Descriptive Statistics
Numerical Summaries
Lecture 4

Lecture objectives

- Summarize data, using measures of central tendency, such as the mean, median, midrange,weighted mean, mode, harmonic mean and geometric mean.

1 Measures of Central Tendency

1.1 The mean

The mean or *arithmetic mean* is the sum of the values, divided by the total number of values, we use the symbol $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

where
n represent total number of values in the sample
For a population, μ is used for the mean.

$$\mu = \frac{\sum_{i=1}^{n} X_i}{N} = \frac{X_1 + X_2 + X_3 + ... + X_N}{N}$$

where
N represents the total number of values in the population.

A parameter is a characteristic or measure obtained by using all the data values from a specific population

Example 1. The data represent the number of days off per year for a sample of individuals selected from nine different countries.
$20, 26, 40, 36, 23, 42, 35, 24, 30$
Find the mean.

Solution

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20+26+40+36+23+42+35+24+30}{9} = \frac{276}{9}$ =30.7
Hence, the mean of the number of days off is 30.7 days.

Example 2. Using the frequency distribution below, find the mean

class limits	frequency
6 − 10	1
11− 15	2
16 −20	3
21 − 25	5
26 − 30	4
31 − 35	3
36 − 40	2
	Total=20

solution

First Find the classes mid point.
Second For each class, multiply the frequency by the midpoint,find it's sum.
Third divide the previous sum by n to get the mean.

$$\bar{x} = \frac{\sum f * x_m}{n}$$

class limits	frequency	Mid pints	$f * x_m$
$6 - 10$	1	8	8
$11 - 15$	2	13	26
$16 - 20$	3	18	54
$21 - 25$	5	23	115
$26 - 30$	4	28	112
$31 - 35$	3	33	99
$36 - 40$	2	38	76
	Total=20		$\sum f * x_m = 490$

$$\bar{x} = \frac{\sum f * x_m}{n} = \frac{490}{20} = 24.5$$

1.2 The Median

The median is the midpoint of the *data array*, which is the ordered data set.

Example 3. The number of rooms in the seven hotels is
713, 300, 618,595, 311, 401, and 292. Find the median

solution

First Arrange the data in order.
292, 300, 311, 401, 595, 618, 713
Second Find the median for the data
if the total data values is odd the median order is $\frac{n+1}{2}$
if the total data values is even the median is the average
of the two midpoint values.$m_1 = \frac{n}{2}$ $m_2 = \frac{n}{2} + 1$
$MD = \frac{x_{m_1} + x_{m_2}}{2}$
the total data in this example is odd so the median order
is $\frac{7+1}{2} = 4$ the median is the $four^{th}$ element which is 401

Example 4. The number of children with asthma during a specific year in eight cities is.
253, 125, 328, 417, 201, 70, ,110,90
Find the median.

solution

First order the data
70, 90,110, 125, 201, 253, 328, 417
the total number of the data is even 8

Second find the two midpoints' order
$m_1 = \frac{8}{2} = 4$ $m_2 = \frac{8}{2} + 1 = 5$
$MD = \frac{x_4 + x_5}{2} = \frac{125+201}{2} = \frac{326}{2} = 163$

Example 5. using the frequency table below find the median

class limits	frequency
$6 - 10$	1
$11 - 15$	2
$16 - 20$	3
$21 - 25$	5
$26 - 30$	4
$31 - 35$	3
$36 - 40$	2
	Total=20

solution

First Find the cumulative frequency.

Second Determine the median class , which is the class with the cumulative frequency exactly greater then the median order.

Third find the median value using the rule
$median(MD) = L + \frac{w}{f_m} * (0.5 * n - cf_b)$
where
L=lower class limit of the class that contains the median
n= total frequency
cf_b =the sum of frequencies (cumulative frequency) for all classes before the median class
f_m =frequency of the class interval containing the median
w= interval width

class limits	frequency	CF
$6 - 10$	1	1
$11 - 15$	2	3
$16 - 20$	3	6
$21 - 25$	5	11
$26 - 30$	4	15
$31 - 35$	3	18
$36 - 40$	2	20
	Total=20	

the median class is the class with the cumulative frequency greater than 10 which is 11 the class is $21 - 25$, applying the median rule we get

$$MD = 21 + \frac{5}{5} * (0.5 * 20 - 6) = 25$$

1.3 The Mode

The value that occurs most often in a data set is called the mode
A data set that has only one value that occurs with the greatest frequency is said to be *unimodal*.If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be *bimodal*. If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be *multimodal*. When no data value occurs more than once, the data set is said to have *no mode*. A data set can have more than one mode or no mode at all

