Plan du cours

Introduction

Chapitre 1 : Planification de l'étude

Section 1 : Conception du projet

Section 2 : Méthode d'échantillonnage

Section 3 : Elaboration du questionnaire

Section 4 : Programmation de la formation

Section 5 : Elaboration du budget

Section 6 : Elaboration du chronogramme

Section 7 : Collecte des données

Chapitre 2 : Traitement des données

Section 1 : Apurement avant la saisie

Section 2 : Apurement après la saisie

Chapitre 3 : Analyse des données

Section 1 : Analyse uni variée

Section 2 : Analyse bi variée

Section 3 : Analyse factorielle

Section 4 : Test de la moyenne et des proportions

Section 5 : Analyse de la régression linéaire

Introduction

À première vue, les gens pensent que le déroulement d'une étude de marché consiste simplement à poser des questions et à compiler les réponses pour obtenir des statistiques. Cependant, il faut faire une étude en respectant les étapes, les procédures et les formules précises pour que les résultats obtenus donnent de l'information exacte et significative. Il faut bien connaître les tâches particulières, leurs liens et leur pertinence pour comprendre le processus complet.

Voici les étapes d'une étude de marché :

➦ Conception et évaluation du projet d'étude ;

➦ Construction de l'échantillon ;

➦ Construction et administration du questionnaire ;

➦ Elaboration du budget et du chronogramme ;

➦ Collecte des données ;

➦ Traitement des données ;

➦ Analyse des données.

Les cinq premières étapes sont consacrées à la planification de l'étude (chapitre 1). Ensuite, nous étudierons successivement les techniques de codification et de redressement des données (chapitre 2) et les outils classiques pour exploiter les informations (chapitre 3) avec les logiciels Excel, Sphinx et SPSS.

Destiné à la pratique professionnelle quotidienne, ce cours est conçu pour être non seulement une référence académique, mais aussi un guide pratique apportant des réponses aux questions les plus posées aux étudiants de master 1 à l'ESGAE.

Nombre d'exemples s'appuyant sur des situations réelles devraient permettre à ces étudiants de mieux comprendre l'utilité des préconisations faites et de les mettre en application avec certaine aisance.

Chapitre 1 PLANIFICATION DE L'ETUDE

La qualité d'une étude de marché dépend du soin apporté à la planification du projet envisagé pour répondre exactement aux besoins de son commanditaire. Parmi les causes d'échecs identifiées dans une étude du marché, la planification figure en bonne place. Raison pour laquelle, la plus grande attention doit être portée à cette phase initiale qui justifie l'importance et précise des objectifs de l'étude (section1), et détermine la méthode d'échantillonnage adaptée à la problématique (section 2). Cette phase permet également d'apprendre les techniques d'élaboration du questionnaire (section 3), de collecte des données (section 4), de formation des agents de terrain et de saisie des données (section 5) et d'élaboration du budget (section 6), tant sur le plan conceptuel que technique, avant que la collecte ne soit réalisée.

1.1Conception du projet

La première tâche de la planification d'une étude de marché est de déterminer le **pourquoi** de cette étude, quel est le **problème** et quels sont les **objectifs** à atteindre. Un énoncé clair des objectifs oriente toutes les étapes ultérieures de l'étude. Ces étapes devraient être planifiées de façon à garantir que les résultats en bout de ligne correspondent aux objectifs à atteindre. La conception d'un projet d'étude doit commencer par l'identification exacte du problème formulé par le commanditaire, afin de déterminer ensuite les informations nécessaires à sa résolution avant de déterminer le protocole technique adapté pour les obtenir. Par bien des aspects, cette conception du projet repose sur une démarche que l'on peut structurer de la manière suivante :

1.1.1 L'identification du contexte de la demande de l'étude

Elle dépend de la compétence de l'analyse et se base sur le concours du commanditaire qui se doit de **fournir toutes les informations nécessaires à la réflexion**, sans se retrancher derrière un prétexte de confidentialité qui risque d'entraver la bonne compréhension du problème à étudier. Ce qui exige une **confiance mutuelle** pour que le commanditaire et le concepteur de l'étude puissent échanger en toute transparence.

L'identification du contexte de l'étude peut se résumer en trois classes d'études: *les études liées à une décision stratégique* (concernent des choix majeurs qui **orientent structurellement les actions de l'entreprise**, tels qu'investir dans un nouveau secteur d'activité, lancer ou non un nouveau produit, se retirer d'un marché, etc.), *les études liées à une décision tactique et opérationnelle* (concernent les décisions liées aux **éléments du marketing-mix**, tels que déterminer le prix de vente d'un produit, analyser le niveau d'**efficacité** d'une publicité, etc.) et

les études liées à une décision politique (il existe deux types d'études dans ce cas ; celles qui visent à **comparer ou appuyer des décisions** et celles qui visent à **confirmer ou confronter des décisions stratégiques** ou tactiques déjà prises. Cette étude n'intervient pas directement dans la décision mais **alimente le débat et fournit les arguments ou des justifications** pour valider le bien fondé de certaines décisions déjà prises. Ce sont des études de contrôle et non de proposition et leur intérêt décisionnel est généralement limité).

1.1.2 L'identification des informations à recueillir

Cette étape se réalise quand les objectifs de l'étude (**général** et **spécifiques**) sont éclaircis. La diversité de ces informations est telle qu'il est illusoire de vouloir en dresser une liste. On peut cependant classer ces informations en trois groupes : *les informations qui concernent l'étude de l'environnement* (identification du marché et analyse de son potentiel, identification des clients et analyse de leur diversité, identification de la concurrence et analyse comparée de leurs forces, identification des intermédiaires et analyse de leur rôle, identification et analyse des facteurs de réussite sur le marché, etc.), *les informations qui concernent la gestion de l'activité marketing et commerciale de l'entreprise* (❶ options stratégiques liées à la décision d'investir ou non sur un marché, de se retirer d'un marché, de remplacer l'offre, de faire évoluer la marque, etc., ❷ décisions stratégiques liées au traitement d'une clientèle : **segmentation**, ❸ décisions tactiques liées au rôle précis concernant la politique d'offre [produit, prix, conditionnement, packaging, etc.], la politique de mise sur le marché [réseau de distribution, force de vente, actions merchandising, etc.], la politique de communication [publicité, actions promotionnelles, etc.], ❹ décisions tactiques liées aux multiples actions des éléments du marketing-mix : en matière de communication, de prix, etc.), *les informations relatives à la connaissance approfondie de la clientèle* (cerner ses attentes, son niveau de satisfaction, etc.).

1.2Méthode d'échantillonnage

Voici l'une des questions les plus souvent posées au réalisateur d'une étude de marché : Combien de personnes doit-on questionner lors d'une étude de marché ? Les gestionnaires sont anxieux d'obtenir une réponse à cette question fondamentale pendant la phase de la planification de l'étude parce qu'elle a des **répercussions directes** sur les considérations opérationnelles, notamment **sur le budget et le temps de réalisation de l'étude**. Il n'y a pas de solution magique ou de recette parfaite pour déterminer la taille de l'échantillon. Il s'agit plutôt d'un processus de compromis au cours duquel les besoins de précision des estimations sont pondérés en tenant compte de diverses contraintes opérationnelles, par exemple, le temps et les ressources disponibles (financière, humaines, etc.).

1.2.1 La sélection d'une base de sondage

La base de sondage est l'outil qu'on utilise pour avoir accès à la population. C'est une liste d'unités (individus ou groupe d'individus) qui couvre toute la population avec une identification de chaque unité. Cette liste doit être exhaustive (sinon défaut de couverture) et sans doublon.
Une base de sondage doit couvrir correctement la population cible, être mis à jour à chaque instant, être de bonne qualité, être aussi stable que possible dans le temps, accessible et facile à utiliser.
Il existe trois types de base de sondage :

➥*Les listes*: c'est une liste physique, notamment, un fichier de données, un imprimé d'ordinateur ou un annuaire téléphonique ; ou une liste conceptuelle, par exemple une liste de toutes les clients qui entrent dans un centre commercial ;

➥*Les bases aréolaires*: c'est une liste géographique dont les unités correspondent à des secteurs géographiques et dont les unités composantes sont des ménages, des fermes, des entreprises, etc.

