Théorie et pratique de l'échantillonnage: l'exemple de l'enquête MONICA

Wietlisbach Vincent

Institut universitaire de médecine sociale et préventive, Lausanne

1. Le projet MONICA

Introduction

Dans les pays développés, les maladies cardio-vasculaires, avec l'infarctus du myocarde en tête, sont responsables de la majeure partie des hospitalisations et des décès. La lutte contre ce fléau s'est organisée à toutes les étapes du processus évolutif des cardiopathies ischémiques: campagnes de prévention pour un mode de vie plus sain, amélioration du système de soins, application de techniques exploratoires et opératoires plus efficientes, mise sur le marché de nouveaux médicaments. L'épidémiologie, quant à elle, peut contribuer à déterminer certains paramètres essentiels de ce processus et, notamment, à mieux en connaître les tenants et les aboutissants. C'est ainsi qu'au stade asymptomatique de la maladie, les études épidémiologiques ont permis d'identifier certains facteurs de risque individuels: tension artérielle, cholestérol, tabac, style de vie, etc; le rôle que jouent ces facteurs, isolément et synergiquement, est cependant loin d'être élucidé. De même, par l'application de critères cliniques uniformes à l'échelle d'une population, l'épidémiologie peut servir à mieux définir le stade final du processus et à dégager de l'occurrence de tous les événements morbides associés ou apparentés l'incidence effective des cardiopathies ischémiques.

C'est dans cette perspective qu'a été conçu le projet MONICA (sigle tiré de «MONItoring of trends and determinants in cardiovascular disease»). Grâce à la collaboration de 39 centres répartis dans 26 pays, ce vaste projet, coordonné par l'OMS, a pour objectif principal de constituer une banque internationale de données standardisées et fiables qui permette une analyse approfondie des «entrées et sorties» dans le processus des maladies cardio-vasculaires. L'étude, d'une durée d'observation de dix ans, porte sur la population adulte de 25 à 64 ans. Chaque centre du projet, qui «couvre» une région géographique donnée, est chargé d'alimenter la banque de données par des opérations de prélèvement définies par le programme [1], dont les deux plus importantes sont:

- la réalisation de trois «examens de santé» de la population – en début, milieu et fin de période – qui consistent à mesurer les niveaux des différents facteurs de risque cardio-vasculaires auprès d'un échantillon aléatoire de sujets tiré de cette population;
- l'enregistrement continu et systématique, en fonction de critères cliniques uniformes et clairement définis, de tous les cas d'infarctus qui surviennent pendant cette période.

La Suisse compte un seul centre MONICA, divisé en deux unités fonctionnelles: l'unité de Lausanne qui couvre les cantons de Vaud et Fribourg, et l'unité de Bellinzone qui couvre le Tessin. Le programme complet du projet MONICA en Suisse a été décrit ailleurs de manière synthétique et détaillée [2]; le *tableau 1* présente le calendrier des opérations relatives à l'enregistrement des infarctus et à la mesure des facteurs de risque. Le projet a démarré en 1985 avec la réalisation du premier examen de santé des populations concernées.

Projet MONICA-CH
Saisie des données d'incidence et de niveau d'exposition

Données saisies	Auprès de qui?	Comment?	Quand?
Relevé des infarctus	Exhaustif	En continu	1984–1993
Mesure des facteurs de risque	Echantillons aléatoires indépendants	Transversal (à 3 reprises)	1. 1984–1985* 2. 1988–1989 3. 1992–1993

*Tab. 1 * Tessin: repoussé en 1985–1986.*

Cet article traite du plan d'échantillonnage qui a été appliqué à cette occasion; il expose d'abord les raisons théoriques et les contraintes pratiques qui ont motivé son choix, puis parle de la manière dont il a pu être réalisé concrètement sur le terrain. La discussion portera sur une appréciation critique de toute la procédure de sélection.

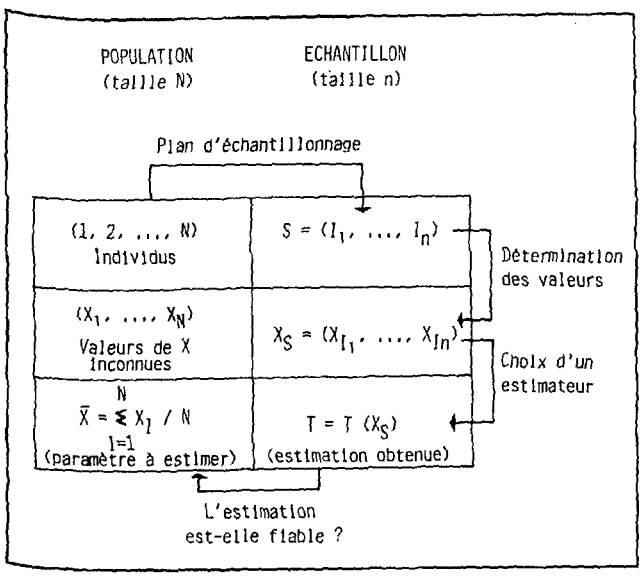
Design de l'étude MONICA

Le «design» du projet MONICA se caractérise par le fait que la mesure de l'incidence de la maladie se fait de manière longitudinale et complète, alors que celle des facteurs de risque, basée sur trois échantillons échelonnés dans le temps, est transversale et partielle. Ces deux types de relevés sont effectués de manière autonome; il n'existe donc pas en principe de système d'identification permettant d'assurer au niveau individuel la liaison entre le registre des infarctus et les fichiers de données relatifs aux examens de santé. Il en

résulte que l'étude MONICA n'a pas le «design» d'une étude prospective classique dans laquelle chaque sujet sélectionné est suivi individuellement, avec mesure périodique des facteurs de risque et relevé systématique des événements morbides.
En raison de ce «design» propre au projet MONICA, les données de l'examen de santé ne peuvent donc être recoupées avec celles du registre d'incidence qu'en fonction des catégories que permettent de définir les caractéristiques socio-démographiques des individus reportées dans les deux types de fichiers. En ce qui concerne les maladies cardio-vasculaires, l'âge et le sexe sont les plus déterminantes de ces caractéristiques, et c'est pour cette raison que les responsables du projet ont décidé de faire de la classe d'âge décennale (25–34 ans, 35–44 ans, etc), spécifique d'un sexe donné, l'unité de base des investigations.

Méthode d'inférence statistique
Comment mettre en relation, à ce niveau d'agrégation des individus, niveau d'exposition aux facteurs de risque et incidence de la maladie? Le nombre d'infarctus enregistrés peut être calculé avec précision pour chaque classe d'âge de la population, mais les facteurs de risque n'y sont mesurés que pour la partie des individus qui font partie de l'échantillon. Il est donc nécessaire de recourir à une méthode d'estimation qui, avec ces données partielles, permet de définir un indicateur fiable du niveau d'exposition globale de chaque catégorie considérée. Il s'agit là d'un problème d'inférence statistique, schématisé dans le *tableau 2*, qui doit être résolu en deux étapes distinctes, à savoir:
- la détermination d'un plan d'échantillonnage selon lequel les individus de l'échantillon (de taille n) sont tirés de la population (de taille N);
- la définition d'un estimateur T qui donne, à partir des seules informations de l'échantillon, l'approximation la plus fiable du niveau d'exposition moyen pour la population en général et pour chacune des catégories de population envisagées en particulier.

