

DESCRIPTIVE STATISTICS

1. Definition:

- Statistics
 - A field of study concerned with the collection, organization and summarization of data.
- Statistical Methods
 - A scientific technique employed for collection, presentation, analysis and interpretation of data.
- Biostatistics
 - Biological field and medicine

2. Uses of statistical methods

- To collect data in the best possible way
 - Designing form
 - Organizing
 - Conducting survey
- To describe the characteristics of a group or a situation
 - Data summary
 - Data presentation
- To analyses data and to withdraw conclusion
 - Scientific, logic
 - Decision making

3. Classification of statistics

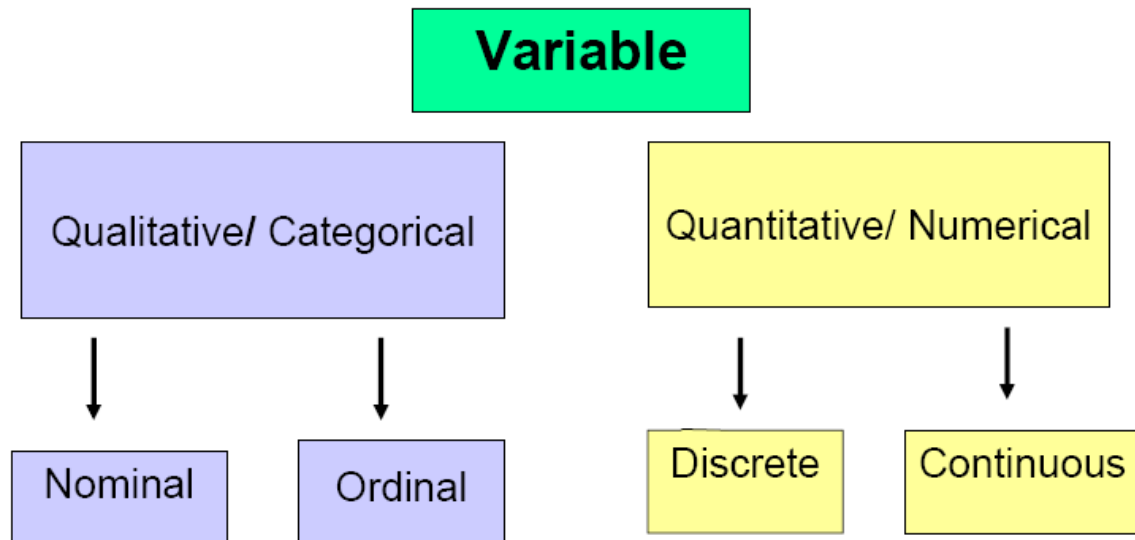
- Descriptive statistics
 - Concerned with collection, organization, enumeration of frequency of characteristics, summarization and presentation of data.

- Describe characteristics of the observed data
 - Type of variable
 - Summary statistics
 - Distribution
 - Graphical presentation
- Inferential statistics
 - Analytical in nature
 - Involve hypothesis testing and confidence interval
 - Allows researcher to infer/ generalize the characteristics of the sample (statistic) to the population (parameter)

