Institut d'Elevage et de Médecine
Vétérinaire des Pays Tropicaux
10, rue Pierre Curie
94704 MAISONS-ALFORT Cedex

Ecole Nationale Vétérinaire
d'Alfort
7, avenue du Général-de-Gaulle
94704 MAISONS-ALFORT Cedex

Institut National Agronomique
Paris-Grignon
16, rue Claude Bernard
75005 PARIS

Muséum National d'Histoire Naturelle
57, rue Cuvier
75005 PARIS

DIPLOME D'ETUDES SUPERIEURES SPECIALISEES
PRODUCTIONS ANIMALES EN REGIONS CHAUDES

SYNTHESE BIBLIOGRAPHIQUE

NOTIONS STATISTIQUES DE BASE DES ENQUETES EPIDEMIOLOGIQUES EN AFRIQUE

par

Jean-Joseph TYBURN

année universitaire 1993-1994

NOTIONS STATISTIQUES DE BASE DES ENQUÊTES EPIDEMIOLOGIQUES EN AFRIQUE

RESUME

Cette synthèse bibliographique a pour objet d'exposer quelques notions statistiques de base utiles pour les hommes de terrain impliqués dans les enquêtes en épidémiologie animale en Afrique. Les trois types d'enquêtes épidémiologiques (descriptive, analytique et évaluative) sont présentées. Dans la planification de celles-ci, une attention particulière devra être accordée au choix de la méthode d'échantillonnage car partir d'un bon échantillon est un fondement absolument indispensable pour l'obtention d'estimations aussi exactes et précises que possible. En Afrique, le choix de l'échantillon et de sa taille est un problème crucial car, plus qu'ailleurs, les considérations économiques y seront contraignantes. Les mesures des indicateurs épidémiologiques devront être également adaptés au contexte des enquêtes. Par conséquent, de bons échantillons sur lesquels des mesures de qualité seront réalisées fourniront des données fiables relatives à la situation d'une population vis-à-vis d'un phénomène étudié. L'étape suivante consistera à les interpréter en utilisant le raisonnement statistique (théories de l'estimation et des tests d'hypothèses) tout en ayant conscience des limites des enquêtes épidémiologiques. L'inférence statistique fournit un cadre rigoureux auquel le raisonnement de l'épidémiologiste doit se référer sans s'y enfermer. Une bonne maîtrise des notions statistiques utilisées en épidémiologie et une utilisation adéquate de ses connaissances permettront à l'épidémiologiste de tirer des conclusions judicieuses des faits observés et ainsi de pouvoir proposer des applications pratiques intéressantes.

Mots clés : Notions de statistique - Enquêtes - Epidémiologie - Afrique - Echantillonnage - Estimation - Mesures - Tests - Causalité.

INTRODUCTION

Cette synthèse bibliographique n'a aucunement la prétention d'être un recueil de formules encore moins un précis épidémiologique. Son ambition se situe à un niveau bien plus modeste : il s'agit d'exposer quelques notions de base de statistique utiles pour la réalisation des enquêtes en épidémiologie animale en Afrique. Ces notions devraient permettre aux personnes travaillant sur le terrain de mieux cerner les principes de l'épidémiologie et de pouvoir ainsi comprendre ses méthodes sans être pour autant des statisticiens de haute lignée. En effet, on ne peut concevoir l'épidémiologie sans la statistique, celle-ci fournissant les méthodes d'observation, d'analyse et d'interprétation de celle-là. Mais, derrière des formules, se cachent toujours des idées dont certaines sont d'une simplicité déconcertante au regard de la complexité apparente des formules les véhiculant.

L'épidémiologie étudie les phénomènes de santé au sein d'une population dans son milieu. On distingue trois phases dans la démarche épidémiologique. L'épidémiologie descriptive consiste à décrire la fréquence et la répartition d'un phénomène (maladie ou facteur de santé) à l'intérieur d'une population dans le temps et dans l'espace. Cette étape est importante afin d'essayer de comprendre les facteurs qui conditionnent et favorisent l'apparition du phénomène, cette partie étant l'épidémiologie analytique que certains préfèrent appeler épidémiologie explicative car c'est surtout l'aspect causal qui prime dans cette démarche. Une fois que le phénomène aura été décrit et compris dans ses causes, il conviendra de mettre en place des moyens d'actions et des dispositifs pour évaluer ces actions : c'est l'épidémiologie opérationnelle. L'efficacité de celle-ci sera fonction de la bonne réalisation des deux premières étapes, l'une descriptive et l'autre explicative. Aussi les enquêtes correspondant à ces deux domaines doivent-elles être conçues et réalisées avec la plus grande rigueur possible.

La statistique est l'un des outils permettant d'asseoir les enquêtes épidémiologiques et leur interprétation sur une base scientifique rigoureuse. Si l'épidémiologie descriptive fait appel à la statistique descriptive et dans une moindre mesure à la statistique inductive, celle-ci constitue le fondement même de l'épidémiologie explicative. En effet, la théorie de l'inférence statistique permet de passer de l'échantillon à la population d'origine, c'est-à-dire qu'à partir d'observations réalisées sur un (des) échantillon(s) on peut estimer avec un certain degré de confiance certaines caractéristiques de la population.Tout ceci constitue l'objet des tests d'hypothèses, de la théorie de l'estimation et des plans d'expériences, choses très utilisées dans les enquêtes, qu'elles soient descriptive, étiologique ou évaluative, étant donné que l'on ne travaille généralement que sur une partie de la population.

Pour une interprétation fiable des observations réalisées, il est certes nécessaire d'employer les bons tests d'une façon

adéquate, mais il faut aussi qu'au préalable les observations soient récoltées correctement grâce à des mesures de qualité. Par conséquent, dans les différentes phases de l'élaboration d'une enquête, chaque point devra être conçu d'une façon rigoureuse. On choisira les indicateurs épidémiologiques et la méthode d'échantillonnage en fonction des objectifs de l'enquête sans oublier le fait que le choix d'une méthode de mesure et l'utilisation inappropriée d'un type de test influenceront les conclusions et par la suite les décisions que les autorités compétentes prendront afin d'améliorer la situation de la population du point de vue de la santé. Nous rappelerons que la santé animale est "l'état de bien-être et d'équilibre entre un organisme et son milieu, lui permettant d'optimiser son potentiel génétique dans des conditions économiquement rentables et dénuées de tout danger ou inconvénient pour l'utilisateur ou le consommateur" (19).

I.LES ENQUÊTES ÉPIDÉMIOLOGIQUES

L'épidémiologie est donc divisée en trois domaines ayant chacun leurs objectifs :

->domaine descriptif : définir l'ampleur et la distribution des phénomènes

->domaine explicatif : dégager les facteurs étiologiques en vue d'orienter les méthodes d'action

->domaine évaluatif : mesurer l'efficacité des mesures d'amélioration.

Pour la réalisation de ces objectifs, des enquêtes devront être menées. Les observations recueillies dans chaque type d'enquête le sont dans le but de construire un schéma cohérent depuis l'évaluation d'un phénomène aux actions entreprises pour le contrôler. C'est ainsi que les données des enquêtes descriptives serviront de base à la conception des études explicatives, lesquelles pourront être remodelées en fonction des résultats de la phase opérationnelle. Si nous prenons l'exemple du contrôle de l'efficacité de la vaccination contre une maladie; devant le constat d'une inefficacité, plusieurs questions se posent. Entre autres, la vaccination a-t-elle été effectuée sur les sujets au bon moment, n'a-t-on pas oublié de prendre en compte un facteur important dans l'apparition de la maladie (une alimentation carencée est un facteur favorisant des maladies et peut expliquer des mauvais résultats de vaccination)? Tous ces éléments font partie des domaines descriptif ou explicatif, ce qui amènerait à les repenser.

1.1.Finalités des enquêtes en épidémiologie descriptive

Lorsque les structures vétérinaires d'un pays sont opérationnelles sur l'ensemble du territoire et les suivis techniques bien organisés, les registres sanitaires peuvent

fournir des informations sur les problèmes rencontrés par l'élevage dans une zone donnée. Aussi est-il nécessaire de consulter ces sources d'informatons, quand elles sont disponibles, avant d'entreprendre des enquêtes sur un sujet donné, ceci afin d'éviter toute étude redondante et de pouvoir le cas échéant proposer quelque chose de nouveau. Malheureusement, dans la plupart des pays africains il n'existe que très peu de structures de collecte de données (organismes de contrôle laitier, abattoirs, équarrissages, groupements techniques d'aide aux éleveurs) et le réseau vétérinaire est peu développé (4, 10). Dans ces conditions, des enquêtes en épidémiologie descriptive seront souvent utiles pour apprécier la répartition de la fréquence d'un phénomène de santé dans le temps et dans l'espace ainsi que son importance économique au sein de la population.
Les enquêtes en épidémiologie descriptive ont deux types de finalités :

->surveillance :
*contrôle sanitaire
*fréquence des maladies

->recherche :
*étiologique
*évaluative

1.1.1.Surveillance :

Contrôle : il s'agit de surveiller l'état sanitaire du cheptel et d'alerter les autorités compétentes en cas de pathologie déclarée. Cela exige des structures de terrain efficaces réparties sur l'ensemble du territoire et des structures de diagnostic capables d'identifier rapidement l'agent pathogène. Ces conditions n'existent pas dans beaucoup de pays africains. N'oublions pas que les applications pratiques de l'épidémiologie coûtent chères.

Surveillance des paramètres de santé d'une population : ceci en vue de la détermination des besoins et des modifications des caractéristiques de l'élevage. Pour cela, les mêmes structures que précédemment sont requises. Cependant, de tels réseaux d'épidémio-surveillance sont difficilement envisageables dans des pays connaissant des difficultés économiques sans le recours à des bailleurs de fonds.

1.1.2.Recherche :

Recherche étiologique : grâce à des études sur le terrain, l'épidémiologie descriptive peut concourir à la formulaton d' hypothèses expliquant les relations observées entre des phénomènes.

Recherche en évaluation : par des comparaisons chronologiques (avant-après) et géographiques (ici-ailleurs), il s'agit de mesurer la diffusion, l'effet et le coût d'une intervention.

Les enquêtes en épidémiologie descriptive peuvent être classées en deux catégories :
- les enquêtes transversales : où un phénomène est observé sur une période de temps limité.
- les enquêtes logitudinales : où la période d'observation est plus longue. Cela exige un suivi régulier des troupeaux, donc des moyens financiers plus important que pour les enquêtes précédentes.

