#### Stat 557 Assignment 2 Name \_\_\_Yifan Zhu\_\_\_\_\_

**Fall 2017**

Reading Assignment: Agresti, Chapters 3 and 4, Section 6.4 and Section 17.2

Written Assignment: Due Tuesday, October 3 in class.

1. The data in the following table were obtained from a sample of 1397 respondents from the population of adults (people at least 18 years old) in the United States in 1982. Assume that simple random sampling was used. Each respondent was cross-classified with respect to opinions expressed on the issues of gun control and imposing the death penalty on criminals convicted of certain violent acts.

|  |  |  |
| --- | --- | --- |
|  | Death Penalty | |
| Gun Registration | Favor | Oppose |
| Favor | 784 | 236 |
| Oppose | 311 | 66 |

1. What is the distribution of the possible 2×2 tables of counts that could be observed in such a survey of 1397 respondents?

It could be a multinomial distribution . in this problem.

1. Report maximum likelihood estimates of the expected counts under the null hypothesis that gun registration opinion is held independently of the death penalty opinion.

Let corresponds to Gun Registration and corresponds to Death Penalty. For and , 1 means favor and 2 means oppose. Then we have

1. Report values of the deviance G² and Pearson statistic X² for testing the independence hypothesis in part (b) against the general alternative. Report degrees of freedom and p-values. State your conclusion.
2. This exercise provides practice with nested models. For a 2×2 contingency table

|  |  |  |
| --- | --- | --- |
|  | Column Factor  j=1 j=2 | |
| Row i=1 | Y11 | Y12 |
| Factor i=2 | Y21 | Y22 |

obtained from a single simple random sample of size n, consider testing the following null hypothesis (call it model A)

Ho: 

for some unknown , against the general alternative (call if model C).

HA:  and 

Model A imposes both independence between the row and column factors and also imposes identical marginal distributions for the row and column factors.

1. Assuming the null hypothesis is true, give a formula for the log-likelihood function.
2. Give a formula for the maximum likelihood estimator  for .
3. Derive formulas for the mle’s for the expected counts in the  table under model A. Use the notation  to denote the mle’s of the expected counts for this model.
4. Compute the value of the deviance statistic,  for testing the null hypothesis that model A is correct against the general alternative, and report its degrees of freedom.
5. Consider another model (call it model B) that only imposes that the marginal distributions for the row and column factors are the same, i.e.,  for i=1, 2. Derive formulas for the mle’s for the expected cell counts (call them ) for this model. Give a formula for the deviance statistic,  for testing the null hypothesis that model B is correct against the general alternative, and report its degrees of freedom.
6. Note that Model A is nested in Model B in the sense that Model A can be obtained from Model B by imposing an additional constraint on the cell probabilities corresponding to independence between the row and column factors. Model B is nested in Model C, the general alternative, that only assumes that the sampling distribution for the counts in the table is a multinomial distribution. Complete the following analysis of deviance table for these three models using the data in the 2×2 table in Problem 1.

**Comparison d.f. deviance value p-value**

────────────────────────────────────────────────

Model A vs Model B

Model B vs Model C

─────────────────────────────────────────────────

Model A vs Model C

Summarize your conclusions from this analysis.

1. Rutherford and Gieger (1910, Phil. Mag. Sixth Ser., 20, 698-704) present data on the numbers of scintillations due to radioactive decay of polonium in each on n=2608 times intervals of length 1/8 minute.

|  |  |  |  |
| --- | --- | --- | --- |
| **Number**  **of**  **Scintillations** | **Number of**  **Time Intervals** | **Expected Counts**  **Poisson Model Neg. Binomial Model** | |
| 0  1  2  3  4  5  6  7  8  9  10  11  12 | 57  203  383  525  532  408  273  139  45  27  10  4  0 |  |  |
| 13  14 | 1  1 |  |  |

