TASK 16: SAMPLE MEANS
Extended Investigation
Unit 4
Topic 4.3: Statistical inference

Course-related information

The concepts and skills included in this investigation relate to the following dot points within the WA Mathematics Specialist syllabus.

4.3.1 examine the concept of the sample mean $\bar{X}$ as a random variable whose value varies between samples where X is a random variable with mean μ and standard deviation σ

4.3.2 simulate repeated random sampling, from a variety of distributions and a range of sample sizes, to illustrate properties of $\bar{X}$ across samples of a fixed size n, including its mean μ its standard deviation $\frac{\sigma}{\sqrt{n}}$ (where μ and σ are the mean and standard deviation of X), and its approximate normality if n is large

4.3.3 simulate repeated random sampling, from a variety of distributions and a range of sample sizes, to illustrate the approximate standard normality of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ for large samples $(n \geq 30)$, where s is the sample standard deviation

4.3.4 examine the concept of an interval estimate for a parameter associated with a random variable

4.3.5 examine the approximate confidence interval $\left(\bar{X}-\frac{zs}{\sqrt{n}}, \bar{X}+\frac{zs}{\sqrt{n}}\right)$ as an interval estimate for the population mean μ, where z is the appropriate quantile for the standard normal distribution

4.3.6 use simulation to illustrate variations in confidence intervals between samples and to show that most but not all confidence intervals contain μ

The ability to choose and use appropriate technology to enhance and extend concept development is essential in the completion of this investigation.

Background information

The Excel instructions used in this extended investigation relate to Excel 2010. It is assumed that students will have access to the *Excel* Add-in Analysis ToolPak. The instructions to access the Analysis ToolPak are provided prior to Task One.

Students need to be able to calculate the population mean and population standard deviation from the distribution parameters of the following distributions: Uniform, Binomial and Normal

Task conditions

The preparation activity assumes that students are familiar with Excel. Students are expected to complete the preparation activity before the in-class validation. Students should **not** be permitted to take their investigative material into the class assessment. It is assumed that students have access to a CAS calculator for the in-class validation.

SAMPLE MEANS

Extended investigation **Part 1: Preparation activity**

Note: For $X \sim U(a,b)$, $\mu = \dfrac{a+b}{2}$ and $\sigma = \dfrac{b-a}{\sqrt{12}}$. For $X \sim B(n,p)$, $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

The distribution formed when the following procedure is carried out is called the **sampling distribution of means**.

- Take a random sample of n independent observations from a population. If the population is finite, sampling should be with replacement to ensure that the observations are independent.
- Calculate the mean of these n sample values. This is known as the sample mean.
- Repeat the procedure until all possible samples of size n have been taken, calculating the sample mean of each one.
- Form a distribution of the sample means.

A sample value may be used to estimate an unknown population parameter by constructing an interval estimate, known as a **confidence interval**. This is an interval that has a specified probability of including the parameter. The probabilities most often used in confidence intervals are 90%, 95% and 99%. If the mean μ of a particular population is unknown, then determining a 95% confidence interval for it would mean constructing the interval (a,b) such that $P(a < \mu < b) = 0.95$.

The interval constructed uses the value of the mean $\bar{x}$ of a random sample of size n taken from the population.

Before constructing a confidence interval for μ, the following questions need to be answered.

- Is the distribution of the population normal?
- Is the variance of the population known?
- Is the sample small or large? (Usually $n \geq 30$ is considered a large sample.)

When calculating confidence intervals it is often the case that the population standard deviation σ is not known. Provided that the sample size is large, the sample standard deviation s may be used as an unbiased estimate for σ.

Open an Excel workbook and look for **Data Analysis** on the **Data** tab.

If the **Data Analysis** command does not appear in the **Data** tab, then follow the instructions below:

1. Click the **File** tab, click **Options**, and then click the **Add-ins** category.
2. In the **Manage** box, select **Excel Add-ins** and then click **Go**.
3. In the **Add-ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.

If you are prompted that the Analysis **ToolPak** is not currently installed on your computer, click **Yes** to install it.

