CHAPTER

9 INTERVAL ESTIMATES FOR PROPORTIONS

SYLLABUS COVERAGE

Syllabus coverage

TOPIC 4.3: INTERVAL ESTIMATES FOR PROPORTIONS

Random sampling

4.3.1 examine the concept of a random sample

4.3.2 discuss sources of bias in samples, and procedures to ensure randomness

4.3.3 use graphical displays of simulated data to investigate the variability of random samples from various types of distributions, including uniform, normal and Bernoulli

Sample proportions

4.3.4 examine the concept of the sample proportion $\hat{p}$ as a random variable whose value varies between samples, and the formulas for the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of the sample proportion $\hat{p}$

4.3.5 examine the approximate normality of the distribution of $\hat{p}$ for large samples

4.3.6 simulate repeated random sampling, for a variety of values of p and a range of sample sizes, to illustrate the distribution of $\hat{p}$ and the approximate standard normality of $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ where the closeness of the approximation depends on both n and p

Confidence intervals for proportions

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

4.3.8 use the approximate confidence interval $\left(\hat{p}-z\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}, \hat{p}+z\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}\right)$ as an interval estimate for p, where z is the appropriate quantile for the standard normal distribution

4.3.9 define the approximate margin of error $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and understand the trade-off between margin of error and level of confidence

4.3.10 use simulation to illustrate variations in confidence intervals between samples and to show that most, but not all, confidence intervals contain p

Video playlists (4):

9.1 Random sampling

9.2 The sampling distribution of sample proportions

9.3 Confidence intervals for proportions

WACE question analysis Interval estimates for proportions

Worksheets (6):

9.2 Sample proportions • Sample proportion calculations • Sample proportion probabilities

9.3 Sample proportion confidence intervals • Margin of error for standard normal variables • Sample sizes

Nelson MindTap

To access resources above, visit **cengage.com.au/nelsonmindtap**

Video playlist
Random sampling

9.1 Random sampling

The language of random sampling

Suppose the administration of a local high school containing 1500 students wanted to collect data from every single student around a particular statistical variable, X. In some cases, this level of large-scale data collection of an entire **population** may be impractical. To do so, a **census** is needed. When a census is conducted to gain data on an entire population, characteristics of that particular population can be calculated, such as the mean or standard deviation of X. These are called **population parameters**.

However, perhaps to be more practical or time efficient, the school administration may consider only collecting data from a **sample** of 100 students from that school around that particular statistical variable. This is then called a **survey** of a sample group of **sample size** 100. From this sample data, **sample statistics** can be calculated, such as the mean and standard deviation, and then used as **point estimates** of the population mean and standard deviation.

In most situations involving data collection, we cannot take a census of a whole population and so population parameters are often unknown. As a result, we must rely on the use of sample statistics obtained from surveys to make inferences about the larger population from which the data was obtained.

WORKED EXAMPLE 1 Identifying population, parameters, sample size and statistics

The first 20 people leaving a convenience store after 6:30 pm were asked how much they spent. The smallest amount was \$5.40, the largest was \$47.60 and the average amount was \$22.40. Identify the

a population **b** sample size **c** sample statistics.

Steps	Working
a Determine the population (the whole group being investigated).	The population is all customers of that particular convenience store.
b Identify the number of people from whom data was collected.	The sample size is 20.
c Identify any characteristics from this information.	Some possible statistics of this sample are: • the minimum amount spent – \$5.40 • the maximum amount spent – \$47.60 • the range – \$42.20 • the mean amount spent – \$22.40.

When data are collected from sample groups, it must be considered as to whether that data gives useful, accurate and reliable information about the population parameters. Things that need to be considered include

- sample size – is it sufficiently large enough to represent the population?
- randomness and bias – are the data free from biases that could affect the reliability of the data being used to estimate population parameters?

When a sample meets these conditions, it can be called a **fair and representative sample** of the population. For example, the sample data collected in Worked example 1 may not be considered a fair and representative sample and, hence, could be considered a **biased sample** as:

- it is only a small sample of 20 customers of all customers of that store
- the data was only collected once, at a particular time of day (after 6:30 pm) on one particular day.

As a result, when collecting data from samples, procedures to ensure randomness and minimise any sources of **bias** need to be considered so that valid conclusions can be made about the population of interest.

Random sampling procedures and sources of bias

9.1

The best way to reduce sampling bias and obtain fair and representative samples is to use a **probability sampling method**. This is a method that involves random selection in which each member of the population or subset of a population has an equally likely chance of being chosen. Commonly used probability sampling methods are listed as follows.

1. **Simple random sampling** – every member of the entire population has an equally likely chance of being selected; for example, through the use of a random number generator to generate n numbers and then those numbers are used to select the sample group.
2. **Systematic sampling** – the population is ordered on the basis of an unbiased characteristic (e.g. alphabetical order based on first name or last name, age, height) and every kth member of the population is chosen to create the sample of size n.
3. **Stratified sampling** – the population is divided into subgroups, called strata, on the basis of common characteristics (e.g. year groups, profession) and then simple random or systematic selections are made from each subgroup proportional to the size of the subgroup.
4. **Cluster sampling** – the population is divided into subgroups, called clusters, on the basis of common characteristics (e.g. location, time) and then simple random or systematic selections are made from each subgroup, but not necessarily in proportion.

Exam hack

If you are asked to describe a method to ensure randomness when sampling, it is not the name of the sampling method that is the important feature but rather the description of the random process.

WORKED EXAMPLE 2 Describing a random sampling process

A company needs to select 100 houses in a particular street of 421 houses to collect data for an employment survey. Describe a sampling procedure that would ensure randomness.

Steps	Working
1 Choose a probability sampling method that ensures each house has an equally likely chance of being selected.	Consider a systematic sampling method. Divide 421 by 100 and round to the nearest integer. $\frac{421}{100} = 4.21 \approx 4$
2 Describe the process.	Use a random number generator to pick a starting house number between 1 and 421; for example, 27. Collect data from every 4th house number, starting from 27 (i.e. 27, 31, 35, 39 … 419) and start the sequence again from 2 until 100 households are surveyed.

Sampling procedures that are not random are called **non-probability sampling methods**, in which each member of the population does not have an equally likely chance of being selected. Commonly used non-probability sampling methods are listed as follows.

1. **Convenience sampling** – the sample is chosen such that the data is conveniently collected from the most accessible data source; for example, surveying family members in the same household about political views.
2. **Quota sampling** – the sample size is pre-determined and no more data is collected once that limit has been reached; for example, the first 10 employees who arrive at work are surveyed about transport methods.
3. **Volunteer sampling** – the sample is collected by asking for volunteers to opt-in to the data collection. An example is a radio show asking for callers to share information on the number of Australian states they have visited.

It is often in these cases that different types of bias will arise and can lead to an under- or over-representation of particular subgroups of a population. Common types of bias are given below.

1 Spatial bias – the sample is non-representative of the population because of the location from which it is collected.

2 Temporal bias – the sample is non-representative of the population because of the time at which it is collected.

3 Self-selection bias – the sample is non-representative of the population because the participants choose to take part voluntarily.

4 Non-response bias – the participants chosen to participate may choose to not give a response.

5 Leading question bias – the data may be collected in a way that encourages a particular response.

Exam hack

Once again, it is not the name of the bias that is the important feature but rather the description of the source of the bias; that is, where has the bias come from and what effect does it have on the data collection process.

WORKED EXAMPLE 3 Discussing sources of bias and ensuring randomness

A legal firm with 2000 employees wants to collect data on employment satisfaction and sends out an all-staff email at 9:00 pm on a Friday night containing a link to an optional survey.

a Identify and explain **two** possible sources of bias with this sampling method.

b Suggest a random sampling procedure that will minimise bias in this data collection.

Steps	Working
a **1** Consider the four main types of bias: spatial, temporal, self-selection, leading question. Identify which of the four are applicable to this situation.	Temporal bias – sending out the survey at 9:00 pm on a Friday night means that there is an increased chance that only employees who are working on the Friday night and over the weekend respond to the survey.
2 Identify the source of bias in the sampling method and give a brief explanation as to why it creates bias.	Self-selection bias – only those who want to respond to the survey will, meaning that there could be an over-representation of those who either are or are not satisfied with their job.
b **1** Choose a probability sampling method that ensures each employee has an equally likely chance of being selected. **2** Describe the process.	List the employees alphabetically by surname and send the email during working hours to every 10th employee in the ordered list to complete a compulsory survey.

Variability of random samples

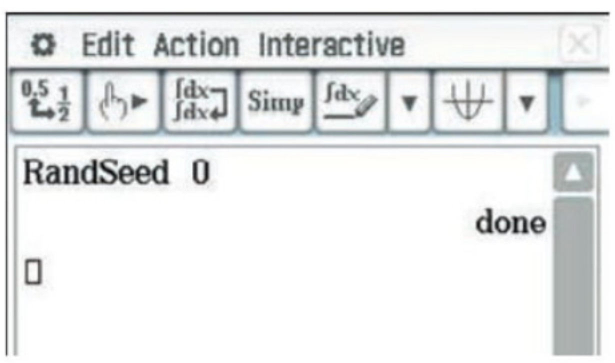
Suppose the legal firm in Worked example 3 randomly surveyed three different sample groups of employees, each of sample size 200. Each of these samples would have produced different sample statistics due to the randomness in the sampling process. This is referred to as the **variability of random samples**. That is, even though all 600 employees came from the same population of 2000, each set of sample data has the potential to provide very different representations of the population.

The variability of samples can be seen through a process of **simulation**, whereby technology can be used to model the events of a random probability experiment or mimic the data collection process of a sample from a population. However, when simulating sample data, assumptions need to be made about the nature of the **population distribution** (also called the **parent distribution**) from which the samples are being taken. Based on our knowledge of random variables, we can simulate sample data from populations that are uniformly, normally, Bernoulli or binomially distributed.

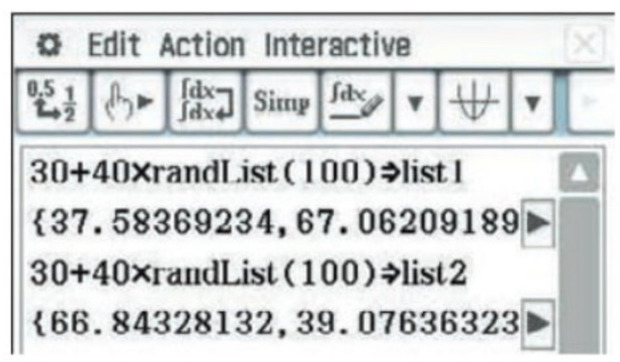
USING CAS 1 Simulating sample data from a uniform distribution

Simulate two different samples of 100 scores from the continuous uniform random variable, $X \sim U[30, 70]$. Compare the mean and standard deviation of the samples to the mean and standard deviation of X.

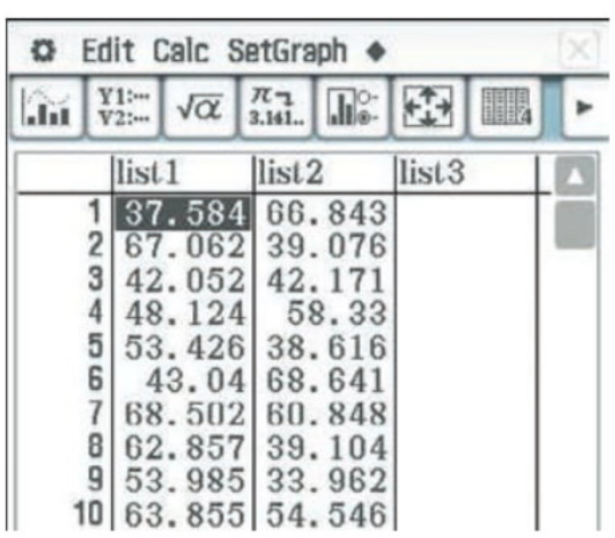
ClassPad



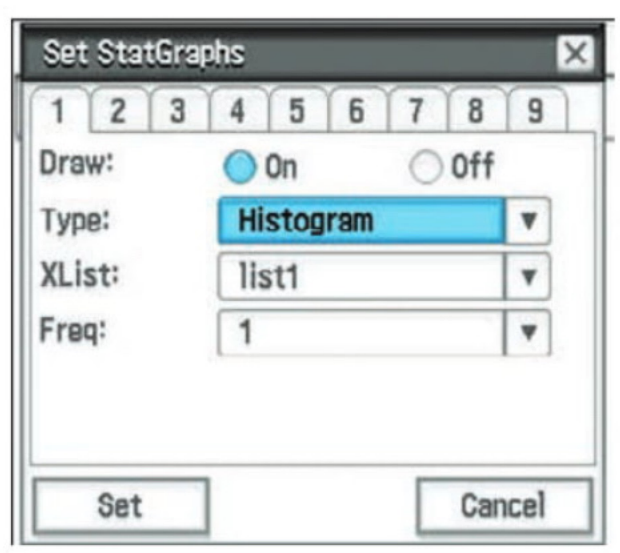
1 Tap **Decimal**.
2 Open the **Keyboard** > **Catalog**.
3 Tap **R** then scroll down to select **RandSeed**.
4 Enter a number from 0 to 9. The default is **0**.
5 Press **EXE**. This sets a new starting point for generating random numbers.



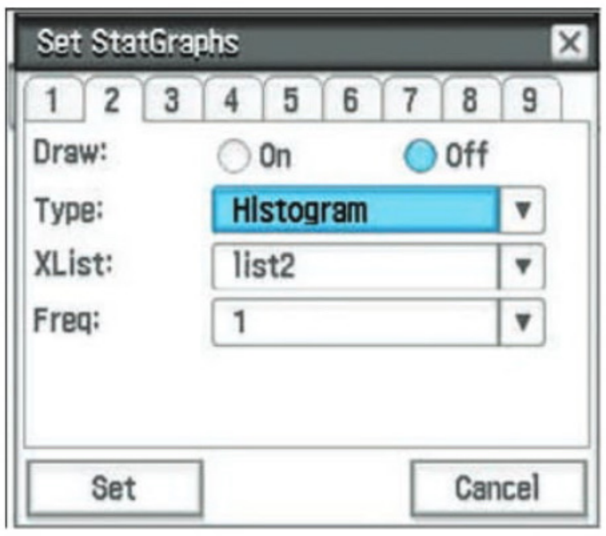
6 Open the **Keyboard** > **Catalog**.
7 Tap **R** then scroll to select **randList**.
8 Use the formula $a + (b - a) \times$ **randList**(m) to generate m values from the distribution $U[a, b]$, as shown above.
9 Store the first sample as **list1**.
10 Repeat for the second sample and store as **list2**.



11 Tap **Menu** > **Statistics**.
12 The randomly generated values will appear in **list1** and **list2**.
13 Tap **SetGraph** > **Setting**.



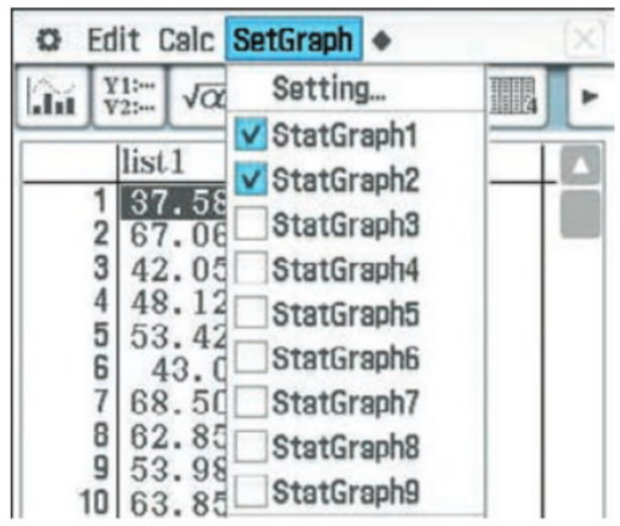
14 Tap on the **1** tab.
15 Change the **Type:** field to **Histogram** and keep the **XList** field as **list1**.



16 Tap on the **2** tab.

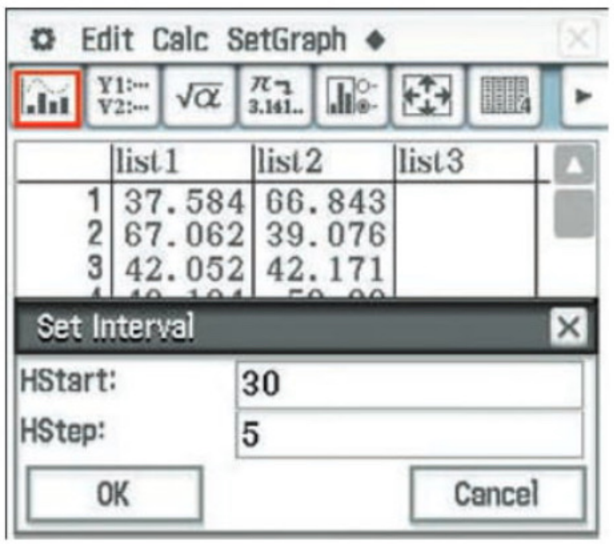
17 Change the **Type:** field to **Histogram** and change the **XList** field to **list2**.

18 Tap **Set**.



19 Tap **SetGraph**.

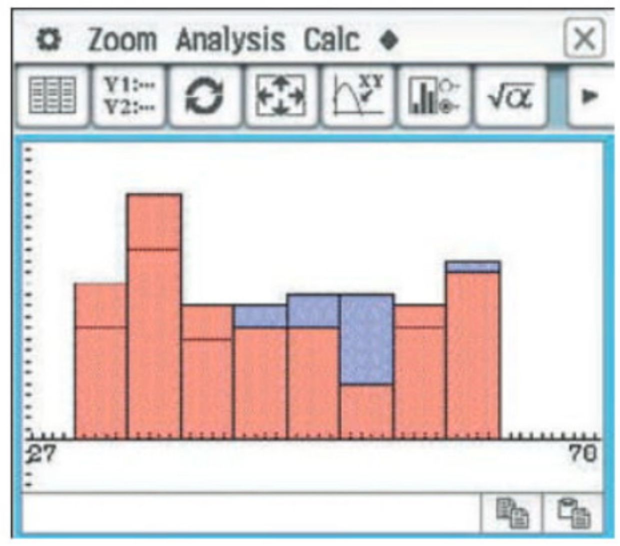
20 Tap to select **StatGraph1** and **StatGraph2**.



21 Tap **Graph**.

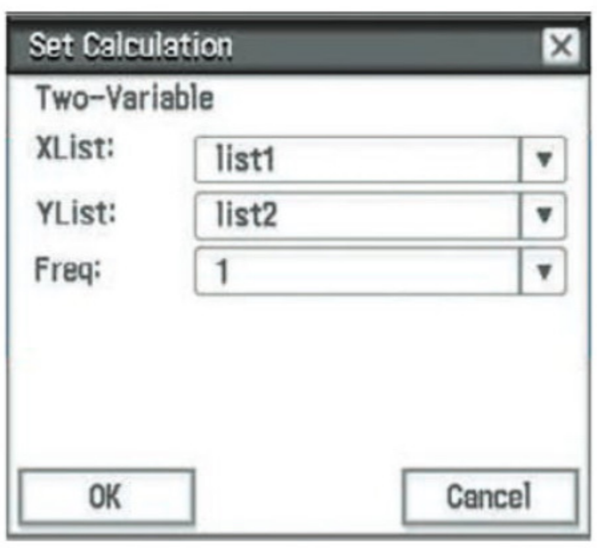
22 Keep the **HStart:** field as **30** and change the **HStep:** field to **5** (optional).

23 Tap **OK**.



24 The histograms of the data will appear in the lower window (the windows have been swapped for this screen).

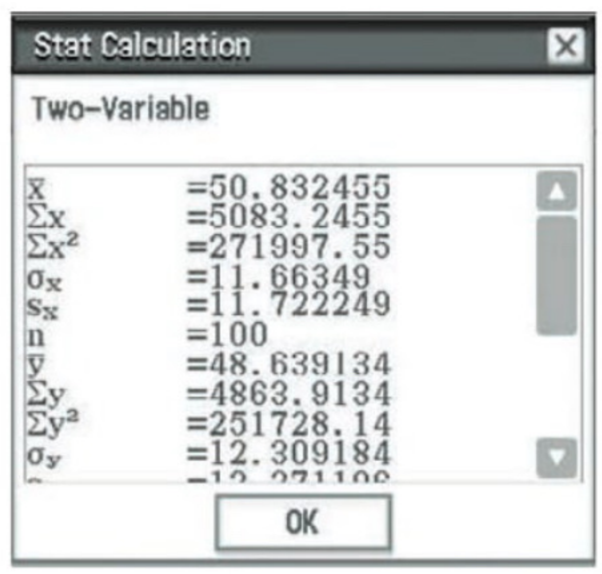
25 Tap the upper window to select **Statistics**.



26 Tap **Calc** > **Two-Variable**.

27 Keep the **XList:** field as **list1** and change the **YList:** field to **list2**.

28 Tap **OK**.



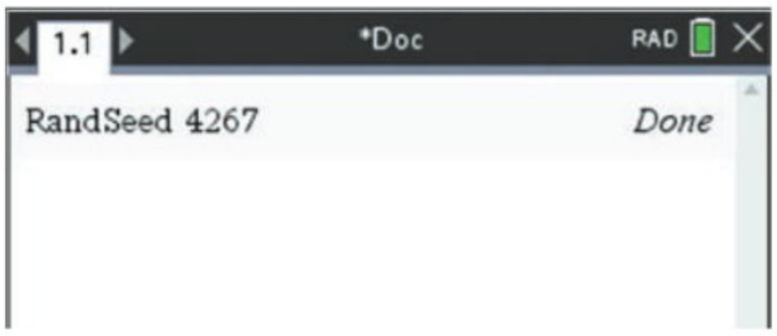
29 The summary statistics will appear in the **Two-Variable** window.

30 Scroll down to view all the statistics.

$\mathrm{E}(X) = \dfrac{30 + 70}{2} = 50$. Both simulated means are approximately 50, but vary.

$\mathrm{SD}(X) = \dfrac{70 - 30}{\sqrt{12}} = 11.55$. Both simulated standard deviations are close to 11, but vary.

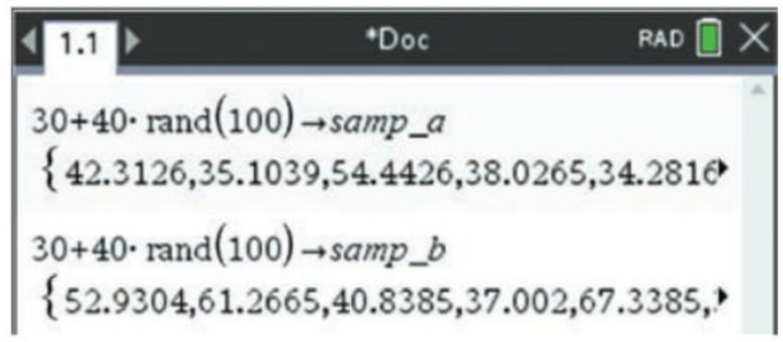
TI-Nspire



1 Press **catalog** > **R** to jump to the functions starting with the letter R.

2 Scroll down and select **RandSeed**.

3 Enter a 4-digit number and press **enter**.
This sets a new starting point for generating random numbers.

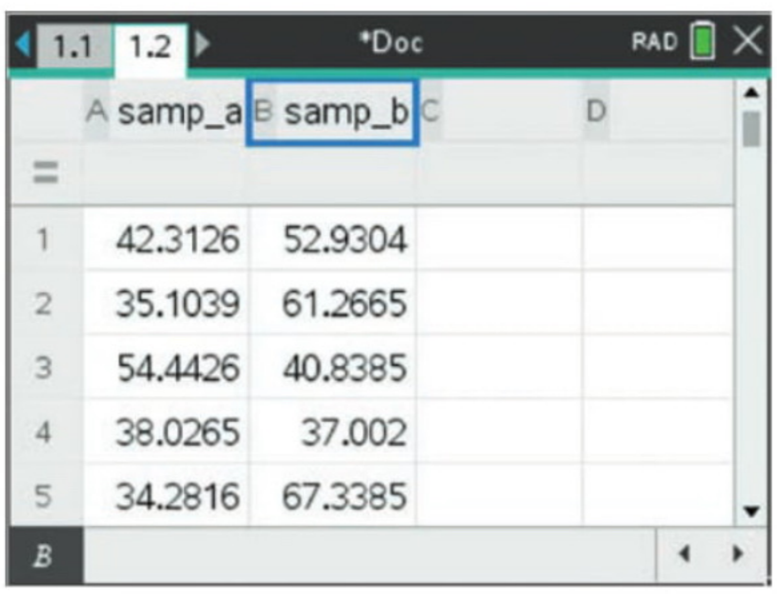


4 Press **catalog** > **rand**.

5 Use the formula $a + (b - a) \times \text{randList}(m)$ to generate m values from the distribution $U[a, b]$, as shown above.

6 Store the first sample as **samp_a**.

7 Repeat for the second sample and store as **samp_b**.



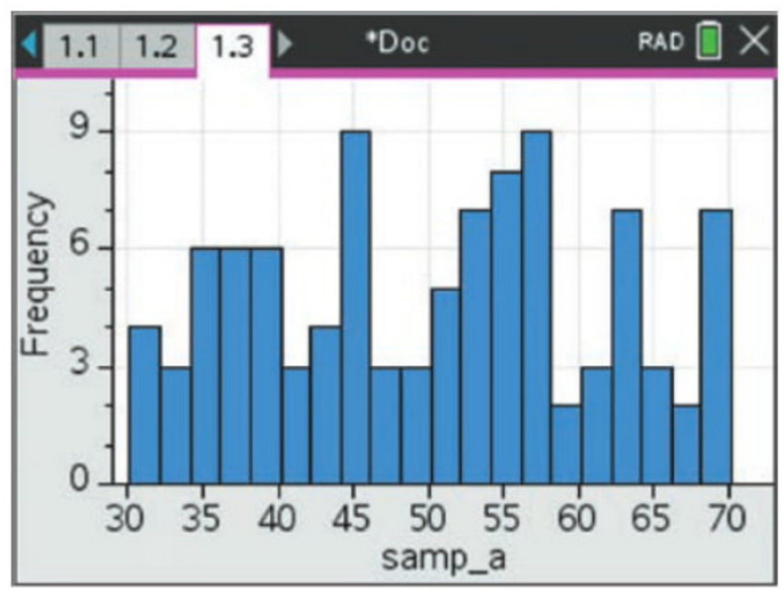
8 Add a **Lists & Spreadsheet** page.

9 Tap in the cell next to the **A**.

10 Press **var** > **Link To:** and select **samp_a**.

11 Tap in the cell next to the **B**.

12 Press **var** > **Link To:** and select **samp_b**.

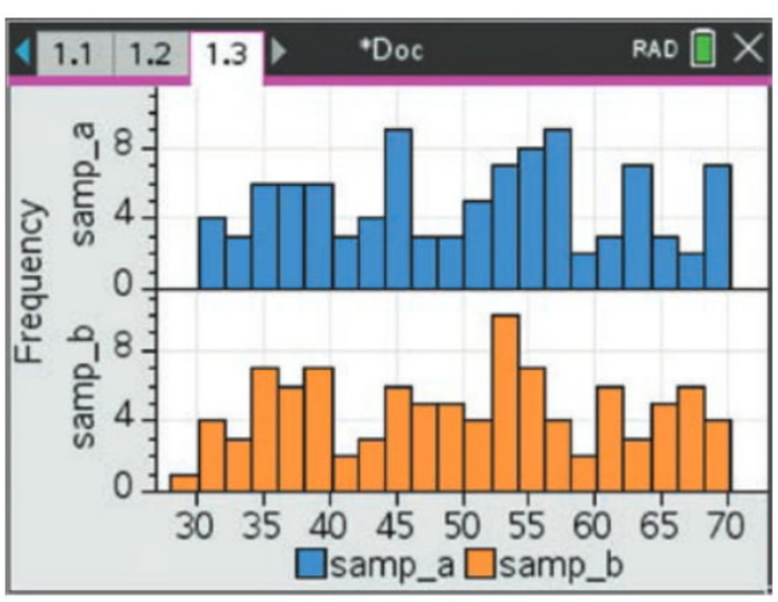


13 Add a **Data & Statistics** page.

14 For the horizontal axis, select **samp_a**.

15 Press **menu** > **Plot Type** > **Histogram**.

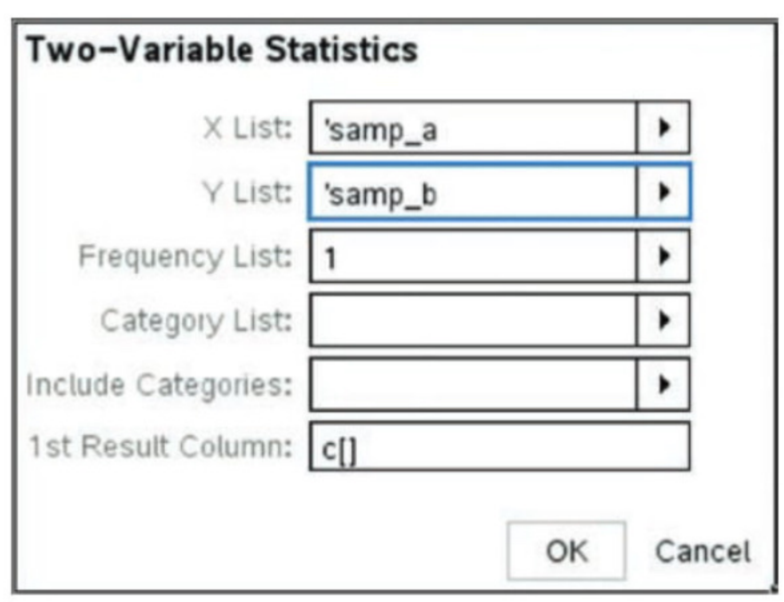
16 The sample a data will be displayed as a histogram.



17 Tap **menu** > **Plot Properties** > **Add X Variable**.

18 Select **samp_b**.

19 The data for both samples will be displayed as histograms.



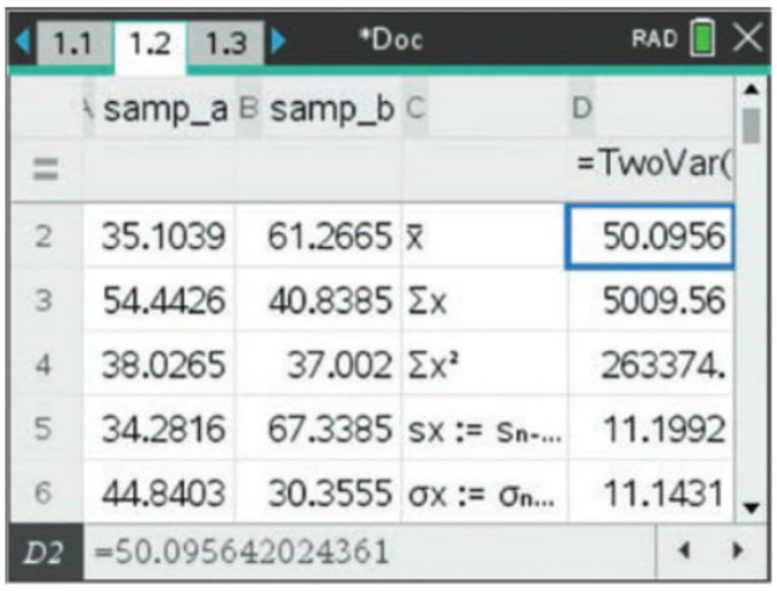
20 Return to the **Lists & Spreadsheet** page.

21 Press **menu** > **Statistics** > **Stat Calculations** > **Two-Variable**.

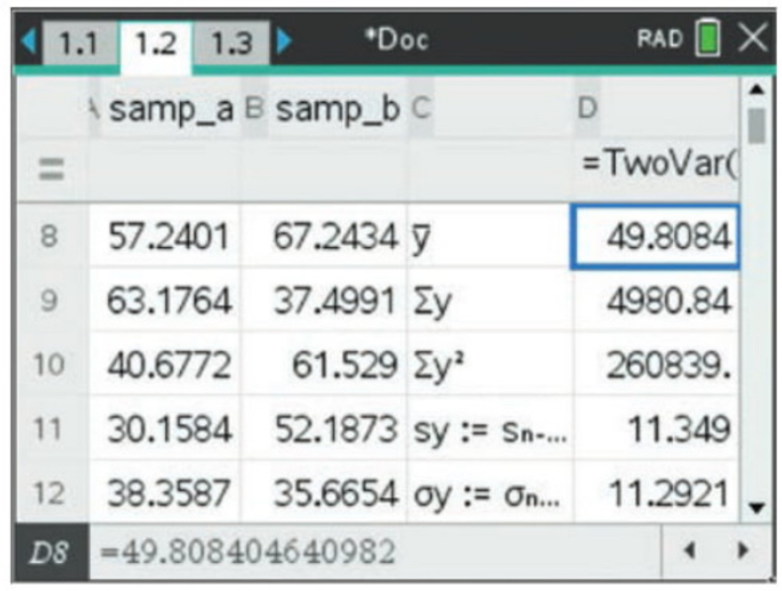
22 In the **X List:** field, press the right arrow and select **samp_a**.

