FACULTE DE MEDECINE DE PARIS

THESE

pour le

DOCTORAT EN MEDECINE

(Diplome d'Etat)

par

Marcel Paul SCHUTZENBERGER

Né le 24.10.1920 à Paris

présentée et soutenue publiquement le 2 Juillet 1948

CONTRIBUTION A L'ETUDE STATISTIQUE DU SEXE A LA NAISSANCE

Président : Monsieur Raymond TURPIN, Professeur.

A

Mon Président de Thèse

Monsieur le Professeur R. TURPIN

Professeur de thérapeutique à la Faculté de Médecine de Paris

Médecin de l'Hôpital Saint Louis

Directeur du Centre de génétique de l'Hôpital Saint Louis

avec l'expression de ma reconnaissance pour le grand honneur qu'il a bien voulu me faire en me confiant le sujet de cette thèse, développement de ses travaux de génétique, en me guidant constamment de ses conseils au cour de sa réalisation et en acceptant d'en préside la soutenance.

A M E S M A I T R E S

DANS LES HOPITAUX

Monsieur le Professeur P.GUILLAUME-LOUIS, Directeur de l'Ecole de Médecine de Tours (I.&.L.) 1938/1941
Monsieur le Docteur D.GAUDEAU, Professeur à l'Ecole de Médecine de Tours - 1939/1940
Monsieur le Docteur F.CHEVE, Professeur à l'Ecole de Médecine de Tours - 1940/1941
Monsieur le Docteur A.TZANCK, Médecin Chef du Centre de transfusion sanguine replié à Blois (L.&.C.) 1939/1940
Monsieur le Docteur J.TUSQUES, Médecin Chef de service à l'Hôpital Psychiatrique de Blois -1941/1942
Monsieur le Docteur A.CARBILLET, Médecin-Directeur de l'Hôpital Psychiatrique d'Yzeure- 1943/1945
Monsieur le Docteur E.SIDI, Médecin Assistant à l'Hôpital Saint Louis.

AU COLLEGE DE FRANCE

Monsieur le Professeur P.LANGEVIN, Membre de l'Institut de France, Directeur de l'Ecole de physique et de chimie industrielle de la Ville de Paris (in memoriam).

A LA FACULTE DES SCIENCES DE PARIS

Monsieur le Professeur A.CHATELET, Ancien Recteur de l'Université de Lille, Professeur d'arithmétique supérieure
Monsieur le Professeur M.FRECHET, Professeur de calcul des probabilités et de statistique

AU CENTRE D'ETUDES DU BOUCHET

Monsieur P.GAVAUDAN, Chef du service de biologie cellulaire des services chimiques de l'Etat.

I N T R O D U C T I O N

Depuis l'antiquité la plus reculée, le problème de la détermination du sexe chez l'animal et chez l'homme a suscité un intérêt considérable, en raison de son importance économique, sociale et politique.

De nombreuses techniques et théories ont été imaginées et il suffirait de citer les aphorismes de la médecine hippocratique sur l'influence masculinisante du côté droit et féminisante du côté gauche, les traditions religieuses des diverses civilisations, les idées sur l'influence de certains régimes alimentaires, les opinions sur le rôle de la configuration astronomique au moment de la conception.

A l'époque moderne, l'utilisation de statistiques de plus en plus rigoureuses et étendues, l'emploi de la méthode expérimentale, le progrès des sciences biologiques ont permis de dégager un certain nombre de faits indiscutables et de fournir des interprétations satisfaisantes de la plupart d'entre eux.

°°°

Le présent travail, exclusivement statistique, a un objectif plus limité; se bornant aux familles humaines, on s'efforcera de prouver la réalité d'une "corrélation" entre naissances immédiatement consécutives, sans préjudice d'ailleurs de l'existence d'autres effets.

Faisant suite à la présentation par Monsieur le Doyen L.BINET, à l'Académie des Sciences d'une note (1) de Monsieur le Professeur R.TURPIN et moi-même, cette thèse apporte-ce qui n'avait pu y figurer- :

le détail des éléments numériques,

la discussion des sources,

le principe des modes de calcul,

l'analyse statistique des principaux résultats,

la construction et le test d'un schéma interprétatif.

En attirant ainsi l'attention sur un phénomène qui semble ne pas avoir été remarqué par d'autres Auteurs, on espère apporter un élément nouveau dont toute théorie se devra de tenir compte, si des recherches ultérieures, reposant sur d'autres sources documentaires, en confirment la généralité et la permanence.

° ° °

Nous tenons à remercier ici tous ceux qui nous ont permis de mener à bien le présent travail, en nous permettant les indispensables recherches dans les archives et le dépouillement des dossiers; en particulier Monsieur KERGALL, Directeur du service de la "Médaille d'Honneur aux Mères" au Ministère de la Santé publique et de la Population, à Messieurs les Présidents des Associations de familles nombreuses, à Messieurs les Membres du personnel administratif de la Fondation COGNACQ-JAY.

---

(1) - Recherches statistiques sur la distribution du sexe à la naissance par MM.R.TURPIN & M.P. SCHÜTZENBERGER ◆ Académie des Sciences- Séance du 1er juin 1948.

Nous remercions également pour une collaboration bénévole, qui nous fut précieuse, Mesdames GOUIN, POUZOL et MERCCHE du Centre de Génétique de l'Hôpital Saint Louis.

Nous exprimons enfin, toute notre gratitude à Monsieur le Docteur Raymond TURPIN, Professeur de thérapeutique à la Faculté de Médecine de PARIS, Directeur du Centre de Génétique à l'Hôpital Saint Louis. Il nous a accueilli dans son service hospitalier de l'Hôpital Saint Louis, il nous a confié ce travail et nous a aidé constamment dans sa réalisation; qu'il soit également remercié pour le très grand honneur qu'il nous a fait en acceptant de présider cette thèse.

B U T , M E T H O D E e t S O U R C E S

BUT

Avant de commencer l'étude de l'enquête statistique actuelle, il peut être bon de rappeler certains points importants de la théorie de la détermination des sexes, en ce qui concerne l'espèce humaine.

Sex ratio : Depuis VILLANI (1300 environ), GRAUNT (1661) et SÜSSMILCH (1761) on sait que la quasi-égalité numérique des sexes à la naissance est un phénomène général et constant; plus précisément il est admis que le nombre de naissances masculines pour CENT féminines est d'environ 106; ce rapport constituant une "véritable caractéristique" de l'espèce humaine comme l'écrivait HALBWACHS.

Il n'est pas besoin d'insister sur le fait que ce taux de masculinité à la naissance n'est pas identique à celui qui s'observerait à la conception, non plus qu'à celui de la population effective, en raison de l'inégale mortalité des individus des deux sexes aux différents âges de la vie (cf.3 et 7.)

D'autre part,cette constance de la sex ratio doit être complétée par la constatation de sa variation et des limites de celles-ci selon les pays, les époques et les milieux; par exemple, de manière significative, la sex ratio en Angleterre a évolué de 1910 à 1919 entre 0,510 et 0,515.

On ne discutera pas ici, car elles sortent du cadre de cette étude, les causes les plus fréquemment invoquées pour rendre

compte de ces fluctuations (influence du changement de régime alimentaire, âge moyen des parents .. etc .. etc.....)

On a aussi avancé que les primipares donnaient plus fréquemment naissance à des garçons et que cette augmentation de la sex ratio se rencontrait également chez les mères âgées; cette hypothèse est encore en suspend et des chiffres contradictoires ont été publiés, (cf.12 et 1.).

Enfin, un fait très important a été établi grâce aux matériaux recueillis par GEISSLER (cf.6.); éliminant les naissances gémellaires, la théorie classique voudrait que la répartition par nombre des garçons des familles de n enfants se fit selon une loi binomiale.

Or, si ce modèle probabiliste est très suffisant dans la pratique, il est possible de prouver en utilisant une base assez large de documentation, qu'il est théoriquement inacceptable.

Notamment FISCHER (cf.4.) a montré que les variances des histogrammes relatifs aux familles de n enfants sont significativement plus grandes que leur valeur théorique dans l'hypothèse d'une répartition binomiale; ce qui revient à dire que les familles extrêmes (c'est à dire celles où l'on observe un fort excès de garçons ou de filles) sont proportionnellement plus nombreuses que dans le cas de cette répartition.

