STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

March 8, 2018

Case Study VI

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

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

- 2 factors
- each @ 2 levels.
- let one be Y and the other X.

$41/286 = \hat{p}_1$

$51/1043 = \hat{p}_2$

Analysis I: Difference between two proportions

Notation: Let

- π_H = the probability of CVD in men with high (H) cholesterol
- π_L = the probability of CVD in men with low (L) cholesterol

$y_H \sim Bin(\quad)$

$y_L \sim Bin(\quad)$.

Hypotheses:

- $H_0 : \pi_H = \pi_L$, $\quad \pi_H - \pi_L = \underbrace{0}_{D_0}$
- $H_a : \pi_H \neq \pi_L$

Test Statistic:

$$\frac{\hat{\pi}_H - \hat{\pi}_L - D_0}{SE(\hat{\pi}_H - \hat{\pi}_L)}$$

$$E(\hat{\pi}_H - \hat{\pi}_L) = \pi_H - \pi_L = 0 \text{ (under } H_0)$$

$$\frac{Est - E(Est)}{SE(Est)} \rightarrow Z$$

Analysis I: Difference between two proportions

Assumption: Depending on level of cholesterol, each person is a Bernoulli trial with chance of developing CVD as:

- π_H with $n_H = 286$ or
- π_L with $n_L = 1043$

Then, for fixed n_H and fixed n_L, the count of the number of people who develop CVD is:

- $y_H \sim$ Binomial $(n_H = 286, \pi_H)$ or
- $y_L \sim$ Binomial $(n_L = 1043, \pi_L)$

y_H H

y_L L

$$\boxed{\hat{\pi}_H = \frac{y_H}{n_H}}$$

$E(y_H) = n_H \pi_H$

$Var(y_H) = n_H \pi_H (1 - \pi_H)$.

Analysis I: Difference between two proportions

Binomial sampling

$$\mathrm{Var}\left(\frac{y_H}{n_H}\right) + \mathrm{Var}\left(\frac{y_L}{n_L}\right)$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\hat{\pi}_H - \hat{\pi}_L) &= \mathrm{Var}(\hat{\pi}_H) + \mathrm{Var}(\hat{\pi}_L) \\
&= \frac{n_H \pi_H (1 - \pi_H)}{n_H^2} + \frac{n_L \pi_L (1 - \pi_L)}{n_L^2} \\
&= \frac{\pi_H (1 - \pi_H)}{n_H} + \frac{\pi_L (1 - \pi_L)}{n_L}
\end{aligned}$$

Assume y_L indep of y_H

$\mathrm{Var}(aX + bY)$
$= a^2 \mathrm{Var}\, X + b^2 \mathrm{Var}\, Y$
$+ 2ab\, \mathrm{Cov}(X, Y)$.

π_H, π_L are parameters.

$\hat{\pi}_H$ $\hat{\pi}_L$

Analysis I: Difference between two proportions

Binomial sampling

- Estimates of population proportions:
  - $\hat{\pi}_H = \frac{41}{286}$, $\hat{\pi}_L = \frac{51}{1043}$
- Estimate of variance of difference:

$$\widehat{\text{Var}}(\hat{\pi}_H - \hat{\pi}_L) = \frac{\hat{\pi}_H(1-\hat{\pi}_H)}{n_H} + \frac{\hat{\pi}_L(1-\hat{\pi}_L)}{n_L}$$

- Note, under H_0, $\pi_H = \pi_L$, then:
  - $\hat{\pi}_{combined} = \dfrac{92}{1329}$ and

$$\widehat{SE}(\hat{\pi}_H - \hat{\pi}_L) = \sqrt{\frac{\frac{92}{1329}(1-\frac{92}{1329})}{286} + \frac{\frac{92}{1329}(1-\frac{92}{1329})}{1043}}$$

$$= \sqrt{\hat{\pi}_c(1-\hat{\pi}_c)\left(\frac{1}{n_H} + \frac{1}{n_L}\right)}$$

CI for $\pi_1 - \pi_2$:

$$\boxed{(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\, SE(\hat{\pi}_1 - \hat{\pi}_2)}$$

$\widehat{SE}(\hat{\pi}_1 - \hat{\pi}_2) =$ (1)

(1)

