

Corrélations entre sujets apparentés et analyse statistique de la variance chez les jumeaux

par

E. DEFRISE-GUSSENHOVEN

INTRODUCTION

Quelques collègues nous ont demandé l'explication des bases statistiques et génétiques d'un article et d'un programme de Christian et al. (1974) consacrés au choix d'une estimation de la variance génétique en biométrie. Nous remercions tout particulièrement ces auteurs de nous avoir communiqué ce programme excellent : il a servi dans plusieurs études biométriques consacrées à des jumeaux belges.

Nous donnons d'abord une analyse intuitive des causes possibles de corrélations positives entre sujets apparentés : doublets père/fils ou jumeaux d'une paire, soit monozygotiques (MZ), soit dizygotiques (DZ).

La deuxième partie explique les bases statistiques du programme, la troisième est consacrée à la décomposition de la variance génétique et la quatrième à la synthèse des analyses statistique et génétique. Nous terminons par quelques remarques sur la validité du modèle.

1. CORRELATIONS ENTRE SUJETS APPARENTES.

1.1. Corrélations entre les tailles père / fils (fig. 1, 2, 3, 4)

Pour séparer les causes des corrélations positives entre les mesures d'un père et celles de son fils, nous faisons appel à une expérience fictive faite sur des souris. Imaginons une population de souris dans laquelle nous distinguons des doublets composés d'un mâle et d'un de ses descendants mâles de première génération et que nous désignerons dans la suite par « père et fils », père et fils d'un même doublet *restant réunis* dans une même niche écologique. Une telle étude serait réalisable, mais si à notre connaissance elle n'a pas été faite, c'est que les généticiens qui utilisent des souris ne font pas en général, de biométrie. Nous représenterons les mesures, par exemple les statures de chaque doublet, par un point (fig. 1, 2, 3, 4), l'abscisse étant la taille du père, l'ordonnée celle du fils. Nous allons varier les conditions de l'expérience fictive et examiner leurs effets sur l'allure des nuages de points représentant les doublets.

1.1.A. *Lignée pure de souris, même milieu pour tous les individus donc pour l'ensemble des doublets (fig. 1)*

S'ils appartiennent à une lignée pure de souris, tous les mâles sont des homozygotes identiques et si, en plus, ils vivent tous dans des conditions égales, il n'y a aucune raison pour qu'un fils ressemble plus à son père qu'à n'importe quel autre mâle; on peut s'attendre, ainsi que le suggère la figure 1, à trouver une corrélation nulle entre les tailles père/fils, puisque, dans cette expérience, tous les individus mâles partagent les mêmes gènes et le même milieu. La droite de régression taille fils / taille père est horizontale.

Cela étant, on peut se demander d'où provient alors la variabilité entre les individus; on peut l'expliquer : a) par les erreurs de mesure; b) par l'inévitable inégalité des conditions de vie in utero et durant l'allaitement. Remarquons cependant que ces causes de variation sont aléatoires et qu'elles affectent indépendamment les mesures du père et celles du fils, n'entraînant aucune corrélation non nulle.

1.1.B. *Population génétiquement variée de souris, même milieu pour tous les individus (fig. 2)*

Lorsque la population est génétiquement variée et que le caractère mesuré est sous la dépendance des gènes, étant donné qu'un fils reçoit la moitié des gènes de son père, il ressemblera plus à celui-ci qu'aux autres individus. Si le père est grand, le fils aura également tendance à dépasser la moyenne, ainsi que le suggère le figure 2, avec une droite de régression taille fils / taille père à pente positive, le coefficient de corrélation étant lui aussi positif. C'est la ressemblance génétique du père et du fils qui en est la cause et non pas une quelconque variation du milieu entre les doublets, car ici nous supposons le milieu égal pour tous. Rappelons en outre que si la population est panmictique, le coefficient de corrélation r a 0,5 comme espérance mathématique. Remarquons que le nuage de points sur la figure 2 est plus étalé que celui de la figure 1 à cause de la diversité génétique de la population.

1.1.C. *Lignée pure de souris, milieux contrastés pour les doublets, père et fils restant réunis (fig. 3)*

On est de nouveau dans le cas d'une population d'homozygotes tous identiques, mais on place les doublets père/fils dans des milieux contrastés. L'étalement plus grand du nuage de points sur la figure 3 que sur la figure 1 est dû aux variations du milieu. Cependant, puisqu'un père et son fils partagent les mêmes avantages (ou désavantages) des conditions de vie, le fils aura tendance à ressembler au père, r et le coefficient de régression taille fils/taille père sont positifs. Cette ressemblance est due aux variations des conditions mésologiques entre les doublets et non pas aux variations du génome des individus.

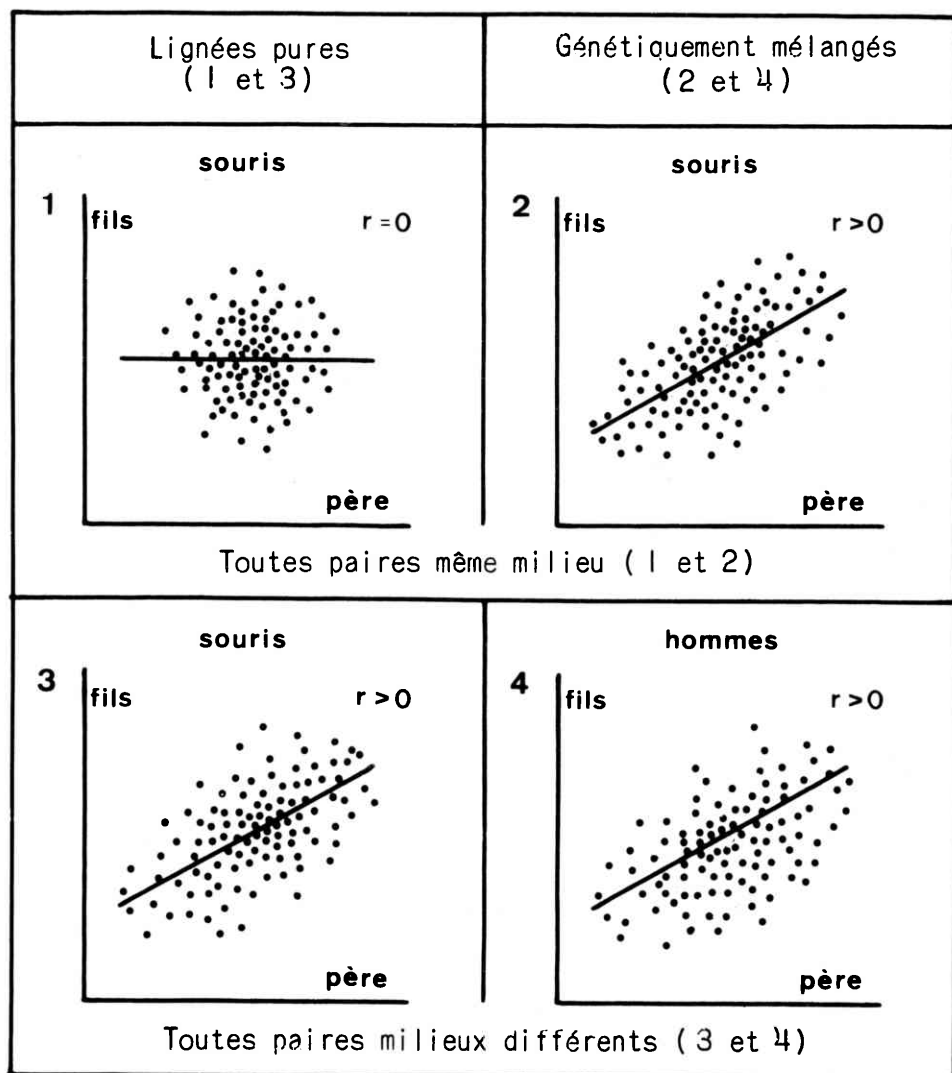


Fig. 1, 2, 3, 4 : Corrélations statures père/fils

Père et fils d'un même doublet restent toujours dans un même milieu.

1.1.D. *Population humaine aux variations mésologiques et génétiques (fig. 4)*

Chez l'homme il n'y a pas de lignées pures. Galton (1885) construisit un graphique, esquissé et simplifié sur la figure 4, chaque point représentant la taille d'un étudiant (ordonnée) et celle de son père (abscisse). La valeur de r est naturellement positive, mais contrairement aux situations représentées sur les figures 2 et 3, on ne peut décider si la ressemblance père/fils est due à la variation des gènes ou à celle du milieu. Ce n'était évidemment pas un hasard si Galton avait pris soin de choisir son échantillon de doublets père/fils dans un milieu assez homogène du point de vue culturel (Université de Cambridge), de sorte qu'on peut penser que la valeur $r = 0,5$ indique plutôt une ressemblance due au génome qu'au milieu. Rappelons que c'est Galton qui a introduit le terme « régression vers la médiocrité » de la taille du fils par rapport à la taille du père.

