Matt Isaac
A01515095
Biostatistics
Homework 3 – Due Monday, 26 Feb

1. Address the research questions below.
    a. Are physicians detecting depression any differently than surveys? (For this question, ignore how physicians detected depression, and assume each physician assessed each patient for depression.)
        i. Two-way table of category counts:

Table of SurveyDep by PhysicianDep			
SurveyDep	**PhysicianDep**		
Frequency Row Pct	**1**	**2**	**Total**
1	9 14.06	55 85.94	64
2	0 0.00	47 100.00	47
Total	9	102	111

        ii. Name of the appropriate test: Two-sample Binomial Test
        iii. Null and alternative hypothesis:
            H_0: $p_{physician} = p_{survey}$
            H_A: $p_{physican} \neq p_{survey}$
        iv. P-value from the test: 0.0073

Statistic	DF	Value	Prob
Chi-Square	1	7.1926	0.0073

        v. Conclusion in context of research: Since the p-value is < 0.05, we reject H_0. This implies that there is a significant difference between the physicians detecting depression and the survey detecting depression.

    b. Are there differences between the survey-based rates of depression (SurveyDep) in the seven different diagnostic groups?
        i. Two-way table of category counts:

Table of Cardiac_dx by SurveyDep			
Cardiac_dx	**SurveyDep**		
Frequency Col Pct	**1**	**2**	**Total**
1	19 29.69	18 38.30	37
2	15 23.44	5 10.64	20
3	14 21.88	13 27.66	27
4	5 7.81	4 8.51	9
5	6 9.38	1 2.13	7
6	3 4.69	4 8.51	7
7	2 3.13	2 4.26	4
Total	64	47	111

ii. Name of the appropriate test: Chi-square General Association test (Could also use Row Mean Scores Differ test – they will be equivalent).

iii. Null and alternative hypothesis:
H_0: There is no association between Cardiac_dx group and survey-based rates of depression.
H_A: The distribution over Survey-based rates of depression differs between the different levels of Cardiac_dx. (i.e. some association/dependence exists).

iv. P-value from the test: 0.3821 (as seen in table below)

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	**Alternative Hypothesis**	**DF**	**Value**	**Prob**
1	**Nonzero Correlation**	1	0.0046	0.9459
2	**Row Mean Scores Differ**	6	6.3788	0.3821
3	**General Association**	6	6.3788	0.3821

v. Conclusion in context of research: Since our p-value is > 0.05, we fail to reject H_0. We do not have reason to believe that there is a linear association between Cardiac_dx group and the survey-based rates of depression.

c. Are there differences between the survey-based rates of depression in male and female patients?
   i. Two-way table of category counts:

2. Table of Gender by SurveyDep			
Gender	**SurveyDep**		
Frequency **Row Pct**	**1**	**2**	**Total**
1	34 55.74	27 44.26	61
2	30 60.00	20 40.00	50
Total	64	47	111

   i. Name of the appropriate test: 2-sample Binomial test
   ii. Null and alternative hypothesis:
   H_0: $p_{male} = p_{female}$
   H_A: $p_{male} \neq p_{female}$
   iii. P-value from the test: 0.6511 (see table below)

Statistic	DF	Value	Prob
Chi-Square	1	0.2045	0.6511

   iv. Conclusion in context of research: Since our p-value is > 0.05, we fail to reject H_0. That is, we have no evidence that depression rates are dependent on gender.

d. Are there differences between the survey-based rates of depression in patients who are married, single, widowed, or divorced?
   i. Two-way table of category counts:

Table of MaritalStatus by SurveyDep			
MaritalStatus	SurveyDep		
Frequency Col Pct	1	2	Total
1	40 62.50	36 76.60	76
2	4 6.25	1 2.13	5
3	12 18.75	9 19.15	21
4	8 12.50	1 2.13	9
Total	64	47	111

ii. Name of the appropriate test: Chi-square Row Mean Scores Differ test (assuming MaritalStatus to be nominal and the 2-category Survey-based depression variable to be ordinal).

iii. Null and alternative hypothesis:
H_0: There is no association between marital status and survey-based rates of depression.
H_A: Mean scores differ among different marital status groups.

iv. P-value from the test: 0.1474 (see table below)

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	2.9304	0.0869
2	Row Mean Scores Differ	3	5.3580	0.1474
3	General Association	3	5.3580	0.1474

v. Conclusion in context of research: Since our p-value is greater than 0.05, we fail to reject H_0. We have no evidence that there is any association between marital status and survey-based rates of depression.

e. Are there differences between the survey-based rates of depression in patients that had a prior cardiac history versus those with a new diagnosis?

i. Two-way table of category counts:

Table of PriorCardHx by SurveyDep			
PriorCardHx	SurveyDep		
Frequency Row Pct	1	2	Total
1	41 62.12	25 37.88	66
2	23 51.11	22 48.89	45
Total	64	47	111

ii. Name of the appropriate test: 2-sample Binomial test

iii. Null and alternative hypothesis:

H_0: $p_{PriorCardHx} = p_{NoPriorCardHx}$

H_A: $p_{PriorCardHx} \neq p_{NoPriorCardHx}$

iv. P-value from the test: 0.2491

Statistic	DF	Value	Prob
Chi-Square	1	1.3286	0.2491

v. Conclusion in context of research: We fail to reject H_0 because our p-value is > 0.05. This indicates that we have found evidence that Depression diagnosis from the survey is independent of PriorCardHx.

vi.

2. In class (Handout #7) we discussed how the two-sample binomial test statistic Z and the chi-square test statistic χ2 for testing independence in a two-by-two table are related. Clearly demonstrate this relationship in general ($Z^2 = \chi^2$), using the generic category counts a, b, c, and d as in the following table:

b. Report the test statistic Z (in terms of category counts a, b, c, and d only).

