Comparing Proportions

Data Analysis Using R (2017)

Wan Nor Arifin (wnarifin@usm.my), Universiti Sains Malaysia
Website: sites.google.com/site/wnarifin

Contents

1 Independent samples

1.1 Chi-squared test (for association)

```
lung_table = read.table(header = F, text = "
20  12
55 113
")
lung_table = as.table(as.matrix(lung_table))
dimnames(lung_table) = list(smoking = c("smoking", "no smoking"), cancer = c("lung cancer",
    "no lung cancer"))
str(lung_table)
```

```
##  'table' int [1:2, 1:2] 20 55 12 113
##  - attr(*, "dimnames")=List of 2
##   ..$ smoking: chr [1:2] "smoking" "no smoking"
##   ..$ cancer : chr [1:2] "lung cancer" "no lung cancer"
```

```
lung_table
```

```
##             cancer
## smoking      lung cancer no lung cancer
##   smoking             20             12
##   no smoking          55            113
```

```r
addmargins(lung_table)
```

```
##             cancer
## smoking      lung cancer no lung cancer Sum
##   smoking             20             12  32
##   no smoking          55            113 168
##   Sum                 75            125 200
```

```r
prop.test(lung_table)  # 2 x k
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  lung_table
## X-squared = 8.9286, df = 1, p-value = 0.002807
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.09688933 0.49834877
## sample estimates:
##   prop 1   prop 2
## 0.625000 0.327381
```

```r
chisq.test(lung_table)  # k x k
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lung_table
## X-squared = 8.9286, df = 1, p-value = 0.002807
```

```r
library(MASS)
smoke = data.frame(sex = survey$Sex, smoke = survey$Smoke)
str(smoke)
```

```
## 'data.frame':    237 obs. of  2 variables:
##  $ sex  : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
##  $ smoke: Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
```

```r
smoke = na.omit(smoke)
str(smoke)
```

```
## 'data.frame':    235 obs. of  2 variables:
##  $ sex  : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
##  $ smoke: Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:2] 70 137
##   .. ..- attr(*, "names")= chr [1:2] "70" "137"
```

```r
smoke_table = table(smoke)
smoke_table
```

```
##         smoke
## sex      Heavy Never Occas Regul
##   Female     5    99     9     5
##   Male       6    89    10    12
```

```r
chisq.test(smoke_table)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  smoke_table
## X-squared = 3.5536, df = 3, p-value = 0.3139
```

1.2 Fisher's exact test

```
fisher.test(lung_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  lung_table
## p-value = 0.002414
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.462825 8.225749
## sample estimates:
## odds ratio
##   3.401333
```

```
fisher.test(smoke_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  smoke_table
## p-value = 0.3105
## alternative hypothesis: two.sided
```

1.3 Chi-squared test for trend

- when grouping is ordinal

```
levels(smoke$smoke)
```

```
## [1] "Heavy" "Never" "Occas" "Regul"
```

```
smoke$smoke1 = factor(smoke$smoke, levels = c("Never", "Occas", "Regul", "Heavy"))
levels(smoke$smoke1)
```

```
## [1] "Never" "Occas" "Regul" "Heavy"
```

```
str(smoke)
```

```
## 'data.frame':    235 obs. of  3 variables:
##  $ sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
##  $ smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
##  $ smoke1: Factor w/ 4 levels "Never","Occas",..: 1 3 2 1 1 1 1 1 1 1 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:2] 70 137
##   .. ..- attr(*, "names")= chr [1:2] "70" "137"
```

```
table(smoke$smoke, smoke$smoke1)
```

```
##
##         Never Occas Regul Heavy
```

```
##   Heavy     0     0     0    11
##   Never   188     0     0     0
##   Occas     0    19     0     0
##   Regul     0     0    17     0
```

```
table(smoke$smoke == smoke$smoke1)
```

```
##
## TRUE
##  235
```

```
smoke_table1 = table(smoke = smoke$smoke1, sex = smoke$sex)
smoke_table1
```

```
##        sex
## smoke   Female Male
##   Never     99   89
##   Occas      9   10
##   Regul      5   12
##   Heavy      5    6
```

```
library(coin)
```

```
## Loading required package: survival
```

```
chisq_test(smoke_table1)  # common X2 test
```

```
##
##  Asymptotic Pearson Chi-Squared Test
##
## data:  sex by smoke (Never, Occas, Regul, Heavy)
## chi-squared = 3.5536, df = 3, p-value = 0.3139
```

```
chisq_test(smoke_table1, scores = list(smoke = 0:3))  # smoke ordinal
```

```
##
##  Asymptotic Linear-by-Linear Association Test
##
## data:  sex by smoke (Never < Occas < Regul < Heavy)
## Z = -1.4775, p-value = 0.1396
## alternative hypothesis: two.sided
```

2 Dependent samples

2.1 McNemar test

- outcome: 2
- repetition: 2

```
# --- PM rating (Agresti, pg409), n=1600
"Data:
            second
first        approve disapprove
  approve        794        150
  disapprove      86        570
"
```

```
## [1] "Data:\n            second\nfirst        approve disapprove\n  approve        794        150\n  d
```

```
pm_table = read.table(header = FALSE, text = "
794 150
86  570
")
pm_table = as.table(as.matrix(pm_table))
dimnames(pm_table) = list(first = c("approve", "disapprove"), second = c("approve", "disapprove"))
str(pm_table)
```

```
##  'table' int [1:2, 1:2] 794 86 150 570
##  - attr(*, "dimnames")=List of 2
##   ..$ first : chr [1:2] "approve" "disapprove"
##   ..$ second: chr [1:2] "approve" "disapprove"
```

```
pm_table
```

```
##             second
## first        approve disapprove
##   approve        794        150
##   disapprove      86        570
```

```
addmargins(pm_table)  # view marginal counts
```

```
##             second
## first        approve disapprove  Sum
##   approve        794        150  944
##   disapprove      86        570  656
##   Sum            880        720 1600
```

```
mcnemar.test(pm_table)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  pm_table
## McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

2.2 Cochran's Q test

- outcome: 2
- repetition: > 2