4. Terms:

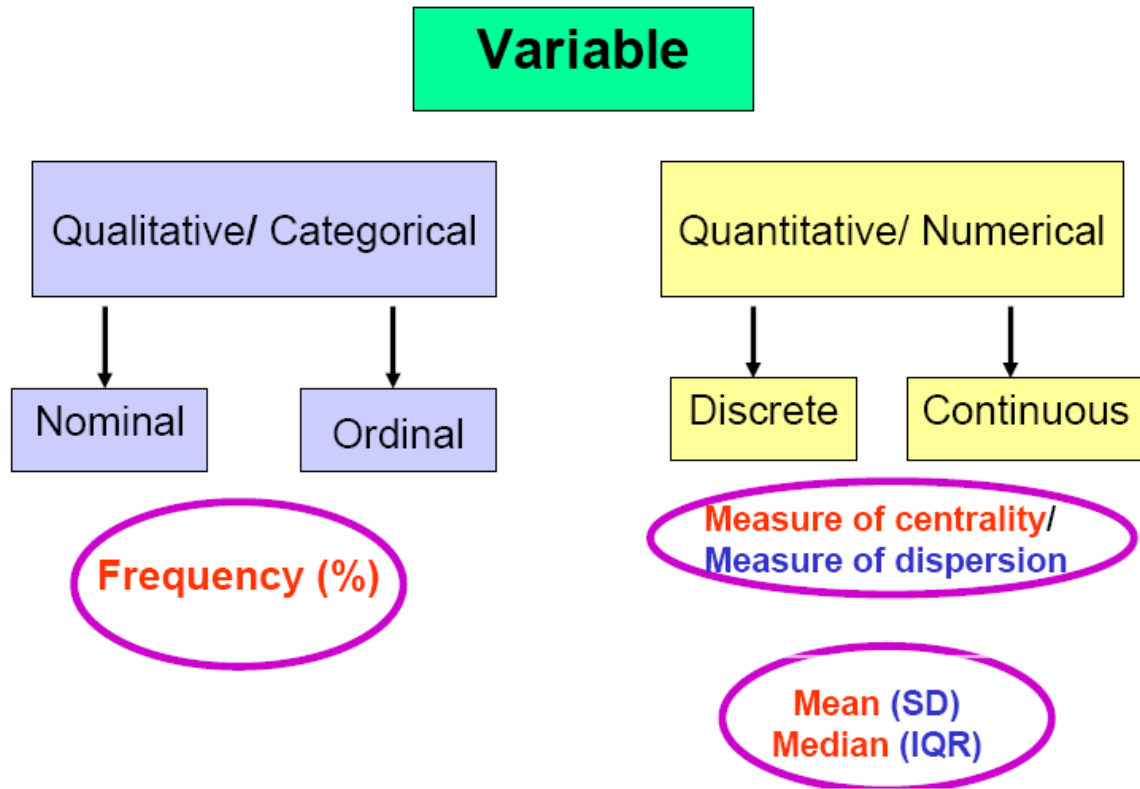
- Population:
 - Full sets of individuals
 - Collection of items objects, things, people
 - Parameter – descriptive measure from population data
- Sample
 - Subset of population
 - Selected to represent the population by sampling technique
 - Statistics – descriptive measure from sample data
- Variable
 - Any characteristics of even/object/person
 - The characteristics being observed/measured
 - E.g. age sex, race, height, weight, etc
- Data
 - The raw or original measurement of statistics
 - Values taken by the characteristics
 - E.g. Malay, female, 155cm

5. Classification of variables



- Discrete
 - Characteristics by gaps or interruptions in the values
 - Values that can be assume only whole numbers
 - Mainly count
 - E.g. no of students, no of teeth extracted
- Continuous
 - No gap or interruption
 - Any value within specified interval
 - Mainly measurement
 - E.g. height, weight, BP, age, etc
- Nominal
 - Unordered categories
 - No implied order among the categories
 - E.g. race, sex, medical diagnosis, etc.
- Ordinal
 - Ordered categories
 - Ranked according to some criteria

- E.g. BP – high, normal, low.



6. Categorical Variables:

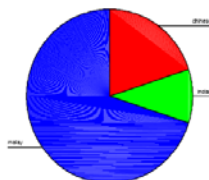
- Data presentation
 - Statistics
 - Frequency
 - Percentage (%)
 - Graphical
 - Pie chart
 - Bar chart

Table 1 Characteristics of respondents

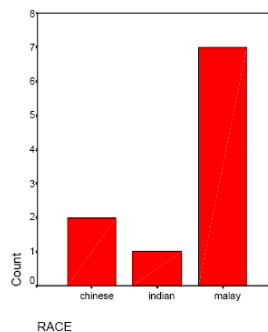
Variables	Frequency (%)
Gender	
Male	30(60)
Female	20(40)
Race	
Malay	33(66)
Chinese	11(22)
Indian	6(12)
IHD	
No	38(76)
Yes	12(24)
BMI group	
Underweight	5(10)
Normal	21(42)
Overweight	24(48)

PIE CHART

- Qualitative data
- Circle represents 360°
- Started at 12 o'clock position & clock wise direction
- Areas represents total frequency / percentages
- Each segment – each category



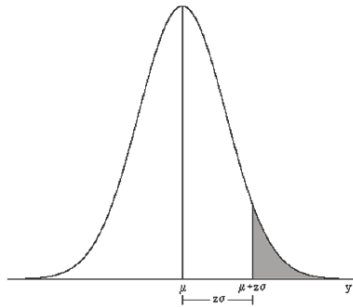
BAR CHART



- Qualitative data
- Equal width
- Bars are separated by an equal space

7. Numerical variables

- Measures of central tendency
 - A measure of centrality



- Mean
 - Arithmetic average
 - Adding all the values in a population/sample and divided by the number of values that are added
 - Affected by the extreme value

