School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT292 **Assignment 2 Solutions**

Note: Solutions must be either typed or written neatly, and questions must be answered in order. Where SAS is used to answer a question, relevant SAS output must be copied as an image file and included with your answer to the question and not at the end of the assignment.

1. **Consider data collected by Brockman (1996) on female horseshoe crabs and the number of male "satellites" residing near them. We will look at a subset of $n = 41$ of these female horseshoe crabs with the best spine condition. For this subset, the numbers of female horseshoe crabs reporting particular numbers of satellites are as shown in the table below.**

Satellites (r)	Frequency (f_r)
0	19
1	3
2	1
3	4
4	7
5	7

Source: Brockman, H.J. (1996). Satellite Male Groups in Horseshoe Crabs, *Limulus polyphemus*, *Ethology 102*(1):1-21.

   a. **Assuming the number of satellites per female horseshoe crab follows a Poisson distribution, estimate the mean number of satellites per female horseshoe crab.**

$$\begin{aligned}\widehat{\lambda} &= \frac{\sum_{i=0}^{5} r \times f_r}{n} \\ &= \frac{0 \times 19 + 1 \times 3 + \cdots + 5 \times 7}{19 + 3 + \cdots + 7} \\ &= \frac{80}{41} \approx 1.95122 \text{ satellites per female horseshoe crab.}\end{aligned}$$

   b. **Suppose we wish to test whether the distribution of the number of satellites per female horseshoe crab is consistent with a Poisson distribution. Can a chi-square goodness-of-fit test be applied to the data as presented in the table, or do certain numbers of satellites need to be grouped? If a grouping of numbers of satellites is necessary, determine an appropriate grouping, showing evidence that a**

chi-square goodness-of-fit test would indeed be appropriate for this grouping.

Expected frequencies are as shown in the table below.

Number of satellites (r)	Observed frequency (f_r)	$P(X = r)$	Expected frequency $\left(\widehat{f}_r = nP(X = r)\right)$
0	19	$\frac{1.95122^0 \exp(-1.95122)}{0!} \approx 0.1421$	$41 \times P(X = 0) \approx 5.82613$
1	3	$\frac{1.95122^1 \exp(-1.95122)}{1!} \approx 0.27727$	$41 \times P(X = 1) \approx 11.36805$
2	1	$\frac{1.95122^2 \exp(-1.95122)}{2!} \approx 0.27051$	$41 \times P(X = 2) \approx 11.09078$
3	4	$\frac{1.95122^3 \exp(-1.95122)}{3!} \approx 0.17594$	$41 \times P(X = 3) \approx 7.21352$
4	7	$\frac{1.95122^4 \exp(-1.95122)}{4!} \approx 0.08582$	$41 \times P(X = 4) \approx 3.51879$
≥ 5	7	$1 - \sum_{r=0}^{4} \frac{1.95122^r \exp(-1.95122)}{r!} \approx 0.04836$	$41 \times P(X \geq 5) \approx 1.98273$

Expected frequencies are less than 5 for both $r = 4$ and $r \geq 5$ satellites, so it would not be appropriate to carry out a chi-square goodness-of-fit test on the data as presented in the table. It is sufficient to combine $r = 4$ and $r \geq 5$ into a new grouping $r \geq 4$ satellites to produce all expected frequencies of at least 5, as demonstrated in the table below.

Number of satellites (r)	Observed frequency (f_r)	$P(X = r)$	Expected frequency $\left(\widehat{f}_r = nP(X = r)\right)$
0	19	$\frac{1.95122^0 \exp(-1.95122)}{0!} \approx 0.1421$	$41 \times P(X = 0) \approx 5.82613$
1	3	$\frac{1.95122^1 \exp(-1.95122)}{1!} \approx 0.27727$	$41 \times P(X = 1) \approx 11.36805$
2	1	$\frac{1.95122^2 \exp(-1.95122)}{2!} \approx 0.27051$	$41 \times P(X = 2) \approx 11.09078$
3	4	$\frac{1.95122^3 \exp(-1.95122)}{3!} \approx 0.17594$	$41 \times P(X = 3) \approx 7.21352$
≥ 4	14	$1 - \sum_{r=0}^{3} \frac{1.95122^r \exp(-1.95122)}{r!} \approx 0.13418$	$41 \times P(X \geq 3) \approx 5.50152$

c. **Test whether the number of satellites per female horseshoe crab is consistent with a Poisson distribution. Be sure to clearly state the null and alternative hypotheses, present the test statistic and its distribution under the null hypothesis, and report the p-value and your conclusion at the $\alpha = 0.05$ significance level.**

