Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

DECLARATION

I understand that this is an individual assessment and that collaboration is not permitted. I have not received any assistance with my work for this assessment. Where I have used the published work of others, I have indicated this with appropriate citation.

I have not and will not share any part of my work on this assessment, directly or indirectly, with any other student.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcdie.libguides.com/plagiarism/ready-steady-write."

I understand that by returning this declaration with my work, I am agreeing with the above statement. Yes

Name: Maria Alonso Garcia

Date: 31 March 2021

TRINITY COLLEGE DUBLIN. School of Computer Science and Statistics
Mid-Term Assignment 2020-21 STU33009: Statistical Methods for Computer Science

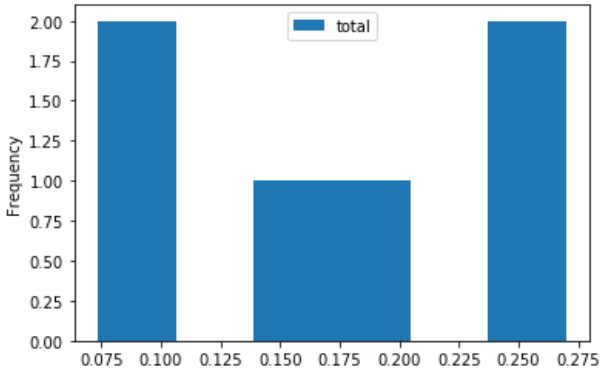
Question 1 (a)

First step. Adding the values of the three columns j for each row i to obtain the items in the shopping basket.

Second step. Create a frequency table: aggregate how many shopping baskets had 1, 2, 3, 4, 5 or 6 items. 3 375 | 4 335 | 2 243 | 5 225 | 6 108| 1 102

Third step. Divide each value by 1388, the sum of all values.
3 0.270173 | 4 0.241354 | 2 0.175072 | 5 0.162104 | 6 0.077810 | 1 0.073487

Fourth step. Plot a histogram.



Question 1 (b)

Calculate the Expected Value.
Step 1. Sum the amount of 0s and 1s there are in the first column / 1388 (total number of rows) 1 0.501441 0 0.498559

Step 2.
Expected Value $E[X] = \sum x_i p(x_i)$
(0.501441 * 1) + (0.498559 * 0) = 0.501441

This value shows it is random that there is an item in the first column or not. There is equal probability 0.501, 0.499 for a 1 or a 0.

Question 1 (c)

Step 1. Calculate the standard deviation. = sqrt(((1-expected_value)^2)/total1)
Total 1 is the total number of 1s there are in column 1 = 696

Mean = expected value = 0.501441
Standard deviation = 0.018897830125432724

Step 2. CLT confidence intervals, lower and upper.
CLT lower = mean - ((2*stDev) / sqrt(total1)) = 0.500
CLT upper= mean + ((2*stDev) / sqrt(total1)) = 0.503

Step 3. Chebyshev confidence intervals, lower and upper.
CHEB lower = mean - (stDev / sqrt(0.05*total1)) = 0.498
CHEB upper = mean + (stDev / sqrt(0.05*total1)) = 0.505

Question 1 (d)

To find the shopping baskets needed to collect data from to estimate the value of P(Zi,1 = 1) to an accuracy of ±1% with 95% confidence, I will use the formula: σ^2 = variance / N
2 σ represents 95% confidence.
I need to rearrange the formula to get N, the number of baskets needed.
N = variance / ((0.01)2 / 2)
N = 10005
I will need 10005 shopping baskets to estimate the value with a confidence of 95%.

Question 2 (a)

Sample mean = $\bar{x}$ = (Σ xi) / n
xi = sample items
n = number of items

Step 1. Calculate how many 0,1,2,3 there are in the second column. This will be n, number of items.
0 330
1 365
2 343
3 350

Step 2. Calculate the sum of sample items. When column 2=0, how many 1s are there in column 1? When column 2=1, how many 1s are there in column 1? … for all z values {0,1,2,3}

2.1 reduced my dataframe to only include the rows in which column 1= 1
2.2 reduced dataframes for when column 2 = z for each z value

Col 1	Col 2	frequency (Col 1 && Col 2)	frequency (Col 2) from Step 1	calculation	Sample mean
1	0	330	330	330/330	1.0
1	1	279	365	279/365	0.764
1	2	87	343	87/343	0.254
1	3	0	350	0/350	0

The calculation for the sample mean is the frequency in which
(column 1=1 && column2=z) / total number of occurrences of z in column 2.

