Class 4: Chapter 2 Contingency Tables

R section EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

Overview

1. Object:
   - Fisher's exact test
   - Cochran Mantel Haenszel statistics
   - Breslow-Day statistic
   - Making cross tables
2. R Example
3. Exercise

Fisher's exact test

Assuming the null hypothesis that men and women are equally likely to study, what is the probability that these 10 studiers would be so unevenly distributed between the women and the men?

	Men	Women	*Row Total*
Studying	***a***	***b***	*a + b*
Non-studying	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

Fisher's exact test

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

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

Cochran–Mantel–Haenszel statistics

In statistics, the Cochran-Mantel-Haenszel test (CMH) is a test used in the analysis of stratified or matched categorical data. We consider a binary outcome variable such as case status (e.g. lung cancer) and a binary predictor such as treatment status (e.g. smoking). The observations are grouped in strata. The stratified data are summarized in a series of 2×2 contingency tables, one for each strata. The ith such contingency table is:

	Treatment	No treatment	Row total
Case	A_i	B_i	N_{1i}
Controls	C_i	D_i	N_{2i}
Column total	M_{1i}	M_{2i}	T_i

The common odds-ratio of the K contingency tables is defined as:

$$R = \frac{\sum_{i=1}^{K} \frac{A_i D_i}{T_i}}{\sum_{i=1}^{K} \frac{B_i C_i}{T_i}},$$

Cochran–Mantel–Haenszel statistics

The null hypothesis is that there is no association between the treatment and the outcome. More precisely, the null hypothesis $H_0 : R = 1$ and the alternative hypothesis is $H_1 : R \neq 1$. The test statistic is:

$$\xi_{CMH} = \frac{\sum_{i=1}^{K}(A_i - \frac{N_{1i}M_{1i}}{T_i})^2}{\sum_{i=1}^{K}\frac{N_{1i}N_{2i}M_{1i}M_{2i}}{T_i^2(T_i-1)}}.$$

It follows a χ^2 distribution with degree of freedom $K-1$ asymptotically under the null hypothesis.

	Treatment	No treatment	Row total
Case	A_i	B_i	N_{1i}
Controls	C_i	D_i	N_{2i}
Column total	M_{1i}	M_{2i}	T_i

Contingency Coefficient

Phi Coefficient is a measure of association based on adjusting chi-square significance to factor out sample size. The range of it is between -1 and 1 for 2-by-2 tables, and is between 0 and min(sqrt(rows - 1), sqrt(columns - 1)). Computationally, phi is the square root of chi-square divided by n, the sample size. The phi coefficient is often used as a measure of association in 2-by-2 tables formed by true dichotomies.

Contingency Coefficient is an adjustment to phi coefficient, intended to adapt it to tables larger than 2-by-2. The contingency coefficient is computed as the square root of chi-square divided by chi-square plus n, the sample size. The contingency coefficient will be always less than 1 and will be approaching 1.0 only for large tables. The larger the contingency coefficient the stronger the association. Some researchers recommend it only for 5-by-5 tables or larger. For smaller tables it will underestimated the level of association.

Cramer's V is the most popular of the chi-square-based measures of nominal association because it is designed so that the attainable upper limit is always 1. Cramer's V equals the square root of chi-square divided by sample size, n, times m, which is the smaller of (rows - 1) or (columns - 1).

Breslow-Day statistic in homogeneity test of odds ratio

In Statistical Methods of Cancer Research; Volume 1 (https://www.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32.pdf) The analysis of case-control studies the authors Breslow and Day derive a statistic to test for the homogeneity of combining strata into an odds ratio (equation 4.30). Given the value of the statistic, the test determines if it is appropriate to combine strata together and compute a single odds ratio.

	Disease present	Disease absent	Totals
Risk factor present (success)	A	B	R1
Risk factor absent (failure)	C	D	R2
Totals	C1	C2	N

the odds ratio for getting a disease with a risk factor compared to not having the risk factor is:

$$\psi = (A * D)/(B * C)$$

Breslow-Day statistic in homogeneity test of odds ratio

if we have multiple contingency tables (for example, we stratify by age group), we can use the Mantel-Haenzel estimate to compute the odds ratio across all I strata:

$$\psi_{mh} = \frac{\sum_{i=1}^{I} A_i D_i / N_i}{\sum_{i=1}^{I} B_i C_i / N_i}.$$

For each contingency table we have $R1 = A + B, R2 = C + D$ and $C1 = A + C$, so we can express the expected odds ratio for that table in terms of the totals:

$$\psi_{mh} = \frac{AD}{BC} = \frac{\tilde{A}(R2 - C1 + \tilde{A})}{(R1 - \tilde{A})(C1 - \tilde{A})}$$

	Disease present	Disease absent	Totals
Risk factor present (success)	A	B	R1
Risk factor absent (failure)	C	D	R2
Totals	C1	C2	N

which gives a quadratic equation for $\tilde{A}$. Let a be the solution to this quadratic equation (only one root gives a reasonable answer).