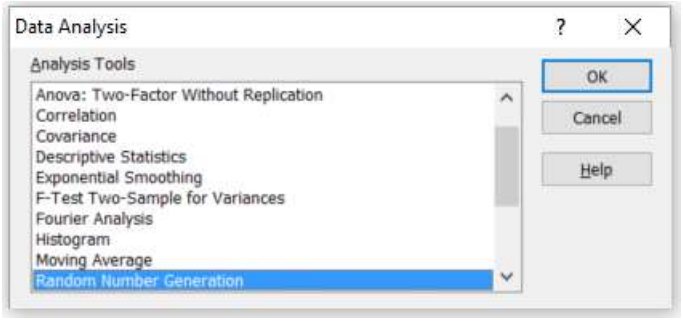
However, there is an issue with MacBook's, the Apple software version of Office and specifically the Excel version 2011 for Mac and 2008 for Mac, do not have the add-in, Data Analysis Tool-Pak.

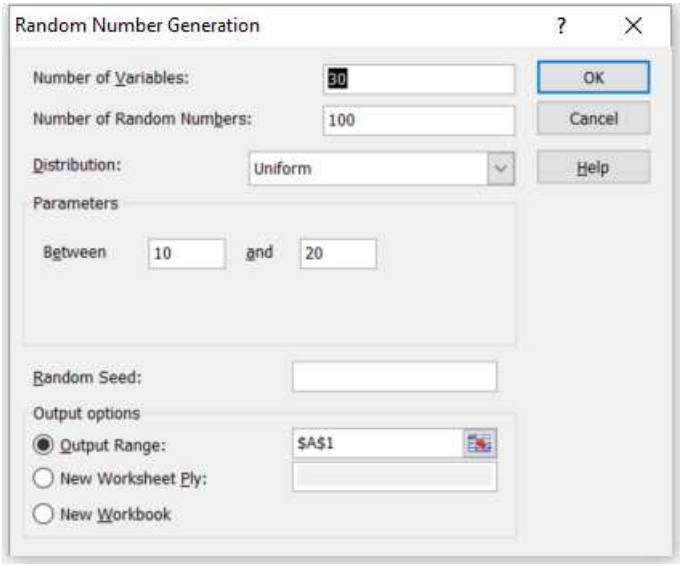
Please refer to the following link for further information about this issue
https://support.microsoft.com/en-au/kb/2431349

Task One:

Step 1: Open an Excel workbook and generate 100 samples each of size $n = 30$ from uniformly distributed numbers between 10 and 20, i.e. $X \sim U(10,20)$.

Method: Click on **Data Analysis** and then double click on **Random Number Generation**.





Number of variables: sample size

Number of Random Numbers: number of samples

Distribution: select distribution

Parameter: enter parameters for distribution

Random Seed: leave blank

Output options

Output Range: insert address of first cell of generated values

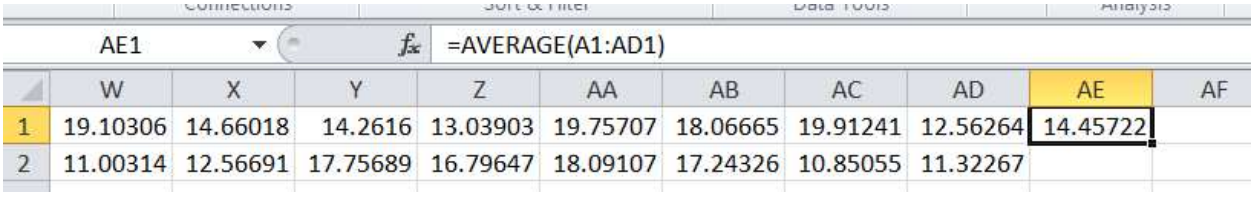
Click **OK**.

Each of the 100 rows contains 30 numbers between 10 and 20.

Step 2a: For each of the 100 samples, calculate the sample mean.

Method: Click in cell AE1 and type **=AVERAGE(A1:AD1)**

Click in cell AE1, grab the small square in the lower right hand corner of the cell and fill down to cell AE100.



