

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

Lecture - 07
Moments, Skewness

Hello and welcome to today's lecture. I would begin by doing a brief recap of what we have discussed in last class mainly ways of quantifying dispersion in a population or a sample right and one of the widely used matrix for characterizing this variation in the data is using standard deviation.

(Refer Slide Time: 00:42)

The image shows handwritten mathematical formulas on a whiteboard. The first formula is the population standard deviation: $\sigma = \frac{\sum (x - \mu)^2}{N}$. The second formula is the sample standard deviation: $s = \frac{\sum (x - \bar{x})^2}{n-1}$. Below these, numerical values are given: $\bar{x} = 75$, $n = 26$, $V = 100$, and $s \approx 10$. At the bottom, a confidence interval is written as $(75 \pm 10) \rightarrow \frac{3}{4} \times 26$.

You either use a sigma to describe standard deviation of a population and this is given by summation of x minus μ whole square by capital n or s is for the sample the summation of x minus \bar{x} whole square by n minus 1 right. So, again once again note this minus 1. So, when you are doing a sample then it is thought of that by dividing by n minus 1 you get a better estimate of standard deviation of the population. So, what exactly is the practical significance of the standard deviation?

(Refer Slide Time: 01:19)

Standard Deviation: Practical Significance

Chebyshev's Theorem:

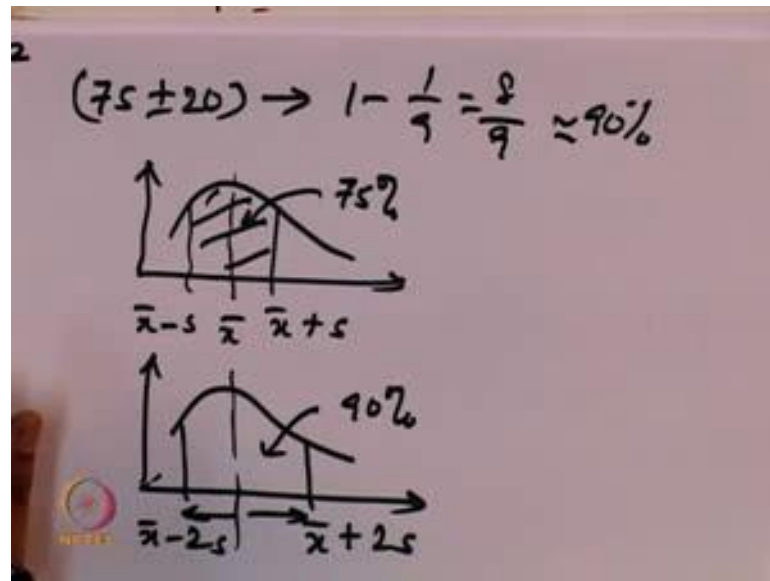
Given a number k (greater than 1) and a set of N measurements, at least $1 - 1/k^2$ of the measurements will lie within k standard deviations of their mean

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

And this we had discussed in last class and which is what Chebyshev's theorem tells us. So, it says that given a number k greater than one and a set of n measurements what you are guaranteed is at least $1 - 1/k^2$ proportion of the measurements will lie within k standard deviations of their mean right. So, if I substitute k equal to 2 then that would become $1 - 1/4$ which is 75 percent of the measurements are expected to lie within one standard deviation of the mean which means 75 percent of the data will lie between μ or $\bar{x} - \sigma$ and $\bar{x} + \sigma$.

So, as an example you have n equal to 26 mean 75 variance is 100. So, in this case if I have \bar{x} is equal to 75 and n is equal to 26 variance is equal to 100. So, I can roughly calculate s is approximately equal to 10. So, within 75 plus minus 10 you have three-fourths of the population. So, three-fourths of the number of variables will actually lie within this range which is minus 65 to 85 similarly I can do the same thing for 2 standard deviations.

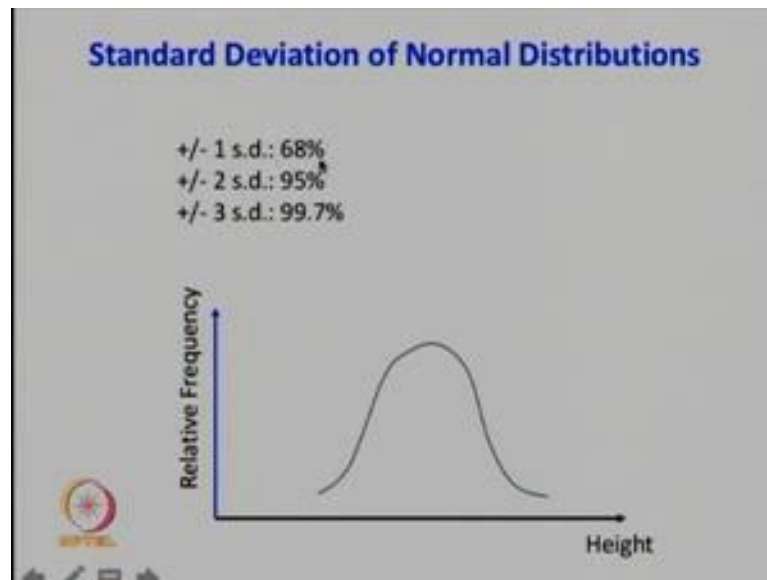
(Refer Slide Time: 02:43)



And so you have 75 plus minus 20 in this case, we will have contain 1 minus 1 by 9 that is 8 by 9 fraction of the population 8 by 9 is roughly 90 percent of the population, but as I had state last week last class that for a generic distribution. So, Chebyshev's theorem is actually a very conservative estimate. So, this is your \bar{x} \bar{x} minus s \bar{x} plus s right. So, they say that this much. So, this is roughly what we calculated is 75 percent and in the generic case \bar{x} plus $2s$ and \bar{x} minus $2s$ this is 90 percent of the population.

