ATMAM Math Methods

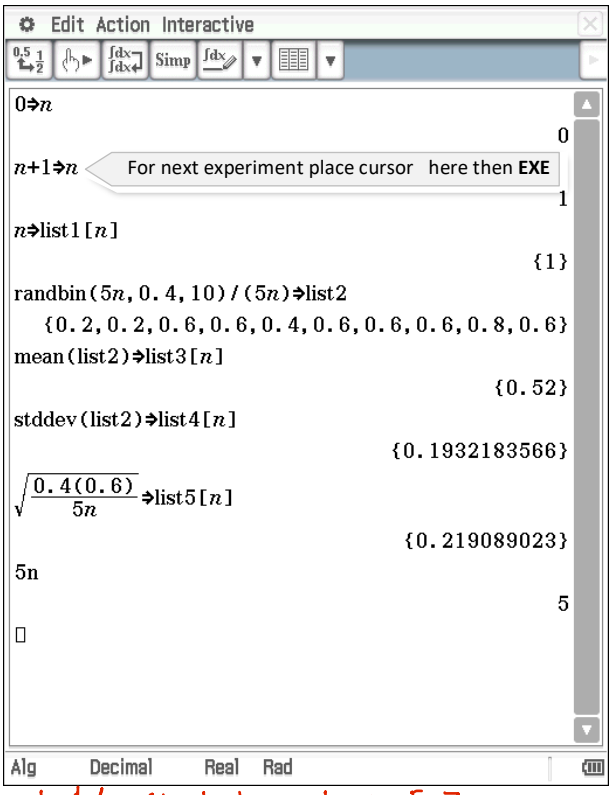
Random Samples and Sampling Distributions

Total: 50 marks

Question 1 **(16 marks)**

A random sampling experiment is being simulated on a CAS calculator to investigate the impact on sample size versus repeated sampling in terms of the convergence of the experimental mean and standard deviation to the population mean and estimated standard deviation according to $\sigma_{\hat{p}} \approx \sqrt{\frac{\sigma^2}{n}}$, where σ is the population standard deviation.

a) The code below has been entered into a CAS calculator. Enter this code into your calculator and give a brief explanation of each line of code, including why the cursor needs to be placed in the second line to create the data for the next experiment. **(5 marks)**



0⇒n Sets the count parameter to ZERO

n+1⇒n Adds 1 to the count and overwrites it to n

n⇒list1[n] Stores the current value of n to the n^{th} row of list1 creating a "count list". This keeps track of the number of experiments.

randbin(5n, 0.4, 10)/(5n)

0.4 is the chance of success = population prop.

10 is the number of samples taken

5n is the sample size of each of those samples

/(5n) without this the function returns the number of success in each sample with it, it returns the sample proportion ($\hat{p}$) of each sample.

mean(list2) ⇒ list3[n]

Calculates the mean of the current (n^{th} experiment) set of sample proportions and stores it to the n^{th} row of list3

stddev(list2) ⇒ list4[n]

Calculates the standard deviation of the current set (n^{th} experiment) of sample proportions and stores it to the n^{th} row of list4

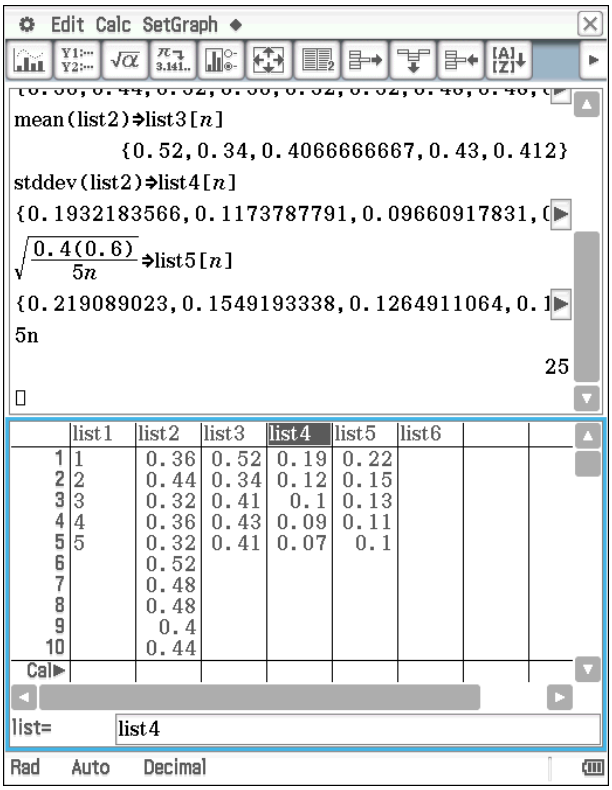
$\sqrt{\frac{0.4(0.6)}{5n}}$ ⇒ list5[n] Calculates the estimated standard deviation for a sample size of 5n of the n^{th} experiment and stores it to the n^{th} row of list 5.

