Hw4

Mohammad Mottahedi

February 6, 2016

```
## Loading required package: grid
```

1. (30 pts) In a study, 1398 randomly-selected children of ages 0-15 were classified according to whether they carried Streptococcus pyogenes and according to the size of their tonsils.

Tonsil size

	Normal	Slightly enlarged	Very enlarged
Carrier	19	29	24
Non-carrier	497	560	269

(a) Analyze this table in an appropriate manner assuming that we are dealing with nominal variables, and report relevant statistics (e.g., X2 and/or G2) and your conclusion.

```
c.table <- array(data = c(19, 497, 29, 560, 24, 269),
    dim = c(2,3), dimnames = list(c("carrier", "non-carrier"),
                                Tonsil_size = c("normal", "slightly enlarge", "very enlarge")))
c.table
```

```
##              Tonsil_size
##               normal slightly enlarge very enlarge
##   carrier         19               29           24
##   non-carrier    497              560          269
```

```
pi.hat.table <- c.table/rowSums(c.table)
pi.hat.table
```

```
##              Tonsil_size
##                  normal slightly enlarge very enlarge
##   carrier     0.2638889        0.4027778    0.3333333
##   non-carrier 0.3748115        0.4223228    0.2028658
```

```
result <- chisq.test(c.table, correct=F)
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  c.table
## X-squared = 7.8848, df = 2, p-value = 0.0194
```

```r
G2 <- 2 * sum(c.table * log(c.table / result$expected))
G2
```

```
## [1] 7.320928
```

```r
# p-value for liklihood ratio
1 - pchisq(G2, 2)
```

```
## [1] 0.02572057
```

both χ^2 and G^2 are large with p-value less 0.05. we can reject the null hypothesis and say the variables are not independent.

(b) How many measures of association are needed to describe the table's departure from independence? Provide estimates and intervals for the relative risks, and interpret the results. Why are you able to make inference about relative risks?

We need $(I-1)(J-1) = 2$ measure of association to describe the tables departure from independence.

Assuming being the carrier is the response and size of tonsil is the predictor. To summarize the relationship between the two variables we can calculate the relative risk of carriers between different tonsil sizes taking normal as the base.

```r
c.table <- array(data = c(19, 497, 29, 560, 24, 269),
    dim = c(2,3), dimnames = list(c("carrier", "non-carrier"),
                                  Tonsil_size = c("normal", "slightly enlarge", "very enlarge")))

p.hat.row1 <- c.table[1,] / colSums(c.table)
p.hat.row2 <- c.table[2,] / colSums(c.table)

table.1 <- array(data = c(p.hat.row1[1],p.hat.row2[1],p.hat.row1[2],p.hat.row2[2]),
    dim = c(2,2), dimnames = list(c("carrier", "non-carrier"),
                                  Tonsil_size = c("normal", "slightly enlarge")))
table.1
```

```
##              Tonsil_size
##                   normal slightly enlarge
##   carrier     0.03682171       0.04923599
##   non-carrier 0.96317829       0.95076401
```

```r
rho1 <- (table.1[1,2]) / (table.1[1,1])
V_logr1 <-  (1-table.1[1,2])/((19+29)*table.1[1,2]) + (1-table.1[1,1])/((497+560)*table.1[1,1])
ci_low1 <- log(rho1) - 1.96 * sqrt(V_logr1)
ci_high1 <- log(rho1) + 1.96 * sqrt(V_logr1)
```

relative risk: $\rho = \frac{P(carrier|slightlyenlarge)}{P(carrier|normal)} = 1.3371459$

$CI = (0.3714658, 4.8132536)$

```r
table.2 <- array(data = c(p.hat.row1[1],p.hat.row2[1],p.hat.row1[3],p.hat.row2[3]),
    dim = c(2,2), dimnames = list(c("carrier", "non-carrier"),
                                  Tonsil_size = c("normal", "very enlarge")))
table.2
```

```
##              Tonsil_size
##                   normal very enlarge
##   carrier     0.03682171   0.08191126
##   non-carrier 0.96317829   0.91808874
```

```r
rho2 <- (table.2[1,2]) / (table.2[1,1])
V_logr2 <-  (1-table.2[1,2])/((19+24)*table.2[1,2]) +(1-table.2[1,1])/((497+269)*table.2[1,1])
ci_low2 <- log(rho2) - 1.96 * sqrt(V_logr2)
ci_high2 <- log(rho2) + 1.96 * sqrt(V_logr2)
```

relative risk: $\rho = \frac{P(carrier|veryenlarge)}{P(carrier|normal)} = 2.2245375$

$CI = (0.7674697, 6.4478987)$

risk being carrier increases with increasing tonsil size.

