Chapter 16 Introduction to Statistics

In our daily life, we often collect data and accumulate information. We would then classify the data, study and analyse them until we can detect some pattern and extract useful information from them. The branch of mathematics specialising in the study and analysis of data in an organized manner is called statistics. This chapter teaches some elementary knowledge of statistics and some useful statiscal methods.

16.1 Population and sample

Let us study the following examples:

1. A department would like to study the body weight of Grade 9 students in a certain region so as to understand more about the health and body growth of Grade 9 students. Therefore, all the Grade 9 students in the region will be the subject of the study. However, as there are so many Grade 9 students in the region, a complete examination of all Grade 9 students individually will take too long and cost too much effort and is therefore not practicable. Therefore, only a sample of the students (e.g. 200 students) can be selected for examination. The averge body weight of the sample can be used to to infer the body weight of all the Grade 9 students in the region.

2. To know the corn production rate of a cetain maize farm (e.g. the average number of ears of corn produced by a maize plant), it can be established that all the maize plant in the farm will be the subject of the study However, there are so many maize plants in the farm that it is not practicable to examine all the maize plant individually. So only a sample of maize plant is selected (e.g. 100 pieces) for the study. Based on the average number of ears of coin produced by this sample of 100 pieces of maize plants, we can infer the average production rate of coins in this maize farm.

3. To examine the casualty radius of certain batch of bombs (e.g. to examine the average casualty radius). This kind of examination is destructive because the bomb has to be detonated before measurement of the casualty radius can be made and that the detonated bomb cannot be used again. Even though there are not many bombs in the batch, we cannot examine every one of them but have to extract a sample (e.g. 10 bombs) for examination. Then we can use the average casualty radius of this sample of 10 bombs to infer the average casualty radius of all the bombs in this batch.

The whole group of objects to be examined is called the **population**, in which the individual object under study is called an **element**. The set of elements being extracted from the population is called a **sample**. The number of elements in the sample is called the **sample size**.

In the first example, the whole group of Grade 9 students in the region is the population. Each student is an element. The 200 students extracted is a sample selected from the population. The sample size is 200.

In the second example, all the maize plants in the maize farm form the population. Each maize plant is an element. The 100 maize plants extracted is a sample selected from the population. The sample size is 100.

In the third example, the whole batch of bombs is the population. Each bomb is an element. The 10 bombs extracted from the batch is a sample selected from the population. The sample size is 10.

As we can see, the object under study in statistics is a quantitative measure of a certain attribute. In the first example, the object under study is the body weight of all Grade 9 students in the region, not the study of the students in general.

It is common to have a lot of elements in a population (such as in Example 1 and Example 2). Sometimes, there can also be a limited number of elements in a population but examination of which is destructive (such as Example 3). In either case, we can only afford to extract and analyse a sample selected from the population, and to use the sample result to predict the corresponding characteristics of the population.

Practice

1. From each of the following questions, describe what is the population, the element, the sample and the sample size:
   (1) To examine the situation of mathematics examination taken by graduates of junior high school in a certain region, the examination result of 200 students are extracted.
   (2) To understand the life of certain batch of light bulbs, 10 light bulbs are extracted for examination.
   (3) To examine the daily number of visitors of a park in the year, the number of visitors of 30 days within the year is extracted for statistics.

2. Give a real life example of examining a population through sampling.

16.2 Average

The results of a Mathematics test of Group 1 students in a class are:

86, 91, 100, 72, 93, 89, 90, 85, 75, 95.

What is the average result of students in this group?

Obviously, the average result is:

$$\frac{86+91+100+72+93+89+90+85+75+95}{10}=87.6$$

In general, if there are ***n*** numbers,

$$x_1,\ \ x_2,\ ...,x_n,$$

then

$$\bar{\boldsymbol{x}}=\frac{\mathbf{1}}{\boldsymbol{n}}(\boldsymbol{x}_1+\boldsymbol{x}_2+\cdots+\boldsymbol{x}_n) \qquad \textbf{(1)}$$

is called the average of these ***n*** numbers, The symbol $\bar{x}$ is read as "*x*-Bar".

For the convenience of writing, sometimes $x_1+x_2+\cdots+x_n$ will be written as $\sum_{i=1}^{n} x_i$, where Σ is the summation sign, read as "sigma" . $\sum_{i=1}^{n} x_i$ is read as $\Sigma-x-i$, where *i* runs from 1 to *n*. The symbol stands for the sum of data from x_1 to x_n. Therefore, Formula (1) can be written as

$$\bar{\boldsymbol{x}}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}\boldsymbol{x}_{\boldsymbol{i}} \qquad \textbf{(1)}$$

Practice

1. Calculate the average of each of the following group of numbers:
   (1) 1.4, 2.8, 5.3, 3.7, 2.3;
   (2) -0.3, 0.2, 0.3, -0.4, 0.6, 0.3, 0.5, -0.4.
2. Re-write the following expression using the summation sign Σ:
   (1) $P_1+P_2+\cdots+P_{10}$;
   (2) $y_1+y_2+\cdots+y_n$.

In a final examination, there were 10,000 candidates. We would like to understand their average result in Mathematics. However, if we add the results of all the candidates and divide it by the total number of candidates, it will take a lot of time and effort and is not feasible. In this case, we can use the method of sampling to infer the average for the population. That is to say, extract a sample of the candidates, calculate the average result for the sample of candidates and use the average result of the sample to infer the average result of all candidates.

The average of all elements in a population is called **population average**. The average of all elements in a sample is called **sample average**. For the result of the population of 10,000 candidates, the average result of all candidates is the population average while the average result of the sampled candidates is the sample average. We usually use the sample average to infer the population average. In general, the larger the sample size, the more accurate the prediction will be. For example, when more candidates are extracted, the sample average will be closer to the population average.

【Example 1】 From the candidates of the final examination, the result of 30 candidates were extracted as follows:

90	84	84	86	87	98	78	82	90	93
68	95	84	71	78	61	94	88	77	100
70	97	85	68	99	88	85	92	93	97

Calculation the sample average (round to the nearest integer).

Solution $\bar{x} = \dfrac{1}{30}(90 + 84 + \cdots + 97) = \dfrac{2562}{30} \approx 85$

i.e. The sample average is 85.

Therefore, it can be inferred that the average result of the candidates taking the examination is 85.

【Example 2】 From a batch of machine parts, 20 pieces are extracted and the weight of the 20 pieces are as follows (Unit: kg):

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

Calculate the sample average (round to the nearest integer).

Solution $\bar{x} = \dfrac{1}{20}(210 + 208 + \cdots + 215) = \dfrac{4129}{20} \approx 206\,\text{kg}.$

i.e. The sample average is 206 kg.

Therefore, it can be inferred that the average weight of each machine part in the whole batch is approximately 206 kg.

As the data values of the sample in this example are large, varying about 200, we can simplify our work by modifying the method as follows:

By substracting 200 from each of the data above, we obtain a new set of data:

10	8	0	5	2	18	6	14	15	7
–5	7	18	–8	2	16	–15	27	–13	15

By calculating the average of this new set of data, we obtain:

$$\bar{x}' = \frac{1}{20}(10 + 8 + \cdots + 15) = \frac{129}{20} \approx 6$$

Adding back the value of 200 subtracted, we obtain the answer of the the average of the sample as:

$$\bar{x} = \bar{x}' + 200 \approx 6 + 200 = 206$$

The results of both methods are the same. We can see that using the modified method, the calculation involves smaller value of data and is thus easier to perform.

In general, if we substract a constant number a [2] from each of the data x_1, x_2, ..., x_n, we will get:

$$x_1' = x_1 - a\,,\quad x_2' = x_2 - a\,, \ldots,\quad x_n' = x_n - a$$

or

$$x_1 = x_1' + a\,,\quad x_2 = x_2' + a\,, \ldots,\quad x_n = x_n' + a\,.$$

Therefore,

$$\begin{aligned}
\bar{x} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\
&= \frac{1}{n}[(x_1' + a) + (x_2' + a) + \cdots + (x_n' + a)] \\
&= \frac{1}{n}[(x_1' + x_2' + \cdots + x_n') + na] \\
&= \frac{1}{n}(x' + x_2' + \cdots + x_n') + \frac{1}{n} \bullet na \\
&= \bar{x}' + a
\end{aligned}$$

That is,

$$\boxed{\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}' + \boldsymbol{a}} \qquad \textbf{(2)}$$

This equation (2) is the one used in the modified method in Example 2.

[2] We usually rounded the sample average to a suitable integer as the constant number *a*.

【**Example 3**】A worker processes a type of machine part. Reviewing his work for the past 30 days, his production quantities were: 2 days with 51 pieces, 3 days with 52 pieces, 6 days with 53 pieces, 8 days with 54 pieces, 7 days with 55 pieces, 3 days with 56 pieces, 1 day with 57 pieces. Calculate the average daily production quantity of the 30 days (round to the nearest integer).

Solution From the 30 data above, 51 appears twice, 52 appears 3 times, 53 appears 6 times, 54 appears 8 times, 55 appears 7 times, 56 appears 3 times, 57 appear once. As this set of data is slightly larger than 50, we can make use of Formula (2) by substituting *a* as 50 to calculate their average.

