Contents

2 Tests for Hardy Weinberg Equilibrium

2.1 Statistical Hypothesis Testing [review]

Statistical Hypothesis Testing Procedure [Review]

- Procedure
  - Identify a hypothesis, an idea you want to test for its applicability to your population.
    - "Hardy-Weinberg equilibrium applies to this population."
    - "The two loci I am studying are independent of each other."
  - Identify and calculate a *test statistic*. Ideally, the statistic should:
    - summarize and accentuate any deviations of the data from what is expected under the hypothesis,
    - have a known sampling distribution under the null hypothesis.
  - Compute or estimate the probability of the observed test statistic under the assumption of the hypothesis.
  - *Reject* the hypothesis if this probability is small.

- Notation
  - **Null hypothesis** (H_0): the hypothesis you wish to test.
  - **Alternative hypothesis** (H_A): when you reject the null hypothesis, you conclude the alternative hypothesis or hypotheses.
  - **Type I error**: the test statistic causes you to (erroneously) reject the hypothesis when it is true.
  - **size**, or **significance level** (α): the probability of a type I error
  - **Type II error**: you accept the hypothesis when it is false
  - β: the probability of a type II error
  - **power** ($1-\beta$): probability that you reject the hypothesis when it is false.
  - Procedure: classically, one decides on the *size* of the test before collecting the data, then selects the most powerful test for the desired size.

	Hypothesis Accepted	Hypothesis Rejected
Hypothesis True	$1-\alpha$	Type I (size, α)
Hypothesis False	Type II (β)	power ($1-\beta$)

2.2 Parametric Tests of HWE

Hardy-Weinberg Equilibrium

Let u and v be alleles at a single locus. Then, HWE implies

$$\begin{aligned} P_{uu} &= p_u^2 \\ P_{uv} &= 2p_u p_v \text{ whenever } u \neq v \end{aligned}$$

where P_{uv} is the population genotype frequency and p_u are the population allele frequencies.

Recall that if two random variables X and Y are independent, then

$$P(X \text{ and } Y) = P(X)P(Y).$$

In English, knowing X tells you nothing about Y and vice versa.

HWE equations imply independence (or no association) among the alleles at a locus.

Hardy-Weinberg Disequilibrium

These equations may not be satisfied in a population where any one (or more) of the Hardy-Weinberg assumptions is (are) violated. When the equations are not satisfied, a **Hardy-Weinberg disequilibrium** applies.

$$\begin{aligned} P_{uu} &\neq p_u^2 \\ P_{uv} &\neq 2p_u p_v \text{ whenever } u \neq v \end{aligned}$$

One could mathematically quantitate this disequilibrium in multiple ways.

- Suppose there is covariation among alleles. Write the disequilibrium in terms of this covariation.
- Consider subtractive disequilibrium: $P_{uu} - p_u^2$ and $P_{uv} - 2p_u p_v$.
- Consider multiplicative disequilibrium once again and convert to log-linear model.

2.2.1 Correlation Between Alleles f

Covariation Between Alleles

Recall the model

$$\begin{aligned} P_{11} &= p_1^2 + p_1(1-p_1)f \\ P_{12} &= 2p_1(1-p_1)(1-f) \end{aligned}$$

What is the meaning of f?

Let X_j, $j = 1, 2$ be an indicator variable indicating whether the jth allele of a random individual is allele 1. Clearly, $E(X_j) = p_1$ by definition and

$$\begin{aligned} \text{Var}\left(X_j\right) &= p_1\left(1-p_1\right) \\ \text{Cov}\left(X_i, X_j\right) &= E\left(X_i X_j\right) - E\left(X_i\right)E\left(X_j\right) = P_{11} - p_1^2 \\ \text{Corr}\left(X_i, X_j\right) &= \frac{\text{Cov}\left(X_i, X_j\right)}{\sqrt{\text{Var}\left(X_i\right)\text{Var}\left(X_j\right)}} \\ &= \frac{P_{11} - p_1^2}{p_1\left(1-p_1\right)} \\ &= \frac{p_1^2 + p_1\left(1-p_1\right)f - p_1^2}{p_1\left(1-p_1\right)} = f. \end{aligned}$$

Covariation Between >2 Alleles (Model 1)

What do you do with f when there are more than 2 alleles? Say, for example, there are 3 alleles u, v, and w. There can be correlation between u and v or u and w, etc.

We can subscript f. Let f_{uv} be the correlation between alleles u and v, where u can equal v. Then,

$$\begin{aligned} P_{uu} &= p_u^2 + p_u\left(1-p_u\right) f_{uu} \\ P_{uv} &= 2p_u p_v\left(1-f_{uv}\right). \end{aligned}$$

But, we are not done. There are relationships among these parameters. If there are n different alleles, there are $n-1$ free allele frequencies p_u and $\frac{n(n+1)}{2}$ correlation coefficients f_{uv}, for a total of $\frac{n^2+3n-2}{2}$ parameters. However, there are only $\frac{n(n+1)}{2}-1$ free genotype counts in the data. How many relationships are there among the parameters?

