HARDY-WEINBERG EQUILIBRIUM

Hardy-Weinberg Law

For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

These are also the results of setting the inbreeding coefficient f to zero.

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

Why would HWE not hold?

- Natural selection.
- LD with trait in trait-only sample.
- Population Structure/Admixture.
- Problems with data.
- etc.

Problems with Data

A SNP with genotype counts $40, 0, 60$ for AA, Aa, aa is likely to cause HWE rejection. What about $4, 0, 6$?

Typing systems may report heterozygotes as homozygotes, as was the likely explanation for
"To justify applying the classical formulas of population genetics in the Castro case, the Hispanic population must be in Hardy-Weinberg equilibrium. In fact, Lifecodes' own data show that it is not. ... Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium: 17 per cent observed homozygotes at D2S44 and 13 per cent observed homozygotes at D17S79 compared with only 4 per cent expected at each locus, indicating, perhaps not surprisingly, the presence of genetically distinct subgroups within the Hispanic sample."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339:501-505.

Population Structure

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the *Wahlund effect*.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Admixture: Departures from HWE

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Population 2 is in HWE, but Population 1 has 51% heterozygotes instead of the expected 49.8%.

Inference about HWE

If $\widehat{f}$ is the MLE of the within-population inbreeding coefficient f, it has a normal distribution for large sample sizes n. It can be transformed into a standard normal variable z by

$$z = \frac{\widehat{f} - f}{\sqrt{\mathrm{Var}(\widehat{f})}}$$

If the true value f is zero, then $\mathrm{Var}(\widehat{f}) = 1/n$, and $X^2 = z^2$ has a chi-square distribution with one degree of freedom:

$$X^2 = \left(\frac{\widehat{f} - 0}{\sqrt{1/n}}\right)^2 = n\widehat{f}^2 \sim \chi 2_{(1)}$$

The HWE hypothesis is rejected at the 5% significance level if $X^2 > 3.84$.

Aside: Inference about HWE

Departures from HWE can be described by the within-population inbreeding coefficient f. This has an MLE that can be written as

$$\hat{f} = 1 - \frac{\tilde{P}_{AB}}{2\tilde{p}_A\tilde{p}_B} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{(2n_{AA} + n_{AB})(2n_{BB} + n_{AB})}$$

and we can use "Delta method" to find

$$\begin{aligned} \mathcal{E}(\hat{f}) &= f \\ \text{Var}(\hat{f}) &\approx \frac{1}{2np_Ap_B}(1-f)[2p_Ap_B(1-f)(1-2f) + f(2-f)] \end{aligned}$$

If $\hat{f}$ is assumed to be normally distributed then, $(\hat{f}-f)/\sqrt{\text{Var}(\hat{f})} \sim N(0,1)$. When H_0 is true, the square of this quantity has a chi-square distribution.

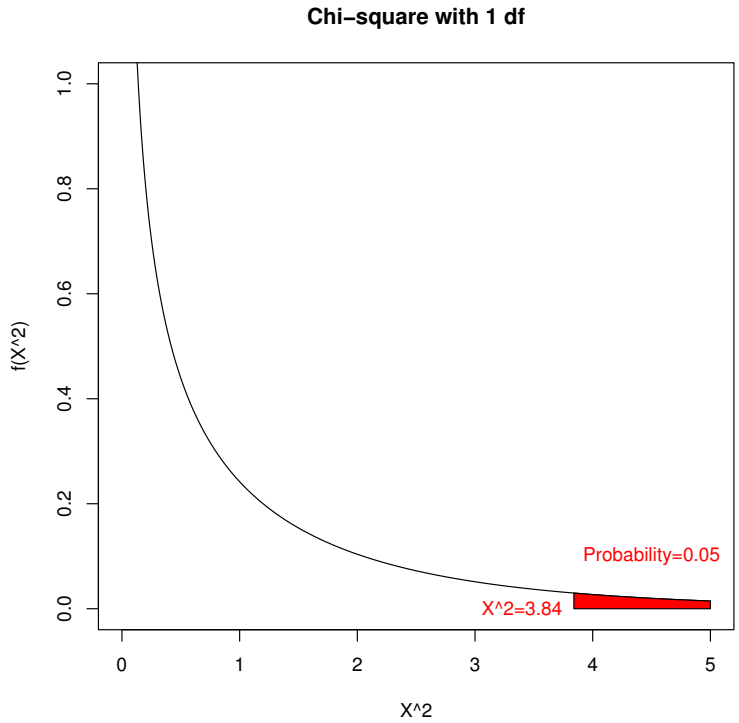
Aside: Inference about HWE

Since $\text{Var}(\widehat{f}) = 1/n$ when $f = 0$:

$$\begin{aligned} X^2 &= \left(\frac{\widehat{f} - f}{\sqrt{\text{Var}(\widehat{f})}} \right)^2 \\ &= \frac{\widehat{f}^2}{1/n} \\ &= n\widehat{f}^2 \end{aligned}$$

is appropriate for testing $H_0 : f = 0$. When H_0 is true, $X^2 \sim \chi^2_{(1)}$. Reject HWE if $X^2 > 3.84$.

Significance level of HWE test



The area under the chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is true. This is the significance level of the test.

Goodness-of-fit Test

An alternative, but equivalent, test is the goodness-of-fit test.

