

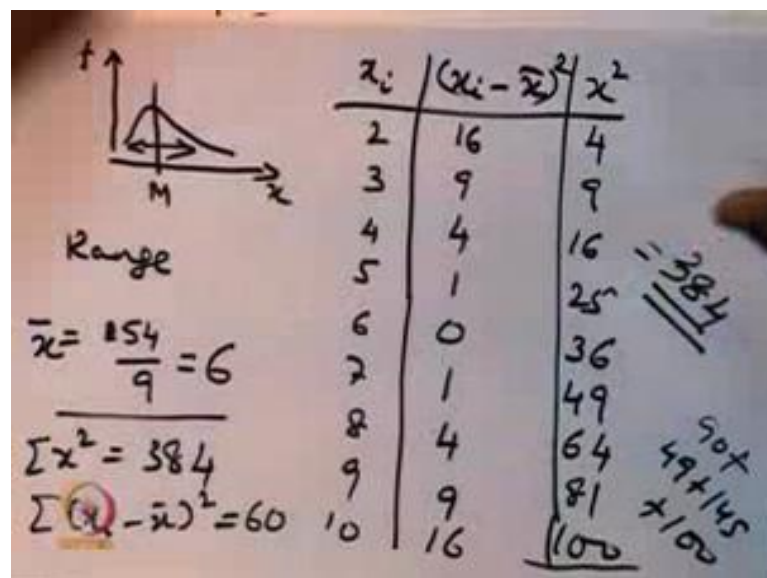
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institution of Technology, Bombay

Lecture – 06
SME, Z-Score, Box plot

Hello and welcome to today's lecture. So, we will continue from where we left off in last lecture. We were discussing the relative, we are discuss the relative advantages or disadvantages of computing matrix of mean median and mode and then we went on to starting today you know discuss ways and means of quantifying the dispersion in the extent of dispersion in the data.

So, just a quick recap between mean median and mode so, as you know mean is much more sensitive to the presence of outliers in the data mode is of course, not sensitive median is also not sensitive mode is primarily used while describing large sets of data while mean and median have can be used for describing both small and large sets of data.

(Refer Slide Time: 01:06)



So, but as we discussed that if you have a population like this you know this is x this is your frequency. So, just computing one value as mean or median or mode is not sufficient there has to be a way of capturing the dispersion in this data and there are various metrics for you know computing this the most simplest is the range and range is nothing but essentially the defined as the maximum and minimum of this data.

(Refer Slide Time: 01:33)

Measures of Variability: Range

Range = maximum – minimum

100, 95, 40, 75, 60

Range should be looked at w.r.t minimum or maximum

So, as you can clearly see in this case the minimum is 40 the maximum is hundred. So, the range is 60. So, as you can also see that in this case, you know your value. So, range has to be thought of in the context of the minimum or the maximum or the average. So, you can have a value of one it would not change your maximum, but will drastically changes the mean. So, you have to think of range in the context of whatever is the minimum and the maximum value.

(Refer Slide Time: 02:05)

Measures of Variability: Standard Deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2}$$
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Variance = square of standard deviation

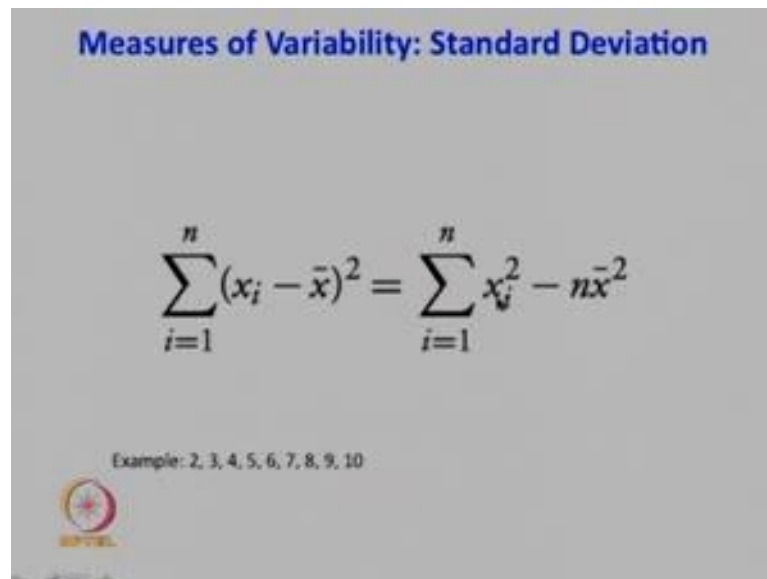
We then discussed about standard deviation. So, standard deviation is essentially the sum of the squares of x values from the average either the population average or the sample average the notable difference is after you add them up you square them up you add them up. So, it does not matter whether the value was lesser or greater than particular average and then you divide by the total number of observations in the population, but in case of calculating the standard deviation of a sample you divide by n minus 1.

