ST3241 Categorical Data Analysis I
Two-way Contingency Tables

2×2 Tables, Relative Risks and Odds Ratios

What Is A Contingency Table (p.16)

- Suppose X and Y are two categorical variables
- X has I categories
- Y has J categories
- Display the IJ possible combinations of outcomes in a rectangular table having I rows for the categories of X and J columns for the categories of Y.
- A table of this form in which the cells contain frequency counts of outcomes is called a *contingency table*.

Example: Belief In Afterlife Data (p.18)

	Belief in Afterlife	
Gender	Yes	No or Undecided
Female	435	147
Male	375	134

- A contingency table that cross classifies two variables is called a *two − way table*.
- A table which cross classifies three variables is called a *three − way table*.
- A two-way table having I rows and J columns is called an $I \times J$ table.

Some Notations, Definitions $\cdots$

- $\pi_{ij} = P[X = i, Y = j]$ = probability that (X, Y) falls in the cell in row i and column j.
- The probabilities $\{\pi_{ij}\}$ form the joint distribution of X and Y.
- Note that,

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1$$

Marginal Distributions (p.17)

- The marginal distribution of X is π_{i+}, which is obtained by the row sums, that is,

$$\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}$$

- The marginal distribution of Y is π_{+j}, which is obtained by the column sums, that is

$$\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$$

- For example, for a 2×2 table

$$\pi_{1+} = \pi_{11} + \pi_{12}, \pi_{+1} = \pi_{11} + \pi_{21}$$

Notations For The Data

- Cell counts are denoted by $\{n_{ij}\}$, with

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

- Cell proportions are

$$p_{ij} = \frac{n_{ij}}{n}$$

- The marginal frequencies are row totals $\{n_{i+}\}$and column totals $\{n_{+j}\}$

Example

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Female	n_{11}=435	n_{12}=147	n_{1+}=582
Male	n_{21}=375	n_{22}=134	n_{2+}=509
Total	n_{+1}=810	n_{+2}=281	n=1091

Example: Sample Proportions

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Female	p_{11}=0.398	p_{12}=0.135	p_{1+}=0.533
Male	p_{21}=0.344	p_{22}=0.123	p_{2+}=0.467
Total	p_{+1}=0.742	p_{+2}=0.258	p=1.00

Conditional Probabilities

- Let Y be a response variable and X be an explanatory variable.
- It is informative to construct separate probability distributions for Y at each level of X.
- Such a distribution consists of *conditional probabilities* for Y given the level of X and is called a *conditional distribution*.

Example: Sample Conditional Distributions

- For females,
  - Proportion of *yes* responses = 0.747
  - Proportion of *no* responses = 0.253
- For males,
  - Proportion of *yes* responses = 0.737
  - Proportion of *no* responses = 0.263

Independence

- Is the belief in afterlife is independent of gender?
- Two variables are statistically independent if all joint probabilities equal the product of their marginal probabilities $\pi_{ij} = \pi_{i+}\pi_{+j}$, for $i = 1, \cdots, I$ and $j = 1, \cdots, J$
- Conditional distributions of Y are identical at each levels of X.

Probability Model For A 2×2 Table

- Poisson Model
  - Each of the 4 cell counts are independent Poisson random variables
- Binomial Model
  - Marginal totals of X are fixed.
  - Conditional distributions of Y at each level of X are binomial.
- Multinomial Model
  - Total sample size is fixed but not the row or column totals.
  - The distribution of 4 cell counts are then multinomial

Comparing Proportions In 2×2 Tables

- Assume that the row totals are fixed and hence we have a binomial model.
- Suppose the two categories of Y are *success* and *failure*.
- Let π_1 = Probability of success in row 1 and π_2 = Probability of success in row 2.
- The difference in probabilities $\pi_1 - \pi_2$ compares the success probabilities in two rows.

