Description of Data I

(Summary and Variability measures)

Objectives:

- **Able to understand how to summarize the data**
- **Able to understand how to measure the variability of the data**
- **Able to use and interpret appropriately the different summary and variability measures**

Team Members: Jawaher Abanumy - Weam Babaier

Team Leaders: Mohammed ALYousef & Rawan Alwadee

Revised By: Maha Alghamdi

Doctor: Dr. Shaffi Ahmed

Resources:

• 436 Lecture Slides + Notes

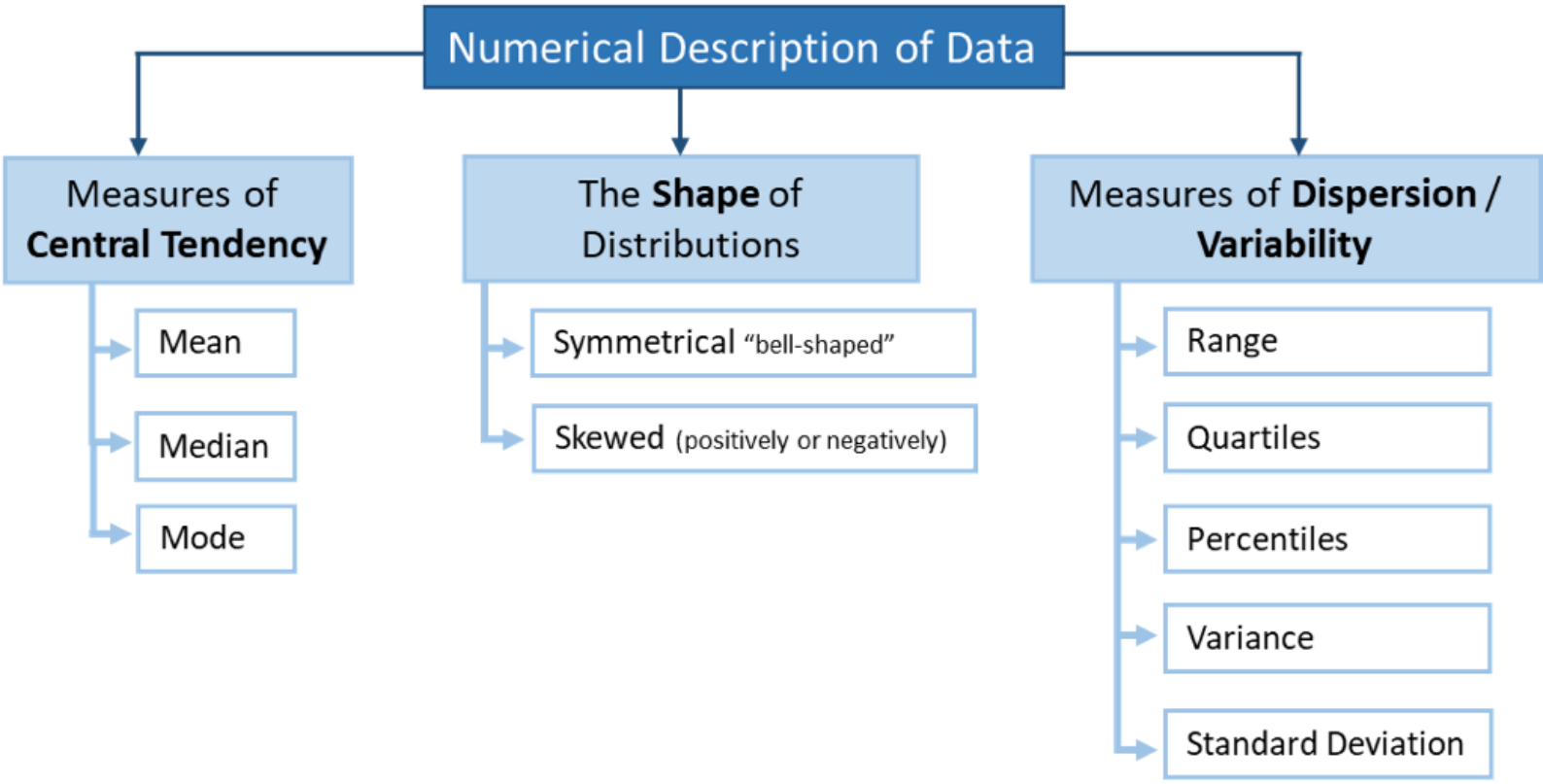
Important – Notes

436researchteam@gmail.com

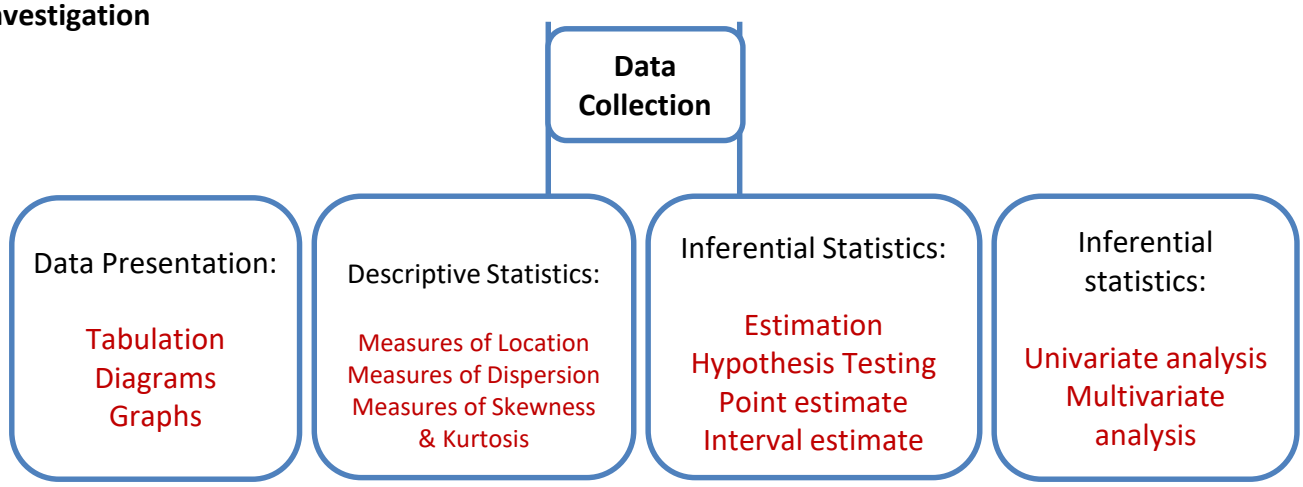
Editing file

Feedback link

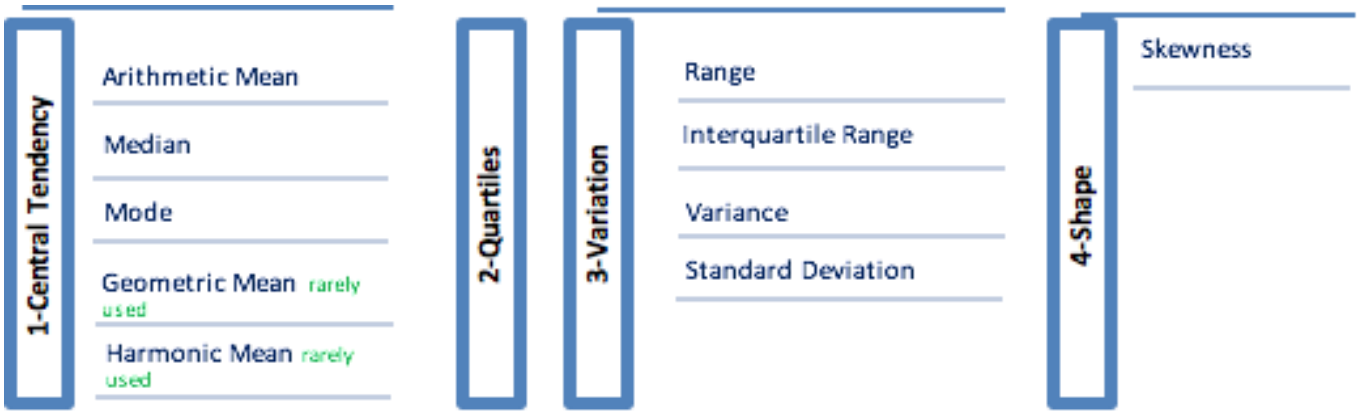
Lecture outline:



Investigation



Summary & Variability Measures



Measures of Central Tendency

- A statistical measure that identifies a single score as representative for an entire distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group
- There are three common measures of central tendency:

1. the mean Sum/total (Average)
2. the median Middle number of all data
3. the mode Most frequent value

Calculating the Mean:

- Calculate the mean of the following data:
  1 5 4 3 2

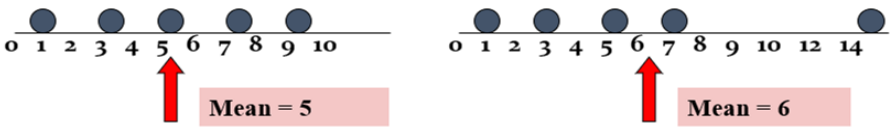
Sum the scores (ΣX):

1 + 5 + 4 + 3 + 2 = 15

I. Divide the sum (ΣX = 15) by the number of scores (N = 5): 15 / 5 = 3
II. Mean Affected by extreme value = $\underline{X}$ = 3

Mean (Arithmetic Mean)

- The most common measure of central tendency
- Affected by extreme values (outliers) *extreme value mean for example is the student mark is between 10-12 but there is 4 student get full mark or the oboset there is 3 student get zero in the exam *



The Median: Q2 is the other name

- The median is simply another name for the 50th percentile
- It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

How To Calculate the Median

- Conceptually, it is easy to calculate the median
- Sort the data from highest to lowest
- Find the score in the middle
  - middle = (N + 1) / 2
  - If N, the number of scores is even, the median is the average of the middle two scores

Median Example

- What is the median of the following scores:
  24 18 19 42 16 12

I. Sort the scores:
   42 24 19 18 16 12
II. Determine the middle score:
   middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5
III. Median = average of 3rd and 4th scores:
   (19 + 18) / 2 = 18.5

-first we arrange the number from high to low
-N= sample size
-here the median because the sample size is double (even) we take the average of the middle two numbers

- What is the median of the following scores:
  10 8 14 15 7 3 3 8 12 10 9

I. Sort the scores:
   15 14 12 10 10 9 8 8 7 3 3
II. Determine the middle score:
   middle = (N + 1) / 2 = (11 + 1) / 2 = 6
III. Middle score = median = 9

the only difference here is that the sample size is single (odd) so after we rearrange them we take the middle one

Median

- Not affected by extreme values

- In an ordered array, the median is the "middle" number

- If n or N is odd, the median is the middle number
- If n or N is even, the median is the average of the two middle numbers (example if n=42 then the median is the average of the 21st and 22nd values)

Measures of Central Tendency

Mean … the most frequently used but is sensitive to extreme scores

e.g. 1 2 3 4 5 6 7 8 9 10
Mean = 5.5 (median = 5.5)

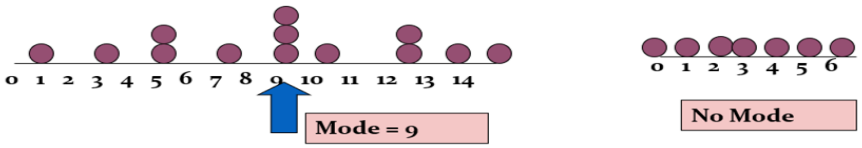
e.g. 1 2 3 4 5 6 7 8 9 20
Mean = 6.5 (median = 5.5)

e.g. 1 2 3 4 5 6 7 8 9 100
Mean = 14.5 (median = 5.5)

here we notice that the mean is changing but the median is constant why? because the mean get affected by extreme value in eg 2 the extreme value is 20 and in eg 3 is 100

Mode

- Value that occurs most often القيمة الأكثر تكرارا
- Not affected by extreme values
- Used for either numerical or categorical(nominal)*data
- There may be no mode if the sample size is low for eg when you take the mark for 3 student only there will be no repetid value
- There may be several modes



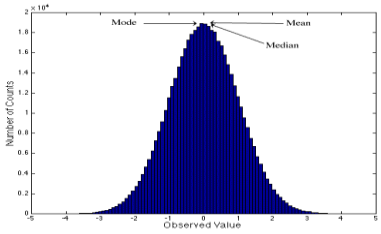
* For example: if we want to count number of people who smoke.

