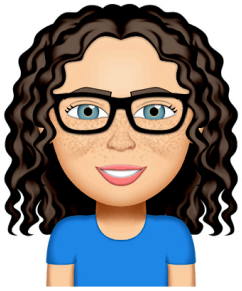
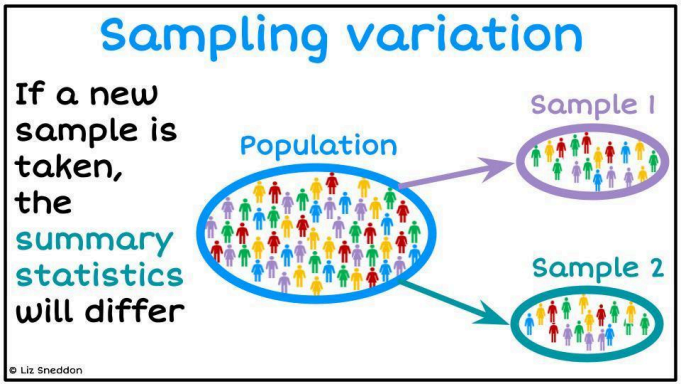
Sampling Variation & Random Variation



By Liz Sneddon

Variation BETWEEN samples (or Sampling Variation)

Each time we collect data in a study (either observational OR experimental), we need to consider the idea of what would happen if we were to collect a new set of data. The variation that we would see **BETWEEN** datasets (sampling the **SAME** variables from the **SAME** population), is called **sampling variation**.



Here is the definition of **sampling variation** from TKI:

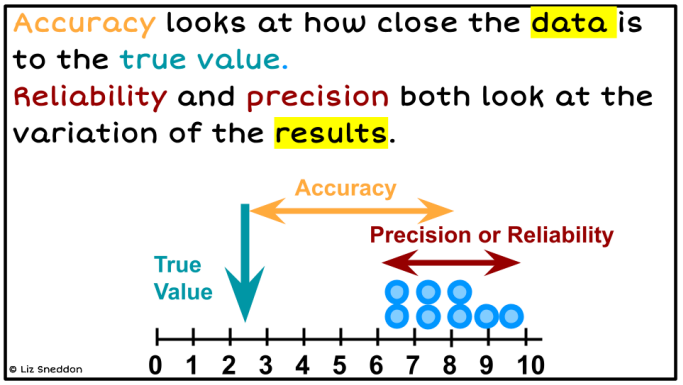
> The *variation* in a *sample statistic* from *sample* to sample.
>
> Suppose a *sample* is taken and a sample statistic, such as a *sample mean*, is calculated. If a second sample of the same size is taken from the same *population*, it is almost certain that the sample mean calculated from this sample will be different from that calculated from the first sample. If further sample means are calculated, by repeatedly taking samples of the same size from the same population, then the differences in these sample means illustrate sampling variation.

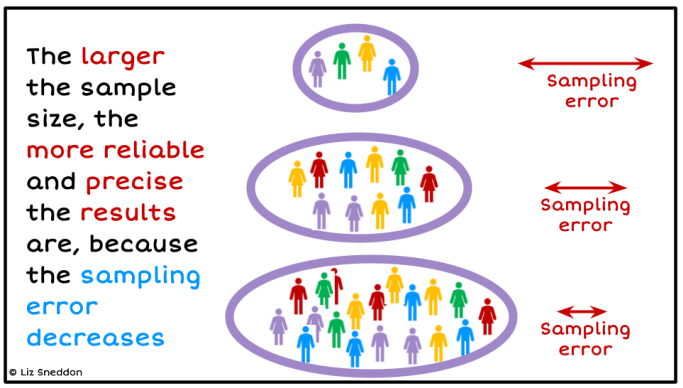
Sample size

Sampling variation **can** be reduced by **increasing** the **sample size**.

The general idea of sample size is that the more information you have, the greater the precision and reliability, the greater the confidence that you can have in the results.

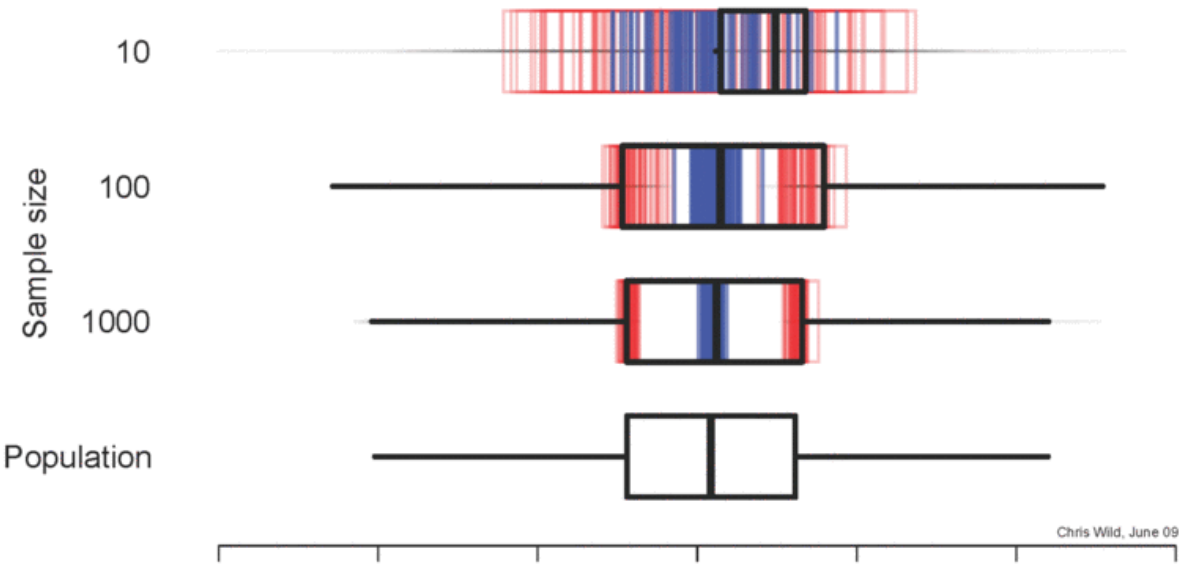
Any estimates we make (such as mean, standard deviation, etc.) have an associated level of uncertainty, which depends on the underlying variation in the distribution as well as the sample size. The more variation there is in the distribution or variable, the greater the uncertainty in our estimate. Similarly, the larger the sample size, the more information we have and so our uncertainty reduces.





Example:

Here is an animation, from Prof. Chris Wild, using boxplots to demonstrate the effect that sample size has on the estimates for the Median, Lower and Upper Quartile:



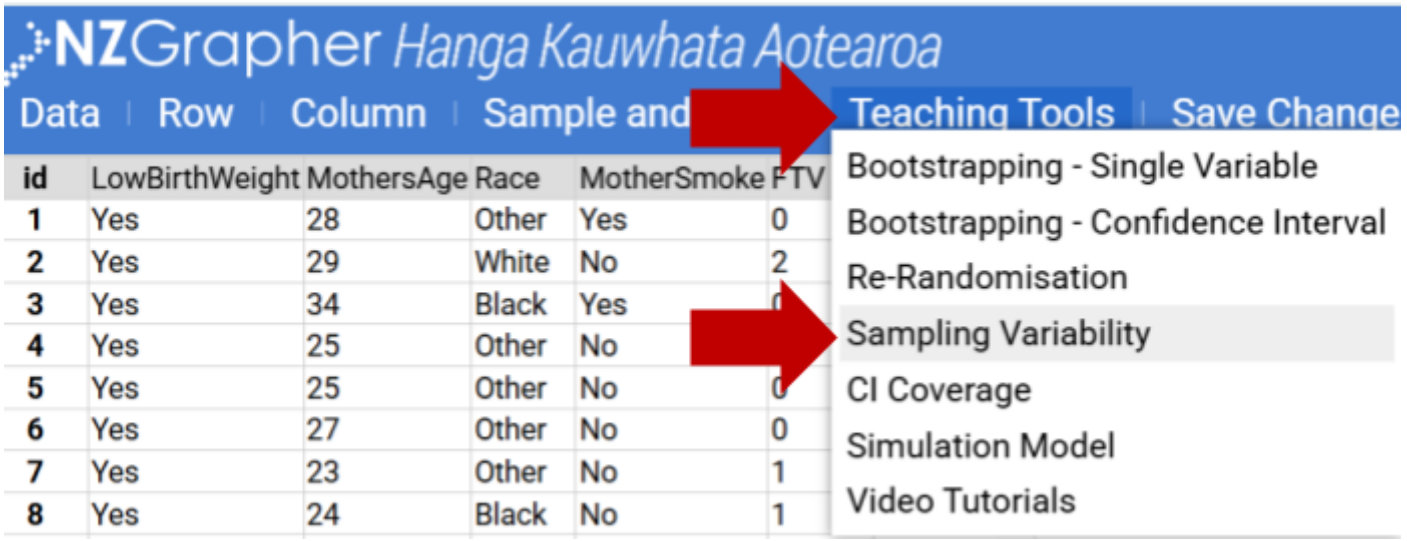
Watch what happens to the median (the blue line) for different sample sizes and notice the sampling variation. This shows us that as the sample size increases, the sampling variation decreases and our estimates for the median become more precise and reliable.

Similarly, we notice that the estimates for the Upper Quartile and Lower Quartile (red lines) are also becoming more precise and reliable as the sample size increases.

If you sample the entire population, then sampling variation disappears, but the variation **WITHIN** a sample/dataset will always remain, as this is the variation inherent in the variable.

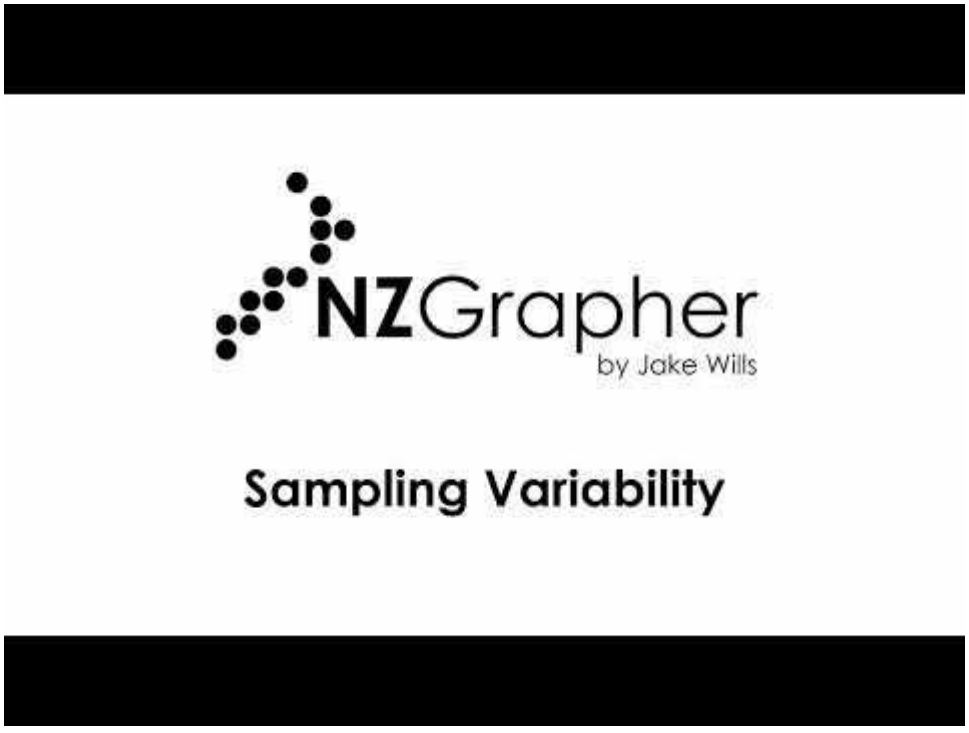
NZGrapher & Sampling Variation

Under the Teaching Tools menu, there is an option for Sampling Variability.



We won't spend time looking at this, but if you haven't explored this before, please explore. It is particularly useful for Inference at Level 1, 2 and 3.

Here's a video that Jake made, talking through how to use the Sampling tool



https://youtu.be/50oBxZw3IIE?si=1m_y_myGggo-xGO5

NZQA Level 3 Probability Concepts Exams

The concept of sampling variation comes up in these exams. Here is an example:

2025 Question 1(d)(ii):

The school claims that, in general, for students playing football, more injuries occur playing on artificial turf than on natural grass.

