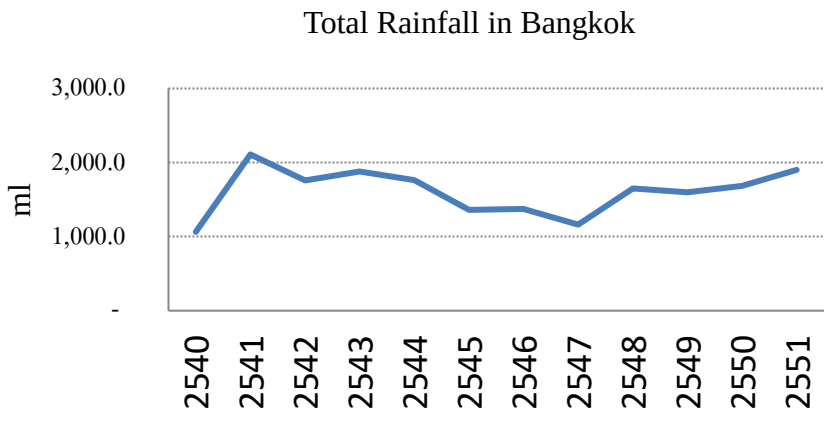
Chapter 1

Descriptive Statistics

1.1 Introduction

People are known to be familiar with statistics, more or less, as the concept of conducting polls or selecting a member of the parliament, for instance, requires the collection of data to ensure accuracy. It is common that accurate data collection can be hard to accomplish; for example, the collection of demographic data for Thailand's entire population (Census). Once data is collected, statistics will gather the latter information to proceed on with accuracy verification, data management, coding, data analysis, and present information in a form of tables or graphs; for example: a line graph depicting total rainfall measured from Bangkok Meteorological Station (Queen Sirikit National Convention Center) from year 2540 BE to 2551 BE.

Statistical procedures consist of statistical methods, which include: data collection, manipulation of data or data processing, data presentation, data analysis, and data interpretation to get an outcome or come to a conclusion.



The Definition of 'Statistics'

In our daily lives, we often see statistics play a major role in life, for instance, what we see in newspaper heads that says 'From the year 2549 BE up to the present, it is found that the average amount of Thai women got married earlier and in their younger years compared to that of the past; where the average age of 23.1 jumped to 22.2 in the present year. It has been found that continuously, women who live away from the municipality got married faster and in their younger years than that of those living in municipality. Furthermore, research has found that before marriage, the average amount of women in the entire country with ages ranging from 15 – 49 and their husbands, receive family planning consultation from health personnel for only about 15%, undergo blood tests for Thalassemia for 19.4%, and tests for HIV for only 20.9%'.

Source: News, 10 June, 2553 BE, National News Agency Public Relations Department

Statistics

Statistics is a general field that involves data collection, data analysis, and data summarization aimed towards decision-making. The word 'statistics' in German, was once translated as 'Political Science' which is derived from a Latin word 'Status'; Romans used the word 'Status' to define position and social status, but later held a meaning of 'State' and was in use for a long period of time before it was replaced by today's definition (for more details, refer to: Book, Stuart Witt, Statistics and Political Science, 1993, Full Quart Press).

1.1.1 Technical Words

Population refers to every unit of information or data of interest

Sample refers to a group or a part of the population

Collecting data from every unit of population is hard to accomplish and is costly. For instance, inquiring the opinions of people eligible for elections regarding Thai politics, can be challenging as the process of inquiring everyone will require both time and money. A study to collect data from every unit of the population (Population unit) is called a 'Census', such as the preparation of the Thai census and housing census, where a survey has been conducted since year 2542 BE. The Ministry of Interior has conducted surveys for a total of 5 times, the National Statistical Office as well conducted a total of 5 times similar to that of the latter, while the final survey has been conducted in the year 2543 BE. The most recent census and housing preparation was in 2553, the 11^{th} time that it has been done. However, in statistical studies, we tend to study **'Sampling'** (obtaining samples), analyzing, and summarizing or concluding the results from data obtained from the samples, which is referred to as **'Descriptive Statistics'**. The process of summarizing and interpreting data obtained as of the population is referred to as **'Inferential Statistics'**.

