Première partie : éléments de base sur les échantillons en statistique

Introduction

Combien de personnes devons-nous interroger pour qu'une enquête soit crédible? Voilà certes la question que tout chercheur se pose avant d'entreprendre une démarche quantitative. On retrouve également la même interrogation dans l'esprit des lecteurs de cette même recherche. Ces derniers peuvent en effet remettre en question les conclusions d'une étude parce qu'un trop petit nombre de personnes, à leurs yeux, ont été interrogées ou encore à cause d'une certaine faiblesse méthodologique dans la constitution de l'échantillon. La question du nombre de répondants n'est, en effet, qu'un des éléments qu'il faut considérer dans le processus d'enquête. En plus de la question du nombre de répondants, il est approprié de s'interroger sur la manière de choisir ces derniers, sur l'utilisation à faire des données recueillies et sur le type d'analyse à privilégier. Toutes ces dimensions auront un impact sur les résultats d'une recherche qui se base sur une enquête. Plus encore, l'application de certaines règles de base permet de répondre avec confiance à la question « Comment, à partir d'un petit nombre, pouvons-nous extrapoler sur l'ensemble? »

Nous comprenons qu'une utilisation abusive et techniquement faible de la statistique et, dans le cas qui nous occupe ici, la manière de recueillir les informations peuvent remettre en question les résultats de travaux qui autrement pourraient contribuer grandement à l'avancement de la connaissance.

L'objectif de ce document est donc de fournir un guide, un support technique, quant à la manière d'utiliser l'outil statistique, notamment en ce qui concerne l'échantillonnage et plus particulièrement les petits échantillons. Les petits échantillons demeurent un sujet peu traité dans les manuels de statistiques. Ce sont surtout les techniques d'analyse à partir d'un nombre

restreint d'observations qui ont fait l'objet d'une réflexion. Cependant, pour comprendre comment appliquer ces techniques, il est essentiel de bien maîtriser les principes de base de l'approche statistique. Nous avons donc jugé opportun de faire un retour sur ces principes avant d'aborder le sujet plus précis de l'utilisation des petits échantillons.

Ce texte s'adresse principalement aux personnes pour qui la statistique ne représente pas leur domaine d'expertise. Nous avons voulu simplifier et démystifier la technique d'échantillonnage afin que tous puissent maximiser l'utilisation de cet outil important. Comme son titre l'indique, ce document est plus un guide qu'un texte technique. C'est pourquoi nous avons limité, dans la mesure du possible, les notations algébriques. Ce texte sert essentiellement à mettre en lumière les principes de base dans l'application de la statistique, notamment au chapitre de l'échantillonnage.

Objectif de la statistique

La statistique sert, entre autres, à décrire, comprendre et estimer (ou projeter) une situation. Les outils statistiques et les techniques d'analyse particulières varieront en fonction de l'objectif poursuivi.

Lorsque l'on travaille à partir d'un échantillon, la statistique descriptive rend simplement compte des observations faites à partir de cet échantillon. On tente de comprendre une situation lorsque l'on est en mesure d'expliquer les différences observées dans l'échantillon au moyen de variables qui expliquent ces différences : par exemple, on observe la présence dans une population de personnes unilingues et de personnes bilingues et les variables servant à expliquer le phénomène peuvent être nombreuses (niveau de scolarité, présence de médias en une autre langue, etc.). Enfin, on projette sur l'ensemble de la population lorsque les données d'un **échantillon** permettent de généraliser sur l'**ensemble de la population**. Il faut donc prendre les moyens pour s'assurer

que l'échantillon est représentatif avant de faire cette opération de généralisation.

Cependant, les statistiques n'auront de sens que si les données utilisées sont crédibles. Pour cela, il faut que le nombre d'observations soit suffisant et fiable. Les personnes ou éléments qui seront utilisés dans ce contexte doivent donc être représentatifs de l'ensemble de la population à l'étude. On parle d'échantillon lorsque l'on étudie une partie – plutôt que la totalité – d'une population dans le cadre d'une enquête. Si l'échantillon est bien construit et, surtout, s'il est représentatif de l'ensemble dont il est issu, il devient alors sécuritaire d'utiliser les données dans une démonstration. En fait, l'opération statistique, par la médiation d'un échantillon, vise à réduire l'erreur de décision que l'on commet lorsque l'on présente un résultat. Cette notion d'erreur demeure omniprésente dans l'application de la statistique.

La démarche statistique (ou quantitative) comporte trois grands axes : la quantification, l'analyse et l'interprétation. Ces trois axes sont intimement liés entre eux. La quantification représente l'étape où on traduit une question scientifique («l'aspirine peut-elle contribuer à réduire les risques d'attaque cardiaque?») en un problème scientifique («les conditions objectives de l'utilisation de l'aspirine dans la prévention des crises cardiaques»). C'est de cette étape qu'il est question dans le présent texte. L'analyse représente l'application de techniques essentiellement mathématiques qui mettent en évidence certaines particularités des valeurs mesurées. Enfin, l'interprétation sert à tirer des conclusions scientifiques des résultats de l'analyse statistique[1].

Les trois étapes mentionnées dans le paragraphe précédent sont intimement liées. Cependant, l'analyse statistique, notamment lorsqu'il est question d'un petit nombre d'observations, est particulièrement délicate. Mais

[1] Ce paragraphe est inspiré de Jacques Allard, *Concept fondamentaux de la statistique*, Montréal, Addison-Wesley, 1992.

que l'on compte sur un grand nombre d'observations ou sur peu d'éléments, les principes d'échantillonnage demeurent les mêmes. C'est pourquoi nous avons cru nécessaire d'insister sur les points fondamentaux de l'échantillonnage. Nous savons par ailleurs que les chercheurs, et les éditeurs de revues scientifiques, préfèrent travailler avec un niveau de confiance élevé par rapport à leurs résultats. On exige souvent un niveau de confiance de 95 %. C'est-à-dire que le chercheur désire être sûr à 95 %, ou 19 fois sur 20, de ne pas commettre d'erreur dans les chiffres qu'il propose. Cela demande une approche rigoureuse, notamment lorsque le chercheur ne dispose pas d'un grand nombre de sujets étudiés.

Décrire, comprendre et projeter

Lorsque l'objectif d'une recherche est simplement de décrire une situation ou un ensemble de données, il est d'usage d'utiliser la statistique descriptive. La statistique descriptive se distingue de l'inférence statistique qui vise, elle, à extrapoler sur la population entière les résultats d'une enquête portant sur un échantillon.

On décrit une situation par le biais de statistiques qui, en un chiffre ou groupe de chiffres, résument une situation, un état ou un problème. Par exemple, l'âge moyen des élèves d'une classe est une valeur qui résume l'âge de tous les membres dans cette classe. La moyenne est la valeur qui permet de trouver un équilibre entre toutes les valeurs d'un groupe (ou d'un groupe d'objets). Lorsqu'il devient impossible d'interroger toutes les personnes qui composent une classe, nous utiliserons une fraction des personnes présentes. Le résultat obtenu sera par la suite utilisé pour représenter tous les membres. L'échantillon, la fraction, sert donc à décrire l'ensemble, la population.

Il existe trois mesures de tendances centrales : la moyenne, la médiane et le mode. Ces mesures servent à représenter l'ensemble de toute la population.

Cependant, l'utilisation de ces mesures peut parfois prêter à confusion. Prenons l'exemple suivant :

Il y a dans une classe 50 % de filles et 50 % de garçons. Il y a donc autant de filles que de garçons. L'âge moyen est de 18 ans. La classe est donc composée d'élèves qui gravitent autour de cette moyenne. L'âge médian, l'âge qui sépare en deux parties égales toutes nos observations, est également de 18 ans. Nous pourrions dire alors que l'âge des élèves est relativement similaire pour tous. L'étude conclut que nous observons une classe relativement homogène au niveau de l'âge. Mais voilà, la classe n'est composée que d'un garçon et d'une fille dont l'âge est respectivement de 22 et de 14 ans. Nous avons effectivement 50 % de garçons et 50 % de filles. L'âge moyen et médian est effectivement de 18 ans. Cependant, ces informations ne rendent pas compte de la réalité.

Même si cet exemple semble extrême, il montre que l'utilisation parfois abusive, et simpliste, des statistiques peut davantage nuire à une bonne compréhension d'une situation qu'à l'éclairer. Comment expliquer cette mauvaise utilisation des chiffres? Premièrement, nous n'avions pas défini notre population. De qui parlions nous? Une classe? Mais une classe de quoi? Pourquoi seulement deux personnes étaient-elles dans la classe? Pourquoi avoir transformé en pourcentage ces deux observations? Il est évident que si nous avions mentionné un garçon et une fille plutôt que 50 % et 50 %, la réaction du lecteur eût été différente. C'est également pourquoi un statisticien, ou une personne sensible aux éléments quantitatifs, demeure prudent lorsqu'il est question de pourcentage. On doit toujours s'interroger sur le nombre total utilisé pour calculer les pourcentages, la moyenne, la médiane ou le mode. Mieux, il faut, dans un premier temps, définir précisément la population étudiée. Il est insuffisant de dire qu'une étude porte sur les classes. De quelle classe s'agit-il? Il faut également préciser si l'étude porte sur les élèves masculins et féminins, en milieu urbain et rural, etc. Dans un tel cas, la simple description sera insuffisante.