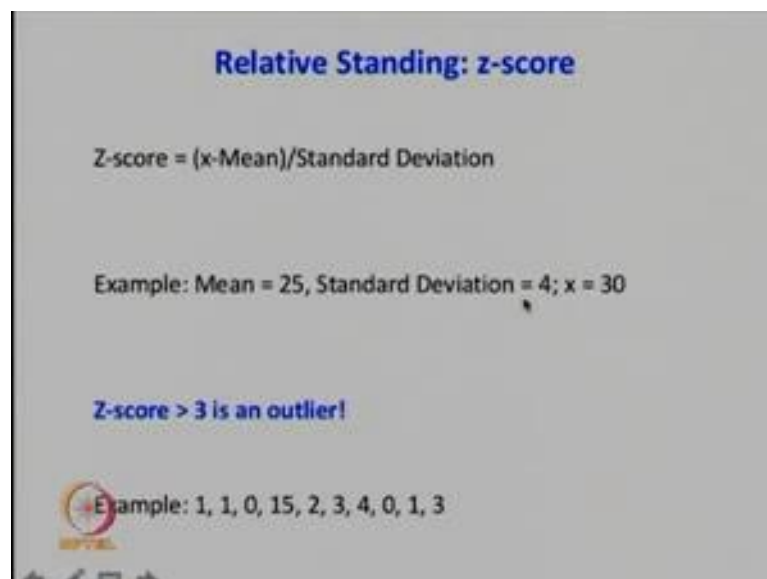
But Chebyshev's theorem is a very conservation approach. So, it does not make any assumptions of how the distribution of data is there.

(Refer Slide Time: 03:45)



In contrast for what you observed for a mound like distribution which is Gaussian distribution, a normal distribution you will see there is 68 percent of the data which we are expected to be there within plus minus 1 standard deviation. So, as opposed to 75 percent predicted by Chebyshev's theorem in normal distribution 68 percent stay within the plus minus 1 standard deviation, but plus minus 2 standard deviations 95 percent of the data is there and 3 standard deviations 99.7 percent of the data is there. So, this also brings us to the concept of zee score or it is relative standing.

(Refer Slide Time: 04:20)



So, what exactly is zee score it is basically defined by x minus mean by standard deviation and you can do this calculation if you know for a particular experiment if your mean is 25 standard deviation is 4 and x is 30 then zee score returns your value of 30 minus 25 by 4 which is 1.1, 1.25. Now you can use zee score to get an estimate of what whether a particular data point is an outlier or not and this can be you know clearly gleaned from this particular example we worked out in last class.

So, what you see if you see look at the data points all the points are clustered between one and 4 except for this one particular value which is 15. So, we can clearly see it seems to us that 15 is outlier or very close to being an outlier you can do this calculation we had worked out what exactly its value is, I do not remember but you can find out whether as per this statement if the zee score comes out to be greater than 3 or not.

(Refer Slide Time: 05:28)

Relative Standing: Percentiles

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

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

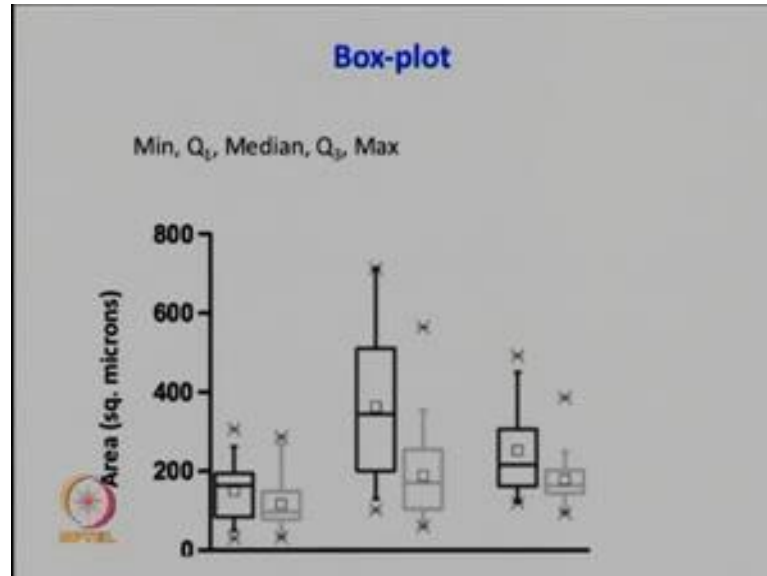
Q_3 : third quartile (at position $0.75 \cdot (n+1)$)

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

Another way of you know characterizing our relative standing is using the concept of percentile. So, p th percentile is the value which is greater than p percent of the measurements. So, 100 percentile is essentially. So, that person who is in the 100 percent or 99 percentile is pretty much better than it has performed better than 99 percent of the population in a class. So, you can use these particular positions to determine how you will calculate the first quartile or third quartile the second quartile is of course, at position point 5 star n plus one is nothing but the median. So, this represents at what 25

percent of the data point is first quartile 75 50 percent of the data point is third quartile and inter quartile range is defined as $Q_3 - Q_1$.

