Genética de Poblaciones Humanas

PAULO ALBERTO OTTO

EDITORIAL UNIVERSITARIA
UNIVERSIDAD NACIONAL DE MISIONES

HUMAN POPULATION GENETICS

(GENÉTICA DE POBLACIONES HUMANAS)

PAULO A. OTTO

Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo
Caixa Postal 11461
05422-970 São Paulo SP

Curso Teórico Práctico de Post-Grado
8 al 14 de Septiembre de 2006
Departamento de Genética
Laboratorio de Citogenética y Genética Humana
Facultad de Ciencias Exactas Químicas y Naturales
Universidad Nacional de Misiones
Posadas, Misiones, República Argentina

I - Teoría

EDITORIAL UNIVERSITARIA DE MISIONES

San Luis 1870
Posadas - Misiones – Tel-Fax: (03752) 428601

Correos electrónicos:
edunam-admini@arnet.com.ar
edunam-direccion@arnet.com.ar
edunam-produccion@arnet.com.ar
edunam-ventas@arnet.com.ar
edunam-prensa@arnet.com.ar

Otto, Paulo Alberto
Genética de poblaciones humanas. - 1a ed. - Posadas : EdUNaM - Editorial Universitaria de la Universidad Nacional de Misiones, 2008.
209 p.; 28x22 cm.

ISBN 978-950-579-113-2

1. Genética de Poblaciones. 2. Genética Humana. I. Título
CDD 616.042

Fecha de catalogación: 15/10/08

ISBN: 978-950-579-113-2
Impreso en Argentina

HARDY-WEINBERG EQUILIBRIUM

Let us consider a population of infinite size, consisting of diploid, sexually-reproducing individuals. In relation to a given autosomal locus where 2 alleles (A and a) are segregating, these individuals will belong to the genotypic classes AA, Aa and aa. Let us suppose that in a given generation the frequencies of these three genotypes, among individuals of both sexes, are d, h, and r respectively and that all matings occur entirely at random. Under this assumption, the probabilities of any individual of the population choosing a mate that is AA, Aa or aa are respectively d, h and r. Since d, h and r are also the probabibilities of the first individual being AA, Aa or aa, the various matings occurring in the population will be given by the cross-products shown in the matrix below:

		males		
		AA	Aa	aa
females	AA	d^2	dh	dr
	Aa	dh	h^2	hr
	aa	dr	hr	r^2

If the generations are discrete and the effects of selection, mutation and migration are considered negligible, that is, each mating pair contributes on average to the next generation with the same offspring number as all other couples, no gene A is transformed by mutation into a and vice-versa, and there is no exchange of genes among individuals belonging to this population and individuals from other populational aggregates, then we obtain the following results:

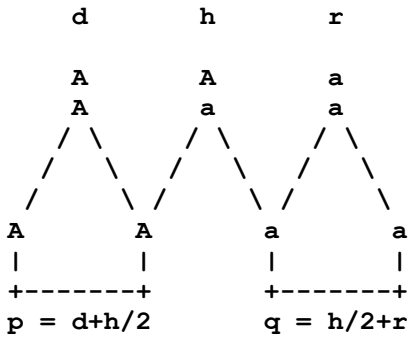
cross. (gen. n)			offspr. genot. frequencies (gen. n+1)		
mal.	fem.	frequencies	AA	Aa	aa
AA	AA	d^2	d^2	0	0
AA	Aa	dh	dh/2	dh/2	0
AA	aa	dr	0	dr	0
Aa	AA	dh	dh/2	dh/2	0
Aa	Aa	h^2	$h^2/4$	$h^2/2$	$h^2/4$
Aa	aa	hr	0	hr/2	hr/2
aa	AA	dr	0	dr	0
aa	Aa	hr	0	hr/2	hr/2
aa	aa	r^2	0	0	r^2

Since the probabilities of a given progeny do not depend upon the sex of its parents (for example, the expected proportions of AA and Aa progeny from crossings AA_m x Aa_f and AA_f x Aa_m are exactly the same), the table above can be simplified to:

cross. (gen. n)		frequencies	offspr. genot. frequencies (gen. n+1) AA	Aa	aa
AA	AA	d^2	d^2	0	0
AA	Aa	2dh	dh	dh	0
AA	aa	2dr	0	2dr	0
Aa	Aa	h^2	$h^2/4$	$h^2/2$	$h^2/4$
Aa	aa	2hr	0	hr	hr
aa	aa	r^2	0	0	r^2

Therefore, the frequency of AA individuals in the following generation is $d^2 + dh + h^2/4 = (d+h/2)^2$,
that of Aa is $dh + 2dr + hr + h^2/2 = 2(d+h/2)(h/2+r)$
and that of aa individuals is $h^2/4 + hr + r^2 = (h/2+r)^2$.

The quantities d+h/2 and h/2+r are respectively the frequencies of the alleles A and a, since each AA individual is represented by two A genes, each heterozygote by one A and one a and each aa homozygote is represented by two a genes:



In fact, if the numbers (or absolute frequencies) of genotypes AA, Aa and aa are D, H and R respectively (a mnemonics to dominant, heterozygote and recessive respectively, in spite of a not being necessarily recessive in relation to A), the numbers of A and a genes are respectively N(A) = 2D + H and N(a) = H + 2R, since each homozygote carries two identical copies of the same gene and a heterozygote has one copy of each allele. Since there are (2D + H) + (H + 2R) = 2D + 2H + 2R = 2N(A) + 2N(a) = 2N genes in the population, the frequencies of the two alleles are given respectively by

P(A) = (2D + H)/2N = 2D/2N + H/2N = D/N + ½ H/N = d + h/2 = p and

P(a) = (H + 2R)/2N = H/2N + 2R/2n = ½ H/N + R/N = h/2 + r = q .

Therefore, if we have a population of infinite size where the frequencies of genotypes AA, Aa and aa are d, h and r respectively and if matings occur at random ('panmixia'), individuals with genotypes AA, Aa and aa will occur after the proportions p^2, 2pq and q^2, where p =

d+h/2 and q = 1-p = h/2+r are the frequencies of the A gene and its allele a in the parental generation.

Obviously after one more generation of random matings the population will still present the same genotypic ratios p^2 : 2pq : q^2, as the following table shows :

cross.(gen. n+1)		frequencies	offspr. genot. frequencies (gen. n+2) AA	Aa	aa
AA	AA	p^4	p^4	0	0
AA	Aa	$4p^3q$	$2p^3q$	$2p^3q$	0
AA	aa	$2p^2q^2$	0	$2p^2q^2$	0
Aa	Aa	$4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
Aa	aa	$4pq^3$	0	$2pq^3$	$2pq^3$
aa	aa	q^4	0	0	q^4

The frequencies of AA, Aa and aa individuals in the generation n+2 are therefore

$$P(AA) = p^4 + 2p^3q + p^2q^2 = p^2(p^2 + 2pq + q^2) = p^2$$

$$P(Aa) = 2p^3q + 4p^2q^2 + 2pq^3 = 2pq(p^2 + 2pq + q^2) = 2pq$$

$$P(aa) = p^2q^2 + 2pq^3 + q^4 = q^2(p^2 + 2pq + q^2) = q^2 .$$

The main conclusion from the analyses shown above is that after one single generation of panmixia, the genotypic frequencies P(AA), P(Aa) and P(aa) are in the ratios p^2, 2pq and q^2, where p and q are the frequencies of a mutually exclusive pair of alleles segregating at an autosomal locus in a breeding population of infinite size. This is the principle, theorem or law of Hardy - Weinberg, named after the two authors who described it quite independently in 1908.

The Hardy-Weinberg principle can be demonstrated straightforwardly using the following argument: the individuals born to random mating pairs result obviously from fertilizations that occur also randomly among gametes produced by male and female individuals from the parental generation. Since the allelic pair under consideration is an autosomal one, among males as well as females from the population genotypes AA, Aa and aa are in the same ratios d, h and r; and males and females will produce gametes A and a in the ratios p = d+h/2 : q = h/2+r respectively. Random union of these gametes result in the offspring genotypes, AA, Aa and aa, that will occur in the ratios p^2 : 2pq : q^2 respectively:

		feminine gametes	
		A	a
		p	q
	A	AA	Aa
masculine	p	p^2	pq
gametes	a	Aa	aa
	q	pq	q^2

Since individuals AA, Aa and aa are now in the ratios p^2, 2pq and q^2, it comes out that the gametes A and a produced by males as well as females from this generation will occur respectively in the frequencies $p^2 + 2pq/2 = p(p+q) = p$

$2pq/2 + q^2 = q(p+q) = q;$

algebraically, all the above is equivalent to the binomial expansion

$[(p^2+pq)+(pq+q^2)]^2 = (p+q)^2 = p^2 + 2pq + q^2.$

Of course a population with Hardy-Weinberg equilibrium has a genotypic distribution $p^2 : 2pq : q^2$, but the inverse is not true (Stark, personal communication, 1983; Li, 1988): it is possible to show that some populations with no panmixia at all have the marginal genotypic distribution $p^2 : 2pq : q^2$.

The evolutionary importance of this simple principle is obvious: in the absence of factors such as mutation, random genetic drift and migration, there exists at the population level a static force that tends to keep genotypic ratios in the proportions $p^2 : 2pq : q^2$, therefore maintaining the population variability throughout time.

The table below (generated by the BASIC code that follows) shows the frequencies of AA, Aa and aa genotypes as functions of the frequency p of the A allele (or q of the a allele).

p	q	P(AA) = p^2	P(Aa) = 2pq	P(aa) = q^2
0.0000	1.0000	0.0000	0.0000	1.0000
0.0500	0.9500	0.0025	0.0950	0.9025
0.1000	0.9000	0.0100	0.1800	0.8100
0.1500	0.8500	0.0225	0.2550	0.7225
0.2000	0.8000	0.0400	0.3200	0.6400
0.2500	0.7500	0.0625	0.3750	0.5625
0.3000	0.7000	0.0900	0.4200	0.4900
0.3500	0.6500	0.1225	0.4550	0.4225
0.4000	0.6000	0.1600	0.4800	0.3600
0.4500	0.5500	0.2025	0.4950	0.3025
0.5000	0.5000	0.2500	0.5000	0.2500
0.5500	0.4500	0.3025	0.4950	0.2025
0.6000	0.4000	0.3600	0.4800	0.1600
0.6500	0.3500	0.4225	0.4550	0.1225
0.7000	0.3000	0.4900	0.4200	0.0900
0.7500	0.2500	0.5625	0.3750	0.0625
0.8000	0.2000	0.6400	0.3200	0.0400
0.8500	0.1500	0.7225	0.2550	0.0225
0.9000	0.1000	0.8100	0.1800	0.0100
0.9500	0.0500	0.9025	0.0950	0.0025
1.0000	0.0000	1.0000	0.0000	0.0000

```
REM PROGRAM FILENAME HARDYWE1.BAS
CLS : DEFDBL A-Z
PRINT "  p            q         P(AA) = p^2  P(Aa) = 2pq  P(aa) = q^2"
PRINT "-------------------------------------------------------------------"
FOR I = 0 TO 20: P = I / 20
PRINT USING " #.####      "; P; 1 - P; P ^ 2; 2 * P * (1 - P); (1 - P) ^ 2
NEXT I
PRINT "-------------------------------------------------------------------"
DO: LOOP WHILE INKEY$ <> " "
```

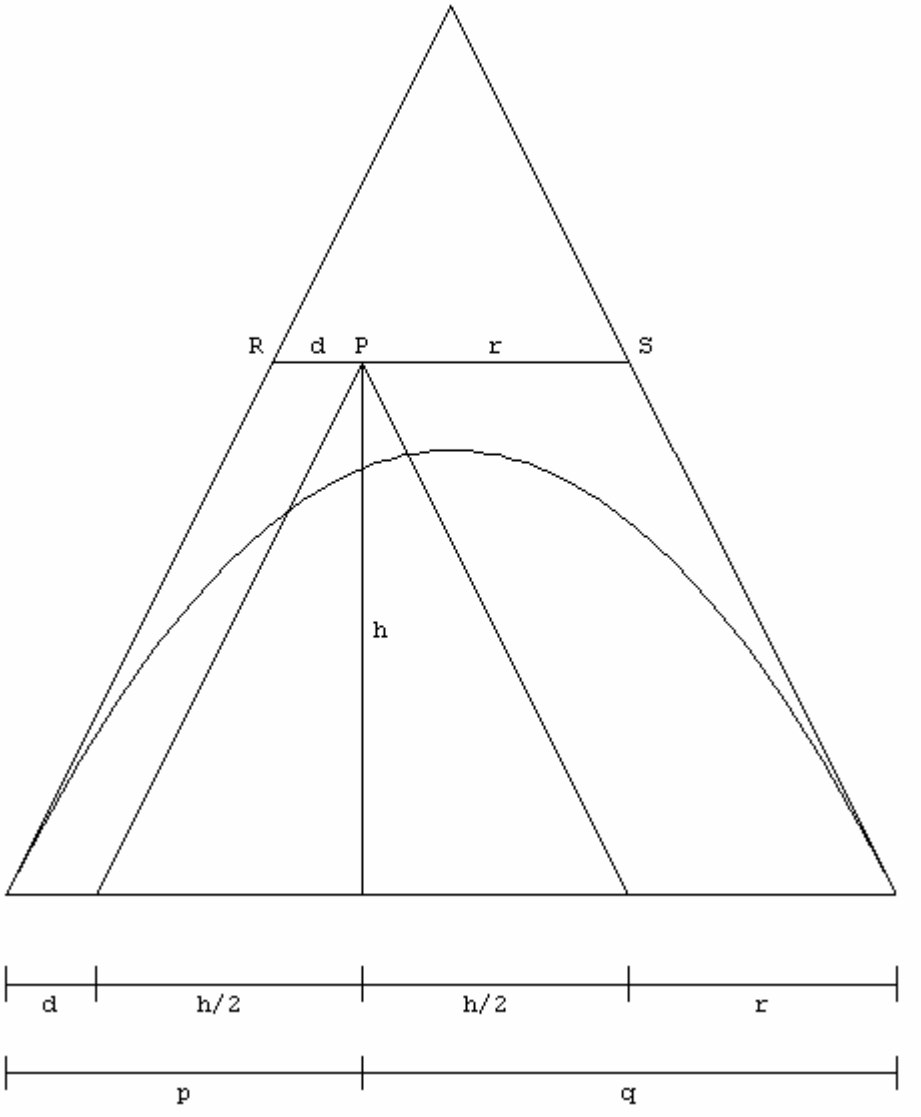
One important property of panmictic populations is that $h^2 = 4dr$. In fact, in these populations, $P(AA) = p^2$, $P(Aa) = 2pq$, $P(aa) = q^2$ and therefore $h^2 = (2pq)^2 = 4p^2q^2 = 4dr = 4.p^2.q^2$.

Another property is that the maximum possible frequency of heterozygotes is 0.5. In fact, if we differentiate $2pq = 2p(1-p) = 2p-2p^2$ in relation to the argument p, we obtain $d[2p(1-p)]/dp = 2-4p$; equating this result to zero, we obtain $2-4p = 0$ and hence $p = 2/4 = 0.5$. This is therefore the value of p that maximizes the function $f(p) = 2p(1-p)$; and for $p = 0.5$ the value of $f(p)$ is also 0.5.

This last property is intuitive: since for $0 < p < 1$ $h = f(p) = 2p(1-p)$ has equal values for complementary values of p adding up to unity and the value of the function is zero for $p = 0$ or $p = 1$, its maximum value takes place when $p = 1-p = 0.5$. And for this value of p the function $2p(1-p)$ has value 0.5. Also, going back to the gamete model we used to demonstrate Hardy-Weinberg equilibrium, it is obvious that the probability of drawing two different gametes (one A from the masculine pool and one a from the feminine one or vice-versa) is at a maximum when the two types of gametes occur within the respective gamete or gene pools with exactly equal frequencies. Therefore, the maximum frequency of heterozygotes in panmictic populations cannot exceed 0.5 or 50%. For

instance, inspecting the sample {AA : 100; Aa : 695; aa : 205} we can assure that the genotypic frquencies are not in Hardy-Weinberg ratios without making any statistical tests, since 695/1000 = 0.695 > 0.5 and this cannot be ascribed to chance fluctuations in a sample of this magnitude.

For the graphical representation of genotypic frequencies one commonly uses a system of triangular coordinates. A very simple system is the isosceles triangle coordinate system (Otto & Benedetti, J. Heredity 1995), the use of which is shown below for the case of a population point P with coordinates d = P(AA) = 0.10, h = P(Aa) = 0.70 and r = P(aa) = 0.20. The perpendicular distance h inside the isosceles triangle of unitary height and basis divides the latter in 2 segments in the proportions p : q, with p + q = 1. This constitutes a clear advantage in relation to the classical representations (Cartesian and equilateral [homogeneous] coordinate systems). Also shown inside the triangle is the Hardy - Weinberg or De Finetti parabola, which represents the set of population points such that d : h : r :: p^2 : 2pq : q^2 .



The figure above was generated by the following Mathematica code:

```
(* TRICOOR2.MA
   Isosceles triang. repres. of genotype freq. *)
Show[
    Plot[2*x*(1 - x), {x, 0, 1},
        Axes -> None,
        DisplayFunction -> Identity],
    Graphics[{
        Line[{{0, 0}, {.5, 1}}],
        Line[{{0, 0}, {1, 0}}],
        Line[{{.5, 1}, {1, 0}}],
        Line[{{0, -.1}, {1, -.1}}],
        Line[{{0, -.08}, {0, -.12}}],
        Line[{{1, -.08}, {1, -.12}}],
        Line[{{.4, -.08}, {.4, -.12}}],
        Line[{{.1, -.08}, {.1, -.12}}],
        Line[{{.7, -.08}, {.7, -.12}}],
        Line[{{0, -.2}, {1, -.2}}],
        Line[{{0, -.18}, {0, -.22}}],
        Line[{{1, -.18}, {1, -.22}}],
        Line[{{.4, -.18}, {.4, -.22}}],
        Line[{{.3, .6}, {.7, .6}}],
        Line[{{.4, .6}, {.4, 0}}],
        Line[{{.4, .6}, {.1, 0}}],
        Line[{{.4, .6}, {.7, 0}}],
        Text["P",{.4, .62}],
        Text["R",{.28, .62}],
        Text["S",{.72, .62}],
        Text["h",{.42, .3}],
        Text["d",{.35, .62}],
        Text["r",{.55, .62}],
        Text["d",{.05, -.12}],
        Text["h/2",{.24, -.12}],
        Text["h/2",{.54, -.12}],
        Text["r",{.85, -.12}],
        Text["p",{.2, -.22}],
        Text["q",{.7, -.22}],
        }],
        DisplayFunction -> $DisplayFunction,
        AspectRatio -> Automatic];
```

For the case of hereditary characteristics determined by autosomal codominant alleles it is possible to test whether the sample drawn from a population is consistent with the Hardy-Weinberg proportions (why authors insist so much on this is a quite different and mysterious problem). This is accomplished using the chi-squared test, the use of which in a real situation is shown below.

In a sample of 230 negroid, unrelated individuals from the city of Rio de Janeiro, Fragoso & Otto (Rev. Med. Est. Guanab. 34 : 59-62 , 1967) determined the haptoglobin types and found the following results:

phenotypes	abs. frequencies
Hp(1-1)	63
Hp(2-1)	117
Hp(2-2)	50

The three phenotypes detected through electrophoresis correspond respectively to the genotypes Hp^1/Hp^1, Hp^1/Hp^2 and Hp^2/Hp^2 determined by the combinations of the two autosomal codominant alleles Hp^1 and Hp^2.

The frequencies of the two alleles in the sample are estimated by direct counting. In fact, the sample above, consisting of 230 individuals, is equivalent to a sample of 2x230 = 460 genes. Since each Hp^1/Hp^1 individual carries two Hp1 genes, each heterozygote carries one Hp^1 gene and one Hp2 gene, and each Hp^2/Hp^2 individual carries two Hp^2 genes, the total number of Hp1 genes in the sample is simply $N(Hp^1)$ = 2x63+117 = 243; and $N(Hp^2)$ is equal to 117 + 2x50 = 217. Therefore the estimate of $p = P(Hp^1)$ is $N(Hp^1)/[N(Hp^1)+N(Hp^2)]$= 243/460= 0.528; the estimate of $q = P(Hp^2)$ is $N(Hp^2)/[N(Hp^1)+N(Hp^2)]$ = 217/460 = 1-p = 0.472.

Of course we cannot know the true frequency of the allele p in the population from which the above sample was drawn. This is not possible even with the sampling of the whole population, that is changing dynamically with time and has its exact genotypic composition submitted to small chance fluctuations varying with time. That is the reason why it is important to calculate the statistical error (standard error) of the estimate p (or q), and that is given in the case of autosomal codominant alleles by the simple formula $s.e.(p) = s.e.(q) = \sqrt{(pq/2N)}$, where N, as before, is the number of sampled individuals. In the above example, s.e.(p) has value 0.023. Since the binomial estimates obtained from samples of the same population will be normally distributed with mean p = 0.528 and s.e. = 0.023, we know, for instance, that the 95% confidence interval of p is given by $0.528 \pm 1.96 \times 0.023$, with limits therefore of 0.483 and 0.573, which permits us to say that the true value of the gene frequency lies between 0.483 and 0.573 with a probability of approximately 95% (that is, we know now that the error we are making when we state this is approximately 5%). More formally, this means that if we take a large number of samples of same size N from the same population, 95% of the confidence intervals thus constructed (i.e., using the parameters obtained from each sample) will contain the true gene frequency.

The expected absolute frequencies E(11), E(12) and E(22) of Hp(1-1), Hp(2-1) and Hp(2-2) phenotypes under the hypothesis of Hardy-Weinberg equilibrium are:

$E(11) = 230xp^2 = 64.18$
$E(12) = 230x2pq = 114.63$
$E(22) = 230xq^2 = 51.18$.

