CHOIX DES POPULATIONS ÉTUDIÉES

J.J. BENET

Ecole nationale vétérinaire - 94704 Maisons-Alfort

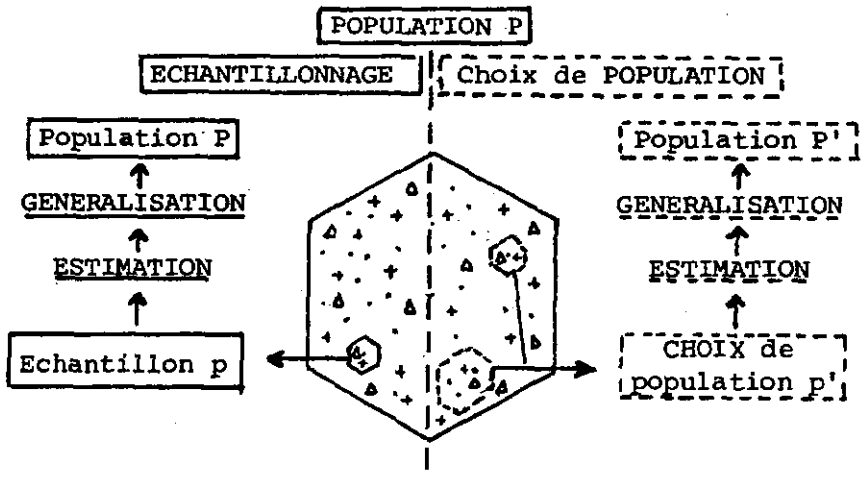
RESUME

Tout travail épidémiologique suppose que l'on extrait d'une population globale un échantillon sur lequel on procèdera aux observations et qui permettra d'extrapoler les conclusions à la population initiale. Pour s'autoriser cette généralisation, il faut satisfaire une règle essentielle : définir les conditions de travail : objectif de l'étude, population initiale, donnée d'observation, etc., constituent des aspects fondamentaux à définir avant d'aborder l'aspect technique du choix proprement dit. Chacune de ces étapes peut être source de biais, dont quelques exemples sont donnés.

Quelle que soit la définition qu'ils donnent à l'épidémiologie, les épidémiologistes ont en commun d'étudier des populations humaines, animales, ou végétales, dont la dimension suppose souvent des méthodes particulières. Il n'est pas possible en effet, de faire porter une étude épidémiologique concernant une population "P" sur son ensemble (bien que la détection systématique des Maladies Légalement Réputées Contagieuses, M.L.R.C., fasse exception, dans un domaine voisin) ; on est obligé de "prélever un échantillon" de population "p", et d'extrapoler les conclusions à l'ensemble "P" dont il est extrait.

En statistique, il s'agit d'un élémentaire problème d'échantillonnage qui permet la constitution d'une population la plus représentative possible de la population initiale (figure 1).

Figure 1 : Principes de l'échantillonnage et du choix des populations.



En épidémiologie pratique, des difficultés multiples conduisent à prendre une latitude certaine par rapport aux règles statistiques : pour marquer cet écart dans la constitution de l'échantillon, plutôt que d'échantillonnage nous préférons parler de choix de populations, dans la mesure où ce n'est plus le hasard seul qui préside à l'élaboration de

cette population à étudier, mais aussi (ou seulement) des considérations pratiques, telles que ... "on n'a pas pu faire autrement !" Il est certain que la population "p'" obtenue dans ces conditions sera notablement distincte de la population idéale théorique "p" que l'on aurait pu obtenir en respectant les règles de la statistique. Le tout est de parvenir, malgré ces contraintes pratiques, à une généralisation satisfaisante, ou plus simplement à une réponse satisfaisante à la question posée.

Il n'y a pas de formule miracle (même statistique), encore moins d'ordinateur magicien (quelle qu'en soit la toute puissance), qui permette de résoudre ce problème. Tout au plus, est-il possible de proposer une "recette" : trouver l'équilibre satisfaisant entre ce que l'on veut faire, et ce que l'on peut faire. Cette formulation simpliste cache un ensemble de questions qui conduisent à la solution, pourvu qu'on y réponde AVANT de commencer le travail de terrain.

Nous verrons ces questions dans le cadre d'enquêtes descriptives, et d'expérimentation ; puis nous montrerons quelques exemples de biais. Au préalable, nous soulignerons les différences existant entre ces types d'enquêtes, et leur importance en ce qui concerne le choix des populations.

I - OBJECTIF DE L'ETUDE

Tout commence donc par une première question : "Que veut-on faire ?" Trois réponses sont possibles :

- "Je veux savoir, connaître..." : il s'agit d'une optique descriptive.
- "Je veux comprendre pourquoi, expliquer..." : l'optique est étiologique.
- "Je veux connaître les résultats du plan de prophylaxie que j'ai mis en place, et comprendre les difficultés rencontrées" : l'optique est mixte pour ces enquêtes d'évaluation de l'efficacité d'un plan de prophylaxie.

Les caractéristiques générales, et différents exemples de ces enquêtes sont d'abord évoqués ici avant leurs conséquences du point de vue du choix des populations.

