STA 111: DESCRIPTIVE STATISTICS

LECTURES NOTES 6

Topic: Measures of Location

By: E. E. E. Akarawak

SCHOOL OF
SCIENCE AND
TECHNOLOGY
PAN-ATLANTIC UNIVERSITY

Lecture Outline

- Measures of Location or Central Tendency
  - Mean of ungrouped and grouped data
  - Median of ungrouped and grouped data
  - Mode of ungrouped and grouped data

Expected Learning Outcome

- At the end of this lecture 1, the students should be able to:

i. Compute different means for ungrouped and grouped data

ii. Compute median for ungrouped and grouped data

iii. Compute the mode for ungrouped and grouped data

Measures of Location

- Measures of location, also known as measures of central tendency, are those values around which the other values lie.
- They describe the center of the distribution of any dataset.
- These values include the means, median and mode.

The Summation Notation

The symbol $\sum$ is the Greek capital letter *sigma*, denoting sum.

The symbol $\sum_{i=1}^{N} X_i$ is used to denote the sum of all the X_i 's from $i = 1$ to $i = N$; by definition

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \cdots + X_N$$

Other ways of using the notation include the following:

i.) $$\sum_{i=4}^{N} X_i = X_4 + X_5 + X_6 + \cdots + X_N.$$

ii.) $$\sum_{i=3}^{6} X_i = X_3 + X_4 + X_5 + X_6.$$

Summation Notation (*contd*)

iii.) $$\sum_{i=4}^{N} X_i^2 = X_4^2 + X_5^2 + X_{6i}^2 + \cdots + X_N^2 .$$

iv.) $$3\sum_{i=4}^{N} X_i = 3(X_4 + X_5 + X_6 + \cdots + X_N) = \sum_{i=4}^{N} (3X_i) .$$

v.) $$\sum_{i=1}^{N} X_i Y_i = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_N Y_N .$$

The Mean

- Mean summarizes a group of figures, smoothing out abnormalities in such a way that it is useful for comparison.
- Mean is the most useful of all the measures of location because it takes all the values under consideration and it is used for further mathematical calculations.
- We have three types of mean. They are:
- (i) The arithmetic mean
- (ii) The harmonic mean
- (iii) The geometric mean

Arithmetic Mean

- The Arithmetic Mean is denoted by $\bar{X}$ (read as X bar). It is the sample mean.
- The population mean (the average of all possible values of a variable is denoted by μ (the Greek letter mu)). The population mean is a constant, that is, not subject to variation but the sample mean varies from sample to sample.
- Because we almost always think of our data as a sample, we will refer to the arithmetic mean as the **sample mean**.
- The arithmetic mean can be obtained for
  - an ungrouped data set.
  - an ungrouped frequency distribution
  - a grouped frequency distribution

Arithmetic Mean of ungrouped data set

- The arithmetic mean is given by:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_\mathrm{N}}{\mathrm{N}}$$

- **Example 1:** Given the following dataset, compute the arithmetic mean.

  2, 3, 1, 0, 5, 1, -9, 5, 7, 10

Solution to Example 1

Mean, $\bar{x} = \dfrac{\sum x}{N}$

where, $\sum x = 2 + 3 + 1 + 0 + 5 + 1 - 9 + 5 + 7 + 10 = 25$ and $N = 10$

so, $\bar{x} = \dfrac{\sum x}{N} = \dfrac{25}{10} = 2.5$

Mean for an Ungrouped Frequency Distribution

$$\overline{X} = \frac{\sum fx}{\sum f} \; ; \quad \text{Where } N = \sum f \text{ is the total frequency}$$

- **Example 2:** If a final examination in a course is weighted 3 times as much as a quiz and a student has a final examination grade of 85 and quiz grades of 70 and 90, the mean grade is:

$$\overline{X} = \frac{(1)(70) + (1)(90) + (3)(85)}{1+1+3} = \frac{415}{5} = 83$$

Example 3:

A die is rolled 60 times. The results are as follows:

4	1	1	6	6	3	2	1	5	3	6	4		
2	5	5	5	6	3	3	2	1	6	6	5		
2	2	3	4	4	1	1	4	5	6	3	1		
3	2	5	4	4	1	1	2	5	4	1	3		
2	4	2	1	3	6	6	5	5	1	1	4		

(a) Prepare a frequency table.

