COMMENT MESURER LA PROFONDEUR GÉNÉALOGIQUE D'UNE ASCENDANCE ?*

Les travaux de démographie historique ou de génétique s'appuient souvent sur des reconstitutions généalogiques. Les résultats de leur exploitation dépendent de l'information obtenue sur l'ascendance des individus. La mesure de paramètres génétiques, tels que le coefficient de parenté entre deux personnes ou le coefficient de consanguinité d'un individu, repose sur l'information disponible, c'est-à-dire sur le nombre d'ascendants identifiés et le nombre de générations sur lequel porte l'analyse. Comparer entre individus d'une même population leurs coefficients de consanguinité respectifs peut mettre en évidence des apparentements plus ou moins lâches au sein d'une même grande famille. Ces coefficients dépendront à la fois du nombre d'ancêtres communs repérés dans l'ascendance mais aussi du nombre de chemins possibles parvenant à ces ancêtres, d'autant plus nombreux que les mariages entre apparentés sont fréquents. Marie-Hélène CAZES** et Pierre CAZES*** *s'interrogent ici sur la possibilité de comparer les résultats obtenus dans deux ascendances.*

I. – Position du problème

Traditionnellement, on définit l'information fournie par une généalogie ascendante par son degré de complétude, qui se mesure par l'*indice de complétude* d'une table d'ascendance C_x.

Celui-ci est le rapport du nombre d'ascendants connus au nombre d'ascendants attendus, à chaque génération x. Il se calcule par la formule :

$$C_x = \frac{\text{nb ascendants connus}}{\text{nb ascendants attendus}}$$

où le nombre d'*ascendants attendus* à la génération x est donné par la formule 2^x, la génération des parents étant la première. On peut aussi calculer

* Nous remercions Gil Bellis, qui est à l'origine de cet article et Daniel Courgeau, dont les suggestions ont permis de l'améliorer considérablement.

** INED.

*** LISE-CEREMADE, Université Paris-Dauphine.

un indice de complétude *cumulé*, rapport de l'ensemble des ascendants connus à l'ensemble des ascendants attendus *jusqu'à* la génération x (Jetté, 1991).

Le doublement du nombre d'ascendants à chaque génération est une notion théorique. Dans la réalité, il arrive souvent que des ascendants apparaissent plusieurs fois dans une table d'ascendance, du fait des apparentements qui s'établissent entre conjoints. Ce sont des *ascendants répétés*.

Un indice est également utilisé, l'*implexe des ascendants*, qui peut s'interpréter comme un indice global de la parenté biologique unissant les ancêtres du probant. Il se calcule par la formule :

$$I_x = \frac{\text{nb ascendants différents}}{\text{nb ascendants attendus}}$$

Plus il est faible, plus la parenté par consanguinité (ou parenté biologique) est élevée.

Ces deux indices ne rendent pas compte du nombre moyen de générations prises en compte. Il est donc utile de définir un indice qui caractériserait la profondeur généalogique d'une ascendance reconstituée. Celui-ci permettrait la comparaison de paramètres entre populations, tel le coefficient moyen de consanguinité. Une telle comparaison ne peut, en effet, se faire qu'à nombre égal de générations. Encore faut-il que le recueil généalogique soit homogène sur l'ensemble de ses branches. Il est fréquent d'obtenir des « arbres » hétérogènes, certaines ramifications remontant très haut tandis que d'autres sont très vite stoppées, dès les générations des aïeux par exemple.

K. Kouladjian (1986) a introduit, dans cet esprit, une mesure d'entropie généalogique, basée sur une formule physique qui désigne la quantité d'information disponible d'un système. Il la définit comme suit : si on transforme une généalogie G en un arbre binaire B, tel que les individus qui paraissent plus d'une fois comme ancêtre sont considérés comme des individus distincts et si p_i représente la probabilité de l'origine du gène provenant du fondateur[1] i dans cet arbre B, alors :

$$S_B = -\sum_i p_i \log p_i$$

$$= \sum_i N_i \frac{1}{2^{N_i}}$$

où log représente le logarithme à base 2 ;

$N_{i,}$ est la génération du fondateur i ;

$$p_i = \frac{1}{2^{N_i}}$$

et la sommation porte sur tous les fondateurs de la généalogie.

[1] Dans une généalogie, un fondateur est un individu dont les deux parents sont inconnus.

S_B est la valeur attendue de la génération des fondateurs et peut servir comme mesure du degré d'enracinement des ascendances.

Sa variance se définit comme suit :

$$V = \sum_i p_i \log^2 p_i - S_B^2$$

$$= \sum_i N_i^2 \frac{1}{2^{N_i}} - S_B^2$$

L'interprétation du paramètre S_B présenté comme la valeur attendue de la génération des fondateurs, n'est cependant pas très claire dès que les fondateurs figurent en de nombreuses générations différentes.

E. Létourneau et F. Mayer (1988) utilisent ce même paramètre comme un coefficient moyen de complétude, afin de synthétiser l'analyse générationnelle dans une généalogie ascendante. Ils le désignent comme *la moyenne de la profondeur généalogique*. D'autres auteurs ont adopté, à leur suite, cette terminologie (De Braekeleer et Bellis, 1994) et utilisent la formulation équivalente suivante :

$$P = \sum_i i \frac{F_i}{T_i}$$

où i est le niveau de génération ;

F_i le nombre de fondateurs à la génération i ;

T_i le nombre d'ascendants attendus à la même génération i ;

et la sommation porte sur les générations.

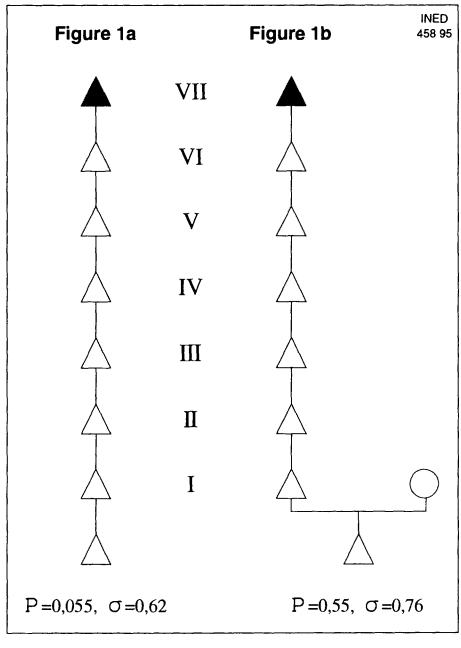
Nous voudrions montrer, dans cet article, que le qualificatif de *profondeur généalogique,* valable dans certaines conditions seulement, n'est pas satisfaisant et que ce paramètre peut conduire à des incohérences en ce qui concerne la mesure de l'information qu'il fournit. Nous proposerons une mesure légèrement différente pour supprimer ces incohérences. Elle devrait mieux répondre au qualificatif de profondeur généalogique et pourra s'appliquer, nous l'espérons, à n'importe quelle généalogie.

