LA PRÉSENTATION DES RÉSULTATS DES RÉGRESSIONS LOGISTIQUES sous forme de probabilités estimées (application à la pratique contraceptive en France)

Laurent Toulemon

La régression logistique est de plus en plus en plus souvent utilisée dans le traitement des enquêtes sociodémographiques. Elles permet de tirer parti, pour des variables qualitatives définissant des catégories ordonnées ou non, des avantages des régressions : estimation et test des « effets spécifiques » de différentes dimensions « explicatives » (exogènes) sur un comportement « à expliquer» (endogène), dans le cadre d'un modèle causal.

Cette méthode est cependant l'objet de deux critiques. D'une part, le schéma causal sous-jacent conduit parfois à des modèles artificiels, quand on soupçonne que certaines dimensions considérées comme explicatives dépendent elles-même du comportement étudié, que les différentes dimensions explicatives ne peuvent être mises sur le même plan pour «séparer» leurs effets, ou que des dimensions explicatives importantes sont omises dans le modèle ; ce premier reproche est souvent fondé, mais il est commun à toutes les méthodes de régression (régression linéaire, analyse de variance, modèles à risques proportionnels, régression logistique) et ne sera pas discuté ici. D'autre part, les régressions logistiques conduisent à des paramètres ininterprétables pour le lecteur profane. Cet inconvénient pourrait être évité si les résultats étaient présentés sous forme de probabilités ajustées, directement comparables aux proportions brutes observés, quand la variable «endogène» est qualitative.

Nous supposerons que la variable endogène est dichotomique : les valeurs possibles sont 0 (absence du comportement étudié) ou 1 (présence du comportement). La régression logistique estime, pour chaque individu, la probabilité du comportement étudié. Les variables exogènes, «dimensions explicatives», sont transformées en blocs de variables explicatives, elles aussi à valeur dans [0,1] [Marpsat, Verger 1991]. Est-il possible de présenter l'effet spécifique de chaque dimension explicative sous forme de probabilités estimées au sein de chaque groupe, sous l'hypothèse que la dimension explicative est seule à avoir un effet sur le comportement étudié ? À cette question, la réponse est résolument «oui», à condition d'accepter l'échelle logistique, dont il faut d'abord montrer la pertinence pour les variables dichotomiques. Les différentes échelles possibles sur les proportions seront comparées dans un premier temps, avant de préciser les conventions nécessaires pour présenter les résultats des régressions

logistiques sous la forme « facile à lire » de probabilités estimées. Les comportements de contraception en France en 1978 et 1988 serviront d'illustration.

L'échelle logistique : échelle naturelle pour les proportions ?

En faisant pour le moment abstraction des méthodes de régression, on peut définir un grand nombre d'*échelles* pour comparer différentes proportions[1]. Soient différents groupes au sein desquels on observe la fréquence d'un comportement (que nous appelons « succès ») à l'aide de la proportion d'individus ayant ce comportement. Assimilons la proportion de succès à la probabilité de succès au sein de chaque groupe, supposé homogène. Pour expliciter les échelles possibles, supposons que la probabilité (notée p) au sein de chaque groupe est fonction d'une *propension au succès* (notée x) qui suit une échelle linéaire : par exemple l'écart entre les propensions 0 et 1 est le même qu'entre 1 et 2, et l'écart entre 0 et 2 est le double de l'écart entre 0 et 1. Chaque échelle sur les proportions peut alors être assimilée à une fonction liant la probabilité p à la propension x supposée suivre une échelle additive.

La comparaison de deux groupes ne pose pas de difficultés, et on conviendra que le groupe où la proportion de succès est la plus élevée est celui pour lequel la propension au succès est la plus forte. La relation entre x et p doit donc être croissante. La question devient plus délicate dès que l'on veut comparer plus de deux proportions : l'écart correspondant aux proportions 10 % et 55 % est-il plus ou moins important qu'entre 1 % et 10 % ? La réponse n'est pas triviale, et des questions de ce type se posent à l'occasion de n'importe quel travail sur des proportions, c'est-à-dire souvent, et chacun a l'habitude d'utiliser implicitement une ou plusieurs échelles sur les proportions, échelles nécessaires pour savoir si une évolution s'accélère ou ralentit, si les écarts entre deux groupes augmentent ou diminuent, etc. Trois types d'échelles seront comparés ici : les échelles linéaires, multiplicatives, et logistiques.

L'échelle linéaire

Si la proportion de succès p dépend linéairement de la propension x, on a (a et b étant des nombres réels) :

(1). Cette discussion reprend un débat qui a eu lieu dans la Revue Française de Sociologie entre 1984 et 1988, au sujet de l'évolution de l'inégalité sociale devant l'école en France, et auquel ont participé Cibois, Combessie, Grémy, Prévot, Merllié. Voir [Vallet 1988] et ses références bibliographiques. Un récent article de Zarca [1993] sur l'héritabilité de l'indépendance professionnelle présente des écarts entre écarts, des évolutions d'écarts ... en utilisant extensivement l'échelle logistique.

$$p = a\,x + b$$

$$\frac{dp}{dx} = a$$

Le paramètre a est positif, pour que p soit une fonction croissante de x. Sous cette contrainte, les valeurs de a et b sont des paramètres de taille et de position, et on ne perd aucune généralité en fixant ($a = 1$) et ($b = 0$).

Une telle échelle implique que l'écart entre deux groupes $[x_2 - x_1]$ se mesure par la *différence* $[p_2 - p_1]$. Par exemple, l'écart entre deux groupes sera de 9 si les proportions sont respectivement 1 % et 10 %, 10 % et 19 %, 50 % et 59 %, ou 94 %... et 103 %, ou encore -8 % et 1 %. Cette échelle convient pour les proportions proches de 50%, mais elle n'est guère satisfaisante pour les proportions proches de 0 % ou de 100 %. On pressent que le passage de 1 % à 10 % correspond à un écart en termes de propension plus forte que le passage de 10 % à 19 %, et on souhaiterait estimer la probabilité correspondant au même écart de 9 entre un groupe où la proportion est 1 % et un groupe où la propension au succès est plus faible, sans aboutir à une proportion négative. De même, la différence entre 81% et 90 % est difficilement recevable comme équivalente à celle séparant 90 % et 99 %.

Les échelles multiplicatives

La proportion p varie de manière exponentielle selon x :

$$p = \exp\,(ax + b)$$

$$\frac{dp}{dx} = a\,p$$

On peut, ici encore, supposer ($a = 1$ et $b = 0$). L'écart $[x_2 - x_1]$ se mesure alors par le *rapport*

$$[\,\exp(x_2) - \exp(x_1)\,] = \frac{p_2}{p_1}.$$

Ainsi des écarts identiques séparent les proportions 1% et 10 %, 10 % et 100 %. La variation de p selon x est satisfaisante pour p proche de zéro (les proportions 1%, 2 %, 4 %, 8 %, 16 % correspondent au même écart), mais perd sa pertinence pour les comportements fréquents.

L'échelle multiplicative n'est pas symétrique, et une autre échelle multiplicative peut être définie à partir de la probabilité d'échec, complémentaire de la probabilité de succès :

$$1 - p = \exp(-a\,x - b)$$

$$\text{soit } p = 1 - \exp(-a\,x - b)$$

$$\frac{dp}{dx} = a\,(\,1 - p\,)$$

Cette échelle est satisfaisante pour les comportements fréquents, mais non pour les comportements rares.

L'échelle logistique

Aucune des échelles précédentes ne permet de comparer des proportions variant sur un large spectre : l'échelle multiplicative semble convenir pour les petites valeurs de p, l'échelle additive pour les valeurs moyennes, et l'échelle multiplicative sur le complément à un pour les grandes valeurs de p. L'échelle logistique combine les avantages des échelles précédentes, là où elles sont pertinentes. Elle est définie par :

$$p = \frac{1}{1 + \exp(-ax - b)}$$

$$\frac{dp}{dx} = a\,p\,(\,1 - p\,)$$

L'équation ci-dessus, illustrée par la figure 1, montre que l'échelle logistique est proche d'une échelle multiplicative en p si p est proche de zéro $\left[\frac{dp}{dx} = ap\right]$ et en (1 - p) si p est proche de 1 $\left[\frac{dp}{dx} = a\,(1 - p)\right]$, tandis qu'elle est presque additive pour les valeurs moyennes de $p\left[\frac{dp}{dx} = \frac{a}{4}\right]$. Cette échelle conduit à définir, en posant (a = 1) et (b = 0), le *logit* d'une proportion p par :

$$x = \text{logit}\,(p) = ln\left(\frac{p}{1 - p}\right)$$

le symbole ln représentant le logarithme népérien. Si p est observé comme le rapport $\frac{n_1}{n}$, où n et n_1 sont des effectifs, on peut estimer le *logit* x par :

$$ln\left(\frac{n_1 + 0{,}5}{n - n_1 + 0{,}5}\right)$$

afin d'obtenir une estimation définie sur l'intervalle [0,1] [Cox 1972]. Nous supposerons que les proportions observées sont comprises strictement entre 0 et 1. L'écart entre deux groupes $[x_2 - x_1]$ s'écrit alors :

$$\text{logit}(p_2) - \text{logit}(p_1) = ln\left[\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}\right]$$

Cette mesure est le logarithme du *rapport des chances* entre les groupes (1) et (2), appelé *odds-ratio* en anglais, et utilisé intensivement par les épidémiologistes. Les rapports des chances sont toujours analysés sur des échelles multiplicatives, ce qui revient à considérer que les *logits*, définis comme leurs logarithmes (en anglais *log-odds*), suivent une échelle linéaire. Les contrastes seront donc les mêmes pour des proportions de 1% et 10%, 10% et 55%, 55% et 93%.

Les justifications de l'échelle logistique sont nombreuses. Cette échelle est proche de la transformation fondée sur la loi normale [Marpsat, Trognon 1992], elle mesure la variation du poids relatif de deux populations dont la croissance est exponentielle [Deville, Naulleau 1982], ou la diffusion d'une épidémie si les rencontres entre individus ont lieu au hasard, selon une loi uniforme du temps. La sensibilité de p selon x est proportionnelle à la variance binomiale de p, une diffusion par contagion conduit à une évolution temporelle logistique, et toute échelle vérifiant quelques contraintes simples d'exhaustivité et de symétrie doit lui ressembler [Cox 1960], [Vallet 1988].