Il faudra déterminer si les deux sexes, si les urbains et les ruraux se comportent de la même façon. À ce moment, l'outil statistique nous permettra de **comprendre** si les élèves se comportent de la même façon ou si, au contraire, ils sont différents et, surtout, pourquoi ils sont différents. Finalement, si les résultats doivent servir à **projeter** sur l'ensemble de la population des élèves, les outils statistiques devraient permettre d'extrapoler et de valider des théories.

Aussi une mauvaise définition du sujet de l'étude, une mauvaise planification ainsi que des choix douteux quant à la sélection des individus peuvent venir invalider les conclusions d'une étude qui autrement pourrait être parfaitement valable. La racine d'une étude statistique réside dans la manière de recueillir l'information. Le nombre de sujets inclus dans l'étude nous permet d'être plus précis et d'examiner plus de questions.

Ce qu'il faut retenir de l'exemple précédent, c'est que la justesse d'une situation dépend largement de la compréhension de la population étudiée. Remarquez que le nombre de personnes retenues pour faire l'évaluation est assurément trop petit. On ne peut en effet tirer de conclusions à partir d'un si petit nombre. Le même principe s'applique pour la sélection des personnes qui feront partie de notre échantillon. La sélection des personnes influencera les résultats que nous obtiendrons et, par conséquent, les conclusions que nous en tirerons. Cela sera d'autant plus vrai lorsque le nombre de personnes interrogées sera relativement petit. Des données aussi simples que la moyenne, la médiane et les pourcentages peuvent nous entraîner vers une fausse perception de la réalité lorsqu'elles sont mal employées.

Comme nous avons pu le constater, une description sommaire, si elle n'est pas contextualisée, peut entraîner une perception qui est loin de la réalité. La même chose peut se produire lorsque l'on veut comprendre une situation.

Pour une étude sociolinguistique, il devient alors important de bien cerner l'objet. Il ne suffit pas d'interroger beaucoup de monde mais d'interroger des personnes en fonction d'un **plan d'analyse statistique prédéfini**.

Population, échantillon et le langage des statistiques

La statistique utilise un langage qui lui est propre. Le statisticien parle algèbre et l'algèbre a ses propres règles de « grammaire ». Bien qu'il ne soit pas nécessaire de parler algèbre pour faire des statistiques, quelques notions de base sont utiles, notamment lors de la manipulation d'un petit nombre d'observations[2]. Mais, comme dans toutes les langues, des exceptions existent. Nous soulignerons ces exceptions dans le texte.

Il est essentiel, dans un premier temps, de bien définir la population à l'étude. Une population, dans le sens statistique du terme, est un ensemble d'individus ou d'unités statistiques qui composent la totalité de l'univers qui est étudié. Si l'étude porte sur les jeunes de 15 à 18 ans, alors la population sera tous les jeunes de 15 à 18 ans. Si l'étude porte sur les jeunes de 15 à 18 ans d'un continent, alors la population sera tous les jeunes de 15 à 18 ans de ce continent et ainsi de suite. Une population peut être très circonscrite ou très générale. Il est cependant primordial, d'un point de vue statistique, de bien connaître cette population afin d'établir le plan échantillonnal optimal. Plus la population aura des caractéristiques différentes, plus il sera difficile d'utiliser des tests statistiques propres aux petits échantillons. L'analyse à partir de petits échantillons sert surtout à examiner des points très précis et non des généralités.

L'échantillon, pour sa part, ne compte que sur une partie de la population. On utilise un échantillon lorsqu'on désire connaître les caractéristiques d'une population mais qu'il est trop difficile, pour des raisons pratiques ou financières,

[2] Essentiellement, la statistique utilise une notation à base de lettres grecques et latines. En général, les lettres grecques représentent les échantillons et les lettres latines les populations.

d'interroger l'ensemble de cette population[3]. Il s'agit donc dans un premier temps de déterminer si l'enquête touchera l'ensemble de la population ou une partie seulement. Il est important de déterminer si les personnes interrogées représentent bien la population générale ou une sous-population. Il est en effet possible que la population se subdivise en sous-populations. Dans un tel cas, l'échantillon devra tenir compte de cette caractéristique. Une sous-population peut être définie en fonction du sujet étudié. Par exemple, si l'étude porte sur l'utilisation de certaines expressions, il est possible que ces expressions soient conditionnées par le lieu de résidence (urbain *vs* rural), l'origine sociale, le sexe, le pays d'origine, etc. Dans un tel cas, nous serions en face de sous-populations ayant des caractéristiques propres. Si l'objectif de la recherche est d'évaluer les différences d'origine sociale ou territoriale du locuteur, on doit alors envisager la constitution d'un échantillon de plus grande taille. Si, au contraire, l'étude se prête bien à un environnement «contrôlé»[4], un plus petit échantillon pourrait alors être utilisé.

Il est donc essentiel de bien définir la population et les sous-populations qui composent notre champ d'étude. Il est également important d'établir si le chercheur désire décrire les caractéristiques d'une population, ou d'une sous-population, de comprendre les raisons du comportement de cette population ou encore de projeter sur l'ensemble de toute la population, ou des sous-populations, le comportement éventuel des personnes qui la composent. La clé réside dans l'échantillonnage et dans la technique d'analyse.

[3] Dans le cas où tous les éléments d'une population sont sélectionnés on parle d'un recensement.

[4] Un exemple d'un environnement contrôlé serait une évaluation qualitative d'une réaction physique (mouvement des yeux, expression générale du visage, etc.) lorsqu'un sujet est soumis à des expressions linguistiques particulières. L'évaluation se fait alors en fonction de la réaction et en rapport avec la connaissance ou l'absence de connaissance préalable de l'expression. En sociolinguistique, un exemple-type serait la technique du *matched guise* ou *locuteur masqué.*

La taille de l'échantillon

Combien de personnes devons-nous interroger? Afin de répondre à cette question, il est essentiel de s'interroger sur la raison d'être de l'enquête. En fait, il existe deux façons d'envisager la question. Voulons-nous brosser un tableau général de la situation (même de façon approximative) ou obtenir des données minutieuses et précises? La taille de l'échantillon est donc tributaire de la précision recherchée. Il est donc plus juste de se demander : «Quelle est la taille de l'échantillon qui assure un niveau de précision acceptable pour les fins de la recherche? ». Plus spécifiquement, il s'agit d'évaluer si la précision recherchée est pour une moyenne («le fumeur moyen a 30 ans»), un total («un million de fumeurs»), une proportion («60 % des fumeurs sont des femmes») ou un effectif («il y aura 1,2 millions de fumeurs dans deux ans»). Il faut donc s'interroger sur les besoins de la recherche afin d'appliquer correctement le test le plus approprié. Le résultat recherché doit-il être précis à un dixième de point près ou serions-nous satisfaits d'un écart de plus ou moins 3 %?

Il existe un principe voulant que la taille de l'échantillon est indépendante de la taille de la population en ce qui concerne l'erreur échantillonnale[5]. Par exemple, un échantillon de 500 personnes pour une école de 3 000 élèves, une ville de 100 000 habitants ou un pays de 25 millions d'habitants offre la même erreur échantillonnale dans chacun de ces trois cas.

La taille de l'échantillon dépend de plusieurs autres facteurs. Nous en examinerons deux. Le premier est la marge d'erreur que nous sommes disposés à tolérer. Le deuxième facteur porte sur la connaissance statistique préalable de

[5] L'erreur d'échantillonnage est la variation observée, et due au hasard, entre les échantillons. À partir d'une même population, il est possible de tirer plusieurs échantillons. Chaque échantillon donnera un résultat différent. Toutefois, l'écart entre les échantillons ne variera qu'à l'intérieur d'une fourchette dont le pourcentage est déterminé par la taille de l'échantillon. Si on examine la note méthodologique des sondages publiés dans les quotidiens et les magazines, on peut lire une formule du genre : «Pour une enquête de cette taille, l'erreur échantillonnale ne s'écartera pas de plus ou moins 4 % (ou 3,5 % ou 5 %, etc.) dix-neuf fois sur vingt.» Cela signifie que, si l'on prend 100 échantillons de la même population, il est possible que cinq de ces échantillons s'écartent de la marge d'erreur de 4 % (ou de 3,5 % ou de 5 %, etc.).

la population enquêtée, connaissance qui peut provenir de recensements, d'enquêtes d'opinion, etc.

Différences de taille d'un échantillon pour une population finie et infinie

Du point de vue statistique, une population finie est une population dont on connaît la taille au départ et elle est généralement petite (par exemple, les 200 élèves d'une école). Une population infinie est une population dont on ne connaît pas la taille exacte ou qui est relativement grande (par exemple, tous les élèves du Québec).

