REVUE DE STATISTIQUE APPLIQUÉE

DANIEL SCHWARTZ

La méthode statistique en médecine : les enquêtes éthiologiques

Revue de statistique appliquée, tome 8, n° 3 (1960), p. 5-27

<http://www.numdam.org/item?id=RSA_1960__8_3_5_0>

LA MÉTHODE STATISTIQUE EN MÉDECINE : LES ENQUÈTES ÉTHIOLOGIQUES

par Daniel SCHWARTZ
Directeur de l'Unité
de Recherches Statistiques de l'Institut National d'Hygiène à l'Institut Gustave Roussy

On se propose de rechercher dans quelle mesure un facteur x intervient causalement dans le déterminisme d'une maladie m au sein d'une population humaine : par exemple l'usage du tabac dans le cas du cancer broncho-pulmonaire.

Ce problème peut théoriquement être abordé :

- soit par la voie expérimentale : examen de 2 groupes comparables, obtenus par tirage au sort, dont l'un sera soumis au facteur x et l'autre non. Cette façon de faire est le plus souvent inapplicable pour des raisons matérielles ou morales; elle ne répond d'ailleurs pas exactement à la question posée : avec ce procédé autoritaire le cancer du poumon pourrait bien frapper, dans le groupe fumeur, des sujets particulièrement vulnérables, qui dans les conditions spontanées ne fumeraient pas.

- soit par la voie de l'observation : on cherche s'il existe, dans la population générale, une association entre l'exposition au facteur x et l'apparition de la maladie m. Cependant l'association ne permet pas de conclure à la causalité, du fait que l'exposition au facteur x est aléatoire, et liée à de nombreux facteurs, parmi lesquels peut se trouver la vraie cause : si l'usage du tabac résulte d'un psychisme déterminé, la consommation de tabac élevée des cancéreux ne serait-elle pas seulement l'indice d'un psychisme particulier, qui serait la cause de ce cancer ? La voie de l'observation ne saurait donc en théorie rien apporter au problème étiologique. En fait elle a permis, dans certains cas qui seront développés en fin de cet exposé, d'aboutir à une forte présomption causale.

Cependant même si, renonçant à l'interprétation causale, on se contente plus modestement d'étudier la relation d'association, il faut préciser d'emblée que cette association à l'état brut est souvent inintéressante : ainsi, dans l'exemple qui vient d'être cité, on doit s'attendre à observer beaucoup plus souvent le cancer broncho-pulmonaire chez les fumeurs que chez les non fumeurs, par le seul fait que l'âge moyen est beaucoup plus élevé dans le premier groupe que dans le second, qui comprend jusqu'aux nouveau-nés ; il n'est intéressant de comparer les fumeurs et les non fumeurs qu'à âge égal ; on arrive ainsi à la notion d'une association corrigée de l'influence de l'âge. Ce problème pourra être abordé, soit en examinant seulement une population d'âge donné (étude en population homogène), soit par une étude en population hétérogène, couvrant un assez large intervalle d'âges, l'influence de l'âge étant éliminée par un procédé statistique.

Les mêmes considérations s'appliquent naturellement à des facteurs autres que l'âge : sexe, peut-être milieu d'habitation, niveau

social ... d'une manière générale à tous les facteurs liés à la consommation de tabac. La liste de ces facteurs peut être longue et inconnue. En fait on peut assez raisonnablement distinguer 2 étapes :

- dans une première étape, on étudie l'association, d'une part à l'état brut, mais simultanément en la corrigeant de l'influence du sexe et de l'âge, et éventuellement d'un nombre extrêmement réduit de facteurs liés fondamentalement à l'exposition au facteur x.

- dans une deuxième étape, on corrige l'association pour tous les autres facteurs liés à l'exposition au facteur x. Cette étape peut être menée plus ou moins loin. Elle est sans fin ...

On peut admettre, sans trop d'arbitraire, que la limite entre ces 2 étapes définit le moment où se termine l'étude du rôle étiologique du facteur, et où commence la recherche d'une interprétation causale.

L'une et l'autre peuvent être abordées : soit par l'examen d'une population homogène, où tous les facteurs en cause sont constants, soit par l'examen d'une population hétérogène, où le rôle de ces facteurs est éliminé par un procédé statistique.

Ces considérations définissent le plan de notre exposé.

I - PRINCIPE DE L'ENQUETE ET MODE D'ECHANTILLONNAGE -

a) Un premier exemple (enquête *prospective*)

Dans cet exemple, on constitue un échantillon représentatif de la population générale ; on note si les sujets sont exposés ou non au facteur x, et on enregistre par la suite tous les cas de la maladie m qui se produisent.

De telles enquêtes ont été conduites notamment pour étudier le rôle de l'obésité et de l'hypertension dans l'étiologie de la maladie coronarienne (18), de la rubéole des femmes enceintes dans la production de malformations chez l'enfant (36), de la consommation de tabac dans le déterminisme de diverses maladies, en particulier le cancer des voies aéro-digestives supérieures (24, 25, 32).

L'échantillon examiné doit être représentatif. Cependant cette exigence n'est pas toujours réalisable, et dans certains cas on a choisi un groupe plus facile à suivre, par exemple la totalité du corps médical (24) ou un groupe de sujets pensionnés de l'Etat (25), faisant l'hypothèse d'une *stabilité* de l'association étudiée (si le tabac est dangereux pour le corps médical, il l'est vraisemblablement d'une manière générale). En tout état de cause, si on ne cherche pas à extrapoler, on dispose au moins de résultats valables pour une population *bien définie*.

L'échantillon peut être représentatif d'une *population homogène* : l'enquête (32) sur les fumeurs portait sur les sujets de sexe masculin, de race blanche, d'âge compris entre 50 et 69 ans. Par contre, l'enquête (24) sur le corps médical portait sur une *population hétérogène* en âges, et l'étude d'association à âge donné nécessita une correction par un des procédés qui seront exposés plus loin.

L'exposition au facteur x est connue par un examen ou un questionnaire, nécessairement très réduit en raison du grand nombre des sujets ; quant à la maladie, elle est enregistrée lorsque c'est possible ; cependant on doit souvent se contenter de l'information du décès, substituant à l'étude d'une maladie l'étude de la mort par cette maladie.

Une enquête "prospective" de ce genre présente d'indiscutables avantages : elle fournit, comme on le verra plus loin, une information complète sur le rôle

du facteur ; elle évite le recours à des groupes témoins critiquables ; l'interrogatoire a lieu à un moment où le sort du sujet n'est pas connu, ce qui en garantit l'impartialité. Enfin il est possible d'étudier simultanément les diverses maladies imputables au facteur.

Mais le nombre de sujets exigé est considérable, dès que la fréquence de la maladie est faible : on a suivi près de 200 000 sujets pendant plusieurs années dans les enquêtes (25) et (32). De telles enquêtes sont donc rarement réalisables.

b) Un deuxième exemple (enquête *rétrospective*).

Dans ce 2e exemple, on constitue un échantillon de sujets atteints de la maladie, et un échantillon qui en est indemne, et on les compare pour la proportion de sujets exposés au facteur x. L'échantillon de *malades* doit être représentatif ; si ceci a pu être réalisé dans de rares cas, notamment dans une enquête sur les cancers et leucémies de l'enfant qui a pu englober *tous les cas* (du moins tous les cas mortels) pendant une période donnée (62), on se contente en général d'un mode de recrutement commode, par exemple des cas rencontrés à l'hôpital et dans certaines villes, admettant ici encore la *stabilité* de l'association.

C'est alors la définition du groupe *témoin* qui devient difficile : il doit être obtenu par tirage au sort parmi les sujets indemnes de la maladie dans " *la population d'où provient l'échantillon de malades*". Si on choisit pour le groupe malade les cas hospitaliers, on admettra que cette population est "la clientèle hospitalière", c'est-à-dire une catégorie de sujets que leur condition (sociale, familiale, psychologique...) rend candidats à l'hôpital en cas de maladie.

En réalité il n'existe pas une clientèle hospitalière en général, mais une clientèle par maladie : plus celle-ci est grave, plus l'hôpital recrute à une grande distance et dans des classes sociales de niveau élevé ; c'est donc la *clientèle spécifique de la maladie m* qu'il faut échantillonner pour constituer le groupe témoin. Pratiquement on forme le groupe témoin avec les cas hospitaliers d'une maladie m' de gravité comparable à m ; ce procédé n'est acceptable que si la maladie m' a frappé cette clientèle "au hasard", c'est-à-dire *si elle ne présente pas de relation avec le facteur x* (ce qui va de soi : il ne faut évidemment pas que la maladie m' soit liée à une sous-exposition ou à une sur-exposition au facteur). On choisira par exemple, pour l'étude d'un cancer, des témoins atteints d'autres cancers, ou d'autres maladies graves ; ou bien, partant du groupe de sujets venus consulter pour une tumeur, dont ils ne savent si elle est bénigne ou maligne, on les divisera après coup en cancers (maladie à étudier) et tumeurs bénignes (témoins), étant à peu près assuré que les mêmes facteurs d'échantillonnage ont dirigé les uns et les autres vers l'hôpital.