The accuracy of the results will depend upon the sampling process. Good samples should represent a population, where the sample unit shall own similar characteristics to those of a population.

Parameter refers to a feature in which may represent a number than can explain a population

Statistics refers to a fact in form of a number that is used to explain a population

	Parameter	Statistic
Mean (or Average)	μ	$\bar{x}$
Standard Deviation	σ	s
Variance	σ^2	s^2
Proportion	p	$\hat{p}$

1.2 Data

We can classify the data types into ‘quantitative’ and ‘qualitative’ data according to the following definitions:

Quantitative data is a numerical data obtained from making measurements or counting.

Qualitative data is information that shows the type, classification, or name.

Example 1 Information on Mercedes Benz cars is as shown in the following table.

Model	Package	Engine (CC)	Engine Power (kilowatt/hoursepower)	Price (Baht)
C 200	Edition C	1,796	135/184	2,290,000
C 250	AMG Plus	1,796	150/204	2,990,000
C 180 Coupe	AMG Dynamic	1,595	115/156	2,990,000
SLK 200	AMG Dynamic	1,796	135/184	3,490,000
E 200	Executive	1,991	135/184	3,390,000
E 300 BlueTEC HYBRID	Executive	2,143	150/204	3,690,000
E 200 Coupe	AMG Dynamic	1,991	135/184	3,790,000
E 200 Cabriolet	AMG Dynamic	1,991	135/184	3,990,000
CLS 250 CDI	AMG Dynamic	2,143	150/204	4,990,000

Note: The information has been announced for use since 25 November, 2556 BE from www.mercedes-benz.co.th

From the above table, it is possible to classify the model number (for example: 180, 250, 200, 300) and car package as qualitative information, whereas engine type, engine power, and costs are all quantitative information.

Scales of Measurement

Levels of measurement was first introduced by a psychologist named Stanley Smith Stevens in the year 1946 (international year format), where it has been published in an article called 'On the theory of scales of measurement'. This article has stated that every measurement procedure in the field of science requires levels of measurement: 1) Nominal scale, 2. Ordinal scale, 3. Interval scale, and 4. Ratio scale, where the details of each level are as follows:

1. Nominal scale

   A nominal scale is a level of measurement that identifies the type or characteristics of the sample unit, for example: faculty, ie. Economics, Science, Laws, Social Science and Humanities, in which were all represented with numerical values of 1, 2, 3, 4, chronologically, despite the correspondence to their importance. Stevens (1946, p.679) has made a statement regarding coding that, "...the use of numerals as names for classes is an example of the assignment of numerals according to rule. The rule is: Do not assign the same numeral to different classes or different numerals to the same class. Beyond that, anything goes with the nominal scale", where the coded numerals are only to replace the group in which the unit belongs to, and therefore, these numerals cannot be used to find the average value or median.

2. Ordinal scale

   An ordinal scale is a level of measurement in which data value or factors can be arranged in order, where the difference in size, be it more or less, cannot be identified. This type of scale measurement enables the assignment of order to factors, for example: 1, 2, or 3. Phrases like least, less, more, most, or dissatisfaction, common, and satisfaction are some of the general forms of ordinal scale.

3. Interval scale

   An interval scale uses numerals to indicate quantity, where the difference of two values is significant, for example: Mr. Somkit scores 20 marks on his test, which is 8 marks more than that of Mr. Sompong who scores only 12 marks on his test; another

example would be that the temperature of 20 degrees Celsius is 10 degrees more than that of 10 degrees Celsius. However, it is not right to assume that 20 degrees Celsius is 2 times more than 10 degrees Celsius as the interval scale method adopts a **Non-true zero** in which 0 degrees Celsius does not refer to a non-existing temperature value, but indicates a freezing point. A score of '0' on a test is not a way of measuring one's intellect either.

Therefore, data obtained through this level of measurement cannot be divided or multiplied, but difference ratio can be calculated. For example: the difference between the temperature of 20 degrees Celsius and 40 degrees Celsius is 2 times more than the difference between the temperature of 35 degrees Celsius and 45 degrees Celsius, consequently.