Then we contrast these expectations with the observed quantities O(11) = 63, O(12) = 117 and O(22) = 50 using the usual chi-squared statistics:

	Hp(1-1)	Hp(2-1)	Hp(2-2)	total
obs. abs. freq. O_i	63	117	50	230
obs. exp. freq. E_i	64.18	114.63	51.18	230
$(O_i-E_i)^2/E_i$	0.022	0.049	0.027	0.098

The formula we just used for obtaining the χ^2 figure is important because through it we can inspect the individual contributions for the total value of the chi-squared statistics and locate the class responsible for the largest deviation contributing to final figure of the statistics. In the case we are not interested in this, the formula above can be simplified to $\chi^2 = \Sigma(O_i-E_i)^2/E_i = \Sigma(O_i^2-2O_iE_i+E_i^2)/E_i = \Sigma(O_i^2/E_i) - 2\Sigma O_i + \Sigma E_i = \Sigma(O_i^2/E_i) - N$, since $\Sigma O_i = \Sigma E_i = N$. This simplified formula is often used in computer programs for calculating the value of the statistics and avoids rounding errors generated by the complete formula $\chi^2 = \Sigma(O_i-E_i)^2/E_i$.

For one degree of freedom (d.f), the chi-squared figure of 0.098 corresponds to a probability between 0.75 and 0.90 favoring the hypothesis just tested. Hence we conclude that the collected data are in accordance with Hardy-Weinberg proportions.

The chi-squared test just performed has 1 d.f. because in order to calculate the expected quantities E(11), E(12) and E(22) necessary to perform the test we used 2 sample parameters: the total number 230 and one gene frequency (p or q). If one is not satisfied with this formal definition of degrees of freedom of a chi-squared statistics for testing Hardy-Weinberg equilibrium, we can show the following: since p and N are used for obtaining the expected values, any single expected value we determine fixes automatically the values of the other two. For instance, if we calculate E(11) as being $Np^2 = 64.18$, the expected number of heterozygotes x is given by $E(12) = 2 \times (64.18 - Np) = 114.63$, because p = frequency of homozygotes + frequency of heterozygotes/2.

In the case of two autosomal codominant alleles (A, a) the usual formula for obtaining the chi-squared value, $\chi^2 = \Sigma[(O_i-E_i)^2/E_i]$, can be simplified using the following algebraic acrobatics.

For the two-allele case the expected numbers of AA , Aa and aa individuals, under the null hypothesis of Hardy-Weinberg equilibrium, are $N(AA) = Np^2$, $N(Aa) = 2Npq$, and $N(aa) = Nq^2$, where p and q are the sample estimates of the frequencies of the gene A and its allele a; these estimates, which actually coincide with the ones obtained through the maximum likelihood method, are obtained by simply counting the total genes of the respective types and then by expressing the counts as the proportions of the total of 2N genes counted: $p = (2D+H)/2N$ and $q = (H+2R)/2N$, where D, H, and R are the numbers of AA, Aa and aa individuals observed among the N sampled ones. Therefore we have:

$$\text{CHI-SQUARED (1 d.f.)} = \Sigma[(O_i-E_i)^2/E_i] = \Sigma(O_i^2/E_i) - N =$$
$$= D^2/Np^2 + H^2/2Npq + R^2/Nq^2 - N =$$
$$= [4ND^2(H+2R)^2+2NH^2(2D+H)(H+2R)+4NR^2(2D+R)^2 -$$
$$- N(2D+H)^2(H+2R)^2]/[(2D+H)^2(H+2R)^2] =$$
$$= N(H^4-8DH^2R+16D^2R^2)/[(2D+H)^2(H+2R)^2] =$$
$$= N\{(H^2-4DR)/[(2D+H)(H+2R)]\}^2 .$$

For D = 63, H = 117, R = 50 and N = 230 (numerical example worked above),

$$\text{CHI-SQUARED (1 d.f.)} = 230x[(13689-12600)/(243x217)]^2 =$$
$$= 230x1185921/2780558361 =$$
$$= 272761830/2780558361 =$$
$$= 0.098 .$$

Hardy-Weinberg law can be generalized in almost all its properties to a series of any number of alleles segregating at an autosomal locus:

$$(p + q + ... + z)^2 = p^2 + 2pq + q^2 + ... + z^2 .$$

For example, let a hypothetical hereditary characteristic be determined by three autosomal alleles A, B, and C. If the frequencies of genotypes AA, AB, AC, BB, BC and CC are respectively a, b, c, d, e and f at generation 0, then the allele frequencies P(A), P(B) and P(C) are given respectively by

$$p = (2a+b+c)/2 , q = (b+2d+e)/2 \text{ and } r = (c+e+2f)/2 .$$

Under the assumption of random matings, the individuals belonging to the next generation will occur in the frequencies

genotypes	frequencies
AA	p^2
AB	$2pq$
AC	$2pr$
BB	q^2
BC	$2qr$
CC	r^2

and the allele frequencies in this population continue to be

$$p^2 + 2pq/2 + 2pr/2 = p(p+q+r) = p$$
$$2pq/2 + q^2 + 2qr/2 = q(p+q+r) = q$$
$$2pr/2 + 2qr/2 + r^2 = r(p+q+r) = r .$$

If we denote by p_i and p_j the frequencies of any two alleles segregating at an autosomal locus, it comes out that the frequency of any genotype, under the assumption of panmixia, is given by

$P(A_iA_j) = (2-\delta_{ij}).p_i.p_j$, where δ_{ij} (Kronecker's delta) is an operator with the property $\delta_{ij} = 1$ if i=j , $\delta_{ij} = 0$ otherwise. Therefore,

$$P(A_iA_i) = (2-1).p_i.p_i = p_i^2$$
$$P(A_iA_j) = (2-0).p_i.p_j = 2p_ip_j.$$

As we commented before, it is intuitive that the chance of a heterozygous individual being produced in a panmictic population is at a maximum when gene frequencies are equal for all the n alleles segregating at an autosomal locus. If there exist n alleles, then under this assumption the frequency of each allele is obviously $p_i = 1/n$ and the frequency of each type of heterozygote is $P(a_i a_j) = 2p_i p_j = 2.1/n.1/n = 2/n^2$. When the number of alleles is n, there exist n(n-1)/2 different types of heterozygotes, and the maximum possible frequency of heterozygotes in a panmictic population is $2/n^2$ x n(n-1)/2 = (n-1)/n. The table below shows the values this frequency takes when n = 2, 3, ..., inf.:

n	1/n	$2/n^2$	n(n-1)/2	(n-1)/n
2	1/2	1/2	1	1/2
3	1/3	2/9	3	2/3
4	1/4	2/16	6	3/4
5	1/5	2/25	10	4/5
...	...	...	...	...
inf.	0	0	inf.	1

It is easy to infere that as the number of alleles increases within a given locus the proportion of heterozygotes in the population also increases.

If in the initial (0) generation a same allele has different frequencies among males and females, in the next generation, under panmixia, males and females will have the same gene frequency, and this has as value the arithmetic mean between parental gene frequencies, since males and females contribute equally to their offspring. In fact, if p' is the allele frequency among males and p" among females at generation 0, it comes out that in the first generation the genotypic distribution among males as well as females will be

P(AA) = p'.p"
P(Aa) = p'.q" + p".q' = p'(1-p")+p"(1-p') = p'+p"-2p'.p"
P(aa) = q'.q" = (1-p').(1-p") = 1-p'-p"+p'.p" ;

since p' ≠ p", then it comes out that P(AA) ≠ p^2, P(Aa) ≠ 2pq and P(aa) ≠ q^2.

Gene frequencies in this first generation are determined as usually:

P(A) = P(AA) + P(Aa)/2 = p'.p" + (p'+p")/2 - p'.p" = (p'+p")/2
P(a) = = (q'+q")/2 .

Therefore we can conclude that different allele frequencies among males and females determine a delay of one generation in the approach of Hardy-Weinberg equilibrium. This property is important to derive the approach to equilibrium in the case of sex-linked genes that we discuss in the lines below.

Let f_n and m_n be the frequencies of a same allele a from the X chromosome among females and males respectively, in a generic generation n. Under the assumption of panmixia, the following recurrence relations are obtained:

(1) $f_{n+1} = (m_n+f_n)/2$
(2) $m_{n+1} = f_n$.

Equation (1) results from the fact that each female receives one X chromosome from her mother and the other from her father. Equation (2) means that the only X chromosome present in males derive from their mothers.

From (1) and (2) we obtain also

$$(3)\quad f_{n+1} - m_{n+1} = (m_n + f_n)/2 - f_n = (f_n - m_n).(-1/2) = (f_n - m_n).r \ , \ r = -1/2 \ .$$

This last equation has the general solution

$$f_n - m_n = (f_0 - m_0).r^n = (f_0 - m_0).(-1/2)^n \ ,$$

which shows that each generation of panmixia halves the absolute value of the initial difference $f_0 - m_0$. Of course when n tends to infinity this difference tends to zero, so that at equilibrium gene frequencies will be the same among females and males: f = m = q.

Equation $f_n - m_n = (f_0 - m_0).r^n$ can be rewritten as

$$f_n = m_n + (f_0 - m_0).r^n = f_{n-1} + (f_0 - m_0).r^n \ .$$

It is easy to verify that

$$f_1 = f_0 + (f_0 - m_0).r$$
$$f_2 = f_1 + (f_0 - m_0).r^2 = f_0 + (f_0-m_0).r + (f_0-m_0).r^2$$
$$f_3 = f_2 + (f_0 - m0).r^3 = f_0 + (f_0-m_0).r + (f_0-m_0).r^2 + (f_0-m_0).r^3$$

and so on. Therefore,

$$f_n = f_0 + (f_0 - m_0).(r^1 + r^2 + r^3 + \ldots + r^n).$$

In the expression above, $r^1 + r^2 + r^3 + \ldots + r^n$ is the sum of the terms of a geometric series with ratio r = -1/2, the solution of which is given by the formula

$$r^1 + \ldots + r^n = (r-r^{n+1})/(1-r) = r(1-r^n)/(3/2) = 2r(1-r^n)/3 = -(1-r^n)/3 \ .$$

Therefore, the general solution of fn is given by

$$f_n = f_0 - (f_0-m_0).(1-r^n)/3 = f_0 - (f_0-m_0)/3 + (f_0-m_0).r^n/3 = (2f_0 + m_0)/3 + (f_0-m_0).(-1/2)^n/3 \ .$$

The limit of this expression, as n tends to infinity, is clearly

$$q = f = m = (2f_0 + m_0)/3.$$

The quantity above is a constant quantity :

$(2f_{n+1} + m_{n+1})/2 = q_{n+1} = [2(m_n+f_n)/2 + f_n]/2 = (2f_n+m_n)/2 = q_n = ... = q$,

representing the average (weighed) gene frequency in the whole population, in any generation. In fact, since 1/3 of all X chromosomes are in males and 2/3 in females, given that in the population there exist equal numbers of males and females, the average (weighed) frequency of the allele in the whole population is

$q_n = 2/3.f_n + 1/3.m_n = q = f = m$,

and this quantity must be a constant given the assumptions of absence of selection, mutation and differential migration.

The above results can be obtained straightforwardly, using the following more formal procedure:

Writing the recurrence equations $f_1 = (f_0+m_0)/2$ and $m_1 = f_0$ in matrix compressed form

$$\begin{pmatrix} f_1 \\ m_1 \end{pmatrix} = WQ_0 = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} f_0 \\ m_0 \end{pmatrix} = RW_dR^{-1}Q_0 = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1/2 \end{pmatrix}\begin{pmatrix} 4/3 & 2/3 \\ 2/3 & -2/3 \end{pmatrix}\begin{pmatrix} f_0 \\ m_0 \end{pmatrix}$$

the general solution $Q_n = RW_d^nR^{-1}Q_0$ is obtained immediately, from which we get

$f_n = (2f_0 + m_0)/3 + (f_0-m_0).(-1/2)^n/3$ and

$m_n = (2f_0 + m_0)/3 - 2(f_0-m_0).(-1/2)^n/3$.

As before, the limit of both expressions, as n tends to infinity, is clearly $q = f = m = (2f_0 + m_0)/3$.

The equilibrium condition for a sex-linked locus is that all its alleles have the same frequencies in males and females. This takes place asymptotically, in an oscillatory manner, since $m_{n+1}-f_{n+1} = -(m_n-f_n)/2$. At equilibrium genotypes are distributed after

genotypes	frequencies
Ay	p
ay	q
AA	p^2
Aa	2pq
aa	q^2

,

that is, the male genotypes (hemizygotes A and a) occur in gene frequencies while the female genotypes AA, Aa and aa follow a typical H-W distribution $p^2 : 2pq : q^2$.

As a numerical example to appreciate the approach to equilibrium, let us consider the following initial population :

$P_0(Ay) = 1.00$
$P_0(ay) = 0.00$
$P_0(AA) = 0.00$
$P_0(Aa) = 0.00$
$P_0(aa) = 1.00$.

From the data above, it comes out that the initial frequencies of the a gene in males and females are respectively $m_0 = 0$ and $f_0 = 1$. Under panmixia, the genotypes in the following generation will occur in the frequencies

$P_1(Ay) = 1-f_0 = 0.00$
$P_1(ay) = f_0 = 1.00$
$P_1(AA) = (1-m_0).(1-f_0) = 0.00$
$P_1(Aa) = (1-m_0).f_0 + m_0.(1-f_0) = 1.00$
$P_1(aa) = m_0.f_0 = 0.00$;

in this first generation gene frequencies are

$m_1 = f_0 = 1.00$ and $f_1 = (m_0+f_0)/2 = 0.5$.

Applying recursively the equations

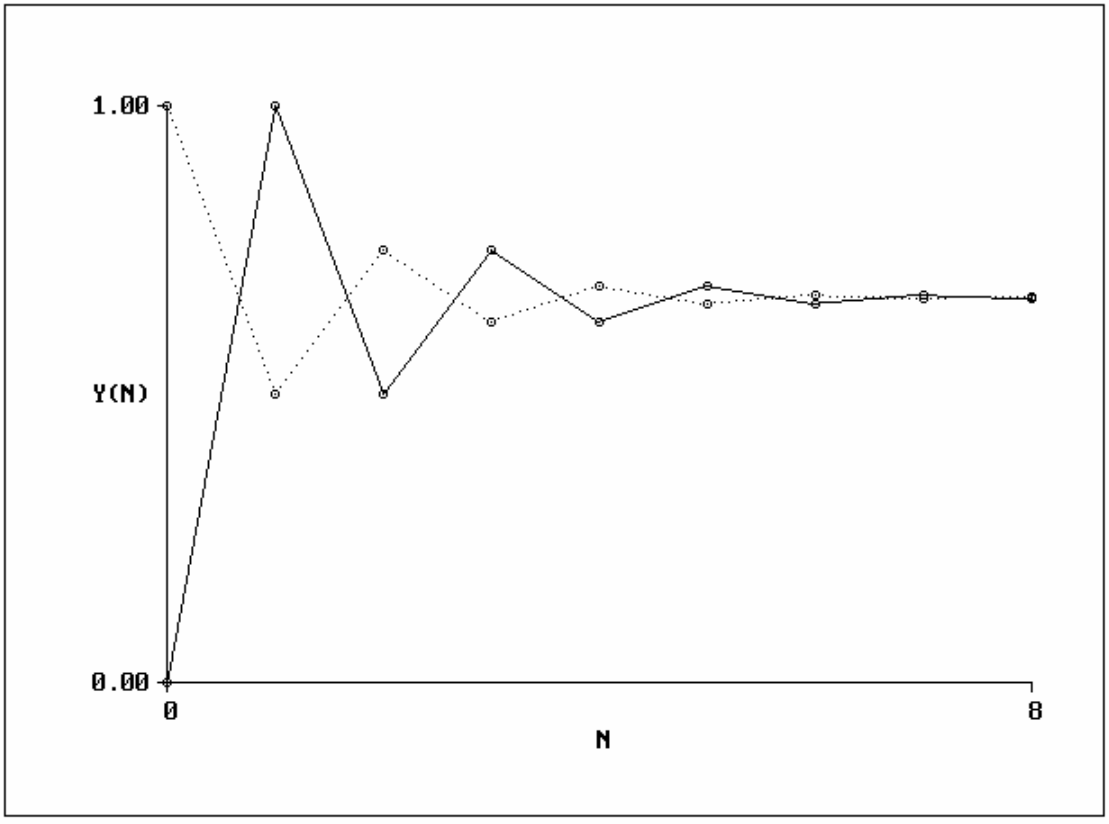
$m_n = P_n(ay)$
$f_n = P_n(Aa)/2 + P_n(aa)$
$P_{n+1}(Ay) = 1-f_n$
$P_{n+1}(ay) = f_n$
$P_{n+1}(AA) = (1-m_n).(1-f_n)$
$P_{n+1}(Aa) = (1-m_n).f_n + m_n.(1-f_n)$
$P_{n+1}(aa) = m_n.f_n$

the values corresponding to other generations are obtained and shown in the table below (followed by the respective BASIC code that generated it):

n	mn	fn	|mn-fn|	Pn(Ay)	Pn(ay)	Pn(AA)	Pn(Aa)	Pn(aa)
0	0.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
1	1.0000	0.5000	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000
2	0.5000	0.7500	0.2500	0.5000	0.5000	0.0000	0.5000	0.5000
3	0.7500	0.6250	0.1250	0.2500	0.7500	0.1250	0.5000	0.3750
4	0.6250	0.6875	0.0625	0.3750	0.6250	0.0938	0.4375	0.4688
5	0.6875	0.6563	0.0313	0.3125	0.6875	0.1172	0.4531	0.4297
6	0.6563	0.6719	0.0156	0.3438	0.6563	0.1074	0.4414	0.4512
7	0.6719	0.6641	0.0078	0.3281	0.6719	0.1128	0.4463	0.4409
8	0.6641	0.6680	0.0039	0.3359	0.6641	0.1102	0.4436	0.4462
9	0.6680	0.6660	0.0020	0.3320	0.6680	0.1115	0.4449	0.4436
10	0.6660	0.6670	0.0010	0.3340	0.6660	0.1109	0.4442	0.4449
11	0.6670	0.6665	0.0005	0.3330	0.6670	0.1112	0.4446	0.4442
12	0.6665	0.6667	0.0002	0.3335	0.6665	0.1111	0.4444	0.4446
13	0.6667	0.6666	0.0001	0.3333	0.6667	0.1111	0.4445	0.4444
14	0.6666	0.6667	0.0001	0.3334	0.6666	0.1111	0.4444	0.4445
15	0.6667	0.6667	0.0000	0.3333	0.6667	0.1111	0.4445	0.4444
16	0.6667	0.6667	0.0000	0.3333	0.6667	0.1111	0.4444	0.4445

```
REM PROGRAM FILENAME HWSEXL01.BAS
DEFDBL A-Z: DEFINT I
INPUT "P0(Ay) = "; PAY
INPUT "P0(ay) = "; PBY
INPUT "P0(AA) = "; PAA
INPUT "P0(Aa) = "; PAB
INPUT "P0(aa) = "; PBB
M = PBY: F = PAB / 2 + PBB
PRINT "-------------------------------------------------------------------"
PRINT " n    mn      fn    |mn-fn|  Pn(Ay)  Pn(ay)  Pn(AA)  Pn(Aa)  Pn(aa)"
PRINT "-------------------------------------------------------------------"
FOR I = 0 TO 16
PRINT USING "##  "; I;
PRINT USING "#.####  "; M; F; ABS(M - F); PAY; PBY; PAA; PAB; PBB
PAY = 1 - F: PBY = F: PAA = (1 - M) * (1 - F)
PAB = (1 - M) * F + M * (1 - F): PBB = M * F
M = PBY: F = PAB / 2 + PBB
NEXT I
PRINT "-------------------------------------------------------------------"
```

The approach to equilibrium can be appreciated by the following graph, where the dashed line indicates the allele frequencies among females and the continuous one the allele frequencies among males, for $m_0 = 0$ and $f_0 = 1$.



HARDY-WEINBERG EQUILIBRIUM WITH OVERLAPPING GENERATIONS

In the lines that follow the reasoning used by Moran (The statistical processes of evolutionary theory, Oxford University Press, Oxford, 1962, pp. 23-24) is adopted.

Let P(t) = frequency of AA individuals at time t
R(t) = frequency of Aa individuals at time t
Q(t) = frequency of aa individuals at time t ;

assuming that in the time interval dt a fraction dt of the population dies and is replaced by a new fraction dt produced by random mating, the equations that follow are obtained :

$$P(t+dt) = P(t) - P(t).dt + [P(t) + R(t)/2]^2.dt =$$
$$= P(t)(1-dt) + [P(t) + R(t)/2]^2.dt$$

$$R(t+dt) = R(t) - R(t).dt + 2[P(t) + R(t)/2][R(t)/2 + Q(t)].dt =$$
$$= R(t)(1-dt) + 2[P(t) + R(t)/2][R(t)/2 + Q(t)].dt$$

$$Q(t+dt) = Q(t) - Q(t).dt + [R(t)/2 + Q(t)]^2.dt =$$
$$= Q(t)(1-dt) + [R(t)/2 + Q(t)]^2.dt.$$

Rearranging the first of the above expressions, we obtain

$$P(t+dt)-P(t) = -P(t).dt + [P(t) + R(t)/2]^2.dt$$

and

$$[P(t+dt)-P(t)]/dt = P[(t+dt)-P(t)]/[(t+dt)-t]$$
$$= -P(t) + [P(t) + R(t)/2]^2;$$

the limit of this expression, as dt tends to zero, is

$$dP(t)/dt = -P(t) + [P(t) + R(t)/2]^2 ;$$

similarly, we obtain

$$dR(t)/dt = -R(t) +2[P(t) + R(t)/2][R(t)/2 + Q(t)]$$
$$dQ(t)/dt = -Q(t) + [R(t)/2 + Q(t)]^2 .$$

If we define p(t) = P(t) + R(t)/2 ,

it comes out that

$$dp(t)/dt = dP(t)/dt + 1/2.dR(t)/dt =$$
$$-[P(t) + R(t)/2] + [P(t) + R(t)/2]^2$$
$$+ [P(t) + R(t)/2][R(t)/2 + Q(t)]$$
$$= -[P(t) + R(t)/2] + [P(t) + R(t)/2][P(t) + R(t) + Q(t)]$$
$$= -[P(t) + R(t)/2] + [P(t) + R(t)/2] = 0 .$$

Therefore, p(t) and q(t) = 1 - p(t) are constant values (p, q).