The Shape of Distributions

Distributions can be either **symmetrical or skewed (**Narrow Space**)**, depending on whether there are more frequencies at one end of the distribution than the other.

Symmetrical Distributions

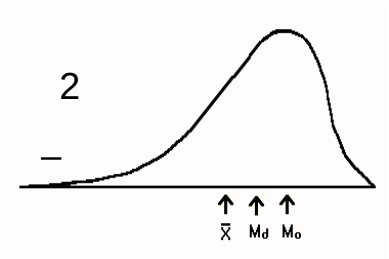
- A distribution is symmetrical if the frequencies at the right and left tails of the distribution are identical, so that if it is divided into two halves, each will be the mirror image of the other SO that mean when you take any point the mean ,median,mode will be the same
- In a symmetrical (normal) distribution the mean, median, and mode are identical.

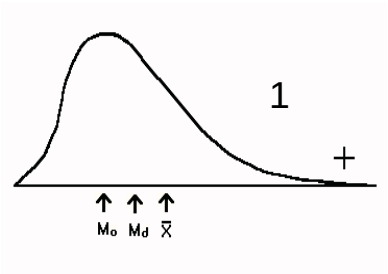


Distributions :(MCQ)

- Bell-Shaped (also known as symmetric" or "normal")
- Skewed: (تضييق)

- negatively (skewed to the left) – it tails off toward smaller values
- positively (skewed to the right to + values) – it tails off toward larger values





*as you go to the right the number increase

pic1:for example the mark of medicine only 5% of the student will get A+ but the majority will be less than that ,the PIC reflet the majority of student in the left(less number)and the student who get high mark on the right
pic2 :for example the student mark in research course the majority of the student get A A+ so they will be on the right of the chart

Skewed Distribution

Median is not effected

Few extreme values on one side of the distribution or on the other.

- Positively skewed distributions: distributions which have few extremely high values (Mean>Median)
- Negatively skewed distributions: distributions which have few extremely low values(Mean<Median)

Choosing a Measure of Central tendency:(very IMP IN MCQ)

- IF variable is Nominal..choose Mode
- IF variable is Ordinal...choose Mode or Median(or both)
- IF variable is Interval-Ratio and distribution is Symmetrical...choose Mode, Median or Mean
- IF variable is Interval-Ratio and distribution is Skewed...choose Mode or Median

Example:

(1) 7,8,9,10,11 n=5, Σ x=45, $\overline{x}$ =45/5=9

(2) 3,4,9,12,15 n=5, Σ x=45, $\overline{x}$ =45/5=9

(3) 1,5,9,13,17 n=5, Σ x=45, $\overline{x}$ =45/5=9

S.D. : (1) 1.58 (2) 4.74 (3) 6.32

Measures of Dispersion Or Measures of variability

Or measures of heterogeneity

Measures of Dispersion

Measures of dispersion summarize differences in the data, how the numbers differ from one another.

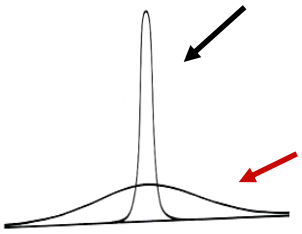
Series I: 70 70 70 70 70 70 70 70 70 70 No variability

Series II: 66 67 68 69 70 70 71 72 73 74 Small variability

Series III: 1 19 50 60 70 80 90 100 110 120 High variability

Measures of Variability

A single summary figure that describes the spread of observations within a distribution.



In this figure, both curves have the same median, but curve 2 (red arrow) has greater variance

Measures of Variability

- Range
  Difference between the smallest and largest observations. The highest mark is 15 and the lowest is 10 the range will be 15-10=5
- Interquartile Range
  Range of the middle half of scores.
- Variance
  Mean of all squared deviations from the mean.
- Standard Deviation
  Rough measure of the average amount by which observations deviate from the mean. The square root of the variance.