(Refer Slide Time: 02:42)

Measures of Variability: Standard Deviation

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Example: 2, 3, 4, 5, 6, 7, 8, 9, 10



So, this is the most important difference between these 2 cases we then came to this by you know we had derived this expression that when you do this x i minus x bar whole square you can this is as good as it is say exactly the same as calculating this particular term which is summation x square minus n times x bar square. So, let us just take this particular example and see whether this holds good or not. So, let me write down the values of x i x i minus x bar. So, x i is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, my x bar I can calculate as. So, 2 is going to be 12 plus 12 plus 12 plus 12; 40, 54 by 1, 2, 3, 4, 5, 6, 7, 8, 9, 9 is equal to 6 my x bar is 6. So, I will plot 2 terms x square and maybe x i minus x bar whole square.

So, this is 4 x i minus x bar is 4 square is 16, 3 is 3 square is 9, 4 is 2 square is 4, 5 is 1 square is 1, 6 is 0 7 gives me 1, 8 gives me 4, 9 gives me 9, 10 gives me 16 and for x square is 3 square 9, 16 25, 36, 49, 64, 81, 100. So, I have to add up all these x bar squares. So, let us see what this gives us 13 and 16: 29, 29 and 25; 54, 54 and 36 is 90,

90 and s, you have 90 plus this is 90 plus 49 plus this is 145 plus 100. So, which is 9; 90 is so 1 94 to 83, 84. So, x summation x square, summation x square is roughly 3 8 is 384 and summation x i minus x bar whole square is 25 plus 4 25 plus 4, 29, 30, 31 plus 35, 9, 44, 60.

(Refer Slide Time: 05:35)

The image shows a handwritten derivation on a chalkboard. On the left side, the following steps are written:

$$\begin{aligned} \sum x^2 &= 384 \\ \sum (x_i - \bar{x})^2 &= 60 \\ \sum x^2 - n\bar{x}^2 & \\ &= 384 - 9 \times 6^2 \\ &= 384 - 324 \quad \begin{matrix} 36 \\ 9 \end{matrix} \\ &= 60 \end{aligned}$$

On the right side, the following relationships are written:

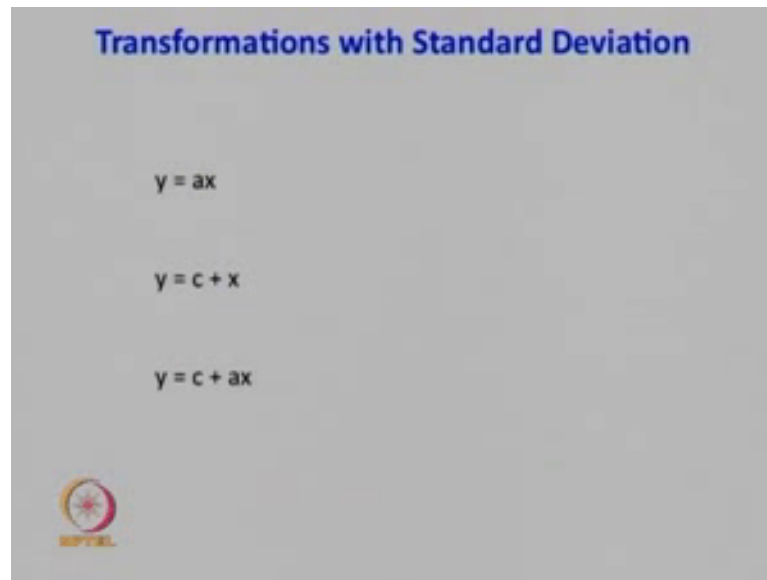
$$\begin{aligned} y &= ax \\ \Delta y &= a \Delta x \\ y &= c + x \\ \Delta y &= \Delta x \\ y &= c + (ax) \\ \Delta y &= a \Delta x \end{aligned}$$

So, I know this value. So, let me again write it down here my x square is 384 and my summation x i minus x bar whole square is 60. So, this is 60 now summation x square minus n times x square average is equal to 384 minus total number is 9 into x bar is 6. So, square. So, 36 and 9, 36 into 9 36 into 9 is 9, 6, 54, 324. So, this is 384 minus 324 is equal to 60. So, just this and this is the same you have confirmed these 2 values are the same now of course, when you are doing it by hand it does not have any value in doing this in that case doing the x by minus x bar root is easier, but in while you are doing these calculations in a computer by writing a program this is much more advantageous.

So, we then discussed about transformations. So, we had come. So, if you have y is equal to a x then you can have s y will simply be a times s x if you have y is equal to c plus x s y is simply equal to s x and in the general case when y is equal to c plus a x s y will still be equal to a s x. So, bottom line is when you have a pre factor which multiplies your variable x to generate your variable y then that also gets reflected in the standard deviation calculation, but when you have a constant been added it has no contribution to the variation. So, let us you know think of the practical significance of your data of what

the standard deviation tells us and that brings us. So, let say if we if you know this is your data x and frequency. So, we want to know. So, you know let us just say that this is your mean and you have a standard deviation.

