**MATH 448 — Lecture 16**

§6.2 (continued) and §6.4: The $\chi^2$ test with several categories; Hardy–Weinberg equilibrium

**The $\chi^2$ test.** A *categorical model* consists of $k$ categories and probabilities $\pi_1, \dots, \pi_k \geq 0$ with $\pi_1 + \cdots + \pi_k = 1$. The data are $n$ independent observations, each falling in category $i$ with probability $\pi_i$. The *observed count* $O_i$ is the number of observations in category $i$, and the *expected count* is $E_i := n\pi_i$. The $\chi^2$ *statistic* is

$$\chi^2 := \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}. \tag{16.1}$$

For an integer $d \geq 1$, the $\chi^2$ *distribution with $d$ degrees of freedom* is the distribution of $Z_1^2 + \cdots + Z_d^2$, where $Z_1, \dots, Z_d$ are independent standard normal random variables. For $0 < \alpha < 1$, the *critical value* $c_{d,\alpha}$ is the number with $P(Z_1^2 + \cdots + Z_d^2 > c_{d,\alpha}) = \alpha$.

| $d$ | 1 | 2 | 3 |
|---|---|---|---|
| $c_{d,.05}$ | 3.841 | 5.991 | 7.815 |
| $c_{d,.01}$ | 6.635 | 9.210 | 11.345 |

*Theorem* (Pearson, Fisher). Suppose the model holds and the probabilities $\pi_i$ depend on $m$ parameters whose values are estimated from the data by maximum likelihood ($m = 0$ if the $\pi_i$ are specified in advance). As $n \to \infty$, the distribution of $\chi^2$ converges to the $\chi^2$ distribution with $d = k - 1 - m$ degrees of freedom.

The $\chi^2$ *test at level* $\alpha$ rejects the model if $\chi^2 > c_{d,\alpha}$. By the theorem, a correct model is rejected with probability approximately $\alpha$ when $n$ is large.

*Example.* A $WwGg \times WwGg$ cross ($W$ round, $G$ yellow) of pea plants gives 556 offspring. Independent assortment gives the phenotypes round yellow, round green, wrinkled yellow, wrinkled green the probabilities $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$, so $k = 4$, $m = 0$, $d = 3$.

| | round yellow | round green | wrinkled yellow | wrinkled green |
|---|---|---|---|---|
| $O$ | 315 | 108 | 101 | 32 |
| $E$ | 312.75 | 104.25 | 104.25 | 34.75 |

By (16.1), $\chi^2 = \dfrac{2.25^2}{312.75} + \dfrac{3.75^2}{104.25} + \dfrac{3.25^2}{104.25} + \dfrac{2.75^2}{34.75} \approx .47 < 7.815 = c_{3,.05}$, so the model is not rejected at level .05.

In this example, the count of round green offspring is binomial with $n = 556$ and probability $\frac{3}{16}$, so its standard deviation is $\sqrt{556 \cdot \frac{3}{16} \cdot \frac{13}{16}} \approx 9.2$, and the observed deviation 3.75 is about .4 standard deviations. The statistic $\chi^2 \approx .47$ sums the four such comparisons. Here $d = 3$ because the four counts add to 556, so three of them determine the fourth, and a correct model gives $\chi^2$ a mean of $d = 3$. The value $c_{3,.05} = 7.815$ means that if the cross were repeated many times with 556 offspring each, a correct model would give $\chi^2 > 7.815$ in about 5% of the repetitions.

**Allele and genotype frequencies.** A population consists of $N$ individuals, each with two copies of a gene, each copy being allele $A$ or $a$. Let $n_{AA}, n_{Aa}, n_{aa}$ be the numbers of individuals with each genotype, so $n_{AA}+n_{Aa}+n_{aa} = N$. The *genotype frequencies* are $f_{AA} := n_{AA}/N$, $f_{Aa} := n_{Aa}/N$, $f_{aa} := n_{aa}/N$. The *allele frequencies* are $p :=$ (number of copies of $A$)$/2N$ and $q :=$ (number of copies of $a$)$/2N = 1 - p$. Since there are $2n_{AA} + n_{Aa}$ copies of $A$,

$$p = \frac{2n_{AA} + n_{Aa}}{2N} = f_{AA} + \tfrac{1}{2} f_{Aa}, \qquad q = f_{aa} + \tfrac{1}{2} f_{Aa}. \tag{16.2}$$

**Hardy–Weinberg law.** The *gene pool* is the set of all $2N$ gene copies in the population. A new generation is formed by *random mating* if each offspring receives two gene copies drawn independently from the gene pool, so that each copy is $A$ with probability $p$ and $a$ with probability $q$. A population is in *Hardy–Weinberg equilibrium* if $(f_{AA}, f_{Aa}, f_{aa}) = (p^2, 2pq, q^2)$.

