Maths Applications: Univariate Data Investigation

Leith Wilkerson

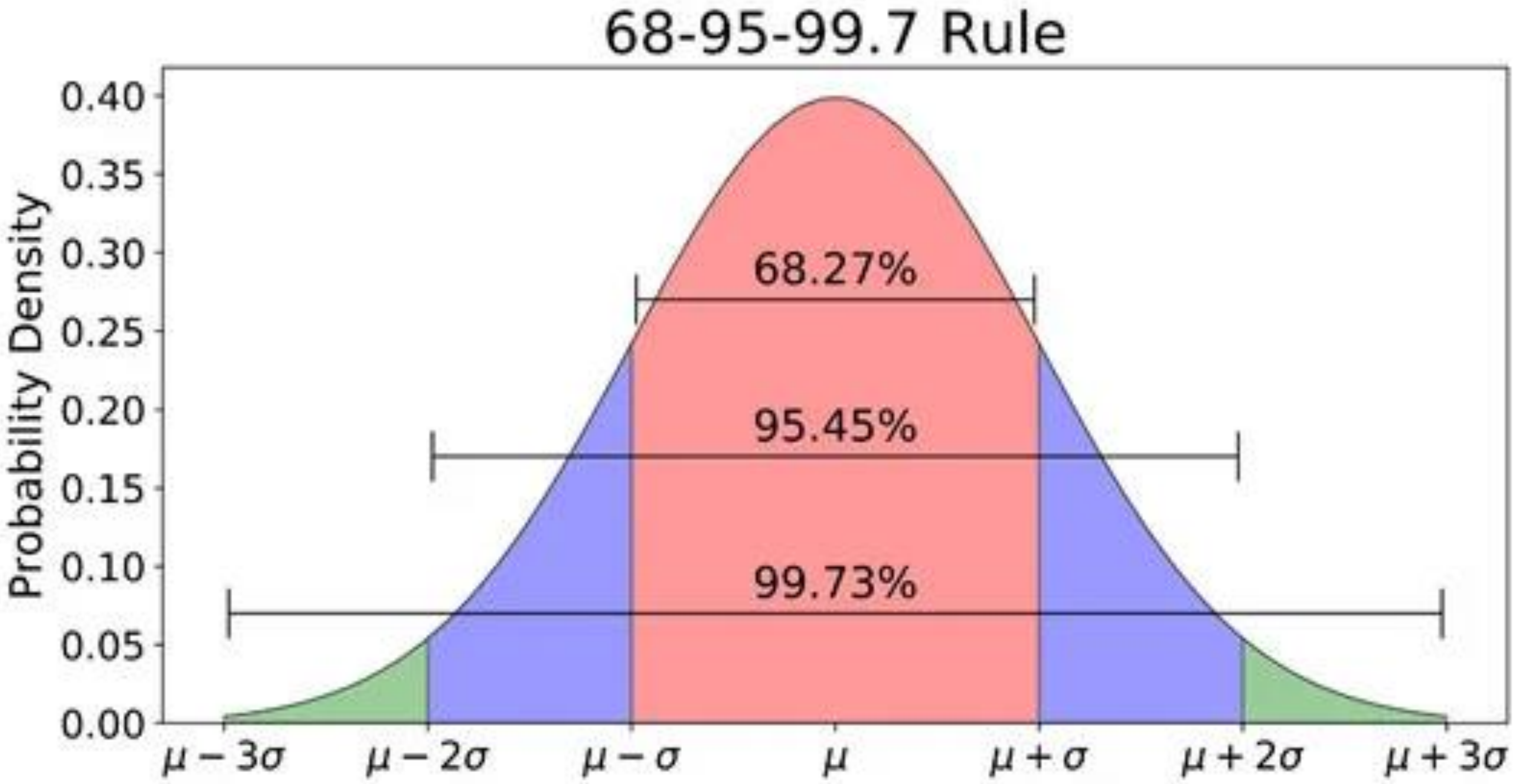
INTRODUCTION

- Normal distribution is a highly helpful tool in statistical analysis, being characterised by its symmetrical, bell curve and has the mean, median and mode all being equal.
- The normal distribution curve follows the "68%, 95%, 99.7% rule", showing that 68%, 95% and 99.7% of data values lie between one, two and three standard deviations (STDEV) from the mean, respectively. Based off this, you can approximate the exact probability of an event occurring in certain normally distributed situations.
- In this investigation, 6 examples of primary and/or secondary, continuous data will be thoroughly analysed to explore the wide uses of the normal distribution. Note that for each example, a sample size of at least 30 is needed to maintain accuracy of results.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation
N = the size of the population
x_i = each value from the population
μ = the population mean

- Mean (μ) = (sum of data) / (number of data sets)
- Median = middle number in numerically ordered data (find mean if number of data sets is even)
- Mode = most reoccurring number in data set
- Standard deviation (σ) = measure of spread: sqrt(($\sum$((data value) - μ)^2) / (number of data sets))


68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ

- For each example displayed in this investigation, the full data set, mean, STDEV and normally distributed bell curve is displayed (keep note that both the median and the mode are the same value as the mean).

- For each curve, the mean is positioned roughly near the centre, while the labelled intervals are split by the STDEV value. This is done to simplify the process of calculating the probability of data in each example falling between a certain number of standard deviations.

- By performing these calculations, you can observe if the probabilities meet near the 68%, 95%, 99.7% rule, determining how "normal" the data set is for each example.