$\pi_H = \pi_L = \pi$

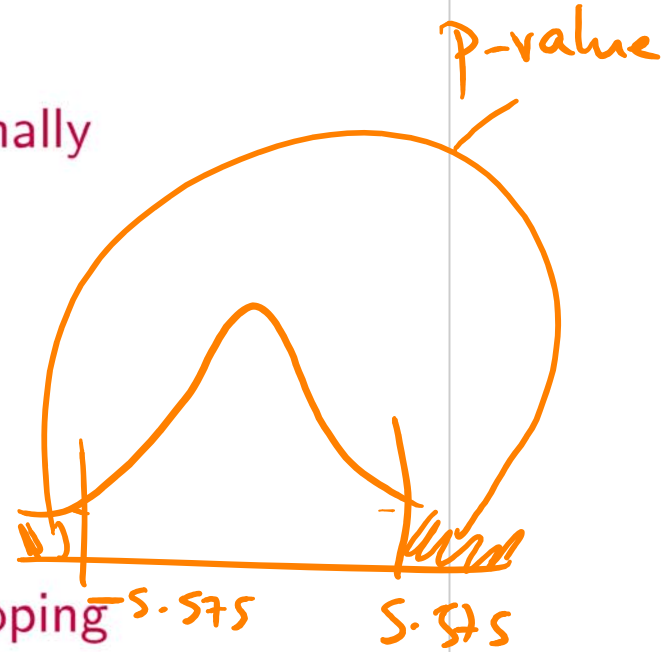
$\hat{\pi}$ – pooled estimate of π

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

... (2)

Analysis I: Summary

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



Analysis II: Contingency Tables

- 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:

- 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 $i = 1, \dots, I$, $j = 1, \dots, J$

- Marginal distribution of C:

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

  - probability an observation falls into row i

- Marginal distribution of D:

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

  -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)
- $H_a : \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$

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:

Analysis II: Test Statistic

Estimated expected cell count:

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

Thus, our test statistic is:

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

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, $y_{i.}$
  - jth column total, $y_{.j}$
  - n
- The row and column total add to n. Overall, we need:
  - $(I-1)$ row totals
  - $(J-1)$ column totals
- Therefore, $df = IJ - (I-1) - (J-1) - 1$

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

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
```

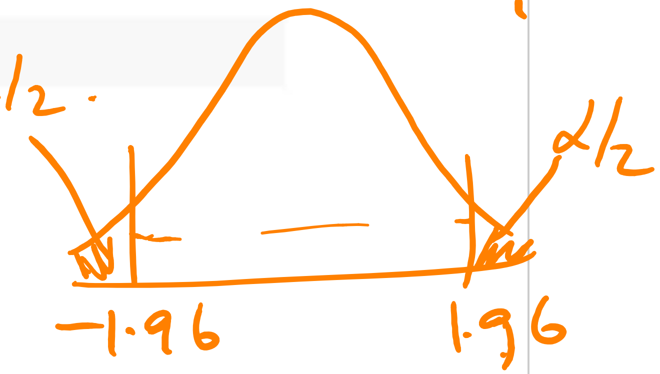
Case Study V: CI for difference of proportions

(*) $\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$

```
(p1.hat=41/(41+245)); (p2.hat=51/(51+992))
```

```
## [1] 0.1433566
```
— $\hat{p}_1$

```
## [1] 0.04889741
```
— $\hat{p}_2$




```
n1=41+245
n2=51+992
conf.level=0.95
(crit.val=qnorm(1-(1-conf.level)/2))
```

$\alpha = 0.05$ $\quad 1-\alpha = 0.95$

$z_{\alpha/2}$

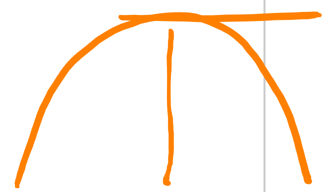
```
## [1] 1.959964
```

```
se.hat=sqrt(p1.hat*(1-p1.hat)/n1+p2.hat*(1-p2.hat)/n2)
c((p1.hat-p2.hat)-crit.val*se.hat,(p1.hat-p2.hat)+crit.val*se.hat)
```

$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p})}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

```
## [1] 0.05178874 0.13712972
```

(*) Two possible values for p_i: (1) Let $p_i = 1/2$
(2) Use $\hat{p}_i$



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

```r
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
```

Same as previous page; without continuity correction

$$Z^2 \sim \chi^2_1$$

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

$$Z \sim N(0,1).$$

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

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

Equivalent procedure (Analysis II)

Case Study VI: With continuity correction

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

(Not used in this course)

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  cvd
## X-squared = 29.633, df = 1, p-value = 5.221e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0495611 0.1393574
## sample estimates:
##     prop 1     prop 2
## 0.14335664 0.04889741
```

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

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cvd
## X-squared = 29.633, df = 1, p-value = 5.221e-08
```

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. (*Hint*: Show that the chi-square statistic is equivalent to

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

Class 16 Summary

- Four Approaches:
  - Analysis I: Difference between 2 proportions
  - Analysis II:
    - 2×2 contingency table
    - $I \times J$ contingency table
  - Analysis III: Fisher's Exact Test
  - Analysis IV: Poisson regression/ Log-linear model
- R functions: `table()`, `prop.test()`, `chisq.test()`
- Next: Analyses III & IV