(b) Calculate the mean

$\sum f$

Solution to Example 3

(a) Frequency table:

X	f	fx
1	13	13
2	9	18
3	9	27
4	10	40
5	10	50
6	9	54
Sum	60	202

(b) The mean:

$$\sum f = 60, \sum fx = 202$$

$$\therefore \bar{X} = \frac{\sum fx}{\sum f} = \frac{202}{60} = 3.37$$

Mean for Grouped Frequency Distribution

$$\bar{X} = \frac{\sum fx}{\sum f}; \quad X = \text{class point}$$

Example 4: In a certain organization, Employees were categorized according to their ages as shown the table. Calculate the mean age of employees in the industry.

Age group	No. of Employees
15 – 19	66
20 – 24	65
25 – 29	56
30 – 34	50
35 – 39	42
40 – 44	37
45 – 49	35
50 – 54	30
55 – 59	24
60 – 64	22

Solution to Example 4

Age group	No. of Employees (f)	Class mark (x)	fx
15 – 19	66	17	1122
20 – 24	65	22	1430
25 – 29	56	27	1512
30 – 34	50	32	1600
35 – 39	42	37	1554
40 – 44	37	42	1554
45 – 49	35	47	1645
50 – 54	30	52	1560
55 – 59	24	57	1368
60 – 64	22	62	1364
Total	**427**		**14709**

$$\bar{X} = \frac{\sum fx}{\sum f}$$

$$= \frac{14709}{427}$$

$$= 34.45$$

Assumed Mean Method

- The assumed mean, *A*, can be any of the observed values of *X*, preferably closer to the center. The formula for is:

$$\bar{X} = \mathrm{A} + \frac{\sum fd}{\sum f}$$; where $d = X - A$

$$\bar{d} = \frac{\sum fd}{\sum f} = \frac{51}{80} = 0.6375$$

Example 5

- The results of a survey on the of number of shops owned in Lagos State are shown in the following table. If 80 business owners were surveyed, compute the mean using the method of assumed mean

Score	1	2	3	4	5	6
Frequency	5	11	22	19	15	8

Solution to Example 5

- Solution: Let the assumed mean be 3 (The value of X with the highest frequency)

Score (x)	F	d = x - A	Fd
1	5	- 2	-10
2	11	-1	-11
3	22	0	0
4	19	1	19
5	15	2	30
6	8	3	24
Total	80		51

$$\bar{d} = \frac{\sum fd}{\sum f} = \frac{51}{80} = 0.6375$$

$$\bar{X} = \mathrm{A} + \frac{(\sum fd)}{\sum f}$$

$= \ 3 + 0.6375$

$= 3.6375$

The Harmonic Mean

- The reciprocal of the arithmetic mean of reciprocals of a set of observations is called the harmonic mean.
- This is useful where two averages are computed under different prevailing conditions e. g. average rates, average speed, etc.
- The harmonic mean is defined as:

$$HM = \frac{1}{\frac{1}{n}\left\{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right\}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example on Harmonic Mean (*contd*)

Example 1:

- Find the harmonic mean of   2, 3, 5, 6, 7.

Solution:

$$HM = \frac{1}{\frac{1}{5}\left\{\frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}\right\}}$$

$$= \frac{5}{1.34} = 3.73$$

Geometric Mean

- The geometric mean of the n observations is the n^th root of their product.
- That is,

$$\bar{x}_G = \sqrt[n]{x_1.x_2.x_3 \cdots x_n} = (x_1.x_2.x_3 \cdots x_n)^{\frac{1}{2}} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

- The log computation is done by:

$$\log \bar{x}_G = \frac{\sum \log x_i}{n}$$

- So that:

$$\therefore \bar{x}_G = anti \log \left(\frac{\sum \log x_i}{n} \right)$$

Example

- Compute the geometric mean of 2, 3, 5, 6, 7.

- **Solution:**

$$\bar{x}_G = (2\times3\times5\times6\times7)^{\frac{1}{5}} = (1260)^{0.2} = 4.169$$

Applications of Harmonic and Geometric Means

- Each type of mean is important depending on its intended purpose.
- The harmonic mean is most applied where the measures are ratable values like cost per litre of petrol consumed over ten years period.
- The geometric mean is most applied in indices of consumer products and prices, where two indices used two different bases of comparison.
- However, the most common measure of location is the arithmetic mean commonly called the mean or average.

