STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

March 13, 2018

Contingency Tables

Class 16 - Case Study VI

Three approaches

Framingham Heart Study

A Project of the National Heart, Lung, and Blood Institute and Boston University

Ref: https://www.framinghamheartstudy.org/index.php

- Learning Objectives
  - Use 4 approaches to analyze Case Study VI data
  - Write out the models used and the assumptions for inference
  - Carry out the inference procedures completely
  - Interpret the respective R outputs

Case Study VI: Framingham Heart Study

- Background: In 1948, in Massachusetts, 5209 healthy mean and women, aged 30-60, were recruited and followed (their descendants are followed too) to examine risk factors for cardiovascular disease (CVD)
- Data considered:
  - $n = 1329$ men
  - X= Cholesterol measurement in 1948
  - Y= After 10 years, did they developed CVD?

X= Cholesterol level (mg/dl)	Y=CVD present	absent	row total
High (≥ 260)	41	245	286
Low (< 260)	51	992	1043
column total	92	1237	1329

$\hat{\mu}_{11} = \frac{286(92)}{1329}$

$\hat{\mu}_{12} = \frac{286(1237)}{1329}$

y_H

y_L

- Q: Is high cholesterol associated with increased risk of CVD?

Analysis I: Summary

Diff btw 2 props.

$H_0: \pi_H = \pi_L$, $H_a: \pi_H \neq \pi_L$

- For large samples, as in our case, proportions are normally distributed by the CLT.
- The test statistic under H_0 is approximately Normally distributed.

$Z \sim N(0,1)$

pnorm
qnorm

- Test Statistic= 5.575.
- p-value=$2P(Z \geq 5.575)$ is very small
- We have strong evidence that the probability of developing CVD is not the same for High and Low cholesterol groups
- Analysis I Approach: "Binomial sampling"
- Underlying distribution of outcome: Binomial

$Y_H \sim B(\quad)$ H

$Y_L \sim B$ L

Analysis II: Contingency Tables (2×2)

- Assume $n = 1329$ is fixed
- Classify the observations in 2 ways:
  1. Cholesterol status: H or L
  2. CVD status: present of absent
- Two categorical variables, each with 2 levels:
  1. C-cholesterol status
  2. D-disease status
- *In general, we have a row factor with I levels and a column factor with J levels*

Analysis II: Contingency Tables

Notation:

Chol CVD status

- Joint distribution of C and D:

$$P(C = i, D = j) = \pi_{ij}$$

- the probability that an observation falls into row i, column j, for $i = 1, \ldots, I$, $j = 1, \ldots, J$

$I = J = 2$

11	12
21	22

$\sum_{ij} \pi_{ij} = 1$

- Marginal distribution of C:

$$P(C = i) = \pi_{i\cdot}$$

$P(C=1) = \pi_{1\cdot}$
$P(C=2) = \pi_{2\cdot}$

$\sum_{i=1}^{2} \pi_{i\cdot} = 1$

- probability an observation falls into row i

- Marginal distribution of D:

$$P(D = j) = \pi_{\cdot j}$$

$\sum_{j=1}^{2} \pi_{\cdot j} = 1$

-probability an observation falls into column j

Analysis II: Contingency Tables

Hypotheses:

- $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ (There is no relationship between C and D)
- $H_a : \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$ (There is an association btw C & D.)

Analysis II: $I \times J$ Contingency Table

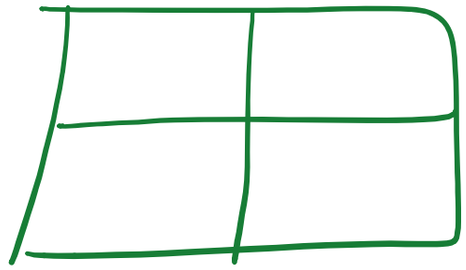
Observed cell counts, and row and column totals:

Row factor	Column factor 1	2	$\cdots$	J	row totals
1	y_{11}	y_{12}	$\cdots$	y_{1J}	$y_{1\cdot} = \sum_{j=1}^{J} y_{1j}$
2	y_{21}	y_{22}	$\cdots$	y_{2J}	$y_{2\cdot} = \sum_{j=1}^{J} y_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$y_{I\cdot} = \sum_{j=1}^{J} y_{Ij}$
col. totals	$\sum_{i=1}^{I} y_{i1}$	$\sum_{i=1}^{I} y_{i2}$	$\cdots$	$\sum_{i=1}^{I} y_{iJ}$	Grand$=\sum_j \sum_i y_{ij}$

Under H_0, we estimate the expected count, μ_{ij} for the (i,j)th cell as:

$$P(XY) = P(X)\,P(Y) \quad \text{if } X \perp Y$$

Analysis II: Test Statistic



Estimated expected cell count:

$$\hat{\mu}_{ij} = n \times \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}$$
$$= n\left(\frac{y_{i\cdot}}{n}\right)\left(\frac{y_{\cdot j}}{n}\right)$$
$$= \boxed{\frac{y_{i\cdot}y_{\cdot j}}{n}}$$

$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$

$\mu_{ij} = n\pi_{ij}$

$\pi_{ij} = \frac{\mu_{ij}}{n}$ (count)

Thus, our test statistic is:

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

(O, E)

$$\frac{(O-E)^2}{E} = \frac{(41-\hat{\mu}_{11})^2}{\hat{\mu}_{11}} + \frac{(245-\hat{\mu}_{12})^2}{\hat{\mu}_{12}} + \frac{(51-\hat{\mu}_{21})^2}{\hat{\mu}_{21}} + \frac{(992-\hat{\mu}_{22})^2}{\hat{\mu}_{22}}$$

Analysis II: Distribution of Test Statistic

- Under H_0, with large samples,

$$X^2 \dot{\sim} \chi^2_{df} \text{ with } df = (I-1)(J-1)$$

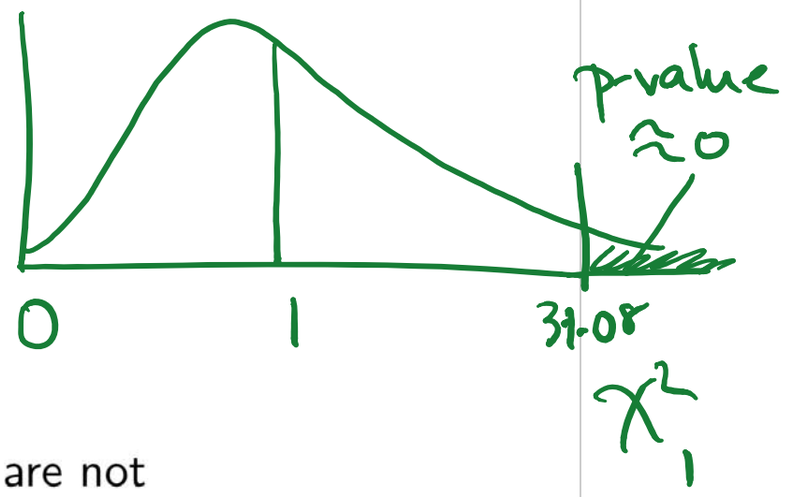
- df = # of cells - # of restrictions on df
- # of restrictions = # of estimates needed to compute T.S.
- To estimate each $\hat{\mu}_{ij}$, we need:
  - ith row total, $y_{i\cdot}$
  - jth column total, $y_{\cdot j}$
  - n
- The row and column totals add to n. Overall, we need:
  - $(I-1)$ row totals
  - $(J-1)$ column totals
- Therefore, $df = IJ - (I-1) - (J-1) - 1 = (I-1)(J-1)$

$4 - (2-1) - (2-1) - 1$

Most $\hat{\mu}_{ij} \geq 5$

Analysis II: R output

- From R output:
  - $X^2 = 31.08$ (a Chi-square statistic)
  - $df = (I-1)(J-1) = 1$ since $I = J = 2$
  - p-value< 0.0001
  - Conc: We have strong evidence that C and D are not independent; CVD status depends on cholesterol level



Equivalence between the 2 approaches

- In the case where $I = J = 2$, the Pearson chi-square test of independence is equivalent to comparing two proportions.
- Show the exact relationship between the test statistics for these two approaches, i.e., show that the chi-square statistic is equivalent to

$$\frac{n(y_{11}y_{22} - y_{21}y_{12})^2}{y_{1\cdot}y_{2\cdot}y_{\cdot 1}y_{\cdot 2}}$$