Replacing these values in the equations for dP(t)/dt, dR(t)/dt and dQ(t)/dt, we obtain

$dP(t)/dt = -P(t) + p^2$
$dR(t)/dt = -R(t) + 2pq$
$dQ(t)/dt = -Q(t) + q^2$.

The solution for $dP(t)/dt = -P(t) + p^2$ is obtained in the lines below.

From $dP(t)/dt = -P(t) + p^2$ we have :

$dP(t)/[P(t)-p^2] = d \ln|P(t)-p^2| = -dt.$

Integrating both sides of $d \ln|P(t)-p^2| = -dt$, that is,

$\int d \ln|P(t)-p^2| = -\int dt$,

we obtain successively

$\ln|P(t)-p^2| = -t + C = -t + \ln C_1$
$\ln[|P(t)-p^2|/C_1] = -t$
$[P(t)-p^2]/C_1 = e^{-t}$
$P(t) = p^2 + C_1.e^{-t}$.

For t= 0 it comes out that

$P(0) = p^2 + C_1.e^0 = p^2 + C_1$

and

$C_1 = P(0)-p^2$.

Therefore, the complete solution of the equation

$dP(t)/dt = -P(t) + p^2$ is

$P(t) = p^2 + [P(0)-p^2].e^{-t}$.

Similarly, we obtain

$R(t) = 2pq + [R(0)-2pq].e^{-t}$
$Q(t) = q^2 + [Q(0)-q^2].e^{-t}$,

where $p = P(0) + R(0)/2$ and $q = 1-p$.

The limits of the above expressions, as t tends to infinity, are clearly

$P = p^2$
$R = 2pq$
$Q = q^2$.

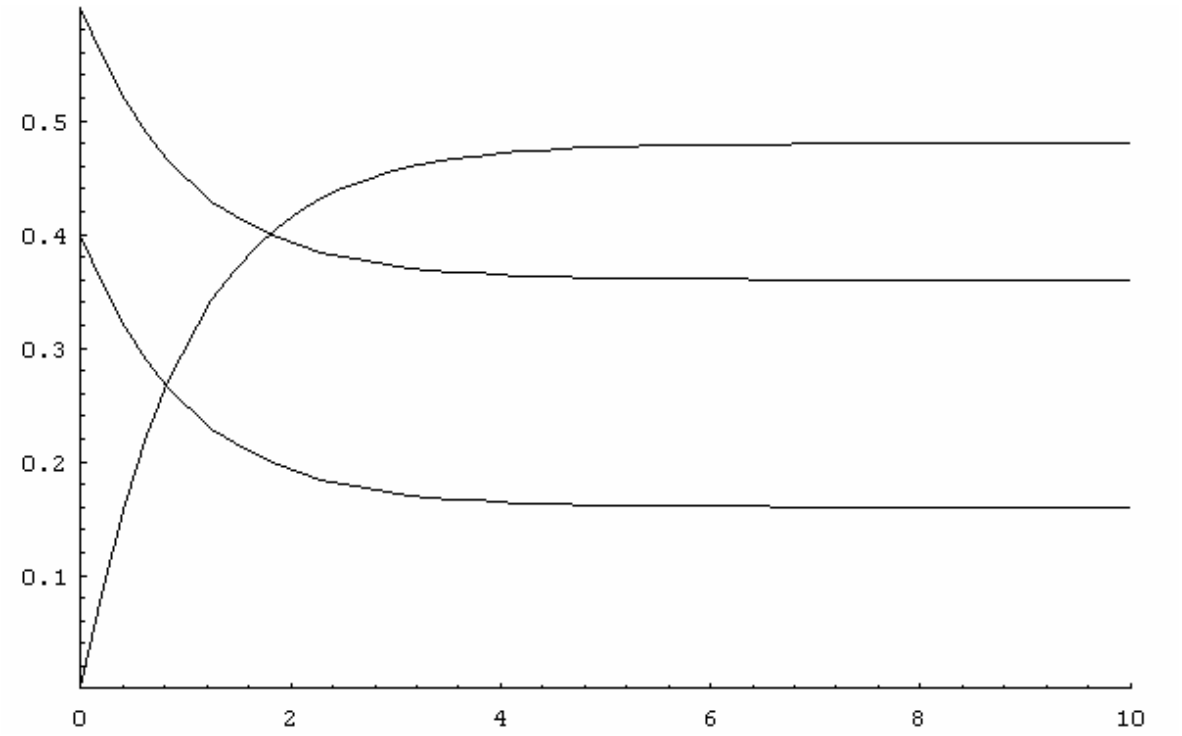
A numerical example of convergence is shown in the table below, followed by the Basic code used for generating it.

```
t            P(t)              R(t)            Q(t)
-----------------------------------------------------
 0         0.40000000       0.00000000       0.60000000
 1         0.24829107       0.30341787       0.44829107
 2         0.19248047       0.41503906       0.39248047
 3         0.17194890       0.45610221       0.37194890
 4         0.16439575       0.47120849       0.36439575
 5         0.19161711       0.47676579       0.36161711
 6         0.16059490       0.47881020       0.36059490
 7         0.16021885       0.47956230       0.36021885
 8         0.16008051       0.47983898       0.36008051
 9         0.16002962       0.47994076       0.36002962
10         0.16001090       0.47997821       0.36001090
11         0.16000401       0.47999198       0.36000401
12         0.16000147       0.47999705       0.36000147
13         0.16000054       0.47999892       0.36000054
14         0.16000020       0.47999960       0.36000020
15         0.16000007       0.47999985       0.36000007
16         0.16000003       0.47999995       0.36000003
17         0.16000001       0.47999998       0.36000001
18         0.16000000       0.47999999       0.36000000
19         0.16000000       0.48000000       0.36000000
20         0.16000000       0.48000000       0.36000000
-----------------------------------------------------
```

```
REM PROGRAM FILENAME HWEQCON1
PRINT "   t          P(t )          R(t)            Q(t)"
PRINT "-----------------------------------------------"
P=.4 : Q=.6 : D0=.40 : H0=0 : R0=.6
FOR T=O TO 20
    Dl=P*P+(D0-P*P)*EXP(-T)
    Hl=2*P*Q+(H0-2*P*Q)*EXP(-T)
    Rl=Q*Q+(R0-Q*Q)*EXP(-T)
    PRINT USING "####";T;
    PRINT USING "     #.#########";Dl;Hl;Rl
NEXT T
    PRINT "-----------------------------------------------"
```

The graph below, generated by the following Matematica code, shows the convergence to equilibrium for AA, Aa and aa genotype frequencies, using the initial numerical values given at the top of the table above.

```
(* hwoverla.ma *)
p = 0.4;
P0 = 0.4;
R0 = 0;
Q0 = 0.6;
P = p^2 + (P0 - p^2) * Exp[-t];
R = 2 * p * (1-p) + (R0 - 2 * p * (1-p)) * Exp[-t];
Q = (1-p)^2 + (Q0 - (1-p)^2) * Exp[-t];
graph = Plot[{P,Q,R},{t,0,10}];
Show[graph, PlotRange -> {0, 0.60},
          AxesOrigin -> {0, 0}]
```


0.5
0.4
0.3
0.2
0.1
0
2
4
6
8
10

FISHER'S PRINCIPLE ON EQUILIBRIUM POPULATIONS

Out of the six possible mating types that occur in a population, four of them (namely, AA × AA, AA × Aa, Aa × aa and aa × aa) reproduce in the offspring exactly the couple genotypic ratios (that is, AA × AA → 1AA:1AA, AA × Aa → 1AA:1Aa, Aa × aa → 1Aa:1aa and aa × aa → 1aa:1aa). Therefore, only the behavior of two crossings (namely AA × aa → 1Aa:1Aa and Aa × Aa → 1AA:2Aa:1aa) has to be analyzed in order to infer any possible equilibrium condition. We start by building the population mating matrix:

	$\mathbf{AA_f}$	$\mathbf{Aa_f}$	$\mathbf{aa_f}$
$\mathbf{AA_m}$	$P(AA_m \times AA_f)$	$P(AA_m \times Aa_f)$	$P(AA_m \times aa_f)$
$\mathbf{Aa_m}$	$P(Aa_m \times AA_f)$	$P(Aa_m \times Aa_f)$	$P(Aa_m \times aa_f)$
$\mathbf{aa_m}$	$P(aa_m \times AA_f)$	$P(aa_m \times Aa_f)$	$P(aa_m \times aa_f)$

The frequency of a given offspring genotype at generation n+1 (v.g., AA) is obtained from the sum of contributions of crossings occurring at generation n, as usual:

$$P_{n+1}(AA) = 1.P_n(AA_m \times AA_f) + \tfrac{1}{2}.[P_n(AA_m \times Aa_f) + P_n(Aa_m \times AA_f)] + \tfrac{1}{4}.P_n(Aa_m \times Aa_f) = P_n(AA \times AA) + P_n(AA \times Aa)/2 + P_n(Aa \times Aa)/4;$$

$$\text{at equilibrium, } P(AA) = 1.P(AA_m \times AA_f) + \tfrac{1}{2}.[P(AA_m \times Aa_f) + P(Aa_m \times AA_f)] + \tfrac{1}{4}.P(Aa_m \times Aa_f) = P(AA \times AA) + P(AA \times Aa)/2 + P(Aa \times Aa)/4 .$$

The frequency of a given parental genotype at generation n (v.g., AA_m) is obtained from the sum of probabilities of crossings in which it participates:

$$P_n(AA_m) = P_n(AA) = P_n(AA_m \times AA_f) + P_n(AA_m \times Aa_f) + P_n(AA_m \times aa_f)$$

$$= P_n(AA \times AA) + P_n(AA \times Aa)/2 + P_n(AA \times aa)/2 .$$

Therefore, a second equilibrium equation for the frequency of the genotype AA can be obtained by adding the elements of the corresponding column or row of the above mating matrix :

$$P(AA) = P(AA \times AA) + P(AA \times Aa)/2 + P(AA \times aa)/2 ,$$

which is equivalent to drop off the subscripts in the equation for $P_n(AA_m)$.

Equating the right sides of the equations

$P(AA) = P(AA\times AA) + P(AA\times Aa)/2 + P(Aa\times Aa)/4$ and
$P(AA) = P(AA\times AA) + P(AA\times Aa)/2 + P(AA\times aa)/2$,

we obtain

$$P(AA\times AA) + P(AA\times Aa)/2 + P(Aa\times Aa)/4 = P(AA\times AA) + P(AA\times Aa)/2 + P(AA\times aa)/2;$$

therefore, the equilibrium condition for any possible population is

$$\mathbf{P(Aa\times Aa) = 2.P(AA\times aa)}$$

The property above (Fisher, 1918) can be used as a short-cut method for determining straightforwardly equilibrium genotype frequencies, avoiding thus the application of tedious algebraic techniques that arise in complicated situations.

SAMPLE ESTIMATES OF GENE FREQUENCIES

1) Two autosomal codominant alleles

$N(AA) = n_1$, $N(Aa) = n_2$, $N(aa) = n_3$, $n_1+n_2+n_3 = N$

Likelihood function: $P = (K/2^{n_2}).(p^2)^{n_1}.(2pq)^{n_2}.(q^2)^{n_3}$
$= K.p^{2n_1+n_2}.q^{n_2+2n_3}$, $K = 2^{n_2}.N!/(n_1!n_2!n_3!)$
$L = \log(P) = (2n_1+n_2).\log(1-q) + (n_2+2n_3).\log q + k$

Max. lik. estimates: $\mathbf{p = (2n_1+n_2)/2N}$, $\mathbf{q = 1-p = (n_2+2n_3)/2N}$
$\mathbf{I(q) = 2N/pq = 2N/[q(1-q)]}$
$\mathbf{var(p) = var(q) = 1/I(q) = pq/2N = q(1-q)/2N}$

2) Two autosomal dominant alleles, A dominant over a

$N(A-) = N(AA)+N(Aa) = n_1$, $N(aa) = n_2$, $n_1+n_2 = N$

Likelihood function: $P = K.(1-q^2)^{n_1}.(q^2)^{n_2}$
$= K.(1-q^2)^{n_1}.q^{2n_2}$, $K = N!/(n_1!n_2!)$
$L = \log(P) = n_1.\log(1-q^2) + 2n_2.\log q + k$

Max. lik. estimates: $\mathbf{q = \sqrt{(n_2/N)}}$, $\mathbf{p = 1-q = 1-\sqrt{(n_2/N)}}$
$\mathbf{I(q) = 4N/(1-q^2)}$
$var(q^2) = q^2(1-q^2)/N = var(q).(dq^2/dq)^2 = 4q^2var(q)$
$\mathbf{var(q) = 1/I(q) = var(q^2)/4q^2 = (1-q^2)/4N}$
$= p^2/4N + pq/2N > pq/2N$

3) Two X-linked codominant alleles

$N(A) = n_1$, $N(a) = n_2$, $n_1+n_2 = N_m$
$N(AA) = n_3$, $N(Aa) = n_4$, $N(aa) = n_5$, $n_3+n_4+n_5 = N_f$

3.1) male sample

Likelihood function: $P = K.p^{n_1}.q^{n_2}$, $K = N_m!/(n_1!n_2!)$
$L = \log(P) = n_1.\log(1-q) + n_2.\log q + k$

Max. lik. estimates: $\mathbf{q = q_m = n_2/N_m}$, $\mathbf{p = p_m = 1-q_m = n_1/N_m}$
$\mathbf{I(q_m) = N_m/[q_m(1-q_m)]}$
$\mathbf{var(q_m) = 1/I(q_m) = q_m(1-q_m)/N_m}$

3.2) female sample

Likelihood function: $P = (K/2^{n_4}).(p^2)^{n_3}.(2pq)^{n_4}.(q^2)^{n_5}$
$= K.p^{2n_3+n_4}.q^{n_4+2n_5}$, $K = 2^{n_4}.N_f!/(n_3!n_4!n_5!)$
$L = \log(P) = (2n_3+n_4).\log(1-q) + (n_4+2n_5).\log q + k$

Max. lik. estimates: $\mathbf{q = q_f = (n_4+2n_5)/2N_f}$,
$\mathbf{p = p_f = 1-q_f = (2n_3+n_4)/2N_f}$
$\mathbf{I(q_f) = 2N_f/[q_f(1-q_f)]}$
$\mathbf{var(q_f) = 1/I(q_f) = q_f(1-q_f)/2N_f}$

3.3) total sample

Likelihood function: $P = (K/2^{n_4}).p^{n_1}.q^{n_2}.(p^2)^{n_3}.(2pq)^{n_4}.(q^2)^{n_5}$
$= K.p^{n_1+2n_3+n_4}.q^{n_2+n_4+2n_5}$,
$K = 2^{n_4}.N_m!N_f!/(n_1!n_2!n_3!n_4!n_5!)$
$L = \log(P) = (n_1+2n_3+n_4).\log(1-q)$
$+ (n_2+n_4+2n_5).\log q + k$

Max. lik. estimates: $q = (n_2+n_4+2n_5)/(N_m+2N_f)$,
$p = 1-q = (n_1+2n_3+n_4)/(N_m+2N_f)$
$I(q) = I(q_m) + I(q_f) = (N_m+2N_f)/[q(1-q)]$
$q \approx [q_m.I(q_m)+q_f.I(q_f)]/[I(q_m)+I(q_f)]$
$var(q) = 1/I(q) = 1/[I(q_m)+I(q_f)] = q(1-q)/(N_m+2N_f)$

4) Two X-linked alleles, A dominant over a

$N(A) = n_1$, $N(a) = n_2$, $n_1+n_2 = N_m$
$N(A-) = N(AA) + N(Aa) = n_3$, $N(aa) = n_4$, $n_3+n_4 = N_f$

4.1) male sample

Likelihood function: $P = K.p^{n1}.q^{n2}$, $K = N_m!/(n_1!n_2!)$
$L = \log(P) = n_1.\log(1-q) + n_2.\log q + k$
Max. lik. estimates: $q = q_m = n_2/N_m$, $p = p_m = 1-q_m = n_1/N_m$
$I(q_m) = N_m/[q_m(1-q_m)]$
$var(q_m) = 1/I(q_m) = q_m(1-q_m)/N_m$

4.2) female sample

Likelihood function: $P = K.(1-q^2)^{n3}.(q^2)^{n4}$
$= K.(1-q^2)^{n3}.q^{2n4}$, $K = N_f!/(n_3!n_4!)$
$L = \log(P) = n_3.\log(1-q^2) + 2n_4.\log q + k$
Max. lik. estimates: $q = q_f = \sqrt{(n_4/N_f)}$, $p = p_f = 1-q_f = 1-\sqrt{(n_4/N_f)}$
$I(q_f) = 4N_f/(1-q_f^2)$
$var(q_f) = 1/I(q_f) = (1-q_f^2)/4N_f$

4.3) total sample

Likelihood function: $P = K.p^{n1}.q^{n2}.(1-q^2)^{n3}.(q^2)^{n4}$
$= K.(1-q)^{n1}.q^{n2+2n4}.(1-q^2)^{n3}$,
$K = N_m!N_f!/(n_1!n_2!n_3!n_4!)$
$L = \log(P) = n_1.\log(1-q) + (n_2+2n_4).\log q$
$+ n_3.\log(1-q^2) + k$
Max. lik. estimates: $q = \{-n_1 + \sqrt{[n_1^2+4(N_m+2N_f)(n_2+2n_4)]}\}/2(N_m+2N_f)$,
$p = 1 - q$
$I(q) = I(q_m)+I(q_f) = [N_m+q(N_m+4N_f)]/[q(1-q^2)]$
$q \approx [q_m.I(q_m)+q_f.I(q_f)]/[I(q_m)+I(q_f)]$
$var(q) = 1/I(q) = 1/[I(q_m)+I(q_f)]$
$= q(1-q^2)/[N_m+q(N_m+4N_f)]$

MAXIMUM LIKELIHOOD ESTIMATE FOR THE FREQUENCY OF DOMINANT AUTOSOMAL ALLELES

Let D and R be the observed numbers of dominant (AA+Aa) and recessive (aa) individuals in a randon sample of G individuals. Assuming panmixia, it comes out that the probability of such a result is given by

$P(D,R) = G![P(1)]^D.[P(2)]^R/(D!R!)$, where $P(1) = 1-q^2$ and $P(2) = q^2$.

Putting $L = \ln P = \text{const.} + D.\ln(1-q^2) + 2R.\ln q$ and $dL/dq = 0$, it comes out that

$$dL/dq = 0 = -2Dq/(1-q^2) + 2R/q$$
$$2Dq^2 = 2R(1-q^2)$$
$$(2D+2R)q^2 = 2R$$
$$q = \sqrt{(R/G)}$$

and

$$d^2L/dq^2 = d(dL/dq)/dq = -2D(1+q^2)/(1-q^2)^2 - 2R/q^2 ;$$

since $R = Gq^2$ and $D = G(1-q^2)$,
at the estimation point $q = \sqrt{(R/G)}$ the second derivative has the numerical value

$$d^2L/dq^2 = - [2Dq^2(1+q^2)+2R(1-q^2)^2]/[q^2(1-q^2)^2] =$$
$$= - [2Gq^2(1-q^2)(1+q^2)+2Gq^2(1-q^2)^2]/[q^2(1-q^2)^2] =$$
$$= - [2G(1+q^2)+2G(1-q^2)]/(1-q^2) = - 4G/(1-q^2);$$

therefore, $\text{var}(q) = - 1/(d^2L/dq^2)$
$= (1-q^2)/4G$.

The result just obtained can be straightforwardly derived using the principle of functional invariance. This is shown in the lines that follow.
Putting $y = q^2$ and $1-y = 1-q^2$, it comes out that
$\text{var}(y) = \text{var}(q^2) = y(1-y)/G$, that is the usual formula for binomial variance); using the property
$\text{var}(y) = (dy/dq)^2 . \text{var}(q)$,
where $dy/dq = 2q$ [and therefore $(dy/dq)^2 = 4q^2$],
we get $\text{var}(y) = q^2(1-q^2)/G = (dy/dq)^2 . \text{var}(q) = 4q^2.\text{var}(q)$;
therefore, $\text{var}(q) = q^2(1-q^2)/4Gq^2 = (1-q^2)/4G$.

We note that $(1-q^2)/4G = (p^2+2pq)/4G = pq/2G + p^2/4G > pq/2G$, as expected.

Numerical example: in a sample of G = 18 randomly collected individuals D = 16 had the dominant phenotype (A- = AA or Aa), while R = 2 exhibited the recessive phenotype corresponding to genotype aa.

Under the ancillary hypothesis of panmixia the expected numbers of dominant individuals are respectively $G(1-q^2)$ and Gq^2, as shown below.

genotypes	expected frequencies	observed numbers	expected numbers
AA + Aa	$p^2 + 2pq = 1 - q^2$	D = 16	$G(1-q^2)$
aa	q^2	R = 2	Gq^2
total	1	G = 18	

The likelihood function is then given by

$P = 153.q^4.(1-q^2)^{16}$ or by $L = \ln(P) = \ln(153) + 4.\ln(q) + 16.\ln(1-q^2)$.