Advantages and Disadvantages of Arithmetic Mean

Advantages of Arithmetic mean:

- It is mathematically easier to work with.
- It takes every individual item into account.
- Since the arithmetic mean uses all the values of all the observations, it is therefore very useful in detecting small differences between sets of observations.
- It can be used for further statistical analysis, because it has smaller standard error than other measures of location

Disadvantages of Arithmetic Mean

- It can easily be influenced by extreme values.
- It may not correspond with an actual item

MEDIAN

- Median is the number in the middle of the rank order. That is, the number that stands in the center when the given data are arranged in order of magnitude from the smallest to the highest or vice versa.
- The median can be interpreted as the item that is half above and half below, then in all cases the median is the value exceeded by 50% of the observations.
- Median will be obtained for:
- (a) a set of data items
- (b) an ungrouped frequency distribution
- (c) grouped frequency distribution

Computing Median for a Given Data Set

- First, arrange the values in order of magnitude, lowest to highest.
- Check the sample size *n*,
- If *n* is odd, select as the value at the center (middle value) as the median.
- If *n* is even, the median is the average of the two middle values.
- If *n* is large (one cannot easily locate value at the center), then determine the median position by $d = \frac{1}{2}(n+1)$,
- Then median is the value at the *d*th position

Computing Median for a Given Data Set (*contd*)

Example 1:

- Given the following set of data, determine the median: 18, 3, 114, 261, 108.

Solution:

- Step 1: Arrange the set in ascending order: 3, 18,108, 114, 261
- Step 2: Determine number of items: $n = 5$ (n is odd)
- Step 3: Obtain $d = \frac{1}{2}(n+1) = \frac{1}{2}(5+1) = 3$ (though not necessary because n is not large)
- Step 4: Then, median is the value at the 3rd position.
- Identify the 3rd value = 108.
- Hence, Median = 108

Computing Median for a Given Data Set (*contd*)

Example 2:

- Given 1, 2, 3, 7, 5, 11, 9, 6; determine the median.

Solution:

- Step 1: Arrange the set in ascending order:

  1, 2, 3, 5, 6, 7, 9, 11
- Step 2: $n = 8$ (n is even)
- Step 3: Obtain $d = \frac{1}{2}(n+1) = \frac{1}{2}(8+1) = 4.5$
- Step 4: Then, median is the value between the 4th and 5th position (central position).
- Identify the 4th value = 5 and 5th value = 6.
- Median = average of the two values = (5+6)/2 = 11/2 = 5.5.
- Hence, Median = 5.5

Computing Median for an Ungrouped Frequency Distribution

- After determining the median position by
- $d = \frac{1}{2}(n+1)$
- Use the cumulative frequency to get to the position and locate the median.
- For the data given below, determine the median.

Height of candidate (cm)	155	156	157	158	159	160	161
Frequency	3	7	10	15	16	9	2

Computing Median for an Ungrouped Frequency Distribution (*contd*)

Solution:

- From the original data, create a cumulative frequency table.

Height of candidate (cm)	155	156	157	158	159	160	161
Frequency	3	7	10	15	16	9	2
Cumulative frequency	3	10	20	35	51	60	62

- n = 62 (even)
- d = (62+1)/2 = 63/2 = 31.5. So, median is the 31.5th value (value between 31st and 32nd positions.
- The median is the average of the two values (31st value and 32nd value).
- $$\text{Median} = \frac{158+158}{2} = 158$$

Computing Median for a Grouped Frequency Distribution

- From the original grouped data, create cumulative frequency and class boundaries.
- Using the cumulative frequency curve, determine the median class as the class that contain the $\frac{1}{2}(n+1)$th value.
- The median can then be computed using the following formula.

$$\text{Median} = L_M + \left(\frac{n}{2} - F_b\right)\frac{C}{f_m}$$

where,

L_M = Lower class boundary of the median class.
n = The total frequency.
F_b = The cumulative frequency before the median class.
C = the class size or class width
f_m = The frequency of the median class

- **Example:** Compute the median of the grouped data below

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	10	95	80	160	250	212	50	25	13	5

- Solution: n=900; $d = \frac{1}{2}(n+1) = \frac{1}{2}(900+1) = 450.5$
- From cumulative frequency in the table, 450.5th value is the class: 40-50.

