NEAR EAST UNIVERSITY

STATISTIC 281

BB

THE ARITHMETIC MEAN

Submitted to : Prof. Dr. Mevlüt Gaglar
Submitted by : Gokhan Hayri YILDIZ (90313)

A MEASURE OF CENTRAL TENDENCY THE ARITHMETIC MEAN

Most of the time when we refer to the "average" of something, we are talking about the arithmetic mean. This is true in cases such as the average winter temperature of New York City, the average life of a flashlight battery, and the average corn yield from an acre of land.

TABLE 1: Downtime of generators at lake Ico Station.

GENERATOR	1	2	3	4	5	6	7	8	9	10
DAYS OUT OF Service	7	23	4	8	2	12	6	13	9	4

Table 1 repeats the data from our chapter opening example. Data in the table represent the number of days the generators are out of service owing to regular maintenance or some mal function. To find the arithmetic mean. we sum the values and divide by the number of observations:

$$\text{Arithmetic mean} = \frac{7+23+4+8+2+12+6+13+9+4}{10} = \frac{88}{10} = 8,8 days$$

In this one-year period, the generators were out of service for an average of 8.8 days. With this figure, the power plant manager has a reasonable single measure of the behaviour of all her generators.

Conventional Symbols:

"Characteristics of a sample are called statistics." To write equations for these measures of frequency distributions. we need to learn the mathematical notations used by statisticians. A sample of a population consists of n observations (o lower - case n) with a mean of $\bar{x}$ (read x- bar).

"Characteristics of a population are called parameters"

The notation is different when we are computing measures or the entire population. That is, for the group containing every element we are describing. The mean of a population is symbolised by μ, which is the Greek letter mu. The number of elements in a population is denoted by the capital italic letter N. Generally is statistics, we use Roman letters to symbolise sample information and Greek. Letters to symbolic population information.

Calculating the Mean From Ungrouped Data

In the example, the average of 8.8 days would be μ (the population mean) if the population of generators is exactly ten. It would be $\bar{x}$ (the sample mean). If the ten generators are a sample drawn from a larger population of generators. To write the formulas for these two means, we combine our mathematical symbols and the steps we used to determine the arithmetic mean. If we add the values of the observations and divide this sum by the number of observations. we will get:

μ : Population mean
Σx : Sum of values of all observer variations
N : Number of elements in the population.

$$\mu = \frac{\Sigma x}{N}$$

AND : $\bar{x}$: Sample mean
Σx : Sum of values of all observations
n : Number of elements in the sample

$$\bar{y} = \frac{\sum x}{n}$$

since μ is the population arithmetic mean, we use N to indicate that we divide by the number of observations or elements in the population. Similarly, $\bar{x}$ is the sample arithmetic mean, and n is the number of observations in the sample. The Greek letter sigma, Σ , indicates that all the valves of x are summed together.

Notice that to calculate this mean. we added every observation separately, in no special order. Statisticians call this ungrouped data. The computation were not difficult, because our sample size was small. But suppose we are dealing with the weight of 5.000 head of cattle and prefer not to add each of our data points separately. Or suppose we have access to only the frequency distribution of the data. not to every individual observation. In these cases, we will need a different way to calculate the arithmetic mean.

Calculating the Mean From Grouped Data

A frequency distribution consists of data that are grouped by classes. Each value of an observation falls some where in one of the classes. Unlike the SAT example. we do not know the separate values of every observation. Suppose we have a frequency distribution (illustrated in Table 3) of average monthly checking - account balances of 600 customers at a branch band. From the information in this table. we can easily compute an estimate of the value of the mean of this grouped data. It is an estimate because we do not use all 600 data points in the sample. Had we used the original. ungrouped data. we could have calculated the actual valve of the

mean - but only after we had averaged the 600 separate valves. For ease of calculation. we must give up accuracy.

