Chapter Six: Descriptive Measures

Lecturer's copy

When you completed this chapter, you will be able to:

- ✓ realise the characteristics or properties of numerical data;
- ✓ understand the concept of central tendency;
- ✓ understand the concept of dispersion;
- ✓ identify the importance of standard deviation in statistics;
- ✓ use the descriptive summary measures as an aid to data analysis and interpretation; and
- ✓ interpret a data set from a Box-and-Whisker Plot.

Reference(s): Mason Chapter 3 and 4, Berenson Chapter 4, Owen Chapter 4 and 9
Exercise(s): Seminars 9, 10 and 11, Mason Chapter3 Exercises 7, 15, 37, 44, 48 and 57, Mason Chapter4 Exercise19, 20, 24, 25, 28, 30, 34 and 40-51

Measures of Central Tendency and Dispersion

Objective: It focuses on special ways to describe a collection of items, particularly how the observations tend to spread out or concentrated in one area.

1. Measure of Central Tendency

Central tendency refers to the *middle point* of a distribution. Measures of central tendency are also called measure of location.

2. Measure of Dispersion

Dispersion refers to the *spread* of the data in a distribution, i.e. the extent to which the observations are scattered.

Another characteristics of data sets which provides useful information is ***skewness***.

Measure of Central Tendency

For a given set of data, the measure of location/central tendency depends on what we definite the term "*middle*". Different definitions give rise to different measures as shown below:

1. Mean / Arithmetic Mean

The arithmetic mean is also called the mean, which is the most commonly used average or measure of central tendency. It is calculated by summing all the observations in a batch of data and then dividing the total by the number of items involved.

a. Ungrouped Data

To find the arithmetic mean for a given data set of N values, X_1, X_2, X_3,...,X_N, we sum the values and divide by the number of observations, N.

Mean = $\dfrac{\sum x_i}{no.\,of\ observations}$ where : ΣX_i is the summation of all observations

i. To calculate the *population mean* $\mu = \dfrac{\sum \mathbf{x_i}}{\mathbf{N}}$

ii. To calculate the *sample mean* $\bar{\mathbf{x}} = \dfrac{\sum \mathbf{x_i}}{\mathbf{n}}$

e.g.1 Find the Mean of the 10 data: 2, 5, 7, 4, 9, 12, 6, 18, 15, and 6.

$$\text{Mean} = \frac{2+5+7+4+9+12+6+18+15+6}{10} = \frac{84}{10} = 8.4$$

Should we use $\mu = 8.4$ *or* $\bar{x} = 8.4$ *?*

b. Grouped Data

A frequency distribution consists of data that are grouped by classes. Each value of an observation falls somewhere in one of the classes.

Say the 10 data in e.g.1 are grouped into classes as following:

Class	Frequency
0 - 4	2
5 - 9	5
10- 14	1
15 - 19	2
Total	10

To find the arithmetic mean of grouped data, we first calculate the *midpoint of each class*. Then multiply each midpoint by the frequency of observations in that class, sum all these results, and divide the sum by the total number of observations.

$$\textbf{Mean} = \frac{\sum \mathbf{f_i m_i}}{\sum \mathbf{f_i}}$$

where : m_i is the midpoint of each class, f_i is the frequency of each class, $\Sigma f_i\, m_i$ is the sum of all the products of class midpoints and class frequency, and $\Sigma\, f_i$ is the number of observation.

e.g.2 Compute the arithmetic mean of the grouped data in e.g.1.

The class midpoints of the classes are 2, 7, 12, and 17. So

$$\text{Mean} = \frac{\sum f_i m_i}{\sum f_i} = \frac{2x2 + 5x7 + 1x12 + 2x17}{2+5+1+2}$$

$$= 85/10 = 8.5$$

Remark : Similarly, we use μ and $\bar{x}$ to represent the population mean and sample mean respectively.

2. Weighted Mean

The weighted mean enables us to calculate an average that takes into account the *importance* of each value to the overall total.

The weighted mean is found by dividing the sum of products of the values and their weighting by the sum of the weights.

$$\textbf{Mean} = \frac{\sum w_i x_i}{\sum w_i}$$

e.g.3 Suppose that there are two job applicants, and their examination results are given. The selection criterion is based on the examination result of the three subjects, English, Maths., and Chinese. And the manager thinks that English is more important than Maths, and Maths is more important than Chinese, so he gives different weighting to these three subjects' result. Which applicant would be selected?

Subject	**Weighting**	Applicant A	Applicant B
English	**3**	75	50
Mathematics	**2**	70	80
Chinese	**1**	55	80

Weighted Mean Result of Applicant A

$$= \frac{\sum w_i x_i}{\sum w_i} = \frac{3x75 + 2x70 + 1x55}{3+2+1} = \frac{420}{6} = 70$$

Weighted Mean Result of Applicant B

$$= \frac{\sum w_i x_i}{\sum w_i} = \frac{3x50 + 2x80 + 1x80}{3+2+1} = \frac{390}{6} = 65$$

Therefore, Applicant A is selected although the *mean* result of applicant A is poorer than that of Applicant B.

