OCCGE-Inf. N° 96 - 1985 - pp. 89 - 99

INFORMATIONS

PRINCIPES GÉNÉRAUX DES SONDAGES

APPLICATION AUX ENQUÊTES DE POPULATION

Par J. VAUGELADE (1)

N° 8 619/85/Doc.Tech.OCCGE

PLAN :

INTRODUCTION

La collecte des données peut s'effectuer sur tous les individus appartenant à la population étudiée : c'est un recensement. Le plus souvent, la collecte est limitée à une fraction de la population, c'est une enquête par sondage et la fraction étudiée de la population est appelée échantillon. Cette note, sans formules mathématiques, explique comment concevoir un sondage.

(1) Démographe ORSTOM - B.P 182 OUAGADOUGOU (Burkina Faso)

1.- *LES TYPES DE SONDAGES*

1.1.- Le sondage par choix raisonné consiste à choisir un échantillon représentatif d'une situation générale, ce choix est subjectif et ne peut être utilisé que pour des enquêtes qui ne visent pas à l'extrapolation des données à l'ensemble de la population.

1.2.- Le sondage par quota impose que la distribution statistique de certains caractères jugés importants soit identique dans l'échantillon et dans la population générale.

Le choix des individus est laissé à l'initiative des enquêteurs. Les caractères contrôlés sont en général le sexe, l'âge, la catégorie socio-professionnelle et le type de milieu (rural, urbain) mais le contrôle est fait séparément sur chaque caractère.

Cette méthode est utilisée pour les enquêtes d'opinion (sondage pré électoraux, publicitaires,...)

C'est une méthode empirique qui présente surtout l'avantage d'un coût moindre qu'un sondage probabiliste. La méthode permet l'extrapolation mais ne permet pas d'évaluer la précision des estimations.

1.3.- Le sondage probabiliste détermine un échantillon par tirage au sort. On peut donc connaître la probabilité d'être sélectionné pour un individu, extrapoler les résultats à l'ensemble de la population et évaluer la précision des résultats. L'évaluation de la précision des résultats nécessite le recours à des formules assez complexes qui ont été écartées de cette note.

1.4.- Les problèmes généraux d'enquêtes : questionnaires, définitions, contrôle des enquêteurs ne relèvent pas des problèmes de sondage, mais ils sont plus importants que la méthode de sondage (voir DESABIE pp. 299-468).

2.- *PRINCIPE DES SONDAGES PROBABILISTES*

2.1.- La précision des estimations dépend de la taille de l'échantillon et non pas du taux de sondage. Ainsi 100 000 personnes pour une région, un pays, un continent ou 100 000 personnes pour le monde entier fournissent une estimation aussi précise. Cependant si on veut une certaine précision au niveau régional par exemple, il faudra que l'effectif de chaque région dépasse un minimum qui dépend des variables étudiées (voir § 6).

2.2.- Il s'agit ici de la précision statistique, cependant les erreurs dûes à une enquête mal organisée et mal contrôlée peuvent être plus importantes que l'incertitude due à l'échantillonnage. Une attitude prudente consiste à préférer un échantillon plutôt petit avec une enquête bien contrôlée qu'un large échantillon si cela doit nuire à la qualité du contrôle.

2.3.- Il peut apparaître plus facile d'étaler le travail sur une plus longue période en ayant recours à un nombre relativement réduit d'enquêteurs qu'il est plus facile de former et de contrôler.

Cependant si les variations saisonnières sont importantes, et si l'enquête est étalée dans le temps, les résultats, pour une région, dépendront de la saison d'enquête dans cette région. Ceci constitue un biais qui empêcherait la comparaison des résultats régionaux.

L'enquête dans chaque région ne doit donc pas être concentrée en une saison mais étalée tout au long de l'année.

2.4.- La théorie des sondages développe deux aspects :

- réduire l'incertitude pour un coût donné,
- une fois l'enquête réalisée, fournir des estimations de l'incertitude due à l'échantillonnage.