Confirmant cette observation et offrant une possibilité d'interprétation, il faudrait pour être complet, exposer les théories selon lesquelles, chaque couple serait prédisposé à engendrer préférentiellement des garçons ou des filles selon certaines particularités de son génotype.(cf.10).

C'est sur la base de ces différents travaux que l'on

discutera les résultats obtenus au cours de cette enquête; c'est à dire successivement:

1°- la sex ratio et sa variation

2°- la répartition du sexe enfonctions du rang des naissances.

3°- l'excès des familles extrèmes

4°- la liaison entre les sexes à la naissance dans une même famille.

A chaque stade de l'étude, on sera amené à formuler une hypothèse probabiliste, permettant de tester l'hypothèse physiologique sous jacente.

Finalement, on construira un modèle statistique, dont on évaluera les paramètres et dont on montrera l'accord avec les faits; cependant qu'on donnera des indications sommaires sur une méthode plus complexe d'analyse de semblables documents.

METHODE

Etant donné le but poursuivi dans ce travail, les dispositions suivantes ont été prises pour la collecte des documents.

1°- Il a semblé nécessaire de se borner aux familles " assez"nombreuses pour que l'on puisse penser avoir écarté au maximum l'influence possible d'une restriction volontaire à la naissance; dans ce sens, il est évident que les familles les plus nombreuses (plus de 8 enfants) apportaient les meilleures garanties; mais la décroissance rapide de leur nombre en fonction du nombre des enfants (coefficient d'environ 0,55 par enfant supplémentaire pour une population non sélectionnée) a nécessité l'abaissement de cette borne inférieure.

Finalement a prévalu une considération d'ordre pratique : la

fixation administrative du chiffre de CINQ Enfants pour la définition de la "famille nombreuse" facilitant la collecte des documents basés sur cette valeur; c'est pourquoi, en définitive, l'enquête a été limitée aux familles de cinq enfants et plus.

2°- Est-il utile de mentionner que n'ont été prises en considération que les familles pour lesquelles les progéniteurs sont censés être les mêmes pour toutes les naissances; "pater est quem nuptiae demonstrant".

3°- On a éliminé toutes les familles comprenant des grossesses multiples; tant en raison de l'impossibilité d'en fixer l'uni ou la pluri-vitellinité (précision pourtant essentielle) qu'en considération de l'alourdissement et de la complexité dans le dépouillement et le traitement mathématique.

4°- On n'a conservé que les familles pour lesquelles était connu le sexe de tous les enfants nés viables, laissant de côté celles où cette condition n'était pas remplie. Une confiance plus grande, à ce point de vue, doit être accordée au stock 1 (dossiers recueillis au Ministère de la Santé Publique et de la Population) dont les documents de base contiennent systématiquement tous les enfants vivants et morts; il n'a pas été toujours possible d'avoir ce renseignement dans les autres stocks (2: région parisienne et 3: grandes familles du Nord) et surtout, hélas, dans le stock 4 (Fondation COGNACQ-JAY).

5°- J'ai du collecter moi-même les matérieux provenant des stocks 2, 3, & et 4, ainsi que la moitié de ceux issus du stock 1; l'autre moitié de celui-ci a été dépouillé par des collaborateurs bénévoles (qui ignoraient à dessein tous les buts

de cette enquête).

Ce point est important en raison de la part de l'influence personnelle de l'observateur sur le choix des documents; il les récolte très souvent (inconsciemment d'ailleurs) avec partialité en faveur de ses thèses, cette remarque faite, il faut indiquer qu'aucune différence significative entre les documents issus des deux moitiés du stock 1, non plus qu'entre les parties de la moitié du stock 1 étudiées par divers collaborateurs n'a été constatée en ce qui concerne les points principaux de cette enquête.

6°- Dans le même esprit on s'est interdit, et également afin d'éviter des duplications possibles, d'introduire dans cette étude des familles issues de source autre que systématique telles que: familles connues personnelement ou citées dans des ouvrages, travaux ..etc.. etc.

7°- Une partie importante (et difficilement déterminable) des familles étudiées sont des familles "en cours" c'est-à-dire qu'une famille de cinq enfants, par exemple, ne représente ici seulement qu'une Famille ayant cinq enfants au moment de l'enquête (1er mars 1948) sans que l'on puisse préjuger si le cinquième sera le dernier enfant ou sera suivi de frères ou de soeurs.

Etant données les sources utilisées, et les conditions matérielles de réalisation il n'a pas été possible d'enregistrer les dates de naissance des enfants successifs, non plus qu'aucune particularité concernant les parents: âge, statut social et économique, habitat, état de santé .. etc ..

Ainsi les données recueillies sont le nombre de famille, de chaque type caractérisé par la succession des naissances masculines ou féminines.

Par exemple:

On a observé 162 familles de 5 enfants dont 2 garçons du type: fille-garçon-fille-fille-garçon- (f.g.f.f.g.) et deux familles de 10 enfants du type : fille-garçon-garçon-garçon-garçon-garçon-fille-garçon-fille-garçon. (f.g.g.g.g.g. f.g.f.g.).

LES SOURCES

STOCK .1: On a utilisé les dossiers constitués au Ministère de la Santé publique et de la Population en vue de l'attribution de la Médaille d'Honneur des familles (promotion 1946 et 1947).

Les irrégularités dans les fréquences relatives des familles de 5, 6, 7, 8, ..etc... enfants s'expliquent aisément du fait qu'il existe.

a- une Médaille de bronze à laquelle ont droit les mères de cinq enfants et plus.

b- une Médaille d'argent pour huit enfants et plus.

c- une Médaille d'or pour dix enfants et plus.

Des considérations diverses font que certaines mères reçoivent la Médaille dès la naissance du nombre minimum d'enfants requis alors que d'autres, ignorées jusque là par le service, se décorées de la Médaille de bronze, par exemple, alors que leur vie obstétricale est largement achevée et qu'elles ont déjà dix enfants.

Dans l'ensemble; les dates de naissance s'étendent d'environ 1900 à 1945 et les 9.895 familles utilisées peuvent être considérées comme un échantillon représentatif de l'ensemble du territoire métropolitain français.

STOCK.2: Les familles de ce stock ont été extraites

des fichiers des associations de famille du département de la Seine; il a été possible d'enquêter dans la plupart des sièges d'arrondissement et d'étudier ainsi 2.912 familles pour lesquelles des remarques chronologiques, semblables à celles faites pour le stock 1; sont valables.

STOCK. 3: On a dépouillé les annuaires dits "Les grandes familles de Roubaix-Tourcoing" (dernière édition 1947) et"Les grandes familles de Lille-Armentières-Cambrai-Douai-Valenciennes (édition 1931).

Un soin tout particulier a été apporté pour éviter qu'une même famille ne soit enregistrée en double et l'on s'est volontairement borné à la dernière génération figurant dans l'annuaire

STOCK.4: On a consulté les dossiers des années 1920,1921, 1922, du Grand Prix COGNACQ-JAY (réservé aux familles de plus de dix enfants) et rassemblé ainsi 147 familles particulièrement intéressantes, en raison du très grand nombre de leurs enfants.

Malheureusement, il semble que les candidats aient souvent tenu à honneur de minimiser le nombre de leurs enfants décédés en bas âge; en raison du travail de vérification supplémentaire qui se trouvait imposé de ce fait, on a du se limiter à ce nombre très restreint de familles.

Remarques générales sur l'échantillonage.

Comme on vient de le montrer, l'échantillonage utilisé ne peut pas, en toute rigueur, prétendre à être représentatif de la composition actuelle de la population française; cependant il semble suffisamment adapté au but poursuivi; l'étude pure-

ment physiologique des naissances dans les familles nombreuses.

D'autre part, divers tests, sur lesquels on ne s'étendra pas, ont permis de constater l'homogeneîte des différentes sources et autorisent ainsi à traiter en bloc les 14.230 familles sans se livrer à des ventilations et sous-ventilations qui alourdiraient les calculs et masqueraient les résultats.