Through this table I can observe that the sample means follow a downward trend with equally sized steps. 1.0 → 0.75 → 0.25 → 0

(1,0) always occurs. Which means that if there is no item in column 2, there will always be an item in column 1.
(1,3) never occurs. Which means that if there are three items in column 2, there will never be an item in column 1.

Question 2 (b)

I used the 95% intervals for CLT and Chebyshev.
To find these I had to calculate the standard deviation for each value of Z. For the mean I used the sample means calculated in part 2(a).

Then I applied the formula for each method, CLT and Chebyshev .

Step 1. Calculate the standard deviation for each value of **Z**
Numerator = sumof(x_i - mean)^2 =
col2zero['sd'] = (col2zero[0] - sample_mean_**z**)**2
Standard deviation =
sd_**z** = np.sqrt((col2zero['sd'].sum()) / total**Z**incol2)

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Step 2. CLT confidence intervals, lower and upper.
CLT lower = mean - ((2*stDev) / sqrt(total**Z**incol2))
CLT upper= mean + ((2*stDev) / sqrt(total**Z**incol2))

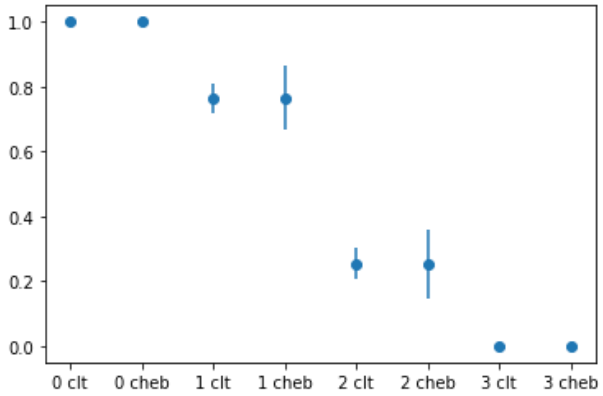
Step 3. Chebyshev confidence intervals, lower and upper.
CHEB lower = mean - (stDev / sqrt(0.05*totalZincol2))
CHEB upper = mean + (stDev / sqrt(0.05*totalZincol2))

Col 1	Col 2	frequency (Col 1 && Col 2)	frequency (Col 2) from Step 1	calculation	Sample mean	CLT lower	CLT upper	Cheb lower	Cheb upper
1	0	330	330	330/330	1.0	1.0	1.0	1.0	1.0
1	1	279	365	279/365	0.764	0.720	0.809	0.665	0.864
1	2	87	343	87/343	0.254	0.207	0.301	0.149	0.359
1	3	0	350	0/350	0	0.0	0.0	0.0	0.0

From this table I can observe that CLT upper and lower bounds are closer to the sample mean than the Chebyshev bounds, in every instance of Z.
This shows that CLT is more accurate than Chebyshev.
This is consistent with what we can expect because CLT is always more precise. CLT Confidence intervals offer a full distribution of X but an approximation of values when N is finite.

On the other hand, Chebyshev will work for any number of N. Therefore it will be more loose and longer difference between the upper and lower bounds.
Chebyshev is an actual bound while CLT is an approximation.

Question 2 (c)



There are two main observations from this error lines plot:
(a) The error lines for Chebyshev are in both cases longer than for CLT. As explained earlier, this is consistent with what I expected because CLT is more precise.

(b) I observe a downward trend from 0 to 3, with an equal separation between the points. This means we could draw a line of best fit and interpolate for values between 0 and 3. For example, if 1.5 items of column 2 were bought, we can estimate that 0.625 of item 1 would be bought. Of course, these are statistically significant but in reality, 1.5 or 0.625 of a product can't be bought. (can't break a packet or food into pieces).

This gradient of -1 gives us information about shopping trends. It proves that buying item in column 1 is dependent on buying item on column 2.

- If 0 items of column 2 are bought, consumers will always buy 1 item of column 1.

For example: assuming consumers always buy butter. If no Kerrygold butter is bought (column 2 product), the consumer will always buy Tesco butter (column 1 product).

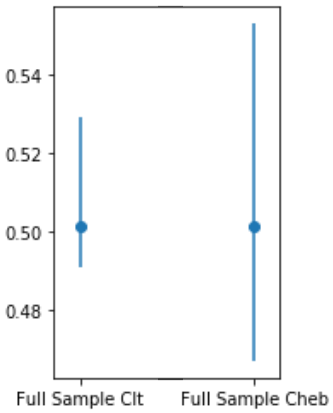
- If 3 items of column 2 are bought, consumers will never buy 1 item of column 1.

