38.5/40

Assignment 2

STAT 292 Applied Statistics 2A

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.

$\lambda = \hat{\lambda} = \frac{0\times19+1\times3+2\times1+3\times4+4\times7+5\times7}{41} = \frac{80}{41} = 1.951219512$

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.

- Calculate the probabilities of each observation:

$P(X = i) = \frac{\hat{\lambda}^x exp^{-\hat{\lambda}}}{x!}, x = 0, 1, ..., \text{and } 0 < \lambda < \infty <\$$

Satellites (r)	Frequency (fr)	r*fr	P(X=i)
0	19	0	0.1421006724
1	3	3	0.2772696047
2	1	2	0.2705069313
3	4	12	0.1759394675
4	7	28	0.08582413049
5	7	35	0.0334923436
>=5	7	35	0.04835919358
n=	41	80	
mean=	1.951219512		

5 is actually all values 5 and over which is calculated using $1 - P(X \leq 4)$

- Get the expected frequencies which is given by $n \times P(X = i)$, the number of observations multipled by the probability of each outcome:

Satellites (r)	Frequency (fr)	r*fr	P(X=i)	Expected Frequency
0	19	0	0.1421006724	5.826127568
1	3	3	0.2772696047	11.36805379
2	1	2	0.2705069313	11.09078419
3	4	12	0.1759394675	7.213518169
4	7	28	0.08582413049	3.51878935
>=5	7	35	0.04835919358	1.982726937
n=	41	80		
mean=	1.951219512			

- All expected frequencies should be greater than 5, if it's less than 5, group observations together until the expected frequencies of the grouped observations are over 5:

In this instance 4 and >=5 are less than 5 so we grouped them into >=4:

Satellites (r)	Frequency (fr)	r*fr	P(X=i)	Expected Frequency
0	19	0	0.1421006724	5.826127568
1	3	3	0.2772696047	11.36805379
2	1	2	0.2705069313	11.09078419
3	4	12	0.1759394675	7.213518169
>=4	14	63	0.1341833241	5.501516287
n=	41	80		
mean=	1.951219512			

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.

$H_0: $ The Population distribution is a Poisson distribution with mean $\lambda = 1.951219512$.

$H_1: $ The Population distribution is not a Poisson distribution with mean $\lambda = 1.951219512$.

The test statistic X^2 is given by $\sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$

$X^2 \approx \frac{(19-5.826127568)^2}{5.826127568} + \frac{(3-11.36805379)^2}{11.36805379} + ... + \frac{(14-5.501516287)^2}{5.501516287}$

$X^2 \approx 59.68871141$

Satellites (r)	Frequency (fr)	r*fr	P(X=i)	Expected Frequency	
0	19	0	0.1421006724	5.826127568	29.78838221
1	3	3	0.2772696047	11.36805379	6.159746034
2	1	2	0.2705069313	11.09078419	9.180949135
3	4	12	0.1759394675	7.213518169	1.431575935
>=4	14	63	0.1341833241	5.501516287	13.1280581
n=	41	80		X2=	59.68871141
mean=	1.951219512				

Distribution under H_0,

$X^2 \sim X^2_{5-1-1} \implies X^2 \sim X^2_3$

```
data;
P_PROB = 1 - cdf("CHISQUARE", 59.68871141, 3);
proc print;
```

Obs	P_PROB
1	6.8512E-13

+1 bonus for p-values

Since how p-value is well below our significance level of $\alpha = 0.05$, there is not sufficient eveidence that our distribution follows a Poisson distribution, thus rejecting our null hypothesis in favour of the alternate hypothesis that our distribution is not Poisson! 20

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.

$\hat{\theta} = \frac{\hat{p_{11}}\hat{p_{22}}}{\hat{p_{12}}\hat{p_{21}}} = \frac{(\frac{n_{11}}{n_{1+}})(\frac{n_{22}}{n_{2+}})}{(\frac{n_{12}}{n_{1+}})(\frac{(n_{21})}{n_{2+}})} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \approx \frac{140 \times 131}{32 \times 32} \approx \frac{18340}{1024} \approx 17.91015625$

A person that has stomach or esophageal cancer is 17.91 times more likely to test positive than negative, likewise a person that doesn't have the cancer is 17.91 times more likely to test negative than positive. In other words a correct result is 17.91 times more likely than a false result. -2.5

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

$log\hat{\theta} \approx log 17.91015625 \approx 2.885367940211003$

$log\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

when reporting CI smallest number goes first

$2.8853679402110039 \pm 1.960 \times \sqrt{\frac{1}{140} + \frac{1}{32} + \frac{1}{32} + \frac{1}{131}} \approx (3.43022130657, 2.34051457385)$

$(exp(3.43022130657), exp(2.34051457385)) \approx (30.88347671, 10.38657985)$

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.

It is appropriate because we're testing if two categorical variables are independent or dependent, in this case we're seeing if the test result versus the actual result are related or unrelated.

