

CHAPTER 1

INTRODUCTION TO STATISTICS

Expected Outcomes

- ✓ Able to define basic terminologies of statistics
- ✓ Able to identify various sampling techniques
- ✓ Able to classify type of data and level of measurement
- ✓ Able to summarise data using measure of central tendency, measure of variation and measure of position
- ✓ Able to conduct exploratory data analysis

CONTENT

- 1.1 Statistical Terminologies**
- 1.2 Statistical Problem Solving Methodology**
- 1.3 Review on Descriptive Statistics**
 - 1.3.1 Measures of Central Tendency**
 - 1.3.2 Measures of Variation**
 - 1.3.2.1 Accuracy and Precision**
 - 1.3.3 Measures of Position**
- 1.4 Exploratory Data Analysis**
 - 1.4.1 Stem and Leaf Plot**
 - 1.4.2 Outliers**
 - 1.4.3 Box Plot**
- 1.5 Normal Probability Plot**

1.1 STATISTICAL TERMINOLOGIES

What is Statistics?

→ is the sciences of conducting studies to collect, organise, summarise, analyse, present, interpret and draw conclusions from data.

Any values (observations or measurements) that have been collected

Basic knowledge of statistics is needed in any disciplines or any field of research or study (in almost all fields of human endeavour) that involve data analysis.

Examples:

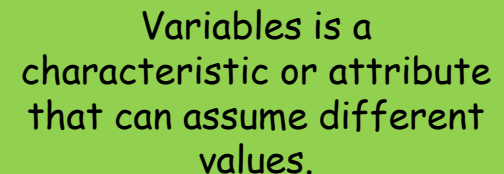
- ✓ In **sports**, statistician may keep records of the number of successful kicks a team scored during a football season.
- ✓ In **public health**, a doctor might be concerned with the number of child who are infected with a H1N1 virus during a certain year.
- ✓ In **education**, an educator might want to know if the performance of students in current semester are better than the previous semester.

Why we Need Statistics?

Knowledge of statistics may help you in:

1. Describing and understanding numerical relationship between variables.

- Is there any significance relationship between **SPM result** and **GPA** achieved by first year student? If yes, will high SPM result become important criteria in choosing new students?



Variables is a characteristic or attribute that can assume different values.

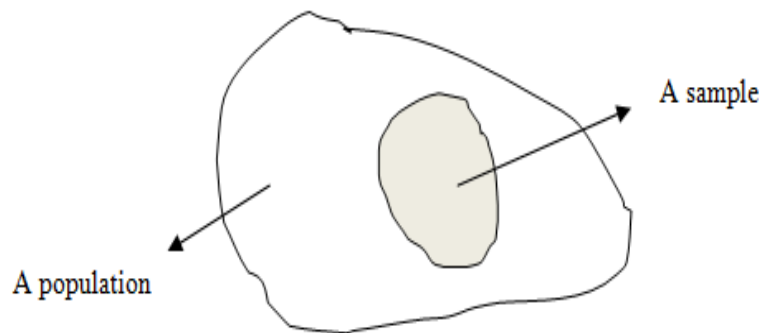
2. Making better decision in the face of uncertainty.

- UMP students claim that there have **no enough time to sleep and study** due to extra curricular activities. The student's committee can use statistical method to show the university how does the extra curricular activities affect the student's performance.

POPULATION AND SAMPLE

Population (N)

A complete collection of measurements, outcomes, objects or individuals under study.



Sample (n)

A subset of the population that is observed

Tangible

finite and the total number of subjects is fixed and could be listed

Ex: all computers in a room, all female students in a university, or all electrical components manufactured in a day, etc.

Conceptual (Intangible)

all values that might possibly have been observed and has an unlimited number of subjects.

Ex: simulated data from computer or instrument, all experimental data such as all measurements of length of metal rod, etc.

Parameter and Statistic

Parameter

A numerical value that represents a certain **population characteristic**

- The **percentage of defective components** in a population of electrical components manufactured in a day
- The **average of height** of students from a population of students in a university, etc.

Statistic

A numerical value that represents a certain **sample characteristic**

- The **percentage of defective components in a sample** of 100 electrical components
- The **average of height** for a sample of female students selected from all students in a university, etc.

Characteristic	Parameter	Statistic
Mean (Average)	μ	\bar{x}
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	π	p

Descriptive and Inferential Statistics

Descriptive statistics

- Includes the process of data collection, data organisation, data classification, data summarisation, and data presentation obtained from the sample.
- Used to describe the characteristics of the sample.
- Used to determine whether the sample represents the target population by comparing sample statistic and population parameter.

EXAMPLE:

Ten thousands parents in Malaysia have chosen Takaful Insurance as their trusted life insurance agency.

Inferential statistics

- Involves a process of generalisation, estimations, hypothesis testing, predictions and determination of relationships between variables.
- Used to describe, infer, estimate, approximate the characteristics of the target population.
- Used when we want to draw a conclusion for the data obtain from the sample.

EXAMPLE:

The death rate of lung cancer was 10 times higher for smokers compared to nonsmokers .

Role of the Computer in Statistics

Two software tools commonly used for data analysis:

1. Spreadsheets

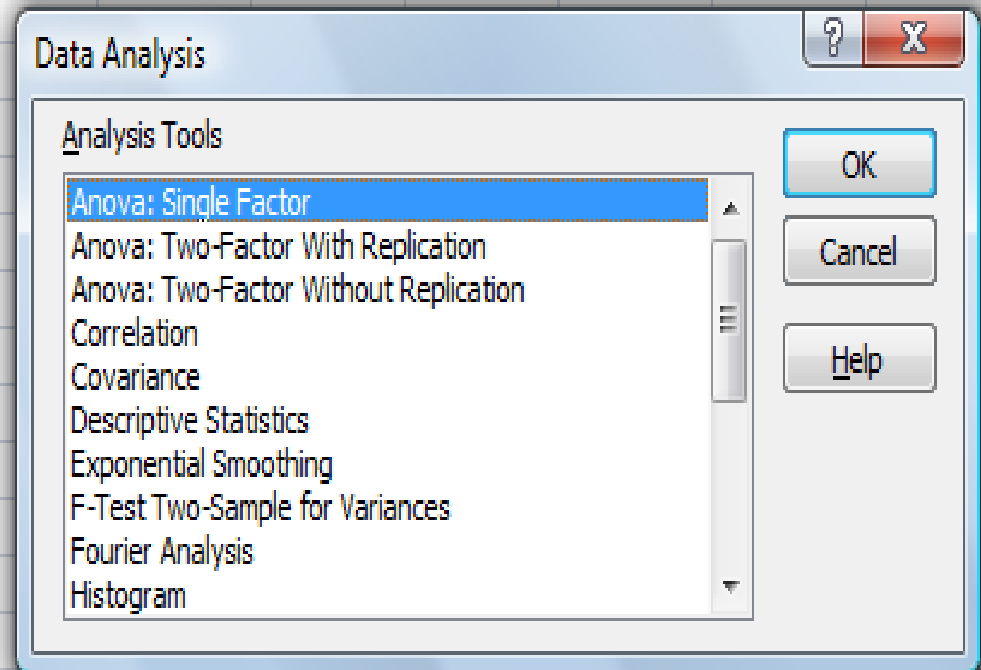
- **Microsoft Excel** & Lotus 1-2-3

2. Statistical Packages

- AMOS, eViews, MINITAB, R, SAS, SmartPLS, SPSS and SPlus

Data Analysis Application Tools in EXCEL

1. Graph and chart
2. Formulas
3. Data Analysis Tools:
File → Options → Add-Ins
→ Analysis ToolPak → ok
→ Data → Data Analysis



1.2: STATISTICAL PROBLEM-SOLVING METHODOLOGY

1. Identify the problem or opportunity

- ❖ Define objective of study.
- ❖ Population or sample? How large?
- ❖ Treatment, experiment, or measurement?