Class Interval	Freq	Cum. Freq.	Class Boundary
0 – 10	10	10	-0.5-10.5
10 – 20	95	105	10.5-20.5
20 – 30	80	185	20.5-30.5
30 – 40	160	345	30.5-40.5
40 – 50	250	595	40.5-50.5
50 – 60	212	807	50.5-60.5
60 – 70	50	857	60.5-70.5
70 – 80	25	882	70.5-80.5
80 – 90	13	895	80.5-90.5
90 – 100	5	900	90.5-100.5

$$\text{Median} = L_M + \left(\frac{n}{2} - F_b\right)\frac{C}{f_m}$$

$$= 40.5 + \left(\frac{900}{2} - 345\right)\frac{10}{250}$$

$$= 40.5 + (105)(0.04) = 44.7$$

Advantages and Disadvantages of Median

- **Advantages of Median**
- (i) Median is not easily influenced by extreme values.
- (ii) It uses only the value of the middle item in the distribution.
- (iii) It is equal to or exceeded by half the values in the distribution and vice versa.
- (iv) It is an actual value occurring in the distribution (unless average of the two middle items).
- (v) It can be computed even if the data is incomplete.

- **Disadvantages of Median**
- (i) It requires the arranging of data in ascending order which can be tedious,
- (ii) It is not suitable for further statistical analysis.

Mode

- Mode is the item or value that occurs most often in a given set of data.

Computing the Mode for a Ungrouped Set of Data

- For a given set of data, the mode is the item that occurs most frequently in the set. That is, the item with the highest frequency or occurrence.

Example:

- Given the following data set: 2, 3, -1, 2, 3, 2, 9, 0, -1, determine the mode.

Solution:

- Item with the highest occurrence is 2. Hence, mode = 2.

Multimodal Distribution

- Note: If two items have the same highest frequency, then both are selected as the mode and the distribution is called a *bimodal* distribution. If more than 2 modes are identified, then it is called a *Multimodal* distribution.

Example:

- Given the following data set: 3, 7, -2, 3, 3, -2, 5, 8, -2. Determine the mode.

Solution:

- Items – 2 and 3 occur equally the most number of times. So, the modes are 3 and –2. This bimodal case

Computing Mode for Ungrouped Frequency Distribution

Example: A die is rolled 60 times. The results are as follows:

4	1	1	6	6	3	2	1	5	3
6	4	2	5	5	5	6	3	3	2
1	6	6	5	2	2	3	4	4	1
1	4	5	6	3	1	3	2	5	4
4	1	1	2	4	4	1	3	2	4
2	1	3	6	6	5	5	1	1	4

(a) Prepare a frequency table.

(b) Find the mode of the distribution

Solution

a) The frequency table is given below:

X	f
1	13
2	9
3	9
4	10
5	10
6	9
	60

b) The value with the highest frequency is 1.
Therefore, the Mode = 1

Computing Mode for Grouped Frequency Distribution

- The mode of a grouped data is determined by:

$$\text{Mode} = L_m + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) C = L_m + \left(\frac{f_1 - f_0}{2 f_1 - f_0 - f_2} \right) C$$

where

L_m = Lower class boundary of the modal class.

Δ_1 = difference between the frequency of the modal class and the class that comes before it (that is, $f_1 - f_0$).

Δ_2 = difference between the frequency of the modal class and the class that comes after it (that is, $f_1 - f_2$).

C = Class size (difference between the upper class boundary and the lower class boundary)

f_0 = the frequency of the class before the modal class, f_1 = the frequency of the modal class, and f_2 = the frequency of the class after the modal class

Example

- Consider the following table of frequency counts and compute the mode of the distribution.

Class interval	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79
Counts (f)	3	11	19	28	31	12	9	7

Solution

- The table

Class interval	F	Class boundary
0 – 9	3	-0.5 – 9.5
10 – 19	11	9.5 – 19.5
20 – 29	19	19.5 – 29.5
30 – 39	28	29.5 – 39.5
40 – 49	31	39.5 – 49.5
50 – 59	12	49.5 – 59.5
60 – 69	9	59.5 – 69.5
70 – 79	7	69.5 – 79.5

- First, determine the modal class (class with the highest frequency.
- The modal class is 40 – 49 hence, the mode is calculated using

Solution (*contd*)

- Therefore,

$$Mode = L_m + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) C$$

$$= 39.5 + \left(\frac{31 - 28}{2(31) - 28 - 12} \right) 10 = 40.86$$

Something to Try:

- Solve using the second formula.

