



L'épidémiologie pour tous

Représentativité d'un échantillon

Rédigée par B. Toma, B. Dufour, E. Garin, J. Rivière, J.J. Bénét et validée par le bureau de l'AEEMA

En [épidémiologie](#), il est souvent question d'« échantillon(s) représentatif(s) ».

Qu'est-ce que cela veut dire ?

En épidémiologie, les [enquêtes](#) sont très souvent réalisées sur des échantillons issus de [populations](#) pour lesquelles on veut obtenir des informations (pourcentage de prévalence d'une maladie, par exemple), mais dont la taille est trop grande pour que tous leurs sujets puissent être étudiés. Ensuite, à partir des résultats obtenus sur l'échantillon soumis à l'étude, on estime la prévalence dans la population. Pour que cette estimation soit valide, il faut qu'elle ne soit pas biaisée (c'est-à-dire faussée). Pour cela, il faut que la composition de l'échantillon ressemble suffisamment à celle de la population. Sous cette condition, la caractéristique étudiée, en l'occurrence, le plus souvent, la fréquence d'une maladie, est semblable dans la population et dans l'échantillon.



*En termes courants, on peut donc considérer qu'un **échantillon est représentatif** d'une population lorsque la fréquence des **caractéristiques de ses sujets** et celle **des sujets de la population** dont il est issu **sont similaires**.*

Il peut s'agir de caractéristiques démographiques intrinsèques (non modifiables), comme l'âge, le sexe, la race, ou extrinsèques (modifiables), comme le lieu de résidence, les habitudes alimentaires, la catégorie socio-professionnelle.

Quand on travaille sur un (des)échantillon(s), a-t-on toujours besoin qu'il(s) soi(en)t représentatif(s) ?

Les objectifs d'une enquête épidémiologique peuvent être variés.

- En [épidémiologie descriptive](#) (cf. Fiche [Epidémiologie descriptive versus épidémiologie analytique](#)), l'objectif peut être quantitatif ou qualitatif :

➔ L'objectif *quantitatif* vise à connaître la fréquence (absolue et/ou relative) d'une maladie dans une population à un instant donné, pendant une période donnée ou son évolution au cours du temps. La fiabilité des résultats d'une enquête descriptive à objectif quantitatif est **conditionnée par la représentativité de l'échantillon étudié**. Par exemple, un sondage réalisé sur des volontaires risque fort de ne pas être représentatif de l'ensemble de la population, la motivation des participants risquant soit de surestimer (intérêt pour bénéficier d'un test de dépistage gratuit en craignant d'être atteint), soit de sous-estimer (crainte d'une décision défavorable en cas de résultat positif) la valeur de l'estimation.

➔ L'objectif *qualitatif* est de savoir si le pourcentage de sujets atteints d'une maladie dans une population est supérieur ou non à un seuil déterminé. Une enquête descriptive à objectif qualitatif n'a pas forcément besoin d'un échantillon représentatif. En effet, dans le cas où seule compte la question de présence (seuil = au moins 1) / absence (= en dessous de

ce seuil, soit 0), pour augmenter les chances de révéler la présence d'une maladie dans une population, l'enquête peut être conduite sur un [échantillon](#) « ciblé » sur les sujets à risque, c'est-à-dire les sujets ayant la plus grande probabilité d'être atteints si la maladie est présente.



Par exemple, dans un troupeau, on prélèvera systématiquement tous les reproducteurs (en particulier tous les mâles) et tous les animaux âgés, pour augmenter la probabilité de détecter au moins un sujet à résultat positif si le troupeau est infecté.

- En [épidémiologie analytique](#) (cf. Fiche [Epidémiologie descriptive versus épidémiologie analytique](#)), l'objectif est de vérifier une hypothèse d'association entre un [facteur de risque](#) supposé et une maladie.

Une enquête analytique est fondée sur la comparaison entre deux échantillons, **dont la qualité majeure est leur comparabilité et non pas la représentativité par rapport à leur population d'origine.**

C'est donc surtout en épidémiologie descriptive, dans le cadre d'enquêtes quantitatives, que la représentativité de l'échantillon est nécessaire.

Comment obtenir qu'un échantillon soit représentatif ?

La méthode la plus sûre et la plus simple (au plan théorique) pour obtenir un échantillon représentatif est **le tirage au sort** (ou tirage aléatoire) au sein de la population d'intérêt. Ceci implique :

- d'avoir défini l'[unité épidémiologique](#) d'intérêt (par exemple, les élevages de bovins dans un département) ;
- de disposer de la *liste exhaustive* et *à jour* des unités épidémiologiques dans la population d'intérêt (dans cet exemple, la liste exhaustive de tous les élevages de bovins dans le département) ;
- de disposer d'une *méthode de tirage au sort* au sein de la liste.

En fonction des situations, par rapport à l'échantillon aléatoire simple (de l'exemple ci-dessus), la constitution d'un échantillon représentatif peut parfois être plus

complexe : échantillon aléatoire à deux degrés, échantillon en grappes, échantillon stratifié, etc. Il existe également d'autres méthodes que le tirage au sort pour obtenir un échantillon représentatif (par exemple, la méthode des quotas).



