In-Class 2. Categorical Data Analysis

Suzer-Gurtekin

January 2025

Overview

Two-Way Tables

Starting with general notation…

Rows=I
Columns=J

	Columns					
Rows	1	2	3	4	5	
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{2+}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{3+}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{4+}
	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n_{+5}	n

Two-Way Tables

In this notation, each cell is n_{ij} where i is the *row* and j is the *column*.

The plus sign denotes marginal totals:

$$n_{i+} = \sum_{j=1}^{J} n_{ij}$$ Sum across columns holding row constant

$$n_{+j} = \sum_{i=1}^{I} n_{ij}$$ Sum across rows holding column constant

$$n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

2×2 Tables

Special Case of IxJ table: I=2 number of rows and J=2 number of columns:
i= 1,2
j= 1,2

Condition	Present	Absent	Row Total
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}
Column Total	n_{+1}	n_{+2}	n

Group Assignments

Group	Student
1	Einolf, Zach Scott
1	Fan, Zhaoyun
1	Mishra, Rohin Prem
1	DesJardins, Grace
2	Adeniyi, Kehinde
2	Lugu, Nicholas Reign
2	LU, Aria
2	Gunderson, Jeremy
3	Wenner, Theodore D
3	Zhou, Zhenjing
3	Kim, Jay
3	Bei, Rongqi
4	Beshaw, Yael Dejene
4	Hoglund, Quentin Michael
4	Jiang, Yujing
4	Jiang, Weishan
5	Popky, Dana
5	Sani, Jamila
5	O'Connell, Greg Al
5	Saucedo, Valeria Castaneda

Group	Student
6	Hussein, Aya Moham
6	Zou, Jianing
6	Wang, Zixin
6	Chakravarty, Sagnik
7	Valmidiano, Megan
7	Glidden, Sarah Acton
7	Sun, Yao
7	Blakney, Aaron
8	Xu, Kailin
8	Linares, Kevin
8	Odei, Doris
8	Nana Mba, Line
9	Zhou, Huan
9	Meng, Lingchen
9	Lin, Xinyu
9	Ge, Feiran
10	Liu, Xiaoqing
10	Lu, Angelina
10	Baez-Santiago, Felix
10	Ma, Ruisi

Group	Student
11	Ding, Yuchen
11	Shrivastava, Namit
11	Kakiziba, Johnia Johansen
11	Cranmer, Evan Koba

Expectations

Active participation in

- Reviewing question/data/method
- Code writing
- Computations
- Interpretation of results
- Select a spokesperson for group discussion

In-Class Problem 1

Starting discussion as a group …

	Disease		
	Yes	No	Row Total
Male	10	40	50
Female	20	30	50
Column Total	30	70	100

1. What is n_{+1}? [Please tell me what this quantity is in plain English]
2. What is n_{2+}?
3. What is $Pr(M, D)$?
4. What is $Pr(D|F)$?

Group Discussion

- Work in groups
- Randomly selected group to go over the solutions to questions 1 and 2
- Randomly selected group to go over the solutions to questions 3 and 4

In-Class Problem 1

	Disease		
	Yes	No	Row Total
Male	10	40	50
Female	20	30	50
Column Total	30	70	100

What is n_{+1}? 30

$$\sum_{i=1}^{I=2} n_{1,1} + n_{2,1} = 10 + 20 = 30$$

What is n_{2+}? 50

$$\sum_{j=1}^{J=2} n_{2,1} + n_{2,2} = 20 + 30 = 50$$

In-Class Problem 1

	Disease		
	Yes	No	Row Total
Male	10	40	50
Female	20	30	50
Column Total	30	70	100

What is $Pr(M, D)$?0.10

$$\Pr(M, D) = \frac{n_{11}}{n} = \frac{10}{100} = 0.10$$

What is $Pr(D|F)$?0.40

$$\Pr(D|F) = \frac{n_{22}}{n_{2+}} = \frac{n_{22}}{\sum_{j=1}^{2} n_{2j}} = \frac{n_{22}}{n_{21} + n_{22}} = \frac{20}{20+30} = \frac{20}{50} = 0.40$$

Two-Way Tables

If the two variables are unrelated, then any cell proportion is the product of the marginal proportions. Using the notation from last time: $\pi_{ij} = \pi_{i+}\pi_{+j}$.

This gives us a method for creating **Expected** counts if we want to test for independence.

Write EXPECTED counts, using our notation:

$$e_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$$

Pearson Chi-Square Statistic

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

where $O_i = n_i$ is the observed count in the i^{th} category, and $E_i = np_{0i}$ is the expected count in the i^{th} category (from H_0).