2. Decide on the method of data collection

- ❖ Internal, external, primary or secondary data
- ❖ Experimental study data
- ❖ Observation, survey, questionnaire

3. Collect the data (sampling techniques)

- ❖ Non-probability data: judgment, voluntary and convenience samples.
- ❖ Probability data: random, systematic, stratified, and cluster samples.

1.2: STATISTICAL PROBLEM-SOLVING METHODOLOGY

4. Classify and summarise the data

- ❖ Classify data as qualitative/categorical/attributes data (nominal or ordinal) or quantitative/numerical data (discrete or continuous)
- ❖ Classify data by its levels of measurement: nominal-level data, ordinal-level data, interval-level data and ratio-level data.
- ❖ Summarise data by graphical: tables, histogram, frequency polygon, ogive, pareto charts, time series graphs, pie charts, stem and leaf plot, and boxplot.
- ❖ Summarise data by descriptive statistics: measure of central tendency, measure of variation and measure of position

5. Present and analyse the data

- ❖ Descriptive statistics – analyse the properties of data and shape of distribution from the graphical summary and descriptive summary.
- ❖ Inferential statistics – confidence interval, hypothesis testing, ANOVA, correlation, regression analysis, etc.

6. Make the decision and conclusion

- ❖ Suggest the best decision, options, solution and conclusion of the study.

Collecting the Data

A. Nonprobability data

- based on **the judgment of the experimenter** i.e. the method that could affect the results of the sample
- 3 basic methods: *Judgment samples, Voluntary samples and Convenience samples*

B. Probability data

- Is one in which **the chance of selection** of each subjects in the population **is known before the sample is picked**
- 4 basic methods : *random, systematic, stratified, and cluster.*

A. Nonprobability Data Samples

1. Judgment samples

- ✓ Based on opinion of one or more expert person.
- ✓ Ex: A political campaign manager intuitively picks certain voting districts as reliable places to measure the public opinion of his candidates.

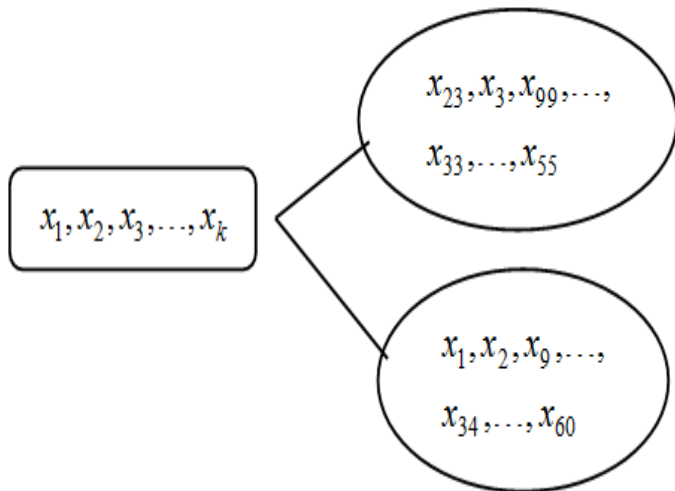
2. Voluntary samples

- ✓ Questions are posed to the public by publishing them over radio or television via phone, short message, email etc.

3. Convenience samples

- ✓ Take an 'easy sample' (most conveniently available). Also called as haphazard or accidental sampling refers to the procedure of obtaining units or people who are most conveniently available.
- ✓ Ex: A surveyor will stand in one location & ask passerby their questions.

B) Probability Data Samples



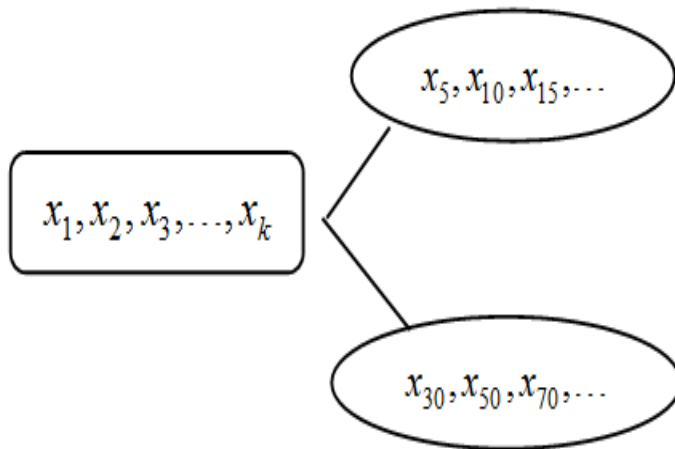
1. Random sampling

- each data is numbered, and then the data is selected using chance or random method. Each data has an equal chance to be selected.

Example:

Suppose a lecturer wants to study the physical fitness levels of students at his/her university. There are 5,000 students enrolled at the university, and he/she wants to draw a sample of size 100 to take a physical fitness test. **She could obtain a list of all 5,000 students, numbered it from 1 to 5,000 and then randomly invites 100 students corresponding to those numbers to participate in the study.**

B) Probability Data Samples



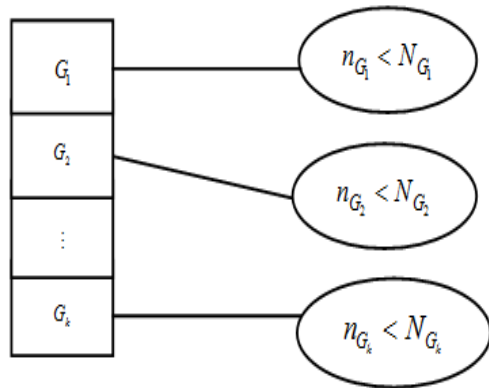
2. Systematic sampling

- Each data is numbered and then the data is selected every k^{th} number where $k=N/n$. The first data is selected randomly between data number 1 and k .

Example:

Suppose a lecturer wants to study the physical fitness levels of students at his/her university. There are 5,000 students enrolled at the university, and he/she wants to draw a sample of size 100 to take a physical fitness test. **She obtains a list of all 5,000 students, numbered it from 1 to 5,000 and randomly picks one of the first 50 voters ($5000/100 = 50$) on the list. If the picked number is 30, then the 30th student in the list should be invited first. Then she should invite the selected every 50th name on the list after this first random starts (the 80th student, the 130th student and so on) to produce 100 samples of students to participate in the study.**

B) Probability Data Samples



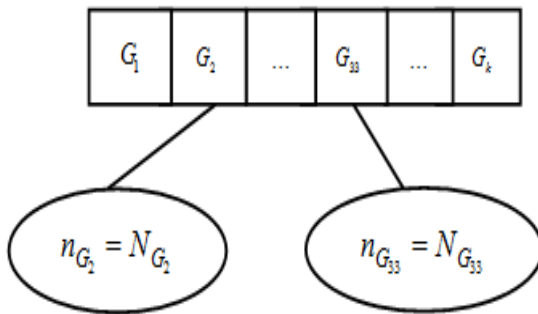
3. Stratified sampling

- the population is divided into groups according to some characteristics that is important to the study, then the sample is selected from each group using random or systematic sampling.

