Estimation of Allele Frequencies.

Consider a locus with two alleles A and B. with frequencies p and $(1-p)$, respectively.

How do we estimate p?

Collect Sample of n individuals and genotype them. (Random Sample from population)

Genotypes	AA	AB	BB
Counts	n_{AA}	n_{AB}	n_{BB}

$$\hat{p} = \frac{2n_{AA} + n_{AB}}{2n}, \qquad n = n_{AA} + n_{AB} + n_{BB}$$

$\hat{p}$ is unbiased estimator of p.

Let $P_{AA} = \text{freq}(AA)$, $P_{AB} = \text{freq}(AB)$

$P_{BB} = \text{freq}(BB)$

Then $p = \frac{1}{2} P_{AA} + P_{AB}$

$$E(\hat{p}) = \frac{1}{2n} E(2n_{AA} + n_{AB})$$

$$= \frac{1}{2n} 2n p_{AA} + n p_{AB}$$

$$= p_{AA} + \frac{1}{2} p_{AB} = p.$$

Maximum Likelihood estimation of p? (MLE)

We need likelihood. (Multinomial).

$$P(n_{AA}, n_{AB}, n_{BB}) = \frac{n!}{n_{AA}!\, n_{AB}!\, n_{BB}!} p_{AA}^{n_{AA}} p_{AB}^{n_{AB}} p_{BB}^{n_{BB}}$$

Then maximum likelihood estimator of

$$\hat{p}_{AA} = \frac{n_{AA}}{n}, \quad \hat{p}_{AB} = \frac{n_{AB}}{n}, \quad \hat{p}_{BB} = \frac{n_{BB}}{n}.$$

Since $\hat{p} = \hat{p}_{AA} + \frac{1}{2}\hat{p}_{AB}$ (i.e. function of mle)

$\hat{p}$ is maximum likelihood estimator of p.

Example : Seed shape

AA	Aa	aa
138	265	126.

A: round seed a = wrinkled seed.

$$\hat{p} = \frac{2*(138)+(265)}{2(529)} = 0.5113$$

freq (round seed) = 51.13%

Multiple alleles:

Suppose there are k alleles denoted by $A_1, A_2, \ldots, A_k$. with $p_i = P(A_i)$, $i=1,\ldots,k$.

Let

n_{ii} = # of individuals homozygote for ith allele.

n_{ij} = # of individuals heterozygotes for ith allele.

Then $\hat{p}_i = \dfrac{2n_{ii} + \sum_{i \neq j} n_{ij}}{2n}$, n : # of individuals.

Example: Genotype count for Pgm locus in mosquito.

Genotype:	11	12	22	13	23	33	14	24	34	44
Counts	9	1	5	0	7	0	0	8	10	0

$$\hat{p}_1 = \frac{2*9+1+0+0}{80} = 0.2375.$$

$$\hat{p}_2 = \frac{2*5 + 1 + 7 + 8}{80} = 0.325$$

$$\hat{p}_3 = 0.2125 \quad , \quad \hat{p}_4 = 0.225$$

Hardy-Weinberg Principle

→ HW states that genotype and allele frequencies remain unchanged (equilibrium) from one generation to next generation.

Consider a single locus with two alleles model

p = freq (A) , $1-p = q$ = freq (B).

If the population is in equilibrium then

$$\text{freq}(AA) = p^2$$

$$\text{freq}(BB) = q^2$$

$$\text{freq}(AB) = 2pq.$$

These frequencies are called HWP.

If individuals are randomly mating

freq (AA) = P(receiving A copy from the father)
* P(receiving A copy from the mother)

$= p \cdot p = p^2$.

freq (BB) = q^2

freq (AB) = P(receiving A copy from father)
* P(receiving B copy from mother)
+ P(receiving B copy from father)
* P(receiving A copy from mother)

$= pq + qp = 2pq$.

Assumption of random mating allows us to show that freq (AA) = $p \cdot p$.

Also, hidden assumption is that population is very large.

Random mating is not practiced?

→ assortative matings : Individual select mates based on certain characteristics. If these characters are associated with A (or B) then then population will not have HWP.

→ isolated populations : Typically smaller islands.

→ migration

→ mutations

→ Inbreeding (form of assortative mating)

Mother copy

Father copy	A	B	
A	p^2+D	$pq-D$	p
B	$qp-D$	q^2+D	$1-p=q$
	p	q	

P_{AA}	P_{AB}
P_{BA}	P_{BB}

$D = P_{AA} - p^2$ is called Hardy-Weinberg disequilibrium.

Testing for HW proportions is typically used to identify errors in genotyping.

Genotype	AA	AB	BB
Observed counts	n_{AA}	n_{AB}	n_{BB}
Expected. (under HWP)	$n\hat{p}_A^2$	$2n\hat{p}_A(1-\hat{p}_A)$	$n(1-\hat{p}_A)^2$.

Where $\hat{p}_A$ is estimated frequency of Allele A.

$$\chi^2 = \sum_{\text{genotypes}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(n_{AA} - n\hat{p}_A^2)^2}{n\hat{p}_A^2} + \frac{(n_{AB} - 2n\hat{p}_A(1-\hat{p}_A))^2}{2n\hat{p}_A(1-\hat{p}_A)} + \left(\frac{n_{BB} - n\hat{p}_B^2}{n(1-\hat{p}_A)^2}\right)^2$$

χ^2 is chi-square random variable with one degree of freedom

notes χ^2 is asymptotic distribution.