We're also assuming the data is collected via random sampling due to not having enough evidence to suggest otherwise.

It is also expected that the expected frequencies for each cell is above 5:

$EF_{ij} = \frac{i_+ \times_+ j}{n}$

$EF_{n_{11}} = \frac{n_{1+} \times n_{+1}}{n} = \frac{172 \times 172}{335} = 88.3104477612$

$EF_{n_{12}} = \frac{n_{1+} \times n_{+2}}{n} = \frac{172 \times 163}{335} = 83.6895522388$

$EF_{n_{21}} = \frac{n_{2+} \times n_{+1}}{n} = \frac{163 \times 172}{335} = 83.6895522388$

$EF_{n_{22}} = \frac{n_{2+} \times n_{+2}}{n} = \frac{163 \times 163}{335} = 79.3104477612$

All expected values are above 5 so we have nothing to suggest a chi-square test of independence would be inappropriate.

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.

need to report these using the right symbols

H_0: **Have stomach and esophageal cancer** and **Tested positive for stomach and esophageal cancer** are independent.

H_1: **Have stomach and esophageal cancer** and **Tested positive for stomach and esophageal cancer** are not independent.

Under H_0, we're assuming they're independent thus: $p_{ij} = P(X = i)P(X = j) = p_{i+}p_{j+}$,

Then the expected frequency for each cell is given by: $\mu_{ij} = np_{ij} = np_{i+}$ where $p_{i+} = \hat{p_{i+}} = \frac{n_{i+}}{n}$ and $p_{+j} = \hat{p_{+j}} = \frac{n_{+j}}{n}$

In other words: $\hat{\mu_{ij}} = n\hat{p}_{i+}\hat{p}_{+j} = n(\frac{n_{i+}}{n})(\frac{n_{+j}}{n}) = \frac{n_{i+}n_{+j}}{n}$

So for all the cells:

$\hat{\mu_{11}} = \frac{n_{1+}n_{+1}}{n} = \frac{172 \times 172}{335} = 88.3104477612$

$\hat{\mu_{12}} = \frac{n_{1+}n_{+2}}{n} = \frac{172 \times 163}{335} = 83.6895522388$

$\hat{\mu_{21}} = \frac{n_{2+}n_{+1}}{n} = \frac{163 \times 172}{335} = 83.6895522388$

$\hat{\mu_{22}} = \frac{n_{2+}n_{+2}}{n} = \frac{163 \times 163}{335} = 79.3104477612$

Which are all well above 5 so it satifies our condition.

Now we calculate the test statistic and the p-value:

Pearson X^2 test:

$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu_{ij}})^2}{\hat{\mu_{ij}}} = \frac{(n_{11} - \hat{\mu_{11}})^2}{\hat{\mu_{11}}} + \frac{(n_{12} - \hat{\mu_{12}})^2}{\hat{\mu_{12}}} + \frac{(n_{21} - \hat{\mu_{21}})^2}{\hat{\mu_{21}}} + \frac{(n_{22} - \hat{\mu_{22}})^2}{\hat{\mu_{22}}}$

$$X^2 = \frac{(140-88.3104477612)^2}{88.3104477612} + \frac{(32-83.6895522388)^2}{83.6895522388} + \frac{(32-83.6895522388)^2}{83.6895522388} + \frac{(131-79.3104477612)^2}{79.3104477612}$$

$$X^2 = 30.2547419743 + 31.9252491999 + 31.9252491999 + 33.6879930207 = 127.793233395$$

The p-value is given by: $P(X^2_{(I-1)(J-1)} > X^2) \approx P(X^2_{(2-1)(2-1)} > 127.793233395) \approx P(X^2_1 > 127.793233395)$

Likelihood ratio X^2 test:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log(\frac{n_{ij}}{\hat{\mu_{ij}}})$$

$$G^2 \approx 2 \times 140 log(\frac{140}{88.3104477612}) + 2 \times 32 log(\frac{32}{83.6895522388}) + 2 \times 32 log(\frac{32}{83.6895522388}) + 131 log(\frac{131}{79.3104477612})$$

$$G^2 \approx 129.019520226 + -61.5282075517 + -61.5282075517 + 131.478792794$$

$$G^2 \approx 137.441897917$$

Since the test statistics for both the Pearson and Likelihood Ratio tests are large it would be difficult to calculate it so we'll use SAS:

Statistics for Table of CANCER by POSITIVE_RESULT

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	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	140
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 335

distribution? -1

Which gives us a test statistic of Pearson: 127.7932 which lines up with what we got by hand, and Likelihood Ratio: 137.4418 which also lines up with what we got.

SAS gives us the p-value for both Pearson and Likelihood Ratio as: <0.001 which is far far below our confidance interview of 0.05, thus we reject the null hypothesis in favour of the alternate hypothesis that a person having cancer is not independent to the result of the test.

16.5
+1bonus for p-values