A - *DIFFERENTS TYPES D'ENQUETES*

1. Optique descriptive

L'enquête est axée sur la population qui constitue l'objet d'étude. La nature des informations diffère selon l'objectif de l'enquête.

a. Surveillance de l'état sanitaire d'une population

. Les premières applications épidémiologiques de ce type ont été la détection des individus atteints de M.L.R.C. Le caractère exhaustif en limite automatiquement le nombre.

. De très nombreuses études sont faites ou à faire pour connaître l'état sanitaire de telle population vis-à-vis de telle maladie.

Il faut souligner la demande cruciale d'informations existant dans ce domaine, et la priorité que l'on doit accorder à ce type d'études, qui permettent de mieux orienter la lutte contre les maladies.

Dans cette optique, il s'agit de connaître la proportion de cheptels, d'individus atteints, infectés par exemple, l'évolution dans le temps,

la répartition et l'évolution dans l'espace, l'impact économique de ces affections, qui devrait être systématiquement envisagé dans l'étude, car l'état de "santé économique" d'un cheptel par exemple fait également partie des préoccupations du spécialiste des maladies animales.

Par extension, on peut être amené à réaliser un véritable bilan de santé d'une population, en dehors de toute manifestation pathologique. Pour cela, il faut trouver des variables que l'on puisse mesurer, afin de quantifier cet état (Exemple : enquête permanente de la Station de Pathologie porcine, Ploufragan).

b. Aide à la planification

L'élaboration, la mise en route d'un plan de prophylaxie sur une grande échelle nécessitent une grande quantité d'informations, qui toutes ne sont pas fournies par la surveillance épidémiologique. Il s'agit en effet d'aboutir à :

. la définition de priorités en fonction de critères économiques, et donc de connaître le coût approximatif du problème étudié ;

. l'évaluation des moyens d'action possibles disponibles, des difficultés éventuelles.

Ainsi, cette démarche peut conduire à apprécier l'importance d'une maladie dans une population, comme dans la surveillance, mais aussi l'impact économique (sous tous ses aspects, directs et indirects) ; par ailleurs, elle doit révéler les obstacles pratiques qu'il faudra résoudre : pratique de la réalisation des prélèvements (qui, quand, où), difficulté provenant de la contention des animaux (... ou absence de...), appréciation des déplacements nécessaires, accueil de professionnels divers.

c. Aide à la recherche

En associant la connaissance de la fréquence d'une maladie et de facteurs de risque correspondants, ce type d'enquête ouvre la voie à des recherches complémentaires, en aidant à la formulation d'hypothèses explicatives. Exemple : syndrome de la truie maigre.

Les différents aspects (surveillance sanitaire, aide à la planification, aide à la recherche) relèvent d'une optique descriptive, bien qu'ils débordent parfois du strict cadre de la pathologie. Les informations fournies sont pourtant indispensables, et dans la mesure du possible, on a intérêt à exploiter au mieux tout système informatif mis en place afin de contribuer à une meilleure approche ultérieure de l'un ou l'autre de ces objectifs.

2. Recherche étiologique

La recherche étiologique a pour but de démontrer une relation de causalité entre facteur (s) et maladie. Elle est donc centrée sur le facteur. Elle est soit explicative, soit pragmatique.

Dans l'attitude explicative, il s'agit des hypothèses causales, permettant de comprendre l'étiopathogénie des maladies. La difficulté provient du fait de l'hétérogénéité des groupes étudiés : le rôle de l'épidémiologiste consiste à les rendre plus comparables, afin d'isoler les facteurs étiologiques susceptibles d'expliquer l'origine de la maladie.

L'attitude pragmatique contourne cette difficulté en déterminant des groupes à haut risque, auxquels on applique des mesures visant à prévenir la maladie.

Prenons l'exemple de la relation entre l'avortement non brucellique chez la vache et un facteur x (par exemple infectieux) dont la responsabilité n'est pas encore démontrée.

La recherche explicative peut conduire à indiquer une bonne corrélation entre le facteur x et l'avortement, mais d'autres facteurs peuvent intervenir, indépendamment ou associés. Il faut, pour lever l'incertitude, une véritable expérimentation, en particulier l'inoculation de l'agent présumé causal... ce qui peut poser des difficultés insurmontables.

D'un abord pratique plus facile la recherche pragmatique consiste (par exemple) à vacciner une partie de la population exposée contre l'agent envisagé, et à comparer les taux avec la population non vaccinée.

A la frontière entre l'optique descriptive et la recherche étiologique se situe l'évaluation des interventions.

