Contingency Tables Part One*

STA 312: Fall 2022

Suggested Reading: Chapter 2

- Read Sections 2.1-2.4
- You are not responsible for Section 2.5

Overview

Contents

1 Definitions

We are interested in relationships between variables

A *contingency table* is a joint frequency distribution.

	No Pneumonia	Pneumonia
No Vitamin C		
500 mg. or more Daily		

A contingency table

- Counts the number of cases in combinations of two (or more) categorical variables

*See last page for copyright information.

- In general, X has I categories and Y has J categories
- Often, X is the explanatory variable and Y is the response variable (like regression).

Cell probabilities π_{ij}
$i = 1, \ldots, I$ *and* $j = 1, \ldots, J$

Passed the Course

Course	Did not pass	Passed	Total
Catch-up	π_{11}	π_{12}	π_{1+}
Mainstream	π_{21}	π_{22}	π_{2+}
Elite	π_{31}	π_{32}	π_{3+}
Total	π_{+1}	π_{+2}	1

Marginal probabilities

- $Pr\{X = i\} = \sum_{j=1}^{J} \pi_{ij} = \pi_{i+}$
- $Pr\{Y = j\} = \sum_{i=1}^{I} \pi_{ij} = \pi_{+j}$

Conditional probabilities

$$Pr\{Y = j | X = i\} = \frac{Pr\{Y = j \cap X = i\}}{Pr\{X = i\}} = \frac{\pi_{ij}}{\pi_{i+}}$$

Passed the Course

Course	Did not pass	Passed	Total
Catch-up	π_{11}	π_{12}	π_{1+}
Mainstream	π_{21}	π_{22}	π_{2+}
Elite	π_{31}	π_{32}	π_{3+}
Total	π_{+1}	π_{+2}	1

- Usually, interest is in the conditional distribution of the response variable given the explanatory variable.
- Sometimes, we make tables of conditional probabilities

Cell frequencies

Passed the Course

Course	Did not pass	Passed	Total
Catch-up	n_{11}	n_{12}	n_{1+}
Mainstream	n_{21}	n_{22}	n_{2+}
Elite	n_{31}	n_{32}	n_{3+}
Total	n_{+1}	n_{+2}	n

For example

Passed the Course

Course	Did not pass	Passed	Total
Catch-up	27	8	35
Mainstream	124	204	328
Elite	7	24	31
Total	158	236	394

2 Study Designs and Models

Estimating probabilities

Should we just estimate π_{ij} with $p_{ij} = \frac{n_{ij}}{n}$?

- *Sometimes.*
- It depends on the study design.
- The study design determines exactly what is in the tables

Study designs

- Cross-sectional
- Prospective
- Retrospective

Cross-sectional design

- Both variables in the table are measured with
  - No assignment of cases to experimental conditions
  - No selection of cases based on variable values
- For example, a sample of n first-year university students sign up for one of three calculus courses, and each student either passes the course or does not.
- Total sample size n is fixed by the design.
- Multinomial model, with $c = IJ$ categories.
- Estimate π_{ij} with p_{ij}
- Estimating conditional probabilities is easy.

Prospective design

- Prospective means "looking forward" (from explanatory to response).
- Groups that define the explanatory variable categories are formed before the response variable is observed.
- Experimental studies with random assignment are prospective (clinical trials).
- Cohort studies that follow patients who got different types of surgery.
- Stratified sampling, like interview 200 people from each province.
- Marginal totals of the explanatory variable are fixed by the design.
- Assume random sampling within each category defined by the explanatory variable, and independence between categories.
- *Product multinomial* model: A product of I multinomial likelihoods.
- Good for estimating *conditional* probability of response given a value of the explanatory variable.

Product multinomial

- Take independent random samples of sizes $n_{1+}, \ldots, n_{I+}$ from I sub-populations.
- In each, observe a multinomial with J categories. Compare.
- Example: Sample 100 entring students from each of three campuses. At the end of first year, observe whether they are in good standing, on probation, or have left the university.
- The π_{ij} are now conditional probabilities: $\pi_{1+} = 1$
- Write the likelihood as

$$\prod_{i=1}^{3} [\pi_{i1}^{n_{i1}} \pi_{i2}^{n_{i2}} (1 - \pi_{i1} - \pi_{i2})^{n_{i3}})]$$

Retrospective design

- Retrospective means "looking backward" (from response to explanatory).
- In a *case control* study, a sample of patients with a disease is compared to a sample without the disease, to discover variables that might have caused the disease.
- Vitamin C and Pneumonia (fairly rare, even in the elderly)
- Marginal totals for the response variable are fixed by the design.
- Product multinomial again
- Natural for estimating conditional probability of explanatory variable given response variable.
- Usually that's not what you want.
- But if you know the probability of having the disease, you can use Bayes' Theorem to estimate the conditional probabilities in the more interesting direction.