La difficulté de trouver un groupe témoin correct conduit en général à choisir plusieurs groupes témoins, qu'on justifie par une comparaison mutuelle : par exemple, dans l'enquête (60) sur le cancer broncho-pulmonaire, les 3 groupes témoins choisis (cancers autres que ceux des voies aéro-digestives supérieures, malades des services de médecine générale, accidentés) ont présenté le même niveau de consommation de cigarettes (alors que celui-ci était beaucoup plus élevé pour les cancers du poumon).

De toute manière, si on suppose que le groupe témoin ne représente pas correctement la population d'où provient le groupe malade, il devient nécessaire de corriger ce biais, soit par l'examen de *populations homogènes* (comparaison des malades et des témoins dans des groupes de milieu d'habitation, niveau social ... donnés), soit en appliquant à l'échantillon de *population hétérogène* les corrections voulues, qui seront exposées plus loin.

Cette partie de l'analyse statistique doit être menée avec soin, si on veut éviter d'appeler association ce qui ne serait en fait que le résultat d'une inégalité d'échantillonnage entre les 2 groupes.

C'est pourquoi on tâche en général de corriger cette inégalité par un *appariement* ; ce procédé , qui est un des avantages possibles de l'enquête rétrospective, consiste à chercher, pour chaque malade interrogé, un témoin comparable eu égard à certaines caractéristiques : celles-ci peuvent être, soit des facteurs d'échantillonnage (milieu d'habitation, niveau social...) soit des facteurs essentiels cités plus haut : sexe, âge ... Ainsi, dans l'enquête (60) sur l'étiologie du cancer broncho-pulmonaire, à chaque malade correspondait un témoin d'âge voisin (même tranche d'âge de 5 ans), interrogé à la même époque et si possible dans le même hôpital ; dans l'enquête sur les cancers de l'enfant (62), les témoins étaient tirés au sort sur les registres de l'état-civil de la commune où était né l'enfant cancéreux, parmi les enfants de même sexe nés à la même date , ce qui assurait l'appariement par sexe, âge et lieu d'habitation.

Il va de soi que l'appariement doit être limité aux seuls facteurs dont le rôle est déjà connu ; dès lors qu'on apparie en fonction d'un facteur, on annule pour ce facteur la différence entre les groupes malade et témoin, de sorte qu'on renonce à toute information sur son rôle dans l'étiologie.

L'exemple ainsi décrit d'enquête rétrospective comporte finalement bien des difficultés de principe ; en contre-partie la conduite de l'enquête est infiniment plus aisée que dans les enquêtes prospectives, car il suffit de réunir un nombre relativement faible de cas. En outre, il devient alors possible de mettre en jeu un questionnaire détaillé, de sorte que ce n'est pas seulement le rôle d'un facteur qui est étudié, mais de plusieurs, voire de tous ceux dont l'influence étiologique est supposée.

Aussi ce genre d'enquête a-t-il tenté de nombreux chercheurs, qui l'ont utilisé pour des maladies aussi variées que la tuberculose (44), la maladie coronarienne (21, 28, 65), la cirrhose du foie (54), les malformations congénitales (45) et surtout le cancer. Dans ce seul domaine, on a ainsi étudié la relation entre la situation de famille et le cancer du sein (30, 37, 59, 63), ou du col de l'utérus (30, 34, 43, 68), l'étiologie du cancer de la vessie en fonction de l'usage du tabac (11, 19, 40) ou d'infections parasitaires (52), du cancer des voies aéro-digestives supérieures en relation notamment avec la consommation du tabac et de l'alcool (38, 50, 55, 56, 58, 69, 71), des cancers gastro-intestinaux en relation avec l'usage des laxatifs (6), du cancer gastrique en relation avec les antécédents héréditaires (64) ou le groupe sanguin (5, 29), de la leucémie chez l'enfant en relation avec une irradiation de la mère pendant la grossesse (62). Rien que pour le cancer broncho-pulmonaire, on peut citer plus de 20 enquêtes rétrospectives (notamment 7, 22, 55, 60, 67 pour ne mentionner que celles qui portent sur au moins 500 cas de cancer et 500 témoins, et 31, 70 pour celles qui portent sur le sexe féminin).

c) Considérations générales - Classification des enquêtes.

Les 2 modes d'enquête qui viennent d'être décrits sont très différents, et l'élément le plus apparent de cette différence est d'ordre chronologique : on s'attache à l'avenir des sujets dans un cas, au passé dans l'autre. Cependant le temps n'est ici qu'un caractère second, et il est bien plus judicieux de classer les enquêtes *d'après le mode d'échantillonnage.*

Nous adopterons à cet effet un modèle, représenté au tableau 2, où les sujets sont classés dans un tableau 2 × 2 en 4 catégories. Il s'agit là d'un modèle simplifié ; en effet :

a) La raison pour laquelle les sujets *non exposés* peuvent contracter la maladie m n'est pas envisagée ; cette raison peut être l'exposition à un facteur y au moins ; il faudrait dans ce cas prévoir un modèle à au moins 3 dimensions (exposition au facteur x, exposition au facteur y, maladie m) ;

b) on pourrait étudier avec plus de précision le rôle du facteur x, en supposant plusieurs degrés d'exposition. Ce point sera parfois pris en considération dans les pages qui suivent ;

c) il faut enfin préciser ce qu'on entend par "sujets atteints de la maladie m". Il peut s'agir de mortalité ou de morbidité, et, dans cette dernière éventualité, soit des nouveaux cas apparus pendant une période donnée, soit des cas existant à un moment donné (*incidence* et *prevalence* de la terminologie anglo-saxonne) ; ces aspects, - et d'autres qu'on peut imaginer - traduisent des moyens différents de mesurer la fréquence d'une maladie dans un groupe. Cette diversité est commune à bien des problèmes d'ordre statistique en médecine, et la source de bien des difficultés. Nous conserverons dans le tableau 2 la terminologie, à dessein vague, de "malades", en sachant qu'il y aurait lieu, pour chaque problème particulier, de formuler au départ une définition plus précise.

Si on adopte le modèle du tableau 2, les 2 variables, exposition au facteur x et atteinte par la maladie m, étant toutes deux aléatoires (puisqu'il s'agit uniquement d'observation, l'expérience étant exclue), c'est la nature de leur distribution - distribution contrôlée ou distribution aléatoire - qui permet de classer les types d'enquête. On obtient alors, avec White et Bailar (66), 3 types :

type 1 = distribution aléatoire pour x et pour m : on constitue un échantillon représentatif de la population étudiée

type 2 = distribution contrôlée pour x, aléatoire pour m : on constitue 2 groupes représentatifs de sujets exposés et non exposés

type 3 = distribution contrôlée pour m, aléatoire pour x : on constitue 2 groupes représentatifs de sujets malades et non malades.

TYPE 1 (ECHANTILLON REPRESENTATIF DE LA POPULATION ETUDIEE)

C'est dans cette catégorie qu'entrent les enquêtes prospectives décrites plus haut. Toutefois le type 1 n'oblige aucunement à suivre les malades dans le futur, on peut très bien dans certains cas se référer au passé ou au présent des sujets. Naturellement, s'il s'agit du cancer du poumon, on ne saurait s'intéresser au passé, car les sujets ayant dans le passé développé cette maladie seront en majorité décédés, ce qui faussera l'échantillonnage ; on ne peut pas davantage s'intéresser au présent, car le nombre de sujets atteints serait trop faible ; force est donc de suivre les sujets dans le futur. Mais dans le cas d'une maladie non mortelle, et fréquente, rien n'empêche de considérer le passé ou le présent : on pourra par exemple étudier la relation entre l'éthylisme et les altérations artérielles du fond d'œil sur un échantillon de taille modeste, le facteur et le signe pathologique étant tous deux largement répandus.

Au type 1 se rattache la catégorie particulièrement intéressante des enquêtes de morbidité, qui indiquent les nouveaux cas de maladie apparus, pendant une période déterminée, dans la *population entière* d'une aire géographique déterminée, et constituent des enquêtes étiologiques possibles lorsque le facteur x est une caractéristique démographique connue par les statistiques de cette population : ainsi a-t-on pu étudier la relation entre les cancers génitaux de la femme et la situation de famille, dans 10 grandes villes des U. S. A. (26) et dans la totalité du Danemark (10).

TYPE 2 (UN GROUPE DE SUJETS EXPOSES, UN GROUPE DE SUJETS NON EXPOSES)

Ce type peut permettre de suivre un nombre de sujets moins considérable que dans le type 1. Avec les notations des tableau 1 et 2, la comparaison des 2 groupes fait intervenir une variance de forme $\frac{m_0(1 - m_0)}{n_{0*}} + \frac{m_1(1 - m_1)}{n_{1*}}$; on peut, se fixant celle-ci, chercher les valeurs de n_{0*} et n_{1*} qui assurent l'effectif total $(n_{0*} + n_{1*})$ minimum. Si on suppose que la fréquence de la maladie ne sera pas beaucoup plus élevée dans le groupe exposé que dans le groupe non exposé $(m_1 = m_0)$, c'est en choisissant des effectifs égaux dans les 2 groupes qu'on obtient le minimum de sujets à suivre.

