Basic Mathematics

Lecture 12

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

Lecturer: Kahenya, N.P

Introduction to lecture 12

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

References

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

Measures of central tendency (Grouped data)

In the previous lecture we dealt with ungrouped data set. In this lecture we shall find the arithmetic mean, mode, and median by calculation and by use of graphs. One can use a histogram to find the mode, while the median can be estimated from a cumulative frequencies curve (ogive).

Example 1: Consider the data set below and hence find the; Arithmetic mean, Arithmetic mean using the assumed mean A, Mode, Median, plot a histogram and estimate the mode, and plot a cumulative frequency curve and estimate the median.

Class	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
f	2	5	7	9	14	8	6	1

Solution:

(i) To calculate the arithmetic mean we use the formula; $\bar{x} = \frac{\sum fx}{\sum f}$

In our case x will be the midpoints of the given classes. Hence, we need a table with the additional rows.

class	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 - 49	Total
f	2	5	7	9	14	8	6	1	52
x	12	17	22	27	32	37	42	47	
fx	24	85	154	243	448	296	252	47	1549

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1549}{52} \approx 29.79$$

(ii) Arithmetic mean, using an assumed mean A

When one needs a quick mean or doesn't have time to calculate one, one can take any class midpoint more so the one at the center (in case the data is normally distributed), and work with it to interpret the data set. This is called a working mean or an assumed mean A. To calculate the arithmetic mean using an assumed mean A we use the formula;

$$\bar{x} = A + \frac{\sum fd}{\sum f}, \text{where } d = x - A$$

We need some more rows for the values of d. We shall use A = 32 as our assumed mean, since it lies almost at the middle of the data set.

class	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 - 49	Total
f	2	5	7	9	14	8	6	1	52
x	12	17	22	27	32	37	42	47	
$d = x - A$	-20	-15	-10	-5	0	5	10	15	
fd	-40	-75	-70	-45	0	40	60	15	-115

$$\bar{x} = A + \frac{\sum fd}{\sum f} = 32 + \frac{-115}{52} \approx 29.79$$

(iii) Mode

Mode of grouped data is not as straight forward as for ungrouped data. In grouped data the mode can be any value in the class interval. Hence, we need a formula to calculate the mode.

$$M_0 = l + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

Where;

l – is the adjusted lower-class limit of the modal class (to get the modal class, find the modal frequency f_0 first).

f_0 – the modal frequency (the highest frequency in the data set)

f_1 – the frequency BEFORE f_0

f_2 – the frequency AFTER f_0

i – the class interval or class size

In our case, the modal frequency f_0 is 14. Hence, the modal class is class 30 – 34 i.e., our mode lies within this class. To improve on accuracy, we need to *adjust* the class i.e., 29.5 – 34. 5. Note that 30 is the **lower-class limit** and 34 is the **upper-class limit.** After adjusting the class, we have 29.5 as our **adjusted lower-class limit** (or lower-class boundary) and 34.5 as **adjusted upper-class limit** (or upper-class boundary). The class interval or class size i is the difference between the adjusted lower-class limit and adjusted upper class limit. In our case;

$$i = 34.5 - 29.5 = 5$$

We now have our values as;

$$l = 29.5, f_0 = 14, f_1 = 9, f_2 = 8, i = 5$$

$$M_0 = l + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i = 29.5 + \frac{14 - 9}{2 \times 14 - 9 - 8} \times 5 = 29.5 + \frac{5}{11} \times 5 \approx 31.77$$

(iv) Median

We need to use a formula to find the median M_d just like the mode,.

$$M_d = l + \frac{\frac{N}{2} - cf}{f} \times i$$

Where;

l – adjusted lower-class limit of the median class

f – frequency of the median class

cf – the cumulative frequency BEFORE the median class

i – the class size or class interval

$\frac{N}{2}$ – the median position

The first step is to find the **median position**. We need to cumulate the frequency to identify the median position and hence the median class. Cumulative frequency (cf) column is like ranking the frequencies to identify the middlemost position.

