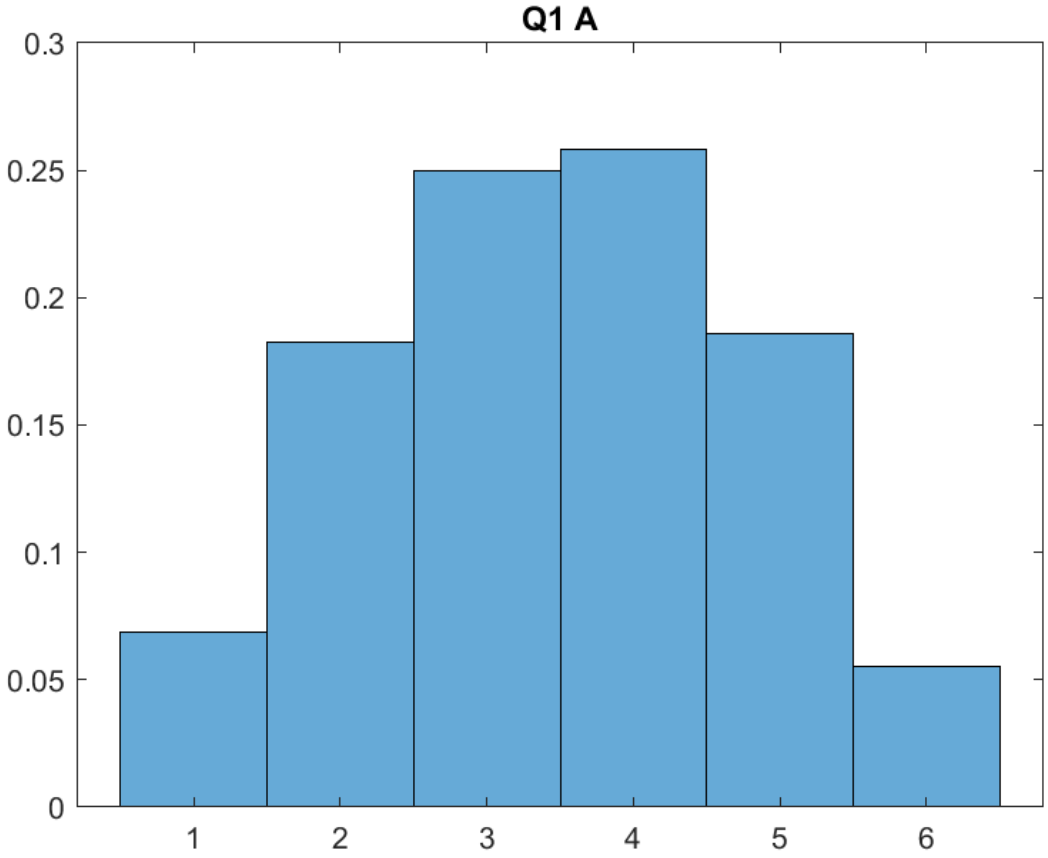
Stats Midterm Assignment - John Sinclair - 16325734

Q1.

(a) I wrote a matlab script that looped through each row in the dataset, summing the values of column 1, 2 and 3 and storing the sum in row i of a new array. I then plotted this array using matlab's histogram function to obtain this plot:



(b) P(Zi,1) = 0.4771

To calculate P(Zi,1 = 1) I wrote a matlab script to count how many occurrences of 1 there was in the first column then divided that number by the number of values in column 1.

(c) CLT: 0.43732 <= P(Zi,1 = 1) <= 0.51745 Cheb: 0.38596 <= P(Zi,1 = 1) <= 0.56881

To derive confidence intervals using chebyshev and CLT I used my value for P(Zi,1 = 1) and calculated the standard deviation using the formula sqrt(prob * (1 - prob)).

I then used the following formulas to calculate the upper and lower bounds for each method.

CLT: -1.96 * (σ / sqrt(n)) + mean <= Y <= 1.96 * (σ / sqrt(n)) + mean

Cheb: mean - (σ / sqrt(n * 0.05)) <= Y <= mean + (σ / sqrt(n * 0.05)

The central limit theorem provided a smaller interval compared to the chebyshev method.

(d) N = 10004

To obtain an accuracy of + or - 1% we know that 2 times the standard deviation = 0.01, and we know σ^2 = variance / N.