Analysis IIb: Formal approach based on MLEs and LRT

Notation: Let

- Y_{ij} be a random variable representing the number of observations in falls row i, column j of 2×2 contingency table, i.e., $I = J = 2$

Observe:

- y_{ij} - observed cell counts

Underlying distribution of $\mathbf{Y} = (Y_{11}, Y_{12}, Y_{11}, Y_{21}, Y_{22})$:

- Multinomial

$$P(\mathbf{Y} = \mathbf{y}) = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}\pi_{11}^{y_{11}}\pi_{12}^{y_{12}}\pi_{21}^{y_{21}}\pi_{22}^{y_{22}}$$

$\pi(1-\pi) = 1$

$P(y=y) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y}$

$\frac{n!}{y!(n-y)!}$ $P(S)$ $P(F)$.

$\sum_i \sum_j \pi_{ij} = 1$

Analysis IIb: Formal approach based on MLEs and LRT

Underlying distribution of $\mathbf{Y}$ is Multinomial$(n, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ where

- $\frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}$ = # of ways of arranging n observations so that y_{11} are in row 1, column 1 and so on
- $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$
- $y_{11} + y_{12} + y_{21} + y_{22} = n$

Multinomial coef:

$$\binom{n}{y_{11}\, y_{12}\, y_{21}\, y_{22}}$$

The log-likelihood is:

$$\log \mathcal{L} = \sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij} \log(\pi_{ij}) + \log \binom{n}{y_{11}y_{12}y_{21}y_{22}}$$

Analysis IIb: ML Estimation

- Maximize $\log \mathcal{L}$ w.r.t. $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ subject to $\sum\sum \pi_{ij} = 1$, we get:

$$\hat{\pi}_{ij} = \frac{y_{ij}}{n}$$

$\hat{\pi} = y/n$

- Under H_0 (independence), $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$:
  - Substitute $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ into $\log \mathcal{L}$
  - Maximize w.r.t. $\pi_{1\cdot}, \pi_{2\cdot}, \pi_{\cdot 1}, \pi_{\cdot 2}$ subject to the constraints $\pi_{1\cdot} + \pi_{2\cdot} = 1$ and $\pi_{\cdot 1} + \pi_{\cdot 2} = 1$

we get:

$$\hat{\pi}_{1\cdot} = \frac{y_{1\cdot}}{n} \qquad \hat{\pi}_{\cdot 1} = \frac{y_{\cdot 1}}{n}$$

$$\hat{\pi}_{1\cdot} = \frac{y_{2\cdot}}{n} \qquad \hat{\pi}_{\cdot 2} = \frac{y_{\cdot 2}}{n}$$

Then $\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}$ and this leads to the same expected counts as X^2.

Analysis IIb: LRT

- To compare multinomial model under assumption of independence ("REDUCED") to model without this assumption ("FULL")
- Test Statistic:

$$
\begin{aligned}
G^2 &= -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right) \\
&= 2\log\mathcal{L}_F - 2\log\mathcal{L}_R \\
&= 2\left\{\sum_j\sum_i y_{ij}\log\left(\frac{y_{ij}}{n}\right) - \sum_j\sum_i y_{ij}\log\left(\frac{y_{i\cdot}}{n}\frac{y_{\cdot j}}{n}\right)\right\} \\
&= 2\sum_j\sum_i y_{ij}\log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)
\end{aligned}
$$

Observed

Expected under H_0

Analysis IIb: Distribution of Test Statistic

- Under H_0,

$$G^2 \sim \chi^2_{df} \text{ with } df = (I-1)(J-1)$$

- $df = df(Unrestricted/FULL) - df(Independence/REDUCED)$
- Unrestricted model: df=# of parameters$(\pi_{ij}) = IJ - 1$
  - Lose 1 df due to constraint $\sum\sum \pi_{ij} = 1$
- Restricted model: df=# of parameters$(\pi_{i\cdot}, \pi'_{\cdot j}s) = I + J - 2$
  - Lose 2 df due to constraints $\sum_i \pi_{i\cdot} = 1$ and $\sum_j \pi_{\cdot j} = 1$
- Therefore, $df = IJ - 1 - (I + J - 2) = (I-1)(J-1)$

To do in R.

