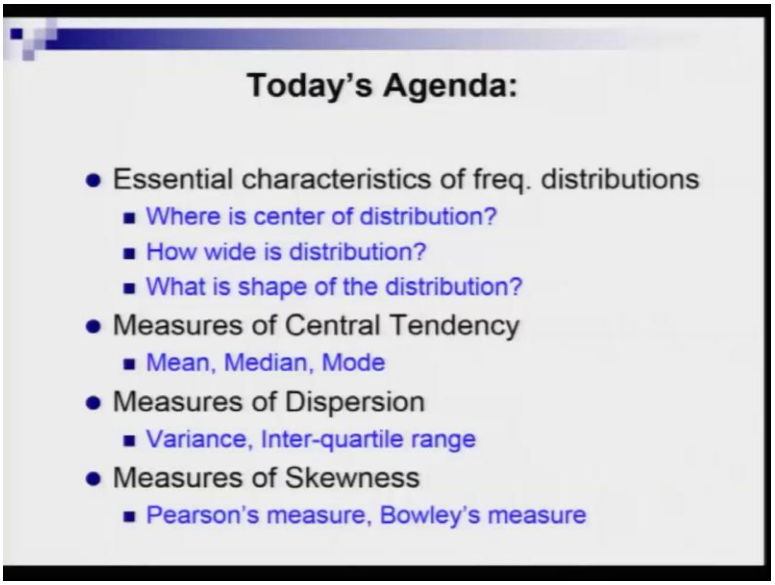
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 04
Summarizing Data through Descriptive Statistics

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So today we are going to continue our discussion on descriptive statistics measures. So let us have a look at today's agenda items.

(Refer Slide Time: 00:31)



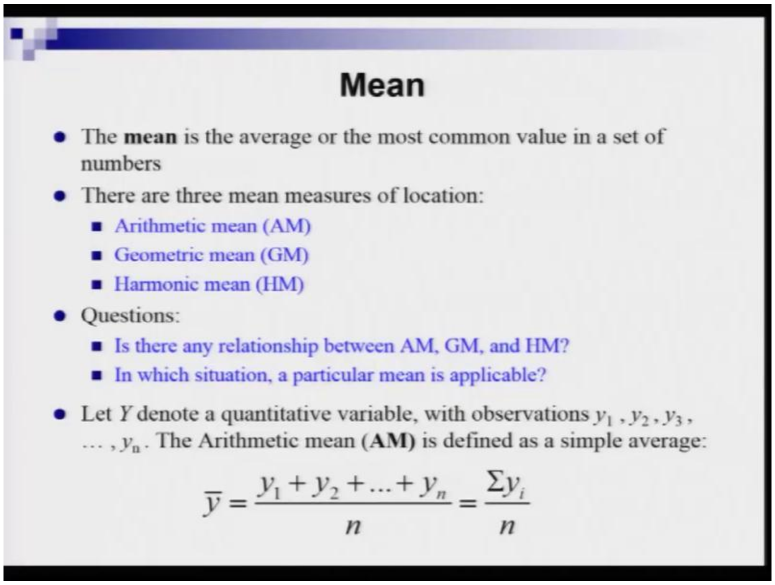
Today's Agenda:

- Essential characteristics of freq. distributions
  - Where is center of distribution?
  - How wide is distribution?
  - What is shape of the distribution?
- Measures of Central Tendency
  - Mean, Median, Mode
- Measures of Dispersion
  - Variance, Inter-quartile range
- Measures of Skewness
  - Pearson's measure, Bowley's measure

So in the last lecture we have started with essential characteristics of a frequency distribution. Then we will look at the measures of these three features of frequency distributions. So we will study mean, median and mode. These are the measures of central tendency. Then we will study variance and inter-quartile range. These are the measures of dispersion. And we will finish today's discussion with measures of skewness namely Pearson's measure and Bowley's measure.

Now the descriptive statistics measures are probably not new concepts to you. Probably you have seen these concepts while you are studying mathematics and statistics at class 10th or 11th or 12th. So I am not going to spend a lot of time on these concepts because I will assume that you have some basic idea about mean, median, mode and variance.

(Refer Slide Time: 01:32)



Mean

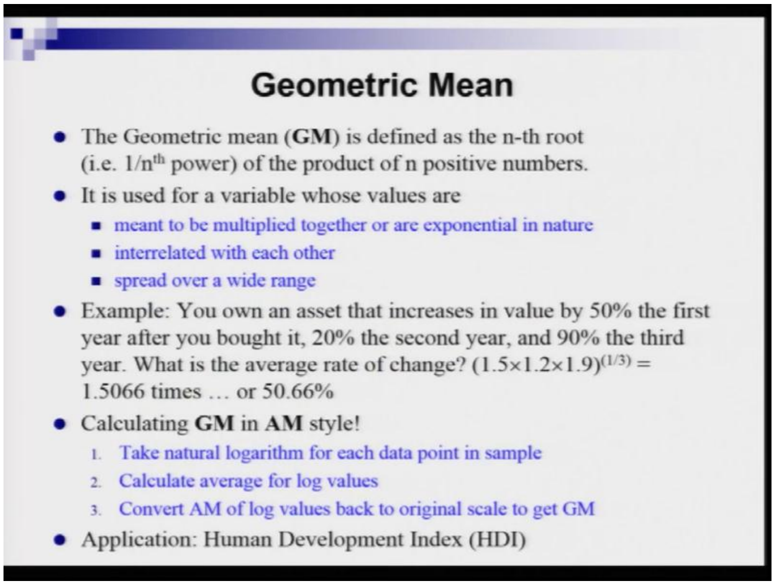
- The **mean** is the average or the most common value in a set of numbers
- There are three mean measures of location:
  - Arithmetic mean (AM)
  - Geometric mean (GM)
  - Harmonic mean (HM)
- Questions:
  - Is there any relationship between AM, GM, and HM?
  - In which situation, a particular mean is applicable?
- Let Y denote a quantitative variable, with observations y_1 , y_2 , y_3 , … , y_n . The Arithmetic mean (**AM)** is defined as a simple average:

$$\overline{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\Sigma y_i}{n}$$

So we will start with mean. What is a mean? Mean is the average or the most common value in a set of numbers. And there are three types of Pythagorean means available; arithmetic mean, geometric mean and harmonic mean. So naturally a question emerges in mind that, is there any relationship between AM, GM and HM? And second pertinent question is that, in which situation a particular mean is applicable? So note that it is not always the case that you will apply the simple average or the arithmetic mean.

