$, soit 32 individus (44 au seuil de 99%). Si on veut que la précision soit au 1/100° près, on doit enquêter 3137 (5257 au seuil de 99%).

2 - La proportion

Si la taille dépasse 30, la taille de l'échantillon est $n \geq (U_\alpha/k)^2((1 - p) / p)$, avec k : précision en % de p, p : proportion observée. Le tableau suivant fixe la taille minimale au seuil de 95%. Pour obtenir la taille au seuil de 99%, il suffit de multiplier l'effectif obtenu pour 95% par le coefficient 1,4

Taille de l'échantillon en fonction de k et p pour une probabilité de 95%

p / k	**0.9**	**0.8**	**0.7**	**0.5**	**0.3**	**0.2**	**0.1**		**0.01**
0.05	91	115	149	292	812	1825	7300		729905
0.1	43	55	71	237	385	865	3457		345745
0.2	19	25	32	62	171	385	1537		153665
0.5	5	7	8	16	43	97	385		38417
0.8	2	2	2	4	11	25	97		9604
1	1	1	1	1	1	1	1		1

Exemple: Une population contient 40% d'analphabètes, on désire que le % dans l'échantillon se trouve dans l'intervalle: p ± 0.01 avec une probabilité de 99%. Quelle est la taille de l'échantillon à enquêter?
Dans ce cas $U_\alpha = 2{,}32$, k = 0,01, p = 0,4, on obtient : $n \geq (2{,}32/0{,}01)^2(1-0{,}4)/0{,}4) = 80736$. Si on veut être significatif seulement à 95% on a : 57624. Si on se limite à une précision de 10% (0,1) seulement on a n = 577 unités.

Lexique

Échantillonnage Aléatoire Simple (SRS)

L'*échantillonnage aléatoire simple* est un type d'échantillonnage de probabilités où les observations sont sélectionnées de façon aléatoire dans une population qui a une probabilité ou une fraction d'échantillonnage connue. Typiquement, on commence avec une liste de *N* observations de la population totale parmi lesquelles on veut tirer un échantillon aléatoire simple (par exemple, une liste des électeurs inscrits sur les listes électorales) ; on peut alors générer un nombre *k* d'observations aléatoires (sans remise) dans l'intervalle de *1* à *N*, et sélectionner les observations respectives dans l'échantillon final (avec une fraction d'échantillon ou une probabilité de sélection connue de *k/N*).

Échantillonnage Aléatoire Stratifié

En général, l'échantillonnage aléatoire est le processus de sélection aléatoire d'observations dans une population pour créer un sous-échantillon qui "représente" les observations dans cette population (voir Kish, 1965 ; voir aussi les rubriques Échantillonnage de Probabilité, Échantillonnage Aléatoire Simple et Échantillons EPSEM ; voir aussi la rubrique Échantillon Représentatif pour une exploration succincte de cette notion souvent mal comprise). Dans l'échantillonnage stratifié, on applique en général des fractions d'échantillon spécifiques (identiques ou différentes) aux différents groupes (strates) dans la population pour tracer l'échantillon. Dans *STATISTICA*, vous pouvez tracer les échantillons aléatoires stratifiés en utilisant les options de la boîte de dialogue Créer un Sous-Ensemble/Échantillonnage Aléatoire.

Sur-échantillonnage de strates particulières pour sur-représenter des événements rares. Dans certaines applications de data mining prédictif, il est souvent nécessaire d'appliquer un échantillonnage stratifié pour sur-échantillonner systématiquement (appliquer une fraction d'échantillonnage supérieure) des "événements rares" d'intérêt particulier. Par exemple, pour un catalogue de vente au détail, le taux

de personnes répondant à des offres particulières du catalogue peut être inférieur à 1%, et lorsque l'on analyse les données historiques (des campagnes d'offres antérieures) pour construire un modèle afin de cibler les acheteurs potentiels plus efficacement, il est souhaitable de sur-échantillonner les anciennes personnes ayant répondu (c'est-à-dire, les personnes interrogées "rares" ayant passé commande dans le catalogue) ; vous pouvez alors appliquer les diverses techniques de construction de modèle pour la classification (voir le Data Mining) pour un échantillon consistant en approximativement 50% de personnes répondant et 50% de personnes ne répondant pas. Sinon, si on traçait un échantillon aléatoire simple pour l'analyse (avec 1% de personnes répondant), pratiquement toutes les techniques de construction de modèle prédiraient alors une simple "non-réponse" pour toutes les observations, et serait correcte (de façon triviale) dans 99% des cas.

Échantillonnage de Probabilité

Dans l'*échantillonnage de probabilité*, chaque observation de la population à partir de laquelle on tire l'échantillon possède une probabilité connue d'être sélectionnée dans l'échantillon ; lorsque cette probabilité est la même pour toutes les observations de la population, l'échantillon est un échantillon de probabilité égal ou échantillon EPSEM (probabilité égale des méthodes de sélection ; voir Kish, 1965, pour plus de détails).

Les échantillons EPSEM ont une certaine probabilité souhaitée ; par exemple, les simples formules pour calculer les moyennes, écart-types, etc., peuvent être appliquées pour estimer les paramètres respectifs dans la population

Échantillonnage par Quota

L'*échantillonnage par Quota* fait en général référence au processus par lequel un chercheur tente d'obtenir dans un échantillon la représentation exacte d'une population en fonction d'une certaine caractéristique démographique qui semble importante (comme le sexe, l'âge, l'ethnie, le revenu, etc.). Par exemple, un chercheur peut s'efforcer à sélectionner un échantillon à partir d'une population pour que l'échantillon soit constitué de exactement 50% d'hommes et 50% de femmes, un certain pourcentage de personnes appartenant à un groupe ethnique, etc. L'objectif de cette pratique est généralement

d'obtenir un type d'échantillon représentatif de la population sous-jacente.
En général, seuls les échantillons de probabilité sélectionnés de façon convenable tels que les échantillons EPSEM garantissent que la population pour laquelle on veut faire une généralisation est correctement "représentée". Référez-vous, par exemple à Kish (1965) pour une présentation détaillée des avantages et caractéristiques des échantillons de probabilité

Échantillon Représentatif

La notion "d'échantillon représentatif" est souvent mal comprise. L'intention globale est de sélectionner un échantillon dans une population pour que les propriétés particulières de cette population puissent être estimées fidèlement à partir de l'échantillon. Par exemple, les experts en sciences politiques peuvent tracer des échantillons dans la population de votants pour prédire avec un certain degré de certitude les résultats d'une élection.
En général, seuls les échantillons de probabilité sélectionnés convenablement tels que les échantillons EPSEM garantissent que la population pour laquelle on veut généraliser est correctement "représentée". D'un autre côté, une notion souvent erronée insinue couramment que pour obtenir une "représentativité", il est souhaitable de tracer un échantillon stratifié en utilisant des "quotas" particuliers (échantillonnage par Quota) où une caractéristique démographique comme l'âge, le sexe, l'ethnie, etc., est bien "équilibrée", pour correspondre précisément à l'habillage de la population sous-jacente. Cette idée est fausse. La précision des estimations (comme la marge des cotes) pour une population calculée à partir d'un tel échantillon s'accroît seulement si les variables dont on attend qu'elles fonctionnent (âge, sexe, ethnie, etc.) sont (fortement) reliées à la variable de sortie d'intérêt (par exemple, le comportement de vote). Toutefois, dans la pratique une telle connaissance *a-priori* est souvent évasive, et s'applique à des méthodes d'échantillonnage par quota qui peuvent aboutir à des résultats très trompeurs.