TABLE 3 : Average monthly balances of 600 customers

CLASS(DOLLARS)	FREQUENCY
0-49.99	78
50.00-99.99	123
100.00-149.99	187
150.00-199.99	82
200.00-249.99	51
250.00-299.99	47
300.00-349.99	13
350.00-399.99	9
400.00-449.99	6
450.00-499.99	4
	600

To find the arithmetic mean of grouped data, we first calculate the mid point of each class. To make the class marks come out in whole cents. we round up. Thus, for example, the class mark for the first class mark for the first class becomes 25.00, rather than 24. 995. Then we multiply each class mark by the frequency of observations in that class, sum all these results, and divide the sum by the total number of observations in the sample. The formula looks like this:

$$\bar{x} = \frac{\sum f(x)}{n}$$

where

* $\bar{x}$ is the sample mean
* $\sum$ is the symbol meaning "the sum of"
* f is the frequency(Number of the observations) in each class
* x represents the class mark for each class in the sample.
* n is the number of observations in the sample

Coding

When we have to do the arithmetic by hand, we can further simplify our calculation of the mean from grouped data. Using a technique called coding, we eliminate the problem of large or incon venient class marks. Instead of using the actual class marks to perform our calculations, we can assign small - value consecutive integers (whole numbers) called codes to each of the class marks. The integer zero can be assigned any where, but to keep the integers small, we will assign zero to the class mark in the middle (or the one nearest to the middle) of the frequency distribution. Then we can assign negative integers to values smaller that that class mark and positive integers to those larger, as follows:

class	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
code(u)	-4	-3	-2	-1	0	1	2	3	4
					↑				
					x_0				

Symbolically, statisticians use. X_0 to represent the class mark that is assigned the code O, and u for the coded class marks. The following formula is used to determine the sample mean sing codes:

$$\bar{x} = x_o + w\frac{\sum (u \times f)}{n}$$

where:

* $\bar{x}$ = mean of sample
* x_o = value of the class mark assigned the code O
* W= numerical width of the class interval
* u= code assigned to each class
* f= frequency or number of observations in each class
* n= total number of observations in the sample

A SECOND MEASURE OF CENTRAL TENDENCY;

The Weighed Mean

The weighted mean enables us to calculate an average that takes into account the importance of each value to the overall total. Consider, for example, the company in Table 10, which uses three grades of labour - unskilled, semiskilled, and skilled - to produce two end products. The company wants to know the average cost of labour for hour for each of the products.

TABLE 10 ; Labour in put in manu facturing process

		Labour hours per unit of output	
Grade of Labour	Hourly wage (x)	product 1	product 2
unskilled	4.00 dollars	1	4
semi-skilled	6.00dollars	2	3
skilled	8.00	5	3

A sample arithmetic average of the labour wage rates would be :

$$\bar{x} = \frac{\sum x}{n} = \frac{4+6+8}{3} = \frac{18}{3} = 6.00 dollars/hour$$

Using this average rate, we would compute the Labour cost of one unit of product 1 to be 6.(1+2+5)= 48 dollars, and of one unit of product 2 to be 6.(4+3+3)=60 doll . But these answers are incorrect.

To be correct, the answers must take in to account the fact that different amounts of each grade of labour are used. We can determine the correct answers in the following manner. For product 1, the total labour cost per unit is

(4x1)+(6x2)+(8x5)= 56 dollars and

since there are eight hours of labour input, the average Labour cost per hour is 56/8=7.00 per hour. For product 2, the total labour cost per unit is (4x4)+(6x3)+(8x3)=58.
For an average labour cost per hour of 58/10 or 5.80 per hour.

Another way to calculate the correct average cost per hour for the two products is to take a weighted average of the cost of the three grades of labour. To do this, we weight the hourly was for each grade by its proportion of the total labour required to produce the product. One unit of product 1, for example, requires eight hours of labour. Unskilled Labour uses 1/8 of this time, semiskilled Labour uses 2/8 of this time, and skilled labour requires 5/8 of this time.If we use these fractions as our weights, then one hour of labour for product 1 costs on average of:

(1/8x4)+(2/8x6)+(5/8x8) = 7.00 dollars/hour

Similarly, a unit of product 2 requires ten labour hours, of which 4/10 is used for unskilled labour, 3/10 for semiskilled labour, and 3/10 for skilled labour. Using these fractions as weights, one hour of labour for product 2 costs:

(4/10x4)+3/10x6)+(3/10x8)=5.80 dollars/hour

Thus, we see that the weigh ted averages give the correct values for the average hourly labour costs of the two products because they take in to account the fact that different amounts of each grade of labour are used in the products. Symbolically, the formula for calculating the weighted average is:

$$\bar{x}_w = \frac{\sum (wx_i x_i)}{\sum w}$$

* $\bar{x}_w$ = the symbol for the weighted mean

* w = weight assigned to each observation (1/8,2/8 and 5/8 for prod.1 in our examp)

* $\sum (wXx)$ = sum of the weight of each element times that element

* $\sum w$ = sum of all the weights

A THIRD MEASURE OF CENTRAL TENDENCY:

The Geometric Mean:

Some times when we are dealing with quantities that change over a period of time, we need to know an average rate of change, such as an average growth rate over a period of several years. In such cases, the simple arithmetic mean

TABLE; 11 Growth of 100 dollars deposit in a saving account

Year	Interest rate	Growth factor	Savings at end of year
1	7%	1,07	107
2	8%	1.08	115.56
3	10%	1.1	125.12
4	12%	1.12	142.37
5	18	1.18	168

Is inappropriate, because it gives the wrong answers. What we need to find is the geometric mean, called simply the G.M.

