DOCUMENT 2.1 : INFORMATIONS COMPLEMENTAIRES SUR LA METHODE D'ENQUETE

1 – Définir le type de variable

Dans notre cas, la variable est quantitative nominale.

Note :
Une variable est qualitative nominale quand ses valeurs sont des éléments d'une catégorie non hiérarchique. C'est-à-dire que ses éléments ne peuvent pas se ranger dans une gradation logique.

Variable	Catégories
Equipement des boites aux lettres	- Autocollant Stop Pub ou autre - Pas d'autocollant

2 – Définir le périmètre d'échantillonnage

Lorsque l'on travaille en statistique, on a rarement la possibilité d'interroger la population au complet. Il faut donc se contenter d'interroger quelques individus pour faire un portrait de la population. Pour s'assurer que notre échantillon est représentatif, il faut étudier les techniques d'échantillonnage.

Dans notre cas, la population est l'ensemble des logements d'un territoire. Ce territoire doit être défini au préalable : il s'agit de la zone qui fera l'objet d'une opération de promotion du Stop Pub. Cette zone peut être découpée en quartier.

3 – Calculer la taille de l'échantillon

L'échantillonnage introduit des erreurs et des imprécisions dans les paramètres tels que la moyenne et les pourcentages. Pour avoir une erreur d'échantillonnage raisonnable, nous calculerons le nombre d'observations nécessaires et évaluerons les erreurs d'échantillonnage.

Notes :
- Plus l'échantillon est important, plus la généralisation sera fiable. Mais, les gains de fiabilité ne sont pas proportionnels à l'augmentation de la taille de l'échantillon.
- La notion de précision est matérialisée par un niveau de confiance et une marge d'erreur. Par exemple : un échantillon défini à un niveau de confiance de 95 % et une marge d'erreur de 3 %, permettra d'extrapoler chaque résultat issu de l'enquête, avec 5 % de risques de se tromper de + ou – 3%.

Calcul de la taille de l'échantillon

Pour faire des analyses statistiques valables, il faut un nombre minimal de 30 observations. Mais cela risque être insuffisant. Les formules mathématiques nous montrent que plus le nombre d'échantillons est élevé, meilleure est la précision statistique. On peut estimer le nombre d'observations nécessaires en fonction de l'intervalle de confiance Z et de la marge d'erreur « E_m » en utilisant la formule ci-dessous :

$$n = \frac{Z^2 p(100-p)}{E_m^2}$$

Où :
n : taille de l'échantillon
p : pourcentage de l'échantillon n'ayant pas de Stop Pub ou équivalent
Z : pondération selon le niveau de confiance choisi
E_m : marge d'erreur

Facteur de correction :

En deçà d'une population de 100 000 logements (N), il faut introduire un facteur de correction à l'aide de la formule suivante :

$$n' = \frac{n}{1+\frac{n+1}{N}} \approx \frac{n}{1+\frac{n}{N}}$$

Où :
n' : taille de l'échantillon corrigé
p : pourcentage de l'échantillon n'ayant pas de Stop Pub ou équivalent
N : taille de la population (nombre de logements)

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

Exemple :
Dans notre cas, il est inutile d'avoir une précision importante. Fixons nous une précision de 5 % (E_m) et un niveau de confiance de 95 % (ce qui donne Z = 1,96 issus de la loir normale). En reprenant les données nationales d'équipement, le pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent peut être estimé à 17 % (p).

On cherche à calculer la taille de l'échantillon avec la formule ci-dessous :

$$n = \frac{Z^2 p(100-p)}{E_m^2}$$

En remplaçant les symboles par les chiffres, on obtient : n = 216. (Si l'on souhaite augmenter la précision, il est nécessaire d'augmenter la taille de l'échantillon. Par exemple : pour une précision à 3 %, l'échantillon doit être de 602 logements.)

Si le nombre de logement (N) est inférieur à 100 000, il faudra appliquer la formule de correction avec la formule ci-dessous :

$$n' = \frac{n}{1+\frac{n+1}{N}} \approx \frac{n}{1+\frac{n}{N}}$$

Exemple, pour Angers : N = 80 773 foyers et donc n' = 215.

Conclusion : pour une population de 80773 foyers, un échantillon comprenant 215 logements est nécessaire pour atteindre les objectifs de précision (5 %) et de niveau de confiance (95 %). (Pour un niveau de confiance de 95 % et une précision de 3%, l'échantillon comprendrai 597 logements).

3 – Réaliser l'enquête sur le terrain

Contrairement à un sondage téléphonique qui peut engendrer de nombreux biais (présence des habitants à l'horaire du sondage, refus de répondre à un sondage, ...), un relevé en porte à porte auprès d'un échantillon des logements apportera des données quantitatives plus fiables et plus rapides.

La méthode statistique proposée est une méthode probabiliste par échantillonnage stratifié. Le principe est de découper la population en sous ensemble, appelés strates (dans notre cas, ce sont des zones géographique ou quartier) et réaliser un sondage dans chacune d'elles.