In-Class Problem 2

The FREQ Procedure

Table of gender by age

gender Frequency Expected Cell Chi-Square	age 0-29	30-39	40-49	50-59	>=60	Total
males	185 180.22 0.1269	207 209.78 0.0368	260 257.46 0.0252	180 178.31 0.016	71 77.237 0.5036	903
females	4 8.7814 2.6034	13 10.222 0.7551	10 12.545 0.5163	7 8.6885 0.3281	10 3.7635 10.335	44
Total	189	220	270	187	81	947

Statistics for Table of gender by age

Statistic	DF	Value	Prob
Chi-Square	4	15.2461	0.0042
Likelihood Ratio Chi-Square	4	12.6670	0.0130
Mantel-Haenszel Chi-Square	1	4.8961	0.0269
Phi Coefficient		0.1269	
Contingency Coefficient		0.1259	
Cramer's V		0.1269	

Sample Size = 947

In-Class Problem 2

Please use the table on the previous page:

1. Write down what is I and J?
2. Write down the table in IxJ notation
3. Data as the table form is saved on canvas website *drunk.dat*:
4. Write down the case level data for this table with the following variable names and give definitions (min/max, value labels):

Case Number   Sex   Age   Disease

5. Using table data calculate a X^2 test of association by hand (or in a spreadsheet), that is, not using R.

In-Class Problem 2

Now let's look at the solution in R. The following code:

```
chisq.test(data.matrix(drunk))
```

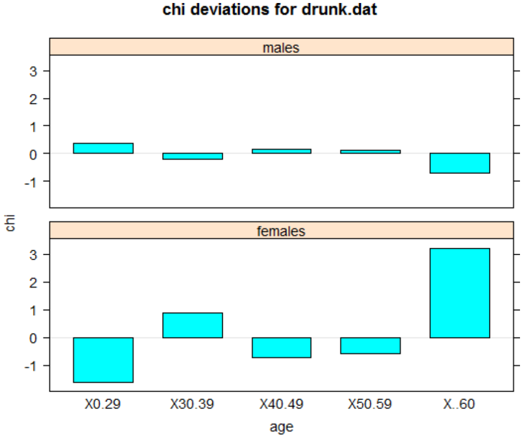
Produces the following output:

```
 Pearson's Chi-squared test

data:  data.matrix(drunk)
X-squared = 15.2461, df = 4, p-value = 0.004217

Warning message:
In chisq.test(data.matrix(drunk)) :
  Chi-squared approximation may be incorrect
```

drunk.r

Example 1

We can also produce a graphic of the chi-square deviations:



Example 2

Now let's look at the example in R. We have a *matrix* with our data:

```
> glasses
           delinq non.del
glasses         1       5
no.glasses      8       2
```

	delinq	non.del	Row Total
glasses	a	b	**a+b**
no.glasses	c	d	c+d
Column Total	**a+c**	b+d	n

What is the probability that the table would be lopsided if wearing glasses was unrelated to delinquency?

Example 2

We can request Fisher's exact test with the following code:

```
fisher.test(glasses)
```

Which produces the following output:

```
 Fisher's Exact Test for Count Data

data:   glasses
p-value = 0.03497
alternative hypothesis: true odds ratio is not equal t
95 percent confidence
 interval: 0.0009525702
 0.9912282442
sample
estimates:
odds ratio
0.06464255
```

OR and RR

	Yes	No	Row Total
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Column Total	π_{+1}	π_{+2}	

	Disease	
	Yes	No
Male	0.15	0.35
Female	0.10	0.40

We want the **conditional probability** that you have disease given that you are male: $PR(D|M) = \pi_{1|1} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}}$.

$$PR(D|M) = \pi_{1|1} = \frac{0.15}{0.15+0.35} = 0.3$$

Relative Risk

Now, define the relative risk.

Relative Risk (Response Category 1) = $\frac{\pi_{1|1}}{\pi_{1|2}}$

For example:

$$\frac{\pi_{1|1}}{\pi_{1|2}} = \frac{\Pr(D|M)}{\Pr(D|F)} = \frac{.3}{.2} = 1.5$$

Please take a moment and compute the conditional probability of having a disease given that you are a female.

Relative Risk

An easy way to estimate the relative risk is:

$$\frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

The distribution of the relative risk is **highly skewed**. Therefore, it is better to estimate the variance on the log scale.

$$\hat{V}\left\{\ln\left(\frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}\right)\right\} = \frac{1-\frac{n_{11}}{n_{1+}}}{n_{11}} + \frac{1-\frac{n_{21}}{n_{2+}}}{n_{21}} = \frac{\Pr(\bar{D}|M)}{n_{11}} + \frac{\Pr(\bar{D}|F)}{n_{21}}$$

This is the variance of the **natural logarithm** of the Relative Risk(Response Category 1).

In-Class Problem 3

Group Exercise

	Disease	
	Yes	No
Male	20	40
Female	30	30

1. What is the relative risk for men relative to women?
2. What is the variance of this estimate? Leave it on the natural log scale, $\hat{V}\left(ln(\hat{RR})\right)$.