II. – Analyse de cas spécifiques

Un probant et ses deux parents connus constituent la plus petite ascendance. Dans ce cas, la valeur de P vaut 1 avec une variance nulle, ce qui semble normal.

Prenons le cas d'une ascendance en ligne directe agnatique (ou utérine, ou cognatique, peu importe), sur 7 générations (figure 1a). La mesure de P nous fournit alors $P = 0{,}055$ et son écart type (racine de la variance) $\sigma = 0{,}62$.

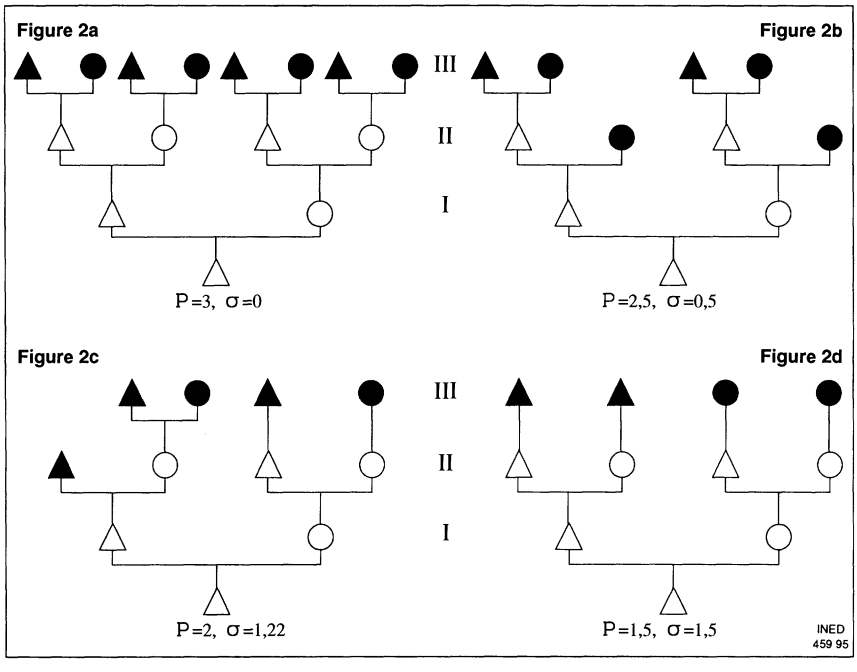
Ainsi P est quasiment nul et la valeur de son écart type est considérable, très supérieure à la moyenne. Si on ajoute à cette ascendance la connaissance de la mère (figure 1b), alors $P = 0{,}55$ et $\sigma = 0{,}76$. Là encore, bien que la génération parentale soit connue, P est de l'ordre de $\frac{1}{2}$ seulement, avec un écart type supérieur à P. Ceci prouve que le calcul de P n'a, dans ces cas, pas beaucoup de sens.



Examinons maintenant une généalogie (figure 2a) portant sur trois générations de parents. Si l'ascendance est complète (les huit arrière-grands-parents sont connus) P fournit la valeur 3 (avec $\sigma = 0$). On peut généraliser cet exemple et observer que, dans le cas d'une ascendance complète sur n générations, le calcul de P correspond bien à la profondeur généalogique attendue n, avec un écart type nul. Les difficultés commencent, bien sûr, quand l'information devient partielle. Examinons les résultats si dans cette ascendance on perd l'identité de quatre des arrière-grands-parents. Ceci peut se réaliser selon trois modalités : ou bien il s'agit de deux couples de parents qui demeurent inconnus (figure 2b) ; ou bien d'un couple et de deux autres parents isolés, quel que soit le sexe (figure 2c) ; ou bien d'un seul parent de chacun des quatre grands-parents, quel que soit leur sexe (figure 2d). Selon le schéma retenu, la valeur de P sera différente passant de 2,5 dans le premier cas, à 1,5 dans le dernier. L'écart type correspondant est minimal dans le premier cas de figure et va en augmentant à mesure que les parents isolés sont plus nombreux. Un individu identifié n'apporte donc pas le même degré d'information selon qu'il complète un couple ou non, ce qui n'est pas satisfaisant.

Le tableau 1 montre comment varie la valeur de P, en fonction du schéma retenu (nombre d'individus identifiés en génération 3 selon qu'ils sont isolés ou faisant partie d'un couple), mettant en évidence les incohérences de la mesure de ce paramètre. Identifier les quatre grands-parents d'un probant revient à avoir une valeur de P égale à 2 (figure 3a), mais si on rajoute l'identification d'un seul parent de chacun des quatre grands-

parents alors on est dans la situation la plus défavorable en matière d'information : cela fait diminuer le paramètre *P* de 0,5 par rapport à la connaissance des quatre grands-parents (figure 2d) ! Remplaçons l'identification des quatre personnes précédentes de la génération 3 par celle de deux personnes isolées et d'un couple (figure 3b, identique à 2c) et nous retrouvons tout juste la valeur de 2 pour *P* (avec cependant un écart type élevé, ce qui relativise beaucoup le sens d'une telle mesure).



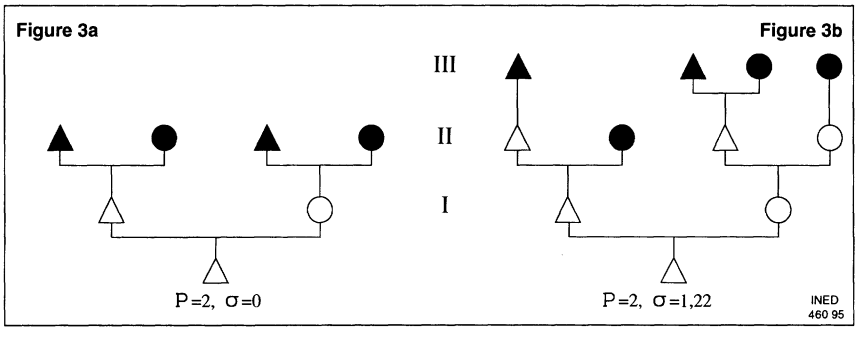


TABLEAU 1. – VALEUR DU PARAMÈTRE *P*
SELON L'INFORMATION POSSÉDÉE EN GÉNÉRATION 3

	Nb d'individus identifiés mais ne faisant pas partie d'un couple		Nb de couples identifiés	*P*	σ
Génération 2			2 couples	2	0
Génération 3	0 individu		0 couple	2	0
	1 individu			1,875	0,78
			1 couple	2,25	0,43
	2 individus			1,75	1,09
	3 individus			1,625	1,32
	1 individu	et	1 couple	2,125	0,93
			2 couples	2,5	0,5
	4 individus			1,5	1,5
	2 individus	et	1 couple	2	1,225
	1 individu	et	2 couples	2,375	0,99
	3 individus	et	1 couple	1,875	1,45
			3 couples	2,75	0,43
	2 individus	et	2 couples	2,25	1,3
	1 individu	et	3 couples	2,625	0,99
			4 couples	3	0

Ainsi, avoir un fondateur en une génération donnée est plus favorable (du point de vue de la valeur de l'indice *P* qui est plus élevé) que si on connaît un seul de ses parents. En revanche, l'identification des deux parents de ce fondateur (qui lui ôte son statut de fondateur) améliore l'information et fait croître *P*.

