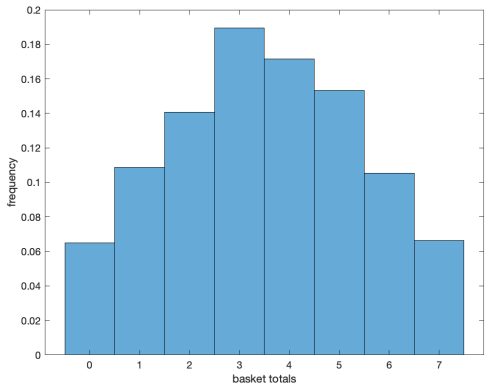
Mid-Term Assignment - Statistical Methods for Computer Science - 18325745 - Ethan Monkhouse
Total Report Pages: **5 pages** | Total Appendix Pages: **5 Pages** | Total Report Word Count: **2677**

Question 1

Part (a)
The PMF of the number of items in a basket is calculated by listing each individual basket total contained in the dataset, and calculating the estimated probability of the event of each total occurring. This is calculated using the Matlab code: sum(dataset,2). We can then use this data to plot the histogram shown below using the code: histogram(a_rowSum,2,'BinMethod','auto','Normalization', 'pdf'). The x-axis lists the various basket total integers, ranging from 1-7. The y-axis lists the estimated probability of the various events occurring. See plot below:



Part (b)
To estimate the probability that the first column in the dataset takes the value 1, we must firstly calculate the number of occurrences of the event where the first column takes the value 1. This is done using a snippet of code which iterates through each row of the dataset and increments a counter if the first column of the row is equal to 1. Code: sum(dataset(:,1)==1). The output value is 732. Once we have this value, we must calculate the probability of this event occurring. To do this we divide the number of occurrences of the event calculated previously, by the population size, which in this case is 1435. This produces an answer of: 0.510105.

P(Zi,1 = 1) = 0.510105

Part (c)
For both the CLT and Chebyshev, the population mean and variance of P(Zi,1 = 1) is required. In terms of calculating the mean of the population we use the following formula: $Population\ mean\ =\ \Sigma\, x/N$, where X is equal to the event where *Zi,1 = 1*. The estimated population mean when P(Zi,1 = 1) is **0.510105.** We then look at calculating the population variance. The variance is calculated using the formula: Population variance = $\Sigma\,(Xi\ -\ \mu)\,/\,N$, we simplify this in Matlab by calculating the variance of the event *Zi,1 = 1,* and then dividing it by the population to calculate the population variance. The estimated population variance when P(Zi,1 = 1) is **0.000174**. We also need the standard deviation for calculating the Chebyshev confidence interval, which is the square root of the population variance. The estimated population standard deviation when P(Zi,1 = 1) is **0.013201.** Now we must calculate the 95% confidence interval using the CLT. We use the following formula: $P(-2\sigma \le X - \mu \le 2\sigma) \approx 0.95,\ when\ X \sim N(\mu, \sigma 2)$ This can be simplified to: $\mu - 2\sigma \le X \le \mu + 2\sigma$. We use the following code to calculate the answer to this formula using the value previously calculated above:

The CLT 95% Confidence Interval is **[0.483703, 0.536507]**.

Now, to calculate a confidence interval using Chebyshev, we use the following formula: $\mu\ -\ \sigma\ /\ \sqrt{0.05N} \le X \le \mu\ +\ \sigma\ /\ \sqrt{0.05N}$. However, when we analyse $\sigma\ /\ \sqrt{0.05N}$, we can note that we have calculated the population variance. Therefore, we adjust this part of the equation to $\sigma\ /\ \sqrt{0.05}$, where σ is the population variance.

Chebyshev 95% Confidence Interval is **[0.451068, 0.569141]**.

When we analyse these results, we notice that the bounds of Chebyshev's confidence interval is larger than that of the CLT confidence interval. This is expected due to the nature of Chebyshev's confidence interval, however compared to the CLT, it

is more accurate in this case as it provides an actual bound. CLT only provides an approximation when N is finite, and in this case it is.

Part (d)
From our previous calculations, we know the population variance when P(Zi,1 = 1) is 0.000174**.** However, as we are looking for the sample required to give the below confidence, we use the sample variance which is calculated using Matlab. **The output was: 0.2501.** Since we are looking to estimate the value of P(Zi,1 = 1) with an accuracy of +- 1% with 95%, we use the CLT and Chebyshev, and the variance.