Example:

Suppose a lecturer wants to study the physical fitness levels of students at his/her university. There are 5,000 students enrolled at the university, and he/she wants to draw a sample of size 100 to take a physical fitness test. **Assume that, because of different lifestyles, the level of physical fitness is different between male and female students. To account for this variation in lifestyle, the population of student can easily be stratified into male and female students. Then she can either use random method or systematic methods to select the participants. As example, she can use random sample to choose 50 male students and use systematic method to chose another 50 female students or otherwise.**

B) Probability Data Samples



4. Cluster sampling

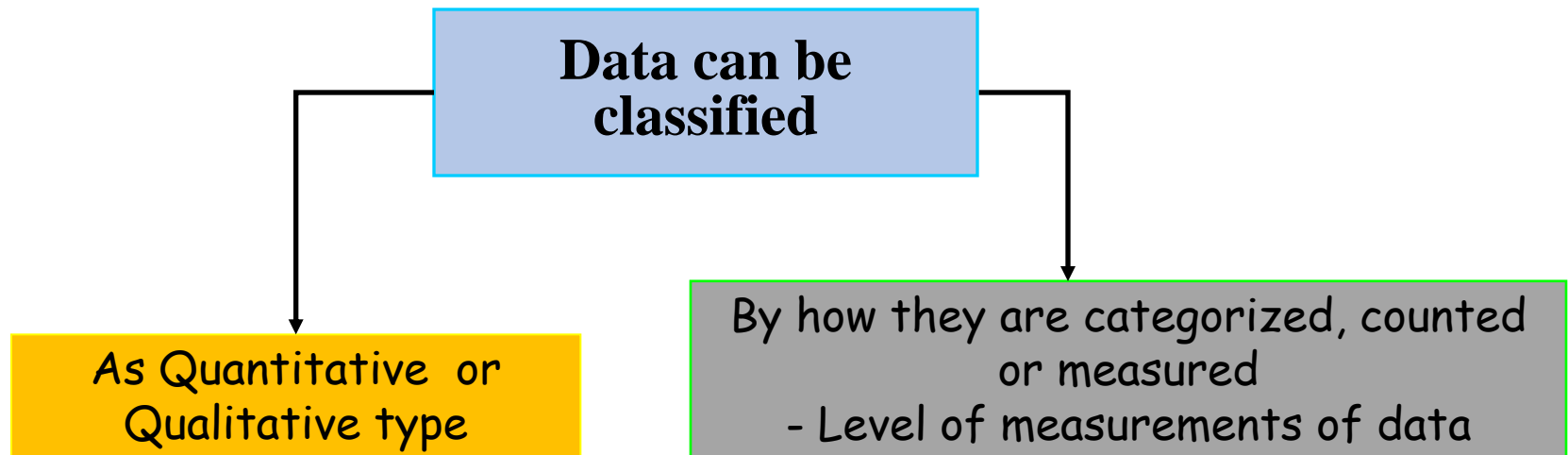
- The population is divided into groups or clusters, then some of those clusters are randomly selected and all members from those selected clusters are chosen. Cluster sampling can reduce cost and time.

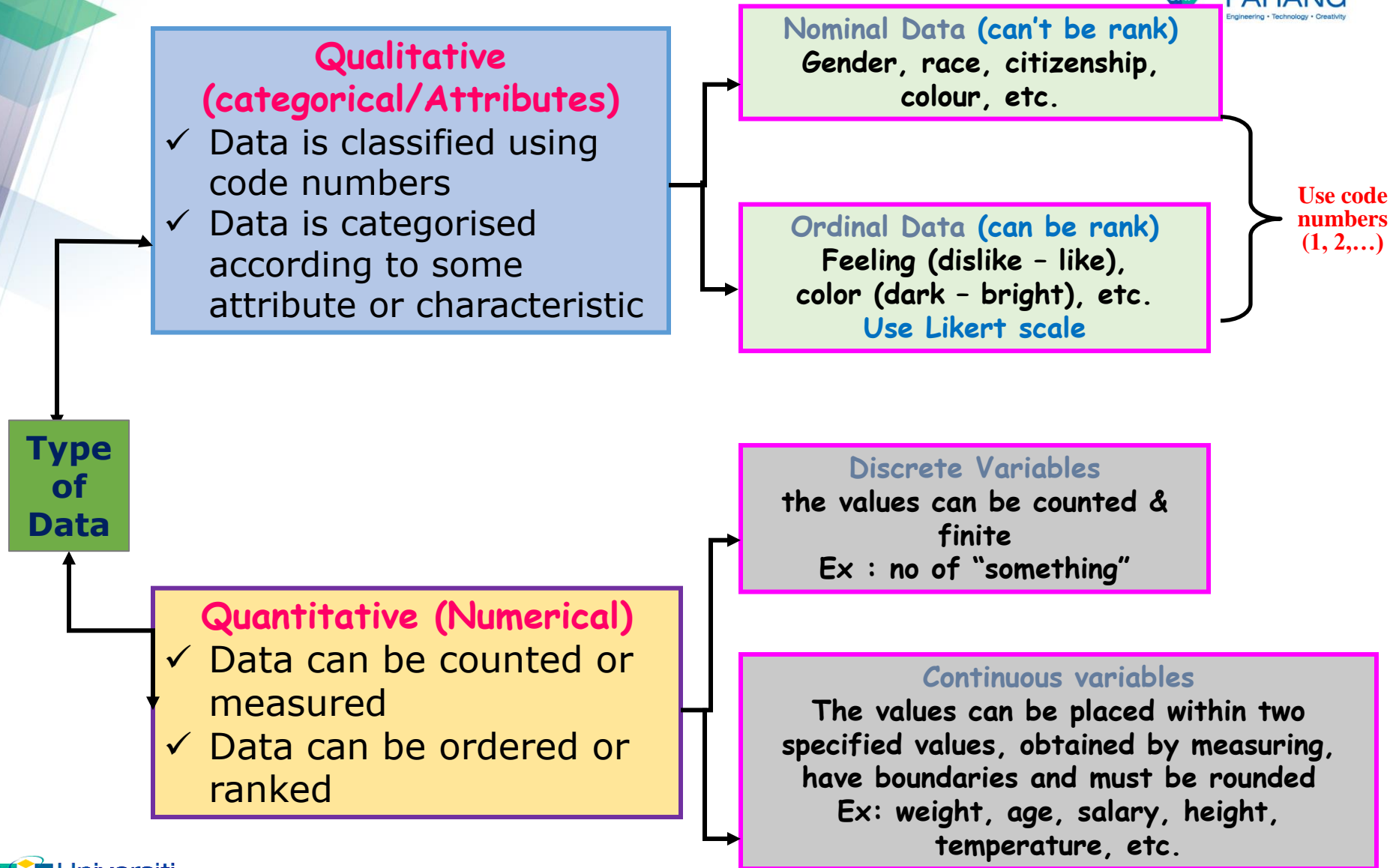
Example:

Suppose a lecturer wants to study the physical fitness levels of students at his/her university. There are 5,000 students enrolled at the university, and he/she wants to draw a sample of size 100 to take a physical fitness test. Assume that, because of different lifestyles, the level of physical fitness is different between 1st year, 2nd year, 3rd year and seniors students. **To account for this variation in lifestyle, the population of student can easily be clustered into that four categories and then he/she can choose any one cluster that consists for example 2nd year students take all of them as the participants.**

Data Classification

- ❑ Data are the values that variables can assume.
- ❑ Variables is a characteristic or attribute that can assume different values.
- ❑ Variables whose values are determined by chance are called random variables.



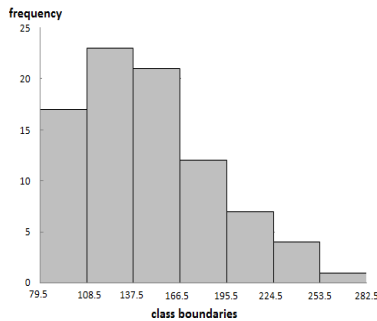


Level of Measurements of Data

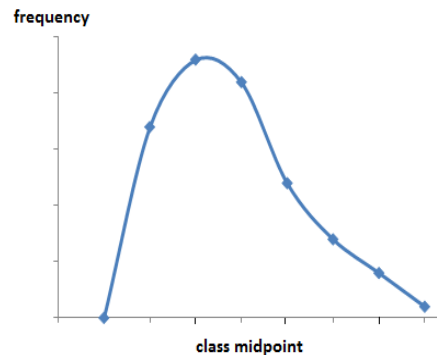
Levels	Descriptions	Examples
Nominal-level	Classifies data into mutually exclusive (non-overlapping), exhausting categories in which no order or ranking can be imposed on the data.	zip code (4, 5, 6,...), gender (female, male), eye colour (blue, brown, green, hazel), political affiliation, religious, affiliation, nationality, etc.
Ordinal-level	Classifies data into categories that can be ranked; however, any specific differences between the ranks do not exist.	grade (A, B, C, D, etc.), judging (first place, second place, etc.), rating scale (poor, good, excellent) etc.
Interval-level	Ranks the data, and precise differences between units of measure do exist; however, there is no meaningful zero.	IQ test temperature
Ratio-level	Possesses all the characteristics of interval measurement, and there exists a true zero.	height, weight, time, salary, age etc.

