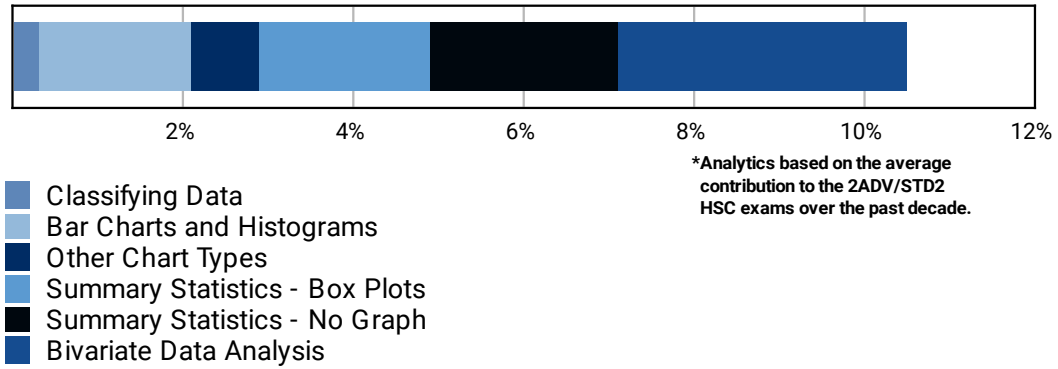
ADVANCED MATHEMATICS
Statistics (Adv), S2 Interpretation and Bivariate Data (Adv)

Classifying Data (Y12)
Bar Charts and Histograms (Y12)
Other Chart Types (Y12)
Summary Statistics - No graph (Y12)
Summary Statistics - Box Plots (Y12)

Teacher: Cathyanne Horvat
Exam Equivalent Time: 90 minutes *(based on allocation of 1.5 minutes per mark)*

S2 Interpretation and Bivariate Data



HISTORICAL CONTRIBUTION

- *S2 Interpretation and Bivariate Data* is a Year 12 topic that didn't previously existed in the Advanced course, although it has a decade long history in the Std2/Gen2 exam.
- *S2 Interpretation and Bivariate Data* has been split into six sub-topics for analysis which are listed in the bar chart above.
- This analysis looks at *Summary Statistics - No Graph (2.2%).*

HSC ANALYSIS - What to expect and common pitfalls

- *Summary Statistics - No Graph (2.2%)* wasn't examined in the 2021 or 2020 Advanced exam, but past *Std2* questions have required students to understand and calculate statistics such as a five-number summary, median, IQR, mean and standard deviation (by calculator) given a simple data set.
- A core competency of calculating a five number summary from a data set is a must, along with understanding mean and median changes when a dataset is adjusted.
- ***Pitfalls:*** The *2019 Std2 exam* tested students on "outlier" calculations after the *2017 Std2 30a* and *2015 Std2 27d* exams exposed a lack of understanding in this area.
- Major issues have been encountered in identifying the IQR of a cumulative frequency table, and finding the mean of grouped data, where students must use the "class centres".

Questions

1. Statistics, STD2 S1 SM-Bank 1 MC

   A survey asked the following question for students born in Australia:

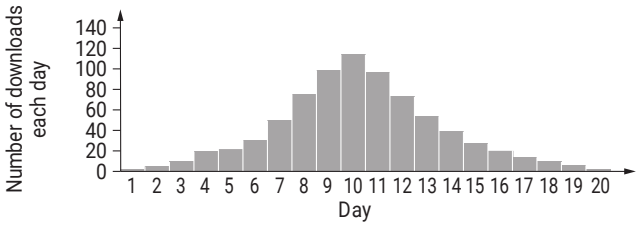
   "Which State or Territory were you born in?"

   How would the responses be classified?

   A. Categorical, ordinal
   B. Categorical, nominal
   C. Numerical, discrete
   D. Numerical, continuous

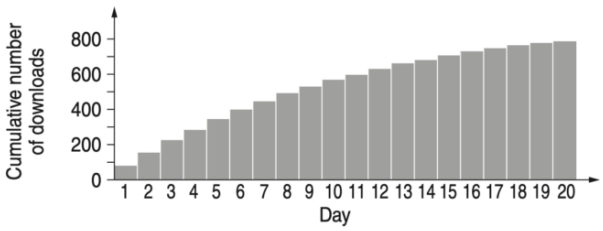
2. Statistics, 2ADV S2 2021 HSC 4 MC

The number of downloads of a song on each of twenty consecutive days is shown in the following graph.

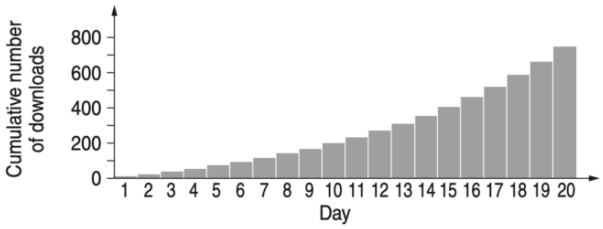


Which of the following graphs best shows the cumulative number of downloads up to and including each day?