So, 2 x sqrt(variance / N)) = 0.01

So, N = variance / $(0.01 / 2)^2$ = 0.2501 / $(0.01 / 2)^2$ = 10004

Q2.

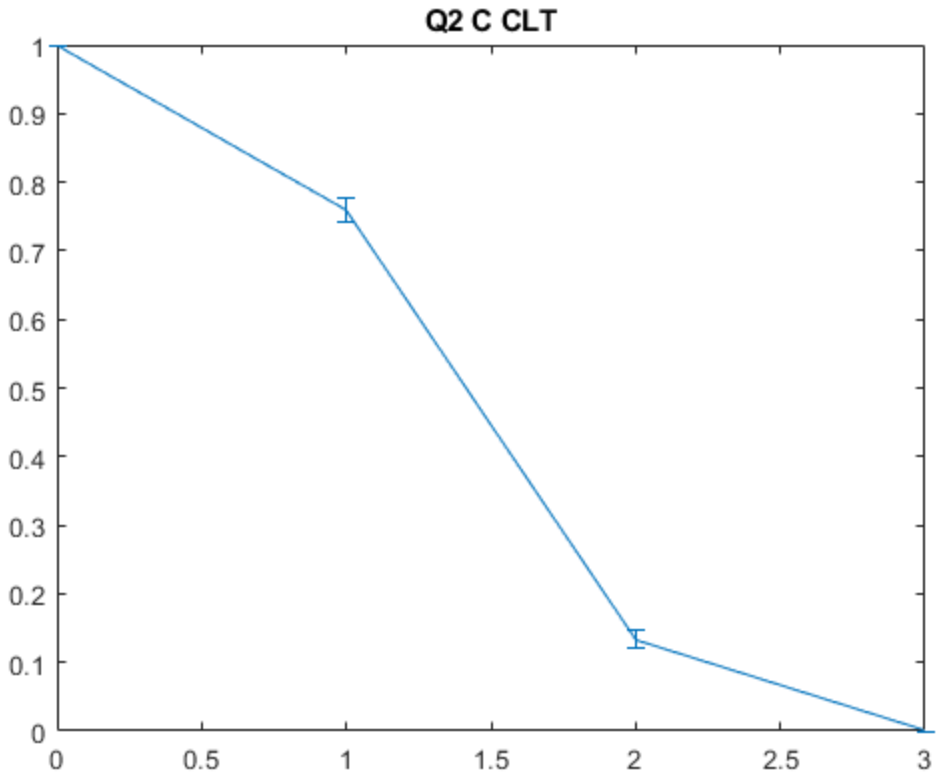
(a) I wrote a matlab script that checks the value z in column 2 and updates the count of z occurrences and if the corresponding value in column 1 was 1 incremented the respective total. Then for each value in column 2 (0, 1, 2, 3) i divided the respective total by the number of occurrences.

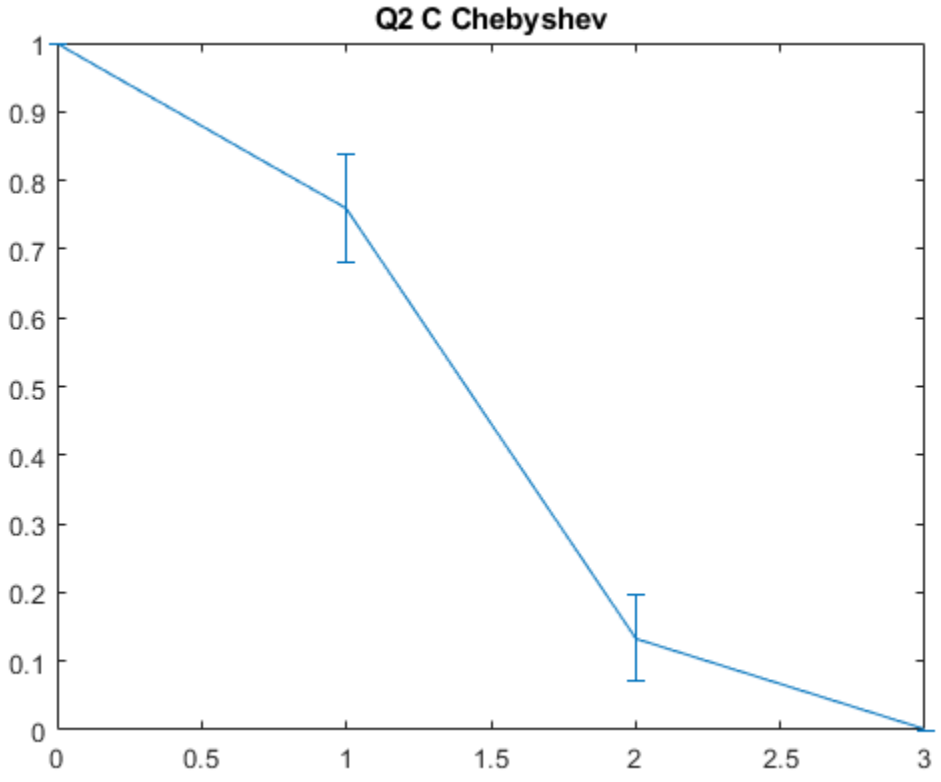
Column 2 Value	Column 1 Mean
0	1
1	0.75862
2	0.13376
3	0

