Basic Mathematics

Lecture 11

Statistics: Measures of central tendency, and variation (Ungrouped data)

Lecturer: Kahenya, N.P

Introduction to lecture 11

This lecture introduces you to basic statistics. It discusses data representation and the measures of central tendency and measures of variation/spread/for ungrouped data.

Intended learning outcomes

At the end of this lecture you will be able to;

(i) Define measures of central tendency and measures of variation.
(ii) Calculate measures of central tendency, and variation for ungrouped data.

References

These lecture notes should be complemented with relevant topics in (Antony & Robert, 2006; Kahenya, 2017; Upton & Cook, 2001).

Introduction

Statistics is the science that deals with the collection, organization, representation, analysis and interpretation of data or information. Data can also be referred to as statistics.

Access to reliable statistics is crucial in development process as it guides or inform good decision making and planning.

Measures of central tendency

Collected data is analyzed using measures of central tendency or measures of variations. Measures of central tendency gives an indication of how the data set is tending to a central trait or behaving towards a common trait or characteristics. Measures of central tendency include arithmetic mean $\overline{x}$, mode m_0, median m_d, harmonic mean $\overline{H}$, geometric mean $\overline{G}$, weighted mean $\overline{w}$ etc.

Arithmetic mean

Arithmetic mean is the average score of the data set.

Arithmetic mean; $\bar{x} = \frac{\text{sum of all scores}}{\text{number of scores}} = \frac{\sum x}{N} = \frac{\sum fx}{\sum f}$

Example 1: Find the arithmetic mean for the following raw data set; x; 4, 5, 7, 11, 21, 16, 20,14

Solution: $\bar{x} = \frac{\sum x}{N} = \frac{4+5+7+11+21+16+20+14}{8} = \frac{98}{8} = 85.75$

Example 2: Calculate the arithmetic mean for the following data set: x: 111, 100,90,90,110,100, 100, 120, 110, 100, 111, 90, 111, 111,110, 111.

Solution: The data can be represented in a frequency distribution table as follows:

Score x	F	fx
90	3	270
100	4	400
110	3	330
111	5	555
120	1	120
	$\sum f = 16$	$\sum fx = 1675$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1675}{16} = 104.6875$$

Mode

Mode is the most popular score in the data set. The mode is the most frequent value or score or item or entry in a data set. Modal frequency is the corresponding frequency of the mode. A data set may have no mode or more than one mode. A data set with two modes is referred to as being bimodal.

The mode is an easier measure to calculate and can give a quick overview of where the data set is tending to. For example, the most popular phone will give a quick idea of what phone model a dealer may need to stock.

Example 1: Find the mode of the data set; x: 111, 100,90,90,110,100, 100, 120, 110, 100, 111, 90, 111, 111,110, 111.

Solution: The data can be represented in a frequency table to have a clearer picture of the distribution of the scores.

Score x	frequency
90	3
100	4
110	3
111	5
120	1

The most occurring scoring is 111. It occurs 5 times while the rest occurs less than five times. Hence, the mode $m_o = 111$, and the frequency 5 is called the ***modal frequency***.

Median

The median is the middlemost item in the data set, if the data set has odd items or the average of the middlemost scores if the data set has even items.

To find the median, one need to rearrange the data in either ascending or descending order.

Example 1: The following were marks obtained by 13 students in a certain quiz;

14, 13,13,15,15,13,16,15,15,11,15, 10, 17. Determine the median mark.

Solution: Arranging the data set in ascending order we get:

10, 11, 13, 13, 13, 14, 15, 15, 15, 15, 15, 16,17

The middle most score will be in the position; $\frac{(N+1)}{2} = \frac{14}{2} = 7^{th}$

Hence 15 is the median since it is the middlemost item, in the middlemost position (7^{th}).

Example 2: The heights of six boys were recorded as follows; 145, 144, 146, 134,157, 148, 147, 148 158, 142. Determine the median height.