	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	19.10306	14.66018	14.2616	13.03903	19.75707	18.06665	19.91241	12.56264	14.45722	
2	11.00314	12.56691	17.75689	16.79647	18.09107	17.24326	10.85055	11.32267		

Step 2b: For each of the 100 samples, calculate the sample standard deviation.

Method: Click in cell AF1 and type **=STDEV.S(A1:AD1)**
Click in cell AF1, grab the small square in the lower right hand corner of the cell and fill down to cell AF100.

Step 3a: Calculate the mean of the sample means, $\overline{X}$.

Method: Click in cell AE101 and type **=AVERAGE(AE1:AE100)**

Step 3b: Calculate the standard deviation of the sample means, $\sigma_{\overline{X}}$.

Method: Click in cell AE102 and type **=STDEV.P(AE1:AE100)**

Step 3c: Record the sample size, the number of samples, the type of distribution (including the distribution parameters), the mean of the sample means and the standard deviation of the sample means.

Step 4: For each sample, construct the 95% confidence interval for the population mean μ.
Since the samples were not taken from a Normal population, the variance of the population is known and the sample size is large, a 95% confidence interval is given by $\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$, where $\overline{x}$ is the sample mean, σ is the population standard deviation and n is the sample size.

Method: Click in cell AG1 and type **=AE1-1.96*(10/SQRT(12))/SQRT(30)**
Click in cell AG1, grab the small square in the lower right hand corner of the cell and fill down to cell AG100.
Click in cell AH1 and type **= AE1+1.96*(10/SQRT(12))/SQRT(30)**
Click in cell AH1, grab the small square in the lower right hand corner of the cell and fill down to cell AH100.

Step 5a: For each sample, test whether or not the population mean lies within the 95% confidence interval.

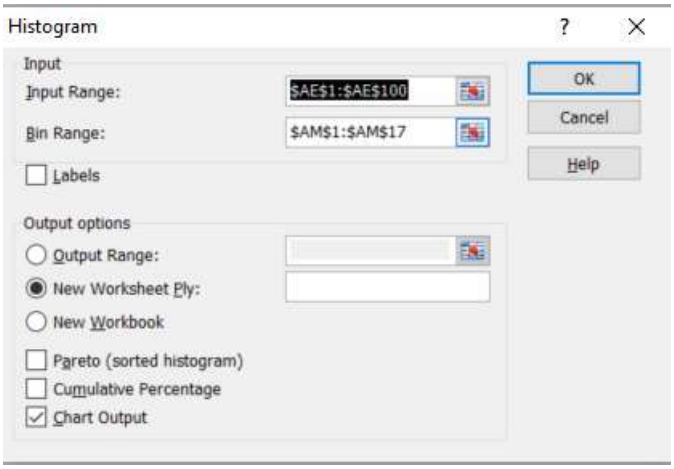
Method: Click in cell AI1 and type **=IF(AND(15>AG1,15<AH1),1,0)**
Click in cell AI1, grab the small square in the lower right hand corner of the cell and fill down to cell AI100.
If the population mean lies within the 95% confidence interval the value 1 is returned, if not the value of 0 is returned.

Step 5b: Determine how many of the 95% confidence intervals contain the population mean.

Method: Click in cell AI101 and type **=SUM(AI1:AI100)**

Step 6: Arrange the values of the sample means that are currently in AE1 to AE100 into intervals
13 to 13.25, 13.25 to 13.5, …, 16.5 to 16.75, 16.75 to 17 and draw a histogram.

Method: Click in cell AM1 and type **13**
Click in cell AM2 and type **=AM1+0.25**
Click in cell AM2, grab the small square in the lower right hand corner of the cell and fill down to cell AM17. (Cell AM17 should show the value 17.)
Click on **Data Analysis** and then double click on **Histogram**.



Input Range: AE1 to AE100
Bin Range: AM1 to AM17

Output options
New Worksheet Ply:

Check Chart Output

Click **OK**.

Click on the word **Bin** and press **Delete**.
To remove the space between the bars, right click on a bar, select **Format Data Series** and change the **Gap Width** to 0%. Select **Border Color** to add a border.