Seul le premier aspect est développé ici. Le deuxième aspect fait appel à une formulation mathématique compliquée.

3.- *LA BASE DE SONDAGE*

3.1.- Pour tirer au hasard des unités, il faut en établir la liste. Cette liste constitue la base de sondage.

Elle doit être :

- sans omission, donc exhaustive
- sans double emploi
- à jour.

S'il est possible d'établir une telle liste pour des entreprises, des écoles,... une telle liste est impossible à établir pour des individus. Dans ce cas, on rassemble les individus en une unité plus grande, et c'est la liste de ces unités qui constitue la base de sondage. Les conditions supplémentaires à celles ci-dessus sont pour des unités collectives.

- Les unités doivent être clairement délimitées et localisées;
- la taille approximative de chaque unité doit être connue;
- si possible, il est utile de connaître d'autres caractéristiques pertinentes de chaque unité, ces caractéristiques permettent la stratification (voir § 4).

3.2.- Pour réduire les coûts de déplacement dans l'enquête, il est utile d'avoir des unités collectives qui ne soient pas dispersées spatialement. On est donc conduit à choisir comme unité collective, des villages, quartiers,... qui sont des unités aréolaires.

Ces unités aréolaires doivent constituer une partition du territoire (une partition nécessite que deux unités ne se recouvrent pas et que chaque partie du territoire appartienne à une unité). On peut bien sûr négliger les zones dont on est sûr qu'elles sont absolument vides de population.

3.3.- Sondage à un degré sans stratification.

Par exemple les 1 691 districts du recensement voltaïque qui ont servi de base de sondage pour l'enquête post-censitaire de 1976 sont constitués de regroupements de villages pour les petits villages, de quartiers pour les gros villages et les villes.

Dans cette liste un district sur 28 a été tiré au hasard (soit 61 districts) chaque district a été enquêté intégralement. Il n'y a eu qu'un seul tirage, celui des districts. On dit que c'est un sondage à un degré.

Dans un tel sondage, la répartition des grappes entre le milieu urbain et le milieu rural, ou entre les différentes régions est laissée au hasard.

Un exemple va montrer que cette répartition peut s'éloigner fortement d'une répartition proportionnelle. Ainsi pour un sondage portant sur 100 grappes avec 80 grappes en milieu rural et 20 grappes en milieu urbain, parmi lesquelles on veut en tirer 10. Le résultat du tirage peut être 10, 9, 8, 7,... grappes en milieu rural et inversement 0, 1, 2, 3,... grappes en milieu urbain. Le tableau suivant présente tous les résultats possibles de tirages.

	Nombre de grappes de l'échantillon		*Probabilité*
	en milieu urbain	*en milieu rural*	
Echantillon non stratifié	*0*	*10*	*0,10*
	1	*9*	*0,27*
	2	*8*	*0 32*
	3	*7*	*0,21*
	4	*6*	*0,08*
	5 et +	*5 et -*	*0,02*
Total			*1,00*
Echantillon stratifié	*2*	*8*	*1,00*

Seul un échantillon sur trois comportera la proportion de population urbaine attendue (deux grappes sur dix)). Par malchance l'échantillon en risque d'être sérieusement déformé. Pour y remédier on considérera d'une part les 80 grappes rurales parmi lesquelles on en tirera 8 et les 20 grappes urbaines parmi lesquelles on en tirera 2, on appelle cette opération une stratification.
Deux strates ont été constituées, une strate urbaine et une strate rurale.

4.- LA STRATIFICATION

4.1. L'exemple 3.3. a montré l'intérêt d'une stratification plus généralement, on a toujours avantage à stratifier.

4.2.- Voici les exemples de stratification :

- selon la région, par exemple en utilisant le découpage administratif en région ou département,
- selon le milieu urbain, rural,
- selon la taille des localités en trois strates par exemple, plus de 20 000 personnes, 5 000 à 19 999 , 4 999 et moins,

- selon la proportion d'agriculteurs par localité par exemple : plus de 50 % et moins de 50 % soit deux strates,
- selon la proportion d'enfants de 0 à 9 ans. Par exemple, plus de 30 %, 25 à 30 %, moins de 25 %. Cette stratification pourrait être conseillée pour une étude sur la fécondité car la proportion d'enfants est en liaison étroite avec la fécondité.