➥*Les bases de sondage multiples*: c'est une combinaison de deux ou plusieurs bases (des listes et des bases aréolaires ou deux listes ou plus). Les bases de sondage multiples sont habituellement utilisées lorsqu'aucune base unique ne peut fournir la couverture nécessaire de la population cible. Lors d'une étude, on peut utiliser la base des entreprises téléphoniques et celle des mairies pour avoir les informations - presque - complètes d'une population.
Un organisme statistique peut habituellement utiliser, approfondir ou créer une base de sondage. La base choisie détermine la définition de la population de l'enquête et peut avoir des répercussions sur les méthodes de collecte des données, de sélection et d'estimation de l'échantillon, ainsi que sur le coût de l'étude et la qualité des résultats. Par exemple, si on utilise une liste de numéros de téléphone pour sélectionner un échantillon de ménages, tous les ménages n'ayant pas le téléphone seront alors exclus de la population observée.

1.2.2 La fixation d'une méthode d'échantillonnage

Il existe deux types de méthodes d'échantillonnage : L'échantillonnage probabiliste ou aléatoire et l'échantillonnage non probabiliste ou empirique. Les échantillons aléatoires sont constitués par tirage au sort dans la population mère pour laquelle on dispose d'une base de sondage. Alors que **pour l'échantillonnage non probabiliste, on n'a pas besoin de la base de sondage**. Une autre différence qui existe entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une « chance » d'être sélectionnée et que cette chance peut être

quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste; dans ce cas, chaque unité incluse à l'intérieur d'une population n'a pas une chance égale d'être sélectionnée.

1.2.2.1 L'échantillonnage probabiliste

Il est plus complexe, prend plus de temps et est habituellement plus coûteux que l'échantillonnage non probabiliste. On distingue 4 méthodes :

➡ L'échantillonnage aléatoire simple ;

➡ L'échantillonnage systématique ;

➡ L'échantillonnage stratifié ;

➡ L'échantillonnage par grappe.

❶ **Echantillonnage aléatoire simple**

L'échantillonnage aléatoire simple est une méthode qui consiste à prélever au hasard et de façon indépendante, n individus ou unités d'échantillonnage d'une population à N individus.

Dans ce cas, chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon. Cette méthode est appliquée quand la population possède des caractéristiques homogènes et que nous détenons une base de sondage.

Le choix du tirage peut se faire avec remise (un individu peut être choisi plusieurs fois) ou sans remise (un individu déjà choisi ne peut l'être de nouveau).

Si l'on note n la taille de l'échantillon et N la taille de la population, on peut tirer :

$\mathbf{N^n}$ échantillon avec remise ou C_N^n échantillon quand on fait un tirage sans remise.

Avantage de cette méthode : On peut espérer un échantillon « représentatif » puisque la méthode donne à chaque individu de la population une chance égale.

Inconvénients : la méthode n'est applicable que lorsqu'il existe une liste exhaustive de toute la population.

❷ **Echantillonnage aléatoire systématique**

L'échantillonnage systématique est une méthode qui exige aussi l'existence d'une liste de la base de sondage où chaque individu est numéroté de 1 jusqu'à N.

Notons n, le nombre d'individus que doit comporter l'échantillon (la taille de l'échantillon). L'entier voisin de $\frac{N}{n}$ sera noté « r » et appelé « raison de sondage » ou « pas de sondage » qui correspond aux nombres d'échantillons possibles que l'on peut tirer. On choisit au hasard un entier naturel d entre 1 et r (cet entier sera le point de départ). L'individu dont le numéro correspond à d est le premier individu. Pour sélectionner les autres, il suffit d'ajouter à d la raison de sondage.

Exemples

On a une population de 400 clients, on veut un échantillon de 100 individus.

De ce fait, r = 400/100=4

On n'a donc que 4 échantillons possibles (E_1, E_2, E_3 et E_4). Si on choisit $r \in [1; 4]$; on aurait :

E_1 = {1, 5, 9,13, 17, …, 397}

E_2 = {2, 6, 10, 14, 18, …, 398}

E_3 = {3, 7, 11, 15, 19, …, 399}

E_4 = {4, 8, 12, 16, 20,…, 400}

Avantages : facile à sélectionner parce qu'un seul individu est choisi au hasard.

On peut obtenir une bonne précision parce que la méthode permet de répartir l'échantillon dans l'ensemble de la liste.

Inconvénients : Les données peuvent être biaisées à cause de la périodicité s'il n'y a pas homogénéité des caractéristiques de la population.

❸ Echantillonnage stratifié

On divise la population en groupes homogènes (appelés strates ou segment en marketing), qui sont mutuellement exclusifs (selon l'âge, le sexe, la province de résidence, le revenu, etc.).

Les variables de stratification doivent être simples à utiliser, facile à observer et étroitement liées au thème de l'étude[1].

Proportionnellement à son importance dans la population, on calcule combien il faut d'individus au sein de l'échantillon pour représenter chaque strate et dans chacune des strates, on sélectionne des échantillons indépendants. On peut utiliser n'importe quelle des méthodes d'échantillonnage. La méthode d'échantillonnage peut varier d'une strate à une autre.

Exemples **:** Choisir par échantillonnage stratifié, 10 clients dans un groupe de 60, en tenant compte du fait que 30 d'entre eux vivent à Bacongo, 18 à Ouenzé et 12 à Talangaï.

Solution **:** On aura dans les 10 clients, 5 de Bacongo, 3 de Ouenzé et 2 de Talangaï.

❹ Echantillonnage par grappe et à plusieurs degrés

On sélectionne au hasard un certain nombre de grappes (des **groupes qui ne sont pas forcément homogènes à l'intérieur**) pour représenter la population.

➡Si on étudie tous les individus des grappes choisies, on parle de sondage par grappe.

➡Si, dans chaque grappe, on tire encore un autre échantillon, on parle de sondage à deux degrés. On peut avoir un sondage à plusieurs degrés (degré supérieur à 2).

[1] Déterminer le prix psychologique, le niveau de satisfaction des clients, le niveau de notoriété d'un produit.

Avantages : La méthode ne nécessite pas forcement une base de sondage puisque seuls les individus inclus dans les grappes comptent. Elle permet de limiter l'échantillon à des groupes compacts ce qui permet de réduire les coûts de déplacement, de suivi et de supervision.
Inconvénients : La méthode peut entraîner des résultats imprécis (moins précis que les méthodes précédentes) puisque les unités voisines ont tendance à se ressembler. Elle ne permet pas de contrôler la taille finale de l'échantillon.

1.2.2.2 L'échantillonnage non probabiliste

On oppose aux méthodes aléatoires les méthodes non aléatoires ou empiriques.
Dans le cas de l'échantillonnage probabiliste, chaque unité a une chance d'être sélectionnée. Dans celui de l'échantillonnage non probabiliste, on suppose que la distribution des caractéristiques à l'intérieur de la population est égale. C'est ce qui fait que le chercheur croit que n'importe quel échantillon serait représentatif et que les résultats, par conséquent, seront exacts.
On ne peut mesurer la fiabilité d'un échantillonnage non probabiliste; la seule façon de mesurer la qualité des données en résultant consiste à comparer certains des résultats de l'étude à l'information dont on dispose au sujet de la population. Encore une fois, rien ne fournit l'assurance que les estimations ne dépasseront pas un niveau acceptable d'erreur. Les statisticiens hésitent à utiliser les méthodes d'échantillonnage non probabiliste parce qu'il n'existe aucun moyen de mesurer la précision des échantillons en découlant.
Cette méthode est utilisée :

➥Pour des études exploratoires;

➥Pour réduire les coûts;

➥Pour l'analyse des petits échantillons ;

➥Quand il est impossible ou non envisageable d'utiliser la méthode aléatoire.