La représentativité de l'échantillon conditionne l'exactitude du résultat (cf. Fiche [Exactitude et précision d'une estimation](#)), mais n'intervient pas dans sa précision (qui est conditionnée, elle, par la taille de l'échantillon).

A quoi ça sert ?

Le résultat relatif à la fréquence d'une maladie, obtenu sur un échantillon représentatif, permet d'estimer la valeur de cette fréquence dans la population (ce qui est l'objectif de l'enquête descriptive) et de la situer au sein d'une fourchette (ou intervalle de confiance), en acceptant un faible risque de se tromper par excès (en général 5 %), c'est-à-dire que la valeur véritable soit située en dehors de cette fourchette. Plus la taille de l'échantillon représentatif est élevée, meilleure est la précision (plus la fourchette de l'estimation est réduite)ⁱ.



*Tout échantillon constitué de manière aléatoire est présumé représentatif de la population d'origine.
(Schwartz : « Un échantillon, si petit soit-il, est représentatif dès lors qu'il résulte d'un tirage au sort »).*

Toutefois, le tirage au sort ne supprime pas le rôle du hasard, générateur de [fluctuations d'échantillonnage](#). Les conséquences du hasard sont d'autant plus grandes (sur le résultat de l'étude) que la taille de l'échantillon est petite (faible précision).

Exemple : soit une population dont 50 % des sujets sont infectés par un agent pathogène donné. Un échantillon de deux sujets tirés au sort est parfaitement représentatif de cette population. Mais, une fois sur quatre, il peut ne comporter aucun sujet infecté ou en comprendre deux et, une fois sur deux, n'avoir qu'un seul sujet infecté (sur les deux). L'écart entre le résultat

obtenu sur ce petit échantillon (aucun infecté ou deux infectés) et la réalité dans la population peut donc être très important (fourchette maximale de 0 % à 100 %). L'augmentation de la taille d'un échantillon représentatif diminue l'effet du hasard et augmente ainsi la précision du résultat. Il en est de même pour un échantillon non représentatif. Toutefois, un échantillon biaisé, même s'il est de grande taille, demeure biaisé et peut conduire à des résultats plus ou moins éloignés de la réalité dans la population.

Un échantillon peut avoir été constitué par tirage au sort, mais si des individus ainsi sélectionnés n'ont pas participé à l'étude, il peut y avoir un risque de biais, lorsque la raison de leur non-participation est liée à la fréquence de la maladie. C'est le cas pour les refus, mais aussi pour la simple absence (hospitalisation pour les personnes, ou intervention sanitaire pour les élevages).

Et si l'échantillon n'est pas représentatif, qu'est-ce que ça change ?

Dans certaines situations d'épidémiologie descriptive quantitative, il peut arriver que l'on ne puisse pas

disposer d'un échantillon représentatif à cause de différentes difficultés matérielles (absence de liste exhaustive des unités de la population, par exemple). Si un échantillon de « convenance » a été utilisé (ce qui est pratiquement toujours le cas pour la faune sauvage, car on ne dispose pas d'une liste de tirage au sort), il faut :

- savoir que l'inférence à la population du résultat obtenu sur l'échantillon ne serait pas judicieuse et **l'estimation de la prévalence au sein d'un intervalle de confiance, pas souhaitable ;**
- rester très prudent avec le résultat obtenu, ne le considérer que comme une information provisoire à vérifier ultérieurement, si possible sur un échantillon représentatif, et indiquer que l'intervalle de confiance qui serait calculé ne serait applicable qu'à un échantillon représentatif ;
- s'il est possible d'estimer le sens (surestimation ou sous-estimation) et l'importance probables du biais introduit par rapport à la fréquence supposée de la maladie dans la population, le signaler dans l'expression du résultat.



- Un échantillon **représentatif** « **ressemble** » suffisamment à la **population** dont il a été tiré pour que l'**estimation** du pourcentage de prévalence soit **fiable**.
- Un des **moyens** d'assurer la représentativité d'un échantillon est le **tirage au sort** (tirage aléatoire simple). Mais d'autres moyens sont possibles.
- Un échantillon constitué sans strict respect d'une méthode adaptée a de grandes chances de **ne pas être représentatif**.
- Il faut **toujours vérifier la représentativité** d'un échantillon en **comparant** ses caractéristiques à celle de la **population** source.
- **Si l'échantillon n'est pas représentatif, on ne peut pas corriger le biais, mais seulement en apprécier l'importance et le sens.**



Toma B. et al. : *Epidémiologie appliquée*, 2018, AEEMA éditeur, 614 p.

Terminologie : des liens hypertextes (en bleu souligné) vous permettent d'accéder aux définitions sur le site de l'AEEMA

Fiches AEEMA : [Epidémiologie descriptive versus épidémiologie analytique](#) ;
[Exactitude et précision d'une estimation](#)