(Refer Slide Time: 07:24)




Transformations with Standard Deviation

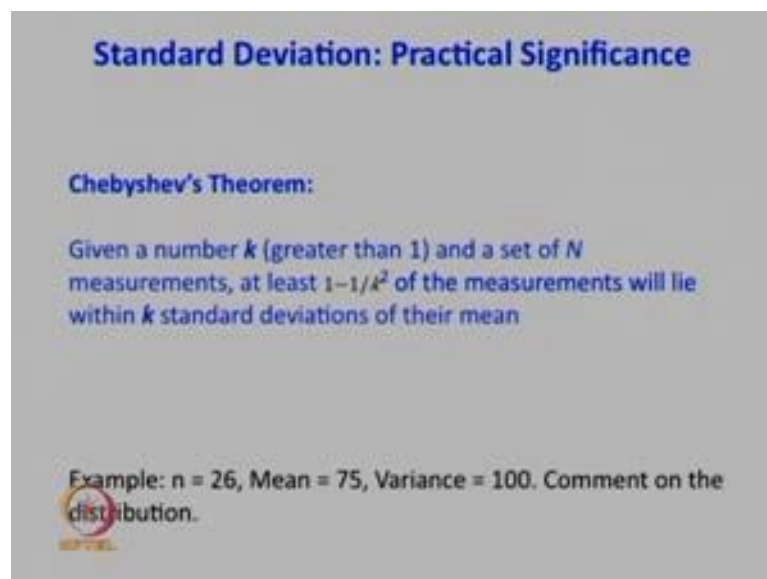
$y = ax$

$y = c + x$

$y = c + ax$



(Refer Slide Time: 07:26)




Standard Deviation: Practical Significance

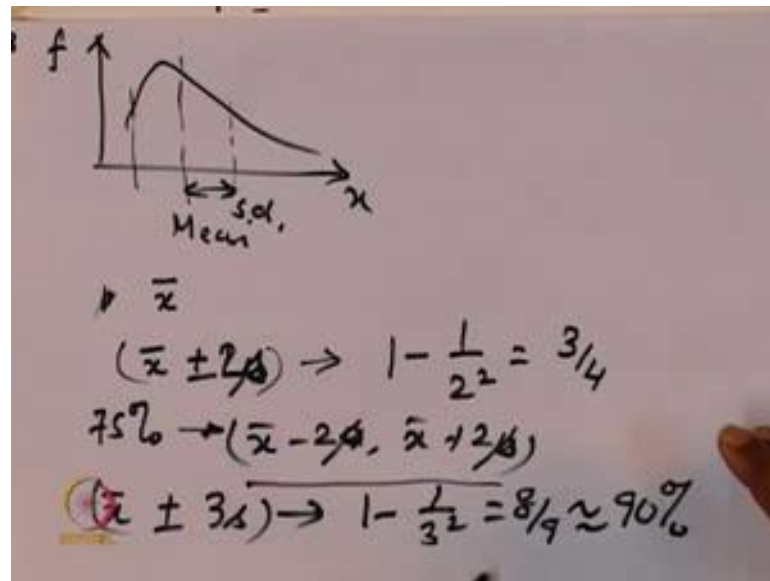
Chebyshev's Theorem:

Given a number k (greater than 1) and a set of N measurements, at least $1 - 1/k^2$ of the measurements will lie within k standard deviations of their mean

Example: $n = 26$, Mean = 75, Variance = 100. Comment on the distribution.



(Refer Slide Time: 07:28)



So, let say this may be your standard deviation. So, we want to know what this standard deviation conveys. So, one theorem I will turn to it is very important it cause Chebyshev's theorem. So, what it states given a number k greater than one and a set of n measurements at least $1 - \frac{1}{k^2}$ of the measurements will lay within k standard deviations of their mean. So, what it tells us is if you have μ or if you have \bar{x} as your; you know mean then $\bar{x} \pm k \cdot \mu$. So, if n is or plus minus 2 times standard deviation. So, we will contain how much of the population $1 - \frac{1}{2^2}$ square is equal to $\frac{3}{4}$. So, 75 percent of the population will lie between $\bar{x} - 2\mu$ and $\bar{x} + 2\mu$ 75 percent of the population will lie between $\bar{x} - 2s$ and $\bar{x} + 2s$ sorry; this is this is $\sigma = 2s$. So, this is this is standard deviation s .


So, 75 percent of the population will lie between $\bar{x} - 2s$ and $\bar{x} + 2s$ similarly if you put k is equal to 3 then $\bar{x} \pm 3s$ you have the value $1 - \frac{1}{3^2}$ equal to $\frac{8}{9}$ which will roughly is 90 percent of your data will lie between 3 standard deviations of your population. So, this is Chebyshev's theorem let us test it with the following example.

(Refer Slide Time: 09:28)

Testing Chebyshev's Theorem

26.1	26	14.5	29.3	19.7
22.1	21.2	26.6	31.9	25
15.9	20.8	20.2	17.8	13.3


Mean = 22



So, you have these values 26.1, 26 14.5, 29.3. So before we go this I think I skipped this value. So, let us come to this example. So, imagine you have 26 observations and the mean of the population is 75.

