wxi-aauh-sxy - 2020-08-25

 ashwini shingare - 09:05

Good morning. Today we are going to observe the next part from descriptive statistics. So let's go through the points. That we discussed yesterday. So descriptive statistics as we

a tells us. Information about our data and information statistics help us to make some predictions based on the samples and the population. So in order to know basic properties of data, we can make use of the measurement of central tendency where we can identify mean mode median mid-range and depending on whether is it a group data or ungrouped data? We can apply the appropriate methods to get the values. Colors in my Audible and is my voice, okay? It is.

 ashwini shingare - 09:10

We are going to discuss major of dispersion that how is the spread of our data with the help of range variance standard deviation and 5 number summary and other methods. There are some more You know images we have that is the mean absolute deviation and we are also going to see how the data can be represented graphically with the help of different plots. So before we start with today's session major of discussion, let us discuss the type of more in our data set. So in yesterday's example, we are observed that for the given age values. There were two modes 25 you and 35 you and then we say that we cannot say that whether a data is a symmetric or not. Okay, so whenever we have Say by model or multimodal data, and if we end with such kind of graphs or the distribution, then in that case, we can say that the distribution is a symmetry because in case of bimodal after having the graph of the same we can observe that two halves are the mirror images of each other so mean and morally median are at the same point and then we are having the two different modes giving the exact mirror image of each other. So in that case one can say that the data distribution is symmetric, okay? So there are many more things we can go and read about it. So for that I request if you are interested, you can just take statistics Facebook and you can have in-depth understanding of all these things. Then let us start with today's topic of discussion that is Major of dispersion. So this mainly tells us the spread of our observations in the example and there are various measures of the dispersion. The commonly used are range variance standard deviation and interquartile range. so before we go ahead we need to know that what is the importance of the major of dispersion so we can observe here that there are the two companies whose samples are taken from the different bottles. And the measurement is in liters. Okay, both the samples are having the same mean. So if we find the mean of both the samples is 1 Okay, but then the bottles from company a with more uniform content than the B. Okay. So in this case, we can say that the dispersion or the variability of the observation from the average is less for a than the sample B. Okay. So here In order to purchase the food juice from a company A or B. Obviously customer feel more confident to buy it from the a because it has very less deviation from the mean value. Okay. The same we can observe with the help of graph. So if we plot it on the if we plot it as a scatter plot or the Dot Plot we can observe that in case of a Yeah, so in case of company a the mean over here for both is one and if the file observations are plotted here which are very much close to the mean value. Okay, whereas for Company B mean remains same that is one but the observations are deviating from the mean value. Okay, and that is why most of the times the visualization technique also help us to make certain decision. So here a range of a sample so we have the N samples from x 1 to X N. Okay arranged in increasing order then we can find out the range as Max minus the mean value or in short x n minus x 1 so range identifies the maximum split and it can be misleading if most of the values are concentrated in the narrow band of values. But there are also relatively small number of extreme values. So in that case variance can be used to know the dispersion of the sample values

 ashwini shingare - 09:15

so we can observe we can calculate variance. by having the summation of all the or dispersions like observation minus the mean square of it and then we can get the variance and then the standard deviation is the square root of variance. Now there are certain things. We need to remember. Standard deviation measures the spread about mean okay, and if it is zero, it means that there is no speed and all the observations have the same value. And if it is greater than 0 then we need to observe whether easy to low standard deviation or the high if it is a low it means the data observations tend to very close to the mean and high standard deviation. That is data sphere over a large number of values. so if this things are very clear then given the data and if we want to do the comparison of two data sets, so one can easily comment with respect to this basic properties. Then another important is coefficient of variation. The example we observed about company and Company B where all the samples were having the same major that is the quantities where in the letters but what if we are having the quantities in the different measurements, okay. So in that case one can go with the coefficient of variation. So that tells us the ratio measure of the speed. So whenever we have the quantities or the variables where the there is a different unit of measurement one can go with the coefficient of variation. so we can observe here the example where the mean and variance of height and weights of 10 standards students are given and the question is which one is More wearing than the other so we can observe that height is in centimeters and weight is in kg. So both these groups are having different unit of measurement. So in that in this case, the mean and Valence with the help of these two, we cannot comment anything. So first, we we can make use of the coefficient of balance which will tell us in percentage. What is the spread of the data? and then we from the Reliance we get the standard deviation and from that we just put values in this formula. That is it is the percentage of standard deviation to mean and then we identify t And the coefficient of variation for weight is greater than that of height so we can one can comment that weight of students is more varying than the height. Okay, so these things we need to remember so if we get the data set with the same unit of measurement, we can go with the standard deviation mean but if we get a Values with different unit of measurem then this percentage measurement or the coefficient variation help us to comment on the given data set. The next we have the mean absolute deviation and the absolute average deviation. So these are Used whenever we have some outliers. Okay. So in that case when can use this to avoid the effect of the outline, but if there are the outlines we can also go ahead with the trim mean and then we can go with the regular standard deviation and the reliance.

