SM-1402 Exercise 1

1. The following data represents the heights of 16 students in centimetres.

162	168	177	147	189	171	173	168
178	184	165	173	179	166	168	165

(a) Find the mean and standard deviation of this data.

Solution: $\sum_{i=1}^{n} x_i = 2733$, so $\bar{x} = 2733/16 = 170.8$. Also, $\sum_{i=1}^{n} x_i^2 = 468261$, and thus $s^2 = 468261/16 - (2733/16)^2 = 89.402$, which means that the sample standard deviation is $s = \sqrt{89.402} = 9.46$.

(b) Using equal class interval widths of 10, tabulate the data by dividing it into 5 classes (i.e. groups) between 140 cm and 190 cm.

Solution:

Height (cm)	Frequency
$140 \leq x < 150$	1
$150 \leq x < 160$	0
$160 \leq x < 170$	7
$170 \leq x < 180$	6
$180 \leq x < 190$	2

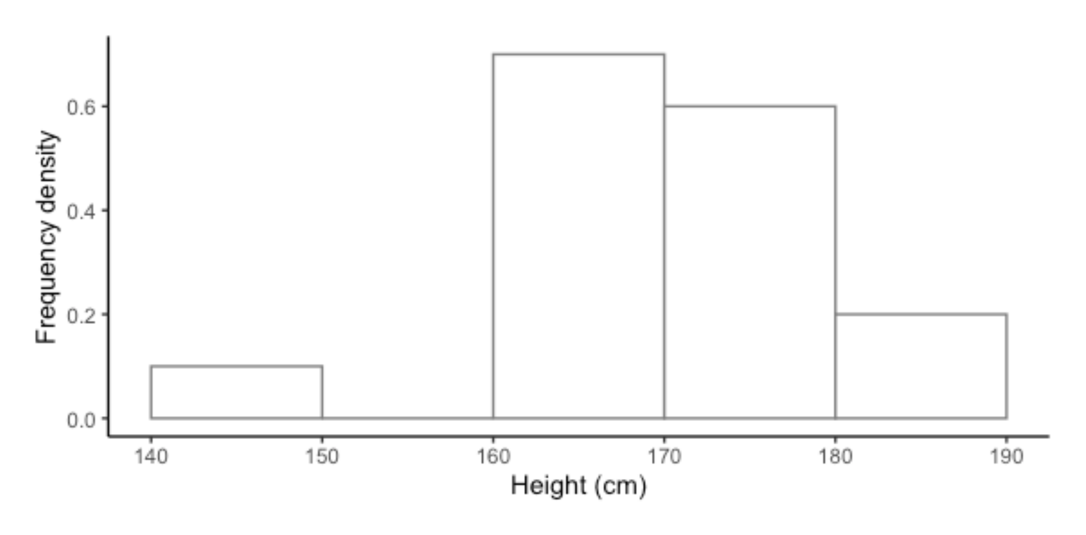
Alternative class intervals that may be used are:

- 140−, 150−, etc.
- 139.5 − 149.5, 149.5 − 159.5, etc.
- $140 < x \leq 150$, $150 \leq x \leq 160$, etc.

Comment: Nowadays when you analyse data using computer software, each software has their own class boundary definition. To be picky about which type of boundary is correct is pretty pointless. Nonetheless, the important thing is to use class boundaries that are 10 in length (as instructed), and to count the frequency correctly depending on the boundaries used.

(c) Draw a histogram for the data.

Solution:



(d) Identify the modal class.

Solution: This is the class corresponding to the highest bar in the histogram, which is $160 \leq x < 170$.

(e) Using the grouped data in (b), find the mean and standard deviation. Compare with the answers obtained in (a) and comment.

Solution: First, figure out what the midpoints of each of the classes are.