Discuss why care should be taken using this data to make this claim.

Answer:

Care should be taken with **generalising** this data to all football players in general, because this data is only collected from **one school, during one season**, from student football players.

Older football players may be more (or less) likely to be injured on different surfaces than younger football players, Injury rates may vary depending on the level of the football players, Injury rates may vary depending on season / weather conditions, Injury rates may differ depending on location e.g. different types of turf or natural grass surfaces.

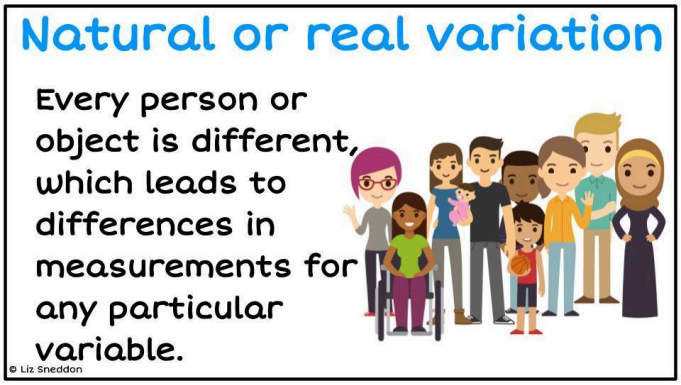
Key concepts:

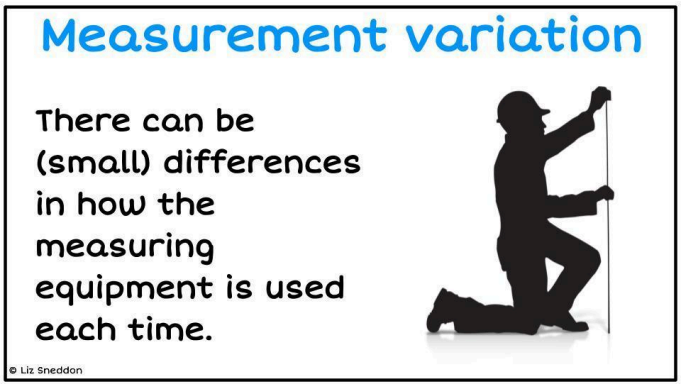
The key concept is whether the data collected is **representative of the population**. Students need to consider **how** the data has been collected, how **accurate** and **reliable** the data is, **sample size**, and whether the results are able to be used to **generalise to a wider population**.

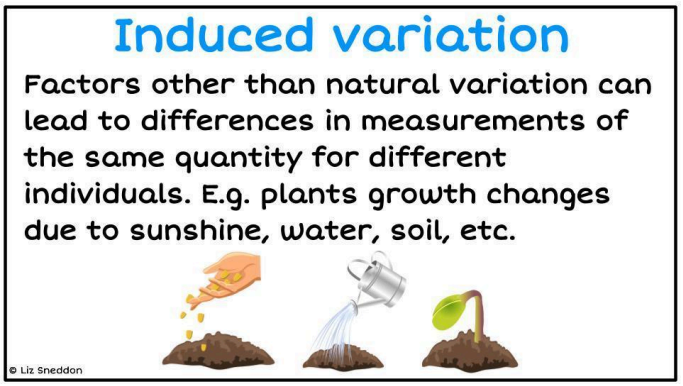
The underlying concept is sampling variation and the differences between samples.

Variation WITHIN samples (or Random Variation)

Our world is full of variation. When we collect data, we can identify key sources of variation that might affect the variables that we are interested in exploring and investigating. There are several different types of variation **WITHIN** a dataset:







With only one data value per person, we need to be aware that variation is always present in each data value and the dataset, and account for this when identifying patterns in the data.

The only way we can **control** the variation **WITHIN** a sample is to make sure that the data is collected accurately as possible and by better experimental design. It is **NOT** possible to eliminate the variation within a sample, as this is the variation that is naturally inherent in the variable.

Example:

Suppose we want to compare the blood pressures of a drug-treated group and a control group on placebo. Blood pressures will vary due to:

- person-to-person variation – blood pressure varies between each person due to factors such as genetics, age, sex, ethnicity, physical build, etc.
- occasion-to-occasion variation – each person's blood pressure differs at different times of the day, due to factors such as stress, exercise, sleep, etc.
- measurement error – the recorded value differs from the person's true pressure due to the tools or process used due to factors such as equipment inaccuracy, and technique errors.

Distributions

A distribution in statistics is a function that shows the possible values for a variable and how often they occur. Each variable has an underlying distribution, which shows the pattern of variation **WITHIN** a variable.

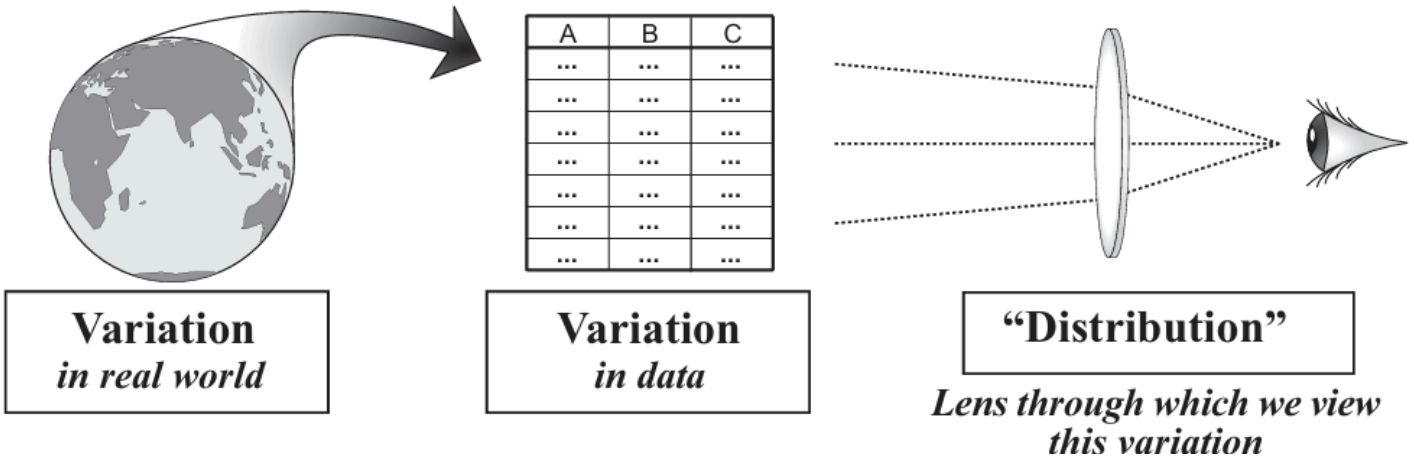


Figure 1: Image from Prof. Chris Wild[1]

The underlying (or theoretical) distribution of a variable is what we are trying to identify when we collect data, so that we understand and apply this knowledge to the real-world.

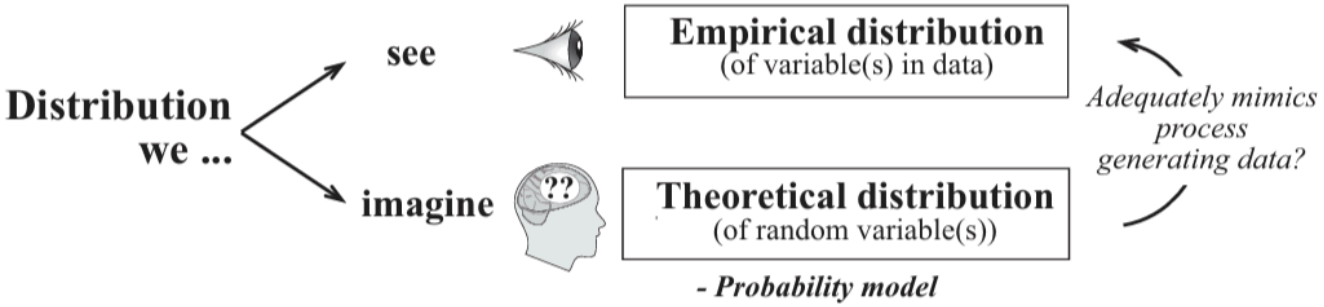
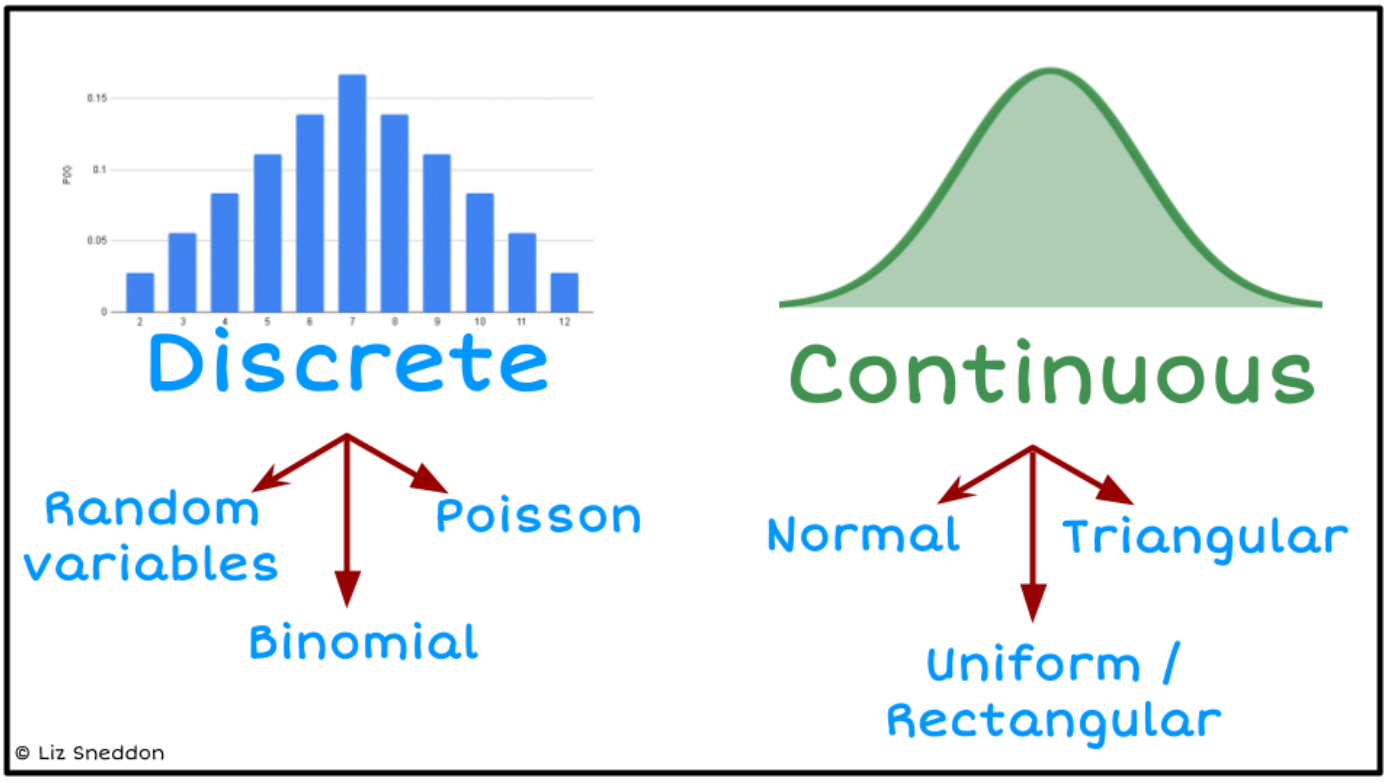


Figure 2: Image from Prof. Chris Wild

[1] https://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Wild.pdf

There are many distributions that students need to be familiar with.

- For Level 3 Probability Concepts, the only distribution that students may be asked about is **discrete uniform distributions** where the outcomes are **equally likely**.
- For Level 3 Probability Distributions, students need to use the distributions shown below.
- It also is relevant to Level 2 Simulation, where students need to use **discrete random variables**.