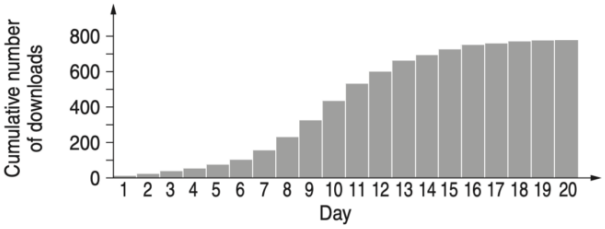
A.



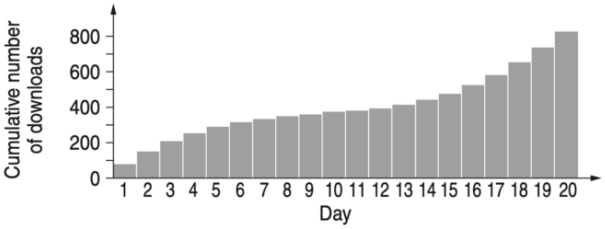
B.



C.

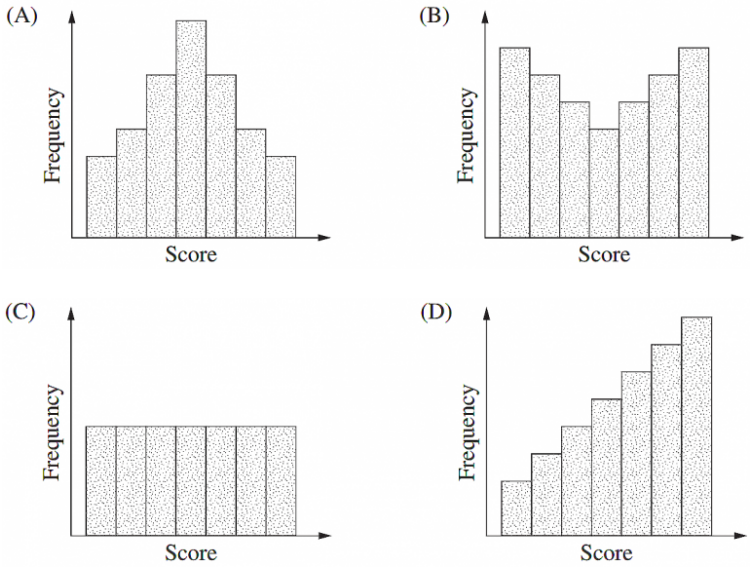


D.



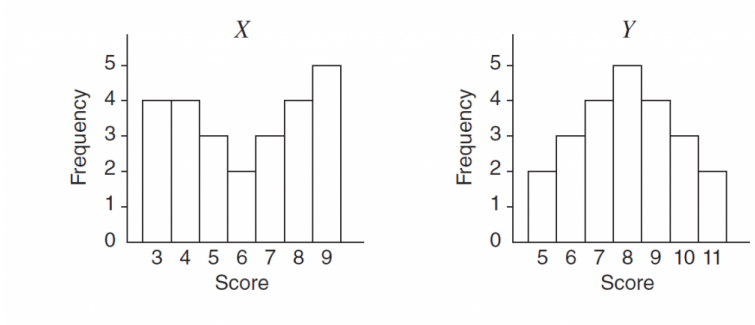
3. Statistics, STD2 S5 2010 HSC 4 MC

Which of the following frequency histograms shows data that could be normally distributed?



4. Statistics, STD2 S1 2011 HSC 11 MC

The sets of data, X and Y, are displayed in the histograms.



Which of these statements is true?

(A) X has a larger mode and Y has a larger range.

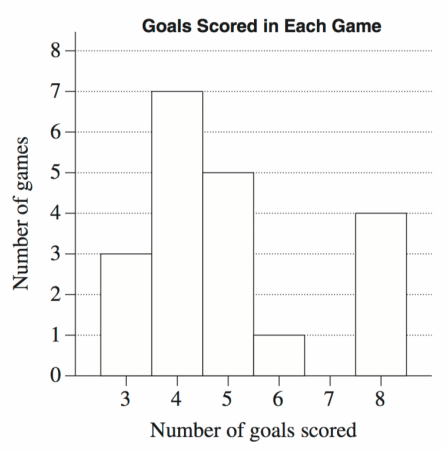
(B) X has a larger mode and the ranges are the same.

(C) The modes are the same and Y has a larger range.

(D) The modes are the same and the ranges are the same.

5. Statistics, STD2 S1 2013 HSC 15 MC

The frequency histogram shows the number of goals scored by a football team in each game in a season.



What is the mean number of goals scored per game by this team?

(A) 4

(B) 4.5

(C) 5

(D) 5.5

---

6. Statistics, STD2 S1 2015 HSC 4 MC

On a school report, a student's record of completing homework is graded using the following codes.

C = consistently
U = usually
S = sometimes
R = rarely
N = never

What type of data is this?

A. Categorical, ordinal

B. Categorical, nominal

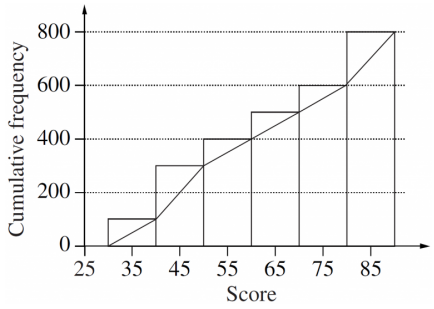
C. Numerical, continuous

D. Numerical, discrete

---

7. Statistics, STD2 S1 2005 HSC 9 MC

A set of data is represented by the cumulative frequency histogram and ogive.