CLT:

$0.01 = 2\sqrt{\frac{\sigma^2}{N}}$, with 95% confidence +- 1% $\quad (0.01)/2 = \sqrt{\frac{(0.2501)}{N}}$

$0.005 = \sqrt{\frac{(0.2501)}{N}} \quad 0.005^{\ 2} = \frac{(0.2501)}{N} \quad N = \frac{(0.2501)}{0.000025}$

N = 10,004

Chebyshev:

$P(|X - \mu| \geq 0.01) \leq 0.05$ *....from the lecture material*

$P(|X - \mu| \geq 0.01) \leq \frac{\sigma^2}{0.01^{\ 2}N}$ *....from the lecture material*

$\frac{\sigma^2}{0.01^{\ 2}N} < 0.05 = \frac{\sigma^2}{0.01^{\ 2}(0.05)} < N$ *....we can deduce this from the previous equations*

$\frac{\sigma^2}{0.01^{\ 2}N} < 0.05 = \frac{0.2501}{0.01^{\ 2}(0.05)} < N$*….we now substitute in the values to the equation*

N = 50,020

$N \geq 10,004$ **- CLT Method**
$N \geq 50,020$ **- Chebyshev Method**

From the above, based on our dataset and values calculated in previous questions, we would need to collect data from **10,004 or more** baskets to ensure an accuracy of +- 1%, with 95% confidence with the CLT method, and **50,020 or more** baskets to ensure an accuracy of +- 1%, with 95% confidence with the Chebyshev method.

Question 2

Part (a)
Before we begin the calculation of the questions asked, we must first import the dataset into Matlab. This calculation is to find the sample mean, which can also be considered the expected value of the event that column 2 = z, given column 1 = 1, which can be shown as E[Zi,1|Zi,2 = z]. We must first calculate the range of z values to be used in the analysis. The output is an array where **z_values = (0;1;2;3).** Now, we need to find the probability of the event that column2=z and column1=1, as well as the probability of the event that column2=z. This is done by iterating through each row, checking if the events are true, and calculating a total sum of the true events. We then divide by the population, which in this case is the number of rows in the dataset. According to Bayes Theorem, we know that:$P(A|B) = P(A \cap B)/P(B)$. We have calculated the right-hand side of the equation.

We receive the following output as a result:

	z	E[Zi,1 ∩ Zi,2 = z]	E[Zi,2 = z]	E[Zi,1|Zi,2 = z]
1	0	0	0.2578	0
2	1	0.0620	0.2411	0.2572
3	2	0.1993	0.2523	0.7901
4	3	0.2488	0.2488	1.0000

For the purpose of this analysis, we have included additional columns relating to the calculation of the sample mean. As you can see, the first column denotes the respective *z value* that is being used. Looking at when z=0, we can deduce that there is no basket where column1 and column2 are both 0, however, there is a 0.2578 probability that column2 is 0, irrespective of column1. As such because of their being no basket where column1 and column2 are 0, the expected value of [Zi,1|Zi,2 = 0]

is going to be 0. Looking at the occurrence of when z = 1 and z = 2, the results are quite similar for both. There are varied expected values for both events, and as such we calculate an expected value between the range of 0 and 1 for both. Lastly, for z = 3, we notice that the expected value of [Zi,1|Zi,2 = 3] is 1, meaning that this event is true every time in the dataset. This is confirmed by the columns 2 and 3 in the table. Both events have a 0.2488 expected value - meaning they are equally likely to occur. If we simplify this in terms of baskets, for every basket, if item 2 has a quantity of 3, it is certain that the same basket contains item 1 with a quantity of 1.

Part (b)

Before we calculate the confidence interval, we must calculate the variance for each sample mean with each z value. To do this we iterate through each z value with a for-loop. We then have a nested for-loop which iterates through each basket in the dataset. If column2 = z and column1 = 1 in the specific basket is true, we know the value is 1. So according to formula for variance = (sumOf(Sample value - sample mean)^2)/N, we calculate this using MatLab. We now have a variance array, which is added to our previous table using the last two lines of code above. We now much calculate the standard deviation, and thus the standard error, as well as the margin of error for use with CLT and Chebyshev. We first define all the necessary arrays to improve the speed of the script. We iterate through each z-value. For each value we calculate the standard error and margin of error using the following formulas:

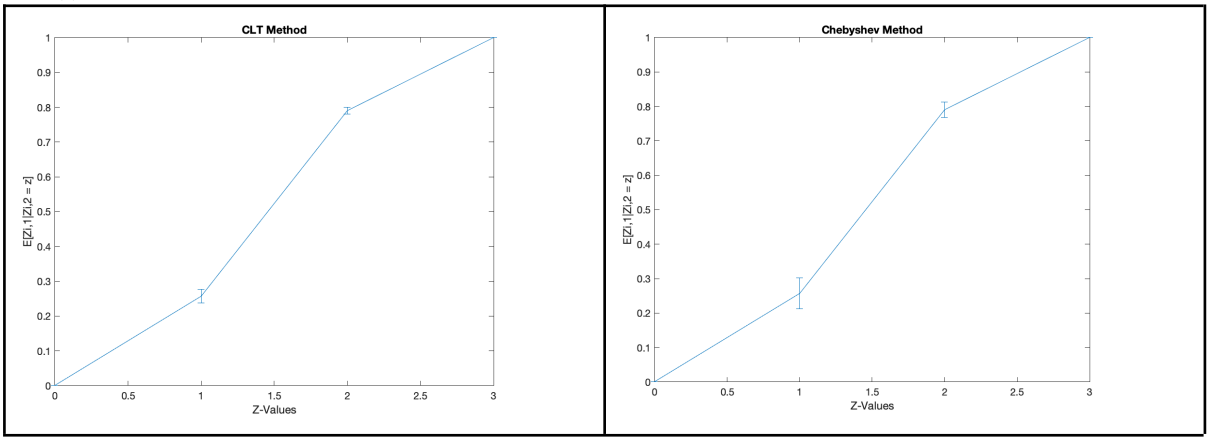
$$\sigma = \sqrt{variance}$$
$$Standard\ Error = \sigma/\sqrt{N}$$
$$Margin\ of\ Error\ (95\%\ Confidence) = \sigma/\sqrt{0.05N}$$

These are calculated using the first two lines of code within the for-loop. Next we calculate the bounds for both CLT and Chebyshev using the following formulas: **CLT:** $P(-2\sigma \leq X - \mu \leq 2\sigma) \approx 0.95$, **Chebyshev:** $\mu - \frac{\sigma}{\sqrt{0.05N}} \leq X \leq \mu + \frac{\sigma}{\sqrt{0.05N}}$. The bounds are calculated in MatLab in the next 4 lines of the for-loop. We add these new arrays to the table. We receive the following table output (please note for the purpose of formatting, the table only shows the extended rows):

CLT 95% Lower Bound	CLT 95% Upper Bound	Chebyshev 95% Lower Bound	Chebyshev 95% Upper Bound
0	0	0	0
0.2373	0.2771	0.2128	0.3017
0.7802	0.7999	0.7680	0.8121
1.0000	1.0000	1.0000	1.0000

Firstly, let us discuss the confidence intervals where z=0. We can observe that the bounds are all zero. This is due to the fact that the expected value is zero for [Zi,1|Zi,2 = 0], thus we can be mathematically 95% confident that 0 will be the expected value of this event. Now looking at the confidence intervals where z=1 and z=2. Similar to the previous question, these values are as expected, and in between the range of 0 and 1. To build on this, we observe that Chebyshev has larger bounds, which is expected as compared to CLT, Chebyshev uses actual bounds - whilst CLT is an estimate. Finally, looking at the confidence intervals where z=3, we can see that the bounds are all 1. Similar to where z=0, were the expected value for [Zi,1|Zi,2 = 3] is 1. Thus we can be mathematically 95% confident that 1 will be the expected value of this event.

Part (c)



Error bars are useful for assessing the certainty of the expected value for a certain z-value. We have generated both CLT and Chebyshev error bars. Looking at the CLT first, we can see that the error bar range is smaller than that of the Chebyshev error bars. This indicates that there is more certainty surrounding the expected value. However, we must remember that the CLT is an estimate, so the bounds are smaller. Looking at the Chebyshev error bar, we can see that compared to the CLT, the range of the error bars is larger, suggesting that there is more uncertainty surrounding the expected value with a corresponding z-value. This is expected, as we know that the bounds of the Chebyshev method is larger than the CLT, as Chebyshev produces actual bounds. Analysing the commonalities between both charts, we can see that the error bar range is larger for when z=1, compared to when z=2. This shows that there is more uncertainty in determining the expected value for when z=1. We should also note that there is no error bar for z=0, and z=3. This is because z=0 has an expected value of 0, and z=3 has an expected value of 1. Thus the bounds for both CLT and Chebyshev are 0 respectively, and do not affect the expected value range with 95% confidence.