(b) Note: Y = E [Zi,1 | Zi,2 = z], I used the same method and formulas from Q1 part c but this time repeated for each value of column 2

Col 2 Value	Y	CLT	Chebyshev
0	1	1 <= Y <= 1	1 <= Y <= 1
1	0.75862	0.72429 <= Y<= 0.79295	0.6803 <= Y <= 0.83694
2	0.13376	0.10645 <= Y <= 0.16106	0.071455 <= Y <= 0.19606
3	0	0 <= Y <= 0	0 <= Y <= 0

(c) The plotted graphs both show a similar downward trend, this indicates a correlation between the value in column 2 and the value in column 1. The CLT confidence intervals are much tighter than the Chebyshev interval bounds and there is no interval at Z = 0 or 3.


Q2 C CLT
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3


Q2 C Chebyshev
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3

(d) From my initial estimates, there was a 47.47% chance item 1 was in the basket. It is clear the presence of item two in the basket is predictive of item 1's presence as there is two cases in which we can have absolute confidence of items ones presence based purely on the quantity of item 2 in the basket; when there is no item 2 in the basket there is a 100% chance item 1 is present and conversely if there are 3 items 2's in the basket we can say with absolute certainty there is no item 1. As in the other two cases, when there are 1 or 2 item 2's in the basket, the probability and confidence intervals are such that we can expect the presence of item one when there is just one items two as well as we expect no item 1 when the basket contains 2 item 2's.

Q3.

(a) When the dataset is reduced to just the first 100 rows the range of my confidence intervals increase. This is because with more data we can be more confident when observing a trend, with less data we cannot say with the same certainty that our observations are fully reflective.

(b) When I re ran my calculations conditioned on column 3 rather than column 2 these were my results:

Col 3 Value	Y	CLT	Chebyshev
0	0.40667	0.36726 <= Y <= 0.44607	0.31676 <= Y <= 0.49657
1	0.46479	0.42478 <= Y<= 0.5048	0.3735 <= Y <= 0.55608
2	0.50336	0.46325 <= Y <= 0.54346	0.41184 <= Y <= 0.59487
3	0.53205	0.49203 <= Y <= 0.57208	0.44072 <= Y <= 0.62338

From the results above we can see that when conditioned on the third column we can no longer predict with any greater confidence than the original mean found in Q1. Column 2 is far more predictive of the expected value in column one than column 3 is.

Code Appendix:

```
data = datasetvalues;
n = 597;
```

Q1 (a)

```
sums = zeros(n,1);
for i=1:n
    sum = data(i,1) + data(i,2) + data(i,3);
    sums(i) = sum;
end
histogram(sums,'Normalization','probability')
title('Q1 A')
```

Q1 (b)

```
tot = 0;
for i=1:n
    tot = tot + data(i,1);
end
prob = tot / n;
q1b = prob
```

Q1 (c)

```
sigma = sqrt(prob*(1 - prob));

clt = sigma / sqrt(n);
cheb = sigma / sqrt(0.05*n);
cltlower = -1.96*clt + prob;
clthigher = 1.96*clt + prob;
cheblower = prob - cheb;
chebhigher = prob + cheb;

names = {'Mean', 'CLT Lower', 'CLT Higher', 'Cheb Lower', 'Cheb
Higher'};
q1c = table(prob, cltlower, clthigher, cheblower, chebhigher,
'VariableNames', names)
```