III. – Une nouvelle mesure de la profondeur généalogique

La source des incohérences que nous venons de relever dans les exemples qui précèdent provient de ce que la présence de « semi-fondateurs » – c'est-à-dire d'individus dont on ne connaît qu'un seul parent – n'est pas prise en compte dans la formule de *P* définie plus haut.

En effet, tout individu d'une généalogie, à commencer par le probant lui-même, peut être :

a) fondateur (ses deux parents sont inconnus) ;

b) ou semi-fondateur (un seul parent est connu) ;

c) ou totalement identifié (ses deux parents sont connus).

Dans les deux premiers cas, l'information généalogique s'arrête ou devient tout au moins partielle. C'est alors que la mesure d'une profondeur généalogique commence à intervenir : à chaque génération, chaque fondateur contribue à *sa* génération (du point de vue de la profondeur généalogique) d'un facteur égal à $\frac{1}{T_i}$ où $T_i = 2^i$ représente le nombre d'ascendants attendus à cette génération et l'ensemble des fondateurs ont une contribution de $\frac{F_i}{T_i}$ (F_i étant le nombre de fondateurs à la génération i). Mais chaque semi-fondateur contribue également à cette génération d'un facteur réduit de moitié, égal à $\frac{1}{2T_i}$ et l'ensemble des semi-fondateurs ont une contribution de $\frac{S_i}{2T_i}$ (S_i étant le nombre de semi-fondateurs à la génération i). Ainsi, à une génération donnée i, chaque fondateur est associé à une probabilité $\frac{1}{T_i}$, chaque semi-fondateur à une probabilité $\frac{1}{2T_i}$ et chaque individu identifié à une probabilité nulle. S'agissant d'événements disjoints, on peut montrer (voir annexe, §1) que la somme des probabilités associées aux fondateurs et aux semi-fondateurs, sur les générations successives, est égale à l'unité. On a :

$$\sum_{i=0}^{n} \left(\frac{F_i}{T_i} + \frac{S_i}{2T_i} \right) = 1$$

Remarque : on montre en annexe que la contribution d'un individu en génération g, qu'il soit fondateur, semi-fondateur ou identifié, et de tous ses ascendants, est égale à $\frac{1}{2^g}$. On peut donc, si on le désire, arrêter la sommation précédente à n'importe quel niveau g de génération ($g \leq n$), auquel cas les individus qui sont en génération g deviennent tous fondateurs.

Si *Ego*, l'individu dont on estime la généalogie, est en génération 0, il peut lui-même être semi-fondateur et sa contribution à la génération 0 sera de $\frac{1}{2}$.

Si on associe, à chaque génération considérée, la suite des chiffres (*0, 1, 2,...., n*), on peut calculer la « moyenne de la profondeur des générations », M_1 comme étant égale à :

$$M_1 = \sum_{i=0}^{n} i \left(\frac{F_i}{T_i} + \frac{S_i}{2T_i} \right)$$

où i est le niveau de génération.

Et la variance de la profondeur des générations s'écrira :

$$V_1 = \sigma_1^2 = \sum_i i^2 \left(\frac{F_i}{T_i} + \frac{S_i}{2T_i} \right) - M_1^2$$

La prise en compte des semi-fondateurs supprime les problèmes posés ci-dessus et rend identique l'information apportée par un individu, qu'il soit isolé ou qu'il complète un couple. Le tableau 1 se modifie alors de la façon suivante (tableau 2).

TABLEAU 2. – VALEUR DU PARAMÈTRE M_1
SELON L'INFORMATION POSSÉDÉE EN GÉNÉRATION 3

	Nb d'individus identifiés mais ne faisant pas partie d'un couple		Nb de couples identifiés	M_1	σ_1
Génération 2			2 couples	2	0
Génération 3	0 individu		0 couple	2	0
	1 individu			2,125	0,33
			1 couple	2,25	0,43
	2 individus			2,25	0,43
	3 individus			2,375	0,48
	1 individu	et	1 couple	2,375	0,48
			2 couples	2,5	0,5
	4 individus			2,5	0,5
	2 individus	et	1 couple	2,5	0,5
	1 individu	et	2 couples	2,625	0,48
	3 individus	et	1 couple	2,625	0,48
			3 couples	2,75	0,43
	2 individus	et	2 couples	2,75	0,43
	1 individu	et	3 couples	2,875	0,33
			4 couples	3	0

On constate ici que M_1 augmente à mesure que s'ajoute l'identification d'un individu en génération 3, comme le bon sens nous le faisait attendre. De plus, les écarts types correspondants ont des valeurs vraisemblables, suffisamment faibles en regard de M_1 pour donner à ce dernier la signification d'une *profondeur généalogique.*

Le tableau 3 permet la comparaison des indices P et M_1 avec leurs écarts types respectifs dans chacune des généalogies examinées jusqu'ici.

Dans la généalogie 1a, le calcul de M_1 montre que 7 générations en ligne directe issues d'un seul parent d'*Ego* équivalent tout juste à la connaissance des deux parents. Le fait de rajouter la mère (généalogie 1b) fait gagner une demi-génération en faisant diminuer l'écart type (ce qui est conforme au bon sens).

TABLEAU 3. – COMPARAISON DES INDICES P ET M_1 POUR QUELQUES GÉNÉALOGIES

	P	σ	M_1	σ_1
Généalogie 1a	0,055	0,62	0,99	1,37
1b	0,55	0,76	1,49	1,07
2a	3	0	3	0
2b	2,5	0,5	2,5	0,5
2c	2	1,225	2,5	0,5
2d	1,5	1,5	2,5	0,5
3a	2	0	2	0
3b	2	1,225	2,5	0,5

Proposons-nous d'observer la mesure de ces indices sur des ascendances un peu plus profondes, dans lesquelles on a fait figurer les fondateurs en noir et les semi-fondateurs en grisé (figures 4, 5, 6).

La généalogie représentée en figure 4 porte sur quatre ou cinq générations. Le calcul de M_1 donne une profondeur généalogique de 4,34 avec un écart type de 0,47. Nous sommes ici très proches du schéma d'une généalogie complète sur quatre générations de parents. La connaissance supplémentaire de onze individus (sur 32 attendus) en génération 5 augmente la profondeur généalogique de 0,34, c'est-à-dire exactement du rapport $\frac{11}{32}$. L'écart type – quoique faible – est déjà notable.

Dans la généalogie de la figure 5, la valeur de M_1 est la même, 4,34 mais l'écart type monte à 1,45. L'intérêt d'une telle mesure apparaît bien sur cet exemple. Si, dans le cas précédent, il était aisé de prédire la profondeur généalogique, il devient très difficile d'en faire autant sur une ascendance fortement hétérogène comme celle de la figure 5. Le calcul de σ_1 est un indicateur du degré d'hétérogénéité de l'ascendance. L'absence d'identification d'individus dans les premières générations (ici, une mère en génération 3) fait baisser sensiblement la profondeur généalogique en supprimant du même coup tous ses ancêtres propres (au nombre de 16 dans la génération 7, la plus haute) et provoque des déséquilibres entre les branches, ce qui joue sur la valeur de σ_1. Les tableaux 4 et 5 fournissent le détail du calcul de M_1 et de sa variance, pour les deux exemples précédents.