Graphical Statistics

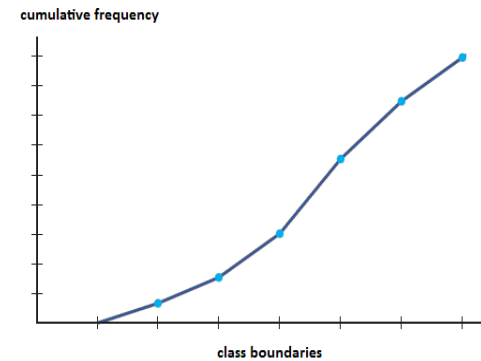
- The purpose of graphs in statistics is to convey the data to the viewer in pictorial form and getting the audience's attention in a publication or a presentation.



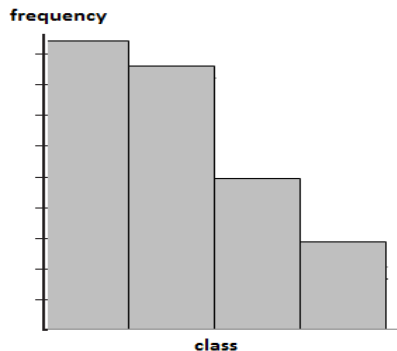
Histogram



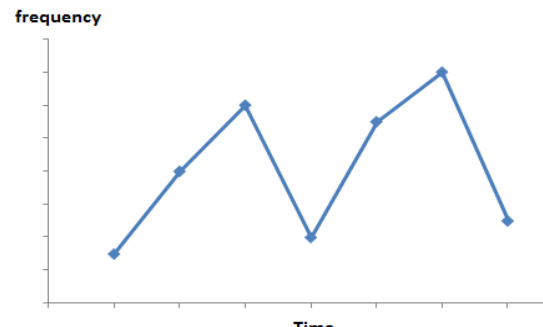
Frequency Polygon



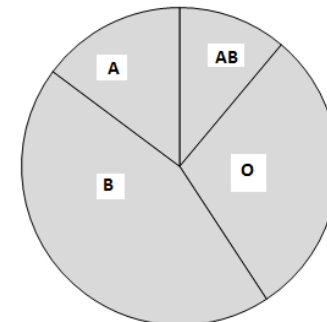
Ogive



Pareto Chart

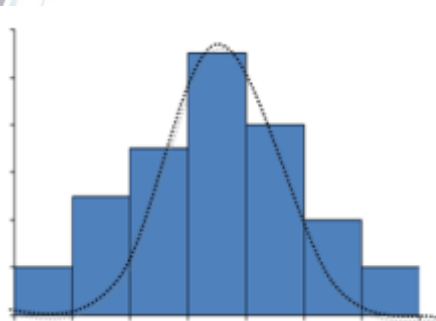


Time Series Graph

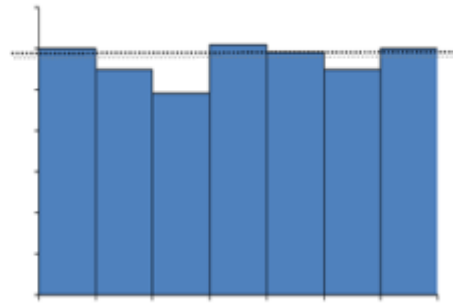


Pie Chart

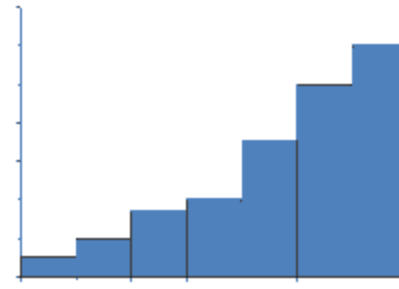
Distribution Shapes for Histogram



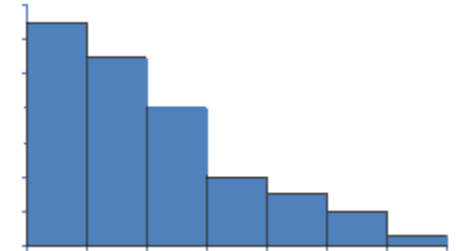
Bell-Shaped



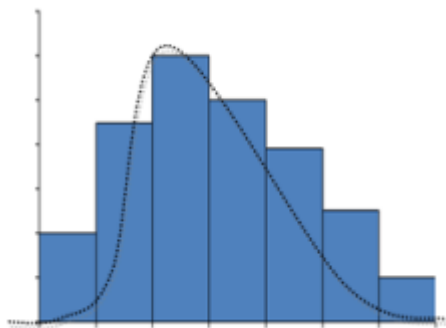
Uniformed



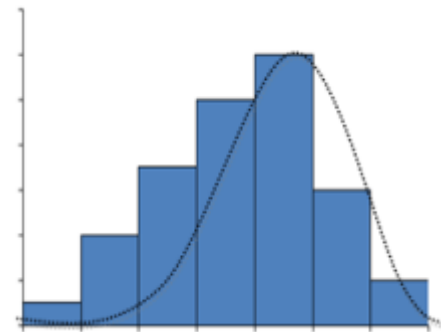
J-Shaped



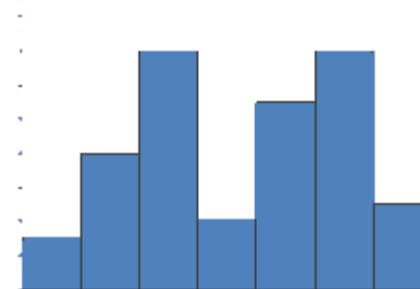
Reverse J-Shaped



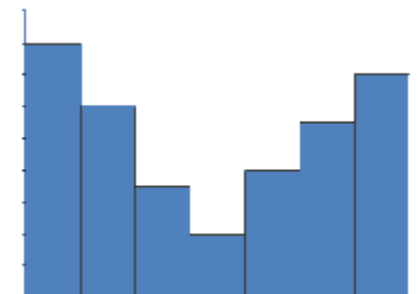
Right Skewed



Left Skewed



Bimodal



U-Shaped

1.3 REVIEWS ON DESCRIPTIVE STATISTICS

- ✓ We can summarise data using *measures of central tendency, measures of variation, and measures of position.*
- ✓ *Measures of central tendency (Measures of average): mean, median, mode, and midrange.*
- ✓ *Measures of variation (measures of dispersion/spread): range, variance, and standard deviation.*
- ✓ *Measures of position (tell where a specific data value falls within the data set or its relative position in comparison with other data values): percentiles, deciles, and quartiles.*

1.3.1 Measures of Central Tendency

Mean

the sum of the values divided by the total number of values.

Population Mean

Sample Mean

$$\mu = \frac{\sum_{i=1}^N x_i}{N}, \quad N \text{ population size}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad n \text{ sample size}$$

Example:

If the data set are 1, 6, 3, 7, 8, 5, then the calculated mean is $\mu = 5$ if it taken from the population and $\bar{x} = 5$ if it taken from the sample.

1.3.1 Measures of Central Tendency

Median

the middle number of n ordered data (smallest to largest)

If n is odd



$$\text{Median(MD)} = x_{\frac{n+1}{2}}$$

If n is even



$$\text{Median(MD)} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Example:

If the data set are 1, 3, 5, 6, 7, then the calculated median is, Median = $x_3 = 5$.

If the data set are 1, 3, 5, 6, 7, 9 then the calculated median is, Median = $\frac{x_3 + x_4}{2} = 5.5$.

1.3.1 Measures of Central Tendency

Mode : the most commonly occurring value in a data series

Example:

If the data set are 1, 6, 3, 7, 3, 8, 5, 3 then the mode is 3.

If the data set are 1, 6, 3, 7, 3, 8, 7, 5, 3, 7 then the mode is 3 and 7.