Dans le cas de l'enquête sur le rôle du tabac, un échantillonnage aléatoire de 200 000 sujets conduit à 30 000 non fumeurs et 170 000 fumeurs ; il est certain qu'en constituant au départ 2 groupes d'effectifs plus équilibrés, on peut, pour une même précision, diminuer le nombre de sujets nécessaire ; cela obligerait, par ailleurs, pour trouver davantage de non fumeurs, à organiser une prospection initiale plus étendue : peut-être serait-ce finalement plus compliqué, ceci dépend des difficultés relatives de la *prospection initiale* et de la *surveillance ultérieure*. La surveillance est en général difficile ; la prospection initiale peut être aisée : s'il s'agit d'étudier la fréquence des cancers génitaux de la femme en fonction du nombre d'enfants, ce dernier renseignement sera facilement disponible, et on aura tout intérêt à constituer 2 groupes d'effectif équivalent de femmes avec ou sans enfants.

Le bénéfice du type 2 est d'autant plus considérable que l'exposition au facteur est plus rare (par exemple exercice d'une profession peu répandue). Si celle-ci se rencontre 1 fois sur 1 000, le coefficient de la variance serait, par millier de sujets, dans le type 1, (toujours dans l'hypothèse $m_0 = m_1$), $\frac{1}{1} + \frac{1}{999} \neq 1$; précision qui peut être obtenue dans le type 2 par $\frac{1}{2} + \frac{1}{2}$, donc avec 4 sujets ; *il suffit donc d'un nombre de sujets 250 fois plus faible.*

A ces gains souvent très considérables, le type 2 permet d'ajouter encore un perfectionnement : lorsque l'exposition au facteur peut être divisée en plus de 2 classes hiérarchisées (0, 1, 2, ... enfants, ou non fumeurs, petits, moyens, grands fumeurs), si on suppose que l'effet du facteur croît en fonction de cette hiérarchie, on peut constituer 2 groupes, correspondant aux valeurs extrêmes (non fumeurs et très grands fumeurs, femmes sans enfants et mères de famille nombreuse). L'écart escompté entre les 2 groupes étant augmenté, on pourra se contenter d'effectifs plus faibles.

Enfin un cas extrême du type 2 est celui où on constitue seulement le groupe exposé au facteur, le groupe non exposé s'identifiant à la population générale : on a suivi, par exemple, un groupe d'ouvriers travaillant dans l'amiante, et comparé la fréquence observée de décès par cancer du poumon à celle de la population générale (23) ; on a de même étudié la mortalité par cancer du poumon chez les sujets gazés, ou souffrant de bronchite chronique (9), la mortalité par cancer de l'estomac chez des personnes achlorhydriques ou atteintes d'anémie pernicieuse (3, 33, 51), la mortalité chez les radiologistes, pour les différentes causes de décès, et en particulier le cancer (17, etc.). Cette méthode suppose naturellement que les sujets exposés constituent, dans la population générale, un groupe suffisamment petit pour qu'on puisse confondre population non exposée et population générale. Par ailleurs, les comparaisons de mortalité ou de morbidité ne s'entendent évidemment qu'à sexe égal, âge égal, éventuellement milieu social égal, etc. ce qui exige les corrections d'usage.

Tableau 1

Echantillon (effectifs)

		malades		
		non	oui	total
exposés	non	n_{00}	n_{01}	n_{0*}
	oui	n_{10}	n_{11}	n_{1*}
Total		n_{*0}	n_{*1}	n

Tableau 2

Population générale (proportions)

		Malades		Total	Proportion de sujets malades dans le groupe
		non	oui		
exposés	non	p_{00}	p_{01}	$p_{0*} = 1 - x$	$\frac{p_{01}}{p_{0*}} = m_0$
	oui	p_{10}	p_{11}	$p_{1*} = x$	$\frac{p_{11}}{p_{1*}} = m_1$
Total		$p_{*0} = 1 - m$	$p_{*1} = m$	1	
Proportion de sujets exposés dans le groupe		$\frac{p_{10}}{p_{*0}} = x_0$	$\frac{p_{11}}{p_{*1}} = x_1$		

TYPE 3 (UN GROUPE DE SUJETS MALADES, UN GROUPE DE SUJETS NON MALADES)

C'est dans cette catégorie qu'entrent les enquêtes rétrospectives décrites plus haut. Elle permet de réduire les effectifs prévus par le type 1, tout comme le type 2, et pour des considérations symétriques, portant cette fois sur les effectifs des groupes malade et témoin. Le gain est obtenu en équilibrant ces effectifs, et il est d'autant plus grand que la maladie, dans la population étudiée, est plus rare : dans le cas du cancer du poumon, il suffit de quelques centaines de sujets dans chacun des groupes malade et témoin pour obtenir la même précision qu'avec 200 000 sujets d'un échantillon aléatoire.

En réalité il arrive souvent, dans les enquêtes de ce genre, que les témoins soient plus faciles à recruter que les malades, de sorte qu'on préfère en

réunir un plus grand nombre ($n_{*0} > n_{*1}$). On se souviendra toutefois qu'il n'est pas opportun d'aller trop loin dans cette voie ; l'expression $\frac{1}{n_{*0}} + \frac{1}{n_{*1}}$ ne diminue plus guère, pour n_{*1} donné, quand n_{*0} devient grand : c'est ainsi qu'entre la valeur atteinte pour $n_{*0} = 3n_{*1}$ (soit $\frac{4}{3}\frac{1}{n_{*1}}$) et pour n_{*0} infini (soit $\frac{1}{n_{*1}}$) la diminution de variance ne compense guère la difficulté de recrutement.

Un cas extrême du type 3 est celui où on constitue seulement le groupe malade, le groupe témoin s'identifiant à la population générale - ceci n'étant possible que si la fréquence d'exposition au facteur est connue pour celle-ci, et si la maladie est suffisamment rare pour qu'on puisse confondre population non malade et population générale : on a comparé par exemple aux données de la population générale la situation de famille observée sur un groupe de 1 200 femmes atteintes de cancer du col utérin (48), ou la fréquence de la mortalité par cancer du sein dans l'ascendance féminine d'un groupe de malades atteintes de ce même cancer (46). Ces comparaisons sont naturellement faites à âge égal, éventuellement à milieu social égal, etc. par les corrections exposées plus loin.

D'une manière générale, dans les enquêtes du type 3, et surtout lorsqu'on craint des biais dans l'échantillonnage du groupe témoin, on devra tenir compte des multiples facteurs d'échantillonnage, pour des raisons qui ont été détaillées dans l'exemple de "l'enquête rétrospective".

d) Conclusion.

Le type 1, avec son échantillonnage représentatif de la population étudiée, est très coûteux en nombre de sujets.

Le type 2 permet de réduire ce nombre, ceci d'autant plus que l'exposition au facteur est une éventualité plus rare.

Le type 3 permet une réduction du même genre, d'autant plus considérable que la maladie est plus rare.

Il va de soi qu'en contre-partie on ne saurait attendre autant des types 2 et 3 que du type 1 : ils ne peuvent donner que des conclusions moins étendues et d'une valeur plus discutable ; c'est ce que précisera le chapitre suivant.

II - TEST ET MESURE DU ROLE ETIOLOGIQUE DE L'EXPOSITION AU FACTEUR -

a) Mesure du rôle étiologique dans la population étudiée, supposée homogène.

Nous nous plaçons dans le cas du modèle simplifié, décrit plus haut, et représenté au tableau 2, où on envisage 4 catégories de sujets, exposés ou non exposés, malades ou non.

Nous supposons en outre, pour commencer, que la population étudiée est *homogène* pour les facteurs essentiels énumérés plus haut, tels que : âge, sexe, niveau social.

Indépendamment des proportions ou probabilités p_{00}, p_{01}, p_{10}, p_{11}, qui définissent entièrement la situation, nous avons fait figurer simultanément au tableau 2 quelques combinaisons de ces probabilités qui, pour simplifier, seront désignées par des symboles plus parlants : m_0, m_1, et m, les probabilités de maladie chez les sujets non exposés, exposés, et globalement ; x_0, x_1, et x les

proportions de sujets exposés au facteur x parmi les sujets indemnes, malades, et globalement.

Si l'exposition au facteur n'intervient pas dans l'étiologie de la maladie m, les probabilités p_{01} et p_{11} sont proportionnelles à p_{00} et p_{10}, ou encore les probabilités m_0 et m_1 sont égales (ainsi d'ailleurs que les proportions de sujets exposés x_0 et x_1). Si elle intervient, il n'en est pas ainsi, les probabilités m_0 et m_1 par exemple sont différentes, en principe dans le sens $m_1 > m_0$.

Si l'exposition au facteur a un rôle étiologique (cette locution ne supposant pas qu'il s'agisse d'une relation causale) on peut se proposer de traduire ce rôle quantitativement.

Il est d'abord certain que le rôle du facteur x est d'autant plus important que le tableau 2 s'écarte davantage du modèle de l'indépendance, c'est-à-dire par exemple que m_1 s'écarte davantage de m_0. On pourra donc mesurer ce rôle par une expression indiquant l'écart entre m_1 et m_0.

