Mathematics

Senior 2 Part II

MELVIN CHIA

Started on 1 January 2023

Finished on ...

Contents

Chapter 18

Statistics

18.1 Basic Concepts

Statistics mainly study how to collect, organize, summarize, and interpret data. It is a branch of mathematics that deals with the collection, analysis, interpretation, and presentation of data. It is used to answer questions about the data and to make decisions based on the data.

Population and Sample

In statistics, a population is the entire group of individuals that we are studying, and the units that form a population are called individuals or elements. A sample is a subset of the population. The number of elements in a sample is called the sample size. For example: select 20 of the 4,000 senior high school mathematics UEC exam papers and record their scores:

72	80	96	20	42
75	60	92	18	53
82	77	53	29	34
57	79	82	90	41

Here, the population is the 4,000 scores, each of which is an element of the population. The sample is the 20 scores, the sample size is 20.

Census and Sample Survey

The way of surveying can be divided into two types: census and sample survey. A census is a survey in which every element of the population is included in the sample. For example: national census. The data collected in a census is more accurate and reliable, but it is very expensive and time-consuming.

A sample survey is a survey in which only a part of the population is included in the sample. Researchers can use a sample survey to estimate the characteristics of the population. For example: a light bulb manufacturer produces a lot of light bulbs, thus it is impossible to test every single light bulb. The manufacturer can randomly select a sample of light bulbs and test them.

18.2 Data Processing

Data that are collected must be processed before they can be analyzed.

Frequency Distribution

When the possible values of a dataset are not too many, we can use a frequency distribution table to organize the data. The frequency distribution table is a table that shows the frequency of each value in a dataset. The frequency of a value is the number of times that value appears in the dataset.

When there are too many possible values, we must group the values into classes. Before grouping the values, we must first determine the range of the values, aka the difference between the largest and smallest values, then determine the number of classes. The number of classes should be determined according to the purpose of the study and the identity of the data. After classifying the data, the range of each group is called the class interval. Typically, the class interval is the same for all classes, and must be greater than the number of classes divided by the range of the data. After the number and interval of the classes are determined, we can arrange the frequency of each class in a frequency distribution table.

Take 100 sample from a population of some kind of component, their weight (in g), are as below:

1.36	1.49	1.43	1.41	1.37	1.40
1.32	1.42	1.47	1.39	1.41	1.36
1.40	1.34	1.42	1.42	1.45	1.35
1.42	1.39	1.44	1.42	1.39	1.42
1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45
1.32	1.48	1.40	1.45	1.39	1.46
1.39	1.53	1.36	1.48	1.40	1.39
1.38	1.40	1.36	1.45	1.50	1.43
1.38	1.43	1.41	1.48	1.39	1.45

1.37	1.37	1.39	1.45	1.31	1.41
1.44	1.44	1.42	1.47	1.35	1.36
1.39	1.40	1.38	1.35	1.38	1.43
1.42	1.42	1.42	1.40	1.41	1.37
1.46	1.36	1.37	1.27	1.37	1.38
1.42	1.34	1.43	1.42	1.41	1.41
1.44	1.48	1.55	1.39		

In the dataset above, the minimum value is 1.27 and the maximum value is 1.55.

$\therefore$ The range of the data is $1.55 - 1.27 = 0.28$.

If we classify the data into 10 classes, then the class interval must be greater than $\frac{0.28}{10} = 0.028$. Thus, we can use a class interval of 0.03.

Let the lower limit of the first class be 1.27, then the lower limit of the second class is $1.27 + 0.03 = 1.30$.

Since all the values in the dataset are of 2 decimal places, the upper limit of the first class is should be 1.29. By the same logic, we can get all the classes: $1.27 - 1.29$, $1.30 - 1.32$, $\cdots$, $1.54 - 1.56$.

Now we can arrange the data into the frequency distribution table:

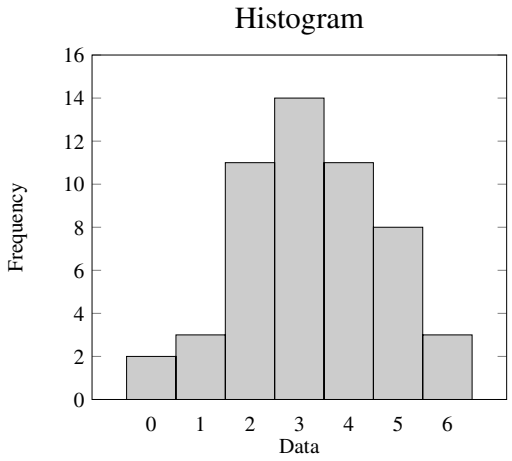
Weight $m(g)$	Frequency
$1.27 - 1.29$	1
$1.30 - 1.32$	4
$1.33 - 1.35$	7
$1.36 - 1.38$	22
$1.39 - 1.41$	24
$1.42 - 1.44$	24
$1.45 - 1.47$	10
$1.48 - 1.50$	6
$1.51 - 1.53$	1
$1.54 - 1.56$	1

In the example above, we assume that the weight of the components is accurate to 2 decimal places. Hence, if a component has a weight of $1.443g$, it is rounded to $1.44g$, thus it belongs to the class $1.42 - 1.44$. Hence, the actual range of the first class $1.27 - 1.29$ is $1.265 \leq m < 1.295$, written as $1.265 - 1.295$, while 1.265 and 1.295 are the boundaries of the first class, 1.265 is the lower boundary and 1.295 is the upper boundary. The mean of the lower boundary and upper boundary of a class is called the class midpoint. For example, the class midpoint of the first class is $\frac{1.265+1.295}{2} = 1.28$.

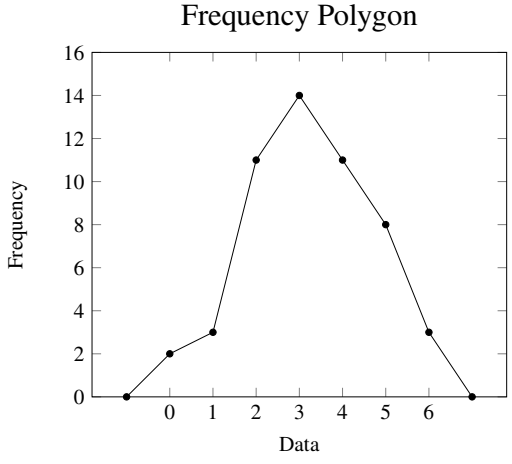
When we are analyzing the data data that have been classified into classes, the midpoint of each class is used as the representative value of the class. Thus, we should try our best to make the data-intensive place the group midpoint when choosing the class interval and boundaries, so that the data can be analyzed more precisely.

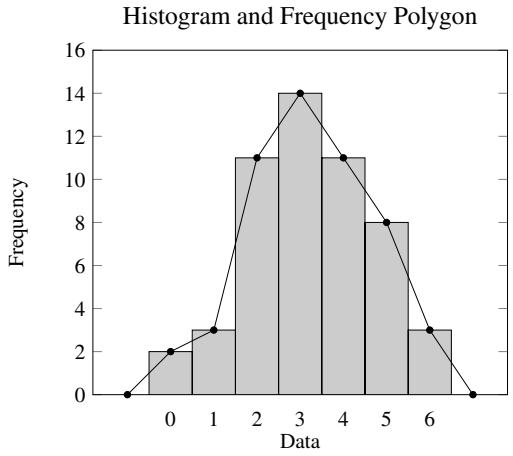
The distribution of frequency can be represented by a histogram or a frequency polygon.

The histogram is a row of continuous bars, the bottom side of each bar on the x-axis. For unclassified data, the bottom side of each bar is marked with the values, while the height of each bar is the frequency of the corresponding value. For classified data, the bottom side of each bar is marked with the boundaries of the corresponding class, while the area of each bar must be proportional to the frequency of the corresponding class. When the class interval of each class is the same, we can use the frequency of each class as the height of the bar.



The frequency polygon is a continuous line graph, the x-axis is the midpoint of each class, and the y-axis is the frequency of each class. To draw a frequency polygon, we plot each point, including the point before the first class and the point after the last class that uses 0 as their frequency, and then connect the points with a continuous line.





18.2.1 Practice 1

There are 105 students in a senior 3 art and commerce class. In a mock exam of UEC, their scores for Mathematics subject are as follows:

35	88	67	32	38	34	45
78	54	58	69	21	90	78
74	43	42	35	57	34	77
89	66	74	71	44	56	48
33	24	73	63	51	59	49
34	55	52	75	72	62	62
44	48	73	49	57	67	80
70	66	54	32	29	35	37
47	41	51	36	46	55	53
60	53	62	39	35	48	42
71	63	70	33	45	42	44
61	59	67	30	42	43	89
96	82	47	63	54	34	45
45	87	28	34	29	77	64
64	50	48	75	33	56	84

(a) Find the range of the data.

Sol.

$$\begin{aligned} \text{Max value} &= 96 \\ \text{Min value} &= 21 \\ \therefore \text{Range} &= 96 - 21 \\ &= 75 \end{aligned}$$

(b) Group the data into 10 classes, draw a frequency distribution table, and find the upper and lower boundary and midpoint of each class.

