Stat 231 Chapter 7: Multinomial Models and Goodness of Fit Tests

Statistics

David Duan, 2019 Fall

Contents

1 Likelihood Ratio Test for the Multinomial Model

1.1 Multinomial Distribution

The *multinomial distribution* is a generalization of the binomial distribution where the number of outcomes is $k \geq 2$ instead of 2. For example, instead of coin toss (with 2 outcomes), we could think of it as modeling the probability of counts of each side for rolling a k-sided die n times.

Def. 1.1.1 For n independent trials each of which leads to a success of exactly one of k categories with each category having a given fixed success probability, the *multinomial distribution* gives the probability of any particular combination of numbers of successes for the various categories.

Prop. 1.1.2 Let $Y_1, Y_2, \ldots, Y_k$ be random variables denoting the number of successes for each category and $y_1, y_2, \ldots, y_k$ be the observed values. The joint probability function is given by

$$f(y_1, y_2, \ldots, y_k; \theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

where $\sum_{j=1}^k y_j = n$ and the multinomial probability θ_j satisfy $0 \leq \theta_j \leq 1$ and $\sum_{j=1}^k \theta_j = 1$.

Ex. 1.1.3 A tool table has 6 pockets, and I practice poll by setting up a rack of 15 balls and trying to pot them in as few shots as possible. The game is repeated 50 times, so we have 750 balls in total. I want to see if balls were distributed across pockets evenly, or if I was more likely to pot balls into some pockets than others.

Pocket	Observed frequency	Expected frequency
Bottom left	156	125
Middle left	86	125
Top left	131	125
Top right	118	125
Middle right	102	125
Bottom right	157	125
Total	**750**	**750**

Let Y_j be the number of balls in pocket j and θ_j be the probability that a randomly selected ball ends up in pocket j for $1 \leq j \leq 6$. The joint distribution of $Y_1, Y_2, \ldots, Y_6$ is multinomial:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_6 = y_6) = \frac{n!}{y_1! y_2! \cdots y_6!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_6^{y_6} \quad 0 < \theta_j < 1, \sum_{j=1}^{6} \theta_j = 1$$

where $y_j = 0, 1, \ldots$ and $\sum_{j=1}^6 y_j = n$.

1.2 Likelihood Function and Maximum Likelihood Estimates

Prop. 1.2.1 For multinomial data $y_1, y_2, \ldots, y_k$, the likelihood function is

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{j=1}^{k} (\theta_j)^{y_j} = \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k} \quad \text{where } 0 < \theta_j < 1, \sum_{j=1}^{k} \theta_j = 1.$$

Prop. 1.2.2 The maximum likelihood estimate of θ_j is

$$\hat{\theta}_j = \frac{y_j}{n}, j = 1, 2, \ldots, k$$

while the maximum likelihood estimator of θ_j is

$$\tilde{\theta}_j = \frac{Y_j}{n}, j = 1, 2, \ldots, k.$$

Remark. There are only $(k-1)$ parameters are to be estimated since $\sum_{j=1}^{k} \theta_j = 1$.

Ex. 1.2.3. The likelihood function is $L(\theta_1, \theta_2, \ldots, \theta_6) = \prod_{j=1}^{6} (\theta_j)^{y_j}$ and the maximum likelihood estimate of θ_j is given by $\theta_j = y_j/6$.

j	y_j	$\hat{\theta}_j$
1	156	0.208
2	86	0.115
3	131	0.175
4	118	0.157
5	102	0.136
6	157	0.209

1.3 Likelihood Ratio Test Statistic

Prop. 1.3.1 Let Y_j be the observed frequency in category j and E_j be the expected frequency in category j if H_0 is true. The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0 = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$ is

$$\Lambda(\theta_0) = -2 \log \left[\prod_{j=1}^{k} \left(\frac{E_j}{Y_j} \right)^{Y_j} \right] = 2 \sum_{j=1}^{k} Y_j \log \frac{Y_j}{E_j},$$

which we can think of in words as $2 \times \text{sum}[\text{observed} \times \log(\text{observed}/\text{expected})]$.