(Refer Slide Time: 09:50)

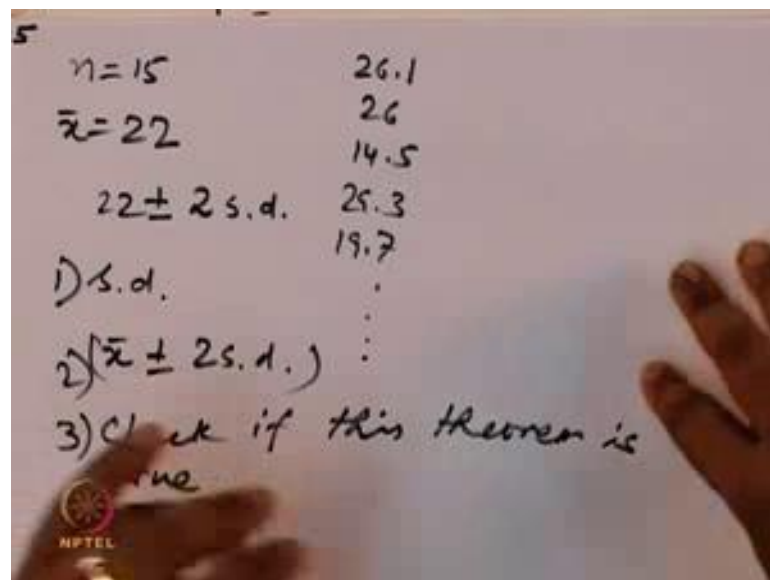
4

$$n = 26$$
$$\bar{x} = 75$$
$$\text{Variance} = 100$$
$$\text{s.d.} = 10$$
$$\bar{x} \pm 2 \text{ s.d.} \Rightarrow (75 \pm 20)$$
$$\underline{55 - 95} \rightarrow \frac{3}{4} \text{ of pop.} = \frac{3}{4} \times 26$$
$$\bar{x} \pm 3 \text{ s.d.} \Rightarrow (75 \pm 30)$$
$$\underline{45 - 105} \rightarrow \frac{8}{9} \times 26$$


So, you have n is equal to 26 x bar is equal to 75 and variance equal to 100. So, standard deviation is equal to 10. So, it tells that if n is equal to 26 then. So, between x bar plus minus 2 times standard deviation which would mean that 75 plus minus 20.

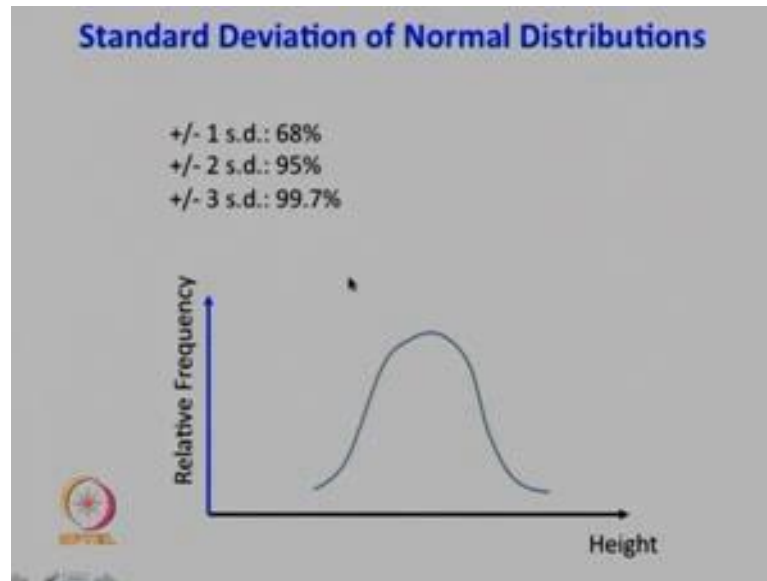
So, between the values 55 to 95 you will have 3 4th of the population that is equal to 3 4th into 26 observations will actually be within this range similarly \bar{x} plus minus 3 standard deviation would give me 75 plus minus 30. So, between the numbers 45 to 105 you will have 8 by 9 into 26 number of the population will lie within this number. So, this allows us to think of how much of the data will lie will represent bulk of the data how much of the fraction of the population represents bulk of the data let us test this particular example as well. So, you have the following population and you can calculate the mean of this population to be roughly 22. So, the numbers of observations are 1, 2, 3, 4, and 5.

(Refer Slide Time: 11:24)



So, in this case n is equal to 15 your \bar{x} is equal to 22. So, let us say. So, it what Chebyshev's theorem predicts is between 22 plus minus 2 times the standard deviation. So, we have not calculated the 2 standard deviation. So, let us see, you have 26.1; 26, 14.5. So, we have to calculate the standard of deviation for this particular problem I will leave it to you for to do it, but you can test. So, you can find out; so you calculate the standard deviation step one calculate \bar{x} plus minus 2 times standard deviation and check. So, this is my and step 3 check if this theorem is true.

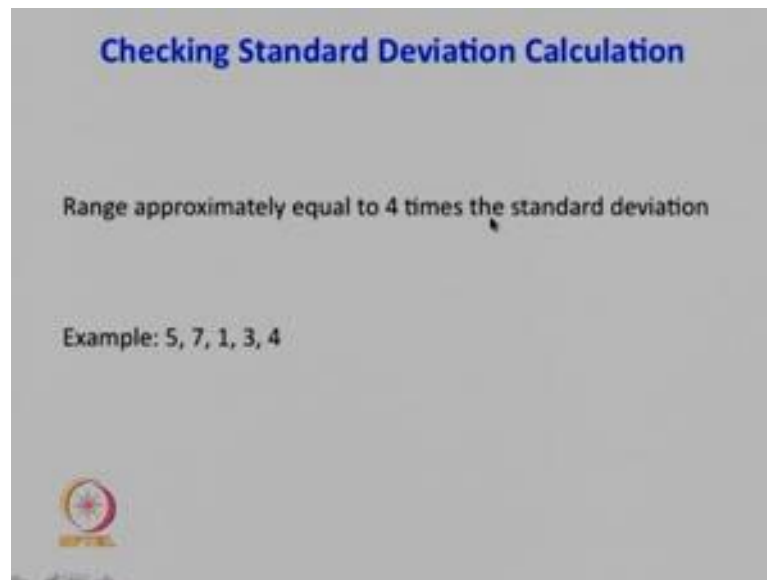
(Refer Slide Time: 12:29)