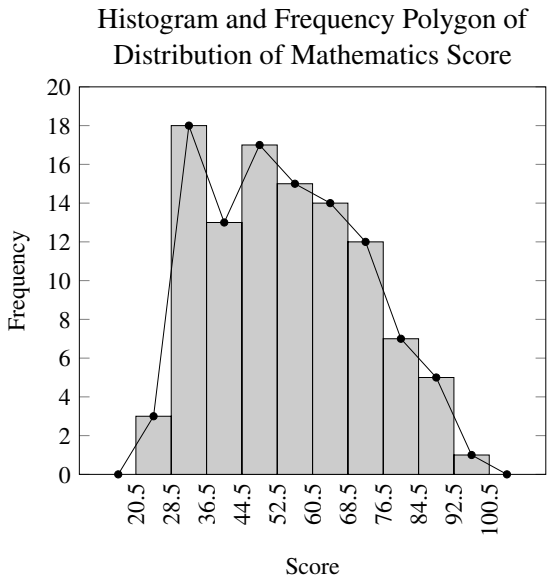
Sol.

$$\begin{aligned} \text{Range} &= 75 \\ \text{Number of classes} &= 10 \\ \text{Class width} &= \frac{75}{10} \\ &= 7.5 \\ &\approx 8 \end{aligned}$$

Score	Lower	Upper	Mid	Freq.
21 - 28	20.5	28.5	24.5	3
29 - 36	28.5	36.5	32.5	18
37 - 44	36.5	44.5	40.5	13
45 - 52	44.5	52.5	48.5	17
53 - 60	52.5	60.5	56.5	15
61 - 68	60.5	68.5	64.5	14
69 - 76	68.5	76.5	72.5	12
77 - 84	76.5	84.5	80.5	7
85 - 92	84.5	92.5	88.5	5
93 - 100	92.5	100.5	96.5	1

(c) Draw a histogram and frequency polygon.

Sol.



Cumulative Frequency Distribution

Summing up the frequency of each class, we obtain the cumulative frequency distribution. Use the upper boundary of each class as the x-axis, and the cumulative frequency as the y-axis, we can draw the cumulative frequency distribution by plotting each point including the point before the first class that uses 0 as its frequency and connect them together. If we split the x-axis and the higest point of the curve into 100 equal

parts, we get the percentage of the cumulative frequency distribution.

18.2.2 Practice 2

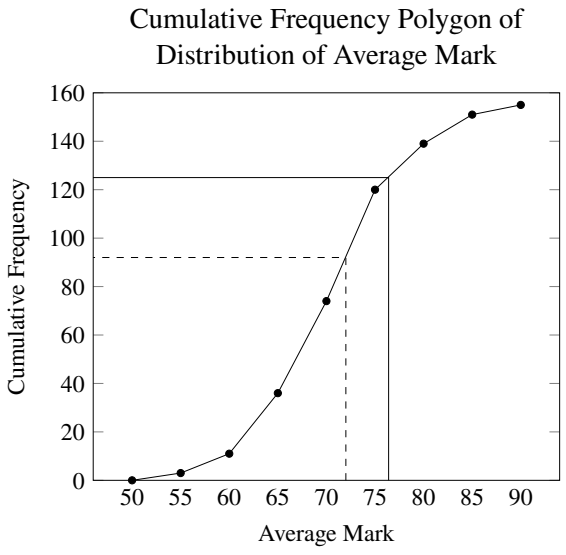
There are 155 students in a senior 3 art and commerce class, and the frequency distribution table of their average marks is shown below:

Average Mark	Frequency
50 - 55	3
55 - 60	8
60 - 65	25
65 - 70	38
70 - 75	46
75 - 80	19
80 - 85	12
85 - 90	4

(a) Make a cumulative frequency distribution table and draw a cumulative frequency polygon.

Sol.

Avg	Freq.	Lower Than	Cumm. Freq.
50 - 55	3	55	3
55 - 60	8	60	11
60 - 65	25	65	36
65 - 70	38	70	74
70 - 75	46	75	120
75 - 80	19	80	139
80 - 85	12	85	151
85 - 90	4	90	155



(b) If the average mark of a student is 72, find his rank in the class.

Sol.

In the graph above, we can see that there are approximately 92 students who have an average mark lower than 72. Therefore, the rank of the student is $155-92 = 63$.

(c) If the top 20% of the class are to be awarded a certificate, find the minimum average mark required for the certificate.

Sol.

$$\begin{aligned}\text{Top } 20\% &= 20\% \times 155 \\ &= 31\end{aligned}$$

Therefore, students with an average mark corresponding to cumulative frequency higher than 124 will be awarded a certificate.

In the graph above, The minimum average mark required for the certificate is 76.

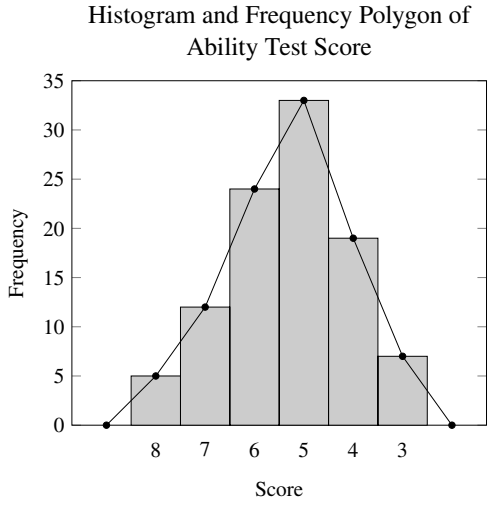
18.2.3 Exercise 18.2

1. A company performed an ability test on 100 job seekers and the results are shown in the following table:

Score	8	7	6	5	4	3
Frequency	5	12	24	33	19	7

Draw a hustogram and a frequency polygon for the data above.

Sol.



2. Take 120 ears of rice from a rice field, the length of each ear is measured (in cm) and the results are as fol-

lowing:

6.5	6.4	6.7	5.8	5.9	5.9
5.2	4.0	5.4	4.6	5.8	5.5
6.0	6.5	5.1	6.2	5.4	5.0
5.0	6.8	6.0	5.0	5.7	6.0
5.5	6.8	6.0	6.3	5.5	5.0
6.4	5.8	5.9	5.7	6.8	6.6
6.0	6.4	5.7	7.4	6.0	5.4
6.5	6.0	6.8	5.3	6.4	5.7
6.7	6.2	5.6	6.0	6.7	6.7
6.0	5.5	6.2	6.1	5.3	6.2
5.8	5.3	7.0	6.0	6.0	5.9
5.4	6.0	5.2	6.0	6.3	5.7
6.8	6.1	4.5	5.4	6.3	6.9
4.9	5.1	5.6	5.9	6.1	6.5
6.6	5.7	5.8	5.8	6.2	6.3
6.5	5.3	5.9	5.5	5.8	6.3
5.2	6.0	7.0	6.4	5.8	6.3
6.0	6.3	5.6	6.8	6.6	4.7
5.7	5.7	5.6	6.3	6.0	5.8
6.3	7.5	6.2	6.4	7.0	6.5

(a) Find the range of the dataset.

Sol.

$$\begin{aligned}\text{Min value} &= 4.0\\ \text{Max value} &= 7.5\\ \therefore \text{Range} &= 7.5 - 4.0\\ &= 3.5\end{aligned}$$

(b) Group the data into 12 classes, make a frequency distribution table, find the upper and lower boundaries and midpoint of each class, and calculate the cumulative frequency.

Sol.

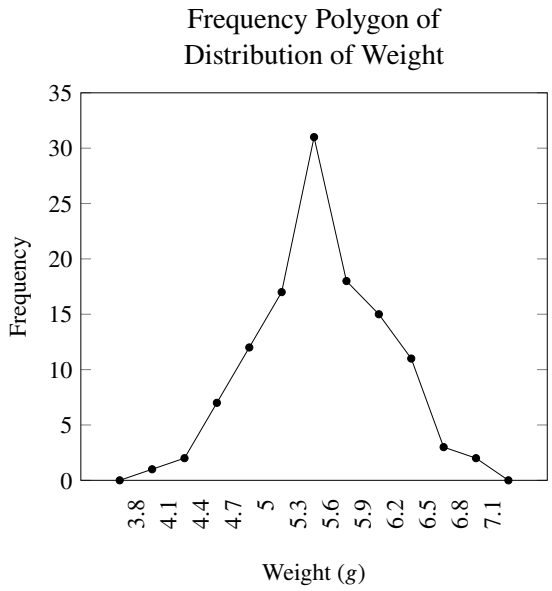
$$\begin{aligned}\text{Range} &= 3.5\\ \text{Number of classes} &= 12\\ \therefore \text{Class width} &= \frac{3.5}{12}\\ &= \frac{3.5}{12}\\ &\approx 0.3\end{aligned}$$

Weight	Lower	Upper	Mid	Freq.
4.0 - 4.2	3.95	4.25	4.10	1
4.3 - 4.5	4.25	4.55	4.40	1
4.6 - 4.8	4.55	4.85	4.70	2
4.9 - 5.1	4.85	5.15	5.00	7
5.2 - 5.4	5.15	5.45	5.30	12
5.5 - 5.7	5.45	5.75	5.60	17
5.8 - 6.0	5.75	6.05	5.90	31
6.1 - 6.3	6.05	6.35	6.20	18
6.4 - 6.6	6.35	6.65	6.50	15
6.7 - 6.9	6.65	6.95	6.80	11
7.0 - 7.2	6.95	7.25	7.10	3
7.3 - 7.5	7.25	7.55	7.40	2

Weight	Freq.	Lower Than	Cum. Freq.
4.0 - 4.3	1	4.3	1
4.3 - 4.6	1	4.6	2
4.6 - 4.9	2	4.9	4
4.9 - 5.2	7	5.2	11
5.2 - 5.5	12	5.5	23
5.5 - 5.8	17	5.8	40
5.8 - 6.1	31	6.1	71
6.1 - 6.4	18	6.4	89
6.4 - 6.7	15	6.7	104
6.7 - 7.0	11	7.0	115
7.0 - 7.3	3	7.3	118
7.3 - 7.6	2	7.6	120

(c) Draw a frequency polygon.