- The following probability notation is used to represent a normally distributed curve with respect to the mean and STDEV value to easily compare data following the same mean and standard deviation to determine how "normally distributed" the data is.

- During the analyses of each data set, standardised score (Z or Z-score) will also be calculated using the following formula. A standard score is a number expressing a certain data value as a number of STDEVs above / below the mean. This is done to identify any potential outliers in a data set and explain any sorts of characteristics of the data set.

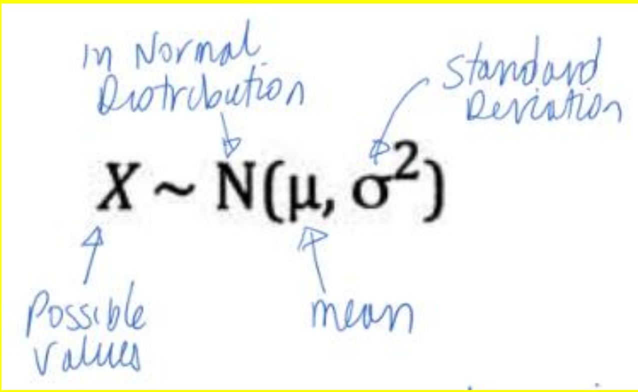
$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

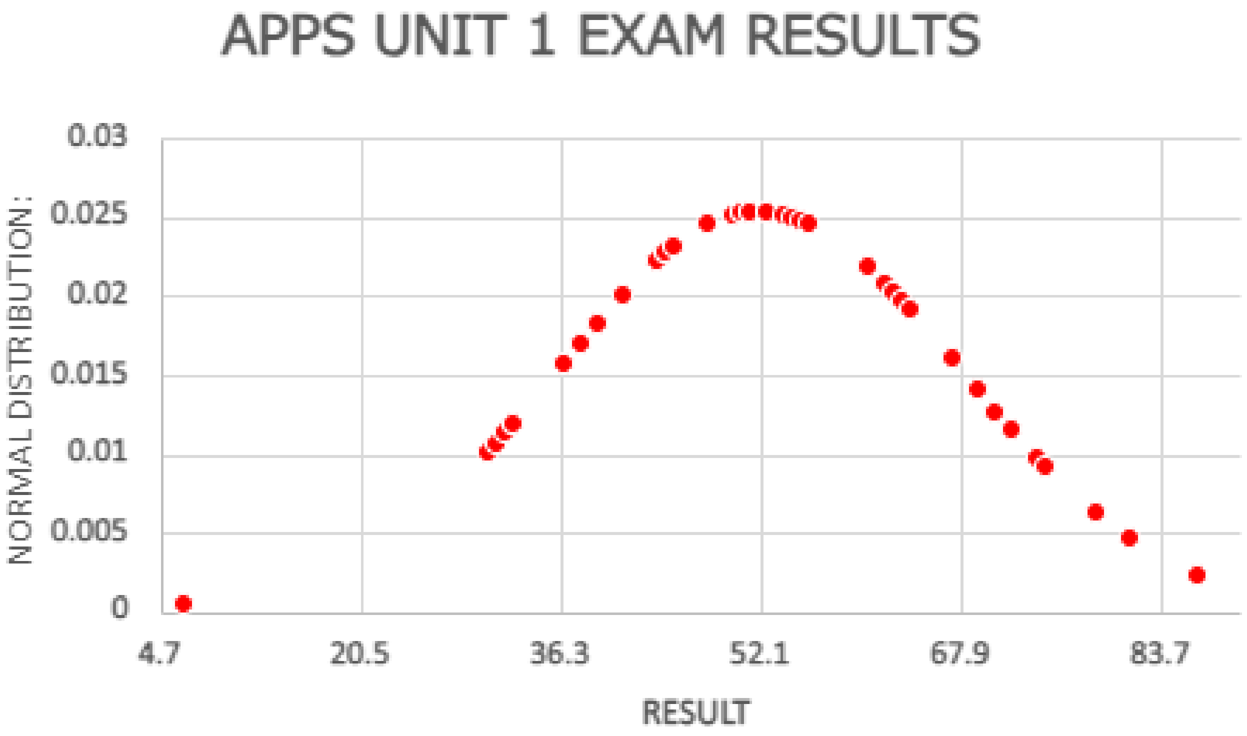
x = observed value

μ = mean of the sample

σ = standard deviation of the sample



(1) EXAM RESULTS



STUDENT:	MARK:	NORMAL DISTRIBUTION:
1	6.7	0.000407
2	30.7	0.01009
3	31.3	0.010615
4	32	0.011242
5	32.7	0.011882
6	32.7	0.011882
7	36.7	0.015702
8	38	0.016956
9	39.3	0.018186
10	39.3	0.018186
11	41.3	0.019989
12	44	0.02214
13	44	0.02214
14	44	0.02214
15	44.7	0.022627
16	44.7	0.022627
17	45.3	0.023016
18	48	0.024414
19	50	0.025027
20	50	0.025027
21	50.7	0.025151
22	50.7	0.025151
23	51.3	0.025217
24	51.3	0.025217
25	52.7	0.025231
26	52.7	0.025231
27	54	0.025068
28	54.7	0.02491
29	55.3	0.024737
30	56	0.024492
31	60.7	0.021773
32	62	0.020749
33	62.7	0.020161
34	63.3	0.01964
35	64	0.019014
36	67.3	0.015896
37	69.3	0.013961
38	70.7	0.012628
39	72	0.011423
40	74	0.009662
41	74.7	0.009078
42	78.7	0.006121
43	81.3	0.004577
44	86.7	0.002296