$\bar{x} = \sum \frac{x_i}{n}$	←	Sample mean
$\mu = \sum \frac{X_i}{N}$	←	Population mean

- Median
 - The middle value of data ordered from the lowest to the highest → arrange all value in order
 - If $n =$ odd number → median is the middle
 - If $n =$ even number → median is the mean of 2 middle observation
 - 50th percentile of a set of observation
 - The middle value of data ordered from lowest to the highest value

- Useful for data with non-normal distribution or skewed data
- Less sensitive to extreme values than the mean
- Median (IQR)
- Mode
 - The most frequent observation
 - Point of maximum concentration
- Measures of dispersion/variability
 - Range = largest value – smallest value
 - Different between the largest and smallest value in a set of observations
 - Give idea about the variability of data
 - Simplest to compute
 - Sensitive to outliers
 - Least useful
 - $R = X_{\max} - X_{\min}$
 - Variance = s^2
 - Total squares of deviation of observations from the mean/number of degree of freedom
 - Average measure of standard deviation of observation from mean sample
 - Measures the amount of variability or spread about/from the mean of a sample
 - $S^2 = \Sigma(x_i - x_{\text{mean}})^2 / n - 1$
 - Standard deviation (SD)
 - A square root of variance
 - The root mean square of the distances (or differences) from mean of sample
 - A better measure of variability of a set of data
 - Smaller SD indicates closer to the mean

- Mean (SD)
- $S = \sqrt{[\sum(x_i - x_{\text{mean}})^2 / n - 1]}$

$$= \sqrt{\frac{\sum 18}{10-1}}$$

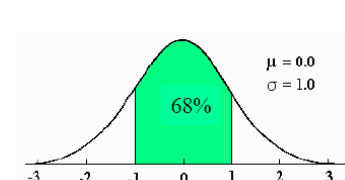
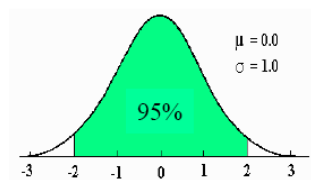
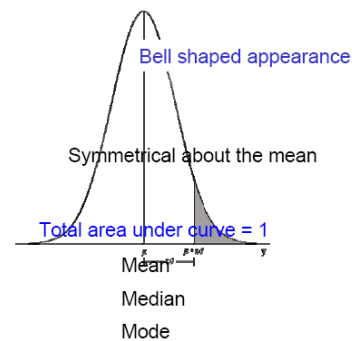
$$= \sqrt{2.0}$$

$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
6-6 = 0	(6-6) = 0
5-6 = -1	(5-6) = 1
8-6 = 2	(8-6) = 4
6-6 = 0	(6-6) = 0
4-6 = -2	(4-6) = 4
6-6 = 0	(6-6) = 0
7-6 = 1	(7-6) = 1
4-6 = -2	(4-6) = 4
6-6 = 0	(6-6) = 0
8-6 = 2	(8-6) = 4
$\sum (x_i - \bar{x})$ = 0	$\sum (x_i - \bar{x})^2$ = 18 ₃₆

- Interquartile range (IQR)
 - $Q_3 - Q_1$
 - Range between 25th and 75th percentile
 - Used along with the median
 - It not affected by outlier
- Percentile = 25th, 50th, 75th, 90th, 95th

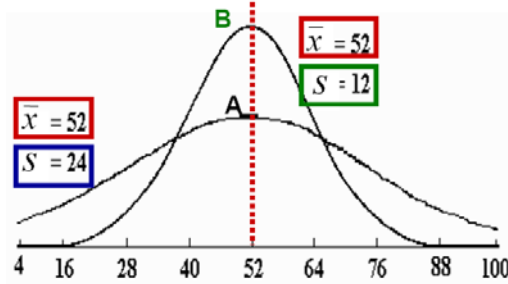
8. Normal distribution

- Characteristic
 - Bell shaped appearance
 - Symmetrical about the mean
 - Mean = median = mode
 - Total area the curve = 1
 - The curve never touch the x line
 - SD usually less than 30% of mean value
- Approximately
 - 68% → 1 SD
 - 95% → 2 SD
 - 99.7% → 3SD

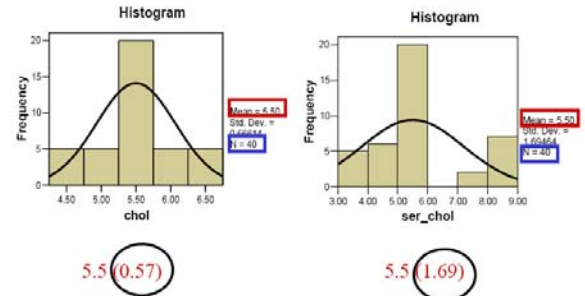


- Mean (SD)