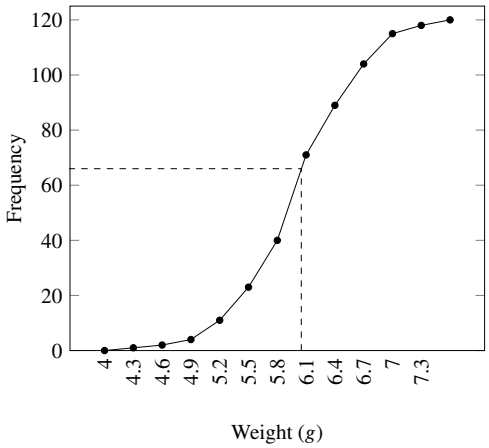
Sol.



(d) Draw a cumulative frequency polygon.

Sol.

Cumulative Frequency Polygon of Distribution of Weight



(e) Find the percentage of the ears of rice whose length is greater than $6cm$.

Sol.

In the diagram above, there are approximately $120-66=54$ ears of rice whose length is greater than $6cm$, which is about $\frac{54}{120}\times 100\% = 45\%$ of the total number of ears of rice.

3. The table below shows the weight distribution of 90 babies (in kg):

Weight	Frequency
1.5 - 2.0	2
2.0 - 2.5	4
2.5 - 3.0	13
3.0 - 3.5	32
3.5 - 4.0	28
4.0 - 4.5	10
4.5 - 5.0	1

(a) Make a cumulative frequency table.

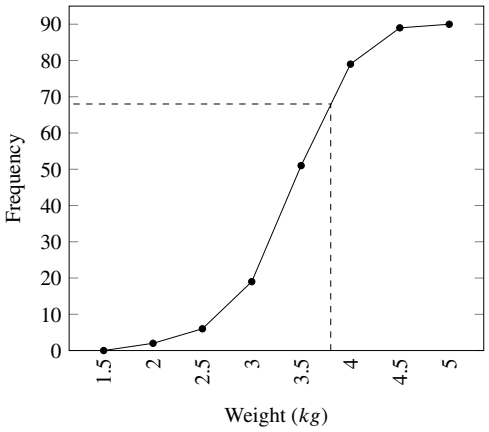
Sol.

Weight	Freq.	Less than	Cum. Freq.
1.5 - 2.0	2	2.0	2
2.0 - 2.5	4	2.5	6
2.5 - 3.0	13	3.0	19
3.0 - 3.5	32	3.5	51
3.5 - 4.0	28	4.0	79
4.0 - 4.5	10	4.5	89
4.5 - 5.0	1	5.0	90

(b) Draw a cumulative frequency polygon.

Sol.

Cumulative Frequency Polygon of Distribution of Weight



(c) Find the percentage of babies whose weight is greater than $3.8kg$.

Sol.

In the diagram above, there are approximately $90-68=22$ babies whose weight is greater than $3.8kg$, which is about $\frac{22}{90}\times 100\% = 24.44\%$ of the total number of babies.

4. The table below shows the average score distribution of 50 students in a class:

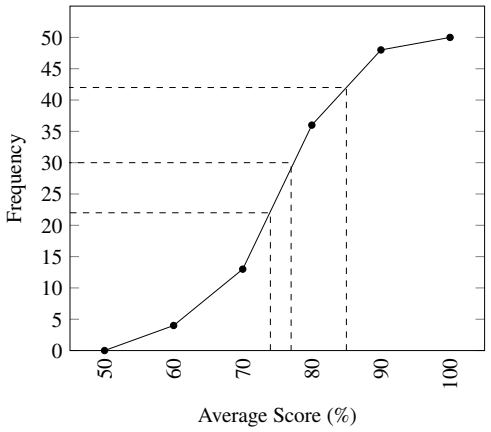
Average Score	Frequency
50.0 - 59.9	4
60.0 - 69.9	9
70.0 - 79.9	23
80.0 - 89.9	12
90.0 - 99.9	2

(a) Make a cumulative frequency table and draw a cumulative frequency polygon.

Sol.

Average Score	Freq.	Less than	Cum. Freq.
50.0 - 59.9	4	60	4
60.0 - 69.9	9	70	13
70.0 - 79.9	23	80	36
80.0 - 89.9	12	90	48
90.0 - 99.9	2	100	50

Cumulative Frequency Polygon of Distribution of Average Score



(b) A student get an average score of 74, find his rank in the class.

Sol.

In the diagram above, there are approximately 22 students whose average score is less than 74, which means that the student is ranked $50-22=28$.

(c) Find the average score of the student who is ranked 20.

Sol.

In the diagram above, the student who is ranked 20 has an average score of about 77.

(d) Find the percentage of students whose average score is greater than 85.

Sol.

In the diagram above, there are approximately $50-42=8$ students whose average score is greater than 85, which is about $\frac{8}{50}\times 100\%=16\%$ of the total number of students.

5. The table below shows the score distribution of 1200 students in UEC accounting exam:

Score	Number of Students
10 - 19	20
20 - 29	60
30 - 39	95
40 - 49	130
50 - 59	340
60 - 69	310
70 - 79	135
80 - 89	80
90 - 99	30

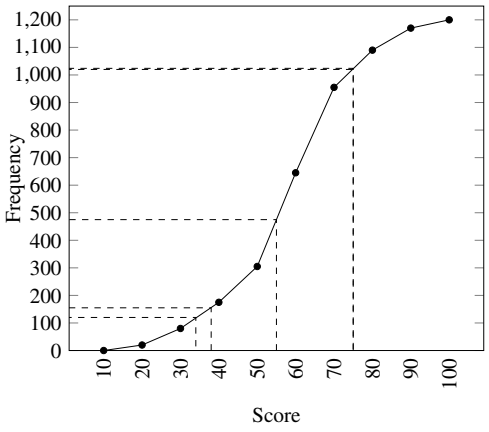
Examinees are categorised into 4 groups based on their score: *Excellent*, *Good*, *Pass*, and *Fail*.

(a) Make a cumulative frequency table and draw a cumulative frequency polygon.

Sol.

Score	Freq.	Less than	Cum. Freq.
10 - 19	20	20	20
20 - 29	60	80	80
30 - 39	95	175	175
40 - 49	130	305	305
50 - 59	340	645	645
60 - 69	310	955	955
70 - 79	135	1090	1090
80 - 89	80	1170	1170
90 - 99	30	1200	1200

Cumulative Frequency Polygon of Distribution of Score



(b) If the passing score is 38, find the percentage of students who pass the exam.

Sol.

In the diagram above, there are approximately $1200-155=1045$ students whose score is greater or equal to 38, which is about $\frac{1045}{1200}\times 100\%=86.67\%$ of the total number of students.

(c) Assume that the minimum score to be categorised as *Excellent* and *Good* is 75 and 55 respectively, find the percentage of students who are categorised as *Excellent* and *Good* respectively.

Sol.

In the diagram above, there are approximately $1200-1024=176$ students whose score is greater or equal to 75, which is about $\frac{176}{1200}\times 100\%=14.67\%$ of the total number of students who are categorised as *Excellent*.

Also, there are approximately $1024 - 475 = 549$ students whose score is greater or equal to 55, which is about $\frac{549}{1200} \times 100\% = 45.75\%$ of the total number of students who are categorised as *Good*.

(d) Find the passing mark if the percentage of students who pass the exam is 90%.

Sol.

If the percentage of students who pass the exam is 90%, then the number of students who pass the exam is 90% of 1200 students, which is 1080 students. That means, there are $1200 - 1080 = 120$ students who fail the exam.

In the diagram above, the passing mark is about 34 given that there are 120 students who fail the exam.

(e) Find the minimum mark of a student who is categorised as *Excellent* if the percentage of students who are categorised as *Excellent* is 15%.

Sol.

If the percentage of students who are categorised as *Excellent* is 15%, then the number of students who are categorised as *Excellent* is 15% of 1200 students, which is 180 students. That means, there are $1200 - 180 = 1020$ students who are not categorised as *Excellent*.

