Class 3: Chapter 2

R section of EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

Overview

1. Object:
   - Odds ratio and Relative risk
   - Pearson chi-squared test
   - Test of trend for ordinal data
2. R Example
3. Exercise

Pearson chi-squared test

Pearson chi-squared test of goodness-of-fit of a set of multinomial probabilities: We begin with a sample of N items each of which has been observed to fall into one of k categories. We can define $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, as the observed numbers of items in each cell. Hence $\sum_{i=1}^{k} x_i = N$

Pearson χ^2 Test of multinomial probabilities

$$\text{Test } H_0 : \pi = (\pi_1, \pi_2, \ldots, \pi_k),\ where \sum_{i=1}^{k} \pi_i = 1$$

$$\text{through } \chi^2 = \sum_{i=1}^{k} \frac{(x_i - E_i)^2}{E_i},\ where\, E_i = N\pi_i$$

R Code for confidence interval of odds ratio

a — The number of individuals who both suffer from exposure and disease.

b — The number of individuals who suffer from disesase but not exposed.

c — The number of individuals who suffer from exposure but are healthy.

d — The number of individuals who neither suffered from exposure nor disease.

```
library(fmsb)
res <- oddsratio(a = 5, b = 10, c = 85, d = 80, conf.level = 0.95)
```

```
##            Disease Nondisease Total
## Exposed          5         85    90
## Nonexposed      10         80    90
## Total           15        165   180
```

```
res
```

```
##
##  Odds ratio estimate and its significance probability
##
## data:  5 10 85 80
## p-value = 0.1787
## 95 percent confidence interval:
##  0.1541455 1.4366513
## sample estimates:
## [1] 0.4705882
```

R Code for confidence interval of relative risk (1)

X — The number of disease occurence among exposed cohort.

Y — The number of disease occurence among non-exposed cohort.

m1 — The number of individuals in exposed cohort group.

m2 — The number of individuals in non-exposed cohort group.

```
library(fmsb)
res <- riskratio(X = 5, Y = 10, m1 = 90, m2 = 90, conf.level = 0.95)
```

```
##            Disease Nondisease Total
## Exposed          5         85    90
## Nonexposed      10         80    90
```

```
print(res)
```

```
##
##  Risk ratio estimate and its significance probability
##
## data:  5 10 90 90
## p-value = 0.1787
## 95 percent confidence interval:
##  0.1779702 1.4047292
## sample estimates:
## [1] 0.5
```

R Code for confidence interval of relative risk (2)

```
detach("package:fmsb", unload = TRUE)
library(epitools)
```

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

```
tapw <- c("Intermediate", "Highest")
outc <- c("Case", "Control")
dat <- matrix(c(2, 29, 35, 64), 2, 2, byrow = TRUE)
dimnames(dat) <- list(`Tap water exposure` = tapw, Outcome = outc)
riskratio(dat, rev = "c", correction = T)
```

```
## $data
##                    Outcome
## Tap water exposure Control Case Total
##       Intermediate      29    2    31
##       Highest           64   35    99
##       Total             93   37   130
##
## $measure
##                    risk ratio with 95% C.I.
## Tap water exposure estimate    lower    upper
##       Intermediate 1.000000       NA       NA
##       Highest      5.479798 1.397111 21.49306
##
## $p.value
##                    two-sided
## Tap water exposure  midp.exact fisher.exact chi.square
##       Intermediate          NA           NA         NA
##       Highest      0.001018658  0.001261178 0.00392597
##
## $correction
## [1] TRUE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

R Code for Pearson chi-squared test

- **The function chisq.test can perform the Pearson chi-squared test of goodness-of-fit of a set of multinomial probabilities. For example, with 3 categories and hypothesized values (0.4, 0.3, 0.3) and observed counts (12, 8, 10),**

```
x <- c(12, 8, 10)
p <- c(0.4, 0.3, 0.3)
chisq.test(x, p = p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 0.22222, df = 2, p-value = 0.8948
```

Test for trend in proportions

```
#x    Number of events
#n    Number of trials
#score  Group score
x <- c(0, 0.5, 1.5, 4, 7)

no <- c(17066, 14464, 788, 126, 37)
yes <- c(48, 38, 5, 1, 1)
patients <- no + yes

chiresult <- prop.trend.test(x = yes, n = patients, score = x)
chiresult
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  yes out of patients ,
##  using scores: 0 0.5 1.5 4 7
## X-squared = 6.5701, df = 1, p-value = 0.01037
```

```
# calculate r
r <- sqrt(chiresult$statistic/(sum(patients) - 1))
print(as.numeric(r))
```

```
## [1] 0.01420229
```

In class exercise

In **Canadian Journal of Sociology** 15 (1), 1990, *page 47*, Smith claimed that the sample showed a close match between the age distributions of women in the sample and all women in Toronto between the ages of 20 and 44. This is especially true in the youngest and oldest age brackets.

Tabel: *Sample and Census Age Distribution of Toronto Women.*

Age	Number in Sample	Percent in Census
20-24	103	18
25-34	216	50
35-44	171	32
Total	490	100

Using the data in Table 1, conduct a chi-square goodness of fit test to determine whether the sample does provide a good match to the known age distribution of Toronto women. Use the 0.05 level of significance.

Take home exercise

To test the hypothesis that a random sample of 100 students major in Public Health has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. There were 39 men in the sample and 61 women observed. Could we still conclude that the gender of students is equal in frequency?

Class over