Sample size and Shape

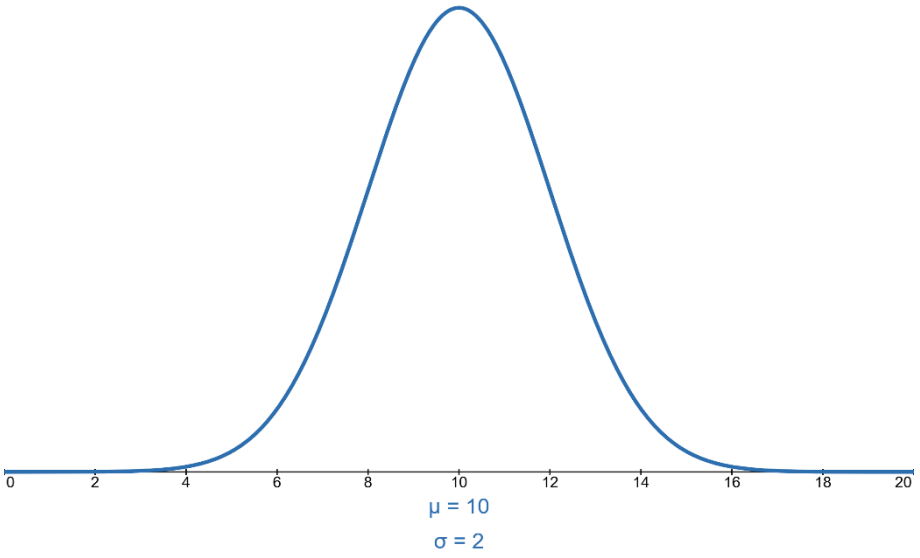
When the sample size of a group is below 30, it becomes much harder to identify the shape of the underlying population distribution. We have seen that outliers have a large effect on the mean, range and standard deviation. In addition, there are times when the data may show clusters or gaps in the data, which as the sample size is small, will also have a large effect on the mean and standard deviation in particular.

In addition, when considering the shape, even if the population distribution is a normal distribution, it is very common for the shape of a small sample to appear skewed.

Example:

The following graphs are random numbers generated from a normal distribution with a mean of 10 and a standard deviation of 2. The graphs are from a variety of sample sizes so we can better understand the connection between the shape and sample size.

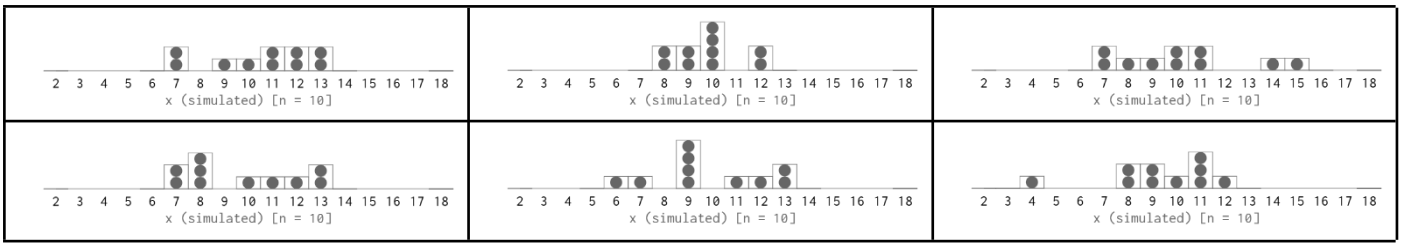
This is the population distribution that the samples below have been sampled from:



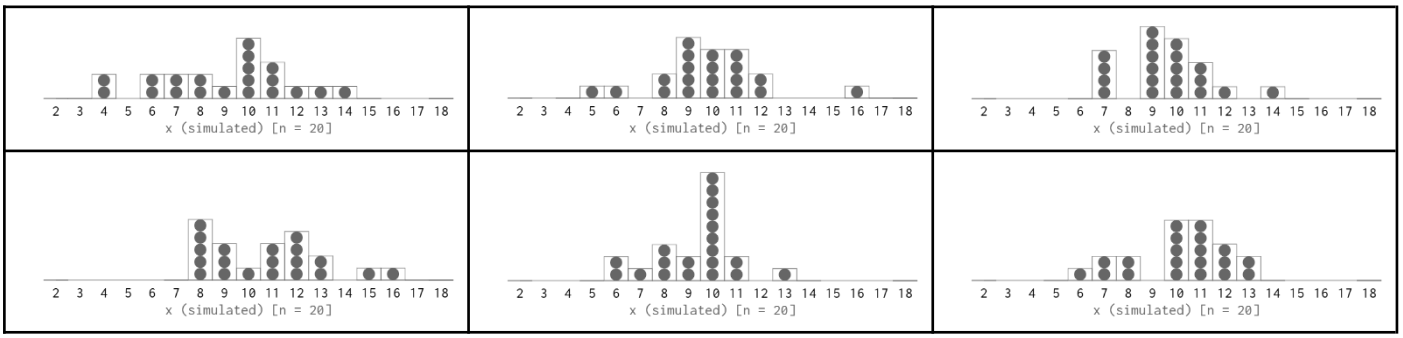
Notice that the centre (mean) is exactly at 10, and with a standard deviation of 2, most of the data lies between 4 and 16.

Now compare the shape, centre and spread for each of the graphs below:

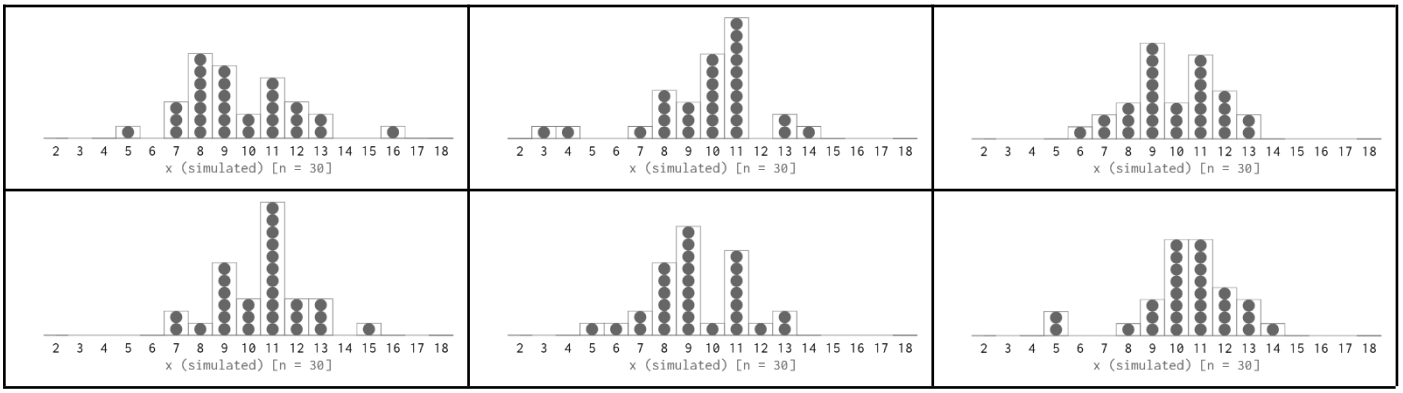
Sample size, $n = 10$:



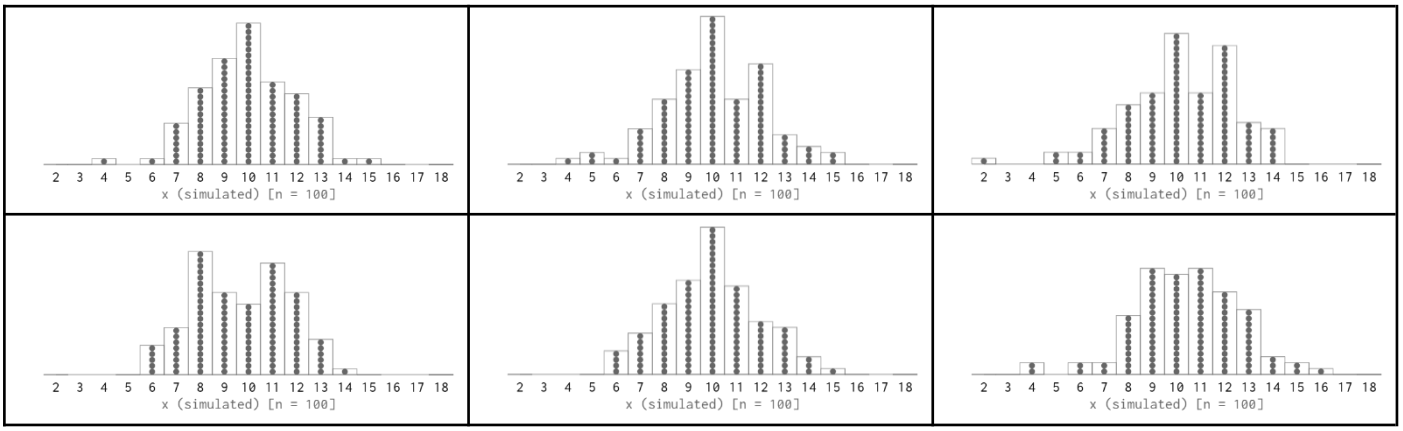
Sample size, $n = 20$:



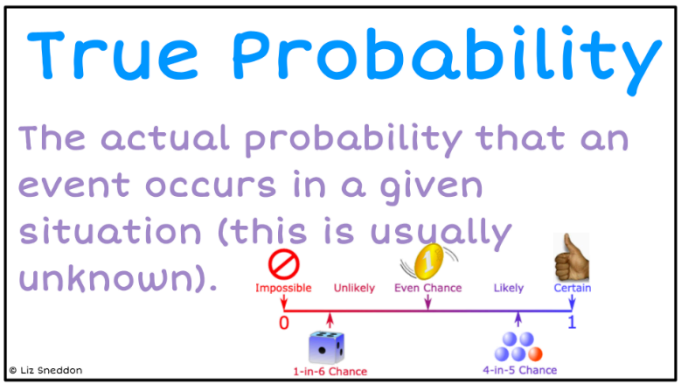
Sample size, $n = 30$:

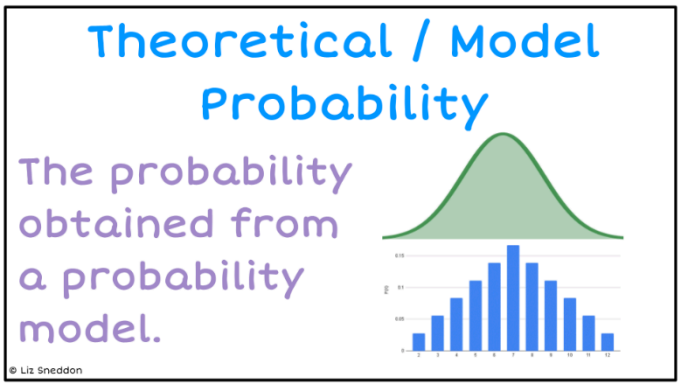


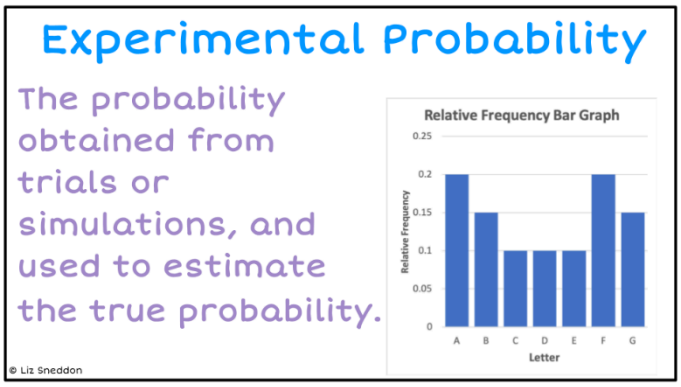
Sample size, $n = 100$:



True, Model and Experimental probabilities







Both the model probability and the experimental probability of an event give ***estimates*** of the true probability.

We frequently use simulation or experimental trials to estimate the true probability. Here is the definition of **simulation** from TKI[2]:

> A technique for imitating the behaviour of a situation that involves elements of chance or a probability activity. The technique uses tools such as coins, dice, random numbers from a calculator, random numbers from random number tables, and random numbers generated by computers.

When a simulation is carried out, results (estimates such as a probability or average) will vary due to random variation when data is generated (e.g. tossing a coin could randomly give a head or a tail).

Example 1:

Students tossed a coin 100 times and recorded that it landed on Heads 48 times.