Conclusions du 1.1.

a) Un coefficient de corrélation positif révèle ou bien des influences mésologiques variées mais égales pour le père et le fils, ou bien des facteurs génétiques égaux pour père et fils ou encore la présence simultanée des deux sortes d'effets, et, sans autre information sur les doublets père/fils et la population à laquelle ils appartiennent, on ne peut décider quelles sont les causes déterminantes de la ressemblance. C'est le cas pour la figure 4.

b) Comme le montre la figure 1, sans variation génétique et mésologique, il n'y a pas de corrélation positive; la figure 1, avec $r = 0$, ne signifie donc pas que la taille des souris ne dépend pas des gènes; elle signifie simplement qu'il n'y a pas de variation génétique dans une lignée pure. Si la population de souris de la figure 2 était génétiquement très polymorphe, r augmenterait à cause de cette variation génétique; de même, si dans la figure 3, les conditions mésologiques dans lesquelles sont plongés les doublets père/fils, devenaient de plus en plus extrêmes, r augmenterait sans cesse.

En général, dans une population panmictique, la ressemblance d'un fils avec son père est conditionnée à la fois par la variation du milieu et par celle du patrimoine génétique. C'est donc en définitive la connaissance de la variance des facteurs génétiques et mésologiques qui est essentielle dans l'analyse des caractères mesurables.

1.2. *Corrélations entre les tailles des jumeaux d'une paire (fig. 5, 6, 7)*

Nous venons de voir que chez l'homme, l'analyse des caractères biométriques à partir de doublets un parent/un enfant présente des difficultés parce qu'il est impossible de séparer les facteurs génétiques et mésologiques qui entrent en jeu. De ce point de vue, les jumeaux présentent de grands avantages. Nous supposons ici que les jumeaux d'une paire sont de même sexe et qu'ils ont été élevés ensemble. Entre

les paires — de même qu'entre les doublets père / fils — existent des variations génétiques et mésologiques, mais chaque paire monozygotique se compose de deux sujets génétiquement identiques, tandis qu'une paire dizygotique comprend deux sujets qui diffèrent génétiquement comme deux frères ou deux sœurs. C'est le contraste entre les deux types de jumeaux qui est exploité.

La figure 5 représente une paire de jumeaux de tailles respectives a et b . Les points 1 et 2 ont comme coordonnées respectivement a, b et b, a . Cette paire a donc comme double image les points 1 et 2, symétriques par rapport à la bissectrice de l'angle formé par les axes du graphique; la distance entre ces points, divisée par $\sqrt{2}$, vaut $|a-b|$, de sorte que plus les jumeaux de la paire diffèrent pour la taille, plus les points 1 et 2 seront éloignés de la bissectrice. Le point d'intersection de celle-ci et du segment qui joint les points 1 et 2 a des coordonnées toutes deux égales à $(a + b) / 2$, la taille moyenne de la paire. D'autre part, si deux jumeaux ont des tailles égales, $a = b$ et les points 1 et 2 sont confondus et situés sur la bissectrice. Cette représentation symétrique supprime l'effet du numéro d'ordre des jumeaux d'une paire. Dans la 2^e partie nous verrons d'ailleurs que dans les calculs, seul le carré de l'écart entre les mesures d'une paire intervient.

1.2.A. *Corrélations des jumeaux monozygotiques (MZ), les jumeaux de chaque paire étant élevés ensemble (fig. 6)*

Les $2n$ points de la figure 6 représentent les tailles de n paires de jumeaux MZ; ils sont peu distants de la bissectrice à cause de la ressemblance des tailles des jumeaux de chaque paire MZ. Si la différence $a-b$ est souvent non nulle, c'est non seulement à cause des erreurs de mesure mais parce que, bien que de même génome et élevés ensemble, les jumeaux d'une paire MZ ne vivent pas exactement dans le même milieu, ni avant, ni après la naissance.

D'autre part, le nuage de points est étiré le long de la bissectrice : en effet, les familles dont sont issues les paires de jumeaux diffèrent par leurs gènes et par leur milieu, et, plus la population d'origine des jumeaux est variée, plus le nuage sera allongé. La corrélation — mesurée pour les jumeaux par R , le coefficient de corrélation intra-classe (formule 20) — croît précisément lorsque les différences intrapaires diminuent en même temps qu'augmentent les différences interpaires, c'est-à-dire les moyennes $(a + b) / 2$ des paires (formule 4).

1.2.B. *Corrélations des jumeaux dizygotiques (DZ), les jumeaux de chaque paire étant élevés ensemble (fig. 7)*

Les figures 6 et 7 sont analogues, mais le nuage de points des paires DZ révèle une corrélation moindre, non parce que les moyennes des paires sont moins dispersées (nous supposons en effet que la population d'origine est la même pour les deux types de jumeaux), mais parce que les différences de taille intrapaires sont plus grandes; en

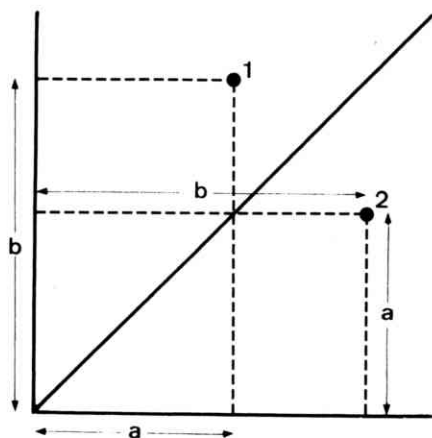


Fig. 5 : Double image d'une paire de jumeaux de statures a et b .
Si $(a - b)^2$ est petit, les jumeaux se ressemblent pour la taille.

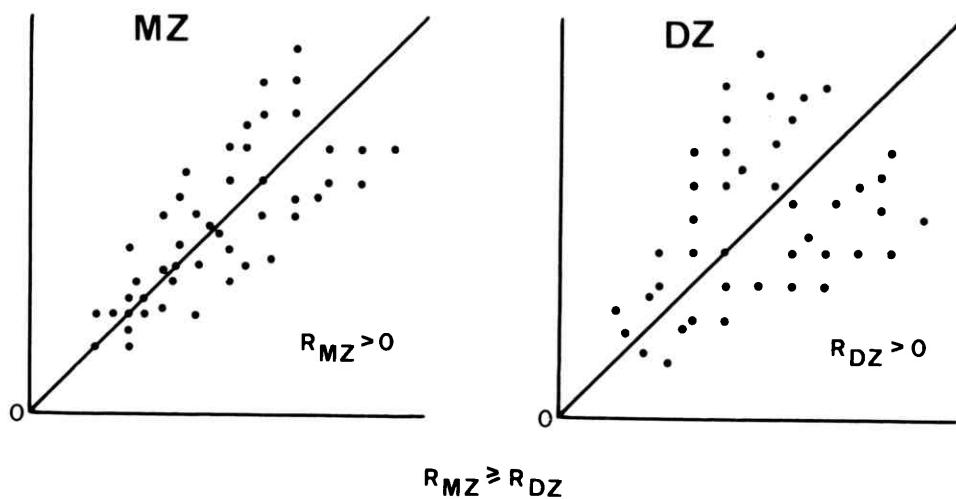


Fig. 6 , 7 : Statures de jumeaux; les paires vivent dans des milieux différents;
les jumeaux d'une paire vivent ensemble.
Les jumeaux MZ sont génétiquement identiques.
Les jumeaux DZ sont génétiquement non identiques.

effet, les jumeaux DZ n'ont qu'une partie de leurs gènes en commun; on sait que pour eux l'espérance mathématique de la fraction de gènes identiques par descendance est égale à $1/2$.