Breslow-Day statistic in homogeneity test of odds ratio

Thus a reasonable test for the adequacy of the assumption of a common odds ratio is to sum up the squared deviation; of observed and fitted values, each standardized by its variance:

$$\chi^2 = \sum_{i=1}^{I} \frac{(a_i - A_i)^2}{V_i}$$

where the variance is:

$$V_i = \left(\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right)^{-1}$$

	Disease present	Disease absent	Totals
Risk factor present (success)	A	B	R1
Risk factor absent (failure)	C	D	R2
Totals	C1	C2	N

If the homogeneity assumption is valid, and the size of the sample is large relative to the number of strata, this statistic follows an approximate chi-square distribution on $I-1$ degrees of freedom and thus a p-value can be determined.

Breslow-Day statistic in homogeneity test of odds ratio

If instead we divide the I strata into H groups and we suspect the odds ratios are homogeneous within groups but not between them, Breslow and Day give an alternative statistic (equation 4.32):

$$\chi^2 = \sum_{h=1}^{H} \frac{\left(\sum_{i \in h} a_i - A_i\right)^2}{\sum_{i \in h} V_i}$$

	Disease present	Disease absent	Totals
Risk factor present (success)	A	B	R1
Risk factor absent (failure)	C	D	R2
Totals	C1	C2	N

where the i summations are over strata in the hth group with the statistic being chi-square with only $H - 1$ degrees of freedom (I assume a different Mantel-Haenzel estimate is computed within each group).

Sample size requirement–Mantel-Fleiss criterion

Peace, Karl E., ED., Statistical issues in drug research and development. Vol. 106. CRC Press, 1989. Page 52.

Halvorsen (1981) and Mehta and Patel (1983) have developed algorithms for calculating the exact significance levels of $r \times c$ tables that are computationally faster than previously proposed algorithms because they do not require total enumeration of the tables.

For a set of fourfold tables, such as

Treatment	Adverse experience Yes	No	
Experimental	n_{h11}	n_{h12}	n_{h1+}
Control	n_{h21}	n_{h22}	n_{h2+}
	n_{h+1}	n_{h+2}	n_h

for $h = 1, 2, \ldots, q$, the Mantel–Haenszel statistic may be appropriate even if the within-stratum sample sizes are small, as long as the combined stratum sample sizes

$$n_{+1+} = \sum_{h=1}^{q} n_{h1+} \quad \text{and} \quad n_{+2+} = \sum_{h=1}^{q} n_{h2+}$$

Sample size requirement–Mantel-Fleiss criterion

$$n_{+1+} = \sum_{h=1}^{q} n_{h1+} \quad \text{and} \quad n_{+2+} = \sum_{h=1}^{q} n_{h2+}$$

are sufficiently large. Mantel and Fleiss (1980) proposed the following criterion for the suitability of the Mantel–Haenszel procedure for a set of fourfold tables:

$$\min\left\{\left[\sum_{h=1}^{q} m_{h11} - \sum_{h=1}^{q} (n_{h11})_{\mathrm{L}}\right], \left[\sum_{h=1}^{q} (n_{h11})_{\mathrm{U}} - \sum_{h=1}^{q} m_{h11}\right]\right\} \geq 5$$

where $m_{h11} = n_{h1+}n_{h+1}/n_h$ is the expected value for n_{h11}, and $(n_{h11})_{\mathrm{L}}$ and $(n_{h11})_{\mathrm{U}}$ are respectively the lowest and the highest possible values for that cell, given that the marginals are fixed. Thus, the criterion requires that the potential variation in the across-strata sum of expected values for a particular cell should be at least 5.0. The criterion, of course, does not depend on which of the four cells is chosen. If the Mantel–Fleiss criterion is not met for a set of fourfold tables, an exact test may be carried out using algorithms like those of Thomas (1975) or Mehta, Patel, and Gray (1985).

R Code for analyzing contingency tables

display the data

```
library(vcdExtra)
data(GSS)
GSS
```

```
##       sex party count
## 1 female   dem   279
## 2   male   dem   165
## 3 female indep    73
## 4   male indep    47
## 5 female   rep   225
## 6   male   rep   191
```

R Code for analyzing contingency tables

display the cross table data: option I

```r
GSStab <- xtabs(count ~ sex + party, data = GSS)
GSStab
```

```
##         party
## sex      dem indep rep
##   female 279    73 225
##   male   165    47 191
```

```r
summary(GSStab)
```

```
## Call: xtabs(formula = count ~ sex + party, data = GSS)
## Number of cases in table: 980
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 7.01, df = 2, p-value = 0.03005
```

R Code for analyzing contingency tables

display the cross table data: option II

```
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 3.6.3
```

```
CrossTable(GSStab)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  980
##
##
##              | party
##          sex |       dem |     indep |       rep | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##       female |       279 |        73 |       225 |       577 |
##              |     1.183 |     0.078 |     1.622 |           |
##              |     0.484 |     0.127 |     0.390 |     0.589 |
##              |     0.628 |     0.608 |     0.541 |           |
##              |     0.285 |     0.074 |     0.230 |           |
## -------------|-----------|-----------|-----------|-----------|
##         male |       165 |        47 |       191 |       403 |
##              |     1.693 |     0.112 |     2.322 |           |
##              |     0.409 |     0.117 |     0.474 |     0.411 |
##              |     0.372 |     0.392 |     0.459 |           |
##              |     0.168 |     0.048 |     0.195 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       444 |       120 |       416 |       980 |
##              |     0.453 |     0.122 |     0.424 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
#CrossTable(GSStab,prop.t=FALSE,prop.r=FALSE,prop.c=FALSE)
```