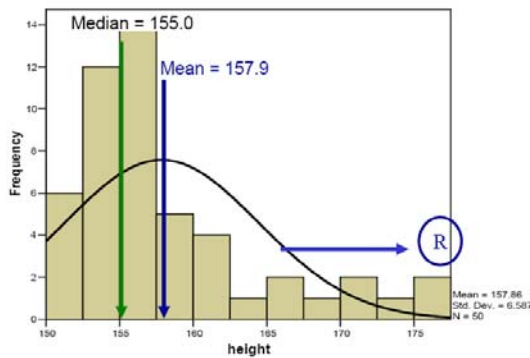
Measures of central tendency
Different standard deviations



MEAN (SD)



POSITIVE SKEWNESS :
Skew to the **right**



NEGATIVE SKEWNESS :
Skew to the **left**

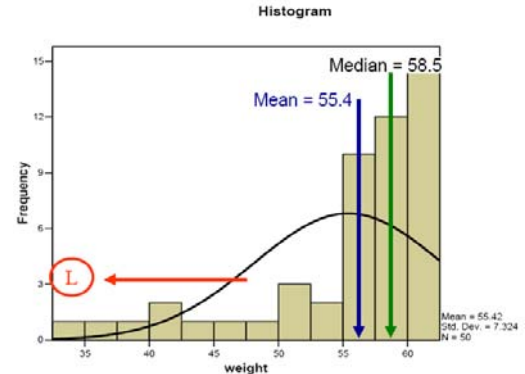


Table 2 Characteristics of the respondents

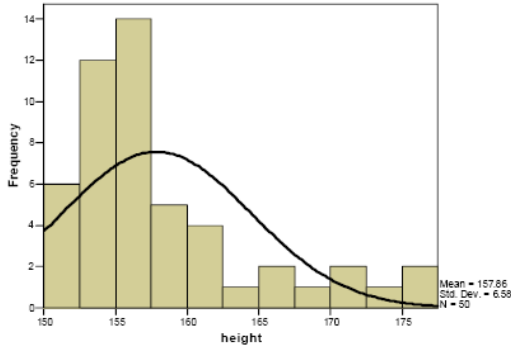
9. Data presentation (numerical data)

- Statistics
 - Mean (SD)
 - Medical (IQR)
- Graphical
 - Histogram
 - Frequency distribution of quantitative data/continuous data
 - Bars represent frequency distribution for each class of interval
 - No spaces between bars

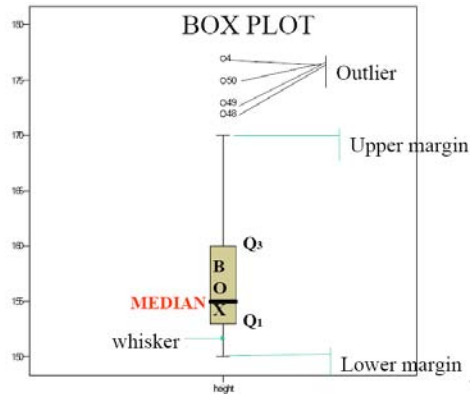
Variables	Mean (SD)	Median (IQR)
Age (years)	57.6 (12.1)	
Weight (kg)		58.5 (6.0)*
Height (m)		155.0 (7.0)**
BMI		22.9(4.1)*

*Skewed to the left
**Skewed to the right

- May have equal/unequal class interval
- Box plot



Histogram



10. Summary

- Categorical data
 - Statistics
 - Frequency (%)
 - Graphs
 - Bar chart
 - Pie chart
- Numerical data
 - Statistics
 - Mean (SD)
 - Median (IQR)
 - Graphs
 - Histogram
 - Box Plot

Table 3 Characteristics of the respondents

Variables	Frequency (%)	Mean (SD)	Median (IQR)
Age (years)		57.6 (12.1)	
Gender			
Male	30(60)		
Female	20(40)		
Race			
Malay	33(66)		
Chinese	11(22)		
Indian	6(12)		
IHD			
No	38(76)		
Yes	12(24)		
Weight (kg)			58.5 (6.0)*
Height (m)			155.0 (7.0)**
BMI	22.9(4.1)		22.9(4.1)*
BMI_gp	5(10)		
	21(42)		
	24(48)		

*Skewed to the left

**Skewed to the right