```
lect = read.csv("lect.csv")  # student's understanding by lecturer
str(lect)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ student      : int  1 1 1 2 2 2 3 3 3 4 ...
##  $ lecturer     : int  1 2 3 1 2 3 1 2 3 1 ...
##  $ understanding: int  1 0 0 0 0 0 0 0 0 0 ...
```

```
lect
```

```
##    student lecturer understanding
## 1        1        1             1
## 2        1        2             0
## 3        1        3             0
## 4        2        1             0
```

```
## 5        2        2             0
## 6        2        3             0
## 7        3        1             0
## 8        3        2             0
## 9        3        3             0
## 10       4        1             0
## 11       4        2             0
## 12       4        3             1
## 13       5        1             1
## 14       5        2             1
## 15       5        3             1
## 16       6        1             0
## 17       6        2             0
## 18       6        3             1
## 19       7        1             0
## 20       7        2             0
## 21       7        3             1
## 22       8        1             1
## 23       8        2             1
## 24       8        3             0
## 25       9        1             0
## 26       9        2             0
## 27       9        3             0
## 28      10        1             1
## 29      10        2             1
## 30      10        3             0
```

```
lect = as.data.frame(lapply(lect, factor))  # have to factor
str(lect)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ student      : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ lecturer     : Factor w/ 3 levels "1","2","3": 1 2 3 1 2 3 1 2 3 1 ...
##  $ understanding: Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
```

```
lect
```

```
##    student lecturer understanding
## 1        1        1             1
## 2        1        2             0
## 3        1        3             0
## 4        2        1             0
## 5        2        2             0
## 6        2        3             0
## 7        3        1             0
## 8        3        2             0
## 9        3        3             0
## 10       4        1             0
## 11       4        2             0
## 12       4        3             1
## 13       5        1             1
## 14       5        2             1
## 15       5        3             1
## 16       6        1             0
## 17       6        2             0
## 18       6        3             1
```

```
## 19       7        1             0
## 20       7        2             0
## 21       7        3             1
## 22       8        1             1
## 23       8        2             1
## 24       8        3             0
## 25       9        1             0
## 26       9        2             0
## 27       9        3             0
## 28      10        1             1
## 29      10        2             1
## 30      10        3             0
```

```
library(coin)
mh_test(understanding ~ lecturer | student, data = lect)
```

```
##
##  Asymptotic Marginal Homogeneity Test
##
## data:  understanding by lecturer (1, 2, 3)
##   stratified by student
## chi-squared = 0.33333, df = 2, p-value = 0.8465
```

2.3 Stuart-Maxwell test

- outcome: 2, > 2
- repetition: 2

```
# My stats lecture understanding level, n=200
"Data:
              after.lecture
before.lecture confused so-so understand
    confused         12     8         80
    so-so            10    10         20
    understand        5     8         47
"
```

```
## [1] "Data:\n              after.lecture\nbefore.lecture confused so-so understand\n    confused
```

```
stats_table = read.table(header = FALSE, text = "
12  8 80
10 10 20
 5  8 47
")
stats_table = as.table(as.matrix(stats_table))
dimnames(stats_table) = list(before.lecture = c("confused", "so-so", "understand"), after.lecture = c("
    "so-so", "understand"))
str(stats_table)
```

```
##  'table' int [1:3, 1:3] 12 10 5 8 10 8 80 20 47
##  - attr(*, "dimnames")=List of 2
##   ..$ before.lecture: chr [1:3] "confused" "so-so" "understand"
##   ..$ after.lecture : chr [1:3] "confused" "so-so" "understand"
```

```
stats_table
```

```
##               after.lecture
## before.lecture confused so-so understand
##     confused          12     8         80
##     so-so             10    10         20
##     understand         5     8         47
```

```
addmargins(stats_table)  # view marginal counts
```

```
##               after.lecture
## before.lecture confused so-so understand Sum
##     confused          12     8         80 100
##     so-so             10    10         20  40
##     understand         5     8         47  60
##     Sum               27    26        147 200
```

```
mh_test(stats_table)  # as nominal
```

```
##
##  Asymptotic Marginal Homogeneity Test
##
## data:  response by
##   conditions (before.lecture, after.lecture)
##   stratified by block
## chi-squared = 68.444, df = 2, p-value = 1.332e-15
```

```
mh_test(stats_table, scores = list(response = 1:3))  # as ordinal
```

```
##
##  Asymptotic Marginal Homogeneity Test for Ordered Data
##
## data:  response (ordered) by
##   conditions (before.lecture, after.lecture)
##   stratified by block
## Z = -8.1438, p-value = 3.831e-16
## alternative hypothesis: two.sided
```

References

Hothorn, T., Hornik, K., van de Wiel, M. A., Winell, H., & Zeileis, A. (2017). *Coin: Conditional inference procedures in a permutation test framework*. Retrieved from https://CRAN.R-project.org/package=coin

Ripley, B. (2017). *MASS: Support functions and datasets for venables and ripley's mass*. Retrieved from https://CRAN.R-project.org/package=MASS