Continuous variation caused by genes with graduated effects
(polygenic inheritance/multifactorial transmission)

STEVEN MATTHYSESE, KENNETH LANGE, and DIANE K. WAGENER

1Mailman Research Center, McLean Hospital, Belmont, Massachusetts 02178; 2Department of Biomathematics, University of California, Los Angeles, California 90024; and 3Western Psychiatric Institute, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

Communicated by Seymour S. Kety, February 26, 1979

ABSTRACT  The classical polygenic theory of inheritance posulates a large number of genes with small, and essentially similar, effects. We propose instead a model with genes of gradually decreasing effects. The resulting phenotypic distribution is not normal; if the gene effects are geometrically decreasing, it can be triangular. The joint distribution of parent and offspring genic value is calculated. The most readily testable difference between the two models is that, in the decreasing-effect model, the variance of the offspring distribution from given parents depends on the parents' genic values. The more the parents deviate from the mean, the smaller the variance of the offspring should be. In the equal-effect model the offspring variance is independent of the parents' genic values.

In the classical polygenic model of inheritance, the contributions of a large number of loci with small but similar effects are summed to produce a phenotypic distribution that is approximately normal (for exact conditions, see ref. 1). It seems more plausible to assume that the contributions of the loci are unequal, with a few major genes and a larger number of minor genes. Indeed, it may not be an exaggeration to suggest that every locus contributes in some way to any given quantitative trait, although the effects of most loci are vanishingly small. An example of unequal gene contributions to a continuously varying trait is heading time in wheat, for which one locus was detected that accounted for about 80% of the additive variance in the trait, and three other loci that jointly accounted for about 14% of the additive variance. The remainder presumably represents the contributions of loci of even smaller effect (2). The common observation of unequal response of different lines to selection for a quantitative trait may be explained in some cases by segregation of major genes of a polygenic character (3). Because the individual genes contributing to a quantitative trait affect the phenotype in different ways, it is not surprising to find their contributions unequal. For example, in Drosophila, wing size is associated predominantly with variance in cell number and to a lesser degree with cell size (4).

We assume an infinite number of loci, each with two possible alleles A, a; the two alleles actually present at each locus are assumed independent of each other and of the alleles at the other loci. The genetic contributions are additive without dominance, X, the representation of the paternal allele and Y, the representation of the maternal allele at the nth locus. The loci are not assumed to have equal effects; the contributions of each allele at the nth locus are δn + a from An (probability p), δn - a from Bn (probability 1 - p). δn is so chosen that the means of X and Y are zero: δn = -2pn an. Assume the an are bounded.

The variance of X (or Y) is σ2 = 4p(1 - 4p)2. There are two possibilities: (i) σ2 = ∞; (ii) σ2 < ∞.

Case i. σ2 = ∞. Let σ2 = σ2 + ... + σ2. The distribution of the normalized sum (X1 + ... + Xn)/Sn tends to the normal distribution with zero mean and unit variance, as n → ∞ (ref. 5, p. 264).

Case ii. Σn=1 σ2 = ∞. Then X1 + ... + Xn converges almost surely to a random variable X (ref. 6, p. 502). X is the limit of the contributions of the paternal alleles only; the distribution function of genic value G = X - Y is the self-convolution of the distribution function of X. In this case, X (and therefore G also) cannot have a normal distribution, because whenever X1 + ... + Xn converges to a limit X, X2 + ... + Xn converges to a limit X'; X = X1 + X'; and X1 and X' are independent. If the sum of two independent random variables is normally distributed, each of them also has a normal distribution (ref. 5, p. 525), whereas X1 is atomic. For example, if pn = 1/2 (all n), the normalized distribution F(x) of X satisfies

|F(x) - N(x)| ≤ 6 Σn=1 a3/3/2 for all x, by the Berry–Esseen theorem (N(x) is the normal distribution with zero mean and unit variance) (ref. 5, p. 544). This quantity can be made vanishingly small by choosing appropriate a's; e.g., a = n for b near and greater than 1/2 or a = a for a near and less than 1.

Classification of limit distributions

According to a remarkable theorem of Jessen and Wintner (7), a convergent convolution of atomic distributions is either (i) atomic, (ii) absolutely continuous, or (iii) continuous but singular. No mixed cases are possible. In other words, either (i) X takes only a discrete set of values; (ii) X has a probability density (not necessarily continuous or bounded); (iii) X has no single value with positive probability, but there is a set of Lebesgue measure zero such that X is in this set with probability 1. In order to construct a model that is suitable for genetic applications, it is reasonable to consider only absolutely continuous limits with densities that are continuous or at least bounded.