Step 7a: For each sample, calculate $z = \dfrac{\overline{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$.

Method: Click in cell AJ1 and type **=(AE1-15)/((10/SQRT(12))/SQRT(30))**
Click in cell AJ1, grab the small square in the lower right hand corner of the cell and fill down to cell AJ100.

Step 7b: Calculate the mean of these z values.

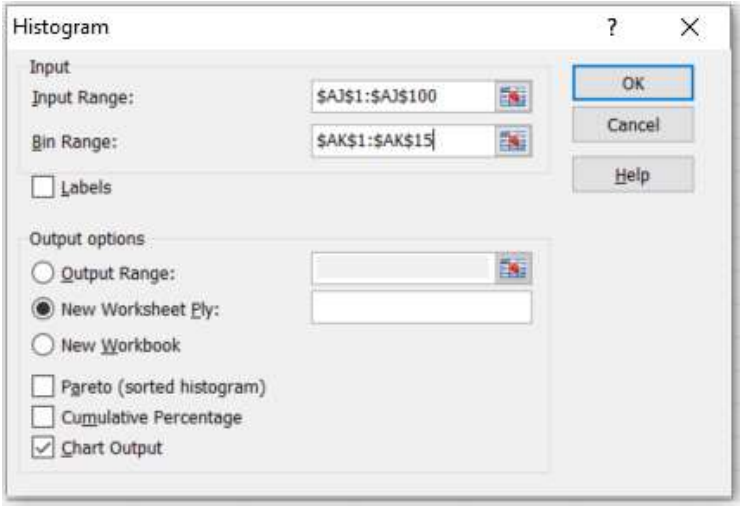
Method: Click in cell AJ101 and type **=AVERAGE(AJ1:AJ100)**

Step 7c: Calculate the standard deviation of these z values.

Method: Click in cell AJ102 and type **=STDEV.P(AJ1:AJ100)**

Step 8: Arrange the values $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ that are currently in AJ1 to AJ100 into intervals -3.5 to -3, -3 to -2.5, -2.5 to -2, …, 2.5 to 3, 3 to 3.5 and draw a histogram.

Method: Click in cell AK1 and type **-3.5**
Click in cell AK2 and type **=AK1+0.5**
Click in cell AK2, grab the small square in the lower right hand corner of the cell and fill down to cell AK15. (Cell AK15 should show the value 3.5.)
Click on **Data Analysis** and then double click on **Histogram**.



Input Range: AJ1 to AJ100

Bin Range: AK1 to AK15

Output options

New Worksheet Ply:

Check Chart Output

Click **OK**.

Click on the word **Bin** and press **Delete**.
To remove the space between the bars, right click on a bar, select **Format Data Series** and change the **Gap Width** to 0%. Select **Border Color** to add a border.

It is often the case that the population standard deviation σ is not known. Provided that the sample size is large, the sample standard deviation s may be used as an unbiased estimate for σ. Steps 4, 5, 7 and 8 will be repeated using s instead of σ.

Step 9: For each sample, construct the 95% confidence interval for the population mean μ using $\left(\bar{x} - 1.96\dfrac{s}{\sqrt{n}}, \bar{x} + 1.96\dfrac{s}{\sqrt{n}}\right)$, where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the sample size.

Method: Click in cell AO1 and type **=AE1-1.96*AF1/SQRT(30)**

Click in cell AO1, grab the small square in the lower right hand corner of the cell and fill down to cell AO100.

Click in cell AP1 and type **= AE1+1.96*AF1/SQRT(30)**

Click in cell AP1, grab the small square in the lower right hand corner of the cell and fill down to cell AP100.

Step 10a: For each sample, test whether or not the population mean lies within the 95% confidence interval.

Method: Click in cell AQ1 and type **=IF(AND(15>AO1,15<AP1),1,0)**

Click in cell AQ1, grab the small square in the lower right hand corner of the cell and fill down to cell AQ100.

If the population mean lies within the 95% confidence interval the value 1 is returned, if not the value of 0 is returned.

Step 10b: Determine how many of the 95% confidence intervals contain the population mean.