Proof. By definition,

$$\begin{aligned}\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right) &= -2\log\frac{\prod_{j=1}^{k}(\theta_0)^{Y_j}}{\prod_{j=1}^{k}(\tilde{\theta})^{Y_j}} \\ &= -2\log\frac{\prod_{j=1}^{k}(1/k)^{Y_j}}{\prod_{j=1}^{k}(Y_j/n)^{Y_j}} \\ &= -2\log\left[\prod_{j=1}^{k}\left(\frac{E_j}{Y_j}\right)^{Y_j}\right] \qquad E_j = n/k \\ &= 2\sum_{j=1}^{k} Y_j \log\frac{Y_j}{E_j}. \qquad \square\end{aligned}$$

Ex. 1.3.2 The expected frequency $e_j = n/k = 125$ here is the same for all categories.

j	y_j	e_j
1	156	125
2	86	125
3	131	125
4	118	125
5	102	125
6	157	125

1.4 Hypothesis Testing

Lemma. 1.4.1 If n is large enough and H_0 is true, then the distribution of Λ is approximately χ^2_{k-1-p} where p denotes the number of parameters we are estimating.

Prop. 1.4.2 For a given data set, given the observed value of the likelihood ratio test statistic $\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\frac{y_j}{e_j}$, the p-value of the observed data can be computed as

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

Remark. As for "large enough", a guideline is to require $e_j \geq 5$ for all j, i.e., the expected counts under H_0 are at least 5 for every category in our data set.

Ex. 1.4.3 For the given data, the observed value of the likelihood ratio test statistic is

$$\lambda(\theta_0) = 2\sum_{j=1}^{6} y_j \log\frac{y_j}{e_i} = 33.571111$$

which is from approximately a $\chi^2_{k-1-p} = \chi^2_5$ distribution.

$$p\text{-value} \approx P(W \geq 33.571111) \approx 2.9 \times 10^{-6} \quad W \sim \chi^2_5$$

Since $p < 0.001$, there is very strong evidence against the null hypothesis that the balls are distributed equally across pockets when I play pool.

1.5 Pearson's Chi-Squared Goodness of Fit Statistic

Alternatively, we could use the *Pearson goodness of fit statistic.*

Def. 1.5.1. The *Pearson goodness of fit statistic* is defined as

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}.$$

with observed value

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}.$$

Prop. 1.5.2 For large n, D and Λ and asymptotically equivalent and have the same asymptotic chi-squared distribution.

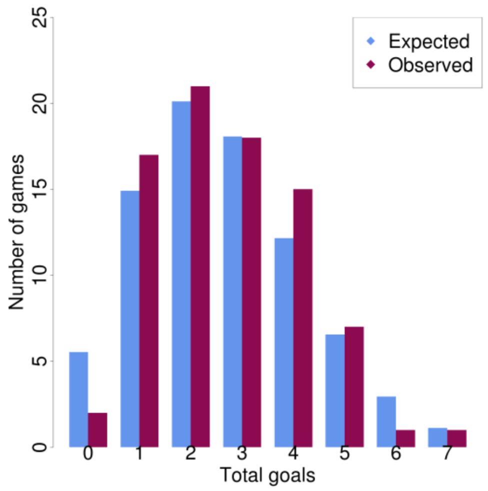
1.6 Full Example

Recall the Poisson distribution is used to model random events in time. Suppose we want to model the distribution of goals using a Poisson distribution.

Goals	0	1	2	3	4	5	6	7+
Games	2	17	21	18	15	7	1	1

We use MLE to calculate the probability the team would score a certain number of goals in a randomly chosen game:

$$P(Y = y) = \frac{\hat{\theta}^y e^{-\hat{\theta}}}{y!}, \quad y = 0, 1, \ldots$$



Let $\theta_0, \ldots, \theta_6$ denote the probability where a game has $0, \ldots, 6$ goals, respectively, and θ_7 denote the probability for a game has ≥ 7 goals. Let Y_j be the number of games from a sample of n with j goals. Then $Y_j \sim \text{Multinomial}(\theta_0, \ldots, \theta_7)$.

Our null hypothesis is H_0 : data are from a Poisson distribution, i.e.,

$$H_0 : \theta_j = \frac{\theta^j e^{-\theta}}{j!}, \quad j = 0, 1, 2, \ldots$$

(Note this is the general form of H_0 for testing a Poisson model. In our set up, θ_7 corresponds to the probability of a game with ≥ 7 goals.)

To test H_0, we need to estimate θ. The MLE for θ is $\hat{\theta} = 2.695$, which is the average number of goals scored per game.