Part (d)

We reference the following results from Question 1, Part (b) & (c):
The CLT 95% Confidence Interval is ***[0.483703, 0.536507].***
Chebyshev 95% Confidence Interval is ***[0.451068, 0.569141].***

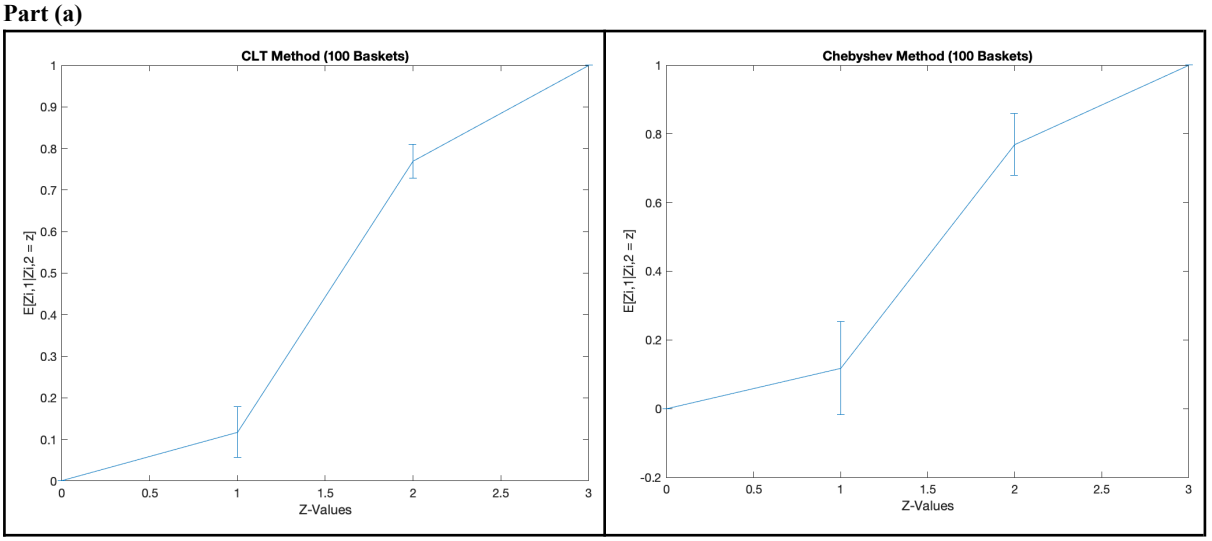
We also reference the following table from the previous part of the question:

CLT 95% Lower Bound	CLT 95% Upper Bound	Chebyshev 95% Lower Bound	Chebyshev 95% Upper Bound
0	0	0	0
0.2373	0.2771	0.2128	0.3017
0.7802	0.7999	0.7680	0.8121
1.0000	1.0000	1.0000	1.0000

We should note that the confidence interval for the event occurring when item 1 is present is *CLT 95% Confidence Interval:* ***[0.483703, 0.536507],*** *and Chebyshev 95% Confidence Interval:* ***[0.451068, 0.569141]***. In comparison to the event E[Zi,1|Zi,2 = z], the expected value is far higher. We can also note that as the quantity of item 2 increases, the expected value E[Zi,1|Zi,2 = z] increases to the bound of 1. In short, when there are no item 2's in the basket, there are no item 1's. When there are item 2's with a quantity of 3, there are always item 1's in the basket. From this we can conclude that the presence of item 2 in the basket is predictive of item 1 being in the basket.

<u>Question 3</u>

Part (a)