Phase 1 **=** Découper le territoire en quartiers homogènes en nombre de logements. Le découpage peut être réalisé à partir des données INSEE tels que le nombre de logement par IRIS (données accessibles dans la rubrique : Données téléchargeables / Base de données infra communales de leur site internet). Par ailleurs, des cartes précisent le découpage infra communale en IRIS pour les communes de 10 000 habitants. Pour accéder aux cartes de positionnement et aux contours des IRIS :
http://www.insee.fr/fr/methodes/default.asp?page=zonages/iris.htm

Exemple d'échantillon :
L'INSEE donne 69 IRIS pour Angers avec pour chacun d'entre eux le nombre de logements. Le nombre total de logements est de 80 773 pour la ville d'Angers. Les IRIS étant trop petits pour un échantillonnage, il serait intéressant de les regrouper en 4 quartiers d'environ 20 200 foyers. L'échantillon par quartier sera alors de 215 / 4 ≈ 54 foyers pour un niveau de confiance de 95 % et une précision de 5% (où 597 / 6 ≈ 100 foyers pour un niveau de confiance de 95 % et une précision de 3%).

Phase 2 **=** Effectuer, de manière aléatoire pour n (ou n' si N < 100 000) logements un relevé en porte à porte du nombre de logements mentionnant un refus de la publicité par rapport au nombre de boites aux lettres total. Tous les habitats doivent être comptabilisés : habitat vertical, habitat individuel, établissement professionnel, ...

Pour rendre le relevé aléatoire, utiliser des astuces comme : choisir les numéros paires d'une rue, espacer les échantillons d'une même rue, relever la première boite aux lettres en haut à gauche des immeubles, … Il s'agira d'éviter les biais.

Phase 3 **=** Calculer le taux d'équipement pour chaque quartier avec la formule suivante : Taux d'équipement des boites aux lettres en autocollants Stop Pub ou équivalent = Nombre de boites aux lettres équipés d'un Stop Pub / Nombre total de boites aux lettres.

Phase 4 **=** Relier cette information au contexte local (action antérieure sur l'un des quartiers par exemple).

4 - Evaluer l'erreur sur le pourcentage

La formule est la suivante : $\Pi = p \pm Z \sigma p$

Où
Π : pourcentage de la population ayant un autocollant Stop Pub ou équivalent
p : pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent
Z : pondération selon le niveau de confiance choisi (dans notre cas, prenons un niveau de confiance raisonnable de 95 %, ce qui donne Z = 1,96)
σ_p : estimation de l'erreur sur le pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent

L'estimation sur l'erreur de pourcentage se calcule comme suit :

$$\sigma_p = \sqrt{\frac{p(100-p)}{n}}$$

σ_p : estimation de l'erreur sur le pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent
p : pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent
n ou n' si la N < 100 000 : taille de l'échantillon

Exemple :
Si suite à la réalisation du relevé en porte à porte, le pourcentage de boites aux lettres ayant un autocollant Stop Pub ou équivalent est estimé à 10 % (p). L'échantillon est de 215 foyers (n') pour une population de 80 773 foyers (N). On cherche l'estimation de l'erreur sur le pourcentage de l'échantillon ayant un autocollant Stop Pub ou équivalent grâce à la formule suivante :

$$\sigma_p = \sqrt{\frac{p(100-p)}{n}}$$

En replaçant les symboles par les chiffres, on obtient : σ_p =2,04. On utilise ensuite la formule :
$\Pi = p \pm Z \sigma_p$

Ainsi, avec une pondération selon le niveau de confiance de 95 %, (Z = 1,96), on peut affirmer que le pourcentage de la population ayant un boites aux lettres avec un autocollant Stop Pub ou équivalent est égal à 10 % ± 4 % avec une confiance de 95 %.
Le taux d'équipement est donc compris à 95 % de chance entre 6 % et 14 %. (Pour un niveau de confiance de 95 % et une précision de 3%, le résultat serait de 10 % ± 2,4 soit un taux d'équipement compris à 95 % de chance entre 7,6 % et 12,4 %.)

Conclusion :

Quel objectif ?
Construire un échantillon tel que les observations pourront être généralisées à l'ensemble de la population (méthode probabiliste).

Quelle Condition ?
Il faut que l'échantillon présente les mêmes caractéristiques que la population cible. En d'autres termes, qu'il soit représentatif. Si ce n'est pas le cas, l'échantillon est biaisé.

Pourquoi un échantillon ?
La population « cible » est généralement trop nombreuse et pour des raisons de coûts, de délais, il est pratiquement impossible d'étudier tous les individus d'une population c'est-à-dire d'effectuer un recensement.

Quelle taille d'échantillon ?
En général, pour trouver la taille nécessaire n' (pour que la marge d'erreur dans l'estimation de la proportion soit inférieur à 5 % et ce, pour un seuil de confiance de 95%), les instituts de sondage utilisent la formule simplifiée :

$$n' = \frac{385}{1+\frac{385}{N}}$$

Pour plus de précision (3% par exemple), vous devrez augmenter la taille de l 'échantillon selon les formules ci-dessus. Sans oublier qu'il est inutile d'avoir l'échantillon le plus large possible. A partir d'un certain seuil, l'augmentation de la taille de l'échantillon n'apporte qu'un gain de précision minime (Lois de Bernoulli).