By subtracting 50 from each of the data 51, 52, 53, 54, 55, 56, 57, we have

1 2 3 4 5 6 7

Their corresponding frequencies of appearance are:

2 3 6 8 7 3 1

Then, the average of the new set of data is:

$$\overline{x}' = \frac{1\times2+2\times3+3\times6+4\times8+5\times7+6\times3+7\times1}{30} = \frac{118}{30} \approx 4$$

According to Formula (2),

$$\overline{x} = \overline{x}' + a \approx 4 + 50 = 54 \text{ pieces},$$

that is to say the average daily production quantity of his work in the 30 days is 54 pieces.

Therefore, we can infer that this worker can produce an average of 54 pieces daily beyond the 30 days under study.

In general, for *n* numbers where x_1 appears f_1 times, x_2 appears f_2 times, ..., x_k appears f_k times (so that $f_1 + f_2 + \cdots + f_k = n$). Then, according to Formula (1), the average of these *n* numbers can be expressed as

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}}{\boldsymbol{n}} \quad (1')$$ [3]

or simplified as

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{k} \boldsymbol{x_i f_i} \quad (1')$$

As in Example 3, when certain samples appear multiple times, it may be easier to make use of Formula (1′) to calculate the sample average.

Practice

1. Extracted 10 trees to measure their heights and the results are as follows (Unit:cm):

   25 41 40 43 22 14 19 39 21 42

   Calculate the sample average (round to the nearest integer).
2. By making use of Formula (2), calculate the average of each of the following groups of data (round to the nearest integer):

   (1) 105, 103, 101, 100, 114, 108, 110, 106, 98, 102;

   (2) 4203, 4204, 4200, 4194, 4204, 4201, 4195, 4199;
3. We sampled 24 male classmates from six classes of the same grade in a high school and measured their heights. The results are as follows (Unit: cm):

   155 157 159 162 162 163 164 164
   165 165 165 166 166 167 167 167
   168 169 169 170 171 171 172 174

   What can be inferred about the average height of the male classmates of this grade (round to the nearest integer)?
4. Within the 40 students in a class, there are 5 students aged 14, 30 students aged 15, 4 students aged 16 and 1 student aged 17. Calculate the average age of the students in this class (round to the nearest 1 decimal place).

Exercise 12

1. The height of the 20 members of a football team in a high school are measured as follows: (Unit: cm):

   170 167 171 168 160 172 168 162 172 169
   164 174 169 165 175 170 165 167 170 172

   Calculate the average height (round to the nearest integer).

[3] This average is called **weighted average**, where f_1, f_2, ..., f_k are called **weights**.

2. To examine the quality of a batch of machine parts, 10 pieces were extracted and their lengths were measured as follows (Unit: cm):

22.36	22.35	22.33	22.35	22.37
22.34	22.38	22.36	22.32	22.35

   (1) In this question, please describe what is the population, the element, the sample and the sample size.

   (2) Calculate the sample average (round to the nearest 2 decimal places). (Hint: To apply Formula (2) for the calculation, you may use $a = 22.30$).

3. In a experimental tree farm, there are many trees. 20 trees were selected and their heights were measured as follows (Unit: cm):

346	294	365	315	339	313	317	305	321	325
315	329	324	329	368	336	366	308	301	362

   (1) In this question, what is the population and what is the sample?

   (2) Calculate the sample average (round to the nearest integer).

4. An experimental farm would like to choose a better quality of hybrid rice. It selects 10 testing locations and records the production quantity of species A and species B, with results as follows:

Species	Production quantity in each testing location (Unit: kg)									
	1	2	3	4	5	6	7	8	9	10
A	779	817	853	793	740	963	794	777	875	864
B	808	771	725	750	750	859	745	740	706	824

   Which species has a higher average production quantity?

5. In a language test, the students in a class achieved the following result: 7 students scored 100, 14 students scored 90, 17 students scored 80, 8 students scored 70, 2 students scored 60 and 2 students scored 50. Calculate the average score of the whole class for this test. (round to the nearest integer)

6. If the averages of x_1, x_2, ..., x_n and y_1, y_2, ..., y_n are $\overline{x}$ and $\overline{y}$ respectively, then what is the average of:

$$x_1 + y_1,\ x_2 + y_2,\ ...,\ x_n + y_n$$

   And why?

16.3 Variance

Two production plants A and B produce the same machine part with diameter specification of 40mm. We extracted from each plant a sample of 10 pieces and measured their diameters. Results are as follows: (Unit: mm)

Plant A	40	39.8	40.1	40.2	39.9	40	40.2	39.8	40.2	39.8
Plant B	40	40	39.9	40	39.9	40.2	40	40.1	40	39.9

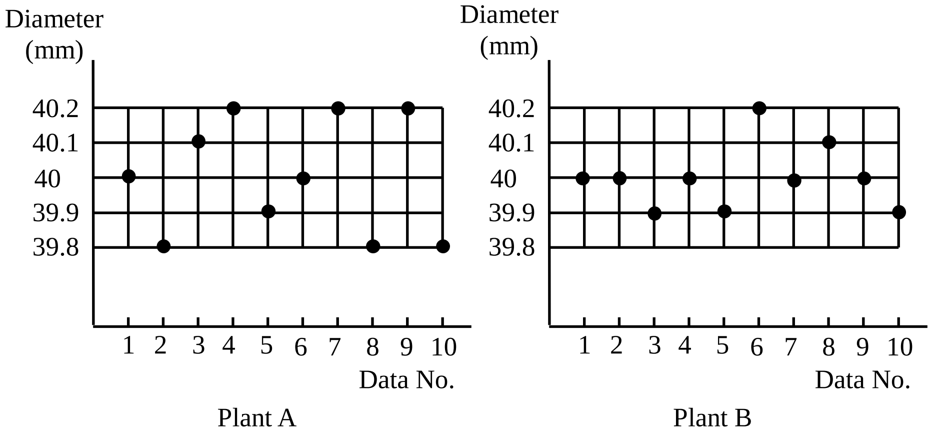
By using Formula (2) and substituting $a = 40$, the average diameter of each of the two production plants are:

$$\overline{x_{\text{A}}} = 40 + \frac{1}{10}[0 + (-0.2) + \cdots + (-0.2)] = 40$$

$$\overline{x_{\text{B}}} = 40 + \frac{1}{10}[0 + 0 + \cdots + (-0.1)] = 40$$

That is to say, the sample average diameter is 40 mm for both production plants. The question is: Based on the fact that the average diameter is the same for both production plants, can we conclude that the quality of production of machine part is the same in both production plants?

The data from the above table are illustrated in diagram 16-1. From the figure, we can see that the diameter of the machine part producted by plant A has a larger variation from the required diameter, being deviated relatively more away from 40mm, while that produced by plant B has a smaller variation from the required diameter, being relatively closer to 40mm. This explains that in meeting the required diameter of the machine part, plant B performs better than plant A.



Plant A Plant B

Diagram 16-1

The question is: Can we quantify in figure the level of deviation of the sample? In this example, the deviations of the diameters of the 10 machine part produced by plant A from the required diameter of 40 as produced by plant A are:

$$0 \quad -0.2 \quad 0.1 \quad 0.2 \quad -0.1 \quad 0 \quad 0.2 \quad -0.2 \quad 0.2 \quad -0.2$$

It may be natural to think of using the sum of the above deviations as an measurement. However, this idea does not work. It is not difficult to verify that the positive deviations and negative deviations will net off each other so that the sum of all the data deviations is zero. To overcome this controversy, there can be a number of possible methods. One method is to square the above deviations and sum up the squares as a measure of the total deviations. As squared numbers cannot take negative values, the measure of the total deviations is always a positive number. However, this will lead to another question. The sum of square of deviations are related to the sample size. The larger the sample size, the greater is the sum of the square of deviations. To eliminate the effect of the sample size, we shall divide the sum of the square of deviations by the sample size *n*. That is to say, we will use the average of the square of deviations (represented by s^2) to determine the level of deviation of the samples. Therefore, for plant A,

$$\begin{aligned} s^2 &= \frac{1}{10}[(40-40)^2+(39.8-40)^2+\cdots+(39.8-40)^2] \\ &= \frac{1}{10}[0^2+(-0.2)^2+\cdots+(-0.2)^2] \\ &= \frac{1}{10}\times 0.26 \\ &= 0.026 \quad (\text{mm}^2) \end{aligned}$$

and for plant B,

$$\begin{aligned} s^2 &= \frac{1}{10}[(40-40)^2+(40-40)^2+\cdots+(39.9-40)^2] \\ &= \frac{1}{10}[0^2+0^2+\cdots+(-0.1)^2] \\ &= \frac{1}{10}\times 0.08 \\ &= 0.008 \quad (\text{mm}^2) \end{aligned}$$

Because 0.026 > 0.008, it explains that the deviation from the required diameter produced by plant A is larger than those producted by plant B.