There are

$$\begin{aligned} &\text{d.f. in parameters} - \text{d.f. in data} \quad = \\ &\frac{n^2+3n-2}{2} - \frac{n(n+1)}{2} + 1 \quad = \quad n \end{aligned}$$

relationships among the model parameters.

To find the relationships among the parameters, recall that

$$p_u \quad = \quad P_{uu} + \frac{1}{2}\sum_{v \neq u} P_{uv}.$$

There are n such relationships, and they will consume all the extra degrees of freedom. Substitute in the expressions for the genotype frequencies to observe

$$f_{uu} = \sum_{v \neq u} \frac{p_v}{1-p_u} f_{uv}.$$

Covariation Between >2 Alleles (Model 2)

Suppose that associations between alleles are not a consequence of specific interactions among alleles. Suppose instead that there is a general association between alleles regardless of allele identity. (We will discuss how this can arise later.) Then, there is one correlation that applies to all pairs of alleles u and v and the applicable equations are again

$$\begin{aligned} P_{uu} &= p_u^2 + p_u\left(1-p_u\right) f \\ P_{uv} &= 2p_u p_v(1-f) \text{ where } u \neq v. \end{aligned}$$

There are $n-1$ free allele frequencies p_u and 1 free correlation f, leading to n free parameters. There are plenty of degrees of freedom $\frac{n(n+1)}{2} - 1$ in the data to cover these parameters.

2.2.2 Additive Disequilibrium D

Additive Disequilibrium

But all is not perfect. Notice, f appears as a multiplier of allele frequencies. We model disequilibrium as a multiplicative factor that expands/shrinks genotype counts away from HWE expectation. We derived a MOM estimate for correlation f previously. In general, estimation of multiplicative factors involves ratios of statistics. Ratios are notoriously difficult statistically.

An easier approach, statistically, is to look for additive disequilibrium. Define

$$\begin{aligned} D_{uu} &= P_{uu} - p_u^2 \\ D'_{uv} &= P_{uv} - 2p_u p_v. \end{aligned}$$

Let $D_{uv} = -2D'_{uv}$, to write more conveniently

$$\begin{aligned} P_{uu} &= p_u^2 + D_{uu} \\ P_{uv} &= 2p_u p_v - 2D_{uv} \text{ whenever } u \neq v. \end{aligned}$$

d.f. for Additive Disequilibrium

Again, there must be n relationships among the parameters to account for the difference in degrees of freedom.

$$\begin{aligned} p_u &= P_{uu} + \frac{1}{2}\sum_{v \neq u} P_{uv} \\ &= p_u^2 + D_{uu} + \sum_{v<u} (p_u p_v - D_{uv}) \\ &= p_u \sum_{v \leq u} p_v + D_{uu} - \sum_{v<u} D_{uv} \\ &= p_u + D_{uu} - \sum_{v<u} D_{uv} \end{aligned}$$

leading to

$$D_{uu} = \sum_{v<u} D_{uv}.$$

Range of Additive Disequilibrium

Because $0 \leq P_{uu}, P_{uv} \leq 1$, we also recognize that the additive disequilibrium can't be just anything. In fact, the total list of constraints is

$$\begin{aligned} -p_u^2 \leq D_{uu} &\leq 1 - p_u^2 \\ p_u p_v - \frac{1}{2} \leq D_{uv} &\leq p_u p_v \\ D_{uu} &= \sum_{v<u} D_{uv} \end{aligned}$$

In the case of just two alleles at a locus 1 and 2, then $D_{11} = D_{12} = D_{22}$, so

$$\begin{aligned} P_{11} &= p_1^2 + D_1 \\ P_{12} &= 2p_1 p_2 - 2D_1 \\ P_{22} &= p_2^2 + D_1 \end{aligned}$$

and

$$\max_{u \in \{1,2\}} -p_u^2 \leq D_1 \leq p_1 p_2.$$

Testing $D_1 = 0$ (HWE)

Testing for HWE is equivalent to testing the null hypothesis

$$H_0 : D_1 = 0.$$

Here, we have restricted to the two allele case.

We need two things:

- An estimate $\hat{D}_1$. Is it close to 0?
- A sampling distribution for the estimate $\hat{D}_1$ to determine whether it is farther from 0 than we would expect by chance.

Estimating D_1

Two free parameters p_1, D_1 and two free pieces of data n_{11}, n_{12} suggests Bailey's method:

$$\begin{aligned} n_{11} &= n\left(p_1^2 + D_1\right) \\ n_{12} &= n\left(2p_1p_2 - 2D_1\right) \end{aligned}$$

can be solved to produce

$$\begin{aligned} \hat{p}_1 &= \frac{2n_{11} + n_{12}}{2n} = \tilde{p}_1 \\ \hat{D}_1 &= \frac{n_{11}}{n} - \tilde{p}_1^2 = \tilde{P}_{11} - \tilde{p}_1^2. \end{aligned}$$

$\hat{D}_1$ Bias

Is our estimate $\hat{D}_1$ unbiased?

$$\begin{aligned} \mathrm{E}\left(\hat{D}_1\right) &= \mathrm{E}\left(\tilde{P}_{11}\right) - \mathrm{E}\left(\tilde{p}_1^2\right) \\ &= P_{11} - p_1^2 - \frac{1}{2n}\left(p_1 + P_{11} - 2p_1^2\right) \\ &= D_1 - \frac{1}{2n}\left[p_1\left(1 - p_1\right) + D_1\right]. \end{aligned}$$

Since $\mathrm{E}\left(\hat{D}_1\right) \neq D_1$ we conclude that the estimator is biased. However, we are encouraged to note that as $n \to \infty$, the bias goes to 0.