Cursor starts at n+1⇒n to ensure that the count is increased by 1 to begin the next (n^{th}) experiment. If the cursor is placed in the first line, the code will not progress beyond experiment 1 and the data in the first row of each list will be continually over written.

After 5 experiments

Use your calculator to run 5 experiments. Your CAS calculator display will show the results similar to those below.

b) For the table below, explain what is shown in each column and in each row , including an explanation as to why list2 has a different number of rows to the other lists. **(5 marks)**



list1 Experiment Number

list2 The current set of 10 sample proportions ($\hat{p}$) from the 5^{th} experiment with sample size = 25. This list is over written with every experiment.

list3 The mean of list2 for each experiment.

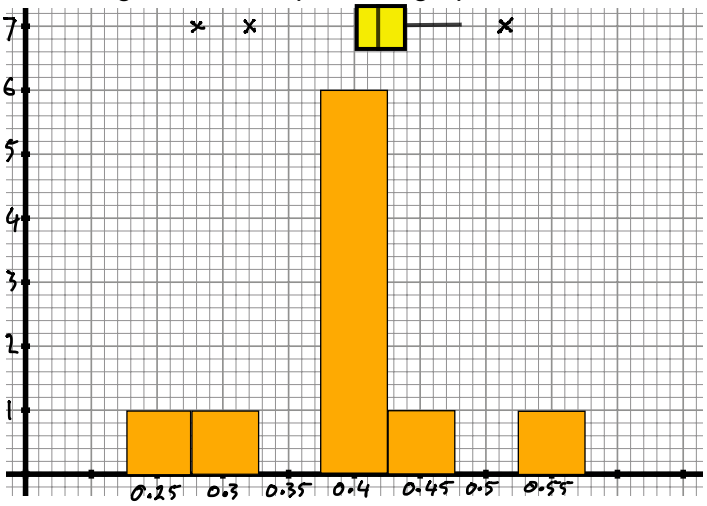
list4 The standard deviation of list 2 for each experiment

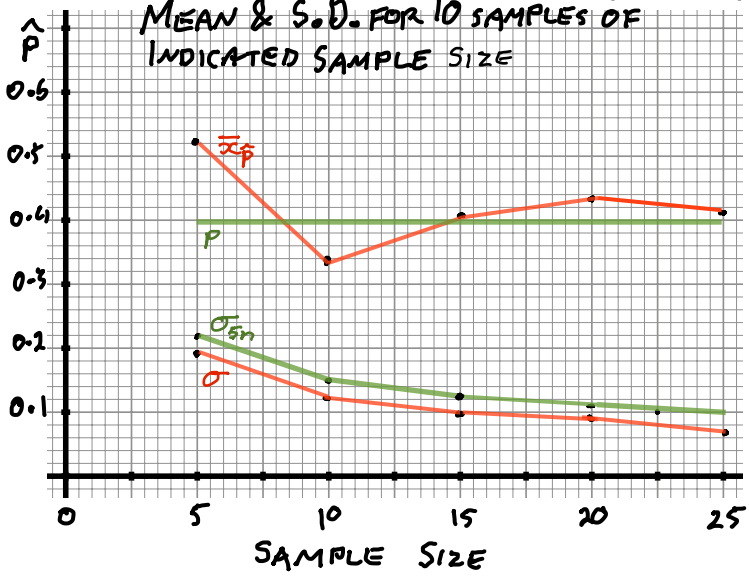
list5 The theoretical estimate of the standard deviation of sample size $5n$ (in this case = 25) using

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

where p = population proportion
n = sample size

c) Use the instructions on the following page to draw graphs of **your** results on the axes below, including: a histogram and boxplot; line graphs of the mean and standard deviation. **(6 marks)**





Comment on your results.