4.3. Plus généralement il est préférable de stratifier avec une variable en liaison forte avec le phénomène étudié. La stratification peut reposer sur plusieurs critères à la fois.

Un fichier des localités comprenant diverses caractéristiques (voir 3.1.) est très utile.

En stratifiant, on effectue un tirage indépendant dans chaque strate, on est sûr alors d'avoir une bonne répartition de l'échantillon entre les strates. Si on garde le même taux de sondage dans chaque strate, en général cela conduit à réduire l'incertitude dûe à l'échantillonnage.

4.4. Une deuxième raison pour la stratification est de modifier le taux de sondage pour certaines parties de la population. Pour obtenir des résultats significatifs pour un domaine d'étude réduit, on peut y augmenter le taux de sondage. C'est souvent le cas des zones urbaines.

Exemple d'un sondage stratifié pour le Burkina

Milieu	*Urbain (5 villes)*	*Rural*	*Total*
Univers : Population en 1975	*326 610*	*5 275 593*	*5 638 203*
Taux de sondage	*1/10*	*1/50*	
Population échantillon	*32 661*	*105 512*	*138 173*

L'échantillon total représente 1/40,8 de la population totale, un échantillon de même taille avec un taux de sondage uniforme de 1/40 aurait conduit à 8 165 personnes en milieu urbain à répartir entre les cinq villes, ce qui peut être jugé insuffisant. Un tel sondage devrait de plus être stratifié selon les régions administratives par exemple.

Cependant on peut retenir le principe que les groupes particuliers devraient être analysés au moyen d'études spéciales et qu'un plan de sondage doit être basé au départ de l'hypothèse de fractions égales de sondage dans les strates. L'adoption d'une solution différente doit être, justifiée par des arguments précis.

4.5.- Un domaine d'étude est un groupe de population qui intéresse l'analyse. Par exemple :

- pour la fécondité on s'intéressera aux femmes de 15 à 49 ans,
- pour l'emploi salarié, on isolera la population urbaine,

- pour étudier l'influence de la religion sur la fécondité, la scolarisation, on étudiera séparément les animistes, les musulmans et les chrétiens. Chaque catégorie de population constitue un domaine d'étude. Normalement un domaine d'étude devrait regrouper une ou plusieurs strates. Cela s'impose pour les domaines géographiques. On est souvent amené à considérer des domaines qui sont inclus dans les strates.

Dans ce cas les effectifs extrapolés peuvent être sérieusement erronés ainsi dans l'exemple cité en 3.3., il y a 27 chances sur 100 pour que la population urbaine de l'échantillon soit de 10 % au lieu de 20 % attendus. Comme domaine d'étude, on préfèrera en général, un domaine inclus dans les strates mais réparti sur plusieurs strates. Dans ce cas, les effectifs extrapolés sont plus pécis.

Dans tous les cas, on peut raisonnablement espérer que les structures (par exemple, la variation de la fécondité selon la religion) soient meilleures que les effectifs.

4.6.- Plus il y a de strates, plus grande est l'efficacité et ceci jusqu'au cas extrême où chaque strate ne comporte qu'une seule unité sélectionnée.

4.6.1. Pour chaque caractère de stratification : il faut définir deux ou trois classes. La troisième classe apporte un gain faible et la quatrième un gain négligeable.

4.6.2. Une façon de choisir les limites y_1 et y_2 de trois classes de manière optimale est d'obtenir :

$$y_1 = \frac{x_1 + x_2}{2}$$

$$y_2 = \frac{x_2 + x_3}{2}$$

où x_1, x_2, x_3 sont les moyennes dans chacune des trois strates.