$\hat{D}_1$ Sampling Variance

Using Fisher's approximation (it applies because $\hat{D}_1$ is a function of proportions), we obtain an approximate sampling variance for $\hat{D}_1$

$$\mathrm{Var}\left(\hat{D}_1\right) \approx \frac{1}{n}\left[p_1^2\left(1 - p_1\right)^2 + \left(1 - 2p_1\right)^2 D_1 - D_1^2\right].$$

To estimate the sampling variance, we substitute in our estimates for p_1 and D_1

$$\mathrm{Var}\left(\hat{D}_1\right) \hat{=} \frac{1}{n}\left[\hat{p}_1^2\left(1 - \hat{p}_1\right)^2 + \left(1 - 2\hat{p}_1\right)^2 \hat{D}_1 - \hat{D}_1^2\right].$$

Since $\hat{D}_1$ is the MLE, we have for large samples that

$$\hat{D}_1 \sim N\left[\mathrm{E}\left(\hat{D}_1\right), \mathrm{Var}\left(\hat{D}_1\right)\right].$$

Testing Additive Disequilibrium: z Test

Testing $H_0 : D_1 = 0$ Using z-Values

Compute the standard normal variate z

$$z = \frac{\hat{D}_1 - \mathrm{E}\left(\hat{D}_1\right)}{\sqrt{\mathrm{Var}\left(\hat{D}_1\right)}}$$

Under the null hypothesis H_0, z approximately follows the standard normal distribution. Compare z against standard normal distribution.

The key is that if $\hat{D}_1$ is very positive or very negative, then z will tend to be far from 0 and your statistic will fall in the tails of the sampling distribution, where it is not expected to fall if the null hypothesis of HWE is true.

Computing z

Under the null hypothesis, we know

$$\begin{aligned} \mathrm{E}\left(\hat{D}_1\right) &\approx 0 \\ \mathrm{Var}\left(\hat{D}_1\right) &\approx \frac{1}{n}\left[\hat{p}_1^2\left(1-\hat{p}_1\right)^2\right]. \end{aligned}$$

so

$$z = \frac{\sqrt{n}\hat{D}_1}{\hat{p}_1\left(1-\hat{p}_1\right)}$$

Note, we have assumed n is sufficiently large that the bias term is negligible.

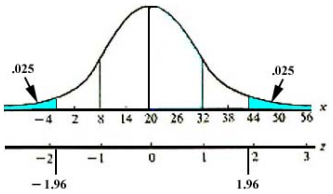
Relevant Alternative Hypotheses

$$\begin{aligned} P_{11} &= p_1^2 + D_1 \\ P_{12} &= 2p_1p_2 - 2D_1 \\ P_{22} &= p_2^2 + D_1 \end{aligned}$$

Depending on your purpose, there may be different alternative hypotheses you consider. Suppose $z = 1.25$, then

- $H_A : D_1 > 0$ or $D_1 < 0$. You have no *a priori* feeling for whether heterozygotes will be over- or under-represented. Use a **two-tailed test**.

```
> 2*pnorm(q=-1.25)
[1] 0.2112995
> 2*(1-pnorm(q=1.25))
[1] 0.2112995
```



One-side hypotheses are appropriate when you suspected heterozygotes would either be under- or over-represented *before* you collected the data.

- $H_A : D_1 > 0$. You suspect that heterozygotes will be under-represented. Use a **one-tailed (right tail) test**.

```
> (1-pnorm(q=1.25))
[1] 0.1056498
```

- $H_A : D_1 < 0$. You suspect that heterozygotes will be over-represented. Use a **one-tailed (left tail) test**.

```
> pnorm(q=1.25)
[1] 0.8943502
```

Testing Additive Disequilibrium: Chi-Square Tests

Chi-Square

An equivalent test is the Chi-Square test for HWE. It depends on comparing z^2 against its sampling distribution, which under the null, is a chi-square distribution with 1 degree of freedom.

$$z^2 = X_1^2 = \frac{n\hat{D}_1^2}{\hat{p}_1^2\left(1-\hat{p}_1\right)^2}.$$

However, note that both positive and negative values of z give the same z^2 statistic. It is not so easy to consider one-sided alternative hypotheses. Since the tests are equivalent, use the z statistic for one-sided tests.

Test Assumptions: The sample size is large so both normality (or chi-square) applies and bias can be ignored.

Chi-Square Goodness-of-Fit

Genotype	11	12	22
Observed (O)	n_{11}	n_{12}	n_{22}
Expected* (E)	$n\hat{p}_1^2$	$2n\hat{p}_1\left(1-\hat{p}_1\right)$	$n\left(1-\hat{p}_1\right)^2$
Observed - Expected	$n\hat{D}_1$	$-2n\hat{D}_1$	$n\hat{D}_1$

*Here, we have made the assumption that n is sufficiently large that the bias terms are 0.