Height (cm)	Frequency (f)	Midpoint (m)	$f \times m$	$f \times m^2$
$140 \leq x < 150$	1	145	145	21025
$150 \leq x < 160$	0	155	0	0
$160 \leq x < 170$	7	165	1155	190575
$170 \leq x < 180$	6	175	1050	183750
$180 \leq x < 190$	2	185	370	68450

Then, the statistics of interest are

$$\sum x = \sum (f \times m) = 145 + 0 + 1155 + 1050 + 370 = 2720$$

and

$$\sum x^2 = \sum (f \times m^2) = 145 + 0 + 1155 + 1050 + 370 = 463800$$

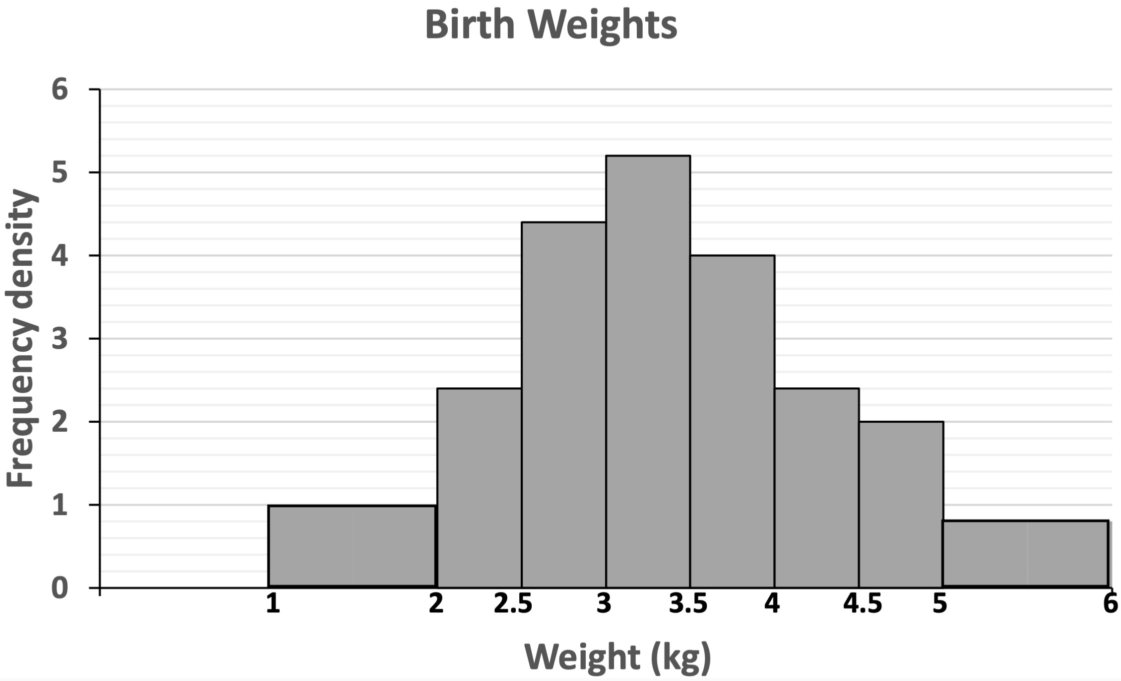
Thus, the mean and variance are

$$\bar{x} = 2720/16 = 170 \qquad s^2 = 463800/16 - 170^2 = 87.5$$

which makes the standard deviation in this case $s = \sqrt{87.5} = 9.35$.

The answers obtained in (b) are different than the ones in (a). We expect the answers in (a) to be more accurate as they come from raw data, while the tabulated data gives an approximation only (using midpoints).

2. The following histogram represents the weights of 60 babies:



6 babies weigh from 4 to 4.5 kg. Calculate the number of babies weighing less than 3 kg.

Solution: Note that the area of a histogram is proportional to the total number of babies.

