MATH 4044 – Statistics for Data Science

Practical Week 12 Solutions

Question 1

Is there a statistically significant relationship between gender and frequency of exercise? The data for this question is stored in a SAS data file called `pulse_rates.sas7bdat` located in `mydata` library on the SAS OnDemand server. Variables in that file are as follows:

Variable	Units
Height	cm
Weight	kg
Age	years
Gender	1 = 'Male', 2 = 'Female'
Smokes	1 = 'Yes', 2 = 'No'
Drinks alcohol	1 = 'Yes', 2 = 'No'
Exercise Frequency	1 = 'High', 2 = 'Moderate', 3 = 'Low'
Pulse	Pulse rate, beats per minute

Check the assumptions and perform an appropriate hypothesis test. Interpret the results.

The test appropriate for this scenario is the chi-square test of independence. As all expected counts in Table 1 are greater than 5, so we can proceed with the test.

The hypotheses are as follows:

H_0: Gender and frequency of exercise are independent

H_1: Gender and frequency of exercise are not independent

α =0.05

From Table 2, the test statistic is χ^2 = 4.4087 with 2 degrees of freedom. The corresponding *P*-value is 0.1103 > 0.05. Therefore, there is no statistically significant relationship between gender and frequency of exercise.

Table of Gender by Exercise					
		Exercise			
		High	Moderate	Low	Total
Gender					
Male	Frequency	11	31	17	59
	Expected	7.5091	31.645	19.845	
	Cell Chi-Square	1.6229	0.0132	0.408	
Female	Frequency	3	28	20	51
	Expected	6.4909	27.355	17.155	
	Cell Chi-Square	1.8775	0.0152	0.472	
Total	Frequency	14	59	37	110

Table 1. Contingency table for Gender by Exercise

Statistic	DF	Value	Prob
Chi-Square	2	4.4087	0.1103
Likelihood Ratio Chi-Square	2	4.6737	0.0966
Mantel-Haenszel Chi-Square	1	3.4635	0.0627
Phi Coefficient		0.2002	
Contingency Coefficient		0.1963	
Cramer's V		0.2002	

Fisher's Exact Test	
Table Probability (P)	0.0043
Pr <= P	0.1201

Table 2. Statistics for the contingency table for Gender by Exercise in Table 1

Question 2

A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, 'Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?' The responses are cross-tabulated based on gender.

		Gender		
		Male	Female	*Total*
Response	Yes	288	378	*666*
	No	61	62	*123*
	Not sure	10	7	*17*
Total		*359*	*447*	*806*

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

The hypotheses are as follows:

H_0: Gender and whether or not Facebook users know how to adjust their privacy settings are independent

H_1: Gender and whether or not Facebook users know how to adjust their privacy settings are dependent

α =0.05

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

From Table 3, all expected counts are greater than 5 and there is no matching of responses, so all necessary conditions for chi-square test of independence are satisfied.

From Table 4, the test statistic is χ^2 = 3.1291 with 2 degrees of freedom. The corresponding *P*-value is 0.2092 > 0.05. Therefore, there is no statistically significant relationship between gender and whether the user knows how to adjust privacy settings.

Table of Response by Gender				
		Gender		
		Female	Male	Total
Response				
No	Frequency	62	61	123
	Expected	68.215	54.785	
	Cell Chi-Square	0.5662	0.705	
NotSure	Frequency	7	10	17
	Expected	9.428	7.572	
	Cell Chi-Square	0.6253	0.7786	
Yes	Frequency	378	288	666
	Expected	369.36	296.64	
	Cell Chi-Square	0.2022	0.2518	
Total	Frequency	447	359	806

Table 3. Contingency table for Response by Gender

Statistic	DF	Value	Prob
Chi-Square	2	3.1291	0.2092
Likelihood Ratio Chi-Square	2	3.1127	0.2109
Mantel-Haenszel Chi-Square	1	2.1090	0.1464
Phi Coefficient		0.0623	
Contingency Coefficient		0.0622	
Cramer's V		0.0623	

Table 4. Statistics for the contingency table for Response by Gender in Table 3.