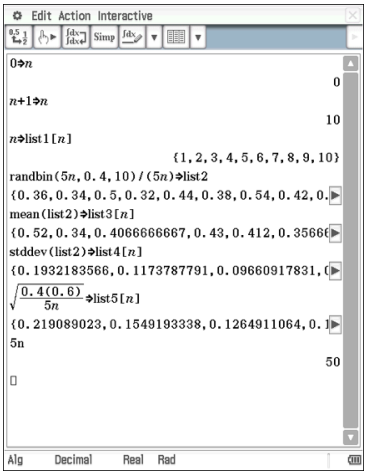
Regards the distribution: at this point it is not symmetrical and is skewed right with $M \neq \mu$. Lack of symmetry prevents regarding this as a normal distribution.

On just 5 data points it would appear that the experimental mean is beginning to converge on the population proportion, fluctuating above and below. The experimental standard deviation is reducing and approximating both the value and behaviour of the theoretical estimate.

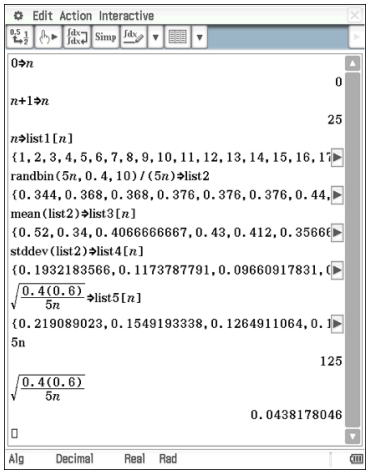
Question 2 **(6 marks)**

CASE 1 *randbin(5n, 0.4, 10)*

After 10 Experiments



After 25 Experiments



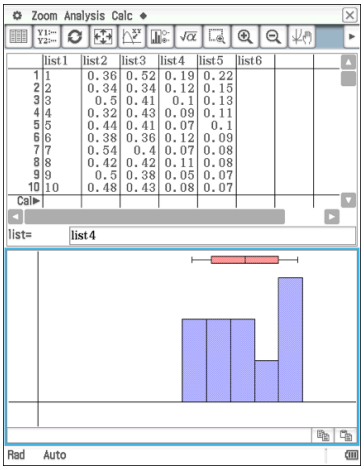
After 50 Experiments

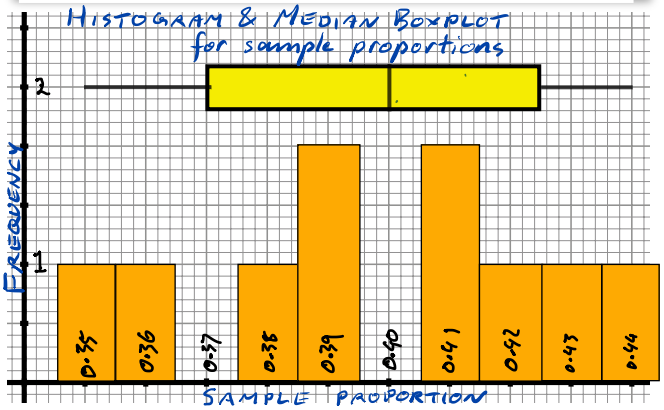
Shown are the results after 10 experiments and 25 experiments.

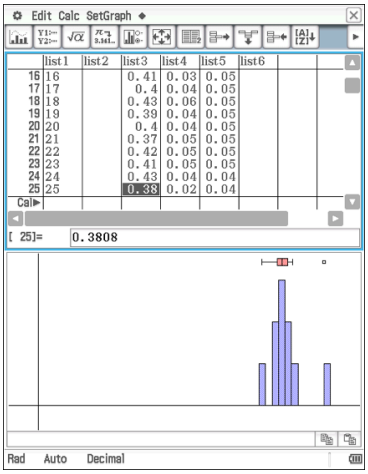
Use your calculator to complete 50 experiments and draw a histogram and boxplot graph and, on the next axes, line graphs of the progressive means and standard deviations respectively.

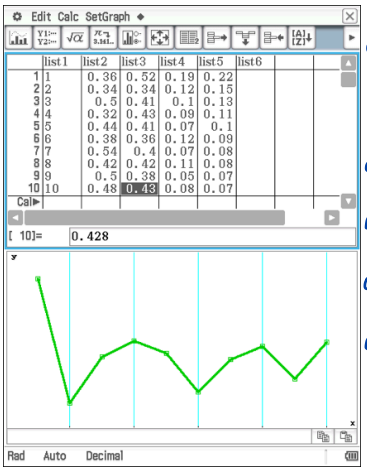
Alternatively – glue screenshots in place, or complete the investigation in MS Word.