23 In the **Y List:** field, press the right arrow and select **samp_b**.

24 Press **enter**.



25 The summary statistics will be displayed in columns **C** and **D**.

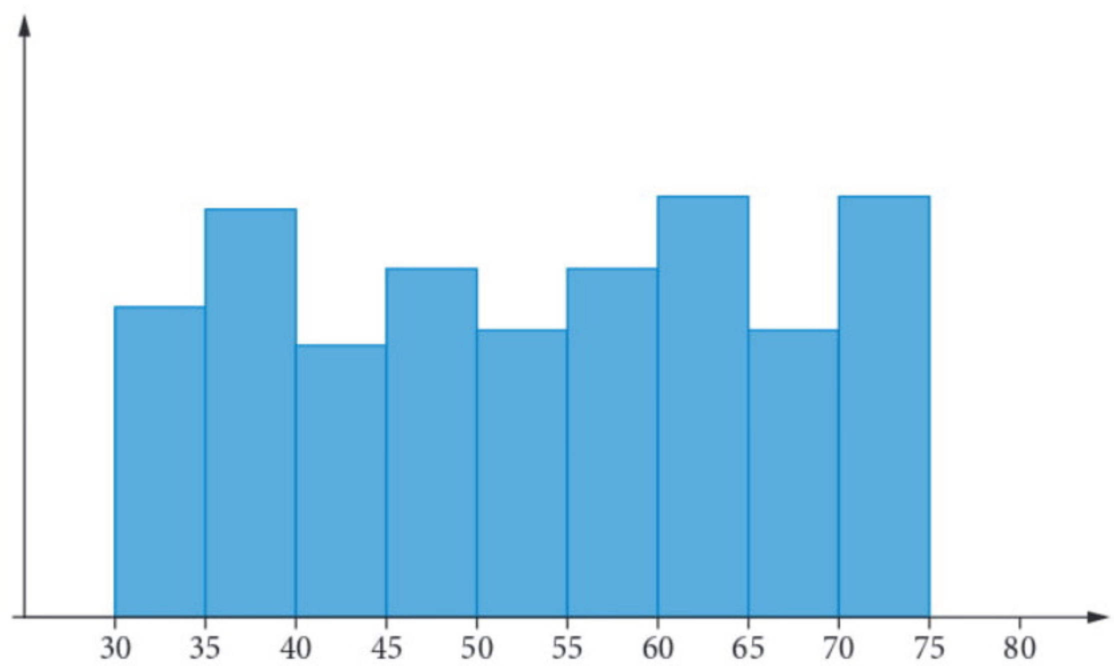


26 Scroll down to view all the statistics.

$E(X) = \dfrac{30 + 70}{2} = 50$. Both simulated means are approximately 50, but vary.

$SD(X) = \dfrac{70 - 30}{\sqrt{12}} = 11.55$. Both simulated standard deviations are close to 11, but vary.

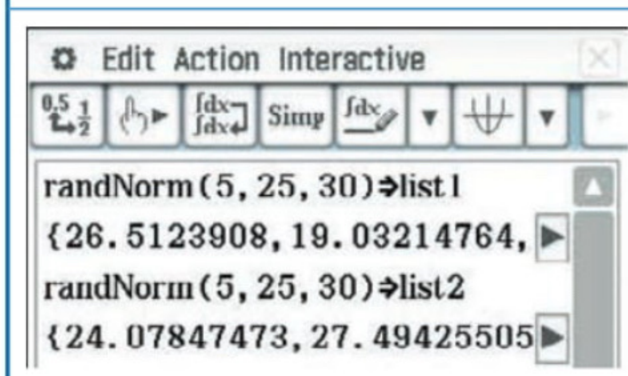
Note that if we were to repeat the above simulation with a greater number of scores, we would *expect* the distribution to become *more uniform*; however, due to the nature of random sampling, we could get a simulation that does not become more uniform. For example, the histogram on the right shows a simulation in which 250 scores were generated.

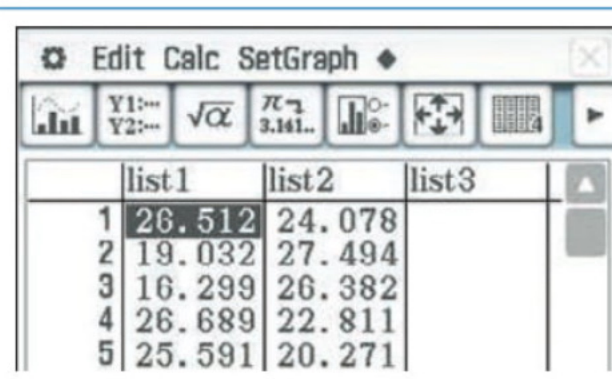


USING CAS 2 Simulating sample data from a normal distribution

Simulate two different samples of 30 scores from the continuous normal random variable, $Y \sim N(25, 5^2)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of Y.

ClassPad





1 Open the **Keyboard** > **Catalog**.
2 Tap **R** then scroll down to select **randNorm**.
3 Use the input **randNorm**(μ, σ, m) to generate m values from the distribution $N(\mu, \sigma^2)$, as shown above.
4 Store the first sample as **list1**.
5 Repeat for the second sample and store as **list2**.
6 Tap **Menu** > **Statistics**.
7 The randomly generated values will appear in **list1** and **list2**.
8 Tap **SetGraph** > **Setting**.

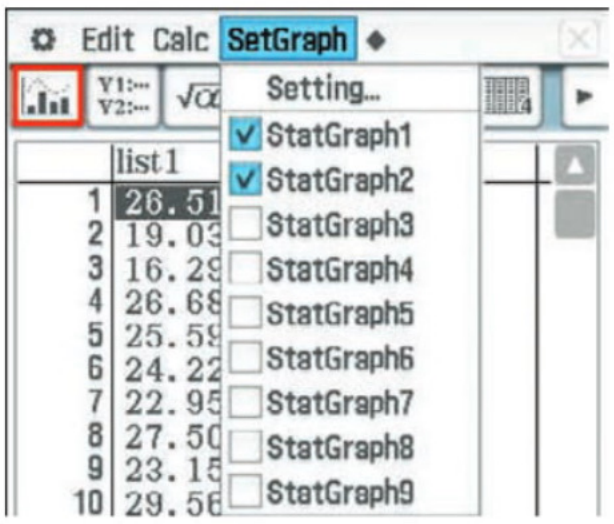
9 Tap tab **1**.

10 Ensure the **Type:** field is **Histogram** and the **XList:** field is **list1**.

11 Tap tab **2**.

12 Ensure the **Type:** field is **Histogram** and the **XList:** field is **list2**.

13 Tap **Set**.



14 Tap **SetGraph**.

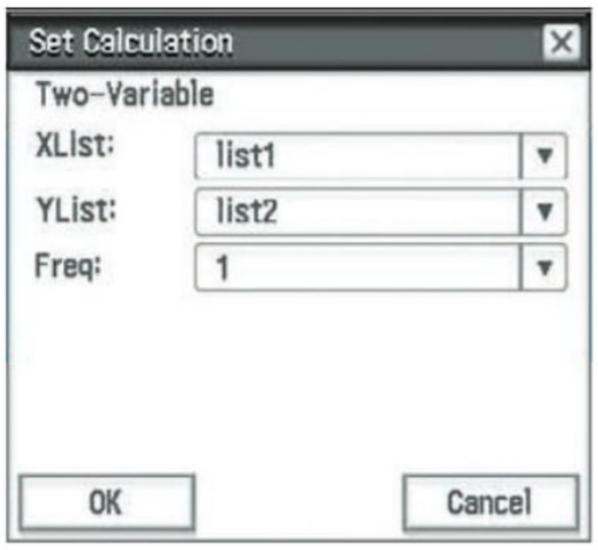
15 Ensure **StatGraph1** and **StatGraph2** are selected.

16 Tap **Graph**.

17 In the **Set Interval** dialogue box, keep the defaults settings and tap **OK**.

18 The histograms of the data will appear in the lower window (the windows have been swapped for this screen).

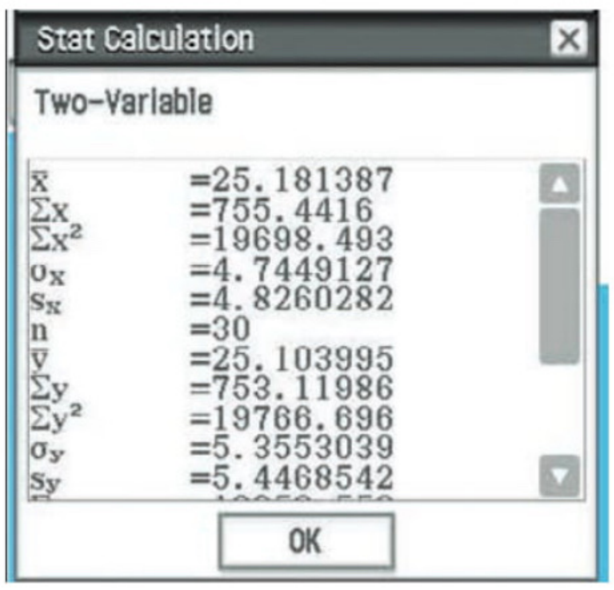
19 Tap the upper window to select **Statistics**.



20 Tap **Calc** > **Two-Variable**.

21 Ensure the **XList:** field is **list1** and the **YList:** field is **list2**.

22 Tap **OK**.



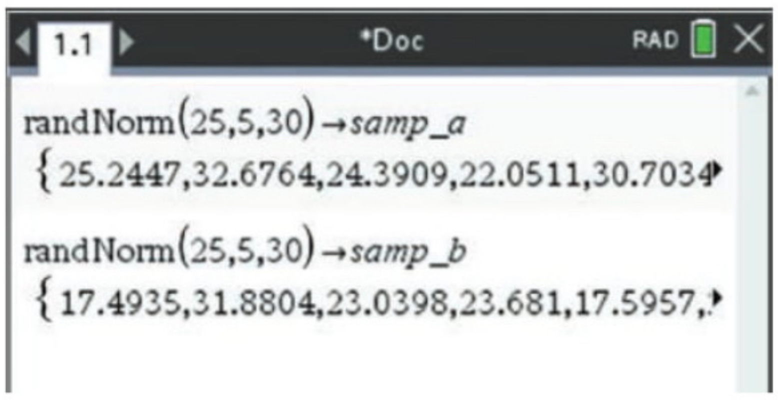
23 The summary statistics will appear in the **Two-Variable** window.

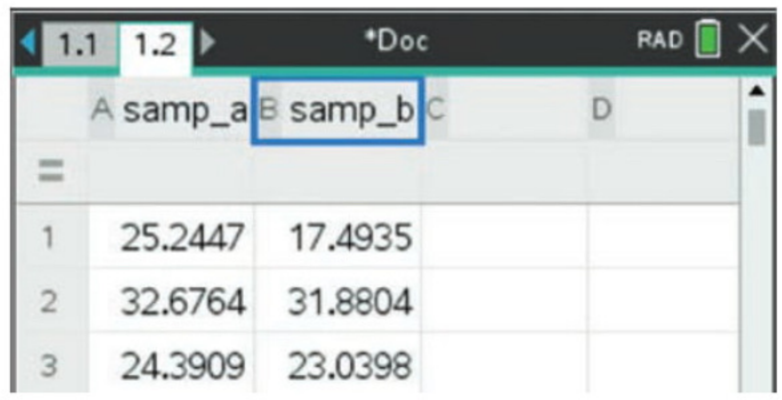
24 Scroll down to view all the statistics.

$E(Y) = 25$. Both simulated means are approximately 25, but vary.

$SD(Y) = 5$. Both simulated standard deviations are close to 5, but vary.

TI-Nspire





1 Press **catalog** > **randNorm**.

2 Use the input **randNorm**(μ, σ, m) to generate m values from the distribution $N(\mu, \sigma^2)$, as shown above.

3 Store the sample as **samp_a**.

4 Repeat for the second sample and store as **samp_b**.

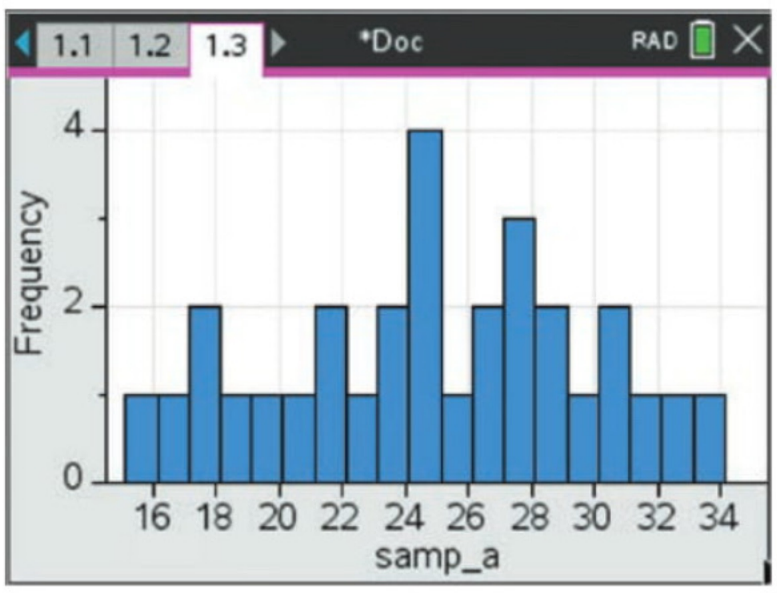
5 Add a **Lists & Spreadsheet** page.

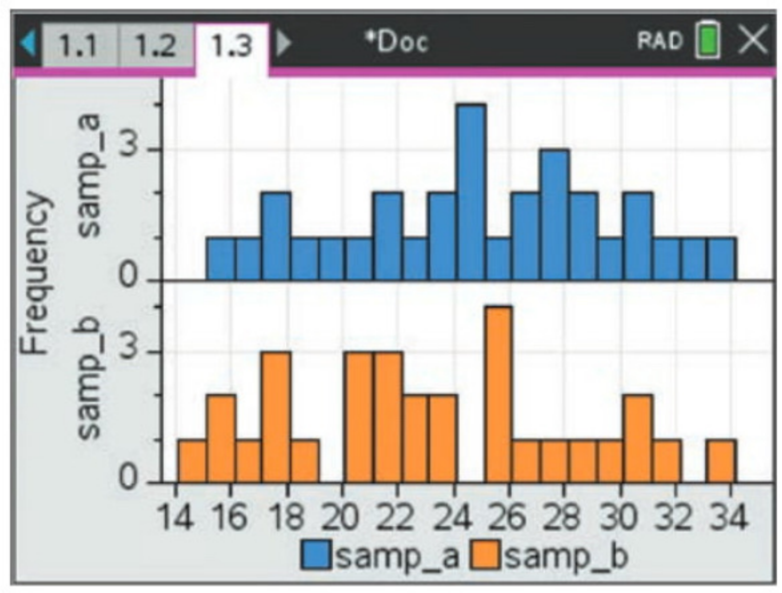
6 Tap in the cell next to the **A**.

7 Press **var** and select **samp_a**.

8 Tap in the cell next to the **B**.

9 Press **var** and select **samp_b**.





10 Add a **Data & Statistics** page.

11 For the horizontal axis, select **samp_a**.

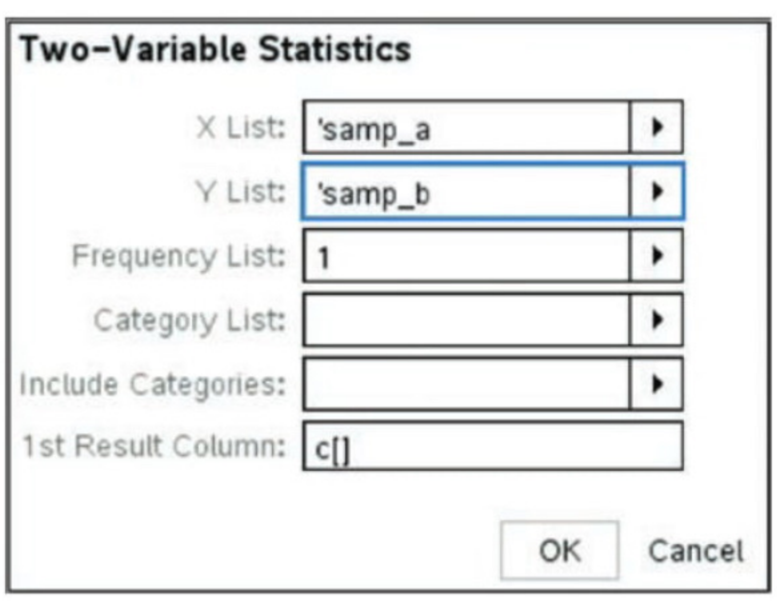
12 Press **menu** > **Plot Type** > **Histogram**.

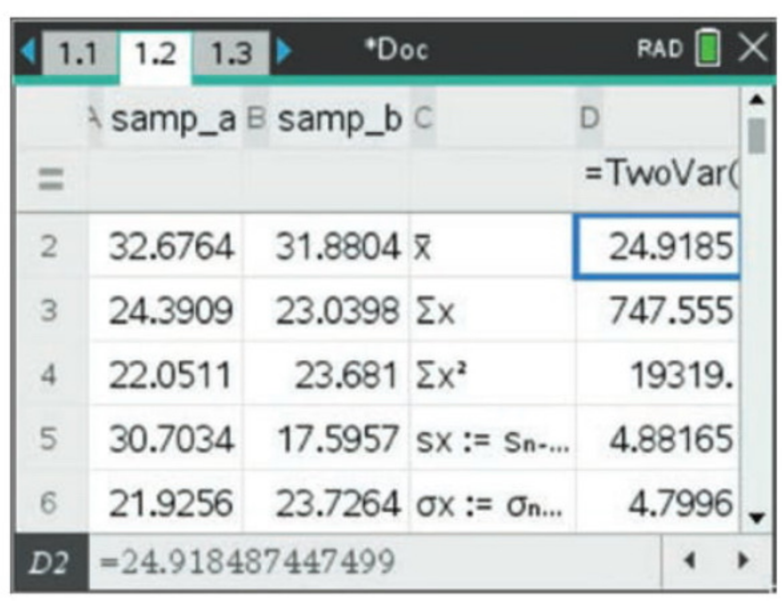
13 The sample a data will be displayed as a histogram.

14 Tap **menu** > **Plot Properties** > **Add X Variable**.

15 Select **samp_b**.

16 The data for both samples will be displayed as histograms.





17 Return to the **Lists & Spreadsheet** page.

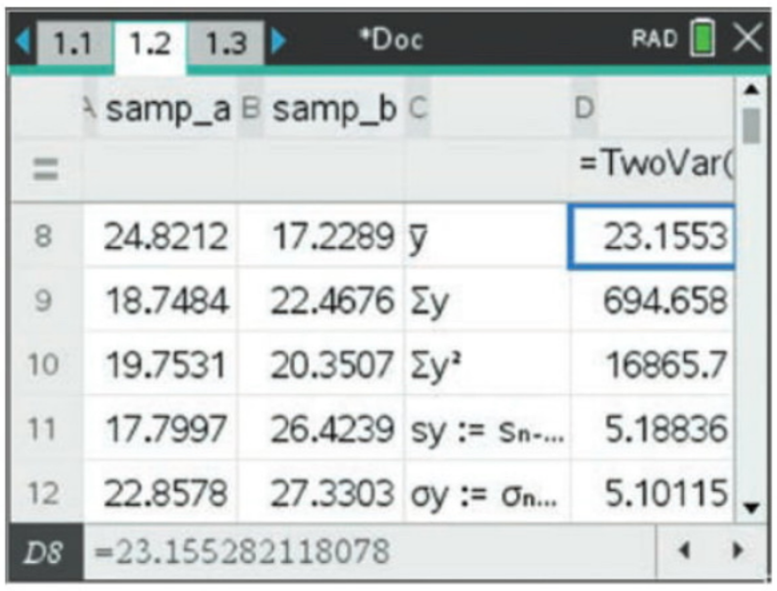
18 Press **menu** > **Statistics** > **Stat Calculations** > **Two-Variable**.

19 In the **X List:** field, press the right arrow and select **samp_a**.

20 In the **Y List:** field, press the right arrow and select **samp_b**.

21 Press **enter**.

22 The summary statistics will be displayed in columns **C** and **D**.

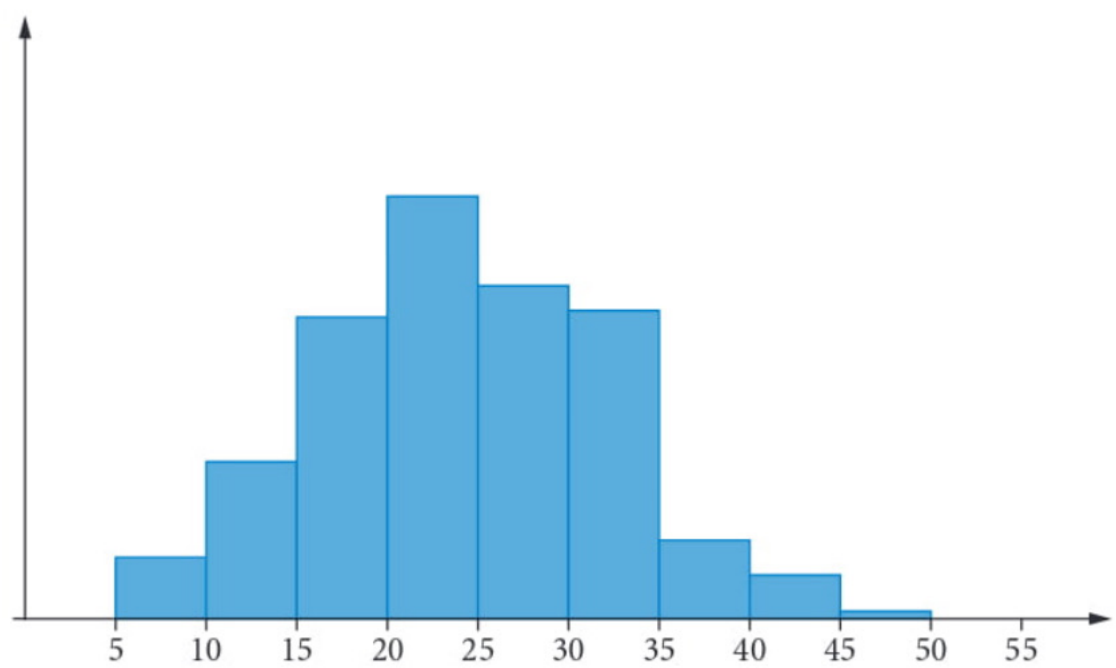


23 Scroll down to view all the statistics.

E(Y) = 25. Both simulated means are approximately 25, but vary.

SD(Y) = 5. Both simulated standard deviations are close to 5, but vary.

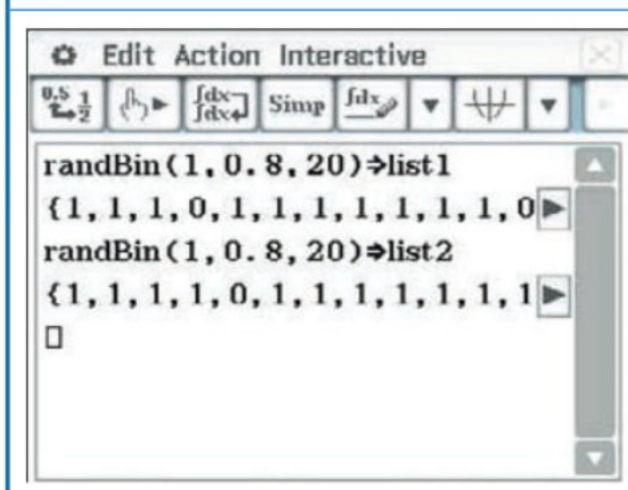
Similarly, if we were to repeat the above simulation with a greater number of scores, we would *expect* the distribution to become *more normal*; however, due to the nature of random sampling, we could get a simulation that does not become more normal. For example, the histogram on the right shows a simulation in which 250 scores were generated.

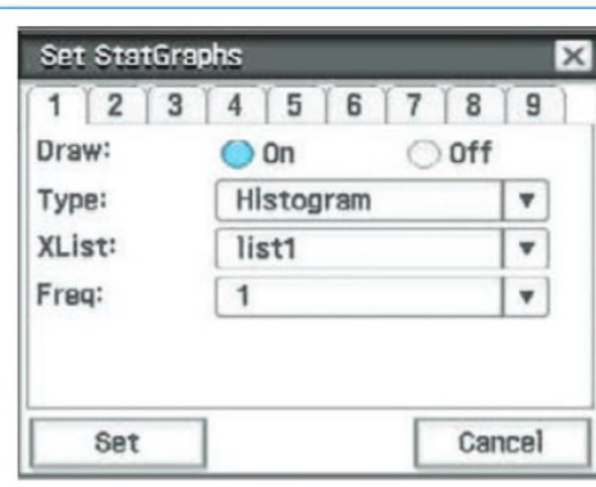


USING CAS 3 Simulating sample data from a Bernoulli distribution

Simulate two different samples of 20 scores from the discrete Bernoulli random variable, $Z \sim \text{Bern}(0.8)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of Z.

ClassPad





1 Open the **Keyboard** > **Catalog**.

2 Tap **R** then scroll down to select **randBin**.

3 Use the input **randBin(1, p, m)** to generate m values from the distribution Bin$(1, p)$, as shown above.

4 Store the first sample as **list1**.

5 Repeat for the second sample and store as **list2**.

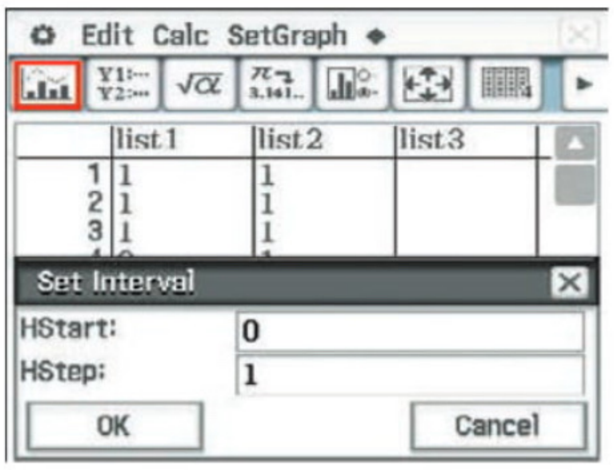
6 Tap **Menu** > **Statistics**.

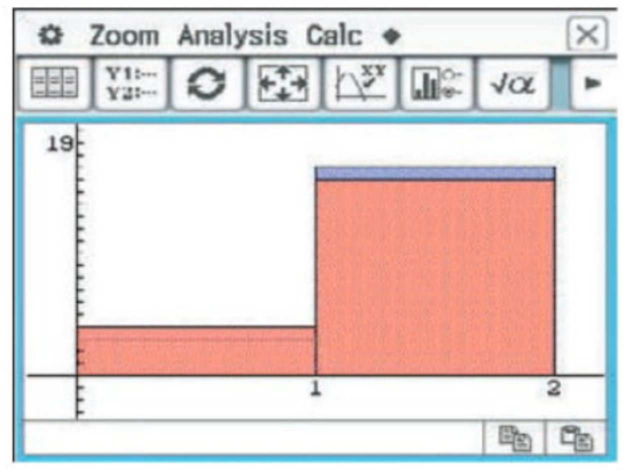
7 Tap **SetGraph** > **Setting**.

8 For tab **1**, ensure the **Type:** field is **Histogram** and the **XList**: field is **list1**.

9 For tab **2**, ensure the **Type**: field is **Histogram** and the **XList**: field is **list2**.

10 Tap **Set**.





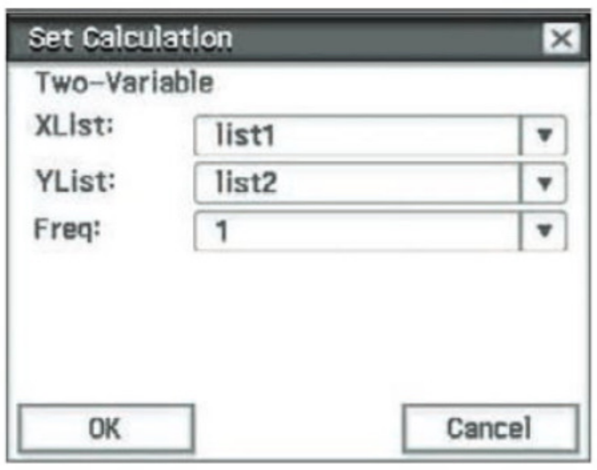
11 Tap **SetGraph** and ensure **StatGraph1** and **StatGraph2** are selected.

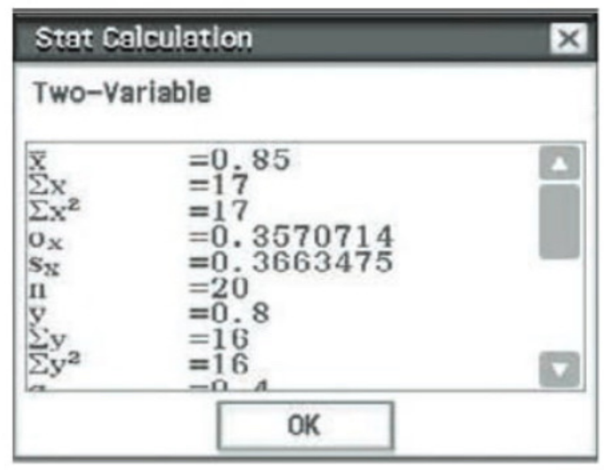
12 Tap **Graph**.

13 In the **Set Interval** dialogue box, change the **HStep**: field to **1** and tap **OK**.

14 The histograms of the data will appear in the lower window (the windows have been swapped for this screen).

15 Tap the upper window to select **Statistics**.





16 Tap **Calc** > **Two-Variable**.

17 Ensure the **XList**: field is **list1** and the **YList**: field is **list2**.

18 Tap **OK**.

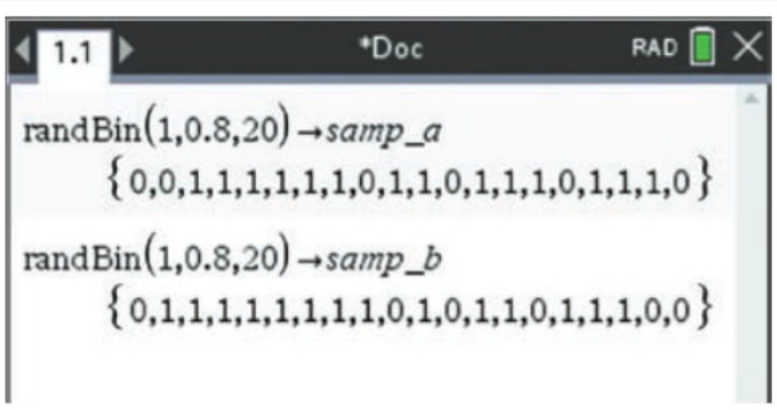
19 The summary statistics will appear in the **Two-Variable** window.

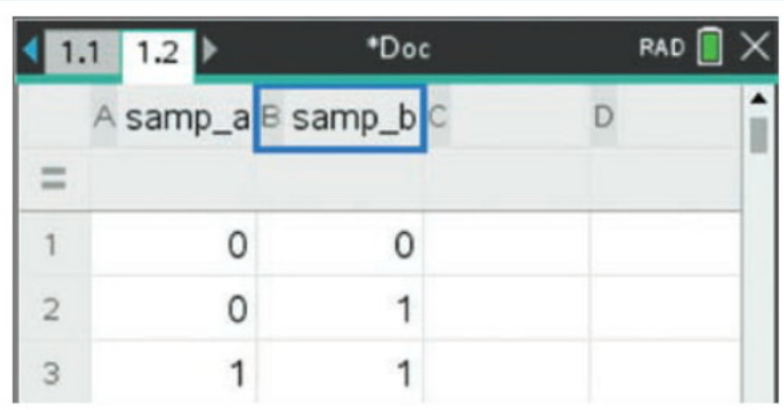
20 Scroll down to view all the statistics.

E(Z) = 0.8. Both simulated means are approximately 0.8, but vary.

SD(Z) = 0.4. Both simulated standard deviations are close to 0.4, but vary.

TI-Nspire





1 Press **catalog** > **randNorm**.

2 Use the input **randBin(1,** p**,** m**)** to generate m values from the distribution Bin(1, p), as shown above.

3 Store the first sample as **samp_a**.

4 Repeat for the second sample and store as **samp_b**.

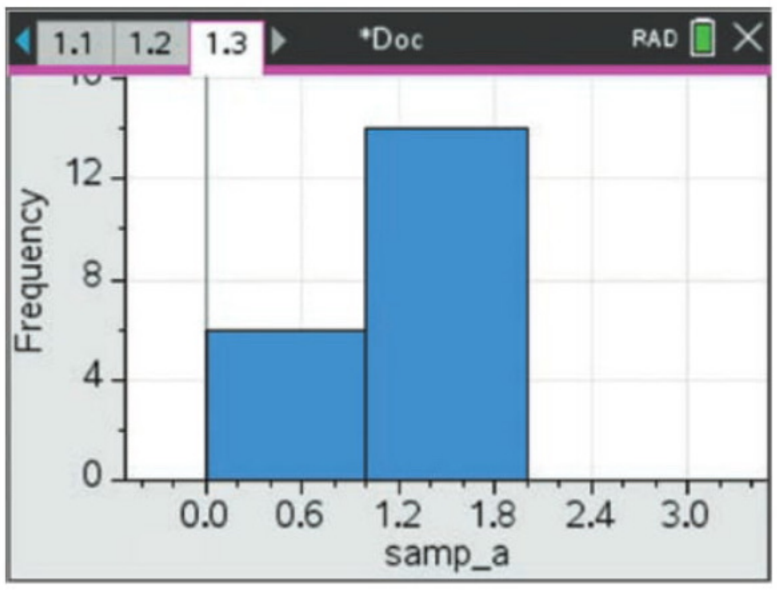
5 Add a **Lists & Spreadsheet** page.