Determination of Mode Using the Histogram

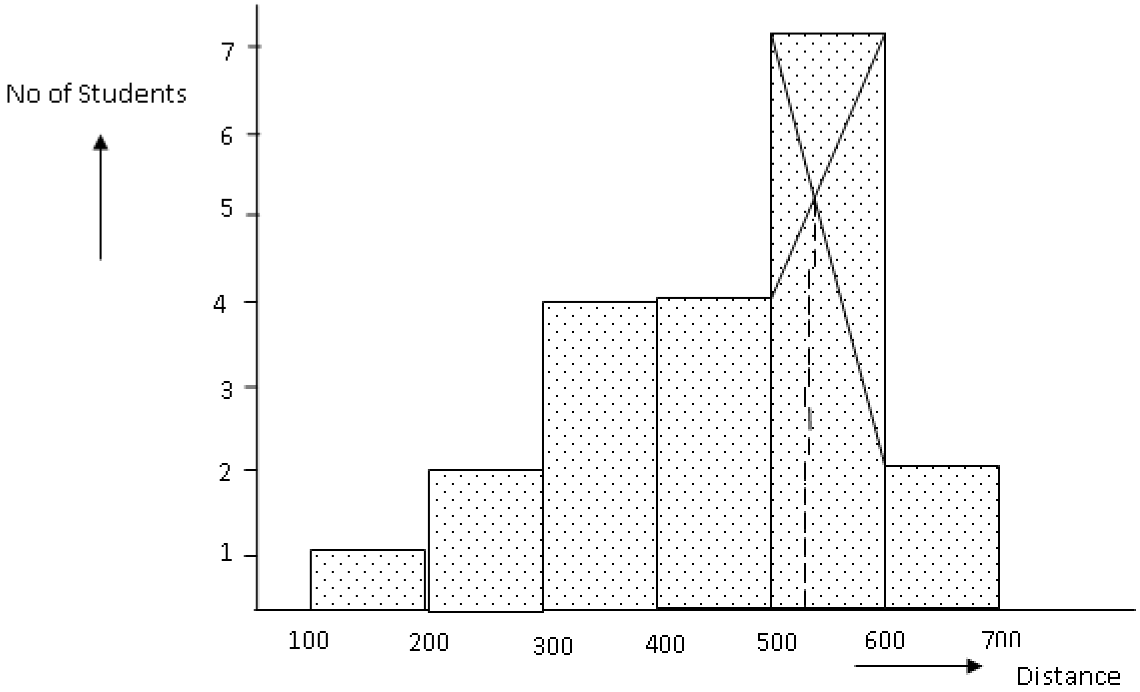
Example:

- The distances (in metres) covered by 20 students daily from their homes to their schools are shown below:
- 231 455 617 440 408 336 567
  546 523 594 680 310 317 363
  280 508 590 540 472 196
- Draw a histogram of the data using classes of 200 – under 300, 300 – under 400, etc. Use the histogram to determine the modal distance.

Determination of Mode Using the Histogram (*Contd*)

- Solution:

Class Limit	F
100 – under 200	1
200 – under 300	2
300 – under 400	4
400 – under 500	4
500 – under 600	7
600 – under 700	2



- From the histogram, the estimated mode is 540 metres.

Advantages and Disadvantages of Mode

- **Advantages of Mode**
- It is easily determined without much mathematical calculation.
- It is not affected by extreme values in the data.
- It may be applied to quantitative as well as qualitative data.

- **Disadvantages of Mode**
- It does not take every single data into consideration.
- It becomes less useful as an average when the distribution has more than a single mode.
- The mode cannot be subjected to further mathematical calculation.

Conclusion on Measures of Location

- Conclusively, looking at the attributes of each of the measures of central tendency above, one can see that they are all useful depending on what one is measuring.
- However, the Arithmetic mean seems to be the best, since it is the only one that represents the entire observations and, therefore, can be used in further statistical computation.