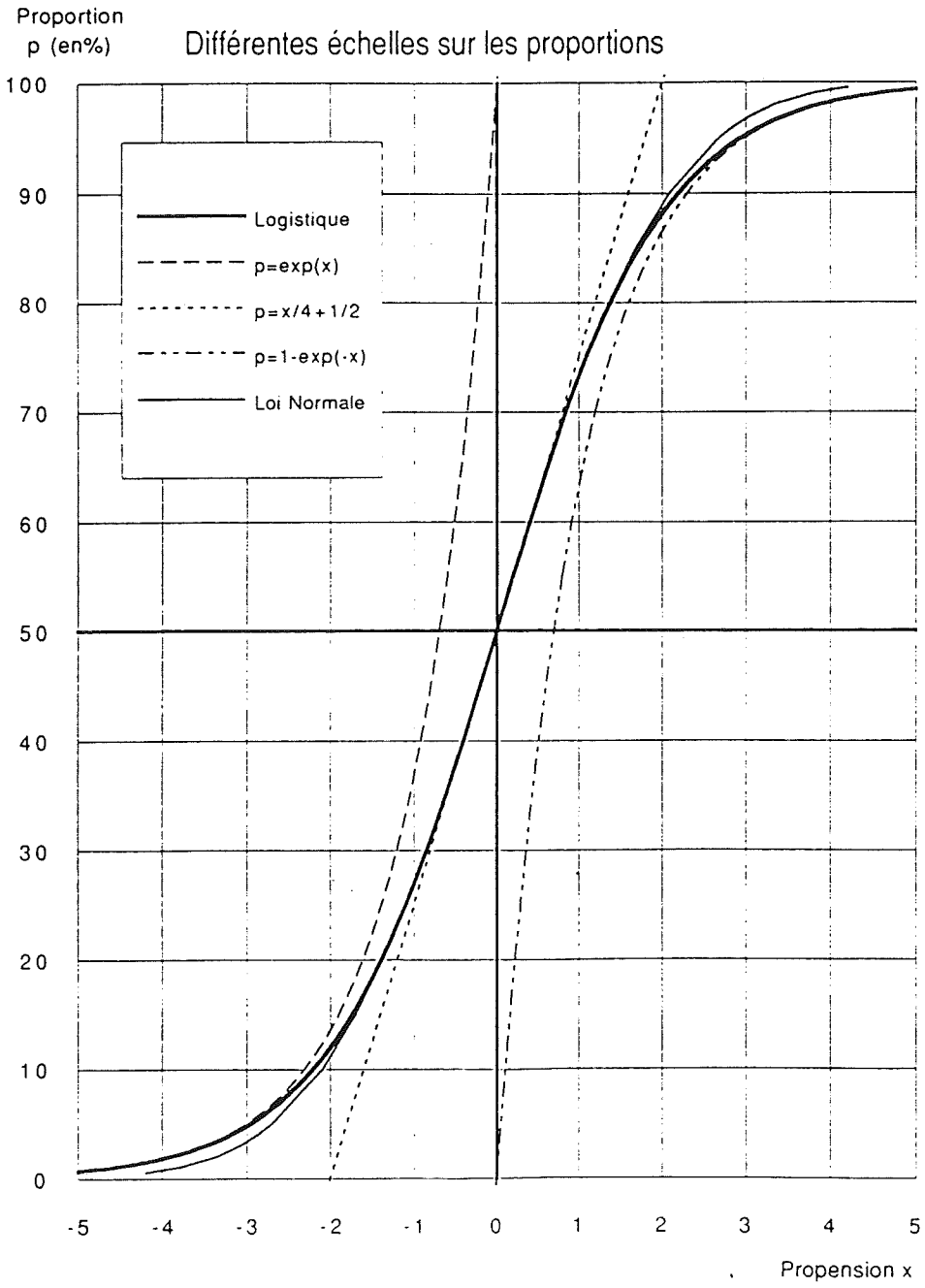
L'échelle logistique apparaît plus « intuitive » si l'on garde à l'esprit qu'elle est presque multiplicative entre 0 % et 20 %, presque additive entre 20 % et 80 %, et presque multiplicative en *(1- p)* entre 80 % et 100 % (*figure 1*). Les trois séries de proportions suivantes (en %) correspondent à des écarts logistiques à peu près égaux à ln(2) = 0,7 et constants, et peuvent servir de jalons pour comparer, en première approximation, différentes proportions :

1- 2 - 4 - 8 - 16 - 30 - 50 - 70 - 84 - 92 - 96 - 98 - 99
2,5 - 5 - 10 - 20 - 35 - 55
45 - 65 - 80 - 90 - 95 - 97,5

Les papiers spécifiques

L'utilisation de l'échelle logistique peut se faire de deux manières équivalentes. La première consiste à calculer le logit de chaque proportion, et à placer ces logits sur un papier millimétré. Les contrastes logistiques apparaissent alors directement (par exemple, une diffusion « à vitesse constante » correspond à un contraste logistique entre deux

Figure 1


Proportion
p (en%)
Différentes échelles sur les proportions
100
90
80
70
60
50
40
30
20
10
0
Logistique
p=exp(x)
p=x/4+1/2
p=1-exp(-x)
Loi Normale
-5
-4
-3
-2
-1
0
1
2
3
4
5
Propension x

périodes successives constant, et la courbe des logits selon le temps est une droite. On dira que la diffusion s'accélère si la courbe est convexe, et qu'elle se ralentit si la courbe est concave).

Il est encore plus simple d'utiliser un papier semi-logit, dont l'échelle est millimétrique en abscisse, et logistique en ordonnée. La *figure 2* présente un tel papier semi-logit. Ce type de papier n'existe pas dans le commerce, mais il peut être remplacé par le papier semi-normal (normal en ordonnées), puisque les fonctions de répartition des lois logistiques et normales sont proches (voir la *figure 1*). Le papier semi-normal « écrase » les contrastes pour les proportions comprises entre 0% et 5% (et entre 95 % et 100 %), ce qui ne modifie pas les interprétations, nécessairement prudentes pour des proportions extrêmes.

La comparaison de deux séries de proportions (par exemple des variations d'un comportement selon l'âge ou toute autre variable, lors de deux observations successives) montre directement si un contraste s'est accru ou a diminué. Le papier logit-logit (*figure3*) correspond à ce type de comparaisons, puisque des écarts logistiques constants se traduisent par une distance à la première diagonale constante. Cependant, ce quadrillage est difficile à manier, et une autre solution, très simple à mettre en pratique, est préférable.

Les courbes d'« iso-contraste logistique », sont des hyperboles[1] dans le plan p_1,p_2, définies par [$logit(p_2) - logit(p_1) = constante$]. Il est alors possible d'effectuer des comparaisons sur un papier millimétré habituel, en traçant les proportions d'hyperboles correspondant à des différences logistiques données. Les hyperboles sont faciles à tracer « à la main », dans la mesure où l'équation [$logit(p_2) - logit(p_1) = a$] correspond aux portions de droite suivantes (*figure 4*) :

$$\left[\frac{p_2}{p_1} = \exp(a)\right] \text{ si } p_1 \text{ et } p_2 \text{ sont prochesde 0.}$$

$$\left[\frac{1-p_1}{1-p_2} = \exp(a)\right] \text{ si } p_1 \text{ et } p_2 \text{ sont prochesde 1, et}$$

$$[p_2 - p_1 = u], \text{ avec } u = 2\frac{\exp(a/2) - 1}{\exp(a/2) + 1}, \text{ si } p_1 + p_2 \text{ est proche de 1,}$$

$$\text{soit } [p_2 - p_1 = \frac{a}{4}], \text{ si de plus a est faible } (p_2 - p_1 < 0{,}3 \; ; \; a < 1{,}2)$$

(1). L'équation [$logit(p_2) - logit(p_1) = a$]définit une portion de la droite [$p_2 = p_1$] si a = 0, et une portion de l'hyperbole d'asymptotes [$p_1 = c + 1$] et [$p_2 = -c$] , d'équation [$(p_1 - (c+1))\ (p_2 + c) = -c(c+1)$], avec $c = \frac{1}{exp(a) - 1}$, si a est négatif (pour a positif, on intervertit p_1 et p_2).

Figure 2

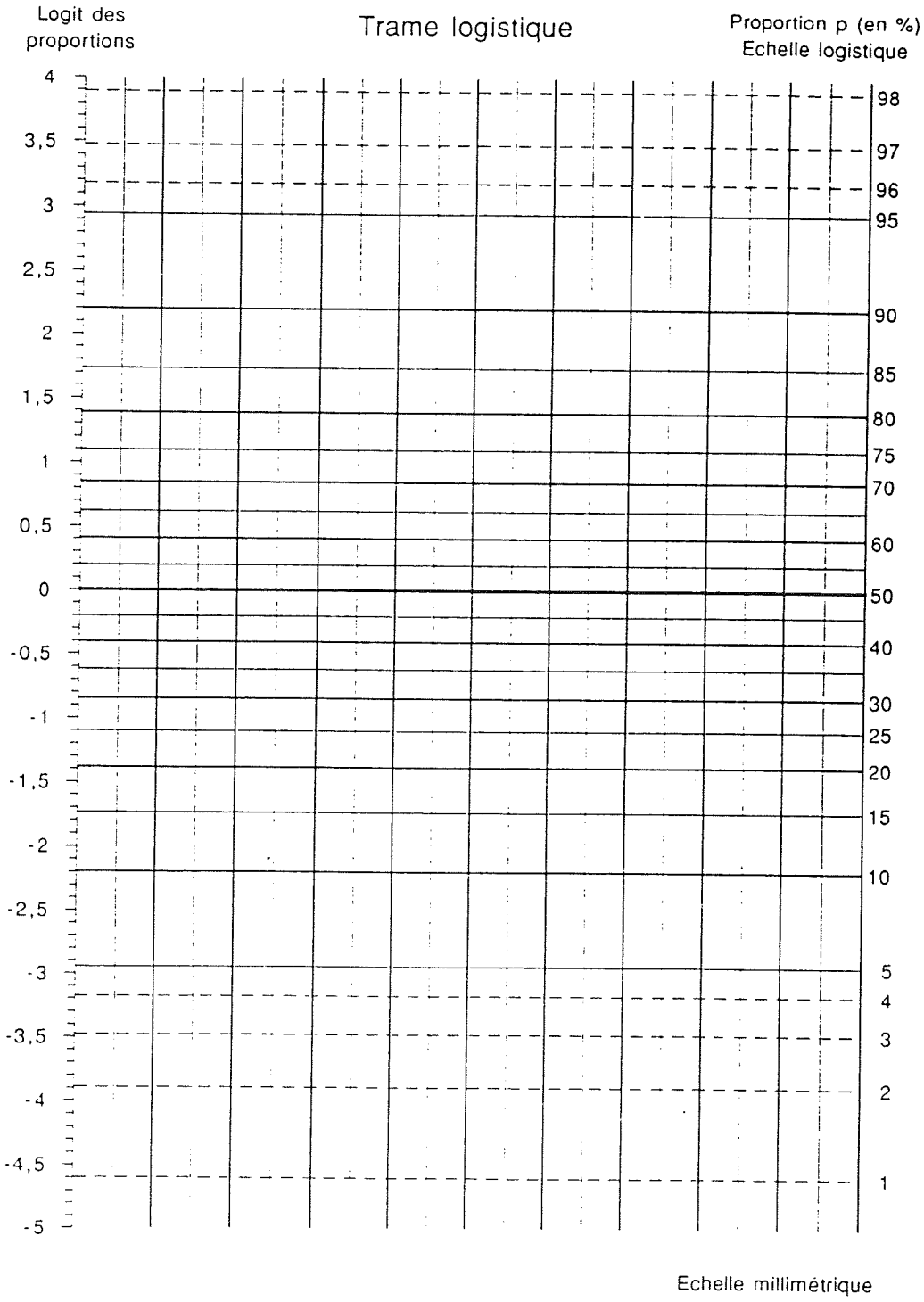

Logit des proportions
Trame logistique
Proportion p (en %) Echelle logistique
4
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
-1
-1,5
-2
-2,5
-3
-3,5
-4
-4,5
-5
98
97
96
95
90
85
80
75
70
60
50
40
30
25
20
15
10
5
4
3
2
1
Echelle millimétrique