Lorsque la population à l'étude est petite[6], la taille de l'échantillon peut être plus petite tout en conservant la même marge d'erreur que pour une population plus grande. Mais nous devons alors appliquer un facteur de correction.

Par exemple, nous avons une population d'élèves de 200 personnes. Nous ignorons les caractéristiques de cette population, donc l'écart-type est inconnu. Nous décidons d'accepter une marge d'erreur de + ou – 5%. Pour déterminer le nombre des répondants nécessaires, nous appliquons ensuite la formule $n = 1 / E^2$, c'est-à-dire la taille de l'échantillon (n) est égale à l'inverse de l'erreur (E) au carré. Comme nous avons déterminé que l'erreur acceptable serait de 5 %, nous avons donc :

$n = 1 / 0{,}05^2$; $n = 1 / 0{,}0025$; $n = 400$.

Nous avons donc besoin de 400 répondants. Avec ce nombre, les résultats ne s'écarteront pas de plus ou moins 5 %, 19 fois sur 20 (ou 95 % des

[6] Pour déterminer si la population est petite, on applique la règle du « 7 ». On multiplie la taille de l'échantillon calculé par 7. Si le résultat est plus petit que la taille de la population, on applique le facteur de correction. Par exemple, si mon échantillon est de 200, il faut que la taille de la population soit d'au moins 1 400.

fois). Comme ma population de 200 élèves est inférieure à la taille de mon échantillon, je dois appliquer le facteur de correction. Ce dernier suit la formule suivante :

$$n' = \frac{N \times n}{N + n}$$

Où n' = l'échantillon corrigé
N = la taille de ma population (ici : 200 élèves)
n = la taille de l'échantillon (ici : 400 répondants)

Donc, ici n' = 200 X 400 / 200 + 400 = 80 000 /600 = 133

Nous avons donc besoin de 133 répondants.

Cette démarche doit être adoptée si nous utilisons des tests statistiques qui extrapolent les résultats à l'ensemble de la population.

Par exemple, si je désire estimer l'âge moyen des répondants, disons 18 ans, et l'extrapoler à l'ensemble de la population, ce résultat peut varier de + ou – 0,9 ans (5 % de 18 ans). Ma moyenne d'âge se situe donc entre 17,1 ans et 18,9 ans.

Remarquez que nous parlons ici du nombre de répondants. Le nombre de personnes à contacter est habituellement plus grand que le nombre de répondants. Car, à moins d'une enquête très ciblée, il est possible que certaines personnes sélectionnées ne se qualifient pas pour notre recherche. Par exemple, le professeur dans une classe ne se qualifierait pas pour faire partie d'une étude sur le comportement linguistique des élèves. Cependant, le professeur fait partie intégrante de la classe. Dans un cas comme celui-là, il est relativement facile d'exclure le professeur du bassin. Ou encore si l'étude porte sur l'utilisation de

certaines expressions françaises chez les élèves dont la langue maternelle n'est pas le français, il est important d'exclure les élèves de langue maternelle française de nos répondants. Ces derniers viendraient fausser les résultats. Si la classe sélectionnée est mixte, élève de langue maternelle française et élèves ayant le français comme langue seconde, et que nous ne pouvons, pour une raison ou une autre, départager l'origine linguistique des élèves, il est alors possible que des personnes qui ne se qualifient pas dans notre population à l'étude soient malgré tout sélectionnées mais il faudra les éliminer[7]. Si le nombre de répondants a été établi à 400 personnes, il faut sélectionner plus de monde dans notre échantillon pour tenir compte de ces élèves que nous devrons exclure de l'échantillon par la suite. De plus, il est possible que certains élèves refusent de participer à l'étude. En d'autres termes, ill faut tenir compte du ***taux d'incidence*** et du ***taux de refus***.

Le taux d'incidence

Dans notre exemple, le taux d'incidence représente le pourcentage des élèves qui satisfont aux critères de l'étude sur le nombre total des élèves dans la classe. Disons que la moitié (50 %) des élèves ont une langue maternelle autre que le français et que nous désirons 400 répondants. Il faudra sélectionner 800 élèves pour obtenir nos 400 répondants puisque, sur les 800 élèves sélectionnés, 400 auront comme langue maternelle le français.

Le taux de refus

Nous savons donc qu'il faut sélectionner 800 élèves pour obtenir nos 400 répondants. Mais est-ce que tous ces élèves vont répondre? S'ils ne répondent pas tous, nous serons en déficit. Il est donc essentiel d'estimer au préalable le nombre d'élèves qui, croyons-nous, refuseront de participer à l'enquête. Cette

[7] Pour ce faire, on peut utiliser des questions-filtres (dans notre exemple, ce serait : quelle est votre langue maternelle?).

estimation est basée à la fois sur l'expérience, le lieu du déroulement de l'enquête et les particularités des questions.

Il est évident que, si l'enquête a lieu dans des écoles où les élèves sont obligés de répondre, le taux de réponse sera probablement de 100% et il n'y aura pas de problème. Cependant, si l'enquête porte, par exemple, sur des expressions religieuses et que certains élèves (disons que notre classe est multiconfessionnelle) refusent de répondre pour des raisons justement religieuses, il est possible, malgré l'obligation faite par la direction de l'école, que des élèves refusent de participer à l'enquête. Aussi, pour aboutir au nombre de répondants calculé au point de départ, il faudra tenir compte de ces refus. Dans notre exemple si 10 % des élèves qui se qualifient pour l'enquête refusent de répondre, il faudra alors sélectionner 90 élèves de plus[8].

Qu'arrive-t-il si la taille de l'échantillon est petite?

Lorsque la taille de l'échantillon est petite, la marge d'erreur est alors plus grande. Sauf qu'il est possible d'utiliser des tests spécifiques pour compenser. On trouve des petits échantillons dans le cas d'études médicales ou lorsque le coût de l'enquête est élevé. Le premier critère à retenir est l'homogénéité de la population à l'étude ainsi que de l'échantillon et le contrôle du champ expérimental. Quand des petits échantillons sont utilisés, c'est surtout la pertinence de l'utilisation de certains tests statistiques qui est importante. Les tests t, dont le plus connu est le t de Student[9], sont à cet égard des plus utiles.

[8] En effet, sur 90 élèves la moitié (45) seront de langue maternelle autre que le français. Nous aurions donc 445 élèves dans notre échantillon. 10 % de ces derniers refuseraient de participer (44,5 élèves) qu'il faut arrondir à 45. Ainsi 445 moins 45 refus nous donne 400 répondants.

[9] Le t de Student fut développé par William Gosset qui travaillait pour la brasserie Guiness. Il était responsable du contrôle de la qualité et devait travailler à partir d'échantillons de la production. Gosset ne pouvait travailler avec de gros échantillons. Gosset a donc développé un test lui permettant de comparer les différents échantillons. Gosset a publié sous le pseudonyme Student, pour étudiant, puisqu'il lui était interdit par son employeur de publier sous son nom.

Le cas particulier des micro-échantillons

On utilise des micro-échantillons lorsque la cueillette de l'information est difficile ou coûteuse ou encore lorsque la population est très petite. Nous retrouvons entre autres cette situation dans des recherches médicales, en biostatistique ou en psychologie. Il est préférable d'avoir recours à au moins deux micro-échantillons, surtout pas à un seul.

L'utilisation de micro-échantillons ne crée pas de problème en soi. Ce sont surtout les tests statistiques utilisés pour valider l'information qui sont importants. En effet, si l'échantillon est petit, l'impact de chacun des éléments de cet échantillon peut être significatif surtout si une variation importante est observée. Il faut donc respecter l'homogénéité des observations et utiliser les tests appropriés. On considère un échantillon comme «micro» lorsque le nombre est inférieur à 20 (certains auteurs disent 30).

Ainsi, à partir de micro-échantillons, il est possible d'effectuer des calculs qui viennent examiner la variabilité des réponses et déterminer si oui ou non il existe un lien ou une association statistique significative. Le choix de ces tests demeure conditionné par le type d'étude effectué et l'objectif recherché. De plus, il est possible de calculer *a posteriori* quelle aurait dû être la taille de l'échantillon au départ à l'aide d'un test de puissance.

Les différents types d'échantillons

Il existe plusieurs méthodes pour sélectionner les personnes qui serviront à répondre aux objectifs d'une recherche. Essentiellement, les échantillons se regroupent en deux grandes familles: les échantillons non probabilistes et les échantillons probabilistes.

Les échantillons non probabilistes

Un échantillon non probabiliste est un échantillon qui n'offre pas à tous les membres de la population une chance égale, ou pré-déterminée, d'être sélectionnés. La probabilité de sélection d'un membre de la population est donc inconnue. Il devient alors impossible de calculer la précision des résultats ainsi obtenus et d'utiliser les résultats pour extrapoler sur l'ensemble de la population. Cette impossibilité réside essentiellement dans le fait qu'il est possible que les répondants peuvent ne pas être représentatifs de la population.