Hammond et Horn, dans l'enquête prospective sur la mortalité en relation avec l'usage du tabac (32), ont utilisé, pour une cause de décès donnée, par exemple le cancer du poumon, le rapport $\frac{m_1}{m_0}$, qui mesure la surmortalité des fumeurs. Berkson (4) pense qu'il vaudrait mieux utiliser la différence $(m_1 - m_0)$. De toute manière, aucune fonction de m_1 et m_0 ne peut à elle seule résumer la situation définie par les 2 données m_1 et m_0 ; il est clair que pour un rapport donné la différence peut être variable, et inversement. M. C. Sheps (61) souligne qu'il est plus intéressant de former telle ou telle fonction de m_1 et m_0 qui ait un sens concret dans un *modèle* donné. Elle propose notamment de faire intervenir la mortalité par cancer du poumon liée en propre à l'usage de la cigarette, soit m_x, et d'écrire la la mortalité chez les fumeurs sous la forme :

$$m_1 = m_0 + m_x - m_0 m_x \qquad (1)$$

C'est là un modèle particulièrement simple, car aux conventions déjà adoptées plus haut (on n'envisage pas que les sujets non exposés au facteur *x* puissent être exposés ou non à d'autres facteurs, ce qui conduirait à un schéma à plus de 2 dimensions), on ajoute une hypothèse supplémentaire : les sujets, qu'ils soient exposés ou non exposés au facteur *x* (tabac), auraient par ailleurs la même probabilité de décès par cancer broncho-pulmonaire pour les "autres causes". Si on adopte ce schéma en première approximation, de (1) on tire :

$$m_x = \frac{m_1 - m_0}{1 - m_0} \qquad (2)$$

Cette fonction de m_1 et m_0 a un sens concret, puisqu'elle mesure la mortalité liée en propre à l'exposition au facteur x, ou encore mortalité qu'on observerait en l'absence des autres causes de cancer broncho-pulmonaire ; c'est surtout dans le cas de la relation causale que cette expression est intéressante : m_x mesure alors *l'effet* propre du facteur x.

On notera que $1 - m_x = \frac{1 - m_1}{1 - m_0}$; ce dernier rapport, qui prend ainsi un sens concret, est le rapport des survies des groupes exposé et non exposé, de sorte que le rapport des survies devient plus intéressant que le rapport des mortalités.

Il va de soi que m_x, pas plus qu'une autre fonction, ne résume m_1 et m_0, et qu'il faut une deuxième information pour définir le couple (m_1, m_0) ; celle-ci

peut être m_0, le couple (m_0, m_x) ayant une valeur plus concrète que le couple (m_0, m_1), puisqu'il exprime le risque en l'absence du facteur, et le risque lié en propre à l'exposition au facteur.

Enfin le couple (m_0, m_x) ne suffit pas encore à résumer la situation décrite par le tableau 2 ; celui-ci est défini par 4 probabilités p_{00}, p_{01}, p_{10}, p_{11}, dont la somme est 1, donc par 3 données indépendantes. On peut alors adjoindre au couple (m_0, m_x) une troisième donnée, par exemple la fréquence de l'exposition au facteur, soit x. La situation serait alors ainsi résumée :

m_0 = probabilité de maladie en l'absence d'exposition au facteur

m_x = probabilité de maladie pour un sujet exposé, en l'absence d'autres causes de la maladie, ou effet propre du facteur dans l'hypothèse causale

x = fréquence de l'exposition au facteur.

On peut naturellement préférer un autre groupe de 3 indices. Il reste que, de toute manière, le rôle étiologique d'un facteur ne saurait être mesuré par un seul indice : c'est là un résultat commun à tout problème de liaison entre 2 variables aléatoires dichotomiques et qu'on rencontre sous une forme similaire quand on veut mesurer le rôle d'un critère en matière de pronostic ou de diagnostic.

Un indice intéressant est *la proportion de cas dus au facteur* (proportion de cancers du poumon dus à l'usage du tabac), soit P. C'est :

$$P = \frac{xm_1 - xm_0}{(1 - x)\, m_0 + xm_1} = \frac{x\,(m_1 - m_0)}{m_0 + x\,(m_1 - m_0)} \qquad (3)$$

ou, en fonction de m_0, m_x, et x,

$$P = \frac{x\,(1 - m_0)\, m_x}{m_0 + x\,(1 - m_0)\, m_x}$$

b) Test et mesure du rôle étiologique d'après l'échantillon. Cas de l'échantillonnage représentatif d'une population homogène.

Pour commencer, nous supposons ici l'échantillonnage correct, c'est-à-dire donnant un échantillon représentatif de la population étudiée dans le type 1, deux groupes représentatifs des catégories exposée et non exposée dans le type 2, malade et témoin dans le type 3.

Nous supposons encore qu'il s'agit, dans chacun de ces cas, d'une *population homogène* en ce qui concerne les caractéristiques essentielles énumérées dès l'introduction de cet exposé, c'est-à-dire de sexe donné, d'âge donné, éventuellement de niveau social ou de milieu d'habitation donné ...

Il s'agit, d'après l'échantillon observé, d'éprouver puis d'estimer le rôle étiologique de l'exposition au facteur.

Le type 1 permet de connaître complètement le rôle étiologique du facteur : on *éprouve* d'abord ce rôle par jugement sur l'échantillon du tableau 1, à l'aide d'un test classique (χ^2 sur le tableau 2×2); il est possible ensuite d'*estimer* m_0 par $\frac{n_{01}}{n_{0*}}$, m_1 par $\frac{n_{11}}{n_{1*}}$, et x par $\frac{n_{1*}}{n}$

Dans le type 2, on a encore des estimations valables de m_0 et m_1, et leur comparaison permet d'*éprouver* le rôle étiologique du facteur. Cette comparaison

de proportions se ramène, ici encore, à un test de χ^2 sur le tableau 2 × 2. La *mesure* du rôle étiologique ne saurait par contre être complète : on a des estimations de m_0 et m_1 comme ci-dessus. Mais la fréquence x de l'exposition n'est pas connue, puisqu'on a choisi arbitrairement les effectifs des groupes exposé et non exposé. Ce mode d'enquête ne permet donc que d'évaluer l'effet du facteur, mais pas sa fréquence. (Notons qu'il est parfois possible de connaître celle-ci par ailleurs, à l'aide de données statistiques générales).

Dans le type 3, une difficulté se présente dès le *test d'association* : on ne peut pas comparer m_1 et m_0, car on ne dispose pas de leurs estimations du fait qu'on a choisi arbitrairement les effectifs des groupes malade et témoin. Par contre on a des estimations correctes de x_1 et x_0, qu'on peut comparer par un test de signification, qui est ici encore un χ^2 sur le tableau 2 × 2. Or il est visible que ce test permet d'éprouver le rôle étiologique du facteur. Si les désignations "malade" ou "non malade" des tableaux 1 et 2 désignent des sujets présentant la maladie pendant l'époque de l'enquête (*), le test d'association est réversible : si $x_1 > x_0$, on a aussi $m_1 > m_0$, c'est-à-dire une fréquence des cas de maladie, dénombrables pendant un intervalle de temps donné, plus élevée dans le groupe exposé, ce qui indique le rôle étiologique du facteur.

La *mesure* de ce rôle est malaisée : on peut estimer seulement x_0 et x_1, ce qui donne comme dans le type 2 deux indices au lieu de 3 ; mais ces indices ne sont guère intéressants, et on ignore m_0, m_1, et x, - à moins naturellement que la fréquence de la maladie dans la population générale ne soit par ailleurs connue par des données statistiques, auquel cas, disposant de 3 données, on peut connaître complètement le rôle du facteur.

Toutefois, lorsque la fréquence de la maladie, sans être connue, est *faible*, on peut tirer de l'enquête des renseignements étiologiques plus intéressants ; si on suppose la maladie rare, tant pour le groupe exposé que pour le groupe non exposé, on a en effet :

$$m_0 = \frac{p_{01}}{p_{00} + p_{01}} \# \frac{p_{01}}{p_{00}}$$

$$m_1 = \frac{p_{11}}{p_{10} + p_{11}} \# \frac{p_{11}}{p_{10}}$$

et

$$\frac{m_1}{m_0} \# \frac{p_{11}/p_{10}}{p_{01}/p_{00}} = \frac{p_{11}}{p_{10}} \frac{p_{00}}{p_{01}} = \frac{p_{11}}{p_{01}} \times \frac{p_{00}}{p_{10}} \qquad (5)$$

expression qui peut être estimée, à partir des données, par

$$r = \frac{n_{11}}{n_{01}} \times \frac{n_{00}}{n_{10}} \qquad (6)$$

Le rapport $\frac{m_1}{m_0}$ a été appelé *risque relatif* par Cornfield (15), qui en a donné l'estimation par la formule (6), ainsi que les limites de confiance. Ce risque relatif r mesure le rapport entre les proportions de sujets présentant la maladie

---

(*) Il s'agit donc, pour reprendre la distinction définie plus haut, des cas existant à un moment donné (en anglais *prevalence*).

donnée, pendant un intervalle de temps déterminé, chez les sujets exposés et non exposés. Dans le cas du cancer broncho-pulmonaire par exemple, le risque relatif des fumeurs, par rapport à celui des non fumeurs, est de l'ordre de 10.

Le risque relatif reste naturellement soumis aux limitations indiquées plus haut pour le rapport $\frac{m_1}{m_0}$; il ne saurait à lui seul résumer m_1 et m_0, et 2 situations étiologiques caractérisées, l'une par $m_0 = 1/1\,000$, $m_1 = 10/1\,000$, l'autre par des proportions 10 fois plus élevées, donnent le même risque relatif $r = 10$ alors que $(m_1 - m_0)$ par exemple est très différente. Mais le risque relatif a pour lui de pouvoir être estimé à partir des données, ce qui n'est le cas ni pour $(m_1 - m_0)$, ni pour m_x...