(c) Find an appropriate partitioning of the total departure from independence, as measured by deviance (G2), for this problem. Give the sub-tables, and show that your partitioning works. Did you learn anything more, inference wise, in comparison to part (a).

we need $(I-1)(J-1) = 2$ partitions.

```r
#first partition
c.table[,c(2,3)]
```

```
##              Tonsil_size
##               slightly enlarge very enlarge
##   carrier                   29           24
##   non-carrier              560          269
```

```r
result <- chisq.test(c.table[,c(2,3)], correct=F)

G2.1 <- 2 * sum(result$observed * log(result$observed/ result$expected))
G2.1
```

```
## [1] 3.537296
```

```r
# p-value for liklihood ratio
1 - pchisq(G2.1, 1)
```

```
## [1] 0.06000322
```

```r
#second partition
part.2 <- array(c(c.table[,1],c.table[,2] + c.table[,3]), dim=c(2,2))
part.2
```

```
##      [,1] [,2]
## [1,]   19   53
## [2,]  497  829
```

```r
result <- chisq.test(part.2, correct=F)

G2.2 <- 2 * sum(result$observed * log(result$observed / result$expected))
G2.2
```

```
## [1] 3.783633
```

```r
# p-value for liklihood ratio
1 - pchisq(G2.2, 1)
```

```
## [1] 0.05175618
```

```r
#sum of two G2
G2.1 + G2.2
```

```
## [1] 7.320928
```

sum of the G^2 form sub-tables add up to sum of G^2 from part a so partitioning works. Looking at the G^2 calculated for the two sub-tables we can't reject the null hypothesis the variables seems independent from each other.

(d) Now consider 'tonsil size' to be an ordinal variable. Re-run the analysis, if necessary, and report the relevant statistics, your conclusions and compare to what you got in parts (a) and (b). If you think it's not necessary to re-run the analysis, then explain why that's the case.

Calculation Pearson and Spearman correlation:

```r
#pearson
pears.res <- pears.cor(c.table, c(1,2), c(1,2,3))
pears.res
```

```
## [1] -0.07172885  7.18760456
```

```r
1 - pchisq(pears.res[2], 1)
```

```
## [1] 0.007340892
```

```r
#spearman
spear.res <- spear.cor(c.table)
spear.res
```

```
## [1] -0.0699344  6.8324769
```

```r
1 - pchisq(spear.res[2], 1)
```

```
## [1] 0.008951505
```

The M^2 value for both case is large and p-value is less than .05. We can reject the null hypothesis of independence and the two variables have a weak linear relationship.

2. (25 pts) Get a dataset from http://lib.stat.cmu.edu/DASL/Stories/EducationalAttainmentbyAge.html, by clicking on the link after Datafile Name and input the dataset into SAS or R or other software and perform the appropriate analysis. What interesting conclusions can you derive about relationship between age and educational attainment?

```
##                        Education Age_Group Count
## 1  Did not complete high school     25-34  5416
## 2  Did not complete high school     35-44  5030
## 3  Did not complete high school     45-54  5777
## 4  Did not complete high school     55-64  7606
## 5  Did not complete high school       >64 13746
## 6         Completed high school     25-34 16431
## 7         Completed high school     35-44  1855
## 8         Completed high school     45-54  9435
## 9         Completed high school     55-64  8795
## 10        Completed high school       >64  7558
## 11            College,1-3 years     25-34  8555
## 12            College,1-3 years     35-44  5576
## 13            College,1-3 years     45-54  3124
## 14            College,1-3 years     55-64  2524
## 15            College,1-3 years       >64  2503
## 16      College,4 or more years     25-34  9771
## 17      College,4 or more years     35-44  7596
## 18      College,4 or more years     45-54  3904
## 19      College,4 or more years     55-64  3109
## 20      College,4 or more years       >64  2483
```

```r
table <- xtabs(data$Count ~ data$Education + data$Age_Group)
table
```

```
##                                 data$Age_Group
## data$Education                  25-34 35-44 45-54 55-64   >64
##   College,1-3 years              8555  5576  3124  2524  2503
##   College,4 or more years        9771  7596  3904  3109  2483
##   Completed high school         16431  1855  9435  8795  7558
##   Did not complete high school  5416  5030  5777  7606 13746
```

```r
#Pearson
pears.res.2 <- pears.cor(table, c(3,4,2,1), c(1,2,3,4,5))
pears.res.2
```

```
## [1]    -0.2987507 11673.5350510
```

```r
1- pchisq(pears.res.2[2],1)
```

```
## [1] 0
```

```r
#Spearman
spear.res.2 <- spear.cor(table)
spear.res.2
```

```
## [1] 3.017687e-01 1.191058e+04
```

```r
1- pchisq(spear.res.2[2],1)
```

```
## [1] 0
```

there is strong correlation between education and age.

3. (30 pts) In 1972, a sample of 1,524 adults reported both their current religious affiliation and their religious affiliation at age 16.

(a) Is there any evidence of a change in the rate of Catholic affiliation over time? Find a confidence interval for the rate of change.