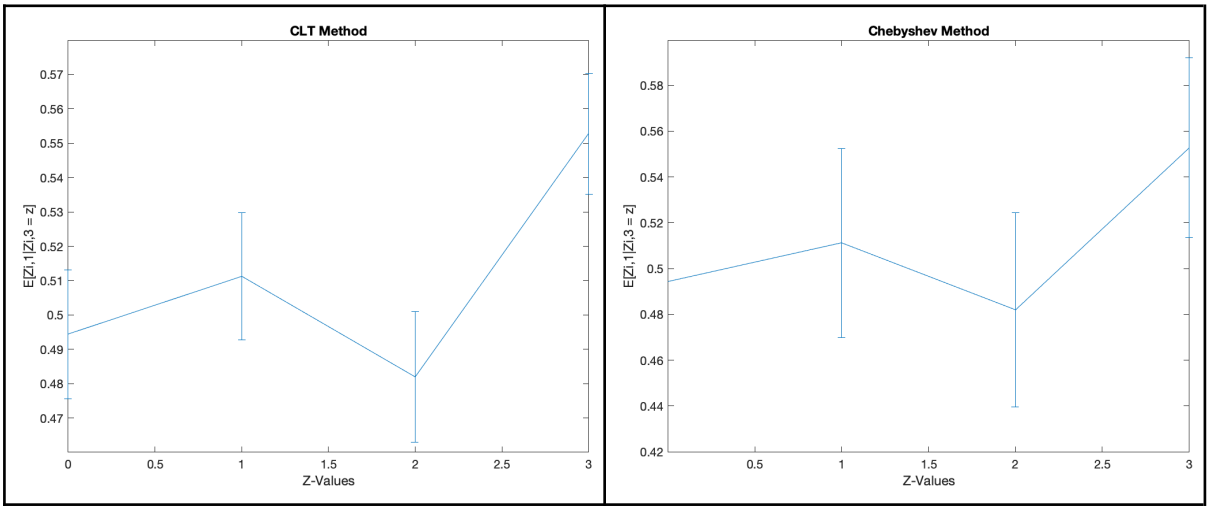


Immediately upon analysis we can deduce that the bounds for the confidence intervals for when z=1 and z=2 have become significantly larger compared to when the entire dataset was used. The reasoning behind this is because the range of the error bars on both charts have increased. Due to this increase, there is increased uncertainty in the expected value E[Zi,1|Zi,2 = z] when z=1 and z=2. As such, we can conclude that by using a smaller dataset, the results are not as accurate. More specifically, the upper and lower bounds for the confidence intervals are required to be larger than before to ensure that they

maintain 95% confidence, as there is a reduction in the data available. As expected, this affects our certainty that the presence of item 2 in the basket is predictive of item 1 being in the basket, however, due to there being no change to the expected value E[Zi,1|Zi,2 = z] when z=0, and z=3, there is still enough data to suggest that the presence of item 2 in the basket is predictive of item 1 being in the basket.

Part(b)

For this question, we remove the constraint on the 100 rows of the dataset, and set it to the entire population. We then replace Zi,2 with Zi,3 throughout the entire script. The following charts are generated:



The first thing to note is that Chebyshev provides larger error bars, which is as expected due to the nature of Chebyshev and the CLT. With Chebyshev, the confidence intervals are larger. We must now consider the overlap in error bars between the error bars where z=0,1,2 for the CLT chart, and where z=0,1,2,3 for the Chebyshev chart. From this, we can conclude that for each z-value in the Chebyshev chart their sample size are all nearly equal. With the CLT chart, the sample size for z=0,1,2 are all nearly equal.

We also notice compared to the second column, the confidence intervals are much larger - in turn resulting in a large range error bar. This indicates that compared to the second column, it would require far more data to give a similar confidence interval for the third column. Additionally, we notice that in the second column z=0, and z=3 were certain with the dataset provided, with the third column z=0, and z=3 are not certain, and have their own individual confidence intervals and error bar.

When it comes to assessing the presence of item 3 in the basket is predictive of item 1 being in the basket, it is a lot more difficult and uncertain. Due to the significant overlap in the error bars, and there is no immediate trend observed, it is not possible to confirm that the presence of item 3 in the basket is predictive of item 1 being in the basket. However, the expected value of E[Zi,1|Zi,3 = z] is significantly higher than what was observed in the analysis of column 2. We can observe that the expected value range is from roughly 0.44 to 0.56. Due to this having a majority above the 0.5 probability mark, we can conclude that the presence of item 3 in the basket is predictive of item 1 being in the basket.

