Chapter 3 Disequilibrium

Testing hypothesis

Hardy-Weinberg disequilibrium

Linkage disequilibrium

Disequilibrium coefficient

Tests for disequilibrium: goodness-of-fit tests,exact tests, permutation tests.

Hardy-Weinberg disequilibrium

In a large random-mating popultion with no selection, mutation, or migration, the gene(allele) frequencies and the genotype frequencies are constant from generation to generation; and, further, there is a simple relationship between the gene frequencies and the genotype frequecies.

Proof of Hardy-Weinberg law: (homework)

The Hardy-Weinberg law for Multiple alleles:

Applications of the Hardy-Weinberg law:

Over a 3-year period detected 5 cases in 55715 babies. The frequency of homozygotes in the sample is 90×10^{-6} or about 1/11000. The Hardy-Weinberg frequency of homozygotes is q^2, so the gene frequency is $q=\sqrt{90\times10^{-6}}=9.5\times10^{-3}=0.0095$

The frequency of carriers is $2q/(1\_q)\cong0.019$

Under H-W equilibrium,

$$P_{uu} = P_u^2 \quad \text{for homozygotes } A_uA_u$$
$$P_{uv} = 2P_uP_v \quad \text{for homozygotes } A_uA_v$$

When there is a disequilibrium,

$$P_{uu} = P_u^2 + P_u(1-P_u)f$$
$$P_{uv} = 2P_uP_v(1-f)\ ,\ u\neq v$$

Where $0\le P_{uu}\le P_u$

$0\le P_{uv}\le \min(2P_u,2P_v)$

$-P_u/(1-P_u)\le f\le 1$, for all u

By introducing indicator variables x_j for the j th allele of a random individual:

$$x_j = \begin{cases}1 & \text{if allele is A}\\ 0 & \text{otherwise}\end{cases}$$

$$\text{Var}(x_j) = p_A(1-p_A)$$

$$\text{Var}(x_j) = p_A(1-p_A)$$

It emerges that f can be regarded as the correlation of x_j and $x_{j'}$. For a locus with k alleles, there are k allele frequencies and $k(k-1)/2$ heterozygotes, suggesting that the $k(k+1)/2$ genotypic frequencies can be expressed in terms of the p_u's and a set of $k(k-1)/2$ fixation indices

$$P_{uu} = P_u^2 + D_{uu}$$
$$P_{uv} = 2P_uP_v - 2D_{uv}, u \neq v$$

f_{uv}, one for each hetrozygote.
Disequilibrium coefficient *D*:

In a two-allele case,

$$P_{AA} = p_A^2 - D_A$$
$$P_{Aa} = 2p_Ap_a - 2D_A$$
$$P_{aa} = p_a^2 + D_A$$
$$\text{Max}[-p_A^2, -p_a^2] \leq D_A \leq p_Ap_a$$

Estimating disequilibrium D_A :

$$\text{MLE: } \hat{D}_A = \tilde{\text{P}}_{\text{AA}} - \tilde{p}_A^2$$

$$\varepsilon(\hat{D}_A) = D_A - \frac{1}{2n}(p_A + P_{AA} - 2p_A^2)$$
$$= D_A - \frac{1}{2n}[p_A(1-p_A) + D_A]$$

$$Var(\hat{D}_A) = \frac{1}{n}[p_A^2(1-p_A)^2 + (1-2p_A)^2 D_A - D_A^2]$$

Testing for Hardg-Weinberg with D_A

Not a real test for H-W equilibrium.

For large samples, the MLE $\hat{D}_A$ is approximately normally distributed

$$\hat{D}_A \sim N[\, E(\hat{D}_A), Var(\hat{D}_A)\,]$$
$$Z = \frac{\hat{D}_A - E(\hat{D}_A)}{\sqrt{Var(\hat{D}_A)}}$$

So that a standard normal variate, Z, can be constructed

Under $H_0 : D_A = 0$,

$$Z^2 = \frac{n\hat{D}_A^2}{\widetilde{P}^2(1-\widetilde{P})^2} \text{ for } n \text{ large.}$$

Goodness-of-fit chi-square test

$$X_A^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}} = Z^2$$

Exact test for HWE

Exact tests are generally used for small sample sizes. However, if there are rare alleles at a locus, expected numbers can be small even in moderate large sample, and exact tests are desirable.