Weight (kg)	Class width	Freq. dens.	Area
$1.0 \le x < 2.0$	1.0	1.0	$1 \times 1 = 1$
$2.0 \le x < 2.5$	0.5	2.4	$0.5 \times 2.4 = 1.2$
$2.5 \le x < 3.0$	0.5	4.4	$0.5 \times 4.4 = 2.2$
$3.0 \le x < 3.5$	0.5	5.2	$0.5 \times 5.2 = 2.6$
$3.5 \le x < 4.0$	0.5	4.0	$0.5 \times 4 = 2$
$4.0 \le x < 4.5$	0.5	2.4	$0.5 \times 2.4 = 1.2$
$4.5 \le x < 5.0$	0.5	2.0	$0.5 \times 2 = 1$
$5.0 \le x < 6.0$	1.0	0.8	$1 \times 0.8 = 0.8$

The total area of the histogram is $1 + 1.2 + 2.2 + 2.6 + 2 + 1.2 + 1 + 0.8 = 12$. Since Area $\propto$ Frequency, and we know that Area $= 12$ and Frequency $= 60$, the proportionality factor must be 5, since $12 \times 5 = 60$.

Thus, the total number of babies born weighing less than 3kg is $(1+1.2+2.2)\times 5 = 22$.

Just to check our answer, the number of babies born weight between 4-5kg is $1.2 \times 5 = 6$, which is the correct answer.

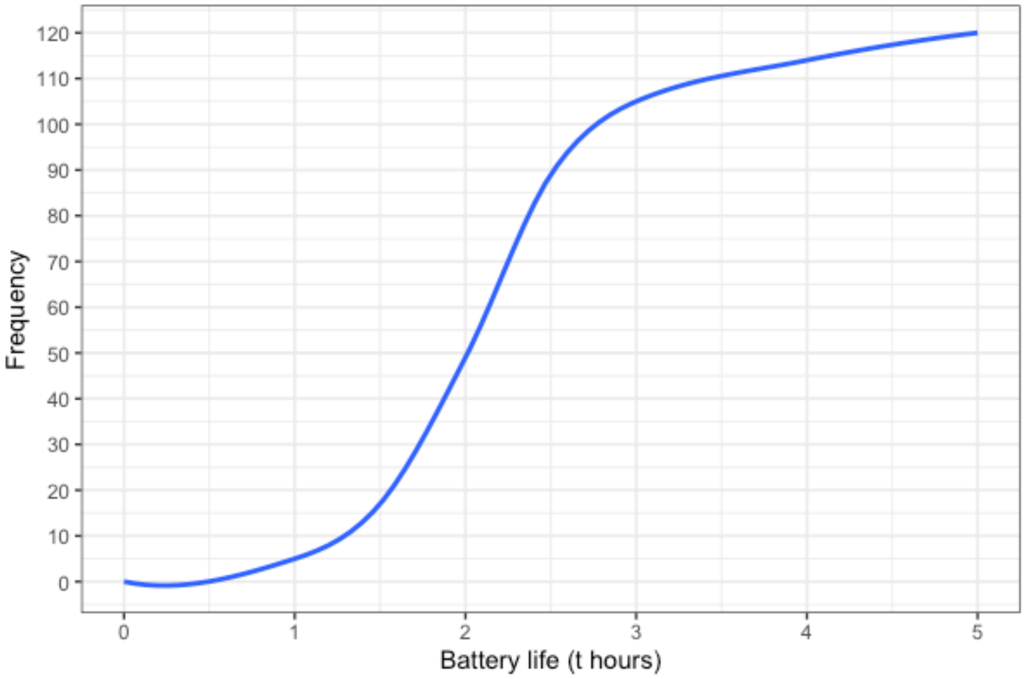
3. A factory producing batteries is interested in finding out the number of hours their batteries lasted. The data from the experiment on the time batteries lasted is represented in the table below:

Battery life (t hours)	Frequency
$0.0 \leq t < 1.0$	5
$1.0 \leq t < 1.5$	12
$1.5 \leq t < 2.0$	32
$2.0 \leq t < 2.5$	40
$2.5 \leq t < 3.0$	16
$3.0 \leq t < 4.0$	9
$4.0 \leq t < 5.0$	6

(a) Draw a cumulative frequency curve.

Solution:

Battery life (t hours)	Frequency
$t < 0.0$	0
$t < 1.0$	5
$t < 1.5$	17
$t < 2.0$	49
$t < 2.5$	89
$t < 3.0$	105
$t < 4.0$	114
$t < 5.0$	120



(b) From the cumulative frequency curve in (a), obtain

i. an estimate of the median.