La *place* du facteur x dans l'étiologie peut également, - toujours dans la même hypothèse de maladie rare et dans le cas du modèle décrit plus haut - être connue. La proportion de cas dus à la maladie étant d'après (3) :

$$P = \frac{x\,(m_1 - m_0)}{(1 - x)\,m_0 + x m_1}$$

on a, pour une maladie rare,

$$x \# p_{10},\quad 1 - x \# p_{00},\quad m_0 \# \frac{p_{01}}{p_{00}},\quad m_1 \# \frac{p_{11}}{p_{10}}$$

de sorte que :

$$P \# \frac{p_{11} - \dfrac{p_{10}\,p_{01}}{p_{00}}}{p_{01} + p_{11}} = \frac{1 - \dfrac{p_{10}\,p_{01}}{p_{00}\,p_{11}}}{\dfrac{p_{01}}{p_{11}} + 1} \qquad (7)$$

expression dépendant seulement de $\frac{p_{10}}{p_{00}}$ et $\frac{p_{01}}{p_{11}}$, qu'on peut estimer d'après l'échantillon.

En faisant intervenir les proportions de sujets exposés, dans les groupes malade :

$$\left(x_1 = \frac{p_{11}}{p_{01} + p_{11}}\right)$$

et non malade :

$$\left(x_0 = \frac{p_{10}}{p_{00} + p_{10}}\right)$$

on peut exprimer P sous les formes :

$$P = \frac{x_1 - x_0}{1 - x_0} \qquad (8)$$

ou

$$P = \frac{x_0\,(r - 1)}{x_0\,(r - 1) + 1} \qquad (9)$$

proposée par Levin (39).

c) Test et mesure du rôle étiologique, dans le cas d'une population hétérogène (élimination de l'influence des tiers facteurs).

Il arrive le plus souvent que la population étudiée soit hétérogène au regard des facteurs déclarés essentiels, tels que sexe, âge, milieu d'habitation... Il s'agit alors, d'après un échantillon reflétant cette hétérogénéité, d'éprouver puis de mesurer le rôle étiologique de l'exposition au facteur.

On peut diviser chacun des facteurs essentiels en classes, par exemple : 5 tranches d'âge, 4 niveaux sociaux, 3 milieux d'habitation (grande ville, petite ville, campagne). Les diverses combinaisons de ces classes constituent c "cellules" (ici $5 \times 4 \times 3 = 60$ cellules).

Chacune de ces cellules est homogène

Une première solution du problème consiste à étudier séparément *chaque cellule*, par les procédés indiqués précédemment, autrement dit à subdiviser l'enquête en c sous-enquêtes ; c'est la seule solution réellement correcte ; elle permet d'observer éventuellement des résultats différents selon les cellules.

Cependant, dès que le nombre des cellules est élevé, les effectifs y deviennent trop faibles pour que ce procédé soit applicable.

On est alors conduit à étudier simultanément les c sous-enquêtes par une analyse d'ensemble, en supposant réalisées certaines hypothèses d'identité des résultats d'une cellule à l'autre. Cette analyse vise ainsi à corriger l'hétérogénéité de la population, en indiquant ce que serait le rôle étiologique de l'exposition au facteur à âge, niveau social, et milieu d'habitation donnés, donc *à éliminer l'influence de ces tiers facteurs*.

La première partie de cette analyse est le *test* du rôle étiologique. L'hypothèse faite pour permettre une étude simultanée des diverses cellules est que, si l'exposition au facteur joue un rôle dans l'étiologie, ceci doit être vrai dans toutes les cellules, et l'hypothèse nulle est l'absence de rôle étiologique dans chacune des cellules.

On est alors ramené à éprouver, par un test unique, l'absence de liaison dans un ensemble de tableaux de contingence 2×2.

Ce problème est justiciable de plusieurs solutions (voir notamment 13, 42) :

1/ On peut utiliser la somme des χ^2, avec la somme des degrés de liberté (c). Ce test présente un inconvénient : il ne tient pas compte du signe de la différence dans chaque cellule.

2/ On peut comparer à 0 la moyenne des χ (avec leur signe) par l'écart-réduit :

$$\frac{(\chi \text{ moyen}) - 0}{1/\sqrt{c}} = \chi \sqrt{c}$$

Ce test est en général meilleur que le précédent, mais il a encore l'inconvénient d'attribuer un même poids aux cellules, quel que soit leur effectif.

3/ Une meilleure solution consiste à donner des poids aux diverses cellules. Adoptons les désignations ci-dessous pour la cellule i (dans le cas du type 3 ; s'il s'agit du type 1 ou 2 on intervertira les termes "malade" et "exposé").

	Malades	Non malades	Différence
Effectif	n_i	n'_i	
Proportion de sujets exposés	p_i	p'_i	d_i
Proportion de sujets non exposés	q_i	q'_i	

Le test de l'égalité à 0 de l'ensemble des d_i peut être effectué en comparant à 0 une combinaison pondérée $\Sigma a_i d_i$, où on calculera les a_i de façon à obtenir le test le plus puissant.

Si on désigne par P_i et Q_i les proportions dans l'ensemble de la cellule, et si on pose $\frac{1}{w_i} = \frac{1}{n_i} + \frac{1}{n'_i}$, Cochran (13) propose comme solution le test :

$$\chi^2 = \frac{(\Sigma w_i d_i)^2}{\Sigma w_i P_i Q_i}$$

avec 1 degré de liberté.

Mantel et Haenszel (49) proposent un test très voisin, modifié pour tenir compte de la correction de continuité.

4/ On peut également, dans chaque cellule, calculer les effectifs théoriques des 4 cases dans l'hypothèse d'indépendance pour cette cellule. On somme ensuite les effectifs des cases homologues de toutes les cellules ; aux 4 effectifs théoriques ainsi obtenus on compare les 4 effectifs observés sur l'échantillon total, par un χ^2 à 1 degré de liberté (6). L'intérêt de ce test, par rapport aux précédents, est qu'il est facilement généralisable au cas où l'exposition au facteur x comporte plus de 2 classes, à condition de prendre le nombre de degrés de liberté voulu.

D'autres tests ont également été proposés. En fait, la neutralisation de variables plus ou moins nombreuses dans la comparaison de 2 groupes est un problème très général, mais il est si important dans le cas des enquêtes médicales qu'il constitue l'élément principal de leur analyse. Ceci explique la variété de tests utilisés.

La plupart des procédés indiqués rappellent la "standardisation par âge" utilisée par les démographes pour comparer "à âge égal" 2 populations dont la distribution d'âge est différente. Aussi sont-ils communément appelés *standardisation par âge, situation sociale, milieu d'habitation*, etc.

La standardisation est appliquée d'une manière relativement empirique, tant par le choix des tiers facteurs retenus (qui peut être restreint ou étendu) que par leur division en classes, et la constitution finale des cellules : il arrive qu'on standardise par rapport à chaque facteur isolément, ou par rapport à des groupes de deux, plutôt que de procéder à une standardisation d'ensemble conduisant à un grande nombre de cellules d'effectif très faible. Il peut arriver également qu'on fasse une étude séparée par sexe, avec pour chaque étude une standardisation pour les autres facteurs. Le choix entre les diverses voies d'approche est une affaire d'opportunité.

Les tests qui viennent d'être décrits, pour complexes qu'ils soient, ne représentent qu'un premier pas : l'épreuve d'association. Si l'exposition au facteur s'avère jouer un rôle étiologique, il reste à le mesurer.

Cette mesure peut être faite dans chaque cellule, mais l'élaboration d'une mesure unique, englobant les résultats de toutes les cellules, soulève des difficultés, car elle n'a de sens que si on suppose une comparabilité de toutes les cellules, qui est rarement vérifiée : dans le cas, par exemple, des enquêtes rétrospectives de type 3, une combinaison pondérée des risques relatifs ne paraît intéressante que si l'espérance mathématique de ces risques est la même dans toutes les cellules, hypothèse peu vraisemblable. Diverses combinaisons pondérées, de nature empirique, ont été proposées (49), mais leur difficulté d'interprétation ne fait que souligner les limitations d'emploi du risque relatif.

Un mot doit être dit enfin des *enquêtes avec appariement* ; l'appariement peut être utilisé, dans les enquêtes de type 2 et 3, pour rendre comparables, vis-à-vis de certains facteurs, les groupes exposé et non exposé ou malade et non malade. C'est donc une méthode visant, dès le stade de l'échantillonnage, a corriger l'hétérogénéité de l'influence de tiers facteurs.

Cependant l'appariement ne remplit complètement sa fonction que si l'analyse statistique en tient compte. On peut utiliser les tests classiques pour la comparaison de deux proportions dans des séries de sujets appariés (voir notamment 49). Ces méthodes sont généralisables au cas où l'exposition au facteur comporte plus de 2 classes (41).

Le gain de précision conféré par l'appariement n'est intéressant que si la variable d'appariement est fortement liée à l'exposition au facteur (12).

d) Validité des résultats.

Les perfectionnements mathématiques apportés à l'analyse statistique ne doivent pas faire perdre de vue diverses erreurs portant sur les données de base, et qui peuvent retirer toute valeur aux conclusions.

Il s'agit d'étudier l'association entre 2 variables x et m.