In the diagram above, the minimum mark of a student who is categorised as *Excellent* is about 75 given that there are 1020 students who are not categorised as *Excellent*.

18.3 Central Tendency

Central tendency is a measure of the central position of a distribution, or a single value that attempts to describe a set of data. The most common measures of central tendency are the mean, median, and mode.

Mean

Mean is also known as arithmetic mean. For n values $x_1, x_2, \ldots, x_n$, the mean is defined as

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{\sum x_i}{n}\end{aligned}$$

For data whose possible values are $x_1, x_2, \ldots, x_n$, and their respective frequencies are $f_1, f_2, \ldots, f_n$, the mean is defined as

$$\begin{aligned}\bar{x} &= \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \\ &= \frac{\sum f_i x_i}{\sum f_i}\end{aligned}$$

For grouped data, we take the mean of each class as the representative value x_i of the class.

Weighted Mean

In some scenario, weighted mean is better than the mean to describe the data.

When calculating the arithmetic mean, each value is given equal weight. However, in some cases, each value in a dataset may not be equally important. For example, the importance of the mark of a student for each subject is weighted according to the number of classes of the subject in a week. Hence, when calculating the average mark of the student, each mark must be multiplied by a value that represents the importance of the subject, and that value is called the weight. The weighted mean is defined as

$$\begin{aligned}\bar{x} &= \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \\ &= \frac{\sum w_i x_i}{\sum w_i}\end{aligned}$$

where x_i are the values and w_i are the weights of x_i.

18.3.1 Practice 3

1. Find the mean of 34, 50, 24, 32, 53, 30, 62, 27.

   Sol.

   $$\begin{aligned}\bar{x} &= \frac{30 + 50 + 24 + 32 + 53 + 30 + 62 + 27}{8} \\ &= \frac{312}{8} \\ &= 39\end{aligned}$$

2. There are three workshop A, B, and C in a factory. Workshop A has 10 workers, their wages are \$35 per day, workshop B has 30 workers, their wages are \$45 per day, and workshop C has 15 workers, their wages are \$55 per day. Find the mean of the wages of the workers in the factory.

   Sol.

   Let the wages of workers be x_1, and the amount of workers be f_1.

x_1	f_1	$x_1 f_1$
35	10	350
45	30	1350
55	15	825
	$\sum f_i = 55$	$\sum f_i x_i = 2525$

$\therefore$ Average wages of workers in the factory is $\frac{2525}{55} =$ \$45.91.

3. A school appoints students to participate in a Math competition. During the competition, candidates must answer 25 questions within an hour. The table below shows the distribution of frequency of the number of questions that those candidates answer correctly:

Answered Correctly	Frequency
1 - 5	3
6 - 10	12
11 - 15	7
16 - 20	8
21 - 25	5

Complete the following table, and find the mean of the number of questions that those candidates answer correctly.

Ans. Correctly	Freq. f_i	Midpoint x_i	$f_i x_i$
1 - 5			
6 - 10			
11 - 15			
16 - 20			
21 - 25			

Sol.

Ans. Correctly	Freq. f_i	Midpoint x_i	$f_i x_i$
1 - 5	3	3	9
6 - 10	12	8	96
11 - 15	7	13	91
16 - 20	8	18	144
21 - 25	5	23	115
	$\sum f_i = 35$	$\sum f_i x_i = 455$	

$\therefore$ The mean of the number of questions that those candidates answer correctly is $\frac{455}{35} = 13$.

18.3.2 Exercise 18.3a

1. Take a sample of 20 from a batch of machine parts, their weight (in g) are as follows:

210	208	200	205	202	218
206	214	215	207	195	207
218	192	202	216	185	227
187	215				

Find the mean weight of these machine parts.

Sol.

$$\begin{aligned}\bar{x} &= \frac{210 + 208 + 200 + \cdots + 215}{20} \\ &= \frac{4129}{20} \\ &= 206.45\end{aligned}$$

2. Given that the mean of a dataset 4, -3, 2, k, 5, 8 is 10, find the value of k.

Sol.

$$\begin{aligned}\frac{4 + (-3) + 2 + k + 5 + 8}{6} &= 10 \\ 16 + k &= 60 \\ k &= 44\end{aligned}$$

3. Given that the mean of x_1, x_2, x_3, x_4, x_5 is 40, and the mean of y_1, y_2, y_3 is 15. Find the mean after combining these two datasets.

Sol.

$$\begin{aligned}\frac{x_1 + \cdots + x_5}{5} &= 40 \\ x_1 + \cdots + x_5 &= 200\end{aligned}$$

$$\begin{aligned}\frac{y_1 + y_2 + y_3}{3} &= 15 \\ y_1 + y_2 + y_3 &= 45\end{aligned}$$

$$\begin{aligned}\bar{xy} &= \frac{x_1 + x_2 + \cdots + y_3}{8} \\ &= \frac{245}{8} \\ &= 30.63\end{aligned}$$

4. A school have 2 senior 3 classes: A and B. In a Chinese language test, the average mark of 49 students in A class in 72, while the average mark for 45 students

in class B is 68. Find the average mark of all students in these two class combined.

Sol.

$$\begin{aligned}\bar{x} &= \frac{72 \times 49 + 68 \times 45}{49 + 45} \\ &= \frac{3528 + 3060}{94} \\ &= \frac{6588}{94} \\ &= 70.09\end{aligned}$$

5. Given that the mean for 8 values are 5. The mean increased by 1.4 after adding two values: x and $3x$. Find the value of x.

Sol.

$$\begin{aligned}\frac{8 \times 5 + x + 3x}{8 + 2} &= 5 + 1.4 \\ \frac{40 + 4x}{10} &= 6.4 \\ 40 + 4x &= 64 \\ 4x &= 24 \\ x &= 6\end{aligned}$$

6. Throwing 6 coin at the same time and record the number of heads. After throwing 100 times, we get the following frequency distribution table:

Number of Heads	Frequency
0	2
1	10
2	24
3	35
4	22
5	6
6	1

Find the mean of the number of heads for each throw.

Sol.

Let the number of heads be x_i and the frequency be f_i.

x_i	f_i	$x_i f_i$
0	2	0
1	10	10
2	24	48
3	35	105
4	22	88
5	6	30
6	1	6
	$\sum f_i = 100$	$\sum x_i f_i = 287$

$\therefore$ The mean of the number of heads for each throw is $\frac{287}{100} = 2.87$.

7. The table below shows the score distribution of 66 students in a Chinese language test:

Score	Frequency
31 - 40	6
41 - 50	12
51 - 60	15
61 - 70	15
71 - 80	8
81 - 90	6
91 - 100	4

Find their mark in average.

Score	Mid x_1	Freq. f_1	$x_1 f_1$
31 - 40	35.5	6	213
41 - 50	45.5	12	546
51 - 60	55.5	15	832.5
61 - 70	65.5	15	982.5
71 - 80	75.5	8	604
81 - 90	85.5	6	513
91 - 100	95.5	4	382
		$\sum f_1 = 66$	$\sum x_1 f_1 = 4073$

$\therefore$ The mark in average is $\frac{4073}{66} = 61.71$.

8. Below are the number of classes and marks for each subject of a junior student:

Subject	Number of Classes	Average Mark
Chinese	7	75
Malay	7	73
English	7	65
Mathematics	7	82
Science	5	86
History	3	73
Geography	3	87

(a) Find his mark in average.

Sol.

$$\begin{aligned}\bar{x} &= \frac{75 + 73 + 65 + 82 + 86 + 73 + 87}{7} \\ &= \frac{541}{7} \\ &= 77.29\end{aligned}$$

(b) Use the number of classes as the weight to find his average mark.

Sol.

$$\begin{aligned}\bar{x} &= \frac{75\times 7 + 73\times 7 + \cdots + 87\times 3}{7+7+7+7+5+3+3}\\ &= \frac{525+511+455+574+430+219+261}{39}\\ &= \frac{2975}{39}\\ &= 76.28\end{aligned}$$

9. The weight of 60 junior 2 students in a school are as follows:

Weight (kg)	Frequency
54 - 56	10
57 - 59	20
60 - 62	x
63 - 65	8
66 - 68	4
69 - 71	y

Given that the mean weight of these students is 60.1 kg, find the value of x and y.

Sol.

$$\text{Total weight} = 60.1\times 60 = 3606$$

Wght (kg)	M. x_1	Freq. f_1	x_1f_1
54 - 56	55	10	550
57 - 59	58	20	1160
60 - 62	61	x	$61x$
63 - 65	64	8	512
66 - 68	67	4	268
69 - 71	70	y	$70y$
		$\sum f_1 = 60$	$\sum x_1f_1 = 3606$

$$\begin{cases} 10+20+x+8+4+y = 60 & (1)\\ 550+1160+61x+512+268+70y = 3606 & (2)\end{cases}$$

$$\begin{aligned}(1) &: 42 + x + y = 60\\ x + y &= 18\\ (2) &: 61x + 70y = 1116\\ (1)\times 61 &: 61x + 61y = 1098\\ (2)-(1) &: 9y = 18\\ y &= 2\\ \text{From } (1) &: x = 16\end{aligned}$$

Median

The median is the middle value of a sorted dataset. The number of values must be equal for both side of the median.