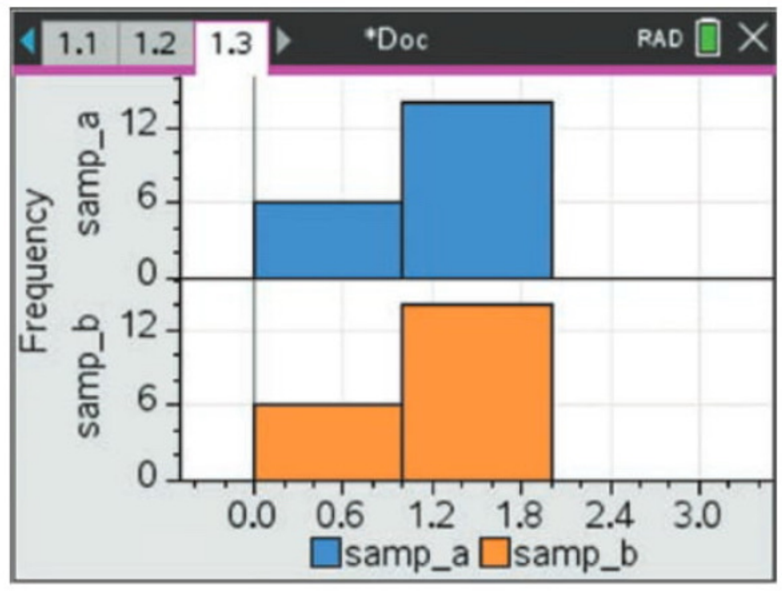
6 Tap in the cell next to the **A**.

7 Press **var** and select **samp_a**.

8 Tap in the cell next to the **B**.

9 Press **var** and select **samp_b**.





10 Add a **Data & Statistics** page.

11 For the horizontal axis, select **samp_a**.

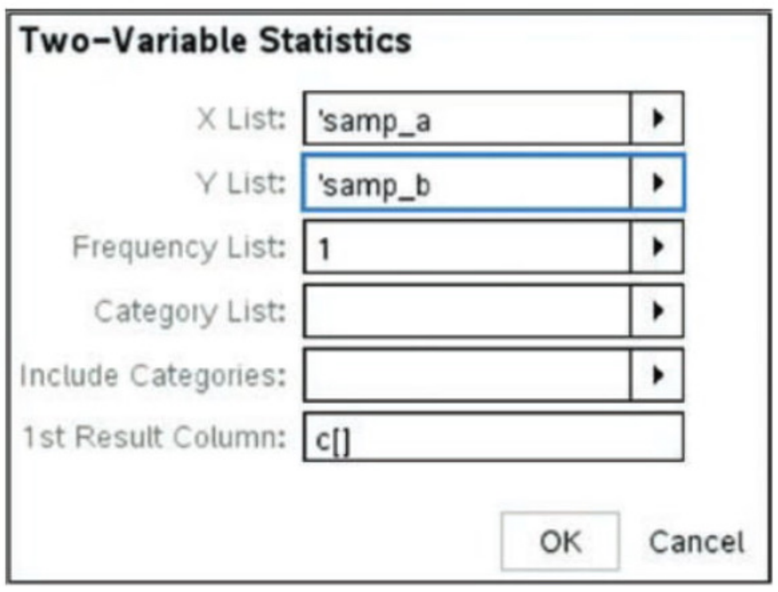
12 Press **menu** > **Plot Type** > **Histogram**.

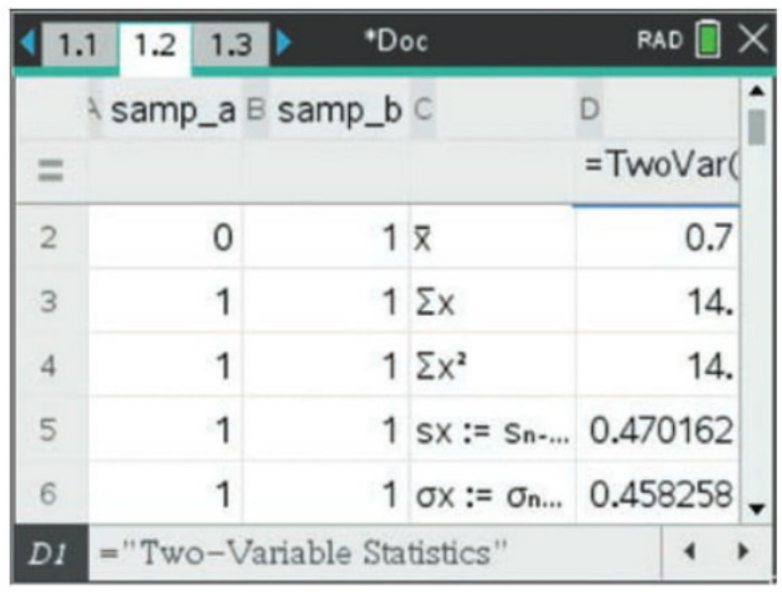
13 The sample a data will be displayed as a histogram.

14 Tap **menu** > **Plot Properties** > **Add X Variable**.

15 Select **samp_b**.

16 The data for both samples will be displayed as histograms.





17 Return to the **Lists & Spreadsheet** page.

18 Press **menu** > **Statistics** > **Stat Calculations** > **Two-Variable**.

19 In the **X List**: field, press the right arrow and select **samp_a**.

20 In the **Y List**: field, press the right arrow and select **samp_b**.

21 Press **enter**.

22 The summary statistics will be displayed in columns **C** and **D**.

23 Scroll down to view all the statistics.

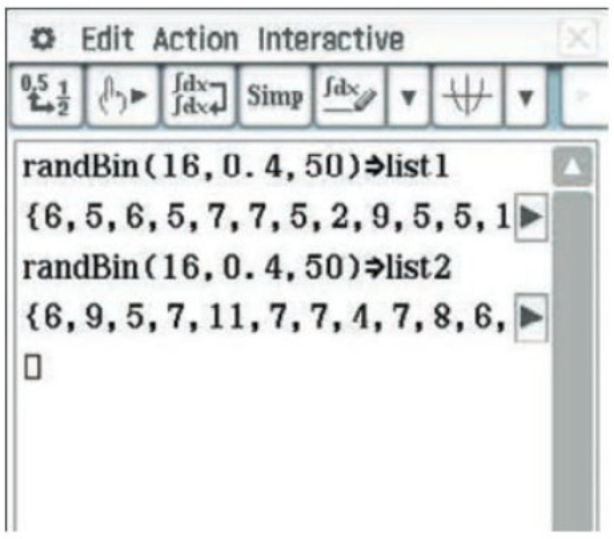
$E(Z) = 0.8$. Both simulated means are approximately 0.8, but vary.

$SD(Z) = 0.4$. Both simulated standard deviations are close to 0.4, but vary.

USING CAS 4 Simulating sample data from a binomial distribution

Simulate two different samples of 50 scores from the discrete binomial random variable, $Z \sim \text{Bin}(16, 0.4)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of Z.

ClassPad



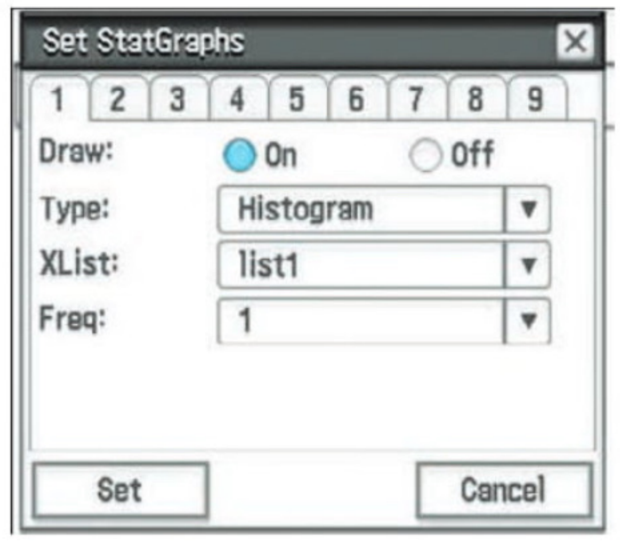
1 Open the **Keyboard** > **Catalog**.

2 Tap **R** then scroll down to select **randBin**.

3 Use the input **randBin**(n, p, m) to generate m values from the distribution $\text{Bin}(n, p)$, as shown above.

4 Store the first sample as **list1**.

5 Repeat for the second sample and store as **list2**.



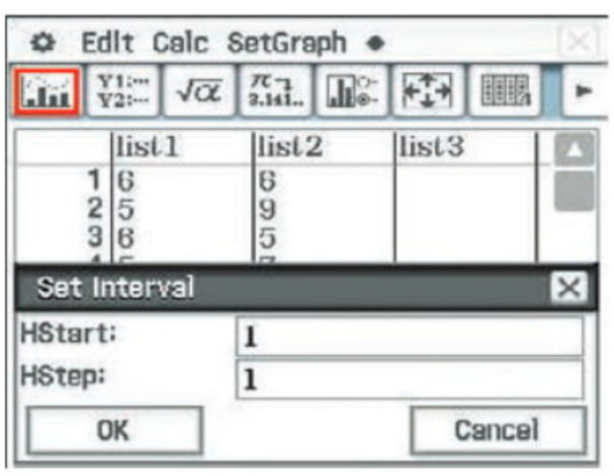
6 Tap **Menu** > **Statistics**.

7 Tap **SetGraph** > **Setting**.

8 For tab **1**, ensure the **Type**: field is **Histogram** and the **XList**: field is **list1**.

9 For tab **2**, ensure the **Type**: field is **Histogram** and the **XList**: field is **list2**.

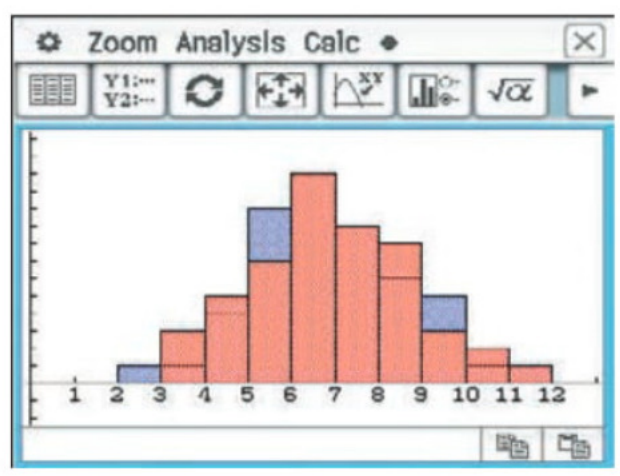
10 Tap **Set**.



11 Tap **SetGraph** and ensure **StatGraph1** and **StatGraph2** are selected.

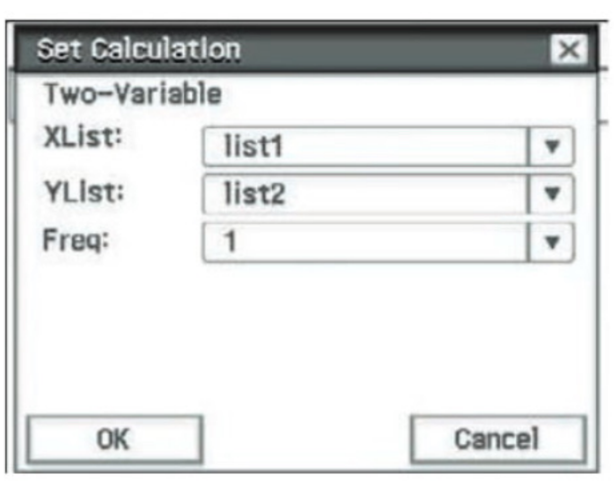
12 Tap **Graph**.

13 In the **Set Interval** dialogue box, change the **HStart**: and **HStep**: fields to **1** and tap **OK**.



14 The histograms of the data will appear in the lower window (the windows have been swapped for this screen).

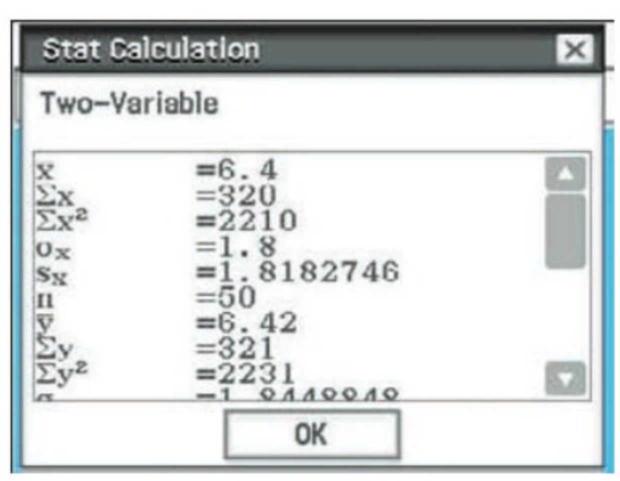
15 Tap the upper window to select **Statistics**.



16 Tap **Calc** > **Two-Variable**.

17 Ensure the **XList**: field is **list1** and the **YList**: field is **list2**.

18 Tap **OK**.



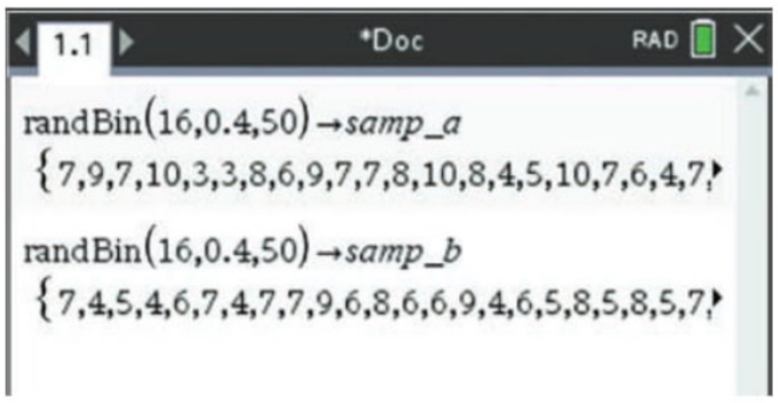
19 The summary statistics will appear in the **Two-Variable** window.

20 Scroll down to view all the statistics.

$E(Z) = 6.4$. Both simulated means are approximately 6.4, but vary.

$SD(Z) = 1.96$. Both simulated standard deviations are approximately 2, but vary.

TI-Nspire

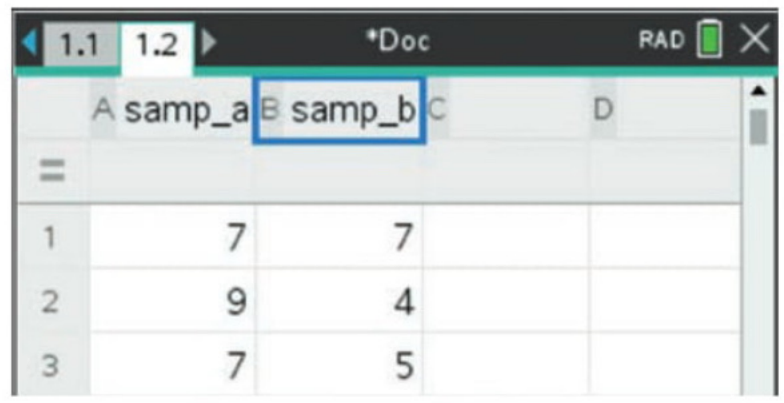


1 Press **catalog** > **randNorm**.

2 Use the input **randBin**$(1, p, m)$ to generate m values from the distribution $\text{Bin}(1, p)$, as shown above.

3 Store the first sample as **samp_a**.

4 Repeat for the second sample and store as **samp_b**.



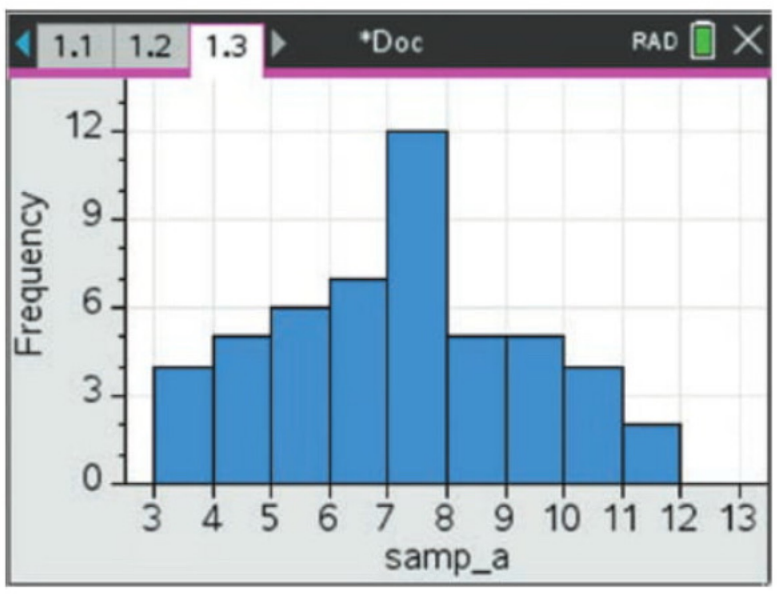
5 Add a **Lists & Spreadsheet** page.

6 Tap in the cell next to the **A**.

7 Press **var** and select **samp_a**.

8 Tap in the cell next to the **B**.

9 Press **var** and select **samp_b**.

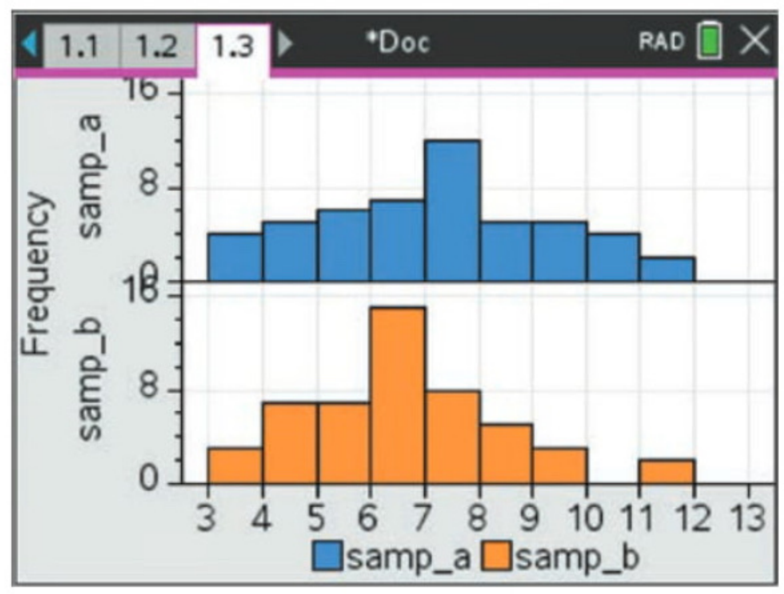


10 Add a **Data & Statistics** page.

11 For the horizontal axis, select **samp_a**.

12 Press **menu** > **Plot Type** > **Histogram**.

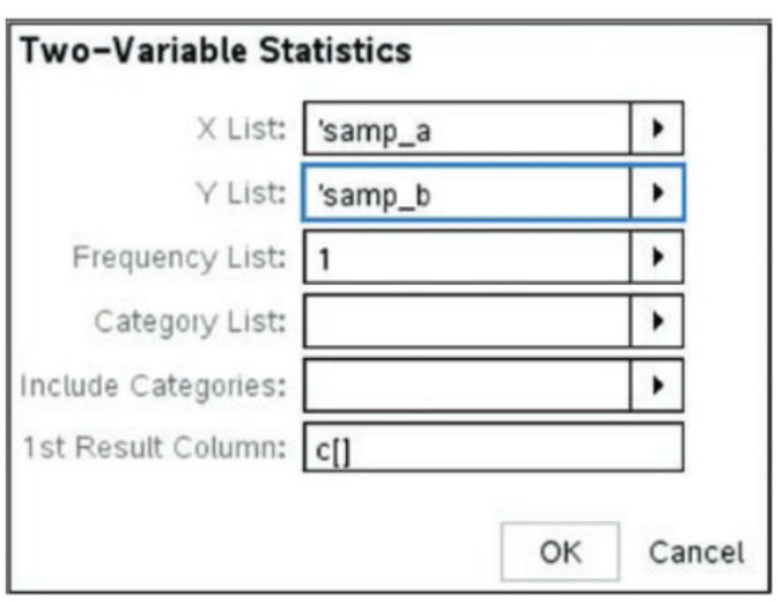
13 The sample a data will be displayed as a histogram.



14 Tap **menu** > **Plot Properties** > **Add X Variable**.

15 Select **samp_b**.

16 The data for both samples will be displayed as histograms.



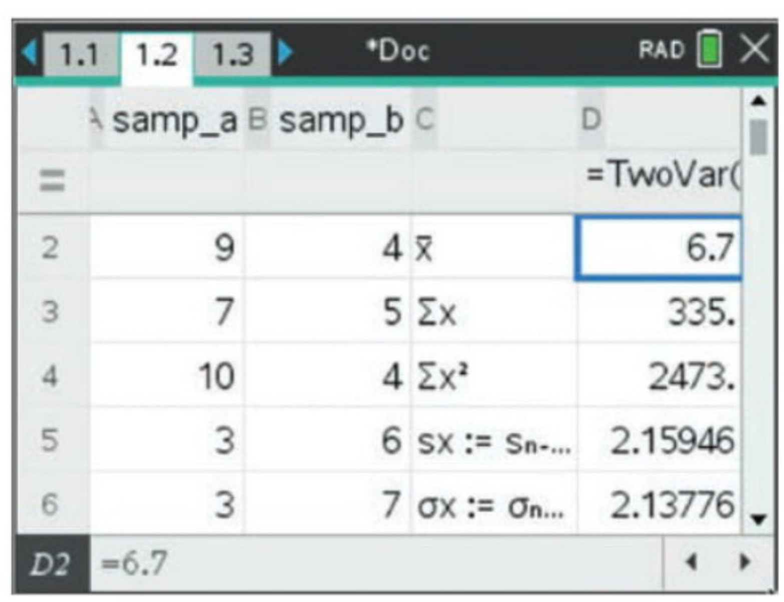
17 Return to the **Lists & Spreadsheet** page.

18 Press **menu** > **Statistics** > **Stat Calculations** > **Two-Variable**.

19 In the **X List**: field, press the right arrow and select **samp_a**.

20 In the **Y List**: field, press the right arrow and select **samp_b**.

21 Press **enter**.



22 The summary statistics will be displayed in columns **C** and **D**.

23 Scroll down to view all the statistics.

$E(Z) = 6.4$. Both simulated means are approximately 6.4, but vary.

$SD(Z) = 1.96$. Both simulated standard deviations are approximately 2, but vary.

Variability of samples and as $n \to \infty$

For each sample of size n taken from a parent distribution defined by the random variable X, the sample statistics and shape of the distribution will vary, but will approximate the parameters and shape of the population distribution.

As $n \to \infty$:

- the mean of a sample will generally tend towards $E(X)$, but can still vary
- the shape of the distribution will better represent the shape of the distribution of X.

EXERCISE 9.1 Random sampling

ANSWERS p. 409

Mastery

1 WORKED EXAMPLE 1 The quality control officer at a local pie factory recorded the weights of six meat pies produced. The weights, to the nearest gram, were 110 g, 105 g, 110 g, 98 g, 101 g and 102 g. Identify the

a population **b** sample size **c** sample statistics.

2 WORKED EXAMPLE 2 A hotel wishes to conduct a survey regarding the quality of room service to be rated on a scale from 1 – Very Poor to 5 – Excellent. Describe a sampling procedure that would ensure randomness.

3 WORKED EXAMPLE 3 For each of the following situations

i identify and explain **one** possible source of bias with this sampling method

ii suggest and describe a random sampling procedure that will minimise bias in this data collection.

a The first 30 people that arrive at the Perth Domestic Airport to catch flights are asked the question 'How many times so far this year have you travelled interstate?'.

b On a Sunday afternoon from 2:00 pm to 3:00 pm, a supermarket store worker asks customers in the self-service checkouts the question 'How many bags have you brought with you today?'.

4 Using CAS 1 Simulate two different samples of 80 scores from the continuous uniform random variable, $X \sim U[15, 75]$. Compare the mean and standard deviation of the samples to the mean and standard deviation of X.

5 Using CAS 2 Simulate two different samples of 50 scores from the continuous normal random variable, $Y \sim N(40, 2.5^2)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of Y.

6 Using CAS 3 Simulate two different samples of 30 scores from the discrete Bernoulli random variable, $Z \sim \text{Bern}(0.35)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of Z.

7 Using CAS 4 Simulate two different samples of 40 scores from the discrete binomial random variable, $T \sim \text{Bin}\left(75, \frac{1}{3}\right)$. Compare the mean and standard deviation of the samples to the mean and standard deviation of T.

Calculator-free

8 © SCSA MM2017 Q12a (4 marks) The Slate Tablet Company produces a variety of electronic tablets. It wants to gather information on consumers' interest in its tablets. In each of the following cases, comment, giving reasons, whether or not the proposed sampling method introduces bias.

a A Slate Tablet Company representative stood outside an electronics store on a Saturday morning and asked people entering the store 'If you were to purchase an electronic tablet would you choose a Slate Tablet or an inferior brand?' (2 marks)

b Fifteen hundred randomly selected mobile phone numbers were telephoned and people were asked 'Which brand of electronic tablet do you prefer?' (2 marks)

9 © SCSA MM2018 Q17c (2 marks) Tina believes that approximately 60% of the mangoes she produces on her farm are large. She takes a random sample of 500 mangoes from a day's picking. Tina decides to select the mangoes for her sample as they pass along the conveyor belt to be sorted. Describe briefly how Tina should select her sample.

10 © SCSA MM2019 Q13bc (4 marks) The proportion of working adults who miss breakfast on week days is estimated to be 40%. Tom takes a random sample of 400 adults to investigate this theory. He obtained his sample by selecting the first 400 workers he met in a busy mall in Perth city during lunchtime.

a Discuss briefly **two** possible sources of bias in Tom's sample. (2 marks)

Amir suggests that a better sampling scheme is to obtain a random sample of 400 voters and contact them by telephone.

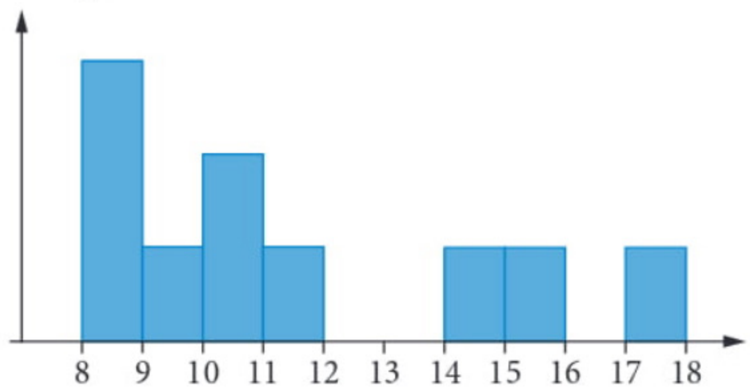
b Outline **one** source of bias in Amir's sampling scheme. (1 mark)

c Which of Tom's or Amir's sampling scheme is better? Provide a reason for your choice. (1 mark)

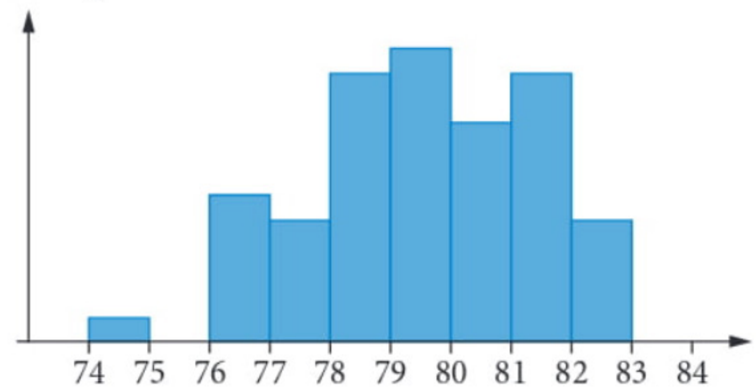
11 © SCSA MM2020 Q14c (4 marks) A suburban council hires a consultant to estimate the proportion of residents of the suburb who use its library. The consultant decides to select the sample by standing on the roadside outside the library at lunchtime and asking a random sample of the passers-by whether they use the library. Identify and explain **two** possible sources of bias with this sampling scheme.

12 (9 marks) CAS is used to simulate three different samples of varying sample sizes, as seen in the histograms provided.

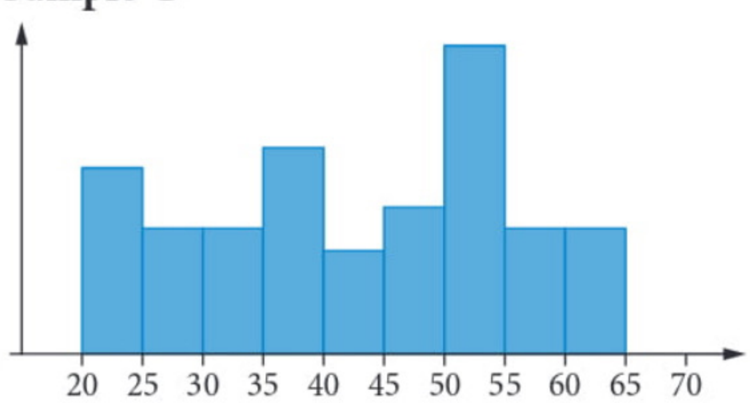
Sample A



Sample B



Sample C



a Explain why it is not likely that these three samples came from the same population. (2 marks)

It is later known that the samples were simulated from a normal random variable X, a uniform random variable Y and a binomial random variable Z.

b **i** Identify the sample that is most likely simulated from the binomial random variable. (1 mark)

ii Justify whether the value of p in $Z \sim \text{Bin}(n, p)$ is less than or greater than 0.5. (2 marks)

c **i** Identify the sample that is most likely simulated from the normal random variable. (1 mark)

ii Estimate the value of $\text{E}(X)$. (1 mark)

d **i** Identify the sample that is most likely simulated from the uniform random variable. (1 mark)

ii Suppose the sample size simulated from this variable doubled. Describe the likely effect on the shape of the distribution. (1 mark)

9.2 The sampling distribution of sample proportions

Video playlist The sampling distribution of sample proportions

Worksheets Sample proportions

Sample proportion calculations

Sample proportion as a random variable, $\hat{p}$

Suppose that a random sample of 500 Australians was taken, and the eye colour of each individual was noted. Now imagine in this sample, it was found that 140 of the 500 had blue eyes. Then it can be said that the **sample proportion** of Australians with blue eyes is $\frac{140}{500}$ or 0.28 or 28%. This sample proportion, denoted as $\hat{p}$, is the relative frequency or experimental probability of a particular 'success condition' out of the sample size n.

Sample proportion as a value

For a sample of size n, the value of the single sample proportion is given by:

$$\hat{p} = \frac{\text{number of observed successes}}{n}$$

Sample proportions can be written as fractions, decimals or percentages.

Now suppose another random sample of 500 Australians found that 108 had blue eyes. Then the sample proportion of Australians with blue eyes for this sample is $\frac{108}{500} = 0.216 = 21.6\%$. As we have previously seen, due to the nature of random sampling, we can expect there to be variability between samples and so we would expect the proportion of Australians with blue eyes to change from sample to sample.

9.2

The purpose of collecting different random samples and examining the sample proportion of Australians with blue eyes is to find a way to estimate the true **population proportion**, p, of Australians with blue eyes because it is impractical and largely impossible to take a census of eye colour of the entire Australian population. As a result, each sample proportion $\hat{p}$ acts as a point estimate for p, which we assume remains constant and does not change like the sample proportions.

WORKED EXAMPLE 4 **Calculating and using sample proportions**

A school has a population of 1080 students. A random sample of 200 students was taken and it was found that 43 were not born in Australia.

a Calculate the sample proportion of students born overseas.

b Hence, use this sample proportion as a point estimate to estimate the total number of students born overseas.

Steps	Working
a Express the number of students born overseas as a fraction of the sample size.	$\hat{p} = \dfrac{43}{200} = 0.215 = 21.5\%$
b **1** Use the proportion to estimate for a population of size 1080.	$0.215 \times 1080 = 232.2$
2 Round to the nearest whole.	It is estimated that 232 students from this school population were born overseas.

In each sample of 500 Australians, we can consider each Australian a Bernoulli random variable or a Bernoulli trial, X_i, for $1 \leq i \leq 500$ where $x_i = 1$ means that the ith Australian surveyed has blue eyes and $x_i = 0$ means that the ith Australian surveyed does not have blue eyes. The distribution can be written as $X_i \sim \text{Bern}(p)$ where p is the probability of success of an Australian having blue eyes. This is the population proportion that is currently unknown. We also know that $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$.

Now when asking each of the 500 Australians, if we assume that each of the trials, $X_1, X_2 \ldots X_{500}$ are identically and independently distributed Bernoulli trials, then we have a binomially distributed random variable X, which represents the number of Australians with blue eyes such that $X \sim \text{Bin}(n, p)$. We should also remember that $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

So, the sample proportion $\hat{p}$ is a random variable that can change from sample to sample, such that the number of observed successes out of a sample of size n is determined by a binomially distributed random variable, X.

Sample proportion as a random variable

Let $X \sim \text{Bin}(n, p)$, where n is the sample size and p is the probability of success (i.e. true population proportion). Then the random variable $\hat{p}$ is defined as

$$\hat{p} = \frac{X}{n}$$

and represents the set of all possible sample proportions that can exist when a sample of size n is taken. It is considered a random variable because its outcomes are the result of a probability experiment.

Imagine we now took lots of different samples of 500 Australians and calculated the sample proportion of Australians with blue eyes for each sample. The distribution of all of the values of $\hat{p}$ is called the **sampling distribution of sample proportions**.

The mean, variance and standard deviation of $\hat{p}$

Given that $\hat{p} = \frac{X}{n}$, we can consider it a binomial random variable scaled by a factor of $\frac{1}{n}$. As a result, we can use the fact that for $X \sim \text{Bin}(n, p)$, $\text{E}(X) = np$ and $\text{Var}(X) = np(1-p)$ to deduce the expected value, variance and standard deviation of $\hat{p}$ as a random variable.