Conclusions du 1.2.

a) Ainsi donc, si une population a fourni deux échantillons de paires de jumeaux de même sexe — par exemple masculins — les uns monozygotiques, les autres dizygotiques, la comparaison des coefficients de corrélation intraclasse RMZ et RDZ (formules 69 et 70) pourra révéler si le caractère biométrique à l'étude est au moins partiellement déterminé par le génome. En effet, si RMZ est significativement supérieur à RDZ et si les jumeaux de chaque paire vivent ensemble, l'écart RMZ-RDZ est essentiellement dû à la différence génétique des jumeaux dizygotiques et non pas aux effets de facteurs mésologiques.

b) Si, dans la population d'origine, la variation des facteurs génétiques et mésologiques augmente, les moyennes $(a + b) / 2$ des paires de jumeaux vont diverger de plus en plus, les nuages de points des figures 6 et 7 s'étireront davantage le long de la bissectrice et les coefficients RMZ et RDZ augmenteront.

c) Lorsque RMZ est peu différent de RDZ, cela n'est pas une preuve de l'absence de facteurs génétiques dans l'élaboration du caractère biométrique étudié; ce fait se produirait par exemple pour une population d'origine très homogène du point de vue génétique.

d) C'est à dessein que nous avons pris comme exemple la taille, caractère pour lequel il n'existe pas d'interaction entre jumeaux d'une paire. Cependant de telles interactions s'observent pour certains traits psychologiques et dans ce cas elles doivent intervenir dans la discussion sur les valeurs respectives de RMZ et RDZ.

2. BASE STATISTIQUE DU PROGRAMME DE CHRISTIAN et al. (1974).

Dans cette deuxième partie, nous nous reportons souvent aux notations et aux données chiffrées du tableau I relatif à la stature de 57 paires de MZ et 39 paires de DZ, toutes masculines. La description de ces échantillons a été donnée par Defrise et al. (1982). Le tableau I est obtenu par le programme que Christian et al. ont bien voulu nous communiquer. Quand il y a lieu, nous indiquons, à côté des formules et des résultats, le numéro correspondant de la ligne du tableau I : ainsi, t.I, 8 signifie «la ligne 8 du tableau I».

2.1. Variances et corrélations, les groupes MZ et DZ étant séparés

L'importance de la comparaison des variances intrapaires et interpaires a été esquissée intuitivement dans les paragraphes 1.2.A et 1.2.B. Ci-dessous suit l'explication statistique.

2.1.A. Notations pour les valeurs observées et calculées

Dans un échantillon de n paires de jumeaux, soient

x_{ij} la mesure du jumeau i ($i = 1$ ou 2) de la $j^{\text{ème}}$ paire, avec $j = 1, 2, \dots, n$,

$$x_{.j} = (x_{1j} + x_{2j}) / 2 = \sum_{i=1}^2 x_{ij} / 2 \quad (1)$$

la moyenne de la $j^{\text{ème}}$ paire,

$$x_{..} = \sum_j \sum_i x_{ij} / 2n = \sum_j (x_{1j} + x_{2j}) / 2n \quad (2)$$

la moyenne des n paires,

$$Q_{\text{tot}} = \sum_j \sum_i (x_{ij} - x_{..})^2 \quad (3)$$

la somme des carrés des écarts à la moyenne, désignée par «total sum of squares» (t.I,8). Cette somme a $2n - 1$ degrés de liberté désignés par DF pour «degrees of freedom» sur le tableau I. Soient encore

$$Q_A = \sum_j \sum_i (x_{.j} - x_{..})^2 = 2 \sum_j (x_{.j} - x_{..})^2 = 2 \sum_j \left(\frac{x_{1j} + x_{2j}}{2} - x_{..} \right)^2 \quad (4)$$

la somme des carrés des écarts interpaires avec $n-1$ DF et «A» pour «among»

$$Q_W = \sum_j \sum_i (x_{ij} - x_{.j})^2 = \sum_j (x_{1j} - x_{2j})^2 / 2 \quad (5)$$

la somme des carrés des écarts intrapaires avec n DF et «W» pour «within»

On a, identiquement,

$$Q_{\text{tot}} = Q_A + Q_W; \quad (6)$$

c'est la relation classique de l'analyse de la variance.

On note encore

$$\text{AMS} = Q_A / (n - 1) \text{ les carrés moyens interpaires,} \quad (7)$$

(«among mean squares», t.I.,6) et

$$\text{WMS} = Q_W / n \text{ les carrés moyens intrapaires,} \quad (8)$$

(«within mean squares», t.1,7).

Par exemple, pour les figures 6 et 7, Q_A et AMS seront d'autant plus grands que le nuage de points sera plus étiré le long de la bissectrice, tandis que Q_W et WMS diminueront dans la mesure où les points se rapprocheront de la bissectrice, c'est-à-dire dans la mesure où les jumeaux se ressembleront davantage pour le caractère étudié.

2.1.B. Notations pour une analyse de la variance

R.A. Fisher (1954) introduisit une analyse destinée à comparer les variances d'enfants, intra-et interfamilles. Nous croyons utile d'exposer le modèle théorique qui sous-tend cette analyse car il permet un examen simple et intuitif des figures 6 et 7.

La valeur observée du jumeau i de la paire j se décompose ainsi :

$$x_{ij} = \mu + f_j + z_{ij} \quad (9)$$

où μ est la moyenne théorique des $2n$ jumeaux, f_j la contribution familiale aux mesures de la $j^{\text{ème}}$ paire et z_{ij} l'écart du jumeau i à $\mu + f_j$, la valeur attendue de x_{ij} pour une famille donnée,

$$\text{avec } \text{cov}(z_{1j}, z_{2j}) = 0 \quad (10)$$

Les distributions de f_j et z_{ij} sont supposées normales avec des variances égales à :

$$V(f_j) = \sigma_f^2, \text{ la variance interfamille et} \quad (11)$$

$$V(z_{ij}) = \sigma^2, \text{ la variance intrapaires} \quad (12)$$

De plus, $\text{cov}(f_j, z_{ij}) = 0$, de sorte que la variance de x_{ij} est

$$V(x_{ij}) = \sigma_f^2 + \sigma^2 = \sigma_{\text{tot}}^2 \quad (13)$$

On trouve aisément, en combinant les formules (4) à (13) :

$$E(\text{AMS}) = 2 \sigma_f^2 + \sigma^2 = \text{variance interpaires et} \quad (14)$$

$$E(\text{WMS}) = \sigma_f^2 = \text{variance intrapaires} \quad (15)$$

On observe encore que

$$E(\text{AMS}) = \sigma_f^2 + (\sigma_f^2 + \sigma^2) = \sigma_f^2 + \sigma_{\text{tot}}^2, \text{ de sorte que}$$

la variance interpaires = la variance interfamilles + la variance totale, (cf formules 59 et 66)

2.1.C. Analyse de la variance de Fisher

D'après les conditions (10) à (13), le rapport

$$F = \frac{\text{AMS}}{\text{WMS}} = \frac{2 \sum (x_{.j} - x_{..})^2 / (n - 1)}{\sum (x_{1j} - x_{2j})^2 / n} \quad (16)$$

a une distribution de Fisher-Snedecor, avec $n-1$ et n degrés de liberté et il peut être testé, l'hypothèse nulle étant $\sigma_f^2 = 0$.

Les espérances mathématiques des termes de F sont respectivement $2\sigma_f^2 + \sigma^2$ et σ^2 , (14) et (15). Ce sont donc bien les grandeurs relatives de σ_f^2 et σ^2 qui entrent en jeu : pour σ^2 petit par rapport à σ_f^2 , F sera grand et si cette valeur de F est significative, on rejette l'hypothèse nulle et on admet que $\sigma_f^2 \neq 0$ et aussi que certains facteurs familiaux agissent sur le caractère analysé. Si, au contraire, la variance interfamille σ_f^2 est minime (ce serait le cas, sans doute rare, d'une population d'origine à faible variation génétique et mésologique) alors F sera proche de 1 et non significatif. Dans un tel cas, les facteurs familiaux éventuels ne peuvent être mis en évidence. Cependant, une valeur de F proche de 1 ne prouve pas nécessairement l'absence de tels facteurs : c'est seulement leur peu de variation dans la population (avec σ_f^2 très petit) qui entraîne la non signification de F .

En utilisant les carrés moyens pour la taille (t.I, 6 et 7), on obtient, d'après (16), $F = 26,10$ pour les 57 paires MZ et $F = 6,74$ pour les 39 paires DZ; ce dernier résultat est noté sur t.I, 19. Ces deux valeurs de F sont hautement significatives avec $P = 0,000000$ et révèlent la présence, pour les deux types de jumeaux, de facteurs mésologiques et / ou génétiques qui différencient les familles et causent la ressemblance des jumeaux de chaque paire.