3 Odds ratio

Meanings of X and Y "unrelated"

- Conditional distribution of $Y|X=x$ is the same for every x
- Conditional distribution of $X|Y=y$ is the same for every y
- X and Y are independent (if both are random)

If variables are not unrelated, call them "related."

Put probabilities in table cells

	$Y=1$	$Y=2$	Total
$X=1$	π_{11}	π_{12}	$\pi_{11}+\pi_{12}$
$X=2$	π_{21}	π_{22}	$\pi_{21}+\pi_{22}$
Total	$\pi_{11}+\pi_{21}$	$\pi_{12}+\pi_{22}$	

$$Pr\{Y=1|X=1\} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}}$$

Conditional distribution of Y given $X = x$
Same for all values of x

	$Y = 1$	$Y = 2$	Total
$X = 1$	π_{11}	π_{12}	$\pi_{11} + \pi_{12}$
$X = 2$	π_{21}	π_{22}	$\pi_{21} + \pi_{22}$
Total	$\pi_{11} + \pi_{21}$	$\pi_{12} + \pi_{22}$	

$$Pr\{Y = 1|X = 1\} = Pr\{Y = 1|X = 2\}$$

$$\begin{aligned}
&\Leftrightarrow \quad \frac{\pi_{11}}{\pi_{11} + \pi_{12}} = \frac{\pi_{21}}{\pi_{21} + \pi_{22}} \\
&\Leftrightarrow \quad \pi_{11}(\pi_{21} + \pi_{22}) = \pi_{21}(\pi_{11} + \pi_{12}) \\
&\Leftrightarrow \quad \pi_{11}\pi_{21} + \pi_{11}\pi_{22} = \pi_{11}\pi_{21} + \pi_{12}\pi_{21} \\
&\Leftrightarrow \quad \pi_{11}\pi_{22} = \pi_{12}\pi_{21} \\
&\Leftrightarrow \quad \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \theta = 1
\end{aligned}$$

Cross product ratio

	$Y = 1$	$Y = 2$	Total
$X = 1$	π_{11}	π_{12}	$\pi_{11} + \pi_{12}$
$X = 2$	π_{21}	π_{22}	$\pi_{21} + \pi_{22}$
Total	$\pi_{11} + \pi_{21}$	$\pi_{12} + \pi_{22}$	

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

Conditional distribution of X given $Y = y$
Same for all values of y

	$Y = 1$	$Y = 2$	Total
$X = 1$	π_{11}	π_{12}	$\pi_{11} + \pi_{12}$
$X = 2$	π_{21}	π_{22}	$\pi_{21} + \pi_{22}$
Total	$\pi_{11} + \pi_{21}$	$\pi_{12} + \pi_{22}$	

$$Pr\{X = 1|Y = 1\} = Pr\{X = 1|Y = 2\}$$

$$
\begin{aligned}
\Leftrightarrow \quad & \frac{\pi_{11}}{\pi_{11}+\pi_{21}} = \frac{\pi_{12}}{\pi_{12}+\pi_{22}} \\
\Leftrightarrow \quad & \pi_{11}(\pi_{12}+\pi_{22}) = \pi_{12}(\pi_{11}+\pi_{21}) \\
\Leftrightarrow \quad & \pi_{11}\pi_{12} + \pi_{11}\pi_{22} = \pi_{11}\pi_{12} + \pi_{12}\pi_{21} \\
\Leftrightarrow \quad & \pi_{11}\pi_{22} = \pi_{12}\pi_{21} \\
\Leftrightarrow \quad & \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \theta = 1
\end{aligned}
$$

X and Y independent

Meaningful in a cross-sectional design

Write the probability table as

$\boldsymbol{\pi} =$

x	$a-x$	a
$b-x$	$1-a-b+x$	$1-a$
b	$1-b$	1

Independence means $P(X=x, Y=y) = P(X=x)P(Y=y)$. If $x = ab$ then

$\boldsymbol{\pi} =$

ab	$a(1-b)$	a
$b(1-a)$	$(1-a)(1-b)$	$1-a$
b	$1-b$	1

And the cross-product ratio $\theta = 1$.

Conversely

x	$a-x$	a
$b-x$	$1-a-b+x$	$1-a$
b	$1-b$	1

If $\theta = 1$, then

$$
\begin{aligned}
& x(1-a-b+x) = (a-x)(b-x) \\
\Leftrightarrow \quad & x - ax - bx - x^2 = ab - ax - bx - x^2 \\
\Leftrightarrow \quad & x = ab
\end{aligned}
$$

Meaning X and Y are independent.

What we have learned about the cross-product ratio θ

- In a 2×2 table, $\theta = 1$ if and only if the variables are unrelated, no matter how "unrelated" is expressed.
  - Conditional distribution of $Y|X = x$ is the same for every x
  - Conditional distribution of $X|Y = y$ is the same for every y
  - X and Y are independent (if both are random)
- It's meaningful for all three study designs: Prospective, Retrospective and Cross-sectional.

Investigate θ a bit more.