Advantages of Mean:

1. The arithmetic mean, as a *single number* representing a whole data set. Its concept is familiar to most people and *clear*.

2. *Every data set has a mean*. It is a measure that can be calculated and that is *unique* because every data set has *one and only one mean*.

3. *Every observation in the data set is taken into account* when calculating the mean. As a result, the mean is a reliable measure, less likely to be determined by chance than other characteristics of a data set.

4. The mean is useful for performing such statistical procedures as *comparing* the means from several data sets (especially when doing the hypothesis testing).

Disadvantages of Mean:

1. While the mean is reliable in that it reflects all the values in the data set, it may also be *affected by extreme values* that are not representative of the rest of the data.

2. It *takes time to compute* the mean because we do use every data point in the calculation.

3. We could not compute a mean value for some data set with *open-ended class*, e.g. "age 50 and above". We have no way of knowledge whether the value is 50, near 50, or far above 50. (In this case, sometimes we have to make assumptions for the upper class limit.)

3. Median

The median is the middle value in an *ordered sequence* of data. If there is no ties, half of the observations will be smaller and half will be larger.

The median is unaffected by any *extreme observations* in a batch of data. Thus whether an extreme observations is present, it is appropriate to use the median rather than the mean to describe a batch of data,

a. Ungrouped Data

To calculate the median from a batch of data collected in its raw form, we must first place the data in numerical order. Such an arrangement is called an ordered array. We then use the positioning point formula

(n + 1) / 2 th observation

to find the place in the ordered array that correspond to the median value.

Rules:

I If the size of the sample is an odd number, the median is represented by the numerical value corresponding to the positioning point, then it is the $(n+1)/2^{th}$ ordered observation.

e.g. 4 Given the data set: 3, 5, 6, 6, 8, 10, and 13. Find the median.

Median = $(7 + 1)/2^{th}$ observation
= 4^{th} observation
= 6

II If the size of the sample is an even number, then the positioning point lies between the two middle observations. The median is then the arithmetic mean of the numerical values corresponding to these two middle observations.

e.g.5 Given the data set : 3, 5, 6, 6, 8, 10, 13, and 17. Find the median.

Median = $(8 + 1)/2^{th}$ observation
= 4.5^{th} observation
= $(4^{th} + 5^{th})/2$
= $(6 + 8)/2$
= 7

III When data has been presented in the form of frequency distribution, the calculation of the median is slightly different.

e.g. 6 Take the following distribution of the number of clerks employed by a sample of twenty-nine insurance companies :

No. of Clerks	Frequency	Less than or equal to	Cumulative Frequency
25	1	25	1
26	2	26	3
27	1	27	4
28	5	28	9
29	**9**	**29**	**18**
30	11	30	29

In the example, the median will be the number of clerks associated with the, $(29 + 1)/2^{th} = 15^{th}$, fifteenth office.

To find the fifteenth office, we construct a cumulative frequency distribution and try to locate the fifteenth value. In our example, nine offices employed up to 28 clerks and eighteen employed up to 29 clerks. The fifteenth office must therefore employ 29 clerks. The median is therefore 29, although which particular office has 29 clerks is not important.

b. Grouped Data

e.g.7 Grouped frequency distribution of the number of clerks employed by a sample of twenty-nine insurance companies.

No. of Clerks	Frequency
25-26	3
27-28	6
29-30	20

the median should be the, $(29+1)/2^{th}$, 15^{th} observations, and the median is contained in the third class, that is (29 -30). In order to determine the single value of the median from the class, use the formula :

$$\textbf{Median} = L_m + \frac{\frac{n}{2} - \text{Cum. Freq.}}{f} w$$

Where :

L_m = Lower boundary of the median class.
n = the total number of observations
n/2 - Cum. Freq. = n/2 minus the cumulative frequency of the class preceding the median class.
f = frequency of the median class.
w = Width of the median class.

If we use the above formula to compute the median, then

L_m = 28.5, n = 29, n/2-cum.freq. = 29/2 - 9 = 5.5, f = 20, and w = 30.5 - 28.5 = 2.

$$\text{Median} = L_m + \frac{\frac{n}{2} - \text{Cum. Freq.}}{f} w$$
$$= 28.5 + (5.5 / 20) \times 2$$
$$= 29.05$$

Advantages of Median:

1. *Extreme* values do not affect the median.

2. Median is *easy to understand and can be calculated from any kind of data*, even for grouped data with open-ended classes such as the frequency distribution, unless the median falls into an open-ended class.

3. We can find the median even when our data are *qualitative descriptions* like colour or sharpness, rather than numbers.

   Suppose we have seven runs of a printing press, the results from which must be rated according to the sharpness of the image. We can array the results from best to worst : extremely sharp, very sharp, very sharp, sharp, sharp, slightly blurred, and very blurred. The median of the seven rating is (7+1)/2, or the 4th rating (*sharp*).