1. Assuming that the model with i.i.d. Poisson(m) random variables is correct, find the maximum likelihood estimate of the common mean count (m).
2. Using your solution to part (a) compute values of the maximum likelihood estimates of the expected counts as indicated in the third column of the table above. Use those values to evaluate the Pearson chi-square test for assessing the fit of the i.i.d. Poisson model. Combine the categories if necessary to keep estimates of expected counts larger than 2. Report values for the Pearson statistic, degrees of freedom, and a p-value. State your conclusion.
3. As an alternative assessment of the i.i.d. Poisson model evaluate Fisher's dispersion index. Report values for the statistic, degrees of freedom, and a p-value. State your conclusion.
4. Construct a 95% confidence interval for the mean number of scintillations per minute. Show how your confidence interval was constructed.
5. Consider a negative binomial model for the radioactive decay data. Report the maximum likelihood estimates of the expected counts in the fourth column of the table. Use the Pearson chi-square statistic to assess the fit of the negative binomial model. Combine categories if necessary to keep estimates of expected counts larger than 2. Report values for the Pearson statistic, degrees of freedom, and a p-value. State your conclusion. (If the algorithm provided in class fails to converge, you do not have to complete this problem, but you should explain why the algorithm failed.)
6. Edwards and Gurland (1961) present the following data on the numbers of accidents sustained by 166 London bus drivers over a period of two years.

|  |  |
| --- | --- |
| Number of  Accidents | Number of  Drivers |
| 0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15 | 15  32  26  29  22  19  9  8  3  1  0  0  1  0  0  1 |

1. Assuming that the model with i.i.d. Poisson(m) random variables is correct, find the maximum likelihood estimate of the mean number of accidents (m) per driver.
2. Using your solution to part (a) compute values of the maximum likelihood estimates of the expected counts for the 16 accident categories (the last category is at least 15 accidents). Use those values to evaluate the Pearson chi-square test for assessing the fit of the i.i.d. Poisson model. Combine categories if necessary to keep estimates of expected counts larger than 2. Report values for the Pearson statistic, degrees of freedom, and a p-value. State your conclusion.
3. As an alternative assessment of the i.i.d. Poisson model evaluate Fisher's dispersion index. Report values for the statistic, degrees of freedom, and a p-value. State your conclusion.
4. Using the Poisson model, even if it is not appropriate for these data, construct a 95% confidence interval for the mean number of accidents per driver. Show how your confidence interval was constructed.
5. Consider a negative binomial model for the accident data. Report values of maximum likelihood estimates for the model parameters.
6. Compute values of maximum likelihood estimates of the expected counts for the negative binomial model for all 16 accident categories. Use the Pearson chi-square statistic to assess the fit of the negative binomial model. Combine categories if necessary to keep estimates of expected counts larger than 2. Report values for the Pearson statistic, degrees of freedom, and a p-value. State your conclusion.
7. Using the negative binomial model, estimate the mean number of accidents per driver. Also report a standard error for your estimate and an approximate 95% confidence interval for the mean number of accidents per driver.
8. In a study of the potential effect of tonsillectomy on the risk of contracting Hodgkin’s disease, Vianna, et al. (1971, Lancet, 1, 431-432) obtained medical records for a sample of 109 patients diagnosed with Hodgkin's disease. A sample of 109 "control" patients was selected from hospital records of patients with no history of Hodgkin's disease or any other malignant disease or chronic illness. The control patients were selected from a set of hospital records that generally matched the composition of the group of patients with Hodgkin's disease with respect to age, sex, race, county of residence, and date of hospital admission. Eight of the patients with Hodgkin's disease and two control patients were not included in the data show below because their tonsillectomy history could not be obtained. The remaining 208 patients were cross-classified into the following 2×2 contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Had**  **Tonsillectomy** | **Did not have**  **Tonsillectomy** | **Totals** |
| **Hodgkin’s**  **Disease** | 67 | 34 | 101 |
| **Controls** | 43 | 64 | 107 |

The value of the Pearson chi-square test for independence is 14.26 on 1 d.f. and the p-value is 0.00016. Vianna, et. al., used the odds ratio  as an approximate measure of relative risk and concluded that tonsillectomy increases the risk of contracting Hodgkin's disease by a factor of nearly 3. They concluded that tonsillectomy removes a protective barrier against Hodgkin's disease.