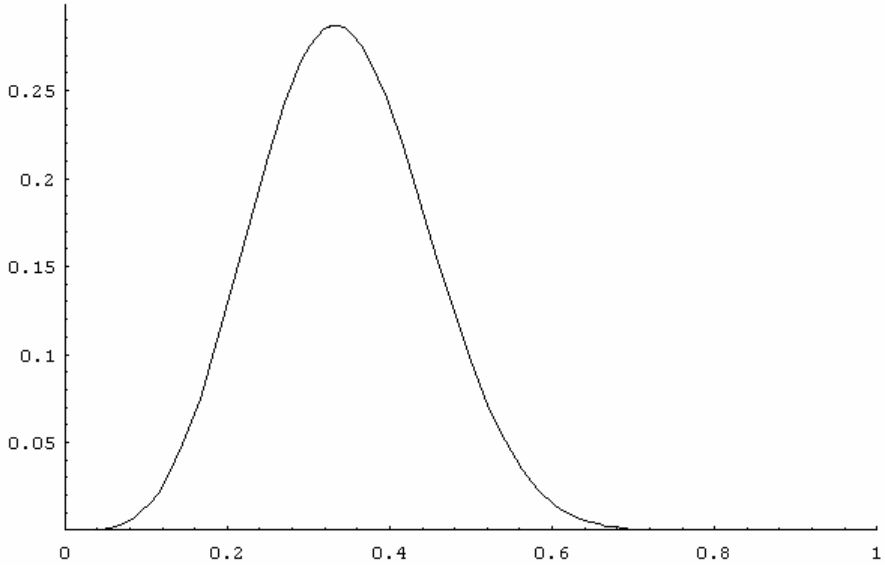
The values P and L take as q varies from 0 to 1 are shown in the table below.

q	P	L
0.000	0.000000	-inf
0.050	0.000919	-6.992541
0.100	0.013027	-4.340708
0.150	0.053818	-2.922154
0.200	0.127395	-2.060466
0.250	0.212810	-1.547356
0.300	0.274056	-1.294424
0.331	0.286860	-1.248761
0.332	0.286903	-1.248S12
0.333	0.286922	-1.248544
0.334	0.286918	-1.248558
0.335	0.286891	-1.248652
0.336	0.286841	-1.248827
0.337	0.286768	-1.249083
0.338	0.286671	1.249420
0.339	0.286552	1.249837
0.350	0.283738	-1.259704
0.400	0.240658	-1.424379
0.450	0.167970	-1.783968
0.500	0.095841	-2.345064
0.550	0.043939	-3.124954
0.600	0.015710	-4.153458
0.650	0.004180	~5.477443
0.700	0.000769	-7.169775
0.750	0.000087	-9.347148
0.800	0.000005	-12.208556
0.850	0.000000	-16.130587
0.900	0.000000	-21.962703
0.950	0.000000	-32.421182
1.000	0.000000	-inf

The maximum value P or L take occurs when $q = \sqrt{(2/18)} = 1/3$; and this is precisely the maximum likelihood estimate of q.

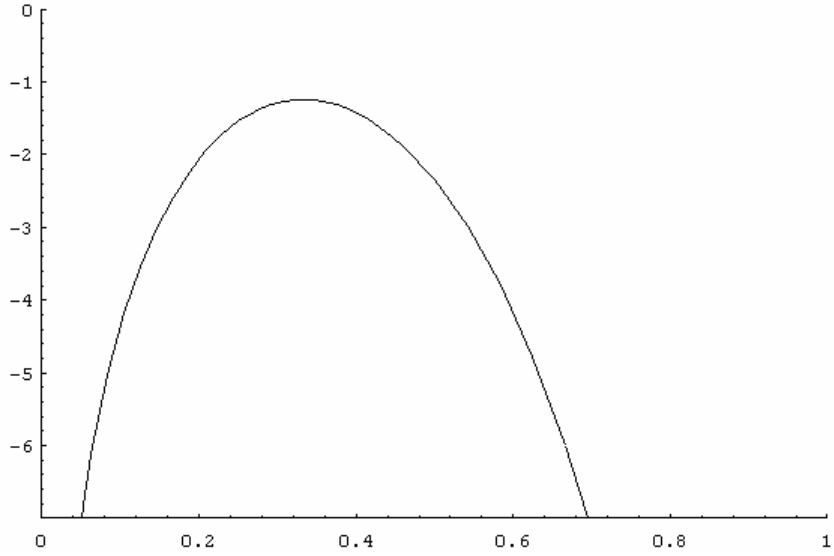
The graphs that follow show the values of P, L and var(q) as functions of q, for D = 16, R = 2 and G = 18. In the last graph the values of var(q) are compared to those obtained using the formula for binomial variance, pq/36. For any q < 1, (1-q^2)/72 > q(1-q)/36, as already stated. The Mathematica codes that generated the graphs are listed below the corresponding figures.

1) Graph of $P = 153.q^4.(1-q^2)^{16}$



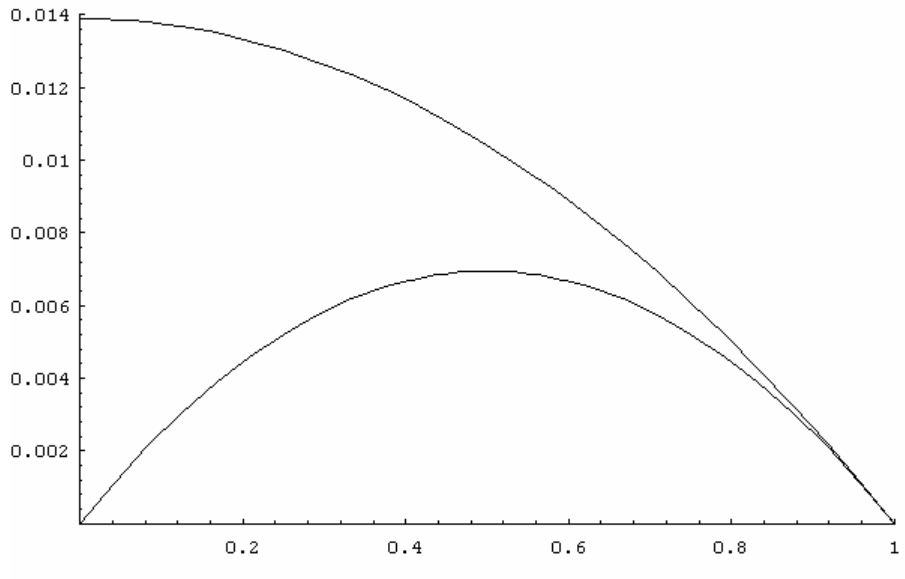

```
(* domlik01.ma *)
P = 153 * q^4 * (1-q^2)^16;
graph = Plot[P,{q,0,1}];
Show[graph, PlotRange -> {0,0.3}, AxesOrigin -> {0,0}]
```

2)Graph of $L = \log(153) + 4.\log(q) + 16.\log(1-q^2)$




```
(* domlik02.ma *)
P = 153 * q^4 * (1-q^2)^16; L = Log[P]; graph = Plot[L,{q,0,1}];
Show[graph, PlotRange -> {-7,0}, AxesOrigin -> {0,-7}]
```

3) Graph of $var'(q) = (1-q^2)/72$ and $var''(q) = q(1-q)/36$




```
(* domlik03.ma *)
var1 = q * (1-q)/36; var2 = (1-q^2)/72;
Plot[{var1,var2},{q,0,1}]
```

GENETIC EQUILIBRIUM IN RELATION TO A PAIR OF LOCI

Let $P_0(AB) = e$
$P_0(Ab) = f$
$P_0(aB) = g$
$P_0(ab) = h$

be the gamete (or haplotype, if the loci are syntenic) composition of a large diploid population at initial generation **0**. The recombination frequency between loci (**A**,**a**) and (**B**,**b**) has a value **r (0.5 ≥ r ≥ 0)**. This means that a coupling heterozygote **AB/ab** produces gametes **AB**, **Ab**, **aB** and **ab** with respective frequencies **(1-r)/2**, **r/2**, **r/2** and **(1-r)/2**, the combined frequency of recombinant gametes (**Ab** and **aB**) being r. Assuming panmixia, the following are the frequencies of possible genotypes in generation **1**:

	AB	Ab	aB	ab	
AB	e^2	ef	eg	eh	e
Ab	ef	f^2	fg	fh	f
aB	eg	fg	g^2	gh	g
ab	eh	fh	gh	h^2	h
	e	f	g	h	1

Therefore, the frequency of **AB** gametes in generation **1** is

$$\begin{aligned}P_1(AB) &= P_1(AB/AB) + P_1(AB/Ab)/2 + P_1(AB/aB)/2 + (1-r)P_1(AB/ab)/2 \\ &\quad + rP_1(Ab/aB)/2 = e^2 + ef + eg + (1-r)eh + rfg \\ &= e2 + ef + eg + eh - r(eh-fg) \\ &= e(e+f+g+h) - r(eh-fg) = e - r(eh-fg) \\ &= P_0(AB) - r[P_0(AB).P_0(ab)-P_0(Ab).P_0(aB)] \\ &= P_0(AB) - r.D_0 \\ &= e^2 + ef + eg + fg + (1-r)eh- fg + rfg \\ &= e(e+f) + g(e+f) + (1-r)(eh-fg) \\ &= (e+f)(e+g) + (1-r)(eh-fg) \\ &= P_0(A).P_0(B) + (1-r)[P_0(AB).P_0(ab)-P_0(Ab).P_0(aB)] \\ &= P_0(A).P_0(B) + (1-r).D_0\end{aligned}$$

Therefore we have

$$P_1(AB) = P_0(AB) - r.D_0 = P_0(A).P_0(B) + (1-r).D_0$$

and, by symmetry,

$$\begin{aligned}P_1(Ab) &= P_0(Ab) + r.D_0 = P_0(A).P_0(b) - (1-r).D_0 \\ P_1(aB) &= P_0(aB) + r.D0 = P_0(a).P_0(B) - (1-r).D_0 \\ P_1(ab) &= P_0(ab) - r.D0 = P_0(a).P_0(b) + (1-r).D_0\ .\end{aligned}$$

Since

$$\begin{aligned}P_1(A) &= P_1(AB) + P_1(Ab) = P_0(AB) + P_0(Ab) = P_0(A) = \ldots = P(A) \\ P_1(a) &= P_1(aB) + P_1(ab) = P_0(aB) + P_0(ab) = P_0(a) = \ldots = P(a) \\ P_1(B) &= P_1(AB) + P_1(aB) = P_0(AB) + P_0(aB) = P_0(B) = \ldots = P(B) \\ P_1(b) &= P_1(Ab) + P_1(ab) = P_0(Ab) + P_0(ab) = P_0(b) = \ldots = P(b)\end{aligned}$$

and

$$P_0(AB) - r.D_0 = P(A).P(B) + (1-r).D_0 = P(A).P(B) + D_0 - r.D_0 ,$$

it comes out that

$$P_0(AB) = P(A).P(B) + D_0 .$$

Comparing this equation with that for $P_1(AB)$,

$$P_1(AB) = P(A).P(B) + (1-r).D_0 ,$$

immediately we get the general solution

$$P_n(AB) = P(A).P(B) + (1-r)^n.D_0$$

and, again by symmetry,

$$P_n(Ab) = P(A).P(b) - (1-r)^n.D_0$$
$$P_n(aB) = P(a).P(B) - (1-r)^n.D_0$$
$$P_n(ab) = P(a).P(b) + (1-r)^n.D_0 .$$

Since at equilibrium, that is when **n** tends to infinity, since $(1-r)^n$ tends to zero as **n** increases,

$$P(AB) = P(A).P(B)$$
$$P(Ab) = P(A).P(b)$$
$$P(aB) = P(a).P(B)$$
$$P(ab) = P(a).P(b) .$$

So we deduce also that at equilibrium obviously the frequencies of the two possible types of double heterozygotes (in coupling and in repulsion) are the same. This is exactly the equilibrium condition, since with equal frequencies of the two possible types of double heterozygotes the production of gametes **AB**, **Ab**, **aB** and **ab** by the whole group of heterozygotes shall be of **1/4** for each gametic class, independent of **r** and as if the loci were unlinked:

	AB	Ab	aB	ab
AB/ab	(1-r)/2	r/2	r/2	(1-r)/2
Ab/aB	r/2	(1-r)/2	(1-r)/2	r/2
average	1/4	1/4	1/4	1/4

We can get the same results using an alternative reasoning, which is shown below. Let us consider again the difference equation

$$P_1(AB) = P(A).P(B) + (1-r).D_0 ;$$

making, in

$$P_2(AB) = P(A).P(B) + (1-r).D_1$$
$$= P(A).P(B) + (1-r).[P_1(AB).P_1(ab)-P_1(Ab).P_1(aB)]$$

the following substitutions

$$P_1(AB) = P_0(A).P_0(B) + (1-r).D_0$$
$$P_1(Ab) = P_0(A).P_0(b) - (1-r).D_0$$
$$P_1(aB) = P_0(a).P_0(B) - (1-r).D_0$$

$$P_1(ab) = P_0(a).P_0(b) + (1-r).D_0$$

we get

$$P_2(AB) = P(A).P(B) + (1-r).(1-r).D_0 = P(A).P(B) + (1-r)^2.D_0$$

and the general solution

$$P_n(AB) = P(A).P(B) + (1-r)^n.D_0$$

already found using the first method.

The method just shown is interesting because it demonstrates clearly that

$$D_1 = (1-r).D_0$$

and therefore that

$$D_n = (1-r)^n.D_0 \ ;$$

when **n** tends to infinity, $(1-r)^n$ tends to zero, so that deleting the subscripts, consistent with equilibrium, yields

$$D = P(AB).P(ab) - P(Ab).P(aB) = 0 \ ,$$

which is (again) the equilibrium condition.

There is a third manner to get the equations that describe the approach to equilibrium (Crow & Kimura, 1970, p.47-48). If we define:

a) $P_n(A_iB_j)$: frequency of the haplotype A_iB_j at generation **n**;
b) $P_{n+1}(A_iB_j)$: same frequency in the next generation;
c) $P_n(A_i) = P(A_i)$: frequency of the **i-th** allele of the **A** locus in any generation or, for sufficiently large populations, the probability of a given allele of the **A** locus being the **i-th** one;
d) $P_n(B_j) = P(B_j)$: frequency of the **j-th** allele of the **B** locus in any generation or, for sufficiently large populations, the probability of a given allele of the **B** locus being the **j-th** one;

we have immediately

$$\begin{aligned} P_{n+1}(A_iB_j) &= P_n(A_iB_j) + r.P(A_i).P(B_j) - r.P_n(A_iB_j) \\ &= (1-r).P_n(A_iB_j) + r.P(A_i).P(B_j); \end{aligned}$$

subtracting from both sides of the above equation the constant quantity $P(A_i).P(B_j)$, we obtain

$$P_{n+1}(A_iB_j) - P(A_i).P(B_j) = (1-r).[P_n(A_iB_j) - P(A_i).P(B_j)]$$

and, therefore, the general solution

$$P_n(A_iB_j) - P(A_i).P(B_j) = (1-r)^n.[(P_0(A_iB_j) - P(A_i).P(B_j)]$$

Since, as we have shown before,

$$P_0(A_iB_j) - rD_0 = P(A_i).P(B_j) + (1-r)D_O \ ,$$

it comes out that

$$P_0(A_iB_j) = P(A_i).P(B_j) + D_0$$

Substituting this in the general solution shown above, we obtain

$$P_n(A_iB_j) - P(A_i).P(B_j) = (1-r)^n.[(P_0(A_iB_j) - P(A_i).P(B_j)]$$
$$= (1-r)^n.[P(A_i).P(B_j) + D_0 - P(A_i).P(B_j)]$$

and

$$P_n(A_iB_j) = P(A_i).P(B_j) + (1-r)^n.D_0$$

which is the solution which we have obtained before.

In general, for any number of syntenic or linked loci (as well as for unlinked loci), at equilibrium

$$P(A_iB_jC_k...) = P(A_i).P(B_j).P(C_k). \ ... \ .$$

The quantity

$$\Delta(A_iB_jC_k...) = P(A_iB_jC_k...) - P(A_i).P(B_j).P(C_k). \ ...$$

is the so-called linkage disequilibrium value for the haplotype $A_iB_jC_k...$. This linkage disequilibrium value may arise as a result of the loci being very near [making thus recombination virtually impossible, as is the case of loci (C,c), (D,d) and (E,e) in Rh blood group system] or as a result of several other factors, such as differential viabilities (or adaptive values) acting on different haplotypes.

The important points to be kept in mind are the following:

1) it is impossible to ascertain linkage using population data, since at equilibrium the population distribution of possible genotypes is exactly the same one observed in relation of two independently inherited loci (that is, situated on different chromosomes); for both cases, this is given by

$$P(AABB) = P(A)^2.P(B)^2$$
$$P(AABb) = 2.P(A)^2.P(B).P(b)$$
$$P(AAbb) = P(A)^2.P(b)^2$$
$$P(AaBB) = 2.P(A).P(a).P(B)^2$$
$$P(AaBb) = P(AB/ab) + P(Ab/aB) = 4.P(A).P(a).P(B).P(b)$$
$$P(Aabb) = 2.P(A).P(a).P(b)^2$$
$$P(aaBB) = P(a)^2.P(B)^2$$
$$P(aaBb) = 2.P(a)^2.P(B).P(b)$$
$$P(aabb) = P(a)^2.P(b)^2 \ ;$$

2) when **r = 1/2**, both types of heterozygotes (**AB/ab** and **Ab/aB**) produce the four possible types of gametes **AB**, **Ab**, **aB** and **ab** with identical frequencies (each one equal to **1/4**); this case corresponds therefore to independent assortment; however, as in the case **r < 0.5**, the population is in an equilibrium state if and only if

$$P(AaBb) = 4.P(A).P(B).P(B).P(b).$$

3) if two loci are separated by a relatively large distance in the chromosome, it is quite probable that the recombination fraction value between the genes from these two loci will approach a value of **1/2,** rendering it difficult or even impossible to demonstrate linkage.

As a numerical example, let us consider the following population, whose gametic composition at generation **0** is the following one:

	B	**b**	
A	**0.2000**	**0.2500**	**0.4500**
a	**0.2000**	**0.3500**	**0.5500**
	0.4000	**0.6000**	**1.0000** ,

that is, $P_0(AB) = 0.20$, $P_0(Ab) = 0.25$, $P_0(aB) = 0.20$, $P_0(ab) = 0.35$, and

$P_0(A) = P(A) = P_0(AB) + P_0(Ab) = 0.45$,
$P0(a) = P(a) = P_0(aB) + P_0(ab) = 0.55$,
$P_0(B) = P(B) = P_0(AB) + P_0(aB) = 0.40$,
$P_0(b) = P(b) = P_0(Ab) + P_0(ab) = 0.60$.

Assuming panmixia and that the recombination frequency is **r = 0.5** (it is therefore irrelevant if the genes are syntenic or not) the following numerical values are obtained for genotype, haplotype and allele frequencies in generations **1 - 10**:

		BB	**Bb**	**bb**			**B**	**b**	
1	**AA**	**0.0400**	**0.1000**	**0.0625**	**0.2025**	**A**	**0.1900**	**0.2600**	**0.4500**
	Aa	**0.0800**	**0.2400**	**0.1750**	**0.4950**	**a**	**0.2100**	**0.3400**	**0.5500**
	aa	**0.0400**	**0.1400**	**0.1225**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				
		BB	**Bb**	**bb**			**B**	**b**	
2	**AA**	**0.0361**	**0.0988**	**0.0676**	**0.2025**	**A**	**0.1850**	**0.2650**	**0.4500**
	Aa	**0.0798**	**0.2384**	**0.1768**	**0.4950**	**a**	**0.2150**	**0.3350**	**0.5500**
	aa	**0.0441**	**0.1428**	**0.1156**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				
		BB	**Bb**	**bb**			**B**	**b**	
3	**AA**	**0.0342**	**0.0981**	**0.0702**	**0.2025**	**A**	**0.1825**	**0.2675**	**0.4500**
	Aa	**0.0796**	**0.2379**	**0.1775**	**0.4950**	**a**	**0.2175**	**0.3325**	**0.5500**
	aa	**0.0462**	**0.1440**	**0.1122**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				
		BB	**Bb**	**bb**			**B**	**b**	
4	**AA**	**0.0333**	**0.0976**	**0.0716**	**0.2025**	**A**	**0.1813**	**0.2687**	**0.4500**
	Aa	**0.0794**	**0.2377**	**0.1779**	**0.4950**	**a**	**0.2188**	**0.3312**	**0.5500**
	aa	**0.0473**	**0.1446**	**0.1106**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				
		BB	**Bb**	**bb**			**B**	**b**	
5	**AA**	**0.0329**	**0.0974**	**0.0722**	**0.2025**	**A**	**0.1806**	**0.2694**	**0.4500**
	Aa	**0.0793**	**0.2377**	**0.1780**	**0.4950**	**a**	**0.2194**	**0.3306**	**0.5500**
	aa	**0.0479**	**0.1449**	**0.1097**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				
		BB	**Bb**	**bb**			**B**	**b**	
6	**AA**	**0.0326**	**0.0973**	**0.0726**	**0.2025**	**A**	**0.1803**	**0.2697**	**0.4500**
	Aa	**0.0792**	**0.2376**	**0.1781**	**0.4950**	**a**	**0.2197**	**0.3303**	**0.5500**
	aa	**0.0481**	**0.1451**	**0.1093**	**0.3025**		**0.4000**	**0.6000**	**1.0000**
		0.1600	**0.4800**	**0.3600**	**1.0000**				

```
-----------------------------------------------------------------------
        BB      Bb      bb                      B       b
 7  AA  0.0325  0.0973  0.0727  0.2025      A   0.1802  0.2698  0.4500
    Aa  0.0792  0.2376  0.1782  0.4950      a   0.2198  0.3302  0.5500
    aa  0.0483  0.1451  0.1091  0.3025          0.4000  0.6000  1.0000
        0.1600  0.4800  0.3600  1.0000
-----------------------------------------------------------------------
        BB      Bb      bb                      B       b
 8  AA  0.0325  0.0972  0.0728  0.2025      A   0.1801  0.2699  0.4500
    Aa  0.0792  0.2376  0.1782  0.4950      a   0.2199  0.3301  0.5500
    aa  0.0483  0.1452  0.1090  0.3025          0.4000  0.6000  1.0000
        0.1600  0.4800  0.3600  1.0000
-----------------------------------------------------------------------
        BB      Bb      bb                      B       b
 9  AA  0.0324  0.0972  0.0729  0.2025      A   0.1800  0.2700  0.4500
    Aa  0.0792  0.2376  0.1782  0.4950      a   0.2200  0.3300  0.5500
    aa  0.0484  0.1452  0.1090  0.3025          0.4000  0.6000  1.0000
        0.1600  0.4800  0.3600  1.0000
-----------------------------------------------------------------------
        BB      Bb      bb                      B       b
10  AA  0.0324  0.0972  0.0729  0.2025      A   0.1800  0.2700  0.4500
    Aa  0.0792  0.2376  0.1782  0.4950      a   0.2200  0.3300  0.5500
    aa  0.0484  0.1452  0.1089  0.3025          0.4000  0.6000  1.0000
        0.1600  0.4800  0.3600  1.0000
-----------------------------------------------------------------------
```