Analysis III: Fisher's Exact Test

		Y	
X	y_1	y_2	row.tot
x_1	a	b	a+b
x_2	c	d	c+d
col.tot	a+c	b+d	n

- A randomization (permutation) test; an exact test
- Appropriate for small sample size
- Assumes that row and column totals are fixed
- Null Hypothesis: Assume equal proportions or independence
- The Hypergeometric distribution is used to calculate the p-value.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

$H_0: \pi_H - \pi_L = 0$.

Analysis IV: Poisson Regression / Log-linear model

- Counts are NOT fixed
- Treat IJ counts as realizations of a Poisson random variable
- The joint distribution of cell counts is

$$P(\mathbf{Y} = \mathbf{y}) = \prod_j \prod_i \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!}$$

$y_{ij} \sim P(\mu_{ij})$.

$P(y_{ij} = y_{ij}) = \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}$

Analysis IV: Log-linear model hypotheses

H_0	*Null*
	Row and Column variables are independent Model with no interaction Additive model "REDUCED" model Eg. Mean # of persons with CVD does NOT dep. on chol. status
H_a	*Alternative*
	Row and column variables are dependent Model with interaction "FULL"/SATURATED model Eg. Mean # of persons with CVD dep. on chol. status

$H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0$

Note: *Row and column variables are treated symmetrically. In contrast, logistic models describe how a categorical response depends on the explanatory variable.*

Analysis IV: Comparing Models

Additive/REDUCED model:

- The probability of being in cell(i,j) is $\pi_{ij} = \pi_{i.}\pi_{.j}$
- Thus, the expected # of obs in each cell is $\mu_{ij} = n\pi_{i.}\pi_{.j}$

$$\log(\mu_{ij}) = \log n + \log \pi_{i.} + \log \pi_{.j}$$

$$\boxed{\log(\mu_{ij}) = \beta_0 + \beta_1 I_{[chol=H]} + \beta_2 I_{[CVD=absent]}}$$

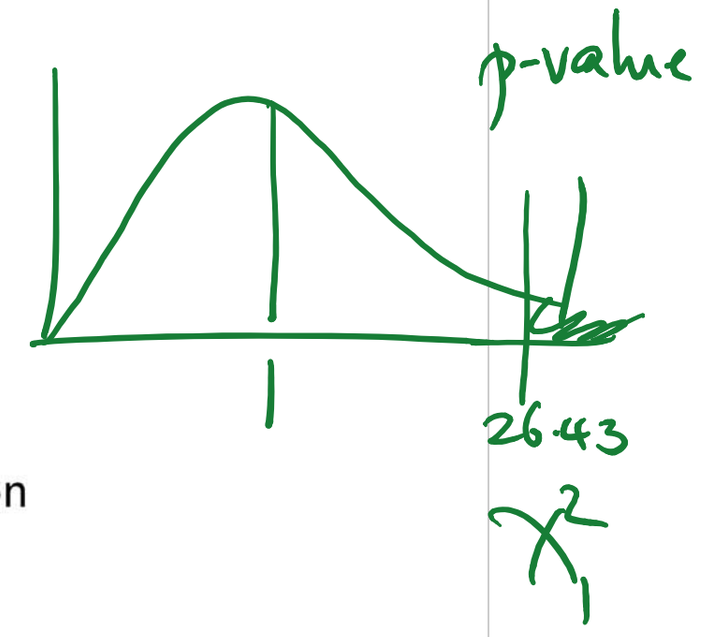
Interaction/SATURATED model:

- The expected # of obs in each cell is $\mu_{ij} = y_{ij}$
- Fits data perfectly. #parameters=# of observed counts=$I \times J$

$$\boxed{\log(\mu_{ij}) = \beta_0 + \beta_1 I_{[chol=H]} + \beta_2 I_{[CVD=absent]} + \beta_3 I_{[chol=H]} * I_{[CVD=absent]}}$$