MEAN:	STANDARD DEVIATION:
52.1	15.8

OF DATA = **44**

OF DATA WITHIN $((\mu-\sigma) \leq X \leq (\mu+\sigma)) = 30 \Rightarrow 30/44 =$ **68.18%**

OF DATA WITHIN $((\mu-2\sigma) \leq X \leq (\mu+2\sigma)) = 42 \Rightarrow 42/44 =$ **95.45%**

EXAM RESULTS ANALYSIS

- The calculations in the previous slide showed that for the exam results data, 68.18% (30/44) of the data values were within one STDEV from the mean ($36.3 \leq X \leq 67.9$), and 95.45% (42/44) of the values were within two STDEV's from the mean ($20.5 \leq X \leq 83.7$).
- This means the data was -0.09% off perfect accuracy from a normal distribution curve showing that 68.27% of data lies within one STDEV from the mean, and that there was perfect accuracy from a normal distribution curve showing that 95.45% of data lies with two STDEVs from the mean.
- Based off this information, it's safe to assume this data set is highly distributed normally.
- There is a significant outlier in the data set, being a 6.7%, meanwhile the closest data values from it is a small cluster of low 30%'s. Being more precise, there is a 24% difference from the outlier and the second-lowest data value. This outlier may have resulted from a major complication during the middle of taking the exam, or didn't bother properly trying.
- There is a major cluster near the mean of the data set, of values from 50-56%, making up 12/44 of data values in the set. This is very common for normally distributed data, especially considering how close from the mean it's located.

LOWEST SCORE = 6.7: Z = (6.7 - 52.1) / 15.8 = -2.87

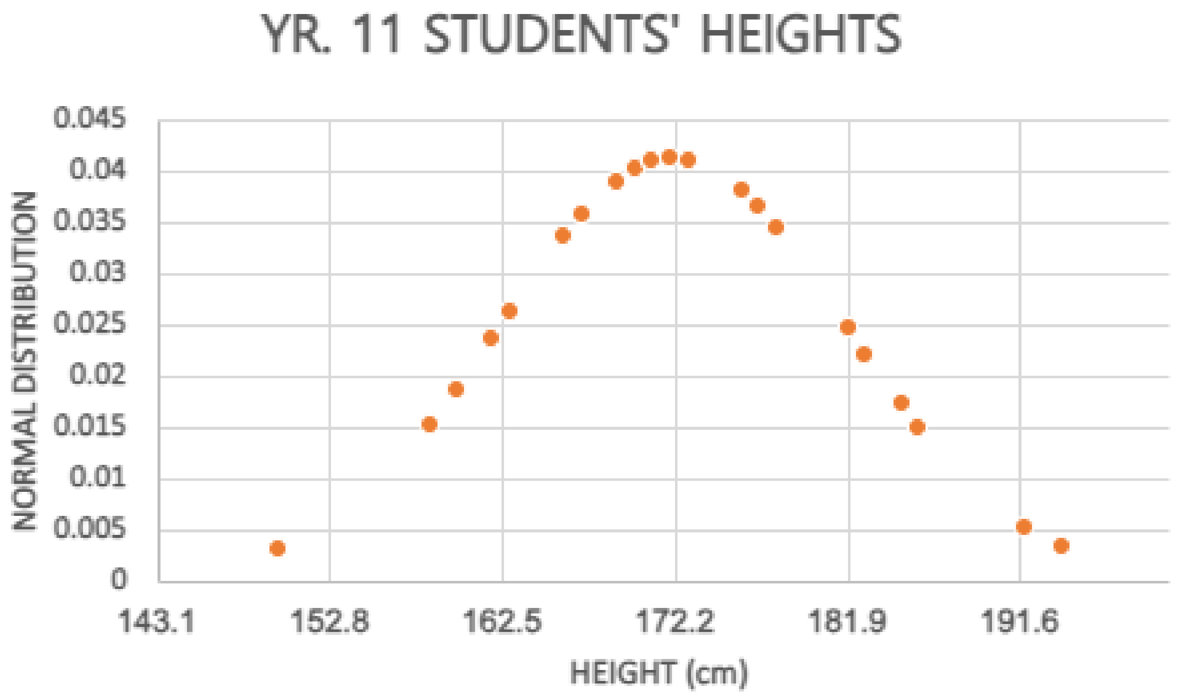
HIGHEST SCORE = 86.7: Z = (86.7 - 52.1) / 15.8 = 2.19

-2.87 represents the standard score for 6.7, the lowest score in the data set, showing this data is significantly low compared to the rest of the data set, having minimal other values near it.

2.19 represents the standard score for 86.7, the highest score in the data set, showing this data is very high compared to the rest of the data set, but wouldn't be considered an outlier.

In summary, the analysis of the exam results showcases how difficult the entire cohort that took the exam believed it was. With a mean being barely above a passing mark, and 18/44 (40.91%) failing the exam, it's safe to assume the exam was overall extremely difficult.