If the number of values is n, when n is odd, the median is the number in $\frac{n+1}{2}$ position.
When n is even, the median is the mean of the number in $\frac{n}{2}$ and $\frac{n}{2}+1$ position.

For grouped data, we can make a cumulative frequency polygon, and the median is the value corresponding to 50% of the percentage of the cumulative frequency.

Let n be the number of values in the dataset, aka $\sum f_1$,

L_m be the lower boundaries of the group of the median,

C_m be the range of the group of the median,

f_m be the frequency of the group of the median,

F_m be the cum. frequency of the group of the median,

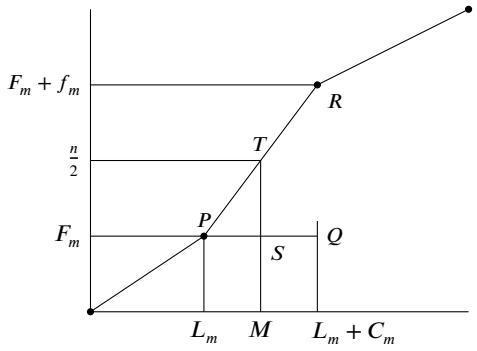


Diagram above shows a part of a cumulative frequency polygon, where R is the point corresponding to the group containing the median, P is the point corresponding to the group before the group containing the median, and M is the median. Since $\Delta PQR \sim \Delta PST$,

$$\therefore \frac{PS}{PQ} = \frac{ST}{QR}$$

$$\text{That is, } \frac{M - L_m}{C_m} = \frac{\frac{n}{2} - F_m}{f_m}$$

We get the following after simplifying the equation:

$$M = L_m + \left(\frac{\frac{n}{2} - F_m}{f_m}\right) C_m$$

18.3.3 Practice 4

1. 10 workers in a factory made the same type of product in a day, the number of products made are as follows:

15	17	14	10	15
19	17	16	14	12

   Find the median of the number of products made by these 10 workers.

   Sol.

   Sort the dataset:

   10 12 14 14 15 15 16 17 17 19

   The median is the mean of the number in $\frac{10}{2} = 5$ and $\frac{10}{2} + 1 = 6$ position, which is $\frac{15+15}{2} = 15$.

2. The table below shows the result of a right eye vision test for 49 students in a class:

Vision	Number of Students
0.2	2
0.3	3
0.4	4
0.5	3
0.6	4
0.8	9
1.0	9
1.2	10
1.5	5

   Find the median of the right eye vision of these students.

   Sol.

Vision	Number of Students	Cum. Frequency
0.2	2	2
0.3	3	5
0.4	4	9
0.5	3	12
0.6	4	16
0.8	9	25
1.0	9	34
1.2	10	44
1.5	5	49

   Since $n = 49$ is odd, the median is the number in the $\frac{49+1}{2} = 25$ position, which is 0.8.

3. The table below shows time distribution of 21 students browsing the Internet:

Time (hours)	Number of Students
1.1 - 1.3	4
1.4 - 1.6	3
1.7 - 1.9	5
2.0 - 2.2	4
2.3 - 2.5	5

   Find the median of the time distribution of these students.

   Sol.

Time	Freq.	Cum. Freq.
1.1 - 1.3	4	4
1.4 - 1.6	3	7
1.7 - 1.9	5	12
2.0 - 2.2	4	16
2.3 - 2.5	5	21

   The median is the number in the $\frac{21}{2} = 10.5$ position, which is $1.7 - 1.9$. $C_m = 0.3$, $L_m = 1.65$, and $f_m = 5$, $F_m = 7$.

   $$\therefore Mean = 1.65 + \frac{10.5 - 7}{5} \times 0.3 = 1.86$$

18.3.4 Exercise 18.3b

1. During a gymnastic competition, there are four judges scoring the performance of each contestant, and the median of these four scores are taken as the final score of the contestant. Given that the scores given by four judges are 9.5, 9.4, 9.8, and 9.4 respectively, find the final score of the contestant.

   Sol.

   Sort the scores:

   9.4 9.4 9.5 9.8

   The median is the mean of the number in $\frac{4}{2} = 2$ and $\frac{4}{2} + 1 = 3$ position, which is $\frac{9.4+9.5}{2} = 9.45$.

2. Following are the weight of 15 boys with same age:

36	35	33	37	35
42	40	38	38	39
40	41	36	38	37

   (a) Find the median of these 15 boys.

   Sol.

Sort the data:

33 35 35 36 36 37 37 38 38 38
39 40 41 42 43

The median is the mean of the number in $\frac{15+1}{2} = 8$ position, which is 38.

(b) Group the data using pattern $33-35, 35-37, \ldots, 41-43$. Then, find the median.

Sol.

Weight (kg)	Frequency	Cum. Frequency
33 - 35	1	1
35 - 37	4	5
37 - 39	5	10
39 - 41	2	12
41 - 43	2	14
43 - 45	1	15

The median is the number in the $\frac{15}{2} = 7.5$ position. $C_m = 2$, $L_m = 37$, $f_m = 5$, and $F_m = 5$.

$$\therefore Median = 37 + \frac{7.5-5}{5} \times 2 = 38$$

3. The table below shows the score distribution of a group of pupils in a minor test:

Score	Number of Pupils
5	4
10	2
15	3
20	x
25	4

Assume that the median is 15, find the possibility value of x.

Sol.

Score	Freq.	Cum. Freq.
5	4	4
10	2	6
15	3	9
20	x	9 + x
25	4	13 + x

$$\frac{13+x+1}{2} \leq 9$$
$$14 + x \leq 18$$
$$x \leq 4$$
$$\therefore 0 \leq x \leq 4$$

Therefore, the possibility values of x are 0, 1, 2, 3, and 4.

4. The following table shows the income of employees in a company:

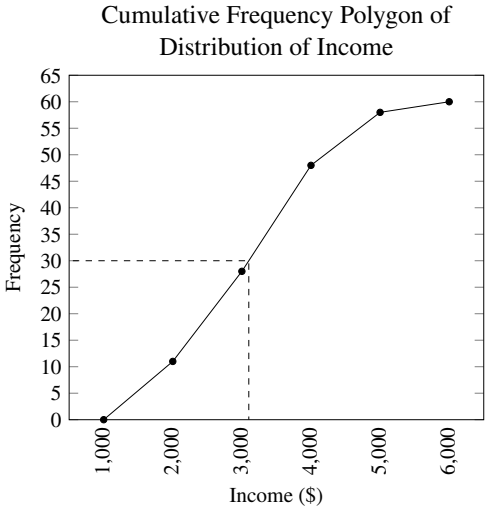
Income ($)	Number of Employees
1000 - 2000	11
2000 - 3000	17
3000 - 4000	20
4000 - 5000	10
5000 - 6000	2

(a) Find the median of their income using cumulative frequency polygon.

Sol.

Income ($)	Freq.	Cum. Freq.
1000 - 2000	11	11
2000 - 3000	17	28
3000 - 4000	20	48
4000 - 5000	10	58
5000 - 6000	2	60

The median is the number in $\frac{60}{2} = 30$ position.



Therefore, the median of their income is \$3100.

(b) Find the median of their income using formula and compare the result with (a).

Sol.

The median is the number in the $\frac{60}{2} = 30$ position, which is 3000 − 4000. $C_m = 1000$, $L_m = 3000$, and $f_m = 20$, $F_m = 28$.

$$\therefore Median = 3000 + \frac{30-28}{20} \times 1000 = 3100$$

Therefore, the median of their income is \$3100, which is the same as (a).

5. The table below shows the distribution of height of 20 students:

Height (cm)	Number of Students
120 - 130	3
130 - 140	4
140 - 150	x
150 - 160	5
160 - 170	6

Find:

(a) The value of x.

Sol.

$$\begin{aligned} x+3+4+5+6 &= 20 \\ x &= 20-18 \\ &= 2 \end{aligned}$$

(b) The median of their height.

Sol.

Height (cm)	Freq.	Cum. Freq.
120 - 130	3	3
130 - 140	4	7
140 - 150	2	9
150 - 160	5	14
160 - 170	6	20

The median is the number in $\frac{20}{2} = 10$ position, which is $150-160$. $C_m = 10$, $L_m = 150$, $f_m = 5$, and $F_m = 9$.

$$\therefore Median = 150 + \frac{10-9}{5} \times 10 = 152$$

Therefore, the median of their height is $152cm$.

6. The table below shows the distribution of wages of workers in a factory:

Wages \$	Number of Workers
40 - 49	4
50 - 59	14
60 - 69	5
70 - 79	x
80 - 89	2

Given that the median is 63.5, find the value of x.

Sol.