3. Evaluation des interventions

Une fois un plan de prophylaxie mis en place, il faut vérifier qu'il atteint son objectif. Le simple suivi des opérations est descriptif, l'étude de l'efficacité relève de l'évaluation d'une relation de cause à effet. On comprendra tout de suite les difficultés qui peuvent surgir en prenant pour exemple l'étude à partir des seules données rétrospectives de l'efficacité du B.C.G. dans la prévention de la tuberculose humaine. L'aspect descriptif montre une diminution des cas de tuberculose humaine parallèlement à l'utilisation du B.C.G. en France. Toutefois, on doit souligner également qu'il existe des interactions concomitantes dues à l'efficacité des traitements antibiotiques, de la pasteurisation du lait de vache. Cet exemple montre bien qu'il n'est pas possible d'établir de relation de cause à effet à partir d'une situation purement descriptive. Une situation toute différente a été mise en place dans l'exemple suivant : l'essai de la prophylaxie de la fièvre aphteuse par abattage des animaux sensibles contaminés a été réalisé en 1957 dans le Finistère, en comparaison avec le reste de la France, et a clairement démontré la supériorité de cette conception par rapport à la vaccination sans abattage.

On voit donc que seule une situation véritablement expérimentale permet d'aboutir à des conclusions sur la relation de cause à effet, et donc sur l'efficacité des mesures entreprises.

La détermination préalable du type exact d'enquête menée est capital du point de vue du choix des populations, de par les conséquences que cela entraîne.

B - CONSEQUENCES SUR LE PLAN DU CHOIX DES POPULATIONS

Priorité = population ou facteur ?

Dans l'optique descriptive, la priorité est donnée à la définition de la population "p", car elle doit être représentative de la population "P" objet de l'étude ; on choisira donc les différentes catégories et leur importance en fonction de la connaissance de la population initiale "P",

qui est un préalable indispensable. Il est certain que des contraintes pratiques peuvent nous amener à nous écarter des exigences statistiques, et ce sera aux dépens de la représentativité de l'échantillon : il ne peut alors être véritablement considéré comme représentatif de la population initiale "P", mais seulement d'une partie ("P") de cette population qu'il importe de définir dans tous ses caractères... Les conclusions de l'observation ne pourront pas être généralisées à la population initiale "P", en revanche, il sera possible d'étendre les résultats à la population plus restreinte (P') dont est extrait l'échantillon (p'), ce qui en pratique peut être tout à fait suffisant.

En ce qui concerne la démarche étiologique, le choix de la population n'est guidé que par un seul souci, celui de définir deux groupes (au moins) les plus ressemblants possibles l'un à l'autre, sans aucun égard pour une représentativité à la population initiale qui n'aurait absolument aucun intérêt. En effet, ce que l'on étudie ici est le facteur dont on suppose qu'il peut avoir un rôle, qu'on l'introduise dans l'une des populations, ou qu'il représente la seule différence véritable entre les deux populations. Dans ces conditions, des résultats différents entre les deux groupes identiques au départ, au facteur près, on pourra déduire le rôle étiologique du facteur envisagé. La représentativité n'apporte aucune valeur supplémentaire à cette conclusion. Tout au plus sera-t-il nécessaire de recommencer l'expérience pour les différentes catégories de population existantes, si l'on suppose qu'elles peuvent avoir une influence sur le plan de l'étiologie. Ces expériences n'auront d'autres exigences statistiques que qualitative : groupes comparables entre eux, et quantitative : nombre de sujets suffisant pour permettre des conclusions statistiques valables.

Illustrons ces différences fondamentales entre les deux types d'enquêtes sur le plan du choix des populations. Dans une enquête descriptive, il s'agit pour une maladie donnée ("x") de connaître la répartition sur l'ensemble du cheptel bovin français. L'échantillonnage tiendra compte des différences géographiques nationales, de la densité d'élevages, de la représentation des races, des classes d'âge, des types d'élevages, etc., afin de constituer un échantillon représentatif de la population bovine française dans son ensemble. En revanche, si pour la même maladie "x" il s'agit de mettre en évidence des facteurs étiologiques, sous forme de facteurs de risque, on n'a pas intérêt à prendre un échantillon représentatif, car la fréquence de la maladie étant relativement limitée, le poids des facteurs de risque serait trop faible pour qu'une conclusion soit possible, à moins de réunir une population bien trop importante ; il vaut mieux sélectionner des élevages atteints de cette maladie, que l'on comparera à d'autres élevages qui auront les mêmes caractéristiques de race, de taille de troupeau, de mode d'élevage, de conditions d'environnement, à un seul facteur près celui de la maladie, et l'on étudiera quelles sont les différences que l'on peut mettre en évidence entre ces deux groupes (atteints, et témoins), et qui peuvent constituer des facteurs de risque. Dans ce dernier cas, la seule représentativité que l'on recherchera sera celle de la population témoin par rapport à la population exposée au risque, et non par rapport à l'ensemble plus vaste de population générale dont ces deux populations sont extraites.

II - CHOIX DES POPULATIONS DANS UNE OPTIQUE DESCRIPTIVE

La formulation de la question envisagée ("je veux savoir...") doit devenir la plus précise possible. Il faut commencer par préciser l'objectif épidémiologique.

A - PRECISION DE L'OBJECTIF EPIDEMIOLOGIQUE

Une formulation telle que : "je veux étudier l'épidémiologie de telle maladie dans telle région" est tout à fait insuffisante.

En fait, toute "quête"épidémiologique constitue une "question" précise dont les termes sont étroitement définis par les conditions de l'enquête. Mais pour savoir si ces conditions permettent précisément de répondre aux voeux du chercheur, il faut tout d'abord formuler précisément le problème, et cela dans les moindres détails.