Figure 3

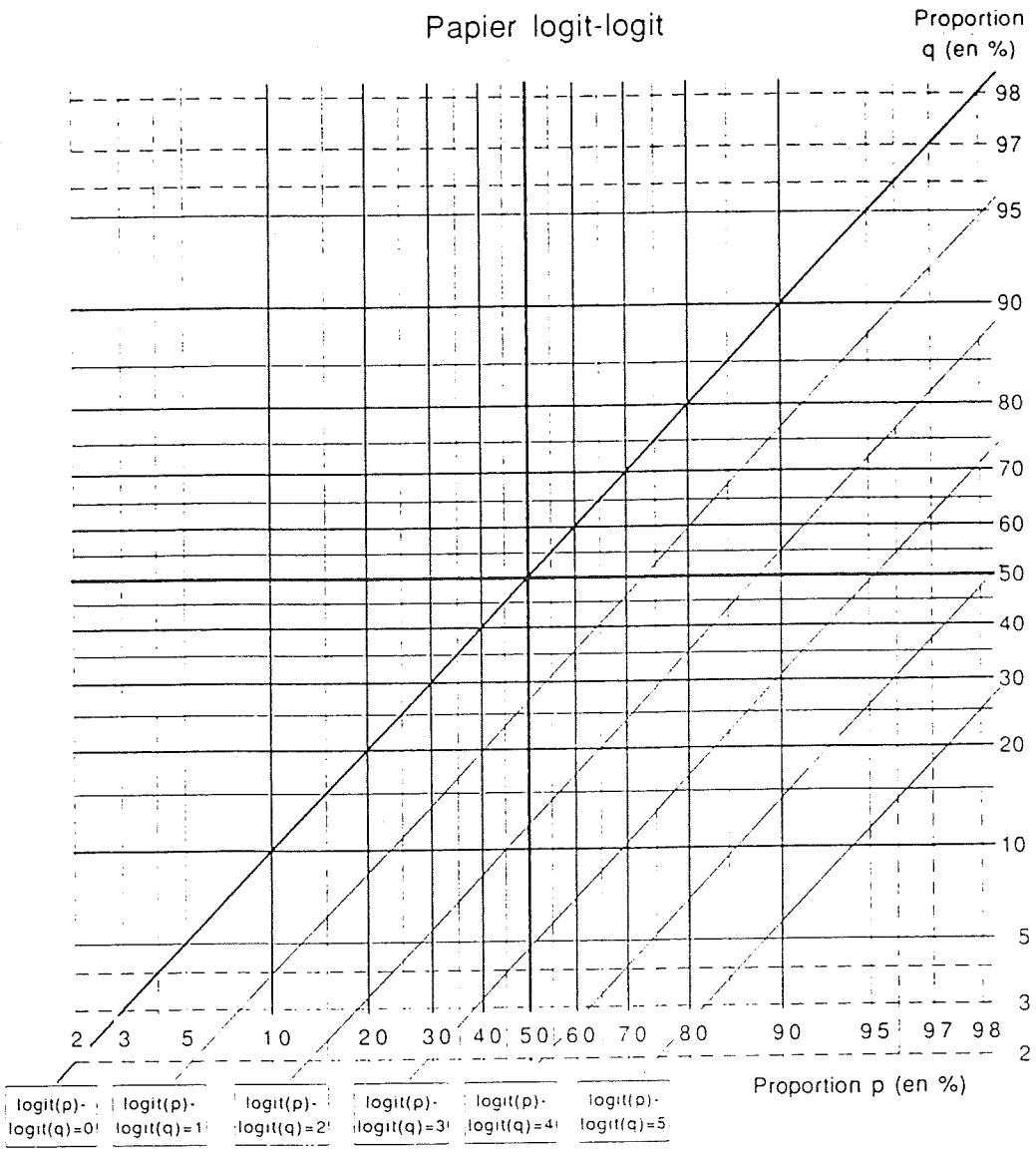
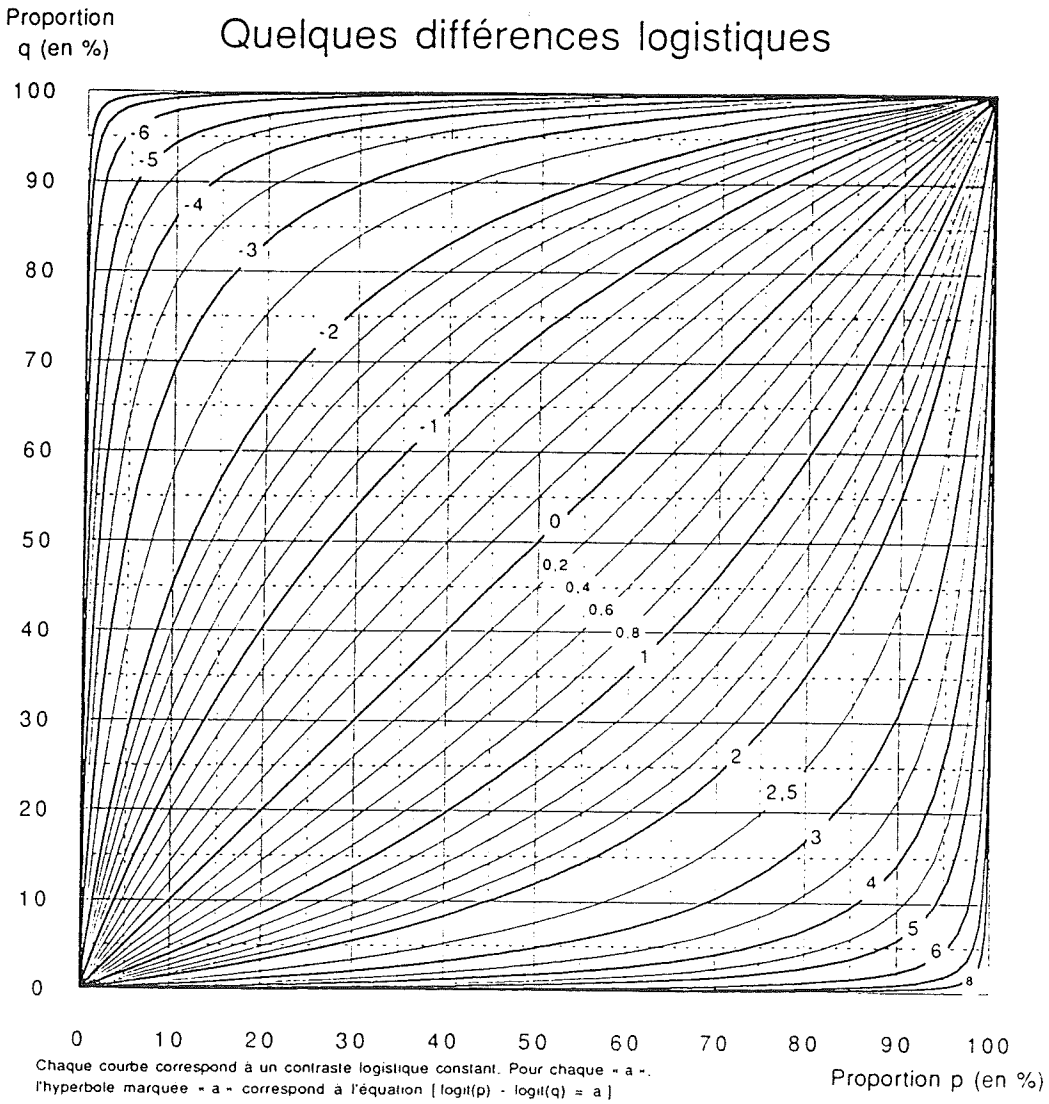

Papier logit-logit
Proportion q (en %)
98
97
95
90
80
70
60
50
40
30
20
10
5
3
2
2 3 5 10 20 30 40 50 60 70 80 90 95 97 98
Proportion p (en %)
logit(p)-logit(q)=0
logit(p)-logit(q)=1
logit(p)-logit(q)=2
logit(p)-logit(q)=3
logit(p)-logit(q)=4
logit(p)-logit(q)=5

Figure 4


Proportion
q (en %)
Quelques différences logistiques
100
90
80
70
60
50
40
30
20
10
0
-6
-5
-4
-3
-2
-1
0
0,2
0,4
0,6
0,8
1
2
2,5
3
4
5
6
8
0
10
20
30
40
50
60
70
80
90
100
Chaque courbe correspond à un contraste logistique constant. Pour chaque « a »,
l'hyperbole marquée « a » correspond à l'équation [logit(p) - logit(q) = a]
Proportion p (en %)

Un exemple : l'utilisation des méthodes médicales de contraception

L'INED a réalisé en 1978, dix ans après la légalisation de la contraception en décembre 1967, une enquête sur les comportements de fécondité et de contraception en France, en collaboration avec l'INSEE, auprès de près de 3 000 femmes âgées de 20 à 44 ans. Une enquête similaire a eu lieu en 1988, en collaboration avec l'INSEE et l'INSERM. Près de 3 200 femmes âgées de 18 à 49 ans ont été interrogées en 1988 [Toulemon, Leridon 1991].

Les données présentées partiellement ici concernent les proportions d'utilisatrices des méthodes médicales de contraception (pilule ou stérilet) parmi les femmes soumises « au risque d'une grossesse non souhaitée », c'est-à-dire les femmes fertiles qui ont des rapports sexuels et ne souhaitent pas devenir enceintes. Les résultats sont issus d'un travail qui décrit les comportements contraceptifs au sein des différents groupes sociaux (définis par la PCS, le diplôme, la taille de la commune d'habitation et l'importance attachée à la religion), les variations dues aux variables démographiques ayant été éliminées grâce à des régressions logistiques [Toulemon, Leridon 1992]. Ce sont ces variations selon les variables démographiques (âge, situation de couple, nombre d'enfants, désir d'un enfant supplémentaire dans l'avenir) qui seront décrites ici, en mettant l'accent sur les conséquences de l'utilisation de l'échelle logistique pour les proportions.

La *figure 5* montre les proportions d'utilisatrices de la pilule (traits pleins), du stérilet (traits pointillés) et de l'une ou l'autre des méthodes médicales (traits grisés) en 1978 et en 1988, sur une échelle additive. Les données de 1978 sont en traits fins, celles de 1988 en traits épais. En 1978, le stérilet est peu utilisé (12%), et les variations apparaissent faibles selon les variables démographiques (entre 4% et 17%). Ce sont surtout les variations des proportions d'utilisatrices de la pilule qui attirent l'œil : seules les femmes les plus jeunes ont massivement recours à la pilule (60% à 20-24 ans, contre 13% à 40-44 ans). La pilule est principalement utilisée par les femmes célibataires, sans enfant. Le stérilet est surtout pratiqué par les femmes d'âge moyen, ce qui résulte de la diffusion du stérilet, d'une génération à l'autre, les femmes de chaque génération utilisant de plus en plus le stérilet au fur et à mesure qu'elles avancent en âge [Toulemon, Leridon 1991].