And there are special types of means like geometric mean and harmonic mean which are applied in special cases. So we all know the definition of arithmetic mean but let us, go through it again in symbolic manner. So if y is my quantitative variable with observations y1, y2 dot dot dot, yn. So there are n number of observations in my dataset. Then the arithmetic mean AM is basically a simple average which is given by y bar and that is basically summation of yi where i ranges from 1 to n, divided by n, the sample size.

(Refer Slide Time: 2:56)



Geometric Mean

- The Geometric mean (**GM**) is defined as the n-th root (i.e. $1/n^{th}$ power) of the product of n positive numbers.
- It is used for a variable whose values are
  - meant to be multiplied together or are exponential in nature
  - interrelated with each other
  - spread over a wide range
- Example: You own an asset that increases in value by 50% the first year after you bought it, 20% the second year, and 90% the third year. What is the average rate of change? $(1.5\times1.2\times1.9)^{(1/3)} =$ 1.5066 times … or 50.66%
- Calculating **GM** in **AM** style!
  1. Take natural logarithm for each data point in sample
  2. Calculate average for log values
  3. Convert AM of log values back to original scale to get GM
- Application: Human Development Index (HDI)

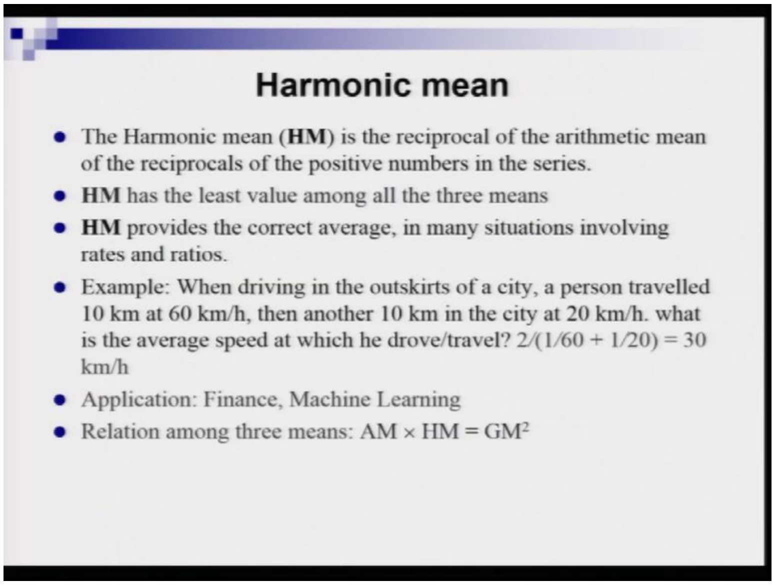
Now we will move on to geometric mean. So geometric mean GM is defined as nth root. It implies one over nth power of the product of n positive numbers. So be very careful here and please pay attention to this important point that here you cannot compute geometric mean of any set of numbers. You have to deal with positive numbers only. So it is a limitation of the concept. But let us proceed with this limitation.

So despite its limitation this is used for variable whose values are either meant to be multiplied together or are exponential in nature or interrelated with each other especially when you are dealing with time series data or financial data or spread over a wide range. So let us study an example and see how it is measured or computed. So suppose you own an asset that increases in value by 50 percent the first year after you bought it. Then at the end of the second year the value further increases by 20 percent. And finally at the end of 3rd year the value further increases by 90 percent.

One may want to know what is the average rate of change in this case. So here note that we are talking about the value being increased by 1.5 times at the end of year 1, 1.2 times by the end of year 2 and 1.9 times by the end of year 3. So if you compute the geometric mean then, you have to multiply these three numbers and then you have to compute the cube root of that. So that will roughly 1.5066. And if you convert that into percentage time, terms then it will be 50.66 percent.

So let us now calculate the geometric mean in arithmetic mean style. So in this case you take natural logarithm for each data point in the sample and then you calculate average for log values. And then you convert that arithmetic mean of log values back to the original scale to get the geometric mean of the original data points. So how to return it back to the original scale? You have to take exponentiation because you have taken natural log here. It is applied in computing Human Development Index which is provided by the United Nations.

(Refer Slide Time: 5:43)