Wages \$	Freq.	Cum. Freq.
40 - 49	4	4
50 - 59	14	18
60 - 69	5	23
70 - 79	x	23 + x
80 - 89	2	25 + x

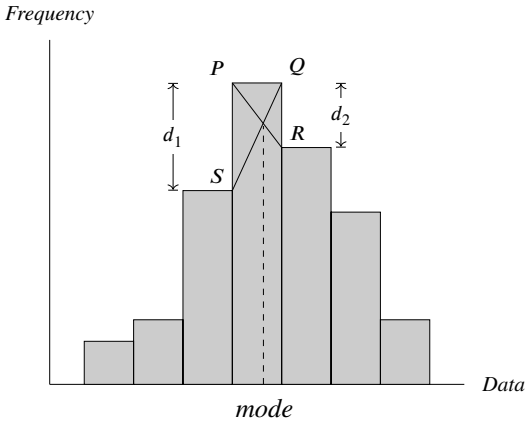
63.5 is in between $60-69$, which is in the $\frac{25+x}{2}$ position. $C_m = 10$, $L_m = 59.5$, $f_m = 5$, $F_m = 18$.

$$\begin{aligned} 59.5 + \frac{\frac{25+x}{2} - 18}{5} \times 10 &= 63.5 \\ \frac{\frac{25+x}{2} - 18}{5} \times 10 &= 4 \\ \frac{\frac{25+x}{2} - 18}{5} &= 0.4 \\ \frac{25+x}{2} - 18 &= 2 \\ \frac{25+x}{2} &= 20 \\ 25 + x &= 40 \\ x &= 15 \end{aligned}$$

Mode

In a set of data, the mode is the value that occurs most frequently. There can be more than one mode in a set of data. If all the values in a dataset occur with the same frequency, then there is no mode for the data.

For grouped data, the mode is the class that has the highest frequency, and there can be more than one mode. Besides that, the mode can also be estimated using histogram. The method is as follows:



The diagram above shows a histogram of a set of data. The class corresponding to the highest rectangle is the mode of the data, and the mode is the x-value of the intersection point of PR and QS.

Unlike median, the formula of mode can be derived from similar triangles. Let:

L be the lower boundaries of the modal class

C be the range of the modal class

d_1 be the difference between the lower boundary of the modal class and the lower boundary of the class immediately before the modal class

d_2 be the difference between the lower boundary of the modal class and the lower boundary of the class immediately after the modal class

then

$$mode = L + \left(\frac{d_1}{d_1 + d_2}\right) C$$

18.3.5 Practice 5

The following table shows the distribution of the score of 36 students in a Mathematics exam:

Score	Number of Students
20 - 29	2
30 - 39	6
40 - 49	10
50 - 59	12
60 - 69	3
70 - 79	2
80 - 89	1

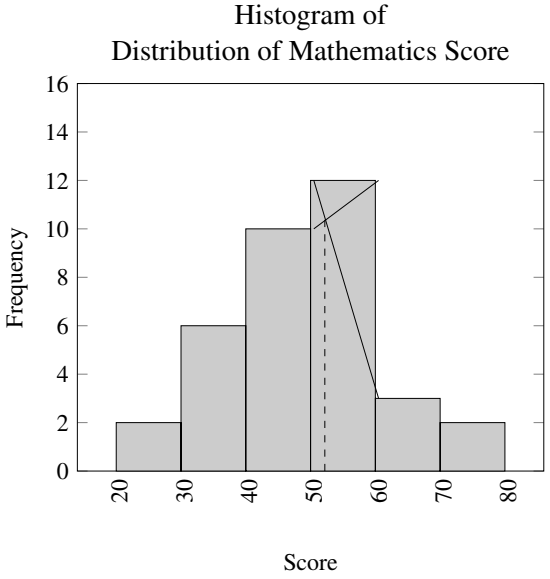
(a) Find the modal class.

Sol.

The modal class is $50 - 59$, which has the highest frequency of 12.

(b) Find the mode of score of the students using histogram.

Sol.



The mode of score of the students is approximately 51.5.

(c) Find the mode of score of the students using formula.

Sol.

$L = 49.5, C = 10, d_1 = 12 - 10 = 2, d_2 = 12 - 3 = 9.$

$$\therefore Mode = 49.5 + \left(\frac{2}{2+9}\right) 10 = 51.32$$

Comparing mean, median and mode

Generally, the mean, median and mode of a set of data are all dirrent, and they are used to describe the data in different ways.

18.3.6 Exercise 18.3c

1. Find the mode of the following data:

   (a) 3 4 3 2 4 5 5 5 4 4

   Sol.

   The mode is 4, which has the highest frequency of 4.

   (b) 7 6 8 8 5 6 6 9 8 5

   Sol.

   The mode is 6 and 8, which has the highest frequency of 3.

   (c) 1.0 1.1 1.0 0.9 0.8 1.2 1.0 0.9 1.1 1.0

   Sol.

   The mode is 1.0, which has the highest frequency of 4.

2. In the sport competition of a high school, the scores of 17 athletes participating in men's high jump are as follows:

Scores (m)	Number of Athletes
1.50	2
1.60	3
1.65	2
1.70	3
1.75	4
1.80	1
1.85	1
1.90	1

Find the mean, median and mode of their scores.

Sol.

$$\begin{aligned} Mean &= \frac{1.50\times 2 + 1.60\times 3 + \cdots + 1.90\times 1}{17} \\ &= \frac{3 + 4.8 + 3.3 + 5.1 + 7 + 1.8 + 1.85 + 1.9}{17} \\ &= \frac{28.75}{17} \\ &= 1.69m \end{aligned}$$

Scores (m)	No. of Athletes	Cum. Frequency
1.50	2	2
1.60	3	5
1.65	2	7
1.70	3	10
1.75	4	14
1.80	1	15
1.85	1	16
1.90	1	17

The median is the number at $\frac{17+1}{2}$ = 9th position, which is $1.70m$.

The mode is $1.75m$, which has the highest frequency of 4.

3. In a Mathematics competition, the scores and the number of students who obtained the scores are as follows:

Scores (%)	Number of Students
10 - 19	20
20 - 29	60
30 - 39	80
40 - 49	40
50 - 59	10

Find the modal class and the mode.

Sol.

The modal clsas is $30-39$, which has the highest frequency of 80.

$L = 29.5, C = 10, d_1 = 80-60 = 20, d_2 = 80-40 = 40.$

$$\therefore Mode = 29.5 + \left(\frac{20}{20+40}\right)10 = 32.83$$

4. Given that the mean of a dataset 3, 5, 8, 6, 8, 10, 5, 3, x, y is 6,

   (a) Prove that $x + y = 12$

   Proof.

   $$\begin{aligned} \frac{3+5+8+6+8+10+5+3+x+y}{10} &= 6 \\ 48 + x + y &= 60 \\ x + y &= 12 \end{aligned}$$

   (b) With that, if

   i. $x = y$

   ii. $x < y$

   Find the mode of the dataset.

5. The mean of a set of data 13, 5, 5, n, 5, 10, 10, 11, 9, n^2 is 7.4,

   (a) Find the possible values of n.

   (b) With that, If

   i. $n > 0$

   ii. $n < 0$

   Find the meadian of the dataset.

6. The following table shows the distribution of scores of a group of students in a competition:

Scores	Number of Students
0	3
1	x
2	4
3	6
4	2

   (a) Assume that the mode is 1, find the minimum value of x.

   (b) Assume that the median is 2, find the maximum value of x.

   (c) Assume that the mean is 1.95, find the value of x.

7. Given thet the mode, median and mean of 5 positive integers are 9, 8, and 7.6 respectively, find these 5 numbers.

8. The following table shows the amount of sales of a brand of shoes in a month:

Shoes Number	Amount of Sales
5	4
6	10
7	11
8	18
9	2

   (a) Find the mean, median, and mode.

   (b) Which of the following central tendency represents the data best? Why?

9. In between 54 examinees in an exam, 15 of them come from cities, 39 of them come from suburbs. Below are the frequency distribution table of their scores:

Scores	City	Suburb
12 - 23	0	1
23 - 34	0	0
34 - 45	0	5
45 - 56	1	6
56 - 67	3	5
67 - 78	4	13
78 - 89	6	4
89 - 100	1	5

   (a) Find the mean, median, and mode of the scores of the examinees from cities and suburbs respectively.

   (b) Find the mean, median, and mode of the scores of all the examinees.

10. The following table shows the distribution of scores of a group of students in a Chinese language test:

Scores x	Number of Students
$40 < x \leq 50$	12
$50 < x \leq 60$	30
$60 < x \leq 70$	35
$70 < x \leq 80$	25
$80 < x \leq 90$	10
$9 < x \leq 100$	3

    Find:

   (a) Mean.

   (b) Modal class and mode.

   (c) Median.

18.4 Measures of Dispersion

The measures of dispersion can be used to describe the spread of the data.

When we're describing a set of data, if we only use the mean, the information provided by the dataset is not enough. For example, given the mean, median, and mode of the average marks of four students in a Mathematics test are all 70 marks, we can't tell the difference between the four students. Thier marks might be similar (e.g. 68, 72, 70, 70) or they might be very different (e.g. 100, 40, 70, 70). The latter case is obviously more spread out than the former case.

The most common measures of dispersion are range, interquartile range, quartile deviation, standard deviation, mean deviation, variance, and standard deviation.

Range

The range of a set of data is the difference between the largest and the smallest value in the dataset.

For grouped data, the range is the difference between the upper limit of the highest class and the lower limit of the lowest class.