TABLEAU N° 1

Origine des documents

Nombre d'enfants par famille	Ministère de la Santé	Région parisienne	Grandes familles du Nord	Fondation COGNACQ-JAY	TOTAL
5	3.309	1.387	432		5.128
6	3.214	740	301		3.255
7	1.398	374	205		1.877
8	1.269	211	139		1.119
9	682	109	81		872
10	527	48	59		634
11	313	27	34	10	384
12	159	12	15	32	218
13	68	3	6	31	108
14	26	3	1	29	59
15	14		2	21	37
16	9		1	12	22
17	5			6	11
18				2	2
19				4	4
Total des familles	9.893	2.914	1.276	147	14.230

TABLEAU N° 2

Répartition numérique des familles
selon le nombre total de leurs enfants et le nombre de garçons

Nombre d'enfants par famille	Nombre de garçons															Nombre total des familles
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
5	144	810	1544	1585	841	204										5.128
6	54	325	717	979	785	330	65									3.255
7	21	111	287	480	529	304	126	19								1.877
8	8	61	180	327	412	406	170	49	6							1.619
9		20	72	155	204	184	140	74	22	1						872
10		5	23	64	120	165	128	78	39	11	1					634
11		3	15	27	51	85	92	61	37	9	3	1				384
12		4	3	18	19	32	50	49	23	16	2	1	1			218
13		2	1	3	9	12	24	19	14	9	13	2				108
14				2	2	6	9	14	11	8	6	1				59
15					2	2	4	6	8	5	5	1	4			37
16					2	3	3	4	3	4	1		2			22
17						1			1	3	2	3	1			11
18									1				1			2
19											1	1		1	1	4
																14.230

S E X R A T I O

Calculée sur les 95.875 naissances, la probabilité, à priori de naissances d'un garçon est de 0, 5006 ± 0,0016, sensiblement plus faible que les valeurs trouvées habituellement.

Aucune autre interprétation plus satisfaisante n'ayant pu être donnée de ce fait, il parait raisonnable de ne voir là qu'une fluctuation d'échantillonage.

Pour comparer maintenant la sex ratio entre les familles d'importance différente, on a employé le test classique de χ^2 correspondant au coefficient de LEXIS et donné par $\chi^2 = \sum N_n \frac{(p_i - \bar{p})^2}{F-1}$ avec K-1 degré de liberté ou N_n est le nombre d'enfants dans les familles de n enfants; p_n la sex ratio correspondante et $\bar{p}$ la valeur des moyennes des p_n; on trouve ainsi $\chi^2 = 26,7$ pour 14 degrés de liberté, ce qui est une valeur nettement significative (0,02 %).

Mais si l'on se borne aux familles de moins de 12 enfants le χ^2 n'est plus que de 7, 7 et par conséquence n'est pas significatif.

Rapprochant ce fait de la non-augmentation de la sex ratio jusqu'à 12 enfants; il semble plausible de ne pas attribuer de signification physiologique à l'écart constaté entre ces 15 valeurs (voir tableau N° 3.).

Et il ne faudrait peut être y voir que l'expression des conditions économiques rendant plus fréquentes, parce que plus aisés, les développements de familles très nombreuses à prédo-

minance masculine dans les milieux ruraux, d'où provient précisément la majeure partie de ces familles exceptionnelles.

TABLEAU N° 3

Valeur caractéristique des histogrammes

Nombre d' enfants	Nombre de familles	Fréquence des naissances masculines	Variance observée	Ecart de la variance (1)
5	3.128	0,5085	1,300	+ 4,6
6	3.255	0,5057	1,595	+ 6,3
7	1.877	0,5061	1.892	+ 6,5
8	1.619	0,5011	2.069	+ 2,7
9	872	0,4980	2.636	+10,7
10	634	0,5178	2.583	+ 1,9
11	384	0,566	2.976	+ 3,8
12	218	0,5022	3.667	+ 8,0
13	108	0,5207	4,990	+ 2,9
14	59	0,5170	3.322	- 1,0
15	37	0,5441	4,695	+ 4,3
16	22	0,4682	5,110	+ 3,7
17	11	0,6112	3,072	± 0,8
18	2	0,5566	8,000	+ 3,5
19	4	0,6316	3,333	- 0,1

(1) Ecart en valeur observée et valeur théorique exprimée en écart type.

REPARTITION DES SEXES SELON LE RANG DE NAISSANCE.

Deux hypothèses ont été avancées à ce sujet:

1°- La probabilité que le premier né soit un garçon est supérieur à la probabilité de naissance d'un garçon dans les grossesses ultérieures.

2°- La proportion des naissances masculines augmente chez les mères d'âge avancé.

Pour tester la première hypothèse, la méthode la plus simple et la plus efficace, est de comparer les fréquences; on trouve ains

Probabilté de naissance d'un garçon à la première grossesse 7.253 / 14.250 = 5097 ± 0,0042.

Probabilité de naissance d'un garçon dans les autres grossesses: $\frac{41.242}{81.495} = 0,5062 \pm 0,0018$.

La différence entre ces deux chiffres, étant de l'ordre de 1 fois son écart type ne peut être considérée comme significative et les documents recueillis n'apportent donc qu'une très faible confirmation de l'hypothèse.

Pour tester la deuxième hypothèse, on a employé une méthode qui semblera peut être un peu artificielle, mais qui permet une utilisation assez précise de l'information contenue dans les documents pour les familles de n enfants, on a considéré que, dans l'hypothèse nulle le nombre des garçons à chaque rang de naissances est réparti selon une loi rectangulaire discrète, autrement dit qu'à chaque rang devait se trouver sensiblement le même nombre de garçon

On introduira donc deux concepts commodes, qui sont les familles de n enfants.

1°- $\bar{i}_n$ rang moyen des garçons . $\bar{i}_n$ se trouvera déplacé vers l'une ou l'autre des extrémités selon que les garçons seront plus ou moins fréquents à la fin OU au début des familles.

Sa valeur est estimée par $i_n = \frac{\Sigma\ i\ f\ i}{N\ g}$

f i = nombre des naissances masculines au i ième rang, Ng = Σ f i = nombre total de garçons dans les familles de n enfants.

Zn sera considéré comme distribué normalement autour de sa valeur moyenne $\frac{n-1}{2}$ avec une variance égale à 1/12 (Ng) - 1 (n^2 - 1)

2°- σ^2_p : variance des rangs qui se trouvera augmentée ou diminuée selon que les garçons sont ou non plus fréquents au début ET à la fin des familles.

σ^2 est estimée par (Σ_i fi - Ng $(\bar{i}_n)^2$ (Ng -1) $^{-1}$ et sera considéré comme distribué normalement autour de sa valeur moyenne :1/12 (N^2-1) avec une variance que pratiquement l'on prendra égale à 1/80 $(Ng)^{-1}$ (n^4-5 n^2 + 4).

Le tableau ci-joint N° 4 montre à simple examen qu'il n'y a que peu de preuves pour écarter l'hypothèse nulle d'une distribution rectangulaire.

Pour les σ^2 où il pourrait y avoir doute, on a combiné les 15 tests par une méthode classique (cf.8.).

L'on sait en effet que l'expression $M^2 = -2\Sigma \log_e p_i$ ou p_i est pour chaque test la probabilité d'observer un écart aussi grand avec la valeur théorique est distribué comme χ^2 avec 2 K degrés de liberté.

Ici M^2 = 39, 7 et 2 K = 30; l'on ne peut donc écarter

l'hypothèse nulle (probabilité comprise entre 0,15 et 0,10.).

Par la suite, on considérera donc comme uniforme la probabilité à priori de naissance d'un garçon à un rang quelconque.

TABLEAU N° 4

Valeurs caractéristiques des distributions rectangulaires

Nombre d' enfants	Rangs moyens	Ecart des rangs moyens	Variance des rangs	Ecart d varian (1)
5	2.015	+ 1,2	2.072	+ 1
6	2.602	+ 1,2	2.502	+ 1,5
7	3.045	+ 1,9	3.045	+ 1,7
8	3.534	+ 1,2	5.320	+ 1,2
9	4.004	+ 5,1	6.727	+ 0,6
10	4.416	- 1,6	8.250	0,0
11	4.966	- 0,5	10.000	0,0
12	5.445	- 0,6	11.694	- 0,1
13	6.026	+ 0,2	14.253	+ 0,5
14	6.312	- 1,0	16.356	- 0,1
15	6.993	0,0	17.975	- 0,7
16	7.550	+ 0,4	22.834	+ 0,4
17	7.467	- 0,1	26.175	+ 1,0
18	9.050	+ 0,5	23.438	-0, 8
19	8.625	- 0,5	33.771	+ 0,9

(1) Ecart entre valeur observée et valeur théorique exprime en écart type.