Under HWE hypothesis,

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} (p_A^2)^{nAA} (2 p_A p_a)^{nAa} (p_a^2)^{naa}$$

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A! n_a!} (p_A)^{nA} (p_a)^{na}$$

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{\Pr(n_{AA}, n_{Aa}, n_{aa} \text{ and } n_A, n_a)}{\Pr(n_A, n_a)}$$

$$= \frac{\Pr(n_{AA}, n_{Aa}, n_{aa})}{\Pr(n_A, n_a)}$$

$$= \frac{n! n_A! n_a! 2^{n_{Aa}}}{n_{AA}! n_{Aa}! n_{aa}! (2n)!}$$

Example：mosquito Pgm data (Table 2.4)

$$n_{11} = 9, n_{1\bar{1}} = 1, n_{\bar{1}\bar{1}} = 30; \; n_1 = 19, n_{\bar{1}} = 61$$

Table 3.1 Exact test for HWE at *Pgm* locus for mosquito data of Table 1.3.

Possible samples 11	$1\bar{1}$	$\bar{1}\bar{1}$	Probability	Cumulative Probability	Disequi–librium	Chi – square
9	1	30*	0.0000	0.0000†	0.1686	34.67†
8	3	29	0.0000	0.0000†	0.1436	25.15†
7	5	28	0.0001	0.0001†	0.1186	17.16†
6	7	27	0.0023	0.0024†	0.0936	10.69†
5	9	26	0.0205	0.0229†	0.0686	5.74†
0	19	21	0.0594	0.0823	−0.0564	3.88†
4	11	25	0.0970	0.1793	0.0436	2.32
1	17	22	0.2308	0.4101	−0.0314	1.20
3	13	24	0.2488	0.6589	0.0186	0.42
2	15	23	0.3411	1.0000	−0.0064	0.05

*Observed sample.
†Causes rejection of HWE at 5% significance level.

Likelihood ratio test for HWE

Likelihood ratio tests for multinomial proportions have been called G tests

Where

$$G^2 = -2\ln\lambda = -2\ln(\frac{L_0}{L_1})$$

And

$$L_1 = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\frac{(n_{AA})^{n_{AA}}(n_{Aa})^{n_{Aa}}(n_{aa})^{n_{aa}}}{n^n}$$

$$L_1 = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\frac{(n_A)^{n_A}(n_a)^{n_a}2^{n_{Aa}}}{(2n)^{2n}}$$

Consequently,

$$G^2 = -2\ln\left[\frac{(n)^n(n_A)^{n_A}(n_a)^{n_a}2^{n_{Aa}}}{(2n)^{2n}(n_{AA})^{n_{AA}}(n_{Aa})^{n_{Aa}}(n_{aa})^{n_{aa}}}\right]$$

Log-linear models

Multiple alleles

When there are k codominant alleles, the $k(k+1)/2$ genotypic frequencies provide $[k(k+1)/2]-1$ degrees of freedom and allow $k-1$ allele frequencies to be estimated and $k(k+1)/2$ disequilibrium coefficients to be estimated and tested for departures from zero.

$$\hat{p}_u = \tilde{p}_u$$

$$\hat{D}_{uv} = \tilde{p}_u \tilde{p}_v - \frac{1}{2}\tilde{P}_{uv}$$

$$\ln L_1 = \text{Constant} + \sum_{\text{u}} n_{uu} \ln(\frac{n_{uu}}{n}) + \sum_{u} \sum_{v \neq u} n_{uv} \ln(\frac{n_{uv}}{n})$$

$$\begin{aligned} 2nVar(\hat{D}_{uv}) &= p_u p_v[(1-p_u)(1-p_v) + p_u p_v] \\ &\quad -[(1-p_u-p_v)^2 - 2(p_u-p_v)^2]D_{uv} \\ &\quad + \sum_{w \neq u,v}(p_u^2 D_{vw} + p_v^2 D_{uw}) - D_{uv}^2 \end{aligned}$$