 ashwini shingare - 09:20

and noted by q1 that And noted by q1. That is And noted by q1 that is 25th percentile. Third is Q3 75th. And the median is Q2. That is the 50th. So this gives us some indication of the center spread and the shape of the distribution. Now in order to know the interquartile range. It is the difference between the third quartile and the first quartile. So that is the major of the spread which tells us that how many What is the range of values okay, which are falling between the q1 and Q3? I cure also helps us to find out the outliers. So here one can go with the file number summary there. the outlier identified the data points which are falling above and below 1.0 IQR. Okay. So those are considered as the outliers. So extreme observations occurring within 1.5. Iql core types are considered as the outliers. So let us observe the file number summary in file number summary. We need to consider the minimum value maximum value along with the median. So the file number summary consists of minimum first quartile q1. Median that is considered as Q2 and Q3 is the third quartile and then we have the maximum value. that is Together all these are considered as fire number summary. Now once the fire number summary is with us one can have a box plot for the same. So here data is represented with a box. Okay, the ends of the Box are the first and the third quartile so we can observe here. So here we have the lower quartile. Here is a upper quartile. Then the median is

marked by a line within the box. So whatever the median we have that is represented as a line in this particular box. Then we can observe the two lines outside the box extended to the minimum and the maximum. So these are known as the whiskers. And they are terminated at the most

observations occurring within 1.5 IQR. So even if you are having the observations beyond the 1.5 IQR and the mean and Max is in that particular range. We do not extend the viscous till that point. Those are extended till 1.5. Iqi. So whatever is the iqure that is Q3 minus q1 into 1.5 till that whichever the data points. We have till the viscous are drawn to indicate the lower and upper extreme values. And then Outlaws are the points which are plotted individually. Beyond this particular whiskers

 ashwini shingare - 09:22

So here we can observe that box plot of unit price data for atoms sold for four branches of particular supermarket and then we can observe that for the branch one. The median price of atoms sold is 80. Okay q1 is 60 Q3 is 100 and there are two outline observations which are plotted individually with value. Say 175 and 202 and these are more than 1.5 times the IQR so Even if we observe that 202 is the maximum value the viscous are not extended till 2 0 2 the viscous are extended till 1.5 into IQR. Okay, and same is the case for the minimum. So at the lower end, there are no outliers. So whatever was the minimum till that the viscous are drawn and then one can observe the Outlast for the given branch and then for other two branches, there are no outline observations. Yeah, now let's take two three

to complete this particular example. So yesterday for this we have identified mean mode median and today just identify the first quartile third quartile. Give the five number summary and if possible show the box plot of the data. Okay. So once you are ready with a box plot, you can turn on your camera and we can go through the box plot for the given data set if possible. Okay, so let us work on this example and then I will move with the next topic.

 ashwini shingare - 09:25

Okay. I hope you have done with some. Points here. I will go ahead with the next point. Okay, so there are different ways to represent the data graphically to have the basic statistical descriptions. So we have observed that we can make use of box plot for the graphic display of FIU number summary.

 ashwini shingare - 09:26

similarly we can have
Similarly, we can have the histogram

 ashwini shingare - 09:26

quantile

 ashwini shingare - 09:27

So can observe the first graph over here has two modes and the next graph that we have here is a unimodal. Okay, so one can make use of one or more visualization technique to more no more about the data set. Scatter plot tells us about the correlation between the two variables. So if we are having a bivariate data, and we need to know that are there any clusters in the data set or are there any outline observations then? The scatter plot helps us to identify these things.