V A R I A N C E S D E S R E P A R T I T I O N S

Soit la répartition de N_n familles de n enfants en familles de g garçons et n-g filles.

C'est une observation déjà faite par FISCHER sur les matériaux de GEISSLER que les familles à forte prédominance masculine ou féminine sont sensiblement plus nombreuses que ne le voudrait une répartition binomiale.

Opérant sur un matériel, d'où les naissances gémellaires ont été éliminées, il est possible, ici, de préciser ce point en comparant la valeur observée de la dispersion de ces distributions à la valeur théorique dans le cas d'une répartition binomiale qui constitue l'hypothèse nulle.

Dans ce cas, on devrait avoir les familles de n enfants

f_g= nombre de familles de g garçons = $\binom{n}{g} p^g q^{n-g} Nn$ où $p = \frac{\sum \varphi f_\varphi}{Nn}$

est la probabilité a priori de naissance d'un garçon.

La variance est estimée par la formule $\frac{1}{Nn-1}\left(\sum \varphi^2 f_\varphi - Nn\, p^2\right)$ et N étant grand, il sera légitime de considérer qu'elle est distribuée comme une variable normale de moyenne $n\,p\,q$ et de variance $\frac{1}{n^2} = n^{-1}\mu_4 + \frac{3-n}{n(n-1)}\mu_2^2$ où μ_2 et μ_4 sont respectivement les moments à la moyenne du 2ième et du 4ième prdre de la répartition binomiale considérée; pratiquement on a pris ici pour valeur de $1/h^2$ l'expression : $1/N \;\; 1/8\,(N)\,{}^{-1}\,(n-1)^{-1}\,(n2 - n + 1)$

Or, comme le montre le tableau 3 dans 12 cas sur 15, la variance observée est supérieure à sa valeur moyenne et dans 10 cas sur 15 cette différence est significative à plus de 1 %.

Il n'est pas besoin d'autre test combinant les résultats des quinze valeurs observées pour pouvoir établir que, d'une manière fort appréciable, les histogrammes sont plus étalés que ne le permettrait l'hypothèse d'une répartition binomiale, qui doit donc être rejetée.

S E Q U E N C E S

Si, comme dans le travail de GEISSLER, l'information dont on dispose se trouvait limitée aux histogrammes, il serait difficile d'aller plus loin et de nombreuses possibilités (entre lesquelles on ne saurait trancher) s'offriraient pour expliquer l'étalement des distributions; notamment en supposant que la probabilité a priori p de naissance d'un garçon constante dans une même famille est répartie d'une façon aléatoire dans la population, on aurait les fréquences suivantes

$$f_g = E\left(p^g \; q^{n-g} \right) \begin{bmatrix} n \\ g \end{bmatrix} \times N_n$$

où E (x) = valeur moyenne de x; et l'on pourrait estimer la variance de p dans la population par l'expression

$$\sigma_p^2 = \begin{bmatrix} n \\ 2 \end{bmatrix}^{-1} \sum \begin{bmatrix} y \\ 2 \end{bmatrix} \delta_y$$

toujours positive quand, et seulement quand, la variance de l'histogramme des familles de n enfants est supérieure à sa valeur théorique.

Mais, dans l'enquête présente, l'information est plus complète et comprend à l'intérieur du total des familles de même composition globale de g garçons et de f filles, la répartition entre les différents arrangements possibles.

On peut donc tester l'hypothèse nulle H_1 selon laquelle tous les arrangements auraient la même fréquence relative d'apparition pour des valeurs données de g et de f.

Ayant déjà examiné dans la seconde partie la possibilité que les garçons ou les filles se rencontrent préférentiellement à des moments privilégiés de la vie obstétricale (premier enfant,

naissances tardives) et n'ayant constaté aucune déviation significative, on est amené à tester un éventuel groupement des naissances de même sexe.

Pour tester H_1 contre cette nouvelle hypothèse, il serait facile d'imaginer des "coefficients d'accumulation " plus ou moins complexes, faisant appel à telle ou telle propriété des suites de tirage aléatoire ou de permutations.

Mais d'abord,de telles expressions seraient surtout valables pour des suites assez longues (soit par exemple 10 à 20 Enfants) et l'on a affaire ici à des suites beaucoup plus courtes (10% seulement des naissances ont été observées dans des familles de onze enfants et plus). Ensuite, et surtout, on risquerait de tomber dans l'erreur fréquente qui consiste à forger des concepts statistiques étrangers au contenu concret du mécanisme aléatoire qu'ils doivent servir à étudier.

On a donc utiliser pour tester h_1 le nombre des "coupures" c'est-à-dire des événements définis par la succession dans une même famille d'une naissance masculine à une naissance féminine la précédant immédiatement ou inversement.

A ce concept se rattache directement le concept de "séquence" masculine" (respectivement féminine) c'est-à-dire de suites de naissances masculines (respectivement féminines) immédiatement consécutives.

Pour une famille de g garçons de f filles, on a immédiatement

$$S_m = S_f + A + B - 1$$

ou Sm désigne le nombre des séquences masculines,

Sf désigne le nombre des séquences féminines,

et A et B sont des indicatrices telles que A = 1 si le premier enfant est un garçon, A = 0 si le premier enfant est une fille,

B = 1 si le dernier enfant est un garçon et B = 0 si le dernier enfant est une fille, et

C = Sm + Sf - 1; où C désigne le nombre de coupures; par exemple dans la famille du type : g.f.f.g.g.f.f.f. (3 garçons et 5 filles on a g = 3; f = 5; B = 0; S_g = 2; S_f= 2; C = 3; et dans la famille du type: f.g.f.g.g.f.g.f.f.g.f. (5 garçons et 6 filles) on a g =5; f =6; A =0; B = 0; S_g =4; S_f = 5; C = 8.

Il est évident que C répond assez exactement a ce que l'on peut entendre plus ou moins intuitivement par "groupage des naissances de même sexe."

L'emploi de la notion de"coupures" (ou"séquences") se recommande par son étroite liaison avec l'estimation de la probabilité conditionnelle d'une naissance masculine quand est connu le sexe de l'enfant précédent; notion elle même interprétable d'une manière concrète sur le plan de la physiologie comme on le montrera plus loin.

Les diverses répartitions liées aux S_g , Sf, et C, ont été étudiées théoriquement par STEVENS (c.f. 9) - (1); et on a complété son travail en calculant l'expression formelle des coefficients de corrélation entre S_g et S_f, ce qui a permis d'obtenir la variance de C.

On a ainsi pour une famille de g garçons et f filles:

$$\text{Prob}\ (\ S_g = u\ ) = \begin{bmatrix} g-1 \\ u-1 \end{bmatrix} \begin{bmatrix} f+1 \\ u \end{bmatrix} \Big/ \begin{bmatrix} g+f \\ g \end{bmatrix}$$

$$\text{valeur moyenne de } S_g = \frac{(g)\ (f+1)}{(g+f)}$$

(1) Il conviendrait, en outre, de citer les travaux (6 et 12) dont je n'ai pu avoir directement connaissance.

variance de Sg = $\frac{(g-1)\ (f+1)\ f_g}{n \qquad (n-1)}$

corrélation entre S_g et $S_f = \sqrt{\frac{g-1}{g+1} \times \frac{f-1}{f+1}}$

Pour C, la distribution est un peu plus compliquée; on a :

Prob (C = 2w) = $\begin{bmatrix} g-1 \\ w \end{bmatrix} \begin{bmatrix} f-1 \\ w-1 \end{bmatrix} + \begin{bmatrix} g-1 \\ w-1 \end{bmatrix} \begin{bmatrix} f+1 \\ w \end{bmatrix} \begin{bmatrix} f+g \\ g \end{bmatrix} - 1$;

Prob (C = 2w +1) = 2 $\begin{bmatrix} g-1 \\ w \end{bmatrix} \begin{bmatrix} f-1 \\ w \end{bmatrix} \begin{bmatrix} f+g \\ g \end{bmatrix} -1$

valeur moyenne de C = $\frac{2\ f\ g}{f\ +\ g}$

Variance de C = $\frac{2\ g\ f\ (2\ g\ f\ -\ g-f)}{(g\ +\ f)2\ (g\ +\ f\ -1)}$

La distribution des S n'est qu'une forme de la distribution hypergéométrique; quand à celle de C, elle est un peu moins maniable.