What is the best approximation for the interquartile range for this set of data?

(A) 25

(B) 30

(C) 35

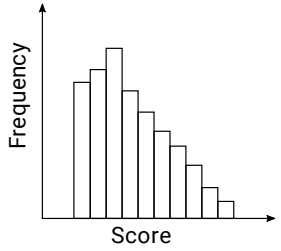
(D) 40

---

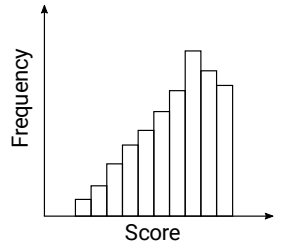
8. Statistics, STD2 S1 2020 HSC 7 MC

Which histogram best represents a dataset that is positively skewed?

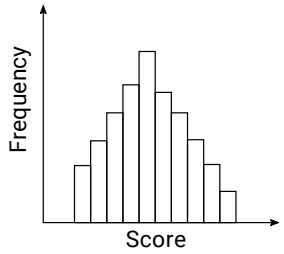
A.



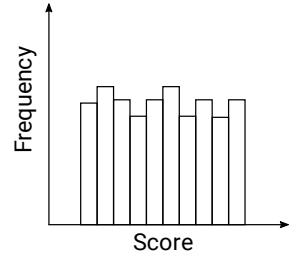
B.



C.

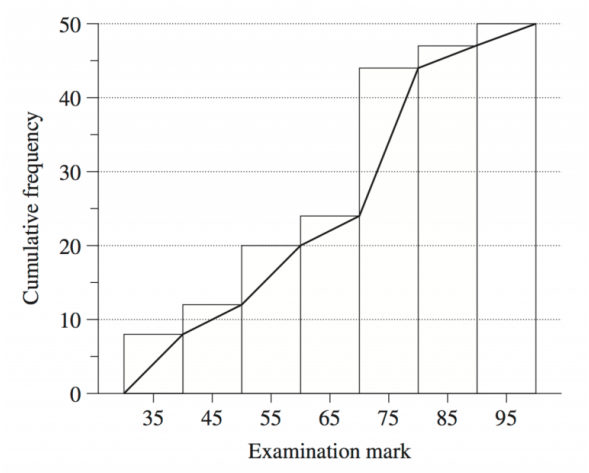


D.



9. Statistics, STD2 S1 2007 HSC 22 MC

A set of examination results is displayed in a cumulative frequency histogram and polygon (ogive).



Sanath knows that his examination mark is in the 4th decile.

Which of the following could have been Sanath's examination mark?

(A) 37

(B) 57

(C) 67

(D) 77

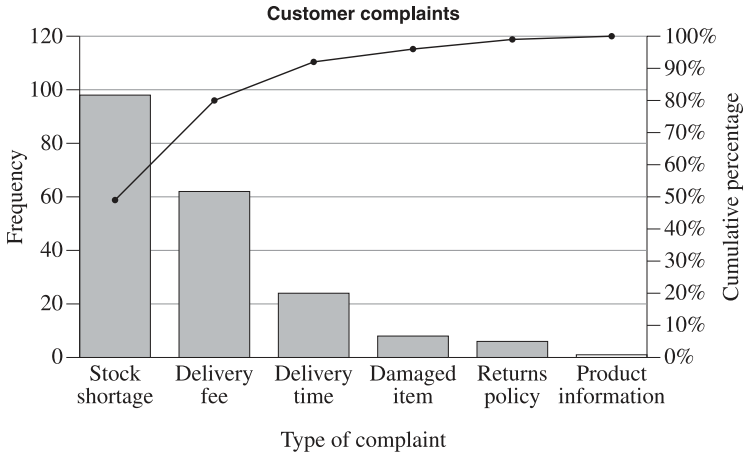
10. Statistics, 2ADV S2 2022 HSC 11

The table shows the types of customer complaints received by an online business in a month.

Type of complaint	*Frequency*	*Cumulative frequency*	*Cumulative percentage*
Stock shortage	98	98	49
Delivery fee	62	***A***	80
Delivery time	24	184	92
Damaged item	8	192	***B***
Returns policy	6	198	99
Product information	2	200	100
Total	200		

a. What are the values of A and B? ***(2 marks)***

b. The data from the table are shown in the following Pareto chart.

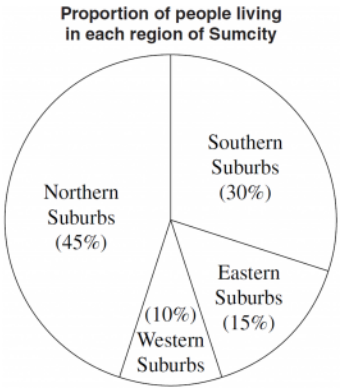


The manager will address 80% of the complaints.