Solution: Since there are 120 batteries in total, the median would be the value of battery life for which there are $120 \times 50\% = 60$ such batteries. Reading the graph, the median is 2.2 hours.

ii. an estimate of the lower and upper quartile.

Solution: Reading the graph's x-axis for which the y-axis are at the $120 \times 25\% = 30$ and $120 \times 75\% = 90$ mark, we get $Q_1 = 1.75$ and $Q_3 = 2.5$ hours.

(c) Calculate an estimate of the interquartile range and interpret the data.

Solution: IQR $= 2.5 - 1.75 = 0.75$ hours. The middle 50% of the distribution of battery life length is between 1.75 hours and 2.5 hours. The maximum battery life is recorded to be 5 hours, and it seems that only a small percentage of all batteries last this long (only 25% of all batteries last longer than 2.5 hours).

4. A student obtained the following marks (in percentage) for their assignments over the course of a year in their studies.

Geography	56	49	63	58	52	50	57	61	
English	61	70	53	60	57	52	48	79	65
Science	68	56	58	73	39	47	55	76	
Mathematics	45	46	42	48	40	45	44	41	47

(a) Find, for each subject, the range and interquartile range.

Solution: The table above, sorted in ascending order:

Geography	49	50	52	56	57	58	61	63	
English	48	52	53	57	60	61	65	70	79
Science	39	47	55	56	58	68	73	76	
Mathematics	40	41	42	44	45	45	46	47	48

Geography

Range $= 63 - 49 = 14$

$Q_2 = (56 + 57)/2 = 56.5$

$Q_1 = (50 + 52)/2 = 51$

$Q_3 = (58 + 61)/2 = 59.5$

IQR $= 59.5 - 51 = 8.5$

English
Range $= 79 - 48 = 31$
$Q_2 = 60$
$Q_1 = (52 + 53)/2 = 52.5$
$Q_3 = (65 + 70)/2 = 67.5$
IQR $= 67.5 - 52.5 = 15$

Science
Range $= 76 - 39 = 37$
$Q_2 = (56 + 58)/2 = 57$
$Q_1 = (47 + 55)/2 = 51$
$Q_3 = (68 + 73)/2 = 70.5$
IQR $= 70.5 - 51 = 19.5$

Mathematics
Range $= 48 - 40 = 8$
$Q_2 = 45$
$Q_1 = (41 + 42)/2 = 41.5$
$Q_3 = (46 + 47)/2 = 46.5$
IQR $= 46.5 - 41.5 = 5$

(b) Which subject is the student most "consistent" in? Explain your answer.

Solution: It seems the most consistent subject is Mathematics, because it has the lowest spread (range and IQR). Note "consistent" here does not mean the subject in which they excelled in, but rather there was not much variation in all marks obtained throughout the year.

(c) What is the student's "best" subject? Explain your answer.

Solution: We could define "best" subject as the subject in which the 'measure of central tendency' is highest (e.g. mean, mode, median, etc.). If we look at the median (Q_2), then the subject that has the highest median is English. Incidentally this is also the subject in which the student obtained the highest score (79) out of all subjects.

5. The height of a group of students is distributed as in the table below:

Height (cm)	151-155	156-160	161-165	166-170	171-175
Frequency	6	9	14	23	8

(a) Would you categorise the data as discrete or continuous? Explain your reasoning.

Solution: Heights are measured in centimetres should be considered as continuous data. That is, it is possible to obtain heights with decimal points e.g. 152.5cm, 166.9 cm etc.

(b) Based on your answer to (a), what do the class boundaries mean?