On peut donner quelques exemples de ces répartitions; ainsi pour les familles de trois garçons et de trois filles on a pour distribution de C:

Valeur de C	1	2	3	4	5
Prob C :	2/20 -	4/20 -	8/20	-4/20 -	2/20

la valeur moyenne de C = $\frac{2 \times 3,3}{3 + 3} = 3$

variance de C = $\frac{2 \times 3 \times 3 \times (2 \times 3 \times 3 - 6}{6 \times 6 \times 5} = 4/10$

Par conséquent,une famille telle que g.f.f.f.g.g. (3 garçons et 3 filles) ou C =2, est "plus groupée" que la moyenne représentée elle-même, par exemple, par g.f.f.g.g.f.; une famille "moins groupée" serait g.f.g.f.f.g. ou C = 4

Pratiquement, la répartition de C étant relativement assez voisine d'une distribution binomiale, on a convenu d'assimiler la répartition du nombre total des coupures dans les N familles

ayant toutes les mêmes nombres g et f de garçons et de filles à une répartition normale dès que N atteint ou dépasse 20.

De fait, l'approximation est encore meilleure qu'on pourrait le penser, car cette valeur élimine la majeure partie des cas où g et f sont très différents et partant, où la distribution de C est fortement asymétrique.

On a ainsi calculé pour toutes les classes de famille de g garçons et de f filles, représentées par plus de 20 cas, la déviation du nombre total des "coupures" à sa valeur théorique, différence exprimée en écart type (Tableau N° 5).

On voit ainsi que sur les 48 classes, la valeur observée dans 36 cas est plus faible que la valeur théorique; combinant ces résultats par la méthode de E.S. PEARSON déjà utilisée pour les variances on trouve un χ^2 de 177,4 pour 96 degrés de liberté, ce qui correspond à une probabilité d'erreur de première espèce inférieure à 1 / 100.000 ième.

Moins rigoureusement, on pourrait employer, comme test, le nombre de cas ou la valeur observée est inférieure à la valeur théorique, en supposant que dans l'hypothèse nulle cette fréquence est 1/2.

On trouve ainsi:

a- Toutes les classes de famille

64 cas plus petit / 43 cas plus grand Probabilité : 0,025

b- Classes représentées par au moins 10 familles

40 cas plus petit / 17 cas plus grand Probabilité : 0,0006

c- Classes représentées par au moins 20 familles

36 cas plus petit / 13 cas plus grand Probabilité : 0,0003

d- Classes représentées par au moins 100 familles

28 cas plus petit / 6 cas plus grand Probabilité : 0,002

e- Classes représentées par au moins 400 familles

10 cas plus petit / 1 cas plus grand Probabilité :0,059

Il n'est pas besoin d'insister sur le caractère un peu sommaire de ce dernier test en raison de la forte asymétrie des répartitions de C pour les classes représentées par un petit nombre de familles.

Grâce au χ^2 et à la faible valeur des probabilités qui viennent d'être calculées on doit donc considérer que les diverses familles de même nombre de garçons et de filles n'ont pas la même fréquence et que sont favorisées (faiblement mais indiscutablement) celles où les enfants de même sexe se trouvent "groupés".

Ce fait est d'une extrême importance, car il prouve le caractère intrinsèque de la corrélation entre naissances successives, corrélation qui pourrait, sans cette démonstration, être attribuée, par exemple, à la variabilité de la probabilité de naissance d'un garçon entre les différentes familles.

TABLEAU N° 5

Nombre de coupures (écart de la valeur observée à la valeur théorique en écart type) pour les familles représentées par plus de 20 Cas.

Nombre de filles	1	2	3	4	5	6	7	8
1				1,2	1,7	0,5	0,7	1,2
2			2,1	1,3	-1,1	1,9	0,9	1,4
3		2,5	2,7	1,2	-0,8	0,5	-0,4	-0,2
4	0,6	1,5	1,0	3,2	0,2	0,6	0,6	-0,1
5	-1,1	1,3	2,1	2,4	-0,5	0,3	-1,0	
6	0,8	0,8	-1,1	0,6	2,0	0,4		
7	0,9	0,1	1,1	1,4	-1,2	1,0		
8	0,8	0,6	-0,3					

CORRELATION ENTRE LES SEXES DANS UNE MEME FAMILLE

Ayant établi par le test précédent l'existence d'une liaison entre les naissances consécutives on en évaluera l'importance afin de tester son influence dans la répartition des familles entre les différents types.

Pour ce faire on représentera provisoirement la probabilité d'apparition d'une famille de g + f enfants par l'expression:

$$p^{A} (1-p)^{1-A} x^{Rg} (1-x)^{R'g} y^{Rf} (1 - y)^{R'f}$$

où A comme précédemment est égal à 1 ou à 0, selon que le premier né est un garçon ou une fille; Rg : nombre des naissances masculines précédées d'une naissance masculine; R' g : nombre des naissances masculines précédées d'une naissance féminine; Rf : nombre des naissances féminines précédées d'une naissance féminine ; R'f : nombre des naissances féminines précédées d'une naissance masculine; p = probabilité a priori d'un premier né masculin; x = probabilité conditionnelle d'un enfant du sexe masculin quand le précédent est du sexe masculin; y = probabilité conditionnelle d'un enfant du sexe féminin quand le précédent est féminin.

Comme il a été annoncé plus haut, les valeurs de Rg et de Rf sont liées au Sg, Sf et C par les relations (bien évidentes) Rg= g-Sg; R' g = Sg - B; Rf = f - Sf; R' f = Sf - (1-B).

Autrement dit, on emploiera provisoirement un schéma de MARKOFF à trois paramètres.

Plus généralement, on étudiera la liaison entre naissances séparées par i autres naissances c'est-à-dire les valeurs de Pgi (respectivement pfi) : probabilité conditionnelle de naissance d'un garçon (respectivement d'une fille) quand un garçon (respectivement une fille) est né précédemment à i naissances d'intervalles.

Dans les paragraphes précédents on avait x = Pgo et y = Pfo; un calcul simple, sur lequel on ne s'étendra pas, montre que l'estimation par la méthode du maximum de vraisemblance (qui est tout indiquée ici où le nombre d'observations est grand) conduit aux valeurs suivantes pour Pg: $\hat{P}_{gi} = \frac{R_{gi}}{N_g - B_{gi}}$; ou Ng représente le nombre total des garçons dans les 14.230 familles; B_{gi} la somme du nombre des garçons aux rangs n-1; n-2.....n-1-i pour les familles de n enfants; Rgi: le nombre total des garçons précédés d'un garçon a i intervalles (Pour i = 0 $R_{gi} = N_g - S_g$); et B i = nombre des derniers nés masculins.

On a, évidemment, pour les naissances féminines des expressior correspondantes.

Les tableaux N° 6 et 9 donnent les éléments numériques résultants du dépouillement des 14.230 familles pour les 4 premières valeurs de i seules intéressantes :

1°- en raison de l'augmentation de la variance des estimations.

2°- en raison du caractère artificiel que présenterait l'assimilation du rapport entre la première et la dernière naissanc dans une famille de cinq enfants à celui qui existe entre naissance de rang J et J + 4 dans les familles de dix enfants, par exemple.

D'autre part, ces tableaux peuvent être considérés comme des tables de contingence où un χ^2 permet de les tester rigoureusement

En ce qui concerne les variances des estimations, un calcul immédiat montre que $\hat{P}gi$ et $\hat{P}fi$ seraient indépendants, si leur valeur vraie était 1/2 et qu'elles auraient alors une variance égale à 1/4 (Ngi-Bgi)-1 ; on admettra que les valeurs actuelles permettent d'employer cette approximation (la différence est de moins 1%).

Enfin, on a calculé le coefficient de corrélation ri entre les sexes des naissances séparées par i intervalles.

L'ensemble de ces résultats est résumé au tableau N° 7.

Or, si les chiffres sont significatifs pour i=0 et i =1; il n'en est pas de même pour i = 2.