Ces variables sont d'abord passibles d'une erreur d'appréciation : on peut classer un sujet comme fumeur alors qu'il ne l'est pas, et inversement ; comme atteint de cancer du poumon alors qu'il en est indemne, et inversement ; de telles erreurs sont inévitables (ne serait-ce que parce qu'un sujet témoin souffre peut-être d'un cancer encore inapparent), mais il importe d'en distinguer 2 catégories :

- les erreurs portant sur *une des variables sans relation avec l'autre* ne sont pas graves : classer quelques sujets dans un groupe au lieu de l'autre revient à atténuer l'écart entre ces groupes et à diminuer la puissance du test, mais ne risque pas d'entraîner à des conclusions erronées.

- beaucoup plus graves sont, par contre, les *erreurs influencées par la liaison à étudier*: si, parce que le sujet est atteint d'un cancer du poumon, lui-même ou l'enquêteur qui l'interroge exagèrent sa consommation de tabac, si inversement le médecin fait intervenir dans les éléments de son diagnostic de cancer une consommation de tabac élevée, alors on risque d'observer une association reflétant uniquement l'idée préconçue. Or la subjectivité des réponses est souvent manifeste : les malades atteints d'un cancer du pharynx se remémorent ou insistent davantage sur les maux de gorge antérieurs, les femmes atteintes d'un cancer du sein ont tendance à exagérer la fréquence des douleurs mammaires dans leur passé, ou des cancers du sein dans leur famille. D'une manière générale, la comparabilité des interrogatoires entre sujets malades et témoins, ou exposés et non exposés, est une des difficultés majeures de l'enquête : comment obtenir qu'une mère, dont l'enfant est mort de leucémie, réponde à l'interrogatoire de la même façon qu'une mère témoin ?

Aussi l'élimination de ce type d'erreur doit-elle être recherchée par tous les moyens.

Il faut d'abord obtenir un diagnostic indépendant du facteur x, ce qui est facile si le diagnostic repose sur des éléments objectifs, par exemple l'histologie pour un cancer.

Il faut ensuite obtenir, pour le facteur x, des informations indépendantes du diagnostic; c'est ici que l'enquête "prospective" décrite plus haut offre des garanties supérieures à toute autre, puisque l'interrogatoire a lieu à un moment où la maladie n'est pas encore déclarée. Dans les autres modes d'enquête, lorsque la maladie est déjà déclarée au moment de l'interrogatoire, l'ignorance du diagnostic par le malade, l'enquêteur, ou les deux, doit être recherchée dans toute la mesure du possible : on interrogera par exemple comme malades et témoins des sujets consultant pour une tumeur qui n'est cataloguée qu'ultérieurement comme maligne ou bénigne ; Doll et Hill (22) ont ainsi apporté un argument important en signalant que la proportion de fumeurs était normale chez des sujets étiquetés "cancer du poumon" au moment de l'interrogatoire, et dont le cancer a été infirmé ultérieurement. Un autre argument important est que certains types histologiques seulement sont liés à l'usage du tabac, et pas d'autres, alors que le type histologique n'est pas connu au moment de l'interrogatoire.

Les considérations précédentes avaient trait aux *erreurs de mesure ;* des réserves analogues doivent être énoncées pour les *erreurs d'échantillonnage*, qu'on doit également subdiviser en 2 catégories :

- *les erreurs d'échantillonnage portant sur une des variables* ne sont pas trop graves : si dans une enquête du type 1 l'échantillon observé n'a pas tout-à-fait le même milieu social que la population étudiée, la consommation de tabac sera peut-être faussée, mais les méfaits éventuels de cette consommation le seront sans doute peu ; si, dans les enquêtes de type 2 et 3 l'échantillonnage des groupes à comparer diffère quelque peu, on pourra corriger ces différences par une standardisation.

- beaucoup plus graves sont par contre les *erreurs d'échantillonnage portant sur l'association même des 2 variables.*

Les enquêtes du type 1 présentent à cet égard une certaine sécurité. Berkson (2) a, il est vrai, imaginé une cause d'erreur possible dans les enquêtes prospectives sur les fumeurs ; cependant un tel biais reste minime (35).

Il n'en est pas de même dans les enquêtes du type 2, et surtout du type 3 ; un premier exemple d'erreur a été signalé par Berkson (1) ; c'est le cas où on étudie l'association entre une maladie m et une autre maladie, jouant le rôle du facteur x, parmi les malades se présentant à l'hôpital. Les sujets souffrant de la maladie m ont une certaine propension à se rendre à l'hôpital. S'ils souffrent *en outre* de la maladie x, cette propension est plus élevée, de sorte que l'échantillon hospitalier de malades (m) montrera une proportion trop élevée de sujets souffrant de la maladie x. Chez les témoins - qui sont des malades souffrant de diverses maladies - ce biais existe également, plus fortement ou moins fortement que pour la maladie m, selon le cas ; la comparaison des 2 groupes peut alors faire apparaître des différences purement artificielles.

La même situation se présente lorsqu'on étudie l'association entre 2 maladies ou signes morbides dans une série d'autopsies : par exemple entre nodules tuberculeux et cancer. Les 4 combinaisons, avec et sans cancer, avec et sans nodules, sont, chez des sujets décédés, différentes de celles qui existent dans la population générale, en raison de leurs taux de mortalité différents ; on peut admettre que, chez les sujets non cancéreux, la présence de nodules augmente

la mortalité, tandis que cet effet est négligeable chez les cancéreux : ainsi apparaîtra illusoirement chez les décédés une association négative, entre nodules tuberculeux et cancer, qui n'existe pas dans la population générale des vivants (2, 47).

De tels biais soulignent une limitation de ce genre d'enquête où malades et témoins *se recrutent d'eux mêmes* par leur venue dans l'échantillon (par la décision de consulter, par la mort, etc.) : c'est qu'il n'est pas possible d'étudier le rôle étiologique d'*un facteur influençant le recrutement*, ou du moins l'influençant inégalement pour les malades et les témoins.

Il faut bien le dire : le statisticien, habitué à constituer un échantillon par des procédés classiques de tirage au sort, risque d'être surpris, voire choqué, en découvrant que dans la plupart des enquêtes médicales on laisse aux sujets la responsabilité de l'autorecrutement. Pour une maladie donnée, le fait d'aller à l'hôpital occasionne déjà une première sélection, dépendant de facteurs sociaux et psychologiques. Les sujets présents un jour donné à l'hôpital constituent une nouvelle sélection, un malade ayant d'autant plus de chances d'être présent que sa durée d'hospitalisation est plus longue (57). Dans le même ordre d'idées, l'échantillon de malades vivants un jour donné constitue également une sélection renforçant la proportion de malades à survie longue (53). La notion de représentativité fait trop souvent place à la notion de commodité, et il arrive, comme le fait remarquer Dorn dans une mise au point récente (27), qu'une enquête du type 3 vise à comparer "deux échantillons sans spécification provenant par une méthode d'échantillonnage inconnue d'une population non identifiée".

Ceci ne doit pas être considéré comme une condamnation des enquêtes du type 3, qui restent le seul moyen facilement réalisable de suggérer des facteurs étiologiques ; mais leurs conclusions doivent être accueillies avec réserve, et soumises, lorsque l'enjeu en vaut la peine, à la confirmation d'enquêtes du type 1, plus rigoureuses mais infiniment plus difficiles à entreprendre.

e) Conclusion.

L'analyse du rôle étiologique d'un facteur peut être complète dans le type 1 ; elle est nécessairement incomplète dans le type 2 et surtout dans le type 3.

La valadité des résultats ne peut d'autre part être garantie que si l'on a pu éviter des erreurs de mesure et des erreurs d'échantillonnage portant précisément sur la liaison à étudier ; à cet égard on peut obtenir une relative sécurité avec les enquêtes du type 1, tandis que les biais sont plus difficilement évitables dans le type 2 et surtout dans le type 3.

III - L'INTERPRETATION CAUSALE -

Après avoir évité les biais et pièges de tout ordre, éliminé le rôle de quelques "tiers facteurs" essentiels (sexe, âge ...), on conclut au rôle étiologique de l'exposition au facteur x. Peut-on interpréter ce rôle en termes de causalité ?

Nous avons souligné dès le départ l'impuissance fondamentale à cet égard de l'enquête d'*observation* ; dans une *expérimentation*, on peut exposer au facteur étudié 2 groupes comparables à tout point de vue, de sorte que toute différence revêt d'emblée une signification causale ; dans l'*observation* l'exposition au facteur est déterminée aléatoirement, en liaison avec d'autres facteurs $x_1, x_2, \ldots$ parmi lesquels peut se trouver la vraie cause : les fumeurs étant plus souvent des citadins et des buveurs de café, la vraie cause du cancer du poumon ne serait-elle pas l'abus du café, ou l'atmosphère polluée des villes ? En outre, dans une

expérience, on peut souvent maintenir la comparabilité entre les groupes exposé et non exposé *après* l'intervention du facteur, tandis que dans les conditions spontanées ces 2 groupes peuvent se différencier systématiquement ; les fumeurs, sujets au catarrhe, devront peut-être davantage se faire radiographier : si les rayons X étaient alors la cause du cancer broncho-pulmonaire, l'usage du tabac serait certes un facteur causal, mais par une voie indirecte dont la signification serait très différente de la causalité directe.