We wish to test the hypotheses

$$\begin{aligned}\mathcal{H}_0 &: \text{The population distribution is Poisson.}\\ \mathcal{H}_1 &: \text{The population distribution is not Poisson.}\end{aligned}$$

The test statistic is given by

$$\begin{aligned}X^2 &= \frac{\left(f_0-\widehat{f}_0\right)^2}{\widehat{f}_0}+\frac{\left(f_1-\widehat{f}_1\right)^2}{\widehat{f}_1}+\cdots+\frac{\left(f_{\geq 4}-\widehat{f}_{\geq 4}\right)^2}{\widehat{f}_{\geq 4}}\\ &\approx \frac{(19-5.82613)^2}{5.82613}+\frac{(3-11.36805)^2}{11.36805}+\cdots+\frac{(14-5.50152)^2}{5.50152}\\ &\approx 29.78838+6.15975+\cdots+13.12806\\ &\approx 59.68871\end{aligned}$$

The distribution for X^2 is χ^2_{k-m-1} where k is the number of unique expected frequencies calculated (5) and m is the number of parameters estimated (1). Thus,

$$\begin{aligned}X^2 &\sim \chi^2_{5-1-1}\\ &\sim \chi^2_3\end{aligned}$$

The p-value is given by

$$\begin{aligned}p\text{-value} &= P\left(\chi^2_{k-m-1} > X^2\right)\\ &\approx P\left(\chi^2_3 > 59.68871\right)\end{aligned}$$

This can be calculated using the following code:

```
data;
P_VALUE = 1 - cdf("CHISQUARE", 59.68871, 3);
proc print;
```

The resulting p-value is approximately 0.

Obs	P_VALUE
1	6.8512E-13

As the p-value is much smaller than any reasonable significance level (*e.g.*, $\alpha = 0.05$, $\alpha = 0.01$), we reject $\mathcal{H}_0$. Thus, there is strong evidence to conclude that the population distribution is not Poisson.

2. **Recall the dataset produced from a study carried out by the European CanCer Organisation and analysed in Assignment 1. In that study, a non-invasive diagnostic test for stomach and esophageal cancers was carried out on 335 people, and cancer statuses and test results for these people were as shown in the table below.**

Have stomach or esophageal cancer?	Tested positive for stomach or esophageal cancer?	
	No	Yes
No	140	32
Yes	32	131

a. **Using an odds ratio, describe and clearly interpret the association between cancer status and test result.**

The estimated odds ratio describing the association between cancer status and test result is given by

$$\begin{aligned}\widehat{\theta} &= \frac{n_{11}n_{22}}{n_{12}n_{21}} \\ &= \frac{140 \times 131}{32 \times 32} \\ &\approx 17.9102\end{aligned}$$

This represents the ratio of the odds of a negative test result for someone who does not have stomach or esophageal cancer to the odds of a negative test result for someone who has stomach or esophageal cancer, so we would estimate that the odds of a negative test result for someone who does not have stomach or esophageal cancer is approximately 17.9102 times that of someone who does have stomach or esophageal cancer.

b. **Obtain a 95% confidence interval for the odds ratio θ calculated in part (a).**

An approximate 95% confidence interval for θ is given by

$$\begin{aligned}&\left(\exp\left(\log\widehat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right),\right. \\ &\left.\exp\left(\log\widehat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right)\right) \\ \implies &\left(\exp\left(\log 17.9102 - 1.96 \times \sqrt{\frac{1}{140}+\frac{1}{32}+\frac{1}{32}+\frac{1}{131}}\right),\right. \\ &\left.\exp\left(\log 17.9102 + 1.96 \times \sqrt{\frac{1}{140}+\frac{1}{32}+\frac{1}{32}+\frac{1}{131}}\right)\right) \\ &\implies (10.3867, 30.8832)\end{aligned}$$

c. **Is it appropriate to carry out a chi-square test of independence for the data presented in the table? Briefly explain why or why not.**

Row and column marginals are as shown in the following table:

	Tested negative	Tested positive	Total
Does not have cancer	140	32	**172**
Has cancer	32	131	**163**
Total	**172**	**163**	**335**

These produce the expected frequencies shown below, the smallest of which is 79.3104, so all are larger than 5. Thus, it is appropriate to proceed with a chi-square test of independence.

