Chapter 7 - Tests based on Multinomial Models

Peter He

December 6 2019

Say we had a die and we wondered if it was fair or not. How do we test this hypothesis? First we need a model under H_0.

1 Multinomial Model

These models can be used to describe situations where n observations fall into k categories. For example, $n = 750$ pool balls falling into $k = 6$ pockets.
Say we had data arise from a Multinomial distribution with joint probability distribution

$$f(y_1, \dots, y_k; \theta_1, \dots, \theta_k) = \frac{n!}{y_1! \dots y_k!} \theta_1^{y_1} \dots \theta_k^{y_k}$$

where $y_i = 0, 1, \dots$ and $\sum_{i=1}^{k} y_i = n$. The probabilities satisfy $0 \le \theta_i \le 1$ and $\sum_{i=1}^{k} \theta_i = 1$. In our die example, $k = 6$. We want to test the hypothesis that all outcomes are equally likely, that is $\theta_i = \frac{1}{k}$.
The Multinomial likelihood function is

$$L(\theta_1, \dots, \theta_k) = \prod_{i=1}^{k} \theta_i^{y_i}$$

ignoring the coeffiecient. The maximum likelihood estimator is $\tilde{\theta}_i = \frac{Y_i}{n}$, and the maximum likelihood estimates are $\hat{\theta} = \frac{y_i}{n}$. To test $H_0 : \theta = \theta_0 = \left(\frac{1}{k}, \dots, \frac{1}{k}\right)$, we can used

$$\Lambda(\theta_0) = -2 \log \left[\frac{L(\theta_0)}{L(\tilde{\theta})} \right]$$

By letting $\tilde{\theta} = \left(\frac{Y_1}{n}, \dots, \frac{Y_k}{n}\right)$, we get

$$\Lambda(\theta_0) = 2 \sum_{i=1}^{k} Y_i \log \left(\frac{Y_i}{E_i} \right)$$

where E_i = the number of items in category i if H_0 is true. In more plain terms, we get

$$2\sum_{i=1}^{k}(\text{observed frequency})\cdot\log\left(\frac{\text{observed frequency}}{\text{expected frequency}}\right)$$

If n is large, $\Lambda(\theta)$ follows a χ^2 distribution with $k-1-p$ degrees of freedom, where p is the number of parameters being estimated. In our die example, $p = 0$. Let $\lambda(\theta_0) = 2\sum_{i=1}^{k} y_i \log\left(\frac{y_i}{e_i}\right)$ be the observed value of the LRT statistic, where $e_i = \frac{n}{k}$. The (approximate) p-value is

$$p = P(\Lambda(\theta_0) \geq \lambda(\theta_0); H_0 \text{ is true}) \approx P(W \geq \lambda(\theta_0)), W \sim \chi^2(k-1)$$

This approximate works well under the additional condition that the expected frequencies under H_0 at all at least 5. If they are all not at least 5, we will make adjustments (see example).

2 Pearson's Goodness of Fit Test Statistic

Aside from using the likelihood ratio test statistic, we can use

$$D = \sum_{i=1}^{k}\frac{(Y_i - E_i)^2}{E_i}$$

Both approaches are approximate, and will have different p-values.

3 Example Problem

The number of accidents in a month is observed over a period of 120 months. The data is as follows:

Number of accidents	Observed Value
0	41
1	40
2	22
3	10
4	6
5	0
6	1
7 or more	0

A Poisson model has been proposed. We must find the expected values. To do so, we must estimate the parameter θ with $\bar{y} = 1.2$. To calculate the expected values, we do $e_i = (120)P(Y = i) = (120)\frac{1.2^j e^{-1.2}}{i!}$ for $i = 0, 1, \ldots$. Now we get

Number of accidents	Observed Value	Probability	Expected Frequency
0	41	0.3	36.1
1	40	0.36	43.4
2	22	0.22	26.0
3	10	0.09	10.4
4	6	0.03	3.1
5	0	0.006	0.75
6	1	0.00125	0.15
7 or more	0	0.00025	0.03

The expected values for some of the rows are below 5, so we adjust:

Number of accidents	Observed Value	Probability	Expected Frequency
0	41	0.3	36.1
1	40	0.36	43.4
2	22	0.22	26.0
3 or more	17	0.12051	14.46
Total	120	1	120

We will use the LRT with $4-1-1=2$ degrees of freedom. We will test H_0 : Data follows a Poisson model. Our observed value will be $\lambda(\theta_0) = 2\sum\limits_{i=1}^{k} y_i \log\left(\frac{y_i}{e_i}\right)$. By using the data from the previous table, we get $\lambda(\theta_0) = 1.993$. So, we have

$$p = P(W \geq 1.993) = 0.369$$

where $W \sim \chi^2(2)$. Since our p-value is greater than 0.05, it appears as though a Poisson model seems appropriate.