Disadvantages of Median:

1. Certain statistical procedures that use the median are more complex than those that use the mean are.

2. Since the median is an *average of position*, we must *array the data* before we can prefer any calculations. This is *time-consuming* for any data set with a large number of elements. Therefore, if we want to use a sample statistics as an estimate of a population location parameter, the mean is easier to use than the median.

4. Mode

The mode is a measure of central tendency that is different from the mean but somewhat like the median because it is not actually calculated by the ordinary processes of arithmetic.

The mode is that value that is *repeated most often* in the data set.

a. Ungrouped Data

e.g. 8 Delivery trips per day :

0	2	5	7	**15**
0	2	5	7	**15**
1	4	6	8	**15**
1	4	6	12	19

the mode tells us that *15 is the most frequent number* of trips.

b. Grouped Data

e.g. 9 Delivery trips per day :

Class (no. of trips)	Frequency	
0 - 3	6	
4 - 7	**8**	⇐ **Modal Class**
8 - 11	1	
12 - 15	4	
16 - 19	1	

To calculate the mode from grouped data:

In order to determine a single value of the mode from this modal class, we use

$$\mathbf{M_o = L_{mo} + \frac{d_1}{d_1 + d_2} w}$$

Where :

L_{mo} = lower boundary of the modal class.
d_1 = frequency of the modal class minus the frequency of the class directly below it.
d_2 = frequency of the modal class minus the frequency of the class directly above it.
w = width of the modal class.

If we use the above equation to compute the mode, then

$L_{mo} = 3.5$, $d_1 = 8 - 6 = 2$, $d_2 = 8 - 1 = 7$, and $w = 7.5 - 3.5 = 4$.

$$M_o = L_{mo} + \frac{d_1}{d_1 + d_2} w$$
$$= 3.5 + [\ 2 / (2 + 7)\ ] \times 4$$
$$= 4.39$$

c. Multimodal Distribution

What happens when we have two different values that each appears the greatest number of items of any values in the data set?

e.g.10 Delivery trips per day :

0	2	5	7	**15**
0	**4**	5	7	**15**
1	**4**	6	8	**15**
1	**4**	6	12	19

notice that both 4 and 15 appear the greatest number of times in the data set. They each appear three times. This distribution, then, has two modes and is called a "*Bimodal Distribution*".

Advantages of Mode:

1. The mode, like the median, can be used as a *central location for qualitative as well as quantitative data*.

2. Also like the median, the mode is not affected by *extreme* values. Even if the high values are very high and the low values are very low, we choose the most frequency value of the data set to be the modal value. We can use the mode no matter how large, how small, or how spread out the values in the data set happen to be.

3. The mode can be used even when one or more of the classes are *open-ended*.

Disadvantages of Mode:

1. The mode is *not used as often* to measure central tendency *as are the mean and median*.

   Too often, there is *no modal value* because the data set contains no values that are repeated. Other times, *every value* is the mode because every value occurring the same number of times. Clearly, the mode is *useless measure* in these cases.

2. When the data set contains two, three, or many modes, they are difficult to *interpret* and compare.

3. In some cases, mode in grouped data *cannot reflect* the mode in raw data.

e.g. 11: In the raw data set : 1, 1, **2, 2, 2, 2, 2, 2,** 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 9, 9. The *mode is 2*, since it repeated 6 times.

But if the data is grouped as :

Class	Frequency	
1 - 3	9	
4 - 6	**12**	⇐ **Modal Class**
7 - 9	6	

The true mode is 2, but in the grouped data, the mode falls into the class (4 - 6).

5. Midrange

$$\textbf{Midrange} = \frac{X_{largest} + X_{smallest}}{2}$$

6. Midhinge

Asides from the measures of central tendency, and dispersion, there also exist some useful measures of "*noncentral*" location, which are often employed when summarising or describing the properties of large batches of quantitative data. These measures are called *Quartiles*.

The *first quartile*, Q_1, is the value such that 25% of the observations are smaller and 75% of the observations are larger.

The *second quartile*, Q_2, is the *median* such that 50% of the observations are smaller and 50% of the observations are larger.

The *third quartile*, Q_3, is the value such that 75% of the observations are smaller and 25% of the observations are larger.

To approximate the quartiles from a population containing N observations, the following positioning point formulas are used :

Q_1 = value corresponding to the (N+1)/4 th observation

Q_2 = median, the value corresponding to the (N+1)/2 th observation

Q_3 = value corresponding the 3(N+1)/4 th observation

Midhinge = (Q_1 + Q_3)/2

Rules:

I. The resulting position point is an integer, the particular numerical observation corresponding to that positioning point is chosen for the quartile.

II. If the resulting positioning point is halfway between two integers, the average of their corresponding values is selected.