2.1.D. Estimateurs sans biais des valeurs théoriques du modèle de la formule (9)

Les valeurs théoriques du modèle (9) peuvent être estimées sans biais comme suit :

$$\hat{\mu} = x_{..}; \quad \hat{\sigma}_f^2 = (\text{AMS} - \text{WMS}) / 2; \quad \hat{\sigma}^2 = \text{WMS}; \quad (17)$$

$$\hat{\sigma}_{\text{tot}}^2 = (\text{AMS} + \text{WMS}) / 2 \quad (18)$$

Pour la taille, on trouve les estimations

$$\hat{\mu}_{\text{MZ}} = 175,585090 \quad \text{et} \quad \hat{\mu}_{\text{DZ}} = 177,535900 \quad (\text{t.I, 3 et 22})$$

$$\hat{\sigma}_{\text{fMZ}}^2 = 38,628325 \quad \text{et} \quad \hat{\sigma}_{\text{fDZ}}^2 = 28,805415 \quad (\text{pas sur t.I})$$

$$\hat{\sigma}_{MZ}^2 = 3,077630 \quad \text{et} \quad \hat{\sigma}_{DZ}^2 = 10,031790 \quad (\text{t.I,7})$$

$$\hat{\sigma}_{\text{tot}MZ}^2 = 41,705955 \quad \text{et} \quad \hat{\sigma}_{\text{tot}DZ}^2 = 38,837205 \quad (\text{pas sur t.I})$$

Ces deux dernières variances, ayant comme espérances mathématiques d'après (13) $\sigma_{fMZ}^2 + \sigma_{MZ}^2$ et $\sigma_{fDZ}^2 + \sigma_{DZ}^2$, sont donc des estimations sans biais de la variance totale de x_{ij} dans chacune des populations d'origine. Le tableau I donne, au contraire, des estimations biaisées de σ_{tot}^2 , du reste peu différentes des précédentes, notamment :

$$Q_{\text{tot}MZ} / (2n_{MZ} - 1) = 41,364110 \quad \text{et} \quad (\text{t.I,23})$$

$$Q_{\text{tot}DZ} / (2n_{DZ} - 1) = 38,463110 \quad (\text{t.I,23})$$

Dans les analyses génétiques, nous préférons les estimations non biaisées ; elles se calculent aisément par la formule (18), grâce aux carrés moyens, qui sont pour la taille :

$$AMZ = 80,334280, \quad ADZ = 67,642620 \quad (\text{t.I,6})$$

$$WMZ = 3,077630, \quad WDZ = 10,031790 \quad (\text{t.I,7})$$

2.1.E. Coefficient de corrélation $(a + b) / |a - b|$

Le coefficient de corrélation Bravais-Pearson entre la somme et la valeur absolue de la différence d'une paire vaut 0,771560 pour les paires MZ et 0,816220 pour les paires DZ (t.I,25). Ces deux coefficients sont hautement significatifs avec $P = 0,000000$ dans les deux cas : plus les tailles sont élevées, plus l'écart intrapaires a tendance à augmenter. Ces coefficients de corrélation ne sont pas analysés du point de vue génétique dans Christian et al. (1974). Pour nos jumeaux, nous avons obtenu des corrélations hautement significatives pour la plupart des mensurations.

Le programme donne aussi (t.I,26) la moyenne des valeurs absolues des différences $a-b$, avec 1,928070 pour les MZ et une valeur plus élevée pour les DZ, 3,682050, comme on pouvait s'y attendre (cf fig. 6 et 7). Ces moyennes de $|a-b|$ n'interviennent pas non plus dans l'analyse génétique.

2.1.F. Coefficients de corrélation intraclasse

Pour les coefficients de corrélation père / fils, on utilise le coefficient de Bravais-Pearson

$$r = \sum (x - \bar{x})(y - \bar{y}) / n s_x s_y, \quad (19)$$

mais pour les jumeaux (fig. 6 et 7), on calcule un coefficient de corrélation spécial, dit «intraclasse», introduit par Fisher en 1925, et dans lequel les différences entre jumeaux n'interviennent qu'en valeur absolue, avec

$$R = (AMS - WMS) / (AMS + WMS) \quad (20)$$

En remplaçant les termes de R par leur espérance mathématique, on obtient, en combinant (13), (14) et (15)

$$2\sigma_f^2 / 2(\sigma_f^2 + \sigma^2) = \sigma_f^2 / \sigma_{\text{tot}}^2, \quad (21)$$

de sorte que R est proche de zéro lorsque $\sigma_f^2 = 0$, c'est-à-dire en l'absence de variation interfamille; R tend vers 1 pour $\sigma^2 = 0$, lorsque les jumeaux de chaque paire ont des mesures égales. Sur les figures 6 et 7, tous les points seraient alors sur la bissectrice.

Notons que comme r, le coefficient R varie entre -1 et +1.

Pour la taille, on trouve (t.I, 24 et 28)

$$R_{\text{MZ}} = 0,926206 \quad \text{et} \quad R_{\text{DZ}} = 0,741696,$$

deux valeurs hautement significatives avec $P = 0,000000$ (t.I,24).

2.2. Tests pour comparer les deux types de jumeaux

Dans le paragraphe 2.1, chaque groupe de jumeaux, MZ ou DZ, était considéré à part. Dans ce nouveau paragraphe, les résultats des deux types de jumeaux sont comparés, but final des analyses gémellaires.

2.2.A. Comparaison des coefficients de corrélation intraclasse R_{MZ} et R_{DZ}

On applique la transformation de Fisher

$$z = \frac{1}{2} \log_e \frac{1 + R}{1 - R} \quad (22)$$

On trouve

$$z_{\text{MZ}} = 1,631018 \quad \text{et} \quad z_{\text{DZ}} = 0,954240 \quad (\text{t.I,29}).$$

Fisher a montré que z_{MZ} et z_{DZ} ont des distributions proches de la normale, avec des moyennes que nous désignons par $\mathcal{G}_{\text{MZ}}$ et $\mathcal{G}_{\text{DZ}}$ et des variances $1/(n_{\text{MZ}} - 3)$ et $1/(n_{\text{DZ}} - 3)$. La variable $z_{\text{MZ}} - z_{\text{DZ}}$ a donc une distribution proche de la normale avec $\mathcal{G}_{\text{MZ}} - \mathcal{G}_{\text{DZ}}$ comme moyenne et $[1/(n_{\text{MZ}} - 3) + 1/(n_{\text{DZ}} - 3)]$ comme variance.

Pour tester l'hypothèse nulle $\mathcal{G}_{\text{MZ}} \geq \mathcal{G}_{\text{DZ}}$ par un test unilatéral, on calcule $z_{\text{MZ}} - z_{\text{DZ}} = 0,676779$ (t.I,29) et

$$[1 / (57 - 3) + 1 / (39 - 3)]^{\frac{1}{2}} = 0,2151658 \quad (\text{pas sur t.I})$$

Le quotient des deux résultats est 3,145383 (t.I,29)

Les tables de la distribution normale $N(0,1)$ indiquent que la probabilité que cette valeur soit dépassée dans le cas de l'hypothèse nulle n'est que de $P = 0,000829$ (t.I,29).

Le coefficient de corrélation des MZ est donc significativement plus grand que celui des DZ, conformément à ce que suggèrent les figures 6 et 7.

2.2.B. Coefficients d'héritabilité basés sur R_{MZ} et R_{DZ}

a) coefficient de Holzinger

$$h_H^2 = (R_{MZ} - R_{DZ}) / (1 - R_{DZ}) = 0,714315 \quad (t.I,35)$$

Le test de nullité est le même que celui de 2.2.A et il donne également $P = 0,000829$ (t.I,35). Si $R_{MZ} = 1$, alors $h_H^2 = 1$. Si, au contraire, $R_{MZ} = R_{DZ}$, $h_H^2 = 0$ et les jumeaux des paires MZ ne se ressemblent pas plus que ceux des paires DZ.

b) coefficient dit « intraclasse »

$$h_I^2 = 2(R_{MZ} - R_{DZ}) = 0,369020 \quad (t.I,36)$$

Cette valeur est également significative avec $P = 0,000829$ (t.I,36).

2.2.C. Test t' (T-Prime, t.I,2) pour comparer les moyennes des jumeaux MZ et DZ

On a tiré du programme de Christian et al.

$$\hat{\mu}_{MZ} = x_{..MZ} = 175,585090 \text{ et } \hat{\mu}_{DZ} = x_{..DZ} = 177,535900 \quad (t.I, 3 \text{ et } 22)$$

Il s'agit de tester

$$x_{..MZ} - x_{..DZ} = -1,950810 \quad (t.I,3), \text{ l'hypothèse nulle étant } \mu_{MZ} = \mu_{DZ}.$$

On utilise le test t' dont la distribution approche celle du t de Student, mais t' présente la particularité de dépendre d'une variance estimée qui est la somme de deux carrés moyens. Nous faisons suivre ici l'explication des calculs de t' et du nombre de degrés de liberté. La démonstration se trouve dans des traités de statistique mathématique comme celui de Scheffé (1959).