Il est très intéressant de noter (tableaux 4 et 5) que la moyenne de la profondeur généalogique peut aussi se calculer par la somme des complétudes sur l'ensemble des générations, à savoir :

$$M_2 = \sum_{i=1}^{n} \frac{O_i}{T_i}$$

où O_i représente le nombre d'individus observés (connus) à la génération i et T_i le nombre d'individus attendus à la génération i.

Et ceci répond au bon sens : si je connais les deux parents de la génération 1, j'ai une profondeur généalogique au moins égale à 1. Si je connais les quatre grands-parents, je monte à 2 ; si j'identifie les huit arrière-grands-parents, je monte à 3 et ainsi de suite. Si je n'identifie que six des arrière-grands-parents, je ne connais la génération 3 qu'aux trois quarts et m'attends donc à avoir une profondeur de 2,75. Nous montrons en annexe (§2) que les deux indices M_1 et M_2 sont, en effet, équivalents. On peut donc obtenir beaucoup plus simplement la moyenne de la profondeur généalogique en calculant, pour chaque génération, la proportion d'individus observés par rapport à leur effectif théorique et en faisant ensuite la somme de ces proportions sur l'ensemble des générations. Cela évite de distinguer « fondateurs » et « semi-fondateurs ». Mais qu'en est-il des variances ?

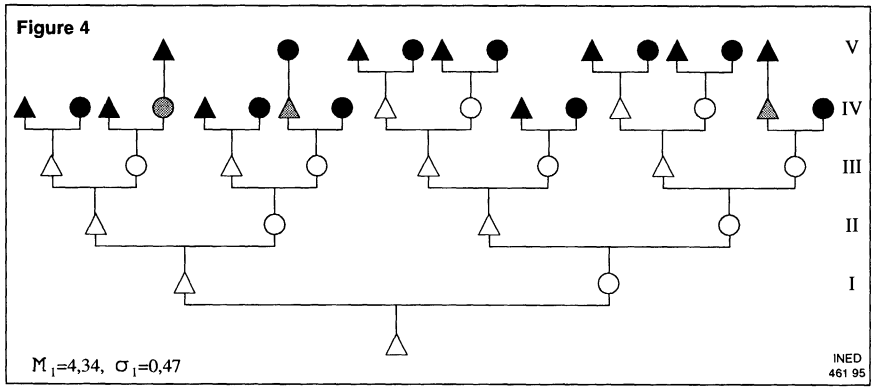
On peut, en effet, associer à ce deuxième indice une variance qui s'exprimerait par :

$$V_2 = \sum_{i=1}^{n} \frac{O_i}{T_i} \left(1 - \frac{O_i}{T_i}\right)$$

Nous montrons en annexe (§3) que V_2 ne fournit que dans certains cas la même valeur que la variance V_1 estimée avec le modèle complet. On ne peut l'utiliser que dans le cas d'une généalogie complète jusqu'en génération $n-1$ (voir annexe). L'avantage de passer par la comptabilisation des semi-fondateurs et des fondateurs, c'est donc de nous fournir, dans tous les cas, la mesure d'une variance associée à la profondeur généalogique.

Le calcul de M_2 est différent de la complétude cumulée, égale à 0,264 dans l'exemple de la figure 5 et qui représente le pourcentage d'ancêtres connus par rapport au nombre théorique d'ancêtres, à un niveau donné de génération (ici, au rang 7). Plus le nombre de générations prises en compte est élevé et plus faible sera la complétude cumulée du fait de l'accroissement du nombre d'ancêtres ; mais en revanche, plus forte sera la profondeur généalogique (surtout si la connaissance des branches est homogène).

Pour illustrer l'écart entre les deux mesures M_1 et P, incluant ou non les semi-fondateurs, nous donnons l'exemple d'une généalogie (figure 6, et détail des calculs dans le tableau 6) où la mesure de P (comptabilisation des fondateurs seulement) aurait donné une profondeur généalogique de 2,63 avec un écart type de 2,37. La prise en compte des semi-fondateurs



TAB. FIG. 4

Génération	Indiv. théor.	Indiv. obs.	Complétude	Complétude cumulée	Nb fond.	Nb semi fond.	Contribution à la moy.*		CTR à la Σ des carrés**	
							Fond.	Semi-fond.	Fond.	Semi-fond.
i	*t*	*o*	*o/t*		*Fi*	*Si*				
1	2	2	1	1	0	0	0	0	0	0
2	4	4	1	1	0	0	0	0	0	0
3	8	8	1	1	0	0	0	0	0	0
4	16	16	1	1	9	3	2,25	0,375	9	1,5
5	32	11	0,344	0,661	11	0	1,719	0	8,594	0
Total	**62**	**41**	**4,344**		**20**	**3**	**3,969**	**0,375**	**17,594**	**1,5**
							4,344		**19,094**	

* La contribution à la moyenne des fondateurs est calculée par i (F_i/t) ; celle des semi-fondateurs par i ($S_i/2t$).
** La contribution à la somme des carrés pour le calcul de la variance est de i^2 (F_i/t) pour les fondateurs et de i^2 ($S_i/2t$) pour les semi-fondateurs.

Figure 5

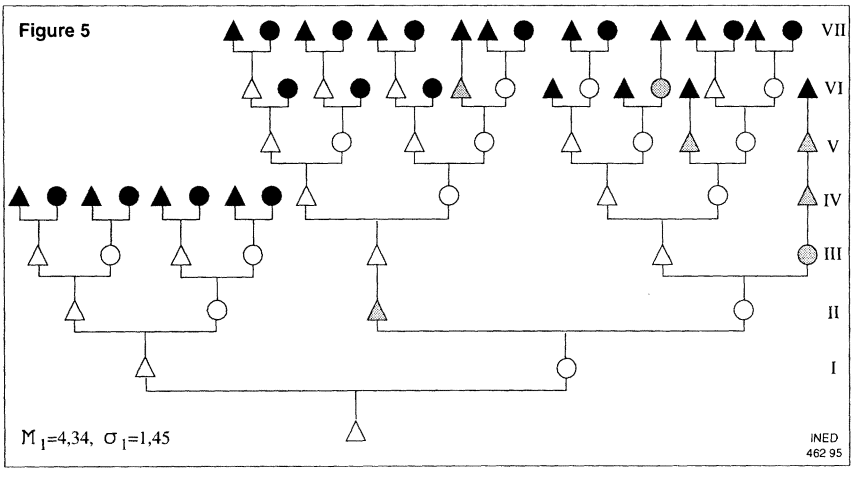
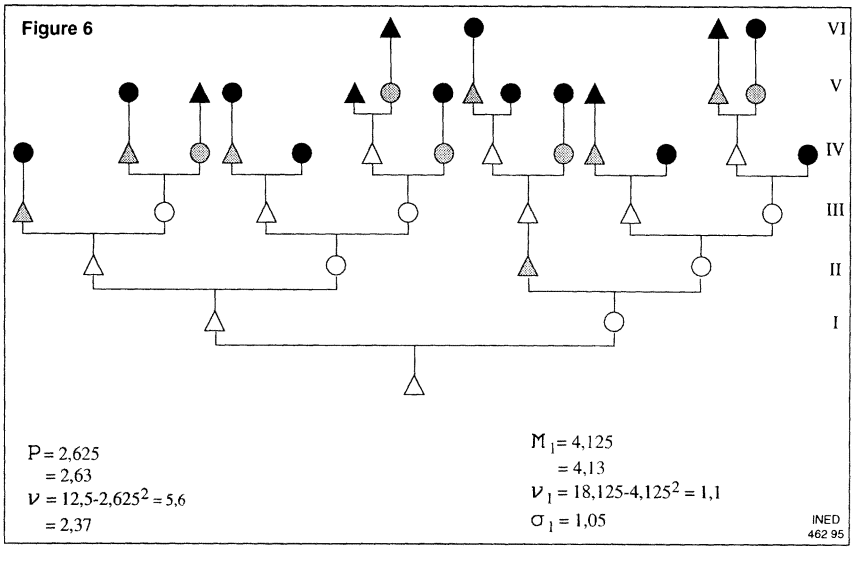