The table above was generated by the following BASIC code:

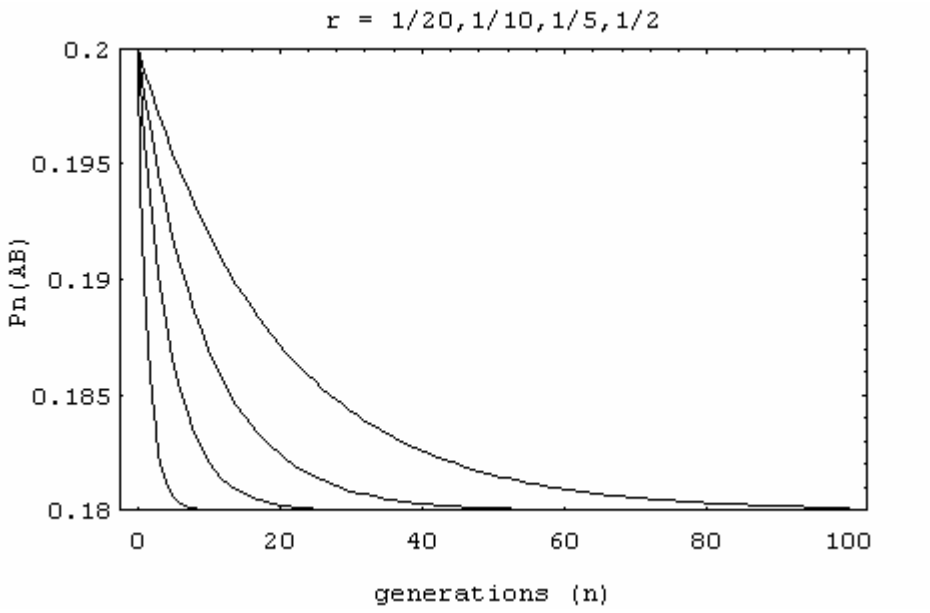
```
REM PROGRAM FILENAME LINKGE01.BAS
REM E = P(AB), F = P(Ab), G = P(aB), H = P(ab)
REM P = P(A) = E+F, Q = P(a) = G+H, S = P(B) = E+G, T = P(b) = F+H
CLS : DEFDBL A-Z
E(0) = .2: F(0) = .25: G(0) = .2: H(0) = .35
P = E(0) + F(0): Q = 1 - P: S = E(0) + G(0): T = 1 - S: R = .5
D = E(0) * H(0) - F(0) * G(0)
FOR I = 1 TO 10
GT1 = E(I - 1) * E(I - 1)'                                GT1 = P(AABB)
GT2 = 2 * E(I - 1) * F(I - 1)'                            GT2 = P(AABb)
GT3 = F(I - 1) * F(I - 1)'                                GT3 = P(AAbb)
GS1 = GT1 + GT2 + GT3'                                    GS1 = P(AA)
GT4 = 2 * E(I - 1) * G(I - 1)'                            GT4 = P(AaBB)
GT5 = 2 * E(I - 1) * H(I - 1) + 2 * F(I - 1) * G(I - 1)' GT5 = P(AaBb)
GT6 = 2 * F(I - 1) * H(I - 1)'                            GT6 = P(Aabb)
GS2 = GT4 + GT5 + GT6'                                    GS2 = P(Aa)
GT7 = G(I - 1) * G(I - 1)'                                GT7 = P(aaBB)
GT8 = 2 * G(I - 1) * H(I - 1)'                            GT8 = P(aaBb)
GT9 = H(I - 1) * H(I - 1)'                                GT9 = P(aabb)
GS3 = GT7 + GT8 + GT9'                                    GS3 = P(aa)
GS4 = GT1 + GT4 + GT7'                                    GS4 = P(BB)
GS5 = GT2 + GT5 + GT8'                                    GS5 = P(Bb)
GS6 = GT3 + GT6 + GT9'                                    GS6 = P(bb)
GST = GS1 + GS2 + GS3'                                    GST = 1
E(I) = P * S + (1 - R) ^ I * D
F(I) = P * T - (1 - R) ^ I * D
G(I) = Q * S - (1 - R) ^ I * D
H(I) = Q * T + (1 - R) ^ I * D
PRINT "-----------------------------------------------------------------------"
PRINT "        BB      Bb      bb                      B       b"
PRINT USING "##  AA  "; I; : PRINT USING "#.####  "; GT1; GT2; GT3; GS1;
PRINT "    A  "; : PRINT USING "#.####  "; E(I); F(I); P
PRINT "    Aa  "; : PRINT USING "#.####  "; GT4; GT5; GT6; GS2;
PRINT "    a  "; : PRINT USING "#.####  "; G(I); H(I); Q
PRINT "    aa  "; : PRINT USING "#.####  "; GT7; GT8; GT9; GS3;
```

```
PRINT "       "; : PRINT USING "#.####  "; S; T; S + T
PRINT "        "; : PRINT USING "#.####  "; GS4; GS5; GS6; GST
NEXT I
PRINT "-------------------------------------------------------------------------"
```

Inspection of the above table shows that Hardy-Weinberg proportions are attained for each locus separately already in the first generation of random mating, as shown by the marginal frequencies of the matrices for genotypic frequencies shown at left. The determinant of the gamete matrix represents the linkage disequilibrium value for haplotypes **AB** or **ab**, which is Δ **= 0.20 x 0.35 - O.20 x 0.25 = 0.07 - 0.05 = 0.02** at generation **0** and Δ **= 0.18 x 0.33 - 0.22 x 0.27 = 0** after an infinite number of generations (the rounded values, with four significant digits or an absolute error equal or less than 0.00005, obtained at generation 10 are already in this situation). The marginals of this last matrix indicate that gene frequencies do not suffer any alteration during the whole process.

In the example just worked, convergence to approximate equilibrium state was fast because **r** was assumed to be **0.5**, the highest value the recombination fraction can take. For other values (such as 0.2, 0.1 and 0.05), convergence takes place slowly, as the following Mathematica graph shows.

```
(* linkge01.ma
   Haplotype ab frequencies
   Pn = P(A).P(B) + (1-r)^n . D
   P0 = 0.20, D = 0.02, r = 0.5, 0.2, 0.1, 0.05
*)
F[n_,r_] := 0.18 + 0.02 * (1-r)^n;
Plot[{F[i,0.5],F[i,0.2],F[i,0.1],F[i,0.05]},{i,0,100},
         PlotPoints->101, Frame -> True,
         PlotLabel->"r = 1/20,1/10,1/5,1/2",
         FrameLabel->{"generations (n)","Pn(AB)"},
         PlotRange -> {0.18,0.20}, AxesOrigin -> {0,0.18}]
```



EXERCISES

1) The following are the frequencies of **Rh** haplotypes in England (Race et al., 1948, apud Race RR & Sanger R, "Blood Groups in Man", Blackwell Scientific Publications, Oxford, 1962):

CDE	**0.0024**
CDe	**0.4205**
CdE	**0.0000**
Cde	**0.0098**
cDE	**0.1411**
cDe	**0.0257**
cdE	**0.0119**
cde	**0.3886**
	1.0000

Estimate: a) the frequencies of alleles **C** and **c**, **D** and **d**, **E** and **e**, and their respective standard errors; b) the haplotype frequencies on the hypothesis of equilibrium; c) the linkage disequilibrium values for each one of the above haplotypes.

2) The following are the results of the testing of a sample of 1400 Hungarians (Rex-Kiss & Horvath, 1966) with 5 Rh anti-sera (anti-C, anti-c, anti-D, anti-E, and anti-e):

REACTION WITH ANTI-					
C	**c**	**D**	**E**	**e**	
+	**-**	**+**	**+**	**+**	**3**
+	**-**	**+**	**-**	**+**	**260**
+	**+**	**+**	**+**	**+**	**182**
+	**+**	**+**	**-**	**+**	**502**
+	**+**	**-**	**-**	**+**	**13**
-	**+**	**+**	**+**	**-**	**37**
-	**+**	**+**	**+**	**+**	**156**
-	**+**	**+**	**-**	**+**	**23**
-	**+**	**-**	**+**	**+**	**6**
-	**+**	**-**	**-**	**+**	**218**
					1400

Estimate the frequencies for the alleles (**C**, **c**), (**D** ,**d**), and (**E**, **e**), with respective standard errors. Estimate the frequencies of the eight haplotypes of Rh system (**CDE**, **CDe**, **CdE**, **Cde**, **cDE**, **cDe**, **cdE** and **cde**). In order to achieve this, you should first write a program based on information contained in pages 53-54 of Mourant, Kopec & Domaniewska-Sobczak's book. Estimate the linkage disequilibrium values for these haplotypes.

CALCULATION OF HAPLOTYPE FREQUENCIES AND OF LINKAGE DISEQUILIBRIUM VALUES FOR LINKED GENE COMPLEXES (E.G., HLA-SYSTEM, Rh-SYSTEM)

This is accomplished by determining, in a population sample, the frequencies of individuals **+/+**, **+/-**, **-/+** and **-/-** who react with two different anti-sera (e.g., anti-sera **anti-A_i** and **anti-B_j**, that is, anti-sera that detect the i-th antigen determined by the one of the alleles--the i-th one--belonging to the **A** locus and the j-th antigen determined by the j-th allele of the **B** locus. **A** and **B** afe **syntenic**, that is, are assumed to be located in the same chromosome). Let us suppose that the results among **N** sampled individuals were the following:

REACTION WITH ANTI-A_i	REACTION WITH ANTI-B_j	
+	+	n_1
+	-	n_2
-	+	n_3
-	-	n_4
		N

The above frequencies can be rearranged as the following contingency table:

		REACTION WITH ANTI-Ai: +	REACTION WITH ANTI-Ai: -	
REACTION WITH ANTI-Bj	+	n_1	n_3	n_1+n_3
	-	n_2	n_4	n_2+n_4
		n_1+n_2	n_3+n_4	N

The frequency of the "double-recessive" **ab/ab** is n_4/N. Since the sample is composed of **N** unrelated individuals, in order to proceed, we assume tacitly that the frequency of **ab/ab** individuals is the square of the **ab** frequency. Therefore the inferred frequency of the haplotype **ab** is

$$P(a\text{-}b) = \sqrt{(n_4/N)}.$$

Under the hypothesis that the linkage disequilibrium value is zero, the expected frequency for the **ab** haplotype is given by the expression

$$P'(a\text{-}b) = (1-p_i)(1-p_j) = q_i q_j ,$$

where $p_i = 1 - q_i$ is the frequency of the A_i allele in the **A** locus and $p_j = 1 - q_j$ the frequency of the B_j allele in locus **B**. The values q_i and q_j are easily estimated from the above contingency table:

$$q_i = \sqrt{[(n_3+n_4)/N]} \text{ and } q_j = \sqrt{[(n_2+n_4)/N]}.$$

If we define the linkage disequilibrium value as $\Delta(a\text{-}b) = P(a\text{-}b) - P'(a\text{-}b)$, it comes out that $\Delta(a\text{-}b) = \sqrt{(n_4/N)} - \sqrt{[(n_3+n_4)(n_2+n_4)]/N}$.

This is the required linkage disequilibrium value between the genes **a** and **b** of loci **A** and **B**.

Numerical exercise:

A sample of 1967 unrelated danes was typed as to antigens A_1 and B_8 of the HLA-system. The results were as follows (Svejgaard et al., 1975, ref. 115 apud Vogel F & Motulsky A, "Human Genetics", Springer Verlag, New York and Berlin, 1979):

REACTION WITH ANTI-A_1	REACTION WITH ANTI-B_8	
+	+	376
+	-	235
-	+	91
-	-	1265
		1967

1) Using a chi-squared test, determine if there is a significant deviation from linkage equilibrium (this can be done by verifying whether there is or is not association between the antigens).

2) What is the frequency, in this population, of the A_1 allele?

3) What is the frequency, in this population, of the B_8 allele?

4) What are the estimated frequencies of the four possible haplotypes (A_1-B_8, A_1-**b**, **a**-B_8, **a-b**) in this population?

5) Under equilibrium conditions, what should be the frequencies of the above haplotypes?

6) What is the value of the linkage disequilibrium between gene A_1 and B_8 from loci **A** and **B** of HLA-system?

Solution of exercise:

1) In the absence of association between the A_1 and B_8 antigens, the expected frequencies of A_1+B_8+, A_1+B_8-, A_1-B_8+, and A_1-B_8- individuals (shown here with the observed) are

	N(e)	N(o)
A_1+B_8+	$(n_1+n_2)(n_1+n_3)/N$ = 145.06	n_1 = 376
A_1+B_8-	$(n_1+n_2)(n_2+n_4)/N$ = 465.94	n_2 = 235
A_1-B_8+	$(n_3+n_4)(n_1+n_3)/N$ = 321.94	n_3 = 91
A_1-B_8-	$(n_3+n_4)(n_2+n_4)/N$ = 1034.06	n_4 = 1265

The value of the chi-squared test is

$$\chi^2 \text{ (1 d.f.)} = \Sigma_i\{[N_i(o)-N_i(e)]^2/N_i(e)\}$$
$$= \Sigma_i\{[(N_i(o)]^2/N_i(e)\} - N$$

$$= (n_1n_4-n_2n_3)^2N/[(n_1+n_2)(n_1+n_3)(n_2+n_4)(n_3+n_4)]$$
$$= 699.35 >>> 3.841 ,$$

indicating thus a very significant association between the two antigens.

2) The frequencies of the gene $\mathbf{A_1}$ and of its "allele" **a** are calculated as follows:

$$P(a) = q_1 = \sqrt{[(n_3+n_4)/N]} = 0.8303$$
$$P(A_1) = p_1 = 1 - q_1 = 1 - \sqrt{[(n_3+n_4)/N]} = 0.1697 .$$

3) And the frequencies of the gene $\mathbf{B_8}$ and of its "allele" **b** as:

$$P(b) = q_2 = \sqrt{[(n_2+n_4)/N]} = 0.8733$$
$$P(B_8) = p_2 = 1 - q_2 = 1 - \sqrt{[(n_2+n_4)/N]} = 0.1267 .$$

4) Since

$$P(a\text{-}B_8) + P(a\text{-}b) = q_1$$
$$P(A_1\text{-}b) + P(a\text{-}b) = q_2$$
$$P(A_1\text{-}B_8) = 1 - [P(A_1\text{-}b) + P(a\text{-}B_8) + P(a\text{-}b)] ,$$

it comes out that the inferred frequencies of the four possible haplotypes are

$$P(a\text{-}b) = \sqrt{(n_4/N)} = 0.8019$$
$$P(a\text{-}B_8) = q_1 - \sqrt{(n_4/N)} = 0.0283$$
$$P(A_1\text{-}b) = q_2 - \sqrt{(n_4/N)} = 0.0713$$
$$P(A_1\text{-}B_8) = 1 - q_1 - q_2 + \sqrt{(n_4/N)} = 0.0984 .$$

5) Under equilibrium conditions, the expected frequencies of these haplotypes should be

$$P'(a\text{-}b) = q_1q_2 = 0.7251$$
$$P'(a\text{-}B_8) = q_1(1 - q_2) = 0.1052$$
$$P'(A_1\text{-}b) = (1 - q_1)q_2 = 0.1482$$
$$P'(A_1\text{-}B_8) = (1 - q_1)(1 - q_2) = 0.0215 .$$

6) The linkage disequilibrium values of these four haplotypes are therefore

$$\Delta(a\text{-}b) = P(a\text{-}b) - P'(a\text{-}b) = +0.0769$$
$$\Delta(a\text{-}B_8) = P(a\text{-}B_8) - P'(a\text{-}B_8) = -0.0769$$
$$\Delta(A_1\text{-}b) = P(A_1\text{-}b) - P'(A_1\text{-}b) = -0.0769$$
$$\Delta(A_1\text{-}B_8) = P(A_1\text{-}B_8) - P(A_1\text{-}B_8) = +0.0769.$$

The following BASIC code performs all the calculations indicated above, as shown by the screen printout appended to it:

```
REM PROGRAM FILENAME HLAHAPL2
REM HLA SYSTEM HAPLOTYPE ESTIMATION
DEFDBL A-Z: CLS : LOCATE 10: C$ = "NO. OF INDIVIDUALS "
INPUT "FIRST  ANTIGEN IDENTIFICATION = "; A$
INPUT "SECOND ANTIGEN IDENTIFICATION = "; B$: PRINT
PRINT C$ + A$ + "(+)/" + B$ + "(+) = "; : INPUT "", N1
PRINT C$ + A$ + "(+)/" + B$ + "(-) = "; : INPUT "", N2
PRINT C$ + A$ + "(-)/" + B$ + "(+) = "; : INPUT "", N3
PRINT C$ + A$ + "(-)/" + B$ + "(-) = "; : INPUT "", N4: CLS
```

```
PRINT "   " + C$ + A$ + "(+)/" + B$ + "(+) = "; : PRINT USING "#####"; N1
PRINT "   " + C$ + A$ + "(+)/" + B$ + "(-) = "; : PRINT USING "#####"; N2
PRINT "   " + C$ + A$ + "(-)/" + B$ + "(+) = "; : PRINT USING "#####"; N3
PRINT "   " + C$ + A$ + "(-)/" + B$ + "(-) = "; : PRINT USING "#####"; N4
N = N1 + N2 + N3 + N4: PRINT "   " + C$ + "TESTED      = ";
PRINT USING "#####"; N: PRINT "GENE FREQUENCIES"
Q1 = SQR((N3 + N4) / N): P1 = 1 - Q1: Q2 = SQR((N2 + N4) / N): P2 = 1 - Q2
PRINT "   P(" + A$ + ") = "; : PRINT USING "#.####"; P1
PRINT "   P(" + B$ + ") = "; : PRINT USING "#.####"; P2
PA0B0 = SQR(N4 / N): PA0B1 = 1 - P1 - PA0B0: PA1B0 = 1 - P2 - PA0B0
PA1B1 = P1 + P2 + PA0B0 - 1: PRINT "INFERRED HAPLOTYPE FREQUENCIES"
PRINT "   P(" + A$ + "/" + B$ + ") = "; : PRINT USING "#.####"; PA1B1
PRINT "   P(" + A$ + "/ -) = "; : PRINT USING "#.####"; PA1B0
PRINT "   P(- /" + B$ + ") = "; : PRINT USING "#.####"; PA0B1
PRINT "   P(- / -) = "; : PRINT USING "#.####"; PA0B0
PRINT "EXPECTED HAPLOTYPE FREQUENCIES"
PRINT "   P(" + A$ + "/" + B$ + ") = "; : PRINT USING "#.####"; P1 * P2
PRINT "   P(" + A$ + "/ -) = "; : PRINT USING "#.####"; P1 * Q2
PRINT "   P(- /" + B$ + ") = "; : PRINT USING "#.####"; Q1 * P2
PRINT "   P(- / -) = "; : PRINT USING "#.####"; Q1 * Q2
PRINT "LINKAGE DISEQUILIBRIUM VALUES"
PRINT "   D(" + A$ + "/" + B$ + ") = "; : PRINT USING "#.####"; PA1B1 - P1 * P2
PRINT "   D(" + A$ + "/ -) = "; : PRINT USING "#.####"; PA1B0 - P1 * Q2
PRINT "   D(- /" + B$ + ") = "; : PRINT USING "#.####"; PA0B1 - Q1 * P2
PRINT "   D(- / -) = "; : PRINT USING "#.####"; PA0B0 - Q1 * Q2
```

```
   NO. OF INDIVIDUALS A1(+)/B8(+) =   376
   NO. OF INDIVIDUALS A1(+)/B8(-) =   235
   NO. OF INDIVIDUALS A1(-)/B8(+) =    91
   NO. OF INDIVIDUALS A1(-)/B8(-) =  1265
   NO. OF INDIVIDUALS TESTED      =  1967
GENE FREQUENCIES
   P(A1) = 0.1697
   P(B8) = 0.1267
INFERRED HAPLOTYPE FREQUENCIES
   P(A1/B8) = 0.0984
   P(A1/ -) = 0.0713
   P(- /B8) = 0.0283
   P(- / -) = 0.8019
EXPECTED HAPLOTYPE FREQUENCIES
   P(A1/B8) = 0.0215
   P(A1/ -) = 0.1482
   P(- /B8) = 0.1052
   P(- / -) = 0.7251
LINKAGE DISEQUILIBRIUM VALUES
   D(A1/B8) = 0.0769
   D(A1/ -) = -.0769
   D(- /B8) = -.0769
   D(- / -) = 0.0769
```

LINKAGE DISEQUILIBRIUM CALCULATIONS

In the lines that follow the notation used by Hill (Hill WG. Estimation of linkage disequilibrium in randomly mating populations. **Heredity 33 :** 229-239, 1974) and Weir & Cockerham (Weir BS & Cockerham CC. Estimation of linkage disequilibrium in randomly mating populations. **Heredity 42 :** 105-111, 1979) is retained whenever possible.