	Response 1	Response 2	
Group 1	a (or X_1)	b (or $X_1 – n_1$)	a + b (or n_1)
Group 2	c (or X_2)	d (or $X_2 – n_2$)	c + d (or n_2)

From handout 7,

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{Var(\hat{p}_1 - \hat{p}_2)}}$$

Let $p_i = P\{Response = 1\ in\ group\ i\}$, estimated by $\widehat{p_i} = \frac{X_i}{n_i}$.

Also, from handout 7, $\sqrt{Var(\hat{p}_1 - \hat{p}_2)} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$,

where p is estimated by

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{a+c}{a+b+c+d}$$

Thus,

$$Z = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\sqrt{\left(\frac{a+c}{a+b+c+d}\right)\left(1 - \frac{a+c}{a+b+c+d}\right)\left(\left(\frac{1}{a+b}\right) + \left(\frac{1}{c+d}\right)\right)}}$$

c. Report the test statistic χ2 (in terms of category counts a, b, c, and d only).

$$\chi 2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where $E_i = \frac{(row\ total)(column\ total)}{table\ total}$.

Thus,

$$\chi^2 = \left(\frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d}\right)^2}{\frac{(a+b)(a+c)}{a+b+c+d}}\right) + \left(\frac{\left(b - \frac{(a+b)(b+d)}{a+b+c+d}\right)^2}{\frac{(a+b)(b+d)}{a+b+c+d}}\right) + \left(\frac{\left(c - \frac{(c+d)(a+c)}{a+b+c+d}\right)^2}{\frac{(c+d)(a+c)}{a+b+c+d}}\right) + \left(\frac{\left(d - \frac{(c+d)(b+d)}{a+b+c+d}\right)^2}{\frac{(c+d)(b+d)}{a+b+c+d}}\right)$$

d. Show that $Z^2 = \chi^2$.

The Maple code on the page below shows the equality of Z^2 and χ^2.

$$> Z := \frac{\left(\frac{a}{a+b} - \frac{c}{c+d}\right)}{\sqrt{\left(\frac{(a+c)}{a+b+c+d}\right)\cdot\left(1 - \frac{(a+c)}{a+b+c+d}\right)\cdot\left(\frac{1}{a+b} + \frac{1}{c+d}\right)}}$$

$$Z := \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\sqrt{\frac{(a+c)\left(1 - \frac{a+c}{a+b+c+d}\right)\left(\frac{1}{a+b} + \frac{1}{c+d}\right)}{a+b+c+d}}} \quad (1)$$

$$> X2 := \frac{\left(a - \left(\frac{(a+b)\cdot(a+c)}{(a+b+c+d)}\right)\right)^2}{\frac{(a+b)\cdot(a+c)}{(a+b+c+d)}} + \frac{\left(b - \left(\frac{(a+b)\cdot(b+d)}{(a+b+c+d)}\right)\right)^2}{\frac{(a+b)\cdot(b+d)}{(a+b+c+d)}} + \frac{\left(c - \left(\frac{(c+d)\cdot(a+c)}{(a+b+c+d)}\right)\right)^2}{\frac{(c+d)\cdot(a+c)}{(a+b+c+d)}} + \frac{\left(d - \left(\frac{(c+d)\cdot(b+d)}{(a+b+c+d)}\right)\right)^2}{\frac{(c+d)\cdot(b+d)}{(a+b+c+d)}}$$

$$X2 := \frac{\left(a - \frac{(a+b)\,(a+c)}{a+b+c+d}\right)^2 (a+b+c+d)}{(a+b)\,(a+c)} + \frac{\left(b - \frac{(a+b)\,(b+d)}{a+b+c+d}\right)^2 (a+b+c+d)}{(a+b)\,(b+d)} + \frac{\left(c - \frac{(c+d)\,(a+c)}{a+b+c+d}\right)^2 (a+b+c+d)}{(c+d)\,(a+c)} + \frac{\left(d - \frac{(c+d)\,(b+d)}{a+b+c+d}\right)^2 (a+b+c+d)}{(c+d)\,(b+d)} \quad (2)$$

$> simplify(Z^2)$

$$\frac{(a\,d - c\,b)^2\,(a+b+c+d)}{(b+d)\,(a+c)\,(a+b)\,(c+d)} \quad (3)$$

$> simplify(X2)$

$$\frac{(a\,d - c\,b)^2\,(a+b+c+d)}{(b+d)\,(a+c)\,(a+b)\,(c+d)} \quad (4)$$

$>$

Appendix: SAS code

```
/* Generated Code (IMPORT) */
/* Source File: depression.csv */
/* Source Path: /home/mattisaac0/BioStatistics */
/* Code generated on: 2/21/18, 9:42 PM */
%web_drop_table(WORK.depression);
FILENAME REFFILE '/home/mattisaac0/BioStatistics/depression.csv';
PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.depression;
	GETNAMES=YES;
RUN;
* proc print data = depression;
%web_open_table(WORK.depression);

proc print data = depression;
run;

/* Code for problem 1(a)*/
* Frequency Table;
proc freq data = depression;
tables SurveyDep * PhysicianDep / chisq nopercent nocol;
run;

/* Code for problem 1(b)*/
proc freq data = depression;
tables Cardiac_dx * SurveyDep / chisq nopercent norow cmh;
run;

/* Code for problem 1(c)*/
proc freq data = depression;
tables gender * SurveyDep / chisq nopercent nocol;
run;

/* Code for problem 1(d)*/
proc freq data = depression;
tables MaritalStatus * SurveyDep / chisq nopercent norow cmh;
run;

/* Code for problem 1(e)*/
proc freq data = depression;
tables PriorCardHx * SurveyDep / chisq nopercent nocol;
run;
```