Solution: The scores are 10. As earlier pointed out, when the number of scores is even, the average of the two middlemost items is taken to be the median. Hence, we need to first rearrange the data in ascending order to get: 134, 142, 144, 145, 146, 147, 148, 148, 157, 158. The middle most score is in the position $\frac{N+1}{2}$ or the average of the scores in positions;

$$\frac{N}{2} \text{ and } \frac{N}{2} + 1$$

Note that $\frac{N+1}{2} = \frac{11}{2} = 5.5^{th}$ (there is no score here). Hence, we find the score at position;

$\frac{N}{2} = \frac{10}{2} = 5^{th}$ i.e. 146 and the item in position $\frac{N}{2} + 1 = 6^{th}$ i.e. 147

Hence the median is the average of 146 and 147 i.e. $m_d = \frac{146+147}{2} = 146.5$

Weighted Mean

The weighted mean $\overline{w}$ is calculated by multiplying each score $x_1, x_2, x_3, \cdots, x_n$ by its corresponding weight $w_1, w_2, w_3, \cdots, w_n$ and dividing the sum of the products by the sum of the weights i.e.

$$\overline{w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

Example 1: A student obtained the following grades in a semester exam. Calculate the corresponding GPA (Grade Point Average);

Course	**Credits, w**	**Grade, x**
Industrial attachment	8	B +
Linear Algebra	3	A -
Basic Mathematics	2	A
Statistics	3	B +
Vectors	3	B -
Calculus	3	D +
Study Skills	2	C +

Note: A = 12 points, A^- = 11points, B^+ =10 points, B = 9 points, B^- 8 points, C^- = 7 points, ..., E = 0 points.

Solution: Our GPA which is a weighted mean is given as;

$$\overline{w} = \frac{8(10) + 3(11) + 2(12) + 3(10) + 3(8) + 3(4) + 2(7)}{8 + 3 + 2 + 3 + 3 + 3 + 2} = \frac{217}{24} \approx 9.0417 \text{ (Grade B)}$$

Measures of Variation/Dispersion/Spread

Measures of central tendency can be complemented with measures of variation to make the interpretation of data more meaningful. In some situation the measures of central tendency fail to give a clear analysis of the data.

Measures of variation that will be covered in this lesson are range R, interquartile range IQR, Quartile deviation QD, variance σ^2, and standard deviation σ

Example 1: Consider the following marks obtained by 10 students in two math quizzes, Quiz 1 and Quiz 2. Each quiz was marked out of 20 as shown below. Measures of central tendency (mean, mode, and median) are also as shown in the table.

Student	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**Mean**	**Mode**	**Median**
Quiz 1	0	0	4	1	4	4	19	12	1	4	**4.9**	4	4
Quiz 2	4	4	11	11	4	4	7	2	1	1	**4.9**	4	4

The two quizzes have the same mean, mode and median. Hence, it will prove difficult to conclude in which quiz did the students perform better.

Therefore, we need to use measures of variation to get a clearer picture of the performance. We can therefore calculate the measures of variation and compare the performance of the two quizzes. We need to calculate the amount of variation in each data. To do so we find the following measures of dispersion/variation:

a) Range R; inter-quartile range IQR
b) Quartile deviation QD
c) Mean deviation, and Standard deviation

Range

One of the popular and easier to find measure of variation is range. Range is the difference between the highest value and the lowest value in the data set.

Range R = Highest value – Lowest value

For example, in the above example the range for Quiz 1 is 19, while for Quiz 2 is 10.

Interquartile range IQR

One can divide a data set in quartiles or into four equal parts when arranged in ascending order. This way you end up with first quartile (lower quartile) Q1, the second quartile (middle quartile) Q2, and the third quartile (upper quartile) Q3.