La méthode d'échantillonnage non-probabiliste est utilisée lorsqu'il n'est pas possible de constituer une liste exhaustive (base de sondage) de toutes les unités du sondage.
Il existe plusieurs méthodes d'échantillonnage non probabiliste, les plus connues sont les suivantes : la méthode de quotas, la méthode de volontaires, la méthode de boule neige, la méthode d'échantillonnage ciblé et la méthode de panel ouvert.

❶La méthode de quotas

La méthode de quotas est une technique d'échantillonnage qui s'impose à chaque fois qu'il faut utiliser un échantillon représentatif alors qu'il n'est pas facile de disposer d'une liste de sondage et que la structure de la population à étudier est connue. Elle repose non plus sur le hasard, au

sens statistique du terme, mais sur le raisonnement de l'analyste qui choisit des éléments composants l'échantillon en fonction de critères de sélection jugés pertinents en regard du problème à résoudre. Les critères de sélection sont appelés « variable de quotas ou de contrôle». Cette méthode consiste à construire un échantillon représentatif à partir de certains critères clés appelés « variable de quota ou de contrôle ».

Cette méthode est **largement utilisée dans les enquêtes d'opinion et les études de marché** notamment parce qu'il ne suppose pas de liste des individus de la population. On parle aussi d'**échantillonnage dirigé ou par choix raisonné**. On demande aux enquêteurs de faire un nombre d'entrevues dans divers groupes établis en tenant compte de la répartition de l'échantillon à partir des variables de quotas. L'enquêteur doit respecter son quota.

Il faut noter qu'on ne doit **retenir qu'un nombre restreint de quotas** (au maximum trois variables de quotas).

Avantages : L'échantillonnage par quotas est généralement moins coûteux que l'échantillonnage aléatoire. Il est également facile à administrer.

Inconvénients : Certaines unités peuvent n'avoir aucune chance d'être sélectionnées (voir la méthode de la **théorie des valeurs extrêmes**).

❷ La méthode de volontaires ou de convenance

On prélève l'échantillon à partir d'un groupe de volontaires (Exemples : expériences médicales ou psychologiques). Dans ce cas, l'échantillon est composé de toutes les personnes que l'analyse considère concernées (population cible) et volontaires (acceptent d'y participer) par le sujet de l'étude.

Avantages : Cette méthode est très facile à appliquer car l'enquêteur ne se préoccupe pas de convaincre les enquêtés. Il ne s'intéresse qu'à ceux qui ont accepté de répondre.

Inconvénients : échantillon biaisé car aucune stratégie de sélection n'est suivie pour s'assurer que l'échantillon ainsi constitué est bien représentatif de la population cible.

❸ L'échantillonnage ciblé ou échantillonnage à la place

Cette méthode consiste à obtenir un échantillon d'éléments recrutés en des lieux où la probabilité de rencontrer les personnes concernées par la thématique de l'étude est très élevée[2].

Avantages : Ce type d'échantillon est largement utilisé pour travailler sur des cibles structurellement bien identifiées mais dispersées ou de faible effectif.

Inconvénients : échantillon biaisé.

[2]Exemple d'une étude auprès des prostitués ou auprès des enfants de la rue.

❹ **La méthode de boule de neige**

C'est une méthode dont l'échantillon est composé d'individus recommandés par des personnes déjà enquêtées.

Avantages : Ce type d'échantillon est largement utilisé en milieu industriel. Il est également utile pour atteindre des populations très particulières ou pour pénétrer des milieux fermés.

Inconvénients : difficulté d'obtention d'un échantillon représentatif si vous n'avez pas pu avoir des recommandations.

❺ **La méthode d'access panel ou de panel ouvert**

L'access panel est un échantillon composé d'individus volontaires qui ont été recrutés au cours des études précédemment réalisées. L'usage de cette méthode est fréquent pour les études par Internet.

Avantages : Son principale avantage est de ne pas avoir à rechercher au hasard des personnes à interrogées, mais de sélectionner sur une liste celles qui répondent aux caractéristiques recherchées et de les contacter.

Inconvénients : Si la base n'est pas actualisée, on risque de ne pas retrouver les enquêtés au cas où ils avaient changé d'adresse ou de numéro de téléphone.

1.3Elaboration du questionnaire

Si toutes les phases de l'étude sont importantes et doivent être exécutées avec beaucoup de soin, la réussite finale de l'étude est fonction de la nature et de la qualité des données recueillies. Des données de qualité exigent essentiellement **un questionnaire bien conçu et des procédures uniformes pour le remplir**. Le questionnaire est au cœur du processus de collecte des données. Il a des répercussions importantes sur la qualité des données parce qu'il constitue le moyen de collecte des données. Il a aussi des répercussions sur l'image de marque que l'organisme statistique projette dans le public.

Les questions posées doivent être conformes à l'énoncé des objectifs de l'étude et permettre la collecte d'information utile pour l'analyse des données. **Elles doivent répondre à tous les besoins d'information**, mais chaque question devrait avoir une justification explicite pour être inscrite dans le questionnaire. Il faut savoir pourquoi chaque question est posée et à quoi servira cette information. La formulation de la question doit être claire. **Les questions doivent être réparties en séquences logiques pour le répondant**. Les questions doivent être formulées pour être faciles à comprendre et permettre au répondant d'y répondre précisément. Enfin, **le questionnaire devrait être mis à l'essai avant son application**, à l'aide de groupes de discussion, d'un prétest et d'autres méthodes décrites dans ce chapitre.

Le type d'informations que nous obtiendrons lors des entrevues avec les enquêtés dépendra de la nature des questions que nous leur poserons et de la manière dont nous les poserons.

La rédaction d'un questionnaire doit répondre à de nombreux critères : les questions doivent être formulées de façon claire et simple ; elles doivent être ordonnées d'une façon logique ; elles doivent enfin donner lieu à des réponses précises et objectives.

La longueur du questionnaire et sa facilité d'utilisation sur le terrain sont aussi des critères importants.

Pour concevoir un questionnaire, il faut respecter les étapes suivantes :

1.3.1 *La consultation avec les experts du domaine, les utilisateurs des données et les répondants*

Le statisticien ne connait pas tout, mais il est recherché partout. Dans son travail, il est dans l'obligation de travailler avec des gens qui maitrise le domaine dont il effectue l'étude. Le processus de consultation avec les utilisateurs des données commence lors de la formulation des objectifs de l'étude au cours de la phase de planification et continue pendant la conception et l'élaboration du questionnaire. Cette consultation approfondie est particulièrement importante pour les grandes études, sinon toutes, d'un organisme statistique. Une compréhension approfondie de l'utilisation des données devrait permettre à l'organisme statistique d'élaborer un questionnaire bien conçu qui répond aux besoins des utilisateurs.

Il faudrait consulter non seulement les utilisateurs des données, mais aussi les répondants, les experts de la matière de l'étude et ceux qui ont procédé à des études semblables auparavant, avant de formuler la version provisoire du questionnaire. Ils devraient pouvoir donner une rétroaction sur le genre d'information que les répondants peuvent fournir et aider à préciser les concepts à étudier. Rencontrer les répondants peut aider à identifier les questions et les préoccupations importantes pour eux et à obtenir des répercussions sur les décisions pertinentes à la matière du questionnaire. Cette intervention peut aussi aider à identifier les expressions et le langage qu'utilisent les répondants pour décrire les concepts de l'étude, et donner une bonne idée de la façon dont les catégories de questions et réponses devraient être formulées. Ces discussions peuvent se dérouler pendant des consultations approfondies ou en groupe de discussion.