Variability Example: Range

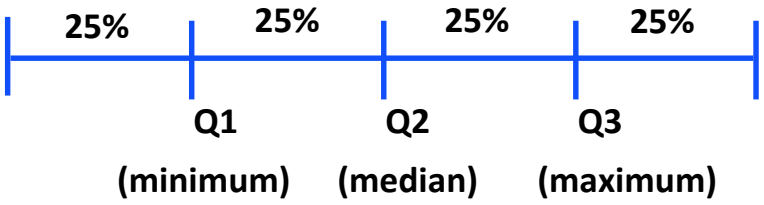
Marks of students
52, 76, 100, 36, 86, 96, 20, 15, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89

Range: 100-15 = 85

Quartiles: [Useful video for better understanding]

Q_1, Q_2, Q_3

divides ranked scores into four equal parts



$$Q1=\frac{n+1}{4}\ \text{th}$$

$$Q2=\frac{2(n+1)}{4}=\frac{n+1}{2}\ \text{th}$$

$$Q=\frac{3(n+1)}{4}\text{th}$$

DON'T MEMORIZE IT

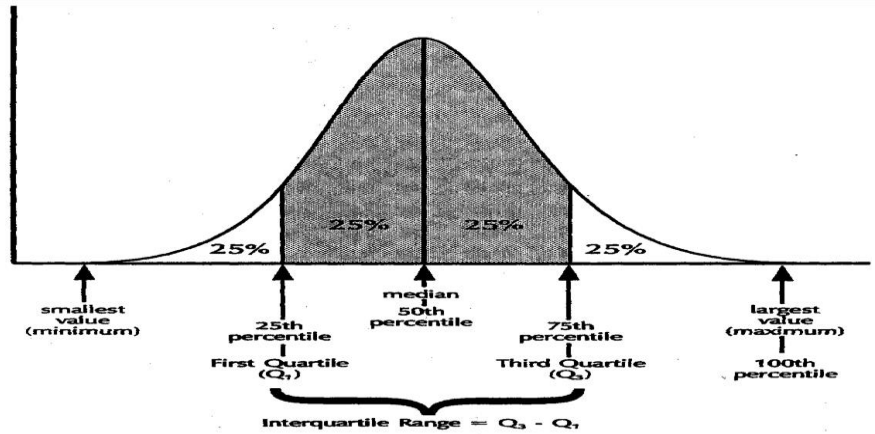
Inter quartile :

$$IQR = Q_3 - Q_1$$

Inter quartile Range:

- The inter quartile range is Q3-Q1
- 50% of the observations in the distribution are in the inter quartile range.
- The following figure shows the interaction between the quartiles, the median and the inter quartile range.

Inter quartile Range



Percentiles and Quartiles: [useful video for better understanding]

- Maximum is 100th percentile: 100% of values lie at or below the maximum
- Median is 50th percentile: 50% of values lie at or below the median
- Any percentile can be calculated. But the most common are 25th (1st Quartile) and 75th (3rd Quartile)

Locating Percentiles in a Frequency Distribution

- A percentile is a score below which a specific percentage of the distribution falls(the median is the 50th percentile.
- The 75th percentile is a score below which 75% of the cases fall.
- The median is the 50th percentile: 50% of the cases fall below it
- Another type of percentile :The quartile lower quartile is 25th percentile and the upper quartile is the 75th percentile

NUMBER OF CHILDREN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	260	26.6	26.6	26.6
	1	161	16.4	16.5	43.1
	2	260	26.6	26.6	69.7
	3	155	15.8	15.9	85.6
	4	70	7.2	7.2	92.7
	5	31	3.2	3.2	95.9
	6	21	2.1	2.1	98.1
	7	11	1.1	1.1	99.2
	EIGHT OR MORE	8	.8	.8	100.0
	Total	977	99.8	100.0	
Missing	NA	2	.2		
Total		979	100.0		

25th percentile → 0 (25% included here)
50th percentile → 2 (50% Included here)
80th percentile → 3 (80% Included here)

Frequency: number of families with the specific number of children. E.g. 161 families have one child
Valid percent: percent when missing data are excluded from calculation. So in this schedule valid percent is calculated after excluding 2 missing data
Cumulative percent: the valid percent + previous valid percent's

Variance: (VERY IMP)

Deviations of each observation from the mean, then averaging the sum of squares of these deviations.