Method: Click in cell AQ101 and type **=SUM(AQ1:AQ100)**

Step 11a: For each sample, calculate $z = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$.

Method: Click in cell AR1 and type **=(AE1-15)/(AF1/SQRT(30))**

Click in cell AR1, grab the small square in the lower right hand corner of the cell and fill down to cell AR100.

Step 11b: Calculate the mean of these z values.

Method: Click in cell AR101 and type **=AVERAGE(AR1:AR100)**

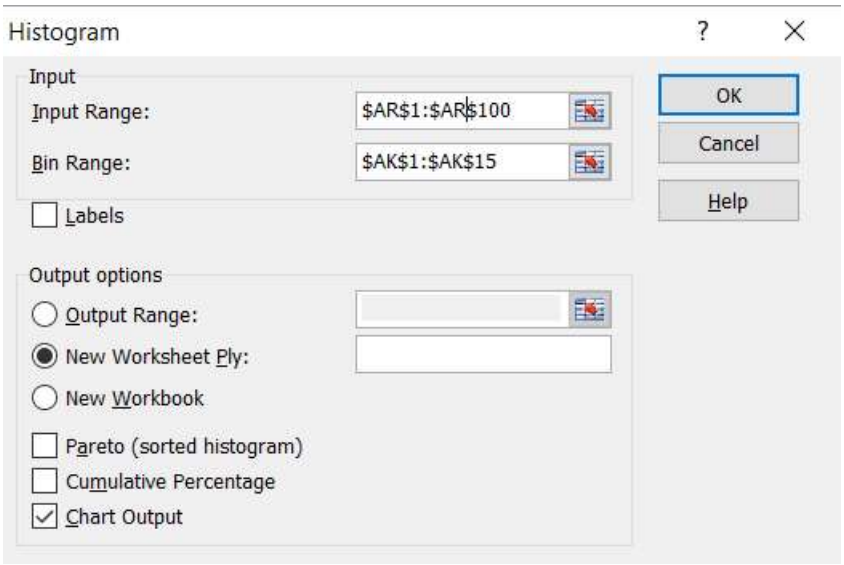
Step 11c: Calculate the standard deviation of these z values.

Method: Click in cell AR102 and type **=STDEV.P(AR1:AR100)**

Step 12: Arrange the values $z = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ that are currently in AR1 to AR100 into intervals

-3.5 to -3, -3 to -2.5, -2.5 to -2, …, 2.5 to 3, 3 to 3.5 and draw a histogram.

Click on **Data Analysis** and then double click on **Histogram**.



Input Range: AR1 to AR100

Bin Range: AK1 to AK15

Output options

New Worksheet Ply:

Check Chart Output

Click **OK**.

Click on the word **Bin** and press **Delete**.

To remove the space between the bars, right click on a bar, select **Format Data Series** and change the **Gap Width** to 0%. Select **Border Color** to add a border.

Task One Questions

1 In Step 2a, the mean $\bar{x}$ of each of the 100 samples of size $n = 30$ was calculated. Do each of your samples have the same sample mean or does this value vary between the samples?

2 (a) Calculate the population mean μ for $X \sim U(10,20)$.

(b) In Step 3a, the mean of the 100 sample means was calculated. Compare your answer with the value of μ.

3 In Step 2b, the standard deviation s of each of the 100 samples of size $n = 30$ was calculated. Do each of your samples have the same sample standard deviation or does this value vary between the samples?

4 (a) Calculate the population standard deviation σ for $X \sim U(10,20)$.

(b) In Step 3b, the standard deviation of the 100 sample means was calculated. Why is the standard deviation of the 100 sample means less than the population standard deviation σ?

5 In Step 4, 95% confidence intervals for μ were constructed using $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ whilst in Step 5 each confidence interval was tested to determine whether or not μ was within the interval. What percentage of the confidence intervals contained μ?

6 In Step 6, a histogram was drawn using the values of the sample means. What do you notice about the shape of this distribution?

7 In Step 7a, the value of $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ was calculated for each sample. In Steps 7b and 7c, the mean and standard deviation of the values of z were calculated. What do you notice about the mean and standard deviation of z?