Veut-on connaître la répartition géographique de la maladie, son évolution éventuelle, son incidence, sa prévalence, son évolution dans le temps... ? Autant de questions dont il faut tenir compte alors au départ dans le protocole, car il faudra prévoir les lieux d'intervention et les moments. A ce stade, il est encore prématuré de répondre précisément à ces questions, qui dépendent étroitement de la définition de la population faisant l'objet de l'étude.

B - DEFINITION DE LA POPULATION OBJET DE L'ETUDE

Aussi bien en statistique qu'en épidémiologie, on entend par "population" un ensemble d'unités de nature très diverses : ensemble de personnes, d'animaux, de végétaux, d'élevages, d'évènements,d'objets abstraits (ensemble des résultats fournis par différents laboratoires...).

* Au départ, il faut déterminer le champ de l'étude, ou pour reprendre la comparaison statistique ce que l'on compte mettre dans l'urne : va-t-on étudier la population dans son ensemble P, ou une partie P', définie alors par quels critères.

Exemple : les chiens vivant dans la région étudiée ou les chiens de moins d'un an vivant dans cette même région.

* Il faut aussi déterminer la nature de l'unité statistique étudiée : individus ou collectivités. Cette distinction est fondamentale, car elle recouvre deux réalités bien différentes sur le plan statistique (exemple : animaux, élevages) même si apparemment elle concerne les mêmes objets d'observation (animaux). De fait, cette distinction n'est pas toujours aussi évidente qu'il peut y paraître, car les individus peuvent être regroupés en collectivités, par le simple fait qu'ils ont quelque chose en commun. Exemple : les prélèvements réalisés à l'abattoir sont des individus mais ce sont des unités collectives pour ceux qui sont réalisés le même jour.

Cette précision n'est pas accessoire car certaines de ces unités peuvent notablement s'écarter de la population moyenne. Exemple : le même jour un important arrivage à l'abattoir provient d'un gros élevage.

* Choix des critères qualitatifs : si la statistique utilise souvent le hasard, cela ne veut pas dire que le choix de population se fasse n'importe comment. Tout "choix" (d'un objet que l'on désire acheter dans un étalage, par exemple) suppose une prise de connaissance préalable des différentes catégories offertes, de leurs différents caractères.

La population ainsi peut être définie en ce qui concerne les animaux par : l'espèce, la race, le sexe, l'âge, les aptitudes de production, mais aussi par leur environnement (type d'élevage, d'alimentation, contexte

pathologique...) ; pour les élevages par : la taille, le type de production (laitier, viande), le mode (intensif, extensif, industriel, fermier), les conditions de logement (stabulation libre, entravée)...

Les critères sont variés mais il s'agit seulement d'en faire la liste, afin de choisir ceux que l'on retiendra, en fonction de l'intérêt épidémiologique qu'ils peuvent avoir. Exemple : bovins de plus de deux ans, élevages d'une taille supérieure à la moyenne départementale.

* On définit aussi les critères d'exclusion, qui sont de nature différente des précédents. Ce ne sont pas comme précédemment des caractères descriptifs de la population qui pourraient avoir un intérêt épidémiologique. Il s'agit de critères très divers selon lesquels on rejette a priori une catégorie. Exemple : vaches ayant mis bas depuis moins de 30 jours (bien qu'elles aient plus de 2 ans) ; élevages dans lesquels on vaccine contre telle maladie (même si leur taille est correcte).

* Il faut ensuite préciser la taille des populations : population totale retenue, catégories définies par les critères choisis.

Enfin on dispose des renseignements indispensables permettant de décrire précisément la population P (ou une sous-population P') dont on veut extraire un échantillon. On comprend combien ce travail préalable est capital, car il permettra de savoir si l'échantillon obtenu correspond bien à la population dont on voulait l'extraire (ou sinon, de quelle manière il s'en écarte). On appréciera ainsi s'il est représentatif.

C - DEFINITION DE LA DONNEE D'OBSERVATION

La donnée d'observation et le choix de la population interfèrent réciproquement. En effet, selon qu'il s'agit d'une observation clinique, d'une mesure, d'un prélèvement (sur le vivant, sur le cadavre), les conséquences ne sont pas les mêmes, et les problèmes pratiques peuvent sérieusement limiter les ambitions initiales, en particulier de représentativité... au point, que l'on peut être amené à procéder à l'inverse de ce que nous avons exposé : en raison de contraintes pratiques irréductibles, seules certaines catégories de données d'observation sont possibles et en fonction de cette donnée, telle population est accessible.

Par exemple, pour une enquête, les organisateurs reconnaissent qu'il leur est impossible de procéder à des visites dans les élevages ; en revanche, il est possible de réaliser un dépistage par exemple sérologique à partir de prélèvements obtenus à l'abattoir, qui constitue une source de données particulièrement privilégiée. Le problème sera alors de constituer un échantillon représentatif de la population étudiée à partir des informations fournies par l'abattoir, ce qui représente un travail particulièrement difficile.

D - GENERALITES SUR LE SONDAGE

Tout est prêt maintenant pour choisir notre population échantillon.