1) ABSENCE OF DOMINANCE

We will begin with the situation in which there is no dominance so that the three possible genotypes given by the alleles of each of the two loci [**(AA, Aa, aa)** and **(BB, Bb, bb)**] are easily distinguishable. Let N_{11}, N_{12}, etc be the observed numbers of genotypes **AB/AB**, **AB/Ab**, etc as shown below:

	BB	Bb	bb	
AA	N_{11}	N_{12}	N_{13}	$N_{1.}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2.}$
aa	N_{31}	N_{32}	N_{33}	$N_{3.}$
	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

That is,

N(AABB) = N(AB/AB) = N_{11}
N(AABb) = N(AB/Ab) = N_{12}
N(AAbb) = N(Ab/Ab) = N_{13}
N(AaBB) = N(AB/aB) = N_{21}
N(AaBb) = N(AB/ab) + N(Ab/aB) = $N_{22} = N'_{22} + N''_{22}$
N(Aabb) = N(Ab/ab) = N_{23}
N(aaBB) = N(aB/aB) = N_{31}
N(aaBb) = N(aB/ab) = N_{32}
N(aabb) = N(ab/ab) = N_{33}

Under panmictic equilibrium, the expected genotype frequencies are

	BB	Bb	bb	
AA	f_{11}^2	$2f_{11}f_{12}$	f_{12}^2	p^2
Aa	$2f_{11}f_{21}$	$2f_{11}f_{22} + 2f_{12}f_{21}$	$2f_{12}f_{22}$	$2p(1-p)$
aa	f_{21}^2	$2f_{21}f_{22}$	f_{22}^2	$(1-p)^2$
	q^2	$2q(1-q)$	$(1-q)^2$	1

where

$f_{11} = f(AB) = f(AABB) + f(AABb)/2 + f(AaBB)/2 + f(AB/ab)/2$
$f_{12} = f(Ab) = f(AAbb) + f(AABb)/2 + f(Aabb)/2 + f(Ab/aB)/2$
$f_{21} = f(aB) = f(aaBB) + f(AaBB)/2 + f(aaBb)/2 + f(Ab/aB)/2$
$f_{22} = f(ab) = f(aabb) + f(Aabb)/2 + f(aaBb)/2 + f(AB/ab)/2$

are the haplotype frequencies to be estimated from the data set and

$p = f(A) = (2N_{11} + 2N_{12} + 2N_{13} + N_{21} + N_{22} + N_{23})/2N$
$= (2N_{1.} + N_{2.})/2N = f_{11} + f_{12}$
$1-p = f(a) = 1 - f(A) = f_{21} + f_{22}$

$q = f(B) = (2N_{11} + 2N_{21} + 2N_{31} + N_{12} + N_{22} + N_{32})/2N$
$= (2N_{.1} + N_{.2})/2N = f_{11} + f_{21}$
$1-q = f(b) = 1 - f(B) = f_{12} + f_{22}.$

Since $N(AaBb) = N(AB/ab) + N(Ab/aB) = N'_{22} + N''_{22} = N_{22}$, N'_{22} can take any value from 0 to N_{22} while N''_{22} varies from N_{22} to 0. Therefore, the lower limit for **f(AB)** is necessarily

$f_1(AB) = (2N_{11}+N_{12}+N_{21})/2N$

while its upper limit is given (also necessarily) by

$f_u(AB) = (2N_{11}+N_{12}+N_{21}+N_{22})/2N.$

In the absence of linkage disequilibrium between the genes from loci **(A,a)** and **(B,b)**, the estimate of **f(AB)** is given simply by

$f_0(AB) = (2N_{11}+N_{12}+N_{21}+N_{22}/2)/2N.$

Since the coefficient of linkage disequilibrium is defined as

$\Delta(AB) = f(AB) - f(A).f(B) = f_{11} - pq,$

it comes out that

$\Delta(AB) = f_{11} - (f_{11}+f_{12})(f_{11}+f_{21})$
$= f_{11}(f_{11}+f_{12}+f_{21}+f_{22}) - (f_{11}+f_{12})(f_{11}+f_{21})$
$= f_{11}.f_{22}-f_{12}.f_{21}$
$= N'_{22}/N - N''_{22}/2N = (N'_{22}-N''_{22})/2N.$

Assuming that the marginal frequencies for both one-locus genoypes [**(AA, Aa, aa)** and **(BB, Bb, bb)**] are in Hardy-Weinberg proportions, the likelihood function is given by

$P = N!/(N_{11}!...N_{33}!).(f_{11}^2)^{N11}.(2f_{11}f_{12})^{N12}.(f_{12}^2)^{N13}$
$. (2f_{11}f_{21})^{N21}.(2f_{11}f_{22}+2f_{12}f_{21})^{N22}.(2f_{12}f_{22})^{N23}$
$. (f_{21}^2)^{N31}.(2f_{21}f_{22})^{N32}.(f_{22}^2)^{N33},$

so that the frequencies f_{11}, f_{12} and f_{21} can be estimated by maximizing the likelihood function in logarithmic form

$$L = \ln P = \text{const.} + \sum X_{ij}.\ln f_{ij} + N_{22}.\ln(f_{11}.f_{22}-f_{12}.f_{21})$$
$$= \text{const.} + X_{11}.\ln f_{11} + X_{12}.\ln f_{12} + X_{21}.\ln f_{21}$$
$$+ X_{22}.\ln f_{22} + N_{22}.\ln(f_{11}.f_{22}-f_{12}.f_{21})$$
$$= \text{const.} + X_{11}.\ln f_{11} + X_{12}.\ln f_{12} + X_{21}.\ln f_{21}$$
$$+ X_{22}.\ln (1-f_{11}-f_{12}-f_{21})$$
$$+ N_{22}.\ln[f_{11}(1-f_{11}-f_{12}-f_{21})-f_{12}.f_{21}],$$

where $X_{11} = 2N_{11} + N_{12} + N_{21}$
$X_{12} = 2N_{13} + N_{12} + N_{23}$
$X_{21} = 2N_{31} + N_{21} + N_{32}$
$X_{22} = 2N_{33} + N_{23} + N_{32}$.

The partial derivatives $\partial L/\partial f_{11}$, $\partial L/\partial f_{12}$ and $\partial L/\partial f_{21}$ are

$$\partial L/\partial f_{11} = X_{11}/f_{11} - X_{22}/(1-f_{11}-f_{12}-f_{21})$$
$$+ N_{22}(1-2f_{11}-f_{12}-f_{21})/[f_{11}(1-f_{11}-f_{12}-f_{21})+f_{12}f_{21}]$$

$$\partial L/\partial f_{12} = X_{12}/f_{12} - X_{22}/(1-f_{11}-f_{12}-f_{21})$$
$$+ N_{22}(f_{21}-f_{11})/[f_{11}(1-f_{11}-f_{12}-f_{21})+f_{12}f_{21}]$$

$$\partial L/\partial f_{21} = X_{21}/f_{21} - X_{22}/(1-f_{11}-f_{12}-f_{21})$$
$$+ N_{22}(f_{12}-f_{11})/[f_{11}(1-f_{11}-f_{12}-f_{21})+f_{12}f_{21}].$$

The estimates f_{11}, f_{12} and f_{21} are obtained by maximizing the function L, that is, they are the solutions of the set of linearly independent equations

$$\{\partial L/\partial f_{11} = 0,\ \partial L/\partial f_{12} = 0,\ \partial L/\partial f_{21} = 0\}.$$

Since it is not possible to obtain explicit solutions for this set of equations, a numerical method as the generalized Newton-Raphson iterative procedure is used:

$$(f_{ij})_{n+1} = (f_{ij})_n + ((-\partial(\partial L/\partial f_{ij})/\partial f_{ij})^{-1}.(\partial L/\partial f_{ij}))_n$$
$$= (f_{ij})_n + ((-\partial^2 L/\partial f_{ij}^2)^{-1}.(\partial L/\partial f_{ij}))_n$$
$$= (f_{ij})_n + ((V_{ij}).(\partial L/\partial f_{ij}))_n,$$

where $(f_{ij})_n$ is the column vector (at the nth iteration)

$(f_{11}, f_{12}, f_{21})^T$,

$(\partial L/\partial f_{ij})_n$ is the column vector, at iteration n, of partial derivatives

$(\partial L/\partial f_{11}, \partial L/\partial f_{12}, \partial L/\partial f_{21})^T$ and

$(-\partial^2 L/\partial f_{ij}2)_n$ is the variance-covariance matrix (also at iteration n)

$$\begin{matrix} V_{11} & V_{12} & V_{13} \\ (V_{21} & V_{22} & V_{23}) \\ V_{31} & V_{32} & V_{33} \end{matrix} = \begin{matrix} \text{VAR}(f_{11}) & \text{COV}(f_{11},f_{12}) & \text{COV}(f_{11},f_{21}) \\ (\text{COV}(f_{12},f_{11}) & \text{VAR}(f_{12}) & \text{COV}(f_{12},f_{21})) \\ \text{COV}(f_{21},f_{11}) & \text{COV}(f_{21},f_{12}) & \text{VAR}(f_{21}) \end{matrix}$$

$$= \begin{matrix} \text{VAR}(f_{11}) & \text{COV}(f_{11},f_{12}) & \text{COV}(f_{11},f_{21}) \\ (\text{COV}(f_{11},f_{12}) & \text{VAR}(f_{12}) & \text{COV}(f_{12},f_{21})) \\ \text{COV}(f_{11},f_{21}) & \text{COV}(f_{12},f_{21}) & \text{VAR}(f_{21}) \end{matrix}$$

$$= \begin{pmatrix} -\partial^2 L/\partial f_{11}^2 & -\partial^2 L/\partial f_{11}\partial f_{12} & -\partial^2 L/\partial f_{11}\partial f_{21} \\ -\partial^2 L/\partial f_{11}\partial f_{12} & -\partial^2 L/\partial f_{12}^2 & -\partial^2 L/\partial f_{12}\partial f_{21} \\ -\partial^2 L/\partial f_{11} df_{21} & -\partial^2 L/\partial f_{12}\partial f_{21} & -\partial^2 L/\partial f_{21}^2 \end{pmatrix}^{-1}$$

The literal values of the second derivatives are:

$$\partial^2 L/\partial f_{11}^2 = - x_{11}/f_{11}^2 - x_{22}/f_{22}^2 - N_{22}(f_{11}^2+2f_{12}f_{21}+f_{22}^2)/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{12}^2 = - x_{12}/f_{12}^2 - x_{22}/f_{22}^2 - N_{22}(f_{21}-f_{11})^2/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{21}^2 = - x_{21}/f_{21}^2 - x_{22}/f_{22}^2 - N_{22}(f_{12}-f_{11})^2/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{11}\partial f_{12} = - x_{22}/f_{22}^2 - N_{22}(f_{11}^2-f_{11}f_{21}+f_{12}f_{21}+f_{21}f_{22})/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{11}\partial f_{21} = - x_{22}/f_{22}^2 - N_{22}(f_{11}^2-f_{11}f_{12}+f_{12}f_{21}+f_{12}f_{22})/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{12}\partial f_{21} = - x_{22}/f_{22}^2 - N_{22}(2f_{11}^2-f_{11})/(f_{11}f_{22}+f_{12}f_{21})^2$$
$$\partial^2 L/\partial f_{12} df_{11} = \partial^2 L/\partial f_{11}\partial f_{12}$$
$$\partial^2 L/\partial f_{21}\partial f_{11} = \partial^2 L/\partial f_{11}\partial f_{21}$$
$$\partial^2 L/\partial f_{21}\partial f_{12} = \partial^2 L/\partial f_{12}\partial f_{21} ,$$

where $\mathbf{f_{22} = 1 - f_{11} - f_{12} - f_{21}}$.

Since, at equilibrium, all double heterozygotes combined (**AB/ab** and **Ab/aB**) produce all types of gametes (**AB**, **Ab**, **aB** and **ab**) in exactly equal proportions, the following trial values of **f11**, **f12** and **f21** are used for the initial evaluation of the matrices $\mathbf{(\partial L/\partial f_{ij})}$ and $\mathbf{(-\partial^2 L/\partial f_{ij}^2)^{-1}}$ at the beginning of the iteration process:

$$\mathbf{f_{11} = (2N_{11}+N_{12}+N_{21}+N_{22}/2)/2N = (2X_{11}+N_{22})/4N}$$
$$\mathbf{f_{12} = (N_{12}+2N_{13}+N_{23}+N_{22}/2)/2N = (2X_{12}+N_{22})/4N}$$
$$\mathbf{f_{21} = (N_{21}+2N_{31}+N_{32}+N_{22}/2)/2N = (2X_{21}+N_{22})/4N}$$

After convergence has occurred to the final estimates $\mathbf{f_{11}}$, $\mathbf{f_{12}}$ and $\mathbf{f_{21}}$, the value of the estimate $\mathbf{f_{22}}$ is then directly obtained from $\mathbf{f_{22} = 1-f_{11}-f_{12}-f_{21}}$. The variances of the estimates $\mathbf{f_{11}}$, $\mathbf{f_{12}}$ and $\mathbf{f_{21}}$ are taken straightforwardly from the variance-covariance matrix at the final evaluation points. The variance of $\mathbf{f_{22}}$ is then calculated after

$$\mathbf{VAR(f_{22}) = VAR(f_{11}) + 2COV(f_{11},f_{12}) + 2COV(f_{11},f_{21})}$$
$$\mathbf{+ VAR(f_{12}) + 2COV(f_{12},f_{21}) + VAR(f_{21}).}$$

Since $\mathbf{f(A) = p = f(AB) + f(Ab) = f_{11} + f_{12}}$
$\mathbf{f(a) = 1-p = f(aB) + f(ab) = f_{21} + f_{22}}$
$\mathbf{f(B) = q = f(AB) + f(aB) = f_{11} + f_{21}}$ and
$\mathbf{f(b) = 1-q = f(Ab) + f(ab) = f_{12} + f_{22}}$,

the consistency of estimates can be tested by verifying the following property discovered by Fisher: the variance of **f(A)**, **VAR(p)** and that of **f(B)**, **VAR(q)**, are the ordinary binomial gene frequency variances

VAR(p) = VAR(1-p) = p(1-p)/2N and
VAR(q) = VAR(1-q) = q(1-q)/2N.

Should the estimates be consistent, then the numeric values thus obtained should match the quantities

$$\mathbf{VAR(p) = VAR(1-p) = VAR(f_{11}+f_{12})}$$
$$\mathbf{= VAR(f_{11}) + 2COV(f_{11},f_{12}) + VAR(f_{12})}$$

and

$$VAR(q) = VAR(1-q) = VAR(f_{11}+f_{21})$$
$$= VAR(f_{11}) + 2COV(f_{11},f_{21}) + VAR(f_{21})$$

taken from the variance-covariance matrix at the final evaluation point.

The linkage disequilibrium value is finally estimated from

$$\Delta(AB) = f_{11} - pq.$$

The logarithmic likelihood function

$$L = \ln P = \text{const.} + \sum X_{ij}.\ln f_{ij} + N_{22}.\ln(f_{11}.f_{22}-f_{12}.f_{21})$$
$$= \text{const.} + X_{11}.\ln f_{11} + X_{12}.\ln f_{12} + X_{21}.\ln f_{21}$$
$$+ X_{22}.\ln f_{22} + N_{22}.\ln(f_{11}.f_{22}-f_{12}.f_{21})$$

can also be expressed as a function of a single variable (one of the haplotype frequencies, v.g. f_{11}), since $f_{12} = p-f_{11}$, $f_{21} = q-f_{11}$ and $f_{22} = 1-p-q+f_{11}$:

$$L = \ln P = \text{const.} + X_{11}.\ln f_{11} + X_{12}.\ln(p-f_{11})$$
$$+ X_{21}.\ln(q-f_{11}) + X_{22}.\ln(1-p-q+f_{11})$$
$$+ N_{22}.\ln[f_{11}(1-p-q+f_{11})+(p-f_{11})(q-f_{11})].$$

The estimate f_{11} is then the solution of the equation obtained by putting $dL/df_{11} = 0$. Hill (1974), using a 'counting method,' found that the estimate f_{11} is the solution of the cubic equation

$$f_{11} = \{X_{11} + N_{22}.f_{11}(1-p-q+f_{11})/[f_{11}(1-p-q+f_{11})+(p-f_{11})(q-f_{11})]\}/2N.$$

As before, the estimate of the linkage disequilibrium value is obtained straightforwardly from

$$\Delta(AB) = f(AB) - f(A).f(B) = f_{11} - pq.$$

Instead of determining the value of $\Delta(AB)$ after estimating the haplotype frequencies, we can get it directly if we remember that under linkage disequilibrium the frequencies of the four haplotypes **AB**, **Ab**, **aB** and ab can be all expressed as a function of Δ and the constants **p** and **q**:

$$f_{11} = pq + \Delta$$
$$f_{12} = p(1-q) - \Delta$$
$$f_{21} = (1-p)q - \Delta$$
$$f_{22} = (1-p)(1-q) + \Delta,$$

where Δ is the linkage disequilibrium value of haplotypes **AB** or **ab** and **p**, **1-p**, **q** and **1-q** are the frequencies of the pairs of alleles **A**,**a** and **B**,**b**:

	A	a	
B	pq + Δ	(1-p)q - Δ	q
b	p(1-q) - Δ	(1-p)(1-q)+Δ	1-q
	p	1-p	1

If the observed absolute frequencies of the genotypes **AB/AB**, ..., **ab/ab** are respectively N_{11}, ..., N_{33} in a total of **N** sampled individuals, under the assumption of panmixia the expected quantities are:

GENOTYPE	OBS.ABS.FREQ.	EXP.ABS.FREQ.
AB/AB	N_{11}	$N(pq+\Delta)^2$
AB/Ab	N_{12}	$2N(pq+\Delta)[p(1-q)-\Delta]$
Ab/Ab	N_{13}	$N[p(1-q)-\Delta]^2$
AB/aB	N_{21}	$2N(pq+\Delta)[(1-p)q-\Delta]$
AB/ab + Ab/aB	N_{22}	$2N\{(pq+\Delta)[(1-p)(1-q)+\Delta]$ $+[p(1-q)-\Delta][(1-p)q-\Delta]\}$
Ab/ab	N_{23}	$2N[p(1-q)-\Delta][(1-p)(1-q)+\Delta]$
aB/aB	N_{31}	$N[(1-p)q-\Delta]^2$
aB/ab	N_{32}	$2N[(1-p)q-\Delta][(1-p)(1-q)+\Delta]$
ab/ab	N_{33}	$N[(1-p)(1-q)+\Delta]^2$

The likelihood function **L = ln P** is clearly

$$L = \text{const.} + X_{11}.\ln(pq+\Delta) + X_{12}.\ln[p(1-q)-\Delta] + X_{21}.\ln[(1-p)q-\Delta] + X_{22}.\ln[(1-p)(1-q)+\Delta] + N_{22}.\ln\{(pq+\Delta)[(1-p)(1-q)+\Delta] + [p(1-q)-\Delta][(1-p)q-\Delta]\},$$

where X_{11}, X_{12}, X_{21} and X_{22} are the summary measures already defined. The allelic frequencies can be treated as constants, and they are easily estimated by an independent direct counting method:

$$p = (X_{11}+X_{12}+N_{22})/2N,\ 1-p = (X_{21}+X_{22}+N_{22})/2N$$
$$q = (X_{11}+X_{21}+N_{22})/2N,\ 1-q = (X_{12}+X_{22}+N_{22})/2N\ .$$

The first derivative $dL/d\Delta$ has literal value

$$dL/d\Delta = X_{11}/(pq+\Delta) - X_{12}/[p(1-q)-\Delta] - X_{21}/[(1-p)q-\Delta] + X_{22}/[(1-p)(1-q)+\Delta] + N_{22}[4\Delta+(1-2p)(1-2q)]/[2\Delta^2+\Delta(1-2q)(1-2p)+2pq(1-p)(1-q)]$$

whereas the second derivative takes value

$$d^2L/d\Delta^2 = - X_{11}/(pq+\Delta)^2 - X_{12}/[p(1-q)-\Delta]^2 - X_{21}/[(1-p)q-\Delta]^2 - X_{22}/[(1-p)(1-q)+\Delta]^2 - N_{22}[8pq(1-p)(1-q)+1-4p(1-q)-4(1-p)q] / [2\Delta^2+\Delta(1-2q)(1-2p)+2pq(1-p)(1-q)].$$

The estimate Δ is the solution of the equation $dL/d\Delta = 0$. Since this equation has no explicit solution, a numerical method such as the Newton-Raphson procedure is used to obtain it:

$$\Delta_{n+1} = \Delta_n - f(\Delta)_n/f'(\Delta)_n = \Delta_n + (dL/d\Delta)_n.[-(d^2L/d\Delta^2)_n]^{-1} = \Delta_n + (dL/d\Delta)_n.VAR(\Delta)_n.$$

Hill (1974) showed that a suitable starting value for iteration is given by

$$f_{11} = \Delta_0 + pq = (X_{11}-X_{12}-X_{21}+X_{22})/4N + 1/2 - (1-p)(1-q)$$

and therefore

$$\Delta_0 = (X_{11}-X_{12}-X_{21}+X_{22})/4N + 1/2 - (1-p)(1-q) - pq.$$

Now, let **Fo** be the observed numbers and **Fe'** and **Fe"** respectively the expected values under the assumptions of $\Delta = \Delta(AB) = 0$ and $\Delta = \Delta(AB) \neq 0$ (estimated after any of the methods just delineated) as follows:

Fo	Fe'	Fe"
N_{11}	$N(pq)^2$	$N(pq+\Delta)^2$
N_{12}	$2Np^2q(1-q)$	$2N(pq+\Delta)[p(1-q)-\Delta]$
N_{13}	$N[p(1-q)]^2$	$N[p(1-q)-\Delta]^2$
N_{21}	$2Np(1-p)q^2$	$2N(pq+\Delta)[(1-p)q-\Delta]$
N_{22}	$4Np(1-p)q(1-q)$	$2N\{(pq+\Delta)[(1-p)(1-q)+\Delta]$ $+[p(1-q)-\Delta][(1-p)q-\Delta]\}$
N_{23}	$2Np(1-p)(1-q)^2$	$2N[p(1-q)-\Delta][(1-p)(1-q)+\Delta]$
N_{31}	$N[(1-p)q]^2$	$N[(1-p)q-\Delta]^2$
N_{32}	$2N(1-p)^2q(1-q)$	$2N[(1-p)q-\Delta][(1-p)(1-q)+\Delta]$
N_{33}	$N[(1-p)(1-q)]^2$	$N[(1-p)(1-q)+\Delta]^2$

For testing if $\Delta \neq 0$ the following G difference test is then used:

$$G = 2\sum\{Fo.\ln(Fo/Fe')\} - 2\sum\{Fo.\ln(Fo/Fe")\}$$
$$= 2\sum\{Fo[\ln(Fo/Fe')-\ln(Fo/Fe")]\}$$
$$= 2\sum[Fo.\ln(Fe"/Fe')]$$

$$= 2X_{11}.\ln(1+\Delta/pq) + 2X_{12}.\ln[1-\Delta/p(1-q)]$$
$$+ 2X_{21}.\ln[1-\Delta/(1-p)q] + 2X_{22}.\ln[1+\Delta/(1-p)(1-q)]$$
$$+ 2N_{22}.\ln\{(1+\Delta/pq)[1+\Delta/(1-p)(1-q)]$$
$$+ [1-\Delta/p(1-q)][1-\Delta/(1-p)q]\} .$$

This statistics has a chi-squared distribution with 1 d.f. The usual statistics (that asymptotically has also a chi-squared distribution with 1 d.f.) is

$$N\Delta^2/p(1-p)q(1-q).$$

2) DOMINANCE

If there is dominance in both linked loci **A,a** and **B,b**, it comes out that, in a panmictic sample of **N** individuals tested with anti-A and anti-B sera

$$
\begin{aligned}
f(A+B+) &= p^2 + 2pq + 2pr + 2qr + 2ps = 2p - p^2 + 2qr &&= f_1\\
f(A+B-) &= q^2 + 2qs &&= 2q - 2pq - q^2 - 2qr = f_2\\
f(A-B+) &= r^2 + 2rs &&= 2r - 2pr - 2qr - r^2 = f_3\\
f(A-B-) &= s^2 &&= (1-p-q-r)^2 = f_4
\end{aligned}
$$

where **p**, **q**, **r** and **s** are the frequencies of haplotypes **AB**, **Ab**, **aB** and **ab**.