(2) STUDENTS' HEIGHTS



STUDENT:	HEIGHT:	NORMAL DISTRIBUTION:	Henry	171	0.040815
Lottie	150	0.002997	Liam	171	0.040815
Tayla	158.5	0.01517	Georgia M	172	0.041119
Georgia L	160	0.018648	Megan	173	0.040988
Annabelle	162	0.023661	Gemma	173	0.040988
Lola	162	0.023661	Lucas	173	0.040988
Kirsty	162	0.023661	Hemroo	176	0.03809
Oliver	163	0.02623	Zac	177	0.036389
Erin	166	0.033529	Leith	178	0.034396
Rebecca	166	0.033529	Deegan	182	0.024688
Lana	167	0.035623	Tayla M	182	0.024688
Holly	167	0.035623	Kale	183	0.022128
Chloe	169	0.03895	Beau	183	0.022128
Horshan	169	0.03895	Ryder	185	0.017219
Matt	170	0.040084	Bailey	186	0.01495
Alyssa	170	0.040084	Zavier	192	0.005121
Sophie	171	0.040815	Tom	194	0.003291

MEAN:	STANDARD DEVIATION:
172.2	9.7

 # OF TOTAL DATA = 33

 # OF DATA WITHIN ((μ-σ) ≤ X ≤ (μ+σ)) = ⇒ 19/33 = 57.58%

 # OF DATA WITHIN ((μ-2σ) ≤ X ≤ (μ+2σ)) = 31 ⇒ 31/33 = 93.94%

STUDENTS' HEIGHTS ANALYSIS

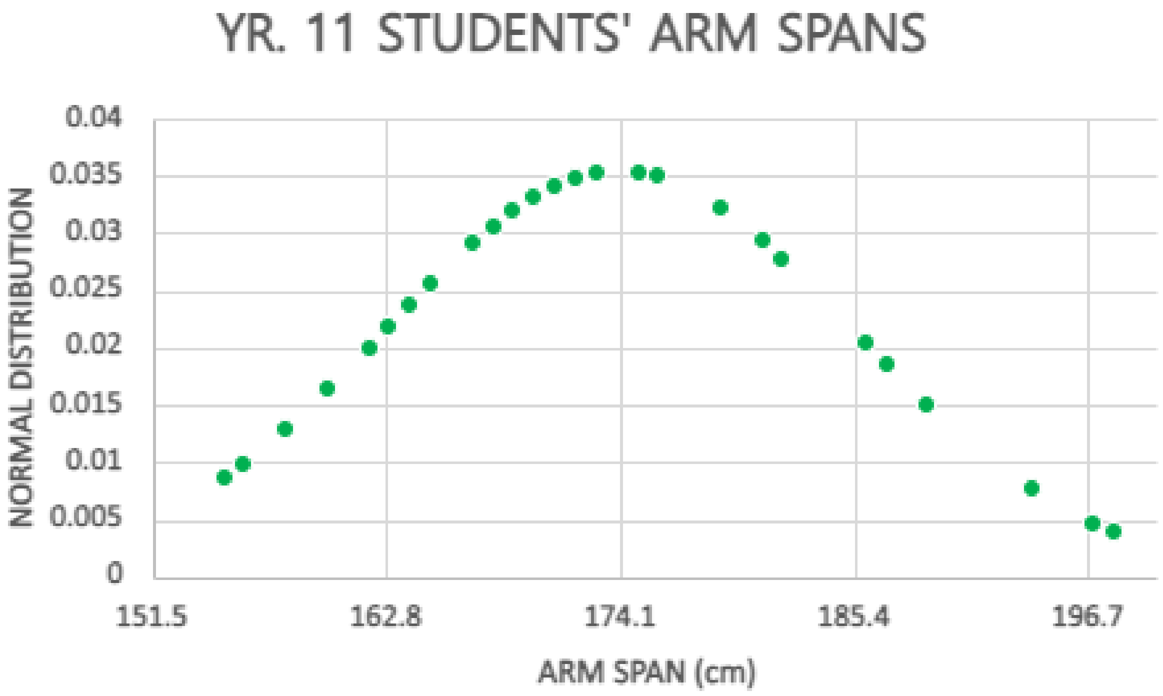
- The calculations in the previous slide showed that for the students' heights data, 57.58% (19/33) of the data values were within one STDEV from the mean ($162.5 \leq X \leq 181.9$), and 93.94% (31/33) of the values were within two STDEV's from the mean ($152.8 \leq X \leq 191.6$).
- This means the data was –10.69% off perfect accuracy from a normal distribution curve showing that 68.27% of data lies within one STDEV from the mean, and that it was also +1.51% off perfect accuracy from a normal distribution curve showing that 95.45% of data lies with two STDEVs from the mean.
- Based off this, you can infer the data set is decently close from being a normal distribution.
- There is a minor outlier, being 150cm. There is a significant 8.5cm difference between that value and the second-lowest value, the largest difference between neighbouring data values.

LOWEST SCORE = 150: Z = (150 – 172.2) / 9.7 = -2.29

HIGHEST SCORE = 194: Z = (194 – 172.2) / 9.7 = 2.25

- -2.87 represents the standard score for 6.7, the lowest score in the data set, showing this data is significantly low compared to the rest of the data set, having minimal other values near it.
- 2.19 represents the standard score for 86.7, the highest score in the data set, showing this data is very high compared to the rest of the data set, but wouldn't be considered an outlier.
- In summary, this analysis of Year 11 students' heights shows how spread out heights can be in people of the same age range (range = 194 – 150 = 44cm), but also showcases the fact that majority of heights lie near the mean, hence its normally distributed characteristics.

