MS Module 22: χ^2 tests – practice problems

(The attached PDF file has better formatting.)

Exercise 22.1: χ^2 When Parameters Are Estimated

The groups of phenotypes, R, S, and T, are in equilibrium if for some θ:

! P(R) = $p_1 = \theta^2$
! P(S) = $p_2 = 2\theta(1–\theta)$
! P(T) = $p_3 = (1–\theta)^2$

A sample from a population has the following number of observations in each group:

! Group R: $n_1 = 145$
! Group S: $n_2 = 235$
! Group T: $n_3 = 120$

The null hypothesis H_0 is that the population is in equilibrium for some parameter θ.

A. What is the maximum likelihood estimate for θ?
B. What are the expected cell counts?
C. What is the χ^2 statistic to test the null hypothesis that the population is in equilibrium?
D. What is the p value to test the null hypothesis that the population is in equilibrium?

Part A: The likelihood is of the observed values given θ is

$$[\pi_1(\theta)]^{n1} \times [\pi_2(\theta)]^{n2} \times [\pi_3(\theta)]^{n3} = [\theta^2]^{n1} \times [2\theta(1–\theta)]^{n2} \times [(1–\theta)^2]^{n3} = 2^{n2} \times \theta^{2n1 + n2} \times 1–\theta)^{n2+2n3}$$

Maximizing the loglikelihood (the natural logarithm of the likelihood) with respect to θ yields

$$\hat{\theta} = (2n_1 + n_2) / [(2n_1 + n_2) + (n_2 + 2n_3)] = (2n_1 + n_2) / 2n, \text{ where } n = n_1 + n_2 + n_3 =$$
$$(2 \times 145 + 235) / (2 \times 500) = 0.525$$

where $n_1 = 145$, $n_2 = 235$, and $n = 500$.

Part B: The expected cell counts are

! Group R: $500 \times 0.525^2 = 137.8125$
! Group S: $500 \times 2 \times 0.525 \times (1 – 0.525) = 249.3750$
! Group T: $500 \times (1 – 0.525)^2 = 112.8125$

Part C: The χ^2 statistic contributions to test the null hypothesis that the population is in equilibrium is

! Group R: $(145 – 137.8125)^2 / 137.8125 = 0.374858$
! Group S: $(235 – 249.375)^2 / 249.375 = 0.828634$
! Group T: $(120 – 112.8125)^2 / 112.8125 = 0.457929$

The χ^2 statistic is 0.374858 + 0.828634 + 0.457929 = 1.661421

Part D: The p value = 1 – the cumulative distribution function of the χ^2 distribution with (3 – 1 – 1) degrees of freedom = 0.197411 (table lookup or spreadsheet function).

Jacob: Why are the degrees of freedom = 3 – 1 – 1 = 1?

Rachel: The scenario has two constraints:

- The sum of the observations in the groups = the total number of observations.
- The observations by group satisfy the proportions:
  - P(R) = $p_1 = \theta^2$
  - P(S) = $p_2 = 2\theta(1\text{–}\theta)$
  - P(T) = $p_3 = (1\text{–}\theta)^2$

Exercise 22.2: Testing for a normal distribution

We draw a sample of 100 points to test whether a population is normally distributed.

Before drawing the sample, we assume the population's mean μ is 8 and its standard deviation σ is 2.

We group the sample values into five groups $(-\infty, k_1)$, (k_1, k_2), (k_2, k_3), (k_3, k_4), (k_4, ∞), which have the same expected number of observations if the population ~ $N(8, 2^2)$.

Summary statistics for the 100 sample values are $\Sigma x_i = 840$ and $\Sigma x_i^2 = 7{,}535.16$.

The number of sample values in the five groups are 16, 18, 19, 21, and 26.

A. What are the values of k_1, k_2, k_3, and k_4?
B. What is the mean of the sample?
C. What is the standard deviation of the sample?
D. What are the percentile bounds for the five groups using the sample mean and standard deviation?
E. What are the expected number of observations in the five groups using the sample mean and the sample standard deviation for the population?
F. What is the χ^2 value to test the null hypothesis?
G. How many degrees of freedom does the χ^2 value have?
H. What is the *p* value to test the null hypothesis?