Goals	Observed y_j	Expected e_j
0	2	5.538
1	17	14.925
2	21	20.112
3	18	18.068
4	15	12.174
5	7	6.562
6	1	2.948
7+	1	1.135

Recall for Λ define in **Prop. 1.3.1**, if $e_j \geq 5$ for all j, then Λ has approximate a chi-squared distribution with $k - 1 - p$ degrees. But two of the expected frequencies (for 6 goals and $7+$ goals) are less than 5 as shown above. To solve this, we collapse adjacent categories with the smallest expected frequencies:

Goals	Observed y_j	Expected e_j
0	2	5.538
1	17	14.925
2	21	20.112
3	18	18.068
4	15	12.174
5+	9	10.645

Since $e_j \geq 5$ for all j, it is safe to consult a chi-squared distribution. The number of categories k has reduced from 8 to 6, so we need to consult $\chi^2_{k-1-p} = \chi^2_4$. The observed value of the likelihood ratio statistic is

$$\lambda(\theta_0) = 2 \sum_{j=1}^{5} y_j \log \frac{y_j}{e_j} = 5.270.$$

The p-value is $P(W \geq 5.270) \approx 0.261 > 0.1$ where $W \in \chi^2_4$. Since $p > 0.1$, there is no evidence against the Poisson model based on these data.

We could also use the Pearson goodness of fit statistic, where we will again compare our value $D = d$ with a χ^2_4 distribution. The observed value is

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j} = 3.498$$

and p-value is $\approx P(W \geq 3.498) \approx 0.478$ where $W \sim \chi^2_4$. Again, there is no evidence against the Poisson model based on these data.

2 Two-Way Tables and Independence Tests

To assess whether two factors or variates appear to be related, we could test the hypothesis that the factors are independent and thus statistically unrelated. Suppose both variates are discrete and take on a fairly small number of possible values.

Suppose that elements in a population in can be classified according to each of two factors A and B. For A, an element can be of any of a mutually exclusive types $A_1, A_2, \ldots, A_a$ where $a \geq 2$ and the same goes for B. If a random sample of n individuals is selected, let y_{ij} denote the number that have type A_i and B_j. We can use a *two-way table* (also called a *contingency table*) to record the observed frequencies:

$A \backslash B$	B_1	B_2	$\cdots$	B_b	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	r_1
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	r_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_a	y_{a1}	$\cdots$	$\cdots$	y_{ab}	r_a
Total	c_1	c_2	$\cdots$	c_b	n

where $r_i = \sum_{j=1}^{b} y_{ij}$, $c_j = \sum_{i=1}^{a} y_{ij}$, and $\sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} = n$.

Let θ_{ij} be the probability that a randomly selected individual is of type (A_i, B_j) and note that $\sum_{i=1}^{a} \sum_{j=1}^{b} \theta_{ij} = 1$. Then $(Y_{11}, Y_{12}, \ldots, Y_{ab})$ follows a multinomial distribution with $k = ab$ classes.

To test independence of the A and B classifications, we test the hypothesis

$$H_0 : \theta_{ij} = \alpha_i \beta_j, \quad 1 \leq i \leq a, 1 \leq j \leq b$$

where $0 \leq \alpha_i < 1$, $0 < \beta_j < 1$, $\sum_{i=1}^{a} \alpha_i = 1$, and $\sum_{j=1}^{b} \beta_j = 1$.

Let $\alpha_i = P(\text{an individual is of type } A_i)$ and $\beta_j = P(\text{an individual is of type } B_j)$, the hypothesis above is the standard definition for independent events:

$$P(A_i \cap B_j) = P(A_i)P(B_j).$$

The expected frequencies are given by

$$e_{ij} = \frac{r_i c_j}{n}, \quad 1 \leq i \leq a, 1 \leq j \leq b.$$

The likelihood ratio test statistic for testing the hypothesis of independence is

$$\Lambda = 2 \sum_{i=1}^{2} \sum_{j=1}^{b} Y_{ij} \log \frac{Y_{ij}}{E_{ij}}$$

with observed value

$$\lambda = 2 \sum_{i=1}^{2} \sum_{j=1}^{b} y_{ij} \log \frac{y_{ij}}{e_{ij}}$$

If $e_{ij} \geq 5$ for all (i, j), then this will approximately follow a χ^2_{k-1-p} distribution, where $k = ab$ and $p = (a-1) + (b-1)$, i.e., $\Lambda \sim \chi^2_{(a-1)(b-1)}$ given n is reasonably large and all expected frequencies are ≥ 5.