(3) STUDENTS' ARM SPANS



STUDENT:	ARM SPAN:	NORMAL DISTRIBUTION:			
Lottie	155	0.008461406	Georgia M	172	0.034700206
Tayla	156	0.00978823	Horshan	172	0.034700206
Georgia L	158	0.012794518	Matt	172	0.034700206
Annabelle	160	0.016208358	Lucas	173	0.035137748
Lola	162	0.019899836	Gemma	175	0.035192827
Kirsty	163	0.021792296	Liam	176	0.034809078
Oliver	164	0.023678561	Megan	176	0.034809078
Holly	165	0.025527394	Leith	179	0.032136654
Chloe	167	0.028980463	Zac	181	0.029299954
Lana	167	0.028980463	Kale	182	0.027650204
Sophie	168	0.030517784	Deegan	186	0.020277401
Rebecca	169	0.03188596	Vincent	186	0.020277401
Alyssa	170	0.033055585	Bailey	187	0.018400849
Maan	170	0.033055585	Hemroo	189	0.014800812
Henry	171	0.034000792	Ryder	189	0.014800812
Erin	171	0.034000792	Beau	194	0.00748832
			Zavier	197	0.004529286
			Tom	198	0.003770878

MEAN:	STANDARD DEVIATION:
174.1	11.3

 # OF TOTAL DATA = 33
 # OF DATA WITHIN ((μ-σ) ≤ X ≤ (μ+σ)) = 20 ⇒ 20/33 = 60.61%
 # OF DATA WITHIN ((μ-2σ) ≤ X ≤ (μ+2σ)) = 31 ⇒ 31/33 = 93.94%

STUDENTS' ARM SPANS ANALYSIS:

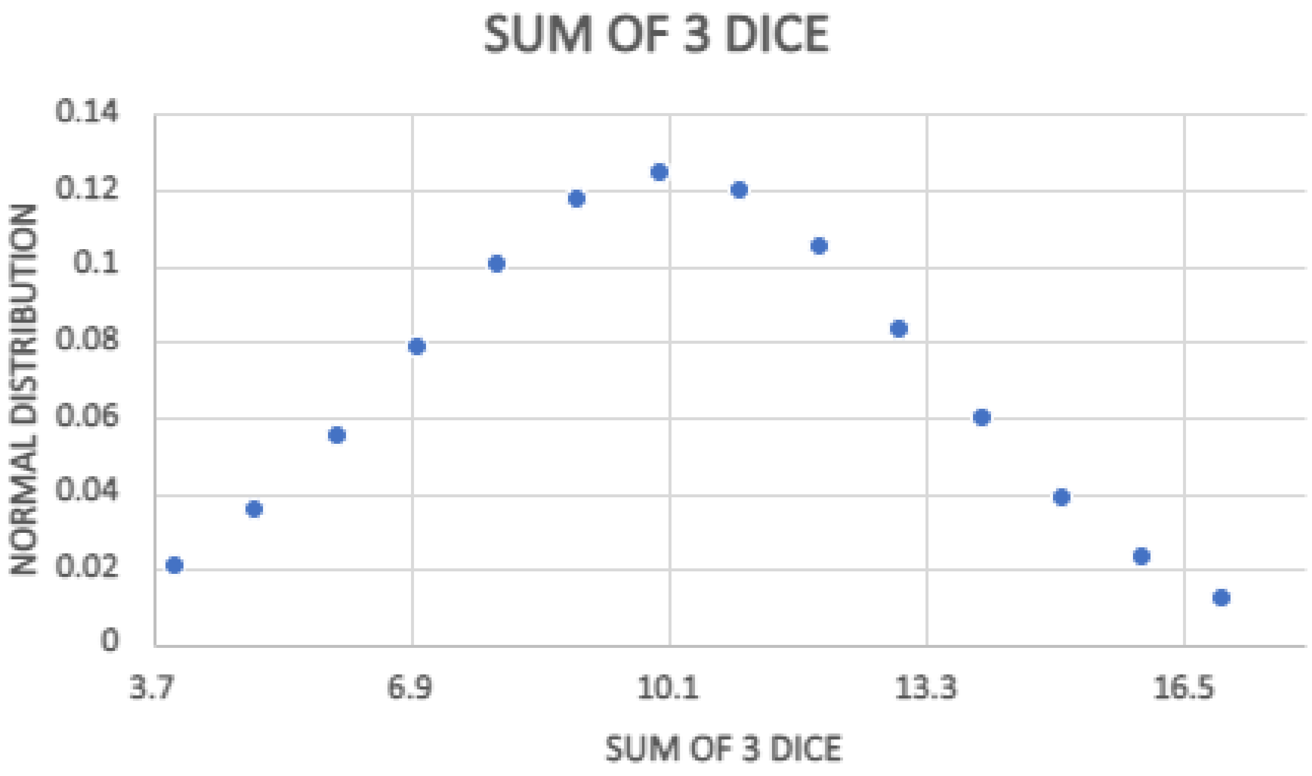
- The calculations in the previous slide showed that for the students' arm spans data, 60.61% (20/33) of the data values were within one STDEV from the mean ($162.8 \leq X \leq 185.4$), and 93.94% (31/33) of the values were within two STDEV's from the mean ($151.5 \leq X \leq 196.7$).
- This means the data was –7.66% off perfect accuracy from a normal distribution curve for 68.27% of data lying within one STDEV from the mean, and that it was also +1.51% off perfect accuracy from a normal distribution curve showing that 95.45% of data lies with two STDEVs from the mean.
- There are no outliers, significant gaps nor clusters in this data set.
- Based off this information, you can infer this data is decently close from being normally distributed.