Harmonic mean

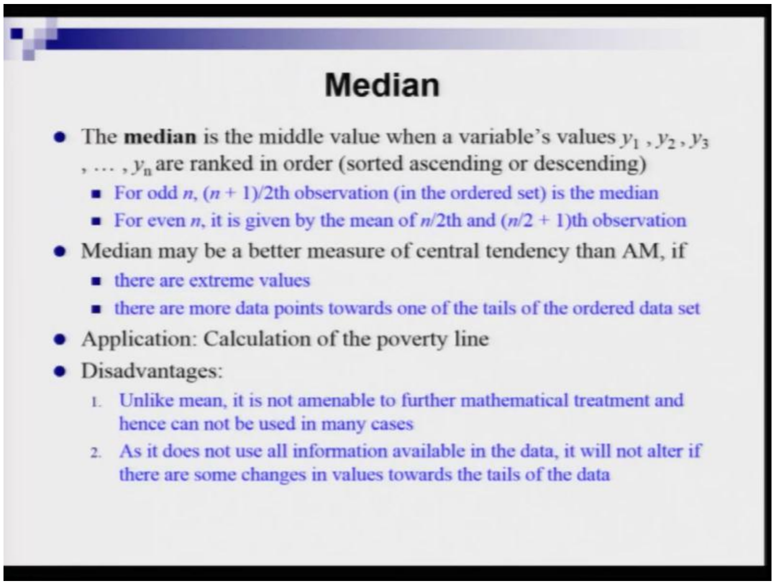
- The Harmonic mean (**HM**) is the reciprocal of the arithmetic mean of the reciprocals of the positive numbers in the series.
- **HM** has the least value among all the three means
- **HM** provides the correct average, in many situations involving rates and ratios.
- Example: When driving in the outskirts of a city, a person travelled 10 km at 60 km/h, then another 10 km in the city at 20 km/h. what is the average speed at which he drove/travel? $2/(1/60 + 1/20) = 30$ km/h
- Application: Finance, Machine Learning
- Relation among three means: $AM \times HM = GM^2$

Now let quickly have a look at the harmonic mean. What is it? Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the positive numbers in the series. Again note that if you are interested in computing harmonic mean it has to be positive numbers only. Now there is an interesting relation here between AM, GM and HM. And from that relationship we know that harmonic mean has the least value among all the three means. We are not interested in proof but, remember this result.

Harmonic mean provides the correct average in many situations involving rates and ratios. So you now know where you have to apply harmonic mean compared to the arithmetic mean. So let us look at the example and it will be clear how to compute harmonic mean in one of these cases. So let us assume that somebody is driving in the outskirts of the city. And the person travelled 10 kilometer at 60 kilometer per hour. Then after entering the city the person drove 10 kilometers more in the city but at a reduced speed of 20 kilometers per hour. What is the average speed at which he drove or travelled?

So you have to compute harmonic mean here. So here the formula could look like 2 divided by 1 over 60 plus 1 over 20 and we get 30 kilometer per hour as the harmonic mean. Now as I mentioned that harmonic means are more suitable for numbers which are expressed in rates or ratios, hence it finds its applications in financial economics. The square of geometric mean is equal to the product of arithmetic mean and harmonic mean.

(Refer Slide Time: 07:37)



Median

- The **median** is the middle value when a variable's values y_1 , y_2 , y_3 , … , y_n are ranked in order (sorted ascending or descending)
  - For odd *n*, (*n* + 1)/2th observation (in the ordered set) is the median
  - For even *n*, it is given by the mean of *n*/2th and (*n*/2 + 1)th observation
- Median may be a better measure of central tendency than AM, if
  - there are extreme values
  - there are more data points towards one of the tails of the ordered data set
- Application: Calculation of the poverty line
- Disadvantages:
  1. Unlike mean, it is not amenable to further mathematical treatment and hence can not be used in many cases
  2. As it does not use all information available in the data, it will not alter if there are some changes in values towards the tails of the data

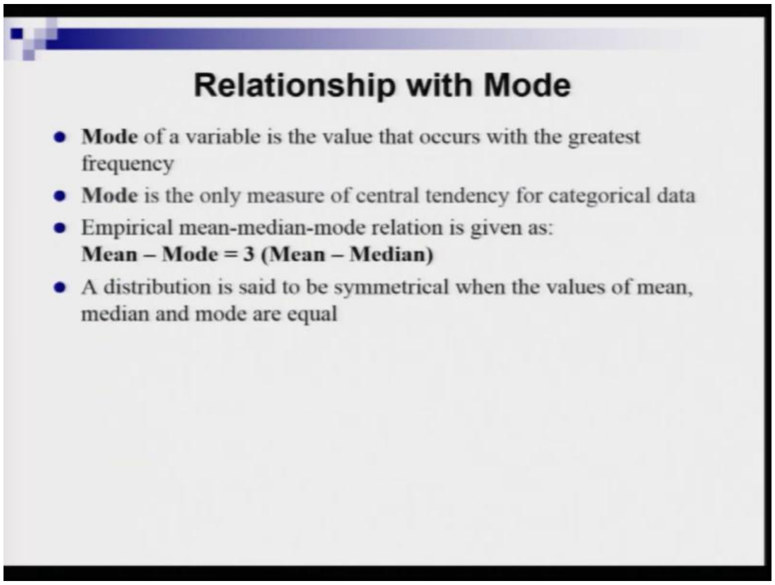
So let me move on to the second measure of central tendency and that is median. So a median is the middle value when variables values y1, y2, y3 etc., are ranked in order sorted ascending or descending. But most of the times we prefer if the numbers are ranked in increasing order. Now there could be two cases; n, the sample size could be an odd number or could be an even number. So if it is odd number then the median is basically the n plus 1 divided by 2th observation in the ordered data set and if n is even then the median is given by the mean of n divided by 2th and n divided by 2 plus 1th observation.

So here in the even case the situation is a bit complicated. You have to find two middle observations in the data set and then you have to take simple average of these two numbers. Now when median is preferred over arithmetic mean as a measure of central tendency? There are two possible cases that I can say. So if there are extreme values and you are aware of that fact and if there are more data points toward one of the tails of the ordered data sheet. So the data is concentrated either towards the lower tail or the upper tail of the data set.

Now where does median find its application? So in development economics, median is used for calculation of poverty line. So is there any disadvantage of median? Yes, there are two major disadvantages of median. And these are the following. So arithmetic mean takes into account all observations while providing you a measure of central tendency or center of the distribution.

But median, as it does not use all information available in the data it will not alter if there are some changes in values towards the tail of the data. And also arithmetic mean is pretty much amenable to further mathematical treatment but median is not. Hence median is not very much useful in many higher level statistical analysis like hypothesis testing.

