STAT4103: Categorical Data Analysis

Semester 2, 2017

Assignment 1

This assignment is due no later than 5pm, August 25, 2017, unless otherwise stated. Show all working out and include all R code and output as part of your submission.

You can hand deliver to me, or the School office (V123), or email me your assignment no later than this date. Late submissions will be penalised 10% of the total value of the assignment for each day it is late.

Question 1 – The Odds Ratio and Edward's Criteria **(20 marks)**

i) Using the delta method, derive an expression for the variance of $\sqrt{\theta}$. *(5 marks)*

ii) Suppose we consider the transformation

$$g(\theta) = \frac{\theta^b - 1}{\theta^b + 1}$$

which satisfies Edward's criteria. Prove that the variance of g(θ) is

$$\mathrm{Var}(g(\theta)) = \frac{b^2}{4}\left[(1 - g(\theta))(1 + g(\theta))\right]^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

(6 marks)

iii) Given the variance of g(θ) derived in part ii), modify the R code `edward.odds.exe` seen in Week 3 lectures to incorporate the calculation of the variance of g(θ). Use this code to calculate the value, and its variance, of Yule's Q, Yule's Y, Digby's H and Edward's J. *(9 marks)*

Question 2 – Common Odds Ratio's for Stratified 2x2 Tables (20 marks)

Suppose we have G stratified 2x2 contingency tables where, for each contingency table, the odds ratio can be determined. However, a "common odds ratio" can also be derived for all G tables.

i) The two most popular "common odds ratio's" are the Mantel-Haenszel (MH) estimate (Mantel and Haenszel, 1959) and the Woolf estimator (Woolf, 1995). Define these odds ratio estimators and describe how they calculate the common odds ratio. *(5 marks)*

ii) Cochran (1954) proposed a similar test statistic to Mantel and Haenszel (1959). As a result a third variation of the common odds ratio is known as the Cochran-Mantel-Haenszel (CMH) statistic. Define this statistic and state how it differs to what Mantel and Haenszel (1959) proposed. *(5 marks)*

iii) Consider Table 6.9 on page 226 of Agresti (2013). It consists of 8 2x2 contingency tables that examine the success or failure of a drug (and a placebo) in a clinical trial across 8 centres. For this data calculate the odds ratio for each of the centres, and determine the MH, Woolfe and CMH statistic. Where zero cell frequencies are present describe how you dealt with those. *(10 marks)*

Agresti, A. (2013), *Categorical Data Analysis* (3rd ed), Chichester: Wiley.

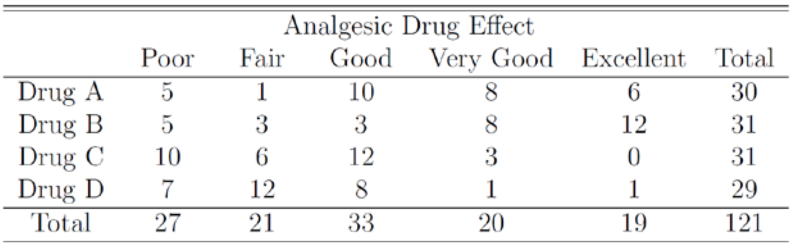
Cochran, W. G. (1954), Some methods of strengthening the common χ^2 tests. *Biometrics*, 10, 417 – 451.

Mantel, N. and Haenszel, W. (1959), Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, 22, 719 – 748.

Woolf, B. (1955), On estimating the relation between blood group and disease. *Annals of Human Genetics (London)*, 19, 251 – 153.

Question 3 – The Divergence Statistic **(30 marks)**

Suppose we consider the following two-way contingency table which was originally seen in Calimlin et. al (1982). The study was aimed at testing four analgesic drugs (randomly assigned the labels A, B, C and D) and their effect on 121 hospital patients. The patients were given a five point scale consisting of the categories Poor, Fair, Good, Very Good and Excellent on which to make their decision. The data is summarised below and should be entered into R as the object `drug.dat`.



	Analgesic Drug Effect					
	Poor	Fair	Good	Very Good	Excellent	Total
Drug A	5	1	10	8	6	30
Drug B	5	3	3	8	12	31
Drug C	10	6	12	3	0	31
Drug D	7	12	8	1	1	29
Total	27	21	33	20	19	121

i) Using `chisq.test()` in R, test the determine whether there is a statistically significant association between drug and their effectiveness. Ensure you specify the hypotheses, chi-squared statistic, degrees of freedom and p-value of your test.