1. Is this a retrospective or a prospective study?
2. Compute an approximate 95% confidence interval for the odds ratio.

(c) A year later, Johnson and Johnson (1972, New England Journal of Medicine, 287, 1122-1125) reported results from a different study of 175 patients treated for Hodgkin's disease at the Radiation Branch of the National Cancer Institute. There was information available on 472 siblings for the 175 patients. The authors chose the closest sibling of the same sex within five years of age of each patient. This matching reduced the data to 85 patient-sibling pairs and the following table was reported.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Had**  **Tonsillectomy** | **Did not have**  **Tonsillectomy** | **Totals** |
| **Hodgkin’s**  **Disease** | 41 | 44 | 85 |
| **Controls** | 33 | 52 | 85 |

The Pearson chi-square test for independence was computed as 1.53 with p-value = 0.22, and the estimated odds ratio is  with 95% confidence bounds (0.80, 2.70). Using these results, Johnson and Johnson claim to have refuted the contention of Vienna, et al., that tonsils provide a lymphoid barrier to Hodgkin's disease.

Which authors, if any, do you agree with? State your reasons. If you think any mistakes were made in either of the analyses, describe the mistakes and explain how the data should be analyzed, even if you do not have enough information to actually perform the analysis.

1. The data in Table 3.14 on page 104 in Agresti's book, Categorical Data Analysis, are

|  |  |  |
| --- | --- | --- |
|  | Cancer  Controlled | Cancer Not  Controlled |
| Surgery | 21 | 2 |
| Radiation Therapy | 15 | 3 |

Assume that the 41 larynx cancer patients were randomly assigned to the two treatments. Use Fisher's exact test to test the null hypothesis that the two treatments are equally effective in controlling the cancer against the alternative that the treatments are not equally effective. Report a p-value and state your conclusion.

1. The following data are fictitious results for 120 individuals who were cross-classified with respect to lung capacity and intensity of smoking habits. Note that the estimates of the expected counts under the null hypothesis of independence do not quite satisfy Cochran’s rule.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Smoking Habit | | | |
| Lung Capacity | None | Occasional | Regular | Heavy |
| Normal | 48 | 26 | 22 | 8 |
| Impaired | 2 | 4 | 6 | 4 |
| TOTALS | 50 | 30 | 28 | 12 |

1. Use the likelihood ratio G2 statistic to test the null hypothesis that lung capacity is independent of smoking habit. Report values for G2, degrees of freedom, and a p-value.
2. Perform an exact conditional test of the null hypothesis in part (a). Report the “exact” p-value. (Use G2 values to order the possible tables.)
3. Repeat Part (a) using the Pearson X2 statistic.
4. Perform an exact conditional test of the null hypothesis in Part (a). Report the “exact” p-value. (Use X2 values to order the possible tables.)
5. Use SAS or R code to simulate 50,000 possible tables of counts, conditioning on the row and column totals, compute an approximate “exact” p-value. Use the value of the Pearson statistic to order the tables. Compare the results with the results in parts (a), (b), (c) and (d).

8. In 1974, the Danish National Institute for Social Science Research interviewed a random sample of Danes between 20 and 69 years old in order to investigate the general welfare in Denmark, The following two tables (Andersen, 1990) cross-classify workers with respect to the physical and psychological demands of the employment. There are separate tables for males and females.

Table 1: Females

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Work is psychologically demanding | | |
|  |  | Usually | Sometimes | Seldom |
| Work is physically demanding | Usually | 100 | 109 | 202 |
|  | Sometimes | 33 | 89 | 179 |
|  | Seldom | 100 | 179 | 542 |

Table 2: Males

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Work is psychologically demanding | | |
|  |  | Usually | Sometimes | Seldom |
| Work is physically demanding | Usually | 113 | 163 | 370 |
|  | Sometimes | 45 | 106 | 280 |
|  | Seldom | 229 | 343 | 568 |

1. Conditional on the gender of the respondents is there any association between attitudes toward physical and psychological demands of employment? For each table, compute values of the Pearson  and the statistics. Report degrees of freedom and p-values. State your conclusions.
2. Use the Goodman-Kruskal gamma statistic to quantify the level of association between attitudes about the physical and psychological demands of work for females and males. Report a standard error for each estimate and use the large sample normal approximation to the distribution of the gamma statistic to construct approximate 95% confidence intervals.
3. The sample size may need to be very large before the sampling distribution of the Goodman-Kruskal gamma  statistic is reasonably well approximated by its limiting normal distribution, especially if  is large. A transformation that approaches its asymptotic normal distribution more rapidly is = 1/2 log. (Note that this is a transformation R.A. Fisher proposed for correlation coefficients.)