1.3.2 L'examen des questionnaires précédents

D'autres enquêtes sont une bonne source d'information pour l'élaboration d'une étude. L'examen des questions posées dans d'autres études sur le même sujet ou un sujet semblable

peut être un bon point de départ lorsqu'il faut formuler une question (c.-à-d. rédiger une question). Lorsque l'on souhaite comparer les résultats de différentes études, il est préférable d'utiliser les mêmes questions. Il faudrait aussi examiner la documentation sur la qualité des données de ces études pour évaluer l'efficacité du questionnaire (par exemple, les problèmes de rédaction des questions, le fardeau de réponse, les taux de refus, etc.).

1.3.3 La rédaction de la version provisoire du questionnaire

Il est important de procéder à la mise à l'essai de toutes les versions (c.-à-d. les versions dans toutes les langues) du questionnaire auprès de répondants bien avant le début de la « vraie » collecte des données. Répondre à une question est un processus complexe. Les répondants doivent d'abord comprendre la question. Ils doivent ensuite faire un effort de mémoire ou fouiller des dossiers pour extraire l'information demandée. Ils doivent aussi réfléchir à la réponse exacte à la question et déterminer s'ils sont disposés à révéler l'information, en tout ou en partie. Ils répondent alors à la question. Chacun de ces processus peut être une source d'erreur.

1.3.4 L'examen du questionnaire auprès des utilisateurs des données

1.3.5 La mise à essai et la révision du questionnaire

Les méthodes appliquées aux mises à l'essai des questionnaires (matière, présentation, etc.) sont habituellement axées sur de petits échantillons subjectifs non probabilistes de répondants tirés de la population cible.

Voici les méthodes décrites dans les sections suivantes :

1.3.5.1 Le prétest

Le prétest (parfois intitulé essai préliminaire) est facile, le coût est raisonnable, et c'est une étape fondamentale de l'élaboration d'un questionnaire. S'il n'y a pas d'autres mises à l'essai du questionnaire, il faudrait au moins faire un prétest. La taille de l'échantillon du prétest peut varier de 20 à 100 répondants ou plus. Si le principal objectif est de repérer des problèmes de rédaction ou de séquence, très peu d'interviews sont nécessaires. Il faut en faire davantage (de 50 à 100) pour déterminer les catégories de réponse aux questions fermées, à partir des réponses aux questions ouvertes du prétest. Lors du prétest, le répondant n'est pas informé, il remplit simplement le questionnaire ou répond à l'interview pour refléter la situation lors de la collecte réelle des données. Le prétest indique seulement là où il y a un problème. Sans aller plus loin,

il ne détermine pas pourquoi il y a un problème ou comment le corriger. La mise à l'essai non officiel n'identifiera peut-être pas non plus tous les problèmes du questionnaire.

Voici à quoi sert le prétest d'un questionnaire : découvrir l'ordre ou la rédaction médiocres des questions, corriger des sauts, fermer certaines questions ouvertes, retrouver les modalités manquantes (NSP), reconnaitre les problèmes des non réponse (ne veut pas répondre) et des hors-champs (ne peut pas répondre), estimer la longueur d'un questionnaire, repérer les erreurs de présentation ou d'instructions du questionnaire.

1.3.5.2 Les groupes de discussion

Un groupe de discussion considère un sujet sélectionné par les participants choisis dans la population d'intérêt. Au cours des premières étapes de l'élaboration du questionnaire, les groupes de discussion peuvent aider à identifier les questions saillantes de la recherche. On fait appel à eux pour évaluer la compréhension du langage et de la rédaction des questions et des instructions de la part du répondant, ainsi que d'autres formulations et mises en forme des questions. Un animateur qui connaît bien les techniques d'interview des groupes et l'objectif de la discussion oriente le groupe de discussion. Chaque groupe comprend habituellement de six à douze personnes et la taille optimale est de sept à neuf personnes. Une séance en groupe de discussion demande habituellement deux heures environ. Le groupe de discussion est enregistré sur bande sonore (et parfois sur bande vidéo) que les observateurs peuvent entendre dans une salle derrière un miroir d'observation. Il est recommandé que ceux qui élaborent le questionnaire observent le groupe de discussion. Les observateurs n'interviennent pas dans la discussion du groupe, mais leurs observations peuvent servir à l'animateur à la fin de la séance du groupe de discussion.

1.3.5.3 L'essai pilote

Une enquête pilote se déroule pour observer toutes les étapes du processus de l'étude, y compris l'administration du questionnaire. Une enquête pilote est une « simulation » qui applique la version finale du plan d'enquête à petite échelle du début à la fin, y compris le traitement et l'analyse des données. Elle permet à l'organisme statistique de considérer les résultats du questionnaire pendant toutes les étapes de l'enquête (collecte, vérification, imputation, traitement, analyse des données, etc.). Le questionnaire est habituellement soumis à des essais approfondis à l'aide des méthodes susmentionnées avant l'enquête pilote.

1.3.6 La touche finale apportée au questionnaire

La conception du questionnaire est un processus itératif : des modifications sont continuellement apportées pendant l'élaboration et la mise à l'essai du questionnaire. Les objectifs et les besoins d'information sont formulés et réévalués, les répondants et les utilisateurs des données sont consultés, la version préliminaire des questions proposées est formulée et mise à l'essai, les questions sont examinées et révisées jusqu'à la formulation de la version finale du questionnaire. Lorsqu'il est décidé qu'il n'y aura pas d'autres modifications apportées au questionnaire, l'étape finale du processus est franchie. La touche finale est alors apportée au questionnaire et il est imprimé ou programmé, selon la méthode de saisie des données appliquée.

Une autre décision dans la conception d'un questionnaire concerne le type de réponse qu'il faut associer à chaque question : une réponse libre pour une question ouverte et une réponse imposée pour une question fermée.

1.3.6.1 La question ouverte

La formulation d'une question ouverte laisse au répondant toute la latitude pour construire librement sa réponse, avec ses propres mots et spontanément : aucune réponse lui est imposée. L'usage d'une question ouverte s'impose dans différents cas de figure :

➥Quand le contenu issu de la spontanéité constitue même l'objet de l'information recherchée. C'est le cas pour une étude qui veut savoir si les répondants ont une bonne connaissance sur un problème, par exemple ;

➥Lorsque la diversité des réponses est beaucoup plus étendue pour établir une liste exhaustive des modalités de la question (la profession d'une personne) ;

➥Quand on n'est pas sûr de connaitre toutes les réponses possibles à la question.

Les différentes formes des questions ouvertes sont :

➥*Question à réponse textuelles* : Qu'est-ce que nous devons ajouter dans notre produit pour qu'il soit meilleur ?

➥*Question à réponse numérique* : Combien de nos savons de toilette utilisez-vous par mois ?

1.3.6.2 La question fermée

La formulation d'une question fermée contient les modalités de réponses attendues entre lesquelles le répondant doit impérativement choisir. Sa rédaction implique donc que l'analyste connaisse les réponses qui peuvent être données et qu'il sélectionne celles qui l'intéressent en regard des objectifs de l'étude.

Il existe plusieurs types de questions fermées :

➥*Question dichotomique* : La question dichotomique est la version la plus simple d'une question fermée. Il s'agit souvent d'une question qui n'a que deux réponses possibles (oui/non, bon/mauvais, masculin/féminin, etc.) ;

➥*Question à choix unique :* c'est une question qui demande au répondant de ne choisir qu'une seule réponse dans la liste. Exemple : Quel est le réseau téléphonique que vous utilisez le plus souvent quand vous appelez à l'étranger ? (MTN, Airtel, Azur, Congo Telecom, **NSP**);

➥*Questions à choix multiples* : c'est une question qui demande au répondant de choisir au moins une réponse dans la liste. Exemple : quels sont les réseaux téléphoniques que vous utilisez ? (MTN, Airtel, Azur, Congo Telecom) ;

➥*Question avec classement (ou à réponse ordonnée)* : La question avec classement est un autre genre de question fermée et elle demande au répondant d'établir l'ordre des catégories de réponse, par exemple : ¨Quelles sont, selon vous, les priorités au Congo ? Veuillez les classer par ordre de priorité en inscrivant « 1 » à la réponse qui serait la plus utile, selon vous, « 2 » à la méthode qui serait la plus utile en second lieu, et ainsi de suite. (lutte contre le **chômage**, lutte contre la **pauvreté**, lutte contre l'**inflation**, lutte contre les inégalités, la santé, l'éducation, l'autosuffisance alimentaire).