Baser une enquête sur des données recueillies auprès d'individus qui connaissent un chercheur est un exemple d'un échantillon non probabiliste. En effet, les personnes que le chercheur connaît ne représentent pas nécessairement l'ensemble de la population. Il faudrait donc conclure d'une telle étude que les personnes qui connaissent le chercheur se comportent d'une certaine manière mais on ne pourrait pas généraliser à l'ensemble de la population.

L'utilisation d'une école, à cause de sa proximité ou parce que le directeur est un intime du chercheur, entre dans la définition d'un échantillon non probabiliste. Le choix d'une classe, parce que le fils ou la fille du chercheur y est inscrit, n'est pas nécessairement représentatif de l'ensemble des classes de l'école. Le choix d'un groupe d'amis à l'intérieur de la classe entre également dans la catégorie d'un échantillon non probabiliste.

L'échantillon volontaire est un autre type d'échantillon non probabiliste. Dans une classe, un échantillon volontaire serait formé de toutes les personnes qui se disent intéressées à participer à l'enquête. Il faut de demander si ces personnes ont les mêmes caractéristiques que celles qui décident de ne pas participer.

Il est toutefois possible et légitime d'utiliser un échantillon non probabiliste pour valider un questionnaire, notamment au niveau de la compréhension des questions ou encore pour calculer le temps d'administration ou de traitement.

Les échantillons probabilistes

Un échantillon est considéré comme probabiliste lorsque la probabilité d'être choisi est connue pour tous les membres d'une population. Il est alors possible d'effectuer des calculs afin de mesurer l'exactitude des résultats de l'enquête.

Il existe plusieurs méthodes d'échantillonnages probabilistes. Nous en examinerons les principales.

L'échantillon aléatoire simple

L'échantillonnage aléatoire simple consiste à sélectionner les répondants au hasard à partir d'une population. Dans ce cas, chaque membre de la population a une chance égale d'être sélectionné. On peut illustrer cette méthode de la façon suivante: le nom de toutes les personnes faisant parti d'une population se retrouvent à l'intérieur d'une immense cuve. Après avoir déterminé le nombre de personnes nécessaires pour l'enquête, on pige au hasard le nombre de noms de cette cuve. C'est le même principe qu'un tirage de tombola où tous les billets participants se trouvent à l'intérieur d'un baril de tirage. S'il y a 10 000 participants, vous avez une chance sur 10 000 que votre billet soit tiré si, bien entendu, chaque participant n'a qu'un seul billet. S'il y a 500 « gagnants », les 500 noms tirés représenteraient ainsi un échantillon. Nous pourrions, à partir de cet échantillon, estimer l'âge, le sexe, le revenu ou toute autre variable pertinente des participants en partant du principe que le hasard sera représentatif de la population à l'intérieur d'une certaine marge. En prenant soin

de remettre dans le baril les 500 premiers gagnants et en répétant l'exercice, nous aurions un deuxième échantillon de 500 personnes. Il est fort probable, toutefois, que l'âge moyen de ce deuxième échantillon sera différent du premier. Mais cet écart, comme nous l'avons vu précédemment, ne variera que du pourcentage déjà établi par la taille de l'échantillon : c'est-à-dire que l'écart se situera à l'intérieur de la marge d'erreur. Dans notre exemple, un échantillon de 500 personnes nous donne une marge d'erreur de + ou – 4,47 % (on peut arrondir à 4,5 %). En fait, l'ensemble de tous les échantillons possibles nous donnerait la vraie moyenne de la population.

L'échantillon aléatoire systématique

Ce type d'échantillon est simple et s'utilise lorsque nous avons affaire à une population captive ou si nous disposons d'une liste des membres qui la composent[10]. Les éléments sont choisis d'une façon systématique selon le nombre de personnes devant être sélectionnées. Par exemple, nous désirons sélectionner 100 personnes et notre liste contient 1 000 noms. Nous prenons un chiffre au hasard entre 1 et 10, puisque 1 000 divisé par 100 donne dix, ceci représente le «pas» de sondage. Disons que nous avons choisi au hasard le chiffre 5. Nous sélectionnons la cinquième personne sur la liste puis toutes les dix personnes qui suivent. Aussi, les personnes choisies seraient la 5^{e} personne de la liste, la 15^{e}, la 25^{e} et ainsi de suite. Toutes les personnes, au départ, avaient une chance égale d'être sélectionnées. En effet, nous aurions pu choisir comme point de départ la deuxième, troisième, dixième personne sur la liste.

L'échantillonnage par grappes

Lorsqu'une population se divise en plusieurs composantes semblables, ou en sous-populations possédant des caractéristiques similaires, on associe ces

[10] Dans un tel cas, nous parlerions d'une base de liste. Le bottin téléphonique, la liste des électeurs, la liste des élèves d'une école représentent des exemples de listes.

populations à des grappes. Dans le cas d'une étude sur des écoles, les écoles d'un quartier formeraient une grappe. Nous tenons pour acquis que toutes les écoles d'un même quartier possèdent les mêmes caractéristiques sociodémographiques. Alors, une école de ce quartier pourrait représenter l'ensemble des écoles de ce quartier. On répèterait la même procédure pour l'ensemble des quartiers. Il est évident que, si un quartier n'est pas homogène, il est hasardeux d'utiliser cette méthode.

La grappe peut être complète ou partielle. C'est-à-dire qu'on peut décider de prendre tous les élèves de l'école ou seulement une partie d'entre eux. Dans ce dernier cas de figure, il faudra évidemment les choisir de façon aléatoire.

Rappelons que l'échantillon par grappes ne peut donc être utilisé que quand la population est homogène et qu'elle peut être sous-divisée. L'échantillon par grappes permet une économie d'échelle, notamment dans les déplacements; c'est une procédure plus particulièrement indiquée dans le cas d'études s'étendant sur un vaste territoire.

L'échantillonnage stratifié

Il y a une relation directe entre la précision des résultats d'une enquête et l'homogénéité de la population à l'étude. Il arrive parfois que des sous-population soient plus homogènes que la population elle-même ou que des éléments d'une population ne représentent qu'eux-mêmes. Dans ce cas, ces sous-populations forment des strates.

Par exemple, dans une étude portant sur l'utilisation de certaines expressions, il est possible que des études préliminaires aient démontré que le niveau d'instruction, l'âge et le revenu influencent le comportement. La sélection des répondants se fera en fonction de ces strates. Les strates peuvent être simples, avec une seule variable, ou complexes, avec plusieurs variables.

Il faut donc retenir que les strates sont formées à partir de certaines caractéristiques de la population et qu'une connaissance préalable de cette dernière est nécessaire. Prenons l'exemple d'une étude qui porte sur les étudiants d'une ville. Supposons qu'il y a quatre universités dans la ville en question. Les quatre universités se partagent deux langues d'enseignement. Il y a deux universités de la langue A, que nous appelons universités A1 et A2, et deux universités de la langue B, que nous appelons B1 et B2. La clientèle des quatre universités est également différente. En effet, les étudiants de l'université A1 sont majoritairement inscrits dans des programmes scientifiques alors que ceux de l'université A2 sont majoritairement inscrits dans des programmes de sciences humaines. Supposons que nous ayons le même phénomène dans les universités B1 et B2. Nous nous retrouvons donc avec une population générale d'étudiants. Cependant, si nous effectuons une recherche sur les étudiants de la ville et que nous ne choisissons que les étudiants de l'université A1, pouvons-nous légitimement utiliser les résultats et en déduire pour l'ensemble des étudiants de la ville? Non, puisque les étudiants de l'université A1 possèdent des caractéristiques qui leur sont propres. Dans notre exemple, ces caractéristiques sont à la fois linguistiques et d'orientation académique. Mais si nous faisions une étude sur l'utilisation des termes techniques, nous pourrions alors affirmer que chacune des universités forme une population qui est homogène.

L'échantillonnage stratifié nécessite une connaissance statistique préalable pouvant provenir des recensements. Ces derniers permettent l'identification des strates et leur pondération.

Quotas

Les éléments d'un échantillon par quota ne sont pas sélectionnés au hasard mais en fonction d'un nombre prédéterminé d'éléments à l'intérieur de certaines catégories. Cette méthode est utilisée par certains instituts de

sondage. Il s'agit de déterminer au départ le nombre de personnes qui offrent certaines caractéristiques. Par exemple, 10 personnes de sexe masculin entre 18 et 24 ans, 15 femmes entre 25 et 34 ans etc. Le sondage par quota n'est pas considéré comme probabiliste. Cependant, lorsqu'il est intégré dans le cadre d'un échantillonnage stratifié, il est possible de travailler à partir de ses résultats de manière statistiquement fiable.

Cette méthode est pratique pour des recherches en marketing; son application dans d'autres champs d'études est plus problématique.

Les échantillons pairés

La méthode de sélection est une étape importante. Cependant, il existe d'autres moyens, notamment lorsqu'il est question de petits échantillons, de maximiser l'utilisation du potentiel des répondants. Le principal est l'échantillon pairé.