Appendix

```
%question_1

%import data
dataset = importdata('/Users/ethanvictormonkhouse/midterm2021.php')
[numberOfRows,numberOfColumns]=size(dataset);
fprintf("The dataset contains %d rows, and %d columns. \n", numberOfRows, numberOfColumns);

%part_a
a_rowSum = sum(dataset,2); %calculate sum of each basket
histogram(a_rowSum,2,'BinMethod','auto','Normalization', 'pdf'); %generate pmf

xlabel('basket totals')
ylabel('frequency')

%part_b
b_probability=sum(dataset(:,1)==1)/numberOfRows; %sum of observations/population
fprintf("P(Zi,1 = 1) is equal to %f \n", b_probability);

%part_c
c_mean=(sum(dataset(:,1)==1)/numberOfRows); %population mean = sum/population
fprintf("The estimated population mean when P(Zi,1 = 1) is %f \n", c_mean); %print
c_var = var(dataset(:,1)==1)/numberOfRows; %population variance = var(x)*(1/population)
fprintf("The estimated population variance when P(Zi,1 = 1) is %f \n", c_var); %print
c_std = sqrt(c_var); %standard deviation = square root of the variance
fprintf("The estimated population standard deviation when P(Zi,1 = 1) is %f \n", c_std); %print

c_clt_lowerBound = c_mean - 2*(c_std); %CLT Lower Bound
c_clt_upperBound = c_mean + 2*(c_std); %CLT Upper Bound
fprintf("CLT 95-Percent Confidence Interval is [%f, %f]. \n", c_clt_lowerBound,c_clt_upperBound); %print

c_chebyshev_lowerBound = c_mean - c_std/sqrt(0.05); %Chebyshev Lower Bound
c_chebyshev_upperBound = c_mean + c_std/sqrt(0.05); %Chebyshev Upper Bound
fprintf("Chebyshev 95-Percent Confidence Interval is [%f, %f]. \n", c_chebyshev_lowerBound,c_chebyshev_upperBound);
%print

%part_d
d_var = var(dataset(:,1)==1) %variance

%question_2

%import data
dataset = importdata('/Users/ethanvictormonkhouse/midterm2021.php')
[numberOfRows,numberOfColumns]=size(dataset);
fprintf("The dataset contains %d rows, and %d columns. \n", numberOfRows, numberOfColumns);

%part_a
z_values = (min(dataset(:,2)):max(dataset(:,2)));
probability_2_true = zeros(max(dataset(:,2))+1,1); %define array
probability_2_1_true = zeros(max(dataset(:,2))+1,1); %define array
for i = z_values %calculate quantities for z
    probability_2_1_true(i+1) = sum(dataset(:,2) == i & dataset(:,1) == 1)/numberOfRows; %probability of column1=1 and
column2=z
   probability_2_true(i+1) = sum(dataset(:,2) == i )/numberOfRows; %probability of column2=z
end
```

```
probability_zi2_given_zi1 = probability_2_1_true./probability_2_true; %probability of column1=1 and column2=z occurring given column1=1

T = table(z_values.', probability_2_1_true, probability_2_true, probability_zi2_given_zi1); %create table
T.Properties.VariableNames = {'z', 'E[Zi,1 ∩ Zi,2 = z]', 'E[Zi,2 = z]', 'E[Zi,1|Zi,2 = z]'} %rename variable columns

%part_b

%calculate variance for each z value
variance = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %iterate through all quantites
   totalSum = 0; %reset totalSum for next quantity
   for j = 1:numberOfRows %iterate through every basket
      if (dataset(j,2) == i) && (dataset(j,1) == 1) %if column2 = z and column1 = 1
          totalSum = totalSum + (1 - probability_zi2_given_zi1(i+1,1))^2; %calculate total sum of expected values minus the expected sample mean, squared
      end
   end
   variance(i+1) = totalSum/numberOfRows; %calculate variance, which is the totalSum divided by the population
end

T = [T table(variance)]; %update table with variance column
T.Properties.VariableNames{'variance'} = 'Variance'; %rename variance column

%calculate variance and confidence intervals
standardError = zeros(max(dataset(:,2))+1,1); %define array
marginError = zeros(max(dataset(:,2))+1,1); %define array
CLTLowerBound = zeros(max(dataset(:,2))+1,1); %define array
CLTUpperBound = zeros(max(dataset(:,2))+1,1); %define array
ChebyshevLowerBound = zeros(max(dataset(:,2))+1,1); %define array
ChebyshevUpperBound = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %iterate through all quantities