Un questionnaire doit être bien ordonné. Il est souvent séparé par des sections. Chaque section est constituée des questions.

Les sections doivent être ordonnées. Il faut commencer par des sections ayant des questions introductives.

Exemple :

1- Nous voulons d'abord obtenir des renseignements généraux sur vous ;

2- Nous voulons maintenant savoir……..

3- Les questions suivantes portent sur ……

4- Pour ne pas trop abuser de votre temps, nous terminerons par deux ou trois petites questions.

Avant de commencer son interview, il faut d'abord se présenter en commençant par :

1- Donner le titre ou le sujet de l'étude (ne pas dire enquête, mais plutôt étude) ;

2- Identifier le commanditaire ;

3- Exprimer l'objectif de l'étude ;

4- Expliquer pourquoi il est important de remplir ce questionnaire (ne pas mentir) ;

5- Souligner comment seront utilisées les données ;

6- Préciser comment le répondant peut avoir accès aux résultats de l'étude ;

7- Indiquer que les réponses seront confidentielles ;

Vous devrez maîtriser votre questionnaire et éviter d'être trop lent.

La première question que l'on pose doit être facile à répondre, fermée et doit rendre le répondant à l'aise (pas stressant). Ne pas commencer par exemple à demander l'âge de l'enquêté.

Poser les questions les plus délicats quand le répondant se sent à l'aise et confiant (gentil, rassuré).

1.4Programmation de la formation

Au moment où notre questionnaire sera prêt à être pré-testé sur le terrain, nous devons avoir rassemblé tout notre personnel pour l'étude. Outre notre équipe de planification de l'étude, il nous faudra un personnel de terrain (enquêteurs, contrôleurs et superviseurs), un personnel de traitement des données (codificateurs et le personnel de saisie de données) et un personnel d'analyse de données.

Les effectifs dont nous aurions besoin dépendront évidemment de la dimension de l'étude et du délai prévu pour la collecte et le traitement des données. Pour choisir l'équipe d'étude, nous devons prendre en compte les facteurs suivants :

➥La personnalité : les membres de l'équipe doivent **être** capable de bien travailler ensemble, de s'adapter, de s'aider les uns les autres et d'avoir une attitude amicale envers les personnes interrogées ;

➥Le niveau d'instruction des enquêteurs ;

➥L'honnêteté des enquêteurs et des contrôleurs ;

➥L'esprit managérial des superviseurs ;

➥La maitrise de la zone de projet

➥L'expérience antérieure des **enquêteurs** ;

➥Les langues et dialectes parlés et écrits ;

➥L'état de santé des enquêteurs ;

Le sexe : si ce facteur est important pour des raisons culturelles dans la composition des équipes d'enquêteurs.

Nous devons prévoir un temps et des moyens suffisants pour apprendre aux **enquêteurs**, aux **contrôleurs** et aux **superviseurs** à accomplir leurs tâches correctement, efficacement et de manière standardisée.

1.4.1 L'enquêteur

C'est un agent de terrain. Il a pour objectif de :

1- Prendre l'information à la base ;

2- Maitriser toutes les questions du questionnaire ;

3- Poser clairement les questions aux enquêtés ;

4- Transmettre dans le questionnaire toutes les réponses avec le maximum de fidélité ;

5- Retourner tous les questionnaires et matériels[3] remis par le contrôleur en parfait état;

6- Respecter toutes les instructions transmises par le contrôleur.

1.4.2 Le contrôleur

C'est agent qui suit et encadre les enquêteurs sur trois aspects :

1- Organisationnel : il organise le travail de son équipe, sensibilise les unités de son échantillon, s'assure que chaque membre de son équipe dispose du matériel nécessaire au bon fonctionnement de l'enquête et transmet les informations importantes à son superviseur ;

2- Technique : il s'assure que l'enquêteur a bien rempli ses questionnaires et a couvert toutes les unités à enquêter de son échantillon ;

3- Pratique : il instaure au sein de son équipe un climat de confiance susceptible de favoriser le bon déroulement du travail. En cas de problème ou de conflit, il est le responsable. Il veille également au bon entretien et à la bonne conservation du matériel de l'enquête.

1.4.3 Le superviseur

C'est le principal responsable de l'enquête dans sa zone d'étude. Il organise le travail concernant toute sa zone, assure le tirage des échantillons de sa zone, sensibilise les chefs de sa zone, transmet les instructions importantes de la coordination à ses contrôleurs, valide les travaux effectués par tous les enquêteurs après vérification des contrôleurs et récupère tout le matériel auprès des contrôleurs.

La formation des intervieweurs doit être soigneusement planifiée pour qu'ils aient tous un rendement uniforme et la même compréhension des concepts de l'enquête.

Un bon programme de formation et un **manuel**[4]clairement rédigé sont essentiels si l'on veut avoir un personnel de terrain efficace.

L'étendu de la formation nécessaire dépend de la complexité de l'échantillon et du questionnaire : les enquêteurs, les contrôleurs, les superviseurs et les agents de saisie doivent comprendre parfaitement toutes les procédures qu'ils auront à appliquer.

[3] Ça peut être des badges, tablettes, sacs, etc.

[4]Ce manuel s'appelle : le manuel de l'enquêteur

1.5Elaboration du budget

Il s'agit ici de déterminer le coût de l'étude : savoir combien va couter cette étude ; de la conception du projet d'étude à la rédaction du rapport de l'étude.

Voici un exemple des rubriques à remplir dans la rédaction du budget d'une étude de marché.

RUBRIQUE	Unité	Quantité	Coût unitaire	Coût total
1. LOGISTIQUE				
Kit de terrain (Sacs, stylo, crayon, gomme, gilets, etc.)	Nombre			
Autres matériel				
2. DOCUMENTS TECHNIQUES				
Questionnaires	Forfait			
Autres documents techniques (manuel de l'enquêteur, etc.)				
3. TRAVAUX PREPARATOIRES				
3.1 Examen de la méthodologie et des autres docs				
Atelier d'adoption de la méthodologie				
Travaux de conception des questionnaires				
Travaux de rédaction des manuels des enquêteurs				
Atelier de finalisation des documents techniques				
Atelier de validation des documents techniques				
3.2 Formation des agents de terrains et de saisie				
Sensibilisation des autorités et médias				
Formation des enquêteurs				
Formation des contrôleurs				
Formation des superviseurs				
Conception des masques de saisie				
Formation des agents de saisie	Homme/jour			
4. COLLECTE DES DONNEES				
Prime des superviseurs de terrain				
Prime des contrôleurs				
Prime des enquêteurs				
5. EXPLOITATION DES DONNEES				
5.1 Saisie des données				
Mise à jour des masques de saisie				
Saisie des données				
5.2 Traitement des données				
Codification des questionnaires				
Redressement des données				
Tabulation des données				
6. REDACTION DU RAPPORT				
Rédaction du rapport de l'étude				
Impression du rapport d'analyse				
Atelier de validation du rapport de l'étude				
Atelier de publication du rapport de l'étude				
7. SOUS-TOTAL				
8. DIVERS ET IMPREVUS	2%			
9. TOTAL GENERAL				

1.6Elaboration du chronogramme

Il est question ici de dire combien de temps va durer l'étude : depuis sa conception jusqu'à la rédaction du rapport de l'étude.