Exact tests with multiple alleles

$$\Pr(\{n_{uv}\}\{n_u\}) = \frac{n! 2^H \prod_u n_u!}{(2n)! \prod_{uv} n_{uv}!}$$

$$\Pr(\{n_{ij}\}\{n_i\}) = \frac{n! 2^{(n_{12}+n_{13}+n_{23})}(n_1)!(n_2)!(n_3)!}{(2n)!(n_{11})!(n_{22})!(n_{33})!(n_{12})!(n_{13})!(n_{23})!}$$

Prmutation version of exact test for HWE

Power of tests for HWE

$$X^2 \sim \chi_1^2 \quad \text{when } \mathrm{H}_0 \text{ true}$$

$$X^2 \sim \chi_{(1,v)}^2 \text{ when } \mathrm{H}_0 \text{ false}$$

$$v = \frac{nD_A^2}{p_A^2(1-p_A)^2}$$

Sex-linked genes: If the gene frequencies among males and among females are different, the population is not in

HW equilibrium. The difference is halved each generation.

$$\bar{p} = \frac{2}{3}p_f + \frac{1}{3}p_m = \frac{1}{3}(2p_f + p_m)$$

$$= \frac{1}{3}(2P + H + R)$$

$$p'_m = p_f$$

$$p'_f = \frac{1}{2}(p_m + p_f)$$

$$p'_f - p'_m = \frac{1}{2}(p_m + p_f) - p_f = -\frac{1}{2}(p_f - p_m)$$

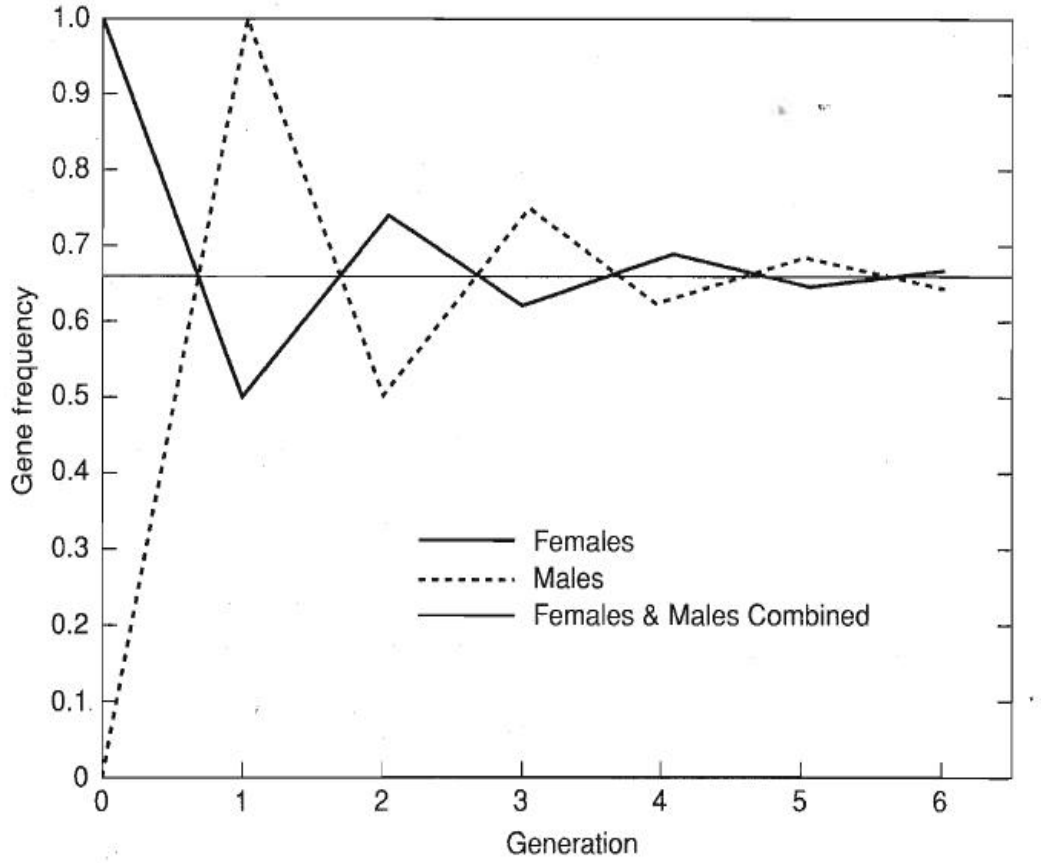

Fig. 1.2. Approach to equilibrium under random mating for a sex-linked gene, showing the gene frequency among females, among males, and in the two sexes combined. The population starts with females all of one sort ($q_f = 1$), and males all of the other sort ($q_m = 0$).