LOWEST SCORE = 155: Z = (155 – 174.1) / 11.3 = -1.69

HIGHEST SCORE = 198: Z = (198 – 174.1) / 11.3 = 2.11

In summary, this analysis of Year 11 students' arm spans goes to show the significant range in lengths when surveying people in the same age range (range = 198 – 155 = 43cm). The ranges from the 68%, 95% rule could've been closer as well in the sample size was bigger; 33 is barely above the minimum of 30. With there being no outliers, significant gaps nor any clusters in data and these conditions, this data set is a very fairly normal distribution.

(4) SUM OF 3 DICE



SUM OF 3 DICE:	NORMAL DISTRIBUTION:
4	0.02026249
5	0.035010444
5	0.035010444
5	0.035010444
6	0.054864426
6	0.054864426
7	0.077978072
7	0.077978072
7	0.077978072
7	0.077978072
7	0.077978072
8	0.100517707
8	0.100517707
9	0.117517106
9	0.117517106
9	0.117517106
10	0.124608604
10	0.124608604
10	0.124608604
10	0.124608604
11	0.119834923
11	0.119834923
11	0.119834923
11	0.119834923
11	0.119834923
11	0.119834923
12	0.104521878
12	0.104521878
12	0.104521878
12	0.104521878
12	0.104521878
13	0.082683611
13	0.082683611
13	0.082683611
13	0.082683611
14	0.059322589
15	0.038601945
15	0.038601945
16	0.02278173
17	0.012194181

MEAN:	STANDARD DEVIATION:
10.1	3.2

 # OF TOTAL DATA = 41

 # OF DATA WITHIN ((μ-σ) ≤ X ≤ (μ+σ)) = 30 ⇒ 30/41 = 73.17%

 # OF DATA WITHIN ((μ-2σ) ≤ X ≤ (μ+2σ)) = 40 ⇒ 40/41 = 97.56%

SUM OF 3 DICE ANALYSIS:

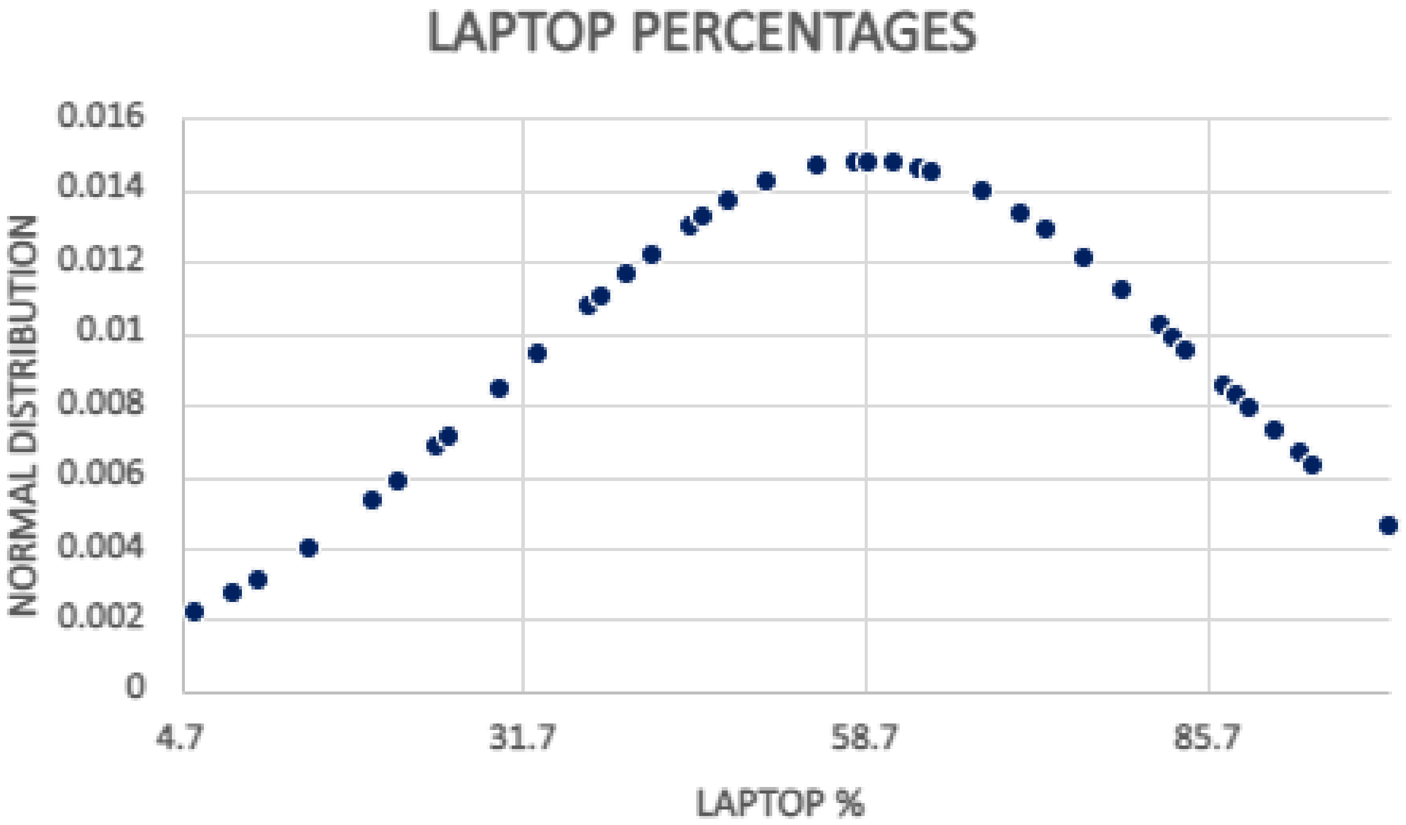
- The calculations in the previous slide showed that for the sum of 3 dice data, 73.17% (30/41) of the data values were within one STDEV from the mean ($6.9 \leq X \leq 13.3$), and 97.56% (40/41) of the values were within two STDEV's from the mean ($3.7 \leq X \leq 16.5$).
- This means the data was +4.9% off perfect accuracy from a normal distribution curve for 68.27% of data lying within one STDEV from the mean, and that it was also +2.11% off perfect accuracy from a normal distribution curve showing that 95.45% of data lies with two STDEVs from the mean.
- Considering this is data coming from a theoretically possible range of (3 – 18), there are no outliers nor gaps in the data set.
- Based off this, you can assume this has very close to normally distributed data.