This method of calculating the average of the square of the deviations (difference between each data and the sample average) is called **sample variance**, i.e.,

$$\boldsymbol{s^2 = \frac{1}{n}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2]}. \qquad (3)$$

Using the Sigma notation, the formula is: conveniently written as:

$$\boldsymbol{s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

When the sample size is very large, the sample variance will be very close to the population variance. It is common to use sample variance to infer the population variance. In this case, by comparing the variances of 2 samples, we are inferring a comparison of the different populations from which the samples are being extracted.

【**Example 3**】Given 2 samples:

A:	9.9	10.3	9.8	10.1	10.4	10	9.8	9.7
B:	10.2	10	9.5	10.3	10.5	9.6	9.8	10.1

Calculate the 2 sample variance respectively.

Solution According to Formula (2) (by substituting $a = 10$), we have

$$\overline{x}_{\text{A}} = 10 + \frac{1}{8}(-0.1 + 0.3 - 0.2 + 0.1 + 0.4 + 0 - 0.2 - 0.3)$$

$$= 10 + \frac{1}{8} \times 0 = 10$$

$$\overline{x}_{\text{B}} = 10 + \frac{1}{8}(0.2 + 0 - 0.5 + 0.3 + 0.5 - 0.4 - 0.2 + 0.1)$$

$$= 10 + \frac{1}{8} \times 0 = 10$$

Therefore,

$$s_{\text{A}}^2 = \frac{1}{8}[(9.9 - 10)^2 + (10.3 - 10)^2 + \cdots + (9.7 - 10)^2]$$

$$= \frac{1}{8}[(-0.1)^2 + 0.3^2 + \cdots + (-0.3)^2]$$

$$= \frac{1}{8}(0.01 + 0.09 + \cdots + 0.09)$$

$$= \frac{1}{8} \times 0.44 = 0.055$$

$$s_{\text{B}}^2 = \frac{1}{8}[(10.2 - 10)^2 + (10 - 10)^2 + \cdots + (10.1 - 10)^2]$$

$$= \frac{1}{8}[0.2^2 + 0^2 + \cdots + 0.1^2]$$

$$= \frac{1}{8}(0.04 + 0 + \cdots + 0.01)$$

$$= \frac{1}{8} \times 0.84 = 0.105$$

Since $s_{\text{A}}^2 < s_{\text{B}}^2$, we know that sample B fluctuates more than sample A.

Under certain circumstances, we will take the square root of the sample variance, i.e.,

$$\boldsymbol{s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}} \qquad (4)$$

It is called the **standard deviation**. It is also used to measure the level of deviation of data in the sample. The units of measure used in sampled data and sample variance are not the same. For example, in the example of the diameter of machine part, sampled data are measured in mm while sample variance was measured in mm^2. However, the units of measure in sampled data and sample standard deviation are the same.

In the example of diameter of machine parts, the 2 sample standard deviations respectively are:

$$s_{\text{A}} = \sqrt{0.026} \approx 0.16 \quad \text{(mm)}$$

$$s_{\text{B}} = \sqrt{0.008} \approx 0.089 \quad \text{(mm)}$$

Practice

1. Calculate the sample variance and sample standard deviationof the following (round to the nearest 1 decimal place)
   (1) -2, 2, 0, -3, -2, 3, 0, 1;
   (2) 28, 24, 25, 23, 27, 24, 22, 24, 25, 28.
2. Worker A and Worker B produce the same type of machine part. We extracted 5 of them from each worker and measured their diameters (Unit: mm) as follows (the diameter required by the specification is 10mm):

Diameter of machine part produced by Worker A	10.05	10.02	9.97	9.96	10.00
Diameter of machine part produced by Worker B	10.00	10.01	10.02	9.97	10.00

Calculate the average and variance of the 2 samples respectively. Explain which worker did a better job in terms of meeting the diameter specification.

16.4 Simplified calculation for variance

The sample variance calculation in Formula (3) may involve finding the square of the deviation of each sampled data with the sample average, which is cumbersome to evaluate. Here is a simplified version for variance calculation.

To simplify the discussion, we assume there are only 3 data x_1, x_2, x_3 in the sample and their average is $\bar{x}$. Then the sample variance is:

$$s^2 = \frac{1}{3}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2].$$

By expanding the brackets, we get:

$$\begin{aligned} s^2 &= \frac{1}{3}[(x_1^2-2x_1\bar{x}+\bar{x}^2)+(x_2^2-2x_2\bar{x}+\bar{x}^2)+(x_3^2-2x_3\bar{x}+\bar{x}^2)] \\ &= \frac{1}{3}[(x_1^2+x_2^2+x_3^2)-2(x_1+x_2+x_3)\bar{x}+3\bar{x}^2] \\ &= \frac{1}{3}[(x_1^2+x_2^2+x_3^2)-2\times3\times(\frac{x_1+x_2+x_3}{3})\bar{x}+3\bar{x}^2] \\ &= \frac{1}{3}[(x_1^2+x_2^2+x_3^2)-2\times3\bar{x}^2+3\bar{x}^2] \\ &= \frac{1}{3}[(x_1^2+x_2^2+x_3^2)-3\bar{x}^2] \end{aligned}$$

In general, if the sample size is *n*, the following formula can be used to calculate the sample variance:

$$\boxed{\boldsymbol{s^2 = \frac{1}{n}[(x_1^2+x_2^2+\cdots+x_n^2)-n\bar{x}^2]}} \qquad (5)$$

or simply

$$\boxed{\boldsymbol{s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}} \qquad (5)$$

In Formula (5), it is a direct computation of the square of each data, and not the square of the deviation of each sampled data with the sample average. Therefore, it involves fewer steps than in Formula (3) and is more convenient for use.

【Example 1】 Calculate the variance of the following sample (round to the nearest 1 decimal place)

$$3,\ -1, 2, 1,\ -3, 3$$

Solution By calculation, we know the sample average is not an integer. It may be inconvenient to use Formula (3) to compute the sample variance. Therefore we will use Formula (5) instead.

$$\begin{aligned} s^2 &= \frac{1}{6}\left[3^2+(-1)^2+2^2+1^2+(-3)^2+3^2-6\times\left(\frac{3-1+2+1-3+3}{6}\right)^2\right] \\ &= \frac{1}{6}\left[9+1+4+1+9+9-6\times\left(\frac{5}{6}\right)^2\right] \\ &= \frac{1}{6}\left(33-6\times\frac{25}{36}\right) \\ &= \frac{1}{6}\times33-\frac{25}{36}\approx 5.5-0.7=4.8 \end{aligned}$$

Practice

Calculate the variance of the following sample (round to the nearest 1 decimal place):

5, 4, 4, 3, 4, 3, 2, 3, 5, 3.

When the data values in the sample are relatively large, the use of Formula (5) may be further refined to alleviate the calculation effort. If the data are relatively close to each other, we can achieve some savings in calculation by subtracting a constant number *a* from each of the sampled data. The choice of the constant number *a* may be the sample average or an integer close to the sample average. For example, if the sample has 3 data x_1, x_2, x_3, by substituting

$$x_1' = x_1 - a,\ \ x_2' = x_2 - a,\ \ x_3' = x_3 - a,$$

Then

$$x_1 = x_1' + a\,,\quad x_2 = x_2' + a\,,\quad x_3 = x_3' + a\,.$$

Following from Formula (2), we have

$$\bar{x} = \bar{x}' + a\,,$$

where $\bar{x}'$ is the average of x_1', x_2', x_3'. Therefore,

$$\begin{aligned}
s^2 &= \frac{1}{3}[(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_2-\bar{x})^2] \\
&= \frac{1}{3}\{[(x_1'+a)-(\bar{x}'+a)]^2 + [(x_2'+a)-(\bar{x}'+a)]^2 + [(x_3'+a)-(\bar{x}'+a)]^2\} \\
&= \frac{1}{3}[(x_1'-\bar{x}')^2 + (x_2'-\bar{x}')^2 + (x_3'-\bar{x}')^2] \\
&= \frac{1}{3}[(x_1'^2 + x_2'^2 + x_3'^2) - 3\bar{x}'^2] \quad \text{(Following from Formula (5))}
\end{aligned}$$

In general, if the sample size is *n*, then

$$\boxed{\boldsymbol{s^2 = \frac{1}{n}[(x_1'^2 + x_2'^2 + \cdots + x_n'^2) - n\bar{x}'^2]}} \qquad (6)$$

in which

$$x_1' = x_1 - a\,,\ x_2' = x_2 - a\,,\ \ldots,\ x_n' = x_n - a\,,$$

where *a* is the constant number which can be the sample average or an integer close to the sample average.

The above formula can also be simplified as

$$\boxed{\boldsymbol{s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i'^2 - n\bar{x}'^2\right)}} \qquad (6')$$

Because we can choose the constant number *a* which is quite close to the sample average $\bar{x}$ to transform the data $x_i' = x_i - a$ ($i = 1$, 2, ..., *n*), each transformed data x_i' is relatively smaller. Therefore when the data values in the sample are large, using Formula (6) to compute the sample variance will be simpler.