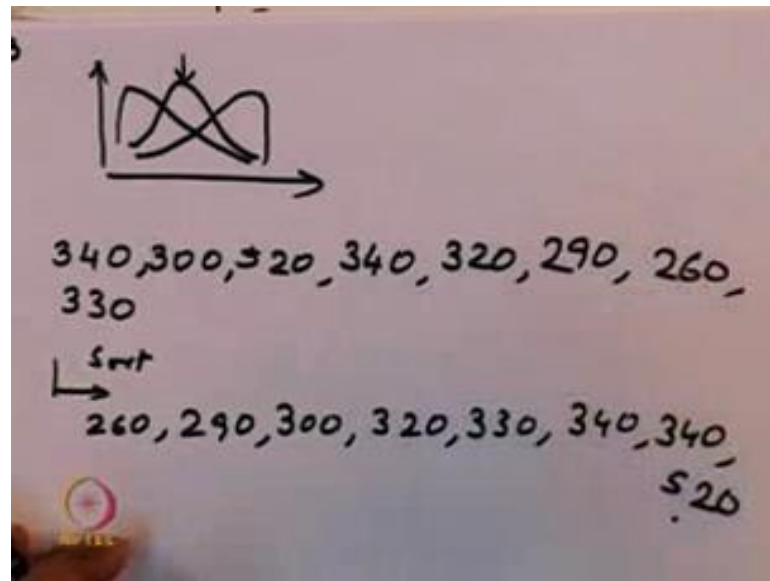
(Refer Slide Time: 06:22)



So using this you know these values one can plot what is called a box plot and in a box plot. So, the lowest value is your minimum this the box outlines. So, you have the Q_1 which is the first quartile Q_2 or the median Q_3 or the third quartile and this is your maximum. So, what you also see are points which may lie outside these definitions of box. So, if you take this point this coincides with the maximum value of the distribution, but this point or for that matter this point really is much outside the box limits. So, these points are examples of outliers and it is perhaps not you know completely surprising that in many experimental data you do have outliers.

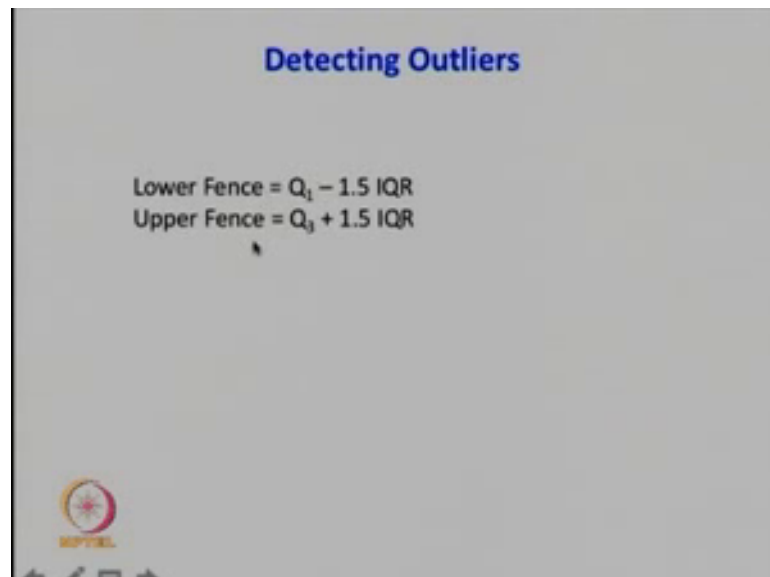
So, this square inside the box actually denotes the median the mean what you see here in this particular population you have variables if you look at the y axis you have variables which vary all the way from around 20 or 30 or 50, all the way to 600. So, when you take an average the effect of that 600 is going to have a much greater effect than a value of 50 which is why in this particular case the mean is slightly shifted above the position of the median. So, you take this particular example it is the other way round where the median is here and the mean is here.