 ashwini shingare - 09:30

So Scatter Plots help us to know whether the variables. Are negatively correlated so graph over here in increases as the value of y increases and that's why one can say that the X and Y variables are positively correlated. whereas this graph hole here indicates the negative correlation because as there is a increase in the value of x there is a decrease in the value of y and that is why this is considered as a negative correlation. And then these three graphs over here represented that there is no correlation between the observations. Because this does not have any. Feature observed like the graph a and graph B over here. The next is the quantile plot, which is used in case of the univariate data distribution. That is if we have only one variable. This tells us the behavior as well as unusual occurrences in the data. So here the values are sorted in the increasing order and each observation is paired with a percentage fi so we can observe here the F values are plotted on the x-axis and then we have the unit price on the y-axis. So here XI is graphed against the FI values. So this tells us that approximately if I percentage of data values are below the X a value.

 ashwini shingare - 09:32

So we can observe here. The median is approximately somewhere 80. So one can comment that 50% of data values are below 80. Okay, so likewise one can have the reading of the contact plant if we have the distribution or the units of atoms sold at different branches, then we can have the quantile control plot so we can observe here that the unit size of atoms sold at Branch 1 versus Branch 2 for each contact and we can observe here that you need price of atoms sold at Branch 1 paint to be lower than those sold at the branch 2. Also here the q1 median and Q3 are plotted. So one can observe that the 50% of atoms sold at Branch one where less than 78 so we can just find out take the projection of medium for the branch one. So it is somewhere 78 or 79 and while 50% of atoms sold at Branch 2 there less than 85. So if we take the projection of median on the branch to somewhere it lies between the 85 so If we are having the values with respect to different branches, then in order to identify or in order to do the comparison of the different branches one can have the quantile content plot. So that is it about the measure of dispersion so we can make use of range where I standard deviation. And the most popular way is the interquartile range 5 number summary and that is represented using the box plot. So now let's go to some mcqs.

 ashwini shingare - 09:33

So first one is if the interquartile range is 0 You can conclude that. The range must also be 0. the mean is also 0 at least 50% of the observations have the same value. And the last one is all the observations have the same value. Kitchen, can you answer this?

KISHAN PARTANI - 09:33
like this opportunity all the observations


ashwini shingare - 09:33
Okay.


ashwini shingare - 09:34
Pretty what do you think? Is this answer, correct?

PREETI - - 09:34
Ma'am, if the integral range is 0 then main should also be 0.


ashwini shingare - 09:34
Okay. Any other comment?

VIKRANT PATIL - 09:34
Ma'am answer should we see?


ashwini shingare - 09:34
At least 50% of the observations have the same value, okay?

VIKRANT PATIL - 09:34
Because ma'am, we are calculating 25% in 1970 percent area in interquartile range. So that both


ashwini shingare - 09:34
Okay.

VIKRANT PATIL - 09:34
cannot comment about rest value and it may be same or even it may not be same.


ashwini shingare - 09:35
Okay. Any other justification? Can you come back and comment on your answer d?

KISHAN PARTANI - 09:35
The land as the interquartile ranges 0 that means the Q3 and q1 must have the same value. That is the median of the lower range as well as the median of the upper and should be equal that may be

TEJAS PRADHAN - 09:35
namsik

KISHAN PARTANI - 09:35
for either C or D.

TEJAS PRADHAN - 09:35
data sorted and if you are having the median of the lower data, which is the 25% and the median of the


ashwini shingare - 09:35
yes, so

TEJAS PRADHAN - 09:35
bound, that is


ashwini shingare - 09:35
yes, D should be correct because as you said if we

TEJAS PRADHAN - 09:35
means that


ashwini shingare - 09:35
50% and then if we

T TEJAS PRADHAN - 09:36
that are also

 ashwini shingare - 09:36
those will be the different values,

T TEJAS PRADHAN - 09:36
So

 ashwini shingare - 09:36
Okay. So the correct option here is D that all the observations have the same value then only we can get the Q 1 3 same.

S You - 09:36
Remembered let's say we have 100 values and right starting from 25th value till it's 75th value. The values are same in this case. Our q1 will be let us say x and Q3 will also be X. So the range will be 0 like Q 3 minus 1 will be 0 so here

 ashwini shingare - 09:36
Yes.

S You - 09:36
but 50% that is. from 25 to 75 that 50%

C CHARMI MEHTA - 09:36
Yes.