The goodness-of-fit chi-square statistic is defined as

$$\begin{aligned} Y_1^2 &= \sum_{\text{genotypes}} \frac{(O-E)^2}{E} \\ &= \frac{\left(n\hat{D}_1\right)^2}{n\hat{p}_1^2} + \frac{\left(-2n\hat{D}_1\right)^2}{2n\hat{p}_1\left(1-\hat{p}_1\right)} + \frac{\left(n\hat{D}_1\right)^2}{n\left(1-\hat{p}_1\right)^2}. \end{aligned}$$

and you can show that $Y_1^2 = X_1^2$.

Standard Cautions About Chi-Square Tests

The fact that the z-test is equivalent to the chi-square goodness-of-fit test allows us to infer some conditions in the data that may lead to invalid inferences. We here state the typical "cautions" that accompany use of the chi-square goodness-of-fit test.

- **Low expected counts $E < 5$ invalidate asymptotics.** Because the expected counts E appear in the denominator, small variation when they are small results in huge changes in X_1^2. It is hard to quantitate the rule, but roughly we hope no more than 20% of the cells have $E < 5$.

- **Data is discrete, but asymptotics assume continuous statistics.** Because the observed data are discrete, but the sampling distribution (normal or chi-square) is continuous, the Yates' correction is recommended to avoid effects of discreteness:

$$X_1^2 = \sum_{\text{genotypes}} \frac{(|O - E| - 0.5)^2}{E}$$

In short, we should be cautious about applying any of the above tests when sample size is low or counts in genotype categories are low.

Test 3: Likelihood Ratio Test

Likelihood Ratio [Review]

Suppose your null hypothesis is

$$H_0 : \phi = \phi_0$$

for some parameter ϕ. Let the maximum likelihood value under H_0 be L_0 and the maximum likelihood value without the restriction on ϕ be L_1. Then, L_0 will always be smaller than L_1 since ϕ_0 may not be the maximum likelihood value of ϕ. However, if the null is true, $\hat{\phi}$ should be very close to ϕ_0 and L_0 will be very close to L_1.

Define the likelihood ratio as

$$\lambda = \frac{L_0}{L_1}.$$

When the null hypothesis is true and the size of ϕ is s, then

$$-2 \ln \lambda \sim \chi^2_{(s)}.$$

Likelihood Ratio Test for HWE

Under the unconstrained model (the alternative hypothesis), the parameters are p_{11}, p_{12}, p_{22} and the data are n_{11}, n_{12}, n_{22}. There are two degrees of freedom in the model and the data, so Bailey's method applies to yield

$$\begin{aligned} \hat{p}_{11} &= \frac{n_{11}}{n} \\ \hat{p}_{12} &= \frac{n_{12}}{n}. \end{aligned}$$

The maximum likelihood under the unconstrained model is

$$L_1 = \frac{n!}{n_{11}!n_{12}!n_{22}!} \left(\frac{n_{11}}{n}\right)^{n_{11}} \left(\frac{n_{12}}{n}\right)^{n_{12}} \left(\frac{n_{22}}{n}\right)^{n_{22}}$$

Under the constrained model,

$$\begin{aligned} \hat{P}_{11} &= \hat{p}_1^2 \\ \hat{P}_{12} &= 2\hat{p}_1 (1 - \hat{p}_1) \end{aligned}$$

with $\hat{p}_1 = \frac{n_1}{2n}$.

The maximum likelihood under the constrained model is

$$L_0 = \frac{n!}{n_{11}!n_{12}!n_{22}!} \left(\frac{n_1}{2n}\right)^{2n_{11}} \left(\frac{2n_1 n_2}{(2n)^2}\right)^{n_{12}} \left(\frac{n_2}{2n}\right)^{2n_{22}}.$$

The test statistic is therefore

$$-2\ln\lambda = -2\ln\left[\frac{n^n n_1^{2n_{11}}(2n_1n_2)^{n_{12}}n_2^{2n_{22}}}{(2n)^{2n}n_{11}^{n_{11}}n_{12}^{n_{12}}n_{22}^{n_{22}}}\right]$$

Cautions. As we have discussed, the asymptotic chi-square distribution applies as sample size increases and as long as parameters are real numbers existing on an interval and with MLES *inside* that interval. In other words the discreteness of the data (and possible parameter values) and the possibility of estimates on the boundary (for example, when counts are 0) also cause problems for this test.

Akaike Information Criterion (AIC)

How can we compare models that are not nested? We will not discuss hypothesis testing in this context, but will discuss ways to rank alternative models from best to worst.

Definition: *AIC*

The AIC provides a measure of goodness of fit of a model. It is

$$AIC = 2k - 2\ln(L)$$

Definition: *AICc*

The corrected AIC includes a correction for small sample size. It is

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

In both cases, k is the number of parameters in the model and L is the maximized likelihood under the model.

- As the model complexity increases, k increases and AIC increases.
- As the model fit improves, L decreases and AIC decreases.
- Thus, AIC is purported to handle the trade-off between model complexity and data fit.
- It is *not* used for hypothesis testing, but for ranking models and identifying better fitting models and worse fitting models.