Figure 6



TAB. FIG. 5

Génération	Indiv. théor.	Indiv. obs.	Complétude	Complétude cumulée	Nb fond.	Nb semi fond.	Contribution à la moy.*		CTR à la Σ des carrés**	
							Fond.	Semi-fond.	Fond.	Semi-fond.
i	*t*	*o*	*o/t*		*Fi*	*Si*				
1	2	2	1	1	0	0	0	0	0	0
2	4	4	1	1	0	1	0	0,25	0	0,5
3	8	7	0,875	0,929	0	1	0	0,188	0	0,563
4	16	13	0,813	0,867	8	1	2	0,125	8	0,5
5	32	9	0,281	0,565	0	2	0	0,156	0	0,781
6	64	16	0,25	0,405	7	2	0,656	0,094	3,938	0,563
7	128	16	0,125	0,264	16	0	0,875	0	6,125	0
Total	**254**	**67**	**4,344**		**31**	**7**	**3,531**	**0,813**	**18,063**	**2,907**
							4,344		**20,97**	

TAB. FIG. 6

Génération	Indiv. théor.	Indiv. obs.	Complétude	Complétude cumulée	Nb fond.	Nb semi fond.	Contribution à la moy.*		CTR à la Σ des carrés**	
							Fond.	Semi-fond.	Fond.	Semi-fond.
i	*t*	*o*	*o/t*		*Fi*	*Si*				
1	2	2	1	1	0	0	0	0	0	0
2	4	4	1	1	0	1	0	0,25	0	0,5
3	8	7	0,875	0,929	0	1	0	0,1875	0	0,5625
4	16	13	0,8125	0,867	4	6	1	0,75	4	3
5	32	12	0,375	0,613	8	4	1,25	0,3125	6,25	1,5625
6	64	4	0,0625	0,333	4	0	0,375	0	2,25	0
Total	**126**	**42**	**4,125**		**16**	**12**	**2,625**	**1,5**	**12,5**	**5,625**
							4,125		**18,125**	

* La contribution des fondateurs est calculée par $i\ (F_i/t)$; celle des semi-fondateurs par $i\ (S_i/2t)$.
** La contribution à la somme des carrés pour le calcul de la variance est de $i^2\ (F_i/t)$ pour les fondateurs et de $i^2\ (S_i/2t)$ pour les semi-fondateurs.

avec le calcul de M_1 ou celui de M_2 fait monter cette valeur à 4,13 avec un écart type bien moindre, de 1,05. Cet exemple montre que les formules M_1 ou M_2 sont plus appropriées que celle de P pour mesurer la profondeur généalogique d'une ascendance. Ainsi, quand on caractérise des individus par leurs coefficients moyens de consanguinité ou de parenté, on devrait toujours leur associer le calcul de l'indice M_1 (ou M_2), bons indicateurs de la quantité d'information contenue dans leurs ascendances respectives.

Marie-Hélène CAZES, Pierre CAZES*

ANNEXE

Dans une généalogie, tout individu est :

— soit fondateur (ses deux parents sont inconnus) ;

— soit semi-fondateur (un seul de ses parents est connu) ;

— soit identifié (ses deux parents sont alors connus).

Dans la suite de cette annexe, nous appellerons :

F_i le nombre de fondateurs à la génération i ;

S_i le nombre de semi-fondateurs à la génération i ;

I_i le nombre d'individus identifiés à la génération i ;

$T_i = 2^i$, le nombre d'individus attendus à la génération i.

Nous allons considérer successivement les expressions :

$$A_n = \sum_{i=0}^{n} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right)$$

$$M_{1n} = \sum_{i=0}^{n} \frac{\mathrm{i}}{2^i}\left(F_i + \frac{S_i}{2}\right)$$

$$M_{2n} = \sum_{i=1}^{n} \frac{(F_i + S_i + I_i)}{2^i}$$

$$V_{1n} = \sum_{i=0}^{n} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) - M_{1n}^2$$

* Avec la collaboration de S. Larrouy.

$$V_{2n} = \sum_{i=1}^{n} \left(\frac{F_i + S_i + I_i}{2^i} \right) \left(1 - \frac{F_i + S_i + I_i}{2^i} \right)$$

Nous allons montrer que :

1) $A_n = 1$, pour tout n ;

2) $M_{1n} = M_{2n}$, pour tout n, auquel cas on notera M_n la valeur commune ;

3) $V_{1n} = V_{2n}$ si, et seulement si, $M_{n-1} = n - 1$, ce qui exige que $M_i = i$, pour tout $i \leq n - 1$.

Les résultats précédents sont vrais trivialement pour $n = 1$.

On va donc supposer qu'à la génération n,

$$A_n = 1 \; ; \; M_{1n} = M_{2n} \; ; \; V_{1n} = V_{2n}$$

et on va montrer par récurrence que :

$$A_{n+1} = 1 \; ; \; M_{1n+1} = M_{2n+1} \; ; \; V_{1n+1} = V_{2n+1}, \text{ si } M_n = n.$$

Quand la généalogie porte sur n générations, on a en dernière génération n, F_n fondateurs et donc, $S_n = 0$ et $I_n = 0$.

Considérons maintenant $n + 1$ générations et appelons F'_i le nombre de fondateurs, S'_i le nombre de semi-fondateurs, I'_i le nombre d'identifiés, à la génération i. Quand on atteint la génération $n + 1$, trois cas sont possibles pour la génération n :

— un nombre F'_n d'ancêtres restent fondateurs, leurs parents demeurant inconnus, avec bien sûr, $F'_n \leq F_n$;

— un nombre S'_n d'ancêtres deviennent semi-fondateurs ;

— un nombre $I'_n = F_n - F'_n - S'_n$ d'ancêtres sont identifiés.