1.2.Finalités des enquêtes en épidémiologie analytique

Nous entrons dans le champ de l'épidémiologie analytique. En fait, il faudrait mieux parler d'épidémiologie explicative; il s'agit, en effet, d'essayer de trouver la ou les cause(s) d'un phénomène. On parle encore d'épidémiologie "causale" ou d'épidémiologie étiologique. Ce domaine de l'épidémiologie peut être subdivisé en deux sous-domaines se complétant mutuellement. L'épidémiologie analytique sensu stricto s'occupe de répertorier et d'analyser tous les éléments explicatifs d'un phénomène tandis que l'épidémiologie synthétique s'attache à reconstruire le phénomène à partir des éléments explicatifs mis en évidence précédemment. Quelle que soit la partie considérée, deux attitudes sont possibles. La première consiste à étudier un phénomène en maîtrisant tous les facteurs extérieurs, c'est l'expérimentation et la seconde à observer les faits sans maîtrise de ces facteurs, c'est l'étude par enquête. Dès à présent, il convient de préciser que négliger la planification d'une enquête, sous prétexte de ne pourvoir maîtriser tous les facteurs, est une attitude qui conduit à l'obtention de résultats inexploitables du fait de l'existence d'une multitude de biais.

En épidémiologie explicative, deux grands types d'enquêtes peuvent être distingués : les unes prospectives, les autres rétrospectives.

1.2.1.Les enquêtes prospectives (enquêtes exposés/non exposés)

On sépare un échantillon d'une population en deux groupes selon l'exposition ou non à un facteur et on étudie la fréquence d'apparition du phénomène au sein de ces deux groupes. En milieu tropical, le suivi de troupeau est un exemple de ce type d'enquête (4). Le suivi de troupeau permet de réaliser des observations fiables car les données sont constatées et relevées par des enquêteurs préalablement formés. Ces enquêteurs peuvent également se rendre compte de la répartition saisonnière du phénomène et de l'impact économique de celui-ci. Par une visite régulière des troupeaux, il sera possible d'évaluer l'acceptation par les éleveurs des nouvelles techniques d'élevage ou des plans de lutte contre une maladie. Cependant, la collecte des données est longue, donc coûteuse. De plus, dans certains cas, ces enquêtes sont irréalisables. En effet, elles ne sont ni adaptées à l'étude des élevages transhumants ou nomades ni à celle d'une maladie dont l'apparition est brutale et dont la répartition régionale se modifie rapidement.

1.2.2.Les enquêtes rétrospectives (enquêtes cas/témoins)

Selon l'apparition de la maladie, deux groupes sont constitués. Sur une certaine période antérieure à cette apparition, la présence du facteur est recherchée au sein de ces deux groupes.

Ce type d'enquête est utilisé dans l'étude des foyers pour la recherche étiologique. Faites sur une période plus courte que les précédentes, elles sont a priori moins chères à réaliser. Cependant, les données sont souvent recueillies en faisant appel à la mémoire des personnes et par conséquent il se pose un problème de fiabilité des données.

La finalité de toutes ces enquêtes est de pouvoir tirer des conclusions quant aux facteurs responsables de l'apparition ou du maintien d'un phénomène de santé au sein d'une population. En pratique, il est souvent difficile d'établir une relation causale entre deux événements à l'aide d'une enquête. Aussi les épidémiologistes adoptent-ils une attitude plus pragmatique. Faute de pouvoir décrire l'étiologie du phénomène, ils se contentent de mettre en évidence des facteurs de risque, "facteur associé à l'augmentation de la probabilité d'apparition ou de développement d'un phénomène..."(19). Mettre en évidence une relation causale n'est possible que si l'épidémiologiste dispose de moyens importants qui ne lui seront accordés que si les conséquences prévisibles de cette découverte sont importantes.

1.3.Les enquêtes d'évaluation

Dans les lignes précédentes, nous avons insisté sur l'aspect financier des enquêtes. Cet aspect permet de justifier, à lui tout seul, les enquêtes d'évaluation. Celles-ci peuvent se faire avant le lancement d'un plan d'action (prise en compte de l'aspect coût-bénéfice d'une intervention) ou après les phases d'application des moyens d'action. Il s'agira alors de déterminer si l'intervention a été profitable au(x) groupe(s) de personnes visé(s) (notion d'utilité) ou si les objectifs définis préalablement ont été atteints (notion d'efficacité). Dans le cadre des études sérologiques, l'évaluation des méthodes de dépistage de la présence d'un germe dans une région donnée ou de celles de diagnostic est très importante. Elle permettra, en effet, de connaître la valeur du test, de la comparer à celle d'autres tests et de se rendre compte de son effet sur l'état de santé du cheptel.

Les enquêtes évaluatives sont du même type que celles de l'épidémiologie descriptive ou explicative. Par exemple, l'évaluation d'une intervention se fera à l'aide d'enquêtes avant-après et ici-ailleurs.

II.OBTENTION DES INDICATEURS ÉPIDÉMIOLOGIQUES ET PRINCIPES GÉNÉRAUX DE L'ÉTUDE PAR ENQUÊTE

Avant toute étude préalable, il est nécessaire de se donner des variables qui nous permettront de caractériser la situation de la population observée vis-à-vis du phénomène que l'on veut étudier. Ces variables sont des indicateurs épidémiologiques. Différents indicateurs peuvent être utilisés.

2.1.Les principaux indicateurs épidémiologiques

2.1.1.Fréquence d'un phénomène

Incidence : cet indicateur permet de se rendre compte d'un changement de la situation d'une population donnée vis-vis d'une maladie.

incidence = nombre de nouveaux cas/nombre d'individus de la population soumise au risque

Le nombre de cas nouveaux est rapporté à la durée d'observation. Celle-ci est souvent égale à une année et ainsi on parle d'incidence annuelle. Grâce à l'incidence, il est possible d'avoir une idée de la forme épidémiologique de la maladie. En effet, une épizootie se caractérise par le fait qu'elle affecte brutalement un grand nombre d'animaux à la fois dans un lieu donné : d'où une incidence forte. Tandis que celle-ci sera faible ou moyenne dans les formes sporadique ou enzootique pour lesquelles la maladie survient à intervalles irréguliers et touche un nombre réduit de sujets pour la première forme ou sévit en permanence dans une région sans qu'il y ait accroissement soudain du nombre de sujets atteints pour la seconde. La détermination de l'incidence exige un réseau dense de surveillance afin de signaler tous les nouveaux cas apparaissant dans le lieu considéré. En outre, cette détermination sera d'autant plus aisée que le phénomène sera facilement observable ou pour une maladie identifiable.

Prévalence : elle donne une image instantanée de la situation et permet de se faire une idée de l'impact du phénomène dans la population.

prévalence = nombre de cas(nouveaux et anciens)/nombre d'individus de la population soumise au risque

Comme pour l'incidence, le nombre de cas est rapporté à la durée d'observation : on parlera de prévalence annuelle si la durée choisie est l'année.

2.1.2.Taux d'atteinte de la population

Ces taux permettent d'apprécier la proportion de la

population touchée par le phénomène.

->taux de morbidité = nombre de malades/nombre d'individus de la population soumise au risque

Dans le cas d'une maladie très contagieuse, ce taux est élevé.

->taux de mortalité = nombre de morts/nombre d'individus de la population soumise au risque

Ce taux nous donne une indication de la gravité de la maladie.

->taux de létalité = nombre de morts/nombre d'individus malades

La détermination de tous ces indicateurs demande des données fiables. Celles-ci ne seront obtenues que si le protocole de collecte de données a été conçu et exécuté rigoureusement.

2.2.Principes généraux de l'étude par enquête

En statistique et en épidémiologie, une enquête est l'ensemble des opérations qui ont pour but de collecter, selon un plan rigoureux, des informations relatives à un groupe d'individus dans son milieu. Les individus sont appelés unités de base ou unités statistiques et l'ensemble des individus auxquels on s'intéresse est la population (3).

Si toutes les unités de la population sont observées individuellement, il s'agit d'enquête complète ou exhaustive, parfois appelée recensement. Si seule une partie de la population est observée, l'enquête est dite partielle ou par échantillonnage, le terme de sondage est aussi utilisé. Pour qu'une enquête puisse fournir des résultats exploitables, il est nécessaire que les points suivants soient résolus lors de sa planification.

->Délimitation du cadre de l'enquête :
*définition de l'unité d'observation : définition précise des animaux (espèce,race,sexe,âge,...) ou du groupe d'animaux (troupeau,élevage,région,...) sur lesquels porte l'enquête.
*définition de la population : ses caractéristiques devront être clairement exprimées.

->Définition des observations à réaliser : les observations seront choisies de façon à répondre précisément au but poursuivi. Les lieux, la date et la durée d'observation doivent être notifiés.

->Choix d'une méthode de collecte des données : en général, le recours à des enquêteurs préalablement formés sera préféré à l'envoi pur et simple de questionnaires. Cela évitera un trop grand nombre de non-réponses, une dépendance vis-à-vis des moyens d'acheminement du courrier et surtout permettra d'obtenir des données plus fiables.

En épidémiologie animale, l'enquête exhaustive est rare car d'une part, il n'est pas toujours possible d'observer tous les individus d'une population et d'autre part, le recensement demande des moyens financiers importants. Pour toutes ces raisons, on s'attachera à estimer par des enquêtes partielles certains paramètres de la population en choisissant une méthode d'échantillonnage et une taille d'échantillon adéquates.

III.L'ESTIMATION ET LES PRINCIPALES MÉTHODES D'ÉCHANTILLONNAGE

3.1.Estimation d'un paramètre d'une population

Le but des enquêtes partielles est de pouvoir tirer des conclusions sur certains paramètres et caractéristiques d'une population à partir des observations faites sur une partie de celle-ci. Aussi est-il important de savoir sous quelles conditions les informations relatives à l'échantillon peuvent être étendues à la population origine; ceci est l'objet de l'inférence statistique dont l'une des parties traite des problèmes de l'estimation.

3.1.1.Rappel de quelques notions de probabilité

Soient Ω une population d'individus notés $\omega(i)$ (i est un indice qui varie de 1 à n, n étant le nombre d'individus de la population) et x une caractéristique individuelle. Nous pouvons définir une application X de Ω dans $\mathbb{R}$ (ensemble des nombres réels) qui à chaque individu $\omega(i)$ associe sa caractéristique $X[\omega(i)]=x(i)$. Cette application est appelée variable aléatoire. Si par exemple, x est le poids d'un individu alors X est la variable aléatoire "poids" d'un individu de la population Ω. Par la suite, par commodité d'écriture, nous désignerons x(i) par x.

X est une variable aléatoire continue car la valeur x n'est pas contenue dans un ensemble fini ou dénombrable. En effet, x peut prendre en théorie n'importe quelle valeur réelle, x est donc contenu dans $\mathbb{R}$, ensemble infini. Si Y est la variable aléatoire "nombre d'animaux malades dans un élevage", c'est une variable discrète : le nombre de malades étant un entier naturel. A une variable aléatoire, il est possible d'associer une probabilité.

3.1.1.1.Définition d'une distribution de probabilité

->Cas d'une variable aléatoire discrète Y :

La probabilité associée à Y peut s'assimiler à la probabilité d'occurence de chaque valeur y que peut prendre Y. La donnée des valeurs de Y et des probabilités pour que Y prenne la valeur y correspondante (P(Y=y)) définit la distribution de probabilité (ou loi de probabilité) de la variable aléatoire Y.