2.2.3 Multiplicative Disequilibrium M

A Multiplicative Model That Works

Let us consider another multiplicative model

$$\begin{aligned} P_{11} &= MM_1^2M_{11} \\ P_{12} &= 2MM_1M_2M_{12} \\ P_{22} &= MM_2^2M_{22} \end{aligned}$$

Here M is the mean effect, M_{11}, M_{12}, M_{22} represent associations between allele frequencies, and M_1, M_2 represent the allele frequency contributions.

Taking logarithms puts this model back in the additive space

$$\begin{aligned} \ln P_{11} &= \ln M + 2\ln M_1 + \ln M_{11} \\ \ln P_{12} &= \ln M + \ln 2 + \ln M_1 + \ln M_2 + \ln M_{12} \\ \ln P_{22} &= \ln M + 2\ln M_2 + \ln M_{22} \end{aligned}$$

Handling Overparameterization

There are, as usual, more parameters than observations. And this time there are multiple ways to deal with the overparameterization. One way is to set

$$\begin{aligned} M_2 M_{12} &= 1 \\ M_2^2 M_{22} &= 1, \end{aligned}$$

then

$$\begin{aligned} P_{11} &= M M_1^2 M_{11} \\ P_{12} &= 2 M M_1 \\ P_{22} &= M. \end{aligned}$$

There is still an extra degree of freedom, but summing all three equations yields

$$1 = M\left(M_1^2 M_{11} + 2M_1 + 1\right) \quad \text{or} \quad M = \frac{1}{1 + 2M_1 + M_1^2 M_{11}}.$$

Estimating Parameters M_1 and M_{11}

Again, Bailey's method applies and the maximum likelihood estimates are

$$\begin{aligned} \hat{M_1} &= \frac{n_{12}}{2n_{22}} \\ \hat{M_{11}} &= \frac{4n_{11}n_{22}}{n_{12}^2} \end{aligned}$$

with the tag-along

$$\hat{M} = \frac{n_{22}}{n}.$$

Substituting these MLEs back into the original multiplicative equations produces the same likelihood under H_A (believe it or not).

$$\begin{aligned} \hat{P}_{11} &= \frac{n_{11}}{n} \\ \hat{P}_{12} &= \frac{n_{12}}{n} \\ \hat{P}_{22} &= \frac{n_{22}}{n} \end{aligned} \qquad L_1 = \frac{n!}{n_{11}!n_{12}!n_{22}!}\left(\frac{n_{11}}{n}\right)^{n_{11}}\left(\frac{n_{12}}{n}\right)^{n_{12}}\left(\frac{n_{22}}{n}\right)^{n_{22}}$$

Log Likelihood Test for Multiplicate Model

HWE implies no interaction term, i.e. $M_{11} = 1$. Under this constraint, we again apply Bailey's method to find

$$\begin{aligned} \hat{P}_{11} &= \left(\frac{n_1}{2n}\right)^2 \\ \hat{P}_{12} &= \frac{n_1 n_2}{2n^2} \\ \hat{P}_{22} &= \left(\frac{n_2}{2n}\right)^2. \end{aligned}$$

It turns out that λ has the same form under this log-linear model as the additive disequilibrium model. So the log-linear model for testing HWE is equivalent to the additive model.

Summary of Tests for HWE

- Normal approximation for MLEs uses the z statistic.
- Chi-square test uses the $X_1^2 = z^2$ statistic and is equivalent to the above test. It is also equivalent to a chi-square goodness-of-fit test (which importantly extends easily to the multi-allele case).
- The likelihood ratio test is widely applicable and flexible when a likelihood function is available.
- The log-linear model uses a multiplicative model and leads to a test equivalent to the likelihood ratio test.
- All of the above tests rely on asymptotic results that require large sample sizes and non-trivial counts in all genotype categories.
- The exact test is useful when the data set is small and particularly when counts in some categories are small.

2.2.4 Multiple Alleles

Tests for Multiple Alleles

When there are more than two alleles at a locus, the general equations

$$\begin{aligned} P_{uu} &= p_u^2 + D_{uu} \\ P_{uv} &= 2p_u p_v - 2D_{uv} \text{ whenever } u < v \end{aligned}$$

apply, with relationship

$$D_{uu} = \sum_{v>u} D_{uv} + \sum_{v<u} D_{vu}$$

and MLEs obtained by Bailey's method (verify this; it is not hard)

$$\begin{aligned} \hat{p}_u &= \tilde{p}_u \\ \hat{D}_{uv} &= \tilde{p}_u \tilde{p}_v - \frac{1}{2}\tilde{P}_{uv}. \end{aligned}$$

We will first consider tests for *complete HWE*, where $D_{uv} = 0$ for all D_{uv}.

Testing Complete HWE

Likelihood Ratio Test

Under complete HWE,

$$H_0 : D_{uv} = 0 \text{ for all } u \neq v,$$

a hypothesis with $\frac{k(k-1)}{2}$ constraints (one for each heterozygote) when there are k alleles. The remaining $k-1$ free parameters p_u have easy-to-obtain MLEs $\hat{p}_u = \frac{n_u}{2n}$.