Midrange

is a rough estimate of the middle & also a very rough estimate of the average and can be affected by one extremely high or low value.



$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

Example:

If the data set are 1, 3, 5, 6, 7, 9 then the calculated midrange is, $MR = \frac{1+9}{2} = 5$.

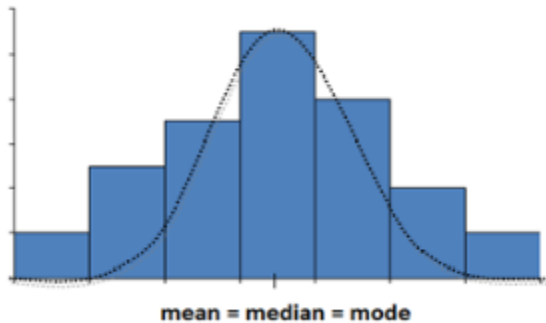
Properties of Mean, Median & Mode

- ✓ The mean is unique, and not necessarily one of the data values.
- ✓ The mean is affected by extremely high or low values and if it occurs, the mean may not be the appropriate average to use. As example, if an extreme value, let say 21 is added to the data set in previous example the new mean value is given by 7.3. This new average value is no longer representing the central of the data set.
- ✓ The mean cannot be computed for an open ended frequency distribution.
- ❑ The median is used when one must find the center or middle value of a data set and to determine whether the data values fall into the upper half or lower half of the distribution.
- ❑ The median is used to find the average of an open-ended distribution.
- ❑ The median is affected less than the mean by extremely high or extremely low values.
- ❖ The mode is used when the most typical case is desired.
- ❖ The mode can be used when the data are nominal, such as religious preference, gender, or political affiliation.
- ❖ The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

Identify Shapes of Data Distribution

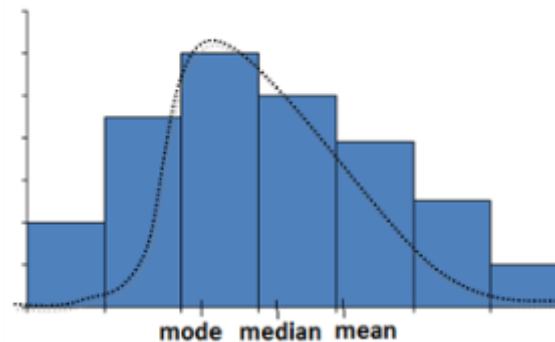
Symmetric

Mean = Median = Mode



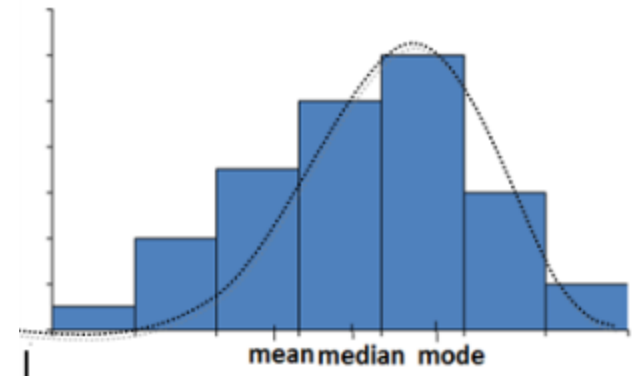
**Positively skewed /
right-skewed**

Mean > Median > Mode



**Negatively skewed /
left-skewed**

Mean < Median < Mode



1.3.2 Measures of Variation/Dispersion

- Used when the central of tendency does not give any meaning or not needed (ex: mean are same for two types of data)
 - ✓ If the mean of x and y are same and dispersion of x is less than dispersion of y , then population x is better than population y
- To measure the variability that exists in a data set
- To learn the extent of the scatter so that steps may be taken to control the existing variation

Range

is the different between the highest value and the lowest value in a data set.
The symbol R is used for the range.



$$R = \text{highest value} - \text{lowest value}$$

Example:

If the data set are 1, 3, 5, 6, 7, 9 then the calculated range is, $R = 9 - 1 = 8$

1.3.2 Measures of Variation/Dispersion

Variance : is the average of the squares of the distance each value is from the mean.

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}, \quad N \text{ population size}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}, \quad n \text{ sample size}$$

Standard Deviation: is the square root of the variance

Population standard deviation, σ

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}, \quad N \text{ population size}$$

Sample standard deviation, s

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}, \quad n \text{ sample size}$$

Example:

Suppose the data set are 1, 6, 3, 7, 8, 5, then the calculated variance is $\sigma^2 = 5.67$ and the standard deviation is $\sigma = 2.38$ if it taken from the population, while the calculated variance is $s^2 = 6.8$ and the standard deviation is $s = 2.61$ if it taken from the sample.

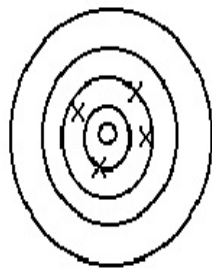
Properties of Variance & Standard Deviation

Smaller	→	More consistent
standard	→	Less dispersed
deviation	→	Less spread
$(\sigma_1 < \sigma_2)$	→	Less variable (small variation)
	→	More precise

Accuracy and Precision

Accuracy is how close a measured value to the 'true' measurements.

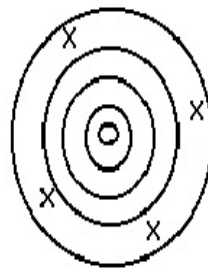
Precision is how close the measured value to each other or how consistent your results are for the same phenomena over several measurements.



