Population et échantillonnage

La problématique étudiée dans ce cours est essentielle car c'est dans la population que l'on sélectionne les sujets inclus dans l'étude, et c'est à partir de cet échantillon que l'on tire les résultats et conclusion. Il convient donc de les choisir correctement.

1. La hiérarchie des populations

La première chose à savoir, c'est que le terme de *population* désigne plusieurs concepts en épidémiologie et recherche clinique et que ces différentes acceptions du terme ne sont pas égales. Le premier type de population est celle qui est concerné par votre question de recherche. On appelle cette population la **population cible**, c'est-à-dire la population pour laquelle on aimerait généraliser les résultats de l'étude. Du fait qu'on cherche souvent à généraliser le plus possible les résultats d'une étude, cette population cible est souvent de contours flous, difficile à définir précisément. Il faudra donc bien souvent isoler une sous-partie précise de cette population cible, dans laquelle on pourra aisément tirer un échantillon. Cette sous-population bien définie est la **population source**, la population à laquelle les chercheurs ont accès et d'où sont issus les personnes incluses. L'ensemble des personnes incluses représente l'échantillon d'étude, aussi appelé **population d'étude**, sous-ensemble de la population source.

exemple : Si vous voulez connaître la prévalence des troubles bipolaires dans la population Nantaise, votre population cible sera la population nantaise, votre population source pourra être la population nantaise inscrite sur l'annuaire (accessible à l'échantillonnage) et votre population d'étude sera l'échantillon finalement inclus.

2. Définir les populations

L'une des premières choses à préciser lors de la construction d'une étude, ce sont les définitions des populations cible et source. Pour cela, on définit des critères d'éligibilité, qui peuvent être positifs (critères d'inclusion) ou négatifs (critères d'exclusion). Ces critères sont de trois ordres : 1. critères spatiaux et temporels 2. critères liés aux expositions ou aux maladies étudiées 3. critères de bon déroulement de l'étude

Si les deux premiers tsont souvent aisés à définir, les critères de bon déroulement de l'étude sont plus difficile à appréhender : il peut s'agir par exemple d'assurer la bonne adhésion des sujets au protocole (sélection de sujets compliants), de limiter les perdus de vue (limiter les sujets dont le suivi sera difficile), de s'assurer d'une faisabilité pratique ou économique (mesures utilisant une méthode peu invasive et peu onéreuse), de limiter des obstacles légaux (inclusion seulement des personnes majeures), etc.

Ces critères doivent aussi présenter certaines caractéristiques résumées par l'acronyme POOLS :

- Précis
- Objectifs

- Opérationnels
- Logiques
- Sensibles et/ou spécifiques

Par **précision**, on entend qu'une définition précise des individus doit être apportée, qui ne prête pas à confusion. Par exemple, on n'inclura pas des patients présentant une fièvre, mais des patients présentant une température auniculaire supérieure à 38°C.
Les critères doivent reposer sur des éléments **objectifs**, et quand cela n'est pas possible sur des outils validés et standardisés (échelles de douleur, de qualité de vie, etc.). L'aspect **logique** des critères renvoie l'utilisation d'opérateurs logiques, comme ET, OU, NON.

Par exemple, on pourra inclure des individus présentant un fièvre supérieure à 38°C ET une toux mais NON présentant des antécédents de BPCO.

Ces définitions peuvent être plus ou moins **sensibles ou spécifiques** (voir le cours sur les erreurs de mesures et les études diagnostiques) et il est possible de définir plusieurs niveaux de définition. Ainsi, dans une étude cas-témoins, une définition sensible définira les cas possibles, une définition spécifique sera utilisée pour les cas certains, et les cas probables se rapporteront à une définition de niveau intermédiaire.

3. Définition du type de question d'étude

Comme on le verra par la suite, une question fondamentale en méthodologie épidémiologique est de savoir distinguer entre les études **descriptives** et **comparatives**. Les études descriptives s'intéressent à l'estimation d'un paramètre dans une population. Connaître la taille moyenne de la population française ou l'incidence du cancer du sein chez la femme de moins de quarante ans en Europe au cours des 20 dernières années sont des exemples. L'idée est de décrire la population sous-jacente. L'important dans ce cas est d'avoir un échantillon **représentatif**, c'est-à-dire semblable à la population cible ––—sur les critères importants tout au moins.

Le but final des études comparatives est d'étudier des liens plusieurs variables : l'exposition aux pesticides est-elle liée à la survenue des hémopathies malignes ? La prise de ce nouveau médicament est-elle associée à de meilleurs résultats thérapeutiques ? L'important est ici que les groupes soient **comparables**. On s'efforcera donc dans ce cas de faire en sorte que le groupe de référence et le groupe de comparaison soient issus de la même population.

4. Calculer le nombre de sujets nécessaires

Afin de limiter suffisamment l'importance de la fluctuation aléatoire et par là même d'augmenter la précision, il est nécessaire d'inclure suffisamment de personnes dans dans votre étude. Cependant, plus il y a de personnes à inclure, plus l'étude sera longue, couteuse et difficile à gérer. Il convient donc d'inclure le minimum nécessaire d'individus. Pour estimer cette taille d'échantillon, il existe une différence entre les études descriptives et comparatives : pour les premières la taille de l'échantillon se base sur la **précision**, pour les secondes sur la **puissance statistique**.

Pour les études descriptives, il faut tout d'abord déterminer avec quelle précision on cherche à estimer le paramètre d'interêt. Cela revient en fait à fixer la taille de l'intervalle de confiance à 95 %.