8 In Step 8, a histogram was drawn using the values of $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$. What do you notice about the shape of this distribution?

9 In Step 9, 95% confidence intervals for μ were constructed using $\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$ whilst in Step 10 each confidence interval was tested to determine whether or not μ was within the interval. What percentage of the confidence intervals contained μ?

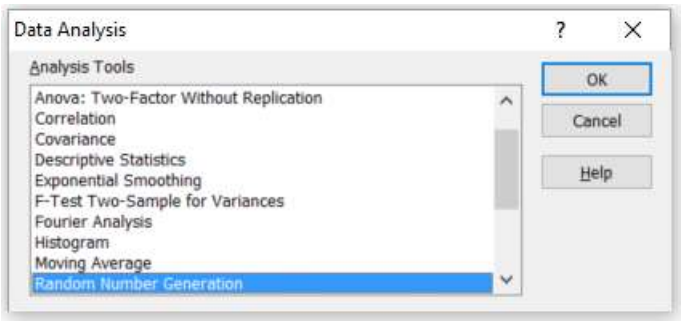
10 In Step 11a, the value of $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ was calculated for each sample. In Steps 11b and 11c, the mean and standard deviation of the values of z were calculated. What do you notice about the mean and standard deviation of z?

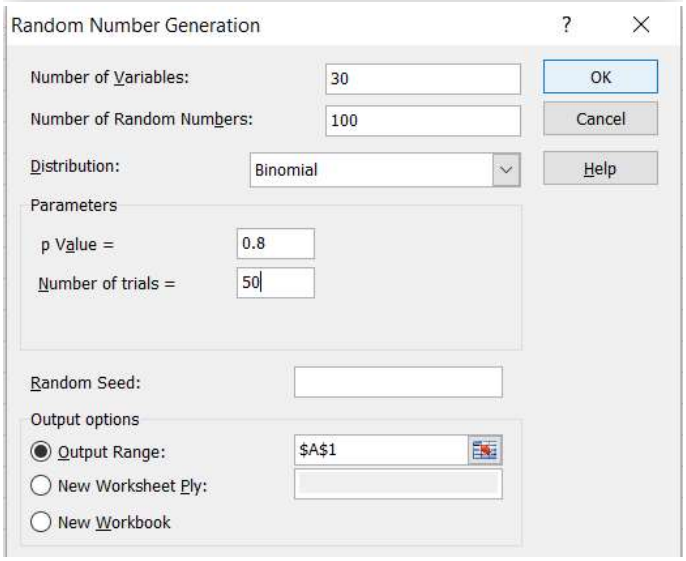
11 In Step 12, a histogram was drawn using the values of $z = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$. What do you notice about the shape of this distribution?

Repeat Task One (include answering the eleven Task One Questions) using different Uniform Distributions, different sample sizes ensuring $n \geq 30$, and different numbers of samples. In Step 6, arrange the sample means into appropriate intervals and then draw the histogram.

Task Two:

Step 1 Using and Excel workbook, generate 100 samples each of size 30 from the binomial distribution $X \sim B(50, 0.8)$.





Step 2 For each of the 100 samples, calculate the sample mean and the sample standard deviation.

Step 3 Calculate the mean of the sample means, $\overline{X}$, and the standard deviation of the sample means, $\sigma_{\overline{x}}$. Record the sample size, the number of samples, type of distribution, the mean of the sample means and the standard deviation of the sample means.

Step 4 For each sample, construct the 95% confidence interval for the population mean μ using $\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$, where $\overline{x}$ is the sample mean, σ is the population standard deviation and n is the sample size.

Step 5 For each sample, test whether or not the population mean lies within the 95% confidence interval $\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$. Determine how many of the 95% confident intervals contain the population mean.

Step 6 Arrange the sample means into appropriate intervals and draw a histogram.

Step 7 For each sample, calculate $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$. Calculate the mean and the standard deviation of the z values.

Step 8 Arrange the values $z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ into intervals -3.5 to -3, -3 to -2.5, -2.5 to -2, …, 2.5 to 3, 3 to 3.5 and draw a histogram.