(Refer Slide Time: 08:02)



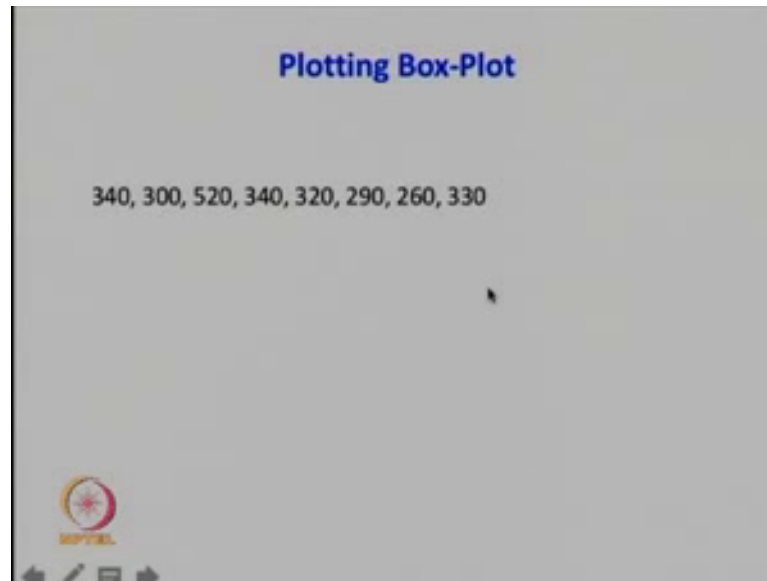
So, based on this, if you were to you know plot it in terms of histograms. So, as opposed to having a distribution like this where your position of mean, median, mode, all coincide you might either shift to the left or to the right. So, the way to detect outliers is using this particular formula.

(Refer Slide Time: 08:14)



So, you can you can construct fence where the lower fence is given by $Q_1 - 1.5$ times inter quartile range and the upper fence is $Q_3 + 1.5$ inter quartile range. So, let us just work out a sample case of how we will actually plot our box plot.

(Refer Slide Time: 08:26)



So let me write down the points you have the points 350, 300, 520, 340, 320, 290, 260 and 330. So, first step of course, is to sort in ascending order. So, my lowest value here is 260 then 290 you can have 300 I have a 320, 320, 330, 340 and one 520. So, we can already see clearly here that as opposed to all of these points which kind of are clustered together this data point seems to be out of the plot.

(Refer Slide Time: 09:36)

4

260, 290, 300, 320, 330, 340, 340, 520

$N = 8$

Median = $\frac{1}{2}(320 + 330) = 325$

Q_1 posn = $\frac{1}{4}(N+1) = \frac{9}{4} = 2.25$

Q_3 posn = $\frac{3}{4} \times 9 = 6.75$

$Q_1 = 290 + 0.25 \times 10$
 $= 292.5$

$Q_3 = 340 + 0.75 \times (340 - 340) = 340$

So, but let us find out do our necessary calculations once again. So, you have to 260, 290, 300, 320, 330, 340, 340; 520. So, my total number of measurements is 1, 2, 3, 4, 5,

6, 7, 8 n is equal to 8. That means my median position is going to be somewhere like this my median value will be half of 320 and 330 equal to 325 the position of Q 1 position will be one times n plus 1 equal to 9 by 4, it was 2.25. So, after 2, this is going to be the position of Q 1 this is your median and Q 3 position is going to be 3 forth into 9 it is 27 by 4 is 6, 424 6.75. So, 1, 2, 3, 4, 5, 6, 7; so Q 3 is going to somewhere here.

So, my Q 1 value will be 290 plus 0.25 times 10 which is going to be 290 plus 2.5 290, 2.5, Q 3 is going to be 6.75; 1, 2, 3, 4, 5, 6 is 340 plus 0.75 into 340. So, it will still get the value. So, Q 3 is in this particular position 175 times 340, no, but the do you know 3 40 minus 340 which is nothing but 340 only. So, we have calculated the values of Q 1 and Q 3.

(Refer Slide Time: 11:44)

5

$$Q_1 = 292.5$$

$$Q_3 = 340$$

$$\Rightarrow IQR = Q_3 - Q_1 = 47.5$$

$$\text{Lower Fence} = Q_1 - 1.5 IQR$$

$$= 292.5 - 1.5 \times 47.5$$

$$\approx 250$$

$$\Rightarrow \text{No lower outliers!!}$$

So, now we need to see our median. So, for this particular distribution I have Q 1 as 290 2.5 Q 3 is equal to 340. So, which would mean that IQR is equal to Q 3 minus Q 1 is 47.5.

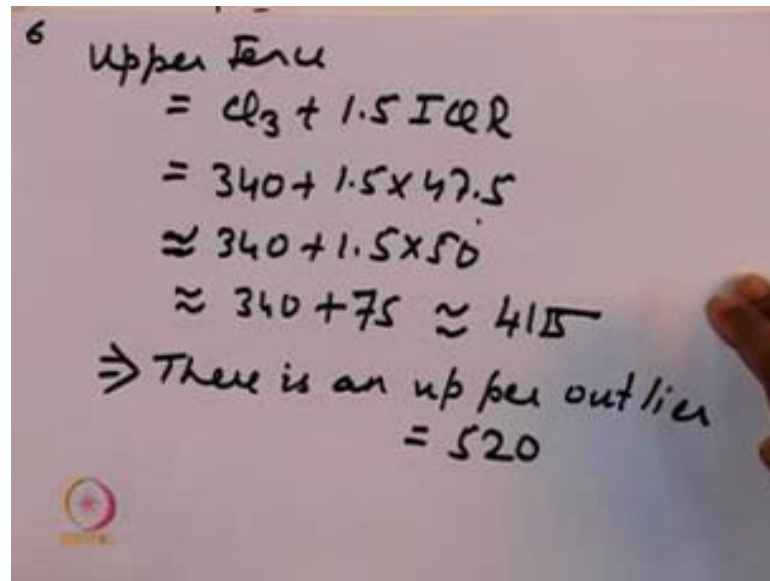
So, now, we know that the lower fence. So, lower fence Q 1 minus 1.5 times IQR. Q 1 is 292 minus 1.5 into 47.5. So, which will be around let us say 1 times 47.5 which will be around 250 I do not know the exact value please calculate the, but as you can see if you look at our points once more the lowest value is 260; that means, that there are no lower outliers. So, this implies that there are no lower outliers I can.