En génération $n + 1$, les S'_n ancêtres seront à l'origine de S'_n fondateurs, les I'_n ancêtres seront à l'origine de $2(F_n - F'_n - S'_n)$ fondateurs.

Le schéma ci-après résume la situation lorsque l'on passe d'une généalogie de n générations à $n + 1$ générations.

En génération $n + 1$,

$$F'_{n+1} = 2\,I'_n + S'_n = 2\,(F_n - F'_n - S'_n) + S'_n = 2\,(F_n - F'_n - \frac{S'_n}{2}) = 2\,\Delta_n$$

avec :

$$\Delta_n = F_n - F'_n - \frac{S'_n}{2}$$

On a aussi :

$F'_i = F_i,\ S'_i = S_i,\ I'_i = I_i$, pour tout i : $0 \le i \le n-1$

$F'_i + S'_i + I'_i = F_i + S_i + I_i$, pour tout i : $0 \le i \le n$

Passage de la génération n à la génération $n+1$			
généalogie de n générations		passage à $n+1$ générations	
en génération n		en génération n	en génération n + 1
	F_n	F'_n	$F'_{n+1} = 2I'_n + S'_n$
	$S_n = 0$	S'_n	$S'_{n+1} = 0$
	$I_n = 0$	$I'_n = F_n - F'_n - S'_n$	$I'_{n+1} = 0$
Total	F_n	F_n	$F'_{n+1} = 2\Delta_n$

1. – Démonstration $A_n = 1$

$$A_n = \sum_{i=0}^{n} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right)$$

Dans le cas d'une généalogie qui s'arrêterait dès la première génération, $n = 1$, les ancêtres attendus à cette génération sont les deux parents d'*Ego*, $T_1 = 2^1 = 2$.

Quelles que soient les conditions de départ (*Ego* totalement identifié, avec ses deux parents connus, ou *Ego* semi-fondateur, avec un seul parent connu), on vérifie aisément que $A_1 = 1$.

Supposons que $A_n = 1$, et calculons A_{n+1} :

$$A_{n+1} = \sum_{i=0}^{n+1} \frac{1}{2^i}\left(F'_i + \frac{S'_i}{2}\right)$$

$$= \sum_{i=0}^{n-1} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{1}{2^n}\left(F'_n + \frac{S'_n}{2}\right) + \frac{2\Delta_n}{2^{n+1}}$$

$$= \sum_{i=0}^{n-1} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{1}{2^n}\left(F'_n + \frac{S'_n}{2} + \Delta_n\right)$$

$$= \sum_{i=0}^{n-1} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{1}{2^n} F_n = \sum_{i=0}^{n} \frac{1}{2^i}\left(F_i + \frac{S_i}{2}\right)$$

$$= A_n = 1$$

A_1 étant égal à 1 et supposant que $A_n = A_1 = 1$, on a encore $A_{n+1} = A_n = 1$, on en déduit que $A_n = 1$, quel que soit n.

2. - *Démonstration* $M_{1n} = M_{2n}$

$$M_{1n} = \sum_{i=0}^{n} \frac{i}{2^i}\left(F_i + \frac{S_i}{2}\right)$$

$$M_{2n} = \sum_{i=1}^{n} \frac{O_i}{T_i} = \sum_{i=1}^{n} \frac{(F_i + S_i + I_i)}{2^i}$$

$$= \sum_{i=0}^{n} \frac{(F_i + S_i + I_i)}{2^i} - 1 \quad (2)$$

où O_i représente le nombre d'individus observés à la génération i et T_i le nombre d'individus attendus à la génération i.

On vérifie aisément que pour $n = 1$, on a bien $M_{11} = M_{21}$ (si les deux parents de *Ego* sont connus, on a $M_{11} = M_{21} = 1$, et si *Ego* est semi-fondateur, on a $M_{11} = M_{21} = \frac{1}{2}$).

Supposons que $M_{1n} = M_{2n}$, alors :

$$M_{1n+1} = \sum_{i=0}^{n+1} \frac{i}{2^i}\left(F'_i + \frac{S'_i}{2}\right)$$

$$= \sum_{i=0}^{n-1} \frac{i}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n}{2^n}\left(F'_n + \frac{S'_n}{2}\right) + \frac{n+1}{2^{n+1}} 2\Delta_n$$

(2) Quand on somme depuis la génération 0, il faut retrancher 1 de la formule du dessus donnant M_{2n} pour enlever la contribution triviale et égale à 1 de *Ego*.

$$= \sum_{i=0}^{n-1} \frac{i}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n}{2^n}\left(F'_n + \frac{S'_n}{2} + \Delta_n\right) + \frac{\Delta_n}{2^n}$$

$$= \sum_{i=0}^{n-1} \frac{i}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n}{2^n} F_n + \frac{\Delta_n}{2^n}$$

$$= \sum_{i=0}^{n} \frac{i}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{\Delta_n}{2^n}$$

$$= M_{1n} + \frac{\Delta_n}{2^n}$$

$$M_{2n+1} = \sum_{i=1}^{n+1} \frac{F'_i + S'_i + I'_i}{2^i} = \sum_{i=1}^{n} \frac{(F_i + S_i + I_i)}{2^i} + \frac{\Delta_n}{2^n}$$

$$= M_{2n} + \frac{\Delta_n}{2^n} = M_{1n} + \frac{\Delta_n}{2^n}$$

d'où

$$M_{1n+1} = M_{2n+1} = M_{1n} + \frac{\Delta_n}{2^n}$$

Les deux formules sont donc équivalentes. Par la suite, on pourra poser $M_{1n} = M_{2n} = M_n$.

3. – Démonstration $V_{1n} = V_{2n}$ *si* $M_{n-1} = n - 1$

Démontrons dans quelles conditions la variance V_{1n} associée à M_{1n} et la variance V_{2n} associée à M_{2n}, sont équivalentes :

$$V_{1n} = \sum_{i=0}^{n} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) - M_n^2$$

$$V_{2n} = \sum_{i=1}^{n} \left(\frac{F_i + S_i + I_i}{2^i}\right)\left(1 - \frac{F_i + S_i + I_i}{2^i}\right)$$

Pour $n = 1$, les deux variances sont nulles si les deux parents d'*Ego* sont connus, mais elles sont toutes deux égales à $\frac{1}{4}$ si *Ego* est semi-fondateur.