True probability =	unknown (it depends on the specific coin, and whether this coin is fair),
Theoretical / Model probability =	0.5 (it is based on a discrete uniform distribution where P(Head) = 0.5)
Experimental probability =	$\frac{48}{100}$ *or* 0.48

As the data was generated by a random process (e.g. tossing a coin), the probabilities in each simulation will vary due to **random variation**.

[2] https://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Glossary/Glossary-page-S

NZQA Level 3 Probability Concepts Exams

The concept of random variation comes up in these exams. Here are several examples:

Example 1: 2025 Question 2(d)

The 2023 School Sports Census of 50 000 students recorded 8880 cricket players. A random sample of 1000 New Zealand secondary school students found that 120 played cricket. Amongst secondary school sports coordinators in New Zealand, it is thought that about 25% of students played cricket.

Compare the **true probability**, **model estimate**, and **experimental estimate** for cricket participation in 2023.

Discuss which estimate is the most reliable for predicting future cricket participation.

Use statistical reasoning to support your answer.

Answer:

$$\text{True probablility from 2023 census data} = \frac{8\ 880}{50\ 000} = 0.1776$$

$$\text{Model probability based on historial percentage} = 0.25$$

$$\text{Experimental probability from random sample in 2023} = \frac{120}{1000} = 0.12$$

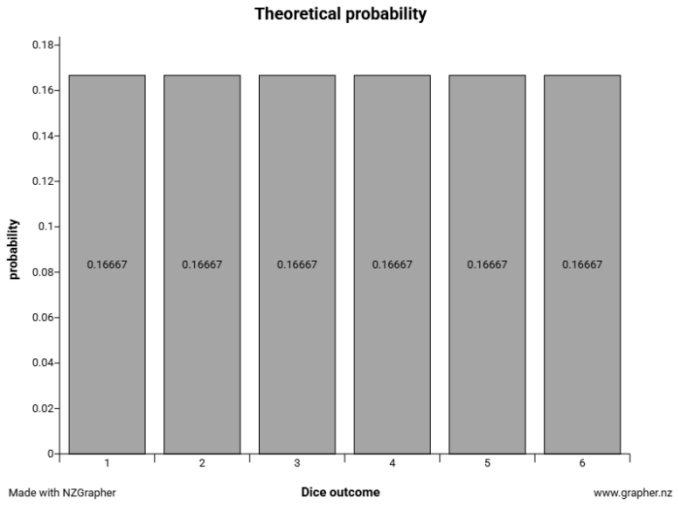
The true probability is based on the entire census and is likely to be the most accurate. The model estimate is based on the opinion of sports coordinators, which isn’t realistic. The experimental estimate may be unreliable because it is based on a small sample size.

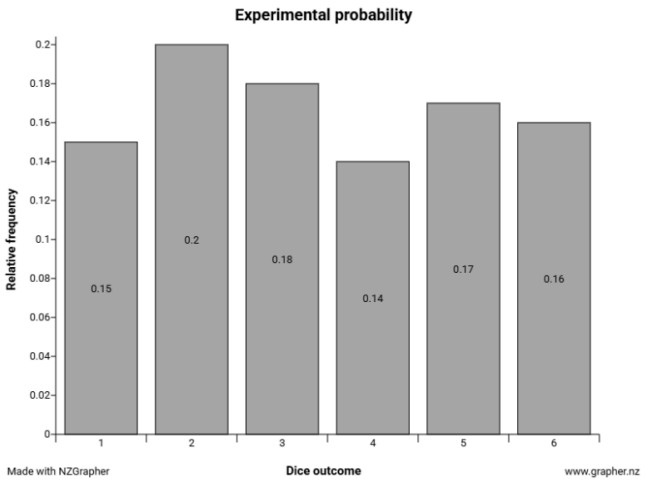
Equally likely and Discrete Uniform distributions

For the Level 3 Probability Concepts exam, students need to be familiar with Discrete Uniform distributions where events are equally likely.

Example 1:

Consider rolling one die. The graphs show the theoretical/model distribution on the left-hand side, and the experimental results on the right-hand side (an experiment was carried out rolling one die 100 times and the outcomes recorded).

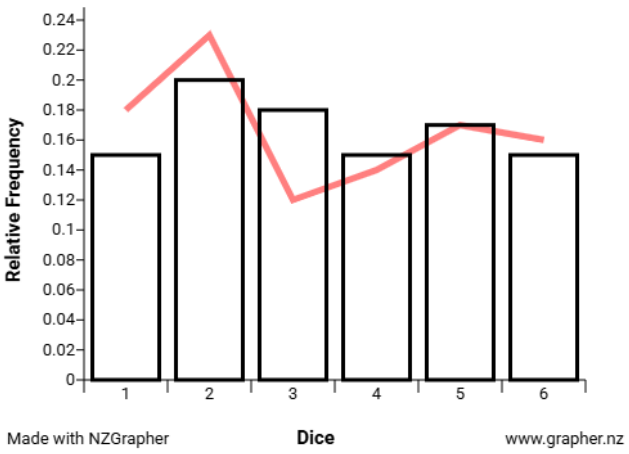


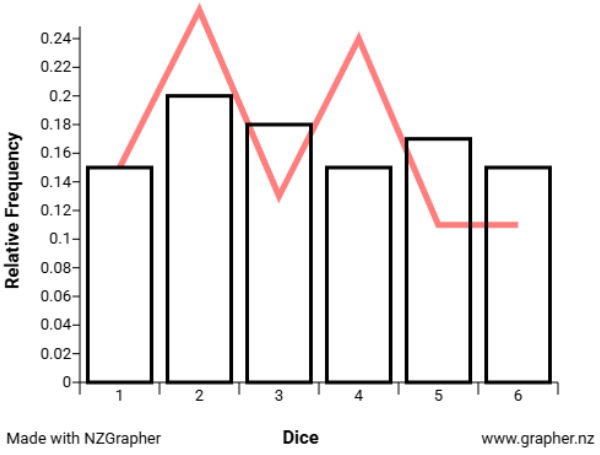


The model / theoretical distribution is a discrete uniform distribution, and shows each outcome is equally likely $\left(P = \frac{1}{6}\right)$. However the experimental probabilities vary, which is due to variation **WITHIN** a sample, or **random chance**.

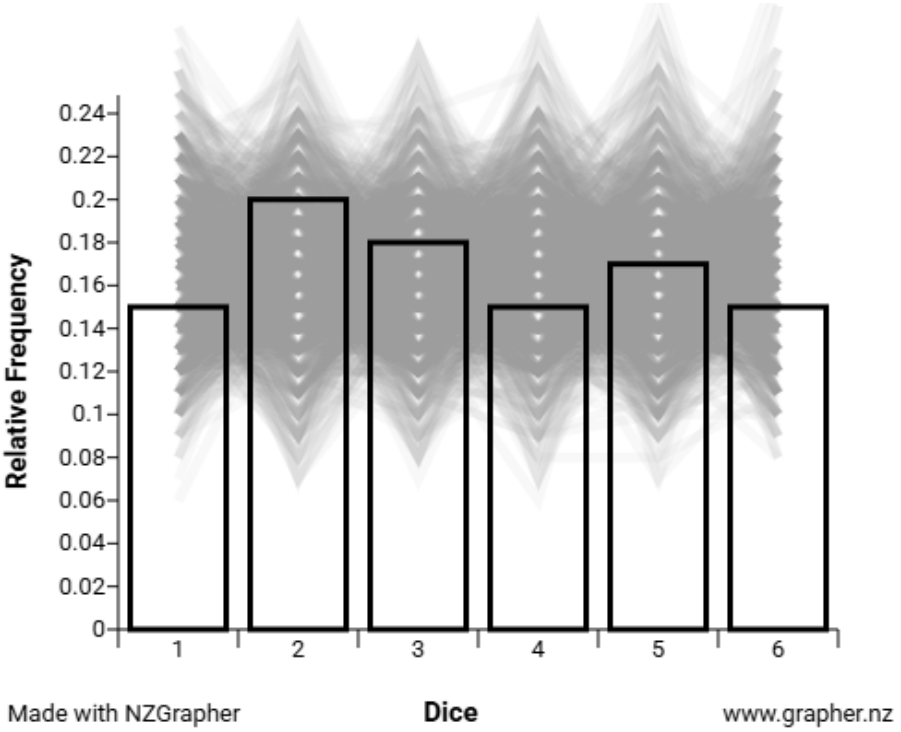
If we were unaware of the theoretical distribution and were only looking at the experimental probability, we need to consider how variation affects the results when trying to identify the underlying theoretical distribution.

Consider what different experimental distributions might look like to better understand variation. In the graphs below, the bar graphs represent the experimental distribution, and the red line represents one simulation based on each die outcome being equally likely.





Here is the experiment when simulated 1000 times (each simulation contains 100 data values). The grey lines show the results of 1000 simulations and give us an idea of **how much variation we can expect just by random chance**. This is random variation or variation **WITHIN** a sample, and the natural variation present when rolling a die.



NZGrapher: Simulation Model

Let's form this simulation model in NZGrapher:

Step 1:

Open the data in Google Sheets:

bit.ly/SimulationData

Select all the data and copy it, along with the heading.

Step 2:

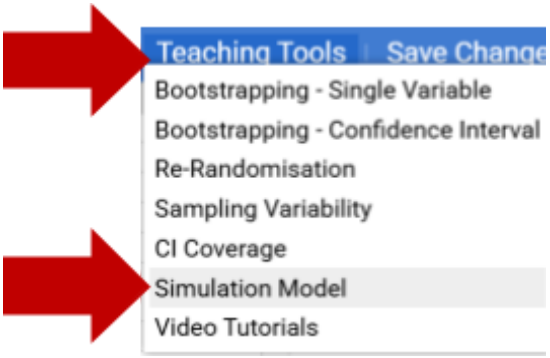
Open NZGrapher.

Go to the File Menu and select **Import from Clipboard**.

(If this is your first time importing data into NZGrapher, a popup window will appear asking you if you want to allow NZGrapher to copy from the clipboard, click the **Allow** button.)

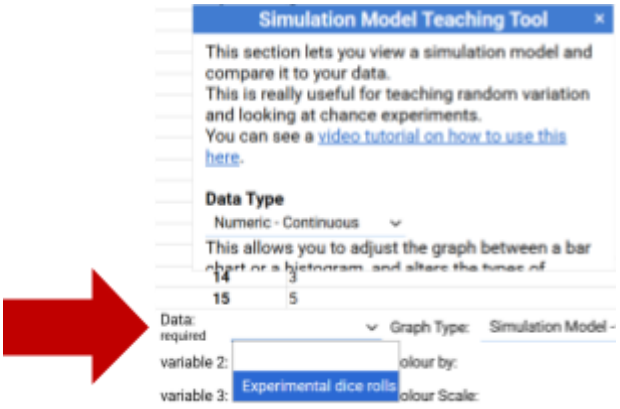
Step 3:

Go to the **Teaching Tools** menu and select the **Simulation Model**.



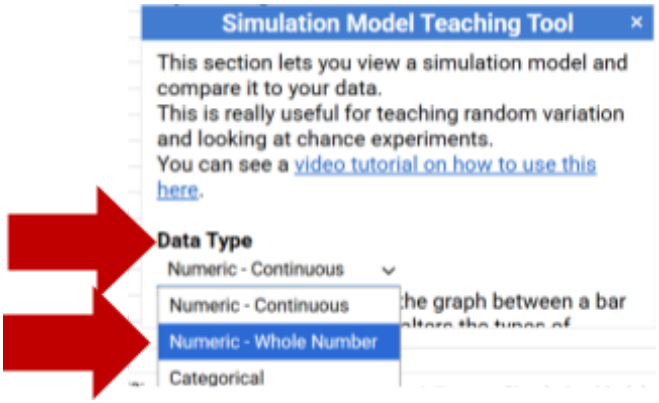
Step 4:

Under the **Data** drop-down list, select **Experimental die rolls**.



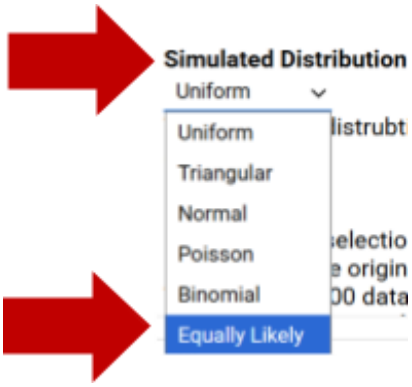
Step 5:

In the Simulation Model Teaching Tool popup box, under **Data Type**, select **Numeric – Whole Number**.