On entend par échantillon pairé un échantillon dont les sujets répondent à deux reprises aux questions. On examine alors la différence entre les résultats. Par exemple, on demande à 20 personnes leur appréciation qualitative d'une expression. Puis, après avoir fourni à ces mêmes personnes une explication de la signification de cette expression, nous leur redemandons une nouvelle appréciation qualitative. Chaque répondant aura alors fourni deux réponses. Nous nous attendons à ce que les réponses soit liées, c'est-à-dire que le niveau d'appréciation de l'expression évolue dans un sens ou dans un autre. Les échantillons pairés sont notamment utilisés pour des recherche très spécifiques et peuvent être fort utiles avec de petits échantillons.

Conclusion

Que ce soit pour des petits échantillons ou pour des échantillons de plus grande taille, il est important de respecter les étapes de la planification méthodologique. Ces étapes sont :

1. Définir la population à l'étude. Il est primordial de bien définir cette population. La définition peut se faire au niveau régional, au niveau de l'âge, du sexe, etc.

2. Établir les objectifs statistiques de l'enquête. Voulons-nous simplement décrire une situation, comprendre les causes de la situation ou extrapoler sur l'ensemble d'une population à partir d'un échantillon?

3. Déterminer le degré de précision recherchée.

4. Établir les contraintes de l'enquête. Taux d'incidence, taux de refus.

5. Établir la méthode d'échantillonnage la plus appropriée en fonction des contraintes :
   a) de temps;
   b) de budget;

6. Coûts

7. Délai imparti

8. Informations disponibles (recensements, enquêtes antérieures, etc.).

Comme nous l'avons mentionné, la taille de l'échantillon peut avoir une influence sur l'erreur que l'on commet lorsque l'on tente de projeter sur l'ensemble de la population les résultats d'une enquête. Il existe cependant un moyen de contourner ce problème lorsque le chercheur est confronté aux contraintes des échantillons, petits ou grands. Il s'agit d'utiliser des tests statistiques appropriés. Le test Mann-Whitney, le t de Student ainsi que les autres tests t sont particulièrement utiles dans ces cas.

L'utilisation de petits échantillons est tout à fait possible et pertinente mais à condition de respecter certains principes. La population doit être homogène et bien définie. L'étude ne doit pas porter sur des différences internes à la population mais plutôt sur des éléments précis du sujet à l'étude.

Deuxième partie : les tests statistiques

La clé: les tests à utiliser

Au-delà de l'échantillonnage, c'est l'utilisation des tests appropriés qui est la clé dans l'utilisation des petits échantillons. Des informations fort pertinentes peuvent être obtenues à partir d'un petit nombre d'observations à condition d'utiliser les tests adéquats.

De plus, on associera pour chacune des catégories de variables une série de tests qui leur sont propres.

Variables (ou échelles) et valeurs

Du point de vue du traitement statistique, les variables peuvent prendre quatre formes. Nous adoptons la définition suivante de variable : «un concept transformé en outil de recherche et présentant des modalités devant servir à classifier les observations s'y rapportant»[11].

Toutefois, il est essentiel de préciser quels sont les types de variables.

Mesurer et quantifier : les différents types de variables[12]

Il existe quatre types de variables: nominale, ordinale, intervalle et métrique.

[11] Alain Gilles, *Éléments de méthodologie et d'analyse statistique pour les sciences sociales*, Montréal, McGraw-Hill, 1994, p. 30.
[12] On trouve dans certains textes l'appellation échelle.

Les variables (ou échelles) nominales

Une variable nominale est de nature qualitative. Une variable est dite nominale lorsque le chiffre associé à chacune des possibilités de réponses à une question est arbitraire. Par exemple, les quartiers d'une ville sont des variables nominales. L'ordre de traitement des quartiers est subjectif. Supposons qu'il y a quatre quartiers dans une ville et que ces quartiers se nomment La Rose, La Fleur, La Tulipe et La Jonquille. Pour des fins de traitements statistiques on attribue une valeur à chacun de ces quartiers, disons La Rose = 1, La Fleur = 2, La Tulipe = 3 et La Jonquille =4. Nous aurions pu attribuer les valeurs suivantes: La Rose = 2, La Fleur = 3, La Tulipe = 4 et La Jonquille =1. En fait, le chiffre associé aux différents quartiers importe peu. C'est ce que l'on appelle une variable nominale. Il suffit de retenir qu'il s'agit en fait d'un NOM. Le sexe (masculin, féminin), les pays, les écoles, le statut matrimonial, la religion sont tous des exemples de variables nominales. En résumé, une variable nominale est une variable dont on ne peut que nommer les catégories. D'un point de vue statistique, il n'est pas possible d'effectuer d'opérations mathématiques avec des données provenant d'une échelle nominale. Nous ne pouvons donc additionner, soustraire, multiplier ou diviser les réponses.

Les variables (ou échelles) ordinales

Les variables ordinales ont les mêmes caractéristiques que les variables nominales, soit de nommer les catégories. Ce sont donc également des variables de nature qualitative. Toutefois, les variables ordinales classent les personnes, les objets ou les événements le long d'un continuum. Les catégories des variables ordinales sont hiérarchisées, elles ont un ORDRE. Un diplôme doctorat est supérieur à un diplôme de premier cycle qui lui-même est supérieur à un diplôme d'études secondaires, etc. Il existe donc un rapport d'ordre, une hiérarchie entre les catégories de la variable. Les échelles ordinales possèdent les deux caractéristiques suivantes:

1) elles sont *antisymétriques*. C'est-à-dire que la catégorie à laquelle on a assigné une valeur, par exemple diplôme universitaire = 4, ne peut pas être plus petite que les catégories qui la précèdent.
2) Elles sont *transitives*. Si un diplôme de fin d'études secondaires (2) est moins élevé qu'un diplôme de premier cycle universitaire (3) qui lui-même est moins élevé qu'un doctorat (4), alors le diplôme de fin d'études secondaires (2) est moins élevé qu'un doctorat (4).[13]

Comme pour les variables nominales, il est possible de compter le nombre de cas pour chacune des valeurs de la variable mais les opérations mathématiques de bases (addition, soustraction, multiplication et division) ne sont pas appropriées.

Les variables d'intervalles

Les variables d'intervalles sont des variables quantitatives. Il est possible ici d'évaluer des différences, ou la distance, entre les points de l'échelle. On attribue toutefois à ce type d'échelle des qualités limitées du point de vue mathématique. Il est en effet possible d'additionner et de soustraire les résultats, mais la multiplication et la division ne sont pas appropriées. La principale caractéristique d'une variable d'intervalle est que le zéro que prend la catégorie de la variable est relatif et parfois arbitraire.

Le quotient intellectuel (QI) est un exemple d'une variable d'intervalle. Une personne qui a un QI de zéro n'est pas totalement dénuée d'intelligence. Il n'y a pas absence totale et complète d'intelligence. Cela est un zéro relatif. Une température de zéro degré centigrade ne signifie pas l'absence de chaleur. De

[13] Les chiffres entre parathèses sont les valeurs que nous aurions pu assigner aux différentes catégories de la variable.

fait, le zéro degré centigrade n'est pas équivalent au zéro degré Fahrenheit. Ces deux zéros sont arbitraires. De plus, nous pouvons affirmer que la différence dans le nombre de degrés sur toute l'échelle est la même. Il y a effectivement la même différence de 5 degrés entre 20 et 25 degrés et entre 10 et 15 degrés. Si on rajoute 5 degrés à 10 degrés, nous obtenons une température de 15 degrés. Toutefois, nous ne pouvons dire qu'une température de 20 degrés est deux fois plus chaude qu'une température de 10 degrés. Au même titre que nous ne pouvons affirmer qu'une personne qui obtient 150 à un test de QI est deux fois plus intelligente qu'une personne qui obtient 75. En conclusion, la distance entre deux points sur l'échelle n'est pas une mesure de proportion.

La variable (ou échelle) de rapport[14]

La principale caractéristique de l'échelle de rapport est qu'elle possède un vrai zéro. Ce zéro représente l'absence de quelque chose. Si une personne n'a pas d'argent, il y a absence totale d'argent. La longueur, le volume, le temps, le revenu, l'âge sont des exemples d'échelles de rapport. Comme le nom l'indique, il est possible d'effectuer des rapports entre les valeurs de la variable. Quelqu'un qui a 20 ans est deux fois plus âgé que quelqu'un qui a dix ans. L'ensemble des opérations mathématiques, l'addition, la soustraction, la multiplication et la division sont alors possibles.

L'ordre de présentation des échelles est importante. On remarquera que les caractéristiques mathématiques, et qualitatives, des échelles sont cumulatives. Les propriétés des différentes échelles sont en effet cumulatives. Une variable de rapport, le revenu, possède en effet toutes les caractéristiques des échelles qui la précèdent. La personne a-t-elle un revenu? Réponse: Oui/ Non (Échelle nominale). Ce revenu est-il Très Important, Assez Important, Peu Important ou Pas du Tout Important? (Échelle ordinale). Ce revenu est-il en haut

[14] On parle également de variable ou d'échelle métrique ou proportionnelle.

ou en bas du seuil de la pauvreté (zéro relatif)? (Échelle d'intervalle). Quel est le revenu réel de l'individu? (Échelle de rapport). Il est donc impossible pour une variable nominale d'avoir les caractéristiques d'une échelle qui lui est supérieure.

