CatData HW3

Matthew Vanaman

03/11/19

All code and work are shown in the appendix.

2.2

(a)

Answer:
Recoding this in a way that makes more sense to me. Will leave X and Y as they are:

- $X(0 = \text{no disease}, 1 = \text{disease})$
- $Y(0 = \text{negative}, 1 = \text{positive})$
- Sensitivity $= \pi_1 = P(Y = 1|X = 1) =$ probability positive diagnosis given disease
- Specificity $= 1 - \pi_2 = 1 - P(Y = 1|X = 0) = 1$ - probability positive diagnosis given no disease

After subtracting away the probability of positive diagnosis given no disease, you are left with probability of negative diagnosis given no disease:

$1 - P(Y = 1|X = 0) = P(Y = 0|X = 0) =$ probability negative diagnosis given no disease

Likewise, the "noise" of the test is captured by 1 - specificity. A good test would want to show that 1 - specificity yields a large probability as an effective test will have a large probability of correctly screening patients that do not have the disease.

(b)

Answer:
$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)}.$$

(c)

Answer:
0.07

(d)

Answer:

	Negative	Positive	Row Margin
No Disease	0.8712	0.1188	0.99
Disease	0.0014	0.0086	0.01

Sensitivity (0.86) is the probability of a positive diagnosis given someone has the disease. That is, 86% of those with the disease will be identified with a positive test result. If 0.01 is the probability of having the disease, then 0.0086 is the probability of having the disease and receiving a positive diagnosis. That is, if I do not know whether or not I have the disease, there is a .0086 probability I will both receive a positive test result and truly have the disease. On the other hand, the specificity (0.88) gives us the probability of negative diagnosis with no disease - that is, if I know I do not have the disease, the test will identify me as negative 88% of the time. Therefore, if I know nothing about my disease, the probability that I will have no disease and receive a negative result is 0.8514. Most of the people with the disease are in the positive diagnosis column, which is a good thing. However, we can see based on the column margin that only about 7% of those who get a positive result actually have the disease. Needless stress!

2.5

(a)

Answer:
Relative risk,though depending on the circumstances, one can sometimes use the odds ratio as an estimate of the relative risk; in those cases, it could be both. If the odds of either men or women getting cancer were small, the odds ratio and risk ratio would be pretty similar.

(b)

(i)

Answer:
0.55

(ii)

Answer:
1.82

2.13

(a)

Answer:
$(-0.018, 0.061)$.

(b)

Answer:
CI for odds ratio: (0.8988938, 1.394697). However, odds ratios are very skewed, and because CIs are inferential statistics, we should probably use the log odds instead. If using the log of the odds, the CI is (-0.1109301, 0.3848729). Interpretation is that if we were to collect 100 samples of this size from this same population, 90 out of 100 of the calculated CIs for the odds (or log odds) ratios will encapsulate the true population parameter, while 10 will fail to do so just due to sampling variability. We of course are counting on our sample having not been one of those 10% of samples that will fail to capture the population parameter due to sampling variability.

(c)

Answer:
$\chi^2 = 0.825, df = 1, p = 0.36$

2.21

(a)

Answer:
No, there are redundances in each of the column. With χ^2, you want to know if two constructs are independent, or if they are not independent, then it is because of some association between the constructs themselves. With this table, dependence might be due to associations between the constructs, but could also be due to subjects being double- or triple-counted across cells. You can't really tell unless the observations in the cells are dependent.

(b)

Answer:
Yes, if there were 100 subjects of each gender, then the numbers in the cells are out of 100.

	Male	Female	Total
Yes	60	75	135
No	40	25	65
Total	100	100	200

2.33

(a)

Three-way contingency table:

	Death Penalty	No Death Penalty
White Killed White	151	19
White Killed Black	9	0
Black Killed White	63	11
Black Killed Black	103	6
Total	326	36

(b)

Andwer:
Victim white: 0.7107189
Victim black: 0.8380567
Could not figure out why my answers did not match the book.

(c)

Answer:
1.15. Yes, because the conditional odds ratio from the partial table indicated that whites are between 29% and 17% less likely to get the death penalty than blacks (depending on race of the victim), while according to the marginal odds ratio (which does not control for victim), whites are 15% *more* likely to recieve the death penalty than blacks. The effect of race of defendenant has switched directions, illustarting Simpson's paradox.

2.36

Answer:
X = Teach a class (yes or no), Y = drink a second up of coffee (yes or no), Z = teaching time (morning or afternoon). Overall, teachers probably find that drinking a second cup of coffee only bears a relationship with teaching when teaching occurs in the morning (marginally associated). If controlling for the time of day that teaching happens, teaching and deciding to drink a second cup of coffee are probably unrelated (conditionally indepedent).

