Contingency Tables

Author: Nicholas Reich and Anna Liu

Course: Categorical Data Analysis (BIOSTATS 743)

Statistical inference for multinomial parameters

Given n observations in c categories, n_j occur in category j, $j = 1, ..., c$. The multinomial log-likelihood function is

$$L(\pi) = \sum_j n_j \log \pi_j$$

Maximizing this gives MLE

$$\hat{\pi}_j = n_j/n$$

Chi-square goodness-of-fit test for a specified multinomial

Consider hypothesis $H_0 : \pi_j = \pi_{j0}, j = 1, ..., c$, - **Chi-square goodness-of-fit statistic (score)**

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}$$

where $\mu_j = n\pi_{j0}$ is called **expected frequencies under** H_0.

- Let X_o^2 denote the observed value of X^2. The P-value is $P(X^2 > X_o^2)$.
- For large samples, X^2 has approximately a chi-squared distribution with $df = c - 1$. The P-value is approximated by $P(\chi^2_{c-1} \geq X_o^2)$.

LRT test for a specified multinomial

- **LRT statistic**

$$G^2 = -2 \log \Lambda = 2 \sum_j n_j \log(n_j / n\pi_{j0})$$

  For large n, G^2 has a chi-squared null distribution with $df = c - 1$.

- When H_0 holds, the goodness-of-fit Chi-squiare X^2 and the likelihood ratio G^2 both have large-sample chi-squared distributions with $df = c - 1$.
- For fixed c, as n increases the distribution of X^2 usually converges to chi-squared more quickly than that of G^2. The chi-squared approximation is often poor for G^2 when $n/c < 5$. When c is large, it can be decent for X^2 for n/c as small as 1 if table does not contain both very small and moderately large expected frequencies.

Distributions of categorical variables: Multinomial

Suppose that each of n **independent and identical** trials can have outcome in any of c categories. Let

$$y_{ij} = \begin{cases} 1 & \text{if trial } i \text{ has outcome in category } j \\ 0 & \text{otherwise} \end{cases}$$

Then $\mathbf{y}_i = (y_{i1}, ..., y_{ic})$ represents a multinomial trial with $\sum_j y_{ij} = 1$. Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, ..., n_c)$ have the *multinomial distribution*. The multinomial pmf is

$$p(n_1, ..., n_{c-1}) = \left(\frac{n!}{n_1! n_2! ... n_c!} \right) \pi_1^{n_1} \pi_2^{n_1} ... \pi_c^{n_c},$$

where $\pi_j = P(Y_{ij} = 1)$

$$E(n_j) = n\pi_j, \qquad Var(n_j) = n\pi_j(1 - \pi_j)$$

$$Cov(n_i, n_j) = -n\pi_i\pi_j$$

Distributions of categorical variables: Poisson

One simple distribution for count data that do not result from a fixed number of trials. The Poisson pmf is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, y = 0, 1, 2, \ldots \quad E(Y) = Var(Y) = \mu$$

For adult residents of Britain who visit France this year, let

- $Y_1=$ number who fly there
- $Y_2=$number who travel there by train without a car
- $Y_3=$number who travel there by ferry without a car
- $Y_4=$number who take a car A poisson model for (Y_1, Y_2, Y_3, Y_4) treats these as independent Poisson random variables, with parameters $(\mu_1, \mu_2, \mu_3, \mu_4)$. The total $n = \sum_i Y_i$ also has a Possion distribution, with parameter $\sum_i \mu_i$.

Distributions of categorical variables: Poisson

The conditional distribution of (Y_1, Y_2, Y_3, Y_4) given $\sum_i Y_i = n$ is *multinomial*$(n, \pi_i = \mu_i / \sum_j \mu_j)$

The Chi-Squared distribution

This is not a distribution for the data but rather a sampling distribution for many statistics.

- The chi-squared distribution with degrees of freedom by df has mean df, variance $2(df)$, and skewness $\sqrt{8/df}$. It converges (slowly) to normality as df increases, the approximation being reasonably good when df is at least about 50.
- Let $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$
- The **reproductive property**: if $X_1^2 \sim \chi^2(\nu_1)$ and $X_2^2 \sim \chi^2(\nu_2)$, then $X^2 = X_1^2 + X_2^2 \sim \chi^2(\nu_1 + \nu_2)$. In particular, $X = Z_1^2 + Z_2^2 + ... + Z_\nu^2 \sim \chi^2(\nu)$ with the standard normal Z's.

An example from r-tutor.com

In the built-in data set survey, the Smoke column records the survey response about the student's smoking habit. As there are exactly four proper response in the survey: "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never", the Smoke data is multinomial. It can be confirmed with the levels function in R.

```r
library(MASS)       # load the MASS package
levels(survey$Smoke)
```

```
## [1] "Heavy" "Never" "Occas" "Regul"
```

```r
(smoke.freq = table(survey$Smoke))
```

```
##
## Heavy Never Occas Regul
##    11   189    19    17
```

Problem and solution

Suppose the campus smoking statistics is as below. Determine whether the sample data in survey supports it at .05 significance level.

Heavy	Never	Occas	Regul
4.5%	79.5%	8.5%	7.5%

We save the campus smoking statistics in a variable named smoke.prob. Then we apply the chisq.test function and perform the Chi-Squared test.

```
smoke.prob = c(.045, .795, .085, .075)
chisq.test(smoke.freq, p=smoke.prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  smoke.freq
```

Testing with estimated expected frequencies

In some applications, the hypothesized $\pi_{j0} = \pi_{j0}(\theta)$ are functions of a smaller set of unknown parameters θ. In this case

- Obtain the ML estimates of expected frequencies: $\hat{\mu}_j = n\pi_{j0}(\hat{\theta})$ by plugging in the ML estimates $\hat{\theta}$ of θ
- Replacing μ_j by $\hat{\mu}_j$ in the definition of X^2 and G^2
- The approximate distributions of X^2 and G^2 are χ^2_{df} with $df = (c-1) - dim(\theta)$.

Example: Pneumonia infections in Calves

A sample of 156 dairy calves born in Okeechobee County, Florida, were classified according to whether they caught pneumonia within 60 days of birth. Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within 2 weeks after the first infection cleared up.

Secondary	Infection	
Primary Infection	Yes	No
Yes	30(38.1)	63(39.0)
No	0	63(78.9)

Probability structure for null hypothesis

The goal of the study was to test whether the probability of primary infection was the same as the conditional probability of secondary infection, given that the calf got the primary infection. Let π be the probability of primary infection, then the null hypothesis states that

Secondary	Infection	-	
Primary Infection	Yes	No	Total
Yes	π^2	$\pi(1-\pi)$	π
No	–	$1-\pi$	$1-\pi$

Example continued

Let n_{ab} denote the number of observations in row a and column b. The ML estimate of π is the value maximizing the kernel of the multinomial likelihood

$$(\pi^2)^{n_{11}}(\pi - \pi^2)^{n_{12}}(1 - \pi)^{n_{22}}$$

The MLE is

$$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22}) = 0.494$$

The score statistic is $X^2 = 19.7$. It follows a Chi-square distribution with $df = c - p - 1 = (3 - 1) - 1 = 1$. The p-value is

$$P(\chi_1^2 > 19.7) = 0.00001$$

Therefore, the primary infection had an immunizing effect that reduced the likelihood of secondary infection.