Homework #4

Dustin Leatherman

May 4, 2019

10

Coffee Drinking and Sexual Habits. Find the p-value for the chi-squared test of independence in the coffee drinking and sexual habits data (Exercise 4)

```
        | Sex. Act |
        | Yes | No |
```

Coffee | 15 | 25 |
No Coffee | 115 | 70 |

```
coffee <- data.frame(Yes=c(15,115), No=c(25, 70), row.names = c("Coffee","No Coffee"))
test <-chisq.test(coffee, correct = F)

test$expected
```

```
##                 Yes       No
## Coffee     23.11111 16.88889
## No Coffee 106.88889 78.11111
```

```
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  coffee
## X-squared = 8.1999, df = 1, p-value = 0.004189
```

Expected values are greater than 5 so chi-square test is valid. The continuity correction is disabled since this is a 2 x 2 table.

p-value = 0.0042

11

Crab Parasites. Find the one-sided p-value for testing whether the odds of a parasite's being present are the same for Dungeness Crab as for Red Crab (Exercise 5), using Fisher's Exact Test.

(Notice that only one hypergeometric probability needs to be computed.)

```
     | Par. pres. |
     | Yes | No   |
```

Red | 5 | 312 |
Dunge. | 0 | 503 |

```
crabs <- data.frame(Yes=c(5,0), No=c(312, 503), row.names = c("Red","Dungeness"))
test <- fisher.test(crabs)

test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  crabs
## p-value = 0.008468
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.464693      Inf
## sample estimates:
## odds ratio
##        Inf
```

p-value = 0.00845

12

Hunting Strategy. For the hunting strategy data of Exercise 6, find the p-value for testing independence by

(a) the chi-squared test of independence, and

(b) Fisher's Exact Test.

```
| Prey sought    |
| Single | Mult. |
```

Alone | 17 | 12 |
Group | 14 | 7 |

```
hunt <- data.frame(Single=c(17,14), Multiple=c(12,7), row.names = c("Alone","Group"))
test.chi <- chisq.test(hunt, correct = F)

test.chi$expected
```

```
##       Single Multiple
## Alone  17.98    11.02
## Group  13.02     7.98
```

```
test.chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  hunt
## X-squared = 0.33468, df = 1, p-value = 0.5629
```

```r
test.fisher <- fisher.test(hunt)
test.fisher
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  hunt
## p-value = 0.7685
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1837663 2.6275878
## sample estimates:
## odds ratio
##  0.7132274
```

Chi-square p-value = 0.5629

fisher test p-value = 0.7685

13

Smoking and Lung Cancer. For the data in Section 18.1.3, test whether the odds of lung cancer for smokers are equal to the odds of lung cancer for nonsmokers, using

a. the excess statistic and
b. Fisher's Exact Test.

```
        | Outcome          |        |
        | Cancer | Control | Totals |
```

Smokers | 83 | 72 | 155 |
Non-smokers | 3 | 14 | 17 |
Totals | 86 | 86 | 172 |

a

```r
cancer <- data.frame(Cancer=c(83,3), Control=c(72,14), row.names = c("Smokers","NonSmokers"))
test.chi <- chisq.test(cancer, correct = F)

excess <- cancer[1,1] - test.chi$expected[1,1]
T <- colSums(cancer) %>% sum
var.excess <- ((colSums(cancer) * rowSums(cancer)) %>% prod) / (T * T * (T - 1))

z <- excess / sqrt(var.excess)

two.sided.p <- 2 * (1 - pnorm(abs(z)))
```

Excess Statistic = 3.8523

Z statistic = 2.8022

p-value = 0.0051

There is convincing evidence that the odds for contracting Lung Cancer is not equal between Smokers and Non-Smokers.

b

```r
fisher.test(cancer)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cancer
## p-value = 0.008823
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.409691 30.094245
## sample estimates:
## odds ratio
##   5.333256
```

There is convincing evidence that the odds for contracting Lung Cancer is not equal between Smokers and Non-Smokers (p-value = 0.0088).

14

Mantel–Haenszel Test for Censored Survival Times: Lymphoma and Radiation Data.

Central nervous system (CNS) lymphoma is a rare but deadly brain cancer; its victims usually die within one year. Patients at Oregon Health Sciences University all received chemotherapy administered after a new treatment designed to allow the drugs to penetrate the brain's barrier to large molecules carried in the blood. Of 30 patients as of December 1990, 17 had received only this chemotherapy; the other 13 had received radiation therapy prior to their arrival at OHSU. The table in Display 19.11 shows survival times, in months, for all patients. Asterisks refer to censored times; in these cases, the survival time is known to be at least as long as reported, but how much longer is not known, because the patient was still living when the study was reported.

The Mantel–Haenszel procedure can be used to construct a test for equality of survival patterns in the two groups. It involves constructing separate 2 X 2 tables for each month in which a death occurred, as shown in Display 19.12. The idea is to compare the numbers surviving and the numbers dying in the two groups, accounting for the length of time until death. The data are arranged in several 2 2 tables, as in Display 19.12. The censored observations are included in all tables for which the patients are known to be alive and are excluded from the others.