Tab. 2 Détermination d'un plan d'échantillonnage et définition d'un estimateur.



2. Théorie de l'échantillonnage

Qu'est-ce qu'un plan d'échantillonnage?
Un plan d'échantillonnage définit toute la procédure de tirage au sort des individus qui composent l'échantillon; il détermine complètement la loi de probabilité sur l'ensemble des échantillons possibles: à chaque échantillon s caractérisé par une combinaison spécifique de n individus choisis dans la population est associée une probabilité de réalisation p(s). Quelques types courants de plans d'échantillonnage sont exposés dans le *tableau 3*. Une fois qu'un tel plan d'échantillonnage est adopté, il est facile de calculer pour tout individu i de la population quelle est sa probabilité a(i) d'être inclus dans l'échantillon: il suffit d'additionner les probabilités p(s) de tous les échantillons qui comprennent l'individu en question. Cependant, l'inverse n'est pas vrai: la connaissance des probabilités individuelles d'inclusion ne suffit pas à déterminer un plan d'échantillonnage. Pour cela, il faut encore connaître les probabilités d'inclusion a(i,j) pour tous les couples d'individus possibles, les probabilités a(i,j,k) pour tous les tiercés possibles, et ainsi de suite.

Quelques propriétés des plans d'échantillonnage

Un plan d'échantillonnage P spécifie la loi de probabilité p(s) de sélection pour l'ensemble des échantillons s possibles

1. *Plan d'échantillonnage équilibré*
   Même probabilité de sélection pour tous les échantillons possibles (a fortiori pour tous les individus)
2. *Plan d'échantillonnage simple*
   Tous les individus sont extraits directement de la population
3. *Plan d'échantillonnage stratifié*
   Les individus sont extraits de différents sous-ensembles (strates) de la population

Tab. 3

Deux approches de l'estimation statistique
En statistique classique, le problème de l'estimation est toujours traité par rapport à une expérience aléatoire simple qu'il est possible de répéter à loisir de manière indépendante et dont on aimerait estimer un paramètre du processus stochastique. Le prototype de ce type d'expérience est le lancer du dé, qu'il faudra effectuer un certain nombre de fois pour déterminer par exemple la probabilité d'obtenir le 6 ou si le dé est équilibré. Dans un tel cadre théorique, il est souvent possible de définir un estimateur optimal qui soit à variance minimale autour du paramètre à estimer.
Lorsqu'on effectue un échantillonnage dans une population humaine, ce sont des individus *différents* qui

sont tirés à chaque fois. Si l'on adopte le point de vue darwiniste, chaque individu peut être considéré comme le produit du hasard lié à la recombinaison génétique. Dans cette approche, la technique de l'échantillonnage ne fait alors que sélectionner un certain nombre d'expériences *déjà réalisées* (mais perpétuellement recommencées dans la population), et permet par ce biais d'estimer certaines caractéristiques «innées» ou «intrinsèques» d'une population. Dans une telle optique, l'approche de la théorie classique est parfaitement justifiée.

Mais qu'en est-il lorsqu'il s'agit d'estimer à l'échelle d'une population des caractéristiques individuelles qui sont généralement acquises dans des circonstances particulières non-reproductibles? Dans de telles conditions, deux individus tirés au sort ne peuvent être considérés comme deux résultats d'une même expérience aléatoire. De même, lorsqu'un plan d'échantillonnage devient complexe et contient plusieurs niveaux de sélection, les tirages de deux individus ne sont plus statistiquement indépendants [3]. Dans les deux cas, le processus stochastique est constitué par la procédure de tirage de tout l'échantillon. Il faut donc définir un nouveau cadre théorique, celui d'une population finie d'individus *hétérogènes* pour laquelle on cherche à estimer la valeur moyenne *actuelle* qu'y prend une certaine variable. Celle-ci est supposée être fixée, au moment de l'enquête, pour chaque individu de la population; les erreurs de mesure, de même que les variations intra-individuelles, ne sont pas prises en compte dans un premier temps.

Les estimateurs optimaux n'existent pas dans un tel cadre théorique; les qualités qui sont requises pour un bon estimateur sont formulées dans le *tableau 4*. Il faut d'abord qu'il soit *centré;* en d'autres termes, cela revient à dire que, si l'on répète un grand nombre de fois le tirage d'un échantillon de taille fixée selon un même procédé aléatoire, les valeurs de l'estimateur obtenues lors de ces tirages successifs doivent osciller autour de la valeur (inconnue) que l'on cherche à estimer. Il faut ensuite qu'il soit *convergent,* c'est-à-dire que sa variance, son écart par rapport à la valeur à estimer, tende vers zéro lorsqu'on procède au tirage d'un échantillon de plus en plus grand. Cette propriété signifie concrètement que l'augmentation de la taille d'échantillon entraîne selon toute vraisemblance un gain de précision dans l'estimation. Il faut enfin que l'estimateur soit *admissible,* tel qu'il n'en existe pas un autre qui soit systématiquement plus fiable, quel que soit l'échantillon que l'on tire de la population.

Tab. 4

Propriétés d'un bon estimateur

1. *Il doit être centré*: sa valeur probable doit être égale à la valeur à estimer

$$E\left(T(Xs)\right) = \sum_s T(Xs)\, p(s) = \bar{X}$$

2. *Il doit être convergent*: sa précision s'améliore si la taille d'échantillon augmente

$$\mathrm{Var}\left(T(Xs)\right) = \sum_s \left(T(Xs) - \bar{X}\right)^2 p(s) \rightarrow O$$

Si $n \rightarrow N$

3. *Il doit être admissible*: il n'existe pas un autre estimateur To systématiquement plus précis, tel que:

$$\mathrm{Var}\left(To(Xs)\right) < \mathrm{Var}\left(T(Xs)\right)$$

Quelles que soient les valeurs de X

p(s) = probabilité qu'un échantillon s donné soit tiré

Quelques théorèmes sur les stratégies d'estimation

Ainsi, la mise en œuvre d'*une stratégie d'estimation* nécessite la réalisation d'un plan d'échantillonnage P et la définition d'un estimateur T calculé sur la base de l'échantillon tiré. Le *tableau 5* présente quelques théorèmes que la théorie statistique permet de formuler à propos des stratégies d'estimation. Les deux premiers théorèmes [4] démontrent les résultats suivants:

- *non-existence d'une stratégie optimale:* il n'y a pas d'estimateur qui, associé à un plan d'échantillonnage donné, produise systématiquement la meilleure estimation pour la valeur moyenne d'une variable. La stratégie optimale à adopter est tributaire des valeurs individuelles de la variable dans la population qui, par définition, sont inconnues a priori.