* Différentes méthodes de sondage

. Le choix peut être empirique

On fixe les catégories de population concernées, et des "quotas", c'est-à-dire le nombre de sujets sur lequel portera l'étude pour chaque catégorie.

Exemple : 20 élevages de 10 à 20 vaches,
20 élevages de 30 à 50 vaches,
20 élevages de 80 à 100 vaches.

On ne cherche pas la représentativité de l'échantillon, c'est-à-dire une image conforme de la population P. On veut seulement disposer d'une observation sur les différentes catégories. Cette solution est tout à fait valable lorsqu'il n'est pas possible d'obtenir une définition suffisamment précise de la population initiale. Bien entendu, l'interprétation doit être limitée en conséquence et tenir compte de cette absence de représentativité.

. Sondage aléatoire

On tire au sort les individus de la population parmi tous ceux de la population. Cela suppose que l'on dispose d'une liste, à partir de laquelle on pourra effectuer ce choix. Cette liste constitue la base de sondage (cf infra).

. Sondage pseudo-aléatoire

Si l'on ne dispose pas de base de sondage, on prend les sujets qui se trouvent dans une situation donnée au regard d'un critère qui n'est pas aléatoire, mais qui est supposé indépendant du phénomène à étudier.

Exemple : on prend dans un village la première ferme devant laquelle on trouve garée une voiture dont le numéro se termine par un trois.

Exemple : on dit à l'enquêteur "vous comptez 100 mètres après l'entrée du village, vous prenez la première ferme à droite ; s'il y a plusieurs routes d'accès, vous arrivez par le nord ; si l'habitat est dispersé, vous retenez le groupe situé près d'un calvaire..., etc.".

Cette dernière méthode élimine la subjectivité de l'enquêteur. Elle est souvent compliquée. La méthode aléatoire est préférable car elle permet d'échapper au choix humain, et de déterminer la précision du sondage.

* Le biais

L'estimation de la variable étudiée est d'une valeur différente de celle de la population concernée. Le problème est de savoir s'il s'agit d'un écart normal lié aux fluctuations d'échantillonnages, ou bien si avec d'autres échantillons, la même erreur serait retrouvée de façon systématique ce qui constitue alors un biais manifeste.

Les biais ont des causes multiples. Ils sont la hantise des chercheurs. Nous en verrons quelques exemples plus loin.

* Tirage au sort

L'usage de pièce de monnaie, ou d'un jeu de dés peut rendre des services... mais il est plus commode d'utiliser des tables de nombre au hasard. On peut les lire dans n'importe quel sens, prendre sur les 5 chiffres seulement ceux que l'on désire (figure 2).

Exemple : sur un effectif de 225 individus on veut sélectionner 20 sujets. On procède à une numérotation des individus de 1 à 225. On lit dans la table les 20 numéros inférieurs ou égaux à 225 (en prenant les trois premiers chiffres par exemple).

Figure 2 : Extrait d'une table de nombres au hasard.

26518	39122	96561	...
36493	41666	27871	...
77402	12994	59892	...
83679	97154	40341	...
71802	39356	02981	...
57494	72484	22676	
73364	38416	93128	
14499	89965	75403	
40747	03084	07734	
42237	59122	92855	

Le tirage au sort, associé à une bonne connaissance de la population initiale est la seule méthode permettant d'assurer la représentativité de l'échantillon.

* Bases de sondage

Pour effectuer ce tirage au sort, il faut disposer d'une liste ou base de sondage. Il en existe de toutes natures selon la population concernée.

Exemple : fichier D.S.V.,
fiche d'étable, registre d'étable,
carte géographique, photographie aérienne,
subdivisées en aires constituant autant d'unités de sondage,
registre de résultats diagnostiques.

E - DIFFERENTES TECHNIQUES DE SONDAGE

Les aspects sont ..."techniques", nous renvoyons donc aux ouvrages spécialisés, après une brève présentation.

* Sondage élémentaire

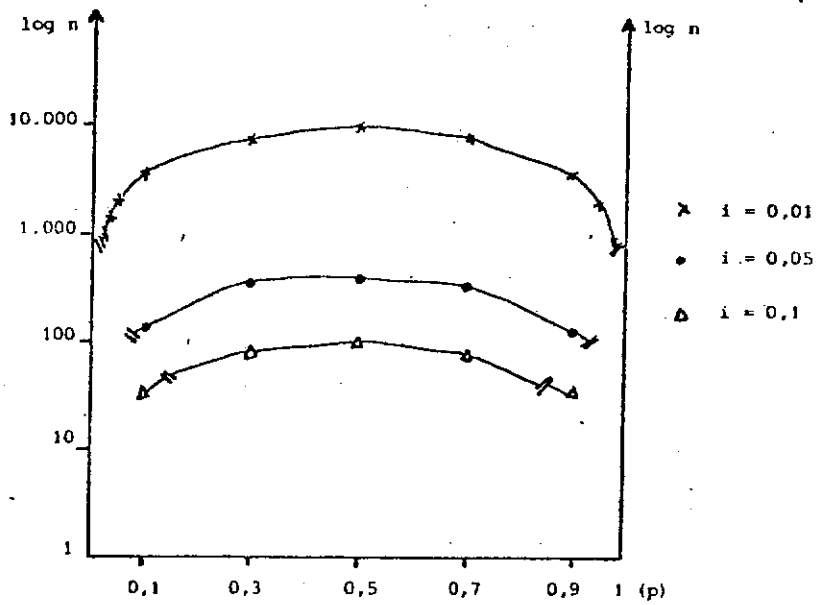
Le sondage élémentaire après tirage au sort permet d'estimer une variable quantitative (titre sérologique, poids de carcasses) ou qualitative (positifs - négatifs ; malades - sains). Il donne l'intervalle de confiance qui permet d'estimer que la moyenne de la population initiale P est située dans une certaine fourchette avec risque d'erreur choisi α