Odds

Denoting the probability of an event by π,

$$\text{Odds} = \frac{\pi}{1-\pi}.$$

- Implicitly, we are saying the odds are $\frac{\pi}{1-\pi}$ "to one."
- if the probability of the event is $\pi = 2/3$, then the odds are $\frac{2/3}{1/3} = 2$, or two to one.
- Instead of saying the odds are 5 to 2, we'd say 2.5 to one.
- Instead of saying 1 to four, we'd say 0.25 to one.
- The higher the probability, the greater the odds.
- And as the probability of an event approaches one, the denominator of the odds approaches zero.
- This means the odds can be any non-negative number.

Odds ratio

- *Conditional Odds* is an idea that makes sense.
- Just use a conditional probability to calculate the odds.
- Consider the *ratio* of the odds of $Y = 1$ given $X = 1$ to the odds of $Y = 1$ given $X = 2$.

- Could say something like "The odds of cancer are 3.2 times as great for smokers."

$$\frac{\text{Odds}(Y=1|X=1)}{\text{Odds}(Y=1|X=2)} = \frac{P(Y=1|X=1)}{P(Y=2|X=1)} \Big/ \frac{P(Y=1|X=2)}{P(Y=2|X=2)}$$

Simplify the odds ratio

	$Y=1$	$Y=2$	Total
$X=1$	π_{11}	π_{12}	π_{1+}
$X=2$	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

$$\begin{aligned}
\frac{\text{Odds}(Y=1|X=1)}{\text{Odds}(Y=1|X=2)} &= \frac{P(Y=1|X=1)}{P(Y=2|X=1)} \Big/ \frac{P(Y=1|X=2)}{P(Y=2|X=2)} \\
&= \frac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}} \Big/ \frac{\pi_{21}/\pi_{2+}}{\pi_{22}/\pi_{2+}} \\
&= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \\
&= \theta
\end{aligned}$$

So the cross-product ratio is actually the odds ratio.

The cross-product ratio *is* the odds ratio

- When $\theta = 1$,
  - The odds of $Y=1$ given $X=1$ equal the odds of $Y=1$ given $X=2$.
  - This happens if and only if X and Y are unrelated.
  - Applies to all 3 study designs.
- If $\theta > 1$, the odds of $Y=1$ given $X=1$ are greater than the odds of $Y=1$ given $X=2$.
- If $\theta < 1$, the odds of $Y=1$ given $X=1$ are less than the odds of $Y=1$ given $X=2$.

Odds ratio applies to larger tables

	Admitted	Not Admitted
Dept. A	601	332
Dept. B	370	215
Dept. C	322	596
Dept. D	269	523
Dept. E	147	437
Dept. F	46	668

The (estimated) odds of being accepted are

$$\widehat{\theta} = \frac{(601)(668)}{(332)(46)} = 26.3$$

times as great in Department A, compared to Department F.

Some things to notice
About the odds ratio

- The cross-product (odds) ratio is meaningful for large tables; apply it to 2x2 sub-tables.
- Re-arrange rows and columns as desired to get the cell you want in the upper left position.
- Combining rows or columns (especially columns) by adding the frequencies is natural and valid.
- If you hear something like "Chances of death before age 50 are four times as great for smokers," most likely they are talking about an odds ratio.

4 Testing Independence

Testing independence with large samples
For cross-sectional data

Passed the Course

Course	Did not pass	Passed	Total
Catch-up	π_{11}	π_{12}	π_{1+}
Mainstream	π_{21}	π_{22}	π_{2+}
Elite	π_{31}	π_{32}	$1-\pi_{1+}-\pi_{2+}$
Total	π_{+1}	$1-\pi_{+1}$	1

Under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

- There are $(I-1)+(J-1)$ free parameters: The marginal probabilities.
- MLEs of marginal probabilities are $\widehat{\pi}_{i+} = p_{i+}$ and $\widehat{\pi}_{+j} = p_{+j}$
- Restricted MLEs are $\widehat{\pi}_{ij} = p_{i+}\,p_{+j}$
- The null hypothesis *reduces* the number of free parameters in the model by $(IJ-1)-(I-1+J-1) = (I-1)(J-1)$
- So the test has $(I-1)(J-1)$ degrees of freedom.

Estimated expected frequencies
Under the null hypothesis of independence

$$
\begin{aligned}
\widehat{\mu}_{ij} &= n\,\widehat{\pi}_{ij} \\
&= n\,\widehat{\pi}_{i+}\widehat{\pi}_{+j} \\
&= n\,p_{i+}p_{+j} \\
&= n\,\frac{n_{i+}}{n}\frac{n_{+j}}{n} \\
&= \frac{n_{i+}n_{+j}}{n}
\end{aligned}
$$

(Row total) × (Column total) ÷ (Total total)

Test statistics
For testing independence

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right) \qquad X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}}$$

With expected frequencies

$$\widehat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n} = \frac{\text{(Row total) (Column total)}}{\text{Total total}}$$

And degrees of freedom

$$df = (I-1)(J-1)$$

Copyright Information