Propriétés d'une stratégie d'estimation

Plan d'échantillonnage P + estimateur T
= stratégie d'estimateur (T, P)

1. *Efficacité d'une stratégie*
   Est définie par la variable de l'estimateur

$$\mathrm{Var}(T, P) = \sum_s (T(Xs) - \bar{X})^2\, p(s)$$

2. *Non-existence d'une stratégie universelle optimale (To, Po)*
   Il existe un plan (T, P) tel que

$$\mathrm{Var}(T, P) < \mathrm{Var}(To, Po)$$

   pour certaines valeurs $(X_1 \ldots X_N)$ de la variable X dans la population

3. *Existence de stratégies admissibles*
   Par exemple, la moyenne d'échantillon et un plan d'échantillon équilibré est une stratégie admissible
   Il existe toujours une autre stratégie (T, P) telle que

$$\mathrm{Var}(\bar{X}s, P\ \text{équil.}) \leqslant \mathrm{Var}(T, P)$$

4. *Avantage de la stratification*

$$\mathrm{Var}(\bar{X}s, P\ \text{équil.}) \leqslant \mathrm{Var}(\bar{X}s, P\ \text{équilibré})$$

stratifié simple

Tab. 5

- *existence de stratégies admissibles:* il est possible de définir des stratégies qui ont la propriété *de n'être jamais les plus inefficaces!*
- *un plan d'échantillonnage équilibré, associé à la moyenne d'échantillon comme estimateur, est une stratégie admissible:* ce dernier théorème justifie le recours à une stratégie d'estimation simple et couramment utilisée dans la pratique.

Echantillonnage simple ou stratifié?

Lorsque la population peut être divisée en un certain nombre de groupes entre lesquels le niveau moyen d'un facteur de risque (ou de toute autre variable étudiée) est nettement différencié, on dit que cette répartition permet de stratifier la population en fonction du facteur de risque considéré et les groupes ainsi constitués sont appelés des strates. Un plan d'échantillonnage stratifié consiste alors non plus à tirer au sort les individus directement dans la population, mais à en tirer un certain nombre dans chacune des strates, comme on tirerait des boules dans plusieurs urnes. Il faut connaître deux théorèmes fondamentaux [5] à propos des stratégies d'estimation basées sur la moyenne d'échantillon et sur un plan d'échantillonnage stratifié:

- la stratégie optimale est obtenue lorsque l'échantillon est composé d'un nombre d'individus de chaque strate proportionnel à la variabilité du facteur étudié à l'intérieur de la strate. (théorème de l'allocation optimale de Neyman). Cependant, un tel plan d'échantillonnage ne peut pas être appliqué strictement, car les variances intra-strates demeurent évidemment inconnues a priori. Cependant, il est possible que l'enquêteur dispose pour ces variances d'estimations tirées d'une étude antérieure.
- un plan d'échantillonnage stratifié, quelle que soit la stratification employée, donne généralement une meilleure estimation que le plan d'échantillonnage simple, pour autant que la taille de l'échantillon soit assez grande.

Ces deux théorèmes parlent en faveur d'un plan d'échantillonnage équilibré et stratifié en fonction d'un critère dont on sait qu'il est habituellement corrélé au facteur de risque étudié.

3. Choix du plan d'échantillonnage

Pour le choix d'un plan d'échantillonnage adapté à la réalisation d'un «examen de santé» valable pour l'ensemble de la population, le modèle de référence est une enquête américaine d'envergure nationale qui est très bien documentée [6].

Une fois soigneusement échafaudées toutes les propriétés idéales que devrait posséder un plan d'échantillonnage, il s'agit ensuite d'examiner à quelles conditions il peut être réalisé concrètement. Le choix définitif du plan d'échantillonnage qui sera appliqué devra essayer de concilier au mieux les exigences d'ordre théorique et les contraintes d'ordre pratique, dont les plus courantes sont exposées ci-dessous.

Contrainte d'ordre économique et détermination de la taille d'échantillon

- C0: *La taille d'échantillon doit être de 3300 pour la région de Vaud et Fribourg, de 2000 pour le Tessin*

D'un point de vue pratique, la principale qualité exigée d'un plan d'échantillonnage est qu'il coûte le moins cher possible, ce qui implique en particulier qu'il atteigne la précision d'estimation voulue avec une taille d'échantillon relativement petite. Dans l'enquête MONICA, ce sont pas seulement les *niveaux* des facteurs de risque qui doivent être estimés correctement, mais aussi leurs *variations* dans le temps. En effet, il est important de pouvoir déterminer dans quelle mesure l'évolution de la morbidité et de la mortalité cardio-vasculaire peut être expliquée par l'évolution concomitante des facteurs de risque.

Le problème est que le protocole du projet MONICA stipule que les trois examens de santé prévus doivent être faits sur des échantillons aléatoires *indépendants*. Cela signifie qu'à chaque fois un nouveau tirage au sort est effectué et qu'il est peu probable qu'un même individu soit sélectionné pour plus d'un examen. Une des raisons de ne pas faire porter deux examens de santé consécutifs sur le même collectif est que, dans ce cas, les individus risquent, à la suite du premier examen, de modifier leurs comportements dans un sens qui ne reflèterait pas forcément la tendance générale de la population. D'un autre côté, il est possible de démontrer [7] que le plan d'échantillonnage basé sur des tirages indépendants est nettement moins économique (en termes de taille d'échantillon) que celui basé sur le tirage unique d'une cohorte examinée à plusieurs reprises.

Le *tableau 6* montre quelle taille d'échantillon est nécessaire si l'on veut être assuré à 80 % de détecter dans la population une variation absolue dx d'un facteur de risque si celle-ci se produit (tout en se préservant à

Nombre de sujets qu'il est nécessaire de convoquer à un «check up» périodique pour déceler de manière fiable* une modification dx d'un facteur de risque dans la population

Taille échantillon	Taux de fumeurs	Tension diastolique	Cholestérol total
N = 100	dx = 19,3 %	dx = 4,0 mm	dx = 15,8 mg %
200	13,7 %	2,8 mm	11,2 mg %
300	11,2 %	2,3 mm	9,1 mg %
400	9,7 %	2,0 mm	7,0 mg %

*Tab. 6. * Test contre l'hypothèse nulle «dx = 0» avec α = .05, β = .20 et nouvelle sélection de sujet à chaque examen.*

95 % contre le fait que l'échantillon indique une variation d'une telle ampleur lorsqu'en réalité, il ne se passe pas grand-chose à l'échelle de la population!). L'option recommandée par les experts du projet MONICA est de fixer à 200 le nombre d'individus à examiner pour chaque classe d'âge décennale spécifique d'un sexe, qui va constituer le plus bas niveau de désagréga-

tion de toutes leurs analyses. Compte tenu de ces exigences de fiabilité statistique et des ressources financières locales, les responsables suisses du projet ont décidé de le faire porter sur la population âgée de 25 à 74 ans pour la région de Vaud et Fribourg, de 35 à 64 ans pour le canton du Tessin. A ces conditions, en supposant un taux de participation de 60 % (hypothèse qui s'est avérée très réaliste), la taille globale de l'échantillon s'est finalement élevée à 3300 personnes pour la région romande et à 2000 pour la région italienne *(tableau 7)*.

Enquête MONICA
Population-cible et taille d'échantillon

Région	Population totale*	Critères de sélection	Population-cible* (N)	Taille échantillon (n)	Fraction d'échantillon (n/N)
	727592		*443034*	*3300*	
VD+FR	VD: 537942 FR: 108790	Résidents 25–74 ans	VD: 334244 FR: 108790	VD: 2500 FR: 800	.0074
Tessin	272867	Résidents 35–64 ans	106598	2000	.0187

*Tab. 7. * Estimation pour le milieu de 1984.*

Contraintes d'ordre administratif

– C1: *L'échantillonnage doit se faire à deux niveaux, d'abord communal, puis individuel*

La réalisation effective du plan d'échantillonnage exige évidemment que l'on se renseigne au préalable sur la disponibilité, l'accessibilité et la structure des fichiers de population indispensables à ce genre d'exercice. Par exemple, pour aucun des cantons participant au projet MONICA (Vaud, Fribourg, Tessin), il n'existe de fichier centralisé des habitants, étant donné que leur contrôle est du ressort des administrations communales. Cet état de fait implique donc que l'échantillonnage doit nécessairement se faire à deux niveaux, d'abord par une sélection appropriée d'un certain nombre de communes, puis par tirage au sort des individus dans leur fichier des habitants respectifs.