4. Ratio scale

A ratio scale is commonly used in the field of science, often seen in forms of: density, length, time, angle, energy, and electric current. This scale of measurement has an **True zero.** For instance, a unit of 0 Kelvin (0K) has a true zero value of -273.15 degrees Celsius, as 0K, in social science, is a point where the particles form a substance with an energy of 0. Some examples include: sales, number of rooms in a hotel, wages, and the total amount of test questions a student has done correctly.

1.3 Frequency Table and Histogram

In this section, students will be able to understand how data explanation and summarization works for a better understanding of the concept. For example: the test scores in Mathematics obtained from a total of 48 students, where the full score is of 40 marks.

40	30	34	15	21	34	32	28	26	25
31	18	32	35	32	29	22	23	31	18
35	39	32	31	25	24	30	15	32	38
33	27	20	31	34	15	31	21	33	29
32	21	29	27	27	22	28	28		

Without finding the median or presenting the above scores in a form of graphs, it is hard to see the patterns or understand the data.

1.3.1 Frequency Distribution and Graphing

Frequency distribution is often presented in a form of tables which consists of class intervals and the counting of data that falls into each class interval which can be equated as follows:

Step 1 Find data range

Range = the highest value of data – the lowest value of data

$$R = X_{max} - X_{min}$$

Step 2 Estimate the number of classes, which normally is around 5 – 20 levels, as introduced by Herbert Sturges (1926) and his 'Sturges' rule' to estimate the number of classes.

$$K = 1 + 3.322 \log N$$

Where N is the amount of all data obtained and K is the number of classes.

Sturges' rule best suits the case where there is normal distribution of data (Rob J Hyndman, 1995). The K value obtained is always rounded to whole number; the number of classes can as well be estimated by $\sqrt{N}$.

Step 3 Find the class width

Class width (I) = Range (R) / number of classes (K)

Note: fraction is always rounded to whole number; the width of a level will have a decimal point equals to the data decimal point.

Step 4 Set a class limit

Set a lower limit for a class interval with the lowest score, where the latter value must cover the data's least value. The least value is often referred to as the lower limit of a class interval with the lowest score. The upper limit can be calculated when the lower limit is added to the width and the number resulting from the addition needs to be subtracted by 1; the following equation can be used:

First lower class limit = (the smallest data value) / (the class width)

- If an outcome is a whole number after division, use this number to multiply with the class width
- If an outcome is not a whole number, round the decimal point to make it a whole number then multiply it with the class width

Step 5 Finding class boundaries

The class boundary's main amount placed in front of the decimal point always exceeds the real data's amount for 1 digit. For instance, if the real amount the data holds is a whole number, the class boundary will have a one-point decimal, where the equation is as follows:

Lower class boundary = (lower class limit + upper class limit of previous class) / 2

Upper class boundary = (upper class limit + lower class limit of the next class) / 2

Step 6 Find the midpoint

Midpoint = (upper class limit + lower class limit) / 2

Step 7 Find the frequency

Find the frequency by counting the amount of observation value that falls in each class interval

Step 8 Find the cumulative frequency

Cumulative frequency is the sum of the frequencies in a certain class interval with every frequencies in lower class intervals.

Step 9 Find the relative frequency

Relative frequency is a frequency of a class interval divided by the sum of all frequencies.

Example 2 From the obtained test marks, make a frequency distribution table

Solution:

Step 1 Find the data range

Step 2 Estimate the number of classes

Step 3 Find the class width

Step 4 Set the lower limit of a class interval with the lowest score, by using the lowest data value as a lower limit of the first class interval

Step 5 Set a class limit

Step 6 Find the midpoint, frequency, cumulative frequency, relative frequency, and relative cumulative frequency

Class Limit	Class Boundaries	Midpoint	Tally	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency

1.3.2 Creating Histogram and Frequency Polygon

Creating a histogram can be done by writing down the upper limit and lower limit of each class interval on the X–axis, while the frequency or the relative cumulative frequency can replace the Y-axis (in most cases the Y–axis is often replaced with frequency)

A frequency polygon can be drawn by joining the midpoint in each class interval, including the one before the first class interval, and the other interval which is after the last interval.

Example 3 From example 2, create a histogram and a frequency polygon

1.4 Measures of Central Tendency

Apart from using histograms and frequency distribution tables, data can be explained through numerals that show the central point and data distribution. Measures of central tendency often use 'Arithmetic Mean' or AM: mean, median, mode, where these 3 central values adopt the same unit with the data.