III. If the resulting positioning point is neither an integer nor a value halfway between two integers, a simple rule of thumb used to approximate the particular quartile is to *round off to the nearest positioning point* and select the numerical value of the corresponding observation.

e.g.12 Given a data set : 2, 4, 7, 8, 9, 15, 17, 20, 22, and 25.

$$\text{Midrange} = \frac{X_{\text{largest}} + X_{\text{smallest}}}{2}$$
$$= (25 + 2)/2$$
$$= 13.5$$

The first quartile, Q_1 = $(10+1)/4^{th}$ observation
= 2.75^{th} observation
≅ 3^{rd} observation
= 7

The second quartile, Q_2 = $(10+1)/2^{th}$ observation
= 5.5^{th} observation
= $(5^{th} + 6^{th})/2$
= (9 + 15)/2
= 12

The Third quartile, Q_3 = $3(10+1)/4^{th}$ observation
= 8.25^{th} observation
≅ 8^{th} observation
= 20

Midhinge = $(Q_1 + Q_3) / 2$
= (20 + 7) / 2
= 13.5

Class Exercise 1:

Given the data set: 2, 4, 4, 6, 7, 7, 9,11, 13, and 20.
Find the mean, median, mode, midrange, and midhinge.

Measure of Dispersion

Two sets of data can have the *same central location but they may be very different if one is more spread out than the other* is.

This is true for the three distributions in figure below. The mean of all three curves is the same, but curve A has less spread (or variability) than curve B, and curve B has less variability than curve C. If we measure only the mean of these distributions, we will miss an important different among the three curves.

錯誤! 連結無效。

Likewise for any data, the mean, median, and the mode tell us only part of what we need to know about the characteristics of the data. To increase our understanding of the pattern of the data, we *must also measure its dispersion* (spread or variability).

The reasons for measuring the dispersion of the distribution:

1. It gives us additional information that enables us to judge the *reliability* of our measure of the central tendency.

   If data are widely dispersion, such as those in curve C in the above figure, the central location is *less representative* of the data as a whole than it would be for data that are more closely centred around the mean, as in curve A.

2. We may wish to compare dispersions of various samples. If, for example, a wide spread of values away from the centre is undesirable or presents an *unacceptable risk*, we need to be able to recognise and avoid choosing those distributions with greatest dispersion.

e.g.13 Financial analysts are concerned about the dispersion of a firm's earnings. Widely dispersed earnings - those varying from extremely high to low or even negative levels - indicate a higher risk to stockholders and creditors than the other firms with relatively stable earnings.

e.g.14 Similarly, a drug that is average in purity but that ranges from very pure to highly impure may endanger lives.

For a given set of data, the most common measures of dispersion are:

1. Range

The range is the *difference between the largest and smallest values* in the data set. In equation form, we have

$$\textbf{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

e.g. 15 Given the data set : 12, 34, 23, 59, 12, 25, 35, 56, 29, 3, 29, and 55.

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$
$$= 59 - 3 = 56$$

Advantages of Range:

The range is easy to understand and to calculate.

Disadvantages of Range:

1. Its usefulness as a measure of dispersion is *limited*. The range considers only highest and lowest values of a distribution, and fails to take account of any other observations in the data set. As a result, it ignores the nature of the variation among all the other observations, and it is *heavily influenced* by *extreme values*.

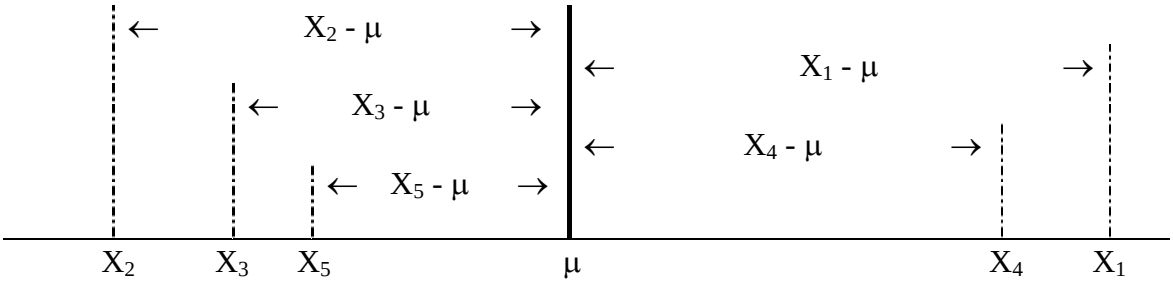
2. *Open-ended distributions have no range* because no "highest" or "lowest" value exists in the open-ended class.

3. The range is *less stable of measures*. For example, in repeated samples taken from the some sources, the range will exhibit more variation from sample to sample than the other measures.

4. As the number of observations is increased, the range generally tends to become *larger*, therefore, to use the ranges to compare the variations in two sets of data is not proper unless they contain the *same number of values*.

2. Deviations

It tells us the *average distance* of any observations in the data set *from the mean* of the distribution.



Negative Deviation ***Positive Deviation***

a. Ungrouped Data

To find the total of the deviations, take their sum and divide it by the number of deviations.

Mean Deviation = $\frac{1}{N}\sum(x_i - \mu)$

Unfortunately, the quantity is equal to *zero*.

Since we are really interested in the *magnitude* of the deviations, *and not in their direction*. We might simply ignore their *signs*, and define a measure of variation in terms of the *absolute values* of the deviation from the mean.