Supposant maintenant que $V_{1n} = V_{2n}$, calculons V_{1n+1} et V_{2n+1}.

$$V_{1n+1} = \sum_{i=0}^{n+1} \frac{i^2}{2^i}\left(F'_i + \frac{S'_i}{2}\right) - M_{n+1}^2$$

$$= \sum_{i=0}^{n-1} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n^2}{2^n}\left(F'_n + \frac{S'_n}{2}\right) + \frac{(n+1)^2}{2^{n+1}} 2\Delta_n - \left(M_n + \frac{\Delta_n}{2^n}\right)^2$$

$$= \sum_{i=0}^{n-1} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n^2}{2^n}\left(F'_n + \frac{S'_n}{2} + \Delta_n\right)$$

$$+ \frac{(2n+1)}{2^n}\Delta_n - \left(M_n^2 + \frac{2M_n\Delta_n}{2^n} + \frac{\Delta_n^2}{2^{2n}}\right)$$

$$= \sum_{i=0}^{n-1} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) + \frac{n^2}{2^n} F_n + \frac{(2n+1)}{2^n}\Delta_n - \left(M_n^2 + \frac{2M_n\Delta_n}{2^n} + \frac{\Delta_n^2}{2^{2n}}\right)$$

$$= \sum_{i=0}^{n} \frac{i^2}{2^i}\left(F_i + \frac{S_i}{2}\right) - M_n^2 + \frac{\Delta_n}{2^n}\left(2n + 1 - 2M_n - \frac{\Delta_n}{2^n}\right)$$

$$= V_{1n} + \frac{\Delta_n}{2^n}\left(1 - \frac{\Delta_n}{2^n} + 2\,(n - M_n)\right)$$

$$V_{2n+1} = \sum_{i=1}^{n+1} \frac{1}{2^i}(F'_i + S'_i + I'_i)\left(1 - \frac{1}{2^i}(F'_i + S'_i + I'_i)\right)$$

$$= \sum_{i=1}^{n} \frac{1}{2^i}(F_i + S_i + I_i)\left(1 - \frac{1}{2^i}(F_i + S_i + I_i)\right) + \frac{1}{2^{n+1}} 2\Delta_n\left(1 - \frac{2\Delta_n}{2^{n+1}}\right)$$

$$= V_{2n} + \frac{\Delta_n}{2^n}\left(1 - \frac{\Delta_n}{2^n}\right)$$

On en déduit que si $V_{1n} = V_{2n}$, alors $V_{1n+1} = V_{2n+1}$ si, et seulement si $M_n = n$, c'est-à-dire si toutes les générations sont complètes jusqu'à la génération n puisque $M_n = n \Leftrightarrow M_i = i$, pour tout $i \leq n$.

Si on s'arrête en génération 1 et si cette dernière est complète, on a :

$$M_1 = 1,\ V_{11} = V_{21} = 0.$$

On déduit donc par récurrence, de façon générale que :

$V_{1n} = V_{2n} \Leftrightarrow M_{n-1} = n-1 \Leftrightarrow M_i = i$, pour tout $i \leq n-1 \Leftrightarrow V_{1i} = V_{2i} = 0$, pour tout $i \leq n-1$.

La dernière implication est évidente puisque :

$$M_i = i = \sum_{j=0}^{i} j \frac{\left(F_j + \frac{S_j}{2}\right)}{2^j} \leq i \sum_{j=0}^{i} \frac{\left(F_j + \frac{S_j}{2}\right)}{2^j} = i$$

Ainsi si $M_i = i$, on a $F_i + \frac{S_i}{2} = F_i = 2^i$, et $F_j + \frac{S_j}{2} = 0$, pour tout $j \leq i-1$.

Les deux calculs V_{1n}, V_{2n}, donnent donc des variances nulles lorsqu'une génération donnée est complète, ce qui implique que les générations précédentes soient nécessairement complètes.

Si la génération 1 est incomplète auquel cas *Ego* est semi-fondateur, alors on a, en s'arrêtant en génération 1 :

$$M_{11} = M_{21} = M_1 = \frac{1}{2} \neq 1,\ V_{11} = V_{21} = \frac{1}{4}$$

Les variances V_1 et V_2 sont égales à ce stade mais comme $M_1 \neq 1$, elles ne sont en général plus égales pour $n \geq 1$.

4. – Remarque générale importante

Considérons un individu fondateur en génération g. La contribution de cet individu à la profondeur généalogique est de $\frac{1}{2^g}$.

Étant donné un semi-fondateur à la génération g, celui-ci a des ascendants dont certains peuvent être semi-fondateurs ou fondateurs ou identifiés. Montrons que la contribution totale de ce semi-fondateur et de ses ascendants est égale à $\frac{1}{2^g}$.

Deux cas sont possibles :

— aucun ascendant de ce semi-fondateur n'est parfaitement identifié. Prenons le cas le plus simple de la figure 7b, celui d'un seul ascendant

fondateur. Dans ce cas, la contribution du semi-fondateur est égale à $\frac{1}{2} \times \frac{1}{2^g}$, celle de son unique ascendant est égale à $\frac{1}{2^{g+1}}$ et la contribution totale de ces deux éléments vaut $\frac{1}{2^g}$.

Par récurrence, quand il n'y a qu'une série de semi-fondateurs entre la génération g et la génération $g + i$ (dernier ascendant fondateur, figure 7d), on démontre facilement que la somme des contributions du semi-fondateur et de ses ascendants est encore égale à $\frac{1}{2^g}$.

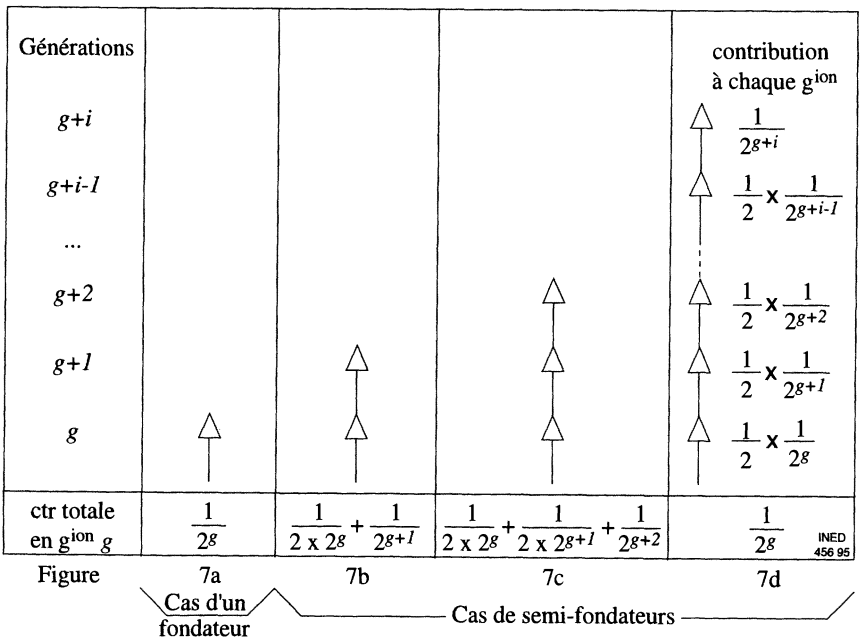


Figure 7.