If the observed numbers of **A+B+**, **A+B-**, **A-B+** and **A-B-** individuals are respectively N_1, N_2, N_3 and N_4 then the estimates **p**, **q**, **r** are the solutions of the set of equations

$\{\partial L/\partial p = 0,\ \partial L/\partial q = 0,\ \partial L/\partial r = 0\}$, where

$$
\begin{aligned}
L &= \sum N_i.\ln f_i\\
&= N_1.\ln f_1 + N_2.\ln f_2 + N_3.\ln f_3 + N_4.\ln f_4\\
&= N_1.\ln(2p-p^2+2qr) + N_2.\ln q + N_2.\ln(2-2p-q-2r)\\
&\quad + N_3.\ln r + N_3.\ln(2-2p-2q-r) + 2N_4.\ln(1-p-q-r).
\end{aligned}
$$

The solutions of the set of equations

$$
\begin{aligned}
\partial L/\partial p &= 2N_1(1-p)/(2p-p^2+2qr) - 2N_2/(q+2s) - 2N_3/(r+2s) - 2N_4/s = 0\\
\partial L/\partial q &= 2N_1 r/(2p-p^2+2qr) + 2N_2 s/(q^2+2qs) - 2N_3/(r+2s) - 2N_4/s = 0\\
\partial L/\partial r &= 2N_1 q/(2p-p^2+2qr) - 2N_2/(q+2s) + 2N_3 s/(r^2+2rs) - 2N_4/s = 0,
\end{aligned}
$$

where $s = 1-p-q-r$, are the obvious ones

$$
\begin{aligned}
p &= f(A) + f(B) + \sqrt{(N_4/N)} - 1\\
&= 1 + \sqrt{(N_4/N)} - \sqrt{[(N_3+N_4)/N]} - \sqrt{[(N_2+N_4)/N]} = 1-q-r-s\\
q &= 1 - f(B) - \sqrt{(N_4/N)}\\
&= \sqrt{[(N_2+N_4)/N]} - \sqrt{(N_4/N)} = \sqrt{[(q+s)^2]} - \sqrt{(s^2)}\\
r &= 1 - f(A) - \sqrt{(N_4/N)}\\
&= \sqrt{[(N_3+N_4)/N]} - \sqrt{(N_4/N)} = \sqrt{[(r+s)^2]} - \sqrt{(s^2)}\\
s &= \sqrt{(N_4/N)} = \sqrt{(s^2)}.
\end{aligned}
$$

The linkage disequilibrium value estimate Δ is obtained directly from

$$
\begin{aligned}
\Delta &= f(AB) - f(A).f(B)\\
&= f(ab) - f(a).f(b)\\
&= \sqrt{(N_4/N)} - \sqrt{[(N_2+N_4)(N_3+N_4)]/N}.
\end{aligned}
$$

For testing the hypothesis $\Delta = 0$ the following chi-squared statistics (with 1 d.f.) is used:

$$
\begin{aligned}
\chi^2 &= N_1^2/[N(1-Q_4^2)(1-Q_3^2)] + N_2^2/[NQ_4^2(1-Q_3^2)]\\
&\quad + N_3^2/[NQ_3^2(1-Q_4^2)] + N_4^2/(NQ_4^2Q_3^2) - N\\
&= N_1^2.N/[(N_1+N_3)(N_1+N_2)] + N_2^2.N/[(N_1+N_3)(N_3+N_4)]\\
&= N_3^2.N/[(N_2+N_4)(N_1+N_2)] + N_4^2.N/[(N_2+N_4)(N_3+N_4)] - N\\
&= (N_1N_4-N_2N_3)^2.N/[(N_1+N_2)(N_1+N_3)(N_2+N_4)(N_3+N_4)],
\end{aligned}
$$

where $Q_3 = 1-f(A) = \sqrt{[(N_3+N_4)/N]}$

$Q_4 = 1-f(B) = \sqrt{[(N_2+N_4)/N]}$.

Therefore, the statistics for testing $\Delta = 0$ is equivalent to test absence of association between antigens **A** and **B** in a 2x2 contingency table. Of course the usual continuity correction can be introduced in the above formula, that then takes the form

$$X^2 = [ABS(N_1N_4-N_2N_3)-N/2]^2.N/[(N_1+N_2)(N_1+N_3)(N_2+N_4)(N_3+N_4)].$$

Alternatively, a **G** test (log-likelihood ratio) can be used (and should be preferred since often numbers occurring in some cells of the table are small):

$$\begin{aligned} G \approx \chi^2 &= 4ND^2/[P_3(2-P_3)P_4(2-P_4)] \\ &= 4N\{\sqrt{(N_4/N)}-\sqrt{[(N_3+N_4)(N_2+N_4)]/N}\}^2/[(1-Q_3^2)(1-Q_4^2)] \\ &= 4N\{\sqrt{(N.N_4)}-\sqrt{[(N_3+N_4)(N_2+N_4)]}\}^2/[(N_1+N_2)(N_1+N_3)]. \end{aligned}$$

If **f(AB) = 0** then it comes out that

$$\begin{aligned} f(A+B+) &= 2qr &&= f_1 \\ f(A+B-) &= q^2 + 2qs &&= f_2 \\ f(A-B+) &= r^2 + 2rs &&= f_3 \\ f(A-B-) &= s^2 &&= f_4, \end{aligned}$$

where **q**, **r** and **s** are the frequencies of haplotypes **Ab**, **aB** and **ab**. The estimates **q**, **r**, **s** of haplotype frequencies **f(Ab)**, **f(aB)** and **f(ab)** are then obtained using the standard ABO blood group system estimation method.

GENETIC VARIABILITY AND ITS ASSESSMENT

Population genetics describes the genetical composition of populations and tries to explain its findings through grossly simplified mathematical models. The unit of measure of population genetics is the "gene" or "allele" frequency, defined as

$$p_i = P(a_ia_i) + \tfrac{1}{2} \Sigma_{j>i} P(a_ia_j) \ ,$$

a parameter with approximate binomial variance $var(p_i) = p_i(1-p_i)/2n$ (which takes place exactly when genotype proportions are in Hardy-Weinberg ratios [$P(a_ia_i) = p_i^2$, $P(a_ia_j) = 2p_ip_j$].

The variance, linearized by the square root transformation, can be used to construct approximate confidence intervals (v.g., 95% c.i.) for the "true" population frequency: $i.c.95\%\ p_i : p_i \pm 1.96 \sqrt{[var(p_i)]} = 1.96\ s.e.(p_i)$.

The mensuration of genetic variability is problematic, since organisms have **4,000 - 50,000** structural loci. After some authors, this problem can be circumvented through "random" samples, but what is a random sample of 4,000 - 50,000 loci?

Several indexes have been proposed to describe genetic variability. One of such indexes is simply the **number of alleles** that segregate in a given locus, with the obvious inconvenience that **k** (the number of detectable alleles) is proportional to **n** (the sample size) : $k \propto n$. The probability of detecting in the population a genotype that constains a rare gene is very small, as shown by the following table (adapted from Evett & Weir, 1998), that lists the required sample sizes (**N**) to detect, with a probability of 95%, genotypes with population frequencies (**P**):

P	n
1	1
0.1	30
0.01	300
0.001	3000
0.0001	30000
...	
P	3 / P

In any case, the number of alleles segregating at a given autosomal locus is an important provider of variability per se. Letting k be the number of such alleles and assuming that all alleles occur with the same frequency, $p_i = ... = p_j = 1/k$, it comes out that $P(a_ia_j) = 2p_ip_j = 2.1/k.1/k = 2/k^2$; since the number of different types of possible heterozygotes is given by $k(k-1)/2$, it follows that the probability of an individual being a heterozygote in such a population is given by the expression $P(het) = 2/k^2 \times k(k-1)/2 = (k-1)/k$. As the following table shows, the value of $(k-1)/k$ converges rapidly to 1.

k	1/k	$2/k^2$	k(k-1)/2	(k-1)/k
2	1/2	1/2	1	1/2
3	1/3	2/9	3	2/3
4	1/4	2/16	6	3/4
5	1/5	2/25	10	4/5
...	...	...	...	...
inf.	0	0	inf.	1

Another useful diversity parameter is the so-called "proportion of polymorphic loci." Polymorphic loci are arbitrarily defined as loci that contain at least two polymorphic alleles (alleles with frequency between 0.99 and 0.01, or between 0.95 and 0.05); genes with frequency larger than 0.99 (or 0.95) are known as monomorphic, in contrast with those with a frequency smaller than 0.01 (or 0.05), known as idiomorphic. The detection of polymorphisms suffers from the restraints associated with the probability of genotype detection commented above.

Another diversity parameter -- this a very important one -- is the index known alternatively as gene diversity or heterozigosity (h, H):

$$h = 1 - \Sigma p_i^2 \rightarrow 2n(1-\Sigma p_i^2)/(2n-1)$$
$$p_i = P(a_i a_i) + \tfrac{1}{2} \Sigma_{j>i} P(a_i a_j)$$
$$H = \Sigma h_j / r \ , \ var(H) = var(h)/r$$
$$var(h) = \Sigma (h_j - H)^2/(r-1)$$

The following table shows the overall results obtained with the analysis of 31 enzymatic loci in the fruit fly ***D. willistoni*** (Ayala et al. 1974) and with 11 proteic loci in the rodent ***S. douglasii*** (Smith & Coss 1984):

	No. of sampled loci	Ave. no. of alleles	proportion of polymorph. loci 5%	proportion of polymorph. loci 1%	heterozygosity H
Dw	31	5.4	14/31	24/31	0.177
Sd	11	2.8	4/11	6/11	0.045

Given the problems mentioned above, there exists a copious literature on the methodology necessary to circumvent them all. Since more than 50% of the loci in most species are monomorphic, one expects to find a large variance between loci using any variability index. This suggests the strategy of surveying a large number of loci instead of a large number of individuals in order to obtain more reliable estimates of H; but of course a reasonable number of individuals analyzed per locus makes the variance within loci smaller and the variance between loci more homogeneous. Mutation, selection, migration, and drift, on the other hand, have an opposite effect, making the variance between loci larger than within loci.

INBREEDING

1) Regular systems of inbreeding

The main effect of inbreeding is an increase in the frequency of homozygotes in the population, with a corresponding decrease in heterozygosity. When inbreeding takes place systematically and exclusively among individuals with a close degree of biological relationship, it leads to the distribution of homozygotes in the gene frequencies and therefore to a complete loss of population heterozygosity. These effects can be appreciated easily when we consider a population of plants with self-fertilization. If we define

$$P_0(AA) = d_0$$
$$P_0(Aa) = h_0$$
$$P_0(aa) = r_0$$

as being the initial frequencies of the three possible genotypes determined by a pair of autosomal alleles **A**, **a** , it comes out that in next generation

$$P_1(AA) = d_1 = d_0 + h_0/4$$
$$P_1(Aa) = h_1 = h_0/2$$
$$P_1(aa) = r_1 = r_0 + h_0/4 \ .$$

Exact general solutions in simple analytical form are easily obtained for these first-order difference equations:

$$P_n(AA) = d_n = d_0 + h_0/2 - h_0/2^{n+1} = p - h_0/2^{n+1}$$

$$P_n(Aa) = h_n = h_0/2^{n+1}$$

$$P_n(aa) = r_n = r_0 + h_0/2 - h_0/2^{n+1} = q - h_0/2^{n+1} \ .$$

The limits (as **n** tends to infinity) of the above expressions are clearly

$$P(AA) = d = d_0 + h_0/4 + h_0/8 + h_0/16 + h_0/32 + \ldots = d_0 + h_0/2 = p$$
$$P(Aa) = h = h_0 - h_0/2 - h_0/4 - h_0/8 - h_0/16 - \ldots = h_0 - h_0 = 0$$
$$P(aa) = r = r_0 + h_0/4 + h_0/8 + h_0/16 + h_0/32 + \ldots = r_0 + h_0/2 = q \ .$$

The frequencies **p** and **q** are constant quantities (therefore independent from **n**), as we show below:

$$p_1 = d_1 + h_1/2 = (d_0+h_0/4) + (h_0/2)/2 = d_0 + h_0/2 = p_0 = \ldots = p$$

and therefore

$$q_1 = 1 - p_1 = q_0 = 1 - p_0 = \ldots = q \ .$$

After a large number of generations (that is, when **n** tends to infinity), the population tends to equilibrium. The process takes place without alterations in gene frequencies and with the heterozygote frequency being halved each generation of self-fertilization.

For other systems of continued and exclusive inbreeding among close relatives (full sibs, double first cousins, quadruple second cousins and octuple third cousins) the population heterozygosity decreases after

$h_{n+2} = h_{n+1}/2 + h_n/4$
$h_{n+3} = h_{n+2}/2 + h_{n+1}/4 + h_n/8$
$h_{n+4} = h_{n+3}/2 + h_{n+2}/4 + h_{n+1}/8 + h_n/16$
$h_{n+5} = h_{n+4}/2 + h_{n+3}/4 + h_{n+2}/8 + h_{n+1}/16 + h_n/32$.

In all these systems the equilibrium frequency of heterozygotes is zero. When crossings occur exclusively among individuals with a biological relationship more distant than that presented by first cousins, the decrease in the population heterozygosity takes place very slowly, and at equilibrium the frequency of heterozygotes tends to a limit different from zero but in all instances smaller than **2pq**, the expected frequency of heterozygotes under a random mating system.

The derivation of the recursion relations shown above is quite cumbersome. In the lines below we show just the derivation of the formula for the heterozygote frequency in a system of matings exclusively among full sibs.

Six different types of matings occur in any population, if we are considering an autosomal locus with two alleles:

a) **AA x AA** matings, whose only progeny is of type **AA**;
b) **AA x Aa** matings, that yield progeny **AA + Aa (1:1)**;
c) **AA x aa** matings, whose only progeny is of type **Aa**;
d) **Aa x Aa** matings, that yield the three possible genotyypes **AA + Aa + aa** in the proportions **1:2:1**;
e) **Aa x aa** matings, that yield progeny **Aa + aa (1:1)**;
f) **aa x aa** matings, whose only progeny is of type **aa**.

If matings are permitted to occur just within sibships, it is not difficult to determine the recursion relations between the matings in two successive generations, using the table shown below:

Matings (n)	Sibships (n+1)	Matings (n+1)	Frequencies
AA x AA	AA (1)	AA x AA	1
AA x Aa	AA + Aa (1:1)	AA x AA	1/4
		AA x Aa	1/2
		Aa x Aa	1/4
AA x aa	Aa (1)	Aa x Aa	1
Aa x Aa	AA + Aa + aa (1:2:1)	AA x AA	1/16
		AA x Aa	1/4
		AA x aa	1/8
		Aa x Aa	1/4
		Aa x aa	1/4
		aa x aa	1/16
Aa x aa	Aa + aa (1:1)	Aa x Aa	1/4
		Aa x aa	1/2
		aa x aa	1/4
aa x aa	aa (1)	aa x aa	1

If we call u_n, v_n, w_n, x_n, y_n, z_n the respective frequencies of **AA x AA**, **AA x Aa**, **AA x aa**, **Aa x Aa**, **Aa x aa**, and **aa x aa** matings in generation n, inspection of the above table shows clearly that

$$u_{n+1} = u_n + v_n/4 + x_n/16$$
$$vn+1 = v_n/2 + x_n/4$$
$$w_{n+1} = x_n/8$$
$$x_{n+1} = v_n/4 + w_n + x_n/4 + y_n/4$$
$$y_{n+1} = x_n/4 + y_n/2$$
$$z_{n+1} = x_n/16 + y_n/4 + z_n \ ,$$

or, in matrix form,

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \\ w_{n+1} \\ x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 1/4 & 0 & 1/16 & 0 & 0 \\ 0 & 1/2 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/8 & 0 & 0 \\ 0 & 1/4 & 1 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 & 1/2 & 0 \\ 0 & 0 & 0 & 1/16 & 1/4 & 1 \end{pmatrix} . \begin{pmatrix} u_n \\ v_n \\ w_n \\ x_n \\ y_n \\ z_n \end{pmatrix}$$

The frequency of heterozygotes in generation **n+1** is obviously

$$h_{n+1} = v_n/2 + w_n + x_n/2 + y_n/2 \ ;$$

and in generations **n+2** and **n+3** ,

$$h_{n+2} = v_{n+1}/2 + w_{n+1} + x_{n+1}/2 + y_{n+1}/2$$
$$= 3v_n/8 + w_n/2 + x_n/2 + 3y_n/8$$

and

$$h_{n+3} = v_{n+2}/2 + w_{n+2} + x_{n+2}/2 + y_{n+2}/2$$
$$= 3v_{n+1}/8 + w_{n+1}/2 + x_{n+1}/2 + 3y_{n+1}/8 =$$
$$= 5v_n/16 + w_n/2 + 3x_n/8 + 5y_n/16 \ ,$$

respectively. Comparing the expressions above we get immediately

$$h_{n+3} = h_{n+2}/2 + h_{n+1}/4 \ .$$

Therefore the recursion equation for the frequency of heterozygotes is

$$h_{n+2} = h_{n+1}/2 + h_n/4.$$

For large values of **n** the heterozygosity of the population in a given generation is **80.9%** of that of the previous generation (in contrast with the rate of **50%** for self-fertilization systems). We obtain this value dividing both sides of the recursion equation

$h_n = h_{n-1}/2 + h_{n-2}/4$ by h_{n-1} ;

we get then

$$h_n/h_{n-1} = 1/2 + h_{n-2}/4h_{n-1} \ ;$$

calling **r** the limit of h_n/h_{n-1} as **n** tends to infinity, it comes out that, for sufficiently large values of **n**,

$r = 1/2 + 1/4r$ or $4r^2 - 2r - 1 = 0$.

The positive root of the above quadratic equation (which is the characteristic equation of the recurrence equation $h_{n+2} - h_{n+1}/2 - h_n/4 = 0$) is

$r = (1+\sqrt{5})/4 = 0.809$.

The other possible solution of the above equation,

$r' = (1-\sqrt{5})/4 = -0.309$,

is non admissible, since **r** is the limit of h_n/h_{n-1} as **n** tends to infinity and h_n is equal to or greater than zero for any value **n** might take; **r** would therefore never have a negative sign.

A numerical example of what happens to the frequency of heterozygotes and to the ratio h_n/h_{n-1} in a population where matings occur only between sibs is shown below (followed by the BASIC code used for generating the table values), taking $h_0 = 1$ and $h_1 = 0.5$ as initial conditions:

n	h_n	h_n/h_{n-1}
0	1.00000	-
1	0.50000	0.50000
2	0.50000	1.00000
3	0.37500	0.75000
4	0.31250	0.83333
5	0.25000	0.80000
6	0.20313	0.81250
7	0.16406	0.80769
8	0.13281	0.80952
9	0.10742	0.80882
10	0.08691	0.80909
11	0.07031	0.80899
12	0.05688	0.80903
13	0.04602	0.80901
14	0.03723	0.80902
15	0.03012	0.80902
16	0.02437	0.80902
17	0.01971	0.80902
18	0.01595	0.80902
19	0.01290	0.80902
20	0.01044	0.80902
...	.....	.....
inf.	0.00000	0.80902

```
REM PROGRAM FILENAME INBREE02.BAS
REM LIMIT OF HN+1/HN IN A SIB MATING SYSTEM
DEFDBL A-Z: CLS : DIM H(20): H(0) = 1: H(1) = .5
PRINT USING "###   "; 0; : PRINT USING "#.#####   "; H(I); : PRINT "   -"
I = 1: GOSUB PRINTOUT
FOR I = 2 TO 20
 H(I) = H(I - 1) / 2 + H(I - 2) / 4
GOSUB PRINTOUT: NEXT I
PRINT "inf.  "; : PRINT USING "#.#####   "; 0; :
PRINT USING "#.#####   "; (1 + SQR(5)) / 4: END
```

```
PRINTOUT:
PRINT USING "###   "; I; : PRINT USING "#.#####   "; H(I); H(I) / H(I - 1)
RETURN
```

The recurrence relation $h_n = h_{n-1}/2 + h_{n-2}/4$ is linear and admits therefore an exact general solution in simple analytical form. This general solution has the form

$$h_n = C_1.r_1^n + C_2.r_2^n ,$$

where

$$r_1 = (1+\sqrt{5})/4 = 0.809$$

$$r_2 = (1-\sqrt{5})/4 = -0.309$$

$$C_1 = (h_1-h_0.r_2)/(r_1-r_2)$$

$$C_2 = (h_0.r_1-h_1)/(r_1-r_2).$$

Since in modulus both r_1 and r_2 are less than unity, it comes out that for large values of n r_1^n and r_2^n tend to zero, and therefore at equilibrium $h = 0$. Since in modulus r_1 is greater than r_2, r_2^n approaches zero faster than r_1^n; consequently, for large n,

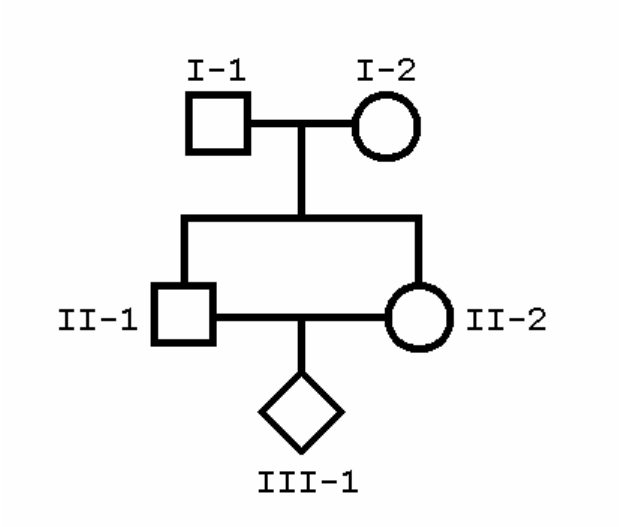
$h_n = C_1.r_1^n$ approximately;

therefore the limit of expression h_n/h_{n-1} as n tends to infinity is really $r_1 = 0.809$:

$$r_1 = \lim_{n\to\infty}(h_n/h_{n-1}) = (C_1/C_1).r_1^n/r_1^{n-1} = r_1.$$

2) Probability of homozygosis for the offspring of consanguineous parents and of sharing of identical genes by two relatives (derivation of the coefficients of inbreeding and common identity)

In panmictic populations, the frequencies of **AA**, **Aa**, **aa** individuals (where **A** and **a** are two alleles segregating at an autosomal locus) are respectively p^2, $2pq$, q^2. These are therefore the probabilities for a child born to unrelated parents of having a genotype respectively **AA**, **Aa**, **aa**. We shall determine now what are the corresponding probabilities for **AA**, **Aa**, **aa** individuals born to relatives. In order to achieve that, we shall consider, for the sake of simplicity, the offspring of a sib mating:



The child (**III-1**) of the above brother (**II-1**) - sister (**II-2**) mating can be homozygous as to the alleles of a given locus by receiving each one from each one of the two different grandparents (homozygosis by independent union of gametes); however, he can also be homozygous by receiving, through both parents (**II-1** and **II-2**), the same gene present in the grandparent **I-1** or **1-2**. This second type of homozygosis is called homozygosis by common descent or autozygosis (this last term having been coined by Cotterman).

