Categorial Data Analysis
Solution #1

---

1. Identify each variable as nominal or ordinal.

   (a) UK political party preference (Labour, Conservative, Social Democrat) Nominal

   (b) Anxiety rating (none, mild, moderate, sever, very severe) Ordinal

   (c) Clinic location (London, Boston, Madison, Rochester, Montreal) Nominal

   (d) Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression) Ordinal

   (e) Favorite beverage (water, juice, milk, soft drink, beer, wine) Nominal

2. Consider the statement, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children." For the 1996 General Social Survey, conducted by the National Opinion Research Center (NORC), 842 replied "yes" and 982 replied "no". Let p denote the population proportion who would reply "yes". Construct a 95% confidence interval for p. Interpret the result. Since the estimated value of probability p is given as

$$\hat{p} = \frac{842}{842 + 982} = 0.4616$$

   and its estimated standard error is

$$\widehat{s.e.}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.4616(1-0.4616)/(842+982)} = 0.0117,$$

   The 95% confidence interval can be obtained as

$$\hat{p} \pm 1.96\widehat{s.e.}(\hat{p}) = 0.4616 \pm (1.96)(0.0117) = (0.4387, 0.4845).$$

   This interval does not include 0.5 and upper bound is less than 0.5. So the less-than-a-half public supports a legal abortion.

3. Suppose that, to collect data in an introductory statistics course, recently I gave the students a questionnaire. One question asked each student whether he or she was a vegetarian. Of $n = 25$ students, 5 answered "yes". For testing $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$, First, note that

	Yes	No	Total
n_i	5	20	25
p_{0i}	0.5	0.5	1
e_i	12.5	12.5	25

(a) Provide the Pearson's chi-squared statistic, X^2

$$X^2 = \sum_{i=1}^{2} \frac{(n_i - e_i)^2}{e_i} = \frac{(5-12.5)^2}{12.5} + \frac{(20-12.5)^2}{12.5} = 9.$$

(b) Provide the likelihood ratio statistic, G^2

$$G^2 = 2\sum_{i=1}^{2} n_i \log\left(\frac{n_i}{e_i}\right) = 2\left\{5\log\left(\frac{5}{12.5}\right) + 20\log\left(\frac{20}{12.5}\right)\right\} = 9.6372$$

(c) Conduct hypothesis test for H_0 with both of X^2 and G^2. (This means you should report p-values for them) Both of statistics have $\chi^2(2-1)$. Based on $\chi^2(1)$, p-value of X^2 is $\Pr(\chi^2(1) \geq 9) = 0.0027$ and p-value of G^2 is $\Pr(\chi^2(1) \geq 9.6372) = 0.0019$. Under $\alpha = 0.05$, both of tests reject the null hypothesis.

4. The below table shows the results of a study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates whether the cancer was controlled for at least two years following treatment.

	Cancer Controlled	Cancer not controlled
Surgery	63	6
Radiation therapy	45	9

Researchers are interested in comparing the proportions of cancer control between two methods. Answer the following questions

(a) Let us denote the proportion of cancer control with surgery by p_1 and that with radiation therapy by p_2. What are $\hat{p}_1$ and $\hat{p}_2$?

$$\hat{p}_1 = \frac{63}{63+6} = 0.9130 \quad \text{and} \quad \hat{p}_2 = \frac{45}{45+9} = 0.8333.$$

(b) Give a 95% confidence interval for $p_1 - p_2$.
Note that

$$\widehat{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{(0.9130)(1-0.9130)}{63+6} + \frac{(0.8333)(1-0.8333)}{45+9}} = 0.0610.$$

Thus, 95% confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\widehat{s.e.}(\hat{p}_1 - \hat{p}_2) = (0.9130 - 0.8333) \pm (1.96)(0.0610) = (-0.0399, 0.1993).$$

(c) Compute the relative risk of cancer control for group "Surgery" relative to group "Radiation therapy" and interpret it.

$$RR = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.9130}{0.8333} = 1.0956.$$

The proportion of cancer control with surgery is 1.0956 times higher than that with radiation therapy.

(d) Compute the odds of cancer control for group "Surgery" and the odds of cancer control for group "Radiation therapy".
The odds for group "Surgery" is

$$\text{odds}_1 = \frac{\hat{p}_1}{(1-\hat{p}_1)} = \frac{0.9130}{1-0.9130} = 10.4923.$$

And the odds for group "Radiation therapy" is

$$\text{odds}_2 = \frac{\hat{p}_2}{(1-\hat{p}_2)} = \frac{0.8333}{1-0.8333} = 4.9988.$$

(e) Provide the odds ratio of cancer control for group "Surgery" relative to group "Radiation therapy". And interpret it.
The odds ratio for group "Surgery" relative to group "Radiation therapy" is

$$\hat{\theta} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{10.4923}{4.9988} = 2.0990.$$

So the odds for group "Surgery" is twice larger than the odds for group "Radiation therapy". Or surgery is twice better than radiation therapy in terms of odds.