Interquartile range IQR, is the difference between the 3rd quartile and the 1st quartile i.e.

$$\text{IQR} = \text{Q}_3 - \text{Q}_1$$

The quartile deviation QD, is the average of the IQR i.e. $\text{QD} = \frac{\text{IQR}}{2}$

Example 1: For quiz 1 we have;

0, 0, [1], 1, 4, [|] 4, 4, [4], 12, 19

Q_1 Q_2 Q_3

The third quartile Q_3 = 4, and the lower quartile Q_1 = 1 then $\text{IQR} = \text{Q}_3 - \text{Q}_1 = 4 - 1 = 3$

Note: 2nd Quartile Q_2 = 4. It is the same as the median.

For quiz 2 we have; 1, 1, [2], 4, 4,[|] 4, 4, [7], 11, 11

The third Q_3 = 7 and the first quartile Q_1 = 2 then IQR = Q_3 - Q_1 = 7 – 2 = 5

Mean deviation

Note that the mode, median, range, and interquartile range involves only two or one score in the data set. The mean deviation involves all the scores of the distribution. Note that the mean deviation is a measure of variation, but it is using a mean which is a measure of central tendency. This is one of the strengths of mean i.e., it can be used to find other measures.

Example 1: Consider the following data for quiz 1; x: 0, 0, 1, 1, 4, 4, 4, 4, 12, 19

The mean $\bar{x} = \frac{\sum x}{N} = \frac{0+0+1+1+4+4+4+4+12+19}{10} = \frac{49}{10} = 4.9$

Next consider the data for quiz 2; x: 1, 1, 2, 4, 4, 4, 4,7, 11, 11

The mean $\bar{x} = \frac{\sum x}{N} = \frac{1+1+2+4+4+4+4+7+11+11}{10} = \frac{49}{10} = 4.9$

Next, we calculate the deviation of each score from the mean i.e.

Quiz 1	x	0	0	1	1	4	4	4	4	12	19
	$d = x - \bar{x}$	-4.9	-4.9	-3.9	-3.9	-0.9	-0.9	-0.9	-0.9	7.1	14.1
Quiz 2	x	1	1	2	4	4	4	4	7	11	11
	$d = x - \bar{x}$	-3.9	-3.9	-2.9	-0.9	-0.9	-0.9	-0.9	2.1	6.1	6.1

The above table shows how far (deviation) is each student from the mean. The negative sign implies the direction. What is important is the absolute deviation, hence we have the third row (for each quiz) in the table below;

Quiz 1	x	0	0	1	1	4	4	4	4	12	19
	$d = x - \bar{x}$	-4.9	-4.9	-3.9	-3.9	-0.9	-0.9	-0.9	-0.9	7.1	14.1
	$\lvert x - \bar{x} \rvert$	4.9	4.9	3.9	3.9	0.9	0.9	0.9	0.9	7.1	14.1
Quiz 2	x	1	1	2	4	4	4	4	7	11	11
	$d = x - \bar{x}$	-3.9	-3.9	-2.9	-0.9	-0.9	-0.9	-0.9	2.1	6.1	6.1
	$\lvert x - \bar{x} \rvert$	3.9	3.9	2.9	0.9	0.9	0.9	0.9	2.1	6.1	6.1

From the table above; Mean absolute deviation (Quiz 1) $= \frac{\sum(x-\bar{x})}{N} = \frac{42.4}{10} = 4.24$

Mean absolute deviation (Quiz 2) $= \frac{\sum(x-\bar{x})}{N} = \frac{28.6}{10} = 2.86$

Remark: In certain situation, the absolute deviation may not be easily handled. Alternative way is to square the deviations and then work out the Mean Squared Deviation as seen in the table below.