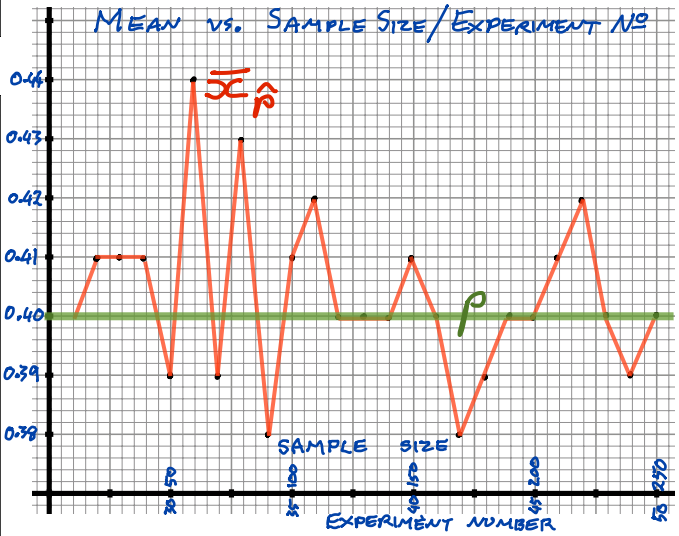
NB when drawing the line graphs by hand, just use the last 25 data points.

