Hardy-Weinberg Proportions (HWP)

Deviation from HWP may arise due to:

- typing errors (e.g., missing heterozygotes)
- assortative mating (e.g., negative assortative mating)
- selection (e.g., heterozygote advantage)
- population structure (e.g., ≥ 2 merged populations)

Tests of fit to HWP

- Chi-square Goodness-of-fit test
- Likelihood Ratio test
- Exact test

Hardy-Weinberg Proportions (HWP)

- Let p_A and p_a be the allele frequencies and p_{AA} , p_{Aa} , and p_{aa} be genotype frequencies

HWP ➔ $p_{AA} = p_A^2 \qquad p_{Aa} = 2p_A p_a \qquad p_{aa} = p_a^2$

- O_i = observed # with the ith genotype (n_{aa}, n_{Aa}, & n_{AA})
- E_i = expected # with the ith genotype, under H_0: HWP

$$X^2_{HW} = \sum (O_i - E_i)^2 / E_i$$

HWP: Chi-square Goodness-of-fit test

- Example: The Pgm locus in a sample of Aedes Aegypti mosquitos

$$O_1 = n_{aa} = 26, \quad O_2 = n_{Aa} = 9, \quad O_3 = n_{AA} = 5$$

$$p_a = (2 \times n_{aa} + n_{Aa}) / 2n = (2 \times 26 + 9) / 80 = 0.7625$$
$$p_A = (2 \times n_{AA} + n_{Aa}) / 2n = (2 \times 5 + 9) / 80 = 0.2375$$

$$E_1 = E_{aa} = n \times p_a^2 = 40 \times (.7625)^2 = 23.25625$$
$$E_2 = E_{Aa} = 2n \times p_A p_a = 80 \times (.7625)(.2375) = 14.48750$$
$$E_3 = E_{AA} = n \times p_A^2 = 40 \times (.2375)^2 = 2.25625$$

HWP: Chi-square G-O-F test

- E_i = expected # with the ith genotype, under H_0: HWP

$$E_1 = E_{aa} = n \times p_a^2 = 40 \times (.7625)^2 = 23.25625$$
$$E_2 = E_{Aa} = 2n \times p_A p_a = 80 \times (.7625)(.2375) = 14.48750$$
$$E_3 = E_{AA} = n \times p_A^2 = 40 \times (.2375)^2 = 2.25625$$

$$X^2_{HW} = \frac{(26 - 23.25625)^2}{23.25625} + \frac{(9 - 14.4875)^2}{14.4875} + \frac{(5 - 2.25625)^2}{2.25625} = 5.738814$$

$$pvalue = 0.0166$$

df = (#categories - 1) - (# indep. allele freqs. estimated from the data)

Definition 12 (Likelihood function, independent data.) Suppose $X_1, \ldots, X_n$ are independent random variables. Then, the likelihood function is defined as

$$L(\psi) = \begin{cases} \prod_{i=1}^n P(X_i = x_i), & \text{if } X_1, \ldots, X_n \text{ are discrete r.v.'s} \\ \prod_{i=1}^n f_{X_i}(x_i), & \text{if } X_1, \ldots, X_n \text{ are continuous r.v.'s.} \end{cases} \quad (3.1)$$

Example 30 (Coin tossing, contd.) The likelihood function for the coin tossing experiment of Example 29 can be computed as $L(\psi) = (1-\psi)^{39}\psi^{61}$, since there are 39 factors $P(X_i = 0) = (1-\psi)$ and 61 factors $P(X_i = 1) = \psi$. A more formal way of deriving this is

$$\begin{aligned} L(\psi) &= \prod_{i=1}^{100} P(X_i = x_i) = \prod_{i=1}^{100} (1-\psi)^{1-x_i}\psi^{x_i} \\ &= (1-\psi)^{100-\sum_{i=1}^{100} x_i}\psi^{\sum_{i=1}^{100} x_i} = (1-\psi)^{39}\psi^{61}, \end{aligned} \quad (3.2)$$

since $\sum_{i=1}^{100} x_i = 61$ is the total number of heads. □

Definition 13 (Maximum likelihood estimator.) The maximum likelihood (ML) estimator is defined as

$$\widehat{\psi} = \arg\max_{\psi \in \Psi} L(\psi),$$

meaning that $\widehat{\psi}$ is the parameter value which maximizes L.

Example 31 (ML-estimator for coin tossing.) If we take the logarithm of (3.2) we get

$$\ln L(\psi) = 39 \ln(1-\psi) + 61 \ln \psi.$$

This function is shown in Figure 3.1. Differentiating this w.r.t. and putting the derivative to zero we get

$$0 = \frac{d \ln L'(\psi)}{d\psi}|_{\psi=\widehat{\psi}} = \frac{61}{\widehat{\psi}} - \frac{39}{1-\widehat{\psi}} \Longleftrightarrow \widehat{\psi} = \frac{61}{100}.$$