La fonction de distribution F de Y est définie par :
F(y)=P(Y≤y), d'où F(y)=Σ P(Y=k) pour l'ensemble des k ≤y

->Cas d'une variable aléatoire continue X :

Reprenons l'exemple de la mesure X des poids des animaux d'une population. D'après ce que nous avons vu X est donc une variable aléatoire continue. Dans un premier temps, intéressons-nous aux poids des individus d'un échantillon de la population. Pour visualiser la distribution des poids, il est possible de construire un histogramme avec en abscisse les classes de poids et en ordonnée le pourcentage d'animaux appartenant à une classe donnée. A présent, passons à la population entière considérée comme infinie. Si nous construisons un histogramme avec un nombre fini de classes, nous allons perdre beaucoup d'informations fournies par les mesures relevées. Pour éviter cela, il convient d'augmenter fortement le nombre de classes et de diminuer énormément la taille de chaque classe; à la limite, pour être précis, il faudrait que le nombre de classes tende vers zéro. Ainsi, nous obtenons une série de petits rectangles (rectangles élémentaires quand leur nombre tend vers l'infini) et en joignant leur sommet par une ligne nous avons une courbe. Cette courbe limite est, en fait, la représentation graphique d'une fonction f, appelée fonction de densité de probabilité associée à la variable aléatoire X.
La fonction de répartition de F est toujours :

$$F(x)=P(X\leq x)=\int_{-\infty}^{x} f(t)\,dt$$

Le signe "∫" est le signe intégrale. Pratiquement, cela correspond à une sommation d'un grand nombre de termes (en théorie infini), termes étant infiniment petits. Ces termes sont les f(t)dt, appelés éléments de probabilité.

En effet, le terme f(t)dt s'apparente à une densité de probabilité. Considérons un accroissement Δt du poids, le pourcentage d'individus dont le poids est compris entre t et t+Δt peut être représenté par F(t+Δt)-F(t).Estimons cette différence quand Δt->0, nous savons que :

$$quand\ \Delta t\to 0,\ \frac{F(t+\Delta t)-F(t)}{\Delta t}\to f(t)\ \ car\ \lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t)}{\Delta t}\to f(t)$$

$$donc\ F(t+\Delta t)-F(t)\approx f(t)\,\Delta t\ quand\ \Delta t\to 0$$

$$ce\ qu'on\ écrit\ F(t+dt)-F(t)=df=f(t)\,dt$$

Par conséquent, la différence F(t+Δt)-F(t) représentant le pourcentage d'individus dont le poids est compris entre t et t+Δt pourra être considérée, quand Δt va tendre vers 0 et en première

approximation, comme la probabilité pour que le poids d'un individu tiré au hasard soit proche de la valeur t.
Les bornes de l'intervalle sont -∞ et x. En effet, il convient de "sommer" tous les éléments de probabilité correspondant à des poids t inférieurs ou égaux à x et X étant une variable pouvant prendre comme valeur un réel quelconque, il faut donc considérer tous les t de -∞ à x.

3.1.1.2.Paramètres d'une distribution de probabilité

Ces paramètres ont pour objet de caractériser de façon simple une distribution de probabilité. Différents types peuvent être distingués et pour chaque type il existe plusieurs paramètres. Nous ne nous intéresserons qu'à deux d'entre eux : un paramètre de position (la moyenne arithmétique) et un paramètre de dispersion (la variance).

La moyenne arithmétique :

C'est l'espérance mathématique de la variable aléatoire correspondante.

->Variable aléatoire discrète Y :
L'espérance mathématique E(Y) est :

$$E(Y) = \sum_{n} yP(Y=y)$$

où n est le nombre de valeurs possibles de Y.
Concrètement, cela signifie que si on réalise un très grand nombre de fois l'épreuve auquelle est associée la variable aléatoire Y, celle-ci prendra en moyenne la valeur E(Y). En outre, si une distribution de valeurs observées (ensemble des couples valeur observée-fréquence d'observation de cette valeur au sein du groupe étudié) est semblable à une distribution de probabilité de moyenne m, alors celle-ci donne l'ordre de grandeur des observations.

->Variable aléatoire continue X :
En considérant l'analogie probabilité/élément de probabilité, il est aisé de trouver la formule donnant l'espérance mathématique de X :

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

La variance :

Elle donne une idée de la dispersion des valeurs autour de la moyenne. Souvent, pour ce faire une idée de la dispersion, on détermine l'écart type qui est la racine carrée de la variance.

->Variable aléatoire discrète y :

$$VAR(Y) = E[(Y-m)^2] \quad où \quad m = E(Y)$$

->Variable aléatoire continue X :

$$VAR(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x)\, dx \quad m = E(X)$$

3.1.2.Importance de la distribution de Laplace-Gauss

Une distribution fondamentale en théorie des probabilité est celle de Laplace-Gauss, appelée aussi distribution normale réduite. Il s'agit d'une distribution dont la densité de probabilité f est de la forme :

$$f(x) = \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-x^2}{2}\right)$$

La courbe de la page 15 est la représentation graphique de cette fonction. Cette courbe présente la caractéristique d'avoir deux points d'inflexion symétriques dont les abscisses sont x=1 et x=-1. La moyenne d'une telle distribution est nulle (on dit qu'elle centrée) et sa variance est égale à 1. La fonction de répartition est :

$$\phi(x) = \int_{-\infty}^{x} t \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-t^2}{2}\right) dt$$

Pour un x donné, la valeur de la fonction de répartition se trouve dans la table de la loi normale réduite, celle-ci figurant dans n'importe quel bon ouvrage de statistique.

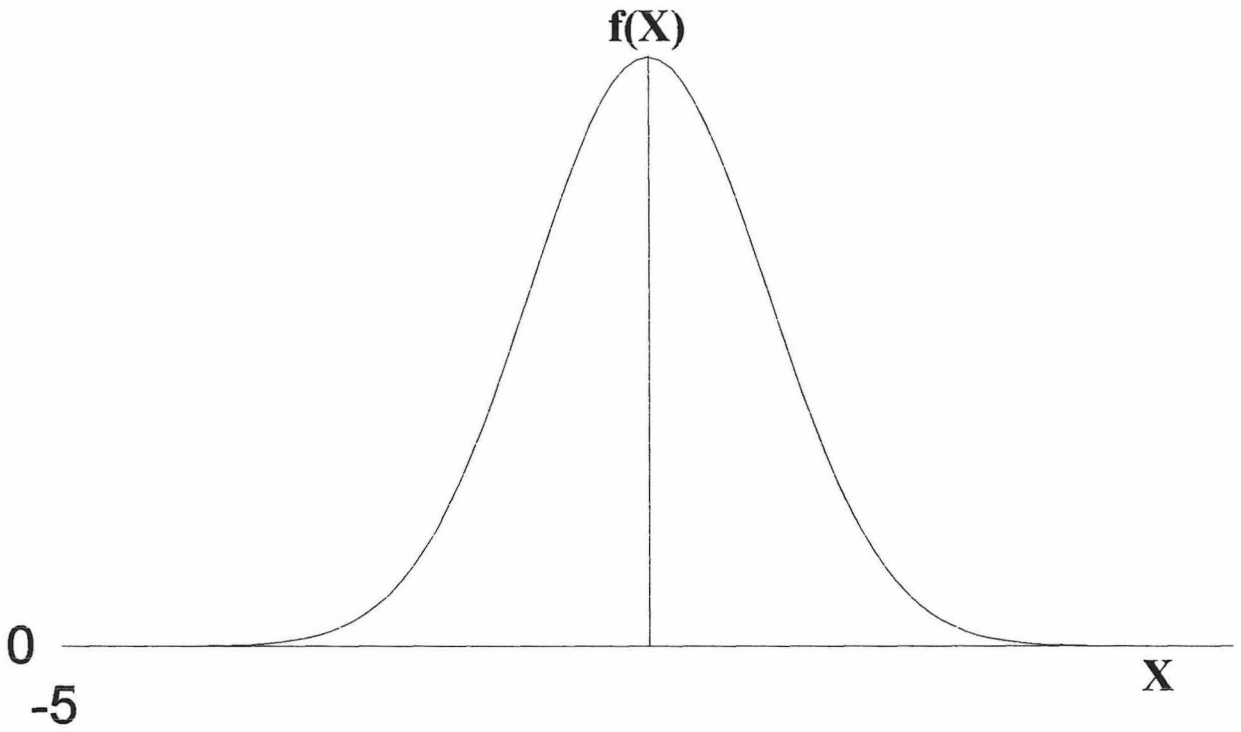
D'une manière générale, une distribution normale de paramètre m et σ, c'est-à-dire, de moyenne m et d'écart-type σ,est une distribution continue dont la densité de probabilité est :

$$f(x) = \frac{1}{\sigma\sqrt{(2\Pi)}} \exp\left[\frac{-1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$$

On dira que X est une variable aléatoire normale de moyenne m et d'écart-type σ, si sa distribution de probabilité est une loi normale de mêmes paramètres. Une propriété importante des lois normales est que si X est une variable aléatoire de moyenne m et d'écart-type σ, alors toute fonction linéaire Y=a+b*X est une variable normale de paramètres :

$$m(Y) = a + b \times m \text{ et } \sigma(y) = |b| \times \sigma(x).$$

REPRESENTATION GRAPHIQUE DE LA LOI NORMALE REDUITE



Par conséquent, si $U=(X-m)/\sigma$, où X est une variable aléatoire normale de paramètres m et σ, alors U est une variable normale dont la moyenne est nulle et l'écart-type égal à un : c'est donc une variable normale réduite. De plus, nous avons :

$$P(a\leq X\leq b)=P(a-m\leq X\leq b-m)=P\{(a-m)/\sigma\leq X\leq(b-m)/\sigma\}=P(A\leq X\leq B)$$
$$\text{où } A=(a-m)/\sigma \text{ et } B=(b-m)/\sigma$$

$$P(a\leq X\leq b)=\phi(b)-\phi(a)$$

Aussi tout calcul fait sur une variable normale se ramène-t-il à un calcul sur la variable normale réduite.

La loi normale occupe une place primordiale en théorie des probabilités et dans les applications pratiques, notamment en biostatistique. L'importance de cette loi vient du théorème central limite stipulant que toute somme de n variables aléatoires indépendantes, même de distribution différente, tend sous certaines conditions vers une variable normale lorsque n tend vers l'infini. En fait, la seule condition nécessaire pour pourvoir utiliser cette propriété est qu'aucune variable ou groupe de variables ne prédomine par rapport aux autres. Concrètement, cela signifie que si l'on s'intéresse à un phénomène dépendant d'un nombre élevé de facteurs indépendants (en théorie d'une infinité de facteurs) ayant des effets du même ordre de grandeur, alors sa distribution est normale.

La loi normale sera également beaucoup utilisée, sinon sous-entendue, dans la théorie de l'estimation. En pratique, chaque fois que l'on est amené à estimer un paramètre d'une population à partir d'un échantillon, il faudra toujours se demander si la distribution d'échantillonnage du paramètre est normale ou pas.