Linkage disequilibrium

The associations between the frequencies for alleles at different loci will be referred to generally as linkage disequilibrium even though they may have nothing to do with linkage. Linkage disequilibrium is needed in several contexts.

Let us consider only the A_1B_1 gametic type. The A_1B_1 gamete can be produced as a non-recombinant from the genotype A_1B_1 / A_xB_x. Or, it can be produced as a recombinant from the genotype A_1B_x / A_xB_1.

$$r' = r(1\text{-}c) + p_A p_B^C$$

$$\begin{aligned} D' &= r'\text{-}p_A p_B \\ &= r(1\text{-}c)\text{-}p_A p_B(1-c) \\ &= (r\text{-}p_A p_B)(1\text{-}c) \\ &= D(1\text{-}c) \end{aligned}$$

$$D'' = D'(1\text{-}c) = D(1-c)^2$$

$$D_t = D_0(1-c)^t$$

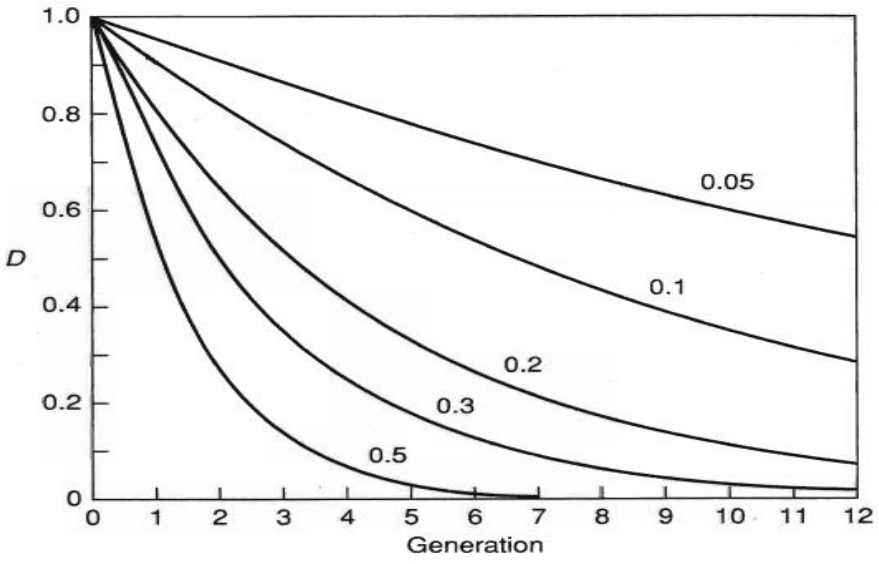

Fig. 1.3. Approach to equilibrium under random mating of two loci, considered jointly. The graphs show the amount of disequilibrium, D, relative to the disequilibrium in generation 0. The five graphs refer to different degrees of linkage between the two loci, as indicated by the recombination frequency shown alongside each graph. The graph marked 0.5 refers to unlinked loci.