Nous reprenons les propos d'Howell[15] qui souligne que «*c'est la variable sous-jacente que nous mesurons …, et non les nombres eux-mêmes, et que c'est elle qui importe dans la définition de l'échelle»*. Utilisons l'exemple tiré du livre de Howell. «Prenons le cas d'un questionnaire sur l'angoisse distribué à un groupe de lycéens. Sans réfléchir on pourrait prétendre qu'il s'agit d'une échelle de rapport sur l'angoisse. On affirmerait qu'une personne obtenant un score de 0 ne présente aucune angoisse et qu'un score de 80 reflète une angoisse deux fois plus grande qu'un score de 40. Même si la plupart des gens trouveraient cette opinion ridicule, il n'en reste pas moins que certains questionnaires permettraient ce type de raisonnement. On pourrait également prétendre qu'il s'agit d'une échelle d'intervalles et que, même si le point zéro est quelque peu arbitraire (l'étudiant obtenant 0 présente tout de même une certaine angoisse, que le questionnaire n'a pu déceler), des différences équivalentes en termes de scores représentent des différences équivalentes en termes d'angoisse. Il serait déjà plus raisonnable d'affirmer que ces scores constituent une échelle ordinale: un score de 95 reflète une plus grande angoisse qu'un score de 85, qui reflète à son tour un plus grande angoisse qu'un score de 75, mais des différences équivalentes en termes de scores ne reflètent pas des différences en termes d'angoisse»[16].

Il faut donc être conscient que c'est l'utilisation des chiffres par le biais de l'analyse qui est importante. Il faut donc répondre aux questions suivantes: pourquoi nous recueillons l'information, comment nous la recueillons et quelle sera la technique d'analyse qui sera utilisée. Tout cela doit être déterminé avant le début de l'enquête et doit être partie intégrante du plan de recherche.

[15] David C. Howell, *Méthodes statistiques en sciences humaines*, Bruxelles, De Boeck, 1998, p. 8.
[16] Howell, *op. cit*., p. 8.

Comme on peut le remarquer, la planification est essentielle à la bonne utilisation des différents tests disponibles selon les circonstances. Car à chaque type d'échelle nous associons un type de test statistique. Les pages qui suivent nous donnent un aperçu sommaire des principaux tests en fonction du type d'échelle.

Tests et variables qualitatives

Nous avons vu qu'il existe deux grandes familles d'échelles: les échelles qualitatives (variables nominales et ordinales) et les échelles quantitatives (variables intervalles et de rapports). Nous examinerons dans un premier temps l'étude des variables qualitatives.

Il est important de souligner que les valeurs d'une variable peuvent se présenter de plusieurs façons. On dira d'une variable qu'elle est dichotomique si cette variable ne peut prendre que deux valeurs. Le sexe du répondant est une variable dichotomique puisqu'il y a deux réponses possible: masculin et féminin. Si les deux variables étudiées sont dichotomiques (par exemple : le sexe; le français comme langue maternelle ou langue seconde), il est possible d'utiliser certains tests ou calculs dont les principaux sont: le calcul de la différence des pourcentages, le Φ (phi) et le Q de Yule.

La différence des pourcentages permet de comparer les observations. Prenons l'exemple suivant:

Tableau 1

Les chiffres représentent le nombre d'observations

	Français langue maternelle	Français langue seconde	Total
Sexe Masculin	25	30	55
Sexe Féminin	25	20	45
Total	50	50	100

On remarque qu'il y a le même nombre de personnes dont la langue maternelle et la langue seconde sont le français. Il existe une petite différence au niveau du sexe. Si l'échantillonnage a été fait selon les normes d'un échantillonnage aléatoire, la différence constatée dans les pourcentages devrait refléter la réalité de l'ensemble de la population.

Le tableau 2 présente les mêmes données, mais sous forme de pourcentage horizontal.

Tableau 2

Les chiffres représentent le pourcentage horizontal par rapport au tableau 1

	Français langue maternelle	Français langue seconde	Total
Sexe Masculin	45,5 %	54,5 %	100 %
Sexe Féminin	55,5 %	44,5 %	100 %

Lorsque l'on transforme les résultats en pourcentage, on remarque qu'il existe un écart de 10 % entre les hommes et les femmes en ce qui concerne la langue et ce même si le nombre de femmes et d'hommes ayant le français comme langue première est le même (25 hommes et 25 femmes). On doit lire ce tableau comme suit: 45,5 % des hommes de notre échantillon ont le français comme langue première comparativement à 55,5 % pour les femmes. On retrouve donc un écart de 10 % (45,5 % - 55,5 %).

Tableau 3

Les chiffres représentent le pourcentage vertical par rapport au tableau 1

	Français langue maternelle	Français langue seconde	Total
Sexe Masculin	50 %	60 %	55 %
Sexe Féminin	50 %	40 %	45 %
Total	100 %	100 %	100 %

Le lecture du tableau 3 se fait comme suit: 50 % des personnes qui ont le français comme langue première sont des hommes. Il y a une différence importante entre les tableaux 2 et 3. Dans le tableau 2, le pourcentage est calculé sur l'ensemble des 100 personnes interrogées alors que dans le tableau trois le calcul se fait sur la base du français langue première ou langue seconde seulement, donc sur 50 personnes pour chacun des cas.

La différence des pourcentages mesure l'écart entre les pourcentages. Ainsi, dans le tableau 3, nous avons un écart de -10% pour les hommes (50 % - 60 %) et de +10 % pour les femmes (50 % - 40 %).

Nous venons de présenter les mêmes données de trois façons différentes. On remarque que l'interprétation varie selon la façon de présenter les données et de lire les tableaux.

Nous avons mentionné préalablement que la statistique utilise un langage qui lui est propre. Dans l'exemple précédent, nous pouvons identifier de la façon suivante chacune des cellules qui composent le tableau:

Tableau 4

Illustration algébrique des cellules

	Français langue maternelle	Français langue seconde
Sexe Masculin	n11	n12
Sexe Féminin	n21	n22

La nomenclature suit la logique suivante : le numéro de la rangée suivi du numéro de la colonne. La première rangée (masculin) et la première colonne (français langue maternelle) forment la cellule 11. La cellule 21 est donc formée de la deuxième rangée et de la première colonne, etc. (Nous avons retenu la notation «n» puisque c'est elle qui sera utilisée dans les formules qui suivront.)

Le nombre des personnes qui font partie de notre échantillon influence les pourcentages. Ainsi, dans l'exemple que nous avons pris, nous avons 50 hommes et 50 femmes. Chaque homme représente 2 % de la population masculine et chaque femme représente 2 % de la population féminine. Mais si nous avions observé 24 femmes et 26 hommes, nous aurions obtenus 48% de femmes et 52 % d'hommes. Par conséquent, plus l'échantillon est petit, plus l'impact de chaque observation particulière est grand. Si l'échantillon est petit, le moindre soubresaut dans l'observation va avoir un impact énorme.Il faut donc être prudent dans l'extrapolation des résultats à partir d'un nombre restreint d'observations. Cela est un problème inhérent aux petits échantillons.

Les tests

Le Φ (phi) est une mesure d'association utilisée pour résumer la relation entre deux variables dichotomiques. Ce test mesure la concentration sur la diagonale. Il est entre autres utile dans le cas où une des deux variables (ou les deux) sont ordinales. Utilisons l'exemple suivant:

Tableau 5[17]

Compétence linguistique et scolarité

Compétence linguistique	Scolarité inférieure	Scolarité supérieure	Total
Unilingue français	93	83	176
Bilingue	91	116	207
Total	184	199	383

Nous posons l'hypothèse qu'il y a plus de personnes bilingues lorsque le niveau de scolarité est élevé.

Calculons maintenant les pourcentages de la même façon que dans les tableaux précédents. Nous allons toutefois intégrer tous les chiffres à l'intérieur d'un seul tableau. C'est la façon dont les données sont présentées dans les tableaux produits par des logiciels de statistiques comme SPSS ou SAS.

[17] Exemple adapté de Alain Gilles, *op. cit*., p. 262. Les chiffres sont fictifs.

Tableau 6

Compétence linguistique et scolarité

(Condensé)

Compétence linguistique	Scolarité inférieure[18]	Scolarité supérieure	Total
Unilingue français	93	83	176
(% - H)	52,9 %	47,1 %	100 %
(% - V)	**50,5 %**	**41,7 %**	
(% - T)	24,3 %	21,7 %	
Bilingue	91	116	207
(% - H)	44 %	56 %	100 %
(% - V)	**49,5 %**	**58,3 %**	
(% - T)	23,9 %	30,1 %	
Total	184	199	383
(% - H et % T)	48,1 %	51,9 %	
(% - V)	**100 %**	**100 %**	

Légende:

(% - H) signifie le pourcentage horizontal (le pourcentage de la ligne). On doit donc lire : 52,9 % des unilingues français ont une scolarité inférieure pour 44 % chez les bilingues.