(Refer Slide Time: 10:02)



Relationship with Mode

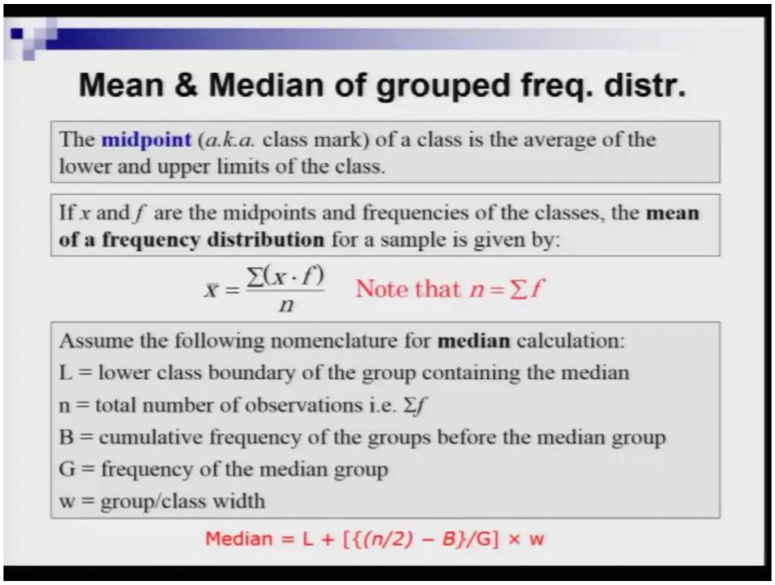
- **Mode** of a variable is the value that occurs with the greatest frequency
- **Mode** is the only measure of central tendency for categorical data
- Empirical mean-median-mode relation is given as:
  Mean – Mode = 3 (Mean – Median)
- A distribution is said to be symmetrical when the values of mean, median and mode are equal

Now quickly have a look at mode. This is a very simple concept compared to mean or median. Mode of a variable is that value that occurs with the highest frequency. Mode is the only measure of central tendency for categorical data. So you see when mode is much more preferred then median or mean. There is an empirical relationship between mean, median and mode. And that is given in the following equation which says that the difference between mean and mode is equal to 3 times the difference between mean and median.

Later we will see that this empirical relationship becomes very handy when we are interested in measuring skewness of frequency distribution, or how skewed the data is. Now at this juncture it is important note that a distribution is said to be symmetrical when the values of mean, median and mode are all equal. We will come back to this point when we will be discussing skewness.

So far we have discussed these measures of central tendency in the case of ungrouped frequency distribution. Now let us have a look how these measures behave when we are dealing with grouped frequency distribution. It implies that we are looking at the intervals, class intervals and frequencies against these class intervals.

(Refer Slide Time: 11:29)



Mean & Median of grouped freq. distr.

The **midpoint** (*a.k.a.* class mark) of a class is the average of the lower and upper limits of the class.

If x and f are the midpoints and frequencies of the classes, the **mean of a frequency distribution** for a sample is given by:

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} \qquad \text{Note that } n = \Sigma f$$

Assume the following nomenclature for **median** calculation:

L = lower class boundary of the group containing the median

n = total number of observations i.e. Σf

B = cumulative frequency of the groups before the median group

G = frequency of the median group

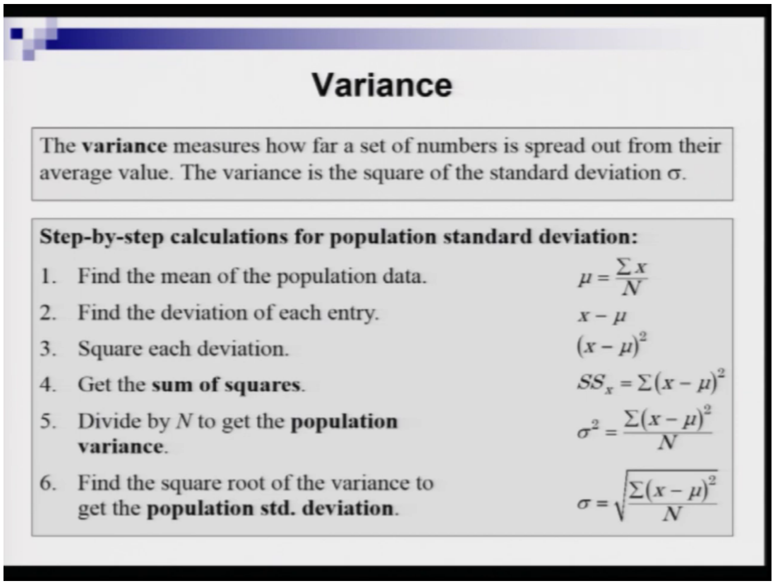
w = group/class width

Median = L + [{(*n*/2) – *B*}/G] × w

So when we are dealing with class intervals we have to compute the mid-point also known as the class mark of a class. And that is basically the average of the lower and upper limits of the class. That is the first step. So once you have got these midpoints, let us denote them by small x and f is the corresponding frequency then mean of a frequency distribution for the sample is to be given by x bar equal to summation x times f whole divided by n. And n, note here it is summation of f. So basically it is the sample size, sum of all frequencies in all possible classes in the table.

Now the calculation of mean is quite simple in the grouped frequency distribution but median is not. For median we have to use complicated formula and that I have shown you here in red as the last equation in the slide. And I have given the nomenclature also for the symbols that I have used the formula. I hope that this nomenclature and then the formula is quite self-explanatory. So with this formula shown let me move to the variance.

(Refer Slide Time: 12:48)



Variance

The **variance** measures how far a set of numbers is spread out from their average value. The variance is the square of the standard deviation σ.