Qu'on ait un fondateur ou un semi-fondateur en génération g, on obtient la même contribution totale de $\frac{1}{2^g}$. Le schéma du semi-fondateur généralisé sur *g+i* générations donne une contribution totale de :

$$\frac{1}{2^{g+1}}\left(1+\frac{1}{2}+\frac{1}{4}+\ldots+\frac{1}{2^{i-1}}+\frac{1}{2^{i-1}}\right)=\frac{1}{2^{g+1}}\left(\frac{1-1/2^i}{1-1/2}+\frac{1}{2^{i-1}}\right)=\frac{1}{2^{g+1}}\left(2\,(1-1/2^i)+\frac{1}{2^{i-1}}\right)=\frac{1}{2^g}$$

— au moins un ascendant de ce semi-fondateur est totalement identifié. Prenons le cas le plus simple de la figure 8c. Dans ce cas, la contribution du semi-fondateur est égale à $\frac{1}{2} \times \frac{1}{2^g}$, la contribution des deux fondateurs en génération $g + 2$ est égale à $\frac{2}{2^{g+2}}$ tandis que la contribution de l'ascendant intermédiaire est nulle.

La contribution totale de ce semi-fondateur et de ses ascendants reste bien égale à $\frac{1}{2^g}$. Par récurrence, on démontre qu'il en est toujours ainsi.

Nous venons de voir qu'un fondateur ou un semi-fondateur avec ses ascendants, en génération g, ont une contribution totale égale à $\frac{1}{2^g}$. Ce ré-

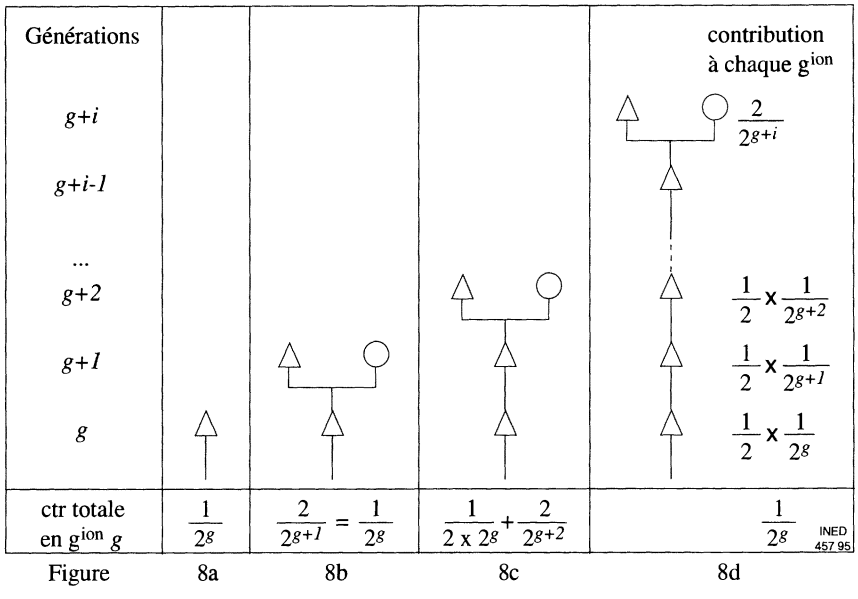


Figure 8.

Qu'on ait un fondateur, un semi-fondateur ou un individu identifié en génération g, on obtient la même contribution totale de $\frac{1}{2^g}$ quel que soit le schéma d'ascendance de l'individu situé en génération g. Le schéma 8d fournit la même contribution que le schéma 7d : en effet, les deux derniers termes de la somme dans le cas 7d : $\frac{1}{2} \times \frac{1}{2^{g+i-1}} + \frac{1}{2^{g+i}}$ sont remplacés par un seul terme dans le cas 8d : $\frac{2}{2^{g+i}}$ qui leur est équivalent.

sultat reste également valable pour un individu identifié en génération g (dont la contribution personnelle est nulle) comme on peut le voir par exemple dans la figure 8b. Ainsi à tout niveau d'une généalogie, en une génération donnée g, chaque individu de cette génération et l'ensemble de ses ascendants ont une contribution totale égale à $\frac{1}{2^g}$.

BIBLIOGRAPHIE

DE BRAEKELEER M. et BELLIS G., (1994), « Généalogies et reconstitutions de familles en génétique humaine », *Dossiers et Recherches,* n° 43, INED, Paris.

JETTÉ R., (1991), *Traité de Généalogie*, Presses de l'Université de Montréal, Québec.

KOULADJIAN K., (1986), « Une mesure d'entropie généalogique », Chicoutimi, SOREP, *Document* III-C-43.

LÉTOURNEAU E. et MAYER F.-M., (1988), « Un modèle d'analyse de généalogie ascendante », *Cahiers Québécois de Démographie*, 17, 2, 213-231.

CAZES (Marie-Hélène), CAZES (Pierre). – **Comment mesurer la profondeur généalogique d'une ascendance ?**

Pour de nombreux travaux de génétique ou parfois de démographie historique, on utilise la reconstitution d'ascendances à partir d'un individu appelé le *probant*. L'information apportée par ces généalogies varie d'un individu à l'autre, en fonction du nombre d'ancêtres qu'on est parvenu à identifier. Les mesures classiques du coefficient de parenté ou de consanguinité vont directement dépendre de cette reconstitution. C'est pourquoi, il est nécessaire de quantifier l'information contenue dans ces ascendances, si l'on veut pouvoir procéder à des comparaisons entre individus. Cet article analyse un indice, supposé exprimer la profondeur généalogique moyenne d'une ascendance. Nous montrons des exemples où l'utilisation de cet indice fournit des résultats incohérents et proposons deux autres indices, avec leurs variances associées, qui suppriment ces incohérences et peuvent s'interpréter vraiment en terme de *profondeur généalogique*.

CAZES (Marie-Hélène), CAZES (Pierre). – **Assessing the genealogical depth of an ancestry**

Many studies in genetics and historical demography rely on family reconstitution, beginning with one individual and his or her ancestry. Information from such genealogies varies depending on the number of the individual's ancestors who can be identified. Traditional assessments of kinship or consanguinity ratios are often based on such reconstitutions. This reinforces the need for information from these genealogies to be quantified in order that individual situations may be compared. In this paper it is shown that an index which is supposed to provide information about the average length of an ancestry may lead to inconsistent results in some cases, and two new indices and their associated variances are introduced which eliminate inconsistencies and can be used to measure average genealogical lengths.

CAZES (Marie-Hélène), CAZES (Pierre). – **Cómo se puede medir la profundidad genealógica de una descendencia ?**

En numerosos trabajos de genética y de demografía historica se utiliza la reconstitución de ascendencias a partir de un individuo de referencia. La información que aportan estas genealogías varía de un individuo a otro, en función del número de antepasados que se identifican. Dado que las medidas clásicas como el coeficiente de parentesco o de consanguinidad dependen directamente de esta reconstitución, es necesario cuentificar la información contenida en ellas para efectuar comparaciones entre individuos. Este artículo analiza un índice que expresa la profundidad genealógica media de una ascendencia. También incluye ejemplos en los que la utilización del índice da resultados incoherentes, y se proponen dos índices alternativos (con sus varianzas) que suprimen estas incoherencias y que se pueden interpretar con mayor exactitud en términos de profundidad genealógica.