【Example 2】 In a wheat farm, there are two types of wheat, Type A and Type B. We extract 10 pieces from each type and measure the height as follows (Unit: cm):

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

Which type of wheat has a relatively more uniform height?

Solution The aim of the question is asking for the comparison of the two sample variances. Because the data values in the sample are relatively large, we use Formula (6) to compute the two sample variances. Steps are as follows:

(1) Find the constant *a* for the substutition $x_i' = x_i - a$.
Because the data in the sample fall around 80, we use $a = 80$.

(2) As illustrated in Table 1 and Table 2, we calculate each of the x_i' and $x_i'^2$ ($i = 1$, 2, ..., 10) for each type and then fill in the table with the sum of x_i' and the sum of $x_i'^2$.

Table 1 (Type A wheat)

x_i	x_i' $(x_i - 80)$	$x_i'^2$
76	–4	16
90	10	100
84	4	16
86	6	36
81	1	1
87	7	49
86	6	36
82	2	4
85	5	25
83	3	9
Total	40	292

Table 2 (Type B wheat)

x_i	x_i' $(x_i - 80)$	$x_i'^2$
82	2	4
84	4	16
85	5	25
89	9	81
79	–1	1
80	0	0
91	11	121
89	9	81
79	–1	1
74	–6	36
Total	32	366

(3) Substitute the relevant data into Formula (6) for calculation:

$$\begin{aligned}
s_{\mathrm{A}}^2 &= \frac{1}{10}\left(\sum_{i=1}^{10} x_i'^2 - 10\bar{x}'^2\right) \\
&= \frac{1}{10}\left[292 - 10\times\left(\frac{40}{10}\right)^2\right] \\
&= \frac{1}{10}(292-160) \\
&= \frac{1}{10}\times 132 \\
&= 13.2
\end{aligned}$$

$$\begin{aligned}
s_{\mathrm{B}}^2 &= \frac{1}{10}\left(\sum_{i=1}^{10} x_i'^2 - 10\bar{x}'^2\right) \\
&= \frac{1}{10}\left[366 - 10\times\left(\frac{32}{10}\right)^2\right] \\
&= \frac{1}{10}(366-102.4) \\
&= \frac{1}{10}\times 263.6 \\
&= 26.36
\end{aligned}$$

Because $s_{\mathrm{A}}^2 < s_{\mathrm{B}}^2$, it can be predicted that Type A wheat has a relatively more uniform length than Type B wheat.

【Example 3】 A farm is trying to test 2 species of rice in 8 testing locations for a comparative testing on their plantation. The production quantities of the 2 species in each of the 8 testing locations are as follows (Unit: kg):

Type A:	804	984	989	817	919	840	912	1001
Type B:	856	932	930	855	872	910	897	918

Which type of rice has a relatively more consistent production quantity?

Solution The aim of the question is asking for the comparison of the two sample variances. Because the data values in the sample are relatively large, we follow the steps in Example 2 to make use of Formula (6) to compute the two sample variances.

(1) Because the data in the samples fall around 900, we use $a = 900$ for the data transformation $x_i' = x_i - a$.

(2) As illustrated in Table 3 and Table 4, we calculate each of the x_i' and $x_i'^2$ and then fill in the table with the sum of x_i' and the sum of $x_i'^2$.

Table 3 (Type A rice)

x_i	x_i' $(x_i - 900)$	$x_i'^2$
804	–96	9216
984	84	7056
989	89	7921
817	–83	6889
919	19	361
840	–60	3600
912	12	144
1001	101	10201
Total	66	45388

Table 4 (Type B rice)

x_i	x_i' $(x_i - 900)$	$x_i'^2$
856	–44	1936
921	32	1024
930	30	900
855	–45	2025
872	–28	784
910	10	100
897	–3	9
918	18	324
Total	–30	7102

(3) Substitute the relevant data into Formula (6) for calculation:

$$\begin{aligned}
s_{\mathrm{A}}^2 &= \frac{1}{8}\left(\sum_{i=1}^{8} x_i'^2 - 8\bar{x}'^2\right) \\
&= \frac{1}{8}\left[45388 - 8\times\left(\frac{66}{8}\right)^2\right] \\
&= \frac{1}{8}(45388-544.5) \\
&= \frac{1}{8}\times 44843.5 \approx 5605
\end{aligned}$$

$$
\begin{aligned}
s_{\mathrm{B}}^{2} &= \frac{1}{8}\left(\sum_{i=1}^{8} x_i'^{2} - 8\overline{x}'^{2}\right) \\
&= \frac{1}{8}\left[7102 - 8\times\left(\frac{-30}{8}\right)^{2}\right] \\
&= \frac{1}{8}(7102 - 112.5) \\
&= \frac{1}{8}\times 6989.5 \approx 874
\end{aligned}
$$

Because $s_{\mathrm{A}}^{2} < s_{\mathrm{B}}^{2}$, it can be inferred that Type B rice has a relatively more consistent production quantity than Type B rice.

We can see that the computation of sample average and variance involve quite a number of steps. For the convenience of computation, it may help to choose an "easier to calculate" sample size, such as 10, 20, 30, etc.

Practice

Calculate the sample variance of the following (round to the nearest 1 decimal place):

(1) 105, 103, 101, 100, 114, 108, 110, 106, 98, 102;

(2) 423, 421, 419, 420, 421, 417, 422, 419, 423, 418.

Exercise 13

1. A and B each fires 10 shots under the same condition. Their scores are recorded as follows:

A:	7	8	6	8	6	5	9	10	7	4
B:	9	5	7	8	7	6	8	6	7	7

Compute the 2 sample variances. According to the computation result, does A or B have a more consistent performance?

2. Machine A and Machine B are used to produce the same machine part. In 10 days, the number of defected items produced by the machines are recorded as follows:

Machine A:	0	1	0	2	2	0	3	1	2	4
Machine B:	2	3	1	1	0	2	1	1	0	1

Compute the sample averages and sample variances. According to the computation result, which machine has a better performance?

3. A farm grows two types of rice, namely Type A and Type B. The average production quantity in the past consecutive 6 years are recorded as follows (Unit: kg):

Type	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
A:	900	920	900	850	910	920
B:	890	960	950	850	860	890

Which type of rice has a more consistent production quantity?

4. The followings are the body lengths of 10 pigs being fed under similar conditions this year and last year (Unit: cm):

Last year:	112	110	110	117	113	122	125	124	119	127
This year:	111	122	115	123	114	115	118	114	116	115

Does the growth of pigs this year or last year produce a relatively more consistent body length?

16.5 Frequency distribution

To understand the body growth situation of high school female students, a measurement of body height was conducted on 60 female students of the same age in a high school. Results are recorded as follows (Unit: cm):

167	154	159	166	169	159	156	166	162	158
159	156	166	160	164	160	157	156	157	161
158	158	153	158	164	158	163	158	153	157
162	162	159	154	165	166	157	151	146	151
158	160	165	158	163	163	162	161	154	165
162	162	159	157	159	149	164	168	159	153

As we understand, the average of this set of data reflects the average height of these students. However, sometimes it may not be sufficient to know only this piece of information. We may like to classify the data into several height-ranges (or classes), and count the number of students in each height-range (or class). Knowing the number of students in each height-range, we can calculate the proportion of students in each height-range. To achieve this, we shall need a procedure to classify the data into height-ranges. The procedure is as follows:

(1) Calculate the range of variation between the largest and smallest data value.

From the above data, the largest data value is 169 and the smallest data value is 146. Their difference is:

$$169-146=23(\text{cm}).$$

Then, we know the range of variation of this set of data is at most 23 (cm).

(2) Determine the number of classes and the class interval.

To divide a set of data into classes, if the data volume is large, we shall consider dividing it into a greater number of classes. When the number of data is within 100, the division into 5 to 12 is usually appropriate.

The class interval means the distance between two end-points of a class. The choice of class interval is inter-related to the the total number of classes.

In this example, if we choose the class interval as 3 cm, we can calculate the number of classes for the batch of data,

$$\frac{\text{Maximum value}-\text{Minimum value}}{\text{Class interval}}=\frac{23}{3}=7\frac{2}{3}$$

This would mean dividing the data into 8 classes.

If the class interval is chosen as 2 cm, then $\frac{23}{2}=11\frac{1}{2}$. This would mean dividing the data into 12 classes, which appears too many.

So we decide to divide the data into 8 classes with a calss interval of 3 cm.

(3) Determine the class interval end-points

Using a class interval of 3 cm, we can pick the lowest data value of 146 cm as the starting point of the first class interval. Accordingly, the 8 class intervals are as follows:

146～149, 149～152, 152～155, 155～158,
158～161, 161～164, 164～167, 167～170.

Now, we are faced with a problem. Some of the data values (e.g. 149, 158, 167) fall at the end points. It is difficult to decide whether to place the data in the lower class or in the higher class. To avoid this situation, we can add 1 decimal point into the division point and to reduce the starting point of the first group a little bit. For example, we can set the starting point of the first group to 145.5 instead of 146. Then the 8 classes will be:

145.5～148.5, 148.5～151.5, 151.5～154.5, 154.5～157.5,
157.5～160.5, 160.5～163.5, 163.5～166.5, 166.5～169.5.

