Contingency Tables

Chi-Square Tests of Independence

Demetris Athienitis

Framework

With a multinomial, $\mu_{ij} = n\pi_{ij}$ and we wish to test

$$\mathrm{H}_0 : \mu_{ij} = \mu_{ij}^0$$

Under the assumption of independence

$$\begin{aligned} \mu_{ij}^0 &= n\pi_{ij} \\ &= n(\pi_{i+})(\pi_{+j}) && \text{by ind.} \end{aligned}$$

and the MLEs under independence, are

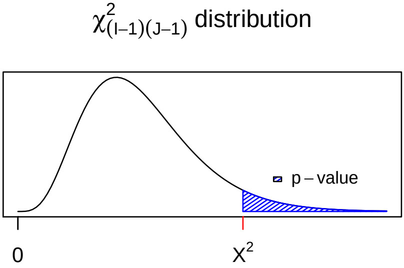
$$\begin{aligned} \hat{\mu}_{ij} &= n\hat{\pi}_{i+}\hat{\pi}_{+j} \\ &= \cancel{n}\frac{n_{i+}}{\cancel{n}}\frac{n_{+j}}{n} \\ &= \frac{(n_{i+})(n_{+j})}{n} \end{aligned}$$

Pearson

The *Pearson chi-square test statistic*, with the condition that $\hat{\mu}_{ij} > 5 \; \forall i, j$ is asymptotically

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \underset{\text{approx.}}{\overset{\text{H}_0}{\sim}} \chi^2_{(I-1)(J-1)}$$

with p-value $P\left(\chi^2_{(I-1)(J-1)} \geq X^2\right)$ (area to the right of the test statistic)



Example (Job Satisfaction)

Data from General Social Survey (1991)

Income	Job Satisfaction				
	Dissat	Little	Moderate	Very	Total
$< 5k$	2	4	13	3	22
5k - 15k	2	6	22	4	34
15k - 25k	0	1	15	8	24
$> 25k$	0	3	13	8	24
Total	4	14	63	23	104

```
> job_test=chisq.test(job); job_test
data:  job
X-squared = 11.524, df = 9, p-value = 0.2415

Warning: Chi-squared approximation may be incorrect
```

Example (continued)

Warning because many expected frequencies are < 5

```
> round(job_test$expected,2)
        Dissat Little Moderate Very
<5        0.85   2.96    13.33 4.87
5k-15k    1.31   4.58    20.60 7.52
15k-25k   0.92   3.23    14.54 5.31
>25k      0.92   3.23    14.54 5.31
```

As p-value is large, with **caution/reservations**, we conclude that we fail to reject the null of independence.

Likelihood Ratio

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

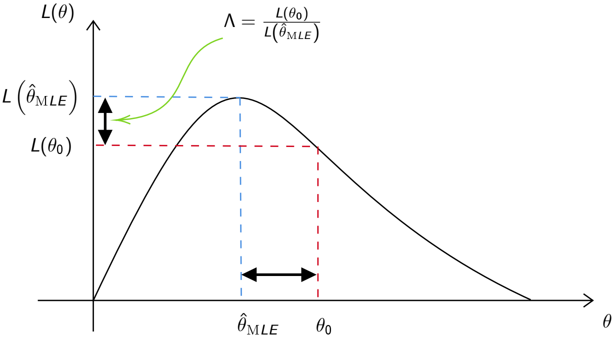
The likelihood ratio is given by

$$\Lambda = \frac{\text{maximum likelihood when } H_0 \text{ is true}}{\text{maximum likelihood when parameters are unrestricted}}$$

So if the ratio is close to 1 it implies that the estimated parameter(s) under the null are close in proximity to the unrestricted MLEs and hence null is plausible.

Likelihood Ratio

For example, $H_0 : \theta = \theta_0$. To determine if the null value θ_0 is plausible, compare it to the maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$, by seeing how close the likelihood functions are at θ_0 and $\hat{\theta}_{\text{MLE}}$.