A necessary and sufficient condition that X have a continuous distribution function is Πn=1 max (pn, 1 - pn) → 0 (ref. 6, p. 537); alternatively, Σn=1 min (pn, 1 - pn) = ∞. For further study of the classification problem, we will assume pn = 1/2 (all n). The limit distribution is then continuous, although not necessarily absolutely continuous. The characteristic function of X is φ(t) = Πn=1 cos (an t) because δn = 0. Wintner (8) discusses such characteristic functions in detail. For example, if an = n for b > 1/2, the corresponding distribution has an infinitely differentiable density (ref. 8, p. 147). If a = a and a = (b/2)1/k (k = 1, 2, ...), then φ(t) = O(|t| - k) (refs. 8, 9), so the characteristic function of G is O(|t| - 2k). G therefore has a continuous density with at least 2k - 2 contin...
uous derivatives (ref. 8, p. 118). We will treat the case $a = \frac{1}{2}$ more fully later, because it leads to a computable density.

Surprisingly, even though the distribution corresponding to $a = \left(\frac{1}{2}\right)^{1/k}$ is absolutely continuous for $k = 1, 2, \ldots, \frac{1}{2} < a < 1$ does not guarantee the existence of a density for $G$. An example is $a = \frac{1}{2}(\sqrt{5} - 1)$, for which $\phi(t)$ does not $\to 0$ as $|t| \to \infty$ (ref. 8, p. 142), so its self-convolution cannot have a density (ref. 5, p. 514).

If $a_n = a^n$, in which $a = 1/k$ ($k = 3, 4, \ldots$), $\phi(t)$ does not $\to 0$ as $|t| \to \infty$ (ref. 8, p. 140), so $G$ is singular. For general $a$ in the range $0 < a < 1/2$, it is known that $X$ is singular (ref. 8, p. 140). This suffices to show that $G$ does not have a bounded or continuous density, because whenever the self-convolution of a symmetric probability distribution $F$ has a bounded density, $F$ itself has a density. Indeed, the characteristic function $\phi$ of $F$ satisfies $\phi^2 \geq 0$ because $F$ is symmetric (ref. 5, p. 499) and furthermore $\phi^2 \leq L$ because $F \ast F$ has a bounded density (ref. 5, p. 510). Consequently $\phi \in \mathbb{L}^2$, and the appropriately normalized inverse Plancherel transform of $\phi$ is the required density. A continuous density for $F \ast F$ is also impossible because $F \ast F$ is concentrated on a finite interval.

Distribution of genic value when $p_n = \frac{1}{2}$ (all $n$)

When $p_n = \frac{1}{2}$ (all $n$), $G$ has characteristic function $\theta(t) = [\prod_{n=1}^\infty \cos(a_n t)]^2$. Because the $a_n$ are bounded, for $|t|$ sufficiently small $\log \theta(t)$ can be expanded as $2 \sum_{n=1}^\infty \log \cos(a_n t)$, and each term has the power series expansion

$$\log \cos(a_n t) = \sum_{n=1}^\infty (-1)^n \frac{n(2n - 1)B_{2n} a_n^2 t^{2n}}{n(2n)!},$$

in which $B_n$ is the $n$th Bernoulli number. This expression is derived by integrating term by term the expansion for the tan function (ref. 9, p. 20). The double series can be rearranged because all terms are negative, and when compared with the expansion of characteristic functions in terms of cumulants (ref. 10, p. 115),

$$\log \theta(t) = \sum_{n=1}^\infty \frac{K_n}{n!} (it)^n,$$

yields for the cumulants:

$$K_{2n} = \frac{2^{2n}(2n - 1)B_{2n}}{n} \sum_{k=1}^{n} a_k^{2n} (n \geq 1),$$

the odd cumulants vanishing. From the cumulants, the moments may be derived (ref. 11, p. 69).

If, in addition, the $a_n$ satisfy $a_n = 2^{-(n+1)}$, $X$ is uniformly distributed in $[-1/2, 1/2]$ (ref. 6, p. 557). The self-convolution of a uniform distribution is triangular (ref. 5, p. 27), so the density function of $G$ is $1 - |x|$ for $|x| \leq 1$ and 0 for $|x| > 1$ (Fig. 1).