Class	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
F	2	5	7	9	14	8	6	1
Cf	2	7	14	23	37	45	51	52

The median position is 26.5^{th} (since N = 52 which is an even number. Note that if the N is a very large number, we ignore the 0.5).

The 26.5^{th} position is at cumulative frequency 37 (here we have positions 24 up to 37).

Hence our median class is class 30 – 34. Which means the l = 29.5. The cumulative frequency BEFORE the median class cf = 23

The median frequency f = 14.

Therefore; $M_d = l + \frac{\frac{N}{2} - cf}{f} \times i = 29.5 + \frac{26.5 - 23}{14} \times 5 = 29.5 + \frac{3.5}{14} \times 5 = 30.75$

Remark: The mean = 29. 79, mode = 31.77, and median = 30.75 lies almost in the same neighborhood. Since the three are measures of *central tendency*, and the data set is almost normally distributed.

(v) Mode using a histogram

Definition 1: A histogram is one way to represent grouped data. A histogram is a bar graph for continuous data set. A histogram can be used to get the mode of the data set.

Procedure of drawing a histogram:

A histogram is drawn by plotting the frequencies against the lower-class boundaries.

The size of the bar is proportional to the frequency i.e., the area of the bars represents the given frequency. Sometimes you may find grouped data with no uniform class intervals. Hence if for instance the class interval is doubled, then the frequency is halved.

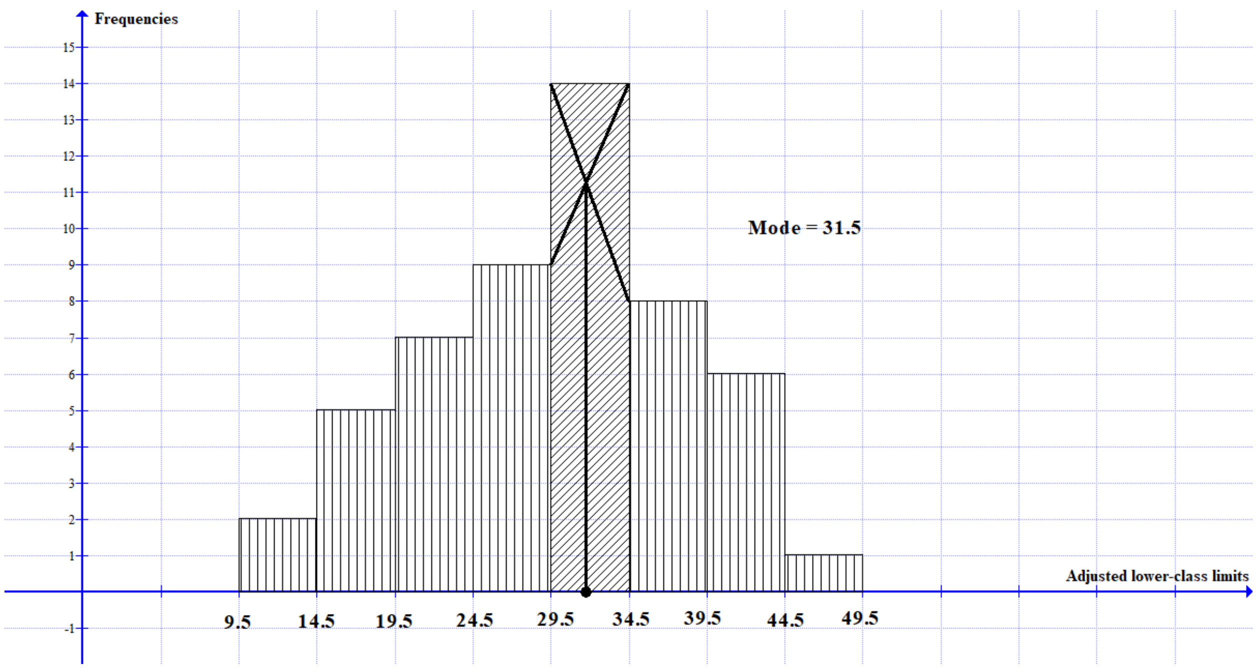
Estimating the mode using a histogram

To estimate the mode, follow the steps below.