(5 marks)

ii) Using R, compare the Monte-Carlo p-value’s when randomly generating 10, 100, 1000 and 10000 contingency tables. *(6 marks)*

iii) Using R, write an R function `divergCR.exe` that calculates the chi-squared value and its Monte-Carlo p-value of each member of the Cressie-Read divergence statistic discussed in the lecture. Provide a figure showing the distribution of each statistic and comment on the relative accuracy of these statistics for `drug.dat`. *(12 marks)*

iv) Write an R function `plotdivergCR.exe` that plots the Cressie-Read divergence statistic versus $\lambda \in [-1, 1]$ for `drug.dat`. From this plot, describe how the members of the divergence statistic (visually) compare and discuss the validity of specific values of λ for the contingency table. *(7 marks)*

Question 4 – Alternatives Measures of Association **(15 marks)**

The following output seen in Week 2 lectures is based on the execution of the R function `monte.study.exe` to Galton’s fingerprint data and provides the Monte-Carlo p-value for each measure of association based on the simulation of 10000 contingency tables. Write R code that verifies these calculations for Galton’s fingerprint data. Provide a comparison of the indices and their distribution of the randomly generated values. Provide an argument that explains which of them provides a good indication of the association and those that don't.

```
> monte.study.exe(fingerprint.dat, 10000)
$Output
             Value MC.P-value
Chi-sq     11.1699 0.031
Belson     48.0305 0.195
Jordan      0.0496 0.275
Var.sq    146.9029 0.113
Phi2        0.1064 0.031
Sakoda      0.3798 0.031
Tschuprow   0.1631 0.031
Cramer      0.2306 0.031

$Sims
[1] 10000
```

Question 5 – The Freeman-Tukey Statistic **(15 marks)**

When $n_{ij} \sim \text{Poisson}\left(np_{i\bullet}p_{\bullet j}\right)$ the assumption is that

$$E\left(n_{ij}\right) = \text{Var}\left(n_{ij}\right) = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

When this is property not satisfied, variance stabilising strategies can be implemented. One such strategy is to consider the transformation proposed by Freeman & Tukey (1950):

$$\sqrt{n_{ij}} + \sqrt{n_{ij} + 1}$$

who showed that this transformation has a mean of $\sqrt{4np_{i\bullet}p_{\bullet j} + 1}$ and a variance of 1.

i) By using this variance stabilisation for large n, prove that Pearson’s chi-squared statistic may be approximated using the Freeman-Tukey statistic. *(10 marks)*

ii) By considering the Freeman-Tukey statistic, derive the asymptotic distribution of $\sqrt{n_{ij}}$.

(5 marks)

Question 6 – Quantile Approximations of the Chi-squared Statistic (15 marks)

When considering an α level of significance in the test of association between two categorical variables with v degrees of freedom, the chi-squared random variable, χ^2_α, has a mean and variance $E(\chi^2_\alpha) = v$ and $Var(\chi^2_\alpha) = 2v$ respectively.

i) By using these results show that, using the central limit theorem and for large n and degrees of freedom,

$$\chi^2_\alpha \approx v + z_\alpha \sqrt{2v}$$

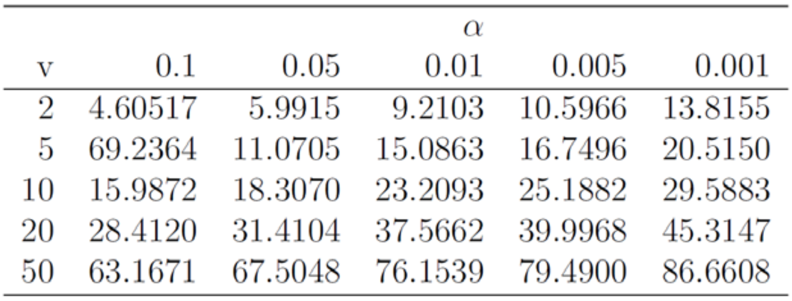
(2 marks)

ii) Wilson and Hilferty (1931) showed that

$$\left(\frac{\chi^2_\alpha}{v}\right)^{1/3} \sim N\left(1 - \frac{2}{9v}, \frac{2}{9v}\right).$$

Use the Central limit theorem to derive an approximation of χ^2_α that may be used as an alternative to $\chi^2_\alpha \approx v + z_\alpha \sqrt{2v}$. *(2 marks)*

iii) The table below provides a summary of the chi-squared statistics, χ^2_α, with a variety of degrees of freedom (v) and α values



			α		
v	0.1	0.05	0.01	0.005	0.001
2	4.60517	5.9915	9.2103	10.5966	13.8155
5	69.2364	11.0705	15.0863	16.7496	20.5150
10	15.9872	18.3070	23.2093	25.1882	29.5883
20	28.4120	31.4104	37.5662	39.9968	45.3147
50	63.1671	67.5048	76.1539	79.4900	86.6608

Using R, construct a table similar to this using the approximation found in part i) and part ii). What can you conclude about the accuracy of these approximations?

(11 marks)