DFSC 5340.02 Assignment 5

Due: Tuesday Nov 10@11:59PM

Total Points: 170 (20 points for question 1 and 30 points for each in questions 2 to 5)

1. Based on current estimates of how well mammograms detect breast cancer, the following Table I shows what to expect for 100,000 adult women over the age of 40 in terms of whether a woman has breast cancer and whether a mammogram gives a positive result (i.e., indicates that the woman has breast cancer).

   (a) Construct the conditional distributions for the mammogram test result, given the true disease status. Does the mammogram appear to be a good diagnostic tool?

   (b) Construct the conditional distribution of disease status, for those who have a positive test result. Use this to explain why even a good diagnostic test can have a high false positive rate when a disease is not common.

Table I.

		Diagnostic Test	
		Positive	Negative
Breast Cancer	YES	860	140
	No	11800	87120

(a)

For Yes,

Positive : 860/1000 = 0.86

For Negative : 140/1000 = 0.14

Total : 860 + 140 = 1000

For No,

Total : 11800 + 87120 = 98920

Positive : 11800/98920 = 0.12

Negative : 87120/98920 = 0.88

The status of true disease is true's conditional distribution on diagonosis test result which is .86 for positive and 0.14 for negative.

Again, the status of the true disease is false's conditional distribution on diagnostic test result is 0.12 for positive and 0.88 for negativ

So, the mammogram appears to be good tool.

(b) Positive Total : 860 + 11800 = 12660

Negative Total : 140 + 87120 = 87260

860/12660 = 0.068

140/87260 = 0.0016

11800/12660 = 0.932

87120/87260 = 0.9984

Conditional distribution for positive test result on true disease status is 0.068 for breast cancer and 0.932 for no breast cancer.

It would be acceptable to have a high false positive than a high false negative rate for a disease that uncommon and life-threatening.

2. The following Table II is from the 2014 General Social Survey, cross-classifies happiness and marital status.

(a) Compute the chi-squared value X^2.

(b) The table also shows, in parentheses, the standardized residuals. Summarize the association by indicating which marital statuses have strong evidence of (i) more, (ii) fewer people in the population in the very happy category than if the variables were independent.

(c) Compare the married and divorced groups by the difference in proportions in the very happy category.

Table II. Happiness and Marital Status

Marital Status	Very Happy	Pretty Happy	Not Too Happy
Married	472 (9.8)	592 (-3.3)	90 (-7.6)
Widowed	49 (-2.4)	120 (.7)	38 (2.2)
Divorced	94 (-3.3)	233 (.5)	84 (4.6)
Separated	12 (-3.2)	47 (.5)	22 (3.7)
Never married	158 (-5)	410 (3.3)	105 (1.2)

(a) Null Hypothesis H_0

Alternative " H_1

We know, $X^2 = \sum \frac{(fo - fe)^2}{fe}$

The test statistic value $X^2 = 135.3$

DF $= (r-1)(c-1) = (5-1)(3-1) = 4(2) = 8$

P-value $=$ CHIDIST (X^2, df)

$=$ CHIDIST $(135.3, 8) = 0.0000$

The Pvalue is low so we can reject null hypothesis. So we can say there is strong evidence of an association between happiness and marital status.

(b) (i) If we see the very happy category highest standardized residual is 9.8. The married marital status has strong evidence in the population in the very happy category than if the variables were independent.

(ii) In the very happiness category the widowed, divorced separated and never married marital status has fewer evidence in the population in the very happy category than if the variables were independent.

(c) Married people total = 472+592+90 = 1154

Divorced " " = 94+233+84 = 411

The proportion of married people who are very happy $\frac{472}{1152}$

$= 0.41$

The proportion of divorced people who are divorced $94/411$

$= 0.23$

$$\text{Z-statistic} = \frac{.41 - .23}{\sqrt{\frac{.41(1-.41)}{1154} + \frac{.23(1-.23)}{411}}}$$

$$= \frac{0.18}{0.0253} = 7.11$$

$$\text{P-value} = 2 \times P(z > z_0) = 2(1 - P(z < z_0)) = 2(1 - P(z < 7.11))$$

$$= 2 \times (1-1) = 0$$

We can see P value is 0 here, so we can reject hypothesis.

There is strong evidence to support that there is significant difference between the proportion of married and divorced people who are very happy.

3. Table III cross-classifies 68,694 passengers in autos and light trucks involved in accidents in the state of Maine by whether they were wearing a seat belt and by whether they were injured or killed. Describe the association using

(a) The difference between two proportions, treating whether injured or killed as the response variable.

(b) The odds ratio

TABLE III.

		Injury	
		Yes	No
Seat Belt	YES	2409	35383
	No	3865	27037

(a) The difference between two proportions treating whether injured or killed :

$$\frac{2409}{2409+35383} - \frac{3865}{3865+27037}$$

$$= -0.06$$

So, we can determine who were injured or killed is 0.06 lower for those who wore seat belts.