LOWEST SCORE = 4: Z = (4 – 10.2) / 3.2 = -1.94

HIGHEST SCORE = 17: Z = (17 – 10.2) / 3.2 = 2.13

In summary, this data set of sum of 3 dice is an accurate representation of the mathematical probabilities of the outcomes for finding the sum of throwing 3 dice, but could be improved by increasing the sample size by a lot. By doing so, this data set is met with a perfect normal distribution, otherwise this data still has a distribution very close to normal.

(5) LAPTOP PERCENTAGES



LAPTOP %:	NORMAL DISTRIBUTION:	61	0.014722127
6	0.002199251	63	0.014589442
9	0.002715067	64	0.014493696
11	0.00310317	64	0.014493696
15	0.003987557	68	0.013924626
20	0.005289679	71	0.013319304
22	0.005866082	73	0.012842065
25	0.006780447	73	0.012842065
26	0.00709638	76	0.012033634
30	0.008398346	79	0.011137739
33	0.009393039	82	0.01018206
37	0.0106975	83	0.009855009
38	0.011013161	84	0.009525388
40	0.011624764	84	0.009525388
42	0.012203189	87	0.00853071
42	0.012203189	88	0.008200265
45	0.012990895	88	0.008200265
46	0.013228264	89	0.007871814
48	0.013659763	91	0.007224062
48	0.013659763	93	0.006593336
51	0.014186837	94	0.006285987
55	0.014637552	100	0.004586405
58	0.014770675	100	0.004586405
59	0.014774728	100	0.004586405

MEAN:	STANDARD DEVIATION:
58.7	27

 # OF TOTAL DATA = 47
 # OF DATA WITHIN ((μ-σ) ≤ X ≤ (μ+σ)) = 28 ⇒ 28/47 = 59.57%
 # OF DATA WITHIN ((μ-2σ) ≤ X ≤ (μ+2σ)) = 47 ⇒ 47/47 = 100%

LAPTOP PERCENTAGES ANALYSIS:

- The calculations in the previous slide showed that for the laptop percentages data, 59.51% (28/47) of the data values were within one STDEV from the mean ($31.7 \leq X \leq 81.7$), and 100% (47/47) of the values were within two STDEV's from the mean ($4.7 \leq X \leq 112.7$).
- This means the data was −8.76% off perfect accuracy from a normal distribution curve for 68.27% of data lying within one STDEV from the mean, and that it was also +4.55% off perfect accuracy from a normal distribution curve showing that 95.45% of data lies within two STDEVs from the mean.
- Keep in mind, in this context the data may be innacurate of a normal distribution, since the maximum possible laptop percentage is 100%, less than within two STDEVs from the mean. This means for there to be any data values more than within two STDEVs from the mean, it must be between 0 – 4.7%, which much more than 95.45% of data wouldn't be able to reach.
- Apart from this, the data would be considered fairly normally distributed.

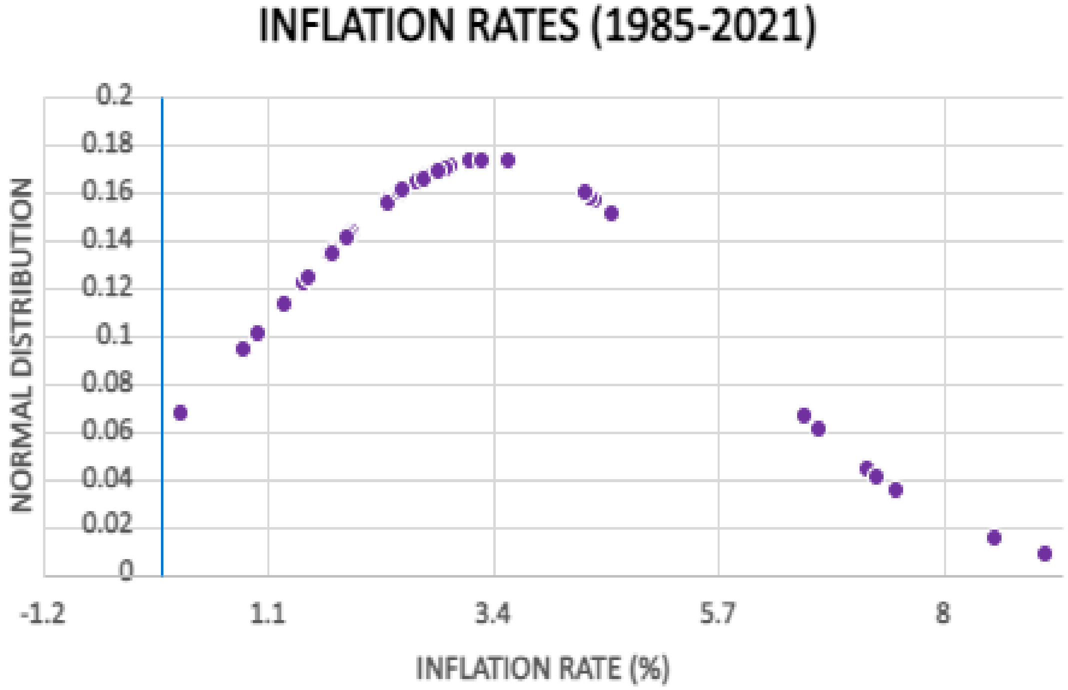
LOWEST SCORE = 6: Z = (6 – 58.7) / 27 = -1.95
HIGHEST SCORE = 100: Z = (100 – 58.7) / 27 = 1.53