Consider, for example, the growth of a savings account. Suppose we deposit 100 dollar initially and let it accrue interest at varying rates for five years. The entry labelled " grow the factor" is equal to :

1+Interest rate/100

The growth factor is the amount by which we multiply the savings at the beg in ning of the year to get the savings at the end of the year. The simple arithmetic mean growth factor would be (1,07+1.08+1.10+1.12+1.18)/5 = 1.11 which core sponds to an average interest rate of 11 percent per year. If the bank gives interest at a constant rate of 11 percent per year, however, a 100 dollar deposit would grow in five years to:

100x1.11x1.11x1.11x1.11x1.11= 168.5

Table 11 shows that the actual figure is only 168.00. dollar Thus, the correct average growth factor must be slightly less than 1.11.

To find the correct average growth factor, we can multiply together the five years growth factors and then take the fifth root of the product - the number that, when multiplied by it

self four times. is equal to the product we started with. The result is the geometric mean growth rate. which his the appropriate average to use here. The formula for finding the geometric mean o f a series of numbers is

$$G.M = \sqrt[n]{\text{Product of all the x values}}$$

n: Number of x values

If we apply this equation to our savings - account problem, we can determine that 1.1093 is the correct average growth factor.

$$G.M = \sqrt[n]{\text{Product of all the x values}}$$

$$= \sqrt[5]{1.07 \times 1.08 \times 1.10 \times 1.12 \times 1.18}$$

$$= \sqrt[5]{1.67\ 79965}$$

= 1.1.093 average growth factor

A FOURTH MEASURE OF CENTRAL TENDENCY : THE MEDIAN.

The median is a measure of central ten dency different from any of the means we have discussed so for. The median is a single value from the data set that measure the central item. in the data. This single item is the middle most or most central item in the set of numbers. Half of the items lie above this point, and the other half lie below it.

Calculating the MEDIAN from Ungrouped Data

To find the median of a data set. first array the data in as cending or descending order. If the data set contains an odd numb

er of items. the middle item of the array is the median. If there is an even number of items. the median is the average of the two middle items. In formal Language. the median is:

Median=the (n+1)/2 th item in a data array

n: Number of items in the array

Suppose we wish to find the median of seven items in a data array. The median is the (7+1)/2=4 th item in the array. If we apply this to our previous example of the times for seven members of a track team. we discover that the fourth element in the array is 4.8 minutes. This is the median time for the track team. Notice that unlike the arithmetic mean we calculated earlier. the median we calculated in Table 12 was not distorted by the presence of the last value (9.0). This value could have been 15. 0 or even 45. 0 minutes. and the median would have been the same!

Table 12. Times for track - team members.

Item in Data array	1	2	3	4	5	6	7
Time in minutes	4.2	4.3	4.7	4.8	5.0	5.1	9.0
				↑			
				median			

Now let's calculate the median for an array with an even number of items. Consider the data shown in Table 13 concerning the number of patients treated daily in the emergency room of a hospital. The data are arrayed in descending order. The median of this data set would be.

median = the (n+1)/2 th item in a data array

=(8+1)/2

= 4.5 th item

Since the median is the 4.5th element in the array, we need to average the fourth and fifth elements. The fourth element in Table 13 is 43, and the fifth is 35. The average of these two elements is equal to (43+ 35)/2 or 39.

There fore, 39 is the median number of patients treated in the emergency room per day during the 8-day period.

TABLE 13. Patients treated in emergency room on 8 consecutive days.

Item in data array	1	2	3	4	5	6	7	8

Number of patients	86	52	49	43	35	31	30	11

median of 39

Calculating the Median From Grouped Data

Often, we have access to data only after it has been grouped in a frequency distribution. We do not, for example, know every observation that led to the construction of Table 14, the data on 600 bank customers originally introduced earlier. Instead, we have ten class intervals and a record of the frequency with which the observation appear in each of the intervals.

TABLE 14 Average monthly balances for 600 customers.