D'autre part, si l'on calcule les probabilités conditionnelles pgl et pfl, qui résulteraient de la seule existence de pgo et pfo; c'est-à-dire la liaison entre événements de rang r et r+2 dans une chaine de MARKOFF simple; on trouve

$$P'gl = p^2 go + (1- pgo) (1 - pfo) = 0,5066$$

$$p'fl = p^2 fo + (1- pgo) (1 - pgo) = 0,4943$$

en utilisant les estimations $\hat{p}o$ pour valeur de p_o.

Les différences entre les p'l et $\hat{p}l$ étant de l'ordre de l'écart type de ces dernières, on voit qu'il n'est pas possible d"écarter l'hypothèse du schéma de MARKOFF sur la base de l'information recueillie, en raison de la faible valeur de ces expériences; et il en serai de même à fortiori pour les p_2.

Autrement dit, si une théorie propose d'interprêter les désaccords entre la répartition observée et la répartition théorique par la seule influence d'une grossesse sur le sexe à la naissance suivante, il n'existe pas de preuve -permettant de l'écarter- que l'on puisse tirer de l'étude des corrélations dans cette enquête.

Or l'exemple bien connu des "free martins" étudiés par LILLIE, dans le bétail, suggère la possibilité d'une imprégnation masculini-

sante (ou peut être féminisante) de l'embryon pouvant entrainer une augmentation de la létalité du foetus ultérieur si celui-ci est de sexe différent.

On ne discutera pas ici les nombreux aspects biologiques de l'hypothèse que l'on avance -en faisant d'ailleurs toutes réserves quand à la possibilité d'existence d'autres actions- et on se bornera à faire ressortir que le fait établi ici "liaison intrinséque entre naissances immédiatement consécutives" ne se trouve complété par l'assertion "suffisant à expliquer les autres corrélations" que relativement au volume de l'information dont on dispose actuellement.

D'autre part, la comparaison entre les valeurs de $\hat{p}_{go}$ et de $\hat{p}_f$ et les fréquences à priori de naissances d'un garçon ou d'une fille montrent également que les différences 0,5206-0,5066 = 0,0140 et 0,5087-0,4934 = 0,0153 ne peuvent être considérées comme significativement différentes; ceci revient à dire que l'amélioration de la prédiction permise par la connaissance du sexe de l'enfant précédent peut être considérée comme égale pour les deux sexes, elle est d'environ 1, 5 %.

Légende du tableau N° 6 ci-contre

les chiffres figurant dans les neuf cases du tableau N° 6 doivent être interprétés comme suit pour les différentes valeurs de i
i = 0 ; i = 1 ; i = 2.

	Nombre de garçons précédés à i intervalles	Nombre de filles précédées à i intervalles	Valeurs marginales
	d'un enfant du même sexe	d'un enfant d'un autre sexe	Nombre total des garçons aux n-1 premiers rangs
Nombre de garçons suivis à i intervalles	(g_g)	(g_f)	(g^0)
	$= R_{gi}$	$= N_g - B_{gi} - R_{gi}$	$= N_g - B_{gi}$
	d'un enfant d'un autre sexe	d'un enfant du même sexe	Nombre total des filles aux n-1 premiers rangs
Nombre de filles suivies à i intervalles	(f_g)	(ff)	(f^0)
	$= N_f - B_{fi} - R_{fi}$	$= R_{fi}$	$= N_f - R_{fi}$
	Nombre total des garçons aux n-i derniers rangs	Nombre total des filles aux n-i derniers rangs	Nombre total des couples étudiés
Valeurs marginales	(0_g)	(0_f)	(00)
	$= N_g - A_{gi}$	$= N_f - A_{fi}$	$= N_g + N_f - A_{gi} - A_{fi}$ $= N_g + N_f - B_{gi} - B_{gi}$

TABLEAU N° 6

Liaisons entre naissances successives

	Nombre de garçons précédés à i intervalles	Nombre de filles précédées à i-1 intervalles	Valeurs marginales
	21398	19703	41101
Pour i = 0	19844	20550	40394
	41242	40253	81495
	17293	16525	33818
Pour i = 1	16801	16540	33347
	34094	33071	67165
	13617	13141	26758
Pour i = 2	15806	12971	26177
	36823	26112	52935

Tableau N° 7

Probabilités conditionnelles et corrélations

	chi carré =	70,25	Probabilité : plus petite que 1/1.000	
	$\hat{p}_{y0}$ =	0,5206	±	0,0025
i = 0	$\hat{p}_{f0}$ =	0,5067	±	0,0025
	p_0 =	0,0294	±	0,0035
	chi carré =	3,80	Probabilité =	5%
	$\hat{p}_{g1}$ =	0,5114	±	0,0028
i = 1	$\hat{p}_{f1}$ =	0,4962	±	0,0028
	p_1 =	0,0075	±	0,0039
	chi carré =	1,04	Probabilité =	30%
	$\hat{p}_{g2}$ =	0,5089	±	0,0031
	$\hat{p}_{f2}$ =	0,4955	±	0,0031
	p_2 =	0,0044	±	0,0043

MODELE PROBABILISTE

Il convient maintenant de faire le point des différentes hypothèses testées successivement; afin de pouvoir construire un modèle: probabiliste de la théorie biologique adoptée; ce modèle devra être le plus simple possible et permettre, après estimation de ses paramètres, de tester l'accord entre la nouvelle théorie et les faits qu'apporterait un retour à l'observation de nouveaux documents.

L'on sait déjà

1° qu'il est possible (malgré les cas exceptionnels des familles de plus de quinze enfants) de considérer comme constante la probabilité à priori de naissance d'un garçon dans les différentes familles et aux différents rangs.

2°- que la seule liaison significative est entre naissances immédiatement consécutives.

Enfin, et surtout, la théorie biologique adoptée implique une influence d'une grossesse sur la grossesse ultérieure.

On est ainsi conduit, tout naturellement, à un schéma de MARKOFF, c'est-à-dire à un tirage aléatoire où la composition de l'urne au ième coup dépend linéairement du résultat du i -1 ième.

Pour éviter de multiplier le nombre des paramètres (qu'on se souvienne de la boutade de SMIRNOV..) on introduira seulement

a- la probabilité à priori de naissance d'un garçon

b- un incrément constant quand les deux naissances sont de même sexe.

Plus précisément -et pour des raisons techniques que l'on exposera plus loin- on posera:

Prob.de naissance d'un enfant du sexe S_2 (quand l'enfant précédent est du sexe S_1) = $1/2 (1 + x\ U + y\ V)$.

où x est une indicatrice égale à + 1 ou -1 selon que S_2 est masculin ou féminin et y une indicatrice égale à

+ 1 si $S_1 = S_2$ (les deux enfants sont du même sexe

0 si S_1 n'existe pas (l'enfant est le premier né de la famille).

- 1 si $S_1 \neq S_2$ (les deux enfants sont de sexes différents).

C'est là,évidemment, le cas le plus simple d'un tirage en chaine.

Grâce à l'expression analytique employée, il est possible de considérer comme nulles les puissances de U et de V supérieures à la deuxième puisque l'on sait que U est de l'ordre de 2% (donc U^2 de l'ordre de 4/ 10.000) et V de l'ordre de 3% (donc V^2 de l'ordre de 1/ 1000)/

Pour cette même raison, 1°= il est possible avec une bonne approximation de réduire l'équation au maximum de vraisemblance donnant les estimations $\hat{U}$ et $\hat{V}$ à des équations linéaires (1)

$$U = \frac{Ng-Nf}{Ng+Nf} = \frac{1265}{95.725} = 0,0132$$

Ng étant le nombre total des naissances masculines et Nf le nombre total des naissances féminines.

---

(1) A dessein, on a laissé de côté l'exposé du traitement mathématique qui n'offre d'ailleurs aucune difficulté spéciale.

$$V = \frac{N + M - 2S}{N - M} = \frac{2401}{81.495} = 0,0295.$$

N étant le nombre total des naissances soit $N = N_g + N_f$; M: le nombre total des familles, S : le nombre total des séquences (ces chiffres se déduisent d'ailleurs immédiatement des calculs des chapitres précédents).

2°- de considérer les estimations $\hat{U}$ et $\hat{V}$ comme quadratiquement indépendantes.

