Stat 24400 Homework 8 Solution

Mar 6, 2016

Total points: 100

1. [20pts] **James Mills Sentences**

   First we estimate θ via MLE.

$$\log(L) = \log\binom{5}{n_i} + \log(\theta)\sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(5 - x_i)$$
$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\log(L) = \frac{n\bar{x}}{\theta} - \frac{5n - n\bar{x}}{1-\theta}$$
$$\Rightarrow \hat{\theta} = \frac{\bar{x}}{5}$$

   We have a total number of $80 \times 0 + 11 \times 1 + 2 \times 51 + 3 \times 42 + 4 \times 20 + 5 \times 4 = 339$ sentences with Class I words. So $\hat{\theta} = \frac{339}{215} \times \frac{1}{5} = 0.315$.

# Sentences with Class I word	0	1	2	3	≥ 4
Observed	87	11	51	42	24
Expected	32.35	74.5	68.61	31.59	7.94

   Notice that we need to group case 4 with case 5 for correction because the expected cell count for case 5 is just 0.67 (< 3.5).
   Under the null hypothesis, the test statistics is

$$X = \sum_{i=1}^{6} \frac{(E_i - O_i)^2}{E_i} \approx 143.655$$

   with $5 - 1 - 1 = 3$ degrees of freedom. The p-value is ≈ 0, so we have strong evidence to reject the null hypothesis: the count of sentences with class I words fit a binomial distribution.

   Grading Scheme: 5 pts for estimating the MLE, 5 pts for calculating the expected count, 5 pts for test statistics and degree of freedom, 5 pts for p-value and conclusion. Deduct 1 pt for not grouping case 4 and case 5.

2. [20pts] **Blood type**

   First of all we notice that $r + p + q = 0.999$, because of a rounding error. We will work under the assumption that this error is negligible. From the numerical values of r, p and q we have that

$$P(\text{“O”}) = r^2 = 0.3364$$
$$P(\text{“A”}) = p^2 + 2pr = 0.3459$$
$$P(\text{“B”}) = q^2 + 2qr = 0.2306$$
$$P(\text{“B”}) = 2pq = 0.0851$$

	O	A	B	AB
Observed	121	120	79	33
Expected	118.7492	122.0942	81.4050	30.0459

Therefore we have that, under the null hypothesis, the test statistics is

$$X = \sum_{i=1}^{4} \frac{(E_i - O_i)^2}{E_i} \approx 0.4401$$

with $4 - 1 - 2 = 1$ degree of freedom (we are removing 2 degrees of freedom for 3 parameters since they sum up to 1). Therefore, the value is ≈ 0.49, and we cannot reject the null hypothesis: the community fits the theory.

Grading Scheme: 5 pts for calculating the expected count, 5 pts for test statistics and degree of freedom, 10 pts for p-value and conclusion.

3. **[20pts] Fingerprint patterns**

Under the assumption of independence, we can build the table of the expected values

	Arches	Loops	Whorls	Totals
Arches	1.81	12.30	4.89	19
Loops	5.81	39.50	15.69	61
Whorls	2.38	16.19	6.43	25
Whorls	10	67.99	27.01	105

So the test statistics is

$$X = \sum_{i=1}^{3} \sum_{i=1}^{3} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 11.1684$$

Under the null hypothesis this is χ^2 with $(3-1) \times (3-1) = 4$ degrees of freedom. Therefore, the pvalue for the test is ≈ 0.0247, so there is some but not conclusive evidence that the traits are independent.

Grading Scheme: 5 pts for calculating the expected count, 5 pts for test statistics and degree of freedom, 10 pts for p-value and conclusion.

4. **[20pts] Death by Horsekick Data**

First we have to estimate θ via MLE. We have a total number of deaths of $1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2 = 196$. Since the total number of events is 280, we have $\hat{\theta} = 196/280 = 0.7$.

Number of Deaths	Frequency count	Expected
0	144	139.04
1	91	97.33
2	32	34.06
3	11	7.95
≥ 4	2	1.61

Under the null hypothesis, the test statistics is

$$X = \sum_{i=1}^{4} \frac{(E_i - O_i)^2}{E_i} \approx 1.97$$

with $5 - 1 - 1 = 3$ degrees of freedom.

Therefore, the pvalue is ≈ 0.42, and we cannot reject the null hypothesis: the death by Horsekick Data fits a Poisson distribution.

Grading Scheme: 5pts for estimating the MLE, 5 pts for calculating the expected count, 5 pts for test statistics and degree of freedom, 5 pts for p-value and conclusion.

5. **[20pts] American roulette**

   (a) The expected outcome that would result if the wheel was fair is $3880/38 \approx 102.1$ for every number. Therefore the Chi square statistics is

   $$X = \sum_{i=1}^{38} \frac{(O_i - 102.1)^2}{102.1} \approx 29.6907$$

   where $O_i = 100$ for $i = 1, ..., 36$ and $O_i = 140$ for $i = 37, 38$. Under the null hypothesis $X \sim \chi^2(37)$, and for $\alpha = 0.05$ we reject the null hypothesis if $X \geq 52.19$, which is clearly not the case. Therefore, we cannot assume that the wheel is unfair.

   (b) We know that, under the Central Limit Theorem approximation

   $$Z \sim \mathcal{N}(204.2105, 193.4626)$$

   Now, at $\alpha = 0.05$, the rejection region is approximately $Z^* \geq 1.64$ where Z^* is the normalized Z. In our case we have $z^* = \frac{280-204.2105}{\sqrt{193.4626}} = 5.4489$. So we reject the null hypothesis and conclude that there are more "0" and "00" than there should be if the wheel was fair.

   (c) In (a) we could not reject the null, in (b) we did. As the hint suggests, this is not a contradiction since the second test is much more focused and ad hoc for the data. The bottom line is: don't test for something after you have seen the data.

Grading Scheme: 8 pts for part (a) (4 pts for the form of Chi-squared test statistics, 4 pts for the conclusion); 8 pts for part (b) (3 pts for application of CLT, 3 pts for rejection region, 2 pts for the conclusion); 4 pts for part (c).