A titre d'information, les cantons de Vaud, de Fribourg et du Tessin sont composés respectivement de 385, 266 et 191 communes.

Contraintes d'ordre logistique

– C2: *Un minimum de 7 personnes doit être tiré par commune*

D'autres contraintes ont dû être imposéees a priori sur le plan d'échantillonnage pour des raisons pratiques d'organisation. La réalisation de l'examen de santé dans les communes sélectionnées nécessite en effet à chaque fois le déplacement de l'équipe médicale, formée de deux personnes, la mise à disposition d'une salle d'examen et son aménagement fonctionnel. Toutes ces opérations coûtent cher en temps et en argent et, par conséquent, il faut éviter qu'elles ne soient répétées trop de fois. C'est pourquoi il a été décidé de tirer un nombre minimal de sujets dans chacune des communes sélectionnées.

Contraintes d'ordre théorique

– C3: *La proportion maximale d'individus à tirer dans une commune ne doit pas excéder un tiers de son effectif total de personnes éligibles pour le projet MONICA*

Si l'on suppose que les habitants d'une petite commune ont acquis une sorte de «spécificité» locale (ce qui peut signifier une faible variance de certaines caractéristiques individuelles), l'allocation optimale de Neyman pour les échantillons stratifiés suggère alors de limiter le nombre maximum d'individus à sélectionner dans de telles communes pour ne pas altérer la représentativité générale de l'échantillon. Certaines enquêtes rapides faites par les instituts de sondage excluent d'emblée les communes de moins de 2000 habitants; certaines méthodes d'estimation, basée sur une moyenne d'échantillon pondérée, attribuent aux effectifs sélectionnés dans les petites communes le poids d'un individu unique. Une option moins radicale a été retenue ici.

– C4: *Le plan d'échantillonnage doit être équilibré*

Cette condition est basée sur le critère d'admissibilité exposé dans la partie théorique; elle implique notamment que le tirage des individus à l'intérieur de chaque commune doit également être équilibré. Le choix d'un plan d'échantillonnage particulier ne peut donc plus porter que sur la manière de sélectionner les communes.

Quelques plans d'échantillonnage envisagables

Le *tableau 8* expose trois manières différentes de procéder à la sélection des communes: tirage simple, tirage proportionnel à la taille de la commune, tirage stratifié. Le premier de ces modes consiste à tirer les communes comme autant de boules distinctes mises dans une urne; toutes les communes ont la même chance d'être choisies et dans chacune d'entre elles, une même proportion d'individus est tirée. Cette procédure risque d'aboutir à une grande concentration géographique si une grande commune est sélectionnée ou à un trop grand nombre de communes à tirer s'il ne sort que des petites communes. Le deuxième mode de sélection impose un contingent fixe d'individus à tirer par commune, ce qui peut entraîner la surreprésentation des petites communes (voir contrainte C3). Aussi, en dernier ressort, le choix du plan d'échantillonnage s'est-il porté sur le tirage stratifié des communes, qui a l'avantage de la souplesse nécessaire pour s'adapter aux contraintes C1 à C4 imposées a priori.

Il faut remarquer que d'autres plans d'échantillonnage ont déjà été appliqués en Suisse, notamment le dénommé «Berner Stichprobenplan» [8] à l'occasion de l'enquête SOMIPOPS (Système des Indicateurs SOCio-médicaux de la POPulation Suisse), première grande enquête de population sur la santé réalisée

ENQUETE MONICA : SELECTION DES COMMUNES

Propriétés de quelques modes de tirage

MODE DE TIRAGE	PROBABILITE DE SELECTION DES COMMUNES	PROBABILITE DE SELECTION D'UN INDIVIDU (A L'INTERIEUR D'UNE COMMUNE)	NOMBRE DE COMMUNES A TIRER	NOMBRE D'INDIVIDUS A TIRER DANS UNE COMMUNE	INCONVENIENTS PROBABLES
SIMPLE	EGALE POUR TOUTES LES COMMUNES	EGALE DANS TOUTES LES COMMUNES	DEPEND DE LA TAILLE DES COMMUNES TIREES	PROPORTIONNEL A LA TAILLE DE LA COMMUNE	SEULEMENT DES GRANDES COMMUNES OU BEAUCOUP DE PETITES
PROPORTIONNEL	PROPORTIONNEL A LA TAILLE	INVERSEMENT PROPORTIONNEL A LA TAILLE DE LA COMMUNE	FIXE	LE MEME POUR TOUTES LES COMMUNES	TROP D'INDIVIDUS SELECTIONNES DANS LES PETITES COMMUNES
STRATIFIE	EGALE POUR LES COMMUNES D'UNE MEME STRATE	EGALE DANS LES COMMUNES D'UNE MEME STRATE	FIXE A PRIORI DANS CHACUNE DES STRATES	DANS CHAQUE STRATE PROPORTIONNEL A LA TAILLE	CHOIX ARBITRAIRE DE LA STRATE

Tab. 8

dans le pays [9]. Ce plan d'échantillonnage équilibré consiste en une sorte de tirage systématique d'individus dans un fichier fictif mettant bout à bout tous les fichiers communaux des habitants, avec un seuil minimal de personnes à tirer par commune. Néanmoins, sa réalisation n'a pas donné une bonne représentativité de l'échantillon par rapport aux variables socio-démographiques, mais il est possible que cette inadéquation soit attribuable à des difficultés et des erreurs intervenues dans l'exécution des différents tirages au hasard.

Choix de la stratification

La répartition des communes ne peut se faire qu'en fonction d'un facteur qui soit connu pour chacune d'entre elles. C'est la taille de la commune, à savoir son nombre total d'habitants, qui a été choisie pour trois raisons. D'abord, il s'agit là d'une donnée démographique publiée chaque année; ensuite, cette variable assure une relative dissémination géographique de l'échantillon; enfin, elle est corrélée plus ou moins à d'autres facteurs associés aux maladies cardio-vasculaires: âge, style de vie, milieu urbain ou rural, etc. Les strates formées en fonction de la taille de la commune sont définies dans le *tableau 9*.

La première strate est constituée par les villes (communes d'au moins 10000 habitants), qui ont toutes été incluses d'office dans l'échantillon, d'une part parce que l'effectif d'une ville représente une fraction trop importante de la population globale pour être laissée de côté, d'autre part parce qu'un fichier urbain des habitants permet de sélectionner en un seul tirage un nombre appréciable d'individus. Les autres strates ont été formées par ordre décroissant de la taille des communes, de manière à ce que chaque strate possède environ le même nombre d'individus. Cette règle permet de concilier au mieux la condition d'une même probabilité de sélection pour tous les individus avec la règle de l'allocation optimale de Neyman.