3°- de prendre comme expression de leur écart type les suivantes

$$\text{écart type de } \hat{U} = \frac{1}{\sqrt{N}} = 0,0032$$

$$\text{écart type de } \hat{V} = \frac{1}{\sqrt{N - M}} = 0,0035$$

On est alors en mesure de représenter, comme on l'a déjà fait au début du chapitre précédant la fréquence des familles de chaque type par l'expression:

$$\left(\frac{1 + A\,U}{2}\right)\left(\frac{1 + U + V}{2}\right)^{R_g}\left(\frac{1 + U - V}{2}\right)^{R'_g}\left(\frac{1 - U + V}{2}\right)^{R_f}$$

$$\left(\frac{1 - U - V}{2}\right)^{R'_f}$$ avec les notations A, Rg et R_f du chapitre précédent.

Toujours, conformément au même principe d'approximation, on pourra calculer les fréquences relatives dans l'ensemble des familles de n = g + f enfants, de la calsse des familles de g garçons et f filles comme

$$N_{gf} = \left[\begin{matrix} g + f \\ g \end{matrix}\right] \times 2^{-(g+f)} \times \left[1 + (g - f)\,U + (g+f-1- 2\,C_{gf})\right] v$$

où C_{gf} désigne la valeur moyenne ou nombre des "coupures".

Pour les premiers histogrammes (g + f = 5, 6, 7, 8) on a comparé les valeurs observées dans les différentes classes aux valeurs théoriques.

a- dans l'hypothèse adoptée d'un tirage en chaine simple de MARKOFF.

b- dans l'hypothèse d'une répartition binomiale.

Les paramètres étant estimés sur la base de toutes les famille il semble logique de prendre pour nombre des degrés de liberté pour les χ^2 relatifs aux familles de n enfants la valeur n et non pas n - et n - 2 comme on devrait le faire si $\hat{U}$ et $\hat{V}$ étaient estimés pour chaque valeur de g + f.

On constate (cf.tableau N° 8) qu' l'hypothèse de MARKOFF rend compte des observations d'une manière beaucoup plus précise sauf en ce qui concerne n = 8) .

Les probabilités des χ^2_1 (schéma binomial) et χ^2_2 schéma de MARKOFF) sont d'ailleurs dans les intervalles suivants:

	n = 5	n = 6	n = 7	n = 8
χ^2_1 =	$0,05 \leqslant 0,10$	$0,10 \leqslant 0,20$	$0,02 \leqslant 0,005$	$0,10 \leqslant 0,20$
χ^2_2 =	$0,30 \leqslant 0,50$	$0,80 \leqslant 0,90$	$0,50 \leqslant 0,70$	$0,05 \leqslant 0,10$

montrant un bon accord entre l'hypothèse et la réalité.

Il est intéressant de remarquer qu'une cause de désaccord, entre les deux hypothèses et l'observation, pour n = 5 est le faible nombre de familles composées exclusivement de filles; peut être faut-il ne voir là que l'expression de causes économiques et sociales qui pousseraient particulièrement à de nouvelles grossesses, le parents n'ayant eu que des enfants du sexe féminin (?.?.?.)

Les chiffres sont trop faibles d'ailleurs pour pouvoir

constater l'augmentation du nombre des familles de 6 enfants du type f,f,f,f,f, g ou f,f,f,f,f,f, qui devrait en résulter.

On pourrait calculer également la valeur théorique de la variance de ces répartitions qui devraient être (pour n enfants)

$$\frac{n}{4} - \frac{n^2}{4} U^2 + \frac{1}{n^{2n}} \times h(n) \hat{V}$$

ou h (n) est une expression complexe égale à

$$h(n) \sum_{J=0}^{n+1} \left[\begin{matrix} n \\ J \end{matrix}\right] \left[4 J^4 - 4 n J^3 + (n^2 - n) J^2\right]$$

Enfin, on trouverait pour valeur théorique du nombre moyen des"coupures" dans les familles de g garçons et f filles:

$$C_{gf} \quad \frac{1 + (g-f) U + (g+f-1-2D_{gf}) V}{1 + (g-f) U + (g+f-1-2C_{gf}) V}$$

où C_{gf} , comme précédemment est le nombre moyen des"coupures" dans l'hypothèse nulle = $\frac{2gf}{g+f}$ et D_{gf} une constante liée à la variance des "coupures" et égale à $\frac{2gf - 1}{g + f - 1}$

Quelques exemples numériques, comme celui des familles de cinq garçons et cinq filles, où la valeur moyenne du nombre des "coupures" devient 4,98 au lieu de 5,00, suffiraient à convaincre combien est faible, pratiquement, l'influence de l'incrément.

Tableau N° 8

Comparaisons entre les histogrammes observés et les schémas probabilistes

Nombre de garçons	Nombres observés	Nombres théoriques dans l'hypothèse binomiale	Nombres théoriques dans l'hypothèse de MARKOFF	
		FAMILLES DE CINQ ENFANTS		
5	204	171	190	
4	841	833	852	chi carré = 9,73
3	1.585	1.624	1.386	
2	1.544	1.581	1.543	chi carré = 5,88
1	810	769	788	
0	144	150	169	
		FAMILLES DE SIX ENFANTS		
6	65	55	63	
5	330	321	336	
4	785	783	776	chi carré = 9,75
3	979	1.017	987	
2	717	743	733	chi carré = 2,24
1	325	289	304	
0	54	47	54	
		FAMILLES DE SEPT ENFANTS		
7	19	16	19	
6	126	109	117	
5	304	320	323	CHI carré = 15,00
4	529	520	507	
3	480	507	493	chi carré = 8,16
2	287	296	208	
1	111	96	104	
0	21	13	10	
		FAMILLES DE HUIT ENFANTS		
8	6	7	5	
7	49	55	60	
6	170	185	190	chi carré = 11,34
5	406	356	351	
4	412	444	432	chi carré = 14,44
3	327	348	343	
2	180	171	176	
1	61	47	52	
0	8	6	9	

DISCUSSION SOMMAIRE DES METHODES EMPLOYEES

On a volontairement employé uene démarche progressive, cherchant à serrer de plus en plus près la réalité par le jeux dialectique des hypothèses biologiques successives, traduites statistiquement en hypothèses nulles et modèles probabilistes et confrontées à l'observation.

Une autre approche eut été possible, partant directement du schéma de MARKOFF et l'éprouvant par une double méthode d'estimation que l'on peut brièvement esquisser.

Il est possible,en effet, d'estimer de deux manières l'incrément V lié à la succession immédiate de deux naissances de même sexe= a-:par la seule connaissance de l'histogramme des répartitions b-: pour chaque classe de familles de g garçons et de f filles par la connaissance de la valeur observée du nombre des "coupures".

a- En employant la première méthode et en se contentant de l'approximation déjà constamment utilisée, on trouve pour valeur de $\hat{V}$ une expression simple, absolument correcte (unbiassed), asymptotiquement équivalente à la solution de l'équation au maximum de vraisemblance à laquelle elle est tangente quand V est un infiniment petit de premier ordre:

$$\hat{V} = \sum_{g} \frac{N_g \; H_g}{K^2 \; M}$$

la sommation étant étendue, pour l'ensemble des M familles de (g + f) = (constante) enfants, à toutes les valeurs de g : les H_g sont liés simplement au nombre moyen des "coupures" et à la probabilité

$$\frac{1+V}{2} \qquad \text{et} \qquad K = \frac{2(g+f-1)}{g+f}$$

La varxiance de cette estimation est alors sensiblement égale à (K M) - 1 .

b- En considérant comme seules connues les familles de g garçon et f filles et la fréquence relative M_c de celles qui présentent C "coupures", on trouve une deuxième estimation, également tangente à l'équation du maximum de vraisemblance, mais affectée d'une erreur systématique, facile à éliminer.

Il serait donc possible de comparer ces deux estimations à l'écart type de leur différence et de tester l'hypothèse nulle de la non existence d'une corrélation résiduelle entre les sexes des naissances dans une même famille.

On ne discutera pas, ici, ces calculs; mais on montrera seulement la perte d'information sur $\hat{V}$, qui résulterait de la seule connaissance des histogrammes pour M familles de g + f enfants; en effet dans les documents employés, chaque famille apportait une information de g + f - 1 ; alors que l'on aurait ici seulement

$$\frac{2(g+f-1)}{g+f}$$

La perte d'information est donc de $\frac{g+f}{2}$.