Entre 1978 et 1988, la diffusion du stérilet est plus forte que celle de la pilule, mais le premier reste peu pratiqué par les femmes jeunes. La faible utilisation du stérilet par les femmes sans enfant (qui correspond à une réticence spécifique des médecins à prescrire le stérilet aux femmes nullipares, en raison du risque de stérilité considéré comme plus important pour les femmes qui ont plusieurs partenaires, et aux conséquences plus dramatiques pour les femmes sans enfant), est compensée en 1988 par une diffusion de la pilule très forte pour les femmes sans enfant ou les mères d'un seul

Figure 5

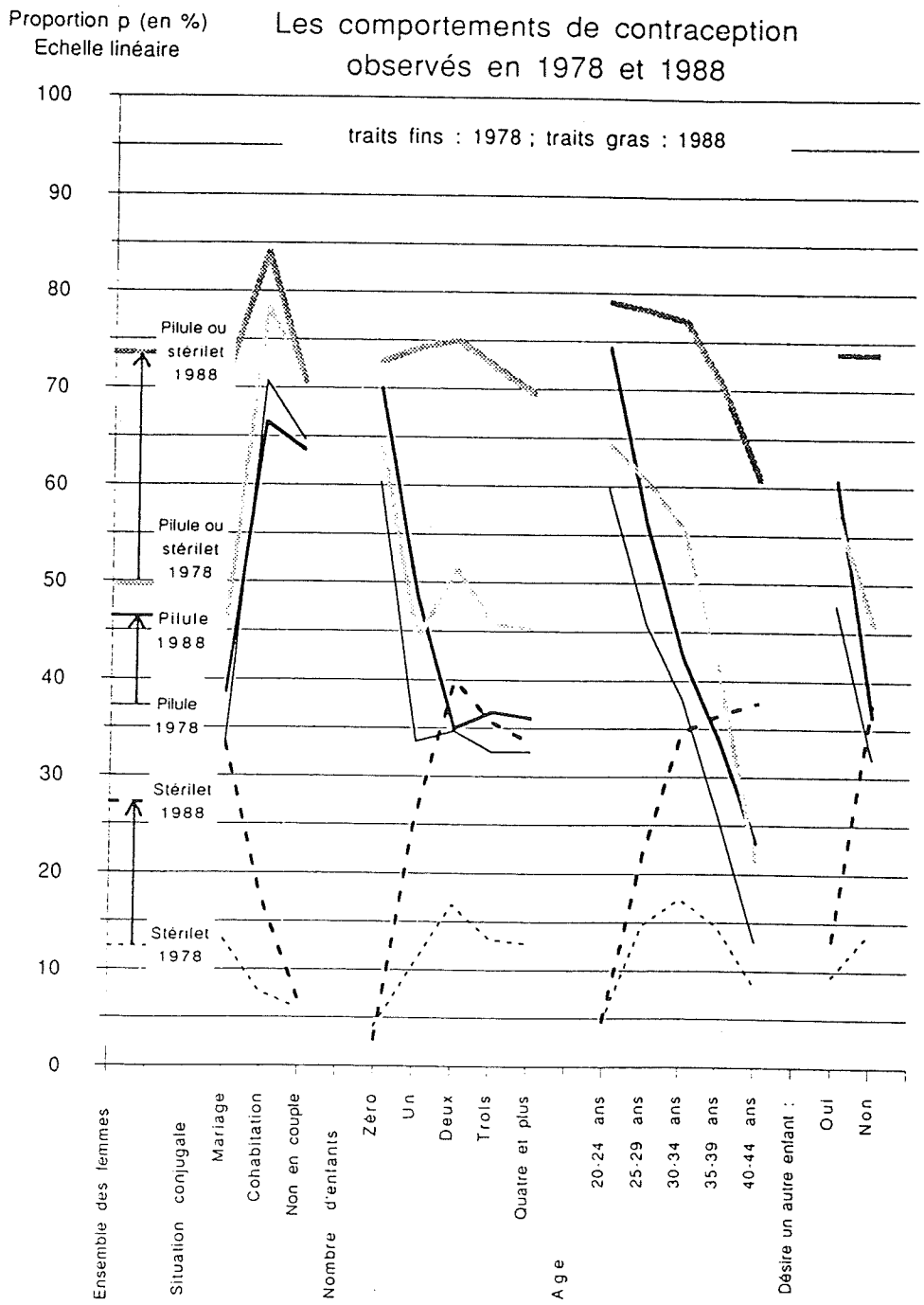
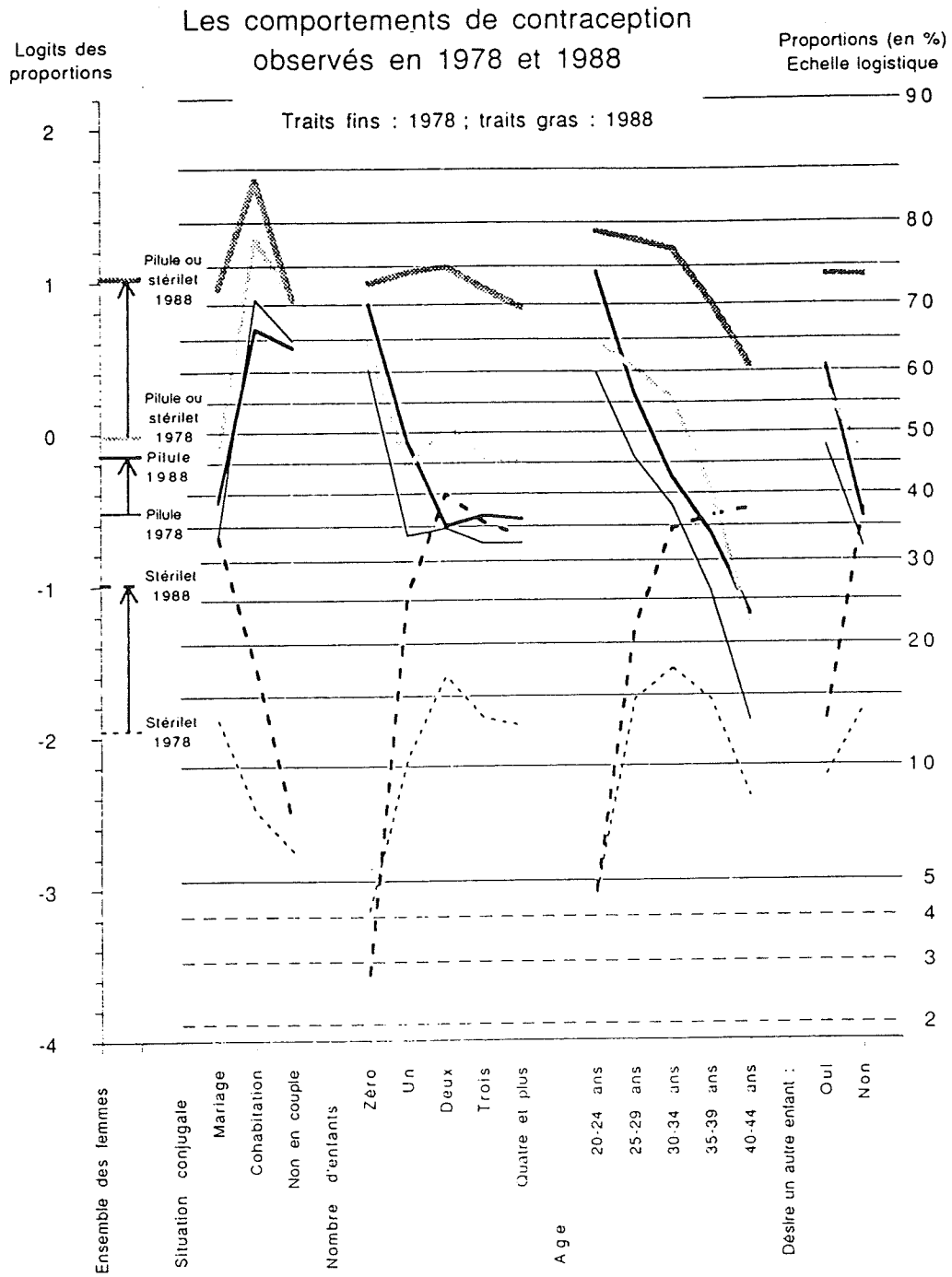

Proportion p (en %)
Echelle linéaire
Les comportements de contraception observés en 1978 et 1988
traits fins : 1978 ; traits gras : 1988
100
90
80
70
60
50
40
30
20
10
0
Pilule ou stérilet 1988
Pilule ou stérilet 1978
Pilule 1988
Pilule 1978
Stérilet 1988
Stérilet 1978
Ensemble des femmes
Situation conjugale
Mariage
Cohabitation
Non en couple
Nombre d'enfants
Zéro
Un
Deux
Trois
Quatre et plus
Age
20-24 ans
25-29 ans
30-34 ans
35-39 ans
40-44 ans
Désire un autre enfant :
Oui
Non

Figure 6


Les comportements de contraception observés en 1978 et 1988
Traits fins : 1978 ; traits gras : 1988
Logits des proportions
Proportions (en %) Echelle logistique
2
1
0
-1
-2
-3
-4
90
80
70
60
50
40
30
20
10
5
4
3
2
Pilule ou stérilet 1988
Pilule ou stérilet 1978
Pilule 1988
Pilule 1978
Stérilet 1988
Stérilet 1978
Ensemble des femmes
Situation conjugale
Mariage
Cohabitation
Non en couple
Nombre d'enfants
Zéro
Un
Deux
Trois
Quatre et plus
Age
20-24 ans
25-29 ans
30-34 ans
35-39 ans
40-44 ans
Désire un autre enfant :
Oui
Non

enfant, et finalement la pratique de l'une ou l'autre des méthodes médicales varie peu selon le nombre d'enfants.

Les mêmes proportions sont présentées sur la *figure 6* selon une échelle logistique (échelle des ordonnées à droite de la figure), qui correspond à une échelle linéaire sur les logits (échelle de gauche). La diffusion du stérilet est plus apparente, tandis que la spécificité des femmes sans enfant est plus spectaculaire : les mères d'un seul enfant ont un comportement proche des mères de deux enfants ou plus.

Comme l'échelle logistique n'est pas additive, on observe simultanément d'une part une compensation entre les pratiques de la pilule et du stérilet selon le nombre d'enfants en 1988 (les proportions d'utilisatrices de l'une ou l'autre des méthodes sont peu différentes selon le nombre d'enfants déjà nés), et d'autre part un écart entre les femmes sans enfant et les autres plus important pour le stérilet : les différences (linéaires) se compensent, mais la rareté du stérilet chez les femmes sans enfant est plus marquée (en termes logistiques) que la plus grande fréquence de la pilule. En termes de comportements, cela correspond à une règle spécifique, imposée par le corps médical, interdisant le stérilet aux femmes sans enfant ; cette règle est d'autant plus facilement respectée qu'elle est compensée par une utilisation plus fréquente de la pilule, méthode utilisée par la moitié des femmes, et dont le niveau de pratique peut donc facilement se modifier.