Quartile, Interquartile Range, and Quartile Deviation

Quartiles are three value Q_1, Q_2, and Q_3 that divide a dataset into four equal parts. Q_2 is the median of the dataset. Q_1 and Q_3 are the medians of the two halves of the dataset, called the lower quartile and the upper quartile respectively.

Assume that the number of data in a sorted dataset is n. If n is odd, then

When n is even, split the dataset into two halves, with $n/2$ data in each half.

When n is odd, split the data into two halves after removing the median, with $(n-1)/2$ data in each half.

The median of the lower half is Q_1 and the median of the upper half is Q_3.

For grouped data, we can make a cumulative frequency polygon. In the percentage of the polygon,

25% of the data is below Q_1.

50% of the data is below Q_2.

75% of the data is below Q_3.

Using the same method of deriving the formula for median, we can derive the formula for upper and lower quartiles.

Let

n be the number of data in the dataset, aka $\sum f_i$

L_k be the lower boundaries of the class of Q_k

C_k be the class range of the class of Q_k

f_k be the frequency of the class of Q_k

F_k be the cumulative frequency of the class of Q_k

then

$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - F_1}{f_1}\right) C_1$$

$$Q_2 = L_2 + \left(\frac{\frac{3n}{4} - F_2}{f_2}\right) C_2$$

The difference between the upper and lower quartiles is called the interquartile range. That is,

$$\textit{Interquartile range} = Q_3 - Q_1$$

The quartile deviation is the interquartile range divided by 2, written as $Q.D.$, that is,

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Since the interquartile range and the quartile deviation are not affected by the outliers, they are more robust than the range, and are more suitable for representing the spread of the data.

18.4.1 Practice 6

1. Find the range, quartiles and interquartile range of the following data:

   (a) 4 8 7 3 3 9 6 5 1 1 2

   (b) 7 6 8 8 5 6 1 9 8

   (c) 1.0 1.1 1.5 0.7 0.8 1.2 1.4 0.9 1.6 1.3

   (d) 3 4 7 2 4 6 5 8

2. The table below shows the cumulative frequency distribution table of the heights of 60 students:

Height (cm)	Cumulative Frequency
150-155	3
155-160	10
160-165	22
165-170	37
170-175	51
175-180	58
180-185	60

   (a) Find the interquartile range of the heights of the students from the cumulative frequency polygon.

   (b) Find the interquartile range of the heights of the students using formula.

18.4.2 Exercise 18.4a

1. Following are the sales of televisions of a shop in 11 days:

   4 9 0 1 3 4 2 5 7 2 3

   Find:

   (a) The range.

   (b) The quartiles and interquartile range.

2. Given a set of data: 1.2, 1.0, 1.1, 1.3, 1.5, 1.7, 1.2, 1.0. Find:

   (a) The range.

   (b) The quartiles and interquartile deviation.

3. The distribution of scores of Mathematics test of 100 senior 1 students from a high school are as follows:

Scores	Number of Students
30 - 40	3
40 - 50	4
50 - 60	13
60 - 70	22
70 - 80	30
80 - 90	23
90 - 100	5

   Find the interquartile deviation of the scores.

Mean Deviation

Let the mean of a set of data $x_1, x_2, \ldots, x_n$ be $\bar{x}$, $|x_i - \bar{x}|$ is the difference between the ith data and the mean, the mean of these n differences are called the mean deviation, and can be

used to calculate the measure of dispersion of the data. That is,

$$Mean\ Deviation = \frac{\sum |x_i - \bar{x}|}{n}$$

If the possible value given data are x_1, x_2, …, x_n, their frequencies are f_1, f_2, …, f_n, respectively, then the mean deviation can be calculated as follows:

$$Mean\ Deviation = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i}$$

For grouped data, we take the midpoints of the classes as the representative value x_i.

18.4.3 Practice 7

Complete the following table, and find the mean and mean deviation of the data.

Lim.	Freq. f_i	Mid. x_i	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert f_i$
50 - 54	2				
55 - 59	3				
60 - 64	6				
65 - 69	9				

18.4.4 Exercise 18.4b

1. Find the mean deviation of the following dataset:

   (a) 7 10 9 12 4 11 3

   (b) 58 65 38 76 43

   (c) 45.0 46.5 47.0 48.0 48.7 48.9 49.5 50.4

2. The table below shows the frequency of the number of questions answered correctly by 26 students in a Mathematics minor test:

Num. of Corr. Ans. Ques.	Num. of Stud.
1	0
2	1
3	1
4	1
5	6
6	8
7	6
8	1
9	1
10	1

   Find the mean deviation fo the number of questions answered correctly.

3. Following are the test scores of 36 students:

77	60	52	73	60	50	70	60	52
68	59	50	72	59	48	66	58	46
60	48	34	61	55	40	62	55	42
63	55	43	65	56	45	65	57	46

   (a) Group the dataset above according to the pattern $[34-38)$, $[38-42)$, $[42-46)$, …, then make a frequency distribution table.

   (b) Find the mean from the frequency distribution table.

   (c) Find the mean deviation from the frequency distribution table.

Variance

Let the mean of a set of data x_1, x_2, …, x_n be $\bar{x}$, $(x_1 - \bar{x})^2$ be the square of the difference between the i^{th} data and the mean, the square of the mean of these n differences are called the variance, written as σ^2, that is,

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

The square root of the variance is called the standard deviation, written as σ, that is,

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

If the possible values of given data are x_1, x_2, …, x_n, their frequencies are f_1, f_2, …, f_n, respectively, then

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$$

For grouped data, we take the midpoints of the classes as the representative value x_i.

The above formula are a bit complicated, so we can sim-

plify the formula:

$$\begin{aligned}\sigma^2 &= \frac{\sum(x_i-\bar{x})^2 f_i}{\sum f_i} \\ &= \frac{\sum x_i^2 f_i - 2\bar{x}\sum x_i f_i + \sum \bar{x}^2 f_i}{\sum f_i} \\ &= \frac{\sum x_i^2 f_i}{\sum f_i} - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2\end{aligned}$$

Hence, when the frequency of value x_i is f_i, Then

$$\sigma^2 = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2$$

$$\sigma = \sqrt{\frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2}$$

When all the frequencies f_i are equal to 1, then

$$\sigma^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$\sigma = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

Compared to mean deviation, the variance and standard deviation do not contain absolute value, so it is more convenient to use them. Furthermore, the variance and standard deviation are more sensitive to the difference between the data and the mean, so they are more commonly used in daily life.

18.4.5 Practice 8

1. Measuring the height of 10 plant seedlings (in cm) in a lab, we get the following data:

   12 6 15 3 12 6 21 15 15 18 12

   Find the standard deviation of the height of the plant seedlings.

2. Complete the following table, then find the standard deviation.

   (a)

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
3	30		
5	35		
7	28		

   (b)

Limit	f_i	$Mid.x_i$	$x_i f_i$	$x_i^2 f_i$
150 - 154	5			
155 - 159	8			
160 - 164	10			
165 - 169	7			
170 - 174	6			
175 - 179	4			

18.4.6 Exercise 18.4c

1. Find the variance and standard deviation of the following dataset:

   (a) 3 6 3 8

   (b) 3 3 4 5 10

   (c) 2 9 10 10 12 2 10 9

2. Find the variance and standard deviation of the data:

   (a)

Values	Frequency
6	35
5	36
4	30

   (b)

Values	Frequency
60	4
70	6
80	2
90	5
100	1

3. Given two sets of data:

A	B
9.9	10.3
10.3	10
9.8	9.5
10.1	10.4
10.4	10.5
10	9.4
9.8	9.8
9.7	10.1

   Find the mean and variance of these two sets of data respectively, and state which set of data is more spread out.

4. Given the Chinese language test scores of two groups of students are as follows:

Group A	Group B
76	82
90	84
84	85
86	89
81	79
87	80
86	91
82	89
85	79
83	74

Find the mean and standard deviation of these two sets of data respectively, and state which set of data is more centered.

5. The table below shows the height distribution of all students of the same grade:

Height (cm)	Frequency
145 - 149	10
150 - 154	36
155 - 159	193
160 - 164	205
165 - 169	240
170 - 174	83
175 - 179	33

Find the mean and standard deviation of the height of all students of the same grade.

6. Following are teh weight distribution of 100 students in a school:

Weight (kg)	Number of Students
45 - 47	3
48 - 50	16
51 - 53	20
54 - 56	32
57 - 59	15
60 - 62	10
63 - 65	4

Find the variance and standard deviation.

7. Given the sum of 10 values is 400, and the sum of their square is 16400. Find the mean and standard deviation of these 10 values.

8. Given 30 values $x_1, x_2, \cdots, x_{30}$, the mean of these values is 5, and the standard deviation is 2. Find $\sum_{i=1}^{30} x_i$ and $\sum_{i=1}^{30} x_i^2$.

9. The mean of 5 values is 10, and it remains the same after adding p to dataset.

   (a) Find the value of p.

   (b) If the sum of square of 5 original values is 558, find the variance of the 6 values after adding p.

10. Given that the mean of 3, 6, 7, 8, 9, 12, 14, 15, x, y is 13, standard deviation is $\sqrt{102}$, find the value of x and y.