It is often the case that the population standard deviation σ is not known. Provided that the sample size is large, the sample standard deviation s may be used as an unbiased estimate for σ. Steps 4, 5, 7 and 8 will be repeated using s instead of σ.

Step 9: For each sample, construct the 95% confidence interval for the population mean μ using $\left(\bar{x} - 1.96\dfrac{s}{\sqrt{n}}, \bar{x} + 1.96\dfrac{s}{\sqrt{n}} \right)$, where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the sample size.

Step 10: For each sample, test whether or not the population mean lies within the 95% confidence interval $\left(\bar{x} - 1.96\dfrac{s}{\sqrt{n}}, \bar{x} + 1.96\dfrac{s}{\sqrt{n}} \right)$. Determine how many of the 95% confident intervals contain the population mean.

Step 11: For each sample, calculate $z = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$. Calculate the mean and the standard deviation of the z values.

Step 12: Arrange the values $z = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ that are currently in AR1 to AR100 into intervals -3.5 to -3, -3 to -2.5, -2.5 to -2, …, 2.5 to 3, 3 to 3.5 and draw a histogram.

Task Two Questions

1 In Step 2, the mean $\bar{x}$ of each of the 100 samples of size $n = 30$ was calculated. Do each of your samples have the same sample mean or does this value vary between the samples?

2 (a) Calculate the population mean μ for $X \sim B(50, 0.8)$.

(b) In Step 3, the mean of the 100 sample means was calculated. Compare your answer with the value of μ.

3 In Step 2, the standard deviation s of each of the 100 samples of size $n = 30$ was calculated. Do each of your samples have the same sample standard deviation or does this value vary between the samples?

4 (a) Calculate the population standard deviation σ for $X \sim B(50, 0.8)$.

(b) In Step 3, the standard deviation of the 100 sample means was calculated. Why is the standard deviation of the 100 sample means less than the population standard deviation σ?

5 In Step 4, 95% confidence intervals for μ were constructed using whilst in Step 5, each confidence interval was tested to determine whether or not μ was within the interval. What percentage of the confident intervals contained μ?

6 In Step 6, a histogram was drawn using the values of the sample means. What do you notice about the shape of this distribution?

7 In Step 7, the value of $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ was calculated for each sample and the mean and standard deviation of the values of z were calculated. What do you notice about the mean and standard deviation of z?

8 In Step 8, a histogram was drawn using the values of $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$. What do you notice about the shape of this distribution?

9 In Step 9, 95% confidence intervals for μ were constructed using $\left(\bar{x} - 1.96\dfrac{s}{\sqrt{n}}, \bar{x} + 1.96\dfrac{s}{\sqrt{n}} \right)$ whilst in Step 10 each confidence interval was tested to determine whether or not μ was within the interval. What percentage of the confidence intervals contained μ?

10 In Step 11, the value of $z = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ was calculated for each sample and the mean and standard deviation of the values of z were calculated. What do you notice about the mean and standard deviation of z?

11 In Step 12, a histogram was drawn using the values of $z = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$. What do you notice about the shape of this distribution?

Repeat Task Two (include answering the eleven Task Two Questions) using different Binomial Distributions, different sample sizes ensuring $n \geq 30$, and different numbers of samples. In Step 6, arrange the sample means into appropriate intervals and then draw the histogram.

Task Three Questions:

In Tasks One and Two, you recorded the sample size, the number of samples, type of distribution, the mean of the sample means and the standard deviation of the sample means.

1. What do you notice about the mean of the sample means, $\bar{X}$, and the mean of the population, μ?
2. What do you notice about the standard deviation of the sample means, $\sigma_{\bar{X}}$, and the standard deviation of the population, σ?
3. Complete the following:
   "*When random samples are taken from a non-normal population with mean* μ *and known variance* σ^2*, the distribution of* $\bar{X}$ *is* _____ *and* $\bar{X} \sim$ _____ *provided that* _____."
4. Complete the following:
   "*When random samples are taken from a non-normal population with mean* μ *and unknown variance, the distribution of* $\bar{X}$ *is* _____ *and* $\bar{X} \sim$ _____ *provided that* _____."