(Refer Slide Time: 12:59)

6

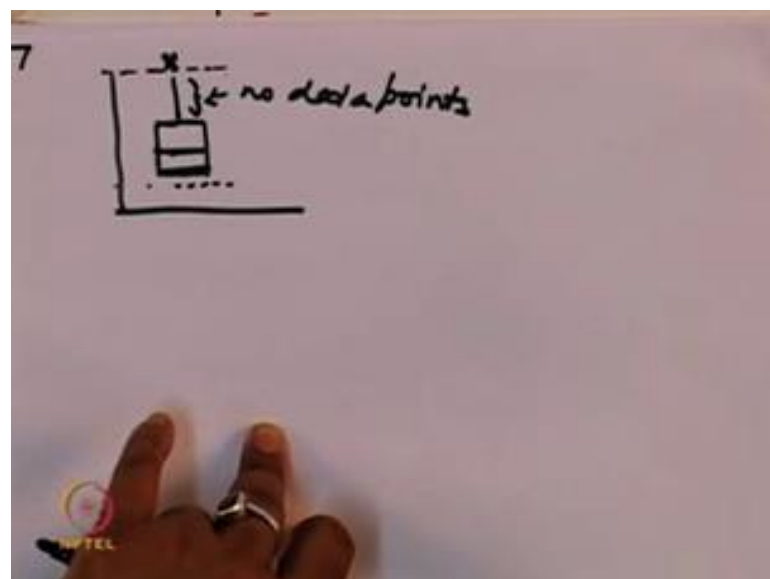
$$\begin{aligned}\text{Upper Fence} &= Q_3 + 1.5 IQR \\ &= 340 + 1.5 \times 47.5 \\ &\approx 340 + 1.5 \times 50 \\ &\approx 340 + 75 \approx 415\end{aligned}$$

\Rightarrow There is an upper outlier
= 520



Similarly, calculate the value of $Q_3 + 1.5$. So, upper fence $Q_3 + 1.5$ times IQR Q_3 is 340 plus 1.5 into 47.5 which will give me a value, if I assume this as 50. So, this is approximately is 340 plus 1.5 times 50. So, approximately is 340 plus 50 or 75 is roughly 415. So, this implies that the number. So, there is an outlier there is an upper outlier and which is nothing which is the value is equal to 520.

(Refer Slide Time: 13:53)

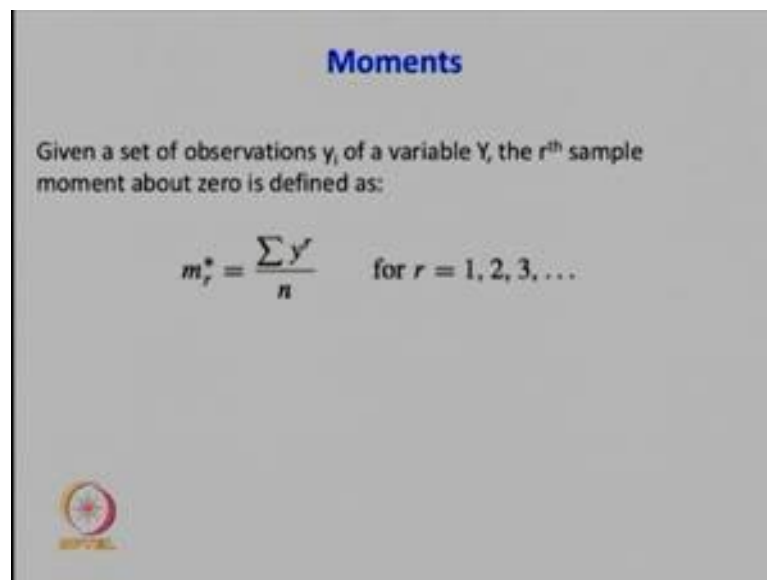


So, if I were to construct the plot if I were to construct the plot my plot would look something like this. So, as you can see that there is no. So, the minimum is 260 and

because, this is your lower fence your upper fence is somewhere here and this value lies much above. So, this means that. So, you have an error bar which sticks out, but this is much outside.

So, there are actually no points here there are no data points in this region no data points in this region because after 340 where are the points because after 340 there is you directly have 520. So, this just means that there is no data point here, but your error, but this shows you up to the maximum time there are no points here. So, with that I will show you how to generate a box plot.