(4) Set up the table of frequency distribution

As set out in column 1 and column 2 of Table 5, we can use a method similar to the counting the votes in polling. For every data item (vote), we put a count in the appropriate class (the total count in each class is called the class frequency). After every data item is counted, we sum up the values to column 3 in Table 5.

Table 5 Frequency distribution table

Group	Cumulative frequency	Frequency	Relative frequency
145.5～148.5	/	1	0.017
148.5～151.5	///	3	0.050
151.5～154.5	~~////~~ /	6	0.100
154.5～157.5	~~////~~ ///	8	0.133
157.5～160.5	~~////~~ ~~////~~ ~~////~~ ///	18	0.300
160.5～163.5	~~////~~ ~~////~~ /	11	0.183
163.5～166.5	~~////~~ ~~////~~	10	0.167
166.5～169.5	///	3	0.050
Total		60	1.000

For each class, the ratio of frequency divided by the sample size is called the **relative frequency** of this class. For example, the relative frequency of the first class is:

$$\frac{1}{60} \approx 0.017.$$

Calculate the relative frequency of each class and fill in column 4 in Table 5. Table 5 is called the **frequency distribution table**. With the frequency distribution table, we will know the proportion of data in each class.

(5) Draw the frequency distribution histogram.

To illustrate the frequency distribution in a vertical graphical way, we usually draw the **frequency distribution histogram**. The frequency distribution histogram for this example is illustrated in diagram 16-2, in which the horizontal axis represents the height while the vertical axis represents the ratio between relative frequency and class interval. It can be easily observed that:

$$\begin{aligned}\text{Area of small rectangle} &= \text{Class interval} \times \frac{\text{Relative frequency}}{\text{Class interval}} \\ &= \text{Relative frequency}\end{aligned}$$

That is to say, area of each small rectangle is equal to the relative frequency of the corresponding class. Consequently, frequency distribution histogram makes use of the area of a graph to reflect the magnitude of the relative frequency in each class. In diagram 16-2, we can also see that,

$$\begin{aligned}\text{Height of small rectangle} &= \frac{\text{Relative frequency}}{\text{Class interval}} \\ &= \frac{1}{\text{Class interval} \times \text{Sample size}} \times \text{Frequency}\end{aligned}$$

Because the product of the class interval and sample size is a constant, $\frac{1}{\text{Class interval} \times \text{Sample size}}$ is also a constant. Therefore, the height of small rectangle and frequency are in proportion. By making use of this relationship, we can determine the relative heights of each small rectangle. While the frequency is k , the height of the corresponding small rectangle will be kh. For example, for the group 148.5~151.5, its frequency is 3, the height of the corresponding small rectangle is $3h$.

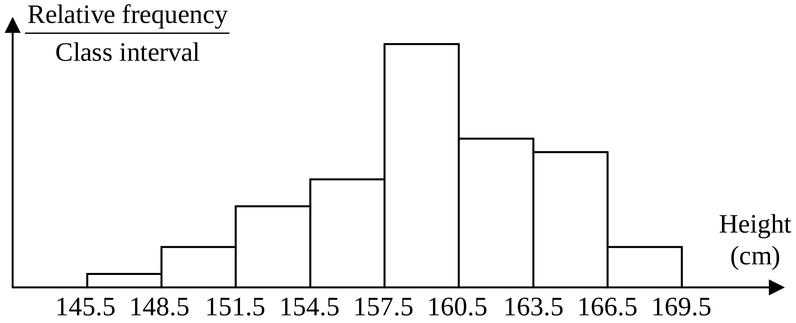


Diagram 16-2

In the frequency distribution histogram, considering that the area of each small rectangle is the relative frequency of the corresponding class and that the sum of relative frequency is 1, we can infer that the sum of the area of all the small rectangles will also be 1.

After understanding about the relative frequency distribution of a sample, we can predict the distribution of the corresponding population. In the above example, the relative frequency of the sample data falling in the group 157.5~160.5cm is 0.3. This explains that for every 100 female students of that age, there will be about 30 students with height between 157.5~160.5cm.

【Example】 A farm testing laboratory would like to understand the distribution of the length of wheat. 100 wheats were picked from the test farm and their lengths were recorded as follows (Unit: cm):

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	6.5	5.3	5.9	5.5	5.8
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	6.3	5.2	6.0	7.0	6.4
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	5.8	6.3	6.0	6.3	5.6
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	6.8	6.6	4.7	5.7	5.7
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	4.35	5.6	6.3	6.0	5.8	6.3

Set out the relative frequency distribution table. Draw the relative frequency histogram.

Solution (1) Calculate the range of variation between the largest value and the smallest value.

In the sample data, the largest value is 7.4. The smallest value is 4.0. The difference is

$$7.4 - 4.0 = 3.4\,(\text{cm}).$$

(2) Determine the class interval and number of classes.

In this example, the range of variation between the largest value and the smallest value is 3.4 cm. If the class interval is set at 0.3 cm, then because $\frac{3.4}{0.3} = 11\frac{1}{3}$, there will be 12 groups.

(3) Determine the class interval end- points.

The division point will be one decimal place more than the data and the starting point of the first group will be adjusted slightly downward from the smallest data value. Then the 12 classes are:

3.95～4.25, 4.25～4.55, 4.55～4.85, 4.85～5.15,
5.15～5.45, 5.45～5.75, 5.75～6.05, 6.05～6.35,
6.35～6.65, 6.65～6.95, 6.95～7.25, 7.25～7.55.

(4) Relative frequency distribution table

Chalk up each data value according to the class it belongs. then count the frequency in each class, calculate the relative frequency and set out the relative frequency distribution table as in Table 6.

Table 6 Relative frequency distribution table

Groupo	Cumulative frequency	Frequency	Relative frequency	Cumulative relative frequency (this will be introduced in the next section)
3.95～4.25	/	1	0.01	0.01
4.25～4.55	/	1	0.01	0.02
4.55～4.85	//	2	0.02	0.04
4.85～5.15	卌	5	0.05	0.09
5.15～5.45	卌 卌 /	11	0.11	0.20
5.45～5.75	卌 卌 卌	15	0.15	0.35
5.75～6.05	卌 卌 卌 卌 卌 ///	28	0.28	0.63
6.05～6.35	卌 卌 ///	13	0.13	0.76
6.35～6.65	卌 卌 /	11	0.11	0.87
6.65～6.95	卌 卌	10	0.10	0.97
6.95～7.25	//	2	0.02	0.99
7.25～7.55	/	1	0.01	1.00
Total		100	1.000	

(5) Draw the relative frequency distribution histogram (The histogram is shown on the diagram 16-3 (a). Steps will be similar to those used for drawing in the previous example)

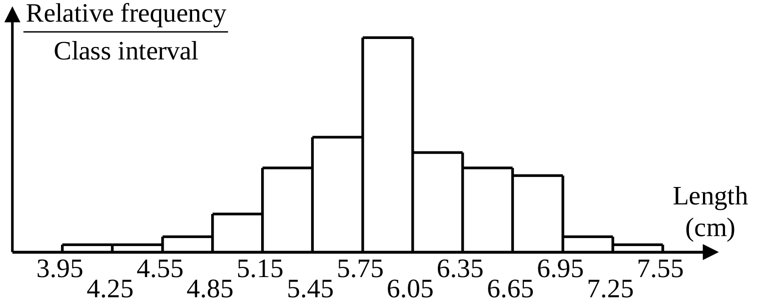


Diagram 16-3 (a)

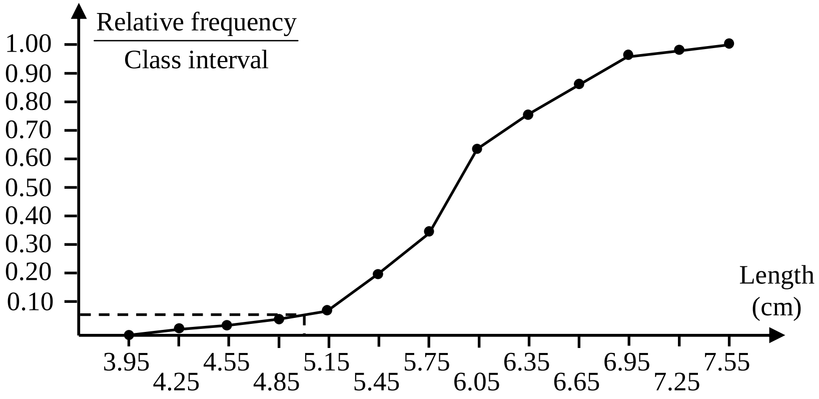


Diagram 16-3 (b)

Now we can base on the frequency distribution of the sample to predict the distribution of the population. For example, it can be observed from Table 6 that the relative frequency of sample data falling in the range of 5.75~6.05 is 0.28. Therefore, we can predict that, 28% of the wheat in this farm will be having the length within the range of 5.75~6.05cm. In using the relative frequency distribution to predict the distribution of the population, we know that the prediction will be more accurate if the sample size is larger.