Step 6:

Scroll down, and under **Simulated Distribution**, select **Equally likely**.



Step 7:

Scroll down, and then you can step through the 1000 simulations (each with a sample size the same as the dataset, in this example, $n = 100$).

Press the **One** button first, to see one simulation, then press **Slow**, and then press **Fast**.

Step 8:

Other useful options:

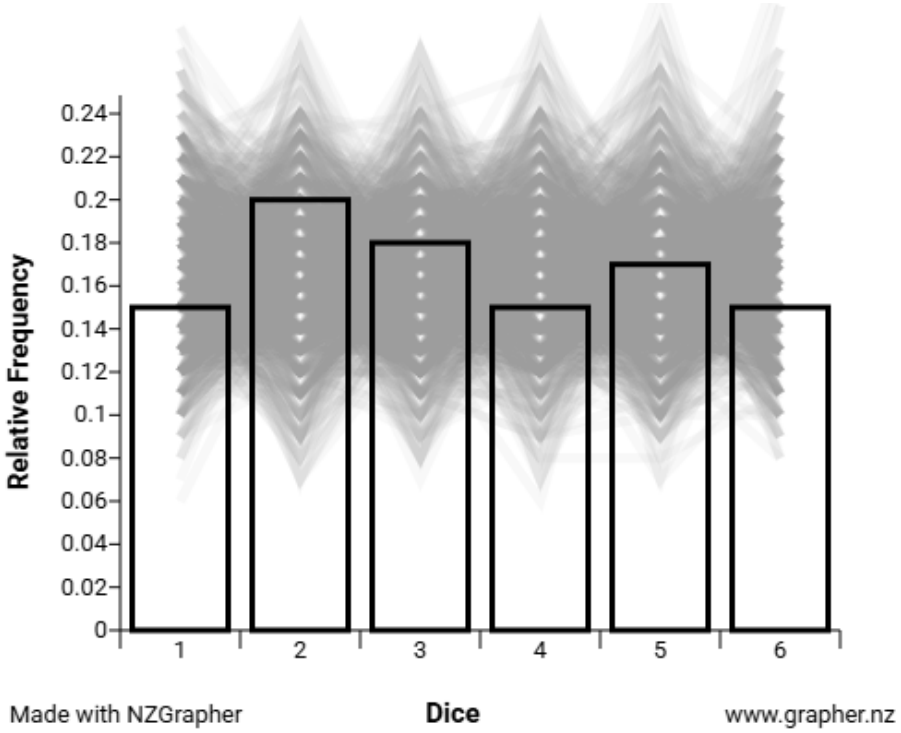
- You can choose to have either the **Frequency** or **Relative Frequency** on the y axis.
- You can make the lines of the boxes **Thick** (easier to see the bar graph).
- I prefer to make the size of the graph **Standard** or **Small** so that the y axis is not compressed.

Comparing results of a simulation against the data

Once we have designed and carried out a simulation, the next step is to compare the results with real-data and decide whether or not the model we used is likely to match the real-life data. We know that random variation is present when we generate data through a simulation, we also know that sampling variation is present when we collect data from a random sample, so for any particular real-life sample that we have, we know that there will be variation in these results. So, we need to consider how much variation is "acceptable" and how much is "unlikely".

Example 1 cont.:

Now look at the graph and form a conclusion about whether the die is fair. Justify with statistical reasoning.



Compare the experimental probability for each outcome (e.g. 1, 2, 3, …) with the simulated random variation. Is each experimental probability within the range of values (e.g. within the grey band)? If it is, then these results are likely to occur just by random chance.

As all outcomes lie within the grey band, it is **LIKELY** that each outcome on the die is **equally likely**.

Exercise 1:

Question 1:

Data was collected on sea turtles about their sex. From 100 sea turtles, researchers identified 73 females and 27 males. Without knowing about the proportion of male and female sea turtles, our assumption is that the sex split is 50:50.

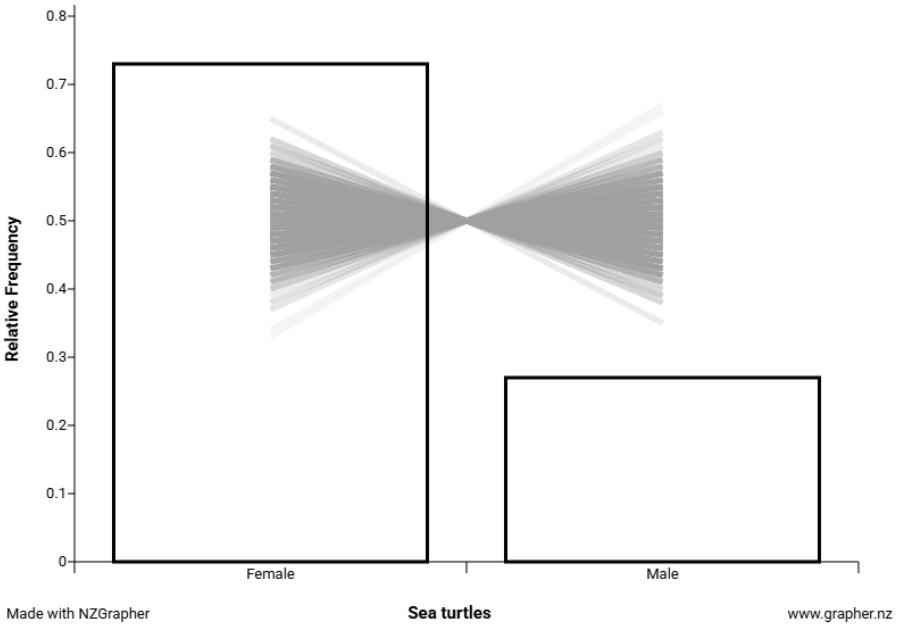
True probabilities:	Unknown
Model / Theoretical probability:	$P(Male) = 0.5,\ P(Female) = 0.5$
Experimental probability:	$P(Male) = \frac{27}{100},\ P(Female) = \frac{73}{100}$

We are going to use the **Sea Turtle** data that you have already copied into NZGrapher to run a simulation model.

Select Data Type: **Categorical** and

Simulated Distributions: **Equally likely outcomes**.

You should get a graph similar to this.



What conclusion(s) can you draw from this graph? Justify using statistical reasoning.

Question 2

Data was collected on the usage of 5 different parking spaces. From 100 cars, each time a car parked in one of the parking spaces, data was recorded. Without knowing about the proportion of cars using each parking space, we assume that each parking space is equally likely to be used.

Parking Space	1	2	3	4	5
Frequency	18	21	37	18	7

True probabilities: Unknown

Model / Theoretical probabilities: $P(Park\ 1) = P(2) = P(3) = P(4) = P(5) = \frac{1}{5}$

Experimental probabilities: $P(Park\ 1) = \frac{18}{100},\ P(2) = \frac{21}{100},\ P(3) = \frac{37}{100},$

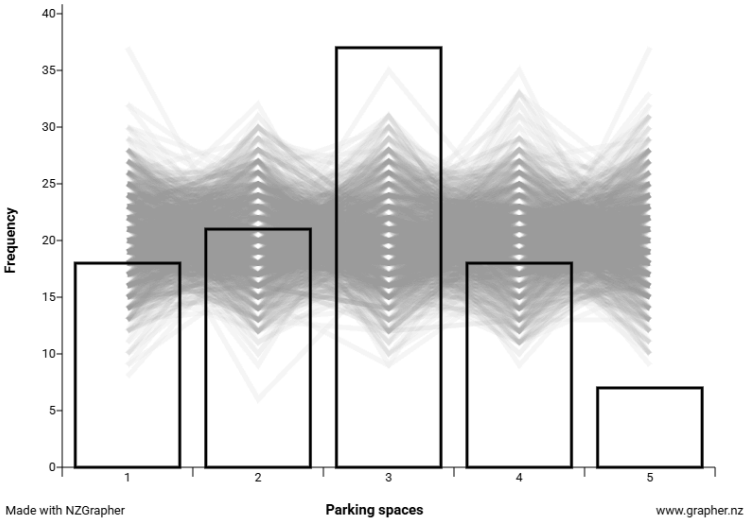
$P(4) = \frac{18}{100},\ P(5) = \frac{7}{100}$

We are going to use the **Parking spaces** data that you have already copied into NZGrapher to run a simulation model.

Select Data Type: **Numeric – Whole number** and

Simulated Distributions: **Equally likely outcomes**.

You should get a graph similar to this.



What conclusion(s) can you draw from this graph? Justify using statistical reasoning.

NZQA Level 3 Probability Concepts Exams

The concept of random variation comes up in these exams. Here are two examples:

Example 1: 2023 Question 3(b)(iii)

(b) The three friends meet for a coffee each in a cafe once a week. They have a system for deciding who will pay for all three coffees which involves each of them flipping a coin. The friend who flips a result different to the other two is considered the "odd one out" and ends up paying for all three coffees.

(iii) The friends have met for coffee 100 times over two years. The friends were surprised that during that whole period of time, they have never had to flip their coins more than five times at any meeting, in order to decide who paid.

Explain, using statistical reasoning, if this is an unusual occurrence.

Answer:

A simulation would allow the group of friends to see the **variation** through the **distribution** of the number (or proportion) of occasions each friend had to pay for the coffees in sets of sample size 100 based on the model of each friend paying of the time.

Friend A could then compare the observed value of 49 (or 0.49) to this simulated distribution to consider the likelihood of the observed number (or proportion) happening.

A decision could be made then as to whether the 49 occasions after 100 coffee meet ups are indeed unusual.

Key concept:

Understanding how random variation when we collect data or run a simulation affects the results.

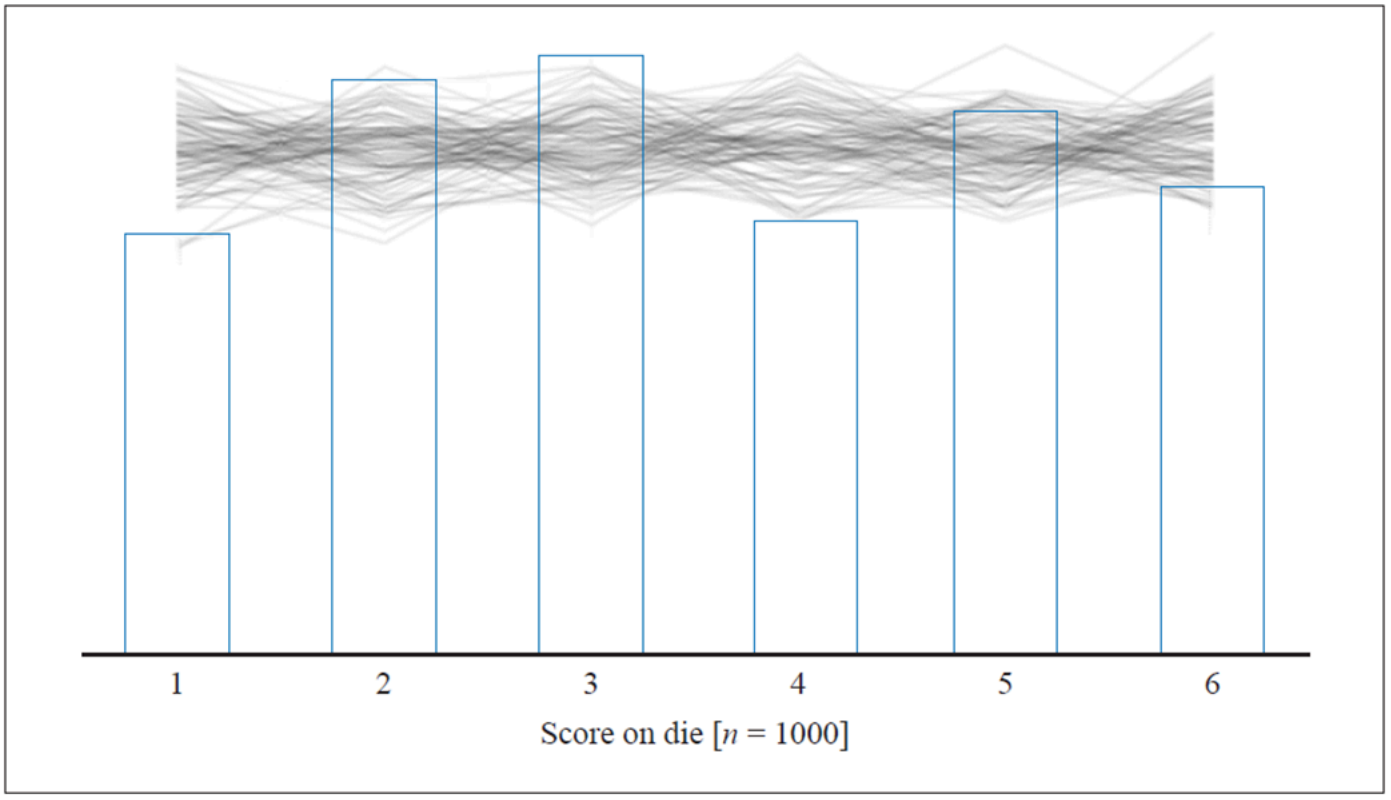
Example 2: 2024 Question 3(b)(ii)

(b) The player is concerned that one of their dice is biased. The outcomes from 1000 rolls of this die are summarised in the table below.