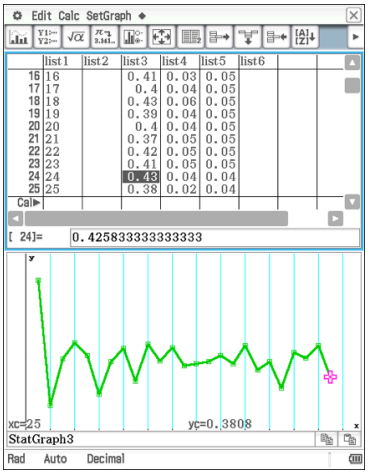


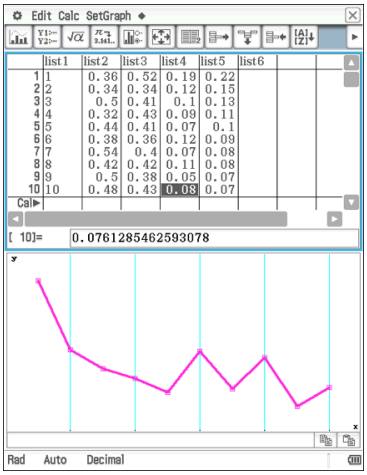


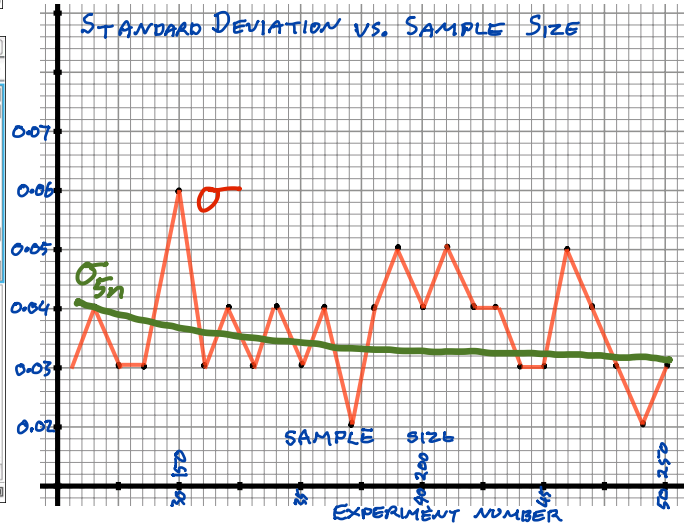


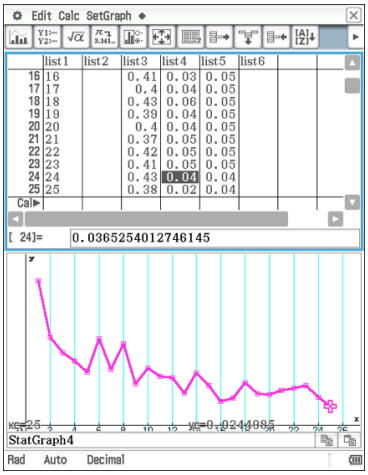








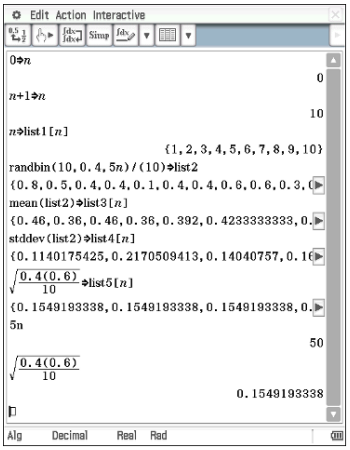




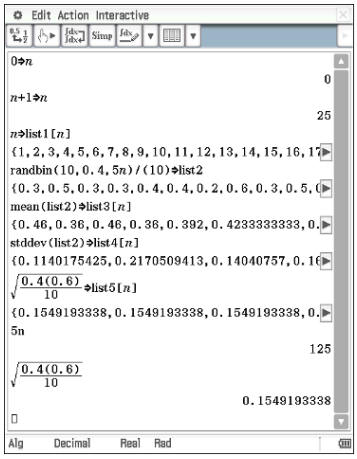
Question 3 **(6 marks)**

CASE 2 *randbin(10, 0.4, 5n)*

After 10 Experiments



After 25 Experiments



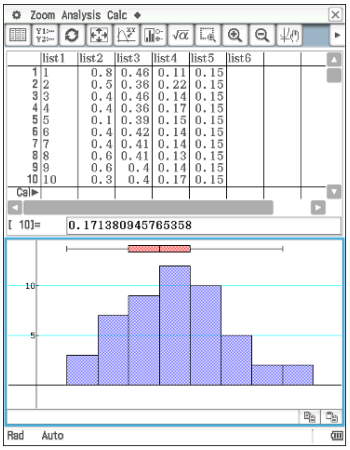
After 50 Experiments

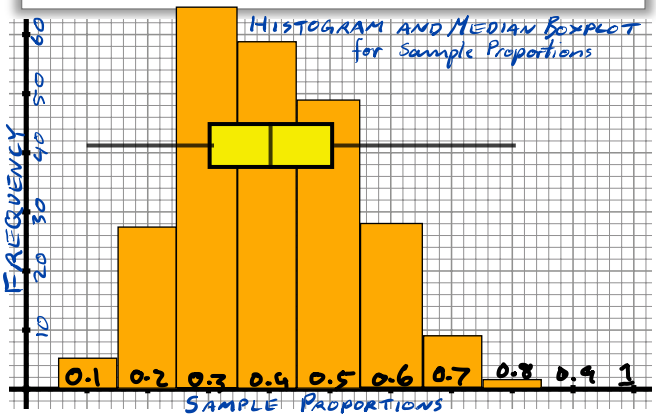
Shown are the results for a sample size of 10 and repeats of $5n$ after 10 experiments and 25 experiments.

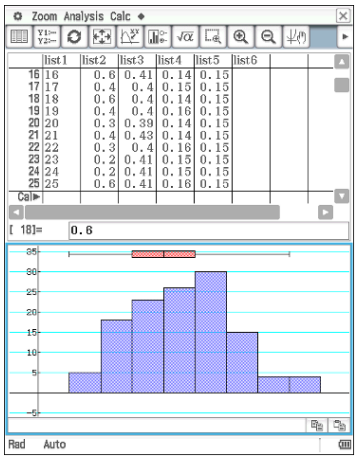
Use your calculator to complete 50 experiments and draw histogram and boxplot graph and, on the next axes, line graphs of the progressive means and standard deviations respectively.

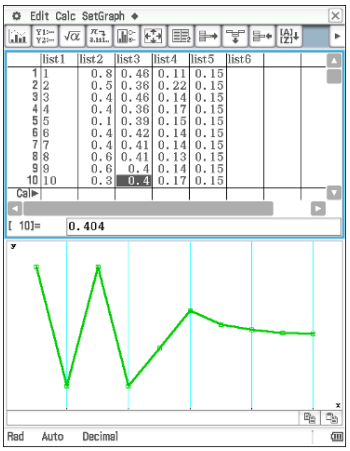
Alternatively – glue screenshots in place, or complete the investigation in MS Word.