So, for distributions which are different than Chebyshev's, but many times you obtain mount like distribution. So, where there is a clear peak there is a tendency for the data to accumulate towards the center of the distribution this is a mount type of distribution and very often the term normal distribution is used. So, what for these kind of distributions what is known as within one standard deviation of the data 68 percent of the data is there within 2 standard deviations of the data 95 percent of the data is there and within 3 standard deviations of the data 99.7 percent of the data is there.

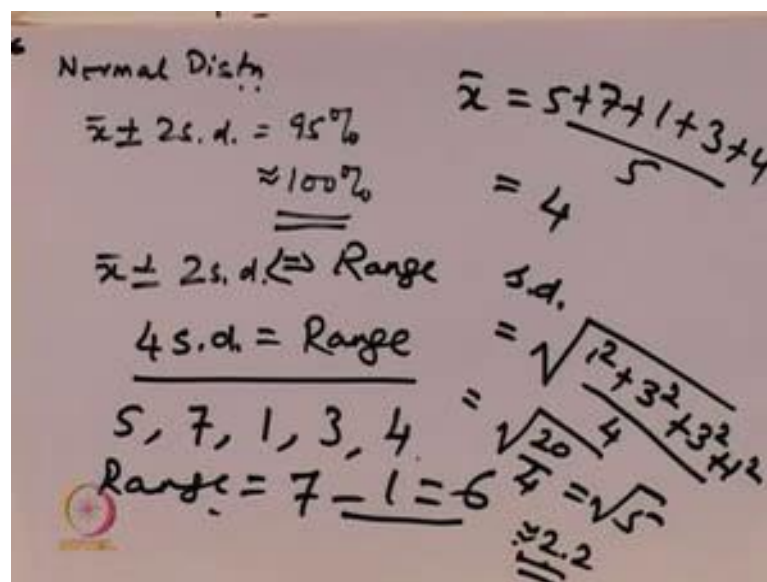
So, you can see that compared to what Chebyshev's theorem predicts which is that within 2 standard deviations you have 75 percent of the data. So, this clearly tells you that Chebyshev's theorem is a much more conservative estimate; that means, it under estimates how much population is because it makes no assumption is to how the population is distributed, but for these particular normal distributions you see that instead of 75 percent predicted by Chebyshev's theorem 95 percent of the population falls between plus minus 2 standard deviations. So, this brings us to a very good exercise that. So, for all practical purposes you can consider 95 percent of the data to be the entire data itself.

(Refer Slide Time: 13:55)



So, when you calculate standard deviation, you can check whether the value of standard deviation that you have calculated is approximately right or not the exercise is very simple you calculate the range.

(Refer Slide Time: 14:12)



So, what normal in case of normal distribution in case of normal distribution \bar{x} plus minus 2 standard deviations is 95 percent of the data. So, we all which we can approximate as hundred percent, if this is 100 percent of the data; that means, the difference; so \bar{x} plus minus 2 standard deviation. If you will simply be is equal to the

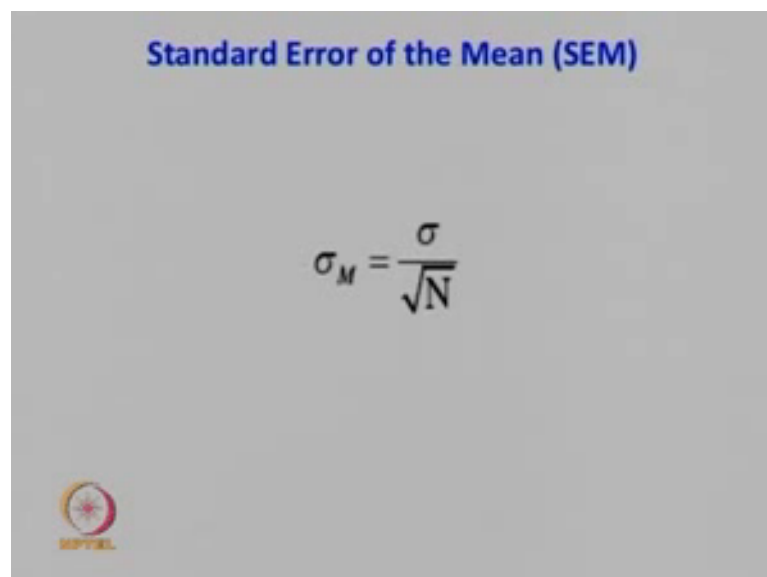
range. So, the range will give me the 2 extremes. So, this would mean that roughly my 4th standard deviation is equal to the range. So, that is a very easy you know way to test.