On peut aussi pour un écart de fluctuation que l'on se fixe (= précision) et pour un risque α que l'on accepte, calculer le nombre de sujets nécessaires (cf annexe I) : on peut ainsi prévoir le budget, la durée d'un programme.

La figure 3 donne le nombre de sujets nécessaires en fonction de différentes fréquences (p) du phénomène étudié. On voit que pour une fréquence de 0,1 le nombre de sujets nécessaires est de 144 pour une précision de 0,05 (soit un intervalle de 0,09 à 0,11), dans les deux cas le risque d'erreur est de 0,05. Dans ce cas très simple où la population est considérée comme homogène, le protocole expérimental varie notablement selon les conditions de rigueur que l'on se fixe ; on peut être ainsi amené à moins de prétentions dans la précision afin que l'observation soit possible ; on peut aussi faire l'économie d'un trop grand nombre d'observations pour des

conditions définies : il faut que le nombre de sujets soit multiplié par 4 pour que la précision soit deux fois meilleure, aussi quelques dizaines d'observations supplémentaires n'apporteront-elles pas grand chose de plus, mais elles représenteront une dépense inutile.

Figure 3 : Nombre de sujets nécessaires pour différentes fréquences (p) du phénomène étudié en fonction de différents degré de précision (i). Les traits de fraction limitent les domaines d'utilisation des courbes au delà desquels la marge d'erreur consentie conduit à accepter de trouver une réponse nulle.



Inversement, ce calcul préalable peut montrer que l'objectif envisagé est irréalisable, étant donné le nombre élevé de sujets indispensables.

Exemple : la prévalence de l'anémie infectieuse des équidés est d'environ 1 p. mille. Si l'on veut estimer cette proportion avec une précision de l'ordre de 0,5 p. mille (soit 1 $\pm$ 0,5 p. mille), le calcul montre qu'il faut... 400.000 chevaux, soit l'ensemble de l'effectif national !

Ces exemples de calcul sont simples parce que l'on a considéré que la population observée est homogène. Le principe demeure toutefois le même pour une population comportant plusieurs catégories ; on a recours alors à la stratification.

* Stratification

La stratification consiste à diviser la population étudiée en sous-populations (ou strates) plus homogènes. Par plus homogène, on entend que la variabilité à l'intérieur des diverses strates est plus faible que pour l'ensemble de la population. A l'intérieur de chaque strate, on effectue un sondage aléatoire qui sera plus précis. On juxtapose ensuite les estimations relatives à toutes les strates, pour obtenir une estimation globale meilleure que ne l'aurait donné un sondage simple dans la population entière.

Les strates sont définies au moyen des caractères de la population : espèce, race, sexe, age, subdivision géographique... On peut utiliser un ou plusieurs critères.

Supposons par exemple que l'on désire estimer le taux d'infection par une maladie donnée chez les ovins et les caprins. Etant donné qu'il y a

environ 10 millions d'ovins en France, et 1 million de caprins, un échantillon représentatif devrait comporter dix fois plus de moutons que de chèvres. Des observations ponctuelles ont montré que la fréquence de la maladie est de l'ordre de 0,05 ; pour obtenir une précision de l'ordre de 0,02, le nombre total de sujets nécessaires est de 475, que nous arrondissons par commodité de calcul à 500.

L'échantillon représentatif comportera donc 450 moutons, et 50 chèvres. Si la précision pour les moutons est bien de 0,02 pour les chèvres, elle n'est plus que de 0,06 en raison du faible nombre d'animaux, ce qui signifie que le résultat peut être égal à zéro, par manque de précision. Il faut donc augmenter le nombre de caprins dans l'échantillon c'est-à-dire les "sur-représenter" en stratifiant selon l'espèce et donc en abandonnant la représentativité : on réunit alors le nombre de sujets nécessaires pour une estimation correcte soit 475 dans chacun des groupes pour obtenir la précision souhaitée.

* Sondages en grappes

Bien que cherchant à estimer une valeur concernant des unités de sondage (animaux) on n'effectue pas toujours directement le choix de ces unités, mais plutôt indirectement, à partir de groupes d'unités (élevages), Chacun de ces groupes constitue une "grappe".

Un sondage en grappe consiste à choisir au hasard un échantillon de ces unités collectives, ou grappes, et à mener l'étude sur tous les individus qui contiennent les grappes tirées. Si l'on effectue un deuxième tirage dans chacune des grappes, il s'agit d'un sondage en grappes à plusieurs degrés.