D'après (9), $x_{ij} = \mu + f_j + z_{ij}$, d'où $x_{..} = \mu + f_{.} + z_{..}$; on en tire, d'après (12) et (13)

$$V(x_{..}) = \sigma_f^2 / n + \sigma_z^2 / 2n = (2\sigma_f^2 + \sigma_z^2) / 2n \quad (23)$$

D'après (14), $2\sigma_f^2 + \sigma^2 = E(AMS)$, d'où l'on tire

$$V(x_{..}) = E(AMS) / 2n \quad (24)$$

dont l'estimateur est

$$\hat{V}(x_{..}) = AMS / 2n \text{ avec } n - 1 \text{ DF} \quad (25)$$

Si l'hypothèse nulle est vraie, alors $x_{..MZ} - x_{..DZ}$ est distribué normalement avec une moyenne nulle et une variance égale à

$$AMZ / 2n_{MZ} + ADZ / 2n_{DZ} \quad (26)$$

L'écart-type de la différence des moyennes est donc

$$\left[\frac{80,334280}{2 \times 57} + \frac{67,642620}{2 \times 39} \right]^{\frac{1}{2}} = 1,253754 \quad \text{et} \quad (\text{pas sur t.1})$$

$$t' = -1,950810 : 1,253754 = -1,555975 \quad (\text{t.1,3}).$$

Le nombre approché de degrés de liberté se calcule comme suit. On pose

$U = AMZ / 2n_{MZ}$ et $W = ADZ / 2n_{DZ}$. Ces deux sommes de carrés ont respectivement $f_U = n_{MZ} - 1$ et $f_W = n_{DZ} - 1$ degrés de liberté d'après (4), (5), (7) et (8). On démontre que le nombre approché de degrés de liberté de $U + W$ est

$$(t.1,3) \quad (U + W)^2 / [U^2/f_U + W^2/f_W] = 86,217441 \quad (27)$$

ce qui permet, avec une interpolation entre 86 et 87 degrés de liberté des tables usuelles de Student de trouver $P = 0,123377$ (t.1,3).

D'après ce test bilatéral, les statures des jumeaux MZ et DZ ne diffèrent pas de façon significative, bien qu'on ne soit pas très loin du seuil de 0,05. Pour d'autres mesures également, les moyennes des MZ sont inférieures à celles des DZ, mais seules certaines différences sont significatives (Defrise et al., 1982).

2.2.D. Test F'' pour comparer les sommes des carrés moyens, c'est-à-dire les estimations des variances totales (cf. formule 18)

On calcule d'abord (SMS «sum of mean squares»)

$$AMZ + WMZ = 2\hat{\sigma}_{\text{totMZ}}^2 = 83,411910 \quad \text{et} \quad (\text{t.1,10})$$

$$ADZ + WDZ = 2\hat{\sigma}_{\text{totDZ}}^2 = 77,674410 \quad (\text{t.1,10})$$

Le calcul approché des degrés de liberté de ces sommes de carrés moyens, se fait comme pour la variance utilisée dans le test t' de 2.2.C (27) .

On obtient ainsi

$$83,411910^2 / \left(\frac{80,334280^2}{56} + \frac{3,077630^2}{57} \right) = 60,286 \approx 60,3 \quad (t.I,10) \text{ et}$$

$$77,414410^2 / \left(\frac{67,642620^2}{38} + \frac{10,031790^2}{39} \right) = 49,055 \approx 49,1 \quad (t.T,10).$$

Comme nous verrons dans la 3^e partie, la comparaison des variances totales s'impose pour l'analyse génétique. On calcule leur rapport, mais de manière à obtenir une valeur de F'' supérieure à 1. Pour la stature, cela donne $F'' = 83,41190 / 77,674410 = 1,073866 \quad (t.I,11)$.

Avec 60,3 et 49,1 degrés de liberté, la valeur approchée de P n'est pas significative ; on obtient $P = 0,801665 \quad (t.I,12)$.

2.3. Quatre estimateurs de GT, la portion de la variance génétique que l'on peut estimer et tester avec les données gémellaires

La définition de GT, la partie cruciale de la variance génétique dans les études des jumeaux, sera donnée dans la 3^e partie. Dans le paragraphe présent, nous expliquons, du point de vue statistique, les estimateurs, les variances et les tests de GT.

2.3.A. Estimateur interpairs, désigné par G_{AT} dans Christian et al. (1974) et par AP sur le tableau I

L'estimation interpairs (« among pair » AP) se calcule comme suit

$$\hat{G}_{AT} = AMZ - ADZ = 12,691660 \quad (t.I,14) \text{ et est testée par}$$

$$F = AMZ / ADZ = 1,187628 \text{ avec } 56 \text{ et } 38 \text{ DF} \quad (t.I,14)$$

$$P = 0,290160, \text{ valeur non significative} \quad (t.I,14)$$

La variance de $\hat{G}_{AT}$ est estimée par

$$\hat{V}(\hat{G}_{AT}) = 2 \left[\frac{AMZ^2}{n_{MZ} + 1} + \frac{ADZ^2}{n_{DZ} + 1} \right] = 451,314014 \quad (t.I,14) \quad (28)$$

En effet, on démontre que si Q est une somme de carrés d'écarts à la moyenne avec d degrés de liberté, alors la variance des carrés moyens Q/d est estimée sans biais par

$$\hat{V}(Q/d) = \frac{2}{d+2} (Q/d) \quad (29)$$

Or, AMZ et ADZ ont respectivement comme degrés de liberté $n_{MZ} - 1$ et $n_{DZ} - 1$ ce qui explique les dénominateurs de (28). La démonstration de (29) est basée sur des propriétés du χ^2 .

2.3.B. *Estimateur intrapaires, désigné par $\hat{G}_{WT}$ dans Christian et al. et par WP sur le tableau I*

L'estimation intrapaires (« within pair », WP) est la plus usuelle. Elle vaut

$$\begin{aligned}\hat{G}_{WT} &= \text{WDZ} - \text{WMZ} = 6,954160 \quad (\text{t.I,15}) \text{ et est testée par} \\ F &= \text{WDZ}/\text{WMZ} = 3,259583 \text{ avec } 39 \text{ et } 57 \text{ DF} \quad (\text{t.I,15}) \\ P &= 0,000025, \text{ valeur hautement significative} \quad (\text{t.I,15})\end{aligned}$$

La variance de $\hat{G}_{WT}$ est estimée par (cf 29)

$$\hat{V}(\hat{G}_{WT}) = 2 \left[\frac{\text{WDZ}^2}{n_{DZ} + 2} + \frac{\text{WMZ}^2}{n_{MZ} + 2} \right] = 5,230191 \quad (\text{t.I,15}) \quad (30)$$

Cette variance est très inférieure à celle du GT interpaires ce qui explique peut-être qu'ici le test est hautement significatif. Il en est souvent de même pour d'autres mesures, de sorte que si le modèle génétique (3^e partie) ne s'y oppose pas on préférera toujours l'estimation intrapaires de GT, $\hat{G}_{WT}$ étant une donnée statistiquement plus précise que $\hat{G}_{AT}$.

2.3.C. *Estimateur de GT, moyenne de $\hat{G}_{AT}$ et $\hat{G}_{WT}$, désigné par $\hat{G}_{CT}$ dans Christian et al. et par AC sur le tableau I*

Ce nouvel estimateur (« among component » AC) est la demi-somme des deux estimateurs précédents. Il vaut

$$\begin{aligned}\hat{G}_{CT} &= (\hat{G}_{AT} + \hat{G}_{WT})/2 \text{ ou} \\ \hat{G}_{CT} &= (\text{AMZ} - \text{ADZ} + \text{WDZ} - \text{WMZ})/2 = 9,822910 \quad (\text{t.I,16}) \quad (31) \\ &\text{et est testé par}\end{aligned}$$

$$\begin{aligned}F'' &= (\text{AMZ} + \text{WDZ})/(\text{ADZ} + \text{WMZ}) = 1,277796 \quad (\text{t.I,16}) \text{ avec} \quad (32) \\ P &= 0,198862, \text{ valeur non significative} \quad (\text{t.I,16}).\end{aligned}$$

Le test F'' est unilatéral car si des facteurs génétiques agissent sur le caractère mesuré, on s'attend à une valeur du numérateur supérieure à celle du dénominateur, comme nous le verrons dans la 3^e partie.

Les degrés de liberté de F'' calculés comme dans la formule (27), sont respectivement 69,3 et 41,5.