Gametic disequilibrium at two loci

$$D_{AB} = P_{AB} - P_A P_B$$

The MLE of D_{AB} is $\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{P}_A \tilde{P}_B$

$$E(\hat{D}_{AB}) = \frac{2n-1}{2n} D_{AB}$$

The large-sample variance is

$$Var(\hat{D}_{AB}) = \frac{1}{2n}[P_A(1-P_A)P_B(1-P_B) + (1-2P_A)(1-2P_B)D_{AB} - D_{AB}^2]$$

A chi-square statistic for $H_0 : D_{AB} = 0$ is

$$X_{AB}^2 = \frac{2n\hat{D}_{AB}^2}{\tilde{P}_A(1-\tilde{P}_A)\tilde{P}_B(1-\tilde{P}_B)}$$

Exact test for gametic disequilibrium

Under the hypothesis of no linkage disequilibrium, the probability of $\Pr(n_{AB})$ is

$$\Pr(n_{AB}) = \Pr(n_{AB}, n_{A\bar{B}}, n_{\bar{A}B}, n_{\bar{A}\bar{B}})$$

$$= \frac{(2n)!(p_A p_B)^{n_{AB}} (p_A p_{\bar{B}})^{n_{A\bar{B}}} (p_{\bar{A}} p_B)^{n_{\bar{A}B}} (p_{\bar{A}} p_{\bar{B}})^{n_{\bar{A}\bar{B}}}}{n_{AB}! n_{A\bar{B}}! n_{\bar{A}B}! n_{\bar{A}\bar{B}}!}$$

The probabilities of the two allel arrays are

$$\Pr(n_A, n_{\bar{A}}) = \frac{(2n)!}{n_A! n_{\bar{A}}!} (p_A)^{n_A} (p_{\bar{A}})^{n_{\bar{A}}}$$

$$\Pr(n_B, n_{\bar{B}}) = \frac{(2n)!}{n_B! n_{\bar{B}}!} (p_B)^{n_B} (p_{\bar{B}})^{n_{\bar{B}}}$$

The conditional probability is

$$\Pr(n_{AB}, n_{A\bar{B}}, n_{\bar{A}B}, n_{\bar{A}\bar{B}} \mid n_A, n_B) = \frac{n_A! n_{\bar{A}}! n_B! n_{\bar{B}}!}{(2n)! n_{AB}! n_{A\bar{B}}! n_{\bar{A}B}! n_{\bar{A}\bar{B}}!}$$

Example：Drosophila restriction site data

Counts		*Xho* I +	*Xho* I −	Total
Bam HI	+	5	6	11
	−	6	0	6
Total		11	6	17

Table 3.4 Linkage disequilibria for restriction sites in *Drosophila* data of Table 1.6.

Gamete*				Probability	Cumulative Probability	D_{AB}	Chi-square
+ +	+ −	− +	− −				
11	0	0	6	0.0001	0.0001†	0.2284	17.00†
10	1	1	5	0.0053	0.0054†	0.1696	9.37†
5	6	6	0‡	0.0373	0.0427†	−0.1246	5.06†
9	2	2	4	0.0667	0.1094	0.1107	4.00†
6	5	5	1	0.2240	0.3334	−0.0657	1.41
8	3	3	3	0.2666	0.6000	0.0519	0.88
7	4	4	2	0.4000	1.0000	−0.0069	0.02

* First symbol is for *Bam*HI, second is for *Xho*I.
† Causes rejection of hypothesis of no disequilibrium at 5% level.
‡ Observed data.