Outcome	1	2	3	4	5	6
Totals	138	189	197	143	179	154

The diagram below shows the results of 1000 trials of a simulation model. The simulation assumed that each outcome on the die was equally likely to occur.

The height of the blue vertical bars shows the relative frequencies of each observed digit outcome on the die, as shown in the table above. The grey band shows the variation expected for each outcome, based on simulating 1000 throws of a fair die.



(ii) Should the player be concerned that one of their dice is biased?

Use the results of the simulation model shown in the diagram on the previous page and the outcomes of the 1000 rolls given in the table, and refer to experimental, theoretical, and true probability as part of your answer.

(iii) Explain what this result means for the chance of throwing a total of 2 using this particular die as one of the pair of dice in game B.

Answer:

(ii) The true probability of each outcome for this die is unknown. The simulation results are an experimental distribution which allows us to see the variation in the numbers / proportion of each outcome from rolling the die for sample size 1000, using a theoretical / model probability of each outcome of 1/6 = 0.167.

We can compare the observed counts / proportions (experimental probabilities of each outcome) to this simulated distribution to consider the likelihood of the observed result happening.

Based on the model, we would expect a 1 to occur with a probability of 0.167.

The observed probability of a 1 for this die is 0.138.

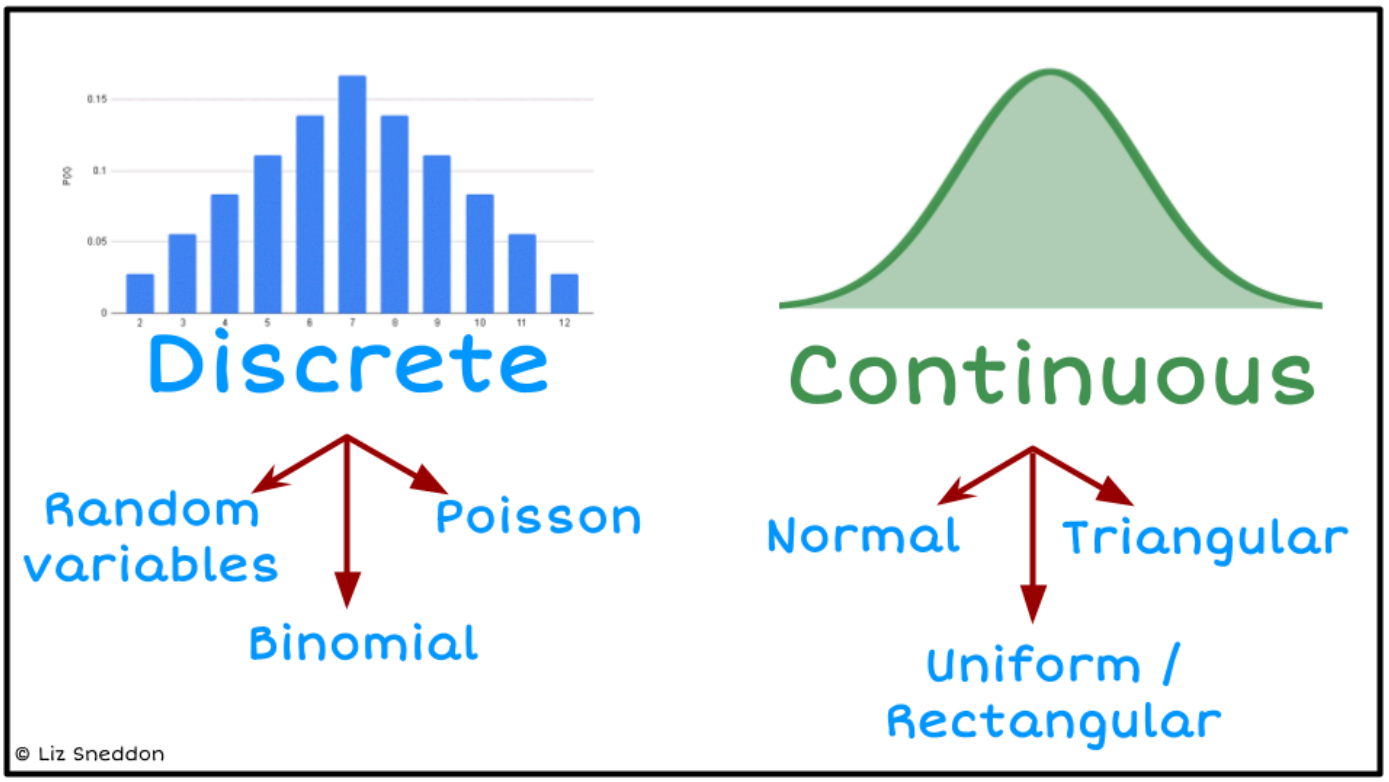
Even taking sampling variability into account, we very rarely see as few '1's in the simulated results, as were observed.

The player should be concerned that their die may be biased.

(iii) If this die is biased, and fewer '1's are occurring than expected, then the probability of rolling a total of 2 will be reduced. This means that it is less likely to roll a total of 2 and lose.

Other distributions

For the Level 3 Probability Distribution exam, students need to be familiar with any of the distributions below.



Students need to explore how well a distribution model fits data. This means students need to understand the role of random variation and be able to consider how much variation a probability model would have just by random chance. Then students can decide whether a particular probability distribution model is an appropriate fit to the data.

In order to explore the fit of a model, we need to identify first if the data is **discrete or continuous**, then we need to identify the **shape** and then the **parameters** in order to decide which model/theoretical probability distribution is appropriate.

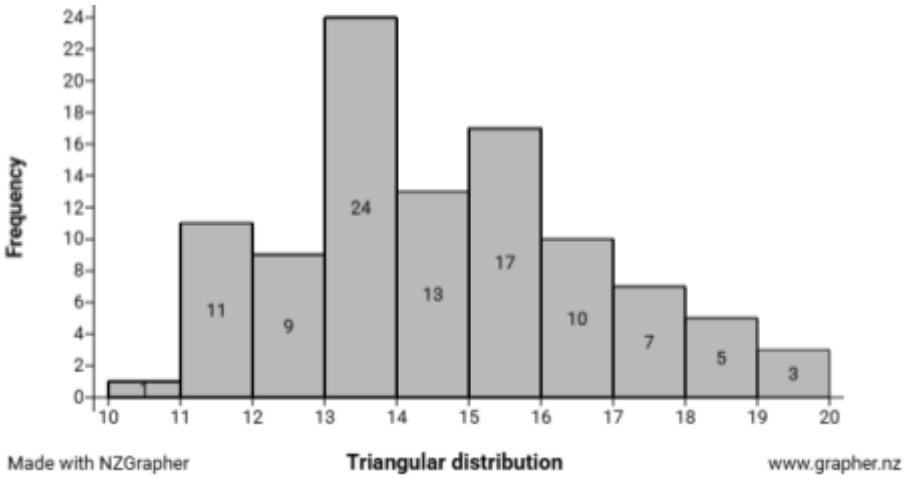
Then we can graph the data and explore how much random variation would be expected under a specific model. Lastly, we can compare data to the model to determine how well that model fits the data.

NZGrapher: Simulation Model

In the Google Sheet data has been randomly generated from a Triangular distribution, with parameters

$a = 10,\ b = 20,\ c = 14.$

This data on the graph is the experimental data.

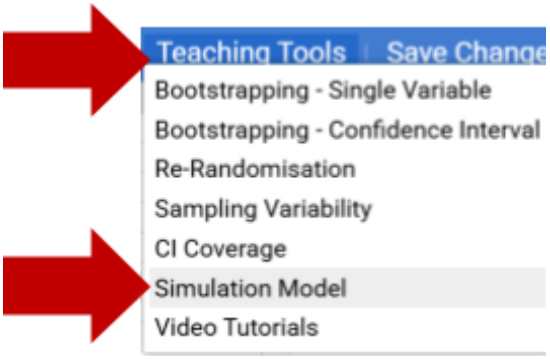


Note: When I draw a histogram of the data in NZGrapher, I need to choose the interval size in order to match the data (or if I'm making an example for my students, to choose the width of the intervals).

Let's form this simulation model in NZGrapher.

Step 1:

Go to the **Teaching Tools** menu and select the Simulation **Model**.

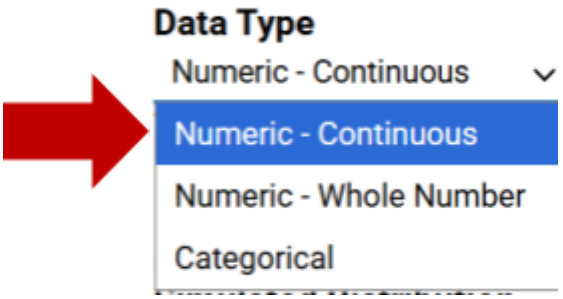


Step 2:

Under the **Data** drop-down list, select **Triangular distributions**.

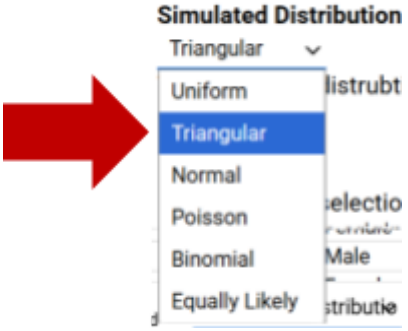
Step 3:

In the Simulation Model Teaching Tool popup box, under **Data Type**, select **Numeric – Continuous**.



Step 4:

Scroll down, and under **Simulated Distribution**, select **Triangular**.



Step 5:

Scroll down to find the **description**. You need to enter the values for the triangular distribution:

$min = 10$, $max = 20$, $mode = 14$.

Description
Based on your selections above we will draw a histogram of the original data.
This data has 100 data points.
We will simulate a triangular distribution (min = 10 , max = 20 , mode = 14) with 100 data points and show this in light grey.
We will do up to 1000 simulations and show the results in the graph.

Step 6:

Scroll down, and then you can step through the 1000 simulations (each with a sample size the same as the dataset, in this example, $n = 100$).

Press the **One** button first, to see one simulation, then press **Slow**, and then press **Fast**.