1.4.1 Mean

A mean is a value that represents the central point of the data; by summing up all the values and divide the outcome by the amount of data obtained.

The following shows the symbols used:

Ungrouped data for data that is not in the form of a frequency distribution table

Population mean is equated as: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample mean is equated as: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Grouped data: for data in a form of frequency distribution table

Population mean is equated as: $\mu = \frac{\sum_{i=1}^{k} f_i x_i}{N}$

Where

k is the amount of class interval

x_i represents the midpoint of level i

f_i represents the frequency of level i

The mean of the sample adopts a similar equation

Example 4 Suppose the data (population) is 22, 25, 30, 30, 37, and 60. Find the population mean

Solution:

Example 5 Mathematics test scores of 48 students are as follows:

40	30	34	15	21	34	32	28
26	25	31	18	32	35	32	29
22	23	31	18	35	39	32	31
25	24	30	15	32	38	33	27
20	31	34	15	31	21	33	29
32	21	29	27	27	22	28	28

Calculate the arithmetic mean from the data given above

Solution:

From example 2, frequency distribution table can be calculated as follows:

Scores	x_i	f_i	$f_i x_i$
15 – 18	16.5	5	82.5
19 – 22	20.5	6	123
23 – 26	24.5	5	122.5
27 – 30	28.5	11	313.5
31 – 34	32.5	16	520
35 – 38	36.5	3	109.5
39 – 42	40.5	2	81
Total		48	1352

Note: Can the arithmetic mean of a frequency distribution table with open end be calculated?

1.4.2 Median

A median is a value located in the center of sorted data and is a value in which reveals that half of all the observation values is lesser than or equals to the latter value. The other half of the remaining data will be more than the median.

Ungrouped data: In case where data is not represented in the form of a frequency distribution table

1. Sort the data from least to most
2. Find the location of the median $\frac{N+1}{2}$ where N is the amount of all data
3. Find the median

   If N is an odd number, the median is a value of data located in $\frac{N+1}{2}$

   If N is an even number, the median is an average value or the mean of two data that surrounds the location of $\frac{N+1}{2}$

Grouped data: In case where the data is represented in the form of a frequency distribution table

1. Calculate the location with the median of $\frac{N}{2}$
2. Consider the class interval with a cumulative frequency of $\frac{N}{2}$ or similar

- If $\frac{N}{2}$ equals to cumulative frequency in any class, the median will equal to the upper class boundary
- If $\frac{N}{2}$ does not equal to cumulative frequency in any class, consider the cumulative frequency in the class before and after the location with a median and use the rule of 3 to find the median.

Or use the following equation

$$\text{Med} = L + \frac{I}{f}\left(\frac{N}{2} - \sum f_L\right)$$

Where

L represents the lower border with a median

I represents the width of a class interval

f represents the frequency of a class with a median

$\sum f_L$ represents the cumulative frequency of a class that is located before the class with a median

Example 6 Suppose the data (population) is 22, 25, 30, 30, 37, and 60. Find the median

Solution:

The location of the median is

Example 7 Find the median of the Mathematics test scores of 48 students

Solution:

Scores	Class Boundaries	f_i
15 – 18	14.5 – 18.5	5
19 – 22	18.5 – 22.5	6
23 – 26	22.5 – 26.5	5
27 – 30	26.5 – 30.5	8
31 – 34	30.5 – 34.5	16
35 – 38	34.5 – 38.5	6
39 – 42	38.5 – 42.5	2
Total		48

Example 8 From example 4, find the median of the test scores obtained from 48 students

Solution:

Scores	Class Boundaries	f_i
15 – 18	14.5 – 18.5	5
19 – 22	18.5 – 22.5	6
23 – 26	22.5 – 26.5	5
27 – 30	26.5 – 30.5	11
31 – 34	30.5 – 34.5	16
35 – 38	34.5 – 38.5	3
39 – 42	38.5 – 42.5	2
Total		48

1.4.3 Mode

A mode is a value that can be obtained by finding the data value with the most frequency or repetition. A mode is an appropriate method for data in the form of countable nouns or ordinances. A mode can find more than 1 value if the data value has an equal value of highest frequency; however, it may not in the case where there is no value with highest frequency that exceeds other values.

