Mid Term Exam

Ellen Chancey

October 6, 2018

Question One

A 55 year old Kansas woman recently received her annual mammogram, and the results of this screening test indicated the presence of breast cancer. The radiologist reading the film has a sensitivity rate for screening mammography of 85% and a specificity rate of 97%. If 1 in 1,000 women (ages 55 and older) have breast cancer, what is the probability that this 55 year old Kansas woman has cancer given the result of her mammogram?

Review probabilities

```
cancerprob <- matrix(c(0.85,0.15,1.0,0.03,0.97, 1.0), byrow = TRUE, nrow = 2)
dimnames(cancerprob) <- list(disease=c("Yes", "No"),
                             diagnosis=c("Yes", "No","Total"))
cancerprob

##         diagnosis
## disease  Yes   No Total
##     Yes 0.85 0.15     1
##     No  0.03 0.97     1
```

Review calculated counts

```
cancer <- matrix(c(8.5,1.5,10,300,9690,9990,308.5,9691.5,10000), byrow = TRUE
, nrow = 3)
dimnames(cancer) <- list(disease=c("Yes", "No", "Total"),
                         diagnosis=c("Yes", "No", "Total"))
cancer

##         diagnosis
## disease   Yes     No Total
##   Yes     8.5    1.5    10
##   No    300.0 9690.0  9990
##   Total 308.5 9691.5 10000
```

The probability that the patient has cancer based on her positive diagnosis is 0.02755

```
# P(disease = yes | diagnosis = yes )
8.5/308.5

## [1] 0.02755267
```

Question Two

In a medical study that compared subjects with non-acute appendicitis and with acute appendicitis in terms of whether they suffered severe right abdominal pain. Such severe pain was reported by 5 of the 15 non-acute cases and by 1 of the 16 acute cases. The doctors believed that greater density of nerve fibres in the non-acute cases could increase the change of such pain. Conduct appropriate test and interpret your results.

```
library(fmsb)
or<- fmsb::oddsratio(5,1,15,16,conf.level = 0.95)

##             Disease Nondisease Total
## Exposed           5         15    20
## Nonexposed        1         16    17
## Total             6         31    37

#exposed is acute vs non acute, disease = yes pain
or

##
##  Odds ratio estimate and its significance probability
##
## data:  5 1 15 16
## p-value = 0.1209
## 95 percent confidence interval:
##   0.5567405 51.0910293
## sample estimates:
## [1] 5.333333
```

The odds ratio for this case indicates that nonacute cases are 5.3 times more likely to experience pain. However, the p value and confidence interval indicate that this is not a strong conclusion. The p value fails to meet a 0.05 threshold, and the lower bound of the confidence interval (0.5567) is very small, and close to zero, which would indicate no association. Therefore, the $H_0: \theta = 0$ cannot be rejected.

Question Three

Suicide has been the subject of increasing study in recent years, as psychologists struggle to understand the reasons that people choose to take their lives. Two questions in such studies concern the gender of the person committing suicide and the method chosen for doing so.Do women commit suicide at a higher rate than men, or is it the other way around? Are there differences by gender in the method used?

Here is a 2 by 4 contingency table that categorizes U.S. suicides in 1983 by gender and method ("hanging" also includes strangulation and suffocation). For instance, of the 28,295 suicides in the U.S. that year, 13,959 were by men using guns.

```
suicide <- matrix(c(13959,3148,3222,14557,21786,2641,2469,709,690,6509,16600,
5617,3931,2147,28295), byrow = TRUE, nrow = 3)
dimnames(suicide) <- list(gender=c("Male", "Female", "Total"),
                          method=c("Guns", "Poison","Hanging", "Other", "Tota
l"))
suicide

##          method
## gender    Guns Poison Hanging Other Total
##   Male   13959   3148    3222 14557 21786
##   Female  2641   2469     709   690  6509
##   Total  16600   5617    3931  2147 28295
```

a. Given that the numbers of men and women in the country are about equal, what does this say about the suicide rate among men as compared with women? Give an explicit numerical answer (for example, "Women are more than 5 times likelier than men to commit suicide," if that's correct).

```
# probability that a suicide is male
maleprob <- 21786/28295
maleprob

## [1] 0.7699594

# probability that a suicide is female
femaleprob <- 6509/28295
femaleprob

## [1] 0.2300406
```

Without knowledge of the overall rate of suicide in the population, the odds ratio cannot be calculated. What can be calculated is conditional probability of gender based on the fact that a suicide occured. If a suicide occured, there is a probability of 0.7699594 that it was a male, while there is only a 0.2300406 that it was a female.

b. Is there a difference between men and women in the chosen method of suicide? Conduct an appropriate hypothesis test.

```
suicide_chi <- chisq.test(suicide)
suicide_chi

## 
##  Pearson's Chi-squared test
## 
## data:  suicide
## X-squared = 14327, df = 8, p-value < 2.2e-16
```

According to the Pearson χ^2 statistic, with H_0: all π_i are the same, H_0 is rejected. There is an association between gender and method of suicide, with an estimate of 14327, and a signifcant p value.

c. What are the risk difference, relative risk, and odds ratio of committing suicides by gun (using gun or not using gun) for male vs female? Provides the 95% CIs.