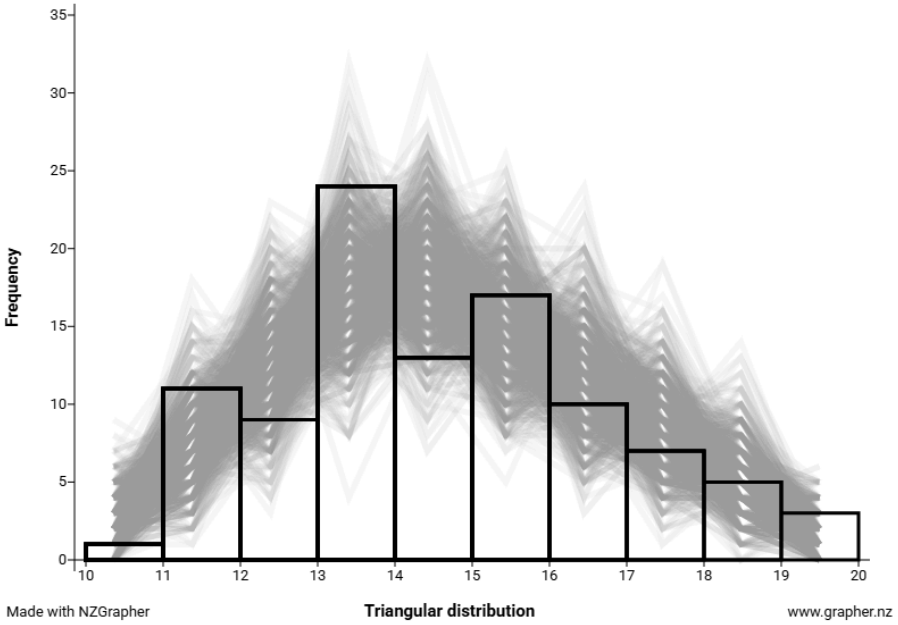
Step through remaining 1000 simulations:

One | Slow | Medium | Fast

Other Controls

Pause | Reset

You should get a graph similar to this.



Now we want to interpret the graph and form a conclusion.

The grey band shows how much random variation we can expect simply due to random chance.

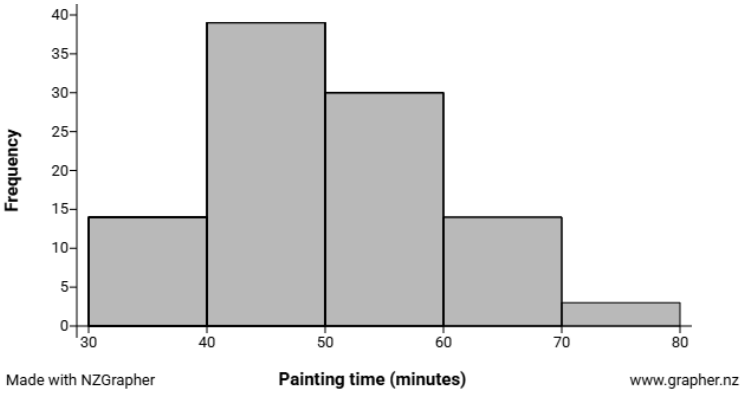
We can see that for each interval, the experimental probabilities are all within the grey bands, meaning that it is possible that the experimental results are all possible to come from a Triangular distribution with $a = 10$, $b = 20$, $c = 14$.

This means that a Triangular distribution with $a = 10$, $b = 20$, $c = 14$ is a good fit to the experimental data.

Example 2:

Data was collected on the length of time it took to paint the walls in a garage (all were 2 car garages) and is shown in the table below.

Paint time (minutes)	Frequency
30-39	14
40-49	39
50-59	30
60-69	14
70-79	3



Step 1: The data is continuous.

Step 2: Looking at the graph, the shape appears triangular, therefore a triangle distribution will be used to model the painting time.

Step 3: Looking at the graph, suitable parameters are: minimum $a = 30$, maximum $b = 80$, mode $c = 45$.

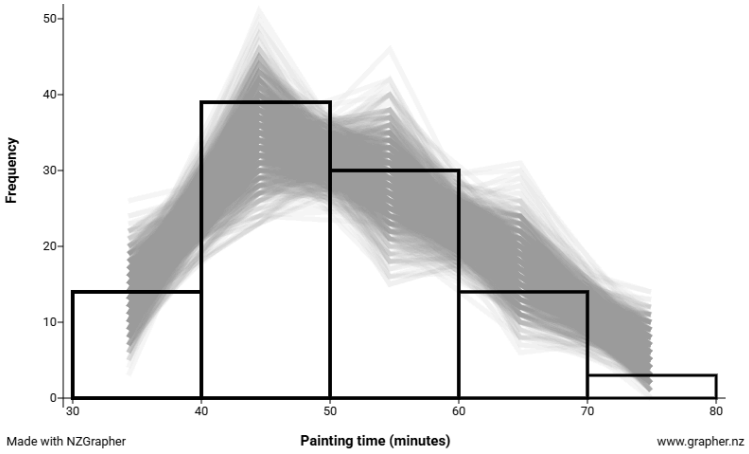
We are going to use the **Painting time** data that you have already copied into NZGrapher to run a simulation model.

Data Type: **Numeric – continuous,**

Simulated Distributions: **Triangular**,

Parameters: min = 30, max = 80, mode = 45.

You should get a graph similar to this.

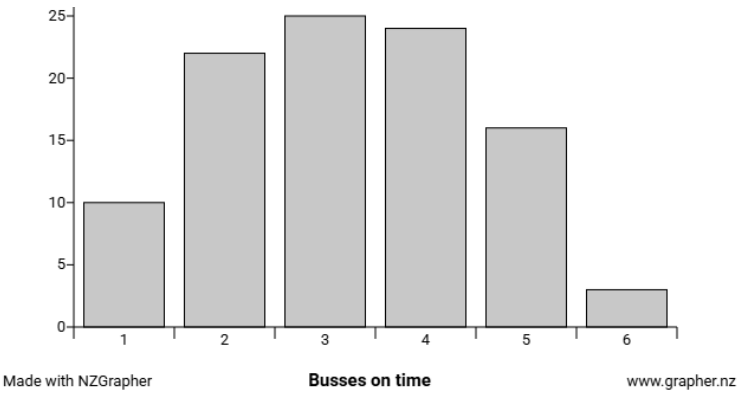


What conclusion(s) can you draw from this graph? Justify using statistical reasoning.

Exercise:

There are 8 buses that bring students to school each day. A student claims that buses only arrive on time 25% of the time. Data was collected on the number that arrived on time and is shown in the table below.

Number of buses on time	Frequency
1	8
2	24
3	28
4	22
5	15
6	3



Step 1: The data is discrete.

Step 2: Model distribution Binomial parameters $p = 0.25$, $n = 8$.

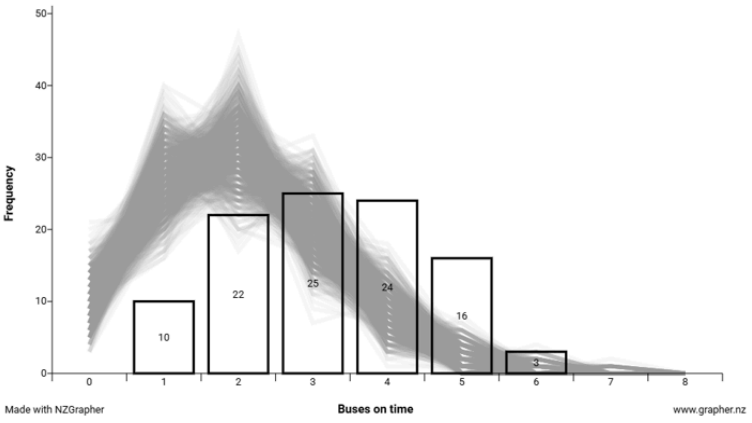
We are going to use the **Busses on time** data that you have already copied into NZGrapher to run a simulation model.

Data Type: **Numeric – Whole number** and

Simulated Distributions: **Binomial**.

Parameters: $p = 0.25$, $n = 8$.

Looking at the graph, the data only starts at 1 and goes up to 5, while the Binomial distribution has a minimum of 0 and a maximum of 8. To change the x axis, select the **More Options** button (bottom right corner), and under **Dot Plots**, change the **Min** to 0 and change the **Max** to 8, and click on the **Update Graph** button. Then run the simulation.



You should get a graph similar to this.

What conclusion(s) can you draw from this graph? Justify using statistical reasoning.

NZQA Level 3 Probability Distribution Exams

The concept of random variation comes up in these exams. Here are two examples:

2024 Question 3(b)(iii):

The sample data shown in Figure 3(a) is run through a simulation model 1000 times, assuming that a Poisson distribution with $\lambda = 2.8$ is used to model the number of belly button washes per 7-day week for the general population. Figure 3(b) shows the results of the simulation model and the original observed data (blue dots).

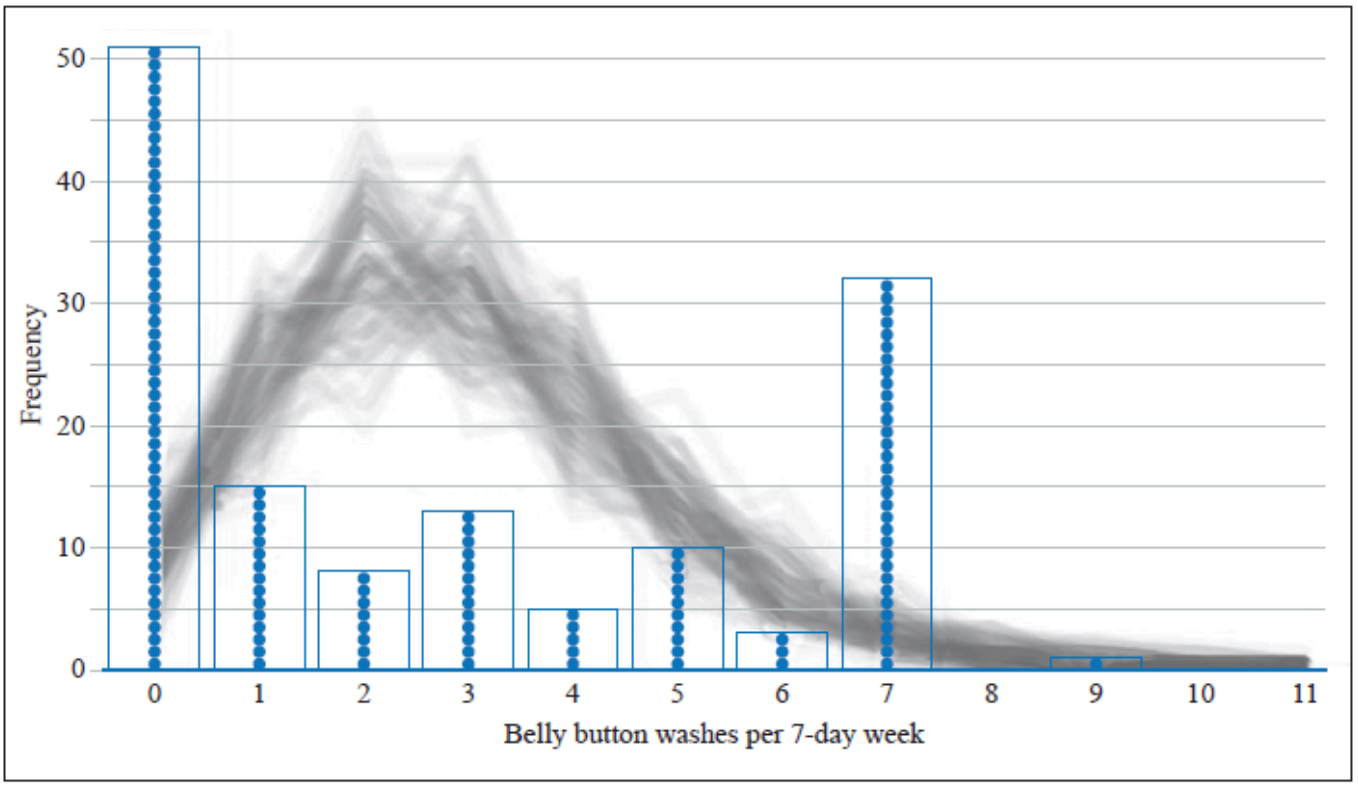


Figure 3(b): Results of the simulation model and original observed data

(iii) Explain what the grey band is showing in Figure 3(b).

(iv) Based on the results of the simulation model and the original observed data (Figure 3(b)), discuss whether the Poisson distribution model (presented opposite) appears to be appropriate for modelling the number of belly button washes per 7-day week.

Support your answer with statistical reasoning.

(v) Propose and justify the use of an alternative model that could be appropriate for modelling the number of belly button washes per 7-day week.

You should state the probability distribution model and the parameters used as part of your answer.

Answers:

(iii) Each grey line shows the expected outcome of one simulation of 138 randomly generated values from a Poisson distribution model with $\lambda = 2.8$. The grey band is made up of 1000 runs of these simulated results. This shows visually the sampling variation that can occur in the model.