Question 3

Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

		Caffeinated coffee consumption					
		<1 cup per week	2-6 cups per week	1 cup per day	2-3 cups per day	>4 cups per day	*Total*
Clinical depression	**Yes**	670	373	905	564	95	*2,607*
	No	11,545	6,244	16,329	11,726	2,288	*48,132*
	Total	*12,215*	*6,617*	*17,234*	*12,290*	*2,383*	*50,739*

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

A chi-square test of independence is appropriate for this scenario.

(b) Write the hypotheses for the test you identified in part (a).

The hypotheses are as follows:

H_0: Incidence of clinical depression and the level of coffee consumption are independent

H_1: Incidence of clinical depression and the level of coffee consumption are dependent

α =0.05

(c) Use SAS to obtain appropriate output for the test identified in part (a). What is the conclusion of this test?

Table of Depression by Consumption							
		Consumption					Total
		1pd	1pw	2to3pd	2to6pw	4pd	
Depression							
No	Frequency	16329	11545	11726	6244	2288	48132
	Expected	16349	11587	11659	6277	2260.6	
	Cell Chi-Square	0.0233	0.155	0.3904	0.1736	0.3331	
Yes	Frequency	905	670	564	373	95	2607
	Expected	885.49	627.61	631.47	339.99	122.44	
	Cell Chi-Square	0.4297	2.8625	7.2084	3.2059	6.1496	
Total	Frequency	17234	12215	12290	6617	2383	50739

Table 5. Contingency table for Depression by Coffee consumption

Statistic	DF	Value	Prob
Chi-Square	4	20.9316	0.0003
Likelihood Ratio Chi-Square	4	21.5560	0.0002
Mantel-Haenszel Chi-Square	1	3.0076	0.0829
Phi Coefficient		0.0203	
Contingency Coefficient		0.0203	
Cramer's V		0.0203	

Table 6. Statistics for the contingency table for Depression by Consumption in Table 5.

From Table 5, all expected counts are greater than 5 and there is no matching of responses, so all necessary conditions for chi-square test of independence are satisfied.

From Table 6, the test statistic is χ^2 = 20.9316 with 4 degrees of freedom. The corresponding *P*-value is 0.0.0003 < 0.05. Therefore, there is a statistically significant relationship between incidence of clinical depression and the level of coffee consumption for women.

Form Table 5, the greatest contributions to the chi-square test statistic come from 2-3 cups per day (7.2084) and more than 4 cups per day (6.1496) for the diagnosed clinical depression category. In those two cases, actual counts were much lower than expected counts. It therefore appears that the incidence of clinical depression was lower for women who consumed at least 2 cups of coffee per day.

(d) One of the authors of this study was quoted on the NYTimes as saying it was 'too early to recommend that women load up on extra coffee' based on just this study. Do you agree with this statement? Explain your reasoning.

Yes I would agree. The study appears to be an observational study, so the causal link between higher coffee consumption and lower incidence of clinical depression cannot be claimed based on these results alone.

Appendix – SAS code

```
proc format;
      value Exercise 1='High' 2='Moderate' 3='Low';
      value Gender 1='Male' 2='Female';
      run;

      proc freq data=work.pulse_rates;
      tables Gender * Exercise / chisq exact expected cellchisq
      nocol norow nopercent;
      format Gender Gender. Exercise Exercise.;
      run;

data facebook;
      input Response $ Gender $ Count;
      datalines;
      Yes Male 288
      No Male 61
      NotSure Male 10
      Yes Female 378
      No Female 62
      NotSure Female 7
;
proc freq data=work.facebook;
      tables Response * Gender / chisq expected cellchisq nocol
      norow nopercent;
      weight Count;
      run;

data coffee;
      input Depression $ Consumption $ Count;
      datalines;
      Yes 1pw 670
      No 1pw 11545
      Yes 2to6pw 373
      No 2to6pw 6244
      Yes 1pd 905
      No 1pd 16329
      Yes 2to3pd 564
      No 2to3pd 11726
      Yes 4pd 95
      No 4pd 2288
;
proc freq data=work.coffee;
      tables Depression * COnsumption / chisq expected cellchisq
      nocol norow nopercent;
      weight Count;
      run;
```