The formula is revised to:

Mean Absolute Deviation (MAD) = $\frac{1}{N}\sum|x_i - \mu|$

It is called *Mean Absolute Deviation (MAD)*. Where $|X_i - \mu|$ is the absolute value of $(X_i - \mu)$, which *convert both the negative and positive deviation to positive deviation*.

b. Grouped Data

$$\textbf{Mean Absolute Deviation (MAD)} = \frac{1}{\sum f_i} \sum f_i \, |m_i - \mu|$$

where m_i is the class midpoint of each class, and f_i is the frequency of each class.

Advantage of MAD:

MAD is a better measure of dispersion than the ranges because it *takes every observation into account*. It *weights each item equally* and indicates how far, on average, each observation lies from the mean.

Disadvantage of MAD:

It is difficult to use in the mathematical operations.

3. Population Variance (σ^2)

Each population has a variance, which is symbolised by σ^2, (sigma squared).

The population variance is similar to the deviation. But it takes *the sum of the squared distances between the mean and each item, and then divides the sum by the total number of elements in the population*. By squaring each distance, it automatically makes every number *positive*.

a. Ungrouped Data

$$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$$

b. Grouped Data

$$\sigma^2 = \frac{1}{\sum f_i} \sum f_i (m_i - \mu)^2$$

4. Population Standard Deviation (σ)

When calculating the range or the MAD, the answers were expressed in the same units as the data itself. For the variance, however, *the units of σ^2 are the squared of the units of the data*.

For example, if the data are in "inches", the variance is in "inches squared". Squared inches are *not easily interpreted*.

For this reason, we have to make a significant change in the variance to compute a useful measure of deviation, one that does not give us a problem with units of measure and thus is less confusing. This parameter is called the *Standard Deviation*. While the variance is expressed in the squared of the units used in the data, *the standard deviation is in the same units as those used in the data*.

The formula for Population Standard Deviation, which is symbolised by σ (sigma), is:

a. Ungrouped Data

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum (x_i - \mu)^2}$$

b. Grouped Data

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{\sum f_i}\sum f_i (m_i - \mu)^2}$$

The square root of a positive number may be either positive or negative. However, when taking the squared root of the variance to calculate the standard deviation, *consider only the positive square root*.

Advantage of Standard Deviation:

Like the Mean Absolute Deviation, it takes into account *every observation* in the data set.

Disadvantages of Standard Deviation:

1. It is not as easy to calculate as the Range.

2. It cannot be computed from open-ended distributions.

3. Extreme values in the data set distort the value of the standard deviation, although to a *lesser extent* than they do the Range.

Interpretation and Uses of the Standard Deviation

The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are location in relation to the mean.

Chebyshev's Theorem:

Given a number k greater than 1, and a set of n measurements, X_1, X_2,...,X_n. there are *at least* $(1 - 1/k^2)$ *of the measurements will lie within k standard deviations of their mean*.

Chebyshev's Theorem stated that *no matter what the shape of the distribution,*

1. The interval $(\mu \pm 2\sigma)$ will contain *at least* 75 % of the measurements.

2. The interval $(\mu \pm 3\sigma)$ will contain *al least* 89 % of the measurements.

3. The interval $(\mu \pm 4\sigma)$ will contain *al least* 94 % of the measurements.

Empirical Rule:

Given a distribution of measurements that is *approximately Symmetrical and Bell-Shaped*, we can say that:

1. The interval $(\mu \pm \sigma)$ will contain approximately 68 % (68.26 %) of the measurements.

2. The interval $(\mu \pm 2\sigma)$ will contain approximately 95 % (95.44 %)of the measurements.

3. The interval $(\mu \pm 3\sigma)$ will contain all or almost all (99.73 %) of the measurements.

e.g.16 Given a data set : 0.04, 0.06, 0.12, 0.14, 0.14, 0.15, 0.17, 0.17, 0.18, 0.19, 0.21, 0.21, 0.22, 0.24, and 0.25. Find the MAD, Variance and Standard Deviation.

Observation		Mean (μ)		Deviation	Absolute Deviation	Squared Deviation
x		**$\Sigma x_i/n$**		**$(x_i-\mu)$**	**$\lvert x_i-\mu \rvert$**	**$(x_i-\mu)^2$**
0.04	-	0.166	=	-0.126	0.126	0.016
0.06	-	0.166	=	-0.106	0.106	0.011
0.12	-	0.166	=	-0.046	0.046	0.002
0.14	-	0.166	=	-0.026	0.026	0.001
0.14	-	0.166	=	-0.026	0.026	0.001
0.15	-	0.166	=	-0.016	0.016	0.000
0.17	-	0.166	=	0.004	0.004	0.000
0.17	-	0.166	=	0.004	0.004	0.000
0.18	-	0.166	=	0.014	0.014	0.000
0.19	-	0.166	=	0.024	0.024	0.001
0.21	-	0.166	=	0.044	0.044	0.002
0.21	-	0.166	=	0.044	0.044	0.002
0.22	-	0.166	=	0.054	0.054	0.003
0.24	-	0.166	=	0.074	0.074	0.005
0.25	-	0.166	=	0.084	0.084	0.007
Σx_i = 2.49					**$\Sigma \lvert x_i - \mu \rvert$ =0.692**	**$\Sigma(x_i-\mu)^2$=0.051**

$$MAD = \frac{1}{N}\sum|x_i - \mu|$$
$$= 0.692/15$$
$$= 0.0461$$

$$\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$$
$$= 0.051 /15$$
$$= 0.0034$$

$$\sigma = \sqrt{\sigma^2}$$
$$= \sqrt{0.0034}$$
$$= 0.58$$

e.g.17 Given a set of data that are grouped as followings, find the MAD, Variance and Standard Deviation.