(f) Give a 95% confidence interval for the odds ratio.
The standard error of $\log\hat{\theta}$ is

$$\widehat{s.e.}(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} = \sqrt{\frac{1}{63}+\frac{1}{6}+\frac{1}{45}+\frac{1}{9}} = 0.5620.$$

The lower bound for 95% confidence interval for odds ratio is

$$\frac{\hat{\theta}}{\exp(1.96\widehat{s.e.}(\log\hat{\theta}))} = \frac{2.0990}{e^{(1.96)(0.5620)}} = 0.6976$$

and the upper bound for 95% confidence interval for odds ratio is

$$\hat{\theta} \times \exp(1.96\widehat{s.e.}(\log\hat{\theta})) = (2.0990)e^{(1.96)(0.5620)} = 6.3153.$$

Thus, (0.6976, 6.3153).

(g) Now we are interested in the independence between method and response. Based on the result of odds ratio, are they independent? Since 95% confidence interval contains 1, we fail to reject the null hypothesis or claim that method and response are independent.

(h) We are still interested in the independence. Conduct X^2 and G^2 tests. Interpret the results. Under the independence assumption or hypothesis, we expect that

$$\begin{aligned} p_{11} &= p_{1\cdot} \times p_{\cdot 1} = \tfrac{63+6}{123} \cdot \tfrac{63+45}{123} = 0.4926, \\ p_{12} &= p_{1\cdot} \times p_{\cdot 2} = \tfrac{63+6}{123} \cdot \tfrac{6+9}{123} = 0.0684, \\ p_{21} &= p_{2\cdot} \times p_{\cdot 1} = \tfrac{45+9}{123} \cdot \tfrac{63+45}{123} = 0.3855, \\ p_{22} &= p_{2\cdot} \times p_{\cdot 2} = \tfrac{45+9}{123} \cdot \tfrac{6+9}{123} = 0.0535. \end{aligned}$$

Based on the above, the expected cell counts are

$$e_{11} = n \times p_{11} = (123)(0.4926) = 60.5898,$$

$$e_{12} = n \times p_{12} = (123)(0.0684) = 8.4132,$$

$$e_{21} = n \times p_{21} = (123)(0.3855) = 47.4165,$$

$$e_{22} = n \times p_{22} = (123)(0.0535) = 6.5805.$$

Thus,

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 1.8008$$

and

$$G^2 = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right) = 1.7869.$$

Both of statistics have the chi-squared distribution with degrees of freedom $(2-1)(2-1) = 1$. Since $\chi_{0.05}(1) = 3.8415$, and X^2 and G^2 are less than this cut-off value, we fail to reject the null hypothesis that two variables are independent. (Or you may report their p-values or confidence intervals by yielding the same conclusion)

5. A study from the University of Texas Southwestern Medical Center examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. Hepatitis C causes about 10,000 deaths each year in the United States, but often lies undetected for years after infection. The data from this study can be summarized in a two-way table, as follows:

	Hepatitis C	No Hepatitis C	Total
Tatto, parlor	17	35	52
Tatto, elsewhere	8	53	61
None	22	491	513
Total	47	579	626

Under the independence hypothesis, we expect that

$$\begin{aligned}
p_{11} &= p_{1+} \times p_{+1} = \tfrac{52}{626} \cdot \tfrac{47}{626} = 0.0062, \\
p_{12} &= p_{1+} \times p_{+2} = \tfrac{52}{626} \cdot \tfrac{579}{626} = 0.0768, \\
p_{21} &= p_{2+} \times p_{+1} = \tfrac{62}{626} \cdot \tfrac{47}{626} = 0.0074, \\
p_{22} &= p_{2+} \times p_{+2} = \tfrac{62}{626} \cdot \tfrac{579}{626} = 0.0916, \\
p_{31} &= p_{3+} \times p_{+1} = \tfrac{513}{626} \cdot \tfrac{47}{626} = 0.0615, \\
p_{32} &= p_{3+} \times p_{+2} = \tfrac{513}{626} \cdot \tfrac{579}{626} = 0.7580
\end{aligned}$$

Thus, the expected cell counts are

$$\begin{aligned}
e_{11} &= n \times p_{11} = (626)(0.0062) = 3.8192, \\
e_{12} &= n \times p_{12} = (626)(0.0768) = 48.0768, \\
e_{21} &= n \times p_{21} = (626)(0.0074) = 4.6324, \\
e_{22} &= n \times p_{22} = (626)(0.0916) = 57.3416, \\
e_{31} &= n \times p_{31} = (626)(0.0615) = 38.4990, \\
e_{32} &= n \times p_{32} = (626)(0.7580) = 474.5080.
\end{aligned}$$

(a) Using X^2, test the hypothesis of independence between tattoo status and Hepatitis C infection status. Report the p-value and interpret.
The Pearson's chi-squared statistic becomes

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 59.4672.$$

Since the degrees of freedom is $(3-1) \times (2-1) = 2$, the cut-off value with $\alpha = 0.05$ is $\chi^2_{0.05}(2) = 5.9915$ which is less than X^2. Therefore, we reject the null hypothesis of independence between tattoo status and Hepatitis C infection.

