TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Mid-Term Assignment 2020-21 STU33009: Statistical Methods for Computer Science

Submitting Your Report

- **Reports must be typed (no handwritten answers please) and submitted on Blackboard.**
- **As a guideline, reports should be about 5 pages in length including all plots (please don't go a lot over this).**
- **You will need to use matlab to calculate values, or alternatively write a short program in python to do this. In either case give the code used as an appendix to the report (it doesn't count towards the page limit), but please keep the code short.**
- **In order to obtain full credit it is essential that you explain/justify how you obtained your results and, where appropriate, that you critically reflect upon them. Simply giving raw numbers as answers will receive few marks as will saying "see code for details" and the like, even if the code contains explanatory comments.**
- **It is mandatory to complete the declaration that the work is entirely your own and you have not collaborated with anyone - the declaration form is available on Blackboard.**

Downloading Data

In this assignment you will analyse the data on shopping behaviour. Start by downloading the following dataset:

- `https://www.scss.tcd.ie/doug.leith/ST3009/midterm2021.php`. Important: You must fetch your own copy of the dataset, do not use the dataset downloaded by someone else. Keep the dataset that you download as I might request it to validate your results.
- The data file consists of rows of data. Each row i corresponds to one supermarket shopping basket and each column j corresponds to one item for sale. The value $Z_{i,j}$ in row i, column j gives how many of the j'th item are in the i'th shopping basket.

Assignment

1. (a) Plot a histogram showing the PMF of the number of items in a basket. Hint: Summing the values in a row gives the number of items in that shopping basket. [5 marks]

(b) Estimate the probability $P(Z_{i,1}=1)$ that the first column in the dataset takes value 1 i.e. that a shopping basket contains an item 1. Briefly explain/discuss your calculation. Hint: Observe that the first column in the dataset only takes values 0 or 1 and recall that for an indicator RV X we have $Prob(X=1)=E[X]$. [5 marks]

(c) Derive a confidence interval for your estimate $P(Z_{i,1}=1)$ using the CLT and Chebyshev Inequality. Explain/discuss your calculation. [5 marks]

(d) Suppose we require to estimate the value of $P(Z_{i,1}=1)$ to an accuracy of ±1% with 95% confidence. How many shopping baskets would we need to collect data from? [5 marks]

2. Your task is to explore whether the presence of item 1 in a shopping basket can be predicted from the presence of other items in the basket. We start with whether item 2 in the basket is predictive of item 1 being in the basket. Since the first column in the dataset only takes values 0 or 1, its conditional expectation $E[Z_{i,1}|Z_{i,2}=z]=P(Z_{i,1}=1|Z_{i,2}=z)$. The sample mean of $Z_{i,1}$ conditioned on $Z_{i,2}=z$ is $\frac{1}{N}\sum_{j\in\{i:Z_{i,2}=z\}} Z_{j,1}$, where $\{i: Z_{i,2}=z\}$ is the set of baskets for which $Z_{i,2}$ equals z i.e. the sum is taken over the baskets with second column equal to z, and $N=|\{i: Z_{i,2}=z\}|$ is the size of this set. This sample mean concentrates on $E[Z_{i,1}|Z_{i,2}=z]$ as the number of shopping baskets observed grows.

(a) Calculate the sample mean of $Z_{i,1}$ conditioned on the second column $Z_{i,2}=z$ for $z=0,1,\ldots$ being each of the different values that the second column takes. Report the values in a table. Briefly explain/discuss your calculation. [5 marks]

(b) Derive confidence intervals for your estimate $E[Z_{i,1}|Z_{i,2}=z]$ using the CLT and Chebyshev Inequality. Explain your working and extend your table from (a) to include these intervals. [5 marks]

(c) Using the matlab errorbar() function, or python equivalent, plot your estimates of $E[Z_{i,1}|Z_{i,2}=z]$ vs z together with their confidence intervals i.e. a plot with z on the x-axis and the estimate of $E[Z_{i,1}|Z_{i,2}=z]$ on the y-axis, together with error bars indicating the confidence interval around this estimate. Discuss. [5 marks]

(d) Compare your estimate of $E[Z_{i,1}|Z_{i,2}=z]$ with your estimate of $E(Z_{i,1})$ from part 1(b)-(c), bearing in mind their confidence intervals. Critically discuss whether the presence of item 2 in the basket is predictive of item 1 being in the basket. [5 marks]

3. (a) Repeat your analysis in 2(d) but now using only the first 100 rows from the dataset (its enough to plot the data, no need to include a table of values). What is the impact on the confidence intervals of using less data, and why? How does that impact what conclusions you can draw from the data? [5 marks]

(b) Now repeat 2(d) but for $E[Z_{i,1}|Z_{i,3}=z]$ i.e. conditioned on the third column $Z_{i,3}=z$. Compare and contrast the behaviour with that observed when conditioning on the second column, again bearing in mind the confidence intervals. [5 marks]