Expected value, variance and standard deviation of $\hat{p}$

Considering $\hat{p} = \frac{1}{n}X$ as a linear change of scale, then:

$$\begin{aligned}\text{E}(\hat{p}) &= \text{E}\left(\frac{1}{n}X\right)\\ &= \frac{1}{n}\text{E}(X)\\ &= \frac{np}{n}\\ &= p\end{aligned}$$

That is, if the population proportion is p, then the expected value (or mean) of all the sample proportions taken from different samples of fixed size n, is p. That is, the mean of the sampling distribution of sample proportions is a good and unbiased estimator of the true population proportion because it is independent of n.

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}\left(\frac{1}{n}X\right)\\ &= \frac{1}{n^2}\text{Var}(X)\\ &= \frac{np(1-p)}{n^2}\\ &= \frac{p(1-p)}{n}\end{aligned}$$

$$\text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

As sample size increases, i.e. $n \to \infty$, $\text{Var}(\hat{p}) \to 0$ and $\text{SD}(\hat{p}) \to 0$. That is, there should be very little variation in the different values of $\hat{p}$ taken from different samples of a significantly large, fixed size n.

WORKED EXAMPLE 5 Expected value, variance and standard deviation of $\hat{p}$ from X with a known p

Given that $X \sim \text{Bin}(50, 0.6)$, determine

a $\text{E}(\hat{p})$ **b** $\text{Var}(\hat{p})$ **c** $\text{SD}(\hat{p})$, correct to three decimal places.

Steps	Working
a Use the fact that $\text{E}(\hat{p}) = p$.	$\text{E}(\hat{p}) = 0.6$
b Use the formula $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$.	$\text{Var}(\hat{p}) = \frac{0.6(0.4)}{50} = 0.0048$
c Use the formula $\text{SD}(\hat{p}) = \sqrt{\text{Var}(\hat{p})}$.	$\text{SD}(\hat{p}) = \sqrt{0.0048} = 0.069$

Let's now revisit the context of Australians with blue eyes. The problem in this example is that although we know the value of $n = 500$, we do not know the true value of p. That is, we do not know the true probability of success of selecting someone from the Australian population with blue eyes. This type of situation whereby p is unknown is most common.

In these situations, we have to use a sample proportion $\hat{p}$ from one specific random sample, hopefully fair and representative, as a point estimate for p. For example, suppose we let the sample proportion obtained from the first sample of 500 Australians, $\hat{p} = 0.28$, be a point estimate for the true value of p. Then we assume that $X \sim \text{Bin}(n, \hat{p})$.

9.2

It then follows that the sampling distribution of sample proportions $\hat{p}$ will have:

$$\text{E}(\hat{p}) = \hat{p}$$

$$\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$$

$$\text{SD}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

So, for our example where $X \sim \text{Bin}(500, 0.28)$, we will have:

$$\text{E}(\hat{p}) = 0.28$$

$$\text{Var}(\hat{p}) = \frac{0.28(0.72)}{500} = 0.0004032$$

$$\text{SD}(\hat{p}) = \sqrt{0.0004032} = 0.0201$$

WORKED EXAMPLE 6 **Expected value, variance and standard deviation of $\hat{p}$ from X with an unknown p**

From a sample of 20 people, it was found that 8 had colds last winter.

a State the sample proportion of people who had colds last winter, $\hat{p}$.

b Using this value of $\hat{p}$ as a point estimate for the true population proportion of people who had colds last winter, determine

i $\text{E}(\hat{p})$ **ii** $\text{Var}(\hat{p})$ **iii** $\text{SD}(\hat{p})$, correct to three decimal places.

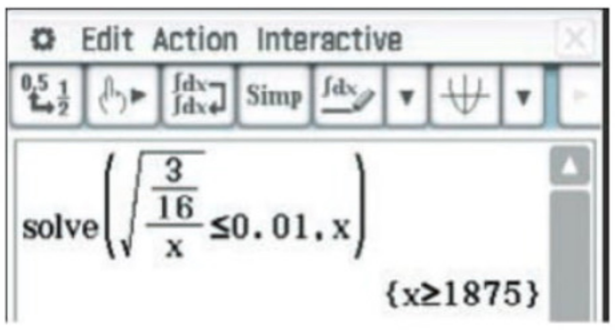
Steps	**Working**
a Express the number of people who had colds last winter as a proportion of the sample size.	$\hat{p} = \frac{8}{20} = 0.4 = 40\%$
b i Use the fact that $\text{E}(\hat{p}) = \hat{p}$.	$\text{E}(\hat{p}) = 0.4$
ii Use the formula $\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$.	$\text{Var}(\hat{p}) = \frac{0.4(0.6)}{20} = 0.012$
iii Use the formula $\text{SD}(\hat{p}) = \sqrt{\text{Var}(\hat{p})}$.	$\text{SD}(\hat{p}) = \sqrt{0.012} = 0.110$

WORKED EXAMPLE 7 **Using standard deviation of $\hat{p}$ to find an unknown parameter**

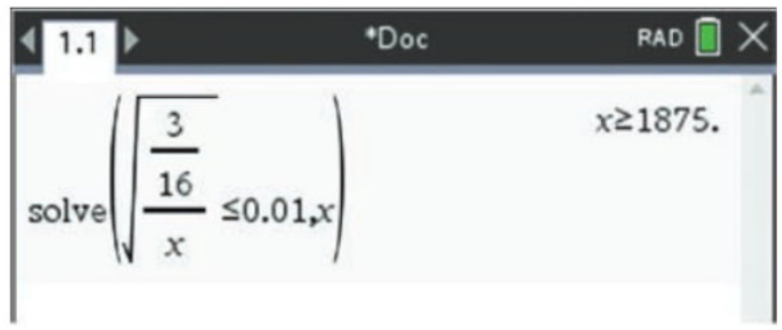
In a large population of fish, the true population proportion of angel fish is known to be $\frac{1}{4}$. Let $\hat{p}$ be the random variable that represents the sample proportion of angel fish for samples of fixed size n taken from the population. Find the smallest integer value of n such that the standard deviation of $\hat{p}$ is less than or equal to 0.01.

Steps	**Working**
1 Establish an inequality involving $\text{SD}(\hat{p})$ using the given values.	$\sqrt{\frac{\frac{1}{4}\left(\frac{3}{4}\right)}{n}} \le 0.01$
2 Solve for n using CAS. (See CAS screens on following page.)	$n \ge 1875$

ClassPad



TI-Nspire



Alone, these calculations involving $\mathrm{E}(\hat{p})$, $\mathrm{Var}(\hat{p})$ and $\mathrm{SD}(\hat{p})$ are just statistics and we cannot do much with them until we know more about the shape of the sampling distribution of sample proportions, $\hat{p}$.

Approximate normality and the central limit theorem

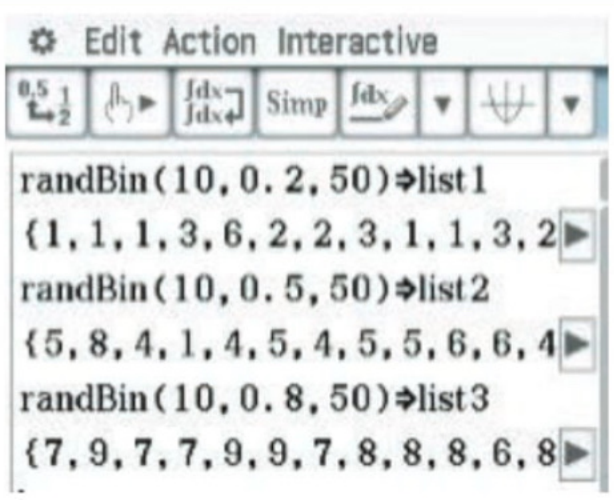
From your work with binomial random variables in Chapter 5, you will recall that the shape of the frequency histogram of a binomial distribution depends on the values of n and p. We can use CAS to simulate binomial distributions with a fixed n value and differing p values to observe the effect.

USING CAS 5 Simulating binomial distributions

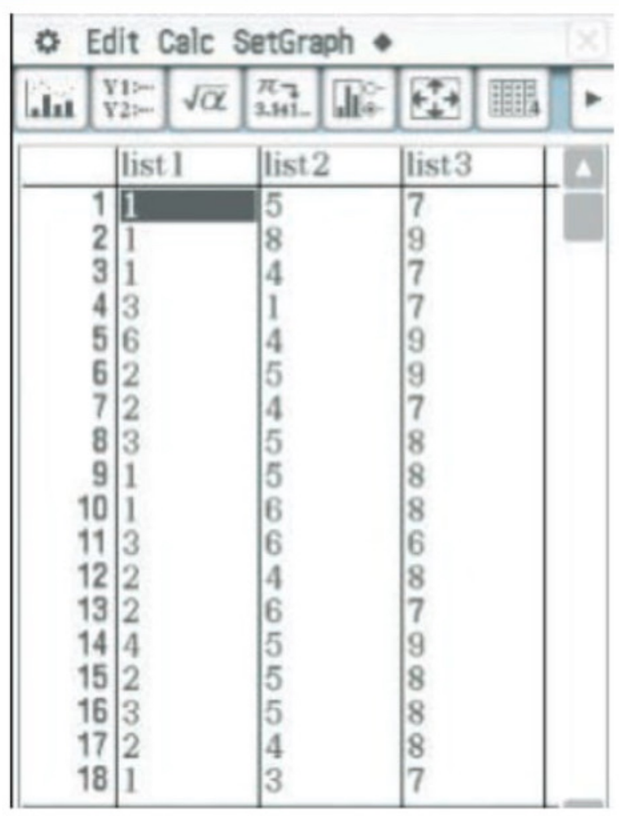
Simulate 50 different observations from each of the following binomial random variables, graphing each of the results.

a $X \sim \text{Bin}(10, 0.2)$ **b** $X \sim \text{Bin}(10, 0.5)$ **c** $X \sim \text{Bin}(10, 0.8)$

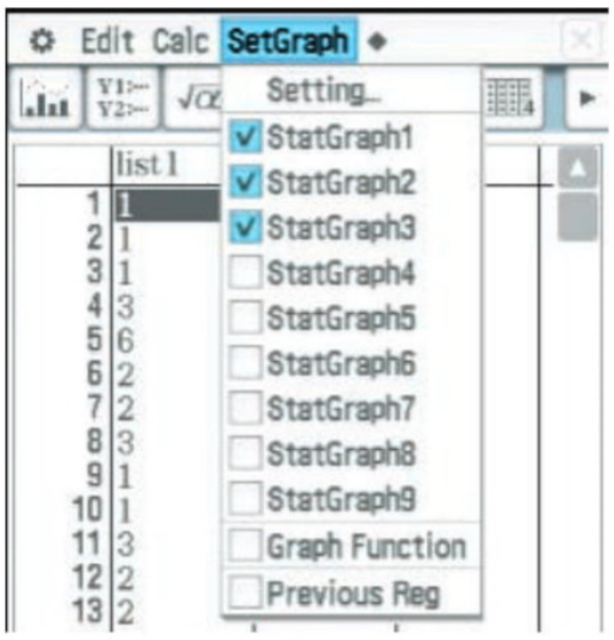
ClassPad



1 Open the **Keyboard** then tap **Catalog** > **R** to jump to the functions starting with **r**.
2 Select **randBin(**.
3 Generate 50 observations of a binomial random variable, with $n = 10$ trials and $p = 0.2$.
4 Store the values into **list1**.
5 Repeat for the second and third set of values, storing the results into **list2** and **list3** respectively.



6 Tap **Menu** > **Statistics**.
7 The values generated will be displayed in **list1**, **list2** and **list3**.
Note that the values are selected randomly so answers will vary.



8 Tap **SetGraph**.

9 Tap to select **StatGraph1**, **StatGraph2** and **StatGraph3**, as shown above. Tick or untick these to view individual graphs.

10 Tap **SetGraph** > **Setting**.

11 For the **Type:** field, select **Histogram**.

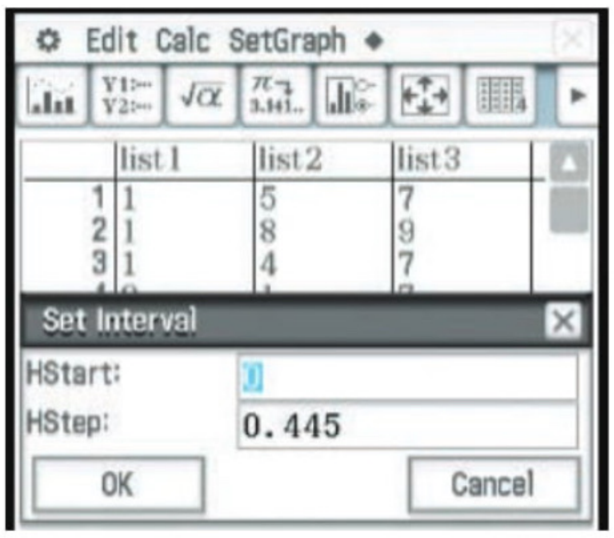
12 Tap tab **2** at the top of the page.

13 Set the **Type**: field to **Histogram**.

14 Change the **XList**: field to **list2**.

15 Repeat for tab **3** by selecting **Histogram** and changing the **XList**: field to **list3**.

16 Tap **Set**.



17 Tap **Graph**.

18 When the **Set Interval** dialogue box is displayed, tap **OK** to accept the default settings.

NOTE: You can change the HStart and HStep if you want to view the histograms on a different scale. For example, try a HStep of 1.

19 Histograms of the three random samples will be displayed.

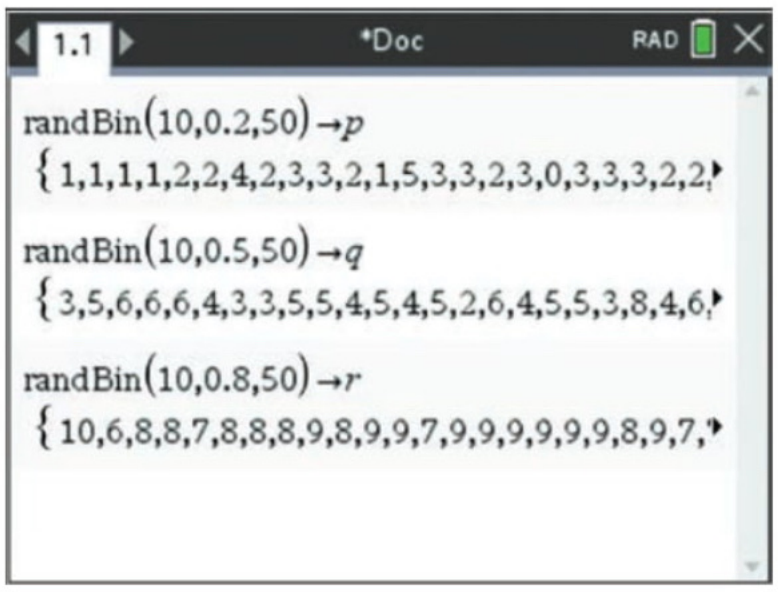
20 Compare the alignment of the histograms with their respective probabilities.

StatGraph1 – purple

StatGraph2 – orange

StatGraph3 – green

TI-Nspire

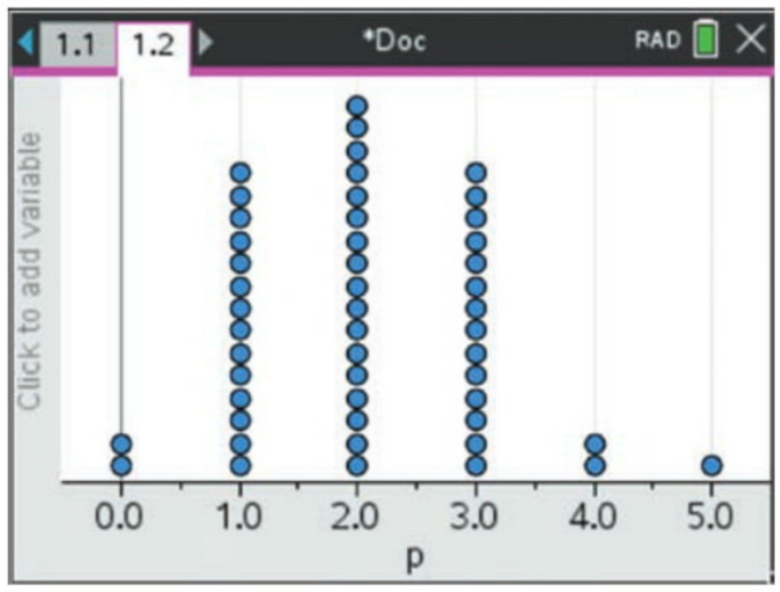


1 Press **menu** > **Probability** > **Random** > **Binomial**.

2 Generate 50 observations of a binomial random variable, with $n = 10$ trials and $p = 0.2$.

3 Press **ctrl** + **var** to store the result in **p**.

4 Repeat for the second and third set of values, storing the results in **q** and **r** respectively.

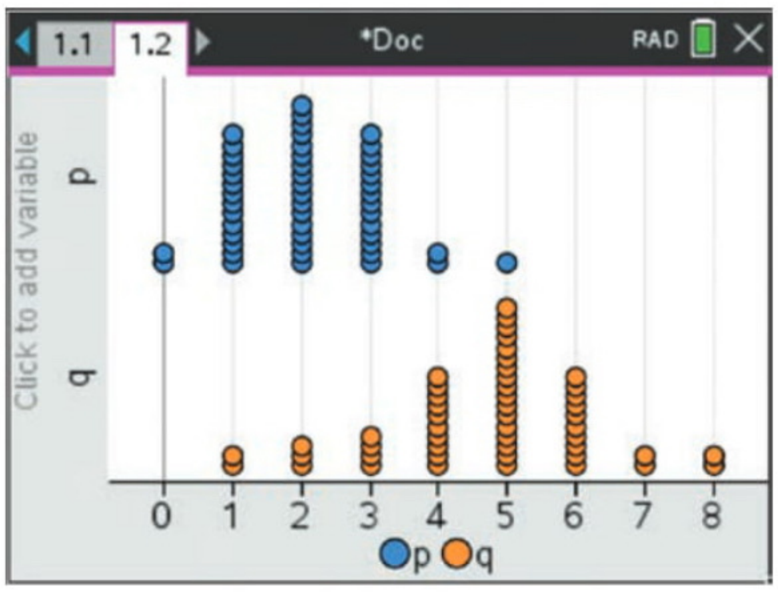


5 Add a **Data & Statistics** page.

6 For the horizontal axis, click to select the variable **p**.

7 A dot plot of the probabilities will be displayed.

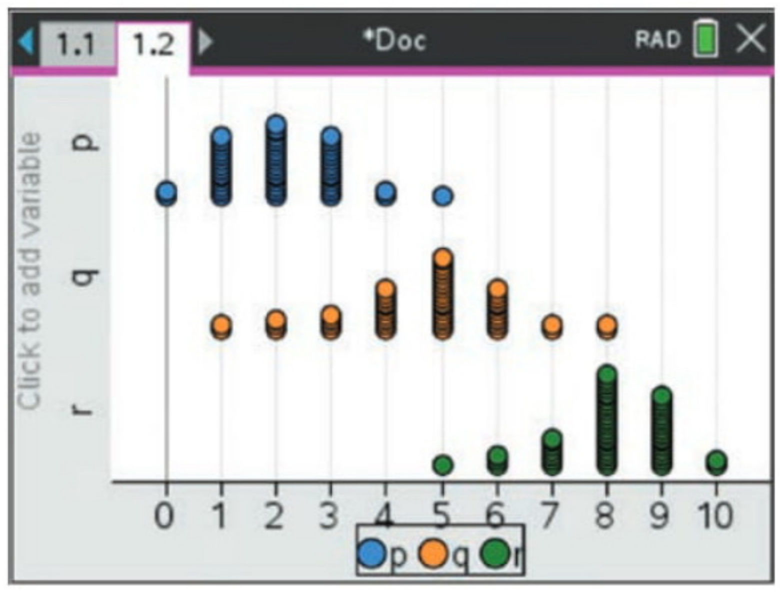
Note that the values are selected randomly so answers will vary.



8 Press **menu** > **Plot Properties** > **Add X Variable**.

9 Select the variable **q**.

10 Parallel dot plots for **p** and **q** will be displayed.



11 Repeat to display the dot plot for the variable **r**.

12 All three dot plots will be displayed.

13 Compare the alignment of the dot plots with their respective probabilities.

From these simulations, we should notice something about the shape of each of the distributions.

- For $X \sim \text{Bin}(10, 0.2)$, the distribution was positively skewed, with an expected value (mean) of $10(0.2) = 2$.
- For $X \sim \text{Bin}(10, 0.5)$, the distribution was approximately symmetrical, with an expected value (mean) of $10(0.5) = 5$.
- For $X \sim \text{Bin}(10, 0.8)$, the distribution was negatively skewed, with an expected value (mean) of $10(0.8) = 8$.

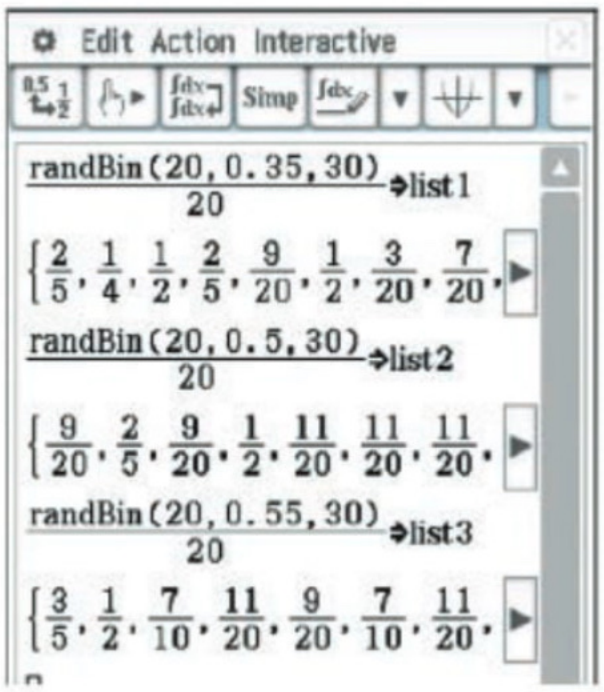
We can also run a simulation exercise for sample proportions by dividing each of the simulated observations from the binomial random variables by the value of n.

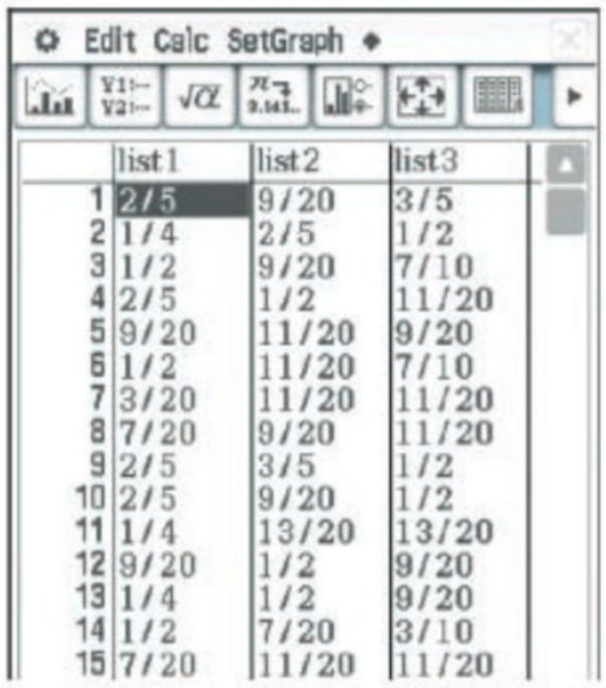
USING CAS 6 Simulating the sampling distributions of sample proportions

Simulate 30 different observations from each of the following binomial random variables, and graph the distributions of each of the corresponding distributions of $\hat{p} = \frac{X}{n}$.

a $X \sim \text{Bin}(20, 0.35)$ **b** $X \sim \text{Bin}(20, 0.5)$ **c** $X \sim \text{Bin}(20, 0.75)$

ClassPad





1 Open the **Keyboard** then tap **Catalog** > **R** to jump to the functions starting with r.

2 Select **randBin(**.

3 Generate 50 observations of a binomial random variable, with $n = 10$ trials and $p = 0.35$, and divide the set by 20.

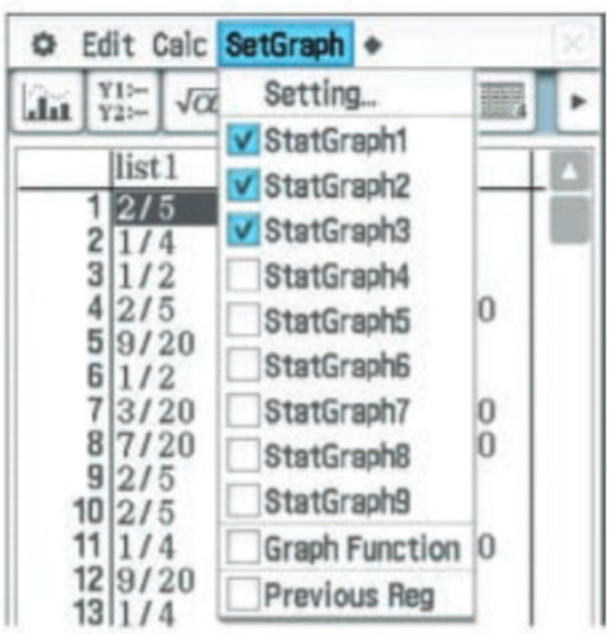
4 Store the values into **list1**.

5 Repeat for the second and third set of values, storing the results into **list2** and **list3** respectively.

6 Tap **Menu** > **Statistics**.

7 The values generated will be displayed in **list1**, **list2** and **list3**.

Note that the values are selected randomly so answers will vary.



8 Tap **SetGraph**.

9 Tap to select **StatGraph1**, **StatGraph2** and **StatGraph3** as shown above. Tick or untick these to view individual graphs.

10 Tap **SetGraph** > **Setting**.

11 For the **Type:** field, select **Histogram**.

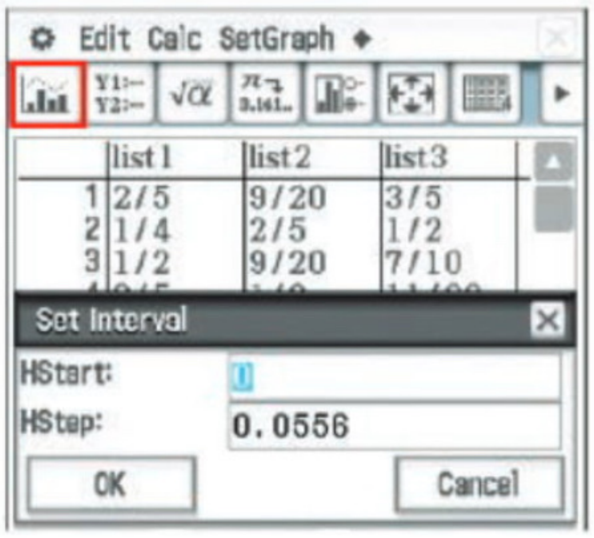
12 Tap tab **2** at the top of the page.

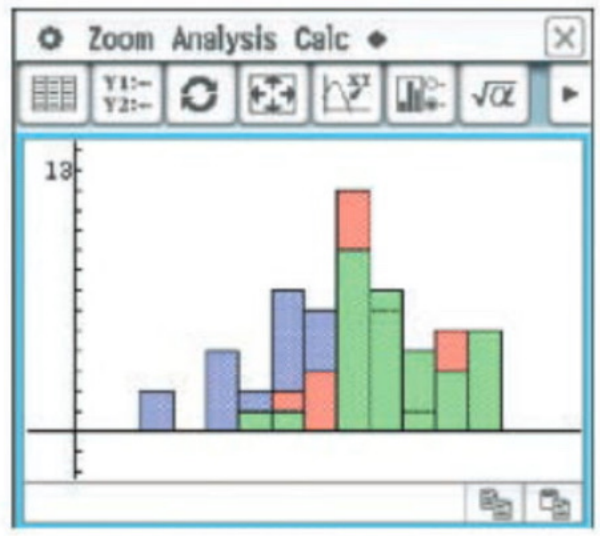
13 Set the **Type:** field to **Histogram**.

14 Change the **XList:** field to **list2**.

15 Repeat for tab **3** by selecting **Histogram** and changing the **XList:** field to **list3**.

16 Tap **Set**.





17 Tap **Graph**.

18 When the **Set Interval** dialogue box is displayed, tap **OK** to accept the default settings.

NOTE: You can change the HStart and HStep if you want to view the histograms on a different scale. For example, try a HStep of 1.

19 Histograms of the three random samples will be displayed.

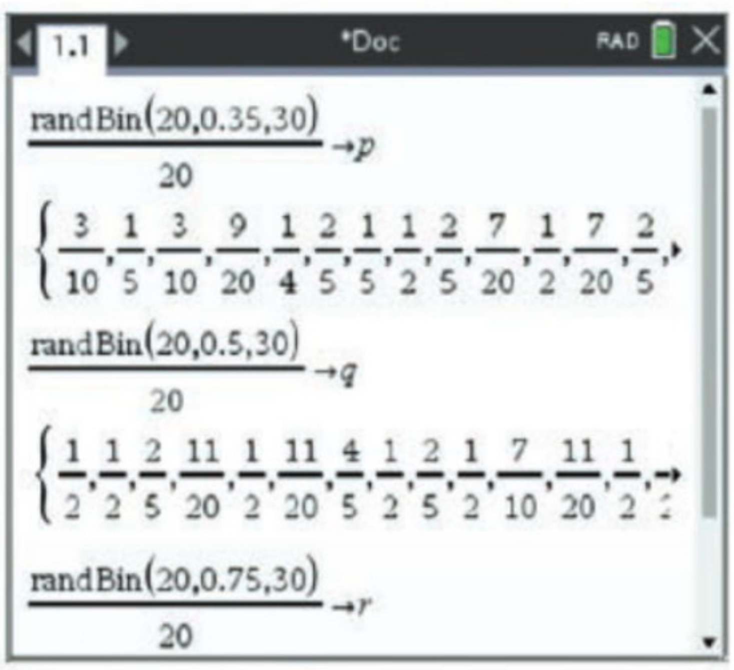
20 Compare the alignment of the histograms with their respective probabilities.

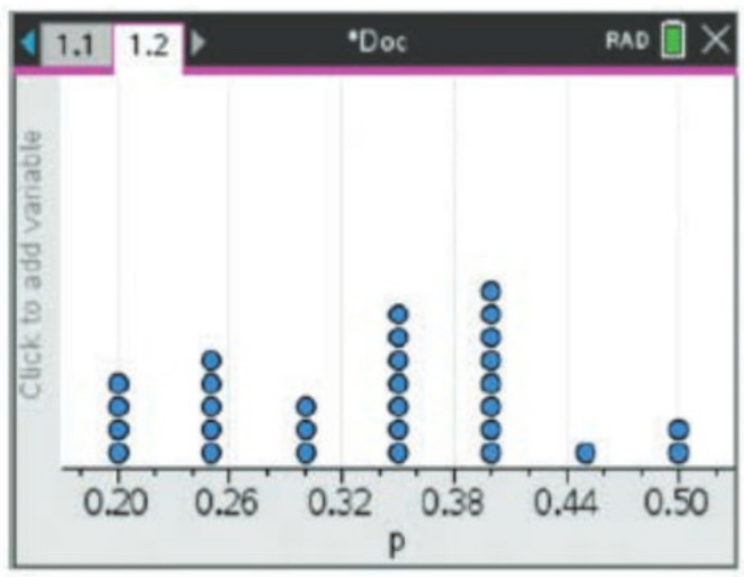
StatGraph1 – purple

StatGraph2 – orange

StatGraph3 – green

TI-Nspire





1 Press **menu** > **Probability** > **Random** > **Binomial**.

2 Generate 50 observations of a binomial random variable, with n = 10 trials and p = 0.35, and divide the set by 20.

3 Press **ctrl** + **var** to store the result in **p**.

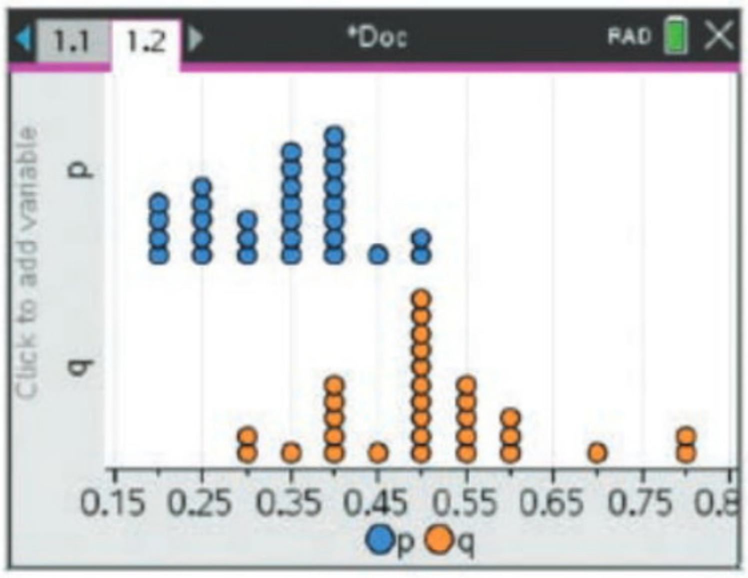
4 Repeat for the second and third set of values, storing the results in **q** and **r** respectively.