Compute the Mantel–Haenszel test statistic (equivalently called the log-rank statistic), and find the one-sided p-value. What can be inferred from the result?

Example

K = 3

| Survives | Dies |
Radiation | 11 | 2 |
No Radiation | 16 | 1 |

```
# Create a vector to transform into an array for input into the tests

months.array <-
  ex1914 %>%
  group_by(Months) %>%
  nest %>%
    mutate(
        A = map(data, ~ c(.$Survived, .$Died))
    ) %>%
    select(A) %>%
    unlist(.) %>%
    array(., dim = c(2,2, length(unique(ex1914$Months))))

BreslowDayTest(months.array)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  months.array
## X-squared = 17.589, df = 16, p-value = 0.3485
```

```
mantelhaen.test(months.array, correct=FALSE, alternative = "greater")
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  months.array
## Mantel-Haenszel X-squared = 3.1583, df = 1, p-value = 0.9622
## alternative hypothesis: true common odds ratio is greater than 1
## 95 percent confidence interval:
##  0.2041897       Inf
## sample estimates:
## common odds ratio
##         0.4405209
```

There is no evidence that the odds ratios are dependent between months after diagnosis (p-value = 0.3485).

There is no evidence that the odds ratio are greater for any number of months after diagnosis (p-value = 0.9622)

15

Bat Conservation International (BCI) is an organization that supports research on bats and encourages people to provide bat roost boxes (much like ordinary birdhouses) in backyards and gardens. BCI phoned people across Canada and the United States who were known to have put out boxes, to determine the sizes, shapes, and locations of the boxes and to identify whether the boxes were being used by bats. Among BCI's findings was the conclusion, "Bats were significantly more likely to move into houses during the first season if they were made of old wood. Of comparable houses first made available in spring and eventually occupied, eight houses made of old wood were all used in the first season, compared with only 32 of 65 houses made of new wood." (Data from "The Bat House Study," Bats 11(1) (Spring 1993): 4–11.)

The researchers required that, to be significant, a chi-square test meet the 5% level of significance.

(a) Is the chi-squared test appropriate for these data?

(b) Evaluate the data's significance, using Fisher's Exact Test.

| Move In | Not Move In |
Old Wood | 8 | 0 |
New Wood | 32 | 33 |

a

The Chi-square test is not an appropriate test to ascertain independence between the type of wood and whether or not a bat moved in. Some of the expected values from a chi-square test are less than 5 so it will not work.

b

```
bats <- data.frame(In=c(8,32), NotIn=c(0,33), row.names = c("Old Wood", "New Wood"))
fisher.test(bats)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  bats
## p-value = 0.006754
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.582328      Inf
## sample estimates:
## odds ratio
##        Inf
```

There is convincing evidence that the odds ratios between Type of Wood and Bat Houses occupied are not the same (p-value = 0.0068).

16

Pursuing an experiment similar to the one in Section 18.1.2, skeptics interviewed the same 800 subjects to determine who knew and who did not know to which group they had been assigned. Vitamin C, they argued, has a unique bitter taste, and those familiar with it could easily recognize whether their pills contained it. If so, the blind was broken, and the possibility arose that subjects' responses may have been influenced by the knowledge of their group status. The following table presents a fictitious view of how the resulting study might have turned out.

Use the Mantel–Haenszel procedure

(a) to test whether the odds of a cold are the same for the placebo and vitamin C users, after accounting for group awareness, and

(b) to estimate the common odds ratio.

a

There is no evidence that the odds ratios are dependent between placebo and Vitamin C users after accounting for group awareness (p-value = 0.8787).

There is no evidence that the odds ratios are different between placebo and Vitamin C users after accounting for group awareness (p-value = 0.3946).

```
knew <-
  ex1916 %>%
  group_by(Knew) %>%
  nest %>%
    mutate(
         A = map(data, ~ c(.$Cold, .$NoCold))
    ) %>%
    select(A) %>%
    unlist(.) %>%
    array(., dim = c(2,2, length(unique(ex1916$Knew))))

BreslowDayTest(knew)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  knew
## X-squared = 0.023293, df = 1, p-value = 0.8787
```

```
mantelhaen.test(knew, correct=FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  knew
## Mantel-Haenszel X-squared = 0.72478, df = 1, p-value = 0.3946
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.811711 1.696612
## sample estimates:
## common odds ratio
##          1.173524
```

b

The estimated common odds ratio is 1.1735.

17

A 1982–1986 study of women in Toronto, Ontario, assessed the added risk of breast cancer due to alcohol consumption. (Data from Rosenberg et al., "A Case–Control Study of Alcoholic Beverage Consumption and Breast Cancer," American Journal of Epidemiology 131 (1990): 6–14.) A sample of confirmed breast cancer patients at Princess Margaret Hospital who consented to participate in the study was obtained, as was a sample of cancer-free women from the same neighborhoods who were close in age to the cases.

The following tables show data only for women in two categories at the ends of the alcohol consumption scale: those who drank less than one alcoholic beverage per month (3 years before the interview), and those that drank at least one alcoholic beverage per day. The women are further categorized by their body mass, a possible confounding factor.