(iv) The simulation shows evidence that the proposed Poisson model is not appropriate to model the number of belly button washes per week. Each number of washes is not appearing in the frequency where variation due to natural variation can explain any discrepancies, as they are not "hidden" in the grey cloud of uncertainty. There is an excess frequency of 'zeros' and 'sevens' – a large proportion of people wash their belly buttons on average once per day, or never! The Poisson distribution does not model this aspect of the observed data.

(v) **Uniform distribution**

$a = 0$

The minimum number of belly button washes is zero.

$7 \leq b \leq 9$

There is no data for more than 9 belly button washes per week.

A uniform model could be appropriate because there is a clear minimum and maximum number of belly button washes per week.

Although the observed data does not appear to be uniformly distributed, taking sampling variation into account, it is possible that the true distribution of belly button washes per week is uniform.

The uniform model assumes that all probabilities are equally likely so the frequencies of each number of belly button washings would be similar.

In the observed data the frequency of 1 to 5 belly button washes is reasonably similar. So, the uniform model is less likely to overpredict the frequency of these belly button washes, unlike the current model which has significantly overpredicted them.

A uniform model may not fit the frequency of the number of 0 and 7 belly button washes, however.

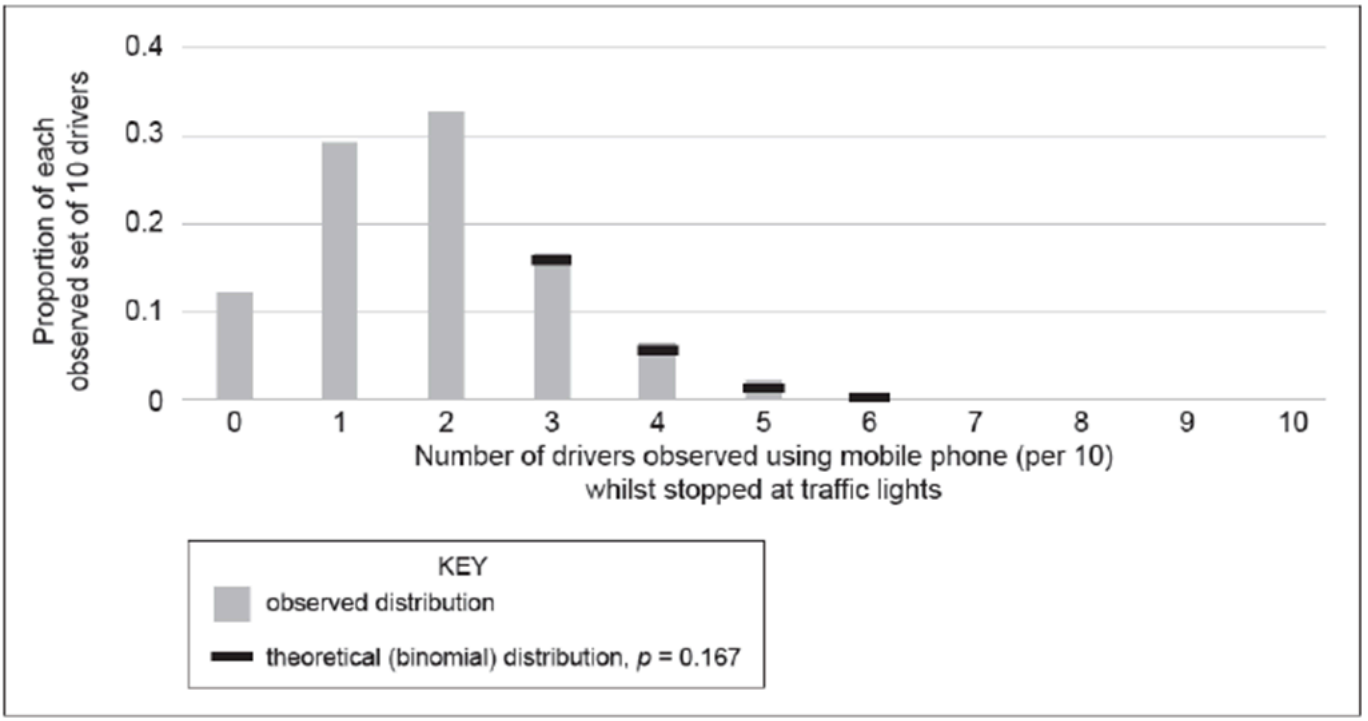
Key concepts:

The key concept is understanding the role that random variation plays in fitting models to data and assessing the fit of that model.

2021 Question 1 (c)(i)

A recent New Zealand study asked drivers to respond anonymously about their mobile phone habits whilst driving. 16.7% of people admitted to using their mobile phone while stopped at a traffic light. A Canterbury researcher was interested to see if the behaviour of drivers on Memorial Avenue was similar to drivers in this New Zealand study.

(i) The graph below shows the observed distribution (grey bars) and a binomial distribution with p=0.167, the theoretical model distribution (shown in black).



Complete the graph by showing the remaining values for the theoretical binomial distribution model.

(ii) Discuss TWO features of the observed and theoretical distributions that justify the suitability of this binomial distribution model for the number of drivers (out of 10) observed using their mobile phone.

Answers:

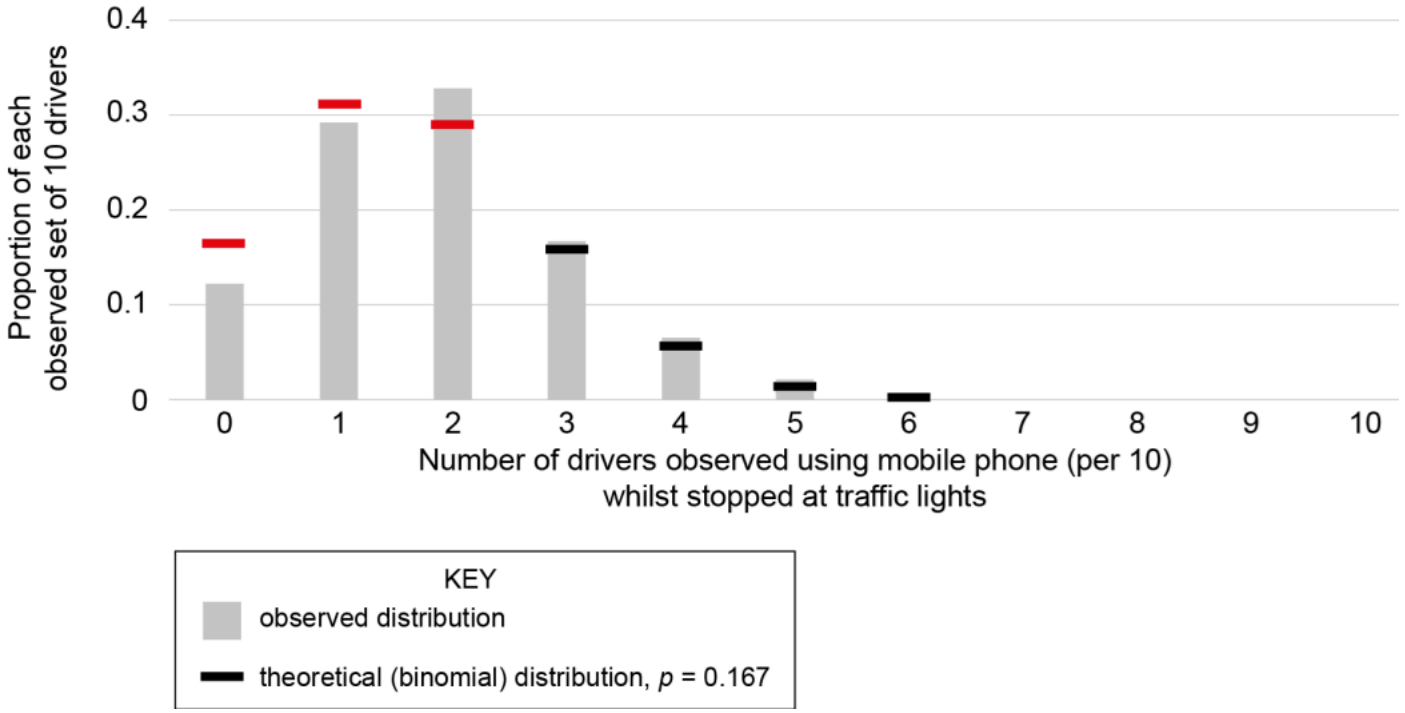
Theoretical probabilities for binomial distribution model using parameters $n = 10$ and $p = 0.167$

These theoretical probabilities are plotted on the graph

$P(X = 0) = 0.161$

$P(X = 1) = 0.322$

$P(X = 2) = 0.291$



(ii) **Discussion:** features of the observed distribution and the theoretical binomial distribution are compared.

Examples of possible comparisons

- The probability of the theoretical distribution for 0 and 1 drivers is more than the probabilities for the observed distribution, while the probability for the theoretical distribution for 2 drivers using mobile phones while stopped at traffic lights is less than the probability for the observed data. Both distributions have very similar probabilities for 3 or more drivers.
- The modal value for the theoretical distribution is 1 driver using a mobile phone while stopped at traffic lights. This is less than the observed distribution which is 2 drivers using a mobile phone.
- The mean of the theoretical distribution is 1.67 drivers using their mobile phone while stopped at traffic lights. This is slightly lower than the mean of the observed distribution which is approximately 1.83 drivers.
- Both the theoretical and observed distributions of the number of drivers using their mobile phones while stopped at traffic lights are positively or right skewed.

Conclusion: the binomial model appears to be a good fit for this situation on Memorial Avenue.

Key concepts:

The key concept is understanding the role that random variation plays in fitting models to data and assessing the fit of that model. When assessing the fit of a model, consider the data type, the shape, the centre, the spread, the parameters and the probabilities.

ANSWERS

Exercise 1:

Question 1:

From the graph we can see that the grey band does NOT cover the experimental probabilities of male $(P(\text{Male}) = 0.27)$ and female sea turtles $(P(\text{Female}) = 0.73)$.

Based on the model that each sex (male/female) is equally likely (p = 0.5), we expect that half the sea turtles would be female and half would be male.

Even taking random variation into account, the proportion of female sea turtles is higher than we expect by chance alone, and the proportion of male sea turtles is lower than we expect by chance alone.

This means that it is NOT likely to get results like these if the sex of sea turtles was equally likely. Therefore, it is much MORE likely that the proportion of female sea turtles is higher than the proportion of male sea turtles.

Question 2

From the graph we can see that the grey band does cover the experimental probability of parking spaces 1, 2 and 4 being used equally, however it does NOT cover the experimental probabilities of parking spaces 3 $(P(3) = 0.37)$ and parking space 5 $(P(5) = 0.07)$.

Based on the model that each parking space (1, 2, 3, 4, 5) is equally likely (p = 0.2), we expect that the probabilities are close to this value.

Even taking random variation into account, the proportion of times that parking space 3 is used is higher than we expect by chance alone, and the proportion of times that parking space 5 is used is lower than we expect by chance alone.

This means that it is NOT likely to get results like these if the parking spaces were equally likely to be used. Therefore, it is much MORE likely that the proportion of times that parking space 3 is used is higher than all other parking spaces, and the proportion of times that parking space 5 is used is lower than all other parking spaces.

Exercise 2:

Question 1

We can see that the overall shape of both the observed data and the Triangular model are right skewed. In addition, the centre of the observed data (mode = 45) is the same as the Triangular model. The spread of both the data and the Triangular model are similar.

When we compare the probabilities, we notice that the probability for paint times for the observed data is within the grey band for the Triangular model. This means that the probabilities are within the range that we could expect due to random chance.

Therefore, the Triangular distribution with parameters $a = 30, b = 80, c = 45$ IS a good fit to this data.

Question 2

The simulation shows that the students' claim of buses being on time 25% of the time is not likely to be valid. We can see that the overall shape of the observed data is approximately normal, while the shape of the Binomial model is right skewed. In addition, the centre of the observed data is further to the right (mode = 3) compared with the centre of the Binomial model is smaller (mode = 2). The spread of both the data and the Binomial model are similar.

When we compare the probabilities, we notice that the probability for 3 buses being late, and 6 buses being late are similar to the Binomial model. However, the probability of one or two buses being late is much lower than we expected, and the probability of 4 or 5 buses is much higher than we expected when compared to the Binomial model.

Therefore, the Binomial distribution with parameters p = 0.25, n = 8 is NOT a good fit to this data.