In relation to the alleles of a given locus, the probability of autozygosis is, for children born to sibs, **1/4**. In fact, the grandparent **I-1** and the grandparent **I-2** present, each one, two alleles at the locus (total of 4 genes) and the probability of autozygosis for individual **III-I** is **1/16** for each one of these genes. The figure of 1/4 is obtained multiplying 1/16 times 4 : **F = 4 x 1/16 = 1/4**.

The probability of **III-1** being autozygous **AA** is **p/4** and that of being autozygous **aa** is **q/4**. In fact, grandparent **I-1** (the same reasoning is valid for grandparent **I-2**) can be **AA**, **Aa** or **aa** with probabilities p^2, **2pq** and q^2, respectively. If the individual **I-1** were **AA** (let us denote his genotype by A_1A_2 in order to differentiate between the gene A_1 received from his father and the gene A_2 received from his mother), the probability that **III-1** is born A_1A_1 is **1/16** and that he is A_2A_2 is also **1/16**; therefore the probability of **III-1** being autozygous **AA** given that the grandparent **I-1** is **AA** is **1/8**. If the grandparent **I-1** is **Aa**, the probabilities of the child **III-1** being born **AA** and **aa** have each the same value of **1/16** (this value of **1/16** arises of course from **1/2 x 1/2 x 1/2 x 1/2**, that is the probability of any gene being transmitted by **I-1** to both **II-1** and **II-2**, and from these to **III-1**). If the grandparent **I-1** is **aa**, the probability of **III-1** being autozygote for any one of these two genes is also **1/8** (as in the case of the **A** allele). Since the probabilities of **I-1** being **AA**, **Aa** or **aa** are p^2, **2pq** and q^2, it comes out that **III-1** has a probability $p^2/8 + 2pq/16 = p/8$ of being autozygote **AA** and a probability $q^2/8 + 2pq/16 = q/8$ of being autozygous **aa**.

Since the grandchild **III-1** can be autozygous **AA** or **aa** by receiving the alleles from the grandparent **I-2**, who has the same probabilities p^2, **2pq** and q^2 of being **AA**, **Aa** or **aa**, it is clear that the probabilities for a

child born to sibs to be an autozygote **AA** or an autozygote **aa** are respectively **p/4** and **q/4**.

We have just verified that the probability of a given locus of the offspring of sibs being autozygous is **p/4 + q/4 = 1/4**. There is, therefore, a probability of **3/4** of not being so; that is, in 3 out of 4 times the locus shall have a constitution **AA** or **aa** (homozygosis by independent union of gametes) or **Aa** (heterozygosis). Therefore the probabilities associated with the genotypes **AA**, **Aa** and **aa** by independent union of gametes are $\mathbf{3p^2/4}$, **3pq/2**, $\mathbf{3q^2/4}$.

Of course we can get these figures using the same reasoning shown some paragraphs above for calculating the chances of autozygosity. In order to differentiate between all genes present in grandparents **I-1** and **I-2**, let $\mathbf{A_1A_2}$, $\mathbf{A_3a_1}$, $\mathbf{a_2a_3}$ be the possible genotypes of individual **I-1**; and $\mathbf{A_4A_5}$, $\mathbf{A_6a_4}$, $\mathbf{a_5a_6}$ the corresponding ones of individual **I-2**.

Individual **III-1** has therefore the following probability of being homozygote by independent union of gametes:

$$\mathbf{P(III\text{-}1 = A_iA_j)}$$
$$\mathbf{= P(A_1A_2) + P(A_4A_5) + P(A_1A_4) + P(A_1A_5) + P(A_1A_6) + P(A_2A_4)}$$
$$\mathbf{+ P(A_2A_5) + P(A_2A_6) + P(A_3A_4) + P(A_3A_5) + P(A_3A_6)}$$

$$= 2.p^2/16 + 2.p^2/16 + 2.p^4/16 + + 2.p^4/16$$
$$+ 2.2.p^3.q/16 + 2.p^4/16 + 2.p^4/16 + 2.2.p^3.q/16$$
$$+ 2.2.p^3.q/16 + 2.2.p^3.q/16 + 2.4.p^2.q^2/16$$
$$= p^2/4 + p^4/2 + p^3.q + p^2.q^2/2$$
$$= p^2/4 + p^2(p^2 + 2pq + q^2)/2$$
$$= p^2/4 + P^2/2 = \mathbf{3.p^2/4}\ .$$

The probabilities shown above are of trivial determination. The only point that deserves some explanation is the factor 2 that multiplies each one of the partial expressions: it arises from the fact that individual **III-1** can be of genotype $\mathbf{A_iA_j}$ through two different paths: the gene $\mathbf{A_i}$ passes through individual **II-1** and the gene $\mathbf{A_j}$ through **II-2** or vice-versa.

By symmetry, the probability of **III-1** being a homozygote aa by distinct origins is

$$\mathbf{P(III\text{-}1 = a_ia_j) = 3.q^2/4}\ .$$

Subbtracting $3.p^2/4 + 3.q^2/4$ from 3/4 we obtain finally the probability of the individual III-1 being heterozygous :

$$\mathbf{P(III\text{-}1 = Aa) = 3.pq/2}\ .$$

The probability of autozygosis for any inbred individual, which takes the value of 1/4 when the parents are brother and sister, is the so-called **coefficient of inbreeding F**. Below we list some values of **F**, given the biological relationship between parents :

Mating	**F**
self-fertilization	**1/2**
parent-child	**1/4**
brother-sister	**1/4**

uncle-niece	**1/8**
double first-cousins	**1/8**
first cousins	**1/16**
first cousins once removed	**1/32**
second cousins	**1/64**
second cousins once removed	**1/128**

Generalizing the situation for any degree of biological relationship between the parents, the probabilities of an inbred individual being **AA**, **Aa**, and **aa** are respectively

$$P(AA) = pF + (1-F)p^2$$
$$P(Aa) = 2(1-F)pq$$
$$P(aa) = qF + (1-F)q^2 \ .$$

The factors **F** and **1-F** in the above formulae can be understood as the partition, among homozygotes, of autozygosis and homozygosis by independent union of gametes or allozygosis (another useful term coined by Cotterman).

The above formulae can be easily rearranged as

$$P(AA) = pF + (1-F)p^2 = p^2 + Fp - Fp^2 = p^2 + Fp(1-p)$$
$$= p^2 + Fpq$$
$$P(Aa) = 2pq - 2Fpq$$
$$P(aa) = q^2 + Fpq \ .$$

These latter representations are useful since they show directly the excess of homozygosis (or alternatively the decrease in the frequency of heterozygotes) in relation to the one existing in panmixia: this has a value of **Fpq**.

We have just defined a useful parameter in population genetics - the coefficient of inbreeding. It must be stressed (again) that this parameter is the probability of autozygosis of a given locus for an inbred individual. Of course it can be understood also as the fraction of genes of an inbred individual that are autozygous: for example, a child born to a couple of first cousins has an inbreeding coefficient of **F = 1/16**; this means that 1/16 of all his or her genes are in autozygotic state.

Another useful inbreeding parameter is the coefficient of common identity **R**. It represents the probability of one randomly chosen gene from one individual being identical by descent to a gene in a second person (if this second person is not biologically related to the individual this probability is zero). Of course this probability means also exactly the total amount of genes which are shared by two related individuals. In the literature this coefficient is sometimes called the coefficient of relationship, but we shall use the name "coefficient of common identity" in order to avoid a confusion that still persists in the specialized literature about the probabilistic meaning of the coefficient of relationship. Perhaps the name "coefficient of relationship" should be used only in the exact and restricted meaning Wright associated with it, that is the (zygotic) coefficient of genetic correlation between two individuals, which can be determined, for example, by the application of rules of path coefficients.

The table below shows some values of **R**, together with the corresponding **F** values for the children of individuals shown in table. It is quite obvious that **R = 2F** for any numerical value.

Relatives	**R**	**F**
Parent-child	**1/2**	**1/4**
Brother-sister	**1/2**	**1/4**
Uncle-niece	**1/4**	**1/8**
Double first cousins	**1/4**	**1/8**
First cousins	**1/8**	**1/16**
1st cousins once rem.	**1/16**	**1/32**
Second cousins	**1/32**	**1/64**
2nd cousins once rem.	**1/64**	**1/128**

The identity **R = 2F** arises from the following: let us choose, in the first individual, a given locus : it hosts, for example, alleles A_i and A_j. The probability that the relative of this individual has the allele A_i is by definition the coefficient of common identity **R**. The same figure (**R**) is true also for the allele $\mathbf{A_j}$. Therefore the chance that a child born to this couple of individuals is $\mathbf{A_iA_i}$ or $\mathbf{A_jA_j}$ (probability of autozygosis or **F**) is

$$\mathbf{F = P(A_iA_i) + P(A_jA_j) = 1.R.1/4 + 1.R.1/4 = R/2 .}$$

3) Applications of F and R in situations of genetic counseling

The probability that an inbred individual has the genotype **aa** is

$$\mathbf{P(aa|F>0) = q^2 + Fpq} \text{ ;}$$

since the frequency of the **aa** genotype among non-inbred individuals is

$$\mathbf{P(aa|F=0) = q^2} \text{ ,}$$

we may define a new parameter, that we shall call relative risk, as the ratio of the two proportions shown above:

$$\mathbf{RR = P(aa|F>0)/P(aa|F=0) = 1 + Fp/q = 1 + F(1-q)/q} \text{ .}$$

For the case of phenylketonuria, for example, **q = 0.008** (since the frequency of affected children, in the offspring of non-consanguineous spouses is $\mathbf{q^2 = 1/15,000}$); if **F = 1/16** (offspring of first cousins), the risk is

$$\mathbf{RR = 8.5} \text{ ,}$$

that is, the frequency of children affected by phenylketonuria is 8.5 times greater among children born to first cousins than to children of unrelated parents.

Frequently consanguineous couples seek genetic counsel in order to learn the risks for their offspring. In the case of first cousins with no record of genetic diseases in their families, the following reasoning can

be used : if all autosomal recessive diseases were produced by pathological alleles with an average frequency of **0.01** (this figure is of course imprecise but is also reasonable), the value of **RR** for any of these diseases should be about **7**. The frequency of recessive diseases at birth among non-inbred children can be estimated roughly in **0.01**. Therefore we deduce that the probability of a child born to a couple of first cousins spouses being affected by any recessive disorder is about **7%**. Other types of diseases (that is, non-recessive conditions) affect children belonging to both groups (consanguineous and non-consanguineous couples) with the same chance, and account for a proportion of about 2%. Therefore the risks for any disease present at birth are of 3% and 9%, respectively for children born to non-consanguineous and to first cousin couples.

One must keep in mind that the above estimates refer only to physical defects and do not include mental retardation. The frequency of this conditionn in the general population has been estimated to be about 1%; a 3 to 4-fold increase of this frequency was observed among children born to first cousin relatives. Including these figures in the risk estimates shown at the end of the last paragraph, we obtain risks of 4% for children of unrelated couples and of 13% for the offspring of first cousin unions.

The table below shows similar risk estimates for children born to several types of consanguineous couples. R_2 is the risk estimate that includes mental retardation and this is the one that should be used for genetic counseling purposes.

Marriage	R_1	R_2
brother-sister	**0.280**	**0.400**
uncle-niece	**0.140**	**0.220**
first cousins	**0.090**	**0.130**
lst cousins cnce removed	**0.060**	**0.085**
second cousins	**0.045**	**0.060**
unrelated persons	**0.030**	**0.040**

Of course the above method contains several simplified assumptions, but it is important because it enables one to calculate genetic risks as a function of **F**.

The coefficient of common identity can also be used in situations of genetic counseling, as we show below.

If we had an estimate of the average number of recessive pathological genes per individual, we could calculate easily the offspring risks. For example, let us suppose that on average each individual has one pathologic recessive gene in heterozygous state. Since the coefficient of common identity has a value **R = 1/8** for first cousins, the risk for their offspring could then be evaluated as being

$$P(aa) = R/4 = 1/32 \approx 3\% ,$$

since 1 is the probability of the first individual having the pathologic gene (we have just stated hypothetically that on average each individual has one pathologic recessive gene in heterozygous state), R is the

probability of this same gene being present in his or her cousin and 1/4 is the compound probability of two heterozygous partners transmitting the same allele to their offspring).

Unfortunately we do not have such estimates. We do have estimates of the average number of lethal equivalents and often in the literature this estimate has been confounded with an estimate of the average number of deleterious genes per person and then used unappropriately (in the manner we have just shown) in genetic counseling of consanguineous couples.

The coefficient of common identity **R** has however some useful applications in the genetic counseling of consanguineous couples, in the case of recorded diseases occurring in the family of the couple. Let us consider, for example, the following case: one albino (and oculo-cutaneous albinism is known to be an autosomal recessive disorder) and his normally pigmented cousin want to know the risk that a child they intend to have will be affected by the disease. Since **R** can be interpreted as the probability of a gene at a given locus in one person being identical by descent at the same locus in a second person and, given that the albino has a genotype $\mathbf{a_1a_2}$ (the subscripts are just to differentiate between the two genes), the probability that his cousin has one of the genes ($\mathbf{a_1}$ or $\mathbf{a_2}$) is **2R = 1/4**. So the risk for a child born to the couple of affection by the disease is **1 x 1/4 x 1/2** (these figures represent respectively the probabilities of the albino transmitting the recessive gene, of the woman being a heterozygote for this gene and, if a heterozygote, of transmitting the same gene). The final figure is **P(aa) = 1/8** or **12.5%**.

4) Average inbreeding coefficient of the population

The average inbreeding coefficient **f** of a population can be understood as the mean value of **F** in a given population:

$$\mathbf{f = \Sigma x_i F_i},$$

where $\mathbf{x_i}$ is the frequency of the class $\mathbf{F_i}$.

For example, let us suppose that in a given population 1000 couples have been randomly sampled, 952 of which were non-consanguineous ones; 32 couples were first degree cousins; and 16 were uncle-niece unions. This situation is summarized in the following table:

F_i	N_i	x_i	x_iF_i
0	952	0.952	0.000
1/16	32	0.032	0.002
1/8	16	0.016	0.002
-	1000	1.000	0.004

It is generally impossible to estimate directly the value of **f** from a population through the determination of the deviations of genotype frequencies from Hardy-Weinberg proportions, because consanguineous marriages occur with a very low frequency in most human populations. However we have the simple and practical method just shown, for which one just need to ascertain the frequencies, in the population, of the different classes of consanguineous matings.

If in a population the frequencies of the different classes of consanguineous matings remain constant from generation to generation

(i.e., if there exists in the population a regular system of inbreeding) the **f** value of the population tends to a constant value and the population is then said to be in an equilibrium state. If consanguineous marriages take place at a low rate as in the numerical example above, the equilibrium inbreeding coefficient will not differ significantly from the average inbreeding coefficient.

Taking as a first example a system of exclusive and continued self-fertilization, the chance of any individual being an autozygote after one generation is

$$f_1 = 1/2 \ ;$$

after two generations, f_n takes the value

$$f_2 = f_1 + (1-f_1)/2 = 3/4 \ ;$$

after three generations,

$$f_3 = f_2 + (1-f_2)/2 = 7/8 \ ;$$

the recursion relation is clearly

$$f_{n+1} = (f_n + 1)/2 \ .$$

Subtracting from the quantity **1** both sides of the equation above we get

$$1 - f_{n+1} = 1 - (f_n + 1)/2 = (1 - f_n).(1/2),$$

which general solution is given by

$$f_n = 1 - (1 - f_0).(1/2)^n.$$

The limiting value f_n takes is clearly **1**, as **n** tends to infinity. This means that after a great number of generations, the population tends to complete autozygosity. A numerical example (with appended BASIC code) is shown in the table below, where initial frequencies of **0.36**, **0.48** and **0.16** have been assumed for the genotypes **AA**, **Aa** and **aa**.

n	d_n	r_n	h_n	p_n	f_n
0	0.36000	0.48000	0.16000	0.60000	0.00000
1	0.48000	0.24000	0.28000	0.60000	0.50000
2	0.54000	0.12000	0.34000	0.60000	0.75000
3	0.57000	0.06000	0.37000	0.60000	0.87500
4	0.58500	0.03000	0.38500	0.60000	0.93750
5	0.59250	0.01500	0.39250	0.60000	0.96875
6	0.59625	0.00750	0.39625	0.60000	0.98438
7	0.59813	0.00375	0.39812	0.60000	0.99219
8	0.59906	0.00187	0.39906	0.60000	0.99609
9	0.59953	0.00094	0.39953	0.60000	0.99805
10	0.59977	0.00047	0.39977	0.60000	0.99902
11	0.59988	0.00023	0.39988	0.60000	0.99951
12	0.59994	0.00012	0.39994	0.60000	0.99976
13	0.59997	0.00006	0.39997	0.60000	0.99988
14	0.59999	0.00003	0.39999	0.60000	0.99994
15	0.59999	0.00001	0.39999	0.60000	0.99997

```
16   0.60000   0.00001   0.40000   0.60000   0.99998
17   0.60000   0.00000   0.40000   0.60000   0.99999
18   0.60000   0.00000   0.40000   0.60000   1.00000
19   0.60000   0.00000   0.40000   0.60000   1.00000
20   0.60000   0.00000   0.40000   0.60000   1.00000
-------------------------------------------------
```

```
REM PROGRAM FILENAME INBREE03.BAS
REM SELF-FERTILIZATION
DEFDBL A-Z: CLS : DIM D(20), H(20), R(20), P(20), F(20)
P = .6: Q = 1 - P: D(0) = P * P: H(0) = 2 * P * Q: R(0) = Q * Q: F(0) = 0
FOR I = 0 TO 20
 D(I) = P - H(0) / 2 ^ (I + 1): H(I) = H(0) / 2 ^ I
 R(I) = Q - H(0) / 2 ^ (I + 1)
 P(I) = D(I) + H(I) / 2: F(I) = 1 - H(I) / H(0)
 PRINT USING "###   "; I;
 PRINT USING "#.#####   "; D(I); H(I); R(I); P(I); F(I)
NEXT I
```

If we begin with a random-mating population, after one generation the frequencies of autozygous and allozygous AA homozygotes are given respectively by

$P'_1(AA) = p^2/2 + 2pq/4 = p/2$ and $P''_1(AA) = p^2/2$,
so that
$P_1(AA) = P'_1(AA) + P''_1(AA) = p/2 + p^2/2$;

and, in the second generation, by

$P'_2(AA) = p/2 + p^2/4 + 2pq/8 = p/2 + p/4 = 3p/4$ and $P''_2(AA) = p^2/4$,
so that
$P_2(AA) = P'_2(AA) + P''_2(AA) = 3p/4 + p^2/4$;

since in generations 0, 1, and 2 the values f_n takes are respectively 0, 1/2, and 3/4, it is easy to see that the equation above can be written as

$$P_n(AA) = P'_n(AA) + P''_n(AA) = f_n.p + (1-f_n).p^2 \ ;$$

evidently,

$$P_n(aa) = P'_n(aa) + P''_n(aa) = f_n.q + (1-f_n).q^2$$

and

$$P_n(Aa) = 2.(1-f_n).pq \ .$$

At equilibrium, **f** tends to a constant value (that is **1** in a self-fertilizing population) and the above equations become

$$P(AA) = f.p + (1-f).p^2$$
$$P(Aa) = 2.(1-f).pq$$
$$P(aa) = f.q + (1-f).q^2 \ .$$

Let us now consider as a second example a system of admixture of self-fertilization and random matings. Let **x** (for example **0.40**) be the fraction of the population that reproduces through self-fertilization and **1-x = 0.60** the fraction that reproduces sexually through panmixia. **x** can be interpreted also, in the above formulation, as being the constant probability that an individual, chosen at random from the population,