Appendix

2.2

(a)

NA

(b)

Work:

$$\frac{\pi_1\gamma}{[\pi_1\gamma + \pi_2(1-\gamma)]} = \frac{P(B|A)P(A)}{P(B)} = \text{Bayes' rule.}$$

In the above equation, P(B) represents P(X=1), or the probability of having the disease. But this is an unknown value expressed as γ, we have to use a particular version of Bayes' rule that takes into account the conditional probabilities associated with P(B):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\widetilde{A})P(\widetilde{A})}.$$

Just replace with the problem's notation:

$$P(X=1|Y=1) = \frac{P(Y=1|X=1)P(X=1)}{P(Y=1|X=1)P(X=1) + P(Y=1|X=0)P(X=0)}.$$

(c)

Work:

We need to solve for $P(X=1|Y=1)$. Lay out the pieces:

- $X(0 = \text{no disease}, 1 = \text{disease}.$
- $P(X=1) = 0.01.$
- $Y(0 = \text{negative}, 1 = \text{positive}.$
- Sensitivity $= \pi_1 = P(Y=1|X=1) = 0.86.$
- Specificity $= 1 - \pi_2 = 1 - P(Y=1|X=0) = 0.88.$
- If $1 - P(Y=1|X=0) = 0.88$, then $P(Y=1|X=0) = 0.12.$
- If $P(X=1) = 0.01$, then $P(X=0) = 1 - P(X=1) = 0.99.$

Plug and chug:

$$P(X=1|Y=1) = \frac{P(Y=1|X=1)P(X=1)}{P(Y=1|X=1)P(X=1)+P(Y=1|X=0)P(X=0)},$$

$$= \frac{0.86 \times 0.01}{(0.86 \times 0.01)+(0.12 \times 0.99)},$$

$$= 0.07.$$

```
# calculations
(0.86 * 0.01) / ((0.86 * 0.01) + (0.12 * 0.99))
```

```
[1] 0.06750392
```

(d)

Work:

```
test <- matrix(c("0.8712","0.0014","0.1188","0.0086",0.99,0.01), ncol=3)
colnames(test) <- c('Negative', 'Positive', "Row Margin")
rownames(test) <- c('No Disease', 'Disease')
test.table <- as.table(test)
knitr::kable(test.table)
```

We start with what we know:
- Sensitivity $= \pi_1 = P(Y=1|X=1) = 0.86$.
- Specificity $= 1 - \pi_2 = 1 - P(Y=1|X=0) = 0.88$.
- $P(Y=1|X=0) = 0.12$.
- $P(X=0) = 0.99$.
- $P(X=1) = 0.01$.

Probability of both having disease and getting positive result:

```
0.01 * 0.86
```

```
[1] 0.0086
```

From there, because we have the margin and one of the cells, we have the other cell in that row:

```
0.01 - 0.0086
```

```
[1] 0.0014
```

Probability of negative diagnosis and not having disease:

```
0.99 * 0.88
```

```
[1] 0.8712
```

Get the last cell:

```
0.99 - 0.8712
```

```
[1] 0.1188
```

```
# check work
0.0086 + 0.0014
```

```
[1] 0.01
```

```
0.1386 + 0.8514
```

```
[1] 0.99
```

2.5

(a)

NA

(b)

(i)

Work:
The risk ratio is just % increase or decrease in risk, relative to 1.00 (1.00 would mean equal risk in both groups because 1.00 indicates that the numerator and denominator of the ratio are the same). If we know the drug group was 45% less likely to get cancer, then we consider what a 45% decrease from 1 would be: 0.55. Say we start with no effect:

$$RR = \frac{1}{1}.$$

But we know the drug group was 45% less likely to get cancer:

$$\frac{1 - 0.45}{1} = 0.55.$$

(ii)

Work:
We found the numerator for the risk ratio in part i. To get the RR for the placebo group, we flip the fraction from the last problem over:

```
1 / (1 - 0.45)
```

```
[1] 1.818182
```

So the placebo group was 82% more likely to get cancer.

2.13

(a)

Work:

Proportions:

```
# Proportion female who believed:
509 / 625
```

```
[1] 0.8144
```

```
# Proportion male who believed:
398 / 502
```

```
[1] 0.7928287
```

Get the standard error:

$$SE = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}},$$

$$= \sqrt{\frac{0.8144(1-0.8144)}{625} + \frac{0.79(1-0.79)}{502}},$$

```
sqrt( (((0.8144 * (1-0.8144)) /625)) + ((0.7928287 * (1 - 0.7928287))  /502 ))
```

```
[1] 0.02385452
```

Next, get the confidence interval:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{0.1/2}(SE),$$

$$(0.8144 - 0.7928) \pm 1.645(0.0239),$$

$$= (-0.01764069, 0.06084069).$$

```
# calculations
#lower
0.8144 - 0.7928 - 1.645 * 0.02385452
```

```
[1] -0.01764069
```

```
#upper
0.8144 - 0.7928 + 1.645 * 0.02385452
```

```
[1] 0.06084069
```