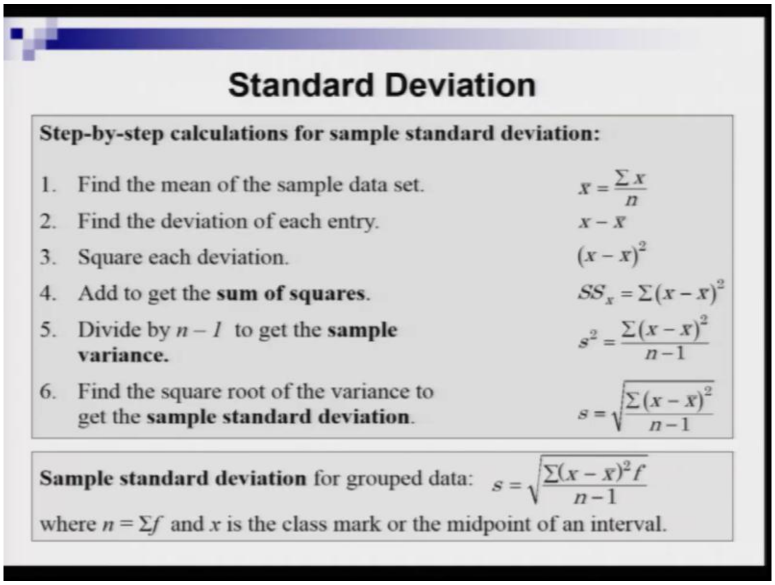
Step-by-step calculations for population standard deviation:

1. Find the mean of the population data. $\mu = \frac{\sum x}{N}$
2. Find the deviation of each entry. $x - \mu$
3. Square each deviation. $(x - \mu)^2$
4. Get the **sum of squares**. $SS_x = \Sigma(x - \mu)^2$
5. Divide by N to get the **population variance**. $\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$
6. Find the square root of the variance to get the **population std. deviation**. $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$

So what is variance? The variance feature of frequency distribution measures how far a set of numbers is spread out from their average value. The variance is the square of the standard deviation measure sigma. Now let us look at step-by-step calculations for population standard deviation. And then we will look at the step-by-step calculation for the sample standard deviation. So you have to find the mean of the population data first and we have discuss the formula, that is basically the simple arithmetic mean.

Then you have to find the deviation for each data point or entry in the sample, that is to be computed by subtracting the population mean that you have just computed in the first step from each data point, of course. And then you square each deviation. Then you get the sum of squares and then finally you divide that sum of squares by capital N which is the population size to get the population variance sigma square. And if you take a positive square root of that then you get population standard deviation.

(Refer Slide Time: 14:04)



Standard Deviation

Step-by-step calculations for sample standard deviation:

1. Find the mean of the sample data set. $\bar{x} = \frac{\Sigma x}{n}$
2. Find the deviation of each entry. $x - \bar{x}$
3. Square each deviation. $(x - \bar{x})^2$
4. Add to get the **sum of squares**. $SS_x = \Sigma(x - \bar{x})^2$
5. Divide by $n - 1$ to get the **sample variance.** $s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$
6. Find the square root of the variance to get the **sample standard deviation**. $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$

Sample standard deviation for grouped data: $s = \sqrt{\frac{\Sigma(x - \bar{x})^2 f}{n-1}}$

where $n = \Sigma f$ and x is the class mark or the midpoint of an interval.

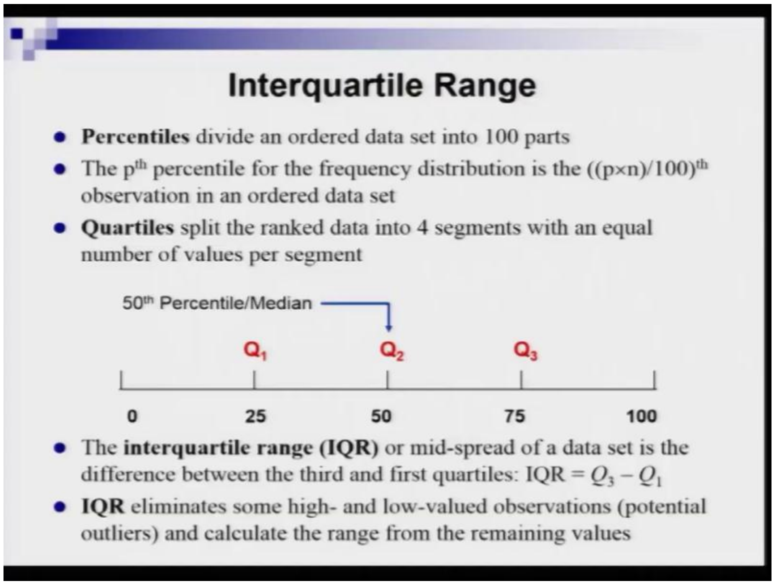
Now we are going to look at the step-by-step calculation for the sample standard deviation. Note here as we are dealing with sample we have to find the mean of the sample. So in the previous formula we just replaced capital N by small n which small n is the sample size here. So next three steps, step number 2, step number 3 and step number 4 are the same from the previous discussion. The difference emerges in step 5 where we need to divide by small n minus 1 to get the sample variance.

Now one may ask why small n minus 1, why not small n? Because there is a concept called degrees of freedom associated with it. So I am not going to discuss this right now. This thing I will discuss later in the course. But remember there is this no distinction between the sample variance and the population variance when you are dividing with the size of the data at hand. For the sample size do not forget to deduct 1 from the sample size.

Then, of course, when you take the square root of the sample variance you get the sample standard deviation. And that is equivalent of step 6 in the population standard deviation case also. So when we are dealing with grouped data how does my formula change? So in the second or the bottom box of the slide I am showing you the revised formula. It is simple.