Gametic disequilibrium with multiple alleles

$$X_T^2 = \sum_{u=l}^{k} \sum_{v=l}^{l} \frac{(2n)\hat{D}_{uv}^2}{\tilde{P}_u \tilde{P}_v}$$

Exact tests with multiple alleles

$$\Pr(\{n_{uv}\} \mid \{n_u\}\{n_v\}) = \frac{\prod_u n_u! \prod_v n_v!}{(2n)! \prod_{uv} n_{uv}!}$$

Variances and covariances of gametic linkage disequilibrium

Gametic disequilibrium at three or four loci

$$D_{ABC} = p_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C$$

$$D_{ABCD} = p_{ABCD} - p_A D_{BCD} - p_B D_{ACD} - p_C D_{ABD} - p_D D_{ABC} - \cdots$$

$$\pi_A = p_A(1-p_A) \ , \ \tau_A = (1-2p_A)$$

$$Var(\hat{D}_{AB}) = \frac{1}{n}[\pi_A \pi_B + \tau_A \tau_B - D_{AB}^2]$$

$$Var(\hat{D}_{ABC}) = \frac{1}{n}[\pi_A \pi_B \pi_C + 6D_{AB}D_{BC}D_{AC} + \tau_A(\tau_B \tau_C D_{BC} - D_{BC}^2) + \cdots$$

Normalized gametic disequilibria $D_{A/B}, D_{AAB}, D_{ABB}, D_{AB}^{AB}$ (Figure 3.1)

Table 3.5 Partitioning of the 9 d.f. available from the 10 genotypes at 2 loci.

Description	d.f.
Allele frequencies p_A, p_B	2
One-locus disequilibria D_A, D_B	2
Gametic disequilibrium D_{AB}	1
Nongametic disequilibrium $D_{A/B}$	1
Trigenic disequilibria D_{AAB}, D_{ABB}	2
Quadrigenic disequilibrium D_{AABB}	1
Total	9

Genotypic disequilibrium at two loci

$$D_{A/B} = p_{A/B} - p_A p_B$$

The trigenic and quadrigenic variance expressions quickly becomes unmanageable.

Composite genotypic disequilibra

When genotypes are scored, it is often not possible to

distinguish between the two double heterozygotes

$AB/\overline{A}\,\overline{B}$ *and* $A\overline{B}/\overline{A}B$

$$\Delta_{AB} = p_{AB} + p_{A/B} - 2p_A p_B = D_{AB} + D_{A/B}$$

$$\Delta_{AABB} = D_{AB}^{AB} - 2D_{AB}D_{A/B}$$

Exact tests

Log-linear tests for linkage disequilibrium

Example: MNS blood group data

		Locus B			
		SS	*Ss*	*ss*	Total
Locus A	MM	n_1=91	n_2=147	n_3=85	323
	MN	n_4=32	n_5=78	n_6=75	185
	NN	n_7=5	n_8=17	n_9=7	29
Total		128	242	167	537

Estimate	*SD(Est.)*	*Test Statistic*
$\widetilde{P}_A = 0.7737$		
$\widetilde{P}_B = 0.4637$		
$\hat{D}_A = 0.0028$	0.0077	$X_A^2 = 0.14$
$\hat{D}_B = 0.0234$	0.0107	$X_B^2 = 4.74^*$
$\hat{\Delta}_{AB} = 0.0273$	0.0090	$X_{AB}^2 = 8.21^*$
$\hat{D}_{ABB} = 0.0063$	0.0026	$X_{ABB}^2 = 5.00^*$
$\hat{D}_{AAB} = 0.0022$	0.0032	$X_{AAB}^2 = 0.46$
$\hat{\Delta}_{AABB} = -0.0034$	0.0020	$X_{AABB}^2 = 3.00$

Multiple tests

Bonferroni procedure

$$\begin{aligned}
\alpha' &= \Pr(\text{at least one test causes rejection} \mid H_0 \text{ true}) \\
&= 1 - \Pr(\text{all test do not cause rejection} \mid H_0 \text{ true}) \\
&= 1 - [\Pr(\text{one test does not cause rejection} \mid H_0 \text{ true})]^L \\
&= 1 - (1-\alpha)^L \\
&\approx L\alpha
\end{aligned}$$

Tests for homogeneity

Often data will be available from several samples, and it is generally desirable to combine such data to perform a goodness-of-fit test on all the information available.

Summary