Picture A



Picture B



Picture C

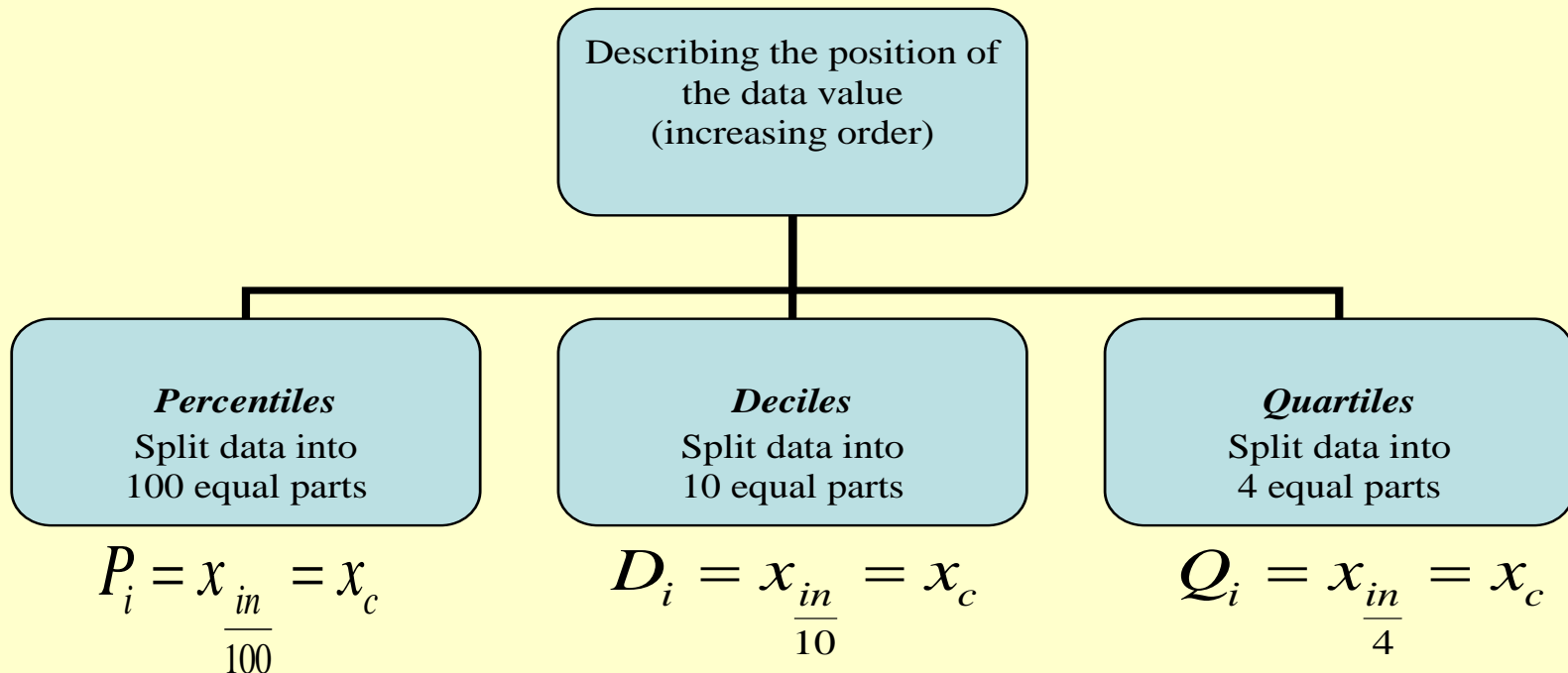


Picture D

- **Picture A** shows a very accurate (close to the mark), but not very precise, since the darts are spread out everywhere.
- **Picture B** shows an example of precision without accuracy (very consistent, but not near the mark).
- **Picture C** shows both inaccuracy and imprecision.
- **Picture D** shows both accuracy and precision.

1.3.3 Measures of Position

- Tell where a specific data value falls within the data set or its relative position in comparison with other data values



If c is not a whole number, **round it up** to the next whole number.

If c is a whole number, then use $Q_i = \frac{x_c + x_{c+1}}{2}$, $D_i = \frac{x_c + x_{c+1}}{2}$, $P_i = \frac{x_c + x_{c+1}}{2}$

EXAMPLE:

The dataset in increasing (ascending) order: 25 26 27 30 31 36 38 40 42 44 45

Quartiles	Percentiles
$Q_1 = x_{\frac{1(11)}{4}} = x_{2.75} \rightarrow x_3 = 27$	$P_{25} = x_{\frac{25(11)}{100}} = x_{2.75} \rightarrow x_3 = 27$
$Q_2 = x_{\frac{2(11)}{4}} = x_{5.50} \rightarrow x_6 = 36$	$P_{50} = x_{\frac{50(11)}{100}} = x_{5.50} \rightarrow x_6 = 36$
$Q_3 = x_{\frac{3(11)}{4}} = x_{8.25} \rightarrow x_9 = 42$	$P_{75} = x_{\frac{75(11)}{100}} = x_{8.25} \rightarrow x_9 = 42$

Summary: Q_1 equivalent to P_{25} ; Q_2 equivalent to P_{50} ; Q_3 equivalent to P_{75} .

EXAMPLE:

The dataset in increasing (ascending) order: 25 26 27 30 31 36 38 40 42 44 45

Deciles	Percentiles
$D_3 = x_{\frac{3(11)}{10}} = x_{3.3} \rightarrow x_4 = 30$	$P_{30} = x_{\frac{30(11)}{100}} = x_{3.3} \rightarrow x_4 = 30$
$D_5 = x_{\frac{5(11)}{10}} = x_{5.5} \rightarrow x_6 = 36$	$P_{50} = x_{\frac{50(11)}{100}} = x_{5.5} \rightarrow x_6 = 36$
$D_7 = x_{\frac{7(11)}{10}} = x_{7.7} \rightarrow x_8 = 40$	$P_{70} = x_{\frac{70(11)}{100}} = x_{7.7} \rightarrow x_8 = 40$

Summary: D_i equivalent to $P_{i(10)}$, where $i=1, 2, 3, 4, 5, 6, 7, 8, 9$.

1.4 EXPLORATORY DATA ANALYSIS

The purpose of **exploratory data analysis** is to examine data in order to find out what information could be discovered. For example:

- Are there any gaps in the data?
- Can any patterns be discerned?

Traditional Method	Exploratory Data Analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range (IQR)

1.4.1 Stem and Leaf Plots

- ❑ A data plot that uses part of a data value as the stem (the leading digit) and part of the data value as the leaf (the trailing digit) to form groups or classes.
- ❑ It retaining the actual data while showing them in graphic form.
- ❑ Arranging the data in order is not essential, but the stem must be arranged in order.
- ❑ We may use the key indicator to define the stem and leaf values. For example; value of 2.1 can be defined as 2|1 where it indicate 2 as the digit (stem) whereas 1 as the decimal (leaf). Therefore the decimal is not written in the stem and leaf plot.
- ❑ Sometime we can construct a mixture model.
- ❑ If the plot is **rotated in horizontal position**, we can see the shape of distribution.
- ❑ The shapes are similar as described by using histogram. Choose more than five stem for a better shape of distribution.

Stem	Leaf
0	2
1	3 1
2	0 3 5
3	1 2 2 2 2 3 5
4	3 4 4 5
5	1 2 7

Key:
3|1 → 3.1

Leaf	Stem	Leaf
3 2	0 5	
9	8 9	
4 1 0	0 6 2 4	
7 6	6 6 7 8 8	
	2 7 1 1 4 4	
9	6 7 8	
	8 2 4	
	8 9	
	9 2	

before exercise after exercise (mixture/ back to back)

Pulse rate 5|9 → 59

1.4.2 Outliers

- An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.
- Outliers can be the result of measurements or observational error.
- When a distribution is normal or bell-shaped, data values that are beyond three standard deviations of the mean can be considered suspected outliers.
- A data value, x is an outlier if

$$x < Q_1 - 1.5(Q_3 - Q_1) \text{ or } x > Q_3 + 1.5(Q_3 - Q_1)$$

Example:

Given 60, 67, 70, 75, 89, 93, 95, 97, 112, 114, 114, 122, 129, 182, 229


$$Q_1 = x_{\frac{1 \times 15}{4}} = x_{3.75} = x_4 = 75$$

$$Q_3 = x_{\frac{3 \times 15}{4}} = x_{11.25} = x_{12} = 122$$

$$Q_1 - 1.5(Q_3 - Q_1) = 75 - 1.5 \times 47 = 4.5$$

$$Q_3 + 1.5(Q_3 - Q_1) = 122 + 1.5 \times 47 = 192.5$$

So, outliers are less than 4.5 or greater than 192.5. Therefore the outliers are 182 thousands and 229 thousands



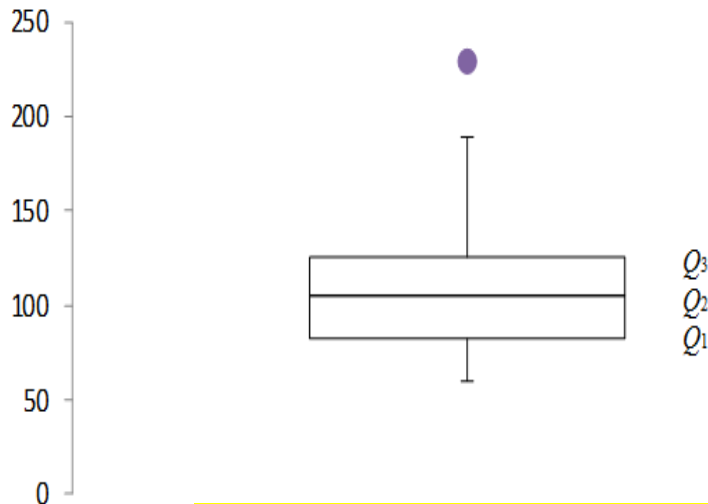
$$\text{IQR} = Q_3 - Q_1$$

1.4.3 Boxplots

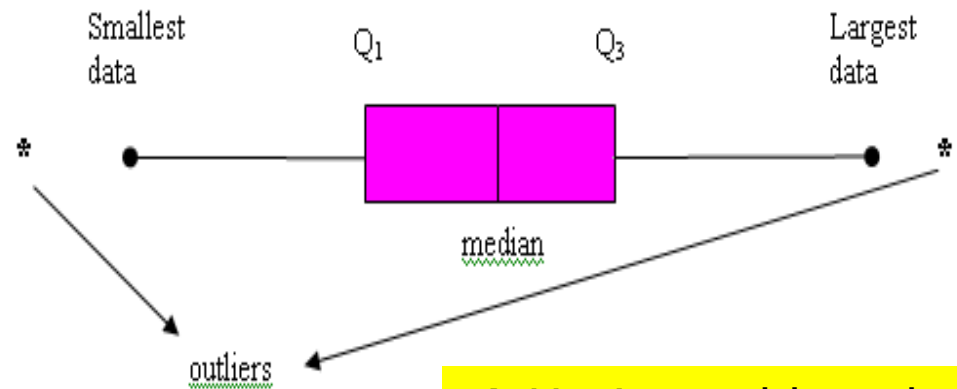
Boxplots (Box and Whiskers plot) are graphical representations of a **five-number summary** of a data set and outliers.