Practice

1. Here is the data set of a sample. The sample size is 50. The class end-points and the relevant frequency of each class are as follows:

53.5～55.5	4	61.5～63.5	10
55.5～57.5	7	63.5～65.5	6
57.5～59.5	9	65.5～67.5	3
59.5～61.5	11		

Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the sample.

Practice

2. Given the following data set of a sample:

25 21 23 25 27 29 25 28 30 29
26 24 25 27 26 22 24 25 26 28

Fill in the following relative frequency distribution table:

Group	Cumulative frequency	Frequency	Relative frequency
20.5～22.5			
22.5～24.5			
24.5～26.5			
26.5～28.5			
28.5～30.5			
Total			

3. To understand the body growth situation of high school students, the heights of 50 male students from a high school were measured and recorded as follows (Unit: cm)

175 168 170 176 167 181 162 173 171 177
179 172 165 157 172 173 166 177 169 181
160 163 166 177 175 174 173 174 171 171
158 170 165 175 165 174 169 163 166 166
174 172 166 172 167 172 175 161 173 167

Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the sample.

*16.6 Cumulative relative frequency distribution[4]

In the example of wheat length in the previous section, the relative frequency for wheat length being less than 4.55 cm is equal to the sum of the relative frequency of the first 2 classes. i.e.,

$$0.01 + 0.01 = 0.02$$

[4] This section is optional.

The relative frequency for wheat length being less than 4.85 cm is the sum of the relative frequency of the first 3 classes, i.e.,

$$0.01+0.01+0.02=0.04.$$

This can be applied similarly to different classes. This concept with the relative frequency being less than a value is called the ***cumulative relative frequency*** of the value. The cumulative relative frequency of each division point in this example is illustrated in the last column of Table 6. This can be used as supplementary information of the relative frequency distribution table.

According to the cumulative relative frequency, we can draw the cumulative frequency distribution graph, as shown in the bottom graph in Figure 16-3. In the graph, the horizontal axis represents the wheat length. The vertical axis represents the cumulative relative frequency. According to each cumulative relative frequency in the table, the graph describes the relevant points. For example, the cumulative relative frequency of point 4.25 is 0.01. In the graph, it is represented by a point with 4.25 on x-axis and 0.01 on y-axis. Then, using line segments to connect each point in sequence, it will form the cumulative relative frequency distribution graph.

With the use of cumulative relative frequency distribution graph for the sample, we can predict the corresponding situation for the population. For example, to predict the propotion of wheats with length less than 5 cm in the piece of farm land, we can locate the point with 5 (cm) on the x-axis and 0.07 on the y-axis from the cumulative relative distribution graph. This tells us that, about 7% of wheat in the farm land have length less than 5 cm.

Practice

According to Table 5, calculate the cumulative frequency of each split point, and draw the cumulative relative frequency distribution graph.

Exercise 14

1. In a batch of wood sticks, the length of 60 pieces of wood sticks were measured and recorded as follows (Unit: mm):

82	202	352	321	25	293	293	86	28	206
323	355	357	33	325	113	233	294	50	296
115	236	357	326	52	301	140	328	238	358
58	255	143	360	340	302	370	343	260	303
59	146	60	263	170	305	380	346	61	305
175	348	264	383	62	306	195	350	265	385

Set out the relative frequency distribution table and draw the relative frequency histogram for the sample.

2. In a batch of machine parts, the deviation between their size and the specified size of 100 pieces were measured and recorded as follows (Unit: 0.1mm):

2	1	0	3	–1	2	1	0	–1	0
3	–2	–1	–2	1	0	0	–4	1	0
1	2	2	0	–2	1	0	–1	0	3
–1	1	0	–1	0	3	1	–2	3	–1
–3	0	0	1	4	0	–2	2	1	2
2	0	–2	0	0	–1	1	4	–2	1
1	1	4	–1	1	–1	0	2	–2	1
0	1	–1	1	0	2	2	1	0	–1
0	–1	3	1	2	–3	1	0	1	1
–2	2	1	2	–1	–2	3	1	–1	0

(1) Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the samples. (Hint: because there are only 9 diffferent values in the samples, and they are consecutive integers, at maximum the data can only be divided into 9 groups.)

(2) Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the sample. (Hint: because there are only 9 diffferent values in the samples, and they are consecutive integers, at maximum the data can only be divided into 9 groups.)

3. A farm is trying to estimate the production quantity of corns before harvest in the autumn. The weight of 100 pieces of corns were extracted and recorded as follows (Unit: g):

50	29	31	32	29	30	38	41	26	20
33	27	34	20	27	17	17	28	35	39
38	35	26	50	38	36	38	25	28	22
32	30	29	48	30	35	18	20	25	24
29	50	23	44	39	27	33	39	24	26
34	23	32	18	39	25	50	33	40	21
35	31	33	30	28	25	27	30	39	24
35	30	38	46	40	18	35	43	24	23
33	34	36	34	40	23	34	37	40	19
39	34	33	37	35	25	34	30	27	15

(1) Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the samples.

*(2) Add in the cumulative relative frequency for each division point to the relative frequency distribution table, and draw the cumulative relative frequency distribution graph.

Chapter Summary

I. This chapter introduces some elementary knowledge and concept of statistics. The essencial idea of statistical method is to extract and analyse a sample selected from the population and use the result of the selected sample to predict the behaviour of the population. The larger the sample size, the more accurate the prediction will be.

II. The population average is an attribute of a population measured by the average value (or average level) of all the data in the population. It is usually estimated by the sample average which is the sum of the attribute of all data in the sample divided by the number of elements in the sample. When the data values of elements in the sample are large, it may be easier to use the transformation $\bar{x} = \bar{x}' + a$ to simplify the calculation.

III. The population variance is another attribute of the population measured by the magnitude of variation in the population. It is usually estimated by the sample variance, or by comparing the sample variance of 2 samples. For a sample with size n and average $\bar{x}$, the formula for the sample variance is:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

The calculation of sample variance is relatively more complicated than the sample average. Without electronic calculators, it is recommended to design a table with columns for filling in the data value, the data deviation from the sample average and square of the data deviation while carrying out the computation.

IV. Population distribution reflects the significance of each element in the population. Sample relative frequency distribution is usually used to predict the population distribution. Relative frequency distribution reflects the significance of the sample data in each of the relevant classes. The following steps are used to obtain the relative frequency distribution of a sample:

(i) compute the range of variation between the largest and smallest value,

(ii) determine the number of classes and class interval,

(iii) determine the division end-points of the class interval,

(iv) set out the relative frequency distribution table,

(v) draw the relative frequency distribution histogram.

Revision Exercise 16

1. In statistics, what is a population? What is a sample of the population? Why do we usually need to use certain attributes of sample to predict the corresponding attributes of the population?

2. Cannons A and B each fires 50 shells at the same target under the same condition. The distances by which the target was missed were measured and recorded as follows:

Distance missed	40 m	30 m	20 m	10 m	0 m
No. of shells fired by Cannon A	0	1	3	7	39
No. of shells fired by Cannon B	1	3	2	3	41

Calculate the sample average of the two cannons respectively. According to the result, which cannon achieved a better accuracy?

3. A fish pond has 100,000 fish of a specific species. 20 of the fish were weighed and recorded as follows (Unit: kg)

2.3 2.1 2.2 2.1 2.2 2.6 2.5 2.4 2.3 2.4
2.4 2.3 2.2 2.5 2.4 2.6 2.3 2.5 2.2 2.3

Calculate the sample average and according to the calculation result, predict the weight (in kg) of all the fish in the pond.

4. To examine the growth of wheat A and wheat B, 10 pieces of wheat from each type were extracted and their heights were measured and recorded as follows (Unit: cm):

Wheat A	12	13	14	15	10	16	13	11	15	11
Wheat B	11	16	17	14	13	19	6	8	10	16

(1) Calculate the average height of each type of wheat respectively.

(2) Which type of wheat has a relatively more uniform height?

5. The results of 10 games participated by two long jump athletes were recorded as follows (Unit: m):

A: 5.85 5.93 6.07 5.91 5.99 6.13 5.98 6.05 6.00 6.19
B: 6.11 6.08 5.83 5.92 5.84 5.81 6.18 6.17 5.85 6.21

Calculate the sample variance respectively and according the the calculation result, predict which athlete has a more consistent result.

6. The size (Unit: m) of 80 of a kind of machine part were measured and recorded as follows:

362.5×1	362.6×2	362.7×2	362.8×3
362.9×3	363.0×3	363.1×5	363.2×6
363.3×8	363.4×9	363.5×9	363.6×7
363.7×6	363.8×4	363.9×3	364.0×3
364.1×2	364.2×2	364.3×1	364.4×1

(1) Set out the relative frequency distribution table and draw the relative frequency distribution histogram for the sample.

*(2) Add in the cumulative relative frequency for each split point to the relative frequency distribution table, and draw the cumulative relative frequency distribution graph.

7. (1) Set $\overline{x}$ as the average of x_1, x_2, ..., x_n and $\overline{y}$ as the average of $3x_1+5$, $3x_2+5$, ..., $3x_n+5$. Prove that $\overline{y}=3\overline{x}+5$.