3.1.3.Distribution d'échantillonnage

A un paramètre γ d'une population, il est possible d'associer une série de valeurs g, g', g" ...observées à partir d'autant échantillons de même effectif et prelevés indépendamment les uns des autres dans des conditions identiques. Ces valeurs observées peuvent être considérées comme les valeurs d'une variable aléatoire G qui dépend des individus présents dans chaque échantillon.

G : fonction de (X1,X2,...,Xn) n étant l'effectif d'échantillon (par exemple : moyenne de la taille des individus d'un échantillon)
Xi: variable aléatoire associée au ième individu tiré de l'échantillon (ex : taille d'un individu)
g : valeur observée de la variable aléatoire G dans un échantillon contenant n individus
xi : valeur observée de la variable aléatoire Xi correspond à une caractéristique individuelle (dans notre exemple, la taille), le paramètre γ dépendant de cette caractéristique (il représente ici la taille moyenne de la population d'origine).

La distribution de G s'appelle la distribution d'échantillonnage de γ. Comme toute distribution, elle a une moyenne et une variance :

$$m(G) = E(G) \quad et \quad VAR(G) = E[(G - m(G))^2]$$

$\sigma(G)$, racine carrée de VAR(G), est appelée erreur standard. Il est à noter que les valeurs de la moyenne et de la variance de G dépendent à la fois de la population d'origine et du mode d'échantillonnage. Souvent, il est possible de prévoir la forme de la distribution d'échantillonnage d'un paramètre connaissant la distribution de la population et la manière dont les échantillons ont été réalisés. En reprenant l'exemple précédent, nous pouvons affirmer que la distribution d'échantillonnage de la moyennne de la taille des individus d'une population, pour un échantillonnage aléatoire et simple (cf définition dans la partie "principales méthodes d'échantillonnage"), est normale si la distribution de la taille dans cette population l'est également. Dans le cas de la moyenne, nous pouvons ajouter que même si la distribution dans la population parente n'est pas normale, celle d'échantillonnage de la moyenne l'est asymptotiquement, c'est-à-dire lorsque l'effectif des échantillons tend vers l'infini (en pratique très grand).

Paramètres de la distribution d'échantillonnage de la moyenne X :

$$E(\overline{X}) = m \quad et \quad s(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

où m est la moyenne de la population
σ est l'écart-type de la population
n : l'effectif d'échantillon et s l'écart-type de la distribution d'échantillonnage, E sa moyenne.

Distribution d'échantillonnage d'un pourcentage :

Pour fixer les idées, supposons que l'on désire étudier la prévalence P d'une maladie au sein d'une population.
N : taille de la population de départ
P : fréquence des malades dans la population, d'où 1-P : fréquence des non-malades.
Nous tirons différents échantillons aléatoires et simples de la population. Soient n l'effectif d'échantillonnage (n étant négligeable devant N) et p la valeur trouvée de la fréquence de la maladie dans un échantillon. Les valeurs p de la prévalence observée pour différents échantillons se distribuent autour de la valeur vraie de la prévalence (P) et l'écart-type de cette distribution est $\sigma = \sqrt{[P*(1-P)/n]}$.

Forme de cette distribution

Cet échantillonnage aléatoire et simple peut être comparé au tirage au sort avec remise de boules de couleur soit noire soit blanche dans une boîte, celle-ci contenant une proportion de P boules blanches et de 1-P boules noires. A la suite de

chaque tirage, la boule tirée est remise dans la boîte après avoir noté sa couleur. En effet, comme n est négligeable devant N, la détection de la maladie sur n individus ne modifiera pas la fréquence de la maladie chez les N-n autres individus, la prévalence de la maladie pourra donc être considérée comme constante au fil du tirage au sort des individus.

La probabilité d'avoir k individus malades dans un échantillon d'effectif n est égale à celle de tirer k boules blanches au terme de n tirages avec remise. Or l'ensemble des valeurs de k associées à leur probabilité suit une loi binomiale B de paramètres E(B) = n*P et VAR(B) = n*P*(1-P)

$Prob(B=k) = C_n^k \, P^k (1-P)^{n-k}$ où C_n^k : *nombre de combinaisons de k parmi*

Une combinaison de k parmi n étant un sous-ensemble de k éléments d'un ensemble en contenant n.
En divisant k par n, nous obtenons la valeur de la prévalence observée. Aussi la distribution des valeurs de p suit-elle une loi binomiale. Or tous les calculs faits sur une loi binomiale de paramètre n*P et n*P*(1-P) peuvent se ramener à ceux réalisés sur une loi normale lorsque n*P > 5 et n*P*(1-P)> 5. Par conséquent,si n est assez grand et la prévalence vraie de la maladie suffisamment éloignée de 0 comme de 1, la distribution d'échantillonnage de la prévalence observée peut être considérée comme normale et a pour paramètres m = P et $\sigma = \sqrt{[P*(1-P)/n]}$.

3.1.4.L'estimation

Le but poursuivi des enquêtes épidémiologiques est d'évaluer des indicateurs au sein d'une population qui puissent permettre de caractériser sa situation vis-à-vis du phénomène étudié et offrir ainsi des renseignements de base sur lesquels pourrait s'établir un plan d'intervention. Cependant, en général, seule une partie de la population est observée et les observations varient d'un échantillon à l'autre. Aussi est-il impératif de savoir dans quelle mesure les paramètres d'une population sont estimables et de déterminer la précision d'une éventuelle estimation.

Soit une population dont la distribution de probabilité dépend d'un paramètre T. Un échantillon de taille n est tiré. Un estimateur du paramètre T est toute fonction G des valeurs observées qui permet d'estimer T : G(X1,...,Xn) et une estimation de T est une valeur numérique prise par G pour l'échantillon considéré. Par exemple, la moyenne de la taille des individus d'un échantillon est une estimation de la moyenne de la taille des individus de la population d'origine. Mais rien ne nous empêche de prendre un autre estimateur de ce paramètre. Aussi face à une multitude d'estimateurs possible, convient-il de faire le bon choix en fonction des qualités requise. Deux qualités sont souvent recherchées : l'exactitude et la précision d'un estimateur.

<u>Absence d'inexactitude (=erreur systématique=biais)</u>

Un estimateur G d'un paramètre T est non biaisé si E(G)=T. Cela signifie que la moyenne des estimations données par plusieurs échantillons est égale à T. Dans de telles conditions, les échantillons sont dits représentatifs.

Précision d'un estimateur

La valeur de E[(G-T)*(G-T)] nous donne une idée de la précision de l'estimateur G de T. Si G est non biaisé alors E(G)=T et donc la valeur précédente est celle de la variance de l'estimateur. Plus la précision est grande, plus la valeur de E[(G-T)*(G-T)] est faible. Par conséquent, pour une précision bonne les différentes estimations de T obtenues par différents échantillons ne sont pas trop éloignées les unes des autres. Un estimateur, même sans biais, qui serait de précision médiocre, serait de peu d'intérêt pour l'estimation du paramètre. En effet, nul ne peut savoir dans pareil cas si une estimation faite est exacte (c'est-à-dire égale à la valeur du paramètre) et si en prenant un autre échantillon l'estimation serait très différente de la première. Si la précision est bonne, nous savons que les estimations obtenues à partir de différents échantillons ne seront guère éloignées les unes des autres et il suffira d'estimer le biais, dans le cas où l'estimateur est biaisé, pour avoir une idée de la valeur du paramètre T.

Il est possible de montrer que pour tout paramètre T, il existe une valeur minimale de E[(G-T)*(G-T)]. En général, dans les problèmes d'estimation on recherchera l'estimateur pour lequel la valeur de E[(G-T)*(G-T)] s'approche le plus du minimum.

Intervalle de confiance :

Toutes ces considérations nous montrent que la seule connaissance de la valeur estimée d'un paramètre est peu utile si elle n'est pas accompagnée d'autres indications. Ces indications sont soit la donnée de l'erreur-standard de la distribution d'échantillonnage de l'estimateur soit celle d'un intervalle autour de la valeur estimée pour lequel on peut dire qu'il y a un certain pourcentage de chances qu'il contienne la vraie valeur du paramètre. Cet intervalle est appelé intervalle de confiance.

Soient G1 et G2 les limites d'un intervalle de confiance. Celui-ci est tel que :

$$P(G1 \leq T \leq G2) = 1-\alpha \quad 1-\alpha : \textit{niveau de confiance}$$

G1,G2 : limites de confiance et T : paramètre à estimer.

Par exemple, si $1-\alpha=0,95$ alors l'intervalle[G1,G2] a 95% chances de contenir la vraie valeur du paramètre T. Attention, dire que T a 95% de chances d'être contenu dans cet intervalle est un non-sens, T ayant une valeur précise est soit à l'intérieur de l'intervalle soit à l'extérieur. α est appelé le risque.

Souvent, les limites G1 et G2 sont choisies de façon à ce que le risque soit divisé en deux parties égales, c'est-à-dire :

$$P(T<G1) = P(T>G2) = \frac{\alpha}{2}$$

En général, α =5% ou 1% ou 0,1%. En biologie, un risque de 5% est souvent pris.

Soit G un estimateur sans biais de T. Supposons que la distribution d'échantillonnage de G est normale ou asymptotiquement normale et que l'échantillonnage est aléatoire et simple. A partir d'un échantillon, nous avons une valeur observée g de G, cette valeur est donc une estimation de T. Exprimons les limites de confiance g1,g2, par rapport à la valeur observée g en prenant un risque α réparti en deux parties égales de part et d'autre de l'intervalle. Nous avons donc : $P(T < G1) = P(T > G2) = \alpha/2$.

Posons G1 = g-d1 et G2 = g+d2. Le problème revient à déterminer les expressions de d1 et de d2.

$$P(T < g-d1) = P(T > g+d2) = \alpha/2$$

$$\text{soit } P(g-T > d1) = P(T-g > d2) = \alpha/2$$

comme G est sans biais E(G) = T, les égalités ci-dessus peuvent être transformées en faisant apparaître la variable normale réduite U :

$U = (g-E(G))/\sigma(G) = (g-T)/\sigma(G)$ où $\sigma(G)$ est l'erreur-standard.

$$\text{D'où } P(U > d1/\sigma(G)) = P(U > d2/\sigma(G)) = \alpha/2$$

$$\text{soit } P(U \leq d1/\sigma(G)) = P(U \leq d2/\sigma(G)) = 1-\alpha/2$$

c'est-à-dire $\phi(d1/\sigma(G)) = \phi(d2/\sigma(G)) = 1-\alpha/2$, ϕ étant la fonction de répartition de la loi normale réduite.
Ainsi $d1/\sigma(G) = d2/\sigma(G) = u(1-\alpha/2)$ où $u(1-\alpha/2)$ est tel que :

$$\int_{-\infty}^{u(1-\alpha/2)} \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-x^2}{2}\right) dx = \phi(1-\alpha/2) = 1-\alpha/2$$

Ainsi, l'intervalle de confiance de T autour d'une valeur observée g est de la forme : $g \pm u(1-\alpha/2)\times\sigma(G)$. Aussi l'estimation et l'intervalle de confiance sont-ils deux notions liées entre elles : plus un estimateur sera exact et précis, plus l'intervalle de confiance sera étroit.