Sample Difference of Proportions

- Let p_1 and p_2 be sample proportions of success for the two rows.
- The sample difference $p_1 - p_2$ estimates $\pi_1 - \pi_2$.
- If the counts in two rows are independent samples, the estimated standard error of $p_1 - p_2$ is

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

Example: Belief in Afterlife (p.16)

- In our example

$$p_1 = n_{11}/n_{1+} = 435/582 = 0.747$$
$$p_2 = n_{21}/n_{2+} = 375/509 = 0.737$$

- Therefore, $p_1 - p_2 = 0.747 - 0.737 = 0.010$
- The estimated standard error

$$\hat{\sigma}(p_1 - p_2) = \sqrt{p_1(1-p_1)/n_{1+} + p_2(1-p_2)/n_{2+}} = 0.02656$$

Example: Aspirin Use And Myocardial Infarction (p. 20)

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

- To find out whether regular intake of Aspirin reduces mortality from cardiovascular diseases

Example: Continued

- In this example, $p_1 = 189/11034 = 0.0171$ and $p_2 = 104/11037 = 0.0094$.
- Thus $p_1 - p_2 = 0.0077$ and the estimated standard error,

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{0.0171 \times 0.9829}{11034} + \frac{0.0094 \times 0.9906}{11037}} = .0015.$$

Confidence Interval (p.19)

- A large sample $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z_{\alpha/2}\hat{\sigma}(p_1 - p_2)$$

  where $z_{\alpha/2}$ denotes the standard normal percentile having a right tail probability equals to $\alpha/2$.

- For the aspirin use example, a 95% C.I. for $\pi_1 - \pi_2$ is $0.0077 \pm 1.96 \times 0.0015 = (0.005, 0.011)$.

Notes (p.21)

- A difference between two proportions of a certain fixed size may have greater importance when both proportions are near 0 or 1 than when they are near the middle of the range.
- e.g. the difference between 0.010 and 0.001 is the same as the difference between 0.410 and 0.401, namely 0.009 but the former one may be more important than the later one.
- Examples of such cases include a comparison of drugs on the proportion of subjects who have adverse reactions when using the drug.

Relative Risk (p.21)

- In 2×2 tables, the relative risk is the ratio of the success probabilities for the two groups π_1/π_2.
- The proportions 0.010 and 0.001 has a relative risk of 10.0 whereas the proportions 0.410 and 0.401 have a relative risk 1.02.

Relative Risk - Continued

- Sample relative risk $= p_1/p_2$.
- Its distribution is heavily skewed and cannot approximated by normal distribution well unless the sample sizes are quite large.
- A large sample confidence interval is given by

$$\exp\left[\log(\frac{p_1}{p_2}) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_{1+}p_1} + \frac{1-p_2}{n_{2+}p_2}}\right]$$

Example: Aspirin Use and MI

- The sample relative risk is 1.818.
- A large sample 95% confidence interval for the relative risk π_1/π_2 is $[1.4330, 2.3059]$.
- The C.I. (0.005, 0.011) for the difference of proportions, $\pi_1 - \pi_2$, makes it seem as if the two groups differ by a trivial amount, but the relative risk shows that the difference may have important public health implications.

Odds Ratio (p.22)

- Within Row 1, the odds of success is

$$Odds_1 = \pi_1/(1 - \pi_1)$$

- Similarly, within Row 2, the odds of success is

$$Odds_2 = \pi_2/(1 - \pi_2)$$

- Odds Ratio

$$\theta = \frac{Odds_1}{Odds_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}$$

Notes (p. 23)

- For example, if $\pi_1 = 0.75, Odds_1 = 0.75/0.25 = 3$.
- Odds are non-negative and values greater than 1 indicates a $success$ is more likely than a $failure$.
- When X and Y are independent, conditional distributions of Rows 1 and 2 are same, that is, $\pi_1 = \pi_2$ and this implies, $\theta = 1$.

Observations

- X and Y are independent $\Leftrightarrow \pi_1 = \pi_2 \Leftrightarrow \theta = 1$.
- If $1 < \theta < \infty$, the odds of success are **higher** in row 1 than in row 2.
- If $0 < \theta < 1$, a success is **less** likely in row 1 than in row 2.

More Observations

- Values of θ farther from 1 (too small or too large) in a given direction indicates stronger level of association.
- If the order of the rows or the order of the columns is reversed (but not both), the new value of θ is the inverse of the original value.
- This ordering is usually arbitrary, so whether we get $\theta = 4.0$ or 0.25 is simply a matter of how we label the rows and columns.