(2) Let s_x be the standard deviation of x_1, x_2, ..., x_n and s_y be the standard deviation of $3x_1+5$, $3x_2+5$, ..., $3x_n+5$. Prove that $s_y=3s_x$.

Annex Revision questions for Junior-high school algebra[5]

1. List the table for classifying real numbers.
2. Calculate:
   (1) $4\times 0.2^2-10\times 1.1^3+3\times\sqrt{5.14}-\sqrt[3]{5.26}$ (round to 3 significant digits);
   (2) $3a^3-[-5a^2+a(-4a^2+2s-1)+4]-7$ (round to the nearest 0.01), where $a=2.35$.
3. (1) If $(a-1)^2+(b+2)^2=0$, *a*, *b* are real numbers, find *a*, *b*;
   (2) If $x^2+2x+y^2-6y+10=0$, where *x* and *y* are real numbers, find *x* and *y*.
4. Calculate each of the following:
   (1) $(a^3-a^2b+ab^2-b^3)(a+b)$;
   (2) $(a+b)^2(a-b)^2-(a^2+b^2)(a^2-b^2)$;
   (3) $(p+q-m-n)(p-q-m+n)$;
   (4) $(x+2y-z)(x-2y+z)-(x+2y+z)^2$;
   (5) $(x^5+x^4+x^3+x^2+x+1)\div(x+1)$;
   (6) $(x^4+x^2+1)\div(x^2-x+1)$.
5. (1) Given $x^2+x-1=0$, solve x^3+2x^2+3;
   (2) If $bc=ad$, prove that $ab(c^2-d^2)=(a^2-b^2)cd$;
   (3) If $a+b+c=0$, prove that $a^3+a^2c+b^2c-abc+b^3=0$;
   (4) Prove that:
   $$a^2(b-c)+b^2(c-a)+c^2(a-b)=-(a-b)(b-c)(c-a).$$
6. Factorise:
   (1) $x^5-x^4+x^3-x^2+x-1$;
   (2) $3x^2-2x-8$;

[5] For use as reference exercise for the revision of junior-high school algebra.

   (3) $4x^2+2x+\dfrac{1}{4}-y^2$; (4) l^4-3l^2+1;
   (5) $a^2-2ab+b^2-6a+6b+5$; (6) $(1-a^2)(1-b^2)-4ab$.
7. Factorise within the set of real numbers:
   (1) x^2+x-1; (2) x^4+x^2-6;
   (3) $6x^4-7x^2-3$; (4) $x^4+3x^3+x^2$.
8. Find value(s) of *x* so that the algebraic equation $\dfrac{-(x-5)}{(3-x)(x+1)}$ satisfies the following condition:
   (1) Equals to 0; (2) Undefined; (3) Equals to 1.
9. Calculate each of the following:
   (1) $\dfrac{x^2+2x+1}{x^3-x}\times\dfrac{x}{x+1}-\dfrac{1}{x+1}$;
   (2) $\dfrac{1}{1-x}+\dfrac{1}{1+x}+\dfrac{2}{1+x^2}+\dfrac{4}{1+x^4}$;
   (3) $\dfrac{x+2}{x+1}-\dfrac{x+3}{x+2}+\dfrac{x+4}{x+3}-\dfrac{x+5}{x+4}$;
   (4) $\dfrac{a^2}{(a-b)(a-c)}+\dfrac{b^2}{(b-a)(b-c)}+\dfrac{c^2}{(c-a)(c-b)}$
10. Prove the following:
   (1) If $\dfrac{a}{b}+\dfrac{b}{a}=x$, $\dfrac{a}{b}-\dfrac{b}{a}=y$, prove that $x^2-y^2=4$;
   (2) If *a*, *b*, *c* are real numbers and $a^2+b^2+c^2-ab-bc-ca=0$, prove that $a=b=c$.
11. Calculate:
   (1) $5\sqrt{28}+3\sqrt{63}-10\sqrt{7}+3\sqrt{\dfrac{1}{7}}$;
   (2) $4\sqrt[6]{125}-3\sqrt[4]{25}+\sqrt{\dfrac{4}{5}}$;
   (3) $(2\sqrt{3}+\sqrt{6})(\sqrt{3}-2\sqrt{6})$;

(4) $\sqrt{(\sqrt{2}-\sqrt{3})^2}+\sqrt{(\sqrt{3}-\sqrt{2})^2}$;

(5) $\sqrt{xy}\left(\sqrt{xy}-3\sqrt{\frac{y}{x}}-2\sqrt{\frac{x}{y}}+4\sqrt{\frac{1}{xy}}\right)$ $(x>0,\quad y>0)$;

(6) $\frac{1}{1-\sqrt[4]{x}}+\frac{1}{1+\sqrt[4]{x}}+\frac{2}{1+\sqrt{x}}+\frac{4}{1+x}$.

12. Simplify $\sqrt{\frac{x^2-6x+9}{x^2+6x+9}}$

13. Solve the following equations:

(1) $\frac{2-x}{6}-\frac{2x-3}{4}=1$; (2) $\frac{3x+6}{8}-\left(\frac{5x}{6}-1\right)=\frac{5}{6}$;

(3) $\frac{1}{3}\left(3x-\frac{10-7x}{2}\right)-\frac{1}{6}\left(2x-\frac{2x+2}{3}\right)=\frac{x}{2}-1$;

(4) $\frac{1}{3}(x-2)-\frac{1}{7}(5x-6)=\frac{22x-63}{105}-\frac{1}{5}(3x-4)$;

(5) $\frac{ax}{b}+\frac{bx}{a}-1=0$ (*a*, *b* are known) ;

(6) $\frac{mx+1}{n}+\frac{nx+1}{m}=1$ (*m*, *n* are known).

14. Solve the following simultaneous equations:

(1) $\begin{cases}\frac{1}{2}(x+11)=\frac{1}{3}(y+13)+2\\ 5x=3y+8\end{cases}$;

(2) $\begin{cases}\frac{2x}{3}+\frac{3y}{4}=\frac{3x-2y}{2}+1\\ \frac{3x}{2}-\frac{4y}{3}=\frac{3x+4y}{6}-1\end{cases}$; (3) $\begin{cases}\frac{3}{x-4}+\frac{4}{y-1}=3\\ \frac{9}{x-4}-\frac{2}{y-1}=2\end{cases}$;

(4) $\frac{2x+y+6}{4}=\frac{4x-3y-7}{6}=\frac{-6x-7y+10}{8}$;

(5) $\begin{cases}2x+3y-4z=4\\ 2y+3z=\frac{17}{12}\\ x+4y=\frac{10}{3}\end{cases}$; (6) $\begin{cases}x+y-z=3\\ z+x-y=1\\ y+z-x=7\end{cases}$

15. Solve the following equation:

(1) $4(x+3)^2-9(x-2)^2=0$;

(2) $\sqrt{3}(y^2-y)=\sqrt{2}(y^2-y)$;

(3) $x^2-(2-2\sqrt{2})x+3-2\sqrt{2}=0$;

(4) $(t+6)(t-6)=2(t-3)$;

(5) $(x+2)^2+(x-1)^2=x^2+6$;

(6) $2(x+5)(x-5)=(x-6)^2$.

16. Given one of the roots of the quadratic equation $x^2-mx+2m=0$ is 1.

(1) Find the value of *m*;

(2) Find the other root.

17. Without solving the equation $2x^2-7x+2=0$, find the sum of the reciprocol of the two roots.

18. (1) When solving fractional equation and radical equation, why is it necessary to verify the root?

(2) Solve the following equations:

(i) $\frac{x-1}{x^2-2x}+\frac{x-2}{x^2-x}-\frac{x}{x^2-3x+2}=0$;

(ii) $\left(\frac{x^2-1}{x}\right)^2+\frac{7}{6}\left(\frac{x^2-1}{x}\right)-4=0$;

(iii) $x^2+3x-\frac{20}{x^2+3x}=8$;

(iv) $\sqrt{5x+4}-\sqrt{2x-1}-\sqrt{3x+1}=0$;

(v) $x^2+3x-2\sqrt{x^2+3x-1}-4=0$;

(vi) $x^4-25x^2+144=0$;

(vii) $(x^2+5x)^2-2x^2-10x-24=0$;

(viii) $(x^2+5x-12)(x^2+5x+2)=32$.

19. Solve the following simultaneous equations:

(1) $\begin{cases}(x+y)^2-3(x+y)=54\\(x-y)^2-5(x-y)=14\end{cases}$; (2) $\begin{cases}a+aq^2=15\\aq+aq^3=30\end{cases}$;

(3) $\begin{cases}\dfrac{3x-5y}{2xy}=\dfrac{4}{3}\\\dfrac{3}{x}+\dfrac{1}{y}=4\end{cases}$; (4) $\begin{cases}\sqrt{\dfrac{x}{y}}+\sqrt{\dfrac{x}{y}}=\dfrac{5}{2}\\x+y=5\end{cases}$;

(5) $\begin{cases}x+xy+y=11\\x^2y+xy^2=30\end{cases}$

20. (1) From $\begin{cases}mu=s(m-1)\\f=s-u\end{cases}$ (where $m\neq 0$), express the relationship of *f*, *s*, *m* in a formula.