(b) Using G^2, test the hypothesis of independence between tattoo status and Hepatitis C infection status. Report the p-value and interpret. The likelihood ratio statistic becomes

$$G^2 = 2 \sum_{i=1}^{3} \sum_{j=1}^{2} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right) = 37.8710.$$

Since the degrees of freedom is $(3-1) \times (2-1) = 2$, the cut-off value with $\alpha = 0.05$ is $\chi^2_{0.05}(2) = 5.9915$ which is less than G^2. Therefore, we reject the null hypothesis of independence between tattoo status and Hepatitis C infection.

6. (**SAS problem**) The table below lists results from a simple random sample of front-seat occupants involved in car crashes (based on data from "Who Wants Airbags?" by Meyer and Finney, *Chance*, Vol. 18 No. 2). Use a 0.05 significance level to test the claim that the fatality rate of occupants is different for those in cars equipped with airbags.

	Occupant Fatalities	Occupant Survivals	Total
Airbag Available	41	11,500	11,541
No Airbag Available	52	9,801	9,853

Using SAS system, answer the followings:

(a) Provide SAS DATA step to create a data set.See the below

(b) Provide relative risk of fatality for group "Airbag Available" relative to group "No Airbag Available" and its 95% confidence interval. Interpret relative risk and its confidence interval. RR is 0.6731 and its 95% confidence interval is $(0.4474, 1.0127)$. The probability of fatality for group "Airbag Available" is 0.6731 times lower than that for group "No Airbag Available". However its evidence is not statistically significant under $\alpha = 0.05$ since 95% confidence interval includes 1.

(c) Provide odds of fatality for each group and interpret them. From the table below, the probability of fatality for group "Airbag Available" is $p_1 = 0.0036$ so the odds of fatality for group "Airbag Available" is $p_1/(1-p_1) = 0.0036$. And the probability of fatality for group "No Airbag Available" is $p_2 = 0.0053$ so the odds of fatality for group "No Airbag Available" is $p_2/(1-p_2) = 0.0053$.

(d) Provide odds ratio of fatality for group "Airbag Available" relative to group "No Airbag Available" and 95% confidence interval. Interpret them. $\theta = 0.6720$ and its 95% confidence interval is $(0.4459, 1.0128)$. The odds of fatality for group "Airbag Available" is 0.6720 times lower than that for group "No Airbag Available" but its evidence is not statistically significant under $\alpha = 0.05$ since 95% confidence interval includes 1.

(e) Provide X^2 and G^2 test results for testing whether airbag status and fatality are associated and interpret them. $X^2 = 3.6544$ and its p-value is 0.0559 and $G^2 = 3.6402$ and its p-value is 0.0564. Both of p-values are close to 0.05 but slightly larger. So under significant level $\alpha = 0.05$, both of tests fail to reject the null hypothesis that airbag use is not associated with fatality from car crashes.

Append necessary SAS code should be attached in the end of the answer of each question.

```
data airbag;
input airbag $ status $ count;
cards;
Yes Fatal 41
Yes Survival 11500
No Fatal 52
No Survival 9801
;
run;
```

```
proc freq data=airbag order=data;
weight count;
tables airbag*status / chisq measures;
run;
```

Table of airbag by status

airbag	status		
Frequency | Percent | Row Pct | Col Pct	Fatal	Survival	Total
Yes	41	11500	11541
	0.19	53.75	53.95
	0.36	99.64	
	44.09	53.99	
No	52	9801	9853
	0.24	45.81	46.05
	0.53	99.47	
	55.91	46.01	
Total	93	21301	21394
	0.43	99.57	100.00

Statistics for Table of airbag by status

Statistic	DF	Value	Prob
Chi-Square	1	3.6544	0.0559
Likelihood Ratio Chi-Square	1	3.6402	0.0564
Continuity Adj. Chi-Square	1	3.2667	0.0707
Mantel-Haenszel Chi-Square	1	3.6542	0.0559
Phi Coefficient		-0.0131	
Contingency Coefficient		0.0131	
Cramer's V		-0.0131	

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.6720	0.4459	1.0128
Cohort (Col1 Risk)	0.6731	0.4474	1.0127
Cohort (Col2 Risk)	1.0017	0.9999	1.0035

Sample Size = 21394