Class	Freq. f_i	Mid-point m_i	$f_i m_i$	Mean $\Sigma f_i m_i / \Sigma f_i$	Deviation $f_i (m_i-\mu)$	Absolute Deviation $f_i \lvert m_i-\mu \rvert$	Squared Deviation $f_i(m_i-\mu)^2$
10-19	2	14.5	29	36.5	-44	44	968
20-29	6	24.5	147	36.5	-72	72	864
30-39	10	34.5	345	36.5	-20	20	40
40-49	8	44.5	356	36.5	64	64	512
50-59	4	54.5	218	36.5	72	72	1296
			$\Sigma f_i m_i$ = 1095			**$\Sigma f_i \lvert m_i - \mu \rvert$ = 272**	**$\Sigma f_i(m_i-\mu)^2$ = 3680**

$$MAD = \frac{1}{\sum f_i} \sum f_i |m_i - \mu|$$
$$= 272 / 30$$
$$= 9.0667$$

$$\sigma^2 = \frac{1}{\sum f_i} \sum f_i (m_i - \mu)^2$$
$$= 3680 / 30$$
$$= 122.6667$$

$$\sigma = \sqrt{\sigma^2}$$
$$= \sqrt{122.6667}$$
$$= 11.0754$$

5&6 Sample Variance (s^2) and Sample Standard Deviation (s)

To compute the Sample Variance and Sample Standard Deviation, we use the same formula as the equations for population, except *replacing μ with $\bar{x}$, and* ***N with (n-1),*** where $\bar{x}$ is the sample mean and n is the no. of data in the sample.

The formulas for Sample Variance are:

a. Ungrouped Data

$$\mathbf{s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

b. Grouped Data

$$\mathbf{s^2 = \frac{1}{\sum f_i - 1} \sum f_i (m_i - \bar{x})^2}$$

and the formula for Sample Standard Deviation is:

$$\mathbf{s = \sqrt{s^2}}$$

e.g.18 Suppose the data set in e.g.16 is sample data, find the variance and the standard deviation.

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$
$$= 0.051 /(15 - 1)$$
$$= 0.051 / 14$$
$$= 0.0036$$

$$s = \sqrt{s^2}$$
$$= \sqrt{0.0036}$$
$$= 0.0603$$

The computational formulas for the population variance and population standard deviation are:

$\sigma^2 = \frac{1}{N}(\sum x_i^{\,2} - n\mu^2)$ and $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}(\sum x_i^{\,2} - n\mu^{2)}}$ respectively.

Try to use the computational formula to redo exercise 16. What is the difference?

What is the computational formula for sample variance and sample standard deviation?

7. Interquartile Range

The interquartile range measures approximately how far from the median we must go either side before we can include *one half of the values* of the data set.

Interquartile Range = $Q_3 - Q_1$

e.g.19 Given the data set: 2, 4, 7, 8, 9, 15, 17, 20, 22, and 25.

The first quartile, Q_1 = $(10+1)/4^{th}$ observation
= 2.75^{th} observation
≅ 3^{rd} observation
= 7

The Third quartile, Q_3 = $3(10+1)/4^{th}$ observation
= 8.25^{th} observation
≅ 8^{th} observation
= 20

Interquartile Range = $Q_3 - Q_1$
= 20 - 7
= 13

Advantage of Interquartile Range

Although they are more complicated to calculate than the range, they *ignore extreme values by using only the middle half of the data*. Thus they have a distinct advantages over the range, which is affected by the extreme values.

Disadvantage of Interquartile Range

Like the range, the interquartile range is *based on only two values* from the data set.

8. The Coefficient of Variation (Relative Dispersion)

Coefficient of variation is a relative measure of dispersion, it is expressed as a *percentage rather than in terms of the units of the particular data*.

It is particularly useful when *comparing the variability of two or more batches of data that are expressed in different units of measurement*.

Coefficient of Variation = (Standard Deviation / Mean) x 100

Population Coefficient of Variation:

$$\mathbf{CV} = \frac{\sigma}{\mu} \times 100 \qquad \mu > 0$$

Sample Coefficient of Variation:

$$\mathbf{CV} = \frac{s}{\bar{x}} \times 100 \qquad \bar{x} > 0$$

e.g.20 In the same set of examinations, Tom and Jack obtained both different means and different standard deviation of marks as follows :

Examination Statistics	Tom	Jack
Mean Marks	48	80
Standard Deviation	1.0	1.5

The Coefficients of Variation are:

Tom's CV = (1 / 48) x 100 = 2.083

Jack's CV = (1.5/80) x 100 = 1.875

In this case, despite Jack having the *larger Standard deviation*, he has the *lower coefficient of variation*. Jack is ***more consistent***.