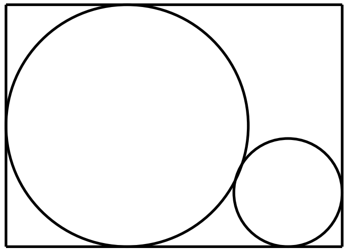
(2) From $\begin{cases}I_k=I_a+I_c\\I_c=mI_k\end{cases}$ (where $m\neq 1$), express the relationship of I_c, *m*, I_a in a formula.

21. Pond A contains 380 m^3 of water. Pond B contains 1500 m^3 of water. Water is pumped into Pond A at the speed 80m^3 per hour while water is pumped into Pond B at the speed 60 m^3 per hour. At what time will both ponds contain the same amount of water?

22. A and B are riding bicycles on a circular runway of 400 m long. The speed of A is faster than B. They started at the same time at the same point. If they ride in opposite direction, they will meet each other every 40 seconds. If they ride in the same direction, they will meet each other every 6 minutes 40 seconds. Find the speeds of A and B.

23. It takes 16 days for A and B to work together to complete a task. If they work together for 4 days and the remaining task to be completed by B alone, it takes 12 days more than A to complete the task alone. Find the time for A and B to complete the whole task working alone individually.

24. In the diagram, there is a metal plate which is 25 cm long and 18 cm wide. A circle touching the 3 sides of the metal plate is now cut from the metal plate. What is the diameter of the largest possible circle which can be cut from the remaining metal plate?



(No. 24)

25. A factory can manufacture 500 pieces of Type A machine part, 600 pieces of Type B machine part, or 750 pieces of Type C machine part in a day. A certain product requires one Type A part, one Type B part and one Type C part for assembly. What is the maximum possible number of products the factory can produce in 30 days? For how many days should the factory manufacture each of the Type A, Type B and Type C machine parts?

26. A pesticide of 40 kg with 15% ingredient is mixed with another pesticide of 50 kg with higher concentration of the ingredient so that the mixture contains 25%~30% (excluding the 25% and 30% marks) ingredient. Find the percentage of ingredient in the pesticide with the higher concentration.

27. Solve the following inequality and to indicate the answer on a numerical axis.

(1) $2(x+1)+\dfrac{x-2}{3}>\dfrac{7x}{2}-1$;

(2) $2x-\dfrac{x-1}{2}<\dfrac{2x-1}{3}+\dfrac{x+1}{6}$;

(3) $\frac{1}{2}\{1-2[x-4(x-1)]\}\le 4x$; (4) $-1\le\frac{3-x}{2}\le 2$;

(5) $-4\le\frac{3x-5}{2}\le -2$; (6) $\begin{cases}3x-5\ge 2x+3\\ \frac{x-5}{2}\le 3x+1\end{cases}$.

28. Indicate on a numerical axis the solution sets which satisfy the following conditions:

(1) $x\le 5$ or $x>-2$; (2) $x\le 5$ and $x>-2$;
(3) $x>5$ or $x>-2$; (4) $x>5$ and $x>-2$;
(5) $x>5$ or $x\le -2$; (6) $x>5$ and $x\le -2$.

29. Calculate:

(1) $(-3x^{\frac{1}{3}}y^{\frac{1}{4}})(2x^{-\frac{2}{3}}y^{-\frac{1}{2}})(-x^{-\frac{2}{3}}y^{\frac{1}{4}})$;

(2) $\left(\frac{m^4n^{-4}}{m^{-1}n}\right)^{-3}\div\left(\frac{m^{-2}n^2}{mn^{-1}}\right)^5$;

(3) $(3x^{\frac{1}{2}}-2y^{-\frac{1}{2}})(3x^{\frac{1}{2}}+2y^{-\frac{1}{2}})$;

(4) $\sqrt[3]{3}\times\sqrt[4]{4}\times\sqrt[6]{6}$;

(5) $\sqrt[3]{x^{-5}y^2\sqrt{x^3y}}$;

(6) $(\sqrt{2x}-3\sqrt[4]{y})(\sqrt{2x}+3\sqrt[4]{y})$.

30. (1) Explain why logarithm can be applied to convert a multiplication operation into a summation opertion, and why logarithm can be applied to convert a division operation into a subtraction operation.

(2) Write the following numbers using scientific notation, and explain the relationship with the characteristics of the common logarithm value.

54890, 0.08351, 4.903, 213.7

31. Given $\log 2=0.3010$, how many digits does $2^7\times 8^{11}\times 5^{11}$ have?

32. Calculate using logarithm:

(1) $\sqrt[8]{\frac{0.3586\times\sqrt[4]{257}}{0.00501}}$; (2) $\frac{32.85^2-12.62^2}{32.85\times 12.64}\div\sqrt{\frac{32.85}{12.64}}$

33. Find the domain of the following independent variables:

(1) $y=\frac{x-5}{x^2-3x+2}$; (2) $y=\sqrt{8-2x-x^2}$

34. Given $y_1=2x-3$, $y_2=-3x+7$, find value(s) of x to fulfill the following requirement:

(1) $y_1>y_2$; (2) $y_1=y_2$; (3) $y_1<y_2$.

Use graph to explain the answer.

35. Solve k when the graph of the function $y=-x^2+2x+k$ intersects the x-axis:

(1) at one point; (2) at two points; (3) never.

Use graph to explain the answer.

36. Given that the minimum value of the function $y=x^2+px+q$ is 4.and that when $x=2$ and $y=5$, find the values of p and q.

37. Assume α is an angle of $0°\sim 180°$. Solve all possible values (if any) of α according to the following condition.

(1) $\tan\alpha=-1$ (2) $\cos\alpha=0$ (3) $\sin\alpha=\frac{1}{2}$

(4) $\cot\alpha=\sqrt{3}$ (5) $\sin\alpha=-\frac{1}{2}$ (6) $\cos\alpha=\sqrt{2}$

38. In a right-angled triangle ABC, AB is the hypotenuse and CD is the altitude of the triangle. CD = 21 cm, AD = 18cm. Solve the triangle. (lengths to be rounded to 2 significant digits, angles to be rounded to the nearest $1°$)

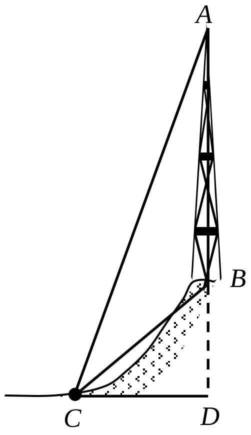
39. In a circle with radius 10.0 cm, a regular pentagon ABCDE is constructed with all vertices touching the circle. Solve the distance between the A (highest point of the pentagon) and B, and the distance between A and C (round to the nearest 0.1 cm).

40. In triangle *ABC*, $B=30°$, $c=150$, $b=50\sqrt{3}$. Find all the sides and angles of the triangle., and comment on the characteristics of the triangle *ABC*?

41. Prove that in $\triangle ABC$, if $\sin^2 A+\sin^2 B=\sin^2 C$, then this triangle is a right-angled triangle.

42. In a paraaellegram *ABCD*, *AB* = 8, *AD* = 5, $\angle A=60°$. The vertex *A* is located at the origin, *AB* is the *x*-axis, *C* lies in the first quadrant, Solve the coordinates of the vertices *A*, *B*, *C* and *D*.

43. As shown in the diagram, there is a television boardcasting tower on a small mountain. The height of the tower *AB* is 50 m. On the ground at point *C*, the angle of elevation measured for B is $40°$ and for *A* is $70°$. Solve the height of the mountain *BD* (round to the nearest 1 m).



(No. 43)

44. Assume *l* is the line in the first and the third quadrnats bisecting the angle between the *x*-axis and the *y*-axis. Find a point on *l* so that it has the same distance from this point to *A*(8, 0) and to *B*(1, -3).

45. Prove that (-4, 3), (8, 8), (13, -4), (1, -9) are the 4 vertices of a square.

46. Given a set of *n* data, x_1 appears f_1 times; x_2 appears f_2 times; ...; x_k appears f_k times ($f_1+f_2+\cdots+f_k=n$); $\overline{x}$ is the average of the *n* data. Prove that:

$$f_1(x_1-\overline{x})+f_2(x_2-\overline{x})+\cdots+f_k(x_k-\overline{x})=0\,.$$

47. The heights of rice stalk of two types of rice were measured and recorded as follows (Unit: cm):

A:	12	9	16	18	14	8	12	10	17	11
B:	12	15	14	16	15	13	12	10	12	10

Which type of rice has a relatively more uniform height of rice stalk?

48. In a production process, the thickness of a sample of 100 units was measured and recorded. Set out the relative frequency distribution table and draw the relative frequency distribution histogram of the sample.

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	1.42	1.43	1.42	1.42
1.42	1.40	1.41	1.37	1.46	136	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	1.37

This chapter is translated to English by courtesy of Mr. Pan Ng and reviewed by courtesy of Mr SIN Wing Sang Edward.