La variance de l'estimateur $\hat{G}_{CT}$ vaut le quart de la somme des variances de $\hat{G}_{AT}$ et $\hat{G}_{WT}$:

$$\hat{V}(\hat{G}_{CT}) = [\hat{V}(\hat{G}_{AT}) + \hat{V}(\hat{G}_{WT})]/4 = 114,136051 \quad (\text{t.I,16}) \quad (33)$$

2.3.D. Estimateur de GT à variance minimum, désigné par $\hat{G}_{MT}$ dans Christian et al. et par « minimum variance » (t.1,17)

Des trois estimations de GT, c'est justement $\hat{G}_{WT}$, avec une très petite variance, pour laquelle le test est significatif. C'est pourquoi Christian et al. ont cherché systématiquement la fonction linéaire de $\hat{G}_{AT}$ et $\hat{G}_{WT}$ qui a une variance minimum. Le calcul de cet extremum et de la variance correspondante est expliqué dans Christian et al. et n'a, de l'avis même des auteurs, qu'un intérêt théorique. Nous ne le reprenons pas et ne donnons ici que le résultat numérique pour la taille. On trouve (t.1,17) $\hat{G}_{MT} = 7,019889$, avec une variance égale à 5,170274, la plus petite des quatre variances calculées, mais à peine inférieure à celle de $\hat{G}_{WT}$.

Résumé de la 2^e partie

Dans la 2^e partie, nous avons développé les formules statistiques qui sont à la base du tableau I élaboré grâce au programme de Christian et al. Nous avons commencé par les groupes séparés de jumeaux monozygotiques et dizygotiques et exposé l'analyse de la variance de Fisher parce que c'est elle qui vient le plus naturellement à l'esprit pour la comparaison des variabilités intra- et interpaires. Ensuite, nous avons explicité les tests statistiques pour la comparaison des deux types de jumeaux, donné les formules des estimateurs de GT, partie de la variance génétique, et les estimations de GT pour la stature. Certaines formules contiennent à la fois l'estimateur et l'estimation, sans que nous insistions sur la distinction entre les deux termes. Dans de tels cas, nous avons simplement indiqué « estimation ».

3. ANALYSE DE LA VARIANCE GENETIQUE.

Nous exposons le modèle génétique destiné à analyser la variance d'une mensuration à l'aide des données gémellaires. Nous suivons l'analyse de Christian et al. (1974) qui se sont eux-mêmes inspirés des modèles de Kempthorne et Osborne (1961) et de Haseman et Elston (1970). Le but de ces études est de découvrir si un caractère mesurable est, du moins partiellement, influencé par le génome.

3.1. Décomposition de la variance totale σ^2_{tot} d'un caractère mesurable

En admettant que les gènes de plusieurs locus contribuent au développement d'un caractère biométrique, la variance d'une population panmictique se décompose comme suit d'après Kempthorne (1954), Cockerham (1954), Haseman et al. (1970), Christian et al. (1974), Defrise (1981) :

$$\sigma^2_{tot} = V(G) + 2 \text{cov}(G, E) + V(E) \quad \text{où} \quad (34)$$

$$V(G) = \sigma_a^2 + \sigma_d^2 + \sigma_i^2 \quad (35)$$

est la composante due à la diversité génétique de la population,

$$\sigma_{ge} = \text{cov}(G, E) \quad (36)$$

la covariance entre les effets génétiques et mésologiques et

$$\sigma_e^2 = V(E) \quad (37)$$

la composante de la variance due aux effets mésologiques.

Rappelons que dans (35)

σ_a^2 est la partie de la variance due aux effets additifs des gènes,

σ_d^2 celle due aux effets de dominance (locus par locus),

σ_i^2 celle due aux effets de l'interaction des gènes des différents locus (les effets épistatiques).

Pour l'analyse gémellaire, on considère en outre

σ_{ge}^* , la covariance entre les effets génétiques sur le caractère biométrique d'un jumeau et les effets mésologiques sur le caractère de son co-jumeau ;

C_{MZ} et C_{DZ} sont les covariances entre les effets mésologiques agissant sur les mesures de deux jumeaux, resp. MZ et DZ. Enfin $\sigma_i^2(1-f)$ est la partie de la variance épistatique qui se manifeste entre paires de jumeaux DZ (Kempthorne, 1954).

3.2. Espérances mathématiques des carrés moyens

On démontre que les espérances des carrés moyens (7,8) sont

$$E(AMZ) = 2(\sigma_a^2 + \sigma_d^2 + \sigma_i^2 + 2\sigma_{ge}) + \sigma_{eMZ}^2 + C_{MZ} \quad (38)$$

$$E(WMZ) = \sigma_{eMZ}^2 - C_{MZ} \quad (39)$$

$$E(ADZ) = 3/2\sigma_a^2 + 5/4\sigma_d^2 + (1+f)\sigma_i^2 + 2(\sigma_{ge} + \sigma_{ge}^*) + \sigma_{eDZ}^2 + C_{DZ} \quad (40)$$

$$E(WDZ) = 1/2\sigma_a^2 + 3/4\sigma_d^2 + (1-f)\sigma_i^2 + 2(\sigma_{ge} - \sigma_{ge}^*) + \sigma_{eDZ}^2 - C_{DZ} \quad (41)$$

Remarquons que les variances totales multipliées par deux sont respectivement.

$$E(AMZ + WMZ) = 2[V(G) + 2\text{cov}(G, E)] + 2\sigma_{eMZ}^2 = 2\sigma_{\text{totMZ}}^2 \text{ et } \quad (42)$$

$$E(ADZ + WDZ) = 2[V(G) + 2\text{cov}(G, E)] + 2\sigma_{eDZ}^2 = 2\sigma_{\text{totDZ}}^2. \quad (43)$$

3.3. Test pour l'égalité des variances σ_{eMZ}^2 et σ_{eDZ}^2

Suivant que les variances dues aux effets mésologiques sont égales ou inégales dans les populations dont sont issus les deux types de jumeaux, une estimation différente de la partie GT de la variance génétique (voir 2.3 et 3.4) s'impose.

Remarquons d'abord que les variances totales des deux types de jumeaux (42) et (43), ne diffèrent précisément que par σ_{eMZ}^2 et σ_{eDZ}^2 . L'hypothèse nulle est donc $\sigma_{eMZ}^2 = \sigma_{eDZ}^2$ et on teste $F_1' = (ADZ + WDZ)/(AMZ + WMZ)$ ou $F_2'' = 1/F_1'$.

Le programme choisit celui des deux F'' qui est supérieur à 1. Le calcul de F'' et des degrés de liberté sont expliqués au 2.2.D. Pour la stature, le test n'est pas significatif avec $P = 0,801665$ (t.l,12); nous admettrons donc dans la suite que, pour la taille,

$$\sigma_{eMZ}^2 = \sigma_{eDZ}^2 = \sigma_e^2 \quad (44)$$

Christian et al. préconisent un seuil de signification de $\alpha = 0,20$, car la puissance du test F'' est faible ce qui est compréhensible étant donné que $V(G) + 2 \text{ cov}(G, E)$, la part commune aux deux termes de F'' , peut être considérable par rapport à σ_e^2 .

3.4. Définition génétique de GT et conditions d'utilisation du modèle

Pour définir génétiquement GT, il faut supposer que $\sigma_{ge}^* = \sigma_{ge}$ et $C_{MZ} = C_{DZ}$; ces conditions ne peuvent pas être testées à l'aide des données gémellaires. Dans ce qui suit nous admettrons souvent avec Christian et al. que ces deux égalités sont satisfaites.

On obtient alors, en définissant

$$GT = 1/2 \sigma_a^2 + 3/4 \sigma_d^2 + (1-f) \sigma_i^2 \quad (45)$$

et en combinant (38), (39), (40), (41)

$$E(\hat{G}_{AT}) = E(AMZ - ADZ) = GT + (\sigma_{eMZ}^2 - \sigma_{eDZ}^2) \quad \text{et} \quad (46)$$

$$E(\hat{G}_{WT}) = E(WDZ - WMZ) = GT - (\sigma_{eMZ}^2 - \sigma_{eDZ}^2). \quad (47)$$

La demi-somme de ces deux expressions vaut

$$E(\hat{G}_{CT}) = GT \quad (48)$$

Comme nous le verrons plus loin (67 et 68), GT est la composante de la variance génétique qui est précisément la cause d'une variance intrapaires plus grande chez les DZ que chez les MZ dès que des effets génétiques agissent sur le caractère mesuré.

3.5. Choix d'un estimateur de GT

Si l'on admet que $\sigma_{eMZ}^2 \neq \sigma_{eDZ}^2$ parce que le test F'' décrit en 2.2.D et 3.2 est significatif avec $P < 0,20$, on prendra

$$\hat{G}_{CT} = (AMZ - ADZ + WDZ - WMZ) / 2, \text{ un estimateur sans biais de GT (48).}$$

Si l'on admet que $\sigma_{eMZ}^2 = \sigma_{eDZ}^2$, le test F'' étant non significatif avec $P > 0,20$, on prendra $\hat{G}_{WT} = WDZ - WMZ$, un estimateur sans biais de GT (46). On préférera $\hat{G}_{WT}$ à $\hat{G}_{AT}$ parce qu'il a en général une plus petite variance (2.3). Remarquons que pour presque toutes les mesures F'' est non significatif, car nos deux échantillons de jumeaux sont issus de la même population.