The **five-number summaries** are:

- The lowest value of data set (minimum)
- Q1 (1st Quartile or 25th percentile)
- The median (2nd Quartile or 50th percentile)
- Q3 (3rd Quartile or 75th percentile)
- The highest value of data set (maximum)



A Vertical boxplot



A Horizontal boxplot

STEP to Construct a Boxplot

- ✓ **STEP1** : Arrange the data
- ✓ **STEP2** : Find the Median
- ✓ **STEP3** : Find Q1 and Q3
- ✓ **STEP4** : Find Outliers

$$x < Q_1 - 1.5(Q_3 - Q_1) \text{ and } x > Q_3 + 1.5(Q_3 - Q_1)$$

- ✓ **STEP5** : Draw a scale for the data on the x axis.
- ✓ **STEP6** : Locate the lowest value, Q1, the median, Q3, the highest value and outliers on the scale.
- ✓ **STEP7** : Draw a box around Q1 and Q3, draw a vertical line through the median, and connect the upper and lower values

Information Obtain from a Boxplot

1. If the median is near the centre of the box, the distribution is approximately symmetric.
 2. If the median falls to the left of the centre of the box, the distribution is positively skewed.
 3. If the median falls to the right of the centre of the box, the distribution is negatively skewed.
 4. If the lines are about the same length, the distribution is approximately symmetric.
 5. If the right line is larger than the left line, the distribution is positively skewed.
 6. If the left line is larger than the right line, the distribution is negatively skewed.
 7. If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared using its central tendency and variability values.
 - ✓ To compare the central tendency measure, use the location of the medians.
 - ✓ To compare the variability, use the length of the interquartile range (IQR).
- ❑ For **symmetric** data, the appropriate measure of central tendency is **mean** and for variability is **standard deviation** or **variance**.
- ❑ For **skewed** data, the appropriate measure of central tendency is **median** and for variability is **interquartile range**.

EXAMPLE

The following mixture stem and leaf plot represent the sample ages of teachers in two schools.

School A	stem	School B
9 7 7 5 5 4	2	2
8 7 6 2 1 1 0	3	3 4 6 7
	4	0 1 3 4 5 7
	5	1 3 4

Given that for School B, $Q_1 = 36$, $Q_2 = 42$, $Q_3 = 47$ and there is no outlier. Draw Boxplots for both schools in the same x -axis. Then compare shapes, averages, and variability of both distributions of age.

SOLUTION, For School A:

$$Q_2 = \frac{x_7 + x_8}{2} = 30.5, \quad Q_1 = x_{\frac{1(14)}{4}} = x_{3.5} \rightarrow x_4 = 27, \quad \text{and} \quad Q_3 = x_{\frac{3(14)}{4}} = x_{10.5} \rightarrow x_{11} = 36$$

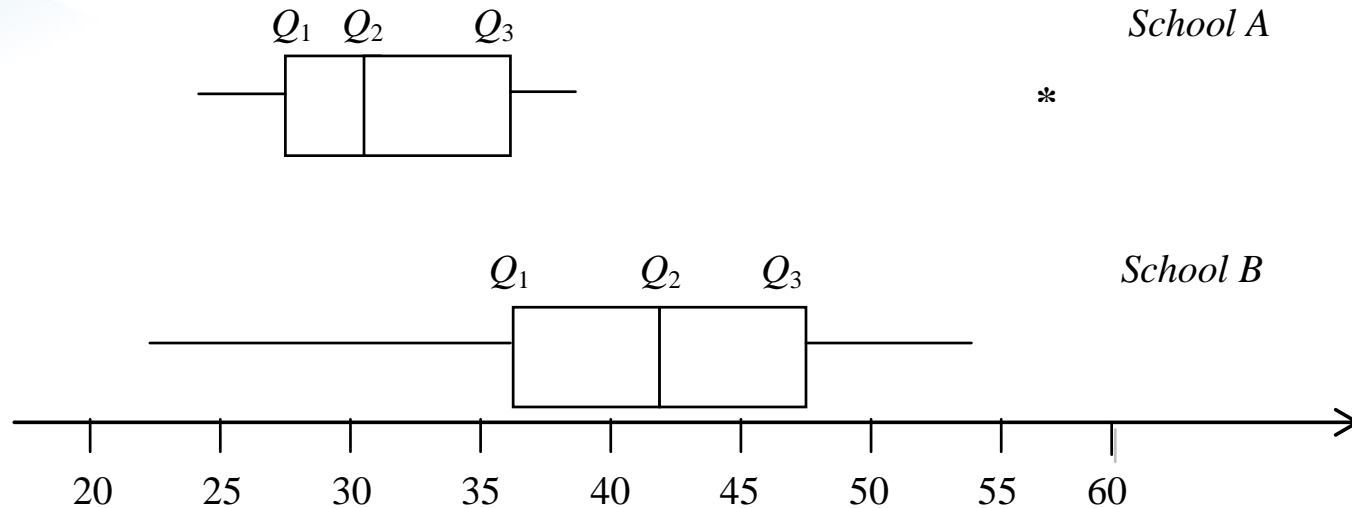
$$Q_1 - 1.5(Q_3 - Q_1) = 27 - 1.5(36 - 27) = 13.5$$

$$Q_3 + 1.5(Q_3 - Q_1) = 36 + 1.5(36 - 27) = 49.5$$

Since $57 > 49.5$, Thus 57 is an outlier.

EXAMPLE: Solution

For **School B**, $Q_1 = 36$, $Q_2 = 42$, $Q_3 = 47$ and there is no outlier.