**Hardy–Weinberg law**

Consider a population that is infinite (so that frequencies equal probabilities), has non-overlapping generations, mates randomly, and has no selection, mutation, or migration. Let $p$ be the allele frequency of $A$. Then the next generation has

$$f'_{AA} = p^2, \qquad f'_{Aa} = 2pq, \qquad f'_{aa} = q^2, \qquad p' = p, \tag{16.3}$$

whatever the genotype frequencies of the first generation. Hence the next generation is in Hardy–Weinberg equilibrium, and so is every later generation.

*Proof.* By independence, $f'_{AA} = p \cdot p$, $f'_{Aa} = p \cdot q + q \cdot p$, $f'_{aa} = q \cdot q$. By (16.2), $p' = p^2 + \frac{1}{2}(2pq) = p(p+q) = p$. Applying the same computation to each later generation gives the rest.

*Example.* A population of $N = 1000$ individuals has $n_{AA} = 300$, $n_{Aa} = 500$, $n_{aa} = 200$. The gene pool has 2000 copies, of which $2 \cdot 300 + 500 = 1100$ are $A$, so $p = .55$ and $q = .45$. Under random mating, an offspring receives each copy as $A$ with probability .55, and the next generation has $f'_{AA} = .55^2 = .3025$, $f'_{Aa} = 2(.55)(.45) = .495$, $f'_{aa} = .45^2 = .2025$, with $p' = .3025 + \frac{1}{2}(.495) = .55$.

**Testing for Hardy–Weinberg equilibrium.** The model has $k = 3$ categories $AA$, $Aa$, $aa$ with probabilities $p^2$, $2pq$, $q^2$, depending on one parameter $p$, so $m = 1$ and $d = 3 - 1 - 1 = 1$. The maximum likelihood estimate of $p$ is $\hat{p} = (2n_{AA} + n_{Aa})/2N$, and the expected counts are $N\hat{p}^2$, $2N\hat{p}\hat{q}$, $N\hat{q}^2$ with $\hat{q} = 1 - \hat{p}$.

*Example.* For the population above, $\hat{p} = .55$ and the expected counts are 302.5, 495, 202.5, compared with the observed 300, 500, 200. Then $\chi^2 = \dfrac{2.5^2}{302.5} + \dfrac{5^2}{495} + \dfrac{2.5^2}{202.5} \approx .10 < 3.841 = c_{1,.05}$, so the model is not rejected at level .05. Here $d = 1$ because the counts satisfy $n_{AA} + n_{Aa} + n_{aa} = 1000$ and $2n_{AA} + n_{Aa} = 1100$, so $n_{AA}$ determines the other two.

**Problem 16.1.** A $WwGg \times WwGg$ cross gives 400 offspring with counts 250, 60, 50, 40 for round yellow, round green, wrinkled yellow, wrinkled green. Test the model of independent assortment (probabilities $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$) at level .05.

**Problem 16.2.** A population has $f_{AA} = .5$, $f_{Aa} = .2$, $f_{aa} = .3$. Find $p$ and $q$, the genotype frequencies after one generation of random mating, and after two.

**Problem 16.3.** A recessive disease affects 1 in 10,000 individuals in a population in Hardy–Weinberg equilibrium. Find $q$ and the frequency of carriers.

**Problem 16.4.** A population of 200 has 80 $AA$, 80 $Aa$, and 40 $aa$ individuals. Test the Hardy–Weinberg model at level .05.

**Problem 16.5.** Show that $2pq = \frac{1}{2} - 2\big(p - \frac{1}{2}\big)^2$. What is the largest possible frequency of heterozygotes in a population in Hardy–Weinberg equilibrium?

*Answers.* 16.1: expected $225, 75, 75, 25$; $\chi^2 \approx 23.1 > 7.815$ with 3 degrees of freedom, so rejected. 16.2: $p = .6$, $q = .4$; $(.36, .48, .16)$ after one generation and after two. 16.3: $q = .01$, carriers $2pq = .0198$, about 1 in 50. 16.4: $p = .6$; expected $72, 96, 32$; $\chi^2 \approx 5.56 > 3.841$ with 1 degree of freedom, so rejected. 16.5: $\frac{1}{2}$, at $p = \frac{1}{2}$.