ESTIMATION AND HYPOTHESIS TESTING READING: Nielsen & Slatkin, pp. 16–18

• Introduction

– Up to now, have treated genotype, gamete, & allele frequencies as known.

- How do we determine what these frequencies are in "reality"?

- How do we determine the validity (or not) of H-W in a study population?

– Solution 1: Sample genotypes of interest from every individual
- No error, but not generally feasible.

– Solution 2: Sample genotypes from a "representative" subset of individuals from the population.
- Generally feasible, but how much error?
- Consider the following scenario:

  – Suppose 10 copies of a "rare" allele exist in a diploid population of 5,000 individuals

  Allele frequency $= \dfrac{10}{2 \times 5000} = \dfrac{10}{10{,}000} = 0.001$

  – Sample 50 individuals from this population. The chance that we do not sample even one copy of this rare allele is $(1 - 0.001)^{2 \times 50} = (0.999)^{100} \approx 90\%$
  - I.e., 90% chance we will not know that this "rare" allele even exists!

- The field of **statistics** deals with such uncertainty.

– Two main (inter-related) concerns addressed by statistics that are of interest to empiricists:

1) **Estimation**
   - What is the frequency of ______ ?

2) **Hypothesis Testing**
   - If I observe this and the world is like so, are my observations usual or not? I.e., Is the world like I think it is?

• Estimating Allele Frequencies

– Data from yellow fever mosquito (*Aedes aegypti*) collected in Ghana by J. Powell [reported in B. Weir "Genetic Data Analysis"]

– Counts of allozyme genotypes from 40 individuals at the Isocitric dehydrogenase (IDH) locus: $\underbrace{N_{11} = 24}_{\text{\# individs. w/ 2 copies of "common" allele}}$; $N_{12} = 16$; $N_{22} = 0$

– Want to compute the frequency of the "2" allele, p_2, in the Ghanaian population.

- <u>Estimate #1:</u> Use allele frequency in sample to infer allele frequency in population:

$$\hat{p}_2 = \frac{N_{12} + 2N_{22}}{2(N_{11} + N_{12} + N_{22})} = \frac{16 + (2 \cdot 0)}{2 \cdot 40} = 0.2 . \qquad (\wedge = \text{"estimate"})$$

- <u>Estimate #2:</u> Assume population is in Hardy-Weinberg equilibrium. Then the frequency of the "22" homozygote is $(p_2)^2$. Using the frequency of 22-homozygotes in sample to infer the frequency in the population, estimate:

$$\hat{p}_2 = \sqrt{\text{observed freq. of "22"-genotype}} = \sqrt{\frac{0}{40}} = 0 \ .$$

- <u>Estimate #3:</u> Use same reasoning to estimate p_1 and use the relation $p_2 = 1 - p_1$:

$$\hat{p}_2 = 1 - \hat{p}_1 = 1 - \sqrt{\text{observed freq. of "11"-genotype}} = 1 - \sqrt{\frac{24}{40}} = 0.23 \quad .$$

– Three estimates (0.2, 0, 0.23) for p_2. Which to use?

• Maximum Likelihood Estimates

– Key question: ***If*** the true value of $p_2 = x$, then what is the probability of observing our data ($N_{11} = 24$, $N_{12} = 16$, $N_{22} = 0$)?

– **Likelihood of the data given** x = Prob[Data | hypothesis $p_2 = x$]

– ”**Maximum Likelihood Estimate (MLE) of** p_2” = the value of p_2 that maximizes the likelihood

- In other words, the maximum likelihood estimate is the hypothesis (value of p_2) which maximizes the probability of observing the data.

– MLE for mosquito data (assume Hardy-Weinberg equilibrium, use multinomial distribution):

- Prob(Data | $p_2 = 0.1) \approx 0.003$
- Prob(Data | $p_2 = 0.2) \approx 0.11$ <— 0.2 closest of these to the maximum likelihood estimate
- Prob(Data | $p_2 = 0.3) \approx 0.014$

– Can use calculus (or computer) to get answer directly: $\hat{p}_2 = 0.2$

– MLE is conceptually simple, but very powerful (and flexible) statistical technique.