Under the alternative hypothesis, all $\frac{k(k-1)}{2}$ disequilibrium D_{uv} parameters are free to vary and all $k-1$ allele frequencies p_u are free to vary for a total of $\frac{k(k+1)}{2}$ free parameters. Notice, there is a one-to-one relationship between paramters D_{uv}, p_u and genotype frequencies P_{uv}. Therefore, the likelihood under the alternative hypothesis is the multinomial distribution

$$\{n_{uv}\} \sim \text{Multi}(n, P_{uv})$$

and MLEs $\hat{P}_{uv} = \frac{n_{uv}}{n}$ are easy to obtain.

Therefore,

$$\lambda = \frac{\ln L(n_{uv}; \hat{p}_u)}{\ln L(n_{uv}; \hat{P}_{uv})}$$

and, asymptotically,

$$-2 \ln \lambda \sim \chi^2_{k(k-1)/2}$$

Chi-Square Goodness-of-Fit

If you are more comfortable with a goodness-of-fit test, that statistic

$$X^2_T = \sum_{u \leq v} \frac{(|n_{uv} - E(n_{uv})| - 0.5)^2}{E(n_{uv})}$$

with

$$\begin{aligned} E(n_{uu}) &= n\tilde{p}_u^2 \\ E(n_{uv}) &= 2n\tilde{p}_u\tilde{p}_v. \end{aligned}$$

follows the same sampling distribution.

Testing Partial HWE

More Complex HW Hypotheses

If you want to test only certain combinations of alleles, the tests are more complicated.

- **Test a single D_{uv}. Example:** If $H_0 : D_{12} = 0$, the likelihood ratio statistic $-2 \ln \lambda$ follows a chi-square with 1 degree of freedom. But Bailey's method does not apply unless $k = 2$ so L_0 is difficult to compute. Iterative methods are required.

  Or, one could apply the z-test or the chi-square test. Details for these tests follow on the next slide.

- **Test multiple, but not all D_{uv}. Example:** If $k = 4$ and $H_0 : D_{12}, D_{34} = 0$, $-2 \ln \lambda$ follows a chi-square with 2 degrees of freedom. Iterative methods are still required.

  This time, there are no easy alternatives. Complex hypotheses cause difficulties for the z-test and related chi-square. The likelihood ratio test handles complex hypotheses quite naturally.

z-Test for Partial HWE

Here, you only need the MLEs under the full alternative model (where Bailey's method applies). For large samples, the MLE $\hat{D}_{uv}$ is approximately normally distributed under H_0 with mean 0 and variance $\text{Var}\left(\hat{D}_{uv}\right)$.

$$z_{uv} = \frac{\hat{D}_{uv}}{\sqrt{\text{Var}\left(\hat{D}_{uv}\right)}}$$

We can again use Fisher's approximation to compute the variance.

$$\begin{aligned} \text{Var}\left(\hat{D}_{uv}\right) = \frac{1}{2n} \Bigg\{ & p_u p_v \left[(1-p_u)(1-p_v) + p_u p_v\right] \\ & - \left[(1-p_u-p_v)^2 - 2(p_u-p_v)^2\right] D_{uv} \\ & + \sum_{w \neq u,v} (p_u^2 D_{vw} + p_v^2 D_{uw}) - D_{uv}^2 \Bigg\} \end{aligned}$$

Under the H_0, the variance is obtained by assuming $D_{uv} = 0$.

2.2.5 Bayesian Hypothesis Testing

OpenBUGS for ABO

```
model {
   # likelihood
   pi[1] <- p.a*p.a + f*p.a*(1-p.a) + 2*p.a*p.o*(1-f)
   pi[2] <- 2*p.a*p.b*(1-f)
   pi[3] <- p.b*p.b + f*p.b*(1-p.b) + 2*p.b*p.o*(1-f)
   pi[4] <- p.o*p.o + f*p.o*(1-p.o)
   x[1:4] ~ dmulti(pi[],n)

   # priors
   a1 ~ dexp(1)
   b1 ~ dexp(1)
   o1 ~ dexp(1)
   p.a <- a1/(a1 + b1 + o1)
   p.b <- b1/(a1 + b1 + o1)
   p.o <- o1/(a1 + b1 + o1)

   f ~ dunif(0,1)

   n <- sum(x[])
}

list(x=c(862, 131, 365, 702))
```

OpenBUGS Results

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
f	0.4078	0.1339	3.030-4	0.06618	0.4365	0.5969	3000	200002
p.a	0.3491	0.02567	5.836e-5	0.2900	0.3527	0.3908	3000	200002
p.b	0.1620	0.01370	3.099e-5	0.1322	0.1632	0.1861	3000	200002
p.o	0.4890	0.03723	8.430e-5	0.4301	0.4832	0.5760	3000	200002
p.a	0.2813	0.0007579	4.079e-5	0.2666	0.2813	0.2962	1000	38002
p.b	0.1293	0.0005368	2.717e-5	0.1190	0.1292	0.1400	1000	38002
p.o	0.5894	0.0008373	4.343e-5	0.5729	0.5894	0.6058	1000	38002

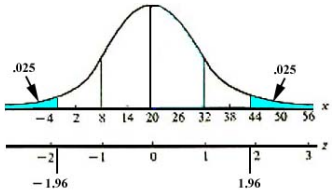
Deviance Information Criterion (DIC)

Model	Dbar	Dhat	DIC	pD
$f > 0$	25.09	22.42	27.76	2.673
$f = 0$	27.79	25.79	29.79	2.002

- $\bar{D}$: posterior mean $-2\ln(L)$
- $\hat{D}$: $-2\ln(L)$ at posterior mean
- pD: measure of model complexity, "effective number of parameters"
- DIC is "analogous" to AIC and can be used to rank models. "A difference of 7-10 units is considered strong evidence in favor of one model over another."

