§ 2: HWE:

- Hardy-Weinberg equilibrium:

p_A = allele frequency of A

Under HWE, $P_{AA} = p_A^2$
$P_{Aa} = 2p_A(1-p_A)$
$P_{aa} = (1-p_A)^2$

I.e. Genotypes are Binom(2, p_A)

Recall Binomial conditions
1. Success or failure outcome
2. See n (here $n=2$) outcomes
3. Outcomes are independent ("Random mating")
4. Infinite population size

- Departures from HWE can be because of
  1. No random mating (eg. pop structure, inbreeding)
  2. Genotyping errors
  3. Small population sizes

- Deviations from HWE

$P_{AA} = p_A^2 + p_A(1-p_A)f$
$P_{Aa} = 2p_A(1-p_A)(1-f)$
$P_{aa} = (1-p_A)^2 + p_A(1-p_A)f$

$f = 0 \Rightarrow$ HWE

- f = "inbreeding frequency" = F_{IS} (measure of relatives mating)
  "fixation index" = F_{ST} measure of population differ
  ↑ 2 different models

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

$$\mathrm{Var}(X_j) = p_A(1-p_A)$$

$$\mathrm{Cov}(X_j, X_j') = \underbrace{E[X_j X_j']} - E[X_j]\,E[X_j']$$

$$= p_{AA} - p_A^2$$

$$= f\,\underbrace{p_A(1-p_A)}_{=\mathrm{Var}(X_j)}$$

$$\Rightarrow f = \mathrm{Cov}(X_j, X_j')$$

- D is equilibrium coefficient:

$D = p_{AA} - p_A^2$ (difference b/t genotype freq. and what is expected at HWE)

$$p_{AA} = p_A^2 + D_A$$
$$p_{Aa} = 2p_A(1-p_A) - 2D_A$$
$$p_{aa} = (1-p_A)^2 + D_A$$

- Note: $\max\{-p_A^2, -(1-p_A)^2\} \le D_A \le p_A(1-p_A)$

proof:

$0 \le p_{AA} \le p_A$ since max occurs at

$$0 \le p_A^2 + D_A \le p_A$$
$$\Rightarrow -p_A^2 \le D_A \le p_A - p_A^2$$

Do same thing for p_{aa} //

Recall

AA	Au	aa	
n_{AA}	n_{Au}	n_{aa}	$n = n_{AA} + n_{Au} + n_{aa}$

$$\hat{p}_A = \frac{2n_{AA} + n_{Au}}{2n}$$

$$\hat{P}_{AA} = \frac{n_{AA}}{n}$$

$$\hat{D}_A = \hat{P}_{AA} - \hat{p}_A^2$$

↳ Can get mean and variance of the sampling dist of $\hat{p}_A$

- By properties of MLE

$$\hat{p}_A \sim N\left(\underbrace{E(\hat{p}_A)}_{\approx 0 \text{ under } H_0: D_A = 0}, \text{Var}(\hat{p}_A)\right)$$

So $$\frac{\hat{D}_A}{\sqrt{\text{Var}(\hat{D}_A)}} \approx N(0,1)$$

↳ Compare to $N(0,1)$

under Null: $E(\hat{D}_A) \approx 0$

$$\text{Var}(\hat{D}_A) \approx \frac{1}{n} p_A^2 (1-p_A)^2$$

$$\chi^2_A = \frac{n \hat{D}_A^2}{\hat{p}_A(1-\hat{p}_A)} \approx \chi^2_1 \quad \text{b/c} \quad \frac{D_A}{\frac{1}{n} p_A(1-p_A)} \approx N(0,1)$$

Approach 2: χ^2 - goodness-of-fit

Observed	n_{AA}	n_{Aa}	n_{aa}
Expected	$n\hat{p}_A^2$	$2n\hat{p}_A(1-\hat{p}_A)$	$n(1-\hat{p}_A)^2$
observed - expected	$n\hat{D}_A$	$-2n\hat{D}_A$	$n\hat{D}_A$

$$\chi^2_A = \sum_{\text{categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \sim \chi^2_1$$