Dans le cas présent, on peut évaluer que pour obtenir la même précision, il eut fallu recueillir 3,1 fois plus de familles,

soit environ 43.000.

Il n'est pas inintéressant, non plus, à la lumière des chiffres calculés de voir rapidement ce que peut apporter comme information un certain ensemble de documents et tout particulièrement le nombre minimum de familles qu'il faudrait rassembler pour pouvoir l'opposer à la valeur de l'incrément V, obtenue ici.

Comme il ne s'agit là que d'indications très sommaires, on prendra deux cas extrêmes pour les familles de trois garçons et de trois filles:

1°- Nombre M ' minimum de famille requis pour établir que V est significativement plus faible que 0,0295 si la valeur de $\hat{V}$ dans l'échantillon est zéro; on a sans peine

$$0{,}0295 \geqslant 3\sqrt{\frac{1}{81.495} + \frac{1}{5\,M'}}$$

soit M' plus grand que 2.388.

2°- Valeur moyenne maxima de V' dans un échantillon de cinquante familles; pour établir que V est significativement plus petit que 0,0295, on a :

$$0{,}0295 - V' \geqslant 3\sqrt{\frac{1}{81.495} + \frac{1-V^2}{250}}$$

soit V' plus petit que - 0,56; soit encore S plus grand que 245; ce qui implique notamment que moins de quinze familles doivent avoir un nombre de "coupures" inférieur à la valeur moyenne 3.

Ces quelques remarques sont suffisantes pour montrer, on l'espère, combien ces effets de "groupage" sont difficiles à étudier sur des faibles populations et combien pése peu dans la discussion un échantillon limité à quelques dizaines de cas même s'il

s'agit de familles très nombreuses: une famille de dix sept enfants par exemple, n'apporte que deux fois plus d'information sur ce point qu'une famille de neuf enfants et quatre fois plus qu'une famille de cinq enfants.

C O N C L U S I O N S

L'étude a porté sur 14.330 familles de [5 à 19 enfants], soit au total 95.725 naissances; étant donné le but poursuivi, une sélection préalable des documents a permis de ne retenir que les familles françaises sans multiparité; divers aspects concrets, tels que le milieu économique et social, l'âge et la santé des parents, l'influence des causes économiques et psychologiques- pouvant favoriser tel ou tel type de famille- ont du être négligés au profit de phénomènes statistiques plus généraux.

° ° °

1°- Les résultats suivant peuvent être considérés comme établis:

L'hypothèse d'une répartition des sexes à la naissance représentable par un mécanisme de tirages successifs indépendants dans une urne à composition constante doit être rejetée.

- en raison de l'étalement trop grand des histogrammes, c'est à dire l'excès des familles à forte prédominance masculine (ou féminine) sur les valeurs théoriques qui résulteraient d'une répartition binomiale (hypothèse a) issue du schéma précédent;

- en raison de la diminution du nombre moyen des séquences de même sexe, traduisant une liaison entre naissances consécutives (hypothèse b) indépendantes du fait précédent.

++++

IIº- Au degré d'approximation, que permet le montant total de l'information recueilli, un modèle statistique de tirage éléatoire en chaine (simple) de MARKOFF a été construit, qui rend compte de l'ensemble des chiffres obtenus d'une manière satisfaisante.

++++

IIIº- Par contre quatre hypothèses (c, d, e, f .) n'ont pu être définitivement confirmées ou infirmées:

- L'amélioration de la prévision due à la connaissance du sexe de l'enfant précédent est différent selon les sexes (hypothèse c).

- La probabilité de naissance d'un garçon ou d'une fille varie selon le rang de celle-ci (hypothèse d.)

- Se surajoutant à la corrélation entre naissances immédiatement consécutives-

- il existe une corrélation résiduelle non nulle entre naissances non immédiatement consécutives (hypothèse e).

- il existe une variation, entre familles, de la probabilité de naissance d'un garçon (hypothèse f).

°
° °

Mais si l'on a tenu à présenter ainsi abstraitement les résultats de cette enquête, afin d'en rendre plus clair l'énoncé, il convient évidemment d'avoir toujours présent à l'esprit le contenu concret, biologique des hypothèses et modèles probabilistes qui seul légitime l'emploi de ces derniers; simultanément, il est nécessaire de voir les limitations internes qu'imposait à ce

travail la nature même des documents rassemblés.

Le rejet de l'hypothèse nulle (hypothèse a) de la répartition binomiale traduit le fait déjà constaté par d'autres auteurs que l'hypothèse d'une détermination du sexe par un mécanisme (inégale vitalité de l'embryon ou fécondation élective .. etc.. etc..) constant d'une grossesse à l'autre et caractérisé par les mêmes paramètres pour tous les couples ne peut être conservée.

L'existence d'une liaison entre naissances immédiatement consécutives (hypothèse b) rendant compte des faits observés, peut traduire deux hypothèses biologiques bien distinctes qui sont:

A- L'hypothèse adoptée pour donner un modèle probabiliste : il existe une influence masculinisante (ou féminisante) d'une grossesse sur une grossesse immédiatement consécutive , égale pour les deux sexes (hypothèse d).

B- Il existe des périodes plus ou moins androphiles(ou gynophiles), qui, pour le volume d'informations recueilli ne se manifestent de manière significative que par leurs actions sur les naissances contigües; mais qui pourraient (et peut être simultanément).

varier de manière systèmatique, constante d'une famille à l' autre et en fonction de l'âge et du passé obstrétrical de la mère (hypothèse e).

avoir une plus ou moins grande durée (3, 4 gestations au plus) étant dues au milieu ou au genre de vie des parents (hypothèse c).

être caractéristique d'un couple donné (et peut être lié à des particularités du génotype des parents) - hypothèse f).

Ce n'est que l'élargissement (enrichissement) de la nature des documents de base, l'observation et la comparaison avec d'autres espèces, où l'expérimentation serait possible; qui pourraient permettre de préciser ces faits; mais surtout d'élucider les mécanismes (letalité ou fécondation élective ou détermination épigammique) par lesquels influent les causes dont on a pu mettre en évidence un des effets les plus marquants.

Le Président de Thèse

R. T U R P I N

VU

Le Doyen

L. B I N E T

VU et PERMIS D'IMPRIMER

Le Recteur:

SARRAIL

REFERENCES

La liste suivante n'a évidemment aucune prétention à être même l'ébauche d'une bibliographie sur un sujet qui a tellement fai couler d'encre et dont les ramifications sont si nombreuses.

On a cité quelques travaux particulièrement représentatifs des thèses étudiées ici statistiquement. En outre, pour certains points de technique statistique, qui risquaient d'être peu familiers aux lecteurs, on a tenu à indiquer les mémoires originaux.

°
° °

1- CIOCCO.A.1938. Variations in the sex ratio at birth in the United States- Human Bio.10. pp 36-44

2- CREW.F.A.E.1937. The sex ratio. Amer.Nat.71. pp 529-559. British Assoc.Report.1937

3- FISCHER?R.A.1944. Statistical Methods for research workers p 129

4- FRIEDA.S.SWED and EISENHART.C. 1943. Tables for testing randonne of grouping in a sequence of altenartives. An.of Math.Stat Vol 14.

5- GEISSLER.A.1889. Beiträge der K.Sächsichen Stat.Bureau 35 pp 1-2

6- KENDALL.M.G. and BABINGTON SMITH B. 1938. Distribution of digite obtained by an observer in using a Randonising machine supp J.R.Stat.Soc. pp 6-21

7- PARKES.A.S. 1938. The factures governing the mamalian sex ratio Sci.Progress 18 pp 426-435

8- PEARSON.E.S. 1938. The probability integral transformation for testing goodness of fit and combining independant test of significance. Biom.10. pp 134

9- STEVENS.W.L. 1939.Distribution of groupe in a sequence of alternatives. Ann.of Eug. Vol IX.

10- TSCHERMAK SEYSENEG.G.A. 1939. Versuch einer Mendelistischen Erklärung des Geschlechteverhältniss des menschlischen Neugeborenen. Bio.Generalis 14 pp 542-551

11- WALD.A. and WOLFOWITZ.J.1940.On a test of wheter two samples are from the same population. Ann.of Math.Stat. Vol XI.

12- WICKSELL.S.D. 1916. Sex proportion and parental age. Kungl Fysiogra. Sälls Handl. 37 . N° 6 .