Dans les publications, les résultats sont souvent présentés, soit comme suit g ±10% par exemple, soit comme cela g ± x à 95% où p est une estimation d'un paramètre. La première présentation est imprécise : le degré de confiance n'est pas donné. S'agit-il de 95%, de 99%, de 99,9% ou d'un autre pourcentage? Comme en

biologie, 95% est souvent choisi comme degré de confiance, on pourrait penser que c'est le cas ici, mais rien ne permet de l'affirmer.

Intéressons-nous à la seconde présentation. Le niveau de confiance est indiqué et on peut donc affirmer qu'il y a 95% de chances que l'intervalle [g + x,g - x] contienne la vraie valeur T du paramètre.

Comme nous avons pu le constater la méthode d'échantillonnage aura une grande influence sur les paramètres et la forme des distributions d'échantillonnage. Cela devra être pris en compte lors de la planification d'une enquête afin de choisir la méthode d'échantillonnage la plus appropriée.

3.2.Les principales méthodes d'échantillonnage

3.2.1.Les méthodes d'échantillonnage

Echantillonnage aléatoire et simple(=complètement aléatoire)

Un échantillonnage aléatoire est un échantillonnage dans lequel tous les individus de la population ont la même probabilité de faire partie de l'échantillon; il est dit simple si les tirages successifs des individus sont indépendants.

Ce mode d'échantillonnage fournit des échantillons représentatifs de la population et évite ainsi les biais de sélection de l'échantillon, c'est-à-dire les erreurs commises dans l'estimation d'un paramètre du fait d'un mauvais choix de l'échantillon (13). En outre, comme nous l'avons vu dans la partie précédente, la théorie de l'estimation fait souvent référence à ce type d'échantillonnage en tant que condition requise à une détermination facile des caractéristiques de la distribution d'échantillonnage d'un estimateur. Dans le cas d'une population parente normale, si l'échantillonnage est complètement aléatoire alors la distribution d'échantillonnage de l'estimateur est normale. La détermination de l'intervalle de confiance d'une estimation est aisée sous de telles conditions, ce qui n'est pas toujours le cas pour les autres méthodes d'échantillonnage. Par conséquent, des échantillons aléatoires et simples peuvent nous permettre d'estimer l'erreur commise par rapport à la vraie valeur du paramètre.

Cependant, pour mettre au point l'échantillonnage aléatoire et simple il est nécessaire de disposer d'une base de sondage fiable, c'est-à-dire d'une liste exhaustive de toutes les unités de base avec leurs caractéristiques. Il faut donc connaître la taille de la population et pouvoir identifier précisément les individus. Or, dans de nombreux pays africains, des informations fiables sur le nombre et le type de troupeaux d'une région donnée sont rares. De plus, la réalisation de ce type d'échantillonnage est coûteuse si l'enquête dure longtemps ou si les unités tirées sont éloignées les unes des autres. Pour toutes ces raisons, ce mode d'échantillonnage sera peu employé dans des études en Afrique.

Echantillonnage systématique(pseudo-aléatoire)

Il consiste à choisir une unité et ensuite à partir de celle-ci, de façon régulière, les autres unités qui vont constituer l'échantillon. Prenons l'exemple d'une enquête sur les habitations d'un village fictif dans lequel celles-ci sont alignées les unes derrière les autres. Un échantillon systématique pourra être constitué de la façon suivante, on choisit d'abord une habitation et ensuite celles qui la suivent et qui sont séparées d'elle par un nombre multiple de k d'habitations.

Pour un même nombre d'observations, cet échantillonnage donne des estimations moins dispersées que l'échantillonnage précédent. En outre, il est plus facile à réaliser lorsque la taille de la population est ipmportante. Cependant, si la population présente une périodicité structurelle (par exemple toutes les habitations séparées de n habitations de celle du choix initial n'ont qu'une fenêtre) semblable à celle du choix des unités prélevées, l'échantillon risque de ne plus être représentatif et la détermination de l'erreur standard est plus difficile. IL est donc impératif de posséder des informations sur la population même si la taille de celle -ci n'est pas connue.

Echantillonnage stratifié

Ici, on subdivise la population en plusieurs parties, les strates. Dans chacune de celles-ci, on choisit les unités qui constitueront l'échantillon, ce choix pourra être fait soit de manière aléatoire soit de manière systématique. Ce système de stratification est variable selon les objectifs poursuivis; c'est ainsi que les strates peuvent correspondre aux régions, aux systèmes de production (extensif/intensif,moderne/traditionnel), à la taille du troupeau, à la race, à l'âge, au sexe etc...

L'échantillonnage stratifié est utile lorsque la population est très hétérogène. La stratification permet de s'assurer que les différentes composantes de cette population seront représentées au sein de l'échantillon. L'estimation d'un paramètre de la population sera d'une précision, pour un même effectif d'échantillon, supérieure (les estimations seront moins dispersées) à celle obtenue lors d'un échantillonnage complètement aléatoire. De plus, on concentre les moyens dont on dispose sur les parties de la population intéressantes.

Echantillonnage à plusieurs degrés

Le principe est de considérer plusieurs types d'unités de base correspondant à autant de niveaux d'échantillonnage. Par exemple, à l'échelle du pays ,le premier niveau pourrait être la région (choix des régions d'enquête), le deuxième celui des troupeaux (quels sont ceux qui vont être étudiés) et le troisième les animaux. Pour constituer l'échantillon, on opère aléatoirement ou systématiquement à chacun de ces niveaux d'échantillonnage. Ce type d'échantillonnage demande moins de déplacements de l'enquêteur et permet donc de réduire les frais de transport par rapport à ceux de l'échantillonnage aléatoire.

Un cas particulier intéressant d'échantillonnage à deux

degrés est celui en grappes. Ici, on étudie exhaustivement les unités de base choisies au premier degré qui correspond à un simple tirage au sort. Son intérêt réside dans le fait de pouvoir étudier tous les animaux d'un même troupeau si celui-ci a été choisi sans pour autant disposer d'une identification de tous les animaux d'une région donnée. Cependant, dans le cas de l'étude de la prévalence d'une maladie, il conviendra de s'intéresser aux différences inter-grappes afin de juger de son impact réel et ce d'autant plus qu'il s'agit d'une maladie infectieuse. En effet, la prévalence peut être tout à fait différente car au sein d'une grappe contaminée (la grappe étant l'unité sélectionnée au premier niveau) il y a de fortes chances que tous les individus soient touchés par l'affection et loin de cette grappe, on peut en trouver une autre saine.

Méthode des quotas

Cette méthode est surtout utilisée dans les sondages d'opinion. Il s'agit, après avoir défini des catégories d'individus, de les remplir avec un nombre fixé d'individus. Ce nombre fixé pour chaque catégorie et la méthode de choix des individus appartenant à une catégorie donnée seront déterminés de façon à ce que l'échantillon ait une composition aussi semblable que possible à celle de la population selon les critères retenus pour former les catégories.

Il y a donc de multiples méthodes d'échantillonnage. Le choix de l'une d'entre elles dépendra des objectifs visés par l'enquête, des critères de qualités requis pour les estimations et aussi des moyens financiers disponibles. En fait, l'enquête doit fournir des informations les plus exactes et précises possibles à un moindre coût. Pour ce faire, il est essentiel que soient déterminés les degrés de qualités voulus des résultats lors de la planification de l'enquête. En effet, étant donnée une méthode d'échantillonnage, des estimations de meilleurs qualités ne peuvent être obtenues avec des effectifs d'échantillonnage réduits et accroître la taille de l'échantillon coûte cher.

3.2.2.La taille de l'échantillon

La précision des estimations est reliée à la taille de l'échantillon et au caractère d'homogénéité de la population parente. En effet, la dispersion des estimations sera d'autant plus faible que l'effectif de l'échantillon sera important et que la population sera moins hétérogène. Dans les conditions pratiques, on se demandera fréquemment, en vue de l'estimation d'un paramètre, quelle est la taille de l'échantillon adéquate pour avoir des degrés d'exactitude ou de précision donnés.

Envisageons le cas de l'estimation de la prévalence d'une affection au sein d'une population (5). Nous supposerons que des études préalables permettent de prévoir que cette prévalence n'est proche ni de 0 ni de 1. Calculons le nombre de sujets nécessaires pour avoir une précision donnée à un niveau de confiance de 95%.

L'estimateur p (fréquence des malades au sein d'un échantillon) est un estimateur sans biais de la prévalence vraie P. Comme nous l'avons vu dans la deuxième partie, les différentes estimations de la prévalence pour des effectifs d'échantillon suffisamment grands se distribuent selon une loi normale autour de P; nous avons effectivement tout intérêt à considérer de grands échantillons car plus leur taille est importante plus les estimations sont proches du paramètre à estimer. L'intervalle de confiance de P est donc de la forme :

$P \pm u(1-\alpha) \times \sigma$ $\sigma = \sqrt{[P \times (1-P)/n]}$ (erreur-standard) et n : effectif d'un échantillon

Pour un niveau de confiance de $1-\alpha = 0,95$, $u(1-\alpha)$ est égale à 1,96 (cf table de la loi normale réduite).

$e = u(1-\alpha) \times \sigma = 1,96 \times \sqrt{[P \times (1-P)/n]}$ d'où :

$$n = [\frac{u(1-\alpha)}{e}]^2 P(1-P)$$

Cette égalité nous montre bien que plus on veut de la précision, c'est-à-dire e faible, plus il faudra augmenter l'effectif de l'échantillon.

Ainsi, lors du choix d'une méthode d'échantillonnage, deux notions doivent être présentes à l'esprit de tout planificateur, à savoir la précision et l'exactitude des estimations. La précision va dépendre de la taille de l'échantillon et de la stratégie d'échantillonnage (échantillonnage aléatoire et simple, stratifié,...)et l'exactitude reflètera la représentativité de l'échantillon étudié. Le principal biais des enquêtes mal conçues ou réalisées est celui de sélection (13). En effet, un mauvais échantillon ne donnera pas nécessairement des résultats aberrants mais conduira à des erreurs systématiques. L'une des principales difficultés des enquêtes épidémiologiques est de les éviter ou tout au moins de les contrôler.

IV.MESURES ET QUALITÉS DE LA MESURE :

Après le choix de la méthode d'échantillonnage, il convient maintenant de passer à l'étape de mesure des indicateurs épidémiologiques. Les mesures doivent répondre à des critères de qualités en rapport avec le contexte et les objectifs de l'enquête.

4.1.Qualités de la mesure

Elles sont de deux types : les unes intrinsèques à la méthode de mesure, les autres dépendant du contexte de l'enquête, notamment de la prévalence du phénomène étudié dans la région considérée.

4.1.1.Qualités intrinsèques

Nous prendrons pour les définir le cas d'un test destiné à révéler les individus affectés par la maladie.