4 Chi-Squared Tests for Independence of Two Variables

We can conduct chi-squared tests to check if two categorical variables are independent. Let's define a model:
Let

- Y_{11} be the number of $A \cap B$ outcomes
- Y_{12} be the number of $A \cap \bar{B}$ outcomes
- Y_{21} be the number of $\bar{A} \cap B$ outcomes
- Y_{22} be the number of $\bar{A} \cap \bar{B}$ outcomes

Then $(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. The null hypothesis is $H_0 : A$ and B are independent, so $H_0 : P(A \cap B) = P(A)P(B)$. Let $P(A) = \alpha$ and $P(B) = \beta$. This

means we can write $H_0 : \theta_{11} = \alpha\beta$, and similarly for the other parameters.
The Multinomial likelihood function, ignoring coefficients, is

$$L(\theta_{11}, \theta_{12}, \theta_{21}, \theta 22) = \theta_{11}^{y_{11}}\theta_{12}^{y_{12}}\theta_{21}^{y_{21}}\theta_{22}^{y_{22}}$$

. The maximum likelihood estimators are $\hat{\theta}_{ij} = \frac{Y_{ij}}{n}$ and the maximum likelihood estimates are $\hat{\theta}_{ij} = \frac{y_{ij}}{n}$.
We can find $L(\tilde{\alpha}, \tilde{\beta})$ under H_0, and after much work, it can be shown that the LRT statistic we desire is

$$2[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)]$$

Note that the expected frequencies, e_{ij} are given by $e_{ij} = \frac{r_i c_j}{n}$. For the test, we have $(a-1)(b-1)$ degrees of freedom. So, our p-value is $p \approx P(W \geq \lambda)$ where $W \sim \chi^2(1)$

4.1 Example - Gender and Hiring

A large corporation has 20 open positions for entry-level accountants. Applications are initially screened to identify those who satisfy the job requirements. There were 200 applicants who qualified, of which 110 were male, 90 were female. Of those hired, 16 were male and 4 were female.

Notice that we have 1 degree of freedom.
Observed Frequencies:

	Hired	Not Hired	Total
Male	16	94	110
Female	4	86	90
Total	20	180	200

Expected Frequencies:

	Hired	Not Hired	Total
Male	$\frac{(110)(20)}{200}$	99	110
Female	9	81	90
Total	20	180	200

Our observed value of $\lambda(\hat{\alpha}, \hat{\beta})$ is

$$\begin{aligned}
&2\left[y_{11}\log\left(\frac{y_{11}}{e_{11}}\right) + y_{12}\log\left(\frac{y_{12}}{e_{12}}\right) + y_{21}\log\left(\frac{y_{21}}{e_{21}}\right) + y_{22}\log\left(\frac{y_{22}}{e_{22}}\right)\right] \\
&= 2\left[16\log\left(\frac{16}{11}\right) + 94\log\left(\frac{94}{99}\right) + 4\log\left(\frac{4}{9}\right) + 86\log\left(\frac{86}{81}\right)\right] \\
&= 6.06
\end{aligned}$$

Thus, our p-value is

$$\begin{aligned}
p &= P(W \geq 6.06) \qquad W \sim \chi^2(1) \\
&= P(Z \geq \sqrt{6.06}) \\
&= 0.0069
\end{aligned}$$

Since our p-value is 0.0069, it appears as though gender and being hired are dependent (in the context of this example).

4.2 Example - Extension to Multiple Classifications

Human blood is classified according to several systems. Two systems are the OAB system and the Rh system. In the former a person is one of four types O, A, B, AB and inthe latter system a person is Rh+ or Rh. To determine whether these two classification systems are genetically independent, a random sample of 300 persons were chosen. Theirblood was classified according to the two systems and the observed frequencies are given in the table below.

Observed Frequencies:

	O	A	B	AB	Total
Rh+	82	89	54	19	244
Rh-	13	27	7	9	56
Total	95	116	61	28	300

Calculating the expected values, we get:

	O	A	B	AB	Total
Rh+	$\frac{(244)(95)}{300} = 77.27$	94.35	49.61	22.77	244
Rh-	17.73	21.65	11.39	5.23	
Total	95	116	61	28	300

The degrees of freedom are $(a-1)(b-1) = 3 \cdot 1 = 3$. Our observed value is $\lambda = 8.447$, and so our p-value is $P(W \geq 8.447) = 0.0376$ where $W \sim \chi^2(3)$. Since our p-value is lower than 0.05, it appears as though the two classification systems are independent of each other.