	Tested negative	Tested positive
Does not have cancer	$\frac{172\times172}{335} \approx 88.3104$	$\frac{172\times163}{335} \approx 83.6896$
Has cancer	$\frac{163\times172}{335} \approx 83.6896$	$\frac{163\times163}{335} \approx 79.3104$

d. **Regardless of your answer to part (c), carry out both Pearson and likelihood ratio chi-square tests of independence to assess whether cancer status and test result are associated. Be sure to clearly state the null and alternative hypotheses, present the test statistic and its distribution under the null hypothesis, and report the p-value and your conclusion at the $\alpha = 0.05$ significance level.**

The hypotheses we wish to test are

$$\mathcal{H}_0 : \text{Cancer status and test result are independent}$$
$$\mathcal{H}_1 : \text{Cancer status and test result are not independent}$$

Pearson and likelihood ratio chi-square test statistics are

$$\begin{aligned} X^2 &\approx \frac{(140-88.3104)^2}{88.3104} + \frac{(32-83.6896)^2}{83.6896} + \cdots + \frac{(131-79.3104)^2}{79.3104} \\ &\approx 127.7932 \\ G^2 &\approx 2 \times \left(140 \log\left(\frac{140}{88.3104}\right) + 32 \log\left(\frac{32}{83.6896}\right) + \cdots + 131 \log\left(\frac{131}{79.3104}\right)\right) \\ &\approx 137.4419 \end{aligned}$$

both of which follow a $\chi^2_{(I-1)(J-1)} = \chi^2_{(2-1)(2-1)} = \chi^2_1$ distribution. Then the p-values corresponding to the Pearson and likelihood ratio chi-square tests are

given by

$$\begin{aligned} p\text{-value} &= P\left(\chi^2_{(I-1)(J-1)} > X^2\right) \\ &\approx P\left(\chi^2_1 > 127.7932\right) \approx 0 \qquad \text{(Pearson)} \\ p\text{-value} &= P\left(\chi^2_{(I-1)(J-1)} > G^2\right) \\ &\approx P\left(\chi^2_1 > 137.4419\right) \approx 0 \qquad \text{(Likelihood ratio).} \end{aligned}$$

These p-values can be produced by running the following code:

```
data;
P_VALUE_PEARSON = 1 - cdf("CHISQUARE", 127.7932, 1);
P_VALUE_LR = 1 - cdf("CHISQUARE", 137.4419, 1);
proc print;
```

Regardless of which of these two tests we use, we arrive at the same conclusion at the $\alpha = 0.05$ significance level. As the p-value is less than any reasonable significance level (and far less than $\alpha = 0.05$), we reject $\mathcal{H}_0$. Thus, there is strong evidence to conclude that cancer status and test result are dependent.

If we carry out the chi-square test in SAS (as well as select the **Measures** option for producing an estimate for the odds ratio and 95% confidence interval), we get the output shown below, which matches what we calculated by hand.

The FREQ Procedure

Table of HAS_CANCER by TEST_POSITIVE

		TEST_POSITIVE		
		No	Yes	Total
HAS_CANCER				
No	Frequency	140	32	172
	Expected	88.31	83.69	
	Percent	41.79	9.55	51.34
	Row Pct	81.40	18.60	
	Col Pct	81.40	19.63	
Yes	Frequency	32	131	163
	Expected	83.69	79.31	
	Percent	9.55	39.10	48.66
	Row Pct	19.63	80.37	
	Col Pct	18.60	80.37	
Total	Frequency	172	163	335
	Percent	51.34	48.66	100.00

Statistics for Table of HAS_CANCER by TEST_POSITIVE

Statistic	DF	Value	Prob
Chi-Square	1	127.7932	<.0001
Likelihood Ratio Chi-Square	1	137.4419	<.0001
Continuity Adj. Chi-Square	1	125.3329	<.0001
Mantel-Haenszel Chi-Square	1	127.4118	<.0001
Phi Coefficient		0.6176	
Contingency Coefficient		0.5255	
Cramer's V		0.6176	

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	17.9102	10.3867	30.8832
Relative Risk (Column 1)	4.1461	3.0145	5.7024
Relative Risk (Column 2)	0.2315	0.1678	0.3193

Sample Size = 335