(Refer Slide Time: 15:04)



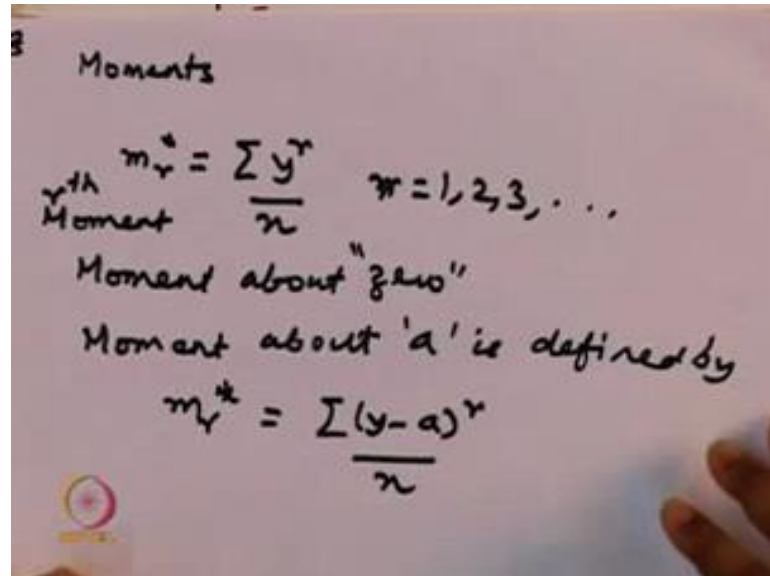
Moments

Given a set of observations y , of a variable Y , the r^{th} sample moment about zero is defined as:

$$m_r^* = \frac{\sum y^r}{n} \quad \text{for } r = 1, 2, 3, \dots$$

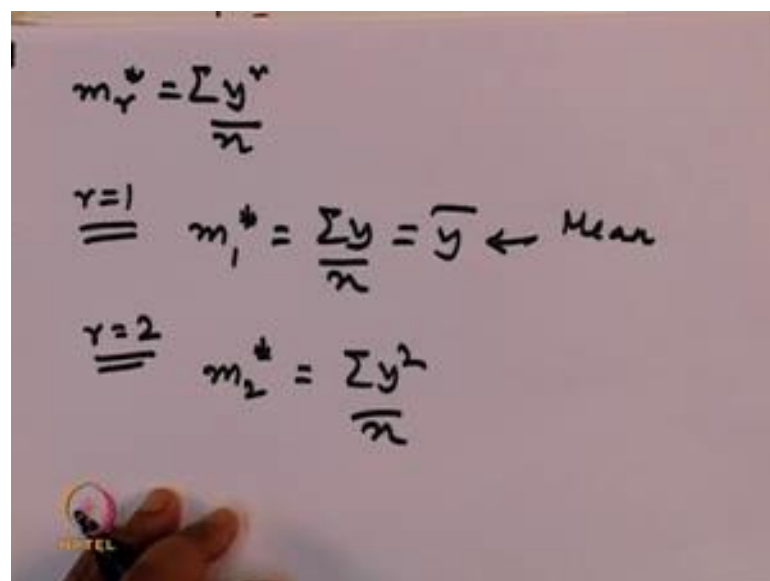


(Refer Slide Time: 15:10)



Now we come to another interesting concept of moments. So, as Pearson was the first statistician to make use of moments to describe data now what is how is that moment defined. So, you have moment about. So, moment about any variable about 0 is defined as summation y to the power r by n where r can have 1, 2, 3, any value. So, clearly, this is moment. So, this is moment about 0. So, in general moment, this is the r -th moment, this is the r -th moment in general moment about a is defined by m_r^* between the generic is y minus a whole to the power r by n .

(Refer Slide Time: 16:18)



So, now let us see what the moments conveyed. So, you have moment about 0 m_r^* is defined by summation y^r by n it is obvious that if I put r equal to one then m_1^* is

equal to summation y by n which is nothing is equal to \bar{y} . So, first moment about 0 is your mean what about r is equal to 2. So, r equal to 2 then why I have a m_2 star about 0 is summation y square by n . So, as you can clearly see that this gives me I know that the way standard deviation or variance is defined you have a term of y minus \bar{y} whole square by n .

So, in other words, if you were to go through the moment. So, the r -th sample moment about the mean would then have this particular value which is m_r is equal to summation y minus \bar{y} whole to the power r by n . So, m_2 about 0 is y square. So, this would mean that if I mean m_r about 0 I put \bar{y} is equal to 0 I have y to the power r .

(Refer Slide Time: 17:30)

10

m_r about Mean

$$m_r = \frac{\sum (y - \bar{y})^r}{n}$$

$$m_1 = \frac{\sum (y - \bar{y})}{n} = \frac{\sum y - \sum \bar{y}}{n}$$

$$= \frac{n\bar{y} - n\bar{y}}{n} = 0$$

So, m_r about the mean so, in that case m_r is defined as summation y minus \bar{y} whole to the power r by n . So, m_1 in this case will be summation y minus \bar{y} by n and this is nothing but summation y minus summation \bar{y} by n . So, summation y is equal to n times \bar{y} and summation \bar{y} n times is nothing but $n \bar{y}$. So, this would give me a value of 0. So, first moment about the mean is 0.