Which types of complaints will the manager address? ***(1 mark)***

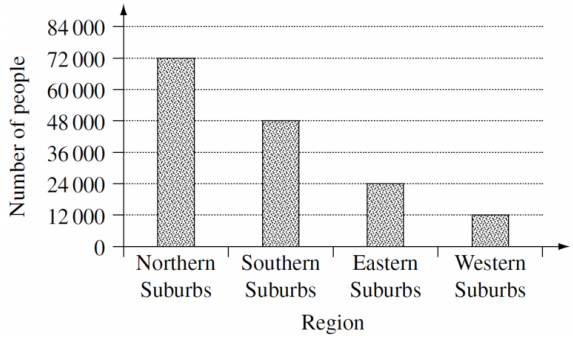
11. Statistics, STD2 S1 2005 HSC 24d

The sector graph shows the proportion of people, as a percentage, living in each region of Sumcity. There are 24 000 people living in the Eastern Suburbs.



i. Show that the total number of people living in Sumcity is 160 000. ***(1 mark)***

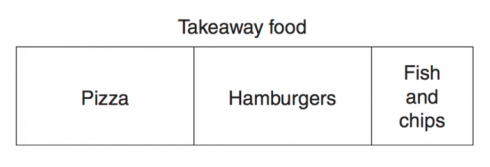
Jake used the information above to draw a column graph.



ii. The column graph height is incorrect for one region.
Identify this region and justify your answer. ***(2 marks)***

12. Statistics, STD2 S1 2008 HSC 23e

In a survey, 450 people were asked about their favourite takeaway food. The results are displayed in the bar graph.



How many people chose pizza as their favourite takeaway food? ***(2 marks)***

13. Statistics, STD2 S1 2019 HSC 19

The heights, in centimetres, of 10 players on a basketball team are shown.

170, 180, 185, 188, 192, 193, 193, 194, 196, 202

Is the height of the shortest player on the team considered an outlier? Justify your answer with calculations. ***(3 marks)***

14. Statistics, STD2 S1 2007 HSC 24d

Barry constructed a back-to-back stem-and-leaf plot to compare the ages of his students.

Ages of students attending Barry's Ballroom Dancing Studio

Females		*Males*
9	1	1 2 3
7	2	0 2 2 2 4 5
5	3	0 0 1 7
5 2	4	6 7
3 2 0	5	2
4 4 2 1	6	4 4

i. Write a brief statement that compares the distribution of the ages of males and females from this set of data. ***(1 mark)***

ii. What is the mode of this set of data? ***(1 mark)***

iii. Liam decided to use a grouped frequency distribution table to calculate the mean age of the students at Barry's Ballroom Dancing Studio.

For the age group 30 - 39 years, what is the value of the product of the class centre and the frequency? ***(2 marks)***

iv. Liam correctly calculated the mean from the grouped frequency distribution table to be 39.5.

Caitlyn correctly used the original data in the back-to-back stem-and-leaf plot and calculated the mean to be 38.2.

What is the reason for the difference in the two answers? ***(1 mark)***

15. Statistics, STD2 S1 2005 HSC 24a

i. Draw a stem-and-leaf plot for the following set of scores.

21 45 29 27 19 35 23 58 34 27 ***(2 marks)***

iii. What is the median of the set of scores? ***(1 mark)***

iv. Comment on the skewness of the set of scores. ***(1 mark)***

16. Statistics, STD2 S1 2015 HSC 27d

In a small business, the seven employees earn the following wages per week:

$300, $490, $520, $590, $660, $680, $970

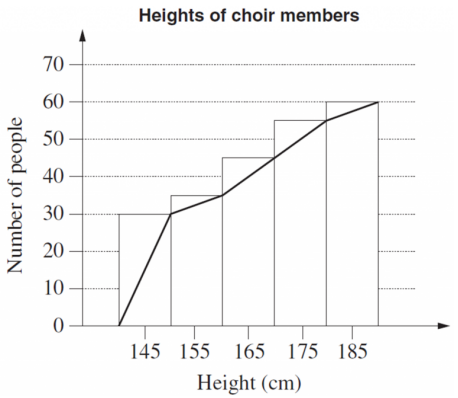
i. Is the wage of $970 an outlier for this set of data? Justify your answer with calculations. ***(3 marks)***

ii. Each employee receives a $20 pay increase.
What effect will this have on the standard deviation? ***(1 mark)***

17. Statistics, STD2 S1 2006 HSC 24c

The heights of the 60 members of a choir were recorded. These results were grouped and then displayed as a cumulative frequency histogram and polygon.

The shortest person in the choir is 140 cm and the tallest is 190 cm.



Draw an accurate box-and-whisker plot to represent the data. ***(3 marks)***

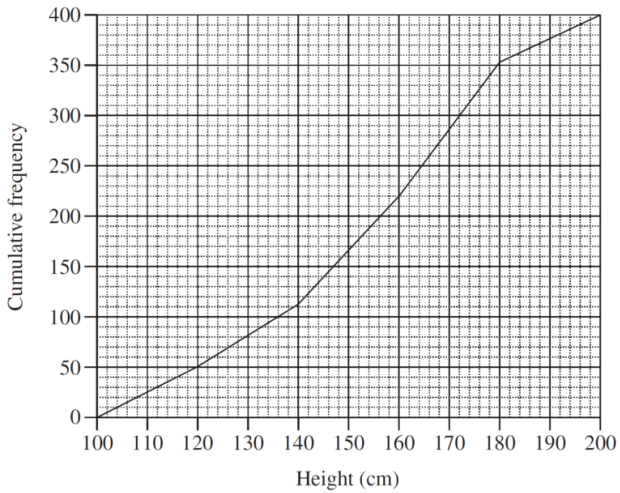
18. Statistics, STD2 S1 2008 HSC 23f

Christina has completed three Mathematics tests. Her mean mark is 72%.