La sensibilité : Se = P(T+/M+)

C'est la probabilité que le test soit positif (T+) sachant que l'animal est affecté (M+). Concrètement, elle correspond à la proportion de vrais positifs (vp : animaux affectés ayant un test positif) par rapport au nombre total d'animaux affectés. Elle indique donc l'aptitude du test à dépister les malades.

La spécificité : Sp = P(T-/M-)

C'est la probabilité que le test soit négatif (T-) sachant que l'individu n'est pas affecté (M-). Autrement exprimée, c'est le rapport des vrais négatifs (vn : animaux non affectés dont le test est négatif) au nombre total d'animaux indemnes. Elle représente donc l'aptitude du test à ne révéler que les individus réellement affectés.

Sensibilité et spécifité sont deux qualités antagonistes : augmenter l'une aura pour conséquence de diminuer l'autre. Par conséquent, on recherchera une qualité plutôt que l'autre selon les objectifs fixés au départ. Prenons le cas du dépistage d'une maladie infectieuse par un test sérologique : le principe est de révéler la présence d'anticorps dans le sang, témoins du passage de l'agent infectieux dans l'organisme (cet agent pouvant être encore présent ou avoir été éliminé). Le test sérologique rend d'utiles services dans les campagnes d'éradication d'une maladie. Considérons deux situations relatives à une telle campagne : la première correspondant au dépistage des malades en début de la campagne et la seconde au même dépistage mais en fin de campagne. Dans la première situation, il est impératif de détecter tous les animaux infectés même si parfois le test donne des résultats faussement positifs. Par conséquent, il convient donc de choisir un test de la meilleure sensibilité possible sans toutefois négliger sa spécificité. A la fin de la campagne, si celle-ci a été efficace, le nombre d'individus reéllement infectés est faible. Bien-sûr il est important de détecter encore les infectés mais on exigera du test qu'il ait en plus une très bonne spécificité. En effet, considérer un individu positif alors qu'il ne l'est pas entraîne des mesures supplémentaires qui coûtent chères, par exemple prolongation de la campagne, changement de stratégie de lutte. Aussi à la fin d'une campagne d'éradication, sera-t-il important de choisir un test de très bonne spécificité et d'une bonne sensibilité (18).

En plus de la sensibilité et de la spécificité d'une méthode, son exactitude et sa précision devront être prises en compte. L'exactitude d'une méthode de mesure indique la faiblesse de la moyenne des différences entre les différentes mesures d'une grandeur et la valeur réelle de celle-ci (plus la différence est faible, plus la méthode est exacte) tandis que la précision

reflète la dispersion des mesures (plus celle-ci est réduite, meilleure est la précision).

4.1.2.Qualités dépendant de la prévalence

Lorsqu'un test est utilisé pour déterminer la présence d'un germe dans une population, il est important de connaître les valeurs prédictives positive et négative du test, c'est-à-dire la probabilité qu'un individu soit effectivement infecté sachant que le test est positif (P(M+/T+)) et la probabilité qu'un individu soit réellement indemne sachant que le test est négatif (P(M-/T-)). Ces probabilités dépendent non seulement de la sensibilité et de la spécificité du test mais aussi de la prévalence de l'infection dans la population étudiée.

Soient Pt : prévalence de l'infection donnée par le test (proportion de sujets positifs par rapport à la population totale) et P : la prévalence vraie. Pt se décompose en vrais positifs, VP, et faux positifs, FP.

Pt = VP + FP avec VP = nombre de vrais positifs/nombre total d'individus

et FP = nombre de faux positifs/nombre total d'individus

1-P est la fréquence des individus indemnes dans la population;

1-P = FP + VN avec VN = nombre de vrais négatifs/nombre total d'individus

d'où FP = 1-P-VN

D'autre part, comme nous l'avons précisé ci-dessus la spécificité Sp peut s'écrire : Sp = VN/(VN+FP)

d'où VN = FP×Sp/(1-Sp)

Nous avons donc : FP = 1-P-[FP×Sp/(1-Sp)]

d'où FP = (1-P)×(1-Sp)

La sensibilité Se est égale à VP/(VP+FN). De là, nous pouvons tirer VP : VP = Se×(VP+FN) = Se×P.

Par conséquent, nous avons : Pt = VP+FP = Se×P+(1-P)×(1-Sp).

Ainsi, la prévalence vraie P est égale à :

$$P=\frac{Pt+Sp-1}{Se+Sp-1}$$

Il est donc possible à partir de la prévalence donnée par le test de calculer la prévalence vraie.

Pour les tests sérologiques, il est important de souligner qu'un résultat sérologique en lui-même n'apporte aucun élément de certitude, seule la mise en évidence de l'agent infectieux permet d'affirmer sa présence dans le groupe d'animaux étudié

dans une région donnée (2). En effet, les anticorps persistant plus ou moins longtemps dans l'organisme, certains restant même présents pendant toute la vie de l'animal, leur présence ne signifie en aucun cas que le germe est responsable des troubles observés. L'animal peut très bien avoir été infecté par l'agent et s'en être débarassé sans garder de séquelles.

4.2.Qualités opérationnelles

Les enquêtes se doivent d'être efficaces. A ce titre, les mesures doivent être facilement réalisables sur le terrain. Si besoin est, la faisabilité de la mesure pourra être assurée par une formation préalable des enquêteurs. De plus, il faut que les éleveurs acceptent les enquêtes et les mesures éventuelles sur leurs animaux. Il apparaît donc essentiel que cette question d'applicabilité de la méthode soit envisagée lors du la mise au point du protocole, une pré-enquête pourrait éviter les mauvaises surprises. Il convient également de s'assurer de la qualité des instruments de mesure. On veillera à leur répétabilité, propriété de donner le même résultat lorsque la mesure est répétée dans des conditions identiques par la même personne et à leur reproductibilité, c'est-à-dire à la propriété de donner des résultats semblables lorsque la mesure est renouvelée dans des conditions différentes (autres personnes, lieu ou date de mesure différentes).

V.LE PROBLÈME DE L'IMPUTATION CAUSALE ET DE L'OBSERVATION

L'épidémiologiste est toujours confronté au même problème: conclure à partir d'observations limitées dans le temps, dans l'espace et dans le champ d'investigation (la population n'étant généralement que peu étudiée dans sa totalité). En effet, quel part de confiance peut-il accorder à une relation observée entre deux événements A et B et dans quelle mesure cette relation est-elle causale, c'est-à-dire peut-il affirmer par exemple que A entraîne B ? Ce problème est d'autant plus délicat dans les enquêtes épidémiologiques que l'observateur ne maîtrise pas tous les facteurs, un facteur de confusion , c'est-à-dire lié à la fois au facteur étudié et au phénomène considéré, peut modifier la perception que l'observateur a d'une éventuelle association. Aussi la mise en évidence d'une association n'implique t-elle pas l'existence d'une relation de cause à effet.

5.1.Relation causale ou facteur de risque

Contrôler tous les facteurs se révèle parfois difficile. Face à ce constat, les enquêtes épidémiologiques seront conçues, réalisées et interprétées selon deux attitudes. A défaut de pouvoir mettre clairement en évidence une(des) cause(s), on s'attachera à rechercher les facteurs qui modifient la probabilité d'apparition d'un événement, autrement dit ses facteurs de risque sensu stricto; certains auteurs ne parlent de facteurs de risque que pour ceux qui augmentent la probabilité

de survenue d'un phénomène (14). Cette attitude est pragmatique: elle vise à prévoir les facteurs de risque. L'autre attitude correspond à une recherche étiologique. Son but est l'identification de la(les) cause(s) du phénomène étudié afin d'aboutir à son explication. Bien entendu, les objectifs étant plus difficiles à atteindre, cela demandera des moyens plus importants.

La notion de relation causale recouvre plusieurs situations. Il est possible de classer les types de relations causales d'après la probabilité de survenue de l'événement y sachant que la cause x est présente ou absente.

->Cas 1 :
Si x absent : P(y) = 0. Il est donc nécessaire que x soit présent pour que y survienne.
Si x présent : P(y) = 1. La présence de x suffit à la réalisation de y.
D'après les deux situations précédentes, x est donc une condition nécessaire et suffisante à la réalisation de y : x représente une cause unique de y.

->Cas 2 :
Si x absent : P(y) ≠ 0. Il n'est pas nécessaire que x soit toujours présent pour que y se réalise.
Si x présent : P(y) = 1. La présence de x suffit à la réalisation de y.
Ainsi x est une condition suffisante mais non nécessaire pour la réalisation de y. Il existe donc des causes parallèles.

->Cas 3 :
Si x absent : P(y) = 0. x est une condition nécessaire à la réalisation de y.
Si x présent : P(y) ≠ 1. La présence de x ne suffit pas pour que y se réalise à chaque fois.
x est donc une condition nécessaire mais non suffisante. Il faut tenir compte d'autres cofacteurs.

->Cas 4 :
Si x absent : P(y) ≠ 0. La présence de x n'est pas toujours nécessaire à la réalisation de y.
Si x présent : P(y) ≠ 1. La présence de x ne suffit pas pour que y se réalise à chaque fois.
Aussi x est-il une condition non nécessaire et non suffisante. Il y a des cofacteurs et des causes parallèles.

En épidémiologie explicative, un des objectifs des enquêtes est de mettre en évidence le type de condition auquel correspond le facteur x isolé. Cependant, sur le terrain les choses ne sont pas aussi simples que cela. Comme nous l'avons écrit plus haut, les facteurs de confusion gênent l'observation. Il faudra donc essayer de les mettre en évidence même si la difficulté paraît grande. Dans la démarche qu'il conviendrait d'adopter lors d'interprétation d'enquête en épidémiologie explicative, trois étapes peuvent être distinguées :
- comparaison "brute" de deux groupes séparés selon l'absence ou la présence d'un facteur

- recherche des facteurs de confusion
- comparaison à nouveau après élimination des facteurs de confusion.

Ces trois étapes font appel à des tests statistiques permettant de mettre à l'épreuve des faits les hypothèses émises.

5.2.Principes généraux des tests d'hypothèses

Un test d'hypothèses ou test de signification sert à répondre à une question à partir des observations réalisées sur un ou des échantillons. Selon le type de question, on distingue différents types de test.

-Le test d'ajustement est destiné à savoir si un échantillon observé peut être considéré comme extrait d'une population donnée.Par exemple, on tire n individus d'une population et on mesure leur poids. Peut-on considérer que la distribution du poids individuel au sein de cette population est normale ?

-Le test d'indépendance : on désire savoir si deux ou plusieurs critères de classification sont indépendants entre eux. Par exemple, pour une population de zébus la production laitière dépend-elle de la longueur des cornes ?

-Le test de conformité : l'échantillon a-t-il une caractéristique (moyenne, variance,...) identique à celle d'une population théorique ?

-Le test d'homogénéité : deux populations peuvent-elles être considérées comme homogènes ?

D'une façon générale pour répondre à une question, il faut que soient déja définis :
-la question,c'est-à-dire les hypothèses
-les critères pour y répondre, autrement dit une règle de décision.

