Lecture 4: Testing for Departures from Hardy-Weinberg Equilibrium

August 31, 2012
Last Time

- Introduction to statistical distributions
- Estimating allele frequencies
- Introduction to Hardy-Weinberg Equilibrium
Today

- Hardy-Weinberg Equilibrium Continued
- Using Hardy-Weinberg: Estimating allele frequencies for dominant loci
- Hypothesis testing
What is a Population?

- **Operational definition:** an assemblage of individuals
- **Population genetics definition:** a collection of randomly mating individuals
- **Why does this matter?**
Measuring diversity

- Allele frequency is same as sampling probability
- Two allele system: frequency of one allele provides frequency of other: p and q
- Homozygotes: individuals with the same allele at both homologous loci
- Heterozygotes: individuals with different alleles at homologous loci
Dominance and Additivity

- **Dominance**: masking of action of one allele by another allele
  - Homozygotes indistinguishable from heterozygotes

- **Additivity**: phenotype can be perfectly predicted from genotype
  - Intermediate heterozygote

- **Codominant**: both alleles are apparent in genotype: does NOT refer to phenotype!
Hardy-Weinberg Law

- Hardy and Weinberg came up with this simultaneously in 1908

- Frequencies of genotypes can be predicted from allele frequencies following one generation of random mating

- Assumptions:
  - Very large population
  - Random mating
  - No selection
  - No migration
  - No mutation
### Hardy-Weinberg Law and Probability

<table>
<thead>
<tr>
<th></th>
<th>A(p)</th>
<th>a(q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(p)</td>
<td>AA ($p^2$)</td>
<td>Aa ($pq$)</td>
</tr>
<tr>
<td>a(q)</td>
<td>aA ($qp$)</td>
<td>aa ($q^2$)</td>
</tr>
</tbody>
</table>

\[ p^2 + 2pq + q^2 = 1 \]
What about a 3-Allele System?

- **Alleles occur in gamete pool at same frequency as in adults**
- **Probability of two alleles coming together to form a zygote is** $A \cap B$

$$
\begin{align*}
A_1A_1 &= p^2 \\
A_1A_2 &= 2pq \\
A_2A_3 &= 2qr \\
A_1A_3 &= 2pr \\
A_2A_2 &= q^2 \\
A_3A_3 &= r^2
\end{align*}
$$

- **Equilibrium established with ONE GENERATION of random mating**
- **Genotype frequencies remain stable as long as allele frequencies remain stable**
- **Remember assumptions!**
Genotype Frequencies Under Hardy-Weinberg

- Frequency of heterozygotes is maximum at intermediate allele frequencies

\[ \frac{d(2pq)}{dq} = \frac{d(2q(1-q))}{dq} \]

\[ = \frac{d}{dq}(2q - 2q^2) = 2 - 4q \]

\[ 0 = 2 - 4q \]

\[ q = 0.5 \]
At extreme allele frequencies, most copies of the minor allele are in heterozygotes, not homozygotes.

Recessive alleles are “hidden” from selection.
Frequencies of genotypes can be predicted from allele frequencies following one generation of random mating.

Allele frequencies remain constant.

Why?
Derivation of Hardy-Weinberg from Genotype Frequencies

<table>
<thead>
<tr>
<th></th>
<th>Moms</th>
<th>Dads</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Genotype Frequency</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>AA</td>
<td>X</td>
<td>X²</td>
</tr>
<tr>
<td>Aa</td>
<td>Y</td>
<td>XY</td>
</tr>
<tr>
<td>aa</td>
<td>Z</td>
<td>ZX</td>
</tr>
</tbody>
</table>

\[
\text{freq}(AA) = \frac{1}{4}Y^2 \\
\text{freq}(Aa) = \frac{2}{4}Y^2 \\
\text{freq}(aa) = \frac{1}{4}Y^2
\]

Aa x Aa = \frac{1}{4} AA + \frac{2}{4} Aa + \frac{1}{4} aa
## Derivation of Hardy-Weinberg from Genotype Frequencies

### Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Parental Mating</th>
<th>Frequency</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa x Aa</td>
<td>$Y^2$</td>
<td>$Y^2/4$</td>
<td>$2(Y^2)/4$</td>
<td>$Y^2/4$</td>
</tr>
</tbody>
</table>

\[
\text{frequency}(Aa) = XY + 2XZ + 2\frac{Y^2}{4} + YZ
\]
\[
= 2\left(\frac{XZ}{2} \cdot \frac{1}{2} \cdot \frac{XY}{2} + \frac{1}{2} YZ + \frac{Y^2}{4}\right)
\]
\[
X + \frac{1}{2} Y = \frac{N_{11}}{N} + \frac{1}{2} \left(\frac{N_{12}}{N}\right) = p
\]
\[
\left(Z + \frac{1}{2} Y\right) = q
\]
\[
= 2pq
\]
How do we estimate genotype frequencies for dominant loci?

First, get genotype frequency for recessive homozygote

- frequency of $A_2A_2 = Z = \frac{N_{22}}{N}$

- $q = \sqrt{q^2} = \sqrt{Z}$
- $p = 1 - q$
- $X = p^2$
- $Y = 2pq$

Assumes Hardy-Weinberg Equilibrium!
Example of calculating allele and genotype frequencies for dominant loci

- *Linanthus parryi* is a desert annual with white and blue flower morphs, controlled by a single locus with two alleles

  - Blue is dominant to white:
    - Blue Flowers: 750  \(B_1B_1\) and \(B_1B_2\)
    - White Flowers: 250  \(B_2B_2\)

- Calculate \(p\), \(q\), \(X\), \(Y\), and \(Z\)
Is this population in Hardy-Weinberg Equilibrium?
Variance of Allele Frequency under Dominance

- Frequency of dominant allele cannot be directly estimated from phenotypes ($A_1A_1$ is identical to $A_1A_2$)

- Frequency of dominant allele ($p$) is estimated from frequency of recessive ($q$)

\[ Z = \frac{N_{22}}{N} \]

\[ q = \sqrt{q^2} = \sqrt{Z} \quad \quad p = 1 - q \]

- Variance of this estimate is therefore

\[ V(\sqrt{Z}) = V(\sqrt{q^2}) \]

- Not the same as $V(q)$!
Derivation of Variance for Dominant Biallelic Locus

By definition:

\[ V(f(x)) = \left( \frac{df(x)}{dx} \right)^2 V(x) \]

\[ V(\sqrt{Z}) = \left( \frac{d\sqrt{Z}}{dZ} \right)^2 \left( \frac{Z(1-Z)}{N} \right) \]

Formula for binomial variance

\[ V(\sqrt{Z}) = \left( \frac{1}{2\sqrt{Z}} \right)^2 \left( \frac{Z(1-Z)}{N} \right) \]

\[ = \frac{Z - Z^2}{4ZN} = \frac{1-Z}{4N} \]

Variance of allele frequency for recessive allele at dominant locus

\[ V(q) = \frac{1-q^2}{4N} \]
Comparison of codominant and dominant variances

\[ V(q) = \frac{q(1-q)}{2N} \]  
Variance of allele frequency for codominant locus

\[ V(q) = \frac{1-q^2}{4N} \]  
Variance of allele frequency for recessive allele at dominant locus

Maximum Variance, Codominance

Maximum Variance, Dominance

Errors in genotype frequency estimates magnified at low allele frequencies

\[ p = 0.5 \]
\[ p = 0.125 \]
Testing for Departures from Hardy-Weinberg Equilibrium
Hypothesis Testing: Frequentist Approach

- Define a null hypothesis, $H_0$:
  - The probability of getting heads on each flip of a coin is $p = 0.5$
- Find the probability distribution for observing data under the null hypothesis (use binomial probability distribution here)
- Calculate the $p$-value, which is the probability of observing a result as extreme or more extreme if the null hypothesis is correct.
- Reject the null hypothesis if the $p$-value is smaller than an arbitrarily chosen level of Type I statistical error (i.e., the probability of rejecting $H_0$, when it is actually correct).
Departures from Hardy-Weinberg

- **Chi-Square test** is simplest (frequentist) way to detect departures from Hardy-Weinberg

- Compare calculated Chi-Square value versus “critical value” to determine if a significant departure is supported by the data
Meaning of P-value

- Probability of a Chi-square value of the calculated magnitude or greater if the null hypothesis is true.
- Critical values are not magical numbers.
- Important to state hypotheses correctly.
- Interpret results within parameters of test.