Pour une prévalence, par exemple, la formule classique pour calculer l'IC95 est $p \pm 1.96 \times \sqrt{\frac{p \times (1-p)}{n}}$ avec **p** représentant la prévalence et **n** la taille de l'échantillon. Si on cherche à obtenir un intervalle de confiance de ± 5 %, il est facile de montrer que $n = \frac{p \times (1-p) \times 1.96^2}{(5\%)^2}$. Trois éléments sont donc essentiels à formuler : la précision, le type d'intervalle de confiance (ici la pondération à 1.96 est dû au choix d'un IC à 95 %) et la prévalence. Notons donc que pour calculer la taille de l'échantillon nécessaire à l'estimation de la prévalence, il faudra avoir une idée de la valeur de cette prévalence…

Pour l'estimation basée sur la puissance, la formule précise dépendra du test statistique utilisé, mais de façon générale, la formule peut s'écrire de cette manière : $f(\alpha, \beta) \times \frac{2 \times s^2}{\delta^2}$. Ici, les éléments à déterminer sont : **α** et **β**, les risques d'erreurs de type I et II, $\mathbf{s^2}$ la variance de la variable estimée et **δ**, la plus petite différence qu'on aimerait pouvoir détecter dans l'étude.

Il existe de nombreuses solutions pour estimer le nombres de personnes à inclure sans faire les calculs soi-même. Voici par exemple un lien vers un outil en ligne : Sampsize.

5. Méthode d'échantillonnage

La dernière étape dans la constitution d'une population d'étude est de choisir la méthode de tirage au sort la plus adaptée à la question de recherche est aux possibilités offertes.

Échantillons non-aléatoires

Pour les échantillons non-aléatoires, plusieurs méthodes existent de l'échantillon de convenance à l'échantillon par quotas. Plusieurs solutions sont expliquées sur le site de statistiques Canada. L'inconvénient principal de ces méthodes est bien sûr le manque de représentativité mais également le manque de validité des statistiques réalisées.

Échantillons aléatoires

Plusieurs possibilités existent pour tirer au sort un échantillon aléatoire simple. Tout d'abord, si nous possédons une liste exhaustive des individus composant la population source, il est assez aisé d'attribuer un numéro à chacun de ces individus et de tirer au sort une séquence de ces numéros dont la longueur est égale à la taille souhaitée de notre échantillon. On peut pour cela utiliser des générateurs de nombres aléatoires comme celui disponible sur random.net. Des méthodes plus pratiques existent aussi sur excel par exemple (pour une version française d'excel, remplacer la fonction ´´rand()´´ par ´´alea()´´). Cette méthode s'appelle un **échantillonnage élémentaire**.

Si en revanche on ne possède pas la liste exhaustive des individus ou que celle-ci est difficile à obtenir ou à utiliser (liste longue et non numérisée, par exemple). On utilisera alors un **échantillonnage systématique**. La première étape d'un échantillonnage systématique est

de déterminer un **pas de sondage**, qui est égal à la taille de la population divisée par la taille souhaitée de l'échantillon. Ensuite, il ne reste plus qu'à choisir aléatoirement le premier individu à inclure en tirant au sort un numéro compris entre 1 et le pas de sondage. Pour plus d'information sur le sondage systématique (et les autres types de sondage), voir cette vidéo.

Il peut arriver que vous vouliez estimer un paramètre pour différentes sous-populations dont la fréquence dans la population source est variable. Il est possible que la population la moins représentée soit peu tirée au sort et que l'estimation du paramètre soit peu précise. On pourra dans ces cas-là réaliser un tirage au sort stratifié, c'est-à-dire tirer au sort avec un taux de sondage différent dans les différentes sous-populations, afin de sur-représenter celles étant les moins fréquentes.

Enfin, il arrive fréquemment que tirer au sort dans l'ensemble de la population source soit difficile à réaliser. Prenons l'exemple d'une étude sur les capacités de lecture d'enfants en CM1 en France. Des listes d'élèves existent dans chaque école mais il peut être fastidieux de les réunir toutes et de les fusionner. De plus, le tirage au sort amènerait à sélectionner un nombre très réduit d'enfants par écoles et nécessiterait aux enquêteurs beaucoup de déplacements entre les écoles. Dans des cas comme celui-ci, il est possible de réaliser tirer aux sort des grappes (ici, par exemple des écoles) puis d'inclure les individus de ces grappes. Plusieurs méthodes d'échantillonnage utilisent ce procédé, comme les échantillonnages par grappe (cluster sampling) et les échantillonnages multi-niveaux. Il est important de noter que les individus issus d'une même grappe (par exemple d'une même classe) seront souvent plus semblables entre eux qu'avec les individus d'autres grappes. On dit que la variance intra-grappe est plus faible que la variance inter-grappe. Des méthodes statistiques avancées sont nécessaires pour prendre cela en compte, mais elles dépassent largement le cadre de ce cours.

Cas des études comparatives épidémiologiques

Dans les études comparatives épidémilogiques, il est souvent difficile de sélectionner les groupes à comparer (cas et témoins ou exposés et non exposés) directement dans la population source. Plusieurs recettes existent et aucune ne fait référence et il faut bien souvent discuter des possibilités au cas par cas. Plusieurs aspects sont à vérifier cependant pour s'assurer que les groupes sont bien comparables : - Si un membre du groupe contrôle (témoins ou non-exposés) étaient finalement malade/exposé, pourrait-il faire partie du groupe d'intérêt ? - La probabilité d'exposition/maladie dans le groupe témoin est-elle la même que celle de la population source ?