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:

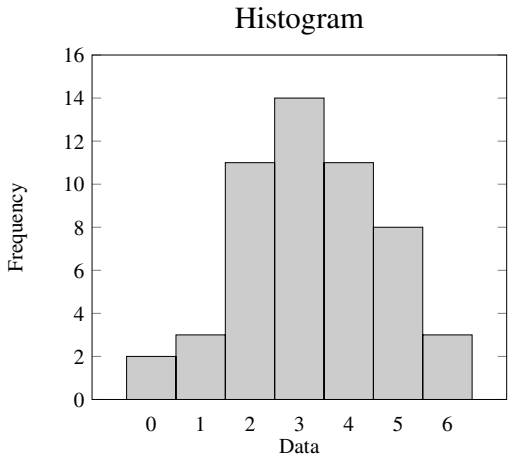
Weight $m(g)$	Frequency
$1.27 - 1.29$	1
$1.30 - 1.32$	4
$1.33 - 1.35$	7
$1.36 - 1.38$	22
$1.39 - 1.41$	24
$1.42 - 1.44$	24
$1.45 - 1.47$	10
$1.48 - 1.50$	6
$1.51 - 1.53$	1
$1.54 - 1.56$	1

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 $\frac{1.265+1.295}{2} = 1.28$.

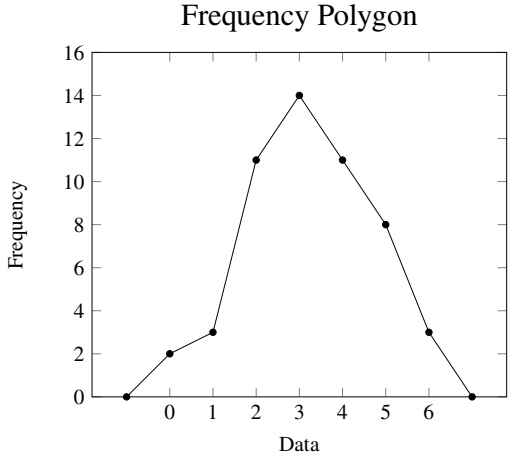
When we are analyzing the data data that have been classified into classes, the midpoint of each class is used as the representative value of the class. Thus, we should try our best to make the data-intensive place the group midpoint when choosing the class interval and boundaries, so that the data can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

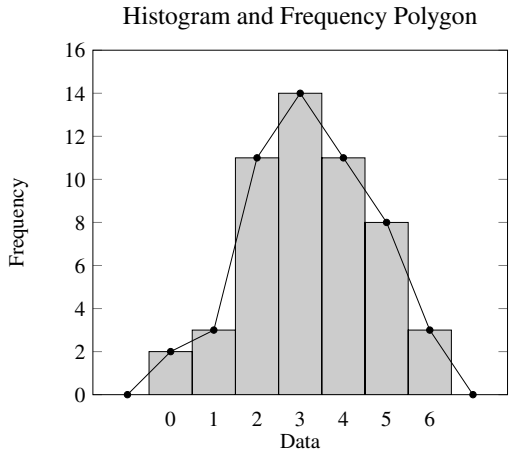
The distribution of frequency can be represented by a histogram or a frequency polygon.

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.





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.

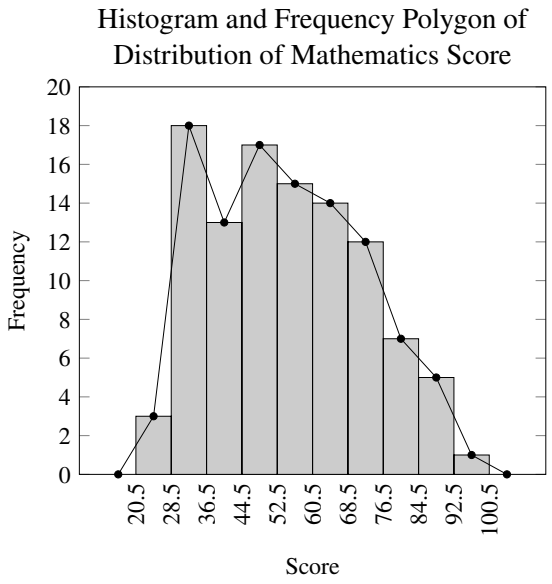
Sol.

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

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.

Practice 2

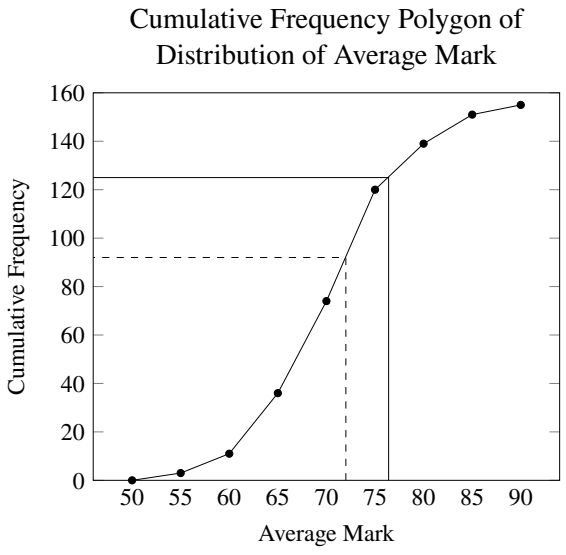
There are 155 students in a senior 3 art and commerce class, and the frequency distribution table of their average marks is shown below:

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.