So you have to first calculate the weighted mean or the grouped mean from the data set. So now this x bar that, I am showing here in the second box, that is basically a weighted mean. So weights are being the frequencies here. And then once this deviation for each entry is computed then you square it and then you multiply that squared deviation with the frequency. And then you sum them. And then, you divide by n minus 1 and then take a square root and you are done. So that is the way you get the sample standard deviation for grouped data.

(Refer Slide Time: 16:35)



Interquartile Range

- **Percentiles** divide an ordered data set into 100 parts
- The p^{th} percentile for the frequency distribution is the $((p \times n)/100)^{th}$ observation in an ordered data set
- **Quartiles** split the ranked data into 4 segments with an equal number of values per segment

50th Percentile/Median
Q_1
Q_2
Q_3
0
25
50
75
100

- The **interquartile range (IQR)** or mid-spread of a data set is the difference between the third and first quartiles: IQR = $Q_3 - Q_1$
- **IQR** eliminates some high- and low-valued observations (potential outliers) and calculate the range from the remaining values

So next we are going to look at another measure of variance and that is interquartile range. But to understand the measure we have to introduce two more concepts which are percentiles and quartiles. So let us have a look at them one by one. So percentiles divide an ordered set in 100 parts. Remember that I have been discussing this. So when a data set is given and you are interested in median computation you have to order the data in increasing, arrange the data in the increasing order.

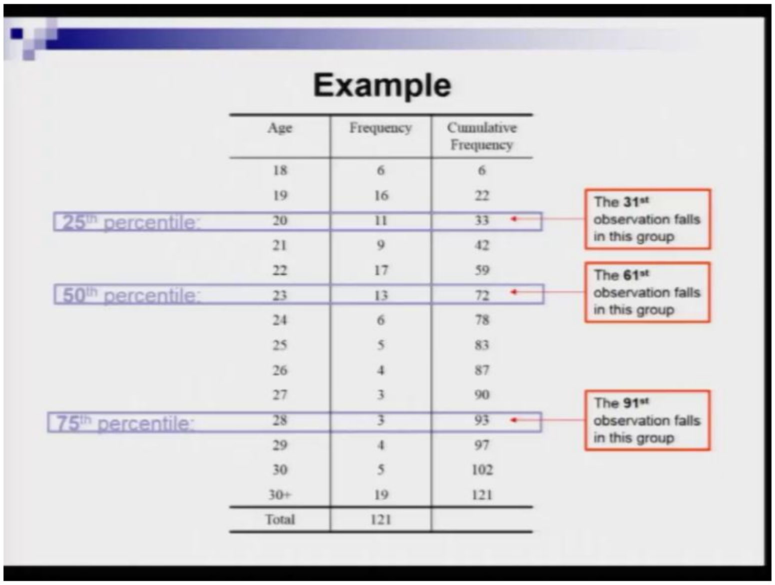
So you do that thing and then you actually go for the percentile and quartiles computation. So the pth percentile for the frequency distribution is p times n divided by 100th observation in an ordered data set provided that it is arranged in increasing order. So next comes quartiles. So quartiles split the ranked data into 4 segments with an equal number of values per segment. Now note from the previous diagram that there is interesting relationship between percentile, quartile and median. So what is median?

Median actually is the number that is splits your data equally half and half. So 50 percent of the observations are below that median value and 50 percent observations in the sample are above the median value. So median is basically a 50th percentile point. And quartile splits the data in 4 segments. So when we are talking about second quartile, so the values which are less than second quartile are actually 50 percent of the sample and the values which are above the second quartile that constitutes the rest 50 percent of the sample.

So the second quartile value is actually the median value. Now the interquartile range concept is based on the quartiles. So interquartile range IQR is also called mid-spread of a data set and this is a difference between third and the first quartile, so Q3 minus Q1. Now IQR eliminates some high and low valued observations and this could be potential outliers in the data. What is outliers?

We will discuss this later. And calculate the range from the remaining values. So it is interesting to note that when you have extreme values in the data set and these extreme values I can say that, they could be outliers. Outlier means that these are way apart from the centre of the data point. And these outliers can emerge from various sources. There could be, data entry issue.

(Refer Slide Time: 19:13)



Example

	Age	Frequency	Cumulative Frequency	
	18	6	6	
	19	16	22	
25th percentile:	20	11	33	The **31st** observation falls in this group
	21	9	42	
	22	17	59	
50th percentile:	23	13	72	The **61st** observation falls in this group
	24	6	78	
	25	5	83	
	26	4	87	
	27	3	90	
75th percentile:	28	3	93	The **91st** observation falls in this group
	29	4	97	
	30	5	102	
	30+	19	121	
	Total	121		

So now let us look at an example. And this example is based on hypothetical data. Let us assume that we entered an educational institution and we have a survey to conduct among the students. And we ask the various questions and we also collected some basic data like their age, their gender, their height, etc. And now we are interested to summarize and represent the data on the age of the students who responded to our survey.

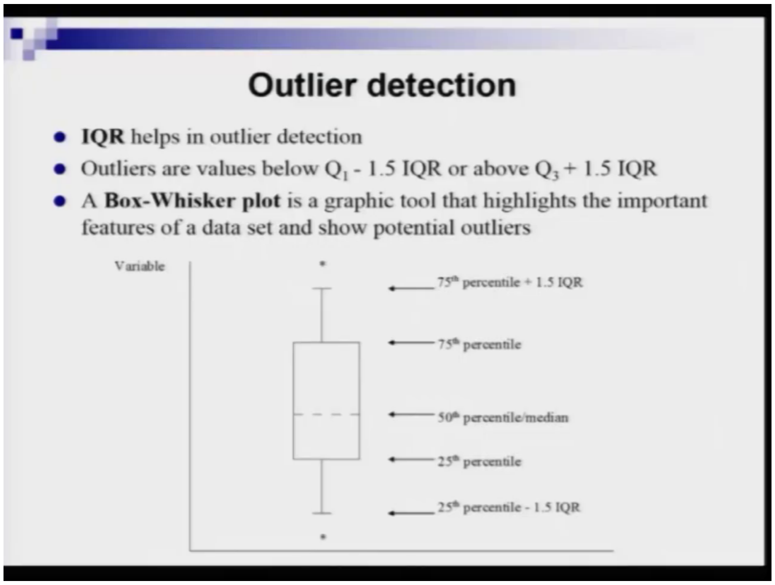
So suppose we surveyed 121 students and we have prepared a frequency distribution. And let us have a look at the frequency distribution and see how this percentiles, etc., could be found from that frequency table. So here the age starts from 18. So of course we do not expect student in higher educational institute aged below 18. So from 18, we see the progressive numbers and then, we say that if there is a student we came across who is aged more than 30 years we do not report their individual numbers.