In-Class Problem 3

	Disease	
	Yes	No
Male	20	40
Female	30	30

What is the relative risk for men relative to women? $\frac{20/60}{30/60} = \frac{.33}{.5} = 0.67$

What is the variance of this estimate? $\hat{V}\left(ln(\hat{RR})\right) = \frac{3}{60}$

Odds Ratio

Within row 1 the *odds* (***not* odds ratio**) that the response is in column 1 instead of column 2 is defined as:

$$Odds_1 = \frac{\pi_{1|1}}{\pi_{2|1}} = \frac{\pi_{11}}{\pi_{12}} \qquad \frac{probability\ of\ you\ are\ in\ column\ 1\ given\ that\ you\ are\ in\ row\ 1}{probability\ of\ you\ are\ in\ column\ 2\ given\ that\ you\ are\ in\ row\ 1}$$

From our example, this could be written:

$$\frac{\Pr(D|M)}{\Pr(\bar{D}|M)} = \frac{\Pr(D|M)}{1 - \Pr(D|M)}$$

	Yes	No	Row Total
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Column Total	π_{+1}	π_{+2}	

Continuing the example, the odds that a man will have the disease are $\frac{.3}{1-.3} = .43$. For women, this odds are $\frac{.2}{1-.2} = .25$

Odds Ratio

Those are the odds. The ratio of the odds is called the *odds ratio*.

$$\theta = \frac{\pi_{1|1}/\pi_{2|1}}{\pi_{1|2}/\pi_{2|2}}$$

	Yes	No	Row Total
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Column Total	π_{+1}	π_{+2}	

Remember that $\pi_{1|1} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}}$. So we can also write it in the following manner:

$$\theta = \frac{\frac{\pi_{11}}{\pi_{11}+\pi_{12}} \Big/ \frac{\pi_{12}}{\pi_{11}+\pi_{12}}}{\frac{\pi_{21}}{\pi_{21}+\pi_{22}} \Big/ \frac{\pi_{22}}{\pi_{21}+\pi_{22}}} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$

Hence the name *cross-product ratio*.

Odds Ratio

We estimate the odds ratio using:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

The variance of $ln(\hat{\theta})$ can be estimated as:

$$\hat{V}\left\{\ln\left(\hat{\theta}\right)\right\} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

In-Class Problem 4

Leg Healing Trial Callum et al. 1992				
	Leg Wound Healing			
	Healed	Not Healed	Total	P_{i+}
Elastic	35	30	65	0.538
Inelastic	19	48	67	0.284

1. Please calculate the **relative risk** and **odds ratio** of healing comparing elastic to inelastic bandages.
2. Please calculate the **variance** of the odds ratio estimate.

In-Class Problem 4

$\hat{\theta} = \frac{35 \times 48}{19 \times 30} = 2.9474$

The variance is determined for $ln(\hat{\theta}) = 1.0809$.

$$\hat{V}\left\{\ln\left(\hat{\theta}\right)\right\} = \frac{1}{35} + \frac{1}{30} + \frac{1}{48} + \frac{1}{19} = 0.13537$$

In-Class Problem 4

Do all the steps on the natural logarithmic scale!

1. $\sqrt{V\left\{\ln\left(\hat{\theta}\right)\right\}} = 0.36793.$
2. $1.96 * 0.36793 = 0.72114.$
3. $1.0809 - 0.72114 = 0.359777398$ and $1.0809 + 0.72114 = 1.802048025$. This is the 95% confidence interval on the logarithmic scale.
4. Exponentiate to get back to the scale of $\hat{\theta}$. Therefore, $(e^{0.359777398}, e^{1.802048025}) = (1.443, 6.062)$.

R Code

We can do this work in R. Here I'm using the *epiR* package:

```
library(epiR)

bandage<-matrix(data=c(35,19,30,48),nrow=2)
bandage<-as.table(bandage)

epi.2by2(bandage)
```

Example 4

Which produces the following output:

```
> epi.2by2(bandage)
Outcome +   Outcome -      Total
Exposed +          35          30          65
Exposed -          19          48          67
Total              54          78         132
Inc risk *       Odds
Exposed +             53.8       1.167
Exposed -             28.4       0.396
Total                 40.9       0.692

 Point estimates and 95% CIs:
----------------------------------------------------------------
 Inc risk ratio                            1.90 (1.22, 2.95)
 Odds ratio                                2.95 (1.43, 6.06)
 Attrib risk *                             25.49 (9.26, 41.72)
 Attrib risk in population *              12.55 (-1.12, 26.22)
 Attrib fraction in exposed (%)           47.33 (18.05, 66.15)
 Attrib fraction in population (%)         30.68 (7.22, 48.21)
----------------------------------------------------------------
Test that OR = 1: chi2(1) = 8.866 Pr>chi2 = 0.00
Wald confidence limits
CI: confidence interval
* Outcomes per 100 population units
```