What mark (out of 100) does she have to get in her next test to increase her mean mark to 73%? ***(2 marks)***

19. Statistics, STD2 S1 2016 HSC 27c

The heights of 400 students were measured. The results are displayed in this cumulative frequency polygon.



Use the polygon to estimate the interquartile range. ***(2 marks)***

20. Statistics, STD2 S1 2017 HSC 30a

A set of data has a lower quartile (Q_L) of 10 and an upper quartile (Q_U) of 16.

What is the maximum possible range for this set of data if there are no outliers? ***(2 marks)***

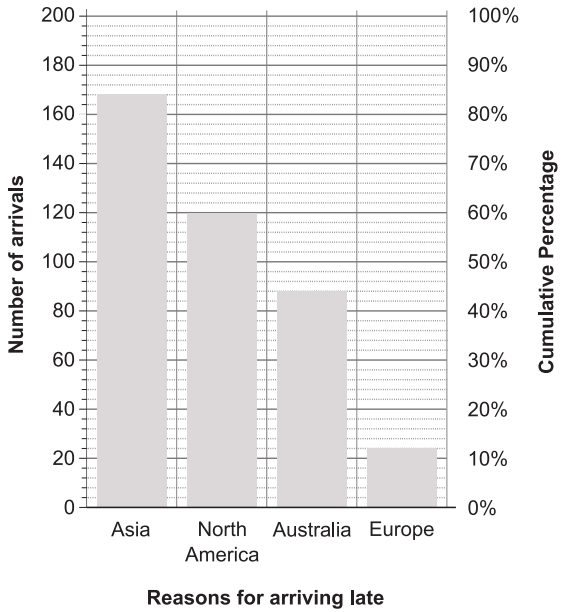
21. Statistics, STD2 S1 EQ-Bank 5

An island resort surveyed 400 guests by asking them on which continent they lived.

The table below shows the data collected.

Continent	Asia	Europe	North America	Australia
Number of guests	168	24	120	88

Complete the Pareto chart below to show the data collected. ***(3 marks)***



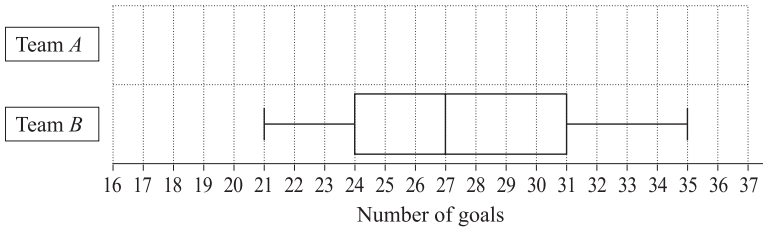
22. Statistics, STD2 S1 2019 HSC 39

Two netball teams, Team A and Team B, each played 15 games in a tournament. For each team, the number of goals scored in each game was recorded.

The frequency table shows the data for Team A.

Number of goals	*Frequency*
19	1
20	0
21	1
22	1
23	1
24	3
25	0
26	4
27	3
28	1

The data for Team B was analysed to create the box-plot shown.



Compare the distributions of the number of goals scored by the two teams. Support your answer with the construction of a box-plot for the data for Team A. ***(5 marks)***

.....

.....

.....

.....

.....

.....

.....

.....

23. Statistics, STD2 S1 2005 HSC 27d

Nine students were selected at random from a school, and their ages were recorded.

Ages		
12	11	16
14	16	15
14	15	14

i. What is the sample standard deviation, correct to two decimal places? ***(2 marks)***

.....

.....

.....

.....

ii. Briefly explain what is meant by the term *standard deviation*. ***(1 mark)***

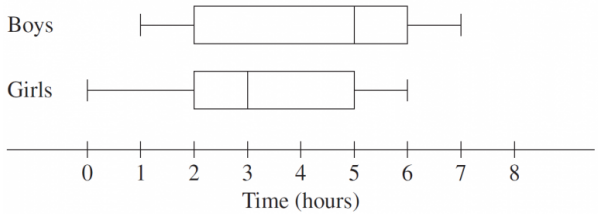
.....

.....

.....

24. Statistics, STD2 S1 2009 HSC 26a

In a school, boys and girls were surveyed about the time they usually spend on the internet over a weekend. These results were displayed in box-and-whisker plots, as shown below.



i. Find the interquartile range for boys. ***(1 mark)***

.....

ii. What percentage of girls usually spend 5 or less hours on the internet over a weekend? ***(1 mark)***

.....

iii. Jenny said that the graph shows that the same number of boys as girls usually spend between 5 and 6 hours on the internet over a weekend.