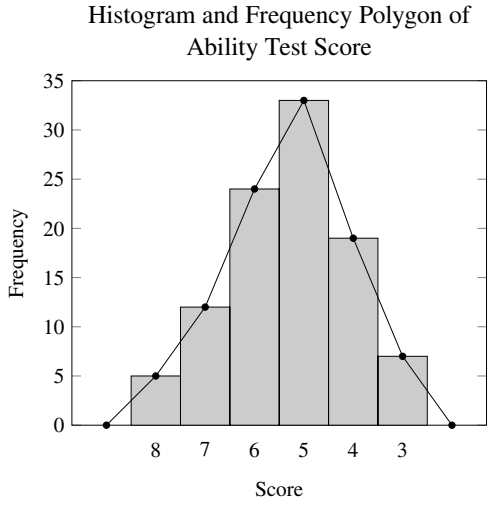
Exercise 18.2

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.

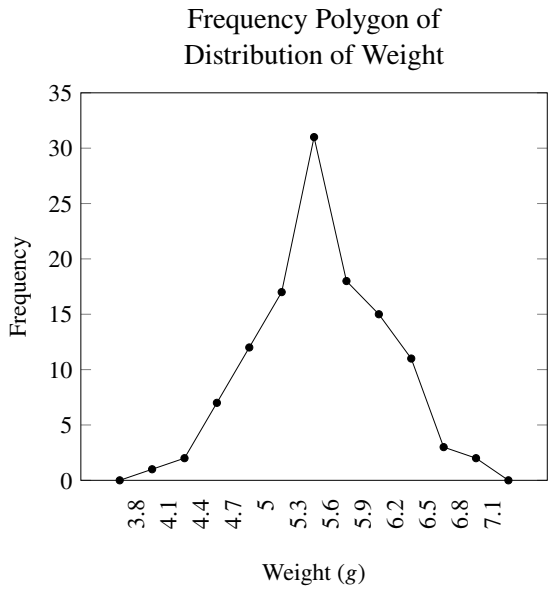
$$\begin{aligned}\text{Range} &= 3.5\\ \text{Number of classes} &= 12\\ \therefore \text{Class width} &= \frac{3.5}{12}\\ &= \frac{3.5}{12}\\ &\approx 0.3\end{aligned}$$

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.

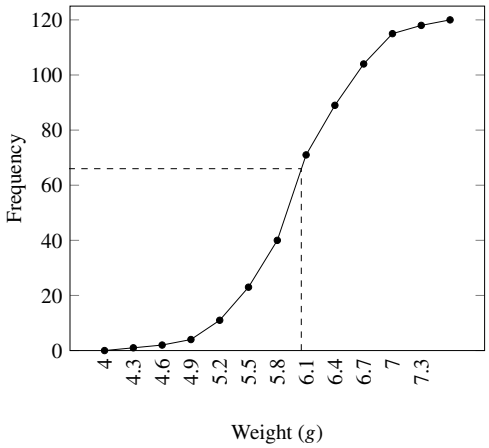
Sol.



(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.

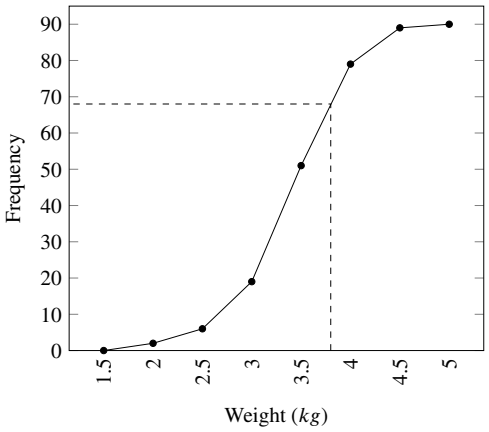
Sol.

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:

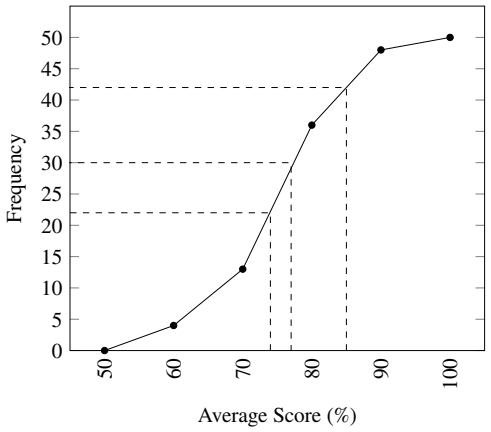
Average Score	Frequency
50.0 - 59.9	4
60.0 - 69.9	9
70.0 - 79.9	23
80.0 - 89.9	12
90.0 - 99.9	2

(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

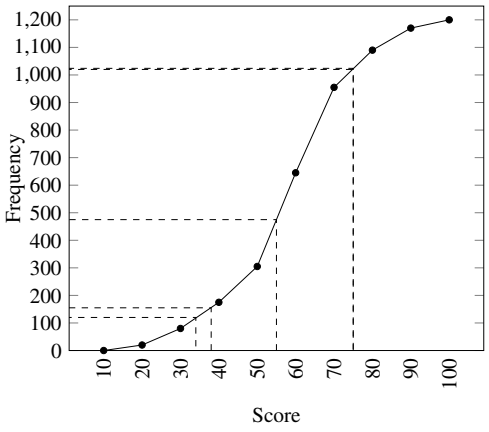
Examinees are categorised into 4 groups based on their score: *Excellent*, *Good*, *Pass*, and *Fail*.

(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.

Practice 3

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

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.

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			

18.3.1 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.

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

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.

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.

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.

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.

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.

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.

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

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

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 $\Delta PQR \sim \Delta PST$,

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

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

We get the following after simplifying the equation:

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

18.3.2 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.

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.

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.

Mode

18.4 Measures of Dispersion

18.5 Coefficient of Variation

18.6 Correlation and Correlation Coefficient

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