   standardError(i+1) = sqrt(variance(i+1))/sqrt(sum(dataset(:,2) == i)); %SE=standard deviation/sqrt(n)
   marginError(i+1) = sqrt(variance(i+1))/(sqrt(sum(dataset(:,2) == i)*0.05)); %ME.95=standard deviation/sqrt(0.05*n)

   CLTLowerBound(i+1) = probability_zi2_given_zi1(i+1) - 2*(standardError(i+1)); %CLT mean - 2(SE)
   CLTUpperBound(i+1) = probability_zi2_given_zi1(i+1) + 2*(standardError(i+1)); %CLT mean + 2(SE)

   ChebyshevLowerBound(i+1) = probability_zi2_given_zi1(i+1) - marginError(i+1); %Chebyshev mean - marginOfError
   ChebyshevUpperBound(i+1) = probability_zi2_given_zi1(i+1) + marginError(i+1); %Chebyshev mean + marginOfError
end

T = [T table(CLTLowerBound, CLTUpperBound, ChebyshevLowerBound, ChebyshevUpperBound)]; %update table with variance column
T.Properties.VariableNames({'CLTLowerBound' 'CLTUpperBound' 'ChebyshevLowerBound' 'ChebyshevUpperBound'}) = {'CLT 95% Lower Bound' 'CLT 95% Upper Bound' 'Chebyshev 95% Lower Bound' 'Chebyshev 95% Upper Bound'}

%part_c


errorbar(z_values, probability_zi2_given_zi1, probability_zi2_given_zi1-CLTLowerBound, CLTUpperBound-probability_zi2_given_zi1); %using CLT method
errorbar(z_values, probability_zi2_given_zi1, probability_zi2_given_zi1-ChebyshevLowerBound, ChebyshevUpperBound-probability_zi2_given_zi1); %using Chebyshev method

%part_d
```

```
p1 = zeros(max(dataset(:,2))+1,1); %define array
p2 = zeros(max(dataset(:,2))+1,1); %define array
p12 = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %calculate quantities for z
   p2(i+1) = sum(dataset(:,2) == i)/numberOfRows; %E[X]
   p1(i+1) = sum(dataset(:,1) == 1)/numberOfRows; %E[Y]
   p12(i+1) = p2(i+1)*p1(i+1); %E[X]E[Y]
end


%question_3

%import data
dataset = importdata('/Users/ethanvictormonkhouse/midterm2021.php');
dataset = dataset(1:100,:)
[numberOfRows,numberOfColumns]=size(dataset);
fprintf("The dataset contains %d rows, and %d columns. \n", numberOfRows, numberOfColumns);

%part_a
z_values = (min(dataset(:,2)):max(dataset(:,2)));
probability_2_true = zeros(max(dataset(:,2))+1,1); %define array
probability_2_1_true = zeros(max(dataset(:,2))+1,1); %define array
for i = z_values %calculate quantities for z
    probability_2_1_true(i+1) = sum(dataset(:,2) == i & dataset(:,1) == 1)/numberOfRows; %probability of column1=1 and
column2=z
   probability_2_true(i+1) = sum(dataset(:,2) == i )/numberOfRows; %probability of column2=z
end

probability_zi2_given_zi1 = probability_2_1_true./probability_2_true; %probability of column1=1 and column2=z
occurring given column1=1

%calculate variance for each z value
variance = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %iterate through all quantites
   totalSum = 0; %reset totalSum for next quantity
   for j = 1:numberOfRows %iterate through every basket
      if (dataset(j,2) == i) && (dataset(j,1) == 1) %if column2 = z and column1 = 1
           totalSum = totalSum + (1 - probability_zi2_given_zi1(i+1,1))^2; %calculate total sum of expected values minus the
expected sample mean, squared
      end
   end
   variance(i+1) = totalSum/numberOfRows; %calculate variance, which is the totalSum divided by the population
end

%calculate variance and confidence intervals
standardError = zeros(max(dataset(:,2))+1,1); %define array
marginError = zeros(max(dataset(:,2))+1,1); %define array
CLTLowerBound = zeros(max(dataset(:,2))+1,1); %define array
CLTUpperBound = zeros(max(dataset(:,2))+1,1); %define array
ChebyshevLowerBound = zeros(max(dataset(:,2))+1,1); %define array
ChebyshevUpperBound = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %iterate through all quantities
   standardError(i+1) = sqrt(variance(i+1))/sqrt(sum(dataset(:,2) == i)); %SE=standard deviation/sqrt(n)
   marginError(i+1) = sqrt(variance(i+1))/(sqrt(sum(dataset(:,2) == i)*0.05)); %ME.95=standard deviation/sqrt(0.05*n)
```

```
    CLTLowerBound(i+1) = probability_zi2_given_zi1(i+1) - 2*(standardError(i+1)); %CLT mean - 2(SE)