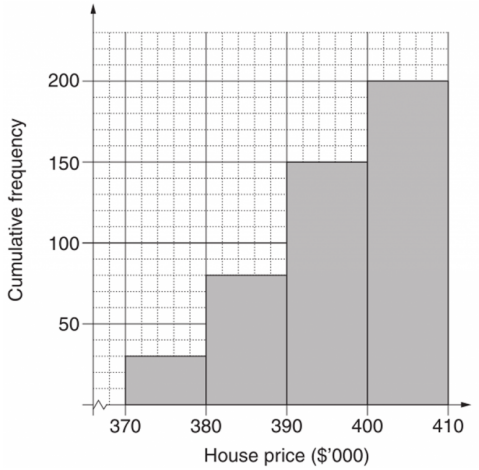
Under what circumstances would this statement be true? ***(1 mark)***

.....

.....

25. Statistics, STD2 S1 2015 HSC 29d

Data from 200 recent house sales are grouped into class intervals and a cumulative frequency histogram is drawn.



i. Use the graph to estimate the median house price. ***(1 mark)***

.....

ii. By completing the table, calculate the mean house price. ***(3 marks)***

Class Centre ($'000)	*Frequency*

.....

.....

Worked Solutions

1. Statistics, STD2 S1 SM-Bank 1 MC

The data is categorical (not numerical) since the name of a State is required.
This data cannot be ordered.
$\Rightarrow B$

2. Statistics, 2ADV S2 2021 HSC 4 MC

The gradient of the cumulative frequency histogram will increase gradually, be steepest at day 10 then decrease gradually.
$\Rightarrow C$

3. Statistics, STD2 S5 2010 HSC 4 MC

Normally distributed data have a frequency histogram graph that is shaped like a bell.
$\Rightarrow A$

4. Statistics, STD2 S1 2011 HSC 11 MC

♦ Mean mark 47%

Mode of $X = 9$
Range of $X = 9 - 3 = 6$
Mode of $Y = 8$
Range of $Y = 11 - 5 = 6$
$\therefore X$ has a larger mode and ranges are the same
$\Rightarrow B$

5. Statistics, STD2 S1 2013 HSC 15 MC

Total number of goals scored

$$= (3 \times 3) + (4 \times 7) + (5 \times 5) + (6 \times 1) + (7 \times 0) + (8 \times 4)$$
$$= 9 + 28 + 25 + 6 + 0 + 32$$
$$= 100$$

Number of games $= 3 + 7 + 5 + 1 + 4 = 20$

$$\therefore \text{ Mean goals per game } = \frac{100}{20} = 5$$

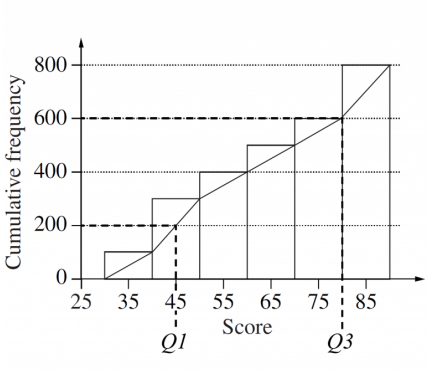
$\Rightarrow C$

6. Statistics, STD2 S1 2015 HSC 4 MC

The data has been grouped into categories and because each category can be ranked, it is ordinal.
$\Rightarrow A$

Worked Solutions

7. Statistics, STD2 S1 2005 HSC 9 MC



$$IQR = Q3 - Q1$$
$$= 80 - 45$$
$$= 35$$

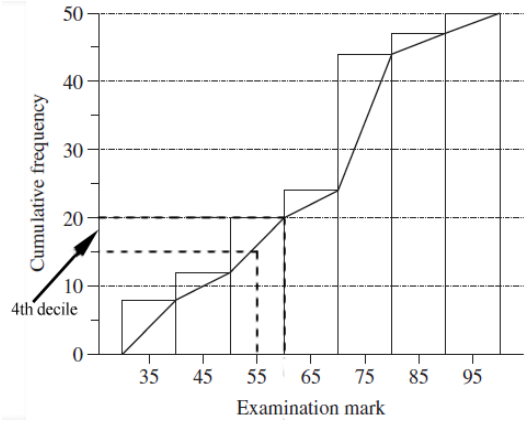
$\Rightarrow C$

8. Statistics, STD2 S1 2020 HSC 7 MC

♦♦ Mean mark 32%.

Positive skew occurs when the tail on the histogram is longer on the right-hand (positive) side.
$\Rightarrow A$

9. Statistics, STD2 S1 2007 HSC 22 MC



4th decile occurs when cumulative frequency is between 15 and 20.
$\therefore$ Examination mark must be between 55 and 60.
$\Rightarrow B$

10. Statistics, 2ADV S2 2022 HSC 11

a. $A = 98 + 62 = 160$

$$\% \text{ Damaged items } = \frac{8}{200} \times 100 = 4\%$$

Cumulative % after damaged items $= 96\%$

$B = 92 + 4 = 96$

b. The right hand side cumulative frequency percentage shows that 80% of all complaints received concern stock shortages and delivery fees.

$\therefore$ The manager will address stock shortages and delivery fees.

---

11. Statistics, STD2 S1 2005 HSC 24d

i. Let the population of Sumcity $= P$

$$15\% \times P = 24\,000$$

$$\therefore P = \frac{24\,000}{0.15}$$

$$= 160\,000 \ldots \text{ as required}$$

ii. Western Suburbs population

$$= 10\% \times 160\,000$$

$$= 16\,000$$

The column graph has this population as 12 000 people which is incorrect.