Likelihood Ratio Test

The *Likelihood Ratio Test (LRT) statistic* is asymptotically

$$G^2 = -2 \log \Lambda \underset{\text{approx.}}{\overset{\mathrm{H}_0}{\sim}} \chi^2_{df}$$

$$\begin{aligned} \text{degrees of freedom} &= \text{no. of parameters in general} \\ &\quad - \text{no. of parameters under } \mathrm{H}_0 \end{aligned}$$

LRT for multinomial

For an $I \times J$ table the likelihood is

$$L(\pi_{ij}; n_{ij}) = \frac{n!}{n_{11}! \cdots n_{IJ}!} \pi_{11}^{n_{11}} \cdots \pi_{IJ}^{n_{IJ}}$$

$$\Rightarrow \Lambda = \frac{\left(\frac{n_{i+} n_{+j}}{n^2}\right)^{n_{ij}}}{\left(\frac{n_{ij}}{n}\right)^{n_{ij}}}$$

Ignoring constants and recalling $\hat{\mu}_{ij} = (n_{i+} n_{+j})/n$,

$$G^2 = 2 \sum_{ij} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$$

with $(I-1)(J-1)$ degrees of freedom...shown next.

LRT for multinomial

- In general, there are IJ groupings in the multinomial with IJ, π_{ij}'s, hence $IJ - 1$ free parameters in general.
- Under H_0, $I - 1$ free π_{i+}'s and $J - 1$ free π_{+j}'s

$$\begin{aligned} df &= (IJ - 1) - [(I - 1) + (J - 1)] \\ &= (I - 1)(J - 1) \end{aligned}$$

Example (Job Satisfaction continued)

```
> library(DescTools)
> GTest(job)

data:  job
G = 13.467, X-squared df = 9, p-value = 0.1426
```

Remark

- No warning message was given for G^2
- As $n \to \infty, X^2 \xrightarrow{d} \chi^2$ faster than $G^2 \xrightarrow{d} \chi^2$, but they are usually similar and asymptotically equivalent, i.e. $X^2 - G^2 \xrightarrow{d} 0$
- These tests treat X and Y as nominal and reordering rows or columns has no effect. Methods for ordinal tests (section 2.5 of textbook as well as author's other textbooks) do exist

Standardized residuals

- Once we have established a dependence in the data, it is of interest to explore where the dependence lies
- Which cells in the table have higher/lower counts than expected (under independence)?
- To explore this, we can look at standardized residuals

Definition (Standardized/Adjusted Residuals)

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}} \overset{H_0}{\sim} N(0, 1)$$

Hence, $|r_{ij}| > 2$ considered significant.

Example (Job Satisfaction continued)

Residuals are:

```
> round(job_test$stdres,4)
         Dissat  Little Moderate    Very
<5       1.4406  0.7305  -0.1606 -1.0792
5k-15k   0.7525  0.8716   0.6005 -1.7726
15k-25k -1.1171 -1.5211   0.2198  1.5098
>25k    -1.1171 -0.1574  -0.7327  1.5098
```

Section 2

Partitioning chi-squared statistics

The sum of two independent chi-squared random variables also follows a chi-squared distribution

Lemma

Let $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ be independent. Then,

$$\chi^2_{\nu_1} + \chi^2_{\nu_2} \sim \chi^2_{\nu_1+\nu_2}$$

The G^2 statistic can be partitioned into separate components to help represent certain aspects of the association.

Partitioning chi-squared statistics

Income	Job Satisfaction				G^2	df
	Dissat	Little	Moderate	Very		
Low					0.30	3
$<$ 5k	2	4	13	3		
5k - 15k	2	6	22	4		
High					1.19	3
15k - 25k	0	1	15	8		
$>$ 25k	0	3	13	8		
Low vs High					11.98	3
$< 15k$	4	10	35	7		
$> 15k$	0	4	28	16		
					13.47	9

Partitioning chi-squared statistics

- Note that the partitioned G^2 values sum to the full table value
- Within low salary or high salary jobs, we see a very small G^2 value
- If we collapse the two low category groups into one, and collapse the two high salary categories into one, then we see a larger G^2 value
- $G^2 = 11.98$ with 3 degrees of freedom gives a p-value of 0.007. Much different story than looking at the full table

We learned

Can test for independence using

- Pearson
- LRT
- and that LRT G^2 may be partitioned