We club everything together in one category 30 plus. That is basically aged more than 30 years. So for corresponding age groups we have frequencies and then of course we have computed the cumulative frequencies. Now cumulative frequency or the ogive that I have discussed in the previous lecture is going to be very handy to find out different percentile values. So let us start with the first quartile or the 25th percentile. So in the last slide I have shown you the formula. So you plug the values in the formula.

So then if you do that then you get the 31st observation falls in this group where age is 20. So you see that we do not observe a number 31 in the cumulative frequency table, we observe 33. So, we have to refer to this particular class or group and then the corresponding is 20. So hence 25th percentile or the Q1 or Quartile 1 for the age distribution is number 20.

Similar approach could be taken to find out the median or the second quartile or the 50th percentile number. And we plug the values of n and p there in the formula that I have shown in the last slide. And we come up with the number 61. The 61st observation falls in a group where I see accumulator frequency is reported as 72. So the corresponding age is 23 and that is the value of my median or 50th percentile. Similarly, we can find the 75th percentile from the table.

(Refer Slide Time: 22:45)



Outlier detection

- **IQR** helps in outlier detection
- Outliers are values below Q_1 - 1.5 IQR or above Q_3 + 1.5 IQR
- A **Box-Whisker plot** is a graphic tool that highlights the important features of a data set and show potential outliers

Variable

75th percentile + 1.5 IQR

75th percentile

50th percentile/median

25th percentile

25th percentile - 1.5 IQR

Now we are going to have a discussion on outlier detection. And outlier detection process is related to the concept interquartile range that we have discussed just sometime back. So let us have a look at how interquartile range could be used to find outliers. For that we have to consult a special type of diagram called Box Plot. So there is a mathematical or statistical result which tells us that outliers are the values below Q1 minus 1.5 times IQR or above Q3 plus 1.5 times IQR. So Q1 and Q3 are the first quartile values and the third quartile values respectively.

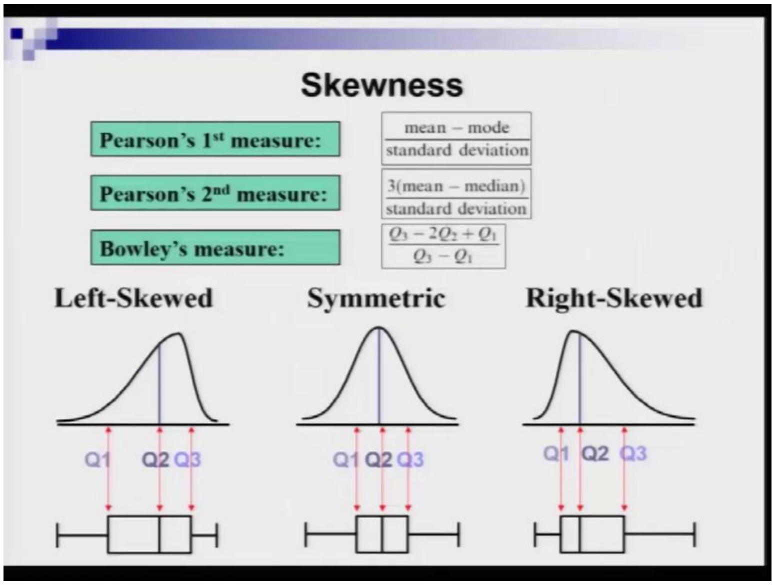
So Box Plot is a graphic tool that highlights the important features of a data set and show potential outliers. Note that along the y axis I am going to measure the variable values. It can be in any units. Now note that in the diagram or in the quadrant there is a large rectangular box. And the lower bottom line of that rectangle actually gives me the 25th percentile or the first quartile value for the variable.

The uppermost line or the border for this rectangular box shows the 75th percentile. It basically marks 75th percentile or the third quartile of my variable. And inside the rectangular box you see a dashed line parallel to x axis. That basically gives me the median or it actually tells me where my 50th percentile value lies.

Now note that two extremes along that vertical line again there are no, two parallel lines to x axis. Now if you look at the upper half or above that rectangular box, the parallel line, line parallel to x axis denotes the upper limit of this outlier detection formula which is 75th percentile or Q3 plus 1.5 times IQR. And the line parallel to x axis which is below the rectangular box gives me the lower limit which is Q1 or 25th percentile minus 1.5 times IQR.

So if there are some observations in the data set which are not in this range that is marked by this Q1 minus 1.5 times IQR and Q3 plus 1.5 times IQR formula then they are declared as suspected or potential outliers. And many softwares use asterix to demarcate these kind of observations. So here I am showing you 2 asterix. These two are the potential outliers as they lie way far from the centre of the data.

(Refer Slide Time: 25:44)



Skewness

Pearson's 1st measure: $\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$

Pearson's 2nd measure: $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

Bowley's measure: $\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$

Next we move on to the last feature of the frequency distribution and that is skewness. So there are two types of skewness measures. One is moment-based measure and the other one is non-moment based measures. So I am going to cover only the non-moment type measures of skewness because, I have not introduced the concept of moment. So let us not get into that. So again we see Professor Karl Pearson, the father figure in the field of mathematical statistics have offered two measures.