Example 6. Find the mode for the following data sets

1. 8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11
2. 10, 31, 30, 84, 20, 18, 62, 77, 33, 52
3. 15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26

solution

1. The mode is 8.
2. There is no mode.
3. The modes are 18 and 24.

Example 7. using the previous frequency table from the last example find the mode

Solution

First Find *the modal class*.which the class with the highest frequency
Second find the mode value using the rule
$mode = L + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} * w$
where
L=lower class limit of the interval that contains the mode(*modal class*)
f_i = the frequency of the modal class.
f_{i-1} = the frequency of the class previous to the modal class.
f_{i+1} = the frequency of the class after the modal class.
w= the width of the modal class.
The modal class is $21 - 25$, applying the previous rule

$$mode = 21 + \frac{5-3}{(5-3)+(5-4)} * 5 = 24.333$$

1.4 The Midrange

The midrange is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$Midrange = MR = \frac{Lowest\ value + Highest\ value}{2}$$

Example 8. For the following dataset find the midrange 2, 3, 6, 8, 4, 1

Solution

$MR = \frac{1+8}{2} = 4.5$
The midrange is 4.5

Example 9. Find the midrange of data for the yearly income for a sample of employees in a company 18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

solution

$$MR = \frac{10 + 34.5}{2} = 22.5$$

Notice that the midrange doesn't represent most of the data , and this is due to the highest value is extremely large.

1.5 The Weighted Mean

the weighted mean of a variable X is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1+w_2X_2+w_3X_3+...+w_nX_n}{w_1+w_2+w_3+...+w_n} = \frac{\sum_{i=1}^{n} w_iX_i}{\sum_{i=1}^{n} w_i}$$

where

$w_1, w_2, w_3, \ldots, w_n$ are the weights and $X_1, X_2, X_3, \ldots, X_n$ are the values. and it's used when, you need to find the mean of a data set in which not all values are equally represented.

Example 10. A student received an A in STAT1 (72), a C in Comp (53), a B in Mang(65), and a D in Math (47). Assuming STAT1 and Comp are 3 hours credit each, Mang and Math are 2 hours credit.what is the student's average

Solution

applying the weighted mean rule we get

$$\bar{X} = \frac{72 * 3 + 53 * 3 + 65 * 2 + 47 * 2}{3 + 3 + 2 + 2} = \frac{599}{10} = 59.9$$

the student's grade average is 59.9.

1.6 Geometric mean

The geometric mean (GM) is dened as the nth root of the product of n values, it useful in nding the average of percentages, ratios, indexes, or growth rates. The formula is

$$GM = \sqrt[n]{(x_1)(x_2)...(x_n)}$$

$$\log GM = \frac{\sum log\, x}{n}$$

For example the GM of 1,3 and 9

$$GM = \sqrt[3]{(1)(3)(9)} = 3$$

Example 11. For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the Geometric mean

Weight of ear head x (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20
Total	160

Solution

First Find *the log of the mid point*
Second find summation of the $frequancy * log\, of\, the\, mid\, point$
Third find the Geometric mean using the rule
$log\, GM = \frac{\sum f_i log\, x_i}{\sum f_i}$
$log\, GM = \frac{324.2}{160} = 2.02625$
$GM = 106.23$

Weight of ear head x (g)	No of ear heads (f)	Mid point	log mid point	f*m
60-80	22	70	1.845	40.59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.176	43.52
Total	160			324.2

1.7 Harmonic mean

The harmonic mean (HM) is dened as the number of values divided by the sum of the reciprocals of each value, this mean is useful for finding the average speed.

$$HM = \frac{n}{\sum \frac{1}{x}}$$

or

$$HM = \frac{\sum_i f_i}{\sum f_i \frac{1}{x_i}}$$

For example HM of 1,4,5 and 2

$$HM = \frac{4}{\frac{1}{1} + \frac{1}{4} + \frac{1}{5} + \frac{1}{2}} = 2.05$$

2 Properties and Uses of Measures of Central Tendency

2.1 The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called *outliers*, and may not be the appropriate average to use in these situations.

2.2 The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

2.3 The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal or categorical, such as gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

2.4 The Midrange

1. The midrange is easy to compute.
2. The midrange gives the midpoint.
3. The midrange is affected by extremely high or low values in a data set.

2.5 The Geometric Mean

1. The geometric mean is more representative of value than the arithmetic mean as it is not affected by extremely high or low values.
2. The geometric mean is the most appropriate measure to calculate the mean ratio and growth rate.
3. The geometric mean can not use if one of the data is zero.

2.6 The Harmonic Mean

1. The harmonic mean is rigidly defined,and defined on all observations.
2. The harmonic mean is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.
3. The harmonic mean is not easily understood and difficult to compute.