What evidence do these data offer that the risk of breast cancer is greater for heavy drinkers (1 drink/day) than for light drinkers (< 1 drink/month)?

```
cases <-
  ex1917 %>%
  group_by(BodyMass) %>%
  nest %>%
    mutate(
         A = map(data, ~ c(.$Cases, .$Controls))
    ) %>%
    select(A) %>%
    unlist(.) %>%
    array(., dim = c(2,2, length(unique(ex1917$BodyMass))))

BreslowDayTest(cases)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  cases
## X-squared = 6.4912, df = 2, p-value = 0.03895
```

```
cases[,,1] %>% fisher.test()
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 0.1174
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8810919 3.3523133
## sample estimates:
## odds ratio
##   1.709275
```

```
cases[,,2] %>% fisher.test()
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  .
## p-value = 0.1154
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.478183 1.081004
## sample estimates:
## odds ratio
##  0.7202373
```

```
cases[,,3] %>% fisher.test()
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 0.3811
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7036234 2.3884921
## sample estimates:
## odds ratio
##   1.299486
```

There is moderate evidence that there is a dependence between the odds ratio of Drinker classification and having Breast Cancer after controlling for Body Mass (p-value = 0.039). This bars a Mantel-Haentszel Test from being run. Each group can be analyzed individually to see if a parallel can be drawn. A Fisher Exact test for each Body Mass type yields no evidence that the odds ratios differ between Drinking Category and contracting Breast Cancer. The estimated odds ratio provided differs to some extent between each Body Mass so while there appears to be some effect that Body Mass has, it is negligble.

19

The table in Display 19.13 shows the numbers of compact sports utility vehicles involved in fatal accidents in the United States between 1995 and 1999, categorized according to travel speed, make of car (Ford or other), and cause of accident (tire-related or other).

From this table, whether the odds of a tire-related fatal accident depend on whether the sports utility vehicle is a Ford, after accounting for travel speed. For this subset of fatal accidents, estimate the excess number of Ford tire-related accidents. (This is a subset of data described more fully in Exercise 20.18.).

```
accidents <-
  ex1919 %>%
  group_by(SpeedCat) %>%
  nest %>%
    mutate(
         A = map(data, ~ c(.$Other, .$Tire))
    ) %>%
    select(A) %>%
    unlist(.) %>%
    array(., dim = c(2,2, length(unique(ex1919$SpeedCat))))
```

```r
accidents[,,1] %>% chisq.test(correct = F) %>% pluck("expected")
```

```
## Warning in chisq.test(., correct = F): Chi-squared approximation may be
## incorrect

##          [,1]      [,2]
## [1,] 170.4639 0.5360502
## [2,] 465.5361 1.4639498
```

```r
accidents[,,2] %>% chisq.test(correct = F) %>% pluck("expected")
```

```
## Warning in chisq.test(., correct = F): Chi-squared approximation may be
## incorrect

##          [,1]      [,2]
## [1,] 242.2703 0.7297297
## [2,] 753.7297 2.2702703
```

```r
accidents[,,3] %>% chisq.test(correct = F) %>% pluck("expected")
```

```
## Warning in chisq.test(., correct = F): Chi-squared approximation may be
## incorrect

##           [,1]     [,2]
## [1,]  99.58741 1.412587
## [2,] 323.41259 4.587413
```

```r
accidents[,,4] %>% chisq.test(correct = F) %>% pluck("expected")
```

```
##          [,1]     [,2]
## [1,] 119.8756  9.12439
## [2,] 261.1244 19.87561
```

```r
accidents[,,1] %>% fisher.test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.06867672        Inf
## sample estimates:
## odds ratio
##        Inf
```

```r
accidents[,,2] %>% fisher.test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1326379       Inf
## sample estimates:
## odds ratio
##        Inf
```

```r
accidents[,,3] %>% fisher.test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 0.1456
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.03988684 2.29526855
## sample estimates:
## odds ratio
##  0.3025972
```

```r
accidents[,,4] %>% fisher.test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 3.077e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.05626406 0.36922967
## sample estimates:
## odds ratio
##   0.151506
```

```r
# swapping columns for easier interpretation
cbind(accidents[,,4][,2], accidents[,,4][,1]) %>% fisher.test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 3.077e-06
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
##   2.708341 17.773337
## sample estimates:
## odds ratio
##   6.600401
```

Some of the expected counts for the cells are less than 5 so we cannot use any test to compare odds ratios between each speed zone. Per speed zone, Fisher's exact Test can be run to ascertain whether the odds ratio is equivalent between Make and Cause of Accident. For speed categories 0-40 mph, 41-55 mph, and 56-65 mp, there is no evidence that odds ratio is different between Ford Sport Utility Vehicles and Tire related accidents (p-value = 1, 1, 0.1456). For the Speed categorry > 66 mph, there is convincing evidence that the odds ratio for Ford Sport Utility Vehicles and Tire related accidents is not 0 (p-value = 3.077e-6).

It is estimated that Non-Ford Cars are 6.6 times more likely to get into Tire Related Accidents than Ford Sport Utility Vehicles.