Ce résultat est obtenu par tâtonnement, ainsi pour l'enquête par sondage de 1961, pour la zone rurale, les strates de tailles suivantes étaient considérées :

Strates selon la taille	*Population globale (recensement administratif)*	*Nombre de localités*	*Population moyenne*
1 - 499	*1 076 076*	*4 862*	*221 = x_1*
500 - 1099	*1 086 809*	*1 490*	*729 = x_2*
1100 et +	*1 284 071*	*703*	*1 827 = x_3*
Total	*3 442 950*	*7 055*	*488*

On a $y_1 = 500$ et $\frac{x_1 + x_2}{2} = \frac{221 + 729}{2} = 475$

$y_2 = 1100$ et $\frac{x_2 + x_3}{2} = \frac{729 + 1827}{2} = 1\,278$

500 et 475 sont peu différents, ainsi que 1 100 et 1 278.

4.6.3.- Les strates très petites peuvent être regroupées pour en faire une plus grande, on peut introduire une nouvelle stratification dans une strate particulière. Le schéma de stratification peut être extrêmement souple. *On doit viser à aboutir à des strates de dimensions à peu près égales.*

5.- *LES SONDAGES A PLUSIEURS DEGRES*

5.1.- Exemple : Enquête démographique par sondage au Burkina 1960-1961

1. Ouagadougou et Bobo-Dioulasso sont exclus et doivent être recensés à part.
2. L'ensemble des 12 centres secondaires sont étudiés au 1/10è
3. Pour le milieu rural, le sondage est au 1/50è.

Le sondage est à deux degrés, au premier degré des localités au deuxième degré des concessions et cela à l'intérieur de chacune des neuf strates géographico-ethniques.

Taille des localités	*Probabilité au 1er degré G*	*Probabilité au 2ème degré H*	*Probabilité d'ensemble f = g h*
1 - 499	*1/50*	*1/1*	*1/50*
500 - 1099	*1/25*	*1/2*	*1/50*
1100 et plus	*1/10*	*1/5*	*1/50*

Dans ce sondage le taux de sondage est uniforme pour le milieu rural.

Les localités constituant les unités primaires, les concessions les unités secondaires. On s'intéresse aux individus qui appartiennent aux concessions tirées dans une localité, ils constituent une *grappe*. Les grappes sont de taille voisine, 365 personnes pour la deuxième strate et 305 pour la troisième strate.

Dans un sondage à trois degrés, on aurait en plus des unités tertiaires.

5.2. Les sondages à plusieurs degrés se justifient dans deux cas.

5.2.1. La base de sondage doit être constituée et il serait trop coûteux de l'établir pour l'ensemble de la population. Ainsi la liste des ménages ne peut être établie pour un pays, par contre la liste peut être établie au moment de l'enquête dans les unités choisies.

Ainsi dans l'enquête décrite en 5.1. la liste des concessions n'a été établie que dans les localités tirées.

5.2.2. On veut éviter la dispersion de l'échantillon pour réduire les coûts de déplacements et faciliter le contrôle.

En effet, l'organisation du travail est importante pour la méthode de sondage, d'habitude on constitue des équipes de deux à quatre enquêteurs avec un contrôleur par équipe ou pour deux équipes. Il faut que chaque lieu d'enquête fournisse du travail pour quelques jours au moins.

Compte tenu du rendement attendu cela détermine le nombre de personnes à enquêter dans un même lieu, ces personnes constituent une grappe.

5.3.- Un sondage à plusieurs degré réduit la précision due au sondage, il ne doit être retenu que si cela est nécessaire.

5.4.- Pour la constitution de la base de sondage, on choisira d'abord l'unité aréolaire la plus petite pour laquelle existe une base de sondage puis l'unité aérolaire finale. On n'introduit un degré supplémentaire que si chaque degré supérieur comporte au moins cinq (dix selon certains auteurs) unités de rang inférieur.

Ainsi, dans le sondage de l'enquête 1960-1961, on aurait pu introduire le quartier (unité secondaire), la concession devenant une unité tertiaire. Cela n'aurait été justifié que si une localité comportait au moins cinq (ou dix) quartiers et chaque quartier au moins cinq (ou dix) concessions.