Enfin, la *figure 7* permet de comparer la diffusion de la pilule au sein de chaque catégorie. Seules les proportions selon l'âge et le nombre d'enfants déjà nés sont portées sur la figure. La diffusion de la pilule apparaît la plus forte pour les âges extrêmes, et pour les mères d'un enfant (écart logistique supérieur à 0,4 entre 1978 et 1988).

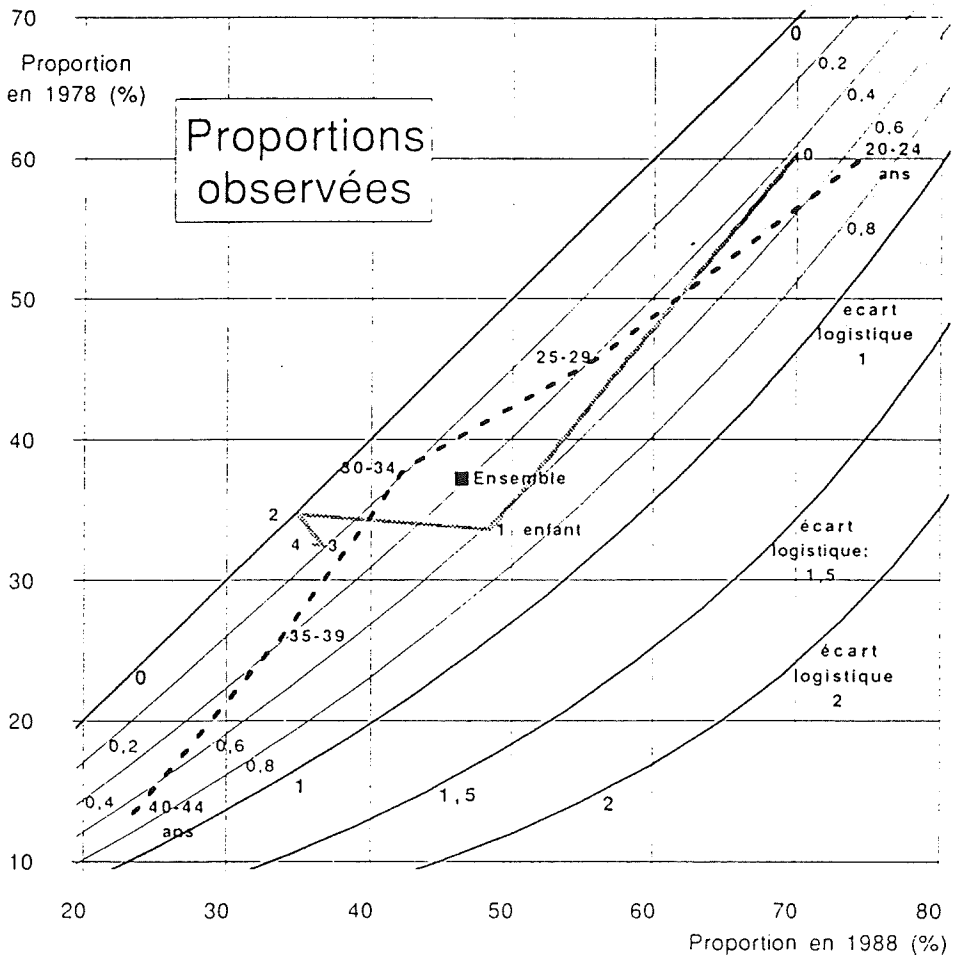
Les évolutions ayant été présentées à grands traits, il est maintenant possible de chercher à remplacer les proportions observées par des proportions qui représentent l'« effet propre » de chaque variable « explicative » sur les comportements de contraception. Par exemple, l'âge et le nombre d'enfants sont des variables très corrélées, et seule une régression permet de « séparer » les variations des comportements en différentes composantes imputables à l'« effet spécifique » de chaque variable.

L'estimation des « effets propres » par la régression logistique

La régression logistique combine les avantages de la régression et de l'échelle logistique. Les régressions permettent d'une part de *tester* l'effet de chaque « dimension explicative », et d'autre part d'*estimer* l'effet « pur » de chaque variable, « toutes choses égales par ailleurs », au sein d'un modèle où la pratique contraceptive ne dépend que des variables démographiques. Seule la question de l'estimation sera abordée ici.

Figure 7

Ecarts logistiques entre les proportions d'utilisatices de la pilule selon l'âge et le nombre d'enfants en 1978 et 1988



L'estimation de l'effet « spécifique » de chaque « dimension explicative » sous forme de probabilités estimées

Le modèle suppose que la *propension* à avoir le comportement étudié dépend de certaines variables « explicatives », selon le modèle linéaire sans interaction classique. Si deux dimensions explicatives interagissent, on les remplace par une dimension explicative qui les combine, afin de se replacer dans le cadre d'un modèle sans interaction. À chaque valeur estimée pour la propension à avoir le comportement étudié, on associe la probabilité dont cette propension est le logit. On cherche à mesurer l'effet « spécifique » de chaque dimension explicative en termes de probabilités *une fois éliminés les effets des autres dimensions.*

L'algorithme de régression calcule les *probabilités les plus vraisemblables*, en fonction des données, pour chaque combinaison de toutes les dimensions explicatives, selon la contrainte fixée par le modèle (absence d'interaction sur les logits). Ces probabilités déduites du modèle sont résumées, pour chaque dimension explicative, sous forme de paramètres. La différence entre les paramètres associés à deux catégories de chaque dimension explicative est définie de manière unique : c'est l'écart logistique « propre » estimé entre ces deux catégories. Mais la valeur de chaque paramètre est obtenue à une constante près pour chaque dimension explicative, constante qui dépend du groupe de référence choisi pour cette dimension explicative. Les probabilités calculées pour chaque combinaison de toutes les dimensions explicatives peuvent aussi fournir des estimations de l'effet « spécifique » de chaque dimension explicative sous forme de *probabilités*, estimées sous les deux contraintes suivantes :

- le contraste logistique (logarithme de l'*odds-ratio*) entre les probabilités estimées pour deux catégories quelconques d'une dimension explicative est égal à la différence des paramètres issus de la régression logistique;

- les probabilités estimées sont *globalement non-biaisées pour chaque dimension explicative*, c'est-à-dire que leur moyenne (pondérée par l'effectif de chaque catégorie) est égale à la proportion moyenne calculée sur l'ensemble de la population observée.

La première contrainte assure la cohérence des probabilités estimées avec les résultats de la régression logistique. Les contrastes entre les probabilités estimées sont le reflet direct de l'« effet propre » de chaque dimension explicative, dans le cadre de la régression logistique. La seconde présente deux avantages. D'une part, elle assure la comparabilité des probabilités estimées avec les proportions observées : si une dimension est seule à avoir un effet spécifique dans la régression, les probabilités estimées pour cette variable seront égales aux proportions observées, et les probabilités estimées

pour les catégories de toutes les autres dimensions explicatives seront égales à la moyenne d'ensemble (pas d'effet spécifique), tandis que la proportion observée pour chaque catégorie ne dépend dans ce cas que de la corrélation de cette catégorie avec la dimension qui a un effet (biais de covariance). D'autre part, elle permet de comparer les niveaux de deux proportions estimées par deux régressions logistiques avec les mêmes dimensions explicatives, correspondant à deux pratiques dont on mesure la fréquence au sein d'une catégorie, « toutes choses égales par ailleurs ».

La contrainte arbitraire sur les paramètres de la régression (groupe de référence) est donc remplacée par une contrainte de cohérence globale : les probabilités estimées pour les différentes catégories d'une dimension explicative sont cohérentes avec la proportion moyenne dans la population, et répondent à la question quelles proportions observerait-on si chaque dimension explicative était seule à avoir un effet ? Les probabilités estimées sont beaucoup plus concrètes que les paramètres du modèle, tout en leur étant équivalentes.

En calculant les contrastes logistiques entre chaque probabilité estimée et la proportion d'ensemble, on obtient de nouveaux paramètres (égaux aux paramètres de départ à une constante près, définie pour chaque dimension explicative) qui mesurent l'écart logistique entre chaque groupe et *l'ensemble de la population, considéré comme groupe de référence*. Cette méthode fournit en pratique des résultats très proches de ceux qu'on obtiendrait en imposant la contrainte que la moyenne pondérée des paramètres doit être nulle.

L'utilisation de la régression comme une *méthode de standardisation* n'est pas nouvelle [Mantel, Stark 1968, Berry 1970, Fleiss 1981]. À la possibilité d'effectuer des tests de signification pour chaque effet s'ajoute l'estimation des effets « spécifiques » de chaque dimension explicative.

Les régressions log-linéaires (échelle additive sur les logarithmes, c'est-à-dire échelle multiplicative) sont très utiles pour décrire les variations des *taux* (taux de mortalité, taux d'incidence) [Brillinger 1986, Hoem 1987, Courgeau, Lelièvre 1989]. En effet, les modèles sur les taux suivent souvent une logique multiplicative (on parle alors de risques proportionnels).

La régression log-linéaire apparaît alors comme une méthode de standardisation plus efficace que les méthodes habituelles (standardisations directe et indirecte), et prend le nom de *standardisation indirecte améliorée* [Hoem 1991]. Comme pour la régression logistique, les paramètres de la régression log-linéaire sont définis à une constante près, et on peut imposer la contrainte que les risques estimés (exponentielles des paramètres) soient *globalement non biaisés*.

La procédure de calcul des probabilités estimées est simple pour la régression log-linéaire, puisque la constante additive sur les paramètres correspond à une constante multiplicative sur les probabilités. Il suffit de multiplier les probabilités estimées à partir

des paramètres par le rapport de leur moyenne pondére à la moyenne d'ensemble (taux global), pour obtenir des taux estimés globalement non biaisés [Breslow, Day 1975, Hoem 1987, Toulemon 1992].

Mise en œuvre pratique du calcul des probabilités estimées par la régression logistique

Pour la régression logistique, il n'existe pas de relation aussi simple que pour les régressions linéaire ou log-linéaire entre les paramètres et les probabilités estimées. Le calcul se fait alors par approximations successives, séparément pour chaque dimension explicative. Les paramètres fournissent les différences logistiques β_i entre chaque catégorie et la catégorie de référence.

a) On fixe d'abord arbitrairement la proportion estimée pour la catégorie de référence et, à partir des différences logistiques, on déduit les proportions estimées pour toutes les catégories. Si $\beta 1$ est la différence logistique entre les groupes 1 et 0 (référence), on a :

$$\text{logit}(p_1) - \text{logit}(p_0) = \beta_1 \Rightarrow p_1 = \frac{p_0 \exp(\beta_1)}{1 - p_0 + p_0 \exp(\beta_1)}$$

b) Ensuite, on calcule la moyenne des probabilités estimées (pondérée par les effectifs de chaque catégorie), qu'on compare à la moyenne empirique pour l'ensemble de l'échantillon. Si la moyenne estimée est trop élevée, on recommence a) avec une proportion estimée plus faible pour la catégorie de référence. En pratique, on essayera des proportions estimées p_0 par itération :

$$p_0(N) = \sum_{i=1}^{N} x_i (0{,}5)^i, \text{ avec } x_i = 0 \text{ ou } 1$$

Au bout de vingt itérations, $p_0(20)$ est précis à 10^{-6} près, ce qui assure une précision des probabilités estimées pour les différentes catégories largement suffisante.