- Identify the highest bar. The mode is in this class.
- Draw a diagonal from the top-left corner of this bar, to touch the top-left corner of the bar opposite this corner.
- Draw a diagonal from the top right corner of this highest bar, to touch the top right corner of the bar opposite this corner.
- Drop a perpendicular from the point of intersection of these diagonals, to the horizontal axis.
- The mode is the point where the perpendicular touches the horizontal axis.

Example 1

Class	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
f	2	5	7	9	14	8	6	1
Lower class boundaries	9.5	14.5	19.5	24.5	29.5	34.5	39.5	44.5



From the histogram the mode is approximately 31.5 (compare with calculated value of 31.77).

(vi) Median using a cumulative frequencies curve

Definition: The cumulative frequencies curve is also referred to as an *Ogive* or an *S- curve*. If the data is normally distributed the curve makes a smooth letter S, hence the name S-curve.

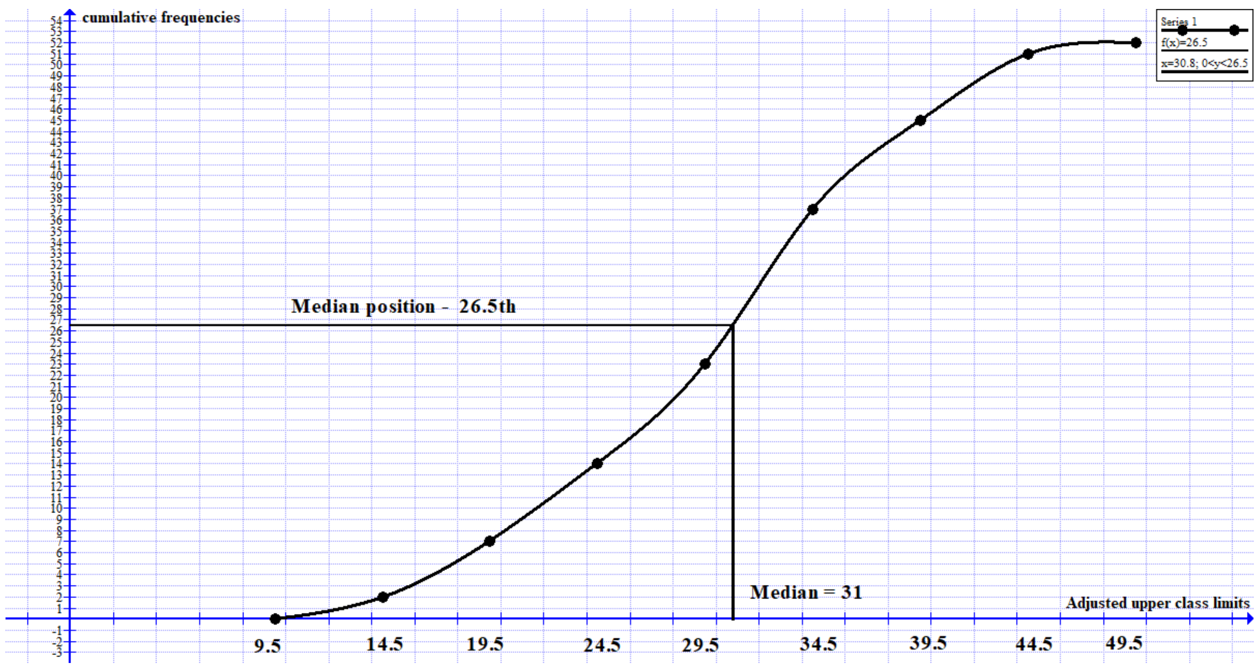
To plot the cumulative frequencies curve, we plot the cumulative frequencies against the adjusted upper-class limits.

If the first class has non-zero frequency, then you need to add another class with zero frequency such that the ogive starts from zero.

Join the points to get a smooth curve. Note that in this example the median position is 26.5th.