Un test d'hypothèses est donc formé d'une part d'hypothèses,au nombre de deux, et d'autre part d'une règle de décision.

Il y a deux hypothèses :
->H° : hypothèse nulle, c'est l'hypothèse d'équivalence. Par exemple, la population A est homogène à la population B.
->H' : hypothèse alternative, celle de non équivalence.

La règle de décision :
on mesure l'écart observé, soit entre certaines caractéristiques de la population théorique et de l'échantillon (tests d'ajustement, d'indépendance et de conformité), soit entre certaines caractéristiques des divers échantillons (tests d'égalité). Ensuite on se place dans le cas où H° est vraie et on calcule la probabilité d'avoir un écart au moins aussi grand que celui observé. La règle de décision consiste, après s'être fixé une limite, à rejeter H° si cette probabilité est inférieure ou égale à cette limite, celle-ci étant appelée le niveau de

signification du test. Rejeter H^*, cela conduit à accepter H'.

Prenons l'exemple de la comparaison de deux moyennes. Une expérimentation a été menée pour étudier l'influence de l'alimentation sur la production de lait. Pour ce faire un ensemble de vaches laitières de même race, de même âge, de même numéro de lactation et élevées dans des conditions identiques a été divisé en deux parties. Au troupeau 1 fut distribué le régime 1 et au 2 le régime 2. Après un certain temps, les moyennes par troupeau furent relevées. Soient $\tilde{m}1$ et $\tilde{m}2$ les valeurs trouvées. La question est de savoir si les régimes sont équivalents du point de vue de la production laitière, c'est-à-dire si $m1 = m2$ où m1 et m2 représentent les moyennes des distributions de la production laitière des vaches de mêmes caractéristiques que celles précédentes et nourries respectivement avec le régime 1 et le régime 2. En effet, $\tilde{m}1$ et $\tilde{m}2$ ne sont que des valeurs observées sur des échantillons, si on en prend d'autres , des valeurs différentes seront probablement obtenues. Le test d'hypothèses se présente donc de la façon suivante :

$$H^* : m1 = m2$$
$$H' : m1 \neq m2$$

$\tilde{m}1$ et $\tilde{m}2$ étant des valeurs observées, se fonder sur elles conduit à des erreurs. Il y en a deux types. L'erreur qui consiste à rejeter une hypothèse vraie est l'erreur de première espèce et le risque de la commettre est le risque de première espèce, souvent noté α . Le deuxième type d'erreur est celle que l'on commet en acceptant une hypothèse fausse (erreur de deuxième espèce) et le risque correspondant est le risque β de deuxième espèce. Représentons ces notions sous forme de tableau où d'un côté on aura la décision prise et de l'autre la réalité.

		DECISION	
		H^*	H'
REALITE	H^*	$1-\alpha$	α
	H'	β	$1-\beta$

$1-\beta$ est la puissance du test, c'est la probabilité de rejeter H^* sachant que H^* est fausse. Le risque α est souvent choisi égal à 5%, 1% ou 0,1%; en biologie il est en général de 5%. Nous nous sommes donc placés dans le cas où H^* était vraie et nous avons calculé la probabilité P d'observer un écart au moins aussi grand que celui observé. Si cette probabilité est inférieure ou égale au niveau de signification, nous rejetons l'hypothèse H^* et nous disons que la différence est significative.

Au lieu de penser en termes de probabilité, de niveau de

signification, on peut aussi raisonner en termes de valeurs limites, c'est-à-dire valeurs que l'écart ne doit pas dépasser pour que H^* ne soit pas rejetée pour des risques α et β donnés.

Revenons à l'exemple précédent. La règle de décision s'écrit

$$\text{rejet de } H^* \text{ si } P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) \leq \alpha$$

où X1 et X2 sont les variables aléatoires correspondant aux moyennes de production laitière d'échantillons de même effectif que les parties 1 et 2 de notre ensemble initial. D'après ce que nous avons vu dans la partie distribution d'échantillonnage, ces variables aléatoires ont une distribution d'échantillonnage normale si la population parente est normale ou asymptotiquement normale sinon (en pratique, cette approximation sera justifiée lorsque l'effectif d'échantillon sera supérieure ou égale à trente). Les paramètres de ces distributions sont : $m(Xi) = mi$ et $\sigma(Xi) = \sigma/\sqrt{n}$, n étant l'effectif d'échantillon. Avant tout test de comparaison de moyennes, il est nécessaire de vérifier que les variances des populations desquelles sont issues les échantillons ne sont pas trop différentes; nous supposerons que cette condition, appelée condition d'homoscédasticité, est satisfaite.

Expression de la valeur limite

X1 et X2 étant deux lois normales (ou asymptotiquement normales) indépendantes, leur différence est une loi normale (ou asymptotiquement normale) de paramètres : $m(X1 - X2) = m1 - m2$ et $\sigma(X1 - X2) = \sigma(X1) + \sigma(X2) = \sigma/\sqrt{(2/n)}$.

$$P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) = P(|X1 - X2| / \sigma(X1 - X2) \geq |\tilde{m}1 - \tilde{m}2| / \sigma(X1 - X2))$$

or

$$U = \frac{(X1 - X2) - (m1 - m2)}{\sigma\sqrt{\left(\frac{2}{n}\right)}}$$

est la loi normale réduite. Donc l'expression devient :

$$U = \frac{X1 - X2}{\sigma\sqrt{\left(\frac{2}{n}\right)}}$$

si $m1 = m2$, c'est-à-dire dans le cas où H^* est supposée vraie. L'expression de la probabilité se ramène donc pour ce cas à :

$$P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) = P(|U| \geq |\tilde{m}1 - \tilde{m}2| / [\sigma/\sqrt{(2/n)}])$$

$$\text{d'où } P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) = 2\,P(U \geq |\tilde{m}1 - \tilde{m}2| / \sigma/\sqrt{(2/n)})$$

$$\text{or } P(U \geq |\tilde{m}1 - \tilde{m}2| / \sigma/\sqrt{(2/n)}) = 1 - P(U \leq |\tilde{m}1 - \tilde{m}2| / \sigma/\sqrt{(2/n)})$$

$$\text{donc } P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) = 2\,[1 - \phi(|\tilde{m}1 - \tilde{m}2| / \sigma/\sqrt{(2/n)})]$$

La règle de décision stipule qu'on rejetera H^* si

$$P(|X1 - X2| \geq |\tilde{m}1 - \tilde{m}2|) \leq \alpha \text{ , soit si}$$

$$\phi(|\tilde{m}1 - \tilde{m}2| / [\sigma/\sqrt{(2/n)}]) \geq 1 - \alpha/2$$

$$\text{donc si } |\tilde{m}1 - \tilde{m}2| / [\sigma/\sqrt{(2/n)}] = u(1 - \alpha/2)$$

L'hypothèse H° sera rejetée si la valeur absolue de la différence observée est supérieure ou égale à la valeur limite $u(1 - \alpha/2) \times [\sigma/\sqrt{(2/n)}]$.

C'est ainsi que l'on dira par exemple que les deux régimes sont différents si $|\tilde{m}1 - \tilde{m}2| \geq L1$: $\alpha 1$ sera le risque d'observer sur les échantillons des valeurs $\tilde{m}1$ et $\tilde{m}2$ telles que $|\tilde{m}1 - \tilde{m}2| \geq L1$ alors qu'en réalité $m1 = m2$ tandis que $\beta 1$ représentera le risque d'observer sur les échantillons des valeurs $\tilde{m}1$ et $\tilde{m}2$ telles que $|\tilde{m}1 - \tilde{m}2| < L1$ alors qu'en fait $m1 \neq m2$. En prenant une autre valeur limite L2 supérieure à L1, la différence observée devra être plus grande pour conclure à une différence significative, il sera donc moins facile de rejeter H° et plus facile d'accepter H°. Aussi aurons-nous $\alpha 2 < \alpha 1$ et $\beta 2 > \beta 1$ ($\beta 2$ et $\alpha 2$: risques associés à la valeur limite L2). Les deux espèces de risque sont antagonistes. Ceci est d'une importance pratique considérable. En effet, à taille d'échantillon donné, lorsque l'on fixe un risque α, on fixe par la même occasion le risque β et plus grand sera α, plus petit sera β.

5.3.De la bonne interprétation des risques

Pour la bonne interprétation d'un test statistique, il est primordial de bien avoir assimilé ce que représentent les deux espèces de risque.

Revenons à notre exemple. Supposons que lors d'une nouvelle expérience réalisée sur un autre échantillon de vaches , nous trouvions que la valeur absolue de la différence des moyennes est inférieure à la valeur limite correspondant au risque 5%. Dire que "les deux moyennes sont égales (d'où acceptation de l'hypothèse H°) avec un risque de 5% de se tromper" est une grave erreur. En effet, le risque 5% est le risque de première espèce, soit la probabilité de rejeter H° alors que H° est vraie. Or notre de décision n'a pas rejeté H°. Le risque qu'il conviendrait de prendre en compte ici est celui de deuxième espèce, c'est-à-dire la probabilité de conclure à l'égalité des moyennes alors qu'en réalité elles sont différentes. Souvent, seul le risque α est donné avec le résultat d'un test et le risque β est occulté (sciemment ou non). Le risque β de certains tests est si élevé qu'il remet en cause leur utilité. En effet, si la probabilité d'accepter une hypothèse fausse est élevée alors que vaut le test considéré ?

Pour un test donné, il est important de considérer la fonction de puissance du test, soit la relation existant entre la probabilité de rejet de H° et le degré de fausseté de l'hypothèse d'équivalence. Celui-ci est dans le cas qui nous préoccupe ici la différence entre les vraies moyennes des populations, soit $D = m1 - m2$. En fait, quand nous écrivons

H° : m1 = m2, ceci est un moyen simple d'écrire que nous voulons voir si m1 et m2 ne sont pas trop éloignées l'une de l'autre. En pratique, nous nous donnerons un seuil Δ tel que si $|m1 - m2| \geq \Delta$ alors nous estimerons que les moyennes sont différentes. D'une façon générale, la puissance d'un test, $1-\beta$ est une fonction croissante du degré de fausseté de l'hypothèse (D), de l'effectif des échantillons et du risque de première espèce. Concrètement, cela signifie que α, β, Δ et n sont liés. Par conséquent, si α, β et Δ sont fixés alors n ne peut pas prendre des valeurs fantaisistes au risque d'ôter toute signification au test.

Afin d'utiliser les tests statistiques d'une façon adéquate, il faut toujours garder présent à l'esprit qu'un test ne pourra jamais affirmer avec certitude un fait. S'il est très pratique pour rejeter des hypothèses, en revanche il ne permettra jamais de conclure avec certitude. En effet, accepter H° ne signifie pas que cette hypothèse est vraie mais que les observations disponibles ne permettent pas de la rejeter au profit de H'. L'être humain a toujours tendance à généraliser un peu trop vite et trop facilement. La statistique discipline cette tendance et indique le degré de confiance qu'il peut accorder à ses conclusions (A. Fagot-Largeault in (12)).