And his first measure is given by difference between mean and mode divided by the standard deviation of the data. He also has offered a second measure which is 3 times difference between mean and median divided by the standard deviation of the data. Now note that these measures actually have come from that empirical relationship between mean, median and mode that I have shown you earlier.

So if there is a difference between mean, median and mode in your data set, that actually tells you that your data is not symmetric, symmetrically distributed. So there will be some skewness. Now how to measure the skewness? You can use that empirical relationship to find out some formula as Professor Pearson suggested. Then there is also one measure proposed by Bowley and that measure is derived from the concepts of quartiles.

And you can see that there is a formula there, a bit complicated. Q3 minus 2 Q2 plus Q1 divided by Q3 minus Q1. And this is the famous Bowley's measure. Now is there any range for these measures? Yes, of course. So the Bowley's measure lies between minus 1 and plus 1. But there are no theoretical limits to the Pearson's measure, both first and second. But in practice the value is rarely very high. The measure of skewness, so when we come to the limit of measures proposed by Pearson there are two sets of results.

If we look at the first measure which is dependent on mean and mode there are no theoretical limits to that measure. But in practice the values is rarely very high. And if we focus on the second measure of skewness proposed by Pearson then that lies between minus 3 and plus 3. So note that in reality we can get three types of shapes for the skewed or non-skewed distributions. So of course, one is symmetric when, there is no skewness. So skewness takes value 0. And then there is positive skewed distribution. And then there is negative skewed distribution.

So now let us have look at the graph to get these concepts clear. So first we are going to talk about the negatively skewed distribution which is also called a left skewed distribution. So note that in the case of left skewed distribution, the mean is less than median and median is less than the mode. Now let us look at the symmetry distribution which is the middle diagram.

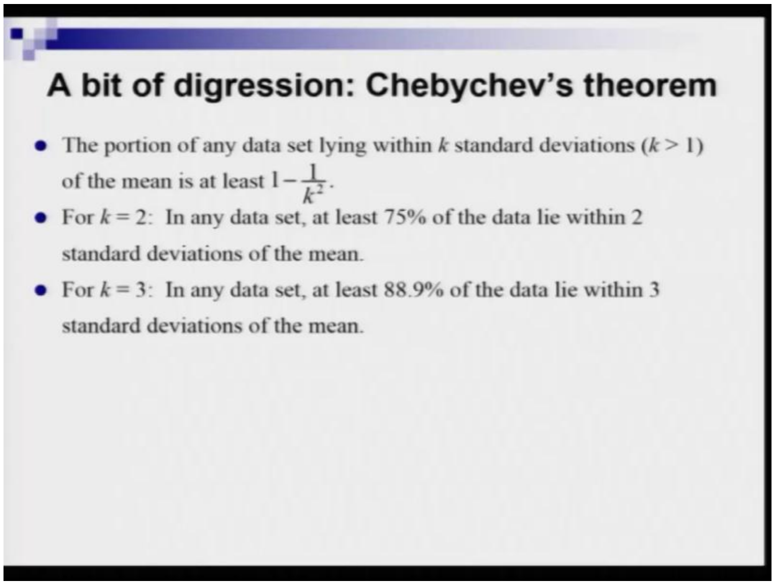
And here, of course, mean, median, mode take the same value. And then the third diagram is showing us the right skewed distribution or it is the positively skewed distribution. That means the skewness measured is positive. And here we note that there is this empirical relation where we find what the mode is higher than the median value and the median is higher than the mean value.

Now also note that these skewed distribution diagrams can be linked through the Box Whisker plot that we had discussed in the last slide also. So these red arrows, for each diagram are actually linking the first quartiles, second quartiles, third quartiles of the distribution to the Box

Whisker Plot. And I think you can see what for the symmetric distribution case, the quartiles are equidistant.

So the portions in the rectangle of the Box Whisker diagram, they are of equal size. But that is not the case when you are talking about a skewed distribution. So we are done with our discussion on basic descriptive statistics measures. But let us end the lecture with a bit of digression. But it is not a very large digression because the result I am going to show you is very much related to the concepts like, variance and mean that we have discussed in today's lecture.

(Refer Slide Time: 30:55)



A bit of digression: Chebychev's theorem

- The portion of any data set lying within k standard deviations ($k > 1$) of the mean is at least $1-\frac{1}{k^2}$.
- For $k = 2$: In any data set, at least 75% of the data lie within 2 standard deviations of the mean.
- For $k = 3$: In any data set, at least 88.9% of the data lie within 3 standard deviations of the mean.

So let us, discuss Chebychev's theorem. I am not going to discuss it in full rigor. I am not also going to show you any proof for the theorem. But I am just going to discuss the main result or the main statement of the theorem. And then I will show you illustrations how one can make use of this theorem or the result. So first bullet point gives you the statement of the Chebychev's theorem.

So Chebychev's theorem says that the portion of any data set lying within k standard deviations of the mean is at least 1 minus 1 over k square where k is definitely greater than 1. So if I now put the values for k equal to 2 or 3 then for k equal to 2 case we see that at least 75 percent of the data lie within 2 standard deviations of the arithmetic mean. And for k equal to 3, for any data set, at least 88.9 percent of the data lie within 3 standard deviations of the arithmetic mean.

Now interestingly I will comment on another point here and then I will conclude today's lecture. And that is that this Chebychev's result does not depend on the symmetry of the distribution. So that holds for any data set, any kind of distribution shape that we have in the data. So we are done with our basic measures of descriptive statistics. In the next lecture we are going to start the discussion on random variables. Thank you.