2.3 Exact tests

Exact Tests



If the probability of the observed sample under the null hypothesis and all less likely samples is small (aka the p-value), then the evidence suggests the data is unlikely to have arisen under the null hypothesis. If one can compute the probability of all possible samples, then obtaining an exact probability (no approximation) is possible.

Exact tests are useful when all the possible observed data can be enumerated practically. This occurs generally when the expected counts are small in some categories (i.e. when the previous tests fail).

The probability of the observed data n_{11}, n_{12}, n_{22} is given by the multinomial distribution

$$P\left(n_{11}, n_{12}, n_{22}\right) = \frac{n!}{n_{11}!n_{12}!n_{22}!} P_{11}^{n_{11}} P_{12}^{n_{12}} P_{22}^{n_{22}}.$$

Exact Tests for HWE

When HWE applies, then

$$P\left(n_{11}, n_{12}, n_{22}\right) = \frac{n!}{n_{11}!n_{12}!n_{22}!} p_1^{2n_{11}} \left(2p_1p_2\right)^{n_{12}} p_2^{2n_{22}}.$$

In addition, the allele counts n_1 and n_2 are binomially distributed (because of independence of alleles and "sampling genotypes is like sampling alleles")

$$P\left(n_1, n_2\right) = \frac{(2n)!}{n_1!n_2!} p_1^{n_1} p_2^{n_2}.$$

Exact tests condition on the marginal allele counts (i.e. n_1, n_2). There has been a long debate whether it is appropriate to condition on these counts when they are themselves unknown prior to data collection. For some motivation, consider that it is the distribution of these alleles among the genotypes that is under question for testing HWE, not how many of each allele there are. In practice, the effect of conditioning is probably insignificant.

The Conditional Probability

The conditional probability, where we condition on the observed allele counts is

$$\begin{aligned} P\left(n_{11}, n_{12}, n_{22} \mid n_1, n_2\right) &= \frac{P\left(n_{11}, n_{12}, n_{22}, n_1, n_2\right)}{P\left(n_1, n_2\right)} \\ &= \frac{P\left(n_{11}, n_{12}, n_{22}\right)}{P\left(n_1, n_2\right)} \\ &= \frac{n!n_1!n_2!2^{n_{12}}}{n_{11}!n_{12}!n_{22}!(2n)!}. \end{aligned}$$

Exact Test for HWE (Example)

Suppose we observe $n_{11} = 10, n_{12} = 1, n_{22} = 2$. Use an exact test to calculate a p-value for rejecting the null hypothesis of HWE.

n_{11}	n_{12}	n_{22}	Probability	Cumul. Prob.
10	1	2	9.1×10^{-5}	9.1×10^{-5}
9	3	1	0.35	0.35
8	5	0	0.63	0.97

The p-value is $p = 9.1 \times 10^{-5}$ since there is no dataset more extreme than the observed.

Exact Tests for Multiple Alleles

The exact test generalizes to multiple alleles. The formula is

$$P\left(\{n_{uv}\} \mid \{n_u\}\right) = \frac{n! 2^H \prod_u n_u!}{(2n)! \prod_{u,v} n_{uv}!}.$$

where H is the total number of heterozygous individuals.

Let's consider the following genotype data for a locus with 4 alleles:

	A_1	A_2	A_3	A_4
A_1	2	16	13	5
A_2	-	15	21	12
A_3	-	-	9	7
A_4	-	-	-	0

GenePOP example.

Exact Tests: Computational Difficulties

So, the computations become quickly overwhelming. Approximate exact tests were developed for this situation.

- GUO, S. and THOMPSON, E. 1992 . Performing the exact test of Hardy Weinberg proportion for multiple alleles. Biometrics 48 pp. 361-372.
- LAZZERONI, L. C. and LANGE, K. 1997. Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. Ann. Statist. 25 pp. 138-168.

Approximate Exact Tests

When it is impossible to enumerate all the possible datasets with the same allele frequencies, approximate methods are needed. One of the simplest uses Monte Carlo.

- Calculate $F = P\left(\{n_{uv}\} \mid \{n_u\}\right)$ for the observed data.
- Set $S = 0$ and put all your genotypes $(13, 66, 31, 16, \ldots)$ in a big vector of length $2n$.

$$13663116\cdots$$

- Permute all alleles and clump successive alleles into genotypes.

$$(36)(16)(36)(13)\cdots$$

- Compute $F^* = P\left(\{n^*_{uv}\} \mid \{n^*_u\}\right)$ for the permuted dataset.
- If $F^* \leq F$, increment S by 1.
- Repeat M times.
- Estimate the p-value as $\frac{S}{M}$.