Mcnamar test: $H_0 : \pi_{12} = \pi_{21}$ $H_A : \pi_{12} \neq \pi_{21}$

```r
c.table.3 <- array(data = c(351, 33, 67, 1073),
    dim = c(2,2), dimnames = list(Affiliation.at.age.16=c("catholic", "non-Catholic"),
                                  Current_Affiliation = c("Catholic", "non-Catholic")))
c.table.3
```

```
##                        Current_Affiliation
## Affiliation.at.age.16 Catholic non-Catholic
##          catholic          351           67
##          non-Catholic       33         1073
```

```r
CrossTable(x=c.table.3, mcnemar=T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1524
##
```

```
##
##                        | Current_Affiliation
## Affiliation.at.age.16 |     Catholic | non-Catholic |    Row Total |
## ----------------------|--------------|--------------|--------------|
##              catholic |          351 |           67 |          418 |
##                       |      573.069 |      193.034 |              |
##                       |        0.840 |        0.160 |        0.274 |
##                       |        0.914 |        0.059 |              |
##                       |        0.230 |        0.044 |              |
## ----------------------|--------------|--------------|--------------|
##          non-Catholic |           33 |         1073 |         1106 |
##                       |      216.585 |       72.955 |              |
##                       |        0.030 |        0.970 |        0.726 |
##                       |        0.086 |        0.941 |              |
##                       |        0.022 |        0.704 |              |
## ----------------------|--------------|--------------|--------------|
##          Column Total |          384 |         1140 |         1524 |
##                       |        0.252 |        0.748 |              |
## ----------------------|--------------|--------------|--------------|
##
##
## McNemar's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  11.56     d.f. =  1     p =  0.0006738585
##
## McNemar's Chi-squared test with continuity correction
## ------------------------------------------------------------
## Chi^2 =  10.89     d.f. =  1     p =  0.0009668483
##
##
```

There's a strong evidence that rate of affiliation has changed.

```r
n <- sum(c.table.3)
d <- 67/sum(c.table.3) - 33/sum(c.table.3)

n12 <- c.table.3[1,2]
n21 <- c.table.3[2,1]

v_d <- 1/n * (n12/n * (1 - n12/n) + n21/n *(1 - n21/n) + 2 * n12 * n21 / n^2)
```

$\hat{d} = \pi_{12} - \pi_{21} = 0.0223097$

$\hat{V}(\hat{d}) = 4.2729052 \times 10^{-5}$

$CI = (0.022267, 0.0223524)$

Current affiliation

Affiliation at age 16	Catholic	Non-Catholic
Catholic	351	67
Non-Catholic	33	1073

(b) For these data, was it beneficial to record religious affiliation for the same individuals at both points in time? In other words, could we have done just as well if we had recorded current religious affiliation for 1,524 individuals, and religious affiliation at age 16 for a separate independent sample of 1,524 other individuals?

```
c.table.4 <- array(data = c(418, 384, 1106, 1140),
    dim = c(2,2), dimnames = list(Affiliation.at.age.16=c("catholic", "non-Catholic"),
                                  Current_Affiliation = c("Catholic", "non-Catholic")))
c.table.4
```

```
##                       Current_Affiliation
## Affiliation.at.age.16 Catholic non-Catholic
##          catholic          418         1106
##          non-Catholic      384         1140
```

```
res <-chisq.test(c.table.4, correct = F)
mcnemar.test(c.table.4, correct = F)
```

```
##
##  McNemar's Chi-squared test
##
## data:  c.table.4
## McNemar's chi-squared = 349.86, df = 1, p-value < 2.2e-16
```

```
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  c.table.4
## X-squared = 1.9561, df = 1, p-value = 0.1619
```

no , it's not beneficial the power of the cross-sectional study will have larger χ^2 value and we can reject the null hypothesis with greater power.

4. (15 pts) Calculate kappa for a 4×4 table having $n_{ii} = 5$ for all i, $n_{i,i+1} = 15, i = 1, 2, 3, n_{41} = 15$, and all other $n_{ij} = 0$. Explain why strong association does not imply strong agreement.

```
c.table.5 <- array(data = c(5,0,0,0,15,5,0,0,0,15,5,0,15,0,15,5),
    dim = c(4,4))
c.table.5
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    5   15    0   15
## [2,]    0    5   15    0
## [3,]    0    0    5   15
## [4,]    0    0    0    5
```

```
chisq.test(c.table.5)
```

```
## Warning in chisq.test(c.table.5): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  c.table.5
## X-squared = 63.98, df = 9, p-value = 2.278e-10
```

```
Kappa(c.table.5)
```

```
##               value     ASE     z  Pr(>|z|)
## Unweighted 0.08571 0.05095 1.682 9.251e-02
## Weighted   0.22995 0.04750 4.841 1.291e-06
```

It's possible to have strong negative association between two variables but in such a case there's no agreement between variables.