Ungrouped data: For data that is not in the form of a frequency distribution table

A mode is a data with highest frequency or data that has the most repetitions

Example 9 Suppose the data (population) is: 22, 25, 30, 30, 37, and 60. Find the mode

Solution:

Example 10 Suppose the data (population) is 1, 2, 3, 4, 5, and 6. Find the mode

Solution:

Example 11 Suppose the data (population) is 1, 1, 2, 2, 3, and 4. Find the mode

Solution:

Example 12 Suppose the data (population) is 1, 1, 2, 2, 3, and 3. Find the mode

Solution:

Example 13 From the random sampling of 100 students, inquiring their faculty, the obtained data are in the table below:

Faculty	Economics	Business	Science	Others
Number of Students	76	14	6	4

Find the appropriate measure of central tendency

Solution:

Grouped data: In case the data is in the form of a frequency distribution table, the mode will be in a class with the highest frequency. It can be calculated with the equation shown below:

$$\text{Mode} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} I$$

Where

L is the lower class boundary of a class the mode is in (the class with the highest frequency)

l is the width of the class the mode is in

f_0 is the frequency of the class before the one the mode is in

f_1 is the frequency of the class the mode is in

f_2 is the frequency of the class after the one the mode is in

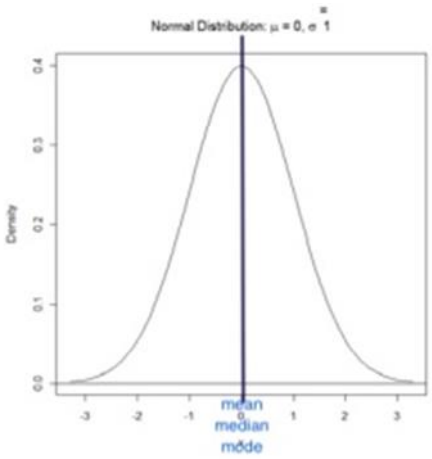
Example 14 From example 5, find the mode of the Mathematics test scores obtained from 48 students

Solution:

Scores	Class Boundaries	f_i
15 – 18	14.5 – 18.5	5
19 – 22	18.5 – 22.5	6
23 – 26	22.5 – 26.5	5
27 – 30	26.5 – 30.5	11
31 – 34	30.5 – 34.5	16
35 – 38	34.5 – 38.5	3
39 – 42	38.5 – 42.5	2
Total		48

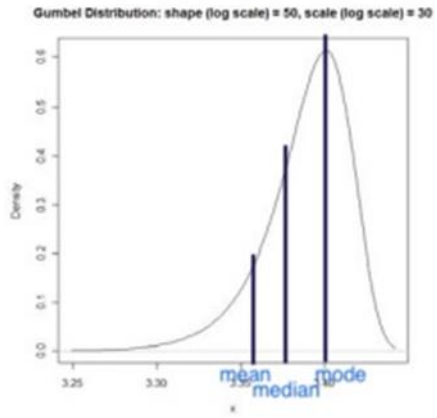
Data distribution

Normally, the arithmetic mean and median is different in value. If population distribution has a tendency of slanting to the left or right, the arithmetic mean will not be equal to the median. Generally, if the data has a biased distribution pattern or extreme highest and/or lowest values detected, the median is often used. In case where population distribution is symmetrical, the mean is equal to the median and the mode as shown below.



Symmetric distribution

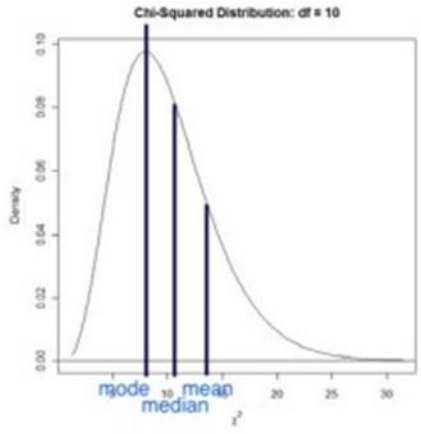
Mode = median = mean



Negative skewed distribution

(or Left skewed distribution)

Mode > median > mean



Positive skewed distribution

(or Right skewed distribution)

Mode < median < mean

1.5 Measure of Location

We now know that the median is a value that divides the obtained data into 2 equal parts, while there are as well other values that classify data into varying parts, for example: quartile positioning value.