Genotype	Observed	Expected	$\frac{(\text{Obs.}-\text{Exp.})^2}{\text{Exp.}}$
AA	n_{AA}	$n\tilde{p}_A^2$	$n\tilde{p}_a^2\hat{f}^2$
Aa	n_{Aa}	$2n\tilde{p}_A\tilde{p}_a$	$2n\tilde{p}_A\tilde{p}_a\hat{f}^2$
aa	n_{aa}	$n\tilde{p}_a^2$	$n\tilde{p}_A^2\hat{f}^2$

The test statistic is

$$X^2 = \sum \frac{(\text{Obs.} - \text{Exp})^2}{\text{Exp.}} = n\hat{f}^2$$

Goodness-of-fit Test

Does a sample of 6 AA, 3 Aa, 1 aa support Hardy-Weinberg?

First need to estimate allele frequencies:

$$\tilde{p}_A = \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa} = 0.75$$

$$\tilde{p}_a = \tilde{P}_{aa} + \frac{1}{2}\tilde{P}_{Aa} = 0.25$$

Then form "expected" counts:

$$\begin{aligned} n_{AA} &= n(\tilde{p}_A)^2 = 5.625 \\ n_{Aa} &= 2n\tilde{p}_A\tilde{p}_a = 3.750 \\ n_{aa} &= n(\tilde{p}_a)^2 = 0.625 \end{aligned}$$

Goodness-of-fit Test

Perform the chi-square test:

Genotype	Observed	Expected	(Obs. $-$ Exp.)^2/Exp.
AA	6	5.625	0.025
Aa	3	3.750	0.150
aa	1	0.625	0.225
Total	10	10	0.400

Note that $\widehat{f} = 1 - 0.3/(2 \times 0.75 \times 0.25) = 0.2$ and $X^2 = n\widehat{f}^2$.

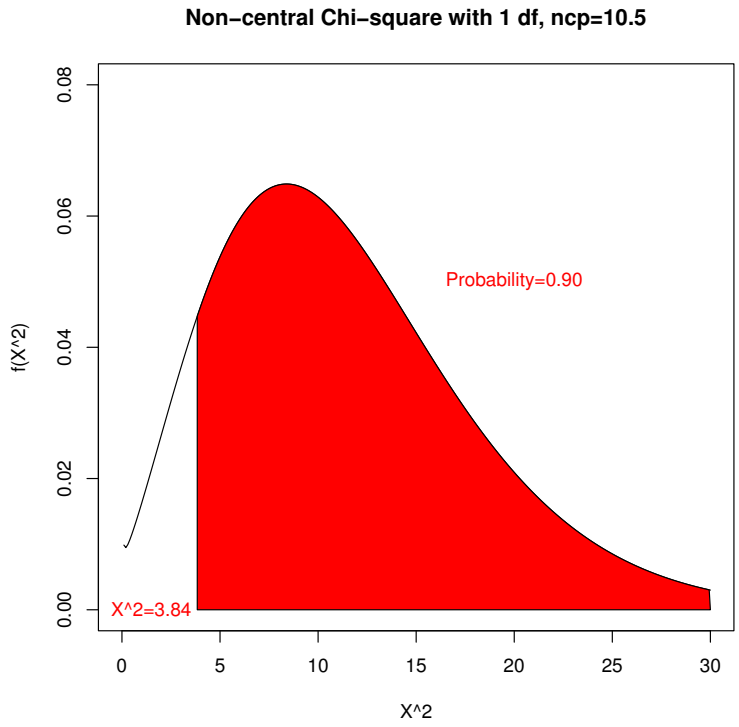
Sample size determination

Although Fisher's exact test (below) is generally preferred for small samples, the normal or chi-square test has the advantage of simplifying power calculations.

When the Hardy-Weinberg hypothesis is not true, the test statistic $n\widehat{f}^2$ has a non-central chi-square distribution with one degree of freedom (df) and non-centrality parameter $\lambda = nf^2$. To reach 90% power with a 5% significance level, for example, it is necessary that $\lambda \geq 10.51$.

```
> pchisq(3.84,1,0)
[1] 0.9499565
> pchisq(3.84,1,10.51)
[1] 0.09986489
> qchisq(0.95,1,0)
[1] 3.841459
> qchisq(0.10,1,10.51)
[1] 3.843019
```

Power of HWE test



The area under the non-central chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is false. This is the power of the test. In this plot, the non-centrality parameter is $\lambda = 10.5$.

Sample size determination

To achieve 90% power to reject HWE at the 5% significance level when the true inbreeding coefficient is f, need sample size n to make $nf^2 \geq 10.51$.

For $f = 0.01$, need $n \geq 10.51/(0.01)^2 = 105,100$.

For $f = 0.05$, need $n \geq 10.51/(0.05)^2 = 4,204$.

For $f = 0.10$, need $n \geq 10.51/(0.10)^2 = 1,051$.

Significance Levels and p-values

The *significance level* α of a test is the probability of a false rejection. It is specified by the user, and along with the null hypothesis, it determines the rejection region. The specified, or "nominal" value may not be achieved for an actual test.

Once the test has been conducted on a data set, the probability of the observed test statistic, *or a more extreme value*, if the null hypothesis is true is the p-*value*. The chi-square and normal tests shown above give approximate p-values because they use a continuous distribution for discrete data.

An alternative class of tests, "exact tests," use a discrete distribution for discrete data and provide accurate p-values. It may be difficult to construct an exact test with a particular nominal significance level.