Voici un exemple du chronogramme d'une étude de marché.

Activités	Octobre				Novembre			
	S1	S2	S3	S4	S1	S2	S3	S4
TDR de l'étude	7 jours							
Conception du questionnaire	7 jours							
Elaboration des manuels des enquêteurs		2 jours						
Recrutement des enquêteurs et contrôleurs		1 jour						
Conception du masque de saisie		1 jour						
Formation des formateurs		1 jour						
Formation des contrôleurs		1 jour						
Formation des enquêteurs		1 jour						
Sensibilisation des autorités et des entreprises		Courrier	Média					
Collecte de données sur le terrain			7 jours	7 jours	6 jours			
Formation à la saisie des données et aux contrôles				1 jour	4 jours			
Saisie des données de terrain					6 jours	3 jours		
Traitement et tabulation des données						2 jours		
Analyse des données et rédaction du draft du rapport						1 jour	4 jours	
Atelier interne de validation du draft du rapport							1 jour	
Atelier externe de validation du rapport								1 jour
Publication du rapport de l'étude								1 jour

1.7Collecte des données

Il existe plusieurs éléments de collecte de données : Papier, tablette, téléphone, ordinateur (internet). La méthode sur support papier est intitulée Interview Papier et Crayon (IPC) et la méthode sur support électronique est intitulée Auto-Interview Assistée par Ordinateur (AIAO) ou Interview sur Place Assistée par Ordinateur (IPAO) ou encore Interview Téléphonique Assistée par Ordinateur (ITAO). La différence entre l'AIAO, l'ITAO et l'IPAO se reconnait au niveau des différentes méthodes de collecte des données.

Voici les méthodes élémentaires de collecte des données :

1.7.1 L'autodénombrement

Le répondant remplit le questionnaire par autodénombrement sans l'aide d'un intervieweur.

Divers moyens peuvent servir à envoyer le questionnaire au répondant et à le retourner à l'expéditeur : le service postal, le télécopieur, un moyen électronique (y compris Internet) ou un enquêteur. (Si le questionnaire est retourné par télécopieur ou sur support électronique, une

ligne sécuritaire ou le chiffrage est alors nécessaire pour garantir la confidentialité des données du répondant). La méthode sur support papier est intitulée Interview Papier et Crayon (IPC) et la méthode sur support électronique est intitulée Auto-Interview Assistée par Ordinateur (AIAO).

1.7.2 Les interviews sur place ou le face-à-face

Un intervieweur aide le répondant à remplir le questionnaire. L'interview se déroule sur place, habituellement à la résidence du répondant ou en milieu de travail, même si elle peut avoir lieu dans un endroit public (p. ex., aéroport, centre commercial). La méthode sur support papier est intitulée Interview Papier et Crayon (IPC) et la méthode assistée par ordinateur est intitulée Interview sur Place Assistée par Ordinateur (IPAO).

1.7.3 Les interviews téléphoniques

Un intervieweur aide le répondant à remplir le questionnaire au téléphone. La méthode sur support papier est intitulée interview papier et crayon (IPC) et la méthode assistée par ordinateur est intitulée Interview Téléphonique Assistée par Ordinateur (ITAO).

Chapitre 2 TRAITEMENT DES DONNEES

Les **valeurs manquantes** ou **aberrantes** sont présentes dans pratiquement toutes les bases de données des applications réelles. Elles peuvent correspondre aux **erreurs de saisie** ou à la **naïveté de l'enquêteur**. La mauvaise gestion de ces valeurs peut conduire à l'induction de modèles erronés et à des analyses fallacieuses.

Le traitement des valeurs manquantes et/ou aberrantes est souvent une tâche exigeante, tant du point de vue méthodologique qu'en termes de calcul. L'objectif principal de ce chapitre est de décrire et de proposer des méthodes de traitement dans chaque cas.

A la fin de ce chapitre, l'étudiant doit être capable de faire la différence entre la **codification** et le **redressement**, entre une **non-réponse** (totale ou partielle) et un **hors-champs** (total ou partiel) ; de reconnaitre une valeur aberrante et d'être apte à traiter les données d'une base afin de la rendre exploitable.

De ce fait, nous allons d'abord présenter les différents contrôles qui se font avant la saisie des données (section 1). Ensuite, nous exposerons les notions du redressement (section 2); notamment celles de la non-réponse (totale et partielle) et des unités atypiques.

2.1Apurement avant la saisie

Dans certains cas pratiques, l'étape de la collecte des données se fait au même moment que celle du masque de saisie du questionnaire. Après la collecte des données, il est souhaitable de passer à la saisie des données sur ordinateur à l'aide des logiciels statistiques de saisie des données. Cependant, **avant la saisie des données, il est souvent recommandé de revoir les données provenant du terrain afin de s'assurer qu'elles ne sont pas de mauvaise qualité**. Parfois, s'il y a lieu, on codifie les questions ouvertes lors de l'étude.

En dehors de la codification, quelques contrôles peuvent être réalisés pour obtenir des données prêtes à être saisies.

Il existe plusieurs contrôles avant la saisie, notamment : le contrôle uni varié, le contrôle de cohérence interne, le contrôle de vraisemblance et le contrôle agrégé.

2.1.1 Le contrôle univarié

Comme le nom l'indique, il s'agit d'un contrôle qui consiste à vérifier les variables séparément via des techniques descriptives simples.

2.1.1.1 Le cas des variables qualitatives

Vérifier les modalités de chaque variable. Par exemple pour la variable sexe, vérifier qu'elle ne contient que deux modalités (1=Masculin 2=Féminin ; par exemple).

2.1.1.2 Le cas des variables quantitatives

Contrôler l'intervalle des modalités de la variable. Par exemple pour le cas d'une étude sur des personnes âgées de 15 à 24 ans, vérifier que la variable âge n'a que des valeurs comprises entre 15 et 24.

2.1.2 Le contrôle de cohérence interne

Dans ce cas, il existe deux types de contrôle :

2.1.2.1 Le contrôle logique

Il consiste à vérifier une variable en fonction des valeurs d'une autre variable. Exemple, si X est la variable « Avoir de l'électricité » et que Y est la variable « Dépense du courant », la variable Y aura une valeur si l'on répond Oui à la variable X. De ce fait, la variable Y dépend de la variable X.

2.1.2.2 Le contrôle algébrique

Ce contrôle consiste à vérifier l'égalité ou l'inégalité entre deux « groupes » de variables. Exemple, si l'on veut vérifier que le nombre de garçons et de filles d'un ménage est égal au nombre total d'enfants vivant dans ce ménage, on peut considérer que X1= nombre de garçons, X2=Nombre de filles, X=X1+X2 et Y=Nombre total des enfants. Dans ce cas, il faudra vérifier si X=Y.

2.1.3 Le contrôle de vraisemblance

Ce contrôle consiste à vérifier si réellement une variable est comprise entre deux autres variables. Exemple, on peut vouloir vérifier si la VA est réellement entre les dépenses et le CA (Dépense<VA<CA).

2.1.4 Le contrôle agrégé

Ce contrôle consiste à vérifier s'il existe des valeurs atypiques au sein d'un **groupe homogène.**

2.2Apurement après la saisie

Le redressement ou contrôle après la saisie consiste à analyser respectivement les quatre (04) étapes suivantes :

➡ faire la différence entre une non-réponse (NR) et un hors-champs(HC) ;

➡ traiter les non-réponses totales (NRT) ;

➡ traiter les données atypiques ;

➡ traiter les non-réponses partielles (NRP).

A la fin de ces quatre (04) étapes, on obtient un fichier dit « apuré ».

Nous verrons dans ce chapitre la distinction entre une non-réponse et un hors-champs.

En effet, on est en présence de valeur manquante lorsque, pour au moins une unité de l'échantillon, au moins une question posée (par un enquêteur, dans un questionnaire…) n'a pas reçu de réponse.