- STANDARD DEVIATION

" ROOT- MEANS-SQUARE-DEVIATIONS"

in exam if they give you the variance=25 and the want the standard deviation SD= $\sqrt{25}$ =5

Standard Deviation

- To "undo" the squaring of difference scores, take the square root of the variance. (in the question if they give you the variance you have to know how to calculate the standard deviation)
- Return to original units rather than squared units.

Quantifying Uncertainty

Standard deviation: measures the variation of a variable in the sample.

-Technically,

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

Example

Data: X = {6, 10, 5, 4, 9, 8} There is variability; N = 6

X	$X - \bar{X}$	$(X - \bar{X})^2$
6	6-7=-1	1
10	10-7=3	9
5	5-7=-2	4
4	4-7=-3	9
9	9-7=2	4
8	8-7=1	1
Total42:	ZERO	Total: 28

Interpretation: All 6 values on average are deviating by 2.16. On average each student is different from other by 2.16.

Mean:

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

Variance:

$$s^2 = \frac{\sum(\bar{X} - X)^2}{N} = \frac{28}{6} = 4.67$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

Calculation of Variance & Standard deviation

- Using the deviation & computational method to calculate the variance and standard deviation.
  Example: 3,4,4,4,6,7,7,8,8,9 ; Given n=10; Sum= 60; Mean = 6

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n}}$$

$$S = \sqrt{\frac{(3-6)^2 + (4-6)^2 + (4-6)^2 + (4-6)^2 + (6-6)^2 + (7-6)^2 + (7-6)^2 + (8-6)^2 + (8-6)^2 + (9-6)^2}{10}}$$

$$S = \sqrt{\frac{40}{10}} = 2.0; \text{var}\,iance = 4$$

X	X²
3	9
4	16
4	16
4	16
6	36
7	49
7	49
8	64
8	64
9	81
Sum: 60	**Sum: 400**

$$S = \sqrt{\frac{n\sum X^2 - (\sum X)^2}{n^2}}$$

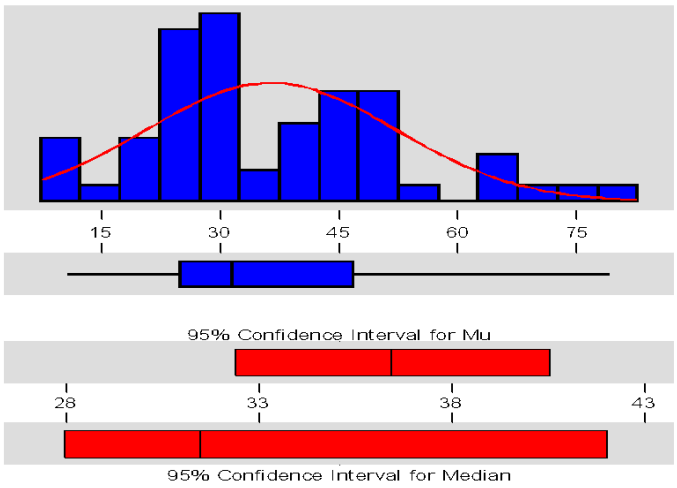
$$S = \sqrt{\frac{10(400) - (60)^2}{10^2}}$$

$$S = \sqrt{\frac{4000 - 3600}{100}}$$

$$S = \sqrt{4.0}$$

$$S = 2.0, \text{var}\,iance = 4$$

Descriptive Statistics



Variable: Age

Anderson-Darling Normality Test

A-Squared:	0.962
P-Value:	0.014
Mean	36.4500
StDev	15.7356
Variance	247.608
Skewness	0.679626
Kurtosis	8.51E-02
N	60
Minimum	11.0000
1st Quartile	25.0000
Median	31.5000
3rd Quartile	46.7500
Maximum	79.0000

95% Confidence Interval for Mu

32.3851	40.5149

95% Confidence Interval for Sigma

13.3380	19.1921

95% Confidence Interval for Median

28.0000	42.0000

Which measure to use ?(very important)

- Distribution of data is symmetric (normal)

  -Use mean & s.D.,

- Distribution of data is skewed (**not** symmetrical)

  -Use median & quartiles

SUMMARY

Flow chart of commonly used descriptive statistics and graphical illustrations