The χ^2 assumption is sensitive to small expected values (as expected value is in the denominator).

Also, since discrete counts are transformed to continuous variable a continuity correction can be used.

$$\chi^2 = \sum_{\text{genotypes}} \frac{(|\text{observed} - \text{expected}| - 0.5)^2}{\text{Expected}}$$

Example.

	AA	Aa	aa
observed	138	265	126
Expected.	138.32	264.36	126.32

$\hat{p} = 0.5153.$

$$\chi^2 = \frac{(138-138.32)^2}{138.32} + \frac{(265-264.36)^2}{264.36} + \frac{(126-126.32)^2}{126.32}$$

$$= 0.003.$$

→ degrees of freedom?

p-value = 0.9555

and conclude that

Don't reject null hypothesis that genotypes are in HWP.

AA	AB	BB
9	1	30
2.25	14.49	23.25

$\hat{p}_A = 0.2375$

$$\chi^2 = \frac{(9-2.25)^2}{2.25} + \frac{(1-14.49)^2}{14.49} + \frac{(30-23.26)^2}{23.26}$$

p-value $= 3.73 \times 10^{-9}$.

degrees of freedom $= \frac{n(n-1)}{2}$, where n is number of alleles.

Exact test for HWP. (a.k.a. Fisher's exact test)

→ Determine the probabilities of all possible samples of the same size as sample at hand assuming the hypothesis is true.

→ Exact test are used for small sample size

The likelihood of counts is

$$Pr(n_{AA}, n_{AB}, n_{BB}) = \frac{n!}{n_{AA}!\, n_{AB}!\, n_{BB}!} (P_{AA})^{n_{AA}} (P_{AB})^{n_{AB}} (P_{BB})^{n_{BB}}$$

Under HWP

$$P_r(n_{AA}, n_{AB}, n_{BB}) = \frac{n!}{n_{AA}!\, n_{AB}!\, n_{BB}!} (p_B^2)^{n_{AA}} (2pq)^{n_{AB}} (q^2)^{n_{BB}}$$

Also n_A, n_B are binomially distributed under HWP

$$\boxed{n_A + n_B = 2n}$$

$$P_r(n_A, n_B) = \frac{(2n)!}{n_A!\, n_B!} p^{n_A} q^{n_B}.$$

$$P(n_{AA}, n_{AB}, n_{BB} \mid n_A, n_B) = \frac{n!\; n_A!\; n_B!\; 2^{n_{AB}}}{n_{AA}!\, n_{AB}!\, n_{BB}!\, (2n)!}$$

Conditional distribution of genotype counts conditions on observed allele frequencies.

Page 100 (Weir, Genetic Data Analyses)

$$n_A = 2n_{AA} + n_{AB}$$

Hardy Weinberg Exact test

A test is performed by computing probabilities under the null hypothesis of all possible genotype combinations that has the same allele frequencies and the total sample size as the observed data. Then, the sum of all probabilities of events less or equal probable to the observed event probability is the exact p-value.

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

Possible samples			Probability	Cumulative Probability	Disequi– librium	Chi – square
11	$1\bar{1}$	$\bar{1}\bar{1}$				
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.

or large chi-square values, and rejection of the hypothesis. The ordering of probabilities is not the same as the ordering on the numbers of heterozygotes. The exact test is two-sided. In this particular example, though, the exact test rejection region consists only of small numbers of heterozygotes.

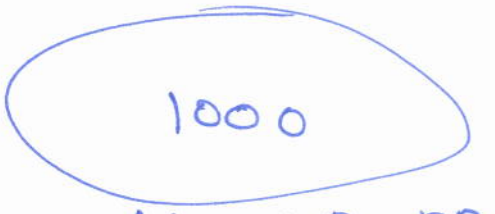
By adding the probabilities for 1, 3, 5, 7, and 9 heterozygotes a rejection region of size 0.0229 is found. In other words, there is a probability of 2.29% of falsely rejecting the hypothesis of HWE when it is rejected with 9 or fewer heterozygotes. This probability is the significance level or probability of a ***type I error***. Adding the next largest probability, for 19 heterozygotes, would give a test of size 8.23%, which would generally be regarded as being too high. Chi-square test statistics are also shown in Table 3.1 and demonstrate that the two procedures differ even for samples as large as 40. The chi-square test rejects for 19 heterozygotes whereas the exact test does not reject. Applying Yates' continuity correction would bring X^2 down to 2.62 for $x = 19$, below the critical value of 3.84, and the two tests would then agree.

This procedure of adding probabilities for all deviations (of the numbers of heterozygotes) from the value expected under the hypothesis regardless of sign has been criticized by Yates (1984). He advocates keeping track of sign. In the present example, the expected number of heterozygotes is $40 \times 2(19/80)(61/80) = 14.5$, and the observed data have fewer heterozygotes than expected, so that the observed disequilibrium is positive. If the rows in Table 3.1 were ordered the same as the numbers of heterozygotes

Admixed Population

$P_A = 0.4$

p-value = 1.0



$P_A = 0.1$

p-value = 1.0

p-value for HWP = 8×10^{-8}

$\Rightarrow$ Admixed population is not in HWP.