The quartile positioning value classifies the data into 4 equal parts, therefore there will be 3 quartile values represented by Q_1, Q_2, and Q_3.

Decile positioning value groups the data into 10 equal parts, and therefore, there will be 9 decile values represented by D_1, D_2,..., D_9, consequently.

Percentile positioning value will sort the data equally into 100 parts, which the 99 values of percentile will be represented by P_1, P_2,...P_{99} consequently.

Sample Definition of Quartile, Decile, and Percentile

Q_1 is a value of the data that proves the following: $\frac{1}{4}$ of all the data is less than or equal to this value, and $\frac{3}{4}$ of all the data is more than this value

Q_2 is a value of the data that proves the following: $\frac{2}{4}$ of all the data is less than or equal to this value, and $\frac{2}{4}$ of all the data is more than this value

Q_3 is a value of the data that proves the following: $\frac{3}{4}$ of all the data is less than or equal to this value, and $\frac{1}{4}$ of all the data is more than this value

D_7 is a value of the data that proves the following: $\frac{7}{10}$ of all the data is less than or equal to this value, and $\frac{3}{10}$ of all the data is more than this value

P_{15} is a value of the data that proves the following: $\frac{15}{100}$ of all the data is less than or equal to this value, and $\frac{85}{100}$ of all the data is more than this value

Finding data positioning values

Ungrouped data: In case where data is not represented in the form of a frequency distribution table

1. Sort the data in an ascending order (from least to most)
2. Find the data location with the following equation

$$Q_i = \frac{i}{4}(N+1), \quad i = 1,2,3$$

$$D_i = \frac{i}{10}(N+1) \quad , i = 1,2,...,9$$

$$P_i = \frac{i}{100}(N+1) \quad , i = 1,2,...,99$$

where N is the amount of all data

3. Find the value of the location obtained in step 2 by using the 'Rule of Three'

Example 15 Suppose the data (population) is 22, 25, 30, 30, 37, 60, and 60. Find Q_2, D_4, P_{75}.

Solution:

Grouped data: In case where data is presented in the form of a frequency distribution table

1. Find the location of data with the following equation

$$Q_i = \frac{i}{4}(N), \quad i = 1,2,3$$

$$D_i = \frac{i}{10}(N) \quad , i = 1,2,...,9$$

$$P_i = \frac{i}{100}(N) \quad , i = 1,2,...,99$$

where N is the amount of all data

2. Consider the class with cumulative frequency that equals to Q_i, D_i, P_i, or other similar values

- If Q_i, D_i, and P_i equals to any cumulative frequency within a class, their values will therefore be equal to the upper class boundary of that particular class
- If Q_i, D_i, and P_i does not equal to any cumulative frequency within a class, consider the cumulative frequency of 2 classes that are located nearby Q_i, D_i, and P_i and can be calculated, to apply the rule of three in order to find the values of Q_i, D_i, and P_i, respectively.

Example 16 The frequency distribution table of the test scores obtained from 48 students. Find the values of Q_3, D_4, and P_{25}, and interpret the latter.

Scores	f_i	Cumulative frequency
14.5 – 18.5	5	5
18.5 – 22.5	6	11
22.5 – 26.5	5	16
26.5 – 30.5	9	25
30.5 – 34.5	11	36
34.5 – 38.5	6	42
38.5 – 42.5	6	48
Total	48	

1.6 Measures of Dispersion

Explaining data through measures of central tendency or median can be insufficient. For example, there are 2 sets of data: 1) 22, 25, 30, 30, 37, 60 and 2) 15, 21, 22, 25, 40, and 81, where these 2 data sets have a similar arithmetic mean value. The mean could also be a representation of the data itself, in which often confuses the evaluator, believing that these 2 data sets are similar. However, in reality, a set of data no.2 has more distribution compared to that of no.1's. Therefore, merely finding the median will not be able to thoroughly explain the data. In statistics, there are a variety of methods upon data explanation. The following are the common methods is use.

1.6.1 Range : R

The range can be found through the difference between the maximum (highest) value and the minimum (lowest) value. This method can quickly measure dispersion; however, in quite a rough way as it only requires 2 values from the data sets obtained.