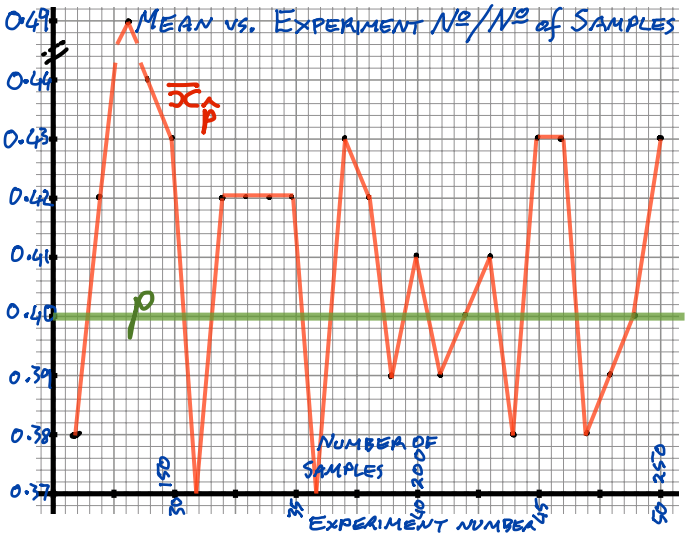
NB when drawing the line graphs by hand, just use the last 25 data points.

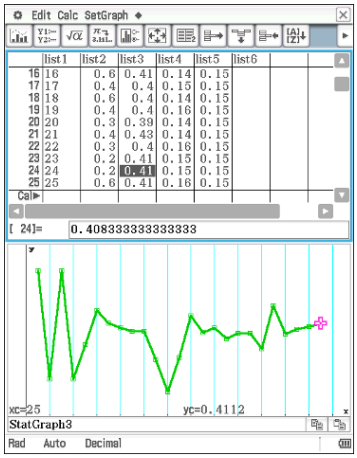


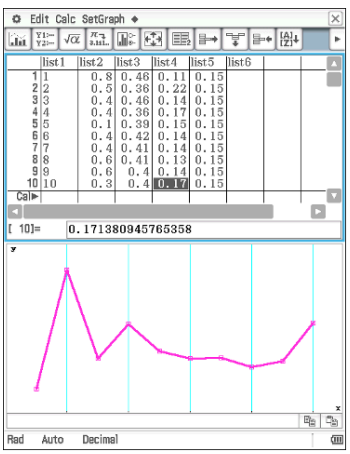


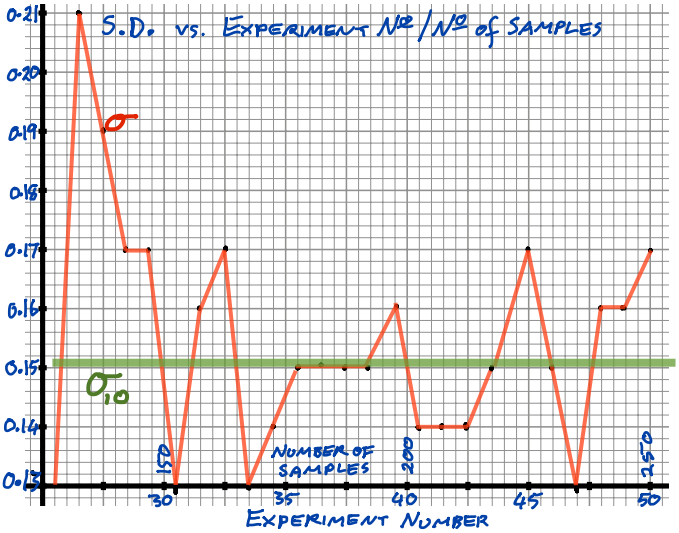


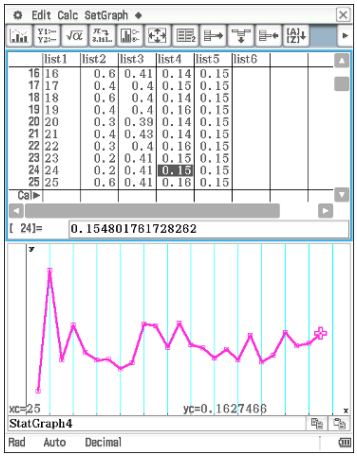












Question 4 **(8 marks)**

Comment on your results from Question 2 and Question 3 regards larger sample sizes and fewer repeats versus smaller sample sizes and a large number of repeats.