5.5.- Un cas important des sondages à plusieurs degrés est celui des sondages autopondérés. C'est-à-dire que chaque individu à la même probabilité d'être choisi et la pondération est la même pour tous.

5.5.1. Une première façon de le réaliser est de tirer les unités primaires, les unités secondaires avec probabilités égales h. La probabilité générale est f = gh. Les probabilités g et h peuvent varier d'une strate à l'autre pourvu que le produit soit constant.
(Voir exemple 5.1.). Cette méthode donne des résultats très imprécis lorsque la taille des unités primaires est très variable.

5.5.2. On peut aussi tirer les unités primaires proportionnellement à la taille (voir 7.3. pour le mode de tirage) et tirer un nombre fixe d'unités secondaires dans chaque unité primaire. Cette méthode fournit des estimations plus précises et sera préférée à la première même si la taille n'est connue qu'approximativement.

5.5.3. On préfèrera la première méthode, si la taille est totalement inconnue ou si l'enquête doit porter sur des unités statistiques très différentes : commerce et population par exemple.

6. *TAILLE USUELLE DES ECHANTILLONS*

6.1. On doit déterminer la taille de l'échantillon pour que dans les tableaux prévus l'effectif de chaque case soit d'au moins trente individus. On peut se contenter de trente individus en moyenne, ce qui revient à un échantillon minimum égal à 30 fois le nombre de cases du tableau le plus détaillé.

6.1.1.- Pour une enquête fécondité, l'habitude est de travailler avec un échantillon de 2 000 à 8 000 femmes.

6.1.2.- Pour construire une table de mortalité l'échantillon doit être d'au moins 100 000 personnes.

6.1.3.- En pratique la taille de l'échantillon influe directement sur le budget. Une fois le questionnaire établi, on peut évaluer le nombre de questionnaires qui peuvent être remplis en une journée de travail.

On peut ainsi faire l'adéquation entre le budget et la taille de l'échantillon. Si le coût est trop élevé, on peut envisager des échantillons emboités : un échantillon plus grand avec un questionnaire réduit et sur un sous-échantillon plus petit, un questionnaire approfondi

6.2. La taille des grappes est importante, il faut éviter de trop grandes grappes, on considère habituellement que les grappes doivent comprendre quelques centaines de personnes au maximum.

7.- TIRAGE DES ECHANTILLONS

7.1. On procède presque toujours par tirage sans remise, c'est-à-dire qu'une unité tirée est otée de la liste pour les tirages ultérieurs.

7.2. Le tirage systématique est la méthode la plus utilisée de tirage sans remise.

7.2.1.- C'est le tirage effectué à intervalle fixe sur une liste à partir d'un point de départ pris au hasard.

7.2.2.- C'est un procédé commode quand on a à tirer un assez grand nombre d'unités sur une liste, par exemple des ménages.

7.2.3.- Si la liste est ordonnée selon un caractère quelconque, taille des localités, proportion d'agriculteurs, cette technique a le même effet qu'un échantillonnage stratifié avec fraction de sondage égale à l'intérieur des strates et on n'a pas besoin de choisir des limites de strates. La liste ne peut être ordonnée que sur une seule variable. Exemple : les individus sont numérés de 1 à .....

a) On veut prendre un individu sur 15 : on choisit un nombre au hasard (c'est-à dire dans une table de nombres aléatoires) entre 1 et 15 soit 6 par exemple.

L'échantillon est constitué des individus de rang 6, 21 (6 + 15) 36 (6 + 15 + 15), 51, 66,...

b) On veut tirer m = 22 individus parmi M = 327. On divise M par m - 1 le quotient est 15, et le reste 12. On choisit un nombre au hasard entre 1 et 12, 3 par exemple. L'échantillon est constitué des individus de rang 3, 18, 33, 48,...

c) Si la liste comprend plusieurs quartiers, il faut continuer la numérotation d'un quartier au suivant. Ainsi si le premier quartier comprend 44 individus, un tirage au 1/10è donne les individus 2, 12, 22, 32, 42. Il reste 2 individus qu'on reporte dans le deuxième quartier on prendra les individus 8 (8 + 2 = 10), 18, 28,... qui sont les individus 52, 62,... de la liste complète.

d) Quand le taux de sondage est faible et si les individus ne sont pas numérotés, on peut réaliser alors un sondage avec des individus équidistants, par exemple tous les 40 cm.