Application aux comportements de contraception

Pour chaque année (1978 et 1988), trois régressions logistiques ont été effectuées, dans lesquelles les variables démographiques « expliquaient » les comportements suivants : utilisation de la pilule, utilisation du stérilet, et pratique de l'une ou l'autre de ces deux méthodes.

En 1978, la pratique de la pilule plus fréquente chez les jeunes femmes correspond à une « effet propre » de l'âge, et les probabilités estimées « à situation conjugale, nombre d'enfants, et désir d'un autre enfant dans le futur égaux » sont très proches des proportions observées (*figure 8*). Par contre, l'utilisation plus fréquente de la pilule par les femmes sans enfant disparait quand sont prises en compte les variations dues à l'âge et à la situation conjugale. « Toutes choses égales par ailleurs », la pilule était surtout utilisée en 1978 par les femmes qui avaient plus de deux enfants. En revanche, la faible pratique du stérilet par les femmes sans enfant correspond à un effet spécifique, et persiste sur les probabilités estimées. L'utilisation d'une méthode médicale (pilule ou stérilet) *augmente* avec le nombre d'enfants déjà nés, quand les autres variables sont prises en compte : si les femmes sans enfant utilisaient plus que les autres une méthode médicale (la pilule), c'est parce qu'elles étaient plus jeunes et plus souvent célibataires que les autres.

Les probabilités estimées ne sont pas additives, et la somme des probabilités estimées d'utiliser la pilule et le stérilet n'est pas égale à la probabilité estimée de pratiquer l'une ou l'autre des méthodes médicales. Cependant, les deux estimations sont très proches en pratique, si les probabilités sont comprises entre 20% et 80%, car l'échelle logistique est presque linéaire dans cette plage de valeurs[1].

En 1988, les méthodes médicales de contraception sont devenues des pratiques courantes. Les variations selon les variables démographiques sont plus faibles qu'en 1978 (*figure 9*). En particulier, les femmes mariées utilisent autant que les autres une méthode médicale. Mais chaque méthode s'est « spécialisée » : la pilule est de plus en plus utilisée par les femmes sans enfant, tandis que la diffusion du stérilet est spectaculaire pour les mères de deux enfants ou plus, et aux âges 30 à 45 : après 40 ans, le stérilet est plus fréquemment utilisé que la pilule, même « toutes choses égales par ailleurs ». Pilule et stérilet sont de plus en plus souvent des méthodes successives dans la « biographie contraceptive » des femmes, la réticence des médecins à prescrire le stérilet aux femmes sans enfant étant de plus en plus affirmée.

(1). Une régression spécifique sur la probabilité d'utiliser soit la pilule, soit le stérilet, est utile pour tester l'effet de chaque variable explicative. Elle pourrait éventuellement être complétée par une régression expliquant le choix de la pilule ou du stérilet, pour les seules utilisatrices d'une méthode médicale. De tels modèles de « choix successifs » ont été essayés, mais les résultats étaient plus confus que pour des régressions séparées pour chaque méthode contraceptive.

Figure 8

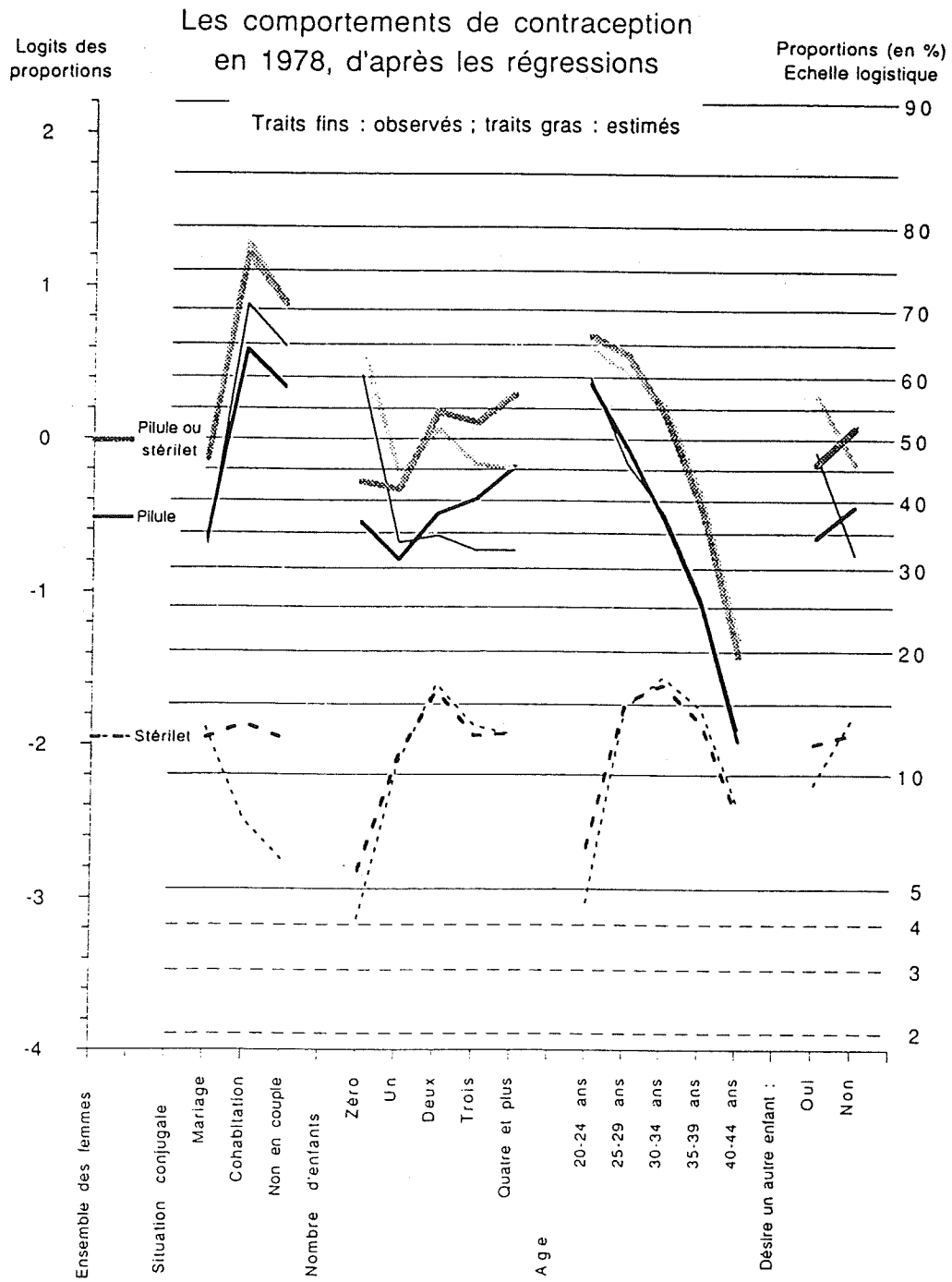
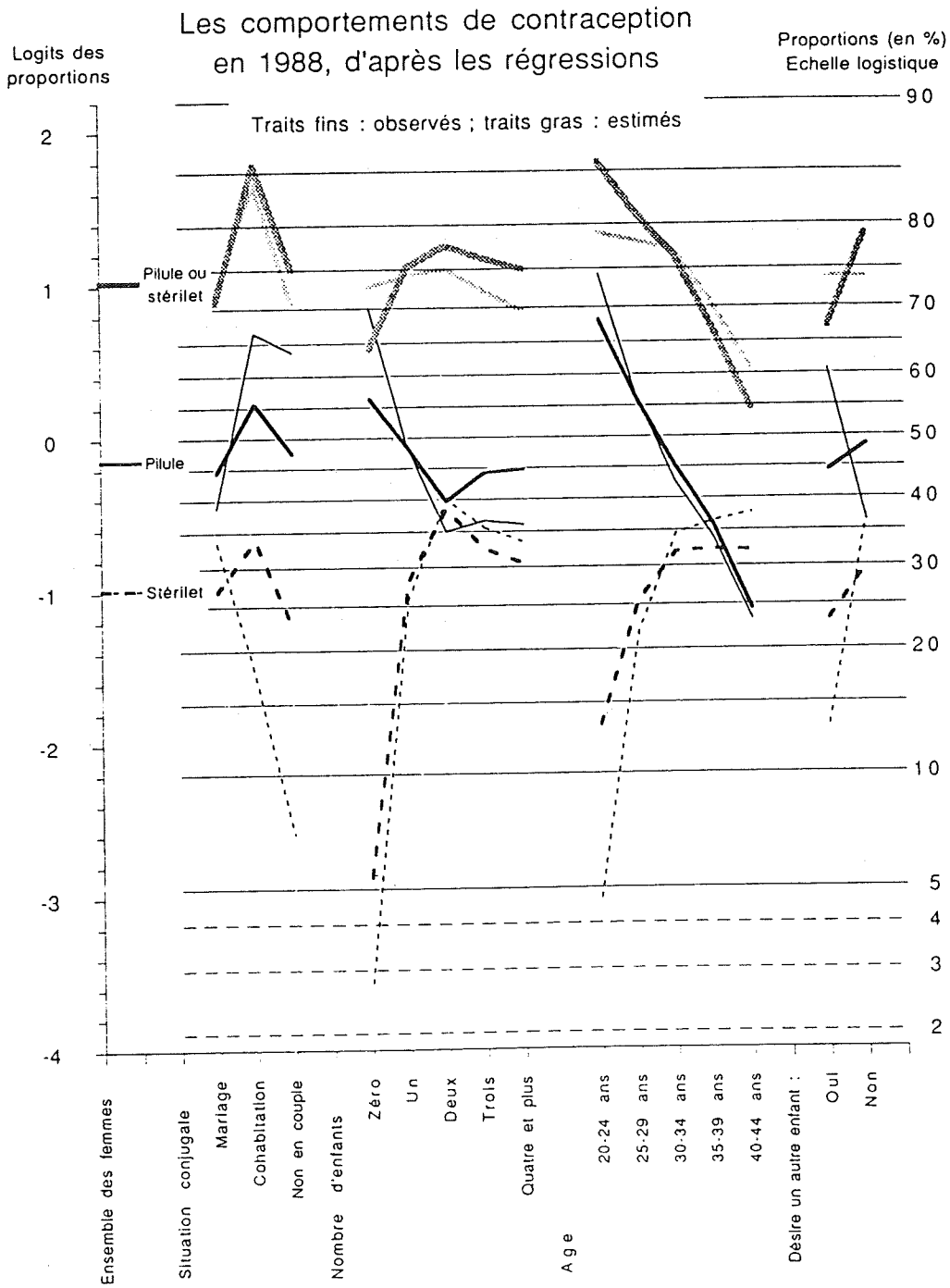

Les comportements de contraception en 1978, d'après les régressions
Traits fins : observés ; traits gras : estimés
Logits des proportions
Proportions (en %) Echelle logistique
2
1
0
-1
-2
-3
-4
90
80
70
60
50
40
30
20
10
5
4
3
2
Pilule ou stérilet
Pilule
Stérilet
Ensemble des femmes
Situation conjugale
Mariage
Cohabitation
Non en couple
Nombre d'enfants
Zéro
Un
Deux
Trois
Quatre et plus
Age
20-24 ans
25-29 ans
30-34 ans
35-39 ans
40-44 ans
Désire un autre enfant :
Oui
Non