(Refer Slide Time: 18:18)

$$m_2 = \frac{\sum (y - \bar{y})^2}{n}$$

$$\rightarrow \text{Population Variance}$$

$$m_3 = \frac{\sum (y - \bar{y})^3}{n}$$

$$y_1 - \bar{y} = -\Delta y$$

$$y_2 - \bar{y} = \Delta y$$

$$f \Delta y^3 + f \Delta y^3 \rightarrow 0$$

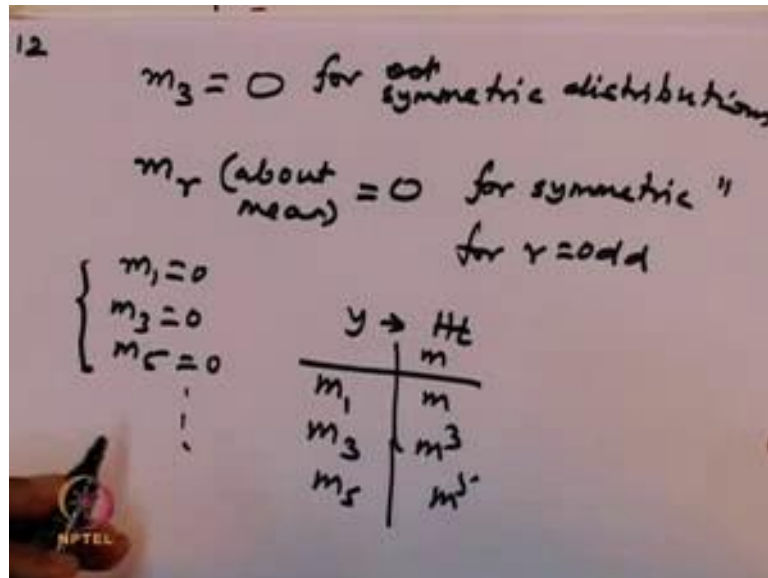
This obviously brings us to the second case that what is the second moment about 0. So, this would be defined by y minus \bar{y} whole square by n . So, as you can see that if this was for a population m_2 is nothing but variance. So, m_2 is variance. So, I am going to make that an approximation because if it is for a sample then you have to be $n - 1$, but this is very simply is equal to the variance. So, m_2 is you can I can say population variance. So, similarly I can calculate this value which is m_3 is equal to summation y minus \bar{y} whole cube to the power n .

Now, let us consider a very symmetric distribution if my distribution was symmetric. So, there is symmetry right in this distribution if I look at how m_3 is defined then I know that for if there is a symmetric would mean that for every value which is to the left of this there is similar value at similar frequency to the right of this right. So, let us say this is \bar{y} this is y_1 and this is y_2 . So, the frequency of y_1 and the frequency of y_2 is symmetric is equal and that is how the distribution is called it is a symmetric distribution in that case. So, if I have for every y_1 .

So, I have 2 things symmetric and let us say this distance is the same. So, y_1 minus \bar{y} is equal to let us say minus Δy and y_2 plus \bar{y} is going to be plus Δy , y_2 minus \bar{y} is going to be plus Δy . So, if I do this summation it just means that for every y_1 which is to the left of \bar{y} . So, at whatever contribution this gives which will be negative in nature if the another point which is equal equidistance in the positive

axis and has same frequency will give me a positive response and anything cubed if you have a negative number its cube is negative if you have a positive number its cube is positive. So, if you add these 2 terms. So, it will be like let us say f times minus Δy cubed plus f times Δy cubed and these 2 terms equate to 0.

(Refer Slide Time: 20:45)



So, this would mean that m_3 . So, so this would mean that m_3 is will return you a value of 0 for odd for symmetric distributions and this is same for any m_r . So, m_r about the mean is going to be 0 for symmetric distributions for r is equal to odd. So, in other words m_1, m_3, m_5, \dots is 0, m_3 is 0, m_5 is 0 and so on and so forth. So, clearly for all symmetric distributions you have the odd moments about the mean return you a value of 0.

Now let us say the variable y that we are measuring is actually some quantity it is not just a number it is a quantity let us say temperature or height. So, m_3 will have. So, each of them have different units. So, if I were to say y represents h then unit of m_1 in terms of meter let us say it is in meters m_3 unit is meter cubed m_5 unit is meter 5th. So, in other words these units are not the same can there be a way of compressing this information and coming up with a non dimensional parameter and that is what that is the what this measure of skewness gives us.

(Refer Slide Time: 22:26)

The image shows handwritten mathematical formulas and two graphs. At the top, the formula for skewness is given as $a_3 = \frac{\sum (y - \bar{y})^3}{[\sum (y - \bar{y})^2]^{3/2}} = \frac{m_3}{m_2^{3/2}}$. Below this, it is noted that $a_3 \rightarrow \text{skewness}$. The first graph shows two overlapping bell-shaped curves that are symmetric about a central vertical line, with the label $a_3 = 0$ next to it. The second graph shows a single bell-shaped curve that is skewed to the right, with the peak labeled 'Mode' and the center of mass labeled 'Mean'. A vertical line is drawn through the center of the curve, labeled 'Median', and the mean is indicated to be to the right of the median. Below the graphs, it is noted that $a_3 = -ve$ with a small graph of a left-skewed distribution.