Example

Consider the following table of genotype counts for a locus with four alleles.

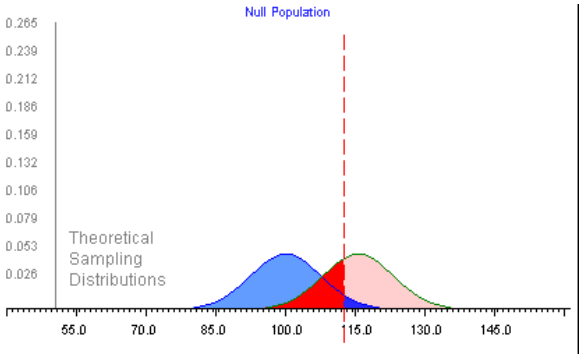
A_1	0			
A_2	3	1		
A_3	5	18	1	
A_4	3	7	5	2
	A_1	A_2	A_3	A_4

Method	Estimate
Exact	0.01744
χ^2	0.02337
MC integration ($M = 1700$)	0.01706

2.4 Power calculations

Power Calculations

Recall that the **power** $(1 - \beta)$ of a statistical test is the probability that you reject the null hypothesis when it is false.



There are two main goals of power analysis:

- to estimate how large a sample you will need to make accurate and reliable statistical conclusions
- to determine the probability that your test will detect effects (e.g. disequilibria) of a certain size in a particular situation

Power Calculation: HWD

Let us suppose you are about to collect data and you are interested in detecting disequilibrium at least of size $D_1 = 0.10$. To detect it means you reject the null that $D_1 = 0$.

The test statistic

$$X_1^2 = \frac{n\hat{D}_1^2}{\hat{p}_1^2\left(1 - \hat{p}_1\right)^2}$$

follows a different distribution depending on whether H_0 is true or not.

H_0 true	H_0 false
$X_1^2 \sim \chi^2_{(1)}$	$X_1^2 \sim \chi^2_{(1,\nu)}$

where $\chi^2_{(1,\nu)}$ is the noncentral chi-square distribution with noncentrality parameter ν.

The Noncentrality Parameter

The noncentral chi-square distribution is an approximate distribution under H_A. The noncentrality parameter is given by

$$\nu = \frac{nD_1^2}{p_1^2\left(1-p_1\right)^2}$$

ν is bigger when D_1 is farther from 0. But note, the approximation is only valid when $\frac{\nu}{n}$ is small, say of order n^{-1}.

If we take the standard significance level $\alpha = 0.05$, then we will reject H_0 if $X_1^2 > 3.84$. With this knowledge one can ask a couple of questions

- How big does D_1 need to be in order to have 90% chance of getting $X_1^2 > 3.84$ and therefore rejecting H_0 for a sample of size $n = 100$ in a population with $p_1 = 0.1$?
- How big does my sample n need to be in order to detect a disequilibrium $D_1 = 0.1$ with 90% probability?

Size of D_1

```
> pchisq( q=qchisq(p=0.05, df=1, lower.tail=F), df=1, ncp=10.5, lower.tail=F )
[1] 0.899799
```

From the above calculation, we see that when $\nu = 10.5$, then X_1^2 will exceed 3.84 with 90% probability.

Rearrange the noncentrality formula to find the smallest D_1 big enough to achieve the desired power:

$$D_1 = p_1(1-p_1)\sqrt{\frac{\nu}{n}} \approx 0.029$$

where we have plugged in our assumed $p_1 = 0.1$ and our assumed sample size $n = 100$.

The trick for computing noncentrality parameters giving a desired power:

```
< ncp <- qchisq(1-beta, df=1, ncp=qchisq(1-alpha, df=1))
```

where `1−beta` is your desired power, and `alpha` is your desired type I error.

Sample Size n

Also, further rearrangement yields

$$n = \frac{\nu p_1^2(1-p_1)^2}{D_1^2}$$

which can be used to compute the size of the sample we'll need to detect a specified disequilibrium D_1.

For example, if $p_1 = 0.1$ as before, the sample size we need to detect $D_1 = 0.01$ is

$$n = \frac{10.5 \times 0.1^2 \times 0.9^2}{0.01^2} = 850.5$$

Power Calculations for Exact Tests

Recall, you need to calculate

$$P\left(\{n_{uv}\} \mid \{n_u\}\right) = \frac{P\left(\{n_{uv}\}\right)}{P\left(\{n_u\}\right)}$$

Under the alternative hypothesis, the genotype frequencies are given by the formulas

$$\begin{aligned} P_{uu} &= p_u^2 + D_{uu} \\ P_{uv} &= 2p_u p_v - 2D_{uv} \text{ whenever } u \neq v, \end{aligned}$$

so you can compute the numerator

$$P\left(\{n_{uv}\}\right) = \frac{n!}{\prod_{u \leq v} n_{uv}!} \prod_{u \leq v} P_{uv}^{n_{uv}}$$

but you can no longer assume the binomial distribution for allele frequencies.

But, of course,

$$n_u = n_{uu} + \frac{1}{2} \sum_{u<v} n_{uv}.$$

so if you know $P\left(\{n_{uv}\}\right)$, you can compute $P\left(\{n_u\}\right)$ by summing over the former.