MATH5945: Categorical Data Analysis

Term 3, 2023

Assignment 1

Submission deadline: Thursday 28 September, 4:05pm

Deliverables: 2 files uploaded to Moodle: (1) PDF file of your worked solutions, and (2) SAS file for **ALL** computations. Files names should be *surname_firstname_z123456789_ASS1*.

Assignment length: There is a **5 page limit** and minimum **12pt font size**. Any pages exceeding this limit or submissions with smaller font sizes will not be marked. Handwritten assignments will not be accepted. This does not include a SAS file of your code. Your document should begin with the Plagiarism Statement below (copy-and-paste it).

SAS code: All computations must be performed using SAS. Your SAS code must run **as is** and I should not need to modify your code in any way to make it work. You may create a library to import data, but any other code should only use the WORK library (you may assume data files of the same name are in my WORK library). SAS should be used for computing only and answers given only within SAS code will not be marked.

Penalties: Failure to adhere to instructions will result in a minimum 5% mark reduction.

Name: _____ Student Number: _____

I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Signed: _____ Date: _____

1. In a study by Brown et al (2023) published in the journal *Traffic Injury Prevention*, data were collected on the proper use of a child restraint system in a motor vehicle. One outcome measure for this study was whether there were errors in the installation or use of a child restraint system. Data collection occurred both before and during the COVID-19 pandemic. Study participants were originally randomly selected from a target population; however, this changed to convenience sampling due to COVID-19 restrictions.

   A 2×2 table for number of vehicles with any errors by the sampling method is given below.

	Sampling Method	
Any Error	Random	Convenience
Yes	123	115
No	12	26

   (a) Create a SAS data set for the child restraint system data.

   (b) Is there evidence that reported child restraint errors are associated with sampling method? Be sure to provide the results for the strength of association (and measure of uncertainty), and the results of an appropriate statistical test.

2. The *probability mass function* (pmf) for the non-central hypergeometric distribution is

$$\mathbb{P}(a) = \frac{\binom{n_1}{a}\binom{n_2}{c}\psi^a}{\sum_{x=a_\ell}^{a_u}\binom{n_1}{x}\binom{n_2}{m_1-x}\psi^x}$$

   where $a_\ell = \max\{0, n_1 - m_2\}$, $a_u = \min\{n_1, m_1\}$ and ψ is the odds ratio. This can be considered a likelihood function for ψ.

   (a) Write out the log-likelihood function for ψ.

   (b) Derive the score function for ψ, $U(\psi) = d \log \mathcal{L}(\psi)/d\psi$.

   (c) Write out the observed score function using the AIDS/HIV data from the lecture notes.

   (d) The MLE for ψ is the solution to $U(\hat{\psi}) = 0$ which cannot be solved directly here. To find a solution, follow these steps:

      i. Compute the odds ratio the *usual* way as a cross-product ratio.

      ii. Create a SAS data file for values for ψ on either side of your estimate from part (i) with an appropriate resolution, and compute the score $U(\psi)$ at each of those values given the observed table.

      iii. Use SGPLOT to create a line graph of $U(\psi)$ versus ψ.

      iv. Print out values from your data set where $U(\psi)$ is *closest* to 0. The value for ψ is your estimated MLE of ψ.

(e) Now, using a similar approach to (d) using the AIDS/HIV data, find the values ψ_ℓ and ψ_u that satisfy the equations for a 95% confidence interval

$$\sum_{x=a_\ell}^{a} \mathbb{P}(x; \psi_u) = 0.025$$
$$\sum_{x=a}^{a_u} \mathbb{P}(a; \psi_\ell) = 0.025$$

(f) Use PROC FREQ to check your results for a 95% confidence interval for ψ (**note** that SAS will not compute the MLE $\hat{\psi}$).