(b) The odds ratio is :

$$\theta = \frac{2403/35383}{3865/27037} = 0.48$$

For those who wear seat belt, the odds of getting injured or killed 0.48 times the odds for those who do not wear seat belt.

4. The 2012 National Survey on Drug Use and Health (NSDUH) estimated that 23% of Americans aged 12 or over reported binge drinking in the past month, and 7% had used marijuana in the past month.

(a) Find the odds of (i) binge drinking, (ii) marijuana use. Interpret.

(b) Find the odds ratio comparing binge drinking to marijuana use. Interpret.

(a) (i) Proportion of current month $(100-23)\% = 77\%$

The odds of binge drinking:

$$\frac{\text{Proportion of past month}}{\text{Proportion of current month}}$$

$$= \frac{\cdot 23}{\cdot 77} = 0.299$$

(ii) The odds of marijuana:

$$\frac{\text{Proportion of past month}}{\text{Proportion of current month}}$$

$[(100-7)\% = 07\%]$

$$= \frac{0.07}{\cdot 93} = 0.075$$

(b) Odds ratio comparing binge drinking to marijuana use:

$$\frac{\cdot 299}{\cdot 075} = 3.97$$

5 According to the U.S. Bureau of Justice Statistics, in 2014 the incarceration rate in the nation's prisons was 904 per 100,000 male residents, 65 per 100,000 female residents, 2805 per 100,000 black residents, and 466 per 100,000 white residents. (Source: www.bjs.gov.)

(a) Find the odds ratio between whether incarcerated and (i) gender, (ii) race. Interpret.

(b) According to the odds ratio, which has the stronger association with whether incarcerated, gender or race? Explain

(a) Given here, the incarceration rate : 904 per 100,000 male

" " " : 65 per 100,000 female

The odd ratio between incarcerated ratio and gender :

$$= \frac{\frac{904}{100,000}}{\frac{65}{100,000}}$$

$$= \frac{904}{65}$$

$$= 13.91$$

(b) The odd ratio between incarcerated ratio and race :

$$= \frac{\frac{2805}{100000}}{\frac{466}{100,000}}$$

$$= \frac{2805}{466}$$

$$= 6.02$$

(b)

According to the odds ratio, incarcerated ratio and gender is stronger than the odds ratio of incarcerated ratio and race.

6. A clinical psychologist wants to choose between two therapies for treating mental depression. For six patients, she randomly selects three to receive therapy A, and the other three receive therapy B. She selects small samples for ethical reasons; if her experiment indicates that one therapy is superior, that therapy will be used on her other patients having these symptoms. After one month of treatment, the improvement is measured by the change in score on a standardized scale of mental depression severity. The improvement scores are 10, 20, 30 for the patients receiving therapy A, and 30, 45, 45 for the patients receiving therapy B.

(a) Using the method that assumes a common standard deviation for the two therapies, show that the pooled s = 9.35 and se = 7.64.

(b) When the sample sizes are very small, it may be worth sacrificing some confidence to achieve more precision. Show that the 90% confidence interval for (μ2 − μ1) is (3.7, 36.3). Interpret.

(c) Estimate and interpret the effect size.

(a)

$\mu_1 = \frac{10+20+30}{3} = 20$

$\mu_2 = \frac{30+45+45}{3} = 40$

Standard Deviation $s_1 = \sqrt{\frac{(10-20)^2+(20-20)^2+(30-20)^2}{2}}$

$= \sqrt{\frac{200}{2}} = 10$

SD $s_2 = \sqrt{\frac{(30-40)^2+(45-40)^2+(45-40)^2}{2}} = \sqrt{\frac{110}{2}}$

$= 8.66$

Pooled $s = \sqrt{\frac{s_1^2+s_2^2}{2}} = \sqrt{\frac{10^2+8.66^2}{2}} = \frac{100+74.99}{2}$

$= 9.35$

So, Pooled $s = 9.35$ (Proved)

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10^2}{3} + \frac{(8.66)^2}{3}}$$

$$= 7.64$$

So, SE is 7.64 (Proved)

(b) We know, $SE = 7.64$ $\quad \bar{\mu}_2 - \bar{\mu}_1 = 40 - 20 = 20$

$n_1 = 3$ $\quad s = 9.35$

$n_2 = 3$

DF is $(3+3-2) = 4$

From table we can see for df = 4 and 90% confidence level t value is 2.13

Confidence interval $= (\mu_2 - \mu_1) \pm t \cdot SE$

$= 20 \pm 2.13 * 7.64 = (3.7, 36.3)$ (Proved)

Interpretation

The P-value is 0.100208, which is greater than 0.05. The result is not significant at $p < .10$ and we can guess, we can reject the hypothesis and go with null hypothesis.

(c) Effect size $= \frac{\mu_2 - \mu_1}{s} = \frac{20}{9.35}$

$= 2.14$ (Answer)

Interpretation

The effect size is big so we can say the therapy will have significantly high effect on depression.