Probability Model vs. Likelihood

X = # Heads in $n = 100$ tosses of a coin

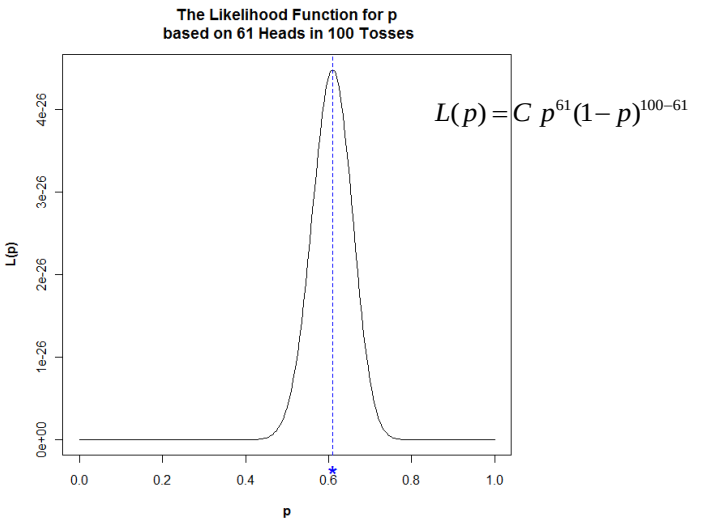
$P(\text{Heads}) = p$

Probability Model:

$$P(x) = C\ p^x(1-p)^{100-x} \qquad x = 0, 1, \ldots, 100$$

Likelihood Function:

$$L(p) = C\ p^x(1-p)^{100-x}$$



HWP: Likelihood Ratio Test

$$L(p_{AA}, p_{Aa}, p_{aa}) = \frac{n!}{n_{AA}!\,n_{Aa}!\,n_{aa}!}\left(p_{AA}\right)^{n_{AA}}\left(p_{Aa}\right)^{n_{Aa}}\left(p_{aa}\right)^{n_{aa}}$$

- L_0 is the likelihood computed under H_0 : HWP

$$p_{AA} = p_A^2 = \left((2n_{AA} + n_{Aa})/2n\right)^2, \ldots$$

- L_1 is computed under H_1 (no restrictions on genotype freqs)

$$p_{AA} = (n_{AA}/n), \ldots$$

- The likelihood ratio chi-square is $S = X^2_{LR} = -2\ln\left(\frac{L_0}{L_1}\right) \sim \chi^2_1$

Fisher's Exact Test

	A	A'	
B	n_{11}	n_{12}	$n_{1.}$
B'	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Fisher showed that the exact probability of a table relating A and B is:

$$P(table) = \frac{\binom{n_{1.}}{n_{11}}}{\binom{N}{n_{.1}}} = \frac{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!}{N!\,n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}$$

HWP: Exact Tests

Observed Table (or arrangement):

	A	a
A	5	
a	9	26
Total	14	26

Procedure:

- Generate all possible tables with fixed margins
- Compute the prob. of each table
- Find the prob. of observing a table as extreme or more extreme than the observed table

Levene (1948) showed that Fisher's formula in the genetics setting is:

$$P(table) = \frac{n!/(n_{AA}!\,n_{Aa}!\,n_{aa}!)}{(2n)!/[(2n-n_a)!\,n_a!]}\,2^{n_{Aa}}$$

HWP: Exact Test

- All possible arrangements of n=40 individuals, with allele frequencies fixed at the observed values.

AA	Aa	aa	Probability*	Cumulative Probability	
9	1	30	0.0000	0.0000	
8	3	29	0.0000	0.0000	
7	5	28	0.0001	0.0001	
6	7	27	0.0023	0.0024	
5	9	26	0.0205	0.0229	←Observed Data
0	19	21	0.0594	0.0823	
4	11	25	0.0970	0.1793	
1	17	22	0.2308	0.4101	
3	13	24	0.2488	0.6589	
2	15	23	0.3411	1.0000	

* conditional probability of the arrangement, given the fixed marginal allele frequencies for A & a (Levene, 1948).

HWP: Highly polymorphic data

- When there are lots of alleles, enumeration of all possible tables is not feasible and Randomization Tests are used instead.

- When simple randomization tests are not feasible, Markov Chain Monte Carlo (MCMC) methods are used to sample the space of possible tables. (e.g., Guo & Thompson, 1992)
  - Tables are sampled probabilistically
  - Transition probabilities between any two tables are a function of the ratio of probabilities of each.

HWP: Testing Individual Genotypes

- Overall tests of HWP do not indicate which individual genotypes may be responsible for deviation at a locus.
  - This information can help point to sources of deviation (e.g., allele dropout, preferential amplification, heterogeneous levels of allelic resolution, population stratification, selection, …)

- Overall tests and ad-hoc tests (e.g., 1-d.f. “GOF” test) of individual genotypes may not be very powerful.

- A modified typing protocol can sometimes alleviate problems revealed by deviation from HWP.

Hardy-Weinberg Disequilibrium Coefficients (D_{ij})

$$p_{AA} = p_A^2 + d_{AA}$$
$$p_{Aa} = 2p_A p_a - 2d_{Aa}$$
$$p_{aa} = p_a^2 + d_{aa}$$

- In the two-allele case there is only one disequilibrium coefficient (i.e., the 3 coefficients above are the same and called d_A)
- Based on the constraints for allele and genotype frequencies,

$$\max(-p_A^2, -p_a^2) \le d_A \le p_A p_a$$

- In the general case, the disequilibrium coefficients are defined as

$$d_{ij} = p_i p_j - \frac{1}{2} p_{ij}$$
$$d_{ii} = p_{ii} - p_i^2$$

Inbreeding Coefficients (D_{ij})

- H-W Disequilibrium coefficients are sometimes parameterized in terms of an measure of inbreeding (f).

$$p_{AA} = p_A^2 + p_A p_a \cdot f$$
$$p_{Aa} = 2p_A p_a \cdot (1 - f)$$
$$p_{aa} = p_a^2 + p_A p_a \cdot f$$