4. SYNTHÈSE DES ANALYSES STATISTIQUE ET GÉNÉTIQUE.

Le lien entre l'analyse de la variance de Fisher (2.1.B et 2.1.C) et l'analyse génétique de Christian et al. (partie 3) est obtenu en égalant les espérances mathématiques des carrés moyens (14), (15), (38), (39), (40), (41).

4.1. Jumeaux MZ

Pour les jumeaux MZ on obtient les variances inter et intrapaires suivantes :

$$E(AMZ) = \sigma_{MZ}^2 + 2 \sigma_{fMZ}^2 = 2[V(G) + 2\sigma_{ge}] + \sigma_{eMZ}^2 + C_{MZ} \text{ et } (49)$$

$$E(WMZ) = \sigma_{MZ}^2 = \sigma_{eMZ}^2 - C_{MZ} \quad (50)$$

En divisant par deux la somme de ces deux expressions, on trouve

$$\sigma_{MZ}^2 + \sigma_{fMZ}^2 = [V(G) + 2\sigma_{ge}] + \sigma_{eMZ}^2 = \sigma_{tot}^2, \text{ la variance totale des MZ } (51)$$

Des formules (49) et (50) on tire aisément

$$\sigma_{MZ}^2 = \sigma_{eMZ}^2 - C_{MZ}, \text{ la variance intrapaires, et } (52)$$

$$\sigma_{fMZ}^2 = V(G) + 2\sigma_{ge} + C_{MZ}, \text{ la variance interfamilles. } (53)$$

Pour serrer de plus près la signification de ces deux égalités, on peut écrire

$$C_{MZ} = \sigma_{eMZ}^2 \rho_{MZ} \quad (54)$$

ρ_{MZ} étant le coefficient de corrélation intraclasse entre jumeaux MZ si seuls les effets mésologiques agissaient. Pour la facilité, nous l'appelons, dans ce texte, coefficient de corrélation «mésologique». On peut alors écrire

$$\sigma_{MZ}^2 = \sigma_{eMZ}^2 (1 - \rho_{MZ}) \quad (55)$$

$$\sigma_{fMZ}^2 = V(G) + 2\sigma_{ge} + \sigma_{eMZ}^2 \rho_{MZ} \quad (56)$$

ce qui montre que pour une corrélation «mésologique» croissante, les variances intrapaires et interfamilles varient en sens inverse : lorsque les jumeaux d'une paire partagent un milieu très semblable (ρ_{MZ} élevé), la variance interfamilles augmente. Les formules (55) et (56) peuvent être rapprochées des analyses intuitives des fig. 6 et 7 (1.2.A) et de la discussion (2.1.C).

Pour clarifier la distinction entre les variances interpaires et interfamilles, on compare les formules (49) et (56)

$$E(AMZ) = 2[V(G) + 2\sigma_{ge}] + \sigma_{eMZ}^2 (1 + \rho_{MZ}), \text{ la variance interpaires } (57)$$

$$\sigma_{fMZ}^2 = [V(G) + 2\sigma_{ge}] + \sigma_{eMZ}^2 \rho_{MZ}, \text{ la variance interfamilles } (58)$$

La variance interpaires est égale à la variance interfamilles augmentée de la variance totale (cf formule 34 et 2.1.B) :

$$E(AMZ) = \sigma_{fMZ}^2 + \sigma_{totMZ}^2 = \text{la variance interpaires} \quad (59)$$

4.2. Jumeaux DZ

Pour des jumeaux DZ, la quantité GT intervient dans les formules ainsi que P_{DZ} , le coefficient de corrélation intraclasse « mésologique » pour les DZ.

On trouve pour les variances inter-et intrapaires

$$E(ADZ) = \sigma_{DZ}^2 + 2\sigma_{fDZ}^2 = 2[V(G) + 2\sigma_{ge}] - GT + \sigma_{eDZ}^2(1 + P_{DZ}) \quad (60)$$

$$E(WDZ) = \sigma_{DZ}^2 = GT + \sigma_{eDZ}^2(1 - P_{DZ}) \quad (61)$$

La demi-somme de ces expressions donne la variance totale σ_{totDZ}^2 .

Les formules 60 et 61 montrent que si la composante génétique GT est grande, la variance intrapaires augmente et la variance interpaires diminue, tandis que si les jumeaux de chaque paire partagent un milieu très semblable (P_{DZ} grand), c'est l'inverse qui se passe.

Si l'on veut comparer les variances intrapaires et interfamilles, on peut écrire

$$\sigma_{DZ}^2 = GT + \sigma_{eDZ}^2(1 - P_{DZ}) \quad (62)$$

$$\sigma_{fDZ}^2 = [V(G) + 2\sigma_{ge}] - GT + \sigma_{eDZ}^2 P_{DZ} \quad (63)$$

dont la somme n'est autre que la variance totale. Ici encore, les effets de GT et P_{DZ} sont de sens contraires sur les variances intrapaires et interfamilles.

Finalement, la comparaison des variances interpaires et interfamilles (60) et (63) conduit à

$$E(ADZ) = 2[V(G) + 2\sigma_{ge}] - GT + \sigma_{eDZ}^2(1 + P_{DZ}) \quad (64)$$

$$\sigma_{fDZ}^2 = [V(G) + 2\sigma_{ge}] - GT + \sigma_{eDZ}^2 P_{DZ} \quad (65)$$

Ici encore, la variance interpaires est égale à la variance interfamilles augmentée de la variance totale (cf formule 34 et 2.1.B) :

$$E(ADZ) = \sigma_{fDZ}^2 + \sigma_{totDZ}^2 = \text{la variance interpaires} \quad (66)$$

4.3. Comparaison des variances intrapaires des MZ et des DZ

Dans ce paragraphe, nous admettons que $\sigma_{eMZ}^2 = \sigma_{eDZ}^2 = \sigma_e^2$;

$$C_{MZ} = C_{DZ} = C \text{ et } \sigma_{ge}^* = \sigma_{ge}.$$

Dans ces conditions, les variances intrapaires valent, respectivement pour les MZ et les DZ :

$$E(WMZ) = \sigma_{MZ}^2 = \sigma_e^2 - C = \sigma_e^2(1 - P) \quad (67)$$

$$E(WDZ) = \sigma_{DZ}^2 = GT + \sigma_e^2 - C = GT + \sigma_e^2(1 - P) \quad (68)$$

La seule différence est GT, cette partie cruciale de la variance génétique; la variance intrapaires des jumeaux DZ sera supérieure à celle des jumeaux MZ dans la mesure où GT dépasse zéro de façon significative. La différence entre deux jumeaux MZ est due uniquement aux effets des facteurs mésologiques, celle entre les jumeaux DZ est augmentée de GT, une partie de la variance génétique.

L'hypothèse nulle $GT = 0$ sera testée par $F = WDZ / WMZ$ et F sera significatif lorsque GT est suffisamment grand par rapport à $\sigma_e^2 - C$. Si la variation mésologique est considérable, si la covariance $C = \sigma_e^2 P$ est faible, il y a peu de chances que GT, même non nul, puisse rendre F significatif. Pour dépister GT, deux conditions doivent donc être remplies :

1. les jumeaux doivent vivre aussi peu séparés que possible pour que C soit grand, c'est-à-dire proche de σ_e^2 ;
2. la variance σ_e^2 , due aux facteurs mésologiques, doit être faible. Il est donc souhaitable que la population dont les jumeaux sont issus tienne sa variation plus des facteurs génétiques que des facteurs mésologiques. Si possible, une population très homogène quant au milieu devrait être choisie.

4.4. Comparaison des coefficients de corrélation intraclasse chez les MZ et les DZ

Suivant les formules (20) et (21), le coefficient de corrélation intraclasse est $R = \sigma_f^2 / \sigma_{tot}^2$. Si on remplace les termes de R par leur espérance, respectivement pour les MZ, (51) et (56), pour les DZ, (63) on obtient

$$([V(G) + 2 \sigma_{ge}] + \sigma_{eMZ}^2 P_{MZ}) / ([V(G) + 2 \sigma_{ge}] + \sigma_{eMZ}^2) \text{ et} \quad (69)$$

$$([V(G) + 2 \sigma_{ge}] - GT + \sigma_{eDZ}^2 P_{DZ}) / ([V(G) + 2 \sigma_{ge}] + \sigma_{eDZ}^2) \quad (70)$$

Dans (69), c'est P_{MZ} , la corrélation «mésologique», fonction du milieu, qui fait augmenter R_{MZ} . Dans (70), R_{DZ} croît aussi avec P_{DZ} , fonction des effets mésologiques, mais décroît avec GT, qui ne dépend que de la variation génétique.