So, let us take the following example let say if you have 5, 7, 1, 3, 4, my range is equal to maximum minus minimum 7 minus 1 is equal to 6 what is my standard deviation my average is. So, my \bar{x} is 5 plus 7 plus 1 plus 3 plus 4 divided by 5, 12, 13, 16, and 20. So, \bar{x} is 4. So, my standard deviation is equal to square root of. So, 5 minus 4 1 square plus 7 minus 4 3 square plus 1 minus 4 3 square plus 3 minus 4 1 square plus 4 minus 4 0 square right by one 2 3 4 5 we have square root of 4 is equal to square root of. So, 9 plus 1; 10, 10 and 10 20 by 4 is equal to square root of 5 which will give you a value of 2.2.

So, your range is 6 you have you know standard deviation is 2.2; that means, that 2.2 into 4 is 8, they are comparable. So, you know that this is this kind of is probably right, but if you had gotten a value of standard deviation is 10 that would have you know Shirley told you that there is something wrong in the calculation. So, this is the; you know this is not exact, but it gives you a way of testing whether the value of calculated is reasonable or not.

Now, when we are you know handling biological data right let say you do an experiment where you measure how far cells move you have huge heterogeneity in the data. So, your error bar is going to be humongous.

(Refer Slide Time: 16:38)



So, one way of reducing this error bar of plotting the error bar is not by plotting the standard of deviation, but by plotting the standard error of the mean and the standard error of the mean is defined by $\sigma_M = \frac{\sigma}{\sqrt{N}}$. So, σ is the standard deviation you divided by square root of N . So, this clearly tells you when your N increases then σ_M will decrease. So, if N is 4 for example, σ_M will be $\frac{\sigma}{2}$ half of σ if N is 100 then σ_M will be $\frac{\sigma}{10}$. So, your standard you know error bar will decrease, but again this is a measure this tells you how confident are you of calculating the mean.

So, one more thing which is very important is the concept of zee score or relative scoring imagine a class of students have taken an exam and always you want to know how you are positioned with respect to the entire population or how we have performed relative to how the class has performed and zee score. So, zee score is defined by given score x minus \bar{x} by standard deviation.

(Refer Slide Time: 17:45)

7

$$z\text{-score} = \frac{x - \bar{x}}{s.d.}$$

$$z\text{-score} = \frac{30 - 25}{4} = 1.25$$

1, 1, 0, 15, 2, 3, 4, 0, 1, 3

↑

$$z\text{-score} = \frac{15 - 3}{3} = 4$$

$$\bar{x} = \frac{30}{10} = 3$$

$$s.d. = 4$$

$$s.d. = 5$$

(Refer Slide Time: 17:55)

Relative Standing: z-score

Z-score = $(x - \text{Mean}) / \text{Standard Deviation}$

Example: Mean = 25, Standard Deviation = 4; $x = 30$

Z-score > 3 is an outlier!

Example: 1, 1, 0, 15, 2, 3, 4, 0, 1, 3

So, zee score is defined by this particular example. So, if we take this example if my mean is 25 my standard deviation is for then if x equal to 30. So, I get a zee score of 30 minus 25 by 4 which is roughly one 30 minus 25 is rough. So, it is actually 1.25 is equal to 1.25, but for practical purposes it is roughly one. So, one thing, the zee score gives you way of testing whether a particular value is an outlier or not and this is an empirical in nature, but what is considered is if your zee score is greater than 3 then that particular value is an outlier let us take the following example. So, you have these values. So, your data is 1, 1, 0, 15, 2, 3, 4, 0, 1, 3. So, for this particular value the zee score.

Let us say we of course, have to calculate the \bar{x} \bar{x} is 2 plus 15, 17, 22, 26, 30, 3, 4, 5, 6, 7, 8, 9, 10. So, \bar{x} is 3 standard deviation. So, in most of the cases will be 0. So, your standard deviation will be you know let say ordered 3 then my zee score of 15 will give me 15 minus \bar{x} is 3 by 3 is giving a value of 4, but I think the standard deviation will probably be standard deviation will probably be order 4 or higher 5 in that case. But what you can clearly see is with this calculation it may not be an exact calculation, but in this case your z score of 15 just by looking at the data when you see most of your points are within 5 or 4. And then this 15 value is sticking out then you know that this is really an outlier and one way to do it is to calculate a zee score and see whether this value is greater than 3 if it is greater than 3 you can completely remove this value and then recalculate your statistics.

So, this helps you to remove outliers which completely mask the information that is contained in your data.

(Refer Slide Time: 20:29)

Relative Standing: Percentiles

'p'th percentile is the value which is greater than p % of the measurements

Q_1 : first quartile (at position $0.25 \cdot (n+1)$)

Q_3 : third quartile (at position $0.75 \cdot (n+1)$)

Inter-quartile Range = IQR = $Q_3 - Q_1$

Another important metric which is widely used is the concept of relative standing or percentiles a p th. So, the definition of a percentile is as follows a p th percentile is the value which is greater than p percent of the measurements. So, you can have based on that you can have first quartile which is at position 25 percent. So, it is defined is the first quartile is the value which is positioned at position point 25 slash n plus one accordingly the third quartile is defined at position point 7 times at n plus 1.