Quiz 1			Quiz 2		
x	d	Squared deviations d^2	x	d	Squared deviations d^2
0	-4.9	24.01	1	-3.9	15.21
0	-4.9	24.01	1	-3.9	15.21
1	-3.9	15.21	2	-2.9	8.41
1	-3.9	15.21	4	-0.9	0.81
4	-0.9	0.81	4	-0.9	0.81
4	-0.9	0.81	4	-0.9	0.81
4	-0.9	0.81	4	-0.9	0.81
4	-0.9	0.81	7	2.1	4.41
12	7.1	50.41	11	6.1	37.21
19	14.1	198.81	11	6.1	37.21
		$\sum d^2 = 330.9$			$\sum d^2 = 120.9$

Mean Squared Deviation for Quiz 1 = $\frac{\sum(x-\bar{x})^2}{N} = \frac{\sum d^2}{N} = \frac{330.9}{10} = 33.09$

Mean Squared Deviation for Quiz 2 = $\frac{\sum(x-\bar{x})^2}{N} = \frac{\sum d^2}{N} = \frac{120.9}{10} = 12.09$

Remark: The mean squared deviation is called **variance.** Its main weakness is that it considers the squares of the units under discussion. You get a value that is very large the even the scores in the data set. We then need to find the square root of the variance i.e., standard deviation which is a preferable measure than variance.

Standard deviation $\sigma = \sqrt{\text{Variance}}$ i.e.

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}} = \sqrt{\frac{d^2}{N}}$$

The standard deviation for Quiz 1 is; $\sigma = \sqrt{33.09} \approx 5.752$

The standard deviation for Quiz 2 is; $\sigma = \sqrt{12.09} \approx 3.477$

Conclusion: We can now compare the performance of the quizzes using the two measures (central tendency and variation) as shown below.

Student	A	B	C	D	E	F	G	H	J	K	Mean	Mode	Median	Range	IQR	σ
Quiz 1	0	0	4	1	4	4	19	12	1	4	4.9	4	4	19	3	5.752
Quiz 2	4	4	11	11	4	4	7	2	1	1	4.9	4	4	10	5	3.477

The measures of variation give a different perspective to the performance. Quiz 2 seems to have performed better going by the measures of variation.

Remark: The formula $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}} \cdots (i)$ can be modified to be used with grouped data.

Since $N = \sum f$ we can have equation (i) as $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}} = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} \cdots (ii)$

Again, note that $\bar{x} = \frac{\sum fx}{\sum f}$

If we expand the numerator for equation (ii) we get;

$$\sigma^2 = \frac{\sum f(x^2 - 2x\bar{x} + \bar{x}^2)}{\sum f} = \frac{\sum fx^2}{\sum f} - 2x\frac{\sum fx}{\sum f} + \bar{x}^2\frac{\sum f}{\sum f}$$

$$= \frac{\sum fx^2}{\sum f} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$\therefore \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$

Exercise

1. Find the mean, mode and median of the following data;

   a) 17, 12, 18, 18, 19, 12, 15, 13, 14,19, 21

   b) 11,14,13,14,22,12,14,21, 22,29

   c) 124,128,118,114,116,126,122,118,116,120, 117, 113

2. The frequency table below shows the weights (kg) of goats in a certain farm.

Weight x	10	11	12	13	14	15	16
Number of goats	3	2	5	7	9	3	1

Find the mode, the mean and the median weight.

1) The number of goals scored by a football team in matches played in a league were recorded as below. Calculate the mean number of goals, the mean absolute deviation, the variance, and the standard deviation.

No. of Goals	0	1	2	3	4	5
No. of matches	3	2	4	1	2	1

2) The marks were obtained some students in a test were recorded as below. Calculate the mean mark, the mean absolute deviation, the variance, and the standard deviation.

Marks	1	2	3	4	5	6	7	8	9	10
No of students	1	4	6	8	3	7	10	9	14	8

3) Calculate the mean, the mean absolute deviation, the variance, and the standard deviation for the data set below. 123, 124, 125, 126, 127, 128, 130, 131, 120, 116, 112, 124, 129.

Bibliography

Antony, C., & Robert, D. (2006). *Foundation Maths*. Prentice Hall.

Kahenya, P. (2017). *Foundation Maths*. LAP Lambert Academic Publishers.

Upton, G., & Cook, I. (2001). *Introduction to Statistics* (2nd ed.). Oxford University Press.