Fisher Test

```
chisq.test(GSStab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  GSStab
## X-squared = 7.0095, df = 2, p-value = 0.03005
```

```
assocstats(GSStab)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 7.0026  2 0.030158
## Pearson          7.0095  2 0.030054
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.084
## Cramer's V        : 0.085
```

```
fisher.test(GSStab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  GSStab
## p-value = 0.03115
## alternative hypothesis: two.sided
```

UNIVERSITY OF MIAMI

Class 4: Chapter 2 Contingency Tables

Min Lu

Mantel Haenszel Test

```
## Agresti book (2007), p. 193.
## Job Satisfaction example.
Satisfaction <-
  as.table(array(c(1, 2, 0, 0, 3, 3, 1, 2,
                   11, 17, 8, 4, 2, 3, 5, 2,
                   1, 0, 0, 0, 1, 3, 0, 1,
                   2, 5, 7, 9, 1, 1, 3, 6),
                 dim = c(4, 4, 2),
                 dimnames =
                   list(Income =
                          c("<5000", "5000-15000",
                            "15000-25000", ">25000"),
                        "Job Satisfaction" =
                          c("V_D", "L_S", "M_S", "V_S"),
                        Gender = c("Female", "Male"))))
## (Satisfaction categories abbreviated for convenience.)
ftable(. ~ Gender + Income, Satisfaction)
```

```
##                     Job Satisfaction V_D L_S M_S V_S
## Gender Income
## Female <5000                           1   3  11   2
##        5000-15000                      2   3  17   3
##        15000-25000                     0   1   8   5
##        >25000                          0   2   4   2
## Male   <5000                           1   1   2   1
##        5000-15000                      0   3   5   1
##        15000-25000                     0   0   7   3
##        >25000                          0   1   9   6
```

```
## Table 6.12 in Agresti book, p. 193.
#mantelhaen.test(Satisfaction)
## See Table 6.13 in Agresti book, p. 196.
```

Mantel Haenszel Test

```
## Agresti book (2007), p. 193.
## Job Satisfaction example.
Satisfaction <-
  as.table(array(c(1, 2, 0, 0, 3, 3, 1, 2,
                   11, 17, 8, 4, 2, 3, 5, 2,
                   1, 0, 0, 0, 1, 3, 0, 1,
                   2, 5, 7, 9, 1, 1, 3, 6),
                 dim = c(4, 4, 2),
                 dimnames =
                   list(Income =
                          c("<5000", "5000-15000",
                            "15000-25000", ">25000"),
                        "Job Satisfaction" =
                          c("V_D", "L_S", "M_S", "V_S"),
                        Gender = c("Female", "Male"))))
## (Satisfaction categories abbreviated for convenience.)
#ftable(. ~ Gender + Income, Satisfaction)
## Table 6.12 in Agresti book, p. 193.
mantelhaen.test(Satisfaction)
```

```
##
##  Cochran-Mantel-Haenszel test
##
## data:  Satisfaction
## Cochran-Mantel-Haenszel M^2 = 10.2, df = 9, p-value = 0.3345
```

```
## See Table 6.13 in Agresti book, p. 196.
```

Breslow Day statistic

```
# SAS file /**** 3.3.2 Coronary artery exmaple ****/
library(metafor)
ai <- c(11,9)
bi <- c(4,9)
ci <- c(10,6)
di <- c(8,21)
res <- rma.mh(ai = ai, bi = bi, ci = ci, di = di, measure = "OR", correct = F)
res
```

```
##
## Fixed-Effects Model (k = 2)
##
## I^2 (total heterogeneity / total variability):  0.00%
## H^2 (total variability / sampling variability): 0.22
##
## Test for Heterogeneity:
## Q(df = 1) = 0.2151, p-val = 0.6428
##
## Model Results (log scale):
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   1.0462  0.4962  2.1086  0.0350  0.0737  2.0186
##
## Model Results (OR scale):
##
## estimate   ci.lb   ci.ub
##   2.8467  1.0765  7.5279
##
## Cochran-Mantel-Haenszel Test:    CMH = 4.5026, df = 1, p-val = 0.0338
## Tarone's Test for Heterogeneity: X^2 = 0.2152, df = 1, p-val = 0.6427
```

```
res$BD
```

```
## [1] 0.2154856
```

```
res$BDp
```

```
## [1] 0.6425014
```

In class exercise

Using the data below, get familiar with manipulating contingency table, and conduct Fisher's exact test to determine whether the data have equal distribution on hair and eye color.

dataset built in R

```
HairEyeColor
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

Take home exercise

Use the "HairEyeColor" data again, conduct Fisher's exact test to determine whether the data have equal distribution on gender and eye color.

UNIVERSITY OF MIAMI

Class over