ENQUETE MONICA

Stratification des communes selon leur taille

	STRATE	TAILLE DE LA PLUS GRANDE COMMUNE	EN % DE LA POP. TOTALE	DANS LA STRATE IL FAUT TIRER ... COMMUNES PARMI LES ...	DANS CHAQUE STRATE, IL FAUT TIRER ENVIRON 1 INDIVIDU SUR ...
VAUD	1.	127'000	48%	9/9	1/135
	2.	9'500	11%	4/9	1/60
	3.	4'800	11%	4/14	1/39
	4.	2'900	10%	4/28	1/19
	5.	1'300	10%	4/68	1/8
	6.	500	10%	8/257	1/4
FRIBOURG	1.	37'000	20%	1/1	1/135
	2.	7'500	19%	4/9	1/60
	3.	1'500	19%	4/44	1/12
	4.	500	19%	5/170	1/3
	5.	Région alémanique	23%	4/42	1/13
TESSIN	1.	17'000	22%	3/3	1/53
	2.	8'500	20%	4/9	1/24
	3.	4'800	20%	4/20	1/11
	4.	1'800	20%	4/48	1/5
	5.	750	18%	10/167	1/3

Tab. 9

Détermination du nombre de communes à tirer dans chaque strate

Le nombre total de communes à tirer a été fixé en fonction des disponibilités de temps, d'argent et de personnel des deux unités fonctionnelles du projet MONICA. La détermination du nombre de communes à tirer dans chaque strate s'est faite de manière à ce que le plan d'échantillonnage respecte au mieux les conditions C0 à C4. Pour que tout individu de la population ait la même chance d'être sélectionné (condition C4), ce qui exige que sa commune soit d'abord tirée au sort dans la strate, puis lui-même dans la commune, il est nécessaire que la relation suivante soit vérifiée:

- le produit de la fraction d'échantillonnage des communes par celle des individus dans les communes doit être égal dans toutes les strates.

Il est intéressant de noter que cette probabilité n'est pas strictement égale à la fraction d'échantillonnage n/N, comme cela eût été le cas dans un plan d'échantillonnage simple; sa valeur n'est déterminée qu'une fois la sélection des communes effectuée.

Détermination du nombre d'individus à tirer dans chaque commune

La seule indication fournie par le plan d'échantillonnage à ce sujet est la *proportion* d'individus à tirer dans chaque commune sélectionnée. Aussi, pour calculer le numérateur de cette fraction, c'est-à-dire le nombre de personnes à tirer, est-il nécessaire de connaître le dénominateur, à savoir le nombre de sujets qui, à ce moment-là, sont éligibles pour l'enquête MONICA. En d'autres termes, le problème est de disposer pour 1984 des effectifs âgés de 25 à 74 ans pour les communes romandes, de 35 à 64 ans pour les communes tessinoises. Dans leur état actuel, les statistiques démographiques ne fournissent pas ces chiffres, dont la détermination la plus récente a été établie par le recensement fédéral de 1980. Ainsi, pour chaque commune, le nombre de personnes éligibles a été estimé par une méthode d'extrapolation, en appliquant à l'effectif total – connu annuellement – la structure par âge observée en 1980 et vieillie de quelques années.

4. Réalisation du plan d'échantillonnage

Il est très important que chaque étape de l'échantillonnage soit réalisée de manière à ce que les procédés de tirage effectivement employés correspondent au plus près à la procédure spécifiée par le plan. Si des inadéquations se produisent à ce niveau, les résultats de l'enquête seront en principe entachés d'un biais d'échantillonnage.

Tous les tirages au sort ont été effectués à l'aide d'un générateur de nombres pseudo-aléatoires installé sur une petite calculatrice programmable.

Tirage des communes

Une fois le tirage des communes dans les différentes strates effectué, il était nécessaire d'obtenir par la voie officielle l'autorisation de chaque conseil communal d'accéder au fichier des habitants. Cette requête n'a été refusée qu'une seule fois, et il a fallu alors procéder au tirage d'une commune de remplacement.

Tirage au sort des individus

Le tirage au sort des individus dans les fichiers communaux est une entreprise plus délicate, étant donné la grande diversité du support et de la structure interne. Certains d'entre eux sont informatisés, les autres sont des cartothèques constituées parfois de fiches individuelles, parfois de «fiches de ménages». Il faut donc particulièrement veiller à ce que la stratégie de tirage appliquée soit adaptée à la structure particulière de chaque fichier et que tous les individus éligibles pour l'enquête MONICA y aient la même probabilité de sélection.

Beaucoup de responsables des fichiers communaux ont proposé un tirage à l'aveugle d'un certain nombre de cartes. Il est certain que cette sorte de «randomisation sauvage» est souvent moins hasardeuse qu'une méthode savante entachée d'un biais systématique plus ou moins subtil. Néanmoins, il ne faut pas confier en principe le tirage au sort aux gestionnaires des fichiers pour éviter que des critères de sélection trop subjectifs entrent en ligne de compte. Il est arrivé plusieurs fois que le préposé à un fichier conseille aimablement au sélectionneur externe de ne pas retenir tel individu choisi parce qu'il n'accepterait jamais de participer à un examen ou parce que son cas serait sans intérêt pour le thème de l'enquête.

Tirage simple ou tirage systématique?

Le tirage simple consiste à générer une suite de nombres aléatoires et à extraire du fichier les individus qui correspondent à ces nombres. Ce mode de tirage exige que chaque individu puisse être repéré dans le fichier à l'aide d'un numéro. Quelques manières de créer un tel système sont exposées plus bas.

Le tirage systématique consiste à extraire tous les $x^{èmes}$ individus du fichier, à commencer par un premier individu dont le rang est tiré au sort entre 1 et x. Une telle procédure a l'avantage de n'avoir besoin que d'un seul nombre aléatoire; il suffit ensuite de progresser dans le fichier et d'en tirer toutes les $x^{èmes}$ fiches. Néanmoins, cette opération peut s'avérer vite fastidieuse si le fichier contient beaucoup de sujets non éligibles (qu'il s'agit alors de sauter) ou si les fiches ne sont pas individuelles (il faut alors compter les individus fiche après fiche).

Un tirage systématique risque de tomber sur une période de tirage qui se recoupe avec un ordre structurel du fichier; ainsi, dans le cas extrême d'un fichier composé uniquement de paires de cartes relatives à des couples (le mari étant placé avant la femme), un tirage systématique ne sélectionnerait dans chaque commune que des individus d'un seul sexe. Un tel risque est cependant relativement minime.

De plus, bien que le tirage systématique soit un plan d'échantillonnage équilibré, il ne permet pas en prin-

cipe d'aboutir exactement à la taille d'échantillon désirée; la taille effective, selon le rang de la première carte choisie, peut s'en écarter d'un individu. Comme cet individu surnuméraire sera toujours tiré ou non tiré en fin de fichier, la catégorie des individus généralement classée à cette place finira par être sous-représentée si le tirage systématique est appliqué à un grand nombre de fichiers pareillement structurés. Un tel phénomène s'est peut-être produit pour l'enquête SOMIPOPS dans laquelle l'échantillon souffrait d'un manque de femmes. Une des explications envisagées en est que le tirage au sort des individus s'est fait dans les registres électoraux des communes, où les femmes sont placées systématiquement après les hommes.
En conséquence de tout cela, le tirage systématique a été pratiqué dans les fichiers homogènes ne contenant que des individus éligibles (par exemple, les listes alphabétiques fournies par certains contrôles des habitants informatisés); dans les fichiers plus complexes, c'est un tirage simple qui a été pratiqué.