5 Add a **Data & Statistics** page.

6 For the horizontal axis, click to select the variable **p**.

7 A dot plot of the probabilities will be displayed.

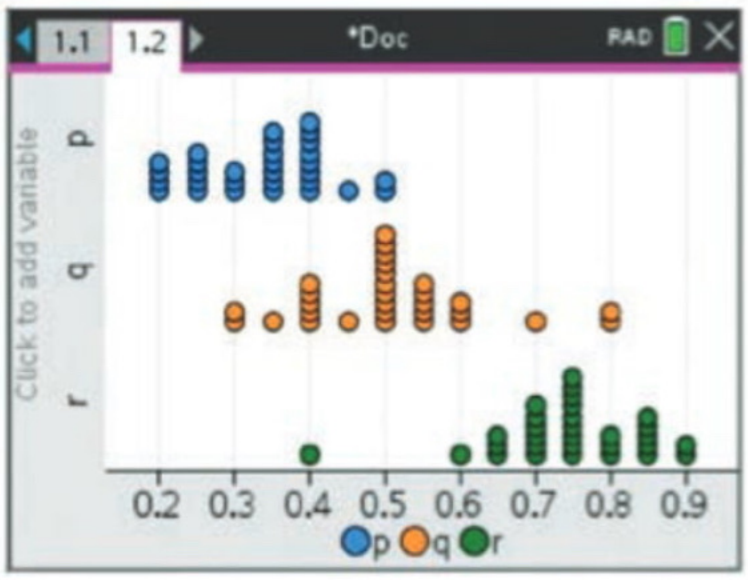
Note that the values are selected randomly so answers will vary.



8 Press **menu** > **Plot Properties** > **Add X Variable**.

9 Select the variable **q**.

10 Parallel dot plots for **p** and **q** will be displayed.



11 Repeat to display the dot plot for the variable **r**.

12 All three dot plots will be displayed.

13 Compare the alignment of the dot plots with their respective probabilities.

The shape of each of the corresponding sampling distributions of sample proportions will be the same as the shape of the binomial distribution, but with a rescaled horizontal axis by a factor of $\frac{1}{n}$.

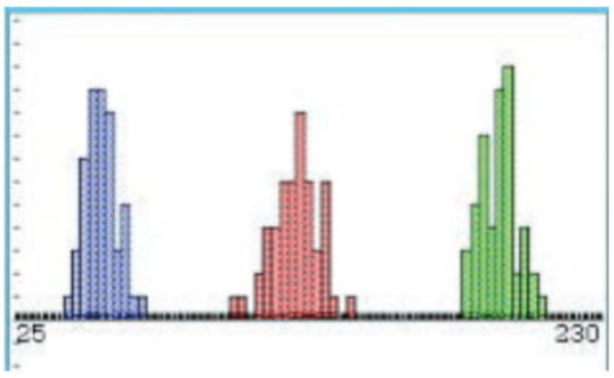
It should be more obvious that these histograms are approximately symmetrical about their expected values of 50, 125 and 200, respectively. Given the apparent symmetry about the mean, we could now consider it appropriate for each of these binomial random variables to be suitably modelled using a normally distributed random variable.

The crucial question is: *when is it appropriate to assume approximate normality*? Let's consider which of the simulations for $n = 10$ gave an approximately symmetrical distribution and hence could be appropriately modelled by a normal distribution.

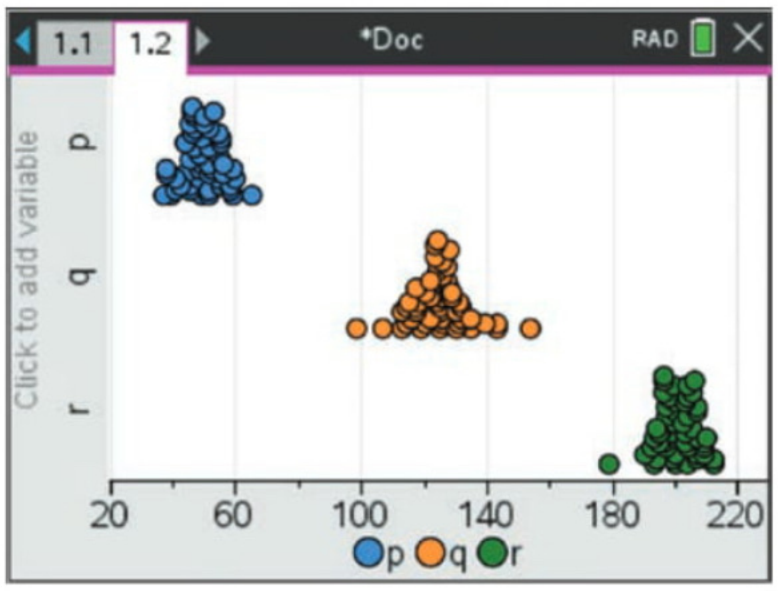
- For $X \sim \text{Bin}(10, 0.2)$, $np = 2$ and the distribution was positively skewed due to $p = 0.2$. The normal distribution is not appropriate!
- For $X \sim \text{Bin}(10, 0.5)$, $np = 5$ and the distribution was approximately symmetrical due to $p = 0.5$. The normal distribution could be appropriate!
- For $X \sim \text{Bin}(10, 0.8)$, $np = 8$ and the distribution was negatively skewed due to $p = 0.8$. The normal distribution is not appropriate!

Suppose we repeat the above simulation exercises with a much larger sample size, say $n = 250$, but the same values of $p = 0.2$, 0.5 and 0.8. Running this simulation once for these three binomial random variables, we might observe three such histograms.

ClassPad



TI-Nspire



The approximate normality of a binomial distribution with $p \approx 0.5$

For a binomially distributed random variable $X \sim \text{Bin}(n, p)$, if $p \approx 0.5$, and n is sufficiently large, the values of x can be approximated by a normally distributed random variable X_N with parameters $\mu = np$ and $\sigma^2 = np(1-p)$. That is,

$$X_N \sim N(np, np(1-p))$$

Exam hack

In these cases, 'sufficiently large' is often considered as $n \geq 30$, but when p is approximately 0.5, some leniency can be given due to the symmetry of the distribution. Remember that you could always check three standard deviations either side of the p value to ensure that the normal approximation is still appropriate!

Now let's consider the cases where n was significantly larger.

- For $X \sim \text{Bin}(250, 0.2)$, $np = 50$ and the distribution was approximately symmetrical even though $p = 0.2$. The normal distribution could be appropriate!
- For $X \sim \text{Bin}(250, 0.5)$, $np = 125$ and the distribution was approximately symmetrical due to $p = 0.5$. The normal distribution could be appropriate!
- For $X \sim \text{Bin}(250, 0.8)$, $np = 200$ and the distribution was approximately symmetrical even though $p = 0.8$. The normal distribution could be appropriate!

The approximate normality of a binomial distribution with a sufficiently large n

For a binomially distributed random variable $X \sim \text{Bin}(n, p)$, if n is sufficiently large enough to cater for a value of p that deviates from 0.5, the values of x can be approximated by a normally distributed random variable X_N with parameters $\mu = np$ and $\sigma^2 = np(1-p)$. That is,

$$X_N \sim N(np, np(1-p)).$$

Some mathematicians like to put a minimum condition on the size of np and $n(1-p)$ instead of having a 'sufficiently large n'. A commonly used restriction is that both $np \geq 10$ and $n(1-p) \geq 10$. That is, there are at least 10 'successful' and 10 'failed' observations. This may vary in different texts.

As a result, when a binomially distributed random variable X is approximately normal, then it can be said the corresponding sampling distribution of sample proportions taken from X, $\hat{p} = \dfrac{X}{n}$, is also approximately normal.

The approximate normality of the sampling distribution of sample proportions $\hat{p}$

For a binomially distributed random variable $X \sim \text{Bin}(n, p)$ that can be approximated by a normally distributed random variable X_N with parameters $\mu = np$ and $\sigma^2 = np(1-p)$, then $\hat{p} = \dfrac{X_N}{n}$ is approximately normal such that:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

When a point estimate is used to estimate p, then

$$\hat{p} \sim N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

A good way of checking whether n was sufficiently large is by ensuring that all values of $\hat{p}$ within three standard deviations of the mean (i.e. approximately 99.7% of $\hat{p}$ values) lie between $0 \leq \hat{p} \leq 1$, as we cannot have negative sample proportions, or sample proportions larger than 1.

The above result is also known as the **central limit theorem** for the sampling distribution of sample proportions.

9.2

WORKED EXAMPLE 8 Describing the sampling distribution of sample proportions

On a Friday afternoon, a random sample of 58 car spaces at Westfield shopping centre carpark were observed and it was found that 24 of the spaces were occupied by a car.

a State the sample proportion of car spaces that were occupied, $\hat{p}$, correct to four decimal places.

b Determine $E(\hat{p})$ and $SD(\hat{p})$, correct to four decimal places.

c Describe the distribution of the random variable $\hat{p}$, justifying your answer.

Steps	Working
a Express the number of occupied car spaces as a proportion of total car spaces observed.	$\hat{p} = \frac{24}{58} = 0.4138$
b 1 Assume the observed sample proportion is a point estimate for the value of p.	Let $\hat{p} = 0.4138$ be a point estimate of p.
2 Use the fact that $E(\hat{p}) = \hat{p}$.	$E(\hat{p}) = \frac{24}{58} = 0.4138$
3 Use the formula $SD(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.	$SD(\hat{p}) = \sqrt{\frac{\frac{24}{58}\left(\frac{34}{58}\right)}{58}} = 0.0647$
c 1 Identify the conditions on n and $\hat{p}$ for the distribution to be considered approximately normal. **2** Give the appropriate mathematical calculations supporting your reasons.	Use any combination of the following reasons • $\hat{p} \approx 0.5$ and n is sufficiently large • $np = 24 \geq 10$ and $n(1-p) = 34 \geq 10$ • $E(\hat{p}) - 3SD(\hat{p}) = 0.2198 > 0$ and $E(\hat{p}) + 3SD(\hat{p}) = 0.6078 < 1$.
3 State the distribution (i.e. normal) and its corresponding parameters.	By the central limit theorem, the distribution of $\hat{p}$ is approximately normal such that $\hat{p} \sim N(0.4138, 0.0647^2)$.

Exam hack

When $\hat{p}$ gives an answer that is a non-terminating decimal, unless you have been asked to round it to a certain number of decimal places, leave it in the exact fractional form. Regardless, you should use the fractional form or the full decimal value in your calculator to avoid any accuracy errors.

Probability problems involving $\hat{p}$

Worksheet Sample proportion probabilities

If it is appropriate to model the sampling distribution of sample proportions by an **approximate normal distribution**, probability calculations can be carried out using the knowledge of normally distributed random variables from Chapter 8.

Note that some texts encourage the use of a **continuity adjustment/correction** to account for the fact that a binomially distributed discrete random variable is being approximated using a normally distributed continuous random variable. In these cases, a default value of 0.5 is added to or subtracted from the discrete integer bounds to 'correct' the error going from a discrete to continuous random variable.

Strictly speaking, this approach is not an expectation of our course, but instead you may be asked to compare the probability obtained when using the approximate normal distribution to the equivalent probability calculation using the binomial distribution.

WORKED EXAMPLE 9 Using an approximate normal distribution

From a random sample of 40 Year 12 students, 19 were found to have heights greater than 180 cm. Let $\hat{p}$ be the sampling distribution of sample proportions of Year 12 students with a height greater than 180 cm.

a Describe the distribution of $\hat{p}$, justifying your answer.

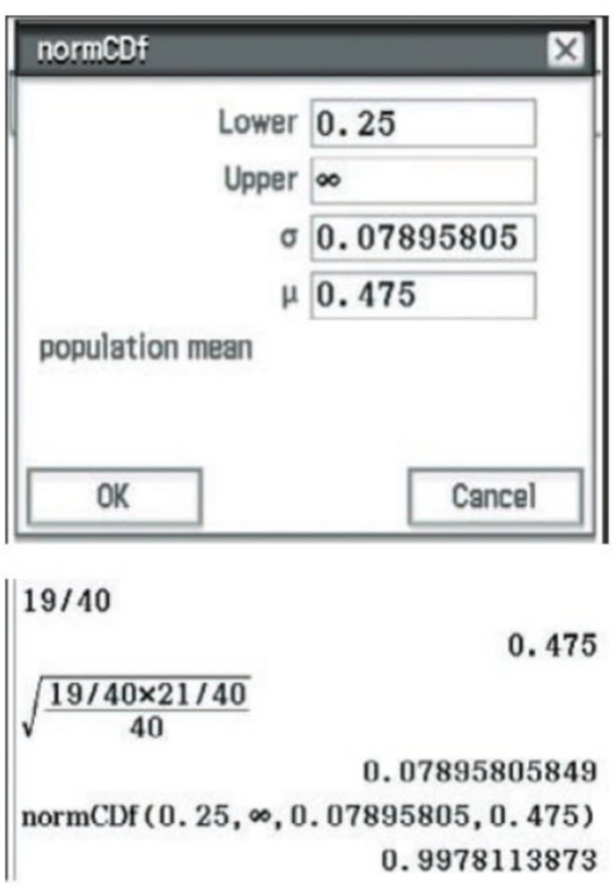
b Hence, determine the probability correct to four decimal places that in a randomly selected sample of 40 Year 12 students

i more than 25% of the students sampled will have a height greater than 180 cm

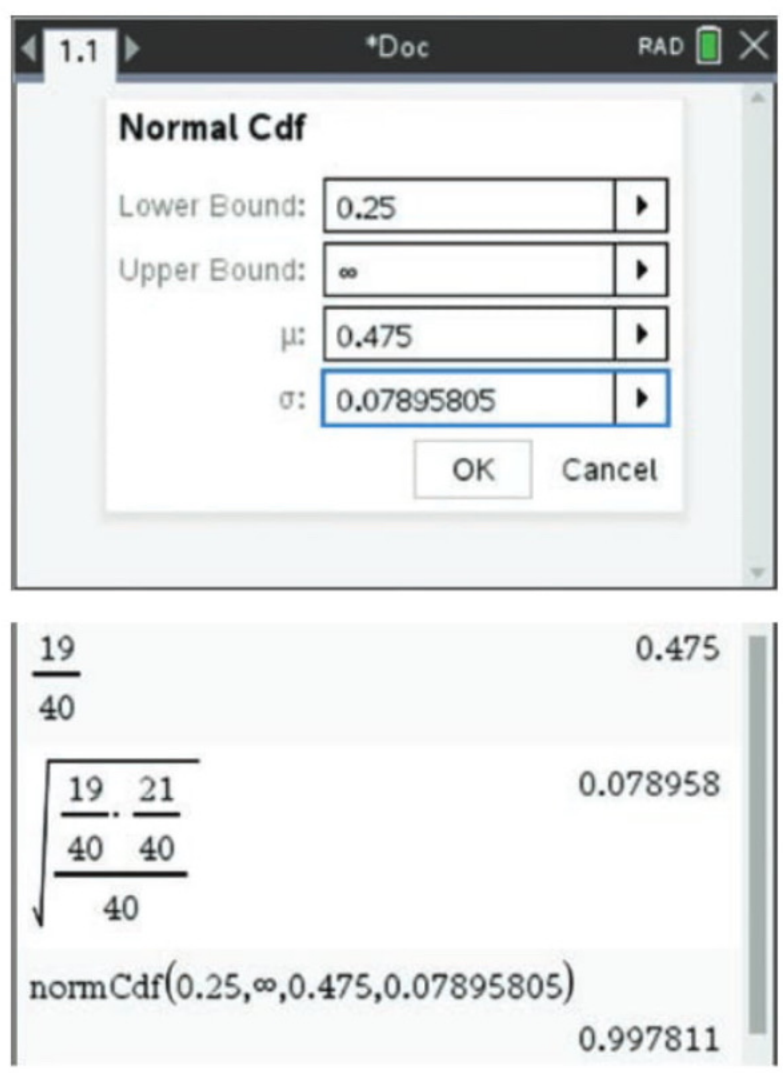
ii between 45% and 65% of the students sampled will have a height greater than 180 cm.

Steps	Working
a 1 Calculate $\hat{p}$ as a point estimate for p.	Let $\hat{p} = \frac{19}{40} = 0.475$ be a point estimate for p.
2 Identify the conditions on n and $\hat{p}$ for the distribution to be considered approximately normal.	Use either of the following reasons: • $\hat{p} \approx 0.5$ and n is sufficiently large • $np = 19 \geq 10$ and $n(1-p) = 21 \geq 10$.
3 Give the appropriate mathematical calculations supporting your reasons.	$\text{SD}(\hat{p}) = \sqrt{\dfrac{\frac{19}{40}\left(\frac{21}{40}\right)}{40}} = 0.0790$
4 State the distribution (i.e. normal) and its corresponding parameters.	By the central limit theorem, the distribution of $\hat{p}$ is approximately normal such that $\hat{p} \sim N(0.475, 0.0790^2)$.
b i 1 Interpret the question as a probability statement.	More than 25% means $\hat{p} > 0.25$.
2 Use CAS to calculate the probability correct to four decimal places.	$\text{P}(\hat{p} > 0.25) = 0.9978$

ClassPad



TI-Nspire



ii 1 Interpret the question as a probability statement.	Between 45% and 65% means $0.45 < \hat{p} < 0.65$.
2 Use CAS to calculate the probability correct to four decimal places.	$P(0.45 < \hat{p} < 0.65) = 0.6109$

ClassPad

```
19/40
                                   0.475
√(19/40×21/40 / 40)
                           0.07895805849
normCDf(0.45,0.65,0.07895805,0.475)
                            0.6109022772
```

TI-Nspire

$$\frac{19}{40} \quad 0.475$$

$$\sqrt{\frac{\frac{19}{40}\cdot\frac{21}{40}}{40}} \quad 0.078958$$

```
normCdf(0.45,0.65,0.475,0.07895805)
                                0.610902
```

When it is not appropriate to model the sampling distribution of sample proportions using an approximate normal distribution, calculations for a binomial distribution should be used.

WORKED EXAMPLE 10 Justifying the sampling distribution of sample proportions

In a sample of 24 Year 12 students, 2 were colour blind.

a State the sample proportion of Year 12 students that are colour blind, $\hat{p}$.

b Explain why it is not appropriate for the sampling distribution of sample proportions to be approximated by a normal distribution.

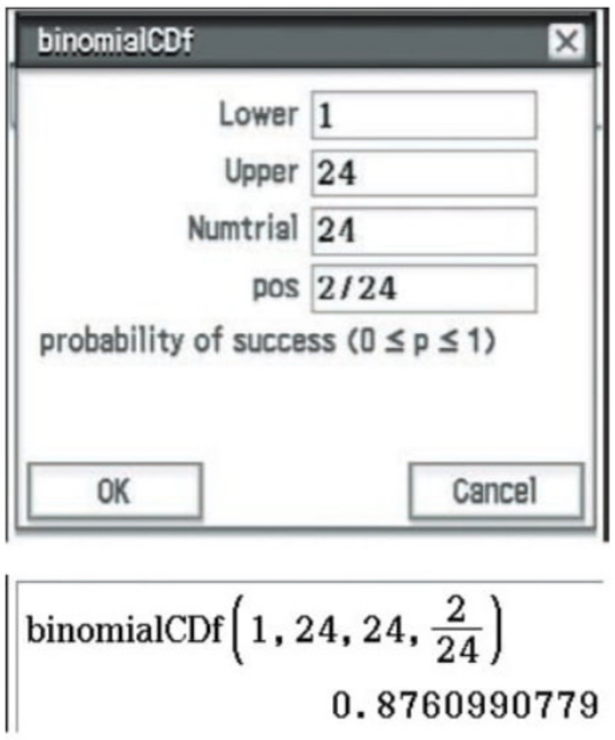
c Hence, estimate the probability that in another sample of 24 Year 12 students, at least 1 student is colour blind.

d Compare the probability in part **c** to the probability if an approximate normal distribution was inappropriately used.

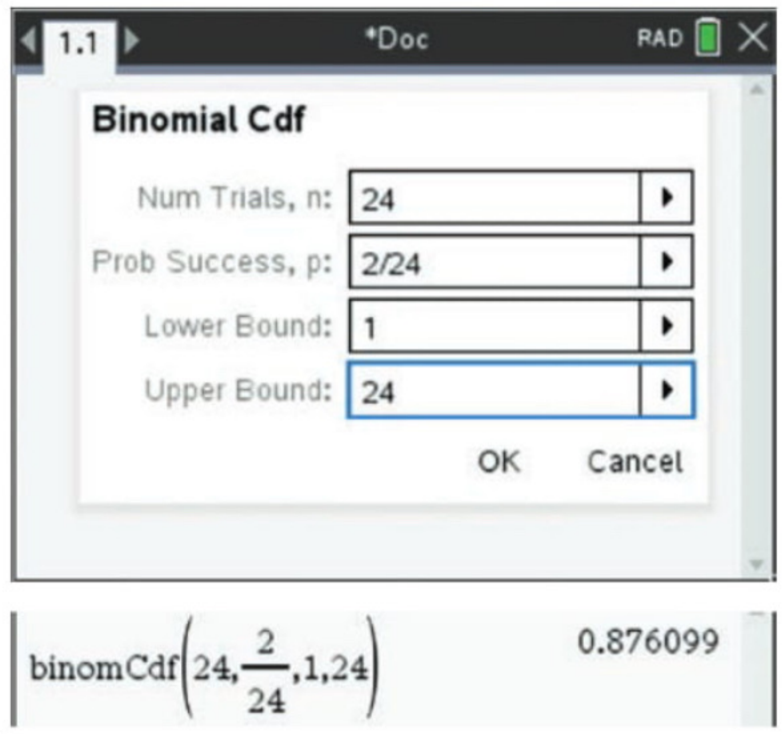
Steps	**Working**
a Express the number of colour blind Year 12 students as a proportion of the sample size.	$\hat{p} = \frac{2}{24} = 0.08\dot{3}$
b Use the values of n and $\hat{p}$ to justify the skewness of the distribution and conclude it is not approximately normal.	Given a very small value of $\hat{p}$ far from 0.5 and an insufficiently large n, the distribution will be positively skewed (i.e. not symmetrical) and, hence, an approximate normal distribution will not be appropriate.
c 1 Define the binomial random variable and its parameters.	Let X be the number of colour blind Year 12 students in a sample of 24. Then $X \sim \text{Bin}\left(24, \frac{2}{24}\right)$.
2 Express the probability using the binomial pdf and complement rule.	$P(x \geq 1) = 1 - P(X = 0)$ $= 1 - \binom{24}{0}\left(\frac{2}{24}\right)^0\left(\frac{22}{24}\right)^{24}$ $= 0.8761$

3 Use CAS to calculate the probability.

ClassPad



TI-Nspire



d 1 Define the approximate normal random variable and its parameters.

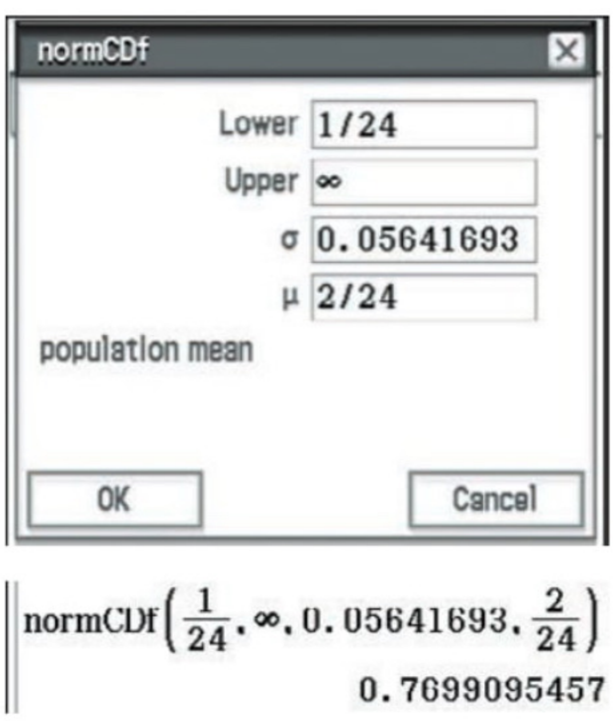
Assume $\hat{p} \sim N\left(\frac{2}{24}, 0.0564^2\right)$.

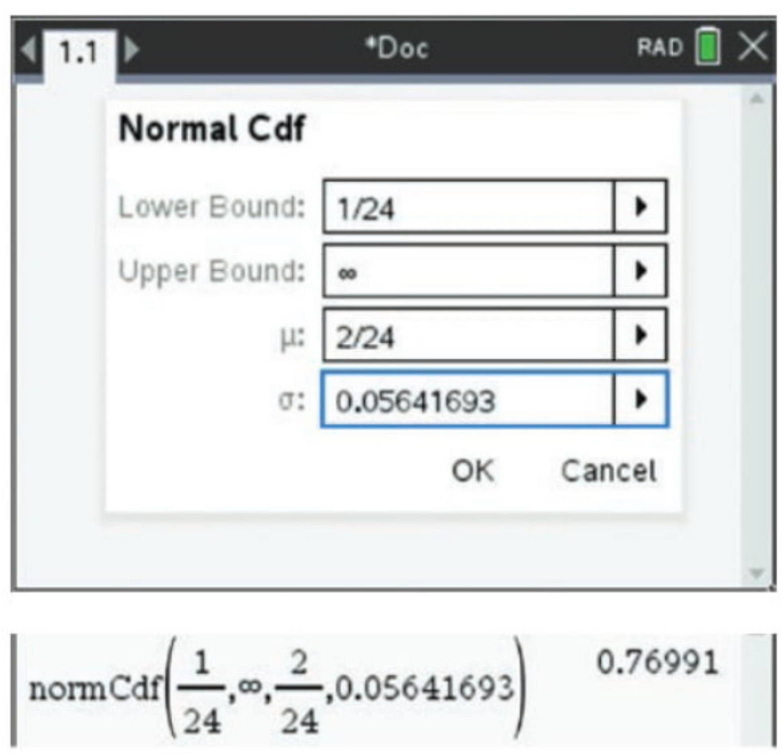
2 Use CAS to calculate the probability.

$\text{P}\left(\hat{p} \geq \frac{1}{24}\right) = 0.7699$

3 Compare the probability from the normal distribution to the probability from the binomial distribution.

The estimate using a non-suitable normal distribution is 0.1062 less than using the binomial distribution.





The standard normal distribution with $\hat{p}$

In some contexts, once an approximate normal distribution has been used to model the sampling distribution of sample proportions, you may be asked to solve problems involving the use of the **approximate standard normal distribution** for the sampling distribution of sample proportions.

The approximate standard normal distribution

For the approximate normal distribution

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

where the value of p is known, then an approximate standard normal distribution $Z \sim N(0,1)$ can be obtained using the linear transformation

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

When the value of p is unknown and a specific sample proportion $\hat{p}_1$ is used as a point estimate for p, then

$$Z = \frac{\hat{p} - \hat{p}_1}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}}$$

WORKED EXAMPLE 11 Finding a z-score using p

Repeated random samples of size 140 are taken from a population with $p = 0.48$ to form the sampling distribution of sample proportions, $\hat{p}$.

a Give one reason why it is appropriate to model $\hat{p}$ using an approximate normal distribution.

b Hence, determine the number of standard deviations that a sample proportion of 0.43 is from the true population proportion.

Steps	Working
a Use the value of n or the value of p to justify the approximate normality of $\hat{p}$.	Given that $p = 0.48 \approx 0.5$ with a sufficiently large $n = 140$, the distribution will be fairly symmetrical and so an approximate normal distribution is appropriate.
b **1** State the expected value and standard deviation of $\hat{p}$.	$\text{E}(\hat{p}) = p = 0.48$
2 Use the standard score formula to find z.	$\text{SD}(\hat{p}) = \sqrt{\frac{0.48(0.52)}{140}}$ $z = \frac{0.43 - 0.48}{\sqrt{\frac{0.48(0.52)}{140}}} = -1.18$
3 Interpret the result as a number of standard deviations from the mean.	So, a sample proportion of 0.43 is 1.18 standard deviations below the population proportion p.

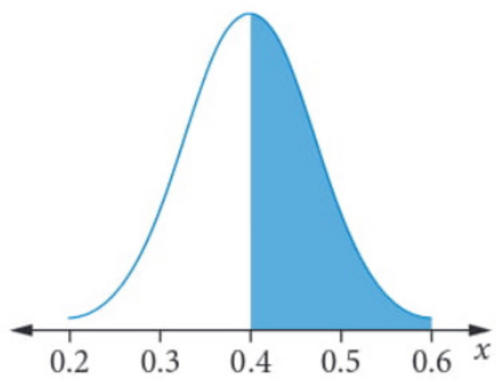
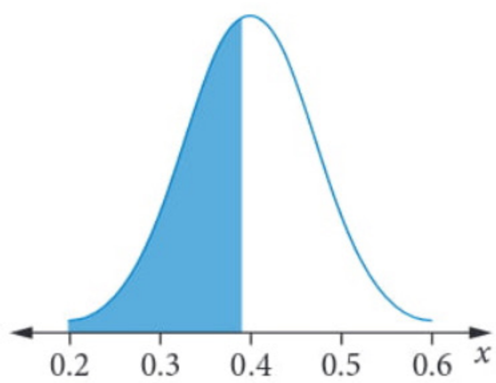
WORKED EXAMPLE 12 Solving for unknowns using $\hat{p}$ and Z

For the approximate normal distribution $\hat{p} \sim N(0.39, 0.002\,379)$, determine the

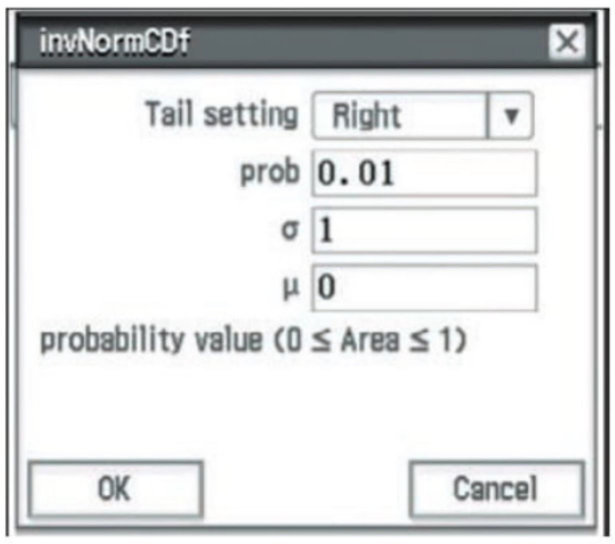
a sample size, n

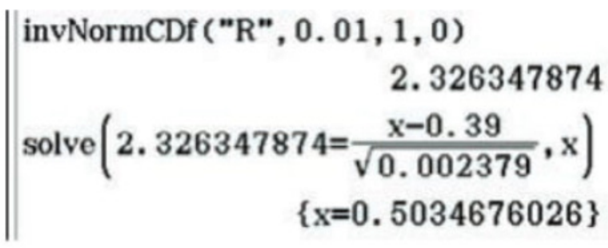
b value of k such that $P(Z < k) = P(\hat{p} > 0.4)$, correct to four decimal places

c value of $\hat{p}$ that corresponds to the value of k such that $P(Z \geq k) = 0.01$, correct to two decimal places.

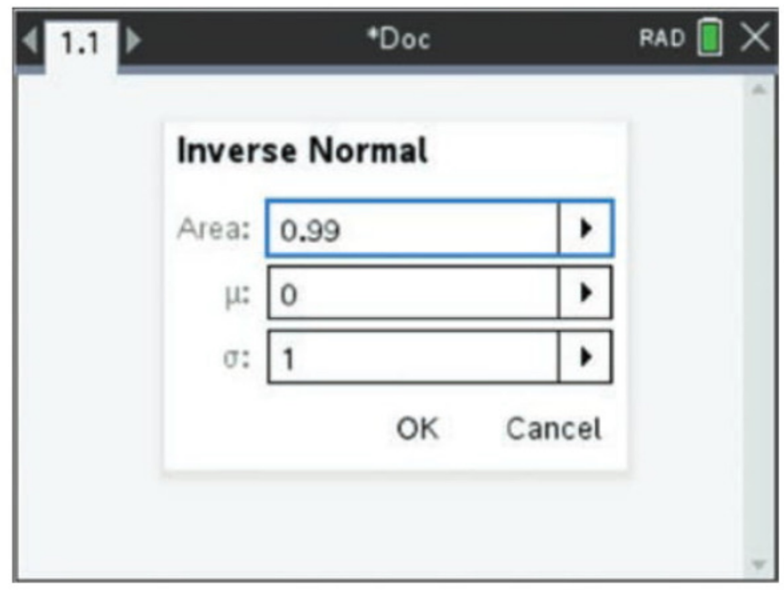
Steps	Working
a 1 Identify the value of $\hat{p}$ used as the point estimate of p.	$\hat{p} = 0.39$
2 Use the formula $\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$ to solve for n.	$0.002\,379 = \frac{0.39(0.61)}{n}$ $n = \frac{0.39(0.61)}{0.002\,379}$ $n = 100$
b 1 Draw a normal curve to represent $P(\hat{p} > 0.4)$.	
2 Use the symmetry about $\hat{p} = 0.39$ to shade the region with equivalent area, $P(\hat{p} < 0.38)$.	
3 Use the standard score formula to find the value of k.	$k = \frac{0.38 - 0.39}{\sqrt{0.002\,379}}$ $= -0.2050$
c 1 Find the z-score, k, corresponding to the 0.99 quantile.	$P(Z \geq k) = 0.01 \Rightarrow k = 2.3263$
2 Use the z-score formula for sample proportions to determine the value of $\hat{p}$.	$z = \frac{\hat{p} - \hat{p}_1}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}}$ $2.3263 = \frac{\hat{p} - 0.39}{\sqrt{0.002\,379}}$ $\hat{p} = 0.50$

ClassPad

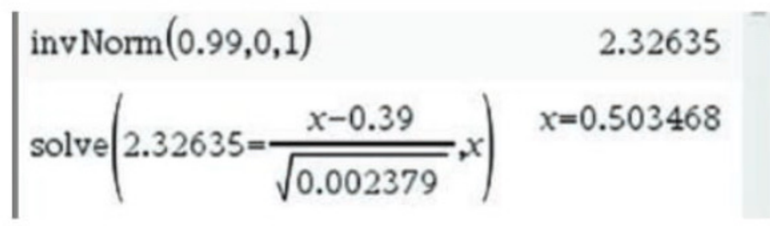




TI-Nspire



NOTE: For the TI-Nspire, be sure to use area = 1 − 0.01 = 0.99.