Data set 1: 22, 25, 30, 30, 37, 60 has a range of: _____

Data set 2: 15, 21, 22, 25, 40, 81 has a range of:_____

Therefore, data set no.2 has a higher dispersion rate than data set no.1

1.6.2 Quartile Deviation: Q.D.

Quartile deviation is a statistical method which measures dispersion through values of positioning that shows the location of Q1 and Q3, with the following equation:

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

It is also a method that measures dispersion roughly, which is similar to finding the range, but a little more precise due to the lack of influence maximum and minimum values have upon the result. Quartile deviation is of the same unit as the data.

1.6.3 Mean Deviation: M.D.

Mean deviation is a statistical method which uses all data values to measure dispersion. The equation works as follows:

In case population data is used: $\mathbf{M.D.} = \dfrac{\sum_{i=1}^{N} |x_i - \mu|}{N}$ (same unit as the data)

In case sample data is used: $\mathbf{M.D.} = \dfrac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

The above is a value that represents dispersion through the average distance of all values from the median. It is a value easy to understand; however, due to mathematical reasoning, the process of statistical analysis rarely uses mean deviation.

1.6.4 Standard Deviation: S.D.

Standard deviation is a statistical method that uses all data values to measure dispersion.

The equation is as follows:

Ungrouped data: In case the data is not represented in the form of a frequency distribution table

Population standard deviation $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2 - N\mu^2}{N}}$

Sample standard deviation $S=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}}$

Grouped data: In case the data is represented in the form of a frequency distribution table

Population standard deviation $\sigma=\sqrt{\frac{\sum_{i=1}^{k}f_i\left(x_i-\mu\right)^2}{N}}=\sqrt{\frac{\sum_{i=1}^{n}f_ix_i^2-N\mu^2}{N}}$

Sample standard deviation $S=\sqrt{\frac{\sum_{i=1}^{k}f_i\left(x_i-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}f_ix_i^2-n\bar{x}^2}{n-1}}$

Where

X_i is the midpoint of class interval i

fi is the frequency of i

k represents the number of classes

If standard deviation is to the power of 2, it will be referred to as a '**variance**', which is a statistical method that measures distribution.

Note: Dividing with n-1 will result in the statistical value of S which has a better characteristic than dividing it with n. Standard deviation applies a similar concept with that of the mean deviation, where it involves all data values with the power of 2s instead of the absolute value. Standard deviation is a statistical method that is commonly used to measure dispersion, as it has numerous statistical characteristics that are beneficial.

Example 17 Suppose the data (population) is 22, 25, 30, 30, 37, 60. Find the standard deviation and variance.

In case it is to be equated manually, use $\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - N\mu^2}{N}}$

Solution:

1.6.5 Coefficient of Variation: C.V.

Coefficient of variation is used to measure data dispersion similar to standard deviation or other statistical methods. However, as the coefficient of variation is a dimensionless number, it can show the level of data dispersion and can be used to compare the dispersion of 2 or more data sets. As for data with different unit or highly different mean, coefficient of variation equation is shown below:

In case data (population) is used

$$C.V. = \frac{\sigma}{\mu} \times 100\%$$

In case sample data is used

$$C.V. = \frac{s}{\bar{x}} \times 100\%$$

Note: Coefficient of variation is often expressed in the form of percentage. Apart from coefficient of variation that uses standard deviation as a basic concept, there are other coefficients that use other measures of distribution. For example, $\frac{X_{max} - X_{min}}{X_{max} + X_{min}}$ but is often disregarded

1.7 Standard Score: Z

Comparing 2 or more data obtained from data group with an arithmetic mean, a standard deviation, or a different unit, can be done by converting the desired value to a standard score.

$$Z = \frac{X - \mu}{\sigma}$$

Note that Z is a value without unit.

Where

Z is the standard score

X is the raw score

μ is the mean

σ is the standard deviation

Example 18 To compare between mathematical and English abilities of 9 students who belong to the same class, the following shows the scores.

	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8	No.9
Mathematics	5	6	8	7	6	9	7	5	4
English	8	7	9	10	4	6	5	8	9

1. Find the subject which the students are able to learn better.
2. Find the subject which the 5^{th} student has ability to learn better.