Next, along the cumulative frequency axis identify the 26.5 point and then draw a line parallel to the horizontal axis. At the point of intersection with the ogive, drop a perpendicular to get the median. In our case about 31 (see Ogive below).

Class	5 - 9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
f	0	2	5	7	9	14	8	6	1
cf	0	2	7	14	23	37	45	51	52
Upper class boundaries	9.5	14.5	19.5	24.5	29.5	34.5	39.5	44.5	49.5



Measures of variation (grouped data)

Example 1: The performance in a math quiz for a group of students was recorded as below. Calculate the interquartile range, and the standard deviation.

Marks	10 -19	20 - 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 - 99
No. of students	1	2	7	10	14	6	5	3	2

Solution:

(i) The interquartile range

Marks	10 -19	20 - 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 - 99
No. of students	1	2	7	10	14	6	5	3	2
Cumulative freq.	1	3	10	20	34	40	45	48	50

Recall that $IQR = Q_3 - Q_1$. To find the third quartile we use the formula;

$$Q_3 = l + \frac{\frac{3}{4}N - cf}{f} \times i$$

This formula is similar with the formula for finding the median (median is the same as Q_2)

The third quartile is at position $\frac{3}{4} \times 50 = 37.5^{th}$ position (at cumulative frequency 40). Hence, the 3rd quartile class is class 60 – 69, with a frequency f = 6.

Hence the adjusted lower-class limit l = 59.5

The cumulative frequency BEFORE the 3rd quartile class is 34. Class size is 10

Therefore, $Q_3 = l + \frac{\frac{3}{4}N - cf}{f} \times i = 59.5 + \frac{37.5 - 34}{6} \times 10 = 59.5 + 5.833 \approx 65.33$

Next, to find the first quartile we use the formula;

$$Q_1 = l + \frac{\frac{1}{4}N - cf}{f} \times i$$

The first quartile is at position $\frac{1}{4} \times 50 = 12.5^{th}$ position (at cumulative frequency 20). Hence, the 1st quartile class is class 40 – 49, with a frequency f = 10.

Hence the adjusted lower-class limit l = 39.5

The cumulative frequency BEFORE the 1st quartile class is 10. Class size is 10.

Therefore, $Q_1 = l + \frac{\frac{1}{4}N - cf}{f} \times i = 39.5 + \frac{12.5 - 10}{10} \times 10 = 39.5 + 2.5 = 42$

Therefore, $IQR = Q_3 - Q_1 = 65.33 - 42 = 23.33$

(ii) Standard deviation

To calculate the standard deviation, we use the formula;

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$$

We therefore need to find the mean $\bar{x}$, x^2, and fx^2 (a table will facilitate this working).

Marks	F	x	Fx	x^2	fx^2
10 – 19	1	14.5	14.5	210.25	210.25
20 – 29	2	24.5	49	600.25	1200.5
30 – 39	7	34.5	241.5	1190.25	8331.75
40 – 49	10	44.5	445	1980.25	19802.5
50 – 59	14	54.5	763	2970.25	41583.5
60 – 69	6	64.5	387	4160.25	24961.5
70 – 79	5	74.5	372.5	5550.25	27751.25
80 – 89	3	84.5	253.5	7140.25	21420.75
90 – 99	2	94.5	189	8930.25	17860.5
	$\sum f = 50$		$\sum fx = 2715$		$\sum fx^2 = 163122.5$

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{\frac{163122.5}{50} - \left(\frac{2715}{50}\right)^2} = \sqrt{3262.45 - 2948.49} = \sqrt{313.96} \approx 17.72$$

Exercise

1. The marks obtained by a group of students in a quiz were recorded as below;

Marks	1- 5	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30	31 - 35
No. of students	2	12	20	15	11	6	4

Calculate the interquartile range, quartile deviation, variance, and the standard deviation.

2. Students' performance in a certain math quiz was recorded as below;

Marks	10 -19	20 - 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 - 99
No. of students	10	13	16	19	27	21	15	11	8

Use the above data to find; Arithmetic mean, Arithmetic mean using an assumed mean, mode, median, mode using a histogram, and median using a cumulative frequencies curve

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.