Class in Dollars	Frequency
0-49.99	78
50.00-99.99	123
100.00-149.99	187 ←median class
150.00-199.99	82
200.00-249.99	51
250.00-299.99	47
300.00-349.99	13
350.00-399.99	9
400.00-449.99	6
450.00-499.99	4

Nevertheless, we can compute the median checkingaccount balance of these 600 customers by determining which of the ten class intervals contains the median. To do this, we must add the frequencies in the frequency column in Table 14 until we reach the (n+1)/2 th item Since there are 600 accounts, the value for (n+1)/2 is 300.5 (the average of the 300th and 301st items). The problem is to find the class intervals containing the 300th and 301st elements. The cumulative frequency for the first two classes is only 78+123= 201. But when we move to the third class interval, 187 elements are added to 201 for a total of 388. Therefore, the 300th and 301st observations must be located in this third class (the interval from 100.00 dollar to 149.99 dollar)

The median class for this data set contains 187 items. If we assume that these 187 items begin at 100.00 dollar and are evenly spaced over the entire class interval from 100.00 dollar to 149.99 dollar, then we can interpolate and find values for the 300th and 301st items. First, we determine that the 300th item is the 99th element in the median class:

301-201 (items in the first two classes)= 99 and that the 301st item is the 100th element in the median class:

301-201= 100

Then we can calculate the width of the 187 equal steps from 100.00 dollar to 149.99 dollar, as follows:

$$\frac{\text{First item of next class-First item of median class}}{187}$$

$$\frac{150.00-100.00}{187} = .267 \text{ in width}$$

Now, if there are 187 steps of 267 dollar each and if 98 steps will take us to the 99th item, then the 99th item is:

(.267x98)+100=126.17

and the 100th item is on assitional step:

126.17+.267=126.44 dollars

Therefore, we can use 126.17 dollar and 126.44 dollar as the values of the 300th and 301st items, respectively.

The actual median for this data set is the value of the 300.5th item;

That is, the average of the 300th and 301st items.

This average is:

(126.17 + 126.44)/2

This figure (126.30 dollar) is the median monthly checking account balance, as estimated from the grouped data in table 14.

A FINAL MEASURE OF CENTRAL TENDENCY; THE MODE

The mode is a measure of central tendency that is different from the mean but some what like the median because it is not actually calculated by the ordinary processes of arithmetic. The mode is that value that is repeated most often in the data set.

As in every other aspect of life, chance can play a role in the arrangement of data. Sometimes chance causes a single unrepresentative item to be repeated often enough to be the most frequent value in the data set. For this reason, we rarely use the mode of ungrouped data as a measure of central ten tendency. Table 15, for example, shows the number of delivery

trips per day made by a Redi-mix concrete plant. The modal value is 15 because it occurs more often than any other value (three times). A mode of 15 implies that the plant activity is higher than 6. 7. The mode tells us that 15 is the most frequent number of trips, but it fails to let us know that most of the values are under 10.

TABLE 15: delivery tripped per day in on 20- day period

TRIPS ARRAYED IN ASCENDING ORDER

0	2	5	7	15
0	2	5	7	15
1	4	6	8	15
1	4	6	12	19

Now let's group this day into a frequency distribution, as we have done in table 16. If we select the class with the most observations, which we can call the modal class, we would choose "4-7" trips.

This class is more representative of the activity of the plant than is the mode of 15 trips per day.

TABLE 16 : Frequency distribution of delivery trips

CLASS IN NUMBER OF TRIPS	0-3	4-7	8-11	12AND MORE
Frequency	6	8	1	5
		†		
		module class		

For this reason,when ever we use the mode as a measure of the central tendency of a data set, we should calculate the mode from grouped data.

The mode in Symmetrical and Skewde Distributions.

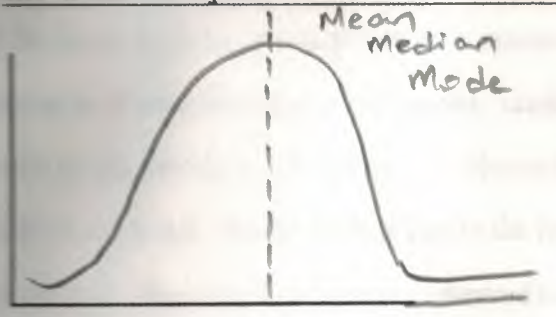

Figure 1
Symmetrical distribution,showing that the mean, median, and mode coincide.