Q2 (a)

```
zs = zeros(4,1);
counts = zeros(4,1);

for i=1:n
    if (data(i,2) == 0)
        if (data(i,1) == 1)
            zs(1) = zs(1)+1;
        end
```

```
        counts(1) = counts(1)+1;
    elseif (data(i,2) == 1)
        if (data(i,1) == 1)
            zs(2) = zs(2)+1;
        end
        counts(2) = counts(2)+1;
    elseif (data(i,2) == 2)
        if (data(i,1) == 1)
            zs(3) = zs(3)+1;
        end
        counts(3) = counts(3)+1;
    elseif (data(i,2) == 3)
        if (data(i,1) == 1)
            zs(4) = zs(4)+1;
        end
        counts(4) = counts(4)+1;
    end
end

means
=[zs(1)/counts(1);zs(2)/counts(2);zs(3)/counts(3);zs(4)/counts(4)];
q2a = table([0;1;2;3], means, 'VariableNames',{'Col 2 Value','Col 1
Mean'})
```

Q2 (b)

```
sigmas = [0;0;0;0];
for i=1:4
    sigmas(i) = sqrt(means(i)*(1 - means(i)));
end

cltlowers = zeros(4,1);
clthighers = zeros(4,1);
cheblowers = zeros(4,1);
chebhighers = zeros(4,1);

clt = zeros(4,1);
cheb = zeros(4,1);

for i=1:4
    clt(i) = sigmas(i)/sqrt(n);
    cheb(i) = sigmas(i) / sqrt(0.05*n);
    cltlowers(i) = -1.96*(sigmas(i)/sqrt(n)) + means(i);
    clthighers(i) = 1.96*(sigmas(i)/sqrt(n)) + means(i);
    cheblowers(i) = means(i) - (sigmas(i) / sqrt(0.05*n));
    chebhighers(i) = means(i) + (sigmas(i) / sqrt(0.05*n));
```

```
end

names = {'Col 2 Value','Col 1 Mean', 'CLT Lower', 'CLT Higher', 'Cheb
Lower', 'Cheb Higher'};
q2b = table([0;1;2;3], means, cltlowers, clthighers, cheblowers,
chebhighers, 'VariableNames', names)
```

Q2 (c)

```
errorbar(means, cheb)
title('Q2 C')
```

Q3 (c)

```
zs = zeros(4,1);
counts = zeros(4,1);

for i=1:n
    if (data(i,3) == 0)
        if (data(i,1) == 1)
            zs(1) = zs(1)+1;
        end
        counts(1) = counts(1)+1;
    elseif (data(i,3) == 1)
        if (data(i,1) == 1)
            zs(2) = zs(2)+1;
        end
        counts(2) = counts(2)+1;
    elseif (data(i,3) == 2)
        if (data(i,1) == 1)
            zs(3) = zs(3)+1;
        end
        counts(3) = counts(3)+1;
    elseif (data(i,3) == 3)
        if (data(i,1) == 1)
            zs(4) = zs(4)+1;
        end
        counts(4) = counts(4)+1;
    end
end

means
=[zs(1)/counts(1);zs(2)/counts(2);zs(3)/counts(3);zs(4)/counts(4)];

sigmas = [0;0;0;0];
for i=1:4
    sigmas(i) = sqrt(means(i)*(1 - means(i)));
```

```
end

cltlowers = zeros(4,1);
clthighers = zeros(4,1);
cheblowers = zeros(4,1);
chebhighers = zeros(4,1);

clt = zeros(4,1);
cheb = zeros(4,1);

for i=1:4
    clt(i) = sigmas(i)/sqrt(n);
    cheb(i) = sigmas(i) / sqrt(0.05*n);
    cltlowers(i) = -1.96*(sigmas(i)/sqrt(n)) + means(i);
    clthighers(i) = 1.96*(sigmas(i)/sqrt(n)) + means(i);
    cheblowers(i) = means(i) - (sigmas(i) / sqrt(0.05*n));
    chebhighers(i) = means(i) + (sigmas(i) / sqrt(0.05*n));
end

names = {'Col 2 Value','Col 1 Mean', 'CLT Lower', 'CLT Higher', 'Cheb
Lower', 'Cheb Higher'};
q3b = table([0;1;2;3], means, cltlowers, clthighers, cheblowers,
chebhighers, 'VariableNames', names)
```