More Observations

- As the odds ratio treats the variables symmetrically, it is unnecessary to identify one classification as a response variable to calculate it.
- When both variables are responses, the odds ratio can be defined using the joint probability as

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

  and called $cross - product\ ratio$.

Sample Odds Ratio (p.24)

- Sample odds ratio is defined as

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

- For the Aspirin Use example, for the Placebo group, the odds of $MI = 0.0174$ and for the Aspirin group, the odds of $MI = 0.0095$.
- The sample odds ratio $= 0.0174/0.0095 = 1.832$.
- The estimated odds are 83% higher for the placebo group.

SAS Codes For Binomial Proportions

```
data veg;
  input habit $ count;
datalines;
Veg 10
Nonveg 15
; run;
proc freq data=veg order=data;
  weight count;
  tables habit / binomial (p=0.5);
run;
```

Output

The FREQ Procedure

habit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Veg	10	40.00	10	40.00
Nonveg	15	60.00	25	100.00

Binomial Proportion for habit = Veg

Proportion	0.4000
ASE	0.0980

SAS Codes For Binomial Proportions

```
95% Lower Conf Limit                 0.2080
95% Upper Conf Limit                 0.5920
Exact Conf Limits
95% Lower Conf Limit                 0.2113
95% Upper Conf Limit                  0.6133
  Test of H0:  Proportion = 0.5
ASE under H0                         0.1000
Z                                   -1.0000
One-sided Pr <  Z                    0.1587
Two-sided Pr > |Z|                   0.3173
Sample Size = 25
```

SAS Codes For Binomial Proportions

```
data aspirin;
input group $ mi $ count;
datalines;
Placebo Yes 189
Placebo No 10845
Aspirin Yes 104
Aspirin No 10933
;
run;
proc freq data=aspirin order=data;
  weight count;
  tables group*mi / measures nopercent norow nocol;
run;
```

SAS Codes For Binomial Proportions

```
The FREQ Procedure
Table of group by mi
group   mi
```

Frequency	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Output

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.8321	1.4400	2.3308
Cohort (Col1 Risk)	1.8178	1.4330	2.3059
Cohort (Col2 Risk)	0.9922	0.9892	0.9953

Sample Size = 22071

R Codes For Binomial Proportions

- For asymptotic test:

```
>veg<-10
>total<-25
>prop.test(veg,total,0.5,correct=F)
```

- For Exact Test:

```
>binom.test(veg,total,0.5)
```

Output For `prop.test`

```
1-sample proportions test without continuity
  correction
data:  veg out of total, null probability 0.5
X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis:  true p is not equal to 0.5
95 percent confidence interval:
0.2340330 0.5926054
sample estimates:
p
0.4
```

Output For `binom.test`

```
                Exact binomial test

data:  veg and total
number of successes = 10, number of trials = 25,
p-value = 0.4244
alternative hypothesis:  true probability of
  success is not equal to 0.5
95 percent confidence interval:
0.2112548 0.6133465
sample estimates:
probability of success
     0.4
```

R Codes For 2×2 Tables

```
>phs<-matrix(c(189,10845,104,10933), ncol=2, byrow=2)
>dimnames(phs)<-list(Group=c("Placebo", "Aspirin"),
MI=c("Yes","No"))
>phs
      MI
Group      Yes   No
  Placebo   189   10845
  Aspirin   104   10933
```

R Codes For Difference In Proportions

```
>prop.test(phs,correct=F)
2-sample test for equality of proportions without
continuity correction
data:  phs
X-squared = 25.0139, df = 1, p-value = 5.692e-07
alternative hypothesis:  two.sided
95 percent confidence interval:
0.004687751 0.010724297
sample estimates:
prop 1 prop 2
0.01712887 0.00942285
```

Relative Risk and Odds Ratio

```
>phs.test<-prop.test(phs,correct=F)
>phs.test$estimate[1]/phs.test$estimate[2]
prop 1 %Relative Risk%
1.817802
>odds<-phs.test$estimate/(1- phs.test$estimate)
>odds[1]/odds[2] %Odds Ratio%
prop 1
1.832054
```