On the 50th experiment both cases process 2500 data points altogether

Case 1: 10 x samples of 250 (with mean & S.D. ultimately calculated from 10 sample proportions)

Case 2: 250 x samples of 10 (with mean & S.D. ultimately calculated from 250 sample proportions)

Progressive Mean Case 1 vs Case 2

From the graphed data the progressive *mean of the sample proportions* ($\bar{x}_{\hat{p}}$) appears to converge onto the true mean or *population proportion* (p) more quickly. Taking the last 5 data points of **Case 1**, $\bar{x}_{\hat{p}}$ is separated by a maximum of 3 hundredths whereas for **Case 2** it is separated by a maximum of 5 hundredths. Across the last 25 data points **Case 2** swings by 11 hundredths between the 27th and 30th data points with an error of 9 hundredths from the mean. **Case 1** has a maximum swing of 5 hundredths between the 31st and 32nd data points, with a maximum error from the mean of 4 hundredths.

Progressive Standard Deviation Case 1 vs. Case 2

The Standard Deviation of **Case 1** and **Case 2** both converge on their theoretical estimates

Case 1 will have $\sigma_{250} \approx 0.031$ (50th experiment)

Case 2 will have $\sigma_{10} \approx 0.155$

Note that for **Case 2** σ_{5n} is fixed at $\approx 0.155 = \sigma_{10}$ since the *sample size* is fixed (hence a horizontal line on the graph). For **Case 1**, however, σ_{5n} is a decreasing function asymptotic to the $x - axis$, decreasing from 0.042 in the 26th experiment to 0.031 in the 50th. This means that for the 26th Experiment the **Case 2 S.D.** is $3.7 \times$ **Case 1 S.D.** and ≈5 times larger by the 50th experiment. Although both **Case 1** and **Case 2** are converging on their predicted S.D. the S.D. for **Case 1** << **Case 2**. Thus on the 50th experiment **Case 2** has $\bar{x}_{\hat{p}} \approx 0.43$, with $\sigma_{10} \approx 0.16$, so 68% of the sample proportions ($\hat{p}$) lie between 0.59 and 0.27; that is, 0.43 ± 0.16 ($\pm 1\sigma$).

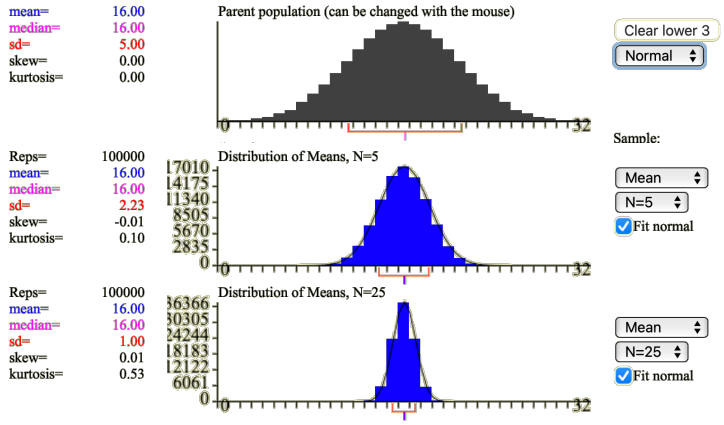
On the other hand that same raw error of 0.16 on $\bar{x}_{\hat{p}} \approx 0.4$ for the 50th experiment for **Case 1,** gives $0.4 \pm 0.16 = 0.4 \pm 5\sigma$ since for **Case 1** $\sigma \approx 0.03$. Since for $X \sim N(1,0)$ $P(-5 < X < 5) = 0.9999994$ then for **Case 1** 99.99994% of the *sample proportions* ($\hat{p}$) will lie between 0.24 and 0.56. This level of accuracy for the *sample proportion* is much greater than for **Case 2**, clearly demonstrating the importance of *sample size* over and above repeated samples in achieving a reliable *sample proportion* that is to be used as an indicator for the true *population proportion*.

Given that level of accuracy at a sample size of just 250, it also points to sample quality and sampling techniques as being the point at which the reliability of a sample proportion breaks down. It therefore underscores how important it is for surveys to be conducted to the highest standard, as poorly conducted surveys will just give highly accurate misrepresentations of the population proportion whatever the sample size.