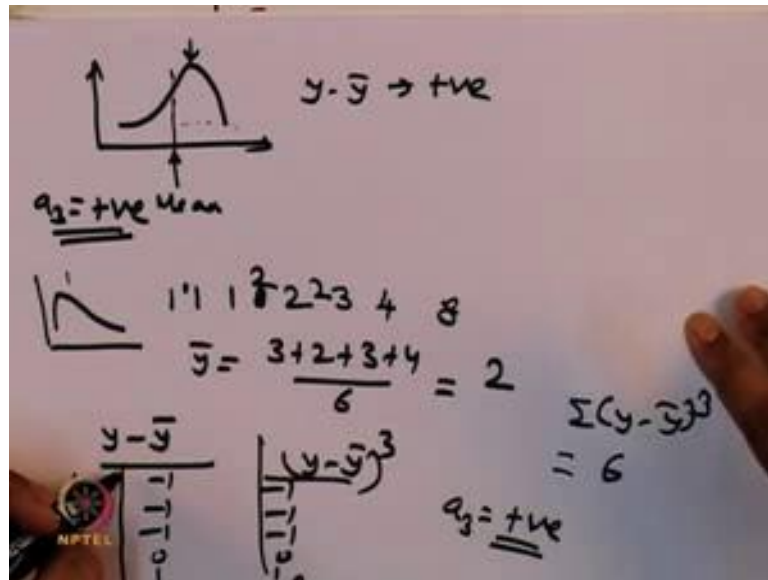
So, skewness is defined in a slightly different way as a_3 is equal to summation of y minus \bar{y} whole cubed by summation of y minus \bar{y} whole square whole to the power of $3/2$. I can again rewrite as m_3 by m_2 whole to the power $3/2$. So, as what you can clearly see that m_3 will have units of meter cubed m_2 will have units of meter square whole to the power $3/2$ will give you immunity of meter cubed and this is after all a number a dimensionless number. So, this parameter a_3 is called a skewness a_3 is called skewness and for any distribution. So, as it is you know obvious from the words a_3 itself. So, for any symmetric distribution any symmetric distribution my skewness a_3 has to be 0. So, it is neither skewed in this direction; nor skewed in this direction. So, this is what skewness is about.

Now, what kind of a value can a_3 be negative if we look at our definition of a_3 ? So, if let us say we take particular distribution this is skewed to the left. So, this is skewed to the right. So, this is going to be my mode my mean will my mean this will be where my mean will lie and this will be where my median will lie. So, what you can clearly see is when I do this computation for a_3 tells me that there are lot many number of points which are less than my mean.

So, this is my \bar{y} value and all for all these values I am going to get this component will return me a negative value and only for few of the others this quantity is going to be returned me a positive value. So, when I actually do this calculation I am going to get a

value of a 3 which is going to be negative. So, a 3 is going to be negative for this kind of distributions. So, we will do one sample calculation to see whether what we think is will remain it like that.

(Refer Slide Time: 24:55)



So, in the other case if it is skewed in the other direction if you have a distribution like this. So, this is your mode this is your; here is where your mean will lie. So, I can clearly see that for all these points which are to the right of the mean $y - \bar{y}$ is going to be positive and as by this token I will get a value of a 3 which is positive. So, let us take a sample example let us take a sample example where we calculate the skewness of a distribution. So, let me write down some numbers which are which kind of portray this picture. So, let us say my variables are 1; 1; 1; sorry, 1.

So, this is doing this particular kind of a case. So, 1, 1, 1, 1, 2, 3, 4, 3, 4, 5, 6, 7, let us do this distribution we have 3 1s, 1 2, 1 3, 1 4. So, your \bar{y} is equal to 3 plus 2 plus 3 plus 4 by 6 is equal to 6 10 by 2. This 2 \bar{y} is 2 now I can calculate my $y - \bar{y}$. So, I have 1, 1, 1, 2, 3, 4. So, for value of 1, it is minus 1 minus 1 minus 1, 0, 1, and 2. So, in this case $y - \bar{y}$ whole cube will give me minus 1 minus 1 minus 1, 0, 1, 2 cube is 8.

So, in this particular case even though the distribution is wise to the left I can see that summation $y - \bar{y}$ whole cube will give me a value of 3 plus 3 4 6. So, in this case it is though it is skewed to the left it is now it is still a 3 is giving me a value which

is kind of positive. So, though, but you can see that if these numbers were much to the left. So, if you had you know a few more of 2 3 4 and you had one number as 8 and you did the; you know you had 2 more of 2 3 more of 2 and do this for this particular distribution you might see that this will slowly become negative.

So with that I conclude today's class. So, what we have done is come up with this metric of skewness. So, starting from standard deviation and going to how we want to do relative standing by using zee score and then from there we went on to see how we can come up with relative matrix of finding out moments and coming up with matrix to characterize the way a distribution is. So, skewness gives us the value for any symmetric distribution skewness will return you a value of 0, but typically if it is biased if most of your data lies to the left of your mean then sometimes this skewness value can be negative. Versus, if your data is to the right it can be positive. With that I conclude today's lecture.

Thank you for your attention.