Tirage dans les fichiers avec indexation numérique des individus (fichiers informatisés)
Dans ce type de fichier, chaque individu peut être repéré rapidement grâce à un numéro d'identification. Il suffit alors de connaître l'intervalle de variation de ces numéros, d'y générer une série de nombres aléatoires et de tirer les individus correspondants aux nombres sortis. Si l'un de ces derniers ne se rapporte à aucun individu ou à un individu non-éligible, il faut passer au nombre suivant.

Tirage dans les fichiers sans indexation numérique (cartothèques)
Si un fichier de ce type est petit, il est possible de tirer au sort le rang des individus à sélectionner, puis de les repérer dans le fichier par simple comptage des fiches. Cependant, si une cartothèque est volumineuse, une telle procédure devient trop longue à exécuter. Il faut alors adopter un autre mode de tirage, comme par exemple de mesurer la longueur totale du fichier, tiroir après tiroir, à l'aide d'une règle graduée, générer aléatoirement des longueurs intermédiaires, repérer en nombre de millimètres (toujours avec la règle) les positions du fichier correspondants et en extraire une fiche par position.

Tirage dans des fichiers non-individuels
Beaucoup de contrôles des habitants communaux sont basés sur les «fiches de ménage» qui contiennent dans le cas d'une famille le mari, la femme et les enfants mineurs. Le tirage comprend nécessairement deux étapes; il porte d'abord sur les fiches, puis sur les individus des fiches. Une manière de procéder un tel tirage est de «jouer» à pile ou face (pile, on prend, face on laisse) tous les individus de toutes les fiches tirées. Une autre manière consiste à choisir aléatoirement le rang de l'individu à prendre, en comptant les individus à partir du premier de la fiche tirée.

Si un individu sélectionné n'est pas éligible, il faut recommencer un nouveau tirage aléatoire indépendant
Dans la ville où s'est effectué le premier tirage, les «échantillonneurs» attitrés (dont l'auteur) ont pu sélectionner les individus en fournissant à l'ordinateur de gestion du fichier une liste aléatoire de numéros d'identification; chaque fois que l'individu tiré n'appartenait pas à la population-cible de l'enquête MONICA, ils faisaient défiler le fichier dans l'ordre croissant des numéros séquentiels jusqu'à tomber sur un candidat apppropríé. Après contrôle de l'échantillon quant à sa répartition par sexe, ils se sont aperçus à leur grande stupeur que les hommes y étaient nettement surreprésentés simplement parce que, dans le fichier, les membres d'une même famille étaient entrés de manière consécutive dans l'ordre mari-femme-enfants!

5. Evaluation du plan d'échantillonnage

D'un point de vue strictement probabiliste, la qualité d'un plan d'échantillonnage proprement dit ne peut être jugée que sur son degré d'adéquation à la population en fonction de variables pour lesquelles la distribution est connue aussi bien pour l'échantillon que pour la population. En pratique, d'autres critères d'appréciation entrent en ligne de compte, comme le côut du plan d'échantillonnage en rapport avec la précision statistique obtenue. Cependant, une telle évaluation nécessiterait une analyse comparative approfondie de plusieurs plans d'échantillonnage qui dépasserait le cadre de cet article.
Les seules caractéristiques en fonction desquelles il est possible de confronter l'échantillon obtenu à la population de départ sont de type socio-démographique: sexe, classe d'âge, nationalité, état-civil. Pour les populations-cible des cantons de Fribourg et du Tessin, la répartition des individus selon ces variables n'y est pas connue pour les années intercensitaires. Cependant, en faisant évoluer les effectifs observés au recensement de 1980 à l'aide de la statistique des décès, ventilée quant à elle par tous les facteurs en question, il est possible d'aboutir à une bonne estimation de leur distribution de 1984. Par exemple, la proportion d'hommes dans la population-cible est ajustée à 49,4% pour la région MONICA de Vaud et Fribourg, à 49,2% pour le Tessin, alors que les échantillons correspondants donnent une proportion estimée de 48,9% et 48,1%. Un test statistique accepte l'hypothèse d'égalité entre les taux estimés et les taux «vrais» à un niveau de signification de 0,6 et 0,3 respectivement.
Le *graphique 1* montre comment l'adéquation entre échantillon et population en fonction de la classe d'âge est réalisée de manière satisfaisante pour chaque sexe et chaque région MONICA. Le *graphique 2* compare selon ce même critère l'enquête MONICA à deux autres enquêtes portant sur les femmes vaudoises, l'une visant leur pratique de l'auto-examen du sein [10], l'autre cherchant à estimer la prévalence des facteurs de risque associés au cancer du sein [11].

L'échantillonnage de ces trois études a été réalisé simultanément selon le même plan pour des raisons de moindre coût de réalisation. Il apparaît très clairement dans le graphique que l'augmentation de la taille d'échantillon entraîne un remarquable gain dans l'adéquation, ce qui est une preuve concluante de fiabilité pour la procédure de sélection décrite dans cet article.
L'étape suivante de l'enquête MONICA consiste à contacter les individus sélectionnés et à les persuader de se soumettre à l'examen de santé. La démarche employée pour atteindre cet objectif est décrite ailleurs [12]. Le *graphique 1* anticipe les résultats obtenus dans cette deuxième phase et montre que les participants contribuent parfois à aggraver, parfois à corriger l'inadéquation entre l'échantillon intial et la population. Aussi le plan d'échantillonnage ne donne-t-il que les conditions initiales d'une bonne étude: la qualité de celle-ci, en fin de compte, dépendra de sa conception globale et de la réalisation du plan d'ensemble.

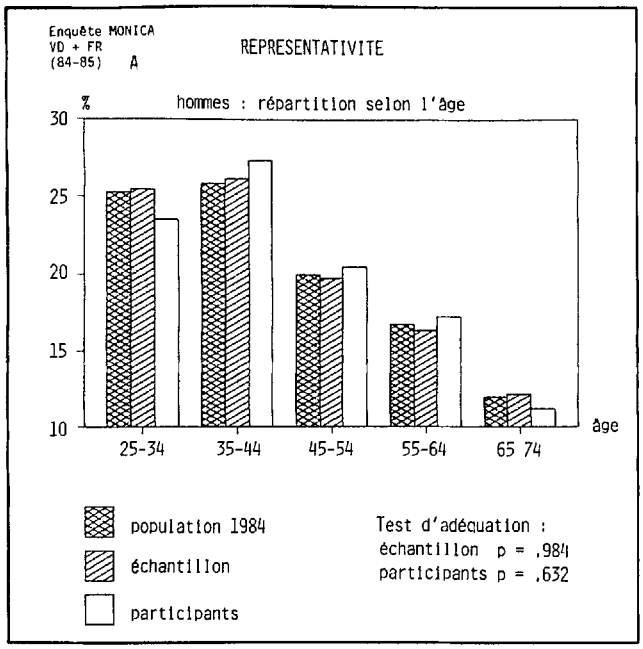


Fig. 1A. Hommes, Vaud et Fribourg. Représentativité selon l'âge.