EXERCISE 9.2 The sampling distribution of sample proportions

ANSWERS p. 409

Recap

1 A media officer for a local football team wanted to collect information about how its supporters travelled to and from matches. Ten different groups of 5 team supporters standing close to each other were surveyed and asked how they travelled to the match.

- **a** Identify the population of interest.
- **b** Identify and explain one source of bias in this sampling.

2 The difference between a statistic and a parameter is that

- **A** a statistic is found from a measurement, but a parameter is a known quantity.
- **B** they measure the same quantity, but a parameter is more reliable than a statistic.
- **C** a statistic is a measure of a sample, while a parameter is a measure of a population.
- **D** a statistic is a measure of a population, while a parameter is a measure of a sample.
- **E** a statistic is found by measuring some aspect of a sample, but a parameter is an approximation of the same measure of the population from which the sample is taken.

Mastery

3 WORKED EXAMPLE 4 A school has a population of 980 high-school students. A random sample of 156 students was taken and it was found that 28 had bought food from the school cafeteria that day.

- **a** Calculate the sample proportion of students who had bought food from the cafeteria.
- **b** Hence, use this sample proportion as a point estimate to estimate the total number of students at this school who buy food from the cafeteria.

4 WORKED EXAMPLE 5 Given that $X \sim \text{Bin}(125, 0.8)$, determine

a $E(\hat{p})$

b $\text{Var}(\hat{p})$, correct to four decimal places

c $SD(\hat{p})$, correct to three decimal places.

5 WORKED EXAMPLE 6 From a sample of 200 people, it was found that 9 had red hair.

a State the sample proportion of people who have red hair, $\hat{p}$.

b Using this value of $\hat{p}$ as a point estimate for the true population proportion of people who have red hair, determine

i $E(\hat{p})$

ii $\text{Var}(\hat{p})$, correct to four decimal places

iii $SD(\hat{p})$, correct to three decimal places.

6 WORKED EXAMPLE 7 A box contains 20 000 marbles that are either blue or red. There are more blue marbles than red marbles. Random samples of 100 marbles are taken from the box. Each random sample is obtained by sampling with replacement. Let $\hat{p}$ be the random variable representing the proportion of blue marbles selected. If the standard deviation of $\hat{p}$ is 0.03, determine the number of blue marbles in the box.

7 Using CAS 5 Simulate 50 different observations from each of the following binomial random variables, graphing each of the results and describing the shape of the distributions.

a $X \sim \text{Bin}(10, 0.4)$ b $X \sim \text{Bin}(10, 0.6)$ c $X \sim \text{Bin}(10, 0.9)$

8 Using CAS 6 Simulate 30 different observations from each of the following binomial random variables, and graph the distributions of each of the corresponding distributions of $\hat{p} = \frac{X}{n}$.

a $X \sim \text{Bin}(100, 0.3)$ b $X \sim \text{Bin}(100, 0.55)$ c $X \sim \text{Bin}(100, 0.7)$

9 WORKED EXAMPLE 8 From a sample of 22 students studying Year 12 Mathematics Methods, 10 students were also studying Chemistry.

a State the sample proportion of Year 12 Mathematics Methods students studying Chemistry, $\hat{p}$, correct to four decimal places.

b Determine $E(\hat{p})$ and $SD(\hat{p})$, correct to four decimal places.

c Describe the distribution of the random variable $\hat{p}$, justifying your answer.

10 WORKED EXAMPLE 9 From a random sample of 500 Australian high-school students, 122 were found to be able to speak a second language. Let $\hat{p}$ be the sampling distribution of sample proportions of high-school students who can speak a second language.

a Describe the distribution of $\hat{p}$, justifying your answer.

b Hence, determine the probability, correct to four decimal places, that in a randomly selected sample of 500 Australian high-school students

i more than 25% of the students sampled can speak a second language

ii between 15% and 25% of the students sampled can speak a second language.

11 WORKED EXAMPLE 10 A stove manufacturer checked the 125 stoves leaving the factory on one day for faults. Eight were found to have faults in the paintwork or other problems that would make them non-sellable items.

a State the sample proportion of non-sellable items, $\hat{p}$.

b Explain why it is not appropriate for the sampling distribution of sample proportions to be approximated by a normal distribution.

c Hence, estimate the probability that in another sample of 125 stoves, more than five are considered non-sellable items.

d Compare the probability in part **c** to the probability if an approximate normal distribution was used.

9.2

12 WORKED EXAMPLE 11 Repeated random samples of size 120 are taken from a population with $p = 0.41$ to form the sampling distribution of sample proportions, $\hat{p}$.

a Give one reason why it is appropriate to model $\hat{p}$ using an approximate normal distribution.

b Hence, determine the number of standard deviations that a sample proportion of 0.5 is from the true population proportion.

13 WORKED EXAMPLE 12 If repeated random sampling from the same population gives sample proportions which have an approximate normal distribution $\hat{p} \sim N(0.56, 0.003\,08)$, determine the

a sample size, n

b value of k such that $P(-k < Z < k) = P(0.5 < \hat{p} < 0.62)$, correct to four decimal places

c value of $\hat{p}$ that corresponds to the value of k such that $P(Z \leq k) = 0.01$, correct to two decimal places.

Calculator-free

14 (4 marks) In a random sample of 75 Australian households, it was found that 12% of them had more than three bedrooms.

a State the number of households in this sample that had more than three bedrooms. (1 mark)

b Give one reason why it may not be considered appropriate to model the sampling distribution of sample proportions of Australian households with more than three bedrooms using an approximate normal distribution. (1 mark)

c Hence, write an expression that can be used to estimate the probability that exactly two houses in a random sample of 10 Australian households have more than three bedrooms. *Do not evaluate this expression.* (2 marks)

15 (3 marks) For random samples of five Australians, let $\hat{p}$ be the random variable that represents the proportion who live in a capital city. Suppose that the value of p, the true population proportion, is known. If $P(\hat{p} = 0) = \frac{1}{243}$, determine the value p.

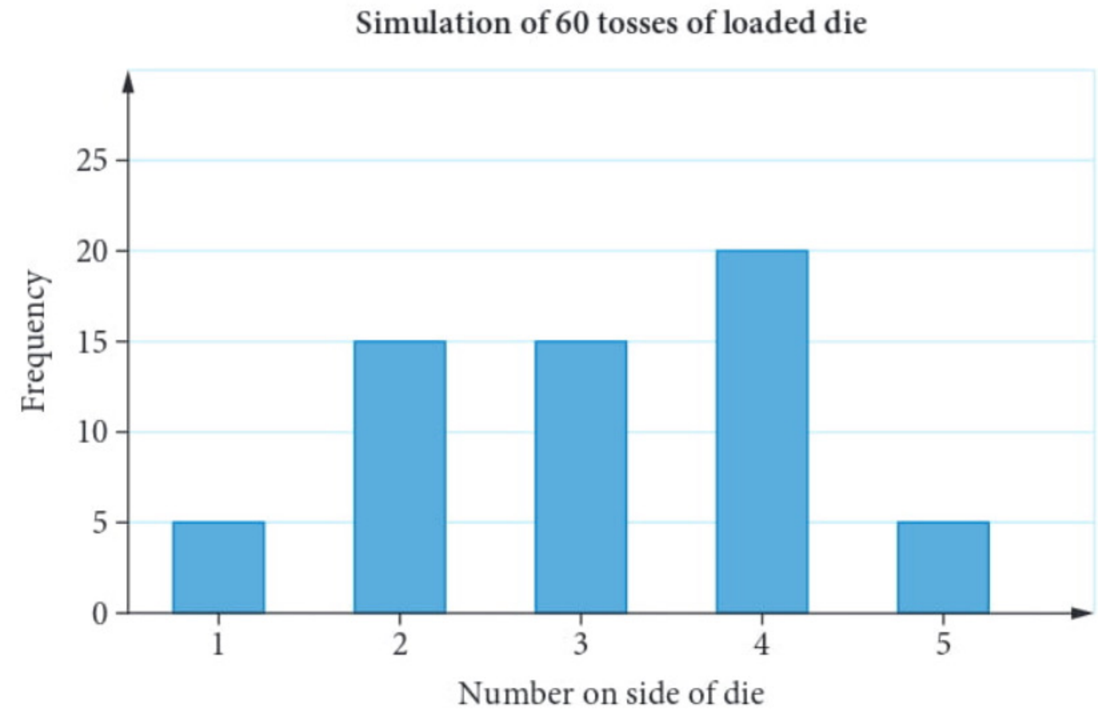
16 (7 marks)

a A random variable $\hat{p}$ representing a sample proportion has a standard deviation of 0.08. If $p = 0.2$, determine the value of n. (3 marks)

b A random variable $\hat{p}$ representing a sample proportion has a standard deviation of 0.04. If $n = 100$ and $p < 0.5$, determine the exact value of p. (4 marks)

Calculator-assumed

17 © SCSA MM2016 Q14abd (6 marks) The simulation of a loaded (unfair) five-sided die rolled 60 times is recorded with the following results.



a Calculate the proportion of prime numbers recorded in this simulation. (2 marks)

b Determine the mean and standard deviation for the sample proportion of prime numbers in 60 tosses, using the results above. (2 marks)

c This simulation of 60 rolls of the die is performed another 200 times, with the proportion of prime numbers recorded each time and graphed. Comment briefly on the key features of this graph. (2 marks)

18 © SCSA MM2017 Q18d MODIFIED (6 marks) Alex is a beekeeper and has noticed that some of the bees are very sleepy. She takes a random sample of 320 bees and finds that 15 of them are indeed so-called *lullabees* that fall asleep easily. It turns out that the true proportion of lullabees is 0.02. Now that Alex knows this, she decides to take a new sample.

a Suppose a new sample of 290 bees was taken. Given that the true proportion of lullabees is 0.02 and assuming an approximate normal distribution for $\hat{p}$, what is the probability that the sample proportion in this new sample is at most 0.03? (3 marks)

b Show one mathematical calculation that suggests an approximate normal distribution for $\hat{p}$ may not be appropriate. (1 mark)

c If Alex takes a larger sample, will the above probability increase or decrease? Explain. (2 marks)

19 © SCSA MM2018 Q17abd (6 marks) Tina believes that approximately 60% of the mangoes she produces on her farm are large. She takes a random sample of 500 mangoes from a day's picking.

a Assuming Tina is correct and 60% of the mangoes her farm produces are large, what is the approximate probability distribution of the sample proportion of large mangoes in her sample? (3 marks)

b What is the probability that the sample proportion of large mangoes is less than 0.58? (2 marks)

A random sample of 500 contains 250 large mangoes.

c On the basis of this data, estimate the proportion of large mangoes produced on the farm. (1 mark)

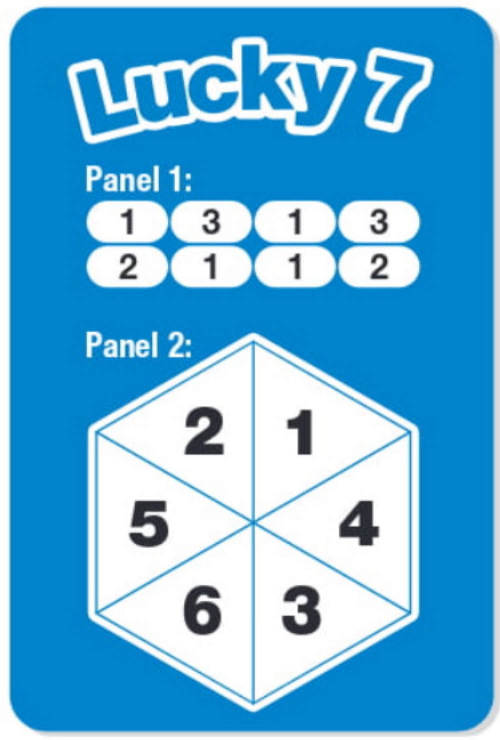
20 © SCSA MM2019 Q13a (4 marks) The proportion of working adults who miss breakfast on weekdays is estimated to be 40%. A study takes a random sample of 400 working adults. For this sample

a what is the (approximate) distribution of the sample proportion of workers who miss breakfast (2 marks)

b what is the probability that the sample proportion of workers who miss breakfast is greater than 44%? (2 marks)

9780170477536

21 © SCSA MM2021 Q10acd MODIFIED (6 marks) A charity organisation has printed 'Lucky 7' scratchie tickets as a fundraiser for use at two special events. The tickets contain two panels. Each ticket has the same numbers as the sample ticket shown below, arranged randomly and hidden within each panel.



A player scratches one section of each panel to reveal a number. The two numbers revealed are then added together. If the total is seven or higher, the player wins a prize.

At the first event, 400 tickets are purchased, and a prize is won on 124 occasions. Let p denote the probability that a prize is won.

a Determine the sample proportion of times that a prize is won at the first event. (1 mark)

It can be shown that the probability p of winning a prize is $\frac{7}{24}$.

b Calculate the mean and standard deviation of the sample proportion of times a prize is won when 400 tickets are purchased. (2 marks)

c At a second event, 400 scratchie tickets are again purchased. If the sample proportion was 0.6 standard deviations from the population proportion, how many prizes were won at the second event? (3 marks)

9.3 Confidence intervals for proportions

Video playlist Confidence intervals for proportions

Worksheets Sample proportion confidence intervals

Margin of error for standard normal variables

Sample sizes

Approximate confidence intervals for p

In Section 9.2, we introduced the idea that a single sample proportion, $\hat{p}$, can act as a point estimate for the true population proportion, p, when p is not known. However, the limitation of a single point estimate does not allow for the possible error in the estimate that arises from the nature of random samples and the variability of samples. As a result, it is common to consider an **interval estimate for p**; that is, a range of possible values that p falls within an interval determined by a sample proportion $\hat{p}$. Because p is unknown in these situations, the true value of $\text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ is also often unknown and so the **standard error** $\text{SD}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is used as a point estimate in its place. The standard error is another term used to describe the standard deviation of the random variable $\hat{p}$ when a single point estimate for p is used.

To calculate an interval estimate for p, first we need to ensure that the distribution of the random variable $\hat{p}$ is approximately normal. Once we have an approximate normal distribution for $\hat{p}$, we know that the mean of $\hat{p}$ is $E(\hat{p}) = \hat{p}$ and the standard error is $SD(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. We can then use a number of standard deviations from the mean, i.e. a z-score, to define a **margin of error**, E, on either side of the point estimate used.

Structure of an interval estimate for p

For an approximately normal sampling distribution of sample proportions formed using $\hat{p}$ as a point estimate for p, an interval estimate for p is of the form

$$\hat{p} - E \leq p \leq \hat{p} + E$$

where E is a margin of error.

The specific type of interval estimate we use in this course is called a confidence interval, or more specifically an **approximate confidence interval**, because the normal distribution being used is only approximate. The term confidence will be explored in more detail soon, but for now let's look at the construction of an approximate confidence interval.

Each approximate confidence interval has a **confidence level**, $100c\%$, where c represents the proportion of $\hat{p}$ values that we want to account for in our estimate. The value of c is used to the calculate the corresponding z-scores such that $P(-z \leq Z \leq z) = c$. For example, if we choose a 95% confidence level, then $P(-z \leq Z \leq z) = 0.95$ gives a value of $z = 1.960$ to three decimal places. Other common confidence levels include

- a 90% confidence level such that $P(-z \leq Z \leq z) = 0.90$ gives a z-score of $z = 1.645$ to three decimal places
- a 99% confidence level such that $P(-z \leq Z \leq z) = 0.99$ gives a z-score of $z = 2.576$ to three decimal places.

The corresponding z-score for any confidence level can be found using the appropriate inverse normal calculation.

Once we have the number of standard deviations z for a given level of c, we can define the value of the margin of error.

Margin of error of an approximate $100c\%$ confidence interval

For a confidence level of $100c\%$ with a corresponding z-score, z, the margin of error of an approximate $100c\%$ confidence interval is given by

$$E = z\,SD(\hat{p}) = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

From this we can define the approximate 100c% confidence interval.

Approximate 100c% confidence interval

For a confidence level of 100c% with a corresponding z-score, z, and a margin of error of $z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$,

an approximate 100c% confidence interval for p is given by

$$\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

90%, 95% and 99% confidence levels are the most commonly used for confidence intervals, with the following standard scores.

Confidence level	90%	95%	99%
Standard score (z-score)	1.645	1.960	2.576

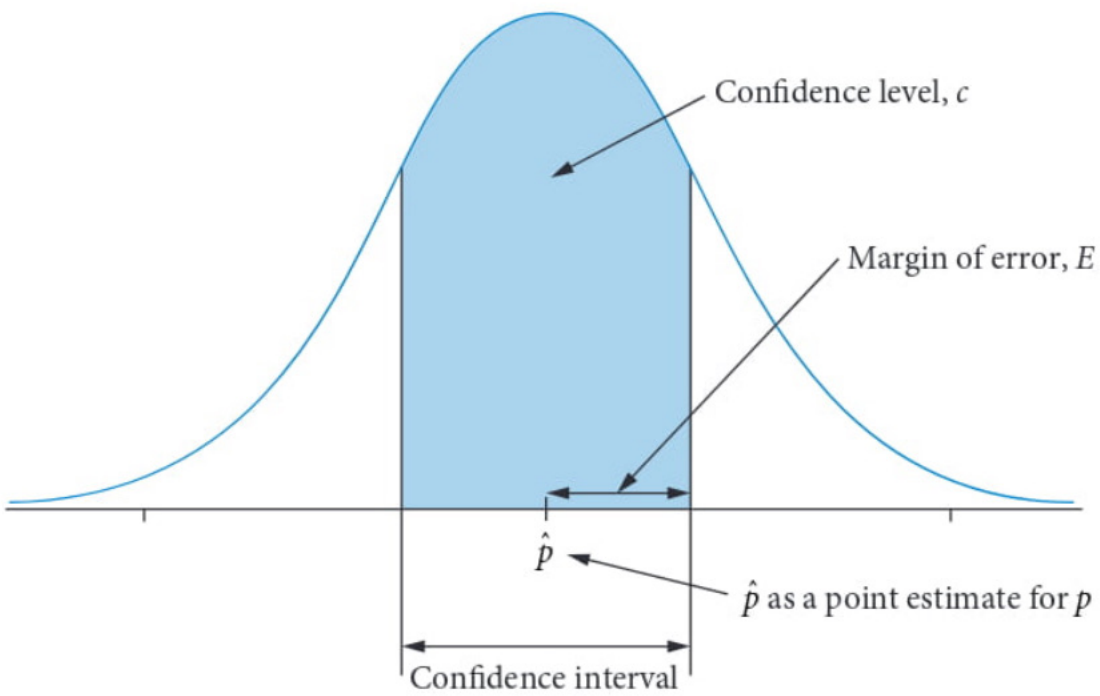
Note that you may also see this written in a different interval notation such as

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

or with open bounds

$$\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$



Exam hack

Always show the line of working involving the construction of the confidence interval with the values of $\hat{p}$, z and n substituted.

Width of an approximate 100c% confidence interval

The width, w, of a 100c% confidence interval is twice the margin of error

$$w = 2E = 2z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Although we still haven't discussed the meaning and interpretation of the word 'confidence' in the context of interval estimates, let's first construct some confidence intervals.

WORKED EXAMPLE 13 Expressing an approximate confidence interval without CAS

A random sample of 40 USBs from a large consignment found that 15 had packaging defects. Let $\hat{p}$ be the random variable for the proportion of USBs with packaging defects in samples of size 40.

a State the distribution of $\hat{p}$. Justify your answer.

b Hence, write expressions for the approximate confidence intervals of p with the following levels of confidence. *Do not evaluate these intervals.*

i 90% confidence level **ii** 95% confidence level **iii** 99% confidence level

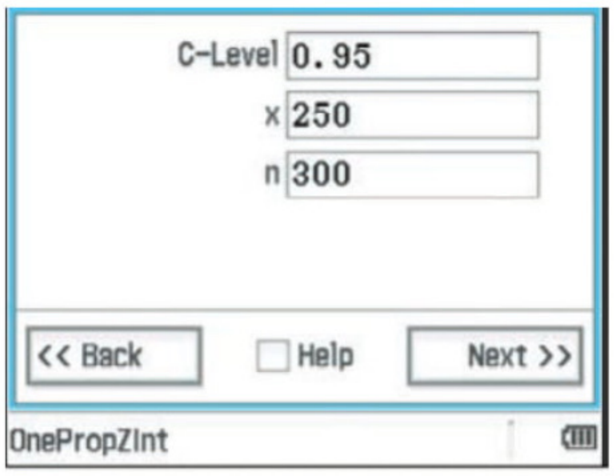
Steps	**Working**
a 1 Calculate $\hat{p}$ as a point estimate for p.	Let $\hat{p} = \frac{15}{40}$ be a point estimate for p.
2 Identify the conditions on n and $\hat{p}$ for the distribution to be considered approximately normal.	Using the fact that $np = 15 \geq 10$ and $n(1-p) = 25 \geq 10$.
3 Give the appropriate mathematical calculations supporting your reasons.	By the central limit theorem, the distribution of $\hat{p}$ is approximately normal such that $\mathrm{E}(\hat{p}) = \frac{15}{40} = \frac{3}{8}$. $\mathrm{SD}(\hat{p}) = \sqrt{\frac{\frac{3}{8}\left(\frac{5}{8}\right)}{40}} = \sqrt{\frac{15}{64 \times 40}} = \frac{\sqrt{15}}{16\sqrt{10}} = \frac{\sqrt{3}}{16\sqrt{2}}$
4 State the distribution (i.e. normal) and its corresponding parameters.	$\hat{p} \sim N\left(\frac{15}{40}, \frac{3}{512}\right)$
b Use the structure of the confidence interval $\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ with each of the appropriate z-scores.	
i $z = 1.645$	$\frac{15}{40} - 1.645\frac{\sqrt{3}}{16\sqrt{2}} \leq p \leq \frac{15}{40} + 1.645\frac{\sqrt{3}}{16\sqrt{2}}$
ii $z = 1.960$	$\frac{15}{40} - 1.960\frac{\sqrt{3}}{16\sqrt{2}} \leq p \leq \frac{15}{40} + 1.960\frac{\sqrt{3}}{16\sqrt{2}}$
iii $z = 2.576$	$\frac{15}{40} - 2.576\frac{\sqrt{3}}{16\sqrt{2}} \leq p \leq \frac{15}{40} + 2.576\frac{\sqrt{3}}{16\sqrt{2}}$

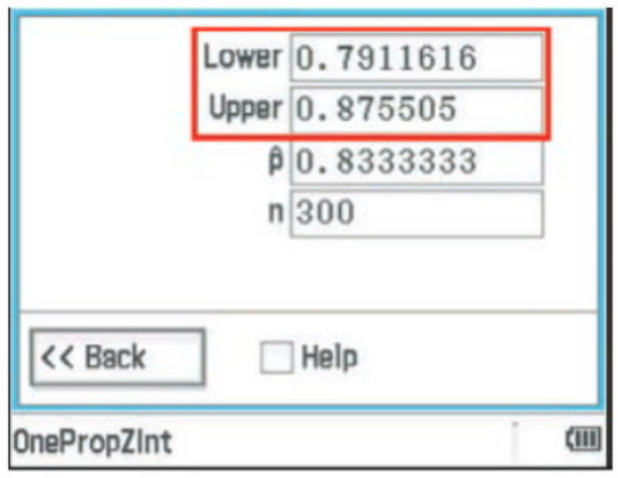
It is likely that you will have the assistance of CAS to evaluate an approximate confidence interval for p.

USING CAS 7 Constructing an approximate confidence interval for p

Three hundred carrot seeds were moistened and placed in an incubator. When they were checked five days later, 250 were found to have germinated. Let $\hat{p}$ be the random variable representing the germination rate of samples of carrot seeds taken from this population. Assuming the approximate normality of $\hat{p}$, determine an approximate 95% confidence interval for the true germination rate of carrot seeds from this population, correct to three decimal places.

ClassPad





1 Open the **Statistics** application.

2 Tap **Calc** > **Interval**.

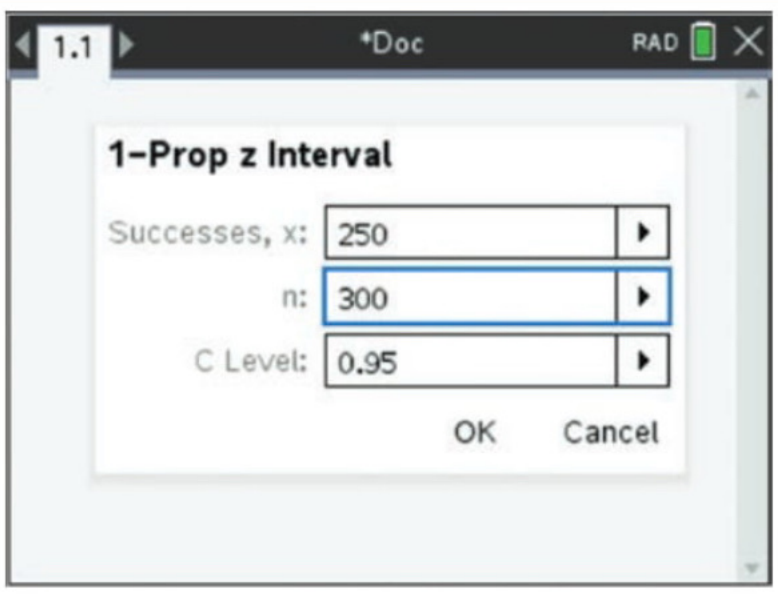
3 In the lower window, select **One-Prop Z Int** from the dropdown menu then tap **Next**.

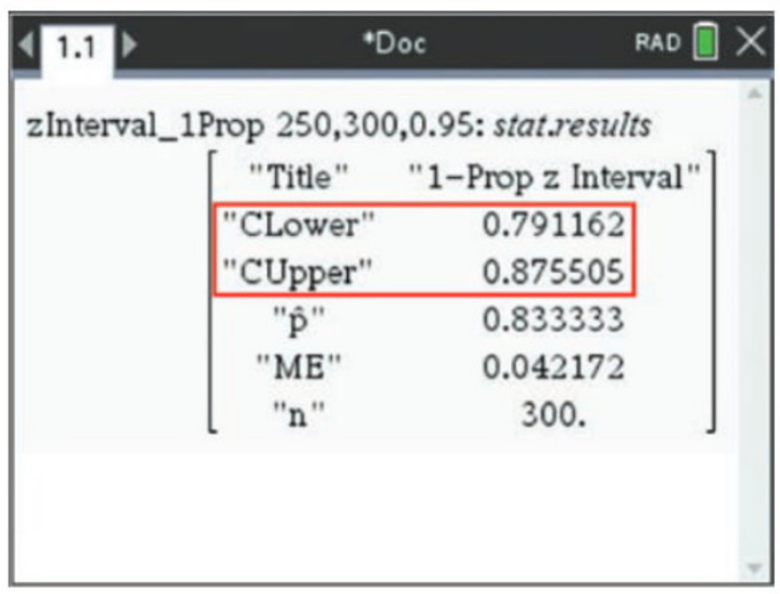
4 Enter the values as shown above then tap **Next**.

5 The values will be displayed in the lower window.

6 The **Lower** and **Upper** values of the confidence interval are highlighted above.

TI-Nspire





1 Press **menu** > **Statistics** > **Confidence Intervals** > **1-Prop z Interval**.

2 Enter the values as shown above.

3 Press **enter**.

4 The *z*-interval table will be displayed.

5 The **CLower** and **CUpper** values of the confidence interval are highlighted above.

The approximate 95% confidence interval for the true germination rate of carrot seeds from this population is $0.791 \leq p \leq 0.876$.

Interpreting confidence intervals and the containment of p

So, here is the big question: *What is the meaning of confidence in the construction of an approximate confidence interval*? The simple answer is that the construction of a single confidence interval serves no purpose other than to act as a single interval estimate for p, but no formal conclusions can be drawn from a single, constructed confidence interval. This is a largely debated concept in statistics, but, for this course, we use the **frequentist interpretation** of a confidence interval.

Frequentist interpretation of confidence intervals

Upon the repeated construction of a large number of approximate $100c\%$ confidence intervals for p, from multiple random samples of sample size n, we can expect (on average) that $100c\%$ of all confidence interval yet to be constructed will contain the true value of p.

The frequentist perspective is about a long-run relative frequency of confidence intervals that are expected to contain the true population proportion.

With this interpretation comes four very important conclusions.

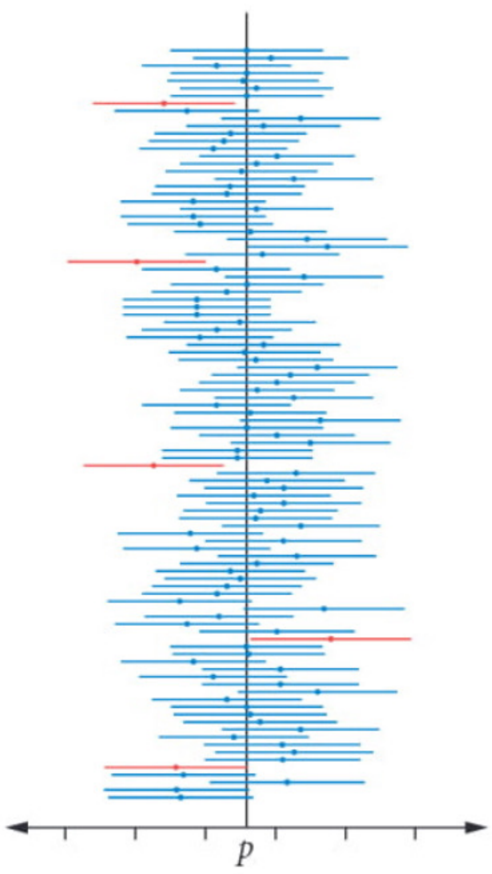
1 Most, but not all, confidence intervals contain p.

2 Because p is unknown, and due to the nature of random sampling, it can never be known for certain whether a confidence interval contains p.

3 Because p is constant, once a confidence interval is constructed, the probability that the given confidence interval contains p is either 0 or 1. It either does not contain p or it does, but we can never know for certain because p is unknown.

4 No single constructed confidence interval is any more or less likely to contain p than any other single constructed confidence interval.

The diagram shows what we could expect to occur if we were to construct 100 different 95% confidence interval from random samples. That is, we would expect (on average) 95 of the 100 to contain the true value of p (shown in blue) and 5 to not contain the true value of p (shown in red). However, this can never be known for certain due to the nature of random sampling but can be shown by simulation.



Exam hack

If you are asked 'which of the following confidence intervals are more likely to contain the true value of p?', it is a trick question! The likelihood that a confidence interval contains p cannot be inferred from a single confidence interval.

WORKED EXAMPLE 14 Discussing containment of p

Boxes of a particular brand of breakfast cereal are labelled with a weight of 485 g. In a random sample of 350 boxes, it is found that 160 were underweight. Let $\hat{p}$ be the random variable representing the sample proportion of underweight boxes of breakfast cereal in samples of size 350.

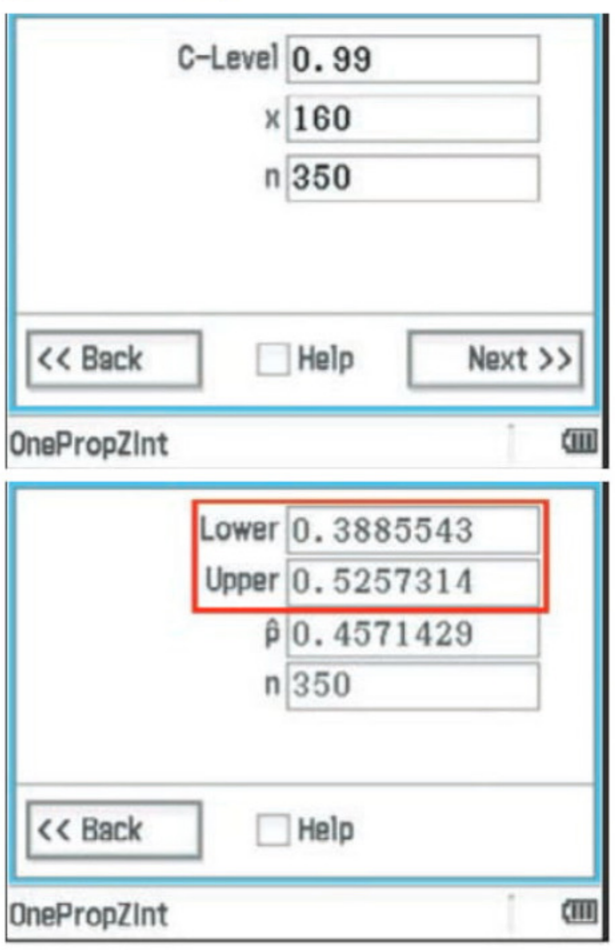
a Assuming the approximate normality of $\hat{p}$, determine an approximate 99% confidence interval for p, correct to three decimal places.

b A second random sample of 350 boxes was taken, the number of underweight boxes was observed and an approximate 99% confidence interval was found to be $(0.372, 0.508)$. Which of the two confidence intervals is more likely to contain the true value of p? Justify your answer.