Au préalable, il faut vérifier que la liste n'est pas plus dense dans certaines parties ce qui pourrait entraîner un biais.

Une méthode voisine, pour une liste constituée de fiches, consiste à prendre une fiche tous les 5 cm par exemple. Si les dossiers sont d'épaisseur variable, on a plus de chance de tomber sur les gros

dossiers on retiendra alors par exemple le troisième dossier qui suit celui sur lequel on est tombé.

Si une liste est constituée de cahiers paginés, dont les lignes sont numérotées à chaque page, on tire un échantillon de pages et à l'intérieur de chaque page un échantillon de lignes.

7.3.- Une autre utilisation de l'échantillon systématique est d'effectuer un tirage proportionnel à la taille des unités.(Voir 5.5.2.).

On effectue le cumul des tailles des unités à tirer soit A le dernier cumul représentant la taille totale et m le nombre d'unités à tirer. Choisir au hasard un point de départ x entre 1 et A/m et sélectionner toutes les unités qui comprennent le x ième individu, le (x + A/m) ième, le (x + 2A/m) ième,.... le (x + (m - 1) A/m ième individu.

Avant d'effectuer ce tirage, il est intéressant de ranger les unités par taille croissante ou décroissante.

N° village	*Nombre d'exploitations*	*Nombre cumulé d'exploitations*
1	*70*	*70*
2	*60*	*130*
3	*54*	*184*
4	*44*	*228*
5	*31*	*259*
6	*30*	*289*
7	*28*	*317*
8	*25*	*342*
...	...	...
124	*10*	*1 590*
125	*10*	*1 600*
126	*9*	*1 609*
127	*7*	*1 616*
128	*6*	*1 622*
129	*5*	*1 627*
130	*5*	*1 632*

Si on a A = 1 632 exploitations pour un sondage au 1/20ème on divise 1 632 par 20 soit 81. On choisit un nombre au hasard entre 1 et 81 soit 55, on prendra les villages qui contiennent les exploitations numérotées 55, 136 (55 + 81), 217 (55 + 81 + 81), 298,...., c'est-à-dire les villages 1, 2, 3, 4, 7, ... 124. En effet le village 3 contient les exploitations entre 131 et 184 donc l'exploitation 136.

Le tirage proportionnel à la taille donne aux gros villages plus de chance d'être tiré, quatre villages sont tirés parmi les huit premiers alors qu'un seul village est tiré parmi les sept derniers.

Si le premier village avait compris plus de 136 exploitations, il aurait été tiré deux fois. Dans le cas du sondage à deux degrés décrit en 5.4. on aurait alors prélevé deux fois le nombre prévu d'unités secondaires (ici des exploitations).

8.- RESUME

Il faut stratifier selon des variables en liaison avec le phénomène étudié.

Pour chaque variable quantitative constituer deux ou trois strates de tailles voisines.

Pour les sondages à deux degrés on préfère tirer les localités proportionnellement à la taille au 1er degré et un nombre fixe d'unités au deuxième degré.

La taille des grappes doit être de quelques centaines de personnes.

REFERENCES BIBLIOGRAPHIQUES

1. DESABIE J. Théorie et pratique des sondages - DUNOD 1971, 483 p.
(Ouvrage de référence en langue française, malheureusement difficile à utiliser. Il fournit une bibliographie en langue anglaise p.470).

2.- WFS : Manuel de sondage - documentation de base Juillet 1976, 82 p.
(Très bon manuel dont la lecture est conseillée. Il est adapté à l'objectif des enquêtes fécondité mais passe en revue tous les problèmes essentiels).