Part A: If the population were ~ N(0.1), the values of k_1, k_2, k_3, and k_4 would be

! -0.841621
! -0.253347
! 0.253347
! 0.841621

Since the population is assumed to be ~ N(8,2), the values of k_1, k_2, k_3, and k_4 are

! -0.841621 × 2 + 8 = 6.316758
! -0.253347 × 2 + 8 = 7.493306
! 0.253347 × 2 + 8 = 8.506694
! 0.841621 × 2 + 8 = 9.683242

Part B: The mean of the sample is $\Sigma x_i / n = 840 / 100 = 8.4$

Part C: The variance of the sample is $(\Sigma x_i^2 - (\Sigma x_i)^2/n)/(n-1) =$

$(7{,}535.16 - 840^2/100)/(100 - 1) = 4.84$

The standard deviation of the sample is $4.84^{0.5} = 2.20$

Part D: If the population is ~ $N(8.4, 2.2^2)$, the percentiles for k_1, k_2, k_3, and k_4 are

! (6.316758 – 8.4) / 2.2 = -0.946928
! (7.493306 – 8.4) / 2.2 = -0.412134
! (8.506694 – 8.4) / 2.2 = 0.048497
! (9.683242 – 8.4) / 2.2 = 0.583292

The bounds for the five groups are

! $(-\infty, -0.946928)$

- (-0.946928, -0.412134)
- (-0.412134, 0.048497)
- (0.048497, 0.583292)
- (0.583292, ∞)

Part E: The expected number of observations in the five groups from the sample of 100 values =

- 100 × Φ (-0.946928) = 17.183763
- 100 × (Φ (-0.412134) – Φ (-0.946928)) = (34.012070 – 17.183763) = 16.828307
- 100 × (Φ (0.048497) – Φ (-0.412134)) = (51.934007 – 34.012070) = 17.921936
- 100 × (Φ (0.583292) – Φ (0.048497)) = (72.015164 – 51.934007) = 20.081157
- 100 × (1 – Φ (0.583292)) = (100 – 72.015164) = 27.984836

Part F: The contribution of each group to the χ^2 statistic is

- $(16 – 17.183763)^2$ / 17.183763 = 0.081548
- $(18 – 16.828307)^2$ / 16.828307 = 0.081581
- $(19 – 17.921936)^2$ / 17.921936 = 0.064849
- $(21 – 20.081157)^2$ / 20.081157 = 0.042043
- $(26 – 27.984836)^2$ / 27.984836 = 0.140775

The χ^2 statistic used to test the null hypothesis that the population is normally distributed =

$$0.081548 + 0.081581 + 0.064849 + 0.042043 + 0.140775 = 0.410796$$

Part G: The χ^2 value has 5 – 1 = 4 degrees of freedom: 5 groups – 1 constraint (the sum of the observations in the five groups = the total observations).

Part H: The *p* value is 1 – the cumulative distribution function of the χ-squared distribution with 4 degrees of freedom at 0.410796 = 0.981584 (table lookup or spreadsheet function).

Question: Why is the *p* value so high?

Answer: The actual number of observations by group are close to the expected number of observations. The slight differences presumably stem from rounding and random fluctuations. The total χ^2 is much less than the degrees of freedom, so we do not reject the null hypothesis that the population is normally distributed.

Exercise 22.3: Phenotypes

The expected proportions of subjects with four phenotypes is 9/16, 3/16, 3/16, and 1/16.

The observed values in an experiment are 895, 280, 305, and 120.

A. What are the expected values in each cell?
B. What is the χ^2 value to test the null hypothesis?
C. What are the degrees of freedom?
D. What is the *p* value to test the null hypothesis?

Part A: The total subjects = 895 + 280 + 305 + 120 = 1600

The expected counts in the four groups are

1. 9/16 × 1600 = 900
2. 3/16 × 1600 = 300
3. 3/16 × 1600 = 300
4. 1/16 × 1600 = 100

Part B: The χ^2 value is the sum of the contributions from the four groups, which are

1. $(895 - 900)^2 / 900 = 0.0278$
2. $(280 - 300)^2 / 300 = 1.3333$
3. $(305 - 300)^2 / 300 = 0.0833$
4. $(120 - 100)^2 / 100 = 4.0000$

The χ^2 value is 0.0278 + 1.3333 + 0.0833 + 4.000 = 5.4444

Part C: The χ^2 test has four cells and one constraint (the total actual values = the total expected values), so the degrees of freedom = 4 – 1 = 3.

Part D: The *p* value = 1 – χ^2 cdf(5.4444, 3) = 0.142 (table lookup or spreadsheet function).