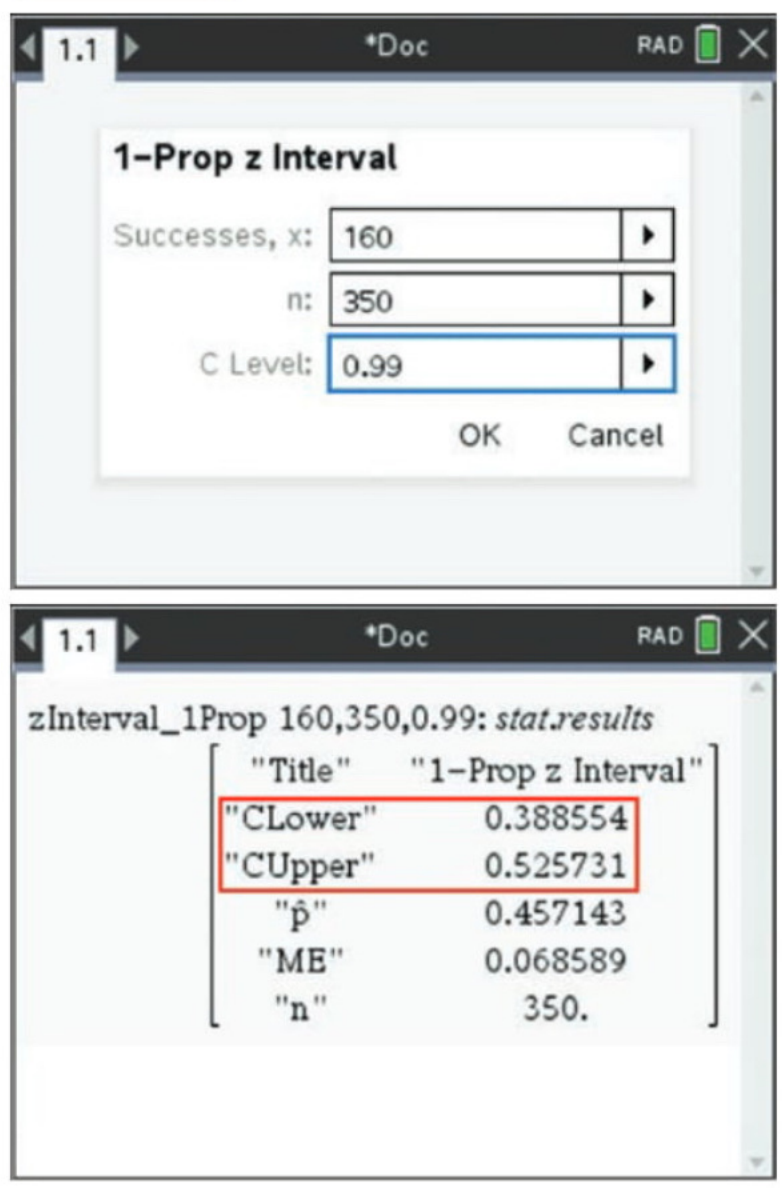
c If a further 500 random samples of 350 boxes were to be taken and approximate 99% confidence intervals were to be constructed for p, how many of the confidence intervals could be expected to contain the true value of p?

Steps	**Working**
a 1 State the distribution of $\hat{p}$.	$\hat{p} \sim N\left(\frac{160}{350}, 0.0266^2\right)$
2 Establish the confidence interval, stating the value of z.	For 99% confidence, $z = 2.576$
3 Use CAS to evaluate the confidence interval.	$\frac{160}{350} - 2.576(0.0266) \le p \le \frac{160}{350} + 2.576(0.0266)$ $0.389 \le p \le 0.526$

ClassPad



TI-Nspire



b Recognise that both confidence intervals have been calculated and, hence, they either do or do not contain p.	Neither one is more likely than the other to contain p, as once observed, the probability that a confidence interval contains p is either 0 or 1. Hence, it cannot be determined.
c 1 Recognise that the 500 confidence intervals are yet to be constructed.	$0.99 \times 500 = 495$
2 Calculate 99% of 500 and estimate the number of confidence intervals expected to contain p.	It can be expected that approximately 495 of the 500 confidence intervals will contain the true value of p.

Another common misconception regarding approximate confidence intervals is that a greater level of confidence means that it is more likely that an observed confidence interval will contain the true value of p. This is not an accurate statement for a single, constructed confidence interval.

Changing confidence levels

If the values of $\hat{p}$ and n remain unchanged, as the confidence level $100c\%$ increases

- the value of z increases, and so
- the value of the margin of error E increases, and so
- the width of the confidence interval increases.

Changing sample size

If the values of $\hat{p}$ and z remain unchanged, as the sample size n increases

- the value of the standard error $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ decreases, and so
- the value of the margin of error E decreases, and so
- the width of the confidence interval decreases.

WORKED EXAMPLE 15 Describing changes to confidence intervals

An approximate 90% confidence interval for p for a random sample of size 400 is found to be $(0.25, 0.33)$. State the effect on the width of the confidence interval constructed

a if the confidence level was increased to 98%, but $\hat{p}$ and n remain unchanged

b if the sample size is reduced to 100, but $\hat{p}$ and z remain unchanged.

Steps	Working
a Describe the sequence of effects if confidence level is increased.	If confidence level is increased, the value of z increases, and so the margin of error increases, increasing the width of the confidence interval.
b Describe the sequence of effects if sample size is reduced.	If sample size is reduced, the standard error increases, and so the margin of error increases, increasing the width of the confidence interval.

Using confidence intervals to calculate unknowns

In some cases, as in Worked example 16, we may be given the lower and upper bounds of a confidence interval. When these bounds are given with sufficient information, the values of $\hat{p}$, E, z, n or $\text{SD}(\hat{p})$ can be calculated.

WORKED EXAMPLE 16 Calculating unknowns given a confidence interval

An approximate 95% confidence interval for p using a random sample of size n is found to be $(0.46, 0.52)$. Calculate the

a sample proportion $\hat{p}$ used to construct the approximate confidence interval

b margin of error of the confidence interval

c standard error of $\hat{p}$ to three decimal places

d sample size used.

Steps	Working
a Use the symmetry of the confidence interval about $\hat{p}$ to calculate the sample proportion.	$\hat{p} = \dfrac{0.46 + 0.52}{2} = 0.49$
b Calculate half of the width of the confidence interval.	$E = \dfrac{0.52 - 0.46}{2} = 0.03$
c Divide the margin of error by the value of z corresponding to the confidence level.	$E = z\,\mathrm{SD}(\hat{p})$ $\mathrm{SD}(\hat{p}) = \dfrac{0.03}{1.960}$ $= 0.015$
d **1** Use the formula $\mathrm{SD}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ to solve for n.	$0.015 = \sqrt{\dfrac{0.49(0.51)}{n}}$ $n = 1066.65$
2 Round to the nearest integer.	$n = 1067$

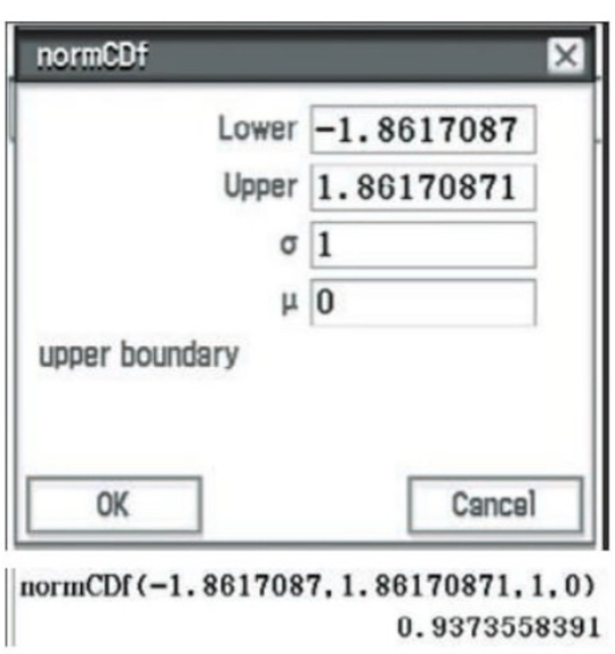
9.3

WORKED EXAMPLE 17 Calculating confidence level given sufficient information

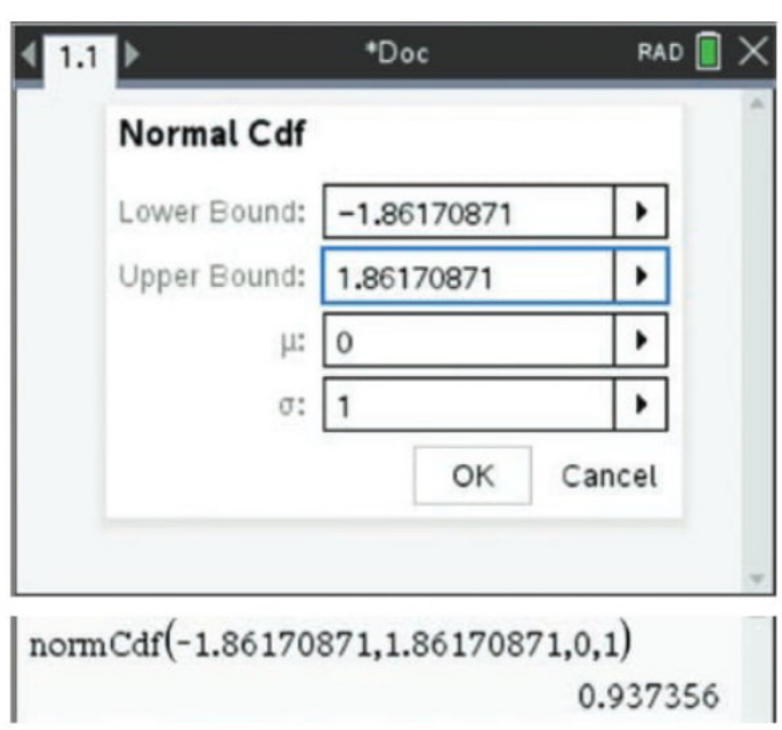
An approximate C% confidence interval for p using a random sample of size 80 is found to be (0.80, 0.94). Find the level of confidence, correct to one decimal place, used to construct the confidence interval.

Steps	Working
1 Find $\hat{p}$ and E.	$\hat{p} = \dfrac{0.80 + 0.94}{2} = 0.87$ $E = \dfrac{0.94 - 0.80}{2} = 0.07$
2 Calculate $\mathrm{SD}(\hat{p})$.	$\mathrm{SD}(\hat{p}) = \sqrt{\dfrac{0.87(0.13)}{80}} = 0.0375\ldots$ $= 0.0376$ to three decimal places
3 Find z using $E = z\,\mathrm{SD}(\hat{p})$.	$z = \dfrac{E}{\mathrm{SD}(\hat{p})} = \dfrac{0.07}{0.0375\ldots} = 1.8617\ldots$
4 Use CAS to find $\mathrm{P}(-1.8617 \le Z \le 1.8617)$.	$\mathrm{P}(-1.8617 \le Z \le 1.8617) = 0.937$
5 State the confidence level as a percentage, rounded to one decimal place.	Therefore, a confidence level of 93.7% was used.

ClassPad



TI-Nspire



Another typical calculation involving confidence intervals is to determine a minimum sample size required to ensure a certain margin of error is obtained in the construction of an approximate confidence interval. However, the issue with these problems is that the sample proportion is not yet known and so the calculation for n cannot be carried out unless there is a given value of $\hat{p}$. It is common for this value to come from historical data and so we can use it as a point estimate for p.

WORKED EXAMPLE 18 Determining a minimum sample size given a point estimate for p

A previous study suggests that about 60% of Year 12 students obtain their driver's licence before they complete Year 12. Determine the minimum sample size that would be needed to obtain an approximate 90% confidence interval for p with a maximum width of 14%.

Steps	Working
1 Calculate E given the width and state the other known values.	$E = \frac{0.14}{2} = 0.07$ $z = 1.645$ $\hat{p} = 0.6$
2 Use the formula $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to solve for n.	$0.07 = 1.645\sqrt{\frac{0.6(0.4)}{n}}$ $n = 132.54$
3 If n is a decimal value, note that the integer smaller than n will give a width larger than 14%. The integer value larger than n will give a width smaller than 14%. Round up!	The minimum sample size to ensure a maximum width of 0.14 is 133 Year 12 students.

If, in a similar problem, the value of $\hat{p}$ is not known, then we assume the worst case scenario and adopt the value of $\hat{p}$ that gives the maximum margin of error.

It can be shown using calculus techniques that the value of $\hat{p}$ that maximises the function $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is 0.5.

$$E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Given that z and n are just constants, then we can simply maximise the function:

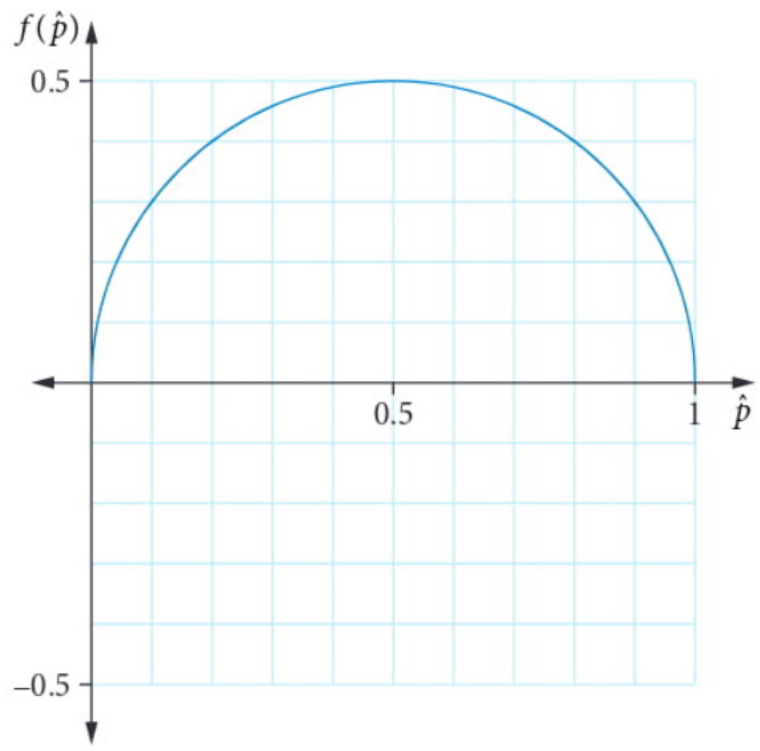
$$\begin{aligned}
f(\hat{p}) &= \sqrt{\hat{p}(1-\hat{p})} = \sqrt{\hat{p}-\hat{p}^2} \\
f'(\hat{p}) &= \frac{1}{2}(\hat{p}(1-\hat{p}))^{-\frac{1}{2}} \cdot (1-2\hat{p}) \\
0 &= \frac{(1-2\hat{p})}{2\sqrt{\hat{p}(1-\hat{p})}} \\
1-2\hat{p} &= 0 \\
\hat{p} &= \frac{1}{2}
\end{aligned}$$

By observing the graph of $f(\hat{p})$ or considering the second derivative, $f''(\hat{p}) = -\frac{\sqrt{\hat{p}(1-\hat{p})}}{4\hat{p}^2(\hat{p}-1)^2}$, we can show that it is indeed a maximum.

$$f''\left(\frac{1}{2}\right) = -2$$

9.3

So, f is concave down at $\hat{p} = \frac{1}{2}$ and so gives a maximum turning point.



WORKED EXAMPLE 19 Determining a minimum sample size when $\hat{p}$ is unknown

A survey wants to establish an interval estimate for the proportion of Year 12 students from single-parent families to within 3% at a confidence level of 99%. Determine the minimum number of Year 12 students who need to be surveyed.

Steps	Working
1 Assume $\hat{p} = 0.5$ and state the known values of E and z.	$\hat{p} = 0.5$ $E = 0.03$ $z = 2.576$
2 Use the formula $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to solve for n.	$0.03 = 2.576\sqrt{\frac{0.5^2}{n}}$ $n = 1843.27$
3 If n is a decimal value, round up to nearest integer value.	At least 1844 Year 12 students should be surveyed.

Population claims, historical data and the comparison of samples

Further to the construction of single approximate confidence intervals and solving for unknown values given sufficient information, there is not much more we can do with confidence intervals without the formal topic of hypothesis testing (which is not in the Year 12 Mathematics Methods course). However, there is a common tendency in WACE exam questions to ask you to infer results from confidence intervals in three different situations, the first two of which are explained below.

1. Compare an approximate confidence interval from a sample to a claimed value of p and suggest whether there is sufficient evidence to suggest that the sample appears to come from the same population.
2. Compare an approximate confidence interval from a sample to a claimed value of p or one that arises from historical data and suggest whether the claim or data is supported by the sample.

Comparing a claimed value of p to an approximate confidence interval for p

If the claimed value of p lies within an approximate confidence interval obtained from a random sample, then

- there is insufficient evidence to suggest the sample came from a different population, or
- there is insufficient evidence to suggest that the claimed value of p shouldn't be accepted.

If the claimed value of p lies outside of an approximate confidence interval obtained from a random sample, then

- there is insufficient evidence to suggest the sample came from the same population, or
- there is insufficient evidence to suggest that the claimed value of p can be accepted based on this sample.

Exam hack

Your responses to these questions should always be written in the context of the question.

WORKED EXAMPLE 20 Comparing a confidence interval to a claimed p

A local fashion designer claims that approximately 70% of Perth residents wear dark-coloured clothing to work. Over the course of a week, a random sample of 120 Perth residents going to work was taken and it was observed that 70 of them were wearing dark-coloured clothing. Use an approximate 95% confidence interval to comment on the fashion designer's claim.

Steps	Working
1 Obtain the distribution of $\hat{p}$.	$\hat{p} \sim N\left(\frac{70}{120}, 0.045^2\right)$
2 Construct a 95% confidence interval using CAS.	$\frac{70}{120} - 1.960(0.045) \le p \le \frac{70}{120} + 1.960(0.045)$
3 Observe the location of the claimed p in relation to the confidence interval.	$0.495 \le p \le 0.672$
4 Conclude appropriately in context of the question.	Given that the claimed $p \approx 0.70$ lies outside of the interval estimate $0.495 \le p \le 0.672$, based on this sample, there is insufficient evidence to suggest that the fashion designer's claim can be accepted.

The third case that may arise is the following:

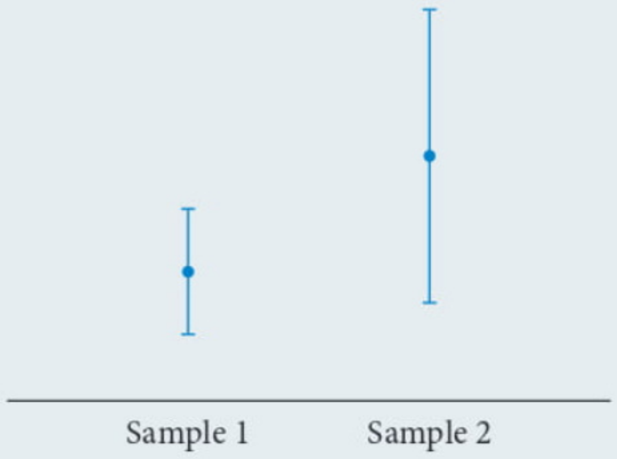
3 Compare approximate confidence intervals from two or more samples after changed conditions and suggest whether the changed conditions have affected the sample proportions.

9.3

Comparing two or more samples

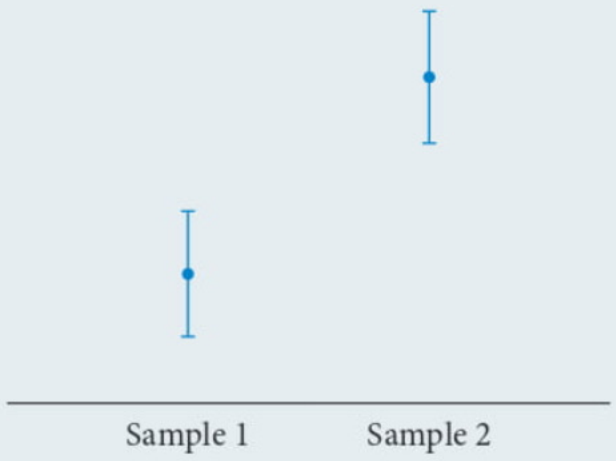
If a second sample proportion $\hat{p}_2$ lies within an approximate confidence interval obtained using a first sample proportion $\hat{p}_1$, OR if the two approximate confidence intervals overlap such that one or both sample proportions are contained within the other, then

- there is insufficient evidence to suggest that the samples came from different populations, or that a changed condition had an impact on the population.



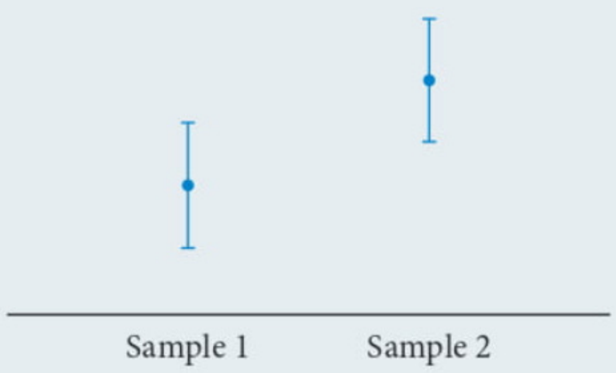
If the approximate confidence interval obtained from $\hat{p}_2$ does not overlap at all with an approximate confidence interval obtained using $\hat{p}_1$, then

- there is sufficient evidence to suggest that the samples may have come from different populations, or that a changed condition may have an impact on the population, but we cannot say for certain.



If the approximate confidence interval obtained from $\hat{p}_2$ partially overlaps an approximate confidence interval obtained using $\hat{p}_1$, such that neither value of $\hat{p}$ is contained within the other, then

- there is insufficient evidence to conclude anything definitive about the two samples.



WORKED EXAMPLE 21 Comparing two samples

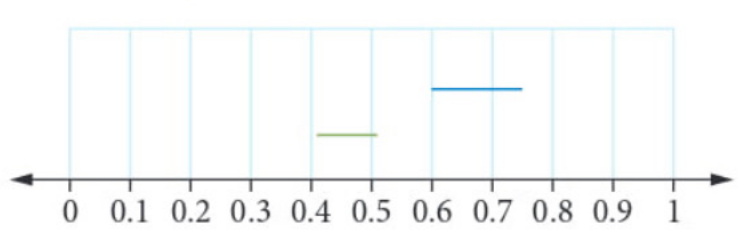
One Saturday, two weeks before a local election, a random sample of 320 people were asked about their voting intentions. It was found that in a two-party preferential vote between Liberal and Labor, 147 indicated they would vote Labor.

a Construct an approximate 90% confidence interval for the true proportion of Labor voters in the electorate, correct to four decimal places.

In the following week, Labor intensified their advertising campaign and in a poll on the following Saturday, it was found that 71 out of 105 people surveyed said they would vote Labor.

b How likely is it to obtain a sample proportion greater than that found in the second random sample when using the sampling distribution of sample proportions of the first sample?

c Perform the necessary calculations to comment on whether the increased advertising campaign improved Labor's polling results.

Steps	Working
a 1 Obtain the distribution of $\hat{p}$ for the first sample.	$\hat{p} \sim N\left(\frac{147}{320}, 0.027\,86^2\right)$
2 Construct a 90% confidence interval using CAS.	$\frac{147}{320} - 1.645(0.028) \le p \le \frac{147}{320} + 1.645(0.028)$ $0.4136 \le p \le 0.5052$ **Exam hack** If you are rounding values on your page, be sure to use the full decimal values in CAS.
b 1 Interpret the question as a probability statement.	For $\hat{p} \sim N\left(\frac{147}{320}, 0.027\,86^2\right)$
2 Use CAS to calculate the probability.	$P\left(\hat{p} > \frac{71}{105}\right) = 3.548 \times 10^{-15}$
3 Comment on the likelihood.	It is extremely unlikely to obtain a sample proportion greater than $\frac{71}{105}$ in the first sampling distribution, as the probability is close to 0.
c 1 Obtain the distribution of $\hat{p}$ for the second sample.	$\hat{p} \sim N\left(\frac{71}{105}, 0.045\,67^2\right)$
2 Construct a 90% confidence interval using CAS.	$\frac{71}{105} - 1.645(0.046) \le p \le \frac{71}{105} + 1.645(0.046)$ $0.6011 \le p \le 0.7513$
3 Draw a diagram representing the intervals and observe the location of the second confidence interval in relation to the first confidence interval.	 The two confidence intervals do not overlap.
4 Conclude appropriately in context of the question.	There is sufficient evidence to suggest that the increased advertising campaign may have improved Labor's polling result, but we cannot say for certain.

WACE QUESTION ANALYSIS

 MM2020 Q12 Calculator-assumed (21 marks)

It is estimated that 20% of small businesses fail in the first year. A business advisory group takes a random sample of 500 new businesses that started in January 2018. An analyst employed by the group suggests the use of the binomial distribution is appropriate in this case.

a What is the probability that at most 120 of the businesses fail in the first year? (2 marks)

b What is the approximate distribution of the sample proportion of small businesses that fail by the end of the year in this sample? Justify your answer. (3 marks)

c What is the probability that the sample proportion of businesses that fail by the end of the year is less than 0.18? (2 marks)

d By January 2019, 90 of the 500 new businesses had failed. Calculate a 95% confidence interval for the proportion of new businesses that fail in the first year. (2 marks)

The business advisory group believes that the proportion of new businesses that fail within a year can be reduced by providing financial advice. They took another random sample of 500 businesses that started in January 2019 and provided them with regular financial advice. In this random sample, at the end of the year 80 businesses had failed.

e Calculate the sample proportion and its margin of error at the 95% confidence level. (2 marks)

f Calculate a 95% confidence interval for the proportion of businesses that failed. What do you conclude regarding the value of the financial advice provided to the new businesses? (4 marks)

g If the sample size was reduced, what would be the effect on the confidence interval? Justify your answer. (2 marks)

h State **two** assumptions that the analyst made in recommending the use of the binomial model in this case and discuss whether they are valid. (4 marks)

Reading the question

- Make note of the given values of n, p and $\hat{p}$ throughout the question.
- Highlight when you are told to use a specific distribution, e.g. binomial distribution.
- Highlight the key high-order command words, e.g. *justify, conclude, discuss.*

Thinking about the question

- Make sure you know when to use the normal distribution rather than the binomial distribution.
- Recall the possible situations and conclusions when comparing confidence intervals for two samples after changed circumstances.

Worked solution (✓ = 1 mark)

a Let the random variable X denote the number of new businesses that fail out of the 500.

$X \sim \text{Bin}(500, 0.2)$

$\text{P}(X \le 120) = 0.9877$

defines a binomial random variable and its parameters ✓

calculates the probability ✓

b Given that n is sufficiently large and $np = 100 \ge 10$ and $n(1-p) = 400 \ge 10$, then $\hat{p}$ is approximately normal, such that $\hat{p} \sim N(0.2, 0.0179^2)$.

uses sample size and p to justify approximate normality ✓

states the mean of the distribution ✓

states the variance of the distribution ✓

9.3

Video WACE question analysis: Interval estimates for proportions

c **$P(\hat{p} < 0.18) = 0.1318$ ✓**

uses the approximate normal distribution to calculate $P(\hat{p} < 0.18)$ ✓

d A new sample proportion is defined and so this must be used in the construction of the confidence interval for p.

$$\hat{p} = \frac{90}{500} = 0.18$$

$z = 1.960$

$$0.18 - 1.960\sqrt{\frac{0.18(0.82)}{500}} \le p \le 0.18 + 1.960\sqrt{\frac{0.18(0.82)}{500}}$$

$0.1463 \le p \le 0.2137$

shows the construction of the confidence interval ✓

obtains the correct bounds of the confidence interval ✓

e A new sample proportion is defined and so this must be used in the construction of the confidence interval for p.

$$\hat{p} = \frac{80}{500} = 0.16$$

$z = 1.960$ ✓

$$\boldsymbol{E = 1.960\sqrt{\frac{0.16(0.84)}{500}} = 0.0321} \ ✓$$

f $0.16 - 0.0321 \le p \le 0.16 + 0.0321$

$0.1279 \le p \le 0.1921$

Comparing $0.1463 \le p \le 0.2137$ and $0.1279 \le p \le 0.1921$, $\hat{p}_1 \in [0.1279, 0.1921]$ and $\hat{p}_2 \in [0.1463, 0.2137]$. Given the overlap of the confidence intervals such that the sample proportions are mutually contained, there isn't sufficient evidence to suggest that the financial advice has reduced the proportional of businesses that fail in the first year.

shows the construction of the second confidence interval ✓

obtains the correct bounds of the second confidence interval ✓

notes the location of the sample proportions in relation to the overlapping confidence intervals ✓

concludes correctly in context of the question ✓

g As n decreases, standard error increases and so margin of error increases and, hence, the width of the confidence interval increases.

describes the effect on the standard error and margin of error ✓

describes the effect on the width of the confidence interval ✓

h The underlying assumptions of a binomial distribution are as follows:

- The success or failure of each business is independent of the success or failure of any other business. This is unlikely to be valid as similar businesses may both fail or both survive, or be affected by the competition within the market of other similar businesses.
- The probability of a business failing in the first year is a fixed, constant value of p. This is unlikely to be valid, as different types of business are expected to have different probabilities of failure, possibly at different times in the year.

states assumption that a binomial distribution involves independent trials in context ✓

explains why this assumption is invalid ✓

states assumption that each trial has a fixed probability of success in context ✓

explains why this assumption is invalid ✓

Exam hack

- When dealing with a large question with multiple marks, check to see whether successive parts of the question rely on earlier answers or whether they can be answered independently.
- Answer to four decimal places to be safe, even if not asked to.
- Always state the distribution and its parameters the first time it is being used.
- Be clear and specific with explanations and justifications and involve the language of the mathematical concepts.

9.3

EXERCISE 9.3 Confidence intervals for proportions

ANSWERS p. 410

Recap

1 Which of the following binomially distributed random variables would give the best approximation of a normal distribution?

A $X \sim \text{Bin}(20, 0.8)$ **B** $X \sim \text{Bin}(15, 0.6)$ **C** $X \sim \text{Bin}(40, 0.25)$

D $X \sim \text{Bin}(40, 0.5)$ **E** $X \sim \text{Bin}(90, 0.01)$

2 Samples of bread rolls from bakeries around Western Australia were weighed. The samples all had the same number of rolls. The mean of the sampling distribution of sample proportions of underweight rolls was found to be about 0.09. The standard deviation of the sampling distribution of sample proportions was found to be about 0.03.

The sample size was approximately

A 3 **B** 9 **C** 10 **D** 90 **E** 100

Mastery

3 WORKED EXAMPLE 13 A random sample of 50 T-shirts from a large consignment found that 25 had embroidery defects. Let $\hat{p}$ be the random variable for the proportion of T-shirts with embroidery defects in samples of size 50.

a State the distribution of $\hat{p}$. Justify your answer.

b Hence, write expressions for the approximate confidence intervals of p with the following levels of confidence. *Do not evaluate these intervals.*

i 90% confidence level

ii 95% confidence level

iii 99% confidence level

4 Using CAS 7 Of 140 randomly sampled songs played on a radio station during 'drive time' over several weeks, it was found that 95 of them were less than 3 minutes long. Assuming the approximate normality of $\hat{p}$, determine an approximate 90% confidence interval for the true proportion of songs during 'drive time' on this radio station that are less than 3 minutes long, correct to three decimal places.

5 WORKED EXAMPLE 14 Blocks of a particular brand of chocolate are labelled with a weight of 250 g. In a random sample of 300 blocks, it is found that 110 were underweight. Let $\hat{p}$ be the random variable representing the sample proportion of underweight blocks of chocolate in samples of size 300.

a Assuming the approximate normality of $\hat{p}$, determine an approximate 95% confidence interval for p, correct to three decimal places.

b A second random sample of 300 blocks was taken, the number of underweight blocks was observed and an approximate 95% confidence interval was found to be $(0.27, 0.37)$. Which of the two confidence intervals is more likely to contain the true value of p? Justify your answer.

c If a further 240 random samples of 300 blocks were to be taken and approximate 95% confidence intervals were to be constructed for p, how many of the confidence intervals could be expected to contain the true value of p?

6 WORKED EXAMPLE 15 An approximate 99% confidence interval for p for a random sample of size 200 is found to be $(0.09, 0.22)$. State the effect on the width of the confidence interval constructed

a if the confidence level was decreased to 97%, but $\hat{p}$ and n remain unchanged

b if the sample size is increased to 300, but $\hat{p}$ and z remain unchanged.

7 WORKED EXAMPLE 16 An approximate 90% confidence interval for p using a random sample of size n is found to be $(0.408, 0.558)$. Calculate the

a sample proportion $\hat{p}$ used to construct the approximate confidence interval

b margin of error of the confidence interval

c standard error of $\hat{p}$ to three decimal places

d sample size used.

8 WORKED EXAMPLE 17 An approximate C% confidence interval for p using a random sample of size 25 is found to be $(0.148, 0.652)$. Find the level of confidence, correct to the nearest percentage, used to construct the confidence interval.

9 WORKED EXAMPLE 18 A previous study suggests that about 28% of Year 12 students work a part-time job during their final year of schooling. Determine the minimum sample size that would be needed to obtain an approximate 95% confidence interval for p with a maximum width of 10%.

10 WORKED EXAMPLE 19

a An advertising company wants to conduct a small survey of consumers to establish a baseline for the proportion of consumers who were aware of a particular brand of ice cream before a marketing campaign. Determine the minimum number of consumers that need to be surveyed to obtain an estimate for p accurate to within 3% at a confidence level of 95%.

b An insurance company wants to conduct a small survey of customers to gain insight into the proportion of customers who have fire and theft cover for their car. Determine the minimum number of customers that need to be surveyed to obtain an estimate for p accurate to within 10% at a confidence level of 90%.

11 WORKED EXAMPLE 20

a A local Perth tour guide claims that approximately 40% of tourists who enquire at the Perth CBD Information Centre ask about attractions in the Perth CBD. Over the course of a week, it was recorded that in a random sample of 75 tourists who enquired at the Information Centre, 36 of them enquired about attractions in the Perth CBD. Use an approximate 95% confidence interval to comment on the tour guide's claim.

b Osborne Park piano tuners north of the Swan River estimate that the proportion of pianos requiring new strings when they are tuned is about 0.30. A piano tuner based in Victoria Park south of the Swan River checked their records and found that of 120 pianos, 50 needed new strings. Use an approximate 95% confidence interval to comment on whether piano tuning demands north and south of the Swan River are the same.