De fait, bien des facteurs apparus, au cours d'une enquête, comme associés à l'apparition d'une maladie, sont sans action causale réelle : le niveau social pour le cancer de l'estomac (14), la presbytie précoce dans le cas de la maladie coronarienne (8), l'âge au mariage, et le nombre d'enfants pour le cancer du col de l'utérus, entrent sans doute dans cette catégorie.

Par contre, dans d'autres cas, comme celui du tabac pour le cancer du poumon, il est possible de justifier une forte présomption de causalité.

Il est d'abord possible d'argumenter contre l'objection du "tiers facteur". Si la "vraie cause" du cancer broncho-pulmonaire est un facteur lié à l'usage du tabac, par exemple l'abus du café, on doit alors observer qu'à consommation de café donnée le rôle du facteur tabac disparaît; cette étude du rôle du tabac à niveau égal pour différents autres facteurs peut être effectuée, ceci par les divers procédés envisagés plus haut (standardisation). L'élimination des "tiers facteurs" entreprise déjà pour certains facteurs essentiels (sexe, âge . . .) dans l'épreuve du rôle étiologique, peut être poursuivie avec plus de détails pour toute une série de facteurs liés à l'exposition au facteur (ceci est nécessaire quel que soit le type d'enquête). Ce travail a été fait, dans le cas du cancer broncho-pulmonaire, et on a observé que la prise en considération de plusieurs dizaines de facteurs ne permettait en aucun cas d'"innocenter" le tabac (20).

Sans doute cette méthode d'exploration est-elle soumise à une sérieuse limitation : elle ne permet d'étudier que les facteurs prévus dans l'interrogatoire des malades ; or la "vraie" cause peut être insoupçonnée. Mais ici intervient un argument d'ordre quantitatif; la liaison entre l'usage du tabac et l'apparition du cancer broncho-pulmonaire étant très forte, il est facile de montrer qu'elle ne peut être "expliquée" par un tiers facteur, que si celui-ci est *à la fois très lié à l'apparition de ce cancer et à l'usage du tabac* (16) ; il est peu probable qu'un tel facteur ait échappé aux nombreuses investigations effectuées. Il n'est pas suffisant, pour invalider l'hypothèse causale, de déceler par exemple un effet de l'hérédité dans l'habitude de fumer : il faudrait encore que la constitution génétique fût *fortement liée* à cette habitude (ce qui n'est pas le cas), et à l'apparition du cancer broncho-pulmonaire (ce qui n'a pas été signalé). D'une manière générale, plus l'association est forte, entre l'exposition au facteur et l'apparition de la maladie, et plus la présomption causale est solide.

Cependant l'élimination de tous les tiers facteurs n'est pas concevable (le fût-elle qu'elle n'apporterait d'ailleurs pas la certitude : à la limite, si les cancéreux et les témoins ne différaient que par l'usage du tabac, on pourrait supposer que c'est le cancer qui conduit les sujets à fumer . . .).

C'est pourquoi d'autres arguments, de divers ordres, doivent être recherchés. Indiquons que, dans le cas du facteur tabac, - indépendamment des confirmations obtenues en laboratoire, *in vitro* et *in vivo*, dont la contribution est toujours essentielle - , on a observé les relations suivantes :

- la probabilité de cancer broncho-pulmonaire est plus élevée chez le fumeur que chez le non fumeur, d'autant plus qu'il fume davantage, selon une loi proportionnelle ; elle diminue si le sujet s'est arrêté de fumer, et d'autant plus qu'il s'est arrêté plus tôt ;

- on rencontre chez les fumeurs une proportion exagérée de cancers de la cavité buccale, du pharynx, du larynx, de l'œsophage, et de la vessie - c'est-à-dire de toutes les localisations directement exposées à la fumée ou à ses dérivés immédiats - et une proportion normale des autres cancers ;

- la probabilité de cancer est augmentée par le fait de respirer la fumée lorsqu'il s'agit du cancer des bronches ou du larynx, elle n'est pas augmentée pour les autres cancers des voies aéro-digestives supérieures.

La convergence parfaite de ces arguments dans le sens de la relation causale ne peut manquer de frapper. Si l'usage du tabac n'est pas la "vraie" cause, il faut qu'il accompagne celle-ci bien fidèlement : présent quand elle est présente, absent si elle disparaît, faible ou fort à sa mesure. L'hypothèse d'un facteur aussi "mimétique" ne saurait être écartée avec certitude, mais elle fait penser à la phrase de l'humoriste : on a découvert l'auteur des pièces de Shakespeare ; c'est un homme qui vivait à la même époque, dans le même village, et qui portait le même nom que lui ...

Ce n'est pas, cependant, sur l'étiologie du cancer broncho-pulmonaire qu'il serait équitable de terminer : si on a pu, dans ce cas, après plus de 25 enquêtes, parvenir à la quasi-certitude, c'est qu'il s'agit d'un cas facile : la probabilité de cancer broncho-pulmonaire est extrêmement faible chez un non fumeur, elle est dix fois plus élevée si le sujet fume, et l'usage du tabac est très répandu ; l'exposition au facteur joue donc dans l'étiologie de cette maladie un rôle considérable.

Mais dans beaucoup d'enquêtes, la situation se présente moins favorablement : alors un travail ardu d'interrogatoire, une analyse statistique complexe pour tenir compte de multiples variables, ne conduisent, en ce qui concerne la relation causale, qu'à des conclusions incertaines : seule est responsable ce cette faible rentabilité la complexité même du sujet.

BIBLIOGRAPHIE

Les références notées ++ portent sur la méthodologie des enquêtes.

Les références notées + correspondent à des enquêtes où sont exposés ou discutés certains points de méthodologie.

++ (1) BERKSON J. - Limitations of the application of fourfold table analysis to hospital data. *Biometrics bulletin* 2, n° 3 : 47-53, 1946.

++ (2) BERKSON J. - The statistical study of association between smoking and lung cancer. *Proceed. of the Staff Meetings of the Mayo Clinic* 30, n° 15, 1955.

(3) BERKSON J., COMFORT M. W., BUTT H. R. - Occurence of gastric cancer in persons with achlorhydria and with pernicious anemia. *Proceed. of the Staff Meetings of the Mayo Clinic* 31 : 583-596, 1956.

++ (4) BERKSON J. - The statistical investigation of smoking and cancer of the lung. *Proceed. of the Staff Meetings of the Mayo Clinic* 34 : 206-224a, 1959.

(5) BILLINGTON B. P. - Gastric cancer - Relationships between abo blood-groups, site, and epidemiology. *The Lancet* : 859-862, 1956.

+ (6) BOYD J. T., DOLL R. - Gastro-intestinal cancer and the use of liquid paraffin. *Brit. J. Cancer* 8 : 231-237, 1954.

(7) BRESLOW L., HOAGLIN L., RASMUSSEN G., ABRAMS H. K. - Occupations and smoking as factors in lung cancer. *Amer. J. Public Health* 44, n° 2, 1954.

(8) BRESLOW L., BUECHLEY R. - Factors in coronary artery disease - Cigarette smoking and exercise *California Medicine* 89, n°3 ; 175-178, 1958.

+ (9) CASE R. A. M., LEA A. J. - Mustard gas poisoning, chronic bronchitis, and lung cancer. *Brit. J. of Prev. & Soc. Med.* 9, n° 2, 1955.

(10) CLEMMESEN J. - Carcinoma of the breast : symposium results from statistical research. *Brit. J. of Radiol.* 21, n° 252 : 583-590, 1948.

(11) CLEMMESEN J., LOCKWOOD K., NIELSEN A. - Smoking habits of patients with papilloma of urinary bladder. *Danish Med. Bull.* 5, n°3 : 123-128, 1958.

++ (12) COCHRAN W. G. - Matching in analytical studies. *Amer. J. of Public Health*, Part I 43, n° 6 : 684-691, 1953.

++ (13) COCHRAN W. G. - Some methods for strengthening the common X^2 tests. *Biometrics* 10, n° 4 : 417-451, 1954.

(14) COHART E. M. - Socioeconomic distribution of stomach cancer in New-Haven. *Cancer* 7, n° 3 : 455-461, 1954.

++ (15) CORNFIELD J. - A method of estimating comparative rates from clinical data. *J. Nation. Canc. Instit.* 2, n° 6 : 1269-1275, 1951.

++ (16) CORNFIELD J., HAENSZEL W., HAMMOND E. C., LILIENFELD A. M., SHIMKIN M. B., WYNDER E.L. - Smoking and lung cancer : recent evidence and a discussion of some questions. *J. of the Nat. Canc. Inst.* 22, n° 1 : 173-203, 1959.

+ (17) COURT BROWN W. M., DOLL R. - Expectation of life and mortality from cancer among british radiologists. *Brit. Med. J.* ii : 181-187, 1958.

(18) DAWBER Th. R., MOORE F. E., MANN G. V. - Coronary heart disease in the Framingham study. *Amer. J. of Public Health :* 4-24, 1957.

(19) DENOIX P. F., SCHWARTZ D. - Tabac et cancer de la vessie. *Bull. du Cancer* 43, n° 4 : 387-393; 1956.

+ (20) DENOIX P. F., SCHWARTZ D., ANGUERA G. - L'enquête française sur l'étiologie du cancer broncho-pulmonaire - Analyse détaillée. *Bull. du Cancer* 45, n° 1 : 1-37, 1958.