S You - 09:36
will also give you the range 0.

 ashwini shingare - 09:36
Okay.

 ashwini shingare - 09:36
Yes.

 ashwini shingare - 09:36
Yes.

P PREETI - - 09:37
Ma'am, like integral to measures the mean central tendency. So it will not affect others. So distribution Matlab 0 means no variability in the

 ashwini shingare - 09:37
Are you saying about the mean and media?

P PREETI - - 09:37
data, but Center of

 ashwini shingare - 09:37
Okay.

P PREETI - - 09:37
be anywhere now. Yes, ma'am.

 ashwini shingare - 09:37
Yes. Okay. So for this question, we will have some data set. Okay, we will work on that and then we can conclude about the answer over here. Okay, so

P PREETI - - 09:37

because

 ashwini shingare - 09:37

option C and D. So I

P PREETI - - 09:37

the 25th and

 ashwini shingare - 09:39

get the data set and then we can conclude for this particular question. Okay. Okay.

let's go ahead with the next identify which of the following is the major of dispersion. So it's quite easy. One more thing. I am observing that very few of you have the profile photo for the vit.edu. So if you can have your recent photograph it will help all the faculty is to identify you and recollect your faces, so that will really help. Okay, so here we can see that. Interquartile range is the measure of dispersion. So let's go ahead with the next.

 ashwini shingare - 09:39

What is primary

 ashwini shingare - 09:40

set of data for which the standard deviation is 0
the major of dispersion. So it's quite easy. One more thing. I am observing that very few of you have the profile photo for the vit.edu. So if you can have your recent photograph it will help all the faculty is to identify you and recollect your faces, so that will really help. Okay, so here we can see that. Interquartile range is the measure of dispersion. So let's go ahead with the next. what is primary characteristics of a set of data for which the standard deviation is 0 Sushant can you answer this?

 ashwini shingare - 09:42

Aditya patil Or you can post your answer in chat box. You can post your answer in chat box for the first one. regarding the characteristics of a data set with standard deviation 0 Yes, so play lots of and most of you have given the correct answer that all values of the variable have the same value. Yeah. Then let us go with the next one. The median is a better major of central tendency than mean if the variable is discrete the distribution is cured. The variable is continuous. The distribution is symmetric.

 ashwini shingare - 09:42

Okay, so I hope most of you are. Giving answer for second one as well. So that is B that is the distribution is cured. So in McQ now, you have a lot of practice first. We need to like do the 50/50 the options which are not at all related and then remaining we can get the correct answer. Yes, now. I want you to read this question carefully and you can answer.

 ashwini shingare - 09:46

okay, so here got the answer for The correct statement and most of you are saying that option D is correct. Yes, so all of the statements given are correct.
That is 1/4 of the members lost less than 2 pounds because first quartile is 2 the middle 50% of the members lost between 2 and 8.5. Yes. Because we have to observe first quartile and the third quartile. The most common weight loss was four pounds that is with respect to more. So with more one can comment the statement three and that's why the option D is correct that all of the above are correct. Now this question is based on box plot.

 ashwini shingare - 09:48

Yes now for this we got the correct answer. That is C2 or the option b. Okay. So in course C2 students perform better. Okay. Now in order to get the answers of first I will give you some values and then I want you to just find out the file number summary or the IQR and then comment on the first one. If the interquartile range is 0 what we can conclude so let me type some values in chat box. Okay.

 ashwini shingare - 09:48

Okay, so I've given only six values. 50% of the observations are same. Okay, and then the remaining 50% are having the different values. And then let's take the one where all the observations are having the same value. All the observations are having the value 10 or 12.

B BALAJI PADAMWAR - 09:48

My mental Fountain range is not 0 now here.

ashwini shingare - 09:48
where the first one here the interquartile range is 0

ashwini shingare - 09:49
Yes, in this case. It won't be 0 but then you have to find out in which situation it will be 0. That is what we have to identify.

B BALAJI PADAMWAR - 09:49
Okay.

ashwini shingare - 09:49
So if we are saying that at least 50% observations have the same value then in that case are we going to get interquartile in 0 so that we have to verify basically?

C CHARMI MEHTA - 09:49
Mama written an example in that interquartile range is 0 so we can comment that 50% observations have same value.

ashwini shingare - 09:49
Okay. I think it's more than 50% now Charmy.