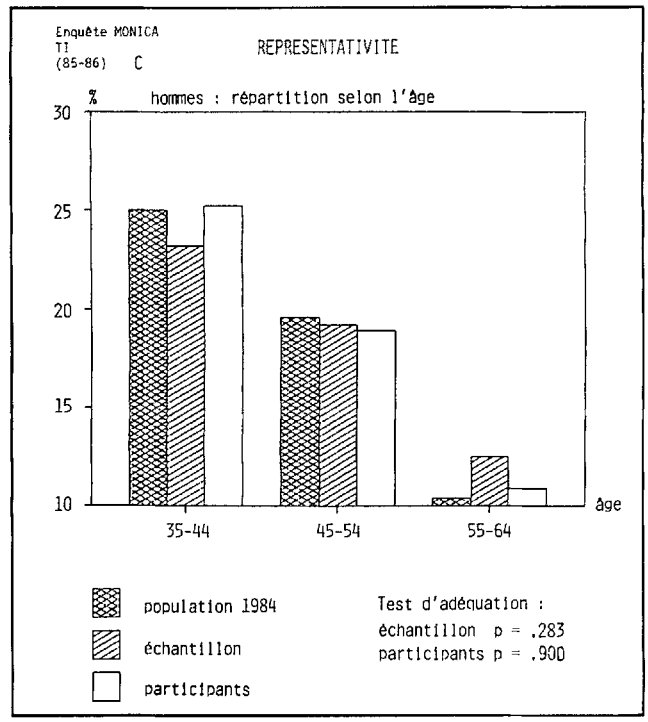


Fig. 1C. Hommes, Tessin. Représentativité selon l'âge.

Fig. 1B. Femmes, Vaud et Fribourg. Représentativité selon l'âge.

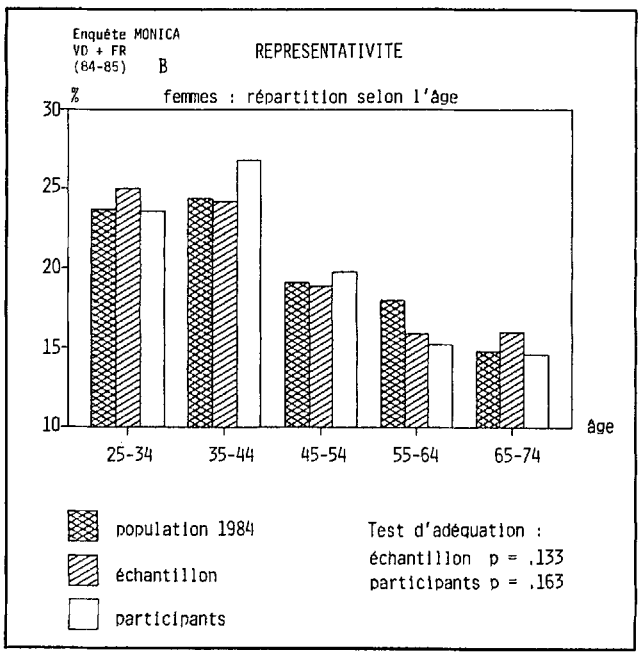
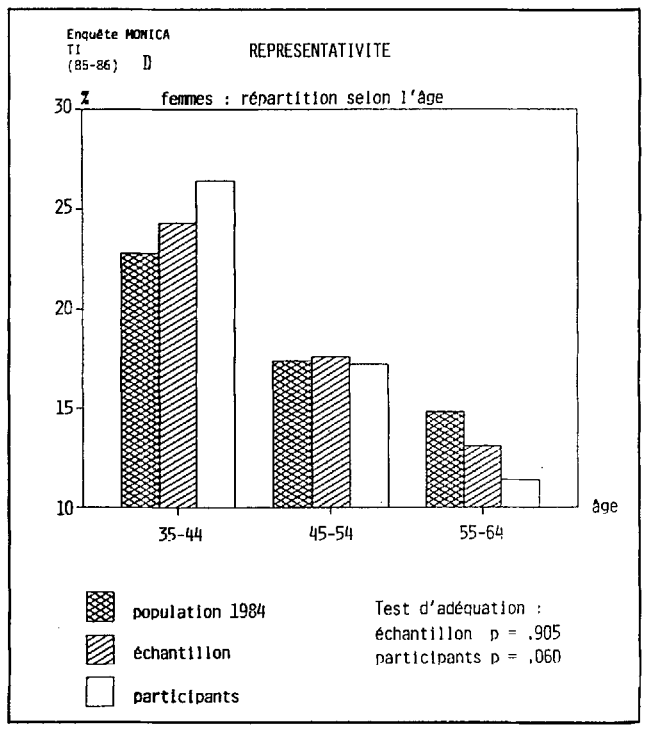


Fig. 1D. Femmes, Tessin. Représentativité selon l'âge.