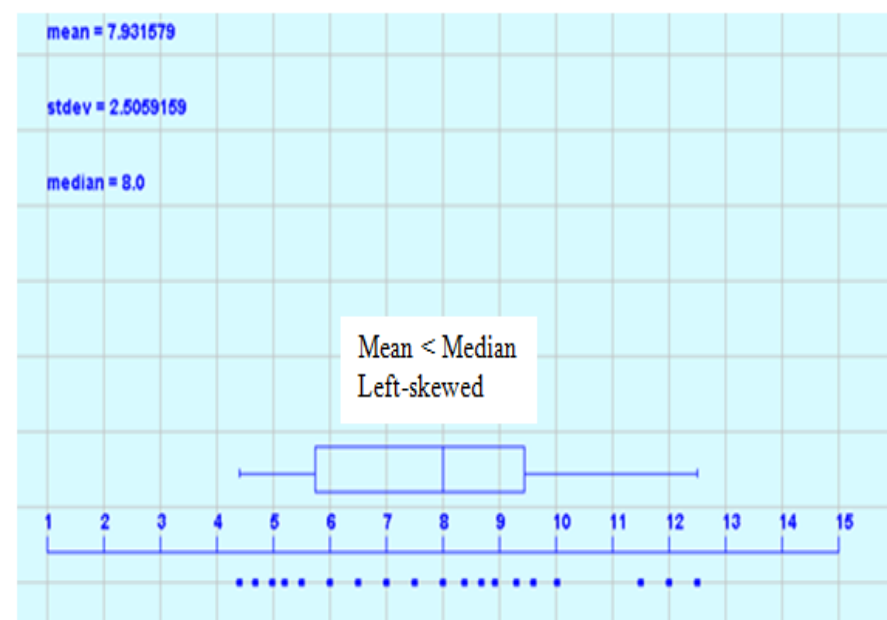
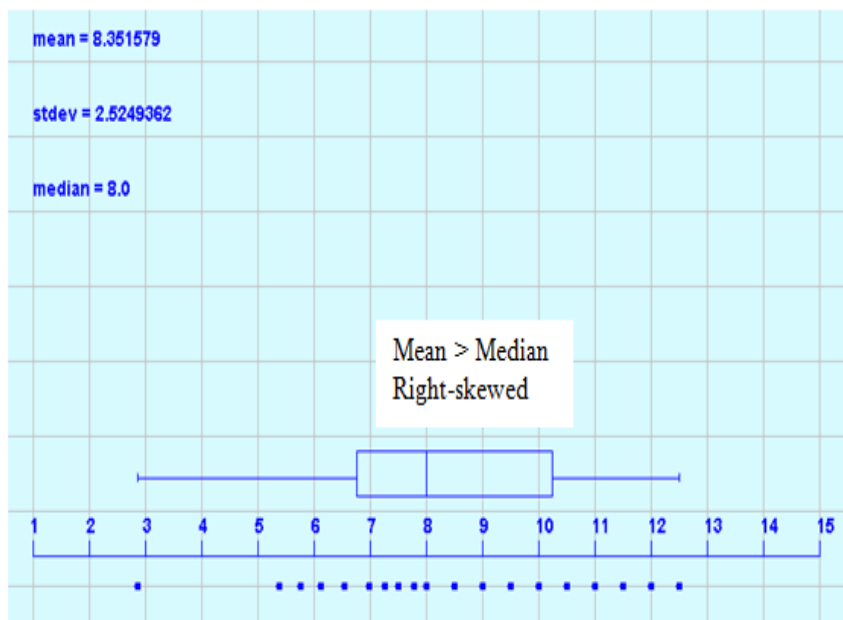
By comparing its shapes, School A has positively skewed distribution, while School B has negatively skewed distribution.

By comparing its averages, School B has higher median compared to School A.

By comparing its variability, School B is more variable compared to School A.

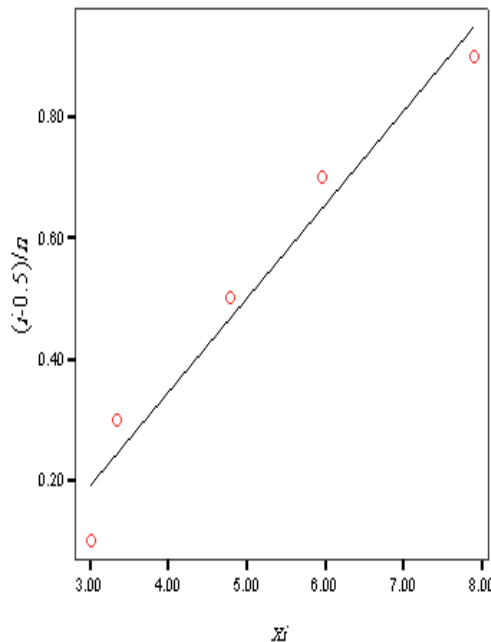
Boxplot for Special Case

- In some cases, we cannot use the general guideline as given above to interpret the boxplot.
- Boxplot is not the best graphical representation to describe a data set if the sample size of the data set is too small.
- The existence of outliers also may affect the boxplot.
- Therefore, in such cases, we have to use the descriptive statistics to identify the distribution of the data set.



1.5 NORMAL PROBABILITY PLOTS

- ✓ To determine whether the sample might have come from a normal population or not.
- ✓ The most plausible normal distribution is the one whose mean and standard deviation are the same as the sample mean and standard deviation.



STEP 1 : Sort the data in ascending order and denote each sorted data as

$$x_i, i = 1, \dots, n.$$

STEP 2 : Numbered the sorted data from i to n .

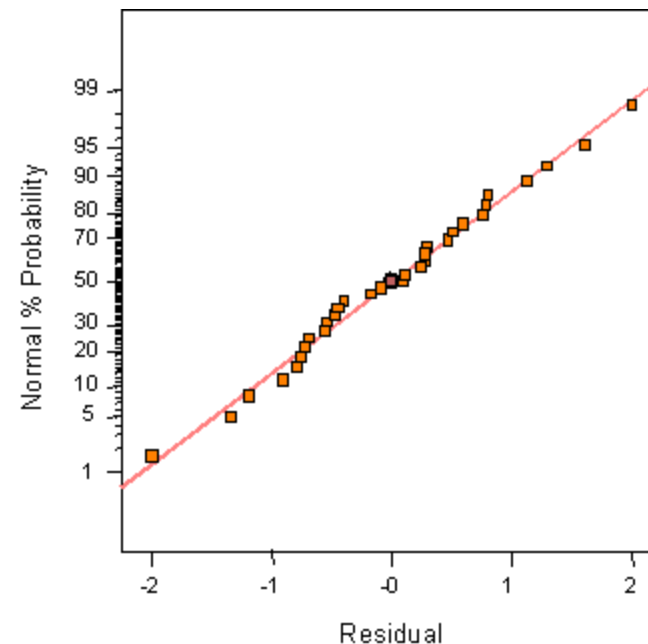
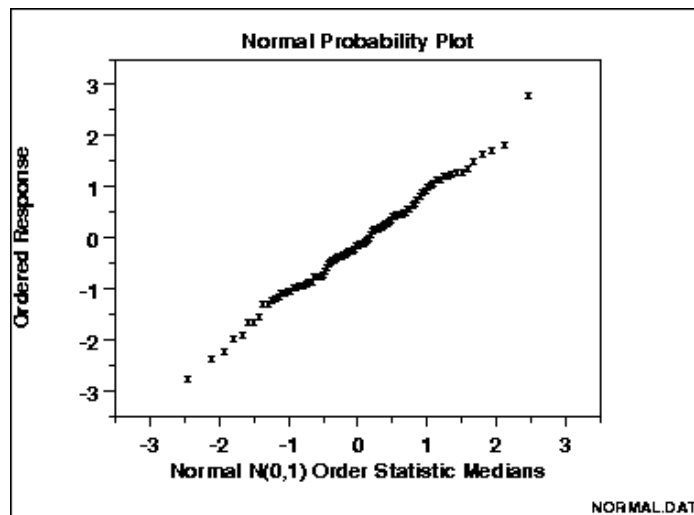
STEP 3 : Calculate the probability value for each x_i using $p_i = \frac{i-0.5}{n}$.

STEP 4 : Plot p_i versus x_i .

If the sample points lie approximately on a straight line, the data is approximately normally distributed.

Testing Normality using Software

- ✓ Other than plot manually, we can obtain it from software such as SPSS, Minitab, Excel, and etc. The normality of the data can be tested by using Kolmogorov Smirnov and Anderson Darling for non-parametric test.



REFERENCES

1. Walpole R.E., Myers R.H., Myers S.L. & Ye K. 2011. *Probability and Statistics for Engineers and Scientists*. 9th Edition. New Jersey: Prentice Hall.
2. Navidi W. 2011. *Statistics for Engineers and Scientists*. 3rd Edition. New York: McGraw-Hill.
3. Triola, M.F. 2006. *Elementary Statistics*. 10th Edition. UK: Pearson Education.
4. Bluman A.G. 2009. *Elementary Statistics: A Step by Step Approach*. 7th Edition. New York: McGraw–Hill.
5. Weiss, N.A. 2002. *Introductory Statistics*. 6th Edition. United States: Addison-Wesley.
6. Sanders D.H. & Smidth R.K. 2000. *Statistics: A First Course*. 6th Edition. New York: McGraw-Hill.
7. Crawshaw, J. & Chambers, J. 2001. *A Concise Course in Advance Level Statistics with Work Examples*, 4th Edition, Nelson Thornes.
8. Satari S. Z. et al. *Applied Statistics Module New Version*. 2015. Penerbit UMP. Internal used.

Thank You

NEXT: Chapter 2 Sampling Distribution and Confidence Interval