- In summary, this data set of students' laptop percentages may be considered a fairly innacurate representation of the normal distribution, since the possible values are only limited to between 0 – 100%, and only the values between 0 – 4.7% can be within two STDEVs from the mean. (For a normal distribution curve, it's calculated that when $X \sim N(58.7, 27^2)$, $P(0 \leq X \leq 4.7) = 0.79\%$, which may be helpful to know when solving some complications.)
- However, there are no outliers nor gaps in data, and the trend of a normal distribution is followed, so this data may still be used as a representation of normally distributed data to an extent.

(6) INFLATION RATES IN AUSTRALIA (1985-2021)

MEAN:	STANDARD DEVIATION:
3.4	2.3

YEAR:	INFLATION RATES (%):	NORMAL DISTRIBUTION:
1985	6.73	0.060813234
1986	9.05	0.00848804
1987	8.53	0.014418152
1988	7.22	0.043670105
1989	7.53	0.034595522
1990	7.33	0.040289352
1991	3.18	0.172661487
1992	1.01	0.101090039
1993	1.75	0.134099556
1994	1.97	0.142969004
1995	4.63	0.150341805
1996	2.62	0.163760156
1997	0.22	0.066693281
1998	0.86	0.094265528
1999	1.48	0.12242231
2000	4.46	0.155976795
2001	4.41	0.157510143
2002	2.98	0.170585162
2003	2.73	0.166247645
2004	2.34	0.155976795
2005	2.69	0.165382524
2006	3.56	0.173033975
2007	2.33	0.155663093
2008	4.35	0.159270683
2009	1.77	0.134933607
2010	2.92	0.169716718
2011	3.3	0.173289298
2012	1.76	0.134517207
2013	2.45	0.159270683
2014	2.49	0.160394644
2015	1.51	0.123752057
2016	1.28	0.113420949
2017	1.95	0.142192762
2018	1.91	0.140620997
2019	0.85	0.093813109
2020	2.86	0.168737826
2021	6.59	0.066292941



 # OF TOTAL DATA = 37

 # OF DATA WITHIN ((μ-σ) ≤ X ≤ (μ+σ)) = 26 ⇒ 26/37 = 70.27%

 # OF DATA WITHIN ((μ-2σ) ≤ X ≤ (μ+2σ)) = 35 ⇒ 35/37 = 94.59%

REFERENCE: https://www.macrotrends.net/countries/AUS/australia/inflation-rate-cpi

INFLATION RATES ANALYSIS:

- The calculations in the previous slide showed that for the Australian inflation rates data, 70.27% (26/37) of the data values were within one STDEV from the mean ($1.1 \leq X \leq 5.7$), and 94.59% (35/37) of the values were within two STDEV's from the mean ($-1.2 \leq X \leq 8$).
- This means the data was –2% off perfect accuracy from a normal distribution curve for 68.27% of data lying within one STDEV from the mean, and that it was also % off perfect accuracy from a normal distribution curve showing that –0.86% of data lies within two STDEVs from the mean.
- While in the middle-left of it there is a small cluster of data between 4.35 - 4.63%, there is a massive gap in data between 3.56 - 6.59% (3.03% range), which may be caused by external inflation behaviours.
- Overall however, you can infer that this data set is still a very accurate representation of the normal distribution.

LOWEST SCORE = 0.22: Z = (0.22 – 3.4) / 2.3 = -1.38

HIGHEST SCORE = 9.05: Z = (9.05 – 3.4) / 2.3 = 2.46

- In summary, this data sets of Australian inflation rates (1985-2021) showcases a very accurate representation of the normal distribution, with the only setback being the large gap taking up more than a STDEV's length of the graph (3.03 > 2.3). This data showcases how the inflation rates of Australia, and most likely as well as most other countries, would be over the course of many decades, being very spontaneous.
- Range = (9.05 - 0.22) = 8.83%

CONCLUSION

- The purpose of this investigation is to explore many different kinds of scenarios, from inflation rates to the sum of dice, that may or may not share data following a similar normally distributed trend.
- This is done to get a clear idea of how many kinds of situations the normal distribution would appear, and through deep analysis, we can see that it can be found almost anywhere.
- Due to how often the normal distribution would appear in our everyday lives, there are many applications of it, ranging from being used as a guideline for scaling WACE subjects, so that getting a final mark in a certain WACE subject can be as fair as possible, to resource allocation.