(% - V) signifie le pourcentage vertical (le pourcentage de la colonne). Les unilingues français forment 50,5 % de toutes les personnes qui ont une scolarité inférieure contre 49,5 % chez les bilingues. Ces chiffres sont en caractères gras.

(% - T) représente le pourcentage sur le total de toutes les personnes qui ont été interrogées. Les unilingues français qui ont une scolarité inférieure représentent 24,3 % de notre échantillon.

[18] Bien entendu, la classification inférieure et supérieure devrait faire l'objet d'une justification dans le cadre théorique. Ce qu'il faut retenir ici, c'est l'ordre de la classification.

Même si nous observons des différences de pourcentage, il est légitime de se demander si ces différences sont significatives ou non mais il faut se méfier des apparences : avant de conclure que les différences sont significatives, il faut les valider. Dans notre exemple, il faut valider l'hypothèse voulant que le niveau d'éducation et la compétence linguistique soient liés. Dans un tel cas, nous nous attendons à ce que les résultats suivent la diagonale unilingue moins de scolarité (cellule n11) et bilingue plus de scolarité (cellule n22). Si on regarde seulement ces deux cellules, on peut avoir l'impression que l'hypothèse est confirmée parce que ce sont ces cellules qui se trouvent à compter le plus d'observations (n11 = 93 observations; n22 = 116 observations). Or, le test va nous montrer que tel n'est pas le cas.

Le Φ nous permet donc d'examiner le niveau de concentration des réponses sur les deux diagonales du tableau.

La formule du Φ est la suivante:

$n_{11}n_{22} - n_{12}n_{21} / \sqrt{n_{1.}} \times \sqrt{n_{2.}} \times \sqrt{n_{.1}} \times \sqrt{n_{.2}}$

> *Remarque sur la notation* : quand le point précède l'indice, cela signifie que le chiffre mis en indice représente le total de la colonne. Ainsi, $n_{.1}$ représente le total de la première colonne. Quand le point suit le chiffre mis en indice, cela signifie que le chiffre représente le total de la rangée. Ainsi, $n_{1.}$ représente le total de la première rangée.

Dans la formule :

n_{11} = 93, soit le nombre d'observations de la cellule 11 Unilingue et scolarité inférieure.

n_{12} = 83, soit le nombre d'observations de la cellule 12. Unilingue et scolarité supérieure.

n_{21} = 91, soit le nombre d'observations de la cellule 21. Bilingue et scolarité inférieure.

n_{22} = 116, soit le nombre d'observations de la cellule 22. Bilingue et scolarité supérieure.

$n_{1.}$ = 176, soit le nombre total d'observations ayant la caractéristique unilingue.

$n_{2.}$ = 207, soit le nombre total d'observations ayant la caractéristique bilingue.

$n_{.1}$ = 184, soit le nombre total d'observations ayant la caractéristique scolarité inférieure.

$n_{.2}$ = 199, soit le nombre total d'observations ayant la caractéristique scolarité supérieure.

Nous avons donc: Φ = (93) (116) - (83) (91) / $\sqrt{}$ (176) (207) (184) (199)

= 10 788 - 7553 / (13,27) (14,39) (13,56) (14,11)

= 3 235 / 36 535,78

= 0,088

Φ varie entre -1 et +1 et ne prendra la valeur maximum (+ ou - 1) que lorsque toutes les données se trouvent sur la diagonale. On obtient un Φ de +1 quand les données se trouvent dans le sens positif de la diagonale, c'est-à-dire quand elles vont de la plus petite à la plus grande. Dans le cas présent, toutes les données devraient être soit dans la cellule 11 (unilingues scolarité inférieure) soit dans la cellule 22 (bilingues et scolarité supérieure). Nous obtiendrions -1 si toutes les données se trouvaient dans la cellule 12 (unilingues et scolarité supérieure) ou 21 (bilingues et scolarité inférieure). Remarquez que le caractère positif ou négatif dépend de la classification, c'est-à-dire de l'ordre dans lequel nous avons placé nos variables. Le résultat ne dicte pas ce qui se passe mais rend compte du comportement en fonction de la classification qui a été effectuée *a priori*. Un résultat de 0 signifierait qu'il y a une relation négligeable entre les deux variables. Dans notre exemple, le résultat de +0,088 indique une relation positive mais faible entre la compétence linguistique et le niveau de scolarité.

En observant le tableau, et en ne se fiant que sur les pourcentages, nous aurions tendance à croire que les unilingues ont une scolarité moins élevée que

les bilingues. En effet, on observe 10 personnes de plus chez les moins scolarisés que chez les scolarisés alors qu'il y a 25 personnes bilingues de plus chez les plus scolarisés. Quant on examine les pourcentages, on remarque que l'écart est de 8,8 % dans les deux cas.

Le Φ est un test de type PRE (Proportion de réduction de l'erreur). Il permet donc de réduire l'erreur de prédiction que l'on peut commettre sur une autre variable lorsque la valeur d'une première variable est connue. La proportion de l'erreur que l'on pourrait commettre est obtenue en prenant le carré de Φ. Dans notre exemple, le Φ^2 est égal à $0{,}088^2$, soit 0,006 ou 0,6%. Il y a donc une variation de moins de 1% entre les deux variables. Cela veut dire que, dans notre exemple, on ne peut pas inférer le niveau de scolarité à partir de la compétence linguistique. Si l'on voulait prédire le niveau de scolarité à partir de la compétence linguistique, ou vice versa, le test montre que je ne réduirais mon erreur de prédiction que de moins de 1 %.

Le Q de Yule

Le Q de Yule examine la congruence dans la combinaison des catégories. Ce calcul examine également la diagonale. Le Q de Yule se calcule comme suit:

$$Q = n_{11}\, n_{22} - n_{12}\, n_{21} / n_{11}\, n_{22} + n_{12}\, n_{21}$$

Dans notre exemple, nous avons;

Q = (93) (116) - (83) (91) / (93) (116) + (83) (91)
= 10 788 - 7553 / 10788 + 7 553
= 3 235 / 18 341
= 0,18

Le résultat varie de -1 à +1. Si les variables sont ordinales, le signe du Q de Yule indique une relation positive ou négative. Plus on s'approche de 1, plus on peut dire qu'il y a une relation de congruence. Un résultat de 0,18 nous indique une relation positive mais faible.[19]

La caractéristique K de Yule

Yule a également développé une mesure de la richesse du vocabulaire basé sur la prémisse que la fréquence d'un mot se caractérise par une distribution de Poisson[20].

La formule est la suivante:

$K = 10^4(\sum i^2V_i - N) / N^2$,
$\sum$ signifie «la somme de»
i =1,2,… et représente les observations
N = le nombre de mots dans le texte
V_1 = Le nombre de mots utilisés une seule fois dans le texte, V_2 le nombre de mots utilisés deux fois dans le texte , …

La statistique du khi carré (χ^2)

Le χ^2 est probablement la statistique la plus connue pour les variables qualitatives. Le résultat du χ^2 s'utilise avec la table de référence du même nom. Le χ^2 examine la relation entre les fréquences observées (Fo) et les fréquences théoriques (Ft)[21]. Les fréquences théoriques s'obtiennent en utilisant le grand total des observations ainsi que les totaux des colonnes et des rangées d'un

[19] La question d'une relation forte ou faible dépend de plusieurs facteurs. Cette interprétation est subjective.
[20] La distribution de Poisson est utilisée dans le cas d'évènements rares et à l'intérieur d'une période déterminée.
[21] On retrouve plusieurs appellations pour les fréquences théoriques dont fréquences attendues et fréquences espérées.

tableau. Le tableau 7 vise essentiellement à comparer les résultats observés en rapport avec les résultats théoriques s'il y avait une indépendance entre les deux variables à l'étude. Quand les fréquences observées sont près des fréquences théoriques, nous dirons que les deux variables sont indépendantes. Le χ^2 est sensible aux nombres de colonnes et de rangées d'un tableau. Le nombre de catégories d'une variable aura donc une influence sur le résultat. Il est donc important lors de la préparation du questionnaire d'avoir cette contrainte à l'esprit.

Reprenons l'exemple de la compétence linguistique pour illustrer l'utilisation du χ^2. Les chiffres entre parenthèses représentent les fréquences théoriques (dont on verra plus loin la façon de les calculer). La démonstration du calcul suit le tableau.

Tableau 7

Compétence linguistique et scolarité

Compétence linguistique	Scolarité inférieure	Scolarité supérieure	Total
Unilingue français	93	83	176
	(84,55)	(91,45)	
Bilingue	91	116	207
	(99,45)	(107,55)	
Total	184	199	383

Pour les colonnes et les rangées, la somme des fréquences théoriques est égale à la somme des fréquences observées. En fait, les fréquences théoriques sont basées sur une distribution parfaite, en fonction du nombre de cas pour chacune des colonnes et des rangées, que nous aurions dû observer si les deux variables étaient complètement indépendantes l'une de l'autre. Le χ^2 va nous permettre de valider si les différences entre les fréquences observées et les

fréquences théoriques entrent dans le domaine du possible ou sont le fruit du hasard.