Il y a **non-réponse (NR)** quand il y a une valeur manquante dans l'échantillon parce que l'enquêté refuse de coopérer ou que l'enquêteur a oublié de remplir la valeur qu'il fallait.

On parle cependant de **hors-champ (HC)** lorsqu'il y a une valeur manquante parce que l'enquêté ne correspond pas à l'unité statistique de l'étude (**hors-champs total : HCT**) ou qu'il ne peut pas répondre à la question posée parce qu'elle ne lui concerne pas (**hors-champs partiel[5] : HCP**).

Pour une unité de l'échantillon, si toutes ou la plupart des variables mesurées sont manquantes, on est en présence de **non-réponse totale (NRT)**. Toutefois, s'il n'y a que quelques variables manquantes pour une unité de l'échantillon quelconque, on parle de **non-réponse partielle (NRP).**

De manière générale : $n = n_R + n_{NR} + n_{HC}$

Avec n, la taille de l'échantillon ; n_R, l'effectif des répondants; n_{NR}, l'effectif des non-répondants; n_{HC}, l'effectif des hors-champs.

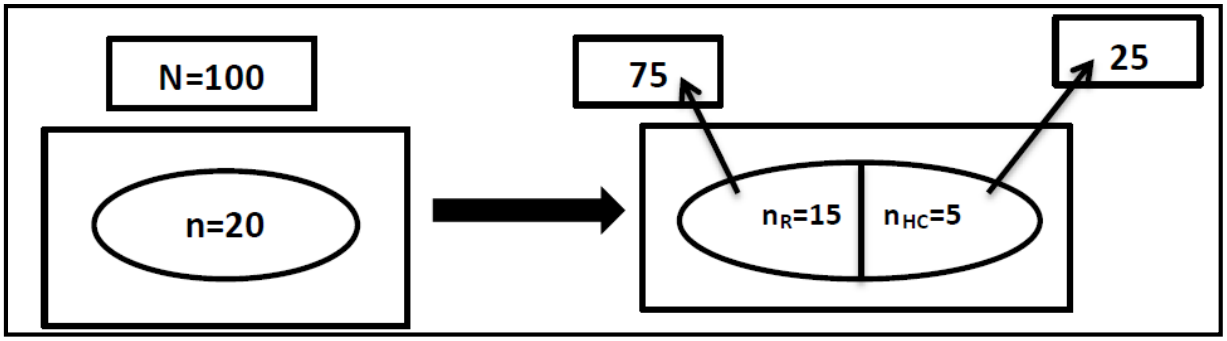
NB : Il est remarqué que le HC n'a aucun effet sur le poids de sondage. Par contre les NR ont des effets sur le poids de sondage.

Exemple :

Si on a les cas suivants :

[5] Cas des sauts dans un questionnaire.

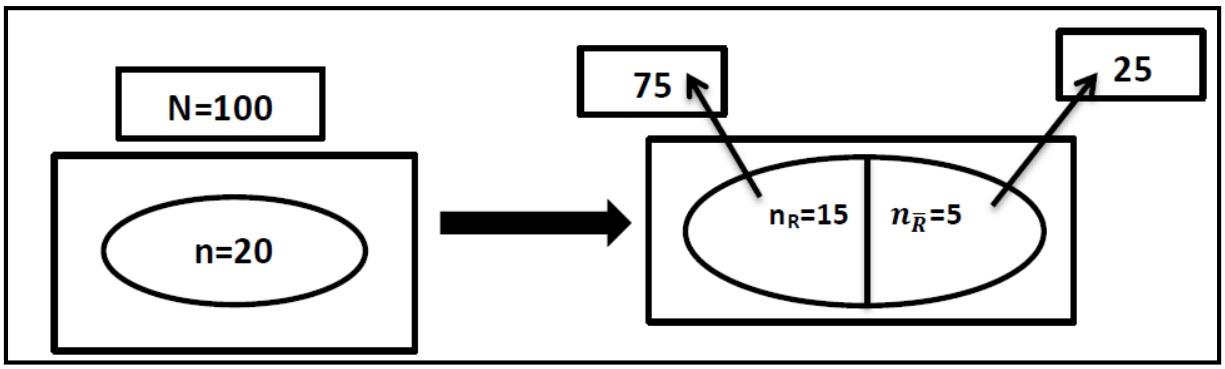
a) <u>Cas des hors-champs</u>



Avant l'étude, on a choisi d'enquêter un échantillon de 20 individus pour une population totale de 100 personnes. De ce fait, le poids de sondage est $w_1 = \frac{100}{20} = 5$. Ce qui signifie que chaque individu dans l'échantillon représente 5 personnes.

Cependant, après l'étude on constate 5 HC. Ce qui représente 25 individus de la population totale. Puisque les hors-champs sont considérés comme des individus ne faisant pas partie des unités statistiques voulues, on considère qu'en réalité il n'y a que 75 individus qui correspondent à la population totale recherchée. Ainsi, le nouveau poids de sondage est $w_2 = \frac{75}{15} = 5$.

b) <u>Cas des non-réponses</u>

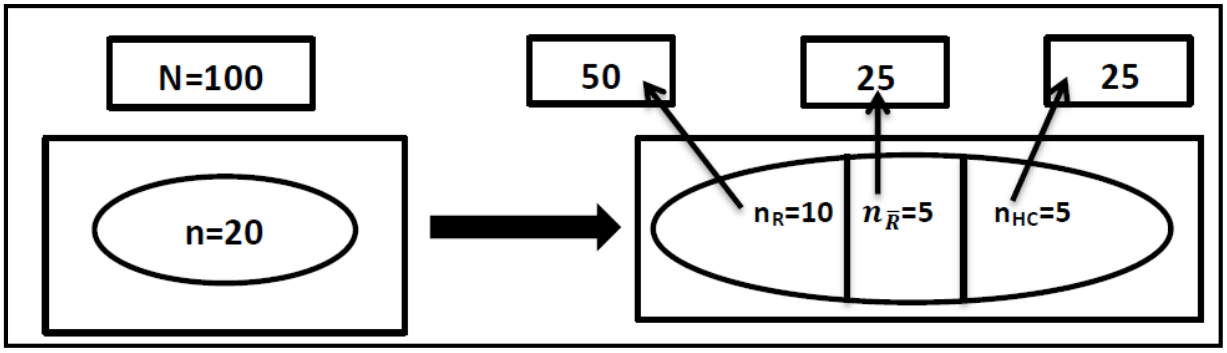


Avant l'étude, on a choisi d'enquêter un échantillon de 20 individus pour une population totale de 100 personnes. De ce fait, le poids de sondage est $w_1 = \frac{100}{20} = 5$.

Cependant après l'étude, on constate 5 NR. Ce qui représente 25 individus de la population totale. Puisque les NR sont considérés comme des individus faisant également partie des unités statistiques voulues, on considère qu'en réalité la population totale est toujours égale à 100 et que les répondants ne sont qu'au nombre de 15 au lieu de 20. Ainsi, le nouveau poids de sondage est $w_2 = \frac{100}{15} = 6{,}7 \neq w_1$.

Donc, chaque personne (les 15 qui ont répondu) représente maintenant 6,7 personnes au lieu de 5 seulement.

c) Cas des non-réponses et des hors-champs



2.2.1 Le traitement de la non-réponse

Nous avons vu précédemment que, d'une part $\boldsymbol{n = n_R + n_{NR} + n_{HC}}$; et d'autre part que les HC n'ont pas d'impact sur le poids de sondage.

Au regard de toutes ces informations, il en découle que dans le traitement des données,

$\boldsymbol{n = n_R + n_{NR}}$**.**

Dans le cas où nous avons des NR, il y a deux possibilités de traitement. Soit on supprime l'individu et on **répondère** (on fait la pondération) **;** soit on garde l'individu et on cherche à remplir les valeurs manquantes ayant causées ces NR. Dans le second cas, on parle de **l'imputation** (extrapolation en français).