(21) DOLGOFF, SCHREK, BALLARD, BAKER - Tobacco smoking as an etiologic factor in disease - Coronary disease and hypertension. *Angiology* 3, n° 4 : 323-334, 1952.

+ (22) DOLL R., HILL A. B. - A study of the aetiology of carcinoma of the lung. *Brit. Med. J.* 2 : 1271, 1952.

+ (23) DOLL R. - Mortality from lung cancer among abestos workers. *Brit. J. Industr. Med.* 12 : 81-86, 1955.

+ (24) DOLL R., HILL A. B. - Lung cancer and other causes of death in relation to smoking. *Brit. Med. J.* ii : 1071, 1956.

(25) DORN H. F. - Tobacco consumption and mortality from cancer and other diseases. *Public Health Reports* 74, n° 7 : 581-593, 1959.

+ (26) DORN H. F. - Morbidity from cancer in the United States. *Publ. Health Monogr.* n° 29, 1955.

++ (27) DORN H. F. - Some problems arising in prospective and retrospective studies of the etiology of disease. *New England J. of Med.* 261 : 571-579, 1959.

(28) ENGLISH, WILLIUS, BERKSON - Tobacco and coronary disease. *J. Amer. Med. Assoc.* 115, n° 16, 1940.

+ (29) FRASER ROBERTS J. A. - Blood groups and susceptibility to disease. *Brit. J. of Prevent & Soc. Med.* 11, n° 3, 1957.

(30) GILLIAM A. G. - Fertility and cancer of the breast and of uterine cervix - Comparisons between pregnancy rates among women with cancer at these and other sites. *J. Nation. Canc. Inst.* 12, n°2 - 287-304, 1951.

+ (31) HAENSZEL W., SHIMKIN M. B., MANTEL N. - A retrospective study of lung cancer in women. *J. Nation. Canc. Inst.* 21, n° 5 : 825-842, 1958.

+ (32) HAMMOND E. C., HORN D. - Smoking and death rates - Report on fourty-four months of follow-up of 187-783 men. *J. Amer. Med. Assoc.* 166 : 1159-1308, 1958.

(33) HITCHCOCK C. R., SULLIVAN W. A., WANGENSTEEN O. H. - The value of achlorhydria as a screening test for gastric cancer. *Gastroenterology* 29, n° 4 : 621-628, 1955.

(34) JONES E. G., MACDONALD I., BRESLOW L. - Study of epidemiologic factors in carcinoma of uterine cervix. *Amer. J. Obst. & Gynec.* 76 : 1-10, 1958.

+ (35) KORTEWEG R. - The significance of selection in prospective investigations into an association between smoking and lung cancer. *Brit. J. Cancer* 10 : 282-291, 1956.

(36) LAMY M., SEROR M. E. - Les embryopathies d'origine rubeolique. *Semaine Hopit. Paris*, n° 36, 1956.

(37) LANE-CLAYPON J. E. - A further report on cancer of the breast, with special referance to its associated antecedent conditions. *Rept. Publ. Health* & M. Subj., n° 32 : 1-189, 1926.

+ (38) LEDERMANN - Cancers - Tabac - Vin & Alcool. *Concours Médical*, n° 11 & 12, 1955.

+ (39) LEVIN M. L. - The occurence of lung cancer in man. *Acta Intern. Union against Cancer* 9, n° 3, 1953.

(40) LILIENFELD A., LEVIN M. L., MOORE G. E. - The association of smoking with cancer of the urinary bladder in humans. *Arch. of Intern. Med.* 98 : 129-135, 1956.

+ (41) LILIENFELD A. M. - Emotional and other selected characteristics of cigarette smokers and nonsmokers as related to epidemiological studies of lung cancer and other diseases. *J. Nation. Canc. Inst.* 22, n°2 : 259-282, 1959.

++ (42) LOMBARD H. L., DOERING C. R. - Treatment of the fourfold table by partial association and partial correlation as it relates to public health problems. *Biometrics* 3 n° 3 ; 123-128, 1947.

(43) LOMBARD, POTTER - Epidemiological aspects of cancer of cervix. II. Hereditary and environmental factors. *Cancer* 3 : 960-968, 1950.

(44) LOWE C. R. - An association between smoking and respiratory tuberculosis. *Brit. Med. J.* : 1082-1086, 1956.

+ (45) MACHT S. H., LAWRENCE P. S. - National survey of congenital malformations resulting from exposure to roentgen radiation. *Amer. J. Roent. & Radiumtherapy* 73, n° 3, 1955.

+ (46) MACKLIN M. T. - Comparison of the number of breast-cancer deaths observed in relatives of breast-cancer patients, and the number expected on the basis of mortality rates. *J. Canc. Inst.* 22, n° 5 : 927-951, 1959.

++ (47) MAINLAND D. - The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. *Amer. Heart J.* 45, n° 5 : 644-654, 1953.

(48) MALIPHANT R. G. - The incidence of the cancer of the uterine cervix. *Brit. Med. J.*, n° 4613 : 978-981, 1949.

++ (49) MANTEL N., HAENSZEL W. - Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nation. Canc. Inst.* 22, n° 4 : 719-748, 1959.

(50) MILLS C. A., MILLS PORTER - Tobacco smoking habits and cancer of the mouth and respiratory system. *Cancer Research* 10, n° 9 : 539-542, 1950.

(51) MOSBECH J., VIDEBAEK A. - Mortality from and risk of carcinoma among patients with pernicious anaemia. *Brit. Med. J.* 2 : 390, 1950.

(52) MUSTACCHI - Cancer of the bladder and infestation with "schistosoma hematobium". *J. Nation. Canc. Inst.* 20 : 825-842, 1958.

++ (53) NEYMAN J. - Statistics - Servant of all sciences. *Science* 122 : 401-406, 1955.

(54) PEQUIGNOT G. - Enquête par interrogatoire sur les circonstances diététiques de la cirrhose alcoolique en France. *Bull. Inst. Nation. Hygiène* 13 : 719-739, 1958.

+ (55) SADOWSKY D. A., CORNFIELD J., GILLIAM A. G. - The statistical association between smoking and carcinoma of the lung. *J. Nation. Canc. Inst.* 13 : 1237-1258, 1953.

+ (56) SANGHVI L. D., RAO K. C. M., KHANOLKAR V. R. - Smoking and chewing of tobacco in relation to cancer of the upper alimentary tract. *Brit. Med. J.* i, n° 4922 : 1111-1114, 1955.

++ (57) SCHWARTZ D., ANGUERA G. - Une cause de biais dans certaines enquêtes médicales : le temps de séjour des malades à l'hôpital. *Comm. Inst. Intern. Statist.* 30è Session - Stockholm 1957.

(58) SCHWARTZ D., DENOIX P. F., ANGUERA G. - Recherches des localisations du cancer associées aux facteurs tabac et alcool chez l'homme. *Bull. du Cancer* 44, n° 2 : 336-361, 1957.

(59) SCHWARTZ D., DENOIX P. F., ROUQUETTE C. - Enquête sur l'étiologie des cancers génitaux de la femme. *Bull. du Cancer* 45, n° 4 : 476-493, 1958.

(60) SCHWARTZ D., DENOIX P. F. - L'enquête française sur l'étiologie du cancer broncho-pulmonaire - Rôle du tabac. *Semaine Hôpit. Paris*, n° 62/7 : 424-437, 1957.

++ (61) SHEPS M. C. - An examination of some methods of comparing several rates or proportions. *Biometrics* 15, n° 1 : 87-97, 1959.

(62) STEWART A., WEBB J., Coll. - Malignant disease in childhood and diagnostic irradiation in utero. *Lancet* 2 : 447, 1956.

(63) STOCKS P. - The epidemiology of carcinoma of the breast. *The Practitioner* 179 : 233-240, 1957.

(64) VIDEBAEK A., MOSBECH J. - Genetic causal factors in cancer of the stomach. *Danish Med. Bull.* 1, n° 7 : 189-193, 1954.

(65) WHITE, SHARBER - Tabac, alcool et angine de poitrine. *J. Amer. Med. Associat.* n° 102 : 655, 1934.

++ (66) WHITE C., BAILAR III J. C. - Retrospective and prospective methods of studying association in medicine. *Amer. J. Public Health & Nat. Health* 46, n° 1 : 35-44, 1956.

(67) WYNDER E. L., GRAHAM E. A. - Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma. *J. Amer. Med. Assoc.* 143, n° 4 : 329-336, 1950.

(68) WYNDER E. L., CORNFIELD J., SHROFF P. - A study of environmental factors in carcinoma of the cervix. *Amer. J. Obst. & Gynec. St-Louis* 68, n° 4 : 1016-1052, 1954.

(69) WYNDER E. L., BROSS I. J., DAY E. - Epidemiological approach to the etiology of cancer of the larynx. *J. Amer. Med. Assoc.* 160 : 1384-1391, 1956.

(70) WYNDER E. L., BROSS I. J., CORNFIELD J., O'DONNELL - Lung cancer in woman - A study of environmental factors. *New England J. Medic.* 225 : 1111-1121, 1956.

(71) WYNDER E. L., BROSS I. J., FELDMAN R. M. - A study of the etiological factors in cancer of the mouth. *Cancer* 10, n° 6 ; 1300-1323, 1957.