Le calcul de la fréquence théorique est simple. Pour chaque cellule, il s'agit de multiplier le total de la rangée par celui de la colonne et de diviser ce produit par le nombre total d'observations.

Ainsi, pour la cellule 11 nous avons observé (Fo) 93 cas. Il y a un total de 176 cas dans la rangée unilingue et un total de 184 qui ont une éducation inférieure. Nous multiplions donc 176 par 184 avec comme résultat 32 384. On divise ce produit par le nombre total d'observations qui, dans ce cas-ci, est de 383. Cela nous donne un résultat de 84,55. Aussi, nous aurions dû observer 84,55 cas dans la cellule 11 s'il y avait indépendance entre les deux variables.

Calcul des fréquences théoriques

Cellule 11 176 X 184 / 383 = 84,55
Cellule 12 176 X 199 / 383 = 91,45
Cellule 21 207 X 184 / 383 = 99,45
Cellule 22 207 X 199 / 383 = 107,55

Le test du χ^2 vise à examiner la différence entre les deux fréquences et à établir si cette différence est significative ou non. Si la différence est significative, nous dirons alors que les deux variables s'influencent (ce qui signifie que, dans notre exemple, il y aurait une association entre la compétence linguistique et le niveau d'éducation).

La formule du χ^2 est la suivante:

$\chi^2 = \sum (Fo - Fe)^2 / Fe$
Fo = La fréquence observée
Fe = La fréquence théorique

La formule se lit comme suit: La somme du carré des différences entre les fréquences observées et les fréquences théoriques divisée par la fréquence théorique.

Dans notre exemple nous aurions donc:

Cellule 11: $(93 – 84,55)^2 / 84.55 = (8,45)^2 / 84,55 = 71,4025 / 84,55 = ,84450$
Cellule 12: $(83 – 91,45)^2 / 91,45 = (8,45)^2 / 91,45 = 71,4025 / 91,45 = ,78078$
Cellule 21: $(91 – 99,45)^2 / 99,45 = (8,45)^2 / 99,45 = 71,4025 / 99,45 = ,71797$
Cellule 22: $(116 – 107,55)^2 / 107,55 = (8,45)^2 / 107,55 = 71,4025 / 107,55 = ,66390$

La somme de chacune des contributions au χ^2 est:

(,84450) + (,78078) + (,71797) + (,66390) = 3,00715

qui représente la valeur du χ^2. Ce chiffre n'a de sens que s'il est examiné en rapport avec la table des valeurs critiques du χ^2 (que l'on trouve en annexe de tous les manuels de statistiques). Si le χ^2 calculé dépasse le seuil critique de la table, nous pouvons croire qu'il y a une association entre les deux variables puisque les différences entre les fréquences observées et les fréquences théoriques sont trop grandes pour être le simple fruit du hasard. Mais avant d'utiliser la table, il faut d'abord décider du seuil de signification que l'on accepte et du degré de liberté.

Le seuil de signification représente le niveau «de confiance» où le résultat calculé peut se produire. Ainsi, si on accepte un seuil de 0,05, ou de 95 %, il est

possible de dire : «je suis confiant à 95 % que ce résultat est représentatif de l'association qui prévaut entre ces deux variables».

Comme nous l'avons mentionné précédemment, le nombre de cellules influence la répartition des fréquences théoriques. Plus il y aura de cellules, plus il y aura de possibilités de réponses et plus l'écart entre les fréquences observées et les fréquences théoriques peut être grand. Il y donc possibilité de plus de différences, même légères, qui influenceront le calcul du χ^2. On peut donc s'attendre à ce que ce dernier soit plus grand. Pour s'adapter à la taille du tableau (par exemple, un tableau de 16 cellules ou un tableau de 32 cellules), on utilise le principe du degré de liberté (dl). Le degré de liberté s'obtient en multipliant le nombre de colonnes moins 1 par le nombre de rangées moins un : (c-1) (r-1). Dans notre exemple, nous avons deux colonnes, éducation inférieure et supérieure, et deux rangées, unilingue et bilingue. Le degré de liberté sera donc de (2-1) (2-1) = 1 X 1 = 1. Nous avons ici un degré de liberté. Les tables de χ^2 sont établies en fonction de divers degrés de liberté.

On constate donc que le choix de la classification, plus précisément le nombre des catégories, aura une influence sur le test.

Dans notre exemple, le χ^2 calculé est de 3,00715. La table des valeurs critiques du χ^2 pour un seuil de ,05 donne 3,84. Notre χ^2 est inférieur au seuil critique : nous invalidons l'hypothèse qu'il y a association entre la scolarité et la compétence linguistique. Les deux variables sont indépendantes l'une par rapport à l'autre.

Le kappa (κ)

Le κ est une mesure d'accord qui s'utilise pour mesurer certaines évaluations faites par deux groupes. On utilise le κ avec des variables ordinales. Supposons que l'on demande à deux évaluateurs d'interroger 300 élèves sur la

compréhension d'une expression. Dans notre exemple, les deux «juges» évaluent les mêmes personnes en même temps ou écoutent les mêmes transcriptions. Les évaluateurs classent les résultats comme: Pas de problème de compréhension (1); Légers problèmes (2) et Beaucoup de problèmes (3)[22]. Nous avons ici deux évaluateurs. Il est légitime de se demander si les évaluateurs ont la même perception ou si, au contraire, une partie de l'évaluation est due au hasard. Le κ permet d'évaluer le niveau d'accord de jugement réel et les accords dus au hasard. Prenons l'exemple suivant:

Tableau 8[23]

Perception de deux évaluateurs

		Évaluateur I			
Évaluateur II		Pas de problème	Légers problèmes	Beaucoup de problèmes	**Total**
	Pas de problème	150	20	30.	**200**
	Légers problèmes	10	30	20	**60**
	Beaucoup de problèmes	0	10	30	**40**
	Total	**160**	**60**	**80**	**300**

Lecture du tableau : les résultats de l'évaluateur I apparaissent dans les colonnes, ceux de l'évaluateur Ii dans les rangées. Si les deux évaluateurs étaient toujours d'accord, il n'y aurait de données que sur la diagonale. Dans la cellule 21, l'évaluateur I a trouvé que 10 n'avaient pas de problème et, pour sa

[22] Le chiffre entre parenthèses est celui qui est assigné pour le traitement informatique.
[23] L'exemple est inspiré du livre de David C. Howell, *op.cit*, p. 184.

part, l'évaluateur Ii considéraient que ces 10 avaient plutôt de légers problèmes. Les deux évaluateurs n'étaient donc pas d'accord sur ces 10 cas.

On peut s'attendre à ce que si l'évaluation des deux juges est similaire elle se retrouve sur la diagonale. On entend par diagonale, ici, les évaluations où les deux évaluateurs ont la même opinion (Pas de problème / Pas de problème; Légers problèmes / Légers problèmes et Beaucoup de problèmes / Beaucoup de problèmes). Il y a en fait 210 cas sur 300 où l'évaluation est la même. Nous serions tentés de dire que les évaluateurs sont d'accord 70 % des fois et que, dans 30 % des cas, ils n'avaient pas la même opinion. Dans le cas des 70 %, il est légitime de se poser la question suivante: est-il possible que les évaluateurs aient été d'accord «par accident»? C'est ce que le κ nous permet de décider.

Le κ applique le même principe que le χ^2. On calcule la fréquence théorique mais uniquement pour les cellules qui se trouvent sur la diagonale. Cela nous donne les résultats suivants:

Cellule 11 200 X 160 / 300 = 106,70
Cellule 22 60 X 60 / 300 = 12,0
Cellule 33 40 X 80 / 300 = 10,7

Il faut examiner le comportement des deux évaluateurs. Ainsi, l'évaluateur I a jugé qu'il n'y avait pas de problèmes 160 fois sur 300 (53 % des fois) alors que l'évaluateur II a jugé qu'il n'y avait pas de problèmes 200 fois sur 300 (67 % des fois). On peut donc s'attendre à ce que les deux évaluateurs classent 10,67 cas dans la catégorie pas de problème. On remarque que la probabilité est égale au calcul de la fréquence théorique. C'est donc sur cette différence entre ce qui a été observé et ce qui, selon les probabilités, aurait dû se produire que le calcul du κ sera effectué.

La formule du κ est :

$$\kappa = \Sigma Fo - \Sigma Fe / N - \Sigma Fe$$

Où: Fo = la fréquence observée
Fe = La fréquence théorique
N = Le nombre total d'observations

Aussi dans notre exemple nous avons: 210 – 129,40 / 30 – 129,4 = 80,6 / 170,6 = ,47

Ce résultat signifie que les évaluateurs étaient d'accord 47 % des fois et non 70 % des fois. L'écart de 23% s'explique par le hasard. Le test montre donc qu'il est possible que les deux évaluateurs sont arrivés à la même conclusion par hasard. (Il n'y a pas de table de référence pour ce test; l'interprétation du résultat est de nature subjective; même s'il ne l'est pas ici, il peut arriver qu'un résultat de 47 % soit intéressant dans certaines recherches.)