(b)

Work:

Use the standard error formula

$$SE = \sqrt{(\frac{1}{n_{00}}) + (\frac{1}{n_{01}}) + (\frac{1}{n_{10}}) + (\frac{1}{n_{11}})},$$

and the CI for log odds:

$$\log\hat{\theta} \pm z_{0.1/2}(SE).$$

```
# calculate standard error
sqrt((1/509) + (1/398) + (1/116) + (1/104) )
```

```
[1] 0.1507092
```

```
# get log odds
(OddsRatio <- (0.8144 / (1 - 0.8144)) / (0.7928 / (1 - 0.7928)))
```

```
[1] 1.146795
```

```
# CI lower
OddsRatio - 1.645 * 0.1507
```

```
[1] 0.8988938
```

```
# CI upper
OddsRatio + 1.645 * 0.1507
```

```
[1] 1.394697
```

```
# what about for log odds?
logodds <- log(OddsRatio)
# CI lower - log
logodds - 1.645 * 0.1507
```

```
[1] -0.1109301
```

```
# CI upper - log
logodds + 1.645 * 0.1507
```

```
[1] 0.3848729
```

(c)

Work:

```
# make a contigency table
belief <- matrix(c("509","398",(509 + 398),"116","104",(116 + 104),
                   (509 + 116),(398 + 104), (907 + 220)),
                 ncol=3)
colnames(belief) <- c('Belief', 'Disbelief', "Row Margin")
rownames(belief) <- c('Female', 'Male', "Column Margin")
belief.table <- as.table(belief)
knitr::kable(belief.table, format = "markdown")
```

n_{00}

```
(502 * 220) / 1127
```

```
[1] 97.99468
```

n_{01}

```
(502 * 907) / 1127
```

```
[1] 404.0053
```

n_{10}

```
(625 * 220) / 1127
```

```
[1] 122.0053
```

n_{11}

```
(625 * 907) / 1127
```

```
[1] 502.9947
```

Following the chi-square formula:

$$\chi^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

```
cell_00 <- (104 - 97.99468)^2 / 97.99468
cell_01 <- (398 - 404.0053)^2 / 404.0053
cell_10 <- (116 - 122.0053)^2 / 122.0053
cell_11 <- (509 - 502.9947)^2 / 502.9947
cell_00 + cell_01 + cell_10 + cell_11
```

```
# check work, get p-value
belief_test <- matrix(c(509,398,116,104), ncol=2)
colnames(belief_test) <- c('Belief', 'Disbelief')
rownames(belief_test) <- c('Female', 'Male')
belief_test <- as.table(belief_test)
knitr::kable(belief_test, format = "markdown")
```

	Belief	Disbelief
Female	509	116
Male	398	104

```
chisq.test(belief_test, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  belief_test
X-squared = 0.82458, df = 1, p-value = 0.3638
```

2.21

(a)

NA

(b)

Work:

```
# make a contingency table
gender <- matrix(c(60,40,(60+40),
                   75,25, (75+25),
                   (60+75), (40+25), 200), ncol=3)
colnames(gender) <- c('Male', "Female", "Total")
rownames(gender) <- c('Yes', 'No', "Total")
knitr::kable(gender)
```

2.33

(a)

Three-way contingency table:

```
deaths <- matrix(c(151, 9, 63, 103, (151+9 +63+103),
                   19, 0, 11,6,(19+0+11+6)),
                 ncol=2)
colnames(deaths) <- c('Death Penalty', "No Death Penalty")
rownames(deaths) <- c('White Killed White', 'White Killed Black',
                      "Black Killed White", "Black Killed Black",
                      "Total")
knitr::kable(deaths)
```

(b)

Work:

Add the 0.5:

```
defendent_white <- matrix(c(19.5, 11.5,0.5,6.5, (19.5+0.5), (11.5+6.5),
                            151.5, 63.5,9.5, 103.5,(151.5 + 9.5),(63.5 + 103.5)),
                          ncol=2)
colnames(defendent_white) <- c('Death Penalty', "No Death Penalty")
rownames(defendent_white) <- c('White Killed White', 'Black Killed White',
                               "White Killed Black", "Black Killed Black",
                               "Total: Def. White", "Total: Def. Black")
knitr::kable(defendent_white)
```

Victim white:

```
(19.5 * 63.5) / (151.5 * 11.5)
```

```
[1] 0.7107189
```

Victim black:

```
(0.5*103.5) / (9.5*6.5)
```

```
[1] 0.8380567
```

(c)

Work:

```
# calculations
(167 * 20) / (161 * 18)
```

```
[1] 1.152519
```

2.36

NA