Analysis IV: Summary of results

- $H_0 : \beta_3 = 0, H_a : \beta_3 \neq 0$
- Test statistic:
  - $G^2 = 26.4298$ follows a Chi-square distribution
  - $df = (I-1)(J-1) = 1$ since $I = J = 2$
- p-value< 0.0001
- Conc.: Strong evidence that CVD status depends on cholesterol status



Class 17 Summary

- Four Approaches:
  - Analysis I: Difference between 2 proportions
  - Analysis II: 2×2 contingency table
    - Pearson's Chi-square test of independence
    - Likelihood Ratio Test
  - Analysis III: Fisher's Exact Test
  - Analysis IV: Poisson regression/ Log-linear model
- R functions: `table()`, `prop.test()`, `chisq.test()`, `fisher.test()`, `glm()`
- Next: Extension to Three- way Tables

STA303/1002 - Class 17 R Markdown

March 13, 2018

Case Study VI: The CVD Data

```
cvd<-matrix(c(41,245,51,992), nrow=2,byrow=TRUE)
dimnames(cvd)<-list(c("High","Low"), c("Present","Absent"))
names(dimnames(cvd))<-c("Cholesterol","Cardio Vascular Disease")
cvd
```

```
##             Cardio Vascular Disease
## Cholesterol Present Absent
##        High      41    245
##        Low       51    992
```

Case Study VI: Difference of Proportions and Pearson's TOI

```
prop.test(cvd,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  cvd
## X-squared = 31.082, df = 1, p-value = 2.474e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05178874 0.13712972
## sample estimates:
##     prop 1     prop 2
## 0.14335664 0.04889741
```

AI

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}_c(1-\hat{\pi}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$Z^2$$

$$Z = \sqrt{31.082} = 5.575$$

```
chisq.test(cvd,correct=FALSE)
```

AII

```
##
##  Pearson's Chi-squared test
##
## data:  cvd
## X-squared = 31.082, df = 1, p-value = 2.474e-08
```

$$Z^2 = X^2 = \sum_i \sum_j \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

Case Study VI: Analysis III

```
fisher.test(cvd)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cvd
## p-value = 2.641e-07
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.050279 5.132098
## sample estimates:
## odds ratio
##   3.251597
```

$$OR = \frac{a/b}{c/d} = \frac{41/245}{51/992}$$

III $\langle H_0 : w_1/w_2 = 1$

$\Updownarrow$

$\pm$ $H_0 : \pi_1 - \pi_2 = 0$

II H_0: Independence btw X & Y.

Case Study VI: Analysis IV

```r
Count=c(41,245,51,992)
CVD=as.factor(c("Present","Absent","Present","Absent"))
CL=as.factor(c("High","High","Low","Low"))
llmod1=glm(Count~CL+CVD, family=poisson) # Additive
summary(llmod1)
```

```
##
## Call:
## glm(formula = Count ~ CL + CVD, family = poisson)
##
## Deviance Residuals:
##      1       2       3       4
##  4.158  -1.317  -2.635   0.678
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.58425    0.05960   93.69   <2e-16 ***
## CLLow        1.29386    0.06675   19.38   <2e-16 ***
## CVDPresent  -2.59866    0.10806  -24.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1658.18  on 3  degrees of freedom
## Residual deviance:   26.43  on 1  degrees of freedom
```

Case Study VI: Analysis IV

```r
llmod2=glm(Count~CL*CVD, family=poisson) #Saturated
summary(llmod2)
```

```
##
## Call:
## glm(formula = Count ~ CL * CVD, family = poisson)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       5.50126    0.06389  86.108  < 2e-16 ***
## CLLow             1.39846    0.07134  19.602  < 2e-16 ***
## CVDPresent       -1.78769    0.16874 -10.595  < 2e-16 ***
## CLLow:CVDPresent -1.18021    0.22156  -5.327 9.99e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.6582e+03  on 3  degrees of freedom
## Residual deviance: 3.1086e-15  on 0  degrees of freedom
## AIC: 35.406
##
## Number of Fisher Scoring iterations: 2
```

Case Study VI: Analysis IV

```
deviance(llmod1)
```

```
## [1] 26.42985
```

```
deviance(llmod2)
```

```
## [1] 3.108624e-15
```