Class Exercise 2:

Given the sample data set: 2, 4, 4, 6, 7, 7, 9,11, 13, and 20.

Find the MAD, Variance, Standard Deviation, Coefficient of Variation, Range, and Interquartile Range.

Other Description of Data Distribution

Distribution of actual data can assume almost any shape or form, but most of those that arise in practice can be described fairly well by one or another of few standard types.

A very important one is the *symmetrical Bell-Shaped Distribution* shown in Figure 1. In many cases, distributions of actual data can be expected to follow this form very closely.

錯誤! 連結無效。

The other two distributions of Figure 2 and 3 can still be called bell shaped, but they certainly cannot be called symmetrical. Distribution of this sorts, having a pronounced "tail" on one side or the other, are said to be *skewed*.

Those, as in Figure 2, with a tail on the right are *positively skewed (right skewed).*

Mathematically, skewness of a distribution can be identified by the relationship among the three central tendency measures, mean, mode and median.

For a data set with **Mean > Median > Mode**, that is *right* skewed.

錯誤! 連結無效。

Those, as in Figure 3, with a tail on the *left* are *negatively skewed (left skewed*), and in this case, **Mean < Median < Mode**.

錯誤! 連結無效。

Furthermore, degree of skewness of a data distribution can be measured by the Coefficient of Skewness:

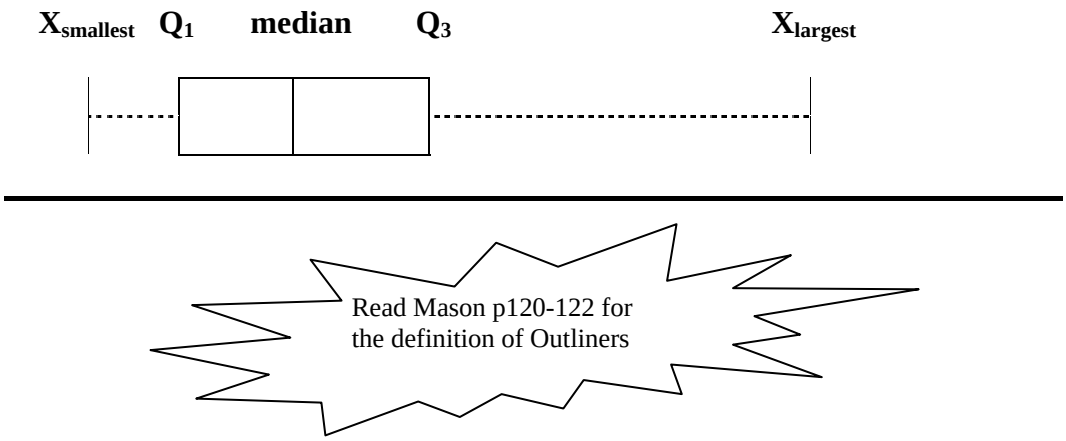
Coefficient of Skewness = $\frac{\textbf{3(mean – median)}}{\textbf{standard deviation}}$, which normally lies between –3 and +3.

The Box-and-Whisker Plot

The Box-and-whisker plot are graphical displays that summarise the main features of a set of data including the *central tendency, dispersion, symmetry, and distances (or tail lengths) to the minimum and maximum values*.

The *edges of a box are drawn at the first quartile (Q_1) and the third Quartile (Q_3)* of the data set so that the *box covers the middle half of the values*. A *line is placed in the box is the median*. The line emanating from the box are known as *whiskers*, and they are extending to cover the values from the box *down to the minimum value* and *up to the maximum value*.

The data displayed graphically by the box plot are known as *five-number-summary*.



e.g.21 For a given set, suppose that we have the following information:

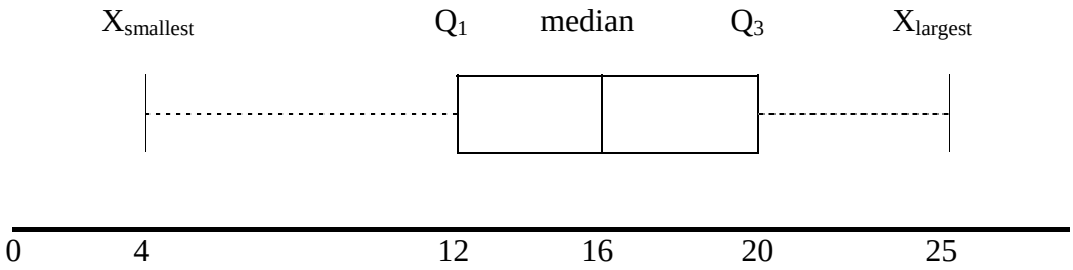
$X_{largest} = 25$, $X_{smallest} = 4$, mean = 14, median = $Q_2 = 16$,

$Q_1 = 12$, $Q_3 = 20$, standard deviation = 3.8765

i) draw a Box-and-Whisker plot for the data set;
ii) determine the coefficient of skewness; and
iii) comment on the skewness.