– Maximum likelihood is also covered in Appendix C of Nielsen & Slatkin, but is presented in a context that we will get to later in the course.

- **Hypothesis Testing**

– We may suspect that the H-W assumptions do not approximate the situation in the *Aedes* population very well.

– Question: How do we (scientifically) go about testing our suspicions that H-W conditions do not hold?

– Answer: Statistically, the best way : assume H-W *does* hold and then try to show that the data do not support this assumption.

- The H-W assumption in this case is called the "**null hypothesis.**"

– **Procedure**:

(1) Determine what data are "expected" under the null hypothesis.

- If p_2 is the true frequency of the "2" allele, then under H-W assumptions "expect" to observe the following numbers of each genotype:

$$\tilde{N}_{11} = 40 \cdot (1 - p_2)^2; \qquad \tilde{N}_{12} = 40 \cdot 2(1 - p_2)p_2; \qquad \tilde{N}_{11} = 40 \cdot p_2^2.$$

- If ($\tilde{N}_{11}, \tilde{N}_{12}, \tilde{N}_{22}$) are "significantly" different from our observations (24, 16, 0), then we can be more confident that our suspicions are true!

- Measure of "different": the ***Chi-square Statistic***, X^2

(2) Compute $X^2 = \dfrac{\left(24 - \tilde{N}_{11}\right)^2}{\tilde{N}_{11}} + \dfrac{\left(16 - \tilde{N}_{12}\right)^2}{\tilde{N}_{12}} + \dfrac{\left(0 - \tilde{N}_{22}\right)^2}{\tilde{N}_{22}}$

- In general, $X^2 = \sum \frac{(\text{Observed number} - \text{Expected number})^2}{\text{Expected number}}$

- If X^2 is "large" then we conclude that H-W assumptions do not hold

– Problem with procedure: need to know p_2 in order to find ($\tilde{N}_{11}, \tilde{N}_{12}, \tilde{N}_{22}$).

– Solution: Use our best estimate of p_2: $\hat{p}_2 = 0.2$:

$$\tilde{N}_{11} = 40 \cdot (1 - 0.2)^2 = 25.6; \quad \tilde{N}_{12} = 40 \cdot 2 \cdot 0.8 \cdot 0.2 = 12.8; \quad \tilde{N}_{11} = 40 \cdot 0.2^2 = 1.6$$

so

$$X^2 = \frac{(24 - 25.6)^2}{25.6} + \frac{(16 - 12.8)^2}{12.8} + \frac{(0 - 1.6)^2}{1.6} = 2.5.$$

- Note: using $\hat{p}_2$ reduces our confidence in X^2 as a measure of discrepancy from the null hypothesis since a large value of X^2 may reflect a bad estimate for p_2 rather than departure from the null hypothesis, H-W.

– Probability that X^2 is "significantly" large or not depends on the chi-square distribution and the "degrees of freedom".

- Degrees of freedom = (number of categories – 1) – (number of estimated parameters)

  – reducing the degrees of freedom for estimated parameters corrects for possibility that X^2 is large due to bad estimates.

– With 3 genotypes (categories) and 1 estimated parameter ($\hat{p}_2$), the value $X^2 = 2.5$ (with $2 - 1 = 1$ degree of freedom) is **not** unusually large under the null hypothesis ($X^2 > 3.9$ are "unusually" large in this case)

– Conclude: Our suspicions that H-W is false are not supported by this data.

– *Careful*: Cannot conclude from this that H-W assumptions do hold (weak inference).

- Can use likelihood to compare hypothesis. The **likelihood ratio test** compares the likelihood of the data under the null hypothesis to it likelihood given the MLE.

  – The likelihood ratio statistic *G* is defined

$$G = 2\ln\left[\frac{\text{Likelihood of the data given the MLE}}{\text{Likelihood of the data given the null hypothesis}}\right]$$

– Turns out that if the null hypothesis is true, G is approximately X^2 distributed.

- the degrees of freedom equal the difference in number of parameters that require estimation between the two hypotheses.

– see **HANDOUT I.4.** Comparing Hypothesis: genes and longevity.