5.4.Les critères de causalité

Les enquêtes épidémiologiques rendent d'énormes services aux responsables des productions animales d'un pays car elles permettent de se rendre compte de la situation de l'élevage du pays. Le problème est de pouvoir tirer des conséquences pratiques des faits rapportés par les enquêtes. Cette difficulté vient en partie du fait que sans une maîtrise de tous les facteurs , il n'est pas possible d'assurer scientifiquement une relation causale entre les faits observés, seule l'expérimentation, où l'observateur provoque l'apparition des faits qu'il désire étudier dans des conditions qu'il maîtrise au moins partiellement, peut aider à atteindre le but visé.

Ainsi, une association statistiquement significative entre deux événements n'implique pas nécessairement qu'il y ait une relation de cause à effet entre eux ou que l'un soit un facteur de risque pour l'autre. La présence éventuelle d'un facteur de confusion peut effectivement gêner la perception de la réalité par l'observateur. L'épidémiologiste, comme, tout chercheur scientifique, se trouve donc confronté au problème d'interprétation des phénomènes observés. Cette difficulté est d'autant plus grande dans les enquêtes que l'observateur n'y a qu'un rôle passif : il n'a pas provoqué le phénomène étudié avec une maîtrise des facteurs extérieurs aussi complète que possible, il se contente d'observer et de noter ses observations en essayant d'éviter les biais éventuels. Comment conclure dans de telles conditions ? Si pour les enquêtes épidémiologiques à visée pragmatique, on se contente de la notion de facteur de risque, cela n'est évidemment pas possible dans la recherche étiologique où il s'agit de mettre en évidence des relations causales.

Pour ce faire, Bradford Hill (8) a proposé un ensemble de critères que devrait satisfaire une association entre des faits

observés afin de pouvoir être considérée comme une relation causale. Nous avons donc un cadre permettant d'aider à la conclusion, mais il ne saurait être dogmatique : une relation de cause à effet particulière peut très bien ne pas répondre à tous ces critères.

Critères de présomption causale de Bradford Hill :

->Absence d'ambiguité temporelle

Un événement A ne peut être la cause de l'événement B que si son apparition précède celle de B. Dans le cas des maladies à temps de latence long, les signes de la maladie B peuvent apparaître sans que l'action de l'agent causal n'ait été vue. Il convient donc de bien choisir la durée d'observation des phénomènes.

->Force de l'association

Les phénomènes étudiés doivent être fortement corrélés et cette corrélation doit être statistiquement significative.

->Effet dose-réponse

C'est le gradient biologique de l'effet. Par exemple, si A est la cause de B : plus A sera d'intensité élevée, plus l'effet B sera important. Cependant, il convient ici de se méfier de l'effet seuil : parfois, B ne va apparaître que si A est d'une intensité supérieure ou égale à celle du seuil d'apparition de B.

->Reproductibilité de l'association

La même association entre les deux événements doit être retrouvée dans les sous-populations. S'il s'agit d'une relation générale, cette association doit également se retrouver dans des populations de caractéristiques différentes de celles de la population de départ.

->Plausibilité biologique des hypothèses et cohérence des observations

Les observations doivent être cohérentes avec celles des autres études épidémiologiques. De plus, les connaissances actuelles sur la biologie au sens large doivent permettrent d'expliquer l'association mise à jour ou du moins ne pas la rejeter complètement. Nous touchons là au problème du dogmatisme en science. Une enquête qui révélerait des faits heurtant les connaissances scientifiques du moment risque d'être rejetée. Pour qu'elle puisse avoir quelques chances d'être prise en compte, elle devra être conçue et réalisée avec le plus de rigueur possible; sans quoi on ne prendra pas la peine d'étudier les résultats obtenus.

->Spécificité de l'association

Prenons le cas d'un facteur causant une maladie. La spécificité de l'association signifie que ce facteur doit être présent chez presque tous les malades et seulement chez eux. Ce critère est mis à mal dans beaucoup de cas. Souvent, pour une maladie infectieuse on aura des porteurs sains, c'est-à-dire que l'agent infectieux est présent dans l'organisme de ces sujets mais n'entraîne pas de signes cliniques. De plus, la réaction du

système immunitaire peut être si intense qu'elle élimine l'agent infectieux tout en déclenchant une pathologie grave.

Les critères de Hill ne sont en aucun cas incontournables. Ils ne font partie que d'un ensemble permettant de conclure à une relation de cause à effet. Dans le domaine scientifique, il est plus facile de réfuter que de mettre en évidence clairement un fait. La démarche scientifique consiste à mettre à l'épreuve des faits, au travers de l'expérimentation ou de l'étude par observation, des hypothèses issues d'observations antérieures. Les résultats de cette épreuve ne peuvent en aucun cas confirmer à eux seuls les hypothèses testées. Ce n'est que l'interprétation de ces résultats à la lumière de l'ensemble des connaissances accumulées qui permettra de confirmer les hypothèses avec une certaine marge d'erreur. Cette longue démarche d'acquisition de connaissances solides doit toujours être présente à l'esprit des personnes désireuses d'utiliser les données des enquêtes explicatives.

CONCLUSION

L'épidémiologie est une science dont les applications pratiques ne manqueront pas de prendre de l'importance dans les années futures. En effet, la lutte efficace contre les maladies monofactorielles, le développement de l'éco-pathologie et des suivis zootechniques et surtout la nécessité d'employer efficacement les fonds alloués à l'amélioration de l'état de santé d'une population animale imposent une démarche rigoureuse dans les processus de mise au point et de réalisation d'un protocole permettant de mieux appréhender la situation d'une population vis-à-vis d'un phénomène étudié. Celui-ci n'est pas nécessairement une maladie. Tout phénomène quantifiable peut être l'objet d'une enquête épidémiologique. Quel que soit le phénomène quantifiable étudié, les principales étapes d'une enquête épidémiologique seront les mêmes. Tout d'abord, il est indispensable de faire le bilan des connaissances sur le thème de l'enquête. Ensuite le cadre de celle-ci devra être bien délimité et ses objectifs seront clairement précisés. Pour répondre aux objectifs, on définira la population étudiée et on choisira les unités de base et la méthode d'échantillonnage. Afin que les données soient interprétables, il est nécessaire que les données soient recueillies rigoureusement. Par la suite, on s'assurera de la validité des données et on indiquera la méthode de recueil de ces données. La discussion des résultats et la conclusion qui pourra en être tirée seront faites en utilisant un raisonnement rigoureux. C'est ainsi que les méthodes d'estimation et les tests statistiques devront être employés à bon escient tout en ayant conscience de leurs limites. Cependant, l'épidémiologie ne peut en aucun cas se résumer à l'application pure et simple de formules statistiques. La statistique fournit à l'épidémiologie une base méthodologique rigoureuse mais ce sont les connaissances de l'épidémiologiste qui lui permettront de choisir, par exemple, telle méthode d'échantillonnage plutôt que d'autres. A cet égard, l'épidémiologiste peut être comparé à un médecin généraliste qui réalise un électro-cardiogramme sur un

patient afin de poser un diagnostic précis. Ce médecin n'a pas besoin d'être un spécialiste en cardiologie, mais il se doit de connaître les principes de base de l'électrocardiographie.

BIBLIOGRAPHIE

1.**BLOCH, N.; DIALLO, I.**
Enquête sérologique dans un pays sahélien, le Niger. Problèmes d'échantillonnage et résultats de la sérosurveillance de la peste bovine. Rev. Elev. Med. vet. Pays trop., 1990, **43** (3) : 305-311.

2.CHARTIER, C.; CHARTIER, F.
Enquête séro-épidémiologique sur les avortements infectieux des petits ruminants en Mauritanie. Rev. Elev. Med. vet. Pays trop., 1988, 41 (1) : 23-34.

3.DAGNELIE, P.
Statistique théorique et appliquée. Tome 1. Duculot, Gembloux, 1992.

4.DOMENECH, J.
Etude de l'épidémiologie des maladies animales en Afrique : stratégies d'approche et rôle des laboratoires vétérinaires. Rev. Elev. Med. vet. Pays trop., 1990, **43** (2) : 149-154.

5.FARVER, T.B.
Disease prevalence estimation in animal populations using two-stage sampling designs. Prev. vet. Med., 1987, 5 : **1-20.**

6.GOLBERG, M. et al.
L'épidémiologie sans peine. 2°éd. Paris, Frison-Roche, 1990.194 p.

7.HAROLD, J.
Theory of Probability, repr. 1948, 1961, Oxford : Clarenton Press

8.HILL, B.
Principes of medical statistics. Oxford Univ. Press, 1971, pp.304-323.

9.JENICEK, M.; CLEROUX, R.
Epidémiologie-Principes.Techniques.Applications. Paris, Maloine, 1984.454 p.

10.LEFEVRE, P.C.
Méthodologies d'enquêtes épidémiologiques en Afrique. Propositions pour la conception et le déroulement des enquêtes. Document n°1. Groupe de travail C.O.R.A.F. sur la santé animale. 27 p. N'Djaména mars 1990.

11.LEFEVRE, P.C.; FAYE, B.
Appui aux enquêtes d'écopathologie menées dans le cadre du programme P.P.R., "Pathologie et productivité des petits Ruminants". Dakar, février 1993.

12.LELLOUCH, J.
Présent et futur de l'épidémiologie. Colloque I.N.S.E.R.M. en l'honneur de D. Schwartz. Paris, Colloque I.N.S.E.R.M., 1988. 121 p.

13.MAURICE, Y.; IDRISS, A.O.
Les biais des enquêtes épidémiologiques vétérinaires dans les pays en développement. Document n°2. Groupe de travail C.O.R.A.F. sur la santé animale. 110 p. N'Djaména mars 1990.

14.MADEC, F.; FOURICHON, C.
Les facteurs de risque en épidémiologie animale. Epidémiol. Santé anim., 1990, 18 : 31-43.

15.PUTT, S.N.H. et al.
Epidémiologie et économie vétérinaire en Afrique. Manuel à l'usage des planificateurs de la santé animale. Addis-Abeba, C.I.P.E.A., 1987.146 p.

16.ROUGEMONT, A.
De la planification sanitaire à la planification des enquêtes épidémiologiques. In : 6 séminaire Y. Biraud. Enseignement de l'épidémiologie pour les pays en voie de développement, Talloires, Centre Européen de la Tufts University, sept. 1983.

17.RUMEAU-ROUQUETTE, C.; BREART, G.; PADIEU, R.
Méthodes en épidémiologie. 3 éd., Paris, Flammarion Médecines-Sciences. 420 p.

18.THRUSFIELD, M.
Veterinary epidemiology. London, Butterworths, 1986. 288 p.

19.TOMA, B. et al.
Glossaire d'épidémiologie animale. Maisons-Alfort, Editions du Point Vétérinaire, 1991. 365 p.

20.WAYNE MARTIN, S.; MEEK, A.H.; WILLEBERG, P.
Veterinary epidemiology. Principles and methods. Ames (USA), Iowa State University Press, 1987. 343 p.