(Refer Slide Time: 21:12)

Example: Calculating Quartiles

16, 25, 4, 18, 11, 13, 20, 8, 11, 9

And you can calculate something called the inter quartile range which is nothing with the difference between Q 3 and Q 1. So, again let us take the following example. So, you want to calculate the quartile.

(Refer Slide Time: 21:16)

8

16, 25, 4, 18, 11, 13, 20, 8, 11, 9

↓ Sort

4, 8, 11, 11, 13, 16, 18, 20, 25

$Q_1 = \frac{1}{4} \times (9+1) = 2.5$

$Q_1 = 8 + 0.5 (11 - 8)$

$Q_3 = \frac{3}{4} \times (9+1) = 7.5$

$Q_3 = 18 + 0.5 (20 - 18)$

So, you have the following example 4, 18, 11, 13, 20, 8, 11, 9. So, first thing will of course, we sort 4, 8, 11, 11, 13, 16, 18, 20, 13, 16, 18, 25, 1, 2, 3, 4, 5, 6, 7, 8, 9, 6, 7, 8, 9, 10, you know. So, 4, 8, 11, 13, 11, 11, 13, 16, and then you have 18 then you have 20 and 25.

So, let us say you have the following measurements 1, 2, 3, 4, 5, 6, 7, 8, 9. So, you have total of 9 numbers. So, for calculating Q 1 you do one 4th into 9 plus one which is 2.5. So, it tells you. So, 2.5 is middle of this 8 and 11 number. So, for calculating Q 1 you have to do 8 plus 0.5 times 11 minus 8. So, this is how you will calculate Q 1 for calculating Q 3 we need 0.75. So, 3 4th time 9 plus 1 is equal to 7.5 position. So, for calculating Q 3, what is the 7th position 1, 2, 3, 4, 5, 6, 7. So, 7.5, this is your Q 3. So, it is going to be 18 plus point half into 20 minus 18. So, this is how you will calculate Q 1 and Q 2 and i in Q 1 and Q 3. So, Q 2 is nothing but your median Q 2 is nothing but your mod.


(Refer Slide Time: 23:20)

Example: Calculating Quartiles

The following data give noise levels measured at 36 different times directly outside of Mumbai CST.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90,
83, 87, 75, 114, 85, 69, 94, 124, 115, 107, 88, 97, 74, 72,
68, 83, 91, 90, 102, 77, 125, 108, 65

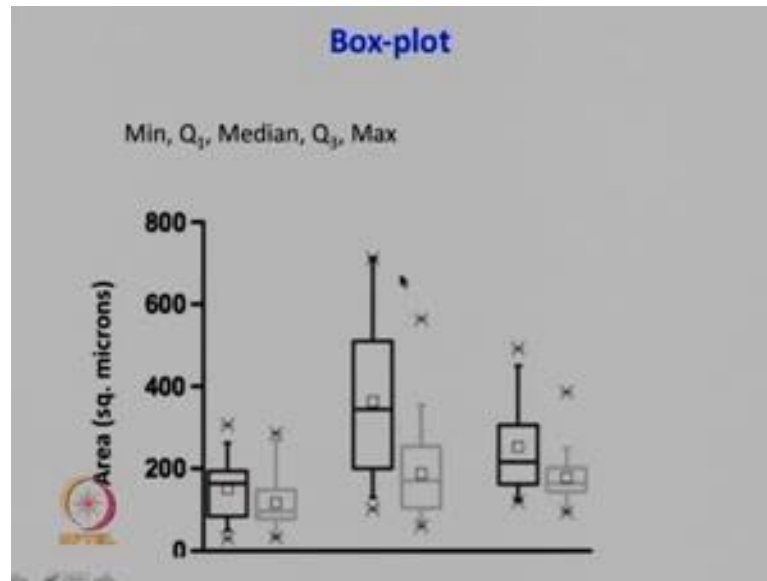
Determine the quartiles.



So, this is another example. So, where you can calculate the same thing, but I am not going to the details, but essentially the same thing you ordered them you find out what this total number of measures you calculate which you know which point is where and then you know n you find out the position and you have to interpolate.

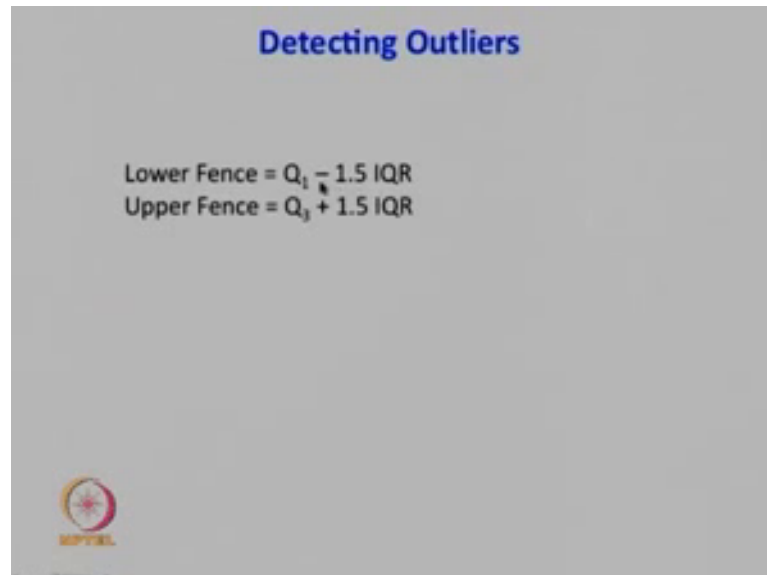
So, depending on if let say hypothetically between these 2 numbers you get a value of position as 2.25 then you have to take the and this is position 2, this is the next position you have to take this position and then 0.25 times the following 2 points. So, what these you know these quartiles once you calculate the quartile one of the way of representing data is using what is widely used is called a box plot. So, this is how a box plot looks, so this value.