Sol]

i) the Box-and-Whisker plot is as follows:



ii) Coefficient of Skewness = $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

$= \frac{3(14-16)}{3.8765} = -1.5478$

iii) the distribution has found to be negatively skewed.

Class Exercise 3:

Refer back to Class Exercises 1 and 2, the given data set are: 2, 4, 4, 6, 7, 7, 9,11, 13, and 20. And we have found that:

$X_{largest} = 20$, $X_{smallest} = 2$, mean = 8.2, median = $Q_2 = 7$, mode = 4 and 7, midrange = 11, $Q_1 = 4$, $Q_3 = 11$, midhinge = 7.5,
MAD = 3.96, variance = 28.0111, standard deviation = 5.2926, CV = 0.6377, range = 18, and Interquartile Range = 7

i) produce a Box-and Whisker plot for the data set;
ii) determine the coefficient of skewness; and
iii) comment on the skewness.

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 1 (Solution)

mean = $\mu = \frac{\sum x_i}{N} = 82 / 10 = 8.2$

median = (n + 1) / 2 th observation
= (10 + 1) / 2 th observation
= 5.5 th observation
= (7 + 7) / 2
= 7

mode = 4 and 7

midrange = ($X_{largest}$ + $X_{smallest}$) / 2
= (20 + 2) / 2
= 11

midhinge = (Q_1 + Q_3) / 2

Q_1 = (N + 1) / 4 th observation
= 11/4 th observation
= 2.75 th observation
≅ 3 rd observation
= 4

Q_3 = 3(N + 1) / 4 th observation
= 3 x 11 / 4 th observation
= 8.25 th observation
≅ 8 th observation
= 11

midhige = (4 + 11) / 2
= 7.5

Class : _____ **Name :** _____ **No. :** _____

Class Exercise 2 (Solution)

mean = $\bar{X} = \frac{\sum x_i}{n}$ = 83 / 10 = 8.3

Observation x		Mean ($\bar{X}$) $\Sigma x_i/n$		Deviation $(x_i - \bar{X})$	Absolute Deviation $\lvert x_i - \bar{X} \rvert$	Squared Deviation $(x_i - \bar{X})^2$
2	-	8.3	=	-6.3	6.3	39.69
4	-	8.3	=	-4.3	4.3	18.49
4	-	8.3	=	-4.3	4.3	18.49
6	-	8.3	=	-2.3	2.3	5.29
7	-	8.3	=	-1.3	1.3	1.69
7	-	8.3	=	-1.3	1.3	1.69
9	-	8.3	=	0.7	0.7	0.49
11	-	8.3	=	2.7	2.7	7.29
13	-	8.3	=	4.7	4.7	22.09
20	-	8.3	=	11.7	11.7	136.89
Σx_i =					**$\Sigma \lvert x_i - \bar{X} \rvert$ = 39.60**	**$\Sigma(x_i - \bar{X})^2$= 252.1**

MAD = $\frac{1}{N}\sum |x_i - \bar{x}|$ = 39.60 / 10 = 3.96

$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ = 252.1 / (10 - 1) = 252.1 / 9 = 28.0111

$s = \sqrt{s^2}$ = $\sqrt{28.0111}$ = 5.2926

Sample Coefficient of Variation = CV = $\frac{s}{\bar{x}}$x100 = 5.2926 = 0.6377

Range = $X_{largest} - X_{smallest}$ = 20 - 2 = 18

From class exercise 1, $Q_1 \cong$ 3 rd observation = 4

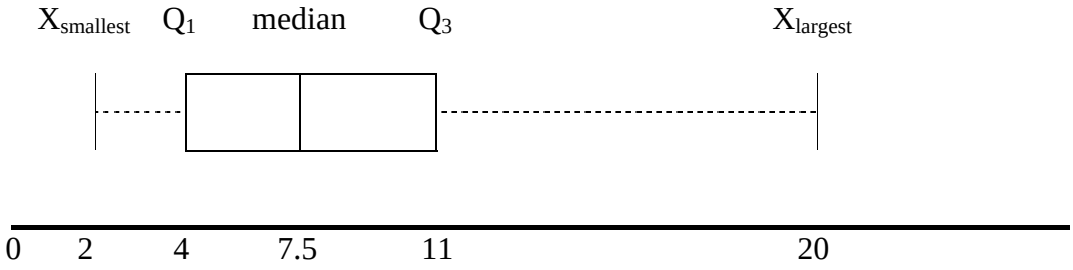
$Q_3 \cong$ 8 th observation = 11

Interquartile Range = $Q_3 - Q_1$ = 11 - 4 = 7

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 3 (Solution)

i)



ii) Coefficient of Skewness $= \dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

$= \dfrac{3(8.2 - 7)}{5.2926} = 0.6798$

iv) the distribution has found to be positively skewed.