χ^2_1: 2 parameters under full ($p_{AA}+p_A$), 1 parameter under null (p), $2-1$

$$= \frac{(n\hat{D}_A)^2}{n\hat{p}_A^2} + \frac{(-2n\hat{D}_A)^2}{2n\hat{p}_A(1-\hat{p}_A)} + \frac{(n\hat{D}_A)^2}{n(1-\hat{p}_A)^2}$$

$$= \frac{n\hat{D}_A^2}{\hat{p}_A^2(1-\hat{p}_A)^2}$$

• Approach 3: Exact tests

① Calculate $P_r(n_{AA}, n_{Au}, n_{uu} \mid n_A, n_u)$

for all possible values of n_{AA}, n_{Au}, n_{uu} given

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

$n_u = n_{Au} + 2n_{uu}$

↑ fixed ↖ vary these

e.g. $n_A = 5$, $n_u = 5$

observed →	n_{AA}	n_{Au}	n_{uu}	$P_r(n_{AA}, n_{Au}, n_{uu} \mid n_A, n_u)$
	2	1	2	p_1
	1	3	1	p_2
	0	5	0	p_3

② Sort probabilities and add up smaller values

$p_3 < p_1 < p_2$

$p_1 + p_3$ = p-value b/c we saw p_1

$$P_r(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!\, n_{Aa}!\, n_{aa}!} \left(p_A^2\right)^{n_{AA}} \left(2p_A(1-p_A)\right)^{n_{Aa}} \left[(1-p_A)^2\right]^{n_{aa}}$$

$$P_r(n_A, n_a) = \frac{(2n)!}{n_A!\, n_a!} (p_A)^{n_A} (1-p_A)^{n_a}$$

only under HWE
$$\begin{cases} n_A \sim Bin(2n, p_A) \\ (n_{AA}, n_{Aa}, n_{aa}) \sim Mult(n, p_A^2, 2p_A(1-p_A), (1-p_A)^2) \end{cases}$$

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

$$= \frac{n!\, n_A!\, n_a!\, 2^{n_{Aa}}}{n_{AA}!\, n_{Aa}!\, n_{aa}!\, (2n)!} \left[\frac{\left(p_A^2\right)^{n_{AA}} \left(p_A(1-p_A)\right)^{n_{Aa}} \left[(1-p_A)^2\right]^{n_{aa}}}{p_A^{n_A} (1-p_A)^{n_a}} \right]$$

$= 1$

$$= p_A^{2n_{AA} + n_{Aa} - n_A} (1-p_A)^{2n_{aa} + n_{Aa} - n_a}$$

$$= p_A^0 (1-p_A)^0$$

$$= 1 \;//$$

- So probabilities do not depend on p_A

- If n is large, too computationally expensive to find all values of n_{AA}, n_{Aa}, n_{aa}

- So take n_A A alleles at n_a a alles, randomly pair, find n_{AA}, n_{Aa}, n_{aa}, calculate prob of this
  - repeat many times
  - proportion tables less probable $\approx$ p-value

- Approach 4: Likelihood ratio test (in genetics, "G"-test, but not related else

$$L_0 = \max_{p_A} P_r\left(n_{AA}, n_{Aa}, n_{aa} \mid p_A^2, 2p_A(1-p_A), (1-p_A)^2\right)$$

$$L_1 = \max_{\underline{p}_{AA}, \underline{p}_{Aa}, \underline{p}_{aa}} P_r\left(n_{AA}, n_{Aa}, n_{aa} \mid \underline{p}_{AA}, \underline{p}_{Aa}, \underline{p}_{aa}\right)$$

Under H_0: $-2\log\left(\frac{L_0}{L_1}\right) \sim \chi^2_1$

(2 parameters in alt − 1 parameter in

$= -2\left(\log(L_0) - \log(L_1)\right)$