Exemple : tirage au sort des élevages choisis, puis des individus retenus.

Exemple : tirage au sort des cantons, puis des élevages, qui seront observés en totalité.

Nous n'avons fait que présenter le principe de ces différentes modalités de sondage, qui doivent être connues de l'épidémiologiste, puisqu'elles conditionnent le protocole de l'enquête. Le traitement statistique peut être réservé au statisticien.

III - CHOIX DES POPULATIONS EN EXPERIMENTATION

L'expérimentation conduit à répartir une population en sous-groupes. Dans le cas le plus simple, il y en a deux, dont l'un est soumis à un type d'intervention et l'autre à un autre type d'intervention, ou pas d'intervention du tout (par intervention on entend toute action risquant de modifier l'état de santé de la population).

La répartition de l'échantillon en deux sous-groupes par tirage au sort est la seule méthode susceptible d'assurer une bonne comparabilité de ces groupes. Tous les deux sont ainsi représentatifs de l'échantillon initial et ne diffèrent entre eux que par les fluctuations dues au hasard. Si l'un des groupes fait (après le tirage au sort) l'objet d'une intervention A et l'autre d'une intervention B, on peut admettre que les différences enregistrées ultérieurement sont bien liées à l'intervention.

La différence est fondamentale par rapport à la situation d'épidémiologie descriptive où l'on peut seulement constater en quoi deux populations pour lesquelles on a estimé les valeurs d'une variable sont différentes (exemple : taux d'atteints) : on constate que la répartition par âge n'est pas la même que l'alimentation, l'environnement... sont différents, on peut éventuellement établir une corrélation (une association) entre certains facteurs et le taux, mais on ne peut en déduire de lien de causalité, de relation de cause à effet entre tel facteur et le taux. Un lien "statistiquement significatif" entre l'avortement non brucellique des vaches et la trace de l'infection par tel agent pathogène ne peut suffire, quelle que soit la force de la liaison à établir la responsabilité de cet agent dans l'avortement.

La répartition aléatoire d'une population en deux sous échantillons permet de traiter tous les problèmes statistiques contingents, en particulier comparaison et nombre de sujets nécessaires.

IV - DIFFERENTS EXEMPLES DE BIAIS

Le biais conduit à une différence systématique entre l'estimation fournie par l'échantillon et la valeur réelle de la population.

A - BIAIS VOLONTAIRE

Le choix peut être fait d'accepter un biais manifeste pour des raisons pratiques.

Exemple : l'estimation de la prévalence sérologique de l'infection du porc par le virus d'Aujeszky aurait pu être recherchée dans un échantillon représentatif de la population porcine ; en fait, la représentativité épidémiologique du phénomène concernant les reproducteurs (agents épidémiologiques majeurs) l'échantillonnage a été biaisé et a comporté surtout des reproducteurs.

Il est évident que les conclusions doivent tenir compte de ce biais.

B - BIAIS INVOLONTAIRE

Le biais involontaire est une cause d'erreur particulièrement banale mais dont les conséquences peuvent être très lourdes. Il provient d'une cause majeure : le choix des populations comporte au moins une étape où le hasard n'est pas intervenu. Si la cause est apparemment simple, s'en préserver nécessite un effort d'attention particulièrement soutenu. On peut toutefois simplifier (mais seulement en schématisant !) en précisant que toutes les étapes de choix des populations peuvent fournir l'occasion d'introduire un (ou plusieurs) biais.

* L'objectif épidémiologique peut avoir été mal défini au départ : on ne peut pas tirer de conclusions explicatives d'une situation purement descriptive. Cette évidence méritait bien d'être soulignée.

* La définition de la population objet de l'étude est facilement sujette au biais.

Le champ de l'étude peut avoir été mal déterminé. Exemple : si l'on compare les taux de rage canine chez les vaccinés, et les non vaccinés, les résultats sont empreints d'une "distorsion" certaine, car les deux sous-

populations ne sont pas soumises au même risque : les animaux vaccinés sont déterminés par le souci que leur maître ont de leur santé, et il s'agit plus volontiers alors de citadins peu exposés, tandis que les chiens ruraux pratiquement pas vaccinés, sont en fait les seuls véritablement exposés au risque.

Les critères d'exclusion peuvent être responsables de biais. Si pour une raison de sécurité on élimine de l'observation les animaux les plus vifs, les plus remuants, on introduit un biais : ces sujets, sélectionnés sur un critère social, sont plus exposés aux contacts, et (peut-être) courent globalement un risque plus élevé de contagion...

D'une façon générale, l'"auto-sélection" est une cause de biais. On désigne par là, l'ensemble des processus qui conduisent des sujets à se rassembler dans une population (région, comportements, activité professionnelle...)

Exemple : la clientèle d'une école vétérinaire est soumise à une autosélection motivée par la proximité, la renommée, les prix des consultations... les résultats ne peuvent pas être extrapolés à ceux de toute la population. De même, les clientèles du lundi, du mercredi ou du samedi ne sont pas les mêmes : les commerçants (disponibles le lundi) n'ont pas les mêmes animaux les mêmes demandes que des enseignants (disponibles le mercredi).