$p<0.05$: The null hypothesis of no significant departure from Hardy-Weinberg equilibrium is rejected.
Alternatives to Chi-Square Calculation

- If expected numbers are very small (less than 5), Chi-square distribution is not accurate
- Exact tests are required if small numbers of expected genotypes are observed
- Essentially a sample-point method based on permutations
  - Sample space is too large to sample exhaustively
  - Take a random sample of all possible outcomes
  - Determine if observed values are extreme compared to simulated values
- Fisher’s Exact Test in lab next time
**Expected Heterozygosity**

If a population is in Hardy-Weinberg Equilibrium, the probability of sampling a heterozygous individual at a particular locus is the **Expected Heterozygosity**: 

- $2pq$
  
  for 2-allele, 1 locus system

OR

- $1-(p^2 + q^2)$ or $1-\sum$(expected homozygosity)

  more general: what’s left over after calculating expected homozygosity

**Homozygosity is overestimated at small sample sizes. Must apply correction factor:**

Correction for bias in parameter estimates by small sample size

$$
H_E = \frac{2N}{2N-1} \left(1 - \sum_{i=1}^{n} p_i^2\right),
$$

$$
H_E = 1 - \sum_{i=1}^{n} p_i^2,
$$
Maximum Expected Heterozygosity

- Expected heterozygosity is maximized when all allele frequencies are equal
- Approaches 1 when number of alleles = number of chromosomes

\[
H_{E_{\text{max}}} = 1 - \sum_{i=1}^{2N} \left( \frac{1}{2N} \right)^2 = 1 - 2N \left( \frac{1}{2N} \right)^2 = \frac{2N - 1}{2N}
\]

- Applying small sample correction factor:

\[
H_E = \frac{2N}{2N - 1} \left( 1 - \sum_{i=1}^{n} p_i^2 \right) = \frac{2N}{2N - 1} \left( \frac{2N - 1}{2N} \right) = 1
\]

Also see Example 2.11 in Hedrick text
Observed Heterozygosity

- Proportion of individuals in a population that are heterozygous for a particular locus:

\[
H_O = \frac{\sum N_{ij}}{N} = \sum H_{ij}
\]

Where \( N_{ij} \) is the number of diploid individuals with genotype \( A_iA_j \), and \( i \neq j \), and \( H_{ij} \) is frequency of heterozygotes with those alleles.

- Difference between observed and expected heterozygosity will become very important soon.

- This is NOT how we test for departures from Hardy-Weinberg equilibrium!
Alleles per Locus

- $N_a$: Number of alleles per locus
- $N_e$: Effective number of alleles per locus

If all alleles occurred at equal frequencies, this is the number of alleles that would result in the same expected heterozygosity as that observed in the population.

\[
N_e = \frac{1}{\frac{N_a}{\sum_{i=1}^{N_a} p_i^2}},
\]
Expected Heterozygosity or Gene Diversity ($H_E$)

- Primary measure of genetic diversity within populations
- Can interpret as probability that two sampled alleles are different

\[ H_E = \frac{2N}{2N - 1} \left( 1 - \sum_{i=1}^{n} p_i^2 \right), \]
Example: Assay two microsatellite loci for WVU football team (N=50)

Calculate $H_e$, $N_a$ and $N_e$

<table>
<thead>
<tr>
<th>Locus A</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.01</td>
</tr>
<tr>
<td>A2</td>
<td>0.01</td>
</tr>
<tr>
<td>A3</td>
<td>0.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus B</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.3</td>
</tr>
<tr>
<td>B2</td>
<td>0.3</td>
</tr>
<tr>
<td>B3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

$$H_e = \frac{2N}{2N-1} \left(1 - \sum_{i=1}^{n} p_i^2\right),$$

$$N_e = \frac{1}{\sum_{i=1}^{N_a} p_i^2}.$$
Measures of Diversity are a Function of Populations and Locus Characteristics

Assuming you assay the same samples, order the following markers by increasing average expected values of $N_e$ and $H_E$:

- RAPD
- SSR
- Allozyme