Le test comparant R_{MZ} et R_{DZ} , décrit dans 2.2.A, sera significatif, avec $R_{MZ} > R_{DZ}$, lorsque GT sera non négligeable et aussi lorsque P_{MZ} sera supérieur à P_{DZ} . Ce dernier cas se présente souvent parce que les jumeaux d'une paire MZ partagent plus volontiers le même milieu que ceux d'une paire DZ.

4.5. Coefficients d'héritabilité des lignes 31, 32, 33 du tableau I

Les coefficients d'héritabilité «among component», «within pair» et «component» s'interprètent aisément si dans les formules t.I, 31, 32, 33, on remplace les termes des fractions par leur espérance mathématique. En combinant les formules (42), (43), (46), (47), (48), on trouve respectivement pour les deux premiers coefficients $2 GT / 1/2 (\sigma_{totMZ}^2 + \sigma_{totDZ}^2)$ et

$$[GT - (\sigma_{eMZ}^2 - \sigma_{eDZ}^2)] / 1/2 (\sigma_{totMZ}^2 + \sigma_{totDZ}^2).$$

Le troisième coefficient est, bien entendu, égal au premier. Dans le cas où $\sigma_{eMZ}^2 = \sigma_{eDZ}^2$, la première expression est le double de la seconde.

5. REMARQUES SUR LA VALIDITE DU MODELE ANALYSE.

La validité du modèle statistique et génétique que nous venons de décrire dépend, nous l'avons vu, de plusieurs conditions. Certaines sont faciles à admettre comme

a) la normalité des variables (10, 11, 12). En effet, les tests de l'analyse de la variance sont robustes et d'autre part, on peut, s'il le faut, normaliser certaines variables comme le poids;

b) l'indépendance de z_{1j} et z_{2j} (10), car la variable f_j épuise à elle seule tout ce qui contribue à la covariance entre x_{1j} et x_{2j} .

D'autres conditions sont plus criticables :

a) l'indépendance de f_j et z_{ij} (11, 12, 13) est contredite par la corrélation positive observée entre $a + b$ et $la - bl$ (t.I,25), c'est-à-dire que l'écart intrapaire tend à augmenter avec la moyenne de la paire (cf. 2.1.E).

b) la nécessité de supposer $C_{DZ} = C_{MZ}$ et $\sigma_{ge}^* = \sigma_{ge}$ (3.4).

Mais à notre avis le plus grand défaut de toutes les analyses gémellaires est de négliger l'interaction entre les effets génétiques et mésologiques. A la difficulté de quantifier les effets mésologiques sur le développement s'ajoute notre ignorance des règles du jeu continu de rétroaction entre les gènes et le milieu, dès la conception. Aucune fonction mathématique des deux types de variables (effets génétiques, effets mésologiques) n'a été proposée (Defrise, 1981).

Nous pensons cependant que le modèle proposé est utile, malgré ses faiblesses, fût-ce pour comparer entre elles les différentes mesures et la composante génétique de leur variance ou pour confronter les résultats de différents échantillons de jumeaux.

BIBLIOGRAPHIE

CHRISTIAN J.C., KE WON KANG et J.A. NORTON Jr.

1974 Choice of an estimate of genetic variance from twin data.
Am. J. hum. Genet., **26** : 154-161.

COCKERHAM C.C.

1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present.
Genetics, **39** : 859-882.

DEFRISE-GUSSENHOVEN, E.

1981 Modèle polygénique pour les caractères mesurables.
Bull. Soc. roy. belge Anthropol. Préhist., **92** : 7-24.

- DEFRISE-GUSSENHOVEN, E., C. SUSANNE et R. ORBAN-SEGEBARTH.
 1982 Biométrie de jumeaux belges âgés de 18 à 25 ans.
Bull. Mém. Soc. Anthropol. Paris, **9** (12) : 177-189.
- DUPAE, E., C. SUSANNE et E. DEFRISE-GUSSENHOVEN.
 1982 Genetic and environmental influences on body measurements of Belgian twins.
Acta. Genet. Med. Gemellol., **31** : 139-144.
- ELSTON, R.C. et C.E. BOKLAGE.
 1978 An examination of fundamental assumptions of the twin method.
 in W.E. NANCE, G. ALLEN et P. PARISI (Edit.) : *Twin research*.
 Part A., New York, Alan R. Liss : 189-199.
- FISHER, R.A.
 1954 *Statistical methods for research workers*.
 London, Oliver and Boyd, 12th ed., 356 pp.
- GALTON, F.
 1885 Regression towards mediocrity in hereditary stature.
Journ. anthropol. Inst. Great Brit. and Ireland, **15** : 246-258.
- HASEMAN, J.K. et R.C. ELSTON.
 1970 The estimation of genetic variance from twin data.
Behav. Genet., **1** : 11-19.
- HAUSPIE, R.C., C. SUSANNE et E. DEFRISE-GUSSENHOVEN.
 1985 Testing for the presence of genetic variance in factors of face measurements of Belgian twins.
Ann. of human Biology, **12** (5) : 429-440.
- KEMTHORNE, O.
 1954 The theoretical values of correlations between relatives in random mating populations.
J. of Iowa Agric. Exper. Station, Ames Iowa, **J.2609** : 153-167.
- KEMPTHORNE, O. et R.H. OSBORNE.
 1961 The interpretation of twin data.
Amer. J. hum. Genet., **13** : 320-339.
- SCHEFFE, H.
 1959 *The analysis of variance*.
 New York, John Wiley and Sons, 477 pp.
- SUSANNE, C., E. DEFRISE-GUSSENHOVEN, P. VAN WANSEEELE et A. TASSIN.
 1983 Genetic and environmental factors in head and face measurements of Belgian twins.
Acta Genet. Med. Gemellol., **32** : 229-238.

Tableau I

S T A T U R E

ANALYSIS OF TWIN DATA									
1 TEST FOR DIFFERENCE BETWEEN MZ AND DZ MEANS									
2	MZ MEAN								
3	175. 585090								
4	ANALYSIS OF VARIANCE								
5	NO OF PAIRS								
6	AMONG MEAN SQUARES (AMS)								
7	WITHIN MEAN SQUARES (WMS)								
8	TOTAL SUM OF SQUARES								
9	TEST OF TOTAL VARIANCE								
10	SUM OF MEAN SQUARES (SMS)								
11	RATIO OF SUM OF MEAN SQUARES								
12	APPROXIMATE PROBABILITY OF F**								
13	ESTIMATES OF GENETIC VARIANCE								
14	AMONG PAIR (AP)								
15	WITHIN PAIR (WP)								
16	AMONG COMPONENT (AC)								
17	MINIMUM VARIANCE								
18	TEST TO EXCLUDE CMZ GREATER THAN (D/								
19	ADZ/WDZ =								
20	PROB =								
21	SAMPLE STATISTICS								
22	MEAN								
23	VARIANCE								
24	INTRA CORR (R)								
25	CORR: (A + B)/A-B/								
26	MEAN ABS DIFFERENCE								
27	INTRACLAS CORRELATION								
28	CORRELATION (R)								
29	Z TRANSFORMATION								
30	HERITABILITY ESTIMATE								
31	AMONG COMPONENT								
32	WITHIN PAIR								
33	COMPONENT								
34	INTRACLAS CORRELATION								
35	1. HOLZINGER								
36	2. * (RMZ-RDZ)								

PROBABILITY
. 123377DF
86. 217441T-PRIME
-1. 555975MZ - DZ
-1. 950810

MZ

DZ MEAN
177. 535900

ANALYSIS OF VARIANCE

NO OF PAIRS

AMONG MEAN SQUARES (AMS)

WITHIN MEAN SQUARES (WMS)

TOTAL SUM OF SQUARES

DF
39DF
38DF
39DF
77

57

80. 334280

3. 077630

4674. 144590

83. 411910

60. 3

F** =

F-RATIO
1. 187628PROBABILITY
. 290160F-RATIO
3. 259583VARIANCE
5. 230191

114. 136051

5. 170274

P =

. 801665

PROBABILITY
(DF = 69. 3, 41. 5)PROBABILITY
. 000025

. 198862

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829PROBABILITY
. 000000PROBABILITY
. 000000PROBABILITY
. 000000

1. 928070

175. 585090

41. 364110

. 926206

. 771560

MZ

. 926206

PROBABILITY
. 000829

PROB