C CHARMI MEHTA - 09:49
Yes, ma'am. but

ashwini shingare - 09:50
Yeah. Okay.

C CHARMI MEHTA - 09:50
Nancy in this also example 10 11 11 12 Or 50% should we actually two and half numbers which is nothing so we can say two or three numbers. So at least 50% are the same and that is

ashwini shingare - 09:50
Okay.

C CHARMI MEHTA - 09:50
0.

ashwini shingare - 09:50
So any other comments on this?

C CHARMI MEHTA - 09:50
M Mam also a 1.5 times. The interquartile range will also always be 0 so all points other than the observations which have same value. They will all be outside of the enter outside 1.5 interquartile M
Mam also a 1.5 times. The interquartile range will also always be 0 so all points other than the observations which have same value. They will all be outside of the enter outside 1.5 interquartile range, so it will not be a part of the box diagram

ashwini shingare - 09:50
Okay.

C CHARMI MEHTA - 09:51
So base. So even that point that all observations are same value should also not be wrong because we will anyway not consider all the other observations, but obviously they exist so it's not 100%

ashwini shingare - 09:51
Okay, so here if we need to choose multiple options we can go with C or and D. But then if it is kind of McQ, you have to answer only one option. That is correct. Then in that case I one should prefer going with d because then you are not sure about wallot observations are and then a generalized one can say that when all observations are same definitely the interquartile range is 0 in case of C. It depends on what range of values we have. Like I gave one example, and then you have given me the another example. So with that we cannot surely say that if the 50% observations have same value we will end up

B BALAJI PADAMWAR - 09:51
Excuse me.

 ashwini shingare - 09:51
interquartile range as 0.

B BALAJI PADAMWAR - 09:52
we should go with C

 ashwini shingare - 09:52
Is this clear

B BALAJI PADAMWAR - 09:52
have given the

 ashwini shingare - 09:52
Yes.

B BALAJI PADAMWAR - 09:52
InterContinental range is

S SAKSHI OSWAL - 09:52
It right.

B BALAJI PADAMWAR - 09:52
in your case you are taking it
in your case, you are taking it vice versa. So means like in that case

C CHARMI MEHTA - 09:52
Yes.

B BALAJI PADAMWAR - 09:52
InterContinental range is

C CHARMI MEHTA - 09:52

B BALAJI PADAMWAR - 09:52
now so it's an

C CHARMI MEHTA - 09:52
Huh?

 ashwini shingare - 09:52
Okay.

 ashwini shingare - 09:52
Okay, but again.

B BALAJI PADAMWAR - 09:52
take

S SUYOG PAWAR - 09:52
same

 ashwini shingare - 09:52
because

B BALAJI PADAMWAR - 09:52
InterContinental range is zero, so

 ashwini shingare - 09:52
values

B BALAJI PADAMWAR - 09:52
that c will be

 ashwini shingare - 09:52

B BALAJI PADAMWAR - 09:52
say these are subset of C.

B BALAJI PADAMWAR - 09:53
No, man, no need not any value like or miss you. I think your missing of places in fact which place the values are there, but

 ashwini shingare - 09:53
Okay.

B BALAJI PADAMWAR - 09:53
But I means like you are missing your case. The condition is

C CHARMI MEHTA - 09:53
and

B BALAJI PADAMWAR - 09:53
are thinking that if 50% values are given same then your you are checking InterContinental range means this is not

C CHARMI MEHTA - 09:53
decays. That is

B BALAJI PADAMWAR - 09:53
rectangle but rectangle is not a square is that condition is just

C CHARMI MEHTA - 09:53
definitely 50% observations will be same but also when all observations are same interquartile range will be 0 so for that these statement. The inverse is true. We cannot comment if we have only the interquartile ranges 0

 ashwini shingare - 09:53
Okay.

 ashwini shingare - 09:53
okay, so

 OJAS NATU - 09:53
a man in

 ashwini shingare - 09:53
Yes, go ahead.

 OJAS NATU - 09:54
we'll have even number of data like digits or numbers. So in that case, we will find the first and the third quartile as the mean of the two numbers so see if you assume that the data is 10 11 11 12, then the first quartile will be something like Or 10.5 and a third call time will be 11.5 while the middle quartile will be 11. But in that case. We can't like assume that. the this thing

 ashwini shingare - 09:54
Okay. Okay, so for this.

 OJAS NATU - 09:54
the