Figure 9


Les comportements de contraception en 1988, d'après les régressions
Traits fins : observés ; traits gras : estimés
Logits des proportions
Proportions (en %) Echelle logistique
2
1
0
-1
-2
-3
-4
90
80
70
60
50
40
30
20
10
5
4
3
2
Pilule ou stérilet
Pilule
Stérilet
Ensemble des femmes
Situation conjugale
Mariage
Cohabitation
Non en couple
Nombre d'enfants
Zéro
Un
Deux
Trois
Quatre et plus
Age
20-24 ans
25-29 ans
30-34 ans
35-39 ans
40-44 ans
Désire un autre enfant :
Oui
Non

La *figure 10* compare les probabilités estimées en 1978 et 1988. L'importance croissante du nombre d'enfants déjà nés sur le choix entre pilule et stérilet apparaît clairement : seules les femmes ayant 0 ou 1 enfant utilisent davantage la pilule en 1988 qu'en 1978, et le stérilet est de plus en plus pratiqué par les femmes ayant au moins un enfant, « toutes choses égales par ailleurs ». Cette interprétation en termes de diffusion au sein de chaque catégorie de la population n'est possible que grâce à l'échelle logistique, qui permet de comparer la diffusion globale entre 1978 et 1988 et les contrastes dus à l'effet propre de chaque dimension explicative. Ces comparaisons ne sont justifiées que si les catégories définies par les différentes dimensions explicatives ne se sont pas trop modifiées entre 1978 et 1988. La baisse de la fécondité aux âges jeunes implique par exemple que les jeunes femmes sont de plus en plus souvent des femmes sans enfant. Mais ce type de précautions vaut aussi pour les proportions observées, et les *figures 6* (proportions observées en 1978 et 1988) *et 10* (probabilités estimées en 1978 et 1988) fournissent deux points de vue complémentaires sur la diffusion des méthodes médicales. Seul le calcul des probabilités estimées à partir des paramètres permet de comparer directement les effets propres calculés par la régression aux contrastes observés. La *figure 11*, construite pour les probabilités estimées sur le même principe que la *figure 7*, montre que la pilule s'est surtout diffusée auprès des femmes ayant 0 ou 1 enfant, « toutes choses égales par ailleurs ».

Les facilités de traitement graphique permettent aujourd'hui d'observer un grand nombre de résultats simultanément sous forme de figures, qui peuvent être considérées comme des *documents de travail*, permettant une vision globale de l'ensemble des résultats. Les résultats peuvent être publiés sous forme de figures similaires à celles présentées ci-dessus, ou sous forme de tableaux.

Les *tableaux 1 et 2* montrent les proportions d'utilisatrices observées pour différentes méthodes contraceptives (pour chaque méthode, une régression spécifique a été effectuée) et les probabilités estimées pour 1978 et 1988. Sur le *tableau 3* figurent les paramètres des régressions logistiques, à une constante près : c'est l'ensemble de la population qui est le « groupe de référence » pour chaque dimension explicative. Les paramètres sont multipliés par 10 et présentés sous forme arrondie, pour faciliter la lecture des effets majeurs.

***

L'échelle logistique se justifie indépendamment de toute régression, et elle peut être considérée comme *l'échelle « naturelle » pour les proportions*, de même que l'échelle multiplicative est « naturelle » pour décrire des prix ou des quantités, dont les évolutions *relatives* sont seules pertinentes. L'utilisation des logits (ou des « rapports des chances ») devrait se généraliser pour la présentation de résultats sous forme de proportions, et entraîner en retour la lecture des proportions sur l'échelle logistique.

Figure 10

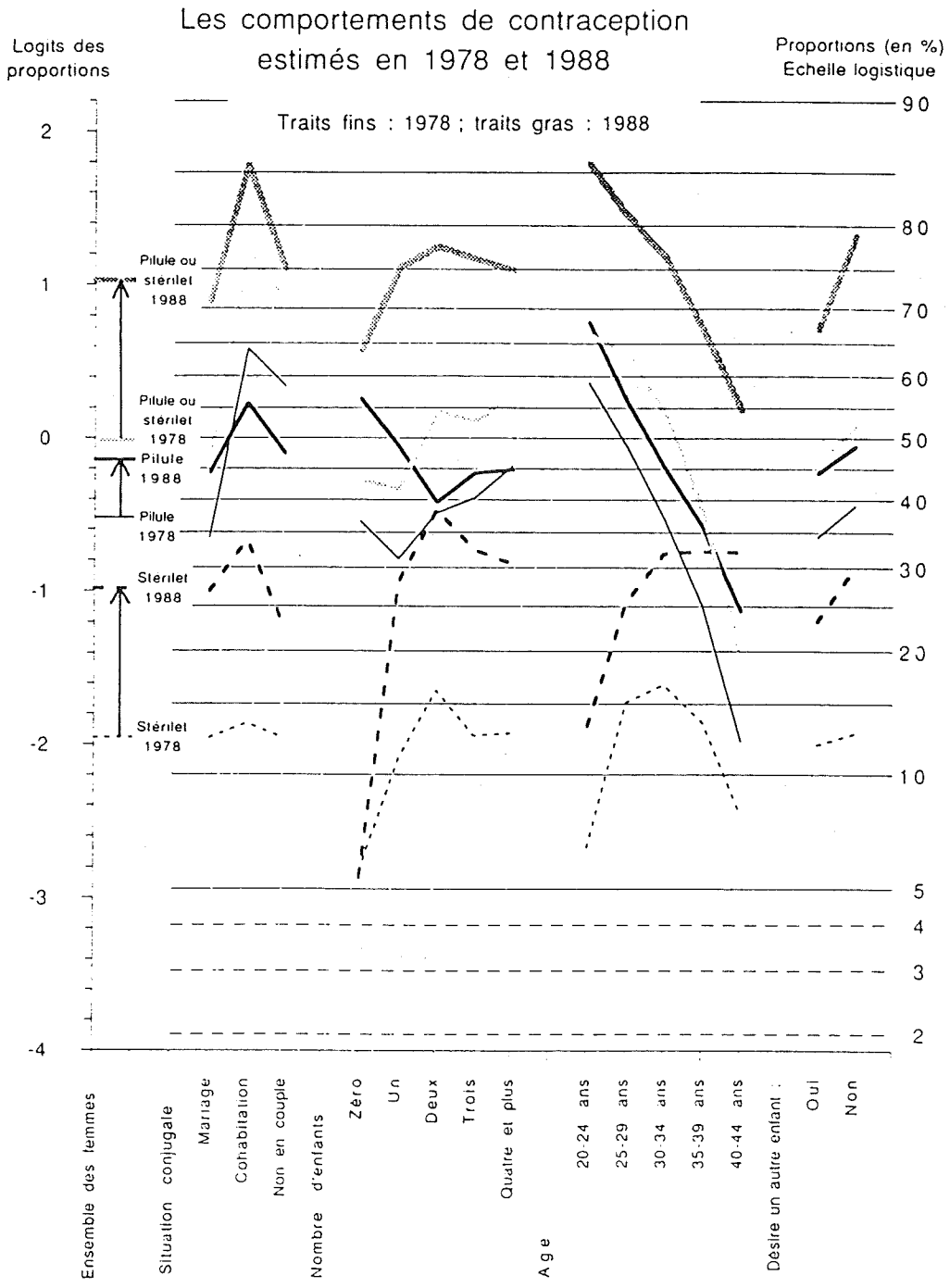
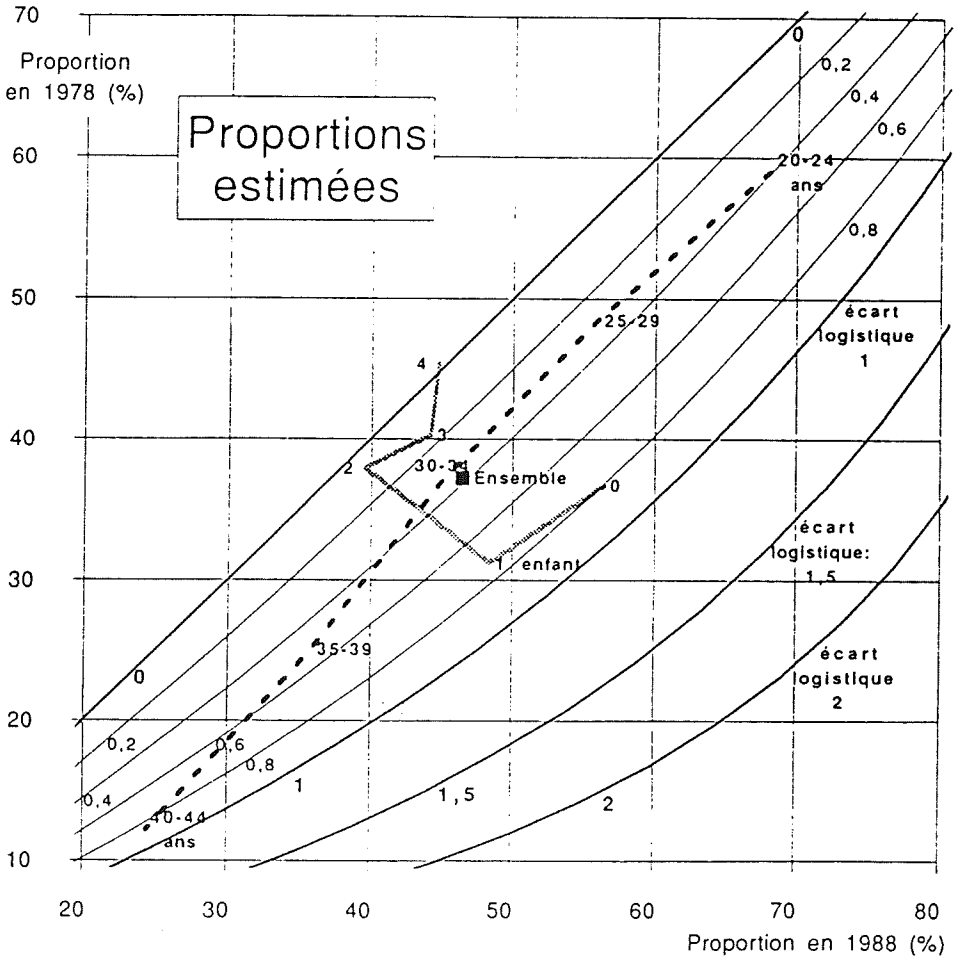

Les comportements de contraception estimés en 1978 et 1988
Traits fins : 1978 ; traits gras : 1988
Logits des proportions
Proportions (en %) Echelle logistique
2
1
0
-1
-2
-3
-4
90
80
70
60
50
40
30
20
10
5
4
3
2
Pilule ou stérilet 1988
Pilule ou stérilet 1978
Pilule 1988
Pilule 1978
Stérilet 1988
Stérilet 1978
Ensemble des femmes
Situation conjugale
Mariage
Cohabitation
Non en couple
Nombre d'enfants
Zéro
Un
Deux
Trois
Quatre et plus
Age
20-24 ans
25-29 ans
30-34 ans
35-39 ans
40-44 ans
Désire un autre enfant :
Oui
Non

Figure 11

Ecarts logistiques entre les proportions d'utilisatices de la pilule selon l'âge et le nombre d'enfants en 1978 et 1988



La présentation des résultats des régressions logistiques sous forme de *probabilités estimées* explicite la spécificité de l'échelle logistique utilisée dans ce type de régression, et permet une comparaison directe des proportions observées et des estimations issues de la régression. Une lecture des estimations issues de la régression logistique, aussi intuitive que la lecture de proportions observées, est alors possible. La procédure proposée pour estimer les probabilités à partir des effets propres permet également de remplacer les paramètres produits par les logiciels courants par des valeurs qui mesurent l'écart logistique « toutes choses égales par ailleurs » entre chaque catégorie et l'ensemble de la population, considéré comme groupe de référence.