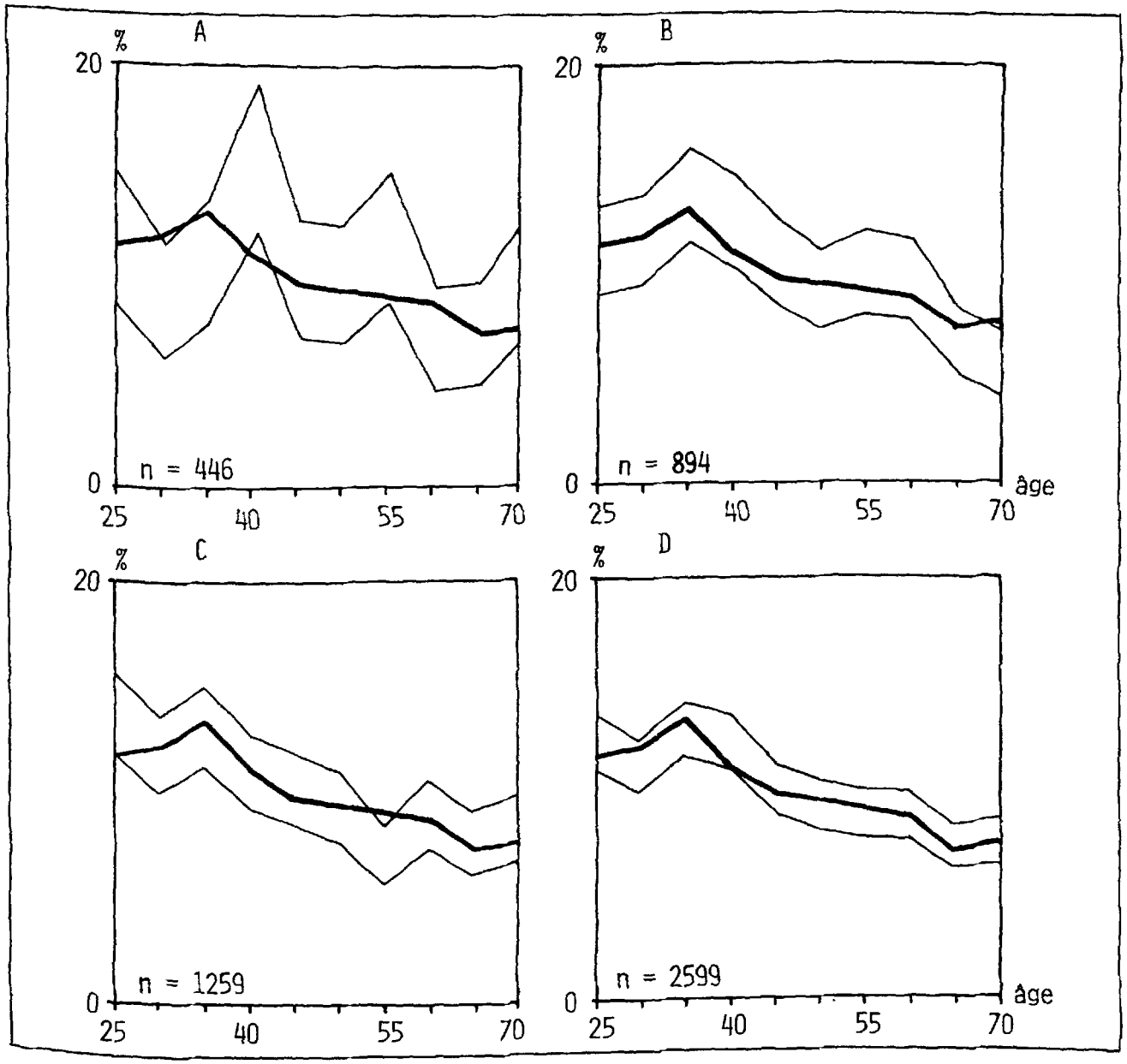


Fig. 2. Vraie distribution en classes d'âge quinquennales (en trait gras: population féminine vaudoise de 1984) et précision de l'estimation obtenue (intervalle de confiance à 95 %) pour trois enquêtes réalisées selon le même plan d'échantillonnage.
A. Enquête «auto-examen des seins». B. Enquête «facteurs de risque pour le cancer du sein». C. Enquête MONICA (Vaud). D. Ensemble des trois enquêtes.

Résumé

Parallèlement à l'enregistrement systématique des cas d'infarctus, le projet MONICA prévoit de mesurer à trois reprises le niveau des facteurs de risque cardio-vasculaires auprès d'un échantillon aléatoire. L'article présente le plan d'échantillonnage de ce premier «examen de santé» MONICA réalisé dans les cantons de Vaud, de Fribourg et du Tessin. Il s'agit d'un plan à deux niveaux, avec tirage stratifié des communes en fonction de leur taille, puis tirage des individus dans les fichiers communaux. Les conditions d'un plan d'échantillonnage efficace dans le cadre plus général de l'inférence statistique sont abordées dans une première partie théorique. Les raisons pratiques (contraintes budgétaires, problèmes de logistique, disponibilité des fichiers administratifs) qui ont motivé le choix de ce plan sont exposées ensuite. Une troisième partie décrit toutes les étapes de sa réalisation, avec les difficultés méthodologiques et concrètes rencontrées. La discussion porte sur une évaluation critique de toute la procédure qui, dans le cadre du projet MONICA, a produit des échantillons dont le degré d'adéquation avec la population est assez élevé.

Zusammenfassung

Theorie und Praxis der Stichprobenbildung: das Beispiel des MONICA-Projektes.

Neben einer systematischen Erfassung der Infarktfälle sieht das MONICA-Projekt eine dreimalige Erfassung der Risikofaktorenprävalenz vermittels einer Zufallsauswahl vor. Der vorliegende Artikel schildert den Stichprobenplan der ersten in den Kantonen

Waadt, Freiburg und Tessin durchgeführten Untersuchung. Es handelt sich um einen zweistufigen Stichprobenplan: zuerst wurde eine Ziehung der Gemeinden – stratifiziert nach Gemeindegrösse – vorgenommen, danach wurden die Individuen auf Grund der Einwohnerregister gezogen. Die Grundbedingungen für einen effizienten Stichprobenplan werden in einem ersten theoretischen Abschnitt diskutiert. Im weiteren werden die konkreten Bedingungen der MONICA-Stichprobenziehung dargestellt (Budget-Limiten, organisatorische Probleme, Aufbau der Einwohnerregister). Ein dritter Teil beschreibt sämtliche Schritte der eigentlichen Stichprobenbildung einschliesslich der aufgetretenen Schwierigkeiten. Die Diskussion nimmt eine Gesamtbeurteilung der Stichprobenziehung im MONICA-Projekt vor, deren Ergebnis recht befriedigend ausfällt.

Summary

Theoretical and practical aspects of sampling: the MONICA-Project.

In parallel with the systematic registration of myocardial infarction, the MONICA-Project attempts to investigate at three different times the prevalence of risk factors for cardiovascular disease in the population. This article presents the sampling plan of the first MONICA survey in the cantons of Vaud, Fribourg and Tessin. The sampling procedure was at two levels: first, a sample of communes stratified according to community size was chosen, and secondly, within these communities, individuals were selected from the population registries. The prerequisites for an efficient sampling plan are discussed on a theoretical level. In addition, the practical constraints (budget, organizational problems, population registry files) are presented. Finally, all steps of the sampling procedure are described including the difficulties encountered. The discussion attempts a critical evaluation of the whole MONICA sampling procedure whose results are largely satisfactory.

Bibliographie

[1] *Tunstall-Pedoe H.* Monitoring trends in cardio-vascular disease and risk factors: the WHO «MONICA» project. WHO Chronicle 1985; *39:* 3–5.

[2] *Rickenbach M, Gutzwiller F, Wietlisbach V, Martin J, Epstein FH.* Switzerland's participation in MONICA. Soz Praeventivmed 1985; *30:* 95–99.

[3] *Kish l, Frankel MR.* Inference from complex sample. J Roy Statistic Soc 1974; *36* (Ser B): 1–37.

[4] *Cassel CM, Sarndal CE, Wretham CE.* Foundations of Inference in Survey Sampling, Chap.3. New York: Wiley and Sons, 1977.

[5] *Cochran WG.* Sampling Techniques, Chap 5. New York: Wiley and Sons, 1977.

[6] U.S. Department of Health, Education and welfare. Plan and initial program of the Health Examination Survey. Vital and Health Statistics: National Center for Health Statistics 1965; Ser 1: 0–4.

[7] *Junod B, Wietlisbach V.* Méthodes et stratégies d'évaluation du programme national suisse de recherche sur la prévention des maladies cardio-vasculaires. Rev Epidém Santé Publ 1981; *29:* 315–25.

[8] *Fritschi P, Meyer R, Schweiter W.* Ein neuer Stichprobenplan für ein gesamtschweizerisches Sample. Rev. Suisse Sociol 1976; *3:* 149–58.

[9] SOMIPOPS Collaborative Group (Gutzwiller F, Leu RE et al). The Swiss Health Survey Project (SOMIPOPS): an example of a data collection from various sources. Soz Praeventivmed 1985; *30:* 76–79.

[10] *Huguenin M, Wietlisbach V, Martin J, Meystre-Agustoni G.* Attitudes et pratiques des femmes vaudoises vis-à-vis de l'examen des seins. Soz Praeventivmed 1985; *30:* 157–61.

[11] *Stalder JB, De Grandi P, Huguenin M, Wietlisbach V.* Prévalence des facteurs de risque du cancer du sein dans la population féminine vaudoise. Délimitations des indications au dépistage. Méd Hyg 1986; *44:* 1526–32.

[12] *Wietlisbach V, Hausser D, Barazzoni F, Rickenbach M.* Enquête MONICA: analyse de la participation. Soz Praeventivmed 1987; *32.*