Question 5 (14 marks) For a variety of distributions (including *normal* and *uniform distributions*), use the **Online Stat Book** interactive tool to investigate the distribution of *sample means* as the *sample size* increases.

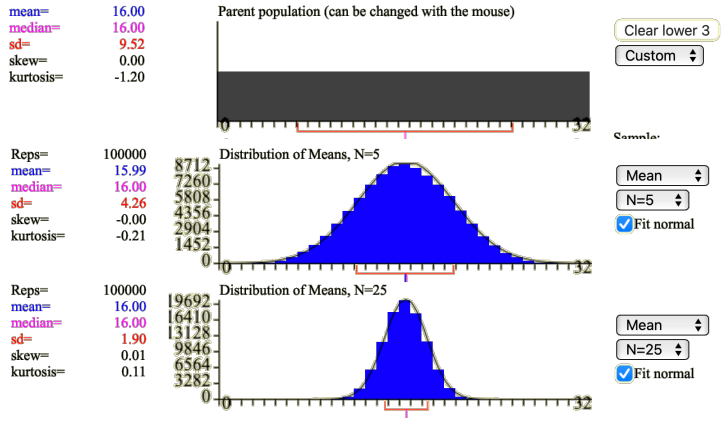
Sampling Distribution of the Sampling Mean: NORMAL DISTRIBUTION

100 000 means of Random Samples with Sample Size, n=5 & n=25



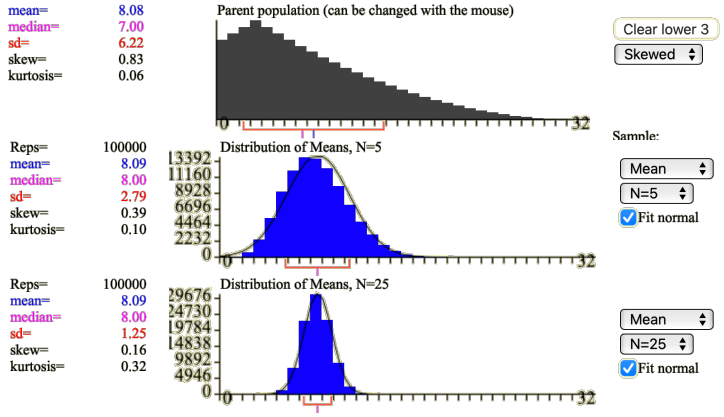
Sampling Distribution of the Sampling Mean: UNIFORM DISTRIBUTION

100 000 means of Random Samples with Sample Size, n=5 & n=25



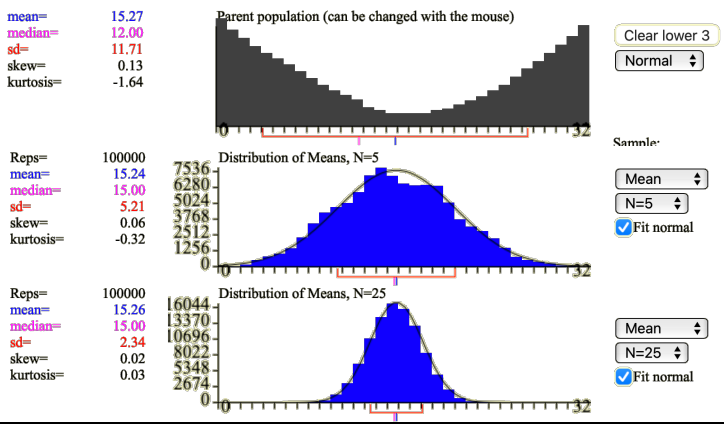
Sampling Distribution of the Sampling Mean: SKEWED DISTRIBUTION

100 000 means of Random Samples with Sample Size, n=5 & n=25



Sampling Distribution of the Sampling Mean: BIMODAL DISTRIBUTION

100 000 means of Random Samples with Sample Size, n=5 & n=25



It is clear from the four cases shown to the left that, regardless of the distribution of the original data set, the Sampling Distribution of the Sampling Mean can be modelled by a Normal Distribution. As the sample size increases and the number of samples increases we can say that the Sampling Distribution of the Sampling Mean is Normally Distributed.

In the case of Sample Proportions, since the Sample Proportion is already a summary statistic, we can equally say that the Sampling Distribution of the Sample Proportions is Normally Distributed for repeated sampling of a large enough sample size. This is true regardless of the nature of the distribution of the population from which the samples are taken.