More complicated tables

ST551 Lecture 24

Charlotte Wickham

2017-11-17

Homeworks

Where to look for comments.

Your Turn: Find all the problems in this statistical summary

against $H_0: \sigma^2 = 9$ $H_a: \sigma^2 > 9$

$H_0: \sigma^2 \le 9$

height for

not

inches^2

(These will all be point deductions!)

no

"There is ~~strong~~ evidence the variance of US males is 9 (t-test of variance, p-value = 0.27~~191099~~). It is estimated the variance of the height of US males is 9.467606. With 95% confidence, the variance in height of US males is between 8.633154 and 10.30206."

two sided

9.5

8.6

10.3

- no unit ← always state units
- written as evidence <u>for</u> the null ← against the null
- Too many decimal places ← estimates & CI's one more digit than measured
- Missing variable of interest ← variable and population should be clear

Two extensions to the 2×2 contingency tables

- More than two categories: Chi-square test
- More than two tables: Mantel-Haenszel Test

More than two categories

More than two categories

We might consider cross classifying our sample of N units on two variables that have more than two categories.

- Is eating breakfast associated with your commute method? 5×2 table

  # rows # cols

  $Y_i = \{$Ate breakfast, Didn't eat breakfast $\}$,

  $G_i = \{$Walk, Bike, Drove alone, Drove with others, Other $\}$
- Is your favorite sport associated with your favorite ice cream flavor? 3×5 table

  $Y_i = \{$Baseball, Basketball, Football, Soccer, Hockey $\}$,

  $G_i = \{$Chocolate, Strawberry, Vanilla $\}$

Chi-square test for $(r \times c)$ tables

Same as in 2×2 case, we can do a Chi-square test.

H_0 : No association between Variable 1 and Variable 2

O_{ij}: observed count in row i, column j

E_{ij}: expected count in row i, column j

$$E_{ij} = \frac{R_i C_j}{N}$$

$$X = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Under null hypothesis $X \dot{\sim} \chi^2_{(r-1)\times(c-1)}$. Reject for large X.

Example

"Table 2.5, from the 2000 General Social Survey, cross classifies gender and political party identification. Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents."

(Agresti 2007)

c=3 cols

r= 2 rows

	Democrat	Independent	Republican	Sum
F	762	327	468	1557
M	484	239	477	1200
Sum	1246	566	945	2757

?chisq.test

Example: Expected counts

	Democrat	Independent	Republican	Sum
F	703.7	319.6	533.7	$\boxed{1557}$
M	542.3	246.4	411.3	1200
Sum	1246	566	945	2757

$$703.7 = \frac{C_1 \times R_1}{N} = \frac{1557 \times 1246}{2756}$$

Example: Cell contributions

	Democrat		Independent		Republican
F	4.835	+	0.1692	+	8.084
M +	6.273	+	0.2196	+	10.49

```
## X-squared
## 30.07015
```

=

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

compare $\chi^2_{(r-1)\times(c-1)}$

$(2-1)\times(3-1) = 2$

Chi-squared test comments

Reference distribution is asymptotically exact.

Like 2×2 case, general rule of thumb: $E_{ij} > 5$ for all i, j.

2×2

More than two tables

Your Turn

Is party preference associated with level of education?

Find the sample odds ratio for these two states?

$$\hat{\omega} = \frac{ad}{bc}$$

Table 4: State 1 OR

education	democrat	rebublican
college	$a =$ 3	$b =$ 27
no college	$c =$ 7	$d =$ 63

$$\hat{\omega}_{\text{Oregon}} = \frac{3 \times 63}{7 \times 27} = 1$$

Table 5: State 2 WA

education	democrat	rebublican
college	$a_2 =$ 63	7
no college	27	3

$$\hat{\omega}_{\text{washington}} = \frac{63 \times 3}{27 \times 7} = 1$$

Your turn

Now combine two tables and find the odds ratio.

Table 6: Combined

education	democrat	rebublican
college	66	34
no college	34	66

$$\hat{\omega}_{combined} = \frac{66 \times 66}{34 \times 34} = 3.77$$

Simpson's paradox

"... in which a trend appears in different groups of data but disappears or reverses when these groups are combined."

https://en.wikipedia.org/wiki/Simpson%27s_paradox

The Mantel-Haenszel procedure attempts to avoid the paradox by combining the individual odds ratios (rather than collapsing the tables and computing a single odds ratio)

Mantel-Haenszel odds ratio

k tables, indexed by $j = 1, \ldots, k$.
k=2

Individual table odds ratio estimates:

$$\hat{\omega}_j = \frac{a_j d_j}{b_j c_j}$$

Combine in a weighted average:

$$\hat{\omega}_{MH} = \sum_{j=1}^{k} \text{weight}_j^* \times \hat{\omega}_j$$

where

$$\text{weight}_j^* = \frac{\text{weight}_j}{\sum \text{weight}_j} \quad \text{and } \text{weight}_j = \frac{b_j c_j}{N_j}$$

Your Turn

Find $\hat{\omega}_{MH}$ for the two tables:

Table 7: State 1

education	democrat	rebublican
college	3	27
no college	7	63

$\text{weight}_1 = \frac{27 \times 7}{100} = 1.89$

$\hat{\omega}_{OR} = 1$

Table 8: State 2

education	democrat	rebublican
college	63	7
no college	27	3

$\text{weight}_2 = \frac{7 \times 27}{100} = 1.89$

$\hat{\omega}_{WA} = 1$

$\text{weight}^*_1 = \frac{1.89}{1.89 + 1.89} = 0.5$

$\text{weight}^*_2 = 0.5$

$\hat{\omega}_{MH} = \text{weight}^*_1\, 1 + \text{weight}^*_2\, 1 = 1$

$= 0.5(1) + 0.5(1) = 1$

Mantel-Haenszel test

a_j	b
c	d

$H_0 : \underbrace{\omega_j = 1}$ for all $j = 1, \ldots, k$

test statistic $X = \dfrac{\left(\sum_{j=1}^{k} (a_j - E(a_j))\right)^2}{\sum_{j=1}^{k} Var(a_j)} = \dfrac{(}{\Sigma}$

$$E(a_j) = \frac{(R_{1j})(C_{1j})}{N_j}$$

$E(a_1)_k$

$E(a_2)$

$$Var(a_j) = \frac{R_{1j} C_{1j} R_{2j} C_{2j}}{N_j^2 (N_j - 1)}$$

Under the null hypothesis $X \dot{\sim} \chi_1^2$. Reject H_0 for large values of X.

In R

$x=$ $y=$

```
mantelhaen.test(df$education, df$party, z = df$state)
```

df education party state

1 College Democrat OR

```
##
##  Mantel-Haenszel chi-squared test without
##  continuity correction
##
## data:  df$education and df$party and df$state
## Mantel-Haenszel X-squared = 0, df = 1,
## p-value = 1
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3649129 2.7403803
## sample estimates:
## common odds ratio
##                  1
```

$P(Y_i = 0)$

$P(Y_i = 1)$

$\frac{ad}{bc}$

$3 \times 4 \times 3$

$1 / \frac{bc}{ad}$

$\mu_Y - \mu_X$

$\mu_X - \mu_Y$

Mantel-Haenszel Cautions

The test assumes the odds ratio is the same in all k tables.

- If this assumption is not met, it's difficult to interpret the p-value, and it doesn't make sense to estimate a common odds ratio.
- The test may fail to reject the null if the odds ratios are different from 1 but in opposite directions.