Joint distribution of parents and offspring

Assume $p_n = \frac{1}{2}$ (all $n$) and $a_n = -2^{-(n+1)}$, other assumptions as above. The joint distribution of the genic values of two parents ($G, G'$) and one child ($G''$), assuming random mating, can be computed explicitly by using characteristic functions. We will show that the conditional distribution of offspring genic value, given the parents' genic values, is trapezoidal. Most important, the genic variance of the offspring is not independent of the parents' genic values, and this constitutes a major difference from the classical "polygenic" model.

Let the contribution of the father's paternal allele at the $n$th locus be $X_n (= \pm 2^{-(n+1)})$ and that of his maternal allele be $Y_n$; the same quantities for the mother will be called $X_n'$ and $Y_n'$. The contributions of the child's paternal and maternal allele will be called $X_n''$ and $Y_n''$. The characteristic function of the random variable $(G_n, G_m, G_n'')$ is $\theta_n(t_1, t_2, t_3) = E \exp[it_1 X_n + it_2 Y_n + it_3 X_n' + it_2 Y_n'].$

For fixed $X_n, Y_n, X_n', Y_n'$ takes on the values $X_n$ and $Y_n$ with equal probability and $Y_n'$ independently takes on the values $X_n'$ and $Y_n'$ with equal probability. Consequently

$$E \left[ \exp[it_3 (X_n + Y_n')] \right] E \left[ \exp[it_3 (X_n' + Y_n)] \right]$$

$$\times E \left[ \exp[it_3 (X_n' + Y_n')] \right] = \frac{1}{4} \exp[it_3 X_n Y_n + it_3 Y_n + it_3 X_n + it_3 Y_n']$$

$$+ \frac{1}{4} \exp[it_3 X_n Y_n + it_3 X_n + it_3 Y_n + it_3 X_n']$$

$$+ \frac{1}{4} \exp[it_3 X_n Y_n + it_3 Y_n + it_3 X_n' + it_3 Y_n'']$$

$$+ \frac{1}{4} \exp[it_3 X_n Y_n + it_3 Y_n + it_3 Y_n' + it_3 X_n']$$

Because $X_n, X_n', X_n''$ and $Y_n, Y_n', Y_n''$ are independent with identical distributions, the expected value of each term is the same, and

$$\theta_n(t_1, t_2, t_3) = \phi_n(t_1 + t_2 + t_3) \phi_n(t_1) \phi_n(t_2),$$

in which $\phi_n$ is the characteristic function of $X_n$. Because of the independence of the alleles at distinct loci, the characteristic function of the random variable $(G_n, G_m, G_n'')$ is the product of the $\theta_n(t_1, t_2, t_3) = \phi(t_1 + t_2 + t_3) \phi(t_1) \phi(t_2)$, in which $\phi$ is the characteristic function of $X$. Setting $t_2 = 0$ yields the characteristic function $\theta(t_1, t_3) = \phi(t_1 + t_3) \phi(t_1) \phi(t_3)$ of $(G, G'')$. Let $u_i$ denote the uniform density in $[-1/2, 1/2]$; $u_i(x) = u(x + \delta)$. Let $g(x_1, x_2) = u_{x_1} * u_{x_2} * u_{x_3}$, in which $u_{x_i}$ denotes the product of $u$ and $u_i$ and * denotes convolution. Consider the Fourier transform of $g(x_3 - x_1 - x_2; x_1, x_2)$.

![Fig. 1. Frequency distribution of genic value in the population according to the geometrically-decreasing-effect model. The curves on the right include independent normally distributed environmental components with standard deviations 0.1, 0.25, and 0.5 (from top to bottom at frequency intercepts).](image)

![Fig. 2. Conditional distribution of offspring genic value for given parents' genic values. $\mu, s, \text{and} s'$ are defined in the text. $h = 1/[2 \max(s, s')]$](image)
Fig. 3. Joint distribution of the genic values of parent and offspring.