---

12. Statistics, STD2 S1 2008 HSC 23e

Number of people who chose pizza

$$= \frac{\text{Length of pizza section}}{\text{Total length of bar}} \times 450$$

$$\approx \frac{7}{18} \times 450$$

$$\approx 175$$

$\therefore$ 175 people chose pizza.

COMMENT: This question required measurement of the actual image on the exam. The same methodology works here.

---

13. Statistics, STD2 S1 2019 HSC 19

$Q_1 = 185,\ Q_3 = 194$

$IQR = 194 - 185 = 9$

Shortest player $= 170$

Outlier height:

$$Q_1 - 1.5 \times IQR = 185 - 1.5 \times 9$$

$$= 171.5$$

$\therefore$ Since $170 < 171.5$, 170 is an outlier.

Mean mark 51%.
COMMENT: The last statement must be made to achieve full marks here!

---

14. Statistics, STD2 S1 2007 HSC 24d

i. More males attend than females and a higher proportion of those are younger males, with the distribution being positively skewed. Female attendees are generally older and have a negatively skewed distribution.

ii. Mode $= 64$ (4 times)

iii. Class centre $= \frac{30 + 39}{2}$

$$= 34.5$$

Frequency $= 5$

$\therefore$ Class centre $\times$ frequency

$$= 34.5 \times 5$$

$$= 172.5$$

iv. The difference in the answers is due to the class centres used in group frequency tables distorting the mean value from the exact data.

15. Statistics, STD2 S1 2005 HSC 24a

i.

Stem	Leaf
1	9
2	1 3 7 7 9
3	4 5
4	5
5	8

ii. 10 scores

$$\therefore \text{Median} = \frac{(5\text{th} + 6\text{th})}{2}$$
$$= \frac{27 + 29}{2}$$
$$= 28$$

iii. The data has a tail that stretches to the right

$\therefore$ Data is positively skewed.

---

16. Statistics, STD2 S1 2015 HSC 27d

i. $300, 490, 520, 590, 660, 680, 970$

♦ Mean mark (i) 39%.

$\text{Median} = 590$

$Q_1 = 490$

$Q_3 = 680$

$IQR = 680 - 490 = 190$

Outlier if \$970 is greater than:

$Q_3 + 1.5x \times IQR = 680 + 1.5 \times 190 = \$965$

$\therefore$ The wage \$970 per week is an outlier.

ii. All values increase by \$20, but so too does the mean.
Therefore the spread about the new mean will not change
and therefore the standard deviation will remain the same.

---

17. Statistics, STD2 S1 2006 HSC 24c

Low $= 140$

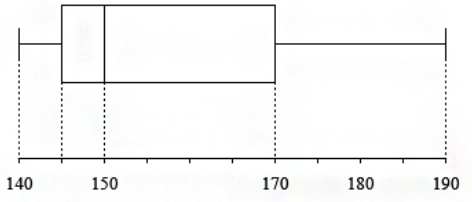
High $= 190$

Median $= 150$ (# People $= 30$)

$Q_1 = 145$ (# People $= 15$)

$Q_3 = 170$ (# People $= 45$)

Box and Whisker



---

18. Statistics, STD2 S1 2008 HSC 23f

Total marks in 3 tests
$= 3 \times 72$
$= 216$

We need 4-test mean $= 73$

i.e. Total Marks (4 tests) $\div 4 = 73$

Total Marks (4 tests) $= 292$

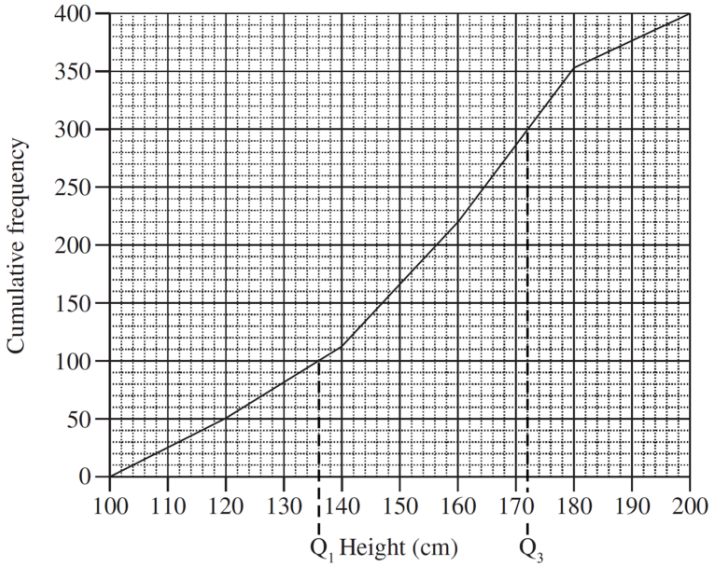
$\therefore$ 4th test score $= 292 - 216$
$= 76$

---

19. Statistics, STD2 S1 2016 HSC 27c

See graph for values:

♦ *Mean mark 46%.*



$$IQR = Q_3 - Q_1$$
$$= 172 - 136$$
$$= 36 \text{ cm}$$

20. Statistics, STD2 S1 2017 HSC 30a

$IQR = 16 - 10 = 6$

If no outliers,

♦♦ *Mean mark 34%.*

$$\text{Upper limit} = Q_U + 1.5 \times IQR$$
$$= 16 + 1.5 \times 6$$
$$= 25$$
$$\text{Lower limit} = Q_L - 1.5 \times IQR$$
$$= 10 - 1.5 \times 6$$
$$= 1$$

$$\therefore \text{ Maximum range} = 25 - 1$$
$$= 24$$

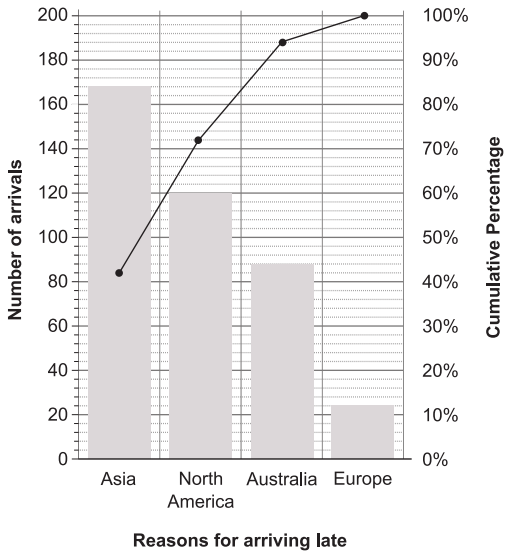
21. Statistics, STD2 S1 EQ-Bank 5

$$\% \text{ Asia} = \frac{168}{400} = 42\%$$

$$\% \text{ North America} = \frac{120}{400} = 30\%$$

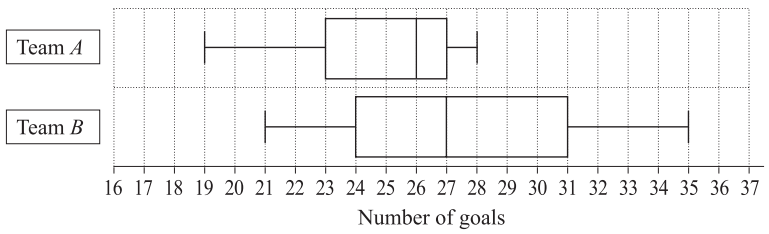
$$\% \text{ Australia} = \frac{88}{400} = 22\%$$

$$\% \text{ Europe} = \frac{24}{400} = 6\%$$



22. Statistics, STD2 S1 2019 HSC 39

Team A: High $= 28$, Low $= 19$, $Q_1 = 23$, $Q_3 = 27$, Median $= 26$



Team A's distribution is negatively skewed while Team B's distribution is slightly positively skewed.
The standard deviation of Team A's distribution is smaller than Team B, as both its IQR and range is smaller.
Team B is a more successful team at scoring goals as each value in its 5-point summary is higher than Team A's equivalent value.

♦♦ *Mean mark 28%.*

---

23. Statistics, STD2 S1 2005 HSC 27d

i. Sample standard deviation

$= 1.6914\ldots$ (by calculator)

$= 1.69$ (to 2 d.p.)

ii. Standard deviation is a measure of how much members of a data group differ from the mean value of the group.

---

24. Statistics, STD2 S1 2009 HSC 26a

i. Interquartile range $= 6 - 2$

$= 4$

ii. Upper quartile $= 5$

$\therefore$ **75% of girls spend 5 or less hours**

♦♦ *Mean mark part ii: 31%*

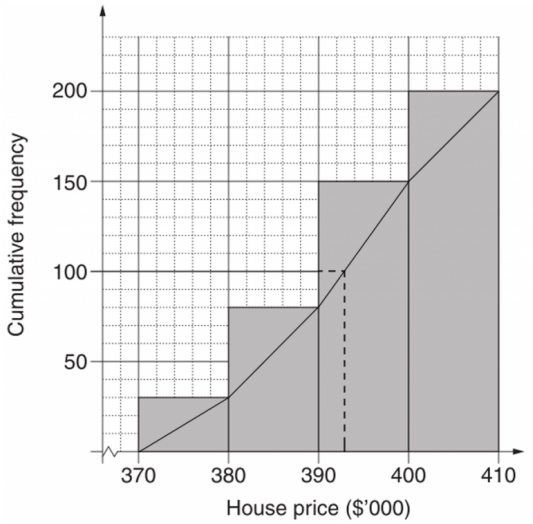
iii. 5-6 hours for girls accounts for 25% of all girls.
5-6 hours for boys accounts for 25% of all boys,
(median to the upper quartile represents 25%.)
$\Rightarrow$ **This will only be the same number if the number of all girls surveyed equals the number of boys surveyed.**

♦♦♦ *Mean mark part iii: 9%*

---

25. Statistics, STD2 S1 2015 HSC 29d

i.



From the graph, the estimated median house price = $392 500

♦♦♦ *Mean marks of 9% for part (i) and 34% for part (ii)!*

ii.

Class Centre ($'000)	*Frequency*
375	30
385	50
395	70
405	50

Mean house price ($'000)

$$= \frac{375 \times 30 + 385 \times 50 + 395 \times 70 + 405 \times 50}{200}$$

$$= \$392$$

$\therefore$ Mean house price is $392 000