Risk Difference

```
gunrd <- fmsb::riskdifference(13959,2641,21786,6509, conf.level = 0.95)

##                    Cases People at risk Incidence rates
## Exposed   1.395900e+04   2.178600e+04    6.407326e-01
## Unexposed 2.641000e+03   6.509000e+03    4.057459e-01
## Total     1.660000e+04   2.829500e+04    5.866761e-01

gunrd

## 
##  Risk difference and its significance probability (H0: The
##  difference equals to zero)
## 
## data:  13959 2641 21786 6509
## p-value < 2.2e-16
## 95 percent confidence interval:
##  0.2214630 0.2485104
## sample estimates:
## [1] 0.2349867
```

The risk difference is 0.2349867 with a sufficiently small p value and 95% confidence interval 0.221463, 0.2485104.

Relative Risk

```
gunrr <- fmsb::riskratio(13959,2641,21786,6509, conf.level = 0.95)

##            Disease Nondisease Total
## Exposed      13959       7827 21786
## Nonexposed    2641       3868  6509

gunrr

## 
##  Risk ratio estimate and its significance probability
## 
## data:  13959 2641 21786 6509
## p-value < 2.2e-16
## 95 percent confidence interval:
##  1.530890 1.628926
## sample estimates:
## [1] 1.579147
```

The relative risk is 1.5791474 with a sufficiently small p value and 95% confidence interval 1.5308896, 1.6289265.

Odds Ratio

```
gunor <- fmsb::oddsratio(13959,2641,(3148+3222+1457),(2469+709+690),conf.leve
l = 0.95)

##             Disease Nondisease Total
## Exposed       13959       7827 21786
## Nonexposed     2641       3868  6509
## Total         16600      11695 28295

# exposed = male, disease = gun, nondisease = not gun
gunor

## 
##  Odds ratio estimate and its significance probability
## 
## data:  13959 2641 (3148 + 3222 + 1457) (2469 + 709 + 690)
## p-value < 2.2e-16
## 95 percent confidence interval:
##  2.468068 2.764375
## sample estimates:
## [1] 2.612023
```

The odds ratio for gun vs other method is 2.6120232, with a pvalue significantly small p value and a 95% confidence interval of 2.4680684, 2.7643746.

Question Four

A study on educational aspirations of high school students (S. Crysdale, Int. J. Compar. Social. 16: 19-36, 1975) measured aspirations with the scale (some high school, high school graduate, some college, college graduate). The student counts in these categories were (9, 44, 13, 10) when family income was low, (11, 52, 23, 22) when family income was middle, and (9, 41, 12, 27) when family income was high.

```
ed_income <- matrix(c(9,44,13,10,11, 52, 23, 22,9, 41, 12, 27), byrow = TRUE,
nrow = 3)
dimnames(ed_income ) <- list(income=c("low", "middle", "high"),
                             educ=c("some hs", "hs grad", "some col", "col grad"
))
ed_income

##          educ
## income    some hs hs grad some col col grad
##   low           9      44       13       10
##   middle       11      52       23       22
##   high          9      41       12       27
```

a. Test independence of educational aspirations and family income using odds ratio. Explain the deficiency of this test for these data.

```
ed_or <- epitools::oddsratio.wald(ed_income)
ed_or
```

```
## $data
##         educ
## income   some hs hs grad some col col grad Total
##   low          9      44       13       10    76
##   middle      11      52       23       22   108
##   high         9      41       12       27    89
##   Total       29     137       48       59   273
##
## $measure
##         odds ratio with 95% C.I.
## income    estimate     lower    upper
##   low    1.0000000        NA       NA
##   middle 0.9669421 0.3672457 2.545917
##   high   0.9318182 0.3369574 2.576839
##
## $p.value
##         two-sided
## income   midp.exact fisher.exact chi.square
##   low            NA           NA         NA
##   middle  0.9509406   0.45641322 0.44712438
##   high    0.8944031   0.06740457 0.07267464
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"

ed_or$measure[2]

## [1] 0.9669421
```

This method is not sufficient because both variables are ordinal, and a linear relationship may be present, and should be tested for.