BIBLIOGRAPHIE

BERRY G., 1970, Parametric Analysis of Disease Incidences in Multiway Tables, *Biometrics*, 26, 572-579.

BRESLOW N. E., DAY N. E., 1975, Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data, *J. Chron. Dis.*, 28, 289-303.

BRILLINGER D. R., 1986, The Natural Variability of Vital rates and Associated Statistics, *Biometrics*, 42, 693-734. With discussion.

COX D., 1972 (1960), *Analyse des données binaires*, Dunod.

COURGEAU D., LELIÈVRE E., 1989, *Analyse démographique des biographies*, Éditions de l'INED.

DEVILLE J. C., NAULLEAU E., 1982, « Les nouveaux enfants naturels et leurs parents », *Économie et Statistique*, 145, 61-81.

FLEISS J., 1981 (1973), *Statistical methods for rates and proportions*, Wiley, New-York.

GAIL M., 1978, The Analysis of Heterogeneity for Indirect Standardized Mortality Ratios, *J. R. Statist. Soc.*, A 141, part 2, 224-234.

HOEM J., 1987, Statistical Analysis of a Multiplicative Model and its Aplication to the Standardization of Vital Rates: A Review, *International Statistical Review*, 55, 2, 119-152.

HOEM J., 1991, La standardisation indirecte améliorée et son application à la divortialité en Suède (1971-1989), *Population*, 46, 6, 1551-1568.

MANTEL N., STARK C., 1968, Computation of indirect-adjusted rates in the presence of confounding, *Biometrics*, 24, 997-1005.

MARPSAT M., TROGNON A., 1992, « L'usage du modèle LOGIT. Présentation générale », *Journées de méthodologie statistique*, Insee, Juin 1992.

MARPSAT M., VERGER D., 1991, *L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs*, Insee, documents de travail de la Direction des statistiques démographiques et sociales, n° F 9110.

TOULEMON L., 1992, Population-type et autres méthodes de standardisation. Application à la mesure du recours à l'avortement selon la PCS, *Population*, 47, 1, 192-204.

TOULEMON L., LERIDON H., 1991, « Vingt années de contraception en France : 1968-1988 », *Population*, 46, 4, 777-812.

TOULEMON L., LERIDON H., 1992, « Maîtrise de la fécondité et appartenance sociale : contraception, grossesses accidentelles et avortements », *Population*, 47, 1, 1-46, 192-204.

VALLET L. A., 1988, « L'évolution de l'inégalité des chances devant l'enseignement. Un point de vue de modélisation statistique », *Revue Française de Sociologie*, XXIX, 395-423, et les articles cités en référence.

ZARCA B., 1993, « L'héritage de l'indépendance professionnelle selon les lignées, le sexe et le rang dans la fratrie », *Population*, 48, 2, 275-306.

Tableau 1 : Utilisation de diverses méthodes contraceptives en 1978 selon la situation conjugale, l'âge et le désir d'enfants : proportions observées, et estimées par le modèle logit (Femmes âgées de 20 à 44 ans, "au risque d'une grossesse non souhaitée")

		Ont consulté un médecin pour contraception		Utilisent une méthode contraceptive		Utilisent la pilule		Utilisent le stérilet		Utilisent pilule ou stérilet		Utilisent le retrait		Utilisent le préservatif		Utilisent une autre méthode	
	Effectif	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé
Ensemble	**2077**	**67**	**67**	**94**	**94**	**37**	**37**	**12**	**12**	**50**	**50**	**26**	**26**	**8**	**8**	**11**	**11**
Situation conjugale																	
Mariage	**1 841**	65	66	94	93	33	34	13	12	46	47	28	28	8	8	11	11
Cohabitation	**118**	83	79	93	95	71	64	8	13	78	77	9	10	2	3	3	4
Non en couple	**118**	78	75	97	98	65	58	6	12	71	71	8	7	7	9	11	14
Nombre d'enfants																	
Zéro	**280**	77	65	93	90	60	37	4	6	64	43	17	32	4	5	7	9
Un	**534**	60	57	92	93	34	31	10	11	44	42	27	28	10	10	12	12
Deux	**690**	68	70	96	96	35	38	17	16	51	54	28	26	7	7	10	10
Trois	**340**	69	74	94	95	32	40	13	13	46	53	26	22	9	9	13	12
Quatre et plus	**233**	63	72	91	92	33	45	13	13	45	57	26	20	6	6	14	11
Âge																	
20-24 ans	**364**	78	82	94	96	60	59	5	6	64	66	17	19	6	6	7	6
25-29 ans	**507**	76	80	95	96	46	48	15	15	61	63	20	19	6	6	7	7
30-34 ans	**484**	73	72	95	95	38	37	17	17	55	54	24	24	9	8	8	8
35-39 ans	**351**	58	54	93	91	26	25	15	14	41	38	30	30	9	9	14	14
40-44 ans	**371**	43	39	91	89	13	12	8	8	21	20	40	40	8	9	21	22
Désire un autre enfant																	
Oui	**691**	71	61	93	91	48	34	9	12	57	46	20	25	8	9	8	13
Non	**1 363**	65	70	94	95	32	39	14	13	46	52	29	26	8	7	12	9
Ne sait pas	**23**	81	72	92	91	39	21	11	13	50	31	14	21	0	0	28	42

Tableau 2 : Utilisation de diverses méthodes contraceptives en 1988 selon la situation conjugale, le nombre d'enfants, l'âge et le désir d'enfants : proportions observées, et estimées par le modèle logit (en %) (Femmes âgées de 20 à 44 ans "au risque d'une grossesse non souhaitée")

En %

		Ont consulté un médecin pour contraception		Utilisent une méthode contraceptive		Utilisent la pilule		Utilisent le stérilet		Utilisent pilule ou stérilet		Utilisent le retrait		Utilisent le préservatif		Utilisent une autre méthode	
	Effectif	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé	Observé	Estimé
Ensemble	1 773	92	92	94	94	46	46	27	27	74	74	7	7	5	5	9	9
Situation conjugale																	
Mariage	1 252	92	92	96	96	39	44	33	27	72	71	9	9	5	5	10	10
Cohabitation	252	96	96	97	97	66	56	18	34	84	86	3	4	3	2	6	6
Non en couple	269	86	87	83	85	64	47	7	22	71	75	2	2	3	3	7	6
Nombre d'enfants																	
Zéro	413	90	90	89	91	70	56	3	5	73	64	4	9	4	6	8	13
Un	389	91	90	95	94	48	48	26	28	74	75	8	7	4	4	9	8
Deux	618	92	93	96	96	35	40	40	38	75	78	6	5	5	5	10	9
Trois	262	93	93	97	96	37	44	36	32	72	76	11	8	5	5	8	6
Quatre et plus	91	91	92	94	94	36	45	33	30	69	75	14	10	1	1	10	8
Âge																	
20-24 ans	333	90	92	90	95	74	68	5	13	79	86	3	4	2	2	6	4
25-29 ans	374	94	94	96	96	56	56	22	26	78	81	5	5	5	4	7	6
30-34 ans	389	94	93	96	95	42	45	35	32	77	76	8	8	4	4	7	6
35-39ans	399	92	91	95	93	34	36	36	32	70	67	7	7	5	5	13	15
40-44 ans	278	87	86	94	91	23	24	38	32	61	55	13	12	6	8	13	17
Désire un autre enfant																	
Oui	712	92	92	93	95	61	44	13	23	74	67	5	8	5	6	9	13
Non	979	92	92	96	94	36	49	37	30	74	79	9	7	4	3	9	6
Ne sait pas	82	87	86	91	91	43	39	27	28	69	67	5	6	6	6	12	12

Tableau 3 : Utilisation de diverses méthodes contraceptives en 1978 et en 1988, selon la situation conjugale, le nombre d'enfants, l'âge et le désir d'enfants : valeurs des paramètres de la régression logit (x10)

	Effectifs		Ont consulté un médecin pour contraception		Utilisent une méthode contraceptive		Utilisent la pilule		Utilisent le stérilet		Utilisent pilule ou stérilet		Utilisent le retrait		Utilisent le préservatif		Utilisent une autre méthode	
	1978	1988	1978	1988	1978	1988	1978	1988	1978	1988	1978	1988	1978	1988	1978	1988	1978	1988
Ensemble	2077	1773	67%	92%	94%	94%	37%	46%	12%	27%	50%	74%	26%	7%	8%	5%	11%	9%
Situation conjugale																		
Mariage	1 841	1 252	-1	0	0	3	-1	-1	0	0	-1	-1	1	2	0	2	0	2
Cohabitation	118	252	6	8	1	6	11	4	1	3	12	8	-12	-7	-10	-7	-11	-5
Non en couple	118	269	4	-5	11	-10	9	0	0	-3	9	1	-15	-13	2	-6	3	-5
Nombre d'enfants																		
Zéro	280	413	-1	-2	-5	-4	0	4	-9	-19	-3	-4	3	3	-4	3	-2	4
Un	534	389	-4	-1	-2	-1	-3	1	-1	0	-3	1	1	1	3	-2	2	-2
Deux	690	618	1	1	5	3	0	-3	3	5	2	2	0	-5	-1	0	-1	0
Trois	340	262	4	3	2	4	1	-1	0	3	1	1	-2	2	2	0	1	-4
Quatre et plus	233	91	3	1	-2	-1	3	-1	0	2	3	1	-4	4	-3	-18	0	-2
Âge																		
20-24 ans	364	333	8	1	6	2	9	9	-7	-9	7	8	-4	-5	-3	-10	-6	-9
25-29 ans	507	374	7	3	4	3	5	4	2	-1	6	4	-4	-3	-2	-1	-5	-5
30-34 ans	484	389	3	3	2	1	0	-1	3	2	2	2	-1	1	1	-1	-4	-4
35-39 ans	351	399	-5	-1	-3	-1	-6	-4	1	2	-5	-3	2	-1	2	1	3	5
40-44 ans	371	278	-12	-6	-6	-5	-15	-10	-5	2	-14	-8	6	6	2	6	9	7
Désire un autre enfant																		
Oui	691	712	-3	0	-4	1	-1	-1	0	-2	-2	-3	-1	1	2	4	2	4
Non	1 363	979	1	0	2	0	1	1	0	1	1	3	0	-1	-1	-4	-2	-5
Ne sait pas	23	82	2	-6	-4	-4	-8	-3	0	1	-8	-3	-3	-2	15	2	18	3