\[
\begin{align*}
\int \int (u_{1x1}) \cdot (u_{2x2}) (x_1 - x_2) \\
\times \exp \{t_1 x_1 + t_2 x_2 + t_3 x_3 \} \, dx_1 \, dx_2 \, dx_3 \\
= \int \int (u_{1x1}) \cdot (u_{2x2}) (s) \exp \{t_1 x_1 + t_2 x_2 + t_3 x_3 \} \\
\times \exp \{t_3 x_3 \} \, ds \, dx_1 \, dx_2 \, dx_3 \\
= \int \int \int (u_{1x1}) \exp \{t_1 + t_3 x_3 \} \, dx_1 \\
\times \int \int \int (u_{2x2}) \exp \{t_2 + t_3 x_2 \} \, dx_2 \\
(A \text{ denoting Fourier transform}) \quad (\text{hence} \quad (u_{1x1}) \cdot (u_{2x2}) = u_{1x1} \cdot u_{2x2})
\end{align*}
\]

and the second, likewise, is \( \phi(-t_2) \phi(t_1 + t_3) \). Because \( \phi(-t) = \phi(t) \), the Fourier transform of \( g(x_2 - x_1 - x_2; x_1, x_2) \) is \( \theta(t, x_1, x_2) \). Consequently, \( g(x_3 - x_1 - x_2; x_1, x_2) \) is the density of the random variable \((G, G')\). It is easy to verify that \( g(x_3 - x_1 - x_2; x_1, x_2) \) is the Lebesgue measure of the intersection \([x_3 - x_1 - x_2; x_1, x_2] \cap [x_3 - x_2 - \frac{1}{2}; x_3 - x_2 + \frac{1}{2}] \cap [-x_2 - \frac{1}{2}; -x_2 + \frac{1}{2}] \); i.e., \( g(x_3 - x_1 - x_2; x_1, x_2) \) is \( 0.1 + \min(x_3 - x_1, x_3 - x_2, 0) \).

It follows that the conditional distribution of offspring genic value given the parents genic values is trapezoidal (Fig. 2), in which \( \mu = \frac{1}{2}(G + G') \) is the mean of the offspring distribution;

\[
\lambda = \frac{\sum a_n^4}{\left( \sum a_n^2 \right)^2}
\]

In particular, if the effects at successive loci decrease geometrically \((a_n = a^n, \lambda = (1 - a^2)/(1 + a^2))\), the results of these calculations are:

\[
E[r_i(V, D) - \lambda] = \frac{2}{n - 1} \left[ \frac{n^2 - n - 2}{n(n - 1)} \right]^{1/2} \left[ 4 - 2\lambda \right]^{1/2}
\]

and

\[
E[r_i(V, \overline{D})] = \frac{2}{n - 1} \left[ \frac{n^2 - n - 2}{n(n - 1)} \right]^{1/2} \left[ \frac{2(n + 1)^2}{n^2} \right] \left[ \frac{n^2 + 6n^2 - 3n + 4}{2n^3} \right]^{1/2}
\]

(Figs. 4 and 5). From a practical point of view, large sibships are needed to test these predictions.

A model of continuous variation based on genes with gradual effects may be useful in the analysis of common diseases without evident Mendelian patterns of inheritance. Gottesman and Shields have proposed that such a model may account for the transmission of schizophrenia: "Some high value genes may

\[ r_n \text{ between the within-sibship variance for an n-sib family,} \]

\[ V = \frac{1}{n - 1} \sum (G_i - \overline{G})^2, \]

in which \( G_i \) are the genic values of the sibs and \( \overline{G} \) their mean, and either the mean-square parental deviation

\[ D = \frac{S}{n} \]

or the square of the within-sibship mean \( (\overline{G})^2 \) itself. A gene frequency of \( \frac{1}{2} \) is assumed at each locus, but the gene effects \( a_n \) are allowed to be arbitrary. Both correlations depend on only two parameters: the sibship size \( n \) and a measure of the rate of decrease of gene effects at successive loci.
be rather specific for some partitionable aspects of the disorder (e.g., catatonia, paranoid features, or some aspect of blood chemistry, brain protein, or neurophysiology); such possibilities would then encourage both formal segregation analyses and searches for linkage relationships to known genetic markers, but for facets of the syndrome." (ref. 15, p. 520).

S.M. was supported by Research Scientist Development Award, U.S. Public Health Service Grant K2-MH-00108, and U.S. Public Health Service Grants MH-16674 and MH-25515. K.L. was supported by National Institutes of Health Special Resource Grant RR-3 and The National Foundation–March of Dimes, Grant 6-55.