(a) Use the delta method to obtain a formula for the large sample variance of  as a function of the large sample variance for .

(b)  also has a limiting normal distribution. Use this fact and the result from part (a) to construct an approximate 95% confidence intervals for the difference between the gamma values for the two tables in Problem 8. Does your confidence interval indicate a significant difference between the levels of association (as measured by gamma) between attitudes toward the physical and psychological demands of work for men and women?

1. Construct a 95% bootstrapped confidence interval for the difference between men and women respected the levels of association (as measured by gamma) between attitudes toward the physical and psychological demands of work. Compare this result to the result from part (b).
2. In a study of disparities between mother and child perceptions of ability, sixth grade children were asked to rate their own academic ability. The mother of each child also was asked to rate the child’s academic ability. Separate tables are reported for white and black children. Each count in a table corresponds to one mother-child pair.

Table 1. White Children

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | Child’s Rating |  |
| Below Average | | | Average | Above Average |
|  | Below Average | 9 | 10 | 5 |
| Mother’s | Average | 26 | 6 | 13 |
| Rating | Above Average | 10 | 17 | 10 |

Table 2. Black Children

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | Child’s Rating |  |
| Below Average | | | Average | Above Average |
|  | Below Average | 10 | 5 | 10 |
| Mother’s | Average | 31 | 10 | 4 |
| Rating | Above Average | 22 | 18 | 9 |

1. To quantify the level of agreement between mother and child perception of academic ability, estimate Cohen’s Kappa and compute a 95% confidence interval for Cohen’s Kappa for each table. Use the large sample normal approximation to the sampling distribution of the kappa statistic to construct the confidence intervals. Is there more than random agreement in either table?
2. Repeat part (a) using bootstrapped confidence intervals. Compare results with part (a).
3. Compute a 95% confidence interval for the difference in the Kappa measures of agreement for the two tables. (Assume the statistics for the two tables are independent.) State your conclusion. Is the level of agreement between ratings given by mother and child the same for white and black families?
4. Cooper, et. al., (1972, see Fleiss) examined data on the proportions of patients diagnosed as schizophrenic by resident hospital psychiatrists in New York and London. The data were obtained from three independent studies. Counts from the three studies are shown in the following tables. The ranges of ages for the patients were not the same for the three studies as also indicated below.

Study 1 (ages 20-34)

|  |  |  |  |
| --- | --- | --- | --- |
| Schizophrenia Diagnosis | | | |
| City | Yes | No | Total |
| New York | 81 | 24 | 105 |
| London | 34 | 71 | 105 |

Study 2 (ages 20-59)

|  |  |  |  |
| --- | --- | --- | --- |
| Schizophrenia Diagnosis | | | |
| City | Yes | No | Total |
| New York | 118 | 74 | 192 |
| London | 69 | 105 | 174 |

Study 3 (35-59)

|  |  |  |  |
| --- | --- | --- | --- |
| Schizophrenia Diagnosis | | | |
| City | Yes | No | Total |
| New York | 82 | 63 | 145 |
| London | 52 | 93 | 145 |

1. Consider the odds of a schizophrenic diagnosis in New York relative to the odds of a schizophrenic diagnosis in London. While not exactly equal, the odds ratios are in the same direction for all three studies, and it may be reasonable to estimate the supposed common odds ratio. Report the value of the MH estimator for the common conditional odds ratio, conditioning on the three studies.
2. Construct and interpret a 95% confidence interval for the common conditional odds ratio.
3. Perform the Breslow-Day test of the null hypothesis that the population odds ratios are the same. Report the value of the test statistics, a p-value, and state your conclusion.
4. Use the COMOR option in the EXACT statement for the FREQ procedure in SAS to compute and “exact” 95% confidence interval for the common conditional odds ratio.
5. If it can be computed, report the value of Zelen’s exact test of homogeneous conditional odds ratios. Report a p-value and state your conclusion.
6. Is estimation of a common conditional odds ratio a reasonable thing to do for these data? Explain.