For example: If 3 plant-based milk cartons are bought, the consumer will never buy 1 item of milk-based milk.

Question 2 (d)

The presence of item 2 in the basket is predictive of item 1 being in the basket. I come to this conclusion because:

Only looking at the values of item 1, there is a random probability of buying or not buying the item. As seen in question 1, the expected value was 0.501. These error bars represent the confidence intervals.

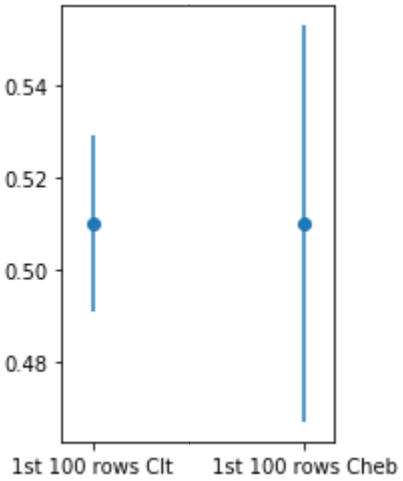


But when looking at the values and plots of Question 2 b-c, we observe that buying or not buying item 1 is dependent on how many of item 2 are bought. A more thorough explanation was given in part c of this question.
Item 2 is predictive of item 1 because if there are 0 items of 2, there will always be 1 item of 1. And if there are three items of 1, there will never be one item of 1.

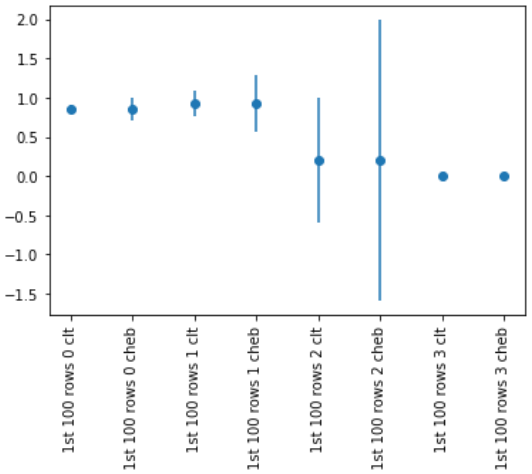
Question 3 (a)

Question 1 b c with 1st 100 rows of data

	Full Data $E(Z_{i,1})$	Only first 100 rows$E(Z_{i,1})$
E[X]	0.501	0.51
StDev	0.019	0.069
cltLower	0.500	0.491
cltUpper	0.502	0.529
chebLower	0.498	0.467
chebUpper	0.505	0.553



Question 2 b c with 1st 100 rows of data --->



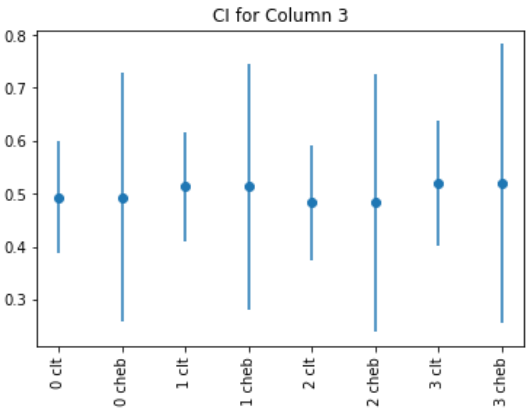
When using less data, the confidence intervals are larger. This occurs because when we have less sample data, we are less sure of the degree of accuracy of the presence of the items.

The larger confidence intervals affects the conclusions we can draw from the data:
With less data, it is less accurate to say if item 2 is predictive of the presence of item 1.

For example, the downward trend from 0 to 3 is now not linear.

For example, when looking at Chebyshev CI for 2 items in column 2, we see that the average is ~0.5 and CI has ranges 2.0 to -1.5. There can be probability 2 all the way to probability -1.5 that if there are two items of column 2, there will be one item of column 1. This is much less accurate than when we had 1388 points of data.

Question 3 (b)



When changing my script to column 3 instead of column 2, I observe many differences in the error bar plot.

The Sample mean for all of them is ~0.5. This means that they are all random, and therefore there is an equal likelihood of happening. ie, if there are two items of column 3, there is an equal chance that there will be one item of column 1 than there will be no item of column 1.

As the sample mean for all of them is ~0.5, there is no upward or downward trend. They are scattered with equal random probability. Therefore the amount of items 3 is not predictive at all of the presence of item 1.

The confidence intervals are wider for both CLT and Chebyshev, for all values of Z. This occurs because the data is random.