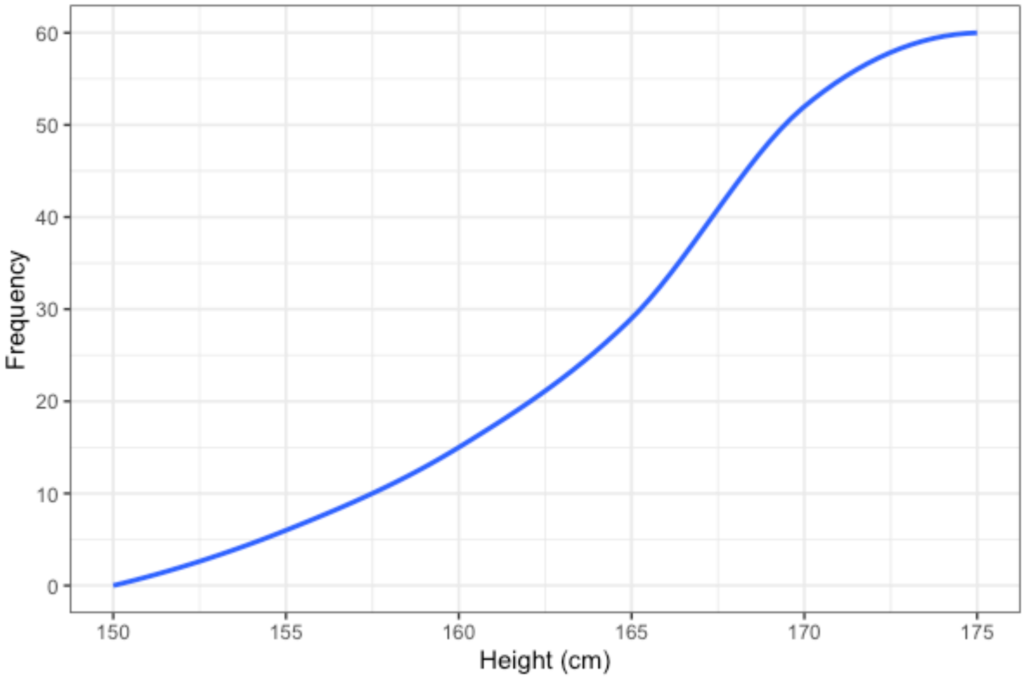
Solution: The class boundaries explicitly refer to these boundaries: $150.5 \leq x < 155.5$, $155.5 \leq x < 160.5$, $160.5 \leq x < 165.5$, $165.5 \leq x < 170.5$, and $170.5 \leq x < 175.5$.

(c) Draw a cumulative frequency curve for the data.

Solution: The cumulative frequency table is

Height (cm)	≤ 150	≤ 155	≤ 160	≤ 165	≤ 170	≤ 175
Frequency	0	6	15	29	52	60

We can now draw the curve:



(d) Use the cumulative curve to obtain an estimate for the interquartile range.

Solution: An estimate for Q_1 (at $25\% \times 60 = 15$ freq.) and Q_3 (at $75\% \times 60 = 45$ freq.) are 160cm and 168cm respectively. So the IQR $= 167 - 160 = 8$ cm.

(e) Estimate the height of the tallest 10% of students.

Solution: The tallest 10% of students (freq. $= 90\% \times 60 = 54$) have height at least 171cm.

(f) Estimate the mean height and standard deviation.

Solution: This table will be helpful:

Height (cm)	Frequency (f)	Midpoint (m)	$f \times m$	$f \times m^2$
$150.5 \leq x < 155.5$	6	153	918	140454
$155.5 \leq x < 160.5$	9	158	1422	224676
$160.5 \leq x < 165.5$	14	163	2282	371966
$165.5 \leq x < 170.5$	23	168	3864	649152
$170.5 \leq x < 175.5$	8	173	1384	239432

The statistics that we are interested in are

$$\sum x_i = \sum (f \times m) = 918 + 1422 + 2282 + 3864 + 1384 = 9870$$

and

$$\sum x_i^2 = \sum (f \times m^2) = 1404454 + 224676 + \cdots + 239432 = 1625680$$

Therefore,

$$\bar{x} = 9870/60 = 164.5 \text{ cm}$$

and

$$s = \sqrt{1625680/60 - 164.5^2} = 5.87 \text{ cm}$$