(Refer Slide Time: 24:04)



So, you have this entire you know range within which you have this value is your median the square value. So, the square value is your average square value is your average this is your median this is your maximum, but this is your 75 third quartile this is your first quartile first quartile median third quartile this is the maximum and this is the minimum.

(Refer Slide Time: 24:48)



But what you can see is these error bars do not x always extend to the maximum of the minimum because you calculate what is called as the lower fence is defined as Q 1 minus 1.5 times inter quartile range and Q 3 is defined you know upper fence is defined as Q 3

plus 1.5 times inter quartile range. So, when you plot. So, for example, in this particular data set what you said even after you have done that there is this point which is a clear outlier, but for this case the extreme point and where you are you know your buck error bars extend is the same point.

So, this would convey that this point is not an external point. So, in this case also you can see the error bar here you can see the error bar here which is a which conveys the message that this point is an outlier the other thing I wanted to look at here is the square point which shows the mean position. So, what you see is in both these cases of the data your mean and median are very close to each other while for this data for example, the mean is biased towards the upward side.

So, this would mean that you have lesser less number of values which populate this portion of the data and greater number of values which populate this portion of data because of which this is higher. So, when you have the same thing. So, when your population is not outside the outlier as in this case of this case. So, your minimum and the error bar would kind of converge, but in these cases in this case here or in this case here these points are clear outliers which you know which lie much outside the bulk of the data.

So, that is how you know you can detect outliers to find out where is your lower fence where is your upper fence and then you know calculate what is your you know, how you want to plot your box plot.

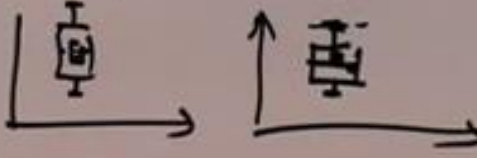
(Refer Slide Time: 26:41)

9
1, 2, 3, 8, 10, 12 IQR = $Q_3 - Q_1$
 $n = 6$ = 3.5
 $Q_1 \Rightarrow 0.25 \times (6+1)$
Position $\frac{7}{4} = 1.75$
 $Q_1 = 1 + 0.75(2-1) = 1.75$
 $Q_3 = 0.75 \times (6+1)$
 $= \frac{3}{4} \times 7 = \frac{21}{4} = 5.25$
 $= 10 + 0.25(12-10) = 10.5$

So, let us take a sample data set let say. So, we do 1, 2, 3, 8, 10, and 12. So, this is your data set. So, so your n is equal to 1, 2, 3, 4, 5, 6. So, your Q 1 is at position. So, position is 0.25 stars 6 plus 1 is equal to 7 by 4 is equal to 1.75. So, my Q 1 is 1.75 positions. So, between 1 and 2 it is going to be 1 plus 0.75 times 2 minus 1 is equal to 1.75. This is the position remember now for Q 3 you have position is 0.75 times 6 plus 1 is equal to 3 4th into 7 21 by 4 which is 5.25. So, fifth number is 10, but only you have 0.25. So, Q 3 is equal to 10 plus 0.25 times 12 minus 10 is 2 so is 5. So, 2 into 0.5, 10.50; Q 3, my IQR is equal to Q 3 minus Q 1 is equal to. So, 10.5 minus 1.75 is I think 8 point. So, 10 8.75 IQR is 8.75.

(Refer Slide Time: 28:27)

10

$$\text{Lower Fence} = Q_1 - 1.5 \times 8.75$$
$$\text{Upper Fence} = Q_3 + 1.5 \times 8.75$$


So, as per this my lower fence has to be $Q_1 - 1.5 \times \text{IQR}$. So, my lower fence my lower fence is $Q_1 - 1.5 \times 8.75$ and upper fence $Q_3 + 1.5 \times 8.75$. So, you can I think none of these points will be outside the range. So, most of these points will lie inside your plot and your box plot if you plot will probably look something like this and so, your median and you know your median. Since each of these values are only repeated once your median will be in between your mean and median will be very close to each other and your error bar would not you know your error bar will encompass all the points because none of the points lie outside so, but it is possible to have plots where your let say your you have an error bar like this. So, where you are you know this point is way outside the maximum to which this goes this is your upper fence which is why you know $Q_3 + 1.5 \times \text{IQR}$.

With that I complete here. So, essentially we have discussed how to calculate standard deviation and then from that what how can you based on mean and standard deviation how to get an idea of how much of the population lies in the bulk and how much of these points are outliers. So, using zee square zee score you can get an idea of which point is an outlier and using box plot you can represent the whole data to show how it looks. So, it of course, conveys the message; which of these points are outliers and not.

With that I thank you for attention, will meet again in the next class.

Thank you.