12 WORKED EXAMPLE 21 In a survey of 30 randomly selected Western Australian government officials, it was found that 40% of officials were in favour of Western Australia adopting a new state flag.

a Construct an approximate 95% confidence interval for the true proportion of government officials in favour of adopting a new state flag, correct to four decimal places.

In a follow-up survey, a few newly proposed designs for the flag were included and it was found that 12 out of 50 people surveyed said they were in favour of adopting a new state flag.

b How likely is it to obtain a sample proportion less than that found in the second random sample when using the sampling distribution of sample proportions of the first sample?

c A marketing officer claimed that the newly proposed designs were not better than the current state flag. Perform the necessary calculations to comment on the marketing officer's claim.

9.3

13 Jacinta tosses a coin five times. Albin suspects that the coin Jacinta tossed is not actually a fair coin and he tosses it 18 times. Albin observes a total of 12 heads from the 18 tosses. Based on this sample, construct the approximate 90% confidence interval for the probability of observing a head when this coin is tossed. Use the z value of 1.645. *Do not evaluate the bounds of the interval estimate.*

Calculator-free

14 © SCSA MM2017 Q4 (3 marks) Two independent samples of different sizes were taken from a population. The first sample had sample size n_1 and the second sample had sample size n_2. The sample proportions of males in the samples were the same. When 99% confidence intervals were calculated for each sample, it was found that the corresponding margin of error in the second sample was half that of the first sample.

What is the ratio of the two sample sizes, $\frac{n_2}{n_1}$?

15 © SCSA MM2018 Q5 (3 marks) A 95% confidence interval for a population proportion based on a sample size of 200 has width w. What sample size is required to obtain a 95% confidence interval of width $\frac{w}{3}$?

Calculator-assumed

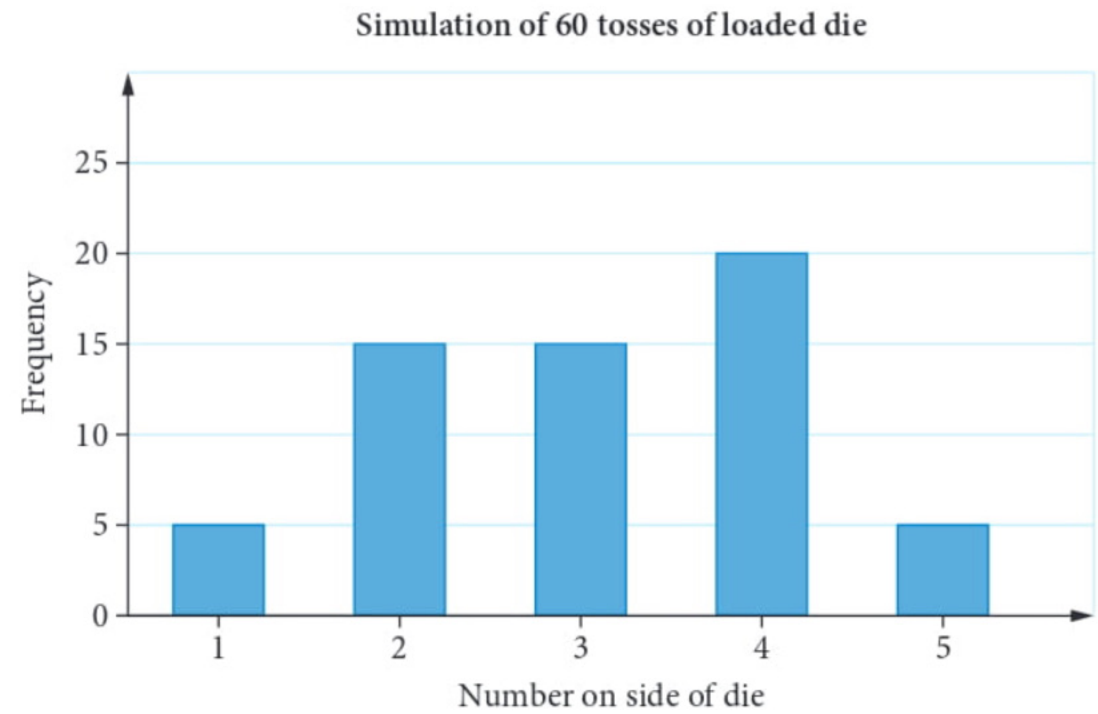
16 (4 marks) A laptop supplier collects a sample of 100 laptops that have been used for six months from a number of different schools and tests their battery life. The laptop supplier wishes to estimate the proportion of such laptops with a battery life of less than three hours. The laptop supplier finds that, in a particular sample of 100 laptops, six of them have a battery life of less than three hours.

a Determine a 95% confidence interval for the supplier's estimate of the proportion of interest. Give values correct to three decimal places. (3 marks)

b Give one reason as to why the confidence interval in part **a** may not be considered reliable. (1 mark)

17 (3 marks) An opinion pollster reported that for a random sample of 574 voters in a town, 436 indicated a preference for retaining the current council. Determine an approximate 90% confidence interval for the proportion of the total voting population with a preference for retaining the current council, correct to three decimal places.

18 © SCSA MM2016 Q14c (3 marks) The simulation of a loaded (unfair) five-sided die rolled 60 times is recorded with the following results.



It has been decided to create a confidence interval for the proportion of prime numbers in this simulation. The level of confidence will be chosen from 90% or 95%.

Explain which level of confidence will give the smallest margin of error. State this margin of error.

19 © SCSA MM2017 Q18abc (6 marks) Alex is a beekeeper and has noticed that some of the bees are very sleepy. She takes a random sample of 320 bees and finds that 15 of them are indeed so-called *lullabees* that fall asleep easily.

a Calculate the sample proportion of lullabees. (1 mark)

b Determine a 90% confidence interval for the true proportion of lullabees, rounded to four decimal places. (3 marks)

c What is the margin of error in the above estimate? (2 marks)

20 © SCSA MM2018 Q13 (10 marks) The proportion of caravans on the road being towed by vehicles that have the incorrect towing capacity is p.

a Show, using calculus, that to maximise the margin of error a value of $\hat{p} = 0.5$ should be used. Note: As z and n are constants, the standard error formula can be reduced to $E = \sqrt{\hat{p}(1-\hat{p})}$. (3 marks)

b A consulting firm wants to determine p within 8% with 99% confidence. How many towing vehicles should be tested at a random check? (3 marks)

c Six months later, the consulting firm carries out a random sampling of towing vehicles. A 99% confidence interval calculated for the proportion of vehicles with incorrect towing capacity is $(0.342, 0.558)$. Determine the number of vehicles in the sample that have an incorrect towing capacity. (4 marks)

21 © SCSA MM2019 Q14 (7 marks)

a What is the minimum sample size required to estimate a population proportion to within 0.01 with 95% confidence? (3 marks)

b Identify **two** factors that affect the width of a confidence interval for a population proportion and describe the effect of each. (4 marks)

9780170477536

22 © SCSA MM2020 Q14ab MODIFIED (6 marks) A suburban council hires a consultant to estimate the proportion of residents of the suburb who use its library.

a The consultant decides to estimate a 95% confidence interval for the proportion to within an error of 0.03. What minimum sample size should be selected? (3 marks)

b If resource limitations dictate that the maximum sample size that can be managed is 500, what is the maximum margin of error in estimating a 99% confidence interval? (3 marks)

23 © SCSA MM2018 Q17efg (6 marks) Tina believes that approximately 60% of the mangoes she produces on her farm are large. She takes a random sample of 500 mangoes from a day's picking and finds that it contains 250 large mangoes.

a Calculate a 95% confidence interval for the proportion of large mangoes produced on the farm, rounded to four decimal places. (3 marks)

b On the basis of your calculations, how would you respond to Tina's belief that the proportion of large mangoes produced is at least 60%? Justify your response. (2 marks)

c What can Tina do to further test her belief? (1 mark)

24 (7 marks) A company supplies schools with whiteboard pens. As a whiteboard pen ages, its tip may dry to the point where the whiteboard pen becomes defective (unusable). The company has stock that is two years old and, at that age, company historical data suggests that 6% of Grade A whiteboard pens will be defective. A box of 100 Grade A whiteboard pens that is two years old is selected and it is found that 10 of the whiteboard pens are defective.

a Determine an approximate 99% confidence interval for the population proportion from this sample, correct to four decimal places. (3 marks)

b Determine an approximate 90% confidence interval for the population proportion from this sample, correct to four decimal places. (3 marks)

c Using the two confidence intervals constructed, comment on whether it appears that the company's historical data is still relevant for their current stock. (1 mark)

25 (4 marks) Rusty's Robotics manufactures sensor components for robots. Prior company data suggests that approximately 5% of all the sensors manufactured are defective. A random sample of 500 sensors is selected and it is found that 40 sensors in this sample are defective.

a Determine an approximate 95% confidence interval for the proportion of defective sensors, correct to four decimal places. (3 marks)

b Comment on whether it appears that the company's historical data is still relevant for their current manufacturing quality. (1 mark)

26 (7 marks) A local entertainment reviewer claims in a newspaper article that approximately 4% of concerts start more than 15 minutes after the scheduled starting time. For the purposes of customer satisfaction, the owners of the local Mathsland Concert Hall decide to review their operation and study the information from 1000 concerts conducted at their venue, collected as a simple random sample. The sample value for the number of concerts that start more than 15 minutes after the scheduled starting time is found to be 43.

a Describe the sampling distribution of sample proportions, $\hat{p}$, for the proportion of interest. Justify your answer. (3 marks)

b Find an approximate 95% confidence interval for the proportion of concerts that begin more than 15 minutes after the scheduled starting time. Give values correct to three decimal places. (2 marks)

c The owners of the Mathsland Concert Hall claim that the reviewer must have visited their concert hall before writing the review. Comment on the validity of such a claim. (2 marks)

27 © SCSA MM2019 Q8 (7 marks) Big Foods is a large supermarket company. The manager of Big Foods wants to estimate the proportion of households that do the majority of their grocery shopping in their stores. A junior staff member at Big Foods conducted a survey of 250 randomly-selected households and found that 56 did the majority of their grocery shopping at a Big Foods store.

a Calculate the sample proportion of households who did the majority of their grocery shopping at Big Foods. (1 mark)

b Determine the 95% confidence interval for the proportion of households who do the majority of their grocery shopping at Big Foods. Give your answer to four decimal places. (3 marks)

c What is the margin of error of the 95% confidence interval? Give your answer to four decimal places. (1 mark)

An independent research company conducted a large-scale survey of household supermarket preferences and estimated that the true proportion of households that conduct most of their grocery shopping at Big Foods was 0.17 (assume that this is indeed the true proportion).

d With reference to your answer to part **b**, does this result suggest that the junior staff member at Big Foods made a mistake? (2 marks)

28 © SCSA MM2021 Q13defg (8 marks) A carnival game involves five buckets, each containing 5 blue balls and 15 red balls. A player blindly selects a ball from each bucket and wins the game if they select at least 4 blue balls. Let X denote the number of blue balls selected. An observer records the outcome of 100 consecutive games and determines the 90% and 95% confidence intervals for the proportion of wins, p. The confidence intervals are $(0.04, 0.16)$ and $(0.05, 0.15)$.

a Which of these intervals is the 95% confidence interval for p? Justify your answer. (2 marks)

b How many wins were observed out of the 100 games? (2 marks)

c Determine what you would expect to happen to the width of the confidence intervals if 400 games had been observed. (2 marks)

d The true proportion of wins does not lie within either of the above confidence intervals. Does this suggest that a sampling error was made? Justify your answer. (2 marks)

29 © SCSA MM2016 Q10 MODIFIED (11 marks) A survey in Western Australia was conducted on the popularity of a calculator known as Type A. Out of 1450 Year 12 students, the survey found that 986 students used the Type A calculator.

9.3

a Determine an approximate 90% confidence interval, to three decimal places, for the proportion of Western Australian Year 12 students who use the Type A calculator, stating any necessary assumptions. (3 marks)

b State the margin of error in this confidence interval. (1 mark)

Another three surveys of Year 12 students were conducted on the use of Type A calculators across Australia.

Survey 2	Survey 3	Survey 4
Type A usage 1772 out of 3221 Year 12 students	Type A usage 1021 out of 1566 Year 12 students	Type A usage 2203 out of 3221 Year 12 students

c Determine approximate 90% confidence intervals for Surveys 2 to 4. (3 marks)

d A data analyst claims that Survey 2 is likely to have been taken outside of Western Australia. Comment on the validity of the analyst's claim. (2 marks)

e Using the sample proportion of Survey 1, determine a sample size that will halve the margin of error for the proportion of Western Australian Year 12 students who use the Type A calculator, with a confidence of 90%. (2 marks)

30 © SCSA MM2016 Q20 MODIFIED (12 marks) A chocolate factory produces chocolates with the machines calibrated such that 80% of chocolates produced are pink. Each box of chocolates contains exactly 30 pieces.

a Identify the probability distribution of X: the number of pink chocolates in a single box. Give the mean and standard deviation of X. (3 marks)

b Determine the probability, to four decimal places, that there are at least 27 pink chocolates in a randomly selected box. (2 marks)

Quality Control collects a sample of 20 boxes of chocolates and finds that 450 chocolates are pink.

c By first stating an appropriate distribution for the sample proportion of pink chocolates, determine an approximate 95% confidence interval for the proportion of pink chocolates in a sample of 20 boxes. (4 marks)

d Quality Control claims that there may be an error with the calibration of the machine. Comment on the validity of the claim. (1 mark)

To check the calibration, Quality Control collects a further three samples and determines an approximate 95% confidence interval each time. It is found that all three contain the value of 0.8.

e Use this finding to account for the results in parts **c** and **d**. (2 marks)

9 Chapter summary

Random sampling

- A **census** collects data from an entire **population** and is used to calculate **population parameters** of a certain characteristic, for example, population mean and standard deviation.
- A **survey** collects data from a **sample** group of a population and is used to calculate **sample statistics** of a certain characteristic, for example, sample mean and standard deviation.
- A sample is **fair and representative** if
  - the sample size is sufficiently large enough to represent the population
  - the data is free from biases that could affect the reliability of it being used to estimate population parameters.
- An unfair and non-representative sample is called a **biased sample**.
- A **probability sampling method** is a data collection process whereby each member of the population has an equally likely chance of being randomly selected. These methods often minimise bias.
- A **non-probability sampling method** is a data collection process whereby each member of the population does not have an equally likely chance of being randomly selected. These methods may introduce different biases.
- Due to the nature of random sampling, there exists **variability in random samples** such that
  - the sample statistics and shape of the distribution will vary, but will approximate the parameters and shape of the population distribution
  - as $n \to \infty$, the mean of a sample will generally tend towards $E(X)$, but can still vary, and the shape of the distribution will better represent the shape of the distribution of X.

The sampling distribution of sample proportions

- For a single sample of size n, the **sample proportion** is $\hat{p} = \dfrac{\text{number of observed successes}}{n}$.
- As a random variable, $\hat{p} = \dfrac{X}{n}$, where $X \sim \text{Bin}(n, p)$ such that n is the sample size and p is the probability of success (i.e. true population proportion).
- The distribution of all possible $\hat{p}$ values has
  - $\text{E}(\hat{p}) = p$
  - $\text{Var}(\hat{p}) = \dfrac{p(1-p)}{n}$
  - $\text{SD}(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}}$
- As $n \to \infty$, $\text{Var}(\hat{p}) \to 0$ and $\text{SD}(\hat{p}) \to 0$, meaning there is very little variation in the different values of $\hat{p}$ taken from different samples of a significantly large, fixed size n.
- For a binomially distributed random variable $X \sim \text{Bin}(n, p)$,
  - if $p \approx 0.5$, with a sufficiently large n (e.g. $n \geq 30$) or
  - n is sufficiently large such that $np \geq 10$ and $n(1 - \text{p}) \geq 10$,

  then X can be modelled by an **approximate normal distribution** of the form:

  $X_N \sim N(np, np(1-p))$.

- If X is approximately normal, then by the **central limit theorem** $\hat{p} = \frac{X_N}{n}$ is approximately normal such that:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- When a **point estimate** is used to estimate p, then

$$\hat{p} \sim N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

- An **approximate standard normal distribution**, $Z \sim N(0, 1)$, can be obtained using the linear transformation
  - $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ when p is known
  - $Z = \dfrac{\hat{p} - \hat{p}_1}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n}}}$ when the value of p is unknown and a specific sample proportion $\hat{p}_1$ is used as a point estimate for p.

Confidence intervals

- An **interval estimate** for p is of the form $\hat{p} - E \le p \le \hat{p} + E$, where E is a **margin of error**.
- An **approximate 100c% confidence interval** for p has the margin of error, $E = z\,\mathrm{SD}(\hat{p}) = z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, where $\mathrm{SD}(\hat{p})$ is called the **standard error**.
- An approximate 100c% confidence interval for p with a corresponding z-score, z, and a margin of error of $E = z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, can be given by any of the following notations:

$$\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- 90%, 95% and 99% confidence levels are the most commonly used for confidence intervals, with the following standard scores.

Confidence level	90%	95%	99%
Standard score (z-score)	1.645	1.960	2.576

- The width, w, of a 100c% confidence interval is twice the margin of error

$$w = 2E = 2z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The frequentist interpretation of confidence intervals says: *Upon the repeated construction of a large number of approximate* $100c\%$ *confidence intervals for* p*, from multiple random samples of sample size* n*, we can expect (on average) that* $100c\%$ *of all confidence interval yet to be constructed will contain the true value of* p*.*
  - Most, but not all, confidence intervals contain p.
  - Because p is unknown and due to the nature of random sampling, it can never be known for certain whether a confidence interval contains p.
  - Because p is constant, once a confidence interval is constructed, the probability that the given confidence interval contains p is either 0 or 1. It either does not contain p or it does, but we can never know for certain because p is unknown.
  - No single constructed confidence interval is any more or less likely to contain p than any other single constructed confidence interval.
- If the values of $\hat{p}$ and n remain unchanged, as the confidence level $100c\%$ increases
  - the value of z increases, and so
  - the value of the margin of error E increases, and so
  - the width of the confidence interval increases.
- If the values of $\hat{p}$ and z remain unchanged, as the sample size n increases
  - the value of the standard error $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ decreases, and so
  - the value of the margin of error E decreases, and so
  - the width of the confidence interval decreases.
- If $\hat{p}$ is unknown, then $\hat{p} = 0.5$ produces the largest margin of error in a confidence interval constructed with a given sample size n.
- Approximate confidence intervals for p constructed from samples can be used to comment on
  - the validity of claimed values of p or a value of p from historical data by observing the location of p with respect to $[\hat{p} - E, \hat{p} + E]$
  - the effect of changed circumstances between samples by observing the location of $[\hat{p}_2 - E, \hat{p}_2 + E]$ in relation to $[\hat{p}_1 - E, \hat{p}_1 + E]$.

Cumulative examination: Calculator-free

Total number of marks: 28 Reading time: 2 minutes Working time: 28 minutes

1 (4 marks) A binomial random variable has mean 20 and variance 4.

a Write two equations in terms of n and p. (2 marks)

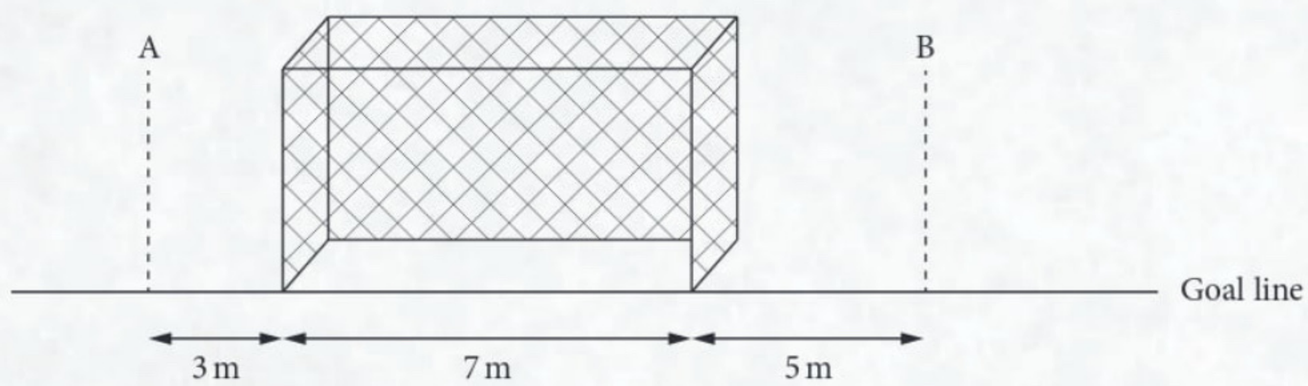
b Find the values of n and p. (2 marks)

2 (5 marks)

a Find $\frac{dy}{dx}$ if $y = x^3 \ln(3x)$. (2 marks)

b Hence find $\int x^2 \ln(3x)\,dx$. (3 marks)

3 © SCSA MM2017 Q2 (6 marks) Michelle is a soccer goalkeeper and has built a machine to help her practise. The machine will shoot a soccer ball randomly along the ground at or near a goal that is seven metres wide. The machine is equally likely to shoot the ball so that the centre of the ball crosses the goal line anywhere between point A three metres left of the goal, and point B five metres right of the goal, as shown in the diagram below.



Michelle sets up a trial run without anyone in the goals. Assume the goal posts are of negligible width.

Let the random variable X be the distance the centre of the ball crosses the goal line to the right of point A.

a Copy and complete the graphical representation of the probability density function for the random variable X. (2 marks)

b What is the probability that the machine shoots a ball so that its centre misses the goal to the left? (1 mark)

c What is the probability that the machine shoots a ball so that its centre is inside the goal? (1 mark)

d If the machine shoots a ball so that its centre misses the goal, what is the probability that the ball's centre misses to the right? (2 marks)

4 © SCSA MM2020 Q7 MODIFIED (13 marks) Consider the function $f(x) = e^{2x} - 6e^x + 8$.

a Determine the coordinates of the x-intercept(s) of f. You may wish to consider the factorised version of f: $f(x) = (e^x - 2)(e^x - 4)$. (3 marks)

b Show that there is only one turning point on the graph of f, which is located at $(\ln(3), -1)$. (3 marks)

c Determine the coordinates of the point(s) of inflection of f. (3 marks)

d Sketch the function f, labelling clearly all intercepts, the turning point and point(s) of inflection. Some approximate values of the natural logarithmic function provided in the table below may be helpful. (4 marks)

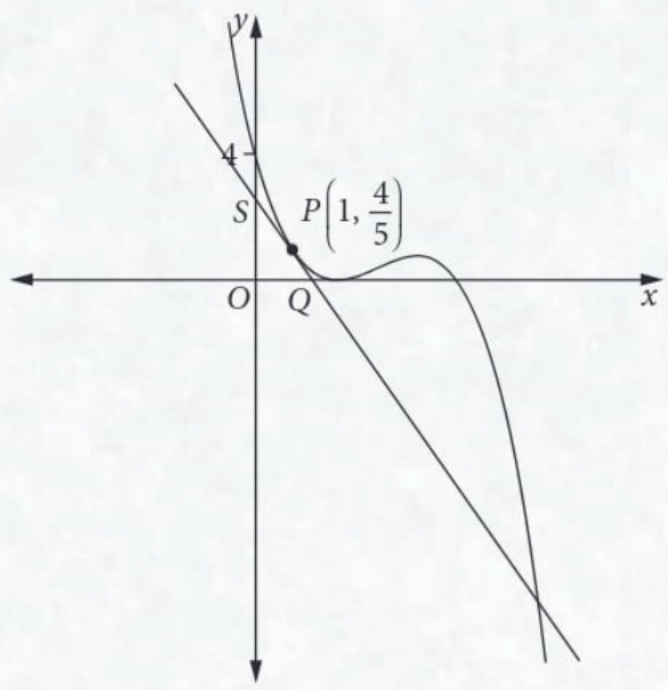
x	1	2	3	4
$\ln(x)$	0	0.7	1.1	1.4

Cumulative examination: Calculator-assumed

Total number of marks: 66 Reading time: 8 minutes Working time: 66 minutes

1 (3 marks) The volume $V\text{cm}^3$ of water in a vessel is given by $V = \frac{1}{6}\pi x^3$, where $x\,\text{cm}$ is the depth of the water in the vessel in cm. By using the increments formula, determine an approximation for the change in depth when the volume of water changes from 200 to $210\,\text{cm}^3$.

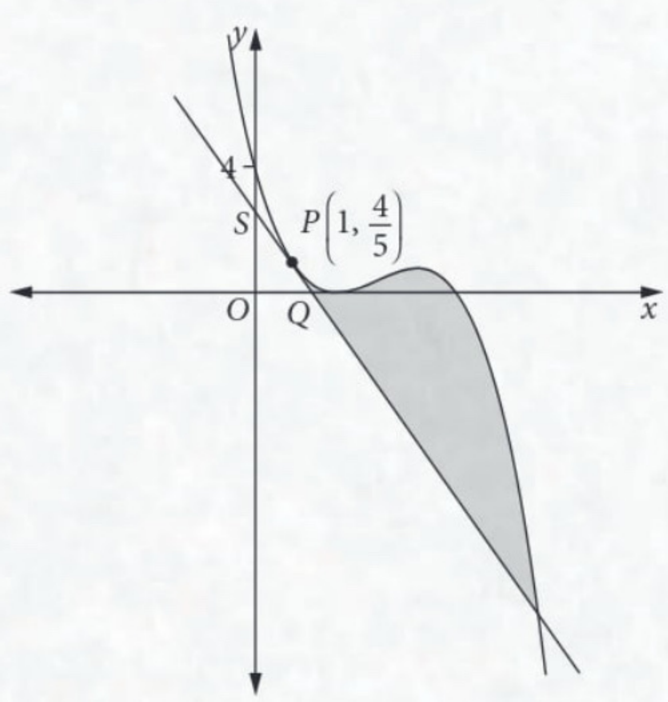
2 (7 marks) Let $f(x) = \frac{1}{5}(x-2)^2(5-x)$. The point $P\left(1, \frac{4}{5}\right)$ is on the graph of f, as shown below. The tangent at P cuts the y-axis at S and the x-axis at Q.



a Write down the derivative $f'(x)$ of $f(x)$. (1 mark)

b **i** Find the equation of the tangent to the graph of f at the point $P\left(1, \frac{4}{5}\right)$. (1 mark)

ii Find the coordinates of points Q and S. (2 marks)



c Find the area of the shaded region in the graph above. (3 marks)

3 (9 marks) Steve, Katerina and Jess are three students who have agreed to take part in a psychology experiment. Each student is to answer several sets of multiple-choice questions. Each set has the same number of questions, n, where n is a number greater than 20. For each question there are four possible options (A, B, C or D), of which only one is correct.

a Steve decides to guess the answer to every question, so that for each question he chooses A, B, C or D at random. Let the random variable X be the number of questions that Steve answers correctly in a particular set.

i What is the probability that Steve will answer the first three questions of this set correctly? (1 mark)

ii Find, to four decimal places, the probability that Steve will answer at least 10 of the first 20 questions of this set correctly. (2 marks)

iii Use the fact that the variance of X is $\frac{75}{16}$ to show that the value of n is 25. (1 mark)

If Katerina answers a question correctly, the probability that she will answer the next question correctly is $\frac{3}{4}$. If she answers a question incorrectly, the probability that she will answer the next question incorrectly is $\frac{2}{3}$.

In a particular set, Katerina answers Question 1 incorrectly.

b Calculate the probability that Katerina will answer questions 3, 4 and 5 correctly. (3 marks)

c The probability that Jess will answer any question correctly, independently of her answer to any other question, is p $(p > 0)$. Let the random variable Y be the number of questions that Jess answers correctly in any set of 25.

If $P(Y > 23) = 6P(Y = 25)$, show that the value of p is $\frac{5}{6}$. (2 marks)

4 (9 marks) Toby is learning to speak Spanish before going to South America for 12 months. While completing an online course, the number of words he learns, w, is modelled by the function.

$$w = 100\ln(t + 1) + 150$$

where t is the number of days after he starts his online course.

Toby needs a very basic vocabulary of 600 words for the trip.

a How many Spanish words did Toby know when he started the course? (1 mark)

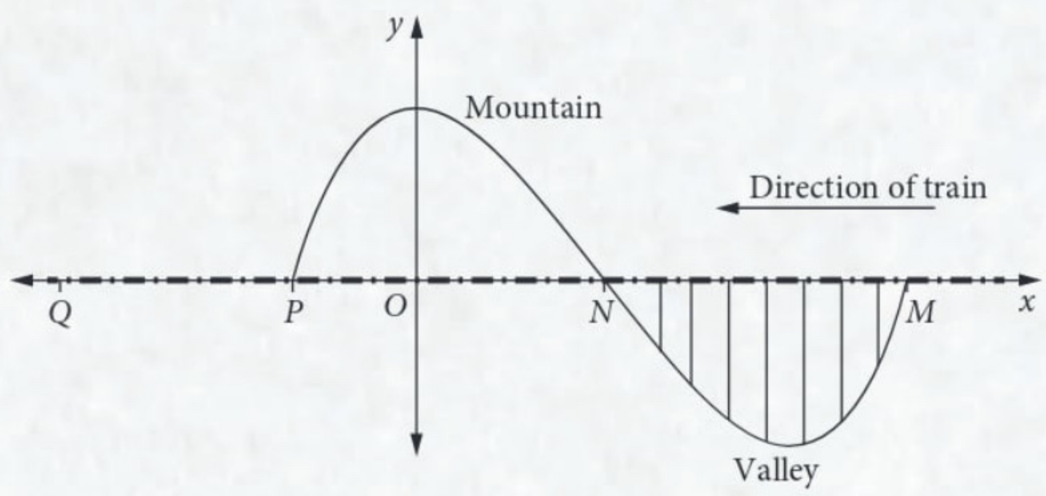
b How many words did Toby learn in the first day? (2 marks)

c How many Spanish words did Toby know after 5 days? (1 mark)

d How long will it take him to learn the very basic vocabulary? (2 marks)

e Write the equation in the form $t = e^{aw-b} - c$. (3 marks)

5 (14 marks) A train is travelling at a constant speed of w km/h along a straight level track from M towards Q. The train will travel along a section of track $MNPQ$.



Section MN passes along a bridge over a valley. Section NP passes through a tunnel in a mountain. Section PQ is 6.2 km long.

From M to P, the curve of the valley and the mountain, directly below and above the train track, is modelled by the graph of

$y = \frac{1}{200}(ax^3 + bx^2 + c)$ where a, b and c are real numbers.

All measurements are in kilometres.

a The curve defined from M to P passes through $N(2, 0)$. The gradient of the curve at N is -0.06 and the curve has a turning point at $x = 4$.

i From this information write down three simultaneous equations in a, b and c. (3 marks)

ii Hence show that $a = 1$, $b = -6$ and $c = 16$. (2 marks)

b Find, giving exact values

i the coordinates of M and P (2 marks)

ii the length of the tunnel (1 mark)

iii the maximum depth of the valley below the train track. (1 mark)

The driver sees a large rock on the track at a point Q, 6.2 km from P. The driver puts on the brakes at the instant that the front of the train comes out of the tunnel at P.

From its initial speed of w km/h, the train slows down from point P so that its speed v km/h is given by $v = k\log_e\left[\frac{(d+1)}{7}\right]$ where d km is the distance of the front of the train from P and k is a real constant.

c Find the value of k in terms of w. (1 mark)

d If $v = \frac{120\log_e(2)}{\log_e(7)}$ when $d = 2.5$, find the value of w. (2 marks)

e Find the exact distance from the front of the train to the large rock when the train finally stops. (2 marks)

6 © SCSA MM2017 Q12bcde MODIFIED (9 marks) A common problem with a particular tablet is screen failure. Historically, the manufacturer of Slate Tablets has found that 1% of its tablet screens will fail within three years. A sample of 200 tablets is taken. Let the random variable X denote the number of tablets that have screen failure within three years in the sample of 200.

a State the distribution of X. (2 marks)

b Determine the probability, to four decimal places, that more than four tablets will have screen failure within three years. (2 marks)

In a random sample of 200 Slate Tablets, four of them had screen failure within three years.

c Calculate an approximate 95% confidence interval for the proportion of tablets that have screen failure within three years. Give your answer to four decimal places. (3 marks)

d Comment on whether the company's historical data still appears relevant for current standards of tablet screen quality. (1 mark)

e The company's quality control department wants the proportion of tablets with faulty screens to be no more than 1%. Based on your confidence interval, decide whether the quality control department is meeting its target. Justify your decision. (1 mark)

7 © SCSA MM2021 Q11 (15 marks) A new political party, the Sustainable Energy Party, is planning to have candidates run in the next election. Researchers have collected data that suggests the proportion of voters likely to vote for the party to be 23%.

One year before the next election, random samples of 400 voters were taken in a particular electorate. Let $\hat{p}$ denote the sample proportion of voters who indicated they would vote for the Sustainable Energy Party at the next election.

a State the distribution of $\hat{p}$. (2 marks)

b Calculate the probability that the proportion of voters likely to vote for the Sustainable Energy Party in a sample of 400 is less than 0.20. (2 marks)

One week before the election, researchers believed that the proportion of voters likely to vote for the party in that same electorate had increased. A random sample of 200 voters was taken at this time, and 55 of them indicated they would vote for the Sustainable Energy Party at the next election.

c Based on this sample, estimate the proportion of voters likely to vote for the Sustainable Energy Party in this electorate. (1 mark)

d For a 99% confidence interval, what is the margin of error of the sample proportion of voters likely to vote for the Sustainable Energy Party in this electorate, based on this sample? (2 marks)

e Based on this sample, calculate a 95% confidence interval for the population proportion of voters likely to vote for the Sustainable Energy Party in this electorate. (3 marks)

f Based on the research, did the proportion of voters likely to vote for the Sustainable Energy Party in this electorate increase in the year leading up to the election? Justify your answer. (2 marks)

g The analysis above models the number of voters likely to vote for the Sustainable Energy Party as binomially distributed. State and discuss the validity of any assumptions for the binomial distribution in this context. (3 marks)