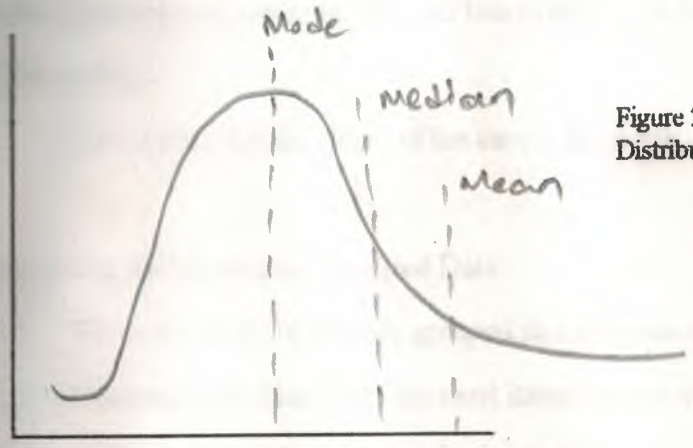

Figure 2
Distribution is skewed to the right

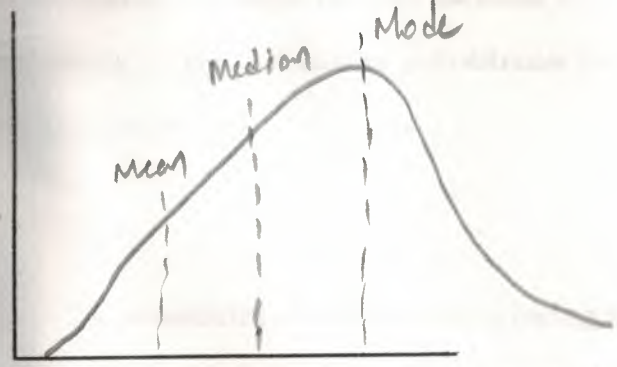

Figure 3
Distribution is skewed to the left

For this reason, when ever we use the mode as a measure of the central tendency of a data set, we should calculate the mode from grouped data.

In the figure 1,where the distribution is symmetrical and there is only one mode, the three measures of central tendency - the mode, median, and mean - coincide with the highest point on the graph. In figure 2, the data set is skewed to the right. Here, the mode is still at the highest point on the graph, but the median lies to the right of this point and the mean falls to the right of the median. When the distribution is skewed to the left, as in Figure 11, the mode is at the highest point on the graph, the median lies to the left of the mode, and the mean falls to the left of the median .

No matter what the shape of the curve, the mode is always located at the highest point.

Calculating the Mode from Grouped Data

When our data are already grouped in a frequency distribution, we must assume that the mode is located in the class with the most items; that is, with the highest frequency. But how can we determine a single value for the mode from this modal class? Two methods are available to vs. The first enables us to estimate the mode from a graph. The second method uses on equation.

To demonstrate these two ways of finding the mode in grouped data Let's us the data in Table 14. First, we can construct a histogram of the data as shown in Figure 12. Then, since the modal class is the tallest rectangle, we can locate the mode in it by:

1) Drawing a line from the top right corner of the tallest rectangle to the top right corner of the rectangle to its immediate left.

2) Drawing a second line from the top left corner of the tallest rectangle to the top left corner of the rectangle to its immediate right.

3) Drawing a line perpendicular to the horizontal axis through the point where the lines drawn in steps 1 and 2 cross.

COMPARING THE MEAN. MEDIAN , AND MODE.

When we work statistical problems, we must decide whether to use the mean, the median, or the mode as the measure of central tendency. Symmetrical distributions that contain only one mode always have the some value for the mean, the median, and the mode, as illustrated in Fig. 9. In these cases, we need not choose the measure of central tendency, because the choice has been made for us.

In a positively skewed distribution (one skewed to the right, such as the one Fig 10), the values are concentrated at the left end of the horizontal axis. Here, the mode is at the highest point of the distribution; the median is to the right of that; and the mean is to the right of both the mode and the median In a negatively skewed distribution, such as in Figure 11, the values are concern trated at the right end of the horizontal axis.

The mode is at the highest point of the distribution, and the median is to the left of that. The mean is to the left of both the mode and the median.

When the population i skewed negatively or positively, the median is often the best measure of location, because it is always between the mean and the mode . The median is not as highly influenced by the frequency of occurrence of a single value as the mode is, nor is it pulled by extreme values as the mean is.

Other wise, there are no universal guidelines for applying the mean, median, or mode as the measure of central tendency for different populations. Each case must be judged independently, according to the guidelines we have discussed.