    CLTUpperBound(i+1) = probability_zi2_given_zi1(i+1) + 2*(standardError(i+1)); %CLT mean + 2(SE)

    ChebyshevLowerBound(i+1) = probability_zi2_given_zi1(i+1) - marginError(i+1); %Chebyshev mean - marginOfError
    ChebyshevUpperBound(i+1) = probability_zi2_given_zi1(i+1) + marginError(i+1); %Chebyshev mean + marginOfError
end

errorbar(z_values, probability_zi2_given_zi1, probability_zi2_given_zi1-CLTLowerBound, CLTUpperBound-probability_zi2_given_zi1); %using CLT method
errorbar(z_values, probability_zi2_given_zi1, probability_zi2_given_zi1-ChebyshevLowerBound, ChebyshevUpperBound-probability_zi2_given_zi1); %using Chebyshev method

%import data
dataset = importdata('/Users/ethanvictormonkhouse/midterm2021.php');
[numberOfRows,numberOfColumns]=size(dataset);
fprintf("The dataset contains %d rows, and %d columns. \n", numberOfRows, numberOfColumns);

%part_b
z_values = (min(dataset(:,3)):max(dataset(:,3)));
probability_3_true = zeros(max(dataset(:,3))+1,1); %define array
probability_3_1_true = zeros(max(dataset(:,3))+1,1); %define array
for i = z_values %calculate quantities for z
     probability_3_1_true(i+1) = sum(dataset(:,3) == i & dataset(:,1) == 1)/numberOfRows; %probability of column1=1 and column2=z
    probability_3_true(i+1) = sum(dataset(:,3) == i )/numberOfRows; %probability of column2=z
end

probability_zi3_given_zi1 = probability_3_1_true./probability_3_true; %probability of column1=1 and column2=z occurring given column1=1

%calculate variance for each z value
variance = zeros(max(dataset(:,2))+1,1); %define array

for i = z_values %iterate through all quantites
    totalSum = 0; %reset totalSum for next quantity
    for j = 1:numberOfRows %iterate through every basket
        if (dataset(j,3) == i) && (dataset(j,1) == 1) %if column2 = z and column1 = 1
            totalSum = totalSum + (1 - probability_zi3_given_zi1(i+1,1))^2; %calculate total sum of expected values minus the expected sample mean, squared
        end
    end
    variance(i+1) = totalSum/numberOfRows; %calculate variance, which is the totalSum divided by the population
end

%calculate variance and confidence intervals
standardError = zeros(max(dataset(:,3))+1,1); %define array
marginError = zeros(max(dataset(:,3))+1,1); %define array
CLTLowerBound = zeros(max(dataset(:,3))+1,1); %define array
CLTUpperBound = zeros(max(dataset(:,3))+1,1); %define array
ChebyshevLowerBound = zeros(max(dataset(:,3))+1,1); %define array
ChebyshevUpperBound = zeros(max(dataset(:,3))+1,1); %define array

for i = z_values %iterate through all quantities
    standardError(i+1) = sqrt(variance(i+1))/sqrt(sum(dataset(:,3) == i)); %SE=standard deviation/sqrt(n)
    marginError(i+1) = sqrt(variance(i+1))/(sqrt(sum(dataset(:,3) == i)*0.05)); %ME.95=standard deviation/sqrt(0.05*n)

    CLTLowerBound(i+1) = probability_zi3_given_zi1(i+1) - 2*(standardError(i+1)); %CLT mean - 2(SE)
    CLTUpperBound(i+1) = probability_zi3_given_zi1(i+1) + 2*(standardError(i+1)); %CLT mean + 2(SE)
```

```
    ChebyshevLowerBound(i+1) = probability_zi3_given_zi1(i+1) - marginError(i+1); %Chebyshev mean - marginOfError
    ChebyshevUpperBound(i+1) = probability_zi3_given_zi1(i+1) + marginError(i+1); %Chebyshev mean + marginOfError
end

errorbar(z_values, probability_zi3_given_zi1, probability_zi3_given_zi1-CLTLowerBound, CLTUpperBound-probability_zi3_given_zi1); %using CLT method
errorbar(z_values, probability_zi3_given_zi1, probability_zi3_given_zi1-ChebyshevLowerBound, ChebyshevUpperBound-probability_zi3_given_zi1); %using Chebyshev method

p1 = zeros(max(dataset(:,3))+1,1); %define array
p3 = zeros(max(dataset(:,3))+1,1); %define array
p13 = zeros(max(dataset(:,3))+1,1); %define array

for i = z_values %calculate quantities for z
    p3(i+1) = sum(dataset(:,3) == i)/numberOfRows; %E[X]
    p1(i+1) = sum(dataset(:,1) == 1)/numberOfRows; %E[Y]
    p13(i+1) = p2(i+1)*p1(i+1); %E[X]E[Y]
end
```