1.8 Other Graphs for Data Explanation

Apart from histograms that we have previously discussed, there are as well other statistical graphs and diagrams that can be used mainly for data explanation. Using graphs for data explanation can be based upon the levels of measurement, for example: the amount of monthly precipitation in Nan province represented by a line graph as the amount of rainfall adopts a ratio scale measurement and requires data collection in intervals. However, to present data in the form of percentage and proportion, for example, of professors with varying academic positions under a department, we may use bar graphs or pie charts as the academic positions are qualitative data that require a level of measurement in the form of an ordinal scale. The salary of families residing in Bangkok can be represented with a histogram, as the salary is in the form of an ordinal scale and is a quantitative data. In this section, we will be discussing other significant charts: stem and leaf plot as well as the box and whisker plot, where these two charts were first introduce by John Wilder Turkey.

Stem and Leaf Plot is a type of chart or graph that uses real data values to plot in the form of a bar graph.

Example: Mathematics test scores obtained from 48 students, where the full score is 40. The scores are arranged from most to least as follows:

15	20	22	26	28	30	31	32	34	38
15	21	23	27	28	30	31	32	34	39
15	21	24	27	29	31	32	32	34	40
18	21	25	27	29	31	32	33	35	
18	22	25	28	29	31	32	33	35	

How to construct a stem and leaf plot:

1. Observe the data; they are 2-digit numerals. Therefore, the tenths will be the ‘stem’ and the roots are the ‘leaf’
2. Use the existing data to write the following:

In this example, the stem is only limited to 2-digit numerals

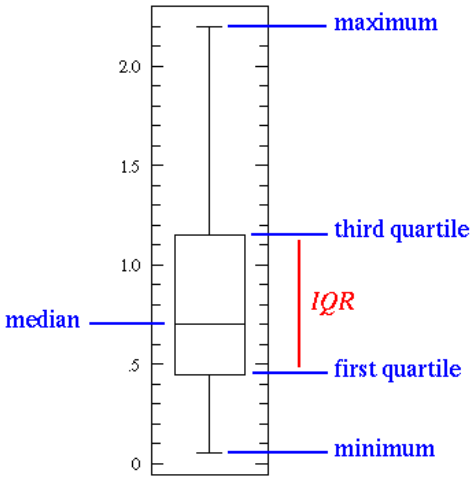
Stem | Leaf

```
1 | 555          (ข้อมูล 3 ตัวคือ 15, 15, 15)
1 |
1 | 88
2 | 0111
2 | 223
2 | 455
2 | 6777
2 | 888999
3 | 0011111
3 | 22222233
3 | 44455
3 |
3 | 89
4 | 0
```

From the stem and leaf plot, the results can be evaluated as having the average score of 32-33, lowest of 15, and the highest of 40 marks.

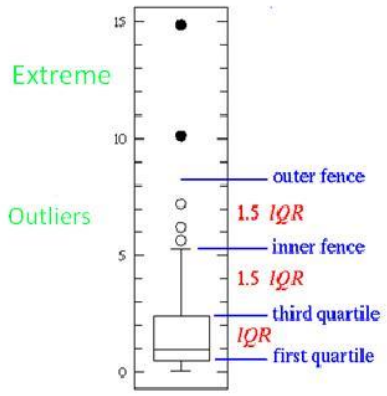
Box and whisker plot is commonly used for data explanation which provides 5 significant statistical values: minimum and maximum data, first quartile, second quartile (median), and third quartile. The following shows the Box and Whisker plot.

The line that has been drawn outwards from the box is called the 'Whisker line', which can be drawn up to 1.5 IQR. If any of the observation value is further away from the end of the box for more than 1.5 IQR, it will be referred to as the 'Outlier'. John Turkey has explained the 2 types of outliers as follows:



Sources : http://www.physics.csbsju.edu/stats/box2.html

1. Highly suspected outlier or 'Extreme' is an observation value that is more than $Q_3 + 3IQR$ or lesser than $Q_1 - 3IQR$.
2. Suspected outlier or Outlier is an observation value that is more than $Q_3 + 1.5IQR$ or less than $Q_1 - 1.5IQR$
   as displayed below



Source: http://www.physics.csbsju.edu/stats/box2.html