b. Find the standardized Pearson residuals. Do they suggest any association pattern? Why?

```
ed_chisq <- chisq.test(ed_income)
ed_chisq$stdres

##         educ
## income      some hs    hs grad   some col   col grad
##   low     0.4061328  1.5828205 -0.1286367 -2.1078423
##   middle -0.1898118 -0.5440627  1.3041565 -0.4031584
##   high   -0.1903291 -0.9459053 -1.2374420  2.4360173
```

The standardized pearson residuals suggest that many cells are significantly different from what is expected under the H_0, and therefore an association exists. This method takes into consideration the ordinal nature of the variables, so an association is clear.

c. Conduct an alternative test that may be more power. Interpret.

```
library("DescTools")
```

```
## Warning: package 'DescTools' was built under R version 3.4.4

##
## Attaching package: 'DescTools'

## The following objects are masked from 'package:fmsb':
##
##     CronbachAlpha, VIF

DescTools::GoodmanKruskalGamma(ed_income, conf.level = 0.95)

##       gamma      lwr.ci      upr.ci
## 0.162547645 0.006716385 0.318378906

# -1 to +1 association, 0 indicates no association
```

The alternative test used here calculated gamma, with an estimate of 0.1625. The 95% CI for gamma is (0.0067,0.3184). This indicates that there is a modest positive association between income and aspirational education.

Question Five

Breast cancer is one of the most common malignancies among women in the U.S. If it is detected early enough-before the cancer spreads-chances of successful treatment are much better. The question is, do screening programs speed up detection enough to matter?
The first large-scale experiment to answer this question was run by a group called the Health Insurance Plan (HIP) of Greater New York, starting in 1963. The subjects (all members of this plan) were 62,000 women age 40 to 64, who were divided at random into two equal groups. In the treatment group, the women were encouraged to come in for annual screening, including a breast examination by a doctor and a mammogram (a kind of X-ray). About 20,200 women in the treatment group did come in for the screening, but 10,800 refused. The control group was offered standard health care that did not include the examination and mammogram. All the women were followed for many years; results for the first 5 years are shown below.

```
cancer_exam <- read.csv("cancer_exam.csv")
print(cancer_exam)

##   treatment death no.death total
## 1  examined    23    20177 20200
## 2   refused    16    10784 10800
## 3   control    63    30937 31000
```

Does screening cause a decrease in breast cancer mortality? Conduct appropriate analysis using given data. You may ignore the fact that some women refused the treatment and use 31000 as the total sample size for treatment group.

```
cancer_exam <- read.csv("cancer_exam2.csv")
cancer_exam
```

```
##   treatment death no.death total
## 1     treat    39    30961 31000
## 2   control    63    30937 31000
```

```
screen <- fmsb::oddsratio(39,63,30961,30937,conf.level = 0.95)
# exposed = male, disease = gun, nondisease = not gun
screen
```

The odds ratio estimate is 0.6186 with a p value of 0.0174and 95% confidence interval of 0.4148, 0.9225 which suggests that there is a strong association between treatment and mortality.

To understand the association better, a standardized Pearson residual will be considered.

```
screen<- matrix(c(39,30961,63,30937), byrow = TRUE, nrow = 2)
dimnames(screen) <- list(treatment=c("exam", "no exam"),
                          mortality=c("yes", "no"))

screen_chisq <- chisq.test(screen)
screen_chisq$stdres
```

```
##          mortality
## treatment       yes        no
##   exam    -2.378311  2.378311
##   no exam  2.378311 -2.378311
```

These results indicate that if the two variables were independent there would have been 2.37 more mortalities in the treatment group than what was observed.

In reality, it is very common that the data is not exactly as planned (missing data, noncompliance etc). Discuss what may be appropriate to do with “Refused” cases.

I would combine the refused and control groups. My rationale for this is that the critical intervention being applied to the test group is the exam and mammogram, not the encouragement to get them. From this perspective, they did not receive the treatment. The argument for treating them as their own group is also worth considering.This method ensures that any potential impact of being encouraged to obtain an exam and/or unwillingness to obtain an exam would be accounted for.

Session Info

```
sessionInfo()
```

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] DescTools_0.99.25 fmsb_0.6.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.13     lattice_0.20-35  mvtnorm_1.0-6    digest_0.6.12
##  [5] rprojroot_1.2    MASS_7.3-47      grid_3.4.1       backports_1.1.0
##  [9] magrittr_1.5     evaluate_0.10.1  stringi_1.1.5    Matrix_1.2-10
## [13] boot_1.3-19      rmarkdown_1.6    epitools_0.5-10  tools_3.4.1
## [17] foreign_0.8-69   stringr_1.2.0    yaml_2.1.14      compiler_3.4.1
## [21] manipulate_1.0.1 htmltools_0.3.6  knitr_1.16       expm_0.999-2
```