Exemple : un échantillon constitué d'éleveurs volontaires pour une enquête est manifestement biaisé par la motivation qui les amène à participer à cette observation ; cette motivation reflète en particulier le souci qu'ils ont de l'état sanitaire de leur élevage, au point de faire un effort pour connaître leur situation sur le point d'enquête qui leur est proposé ; il est donc tout à fait possible qu'ils appliquent des mesures spéciales dans leur élevage visant à les protéger contre certains problèmes, ce qui les écarte de la moyenne des élevages. Les résultats ne seront donc représentatifs que de ce groupe motivé.

Le choix de la donnée d'observation peut être source de biais, certes directement en raison d'une variabilité intervenant entre les différents observateurs selon les critères d'observation, et les conditions d'observation, mais aussi indirectement, en interférant sur le choix de la population. Nous avons vu que des contraintes pratiques insurmontables dans les conditions de travail pouvaient amener à tout d'abord déterminer la donnée d'observation, puis à reconstituer une population à partir de l'échantillon fourni par la récolte de la donnée (exemple : abattoir). Le biais est ici manifeste, dans la mesure où pour une étude descriptive, dans laquelle la définition de la population doit être prioritaire, on a tout d'abord défini la donnée d'observation qui conditionne en quelques sorte la définition de la population.

* Les "non-réponses" peuvent aussi apporter des biais. Les prélèvements perdus (tubes cassés) constituent simplement une perte d'information, toujours préjudiciable à l'homogénéité des résultats. Parfois, on peut par certains artifices de calcul, compenser le déséquilibre des groupes. En revanche, les non-réponses provenant d'une auto-sélection sont plus dangereuses : le refus pur et simple de répondre à un questionnaire est un mode d'auto-sélection, dont les motivations peuvent interférer avec l'objet de l'enquête (sentiment de culpabilité d'un éleveur par rapport à une situation sanitaire défavorable) ; de ce fait, certains sujets sont écartés de l'échantillon et biaisent les résultats.

La détection des biais constitue une partie importante de l'élaboration d'un protocole et de sa discussion. Il est illusoire d'espérer que l'on puisse dans tous les cas les éviter : les contraintes pratiques obligent à s'écarter des exigences statistiques. L'essentiel est de limiter les pièges grossiers, d'avoir connaissance du biais inévitable dans les conditions imposées et d'en tenir compte lors de l'exploitation des résultats.

°°°

CONCLUSION

Dans la plupart des travaux épidémiologiques, le choix des populations est un compromis entre l'objectif que s'est fixé le chercheur, les moyens dont il dispose, les difficultés qu'il rencontre, et les exigences statistiques dont les principes doivent être respectés faute de quoi les résultats ont une valeur strictement limitée à l'observation réalisée, et ne peuvent être extrapolés comme on le souhaite généralement. Cet exposé sur les principes généraux du choix des populations n'est qu'une ébauche ; elle se trouvera considérablement enrichie d'abord par les exposés qui vont suivre, portant sur des réalisations concrètes, et ensuite par la contribution de chacun d'entre nous, au sein d'une discussion que nous n'avions d'autre but que de susciter.

ANNEXE

Nombre de sujets nécessaires pour l'estimation d'une proportion.

Soit à estimer une proportion p (de sujets ou d'élevages infectés dans une population), à partir de l'observation du pourcentage observé p_o sur un échantillon de taille n. La quantité q_o est égale à $(1 - p_o)$. ε_α est la valeur de l'écart réduit lue dans une table d'écart réduit pour le risque α consenti.

L'intervalle de confiance i est :

$$i = \pm \varepsilon_\alpha \sqrt{\frac{p_o \, q_o}{n}}$$

Il représente la précision avec laquelle on connaîtra le pourcentage. Il est facile d'en déduire n :

$$n = \varepsilon_\alpha^2 \cdot \frac{p_o \, q_o}{i^2}$$

Par exemple, pour estimer à 1 p. cent près la valeur de p dont on suppose qu'elle doit être de l'ordre de 10 p. cent, le nombre de sujets nécessaires est de 3.600. La figure 3 permet de déterminer directement le nombre de sujets nécessaires pour estimer un pourcentage, en fonction de la valeur supposée de p, pour différentes précisions choisies, au risque $\alpha \approx 0{,}05$. Cette figure n'a de valeur que pour population homogène.

BIBLIOGRAPHIE

Leech F.B., Sellers K.C.- Statistical Epidemiology in veterinary science. 1 vol., Ch. Griffin and Co, Londres 1979, 158 p.

Rumeau-Rouquette C., Breart G., Padieu R.- Méthodes en épidémiologie. Flammarion, Paris, 1981, 306 p.

Schwabe C.W., Riemann H.P., Franti C.E.- Epidemiology in veterinary science. Lea et Febiger, Philadelphie, 1977, 303 p.

Schwartz D.- Méthodes statistiques à l'usage des médecins et des biologistes. Flammation, Paris, 1981, 318 p.