18.5 Coefficient of Variation

Generally speaking, when we want to compare the variability of two or more sets of data, only comparing the standard deviation of each group is not enough. If the properties or the units of the data are different, the standard deviation of each group must not be comparable. For example, if we want to know whether the deviation of the height of students in a class is larger than that of the weight of students in the same class, we need a relative metric as the standard of comparison, and the coefficient of variation is such a metric. For a non-negative set of value, the definition of coefficient of variation is as follows:

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100\%$$

From the definition, we can see that the coefficient of variation the standard deviation when the mean is 1. Thus, when the coefficient of variation is large, it means that the variability of the data is large, and vice versa.

18.5.1 Practice 9

In a minor test, the full mark of Chinese language test for senior 2 students is 100, its average mark is 70, and the standard deviation is 10, while the full mark of Mathematics test is 70, its average mark is 40, and the standard deviation is 8. Compare the variability of the two tests.

18.5.2 Exercise 18.5

1. The statistics of the height and width if grade 1 students in a primary school are as follows:

	Mean	Standard Deviation
Height (cm)	115.87	4.86
Width (cm)	19.39	2.16

Compare the variability of the height and width of the students.

2. The table below shows the first semester Mathematics exam average mark and standard deviation of five junior 1 classes in a school:

Class	Average Mark	Standard Deviation
A	62	11
B	74	9
C	65	10
D	70	7
E	53	8

Which class has the smallest coefficient of variation?

3. The table below shows the Mathematics exam results of two groups of students A and B:

Group	Marks				
A	60	98	76	84	52
B	88	58	90	69	78

(a) Find the average mark of each group.

(b) Find the standard deviation of each group.

(c) Find the coefficient of variation of each group.

4. The table below shows the price of of papayas and grapes per kilogram in the first half of the year (in $):

Month	Papaya	Grapes
January	3.50	20.00
February	3.00	22.00
March	2.50	24.00
April	3.20	23.00
May	3.60	18.00
June	2.80	21.00

(a) Find the average price and standard deviation of papayas and grapes respectively in the first half of the year.

(b) Which fruit has greater variability in price?

5. The table below shows the distribution of annual average marks of two classes of students A and B:

Marks Range	Class A	Class B
40 - 49	3	4
50 - 59	4	10
60 - 69	10	17
70 - 79	16	14
80 - 89	12	1

Find the coefficient of variation of annual average marks of each class respectively.

18.6 Correlation and Correlation Coefficient

Correlation

In statistics, correlation is a statistical measure of the degree to which two or more variables move in relation to each other. For example, the correlation between the height and weight of a person, the correlation between the price of a stock and the volume of the stock traded.

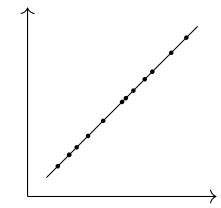
Scatter Plot

A scatter plot is a type of mathematical diagram to show the relationship between two variables. Let two groups of data be $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$, respectively. The scatter plot of the two groups of data is a graph of the points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.

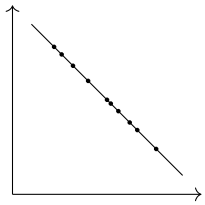
Linear Correlation

If the scatter plot of two groups of data can be approximated by a straight line, then the two groups of data are said to be linearly correlated. According to the trend of the two groups of data, the correlation can be positive, negative, or zero. For example, the weight of a higher person is usually larger, so the correlation between the weight and height of a person is positive. The sales of a product are usually lower when the price of the product is higher, so the correlation between the price of a product and the volume of the product sold is negative. If there is no relationship between the two groups of data, then it is considered zero correlation.

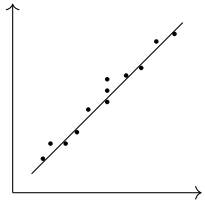
Below are the possible cases of linear correlation:



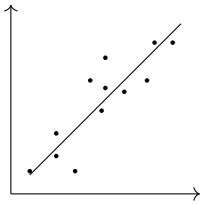
Perfect Positive Correlation



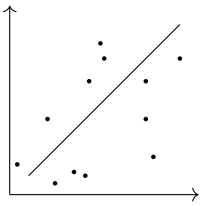
Perfect Negative Correlation



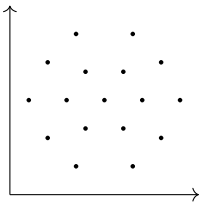
Strong Positive Correlation



Moderate Positive Correlation



Weak Positive Correlation



Zero Correlation

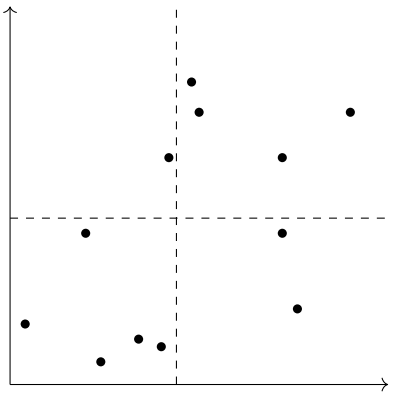
1. If every single point in the scatter plot is on the line of best fit, then it's a perfect positive correlation. If the slope of the line of best fit is positive, then it's a positive correlation. If the slope of the line of best fit is negative, then it's a negative correlation.

2. If the points in the scatter plot are scattered around the line of best fit with non-zero slope, then the closer the points are to the line of best fit, the stronger the correlation is.

3. If the points in the scatter plot are scattered evenly around the whole plot with no obvious pattern, then there is no correlation between the two variables, aka zero correlation.

Correlation Coefficient

Telling the correlation between two variables by looking at the scatter plot is not a very accurate way. To accurately measure the correlation between two sets of data, we need to use a coefficient that can distinguish the strength of the correlation.



Let the mean value of two sets of data be $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be $\bar{x}$ and $\bar{y}$ respectively. Draw two lines $x = \bar{x}$ and $y = \bar{y}$ on the scatter plot of the two sets of data, splitting the plot into four quadrants, as shown in the figure above. Now the origin of the plot is at $(\bar{x}, \bar{y})$. If a point (x_i, y_i) is in the first or the third quadrant, then $(x_i - \bar{x})(y_i - \bar{y})$ is positive. As discussed in the previous section, if the correlation is positive, the points are scattering around the line of best fit with positive slope. Therefore, the points are more likely to be in the first or the third quadrant. That means, there are more positive value of $(x_i - \bar{x})(y_i - \bar{y})$ than negative value, therefore the value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ is positive. The higher the correlation is, the more points are in the first or the third quadrant, the higher the positive value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ is.

On the other hand, if a point (x_i, y_i) is in the second or the fourth quadrant, then $(x_i - \bar{x})(y_i - \bar{y})$ is negative, which means there are more negative value of $(x_i - \bar{x})(y_i - \bar{y})$ than positive value, therefore the value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ is negative. Similarly, he higher the correlation is, the lower the negative value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ is.

Hence, the value and the sign of $\sum (x_i - \bar{x})(y_i - \bar{y})$ can be used to measure the correlation between two sets of data. The value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ will be affected by the measurement unit of the data. To make the value of $\sum (x_i - \bar{x})(y_i - \bar{y})$ independent of the measurement unit, we define the correlation coefficient of two sets of data $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The value of r is always between -1 and 1. If $r = 0$, then there is no correlation between the two sets of data. If $r > 0$, then the correlation is positive. If $r < 0$, then the correlation is negative. The absolute value of r is the strength of the correlation, and is generally divided as follows:

1. $|r| = 1$: perfect correlation
2. $0 < |r| < 0.3$: weak correlation
3. $0.3 \leq |r| < 0.7$: moderate correlation
4. $0.7 \leq |r| \leq 1$: strong correlation

Dividing both the denominator and the numerator of the formula of r by the number of data points n, then the numerator is the mean value of $(x_i - \bar{x})(y_i - \bar{y})$, and the denominator is the product of the standard deviation of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$. Similar to the standard deviation, there is an easier way to calculate the correlation coefficient:

$$r = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\sqrt{(\frac{\sum x_i^2}{n} - \bar{x}^2)(\frac{\sum y_i^2}{n} - \bar{y}^2)}}$$

18.6.1 Practice 10

1. The table below shows the height (in cm) and weight (in kg) of 15 10-year-old children:

Height	Weight
126	41
130	42
110	38
123	36
118	33
130	45
127	34
124	35
116	30
112	32
113	31
121	40
115	34
120	35
118	33

Calculate the correlation coefficient of the height and the weight of the 15 children, and determine on the strength of the correlation.

2. In order to study the relationship between the systolic blood pressure (in $mmHg$) and the age (in *year*) of human, a medical school collected the data of 13 male patients:

Age	Systolic Blood Pressure
51	130
22	141
23	124
31	126
33	117
49	135
58	143
53	138
44	132
55	143
42	133
45	115
25	147

18.6.2 Exercise 18.6

18.7 Statistical Index

Chapter 19

Permutations and Combinations

Chapter 20

Bionomial Theorem

20.1 Bionomial Theorem when n is a Natural Number

20.2 General Form of Bionomial Expansion

Chapter 21

Probability