Si on choisit de faire une repondération, le nouveau poids devient $\boldsymbol{W' = \frac{N}{n_R}}$**,** alors qu'il était avant $\boldsymbol{W = \frac{N}{n}}$**.** Ainsi, $\boldsymbol{W' > W}$.

Au cas où on se penche sur l'imputation, on considère que les individus qui ont des NR sont des receveurs et on cherche des donneurs qui ont répondu « exactement » comme des receveurs. Il faut noter qu'un donneur ne l'est qu'une seule fois.

Les deux méthodes ont chacune d'elles des avantages et des inconvénients dont nous énumérons dans le tableau ci-dessous.

	INCONVENIENTS	**AVANTAGES**
REPONDERATION	Détérioration de la précision des estimations car le poids de sondage n'est plus le même	Facile à manipuler
IMPUTATION	Erreur de sélection car on n'est pas sûr que le receveur devrait réellement donner la même réponse que son donneur s'il nous répondait en toute sincérité	Stabilité de la précision des estimations car le poids de sondage est toujours le même

Les deux techniques reposent sur l'hypothèse que le profil des répondants est le même que celui des non-répondants. La seule différence qui existe entre les deux, c'est que pour la

repondération, la taille de l'échantillon diminue (ce qui fait augmenter le nouveau poids de sondage), alors que pour l'imputation, la taille de l'échantillon reste le même.

A cet effet, existe-t-il une approche préférable à l'autre ? La chose la plus simple à faire est la repondération dans le cas des NRT car imputer toutes les NRT d'une base de données n'est pas un exercice facile. Ainsi, de manière générale, **on répondère dans le cas des NRT et on impute lorsqu'on a des Non-Réponse Partielle (NRP).**

Il existe deux types de méthodes d'imputation : *les méthodes déterministes* (imputation déductive, imputation historique, imputation par la moyenne totale ou par la moyenne de post-strate, imputation par k-plus proche voisin, cold-deck, imputation par ratio, imputation par régression déterministe) et *les méthodes stochastiques* (hot-deck simple, hot-deck multiple, bootsrap).

2.2.2 Le traitement des valeurs atypiques

Une valeur aberrante est définie comme une observation ou un sous-ensemble d'observations qui semble(nt) incohérente(s) par rapport aux autres données de l'ensemble.

Il est possible de faire la distinction entre des valeurs aberrantes **unidimensionnelles** (à une variable) et **multidimensionnelles** (à plusieurs variables). En effet, une observation est une valeur aberrante unidimensionnelle si elle est aberrante par rapport à une seule variable. Une observation est une valeur aberrante multidimensionnelle si elle est aberrante par rapport à deux variables ou plus.

Il est peut-être facile, par exemple, de trouver une personne mesurant deux mètres ou une personne pesant 45 kg, mais quelqu'un qui mesure deux mètres et pèse seulement 45 kg est un exemple de valeur aberrante multidimensionnelle.

Il faut également distinguer les valeurs extrêmes et les valeurs aberrantes. En effet, les valeurs extrêmes peuvent être ou ne pas être des valeurs aberrantes. Cependant, une valeur aberrante est toujours une valeur extrême de l'échantillon.

Chaque étude comprend des valeurs aberrantes pour à peu près chaque variable. Les valeurs aberrantes peuvent provenir de deux manière : l'individu est réellement différent des autres (erreur d'échantillonnage ou erreur due à la méthode d'échantillonnage) ou que la valeur a été mal saisie (une erreur de saisie). De nombreuses raisons expliquent les valeurs aberrantes.

2.2.2.1 Les conséquences des données aberrantes

➡ **Statistique descriptive** : augmentation de la variance, mauvaise orientation de l'axe principale (analyse factorielle)

➡ **Statistique inférentielle** : coefficients biaisés, etc.

2.2.2.2 Les méthodes de détection

a) **Contrôle uni varié** :

➡ **Variable qualitative** : vérifier les modalités de chaque variable : sexe, niveau d'instruction, etc.

➡ **Variable quantitative** : contrôler l'intervalle des modalités de la variable : Exemple : Pour la variable « Total des heures effectuées », une borne maximale (208 heures) est fixée à partir de la convention collective. Les valeurs supérieures à 208 heures sont aberrantes.

b) **Détection graphique** : Pour détecter la présence de valeurs aberrantes On peut utiliser :

- les box plot ;
- les histogrammes ;
- les nuages de points.

c) **Tests de cohérence logique** : On croise des variables. Exemple : « Taille », « Age » et « Poids »

d) **Détermination de plafonds au-delà desquels il est nécessaire de contrôler les réponses.**

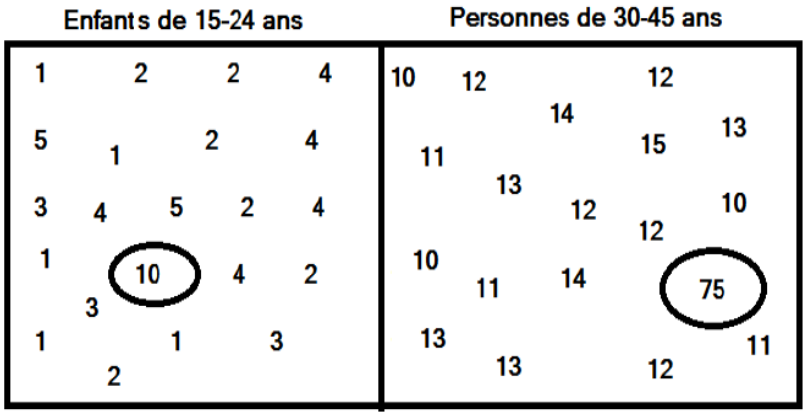
- On cherche les valeurs aberrantes en dehors de $[\bar{X} - 1{,}5(Q_3 - Q_1);\ \bar{X} + 1{,}5(Q_3 - Q_1)]$
- Selon Coulombe et McKay, X_j est une valeur aberrante si $ln(X_j) > \overline{ln(X)} + 3\sigma(ln(X))$

e) **Techniques classiques d'analyses multi variées** (analyse discriminante, analyse factorielle des correspondances, analyse en composantes principales) offrent des possibilités d'identification de valeurs anormales.

Remarque :

- Pour détecter des valeurs aberrantes on peut être amené à calculer de nouvelles variables : Exemples : Total des heures effectuées par employé, total des heures payées par employé ou montant des salaires bruts payés par employé.
- Toute utilisation de méthodes de détection de valeurs aberrantes par ordinateur doit tenir compte des limites des méthodes fournies par les logiciels.
- Une valeur est aberrante si elle engendre un effet de surprise en fonction de ce qu'on attend à partir du modèle. On compare les résultats obtenus à partir du fichier sans la valeur aberrante à ceux obtenus à partir du fichier avec la valeur aberrante.

2.2.2.3 La différence entre inliers et outliers



2.2.2.4 Les méthodes de traitement

Il y a deux (02) possibilités pour traiter les données aberrantes :

- Les valeurs aberrantes pouvant provenir d'erreurs de saisie. Si c'est le cas, on retourne au questionnaire papier quand c'est possible et on corrige. Si on ne retrouve pas le questionnaire, on les supprime et on applique ensuite une des méthodes d'imputation (moyenne, médiane, etc.). Il faut noter que dans la présence d'une mesure aberrante, la médiane des données ne change pas. La médiane est robuste (généralement, il ne varie pas beaucoup) en présence d'un petit nombre de valeurs aberrantes : par contre la moyenne change rapidement.
- Si la valeur a été bien saisie (erreur d'échantillonnage ou due par la méthode d'échantillonnage), on fait les analyses avec et sans elle pour juger s'il est nécessaire de la laisser dans la base.