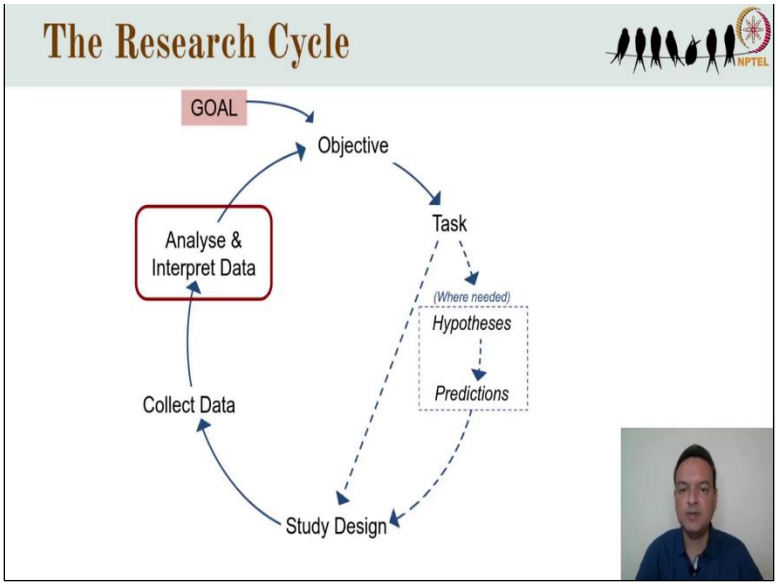
Basic Course in Ornithology
Dr. Suhel Quader
Nature Conservation Foundation

Lecture -26
Studying Bird Populations and Communities Part 1

(Refer Slide Time: 00:32)



Welcome to an introduction to data visualization and analysis. Before we get into the topic at hand, let's remind ourselves where we are in the generalized research cycle. As you remember, our broad motivation for research leads us to articulating some overall goal for a particular project. To meet this goal, we need to frame one or more objectives, often in the form of questions. From those questions are derived different tasks which might also be phrased as questions. Sometimes we can answer the task level question by direct observation,

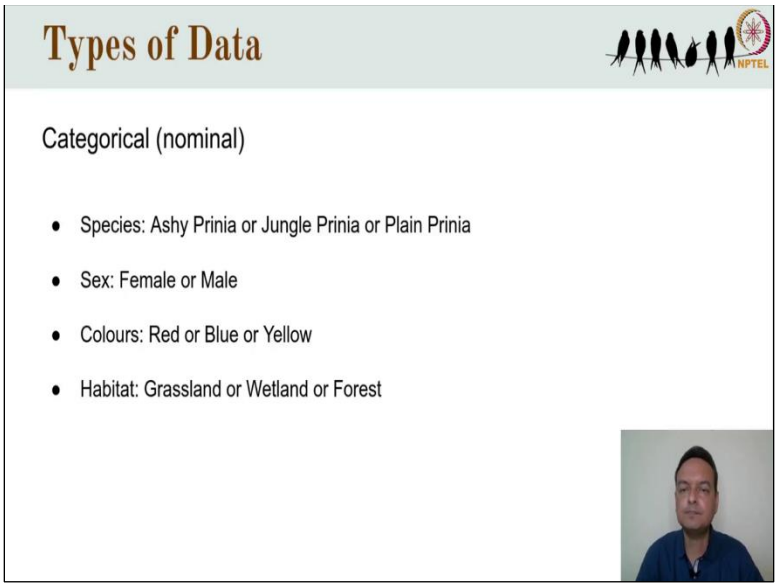
but in other cases where the phenomenon we are interested in can't be directly observed, we need to frame one or more hypotheses from which we deduce observable predictions that can be tested through observation or experiment. I spoke about this process in the lecture on methods of science and posing research questions. Once you have a clear idea of what exactly needs to be done, it is time to design your study keeping a number of key principles in mind which I have outlined in the lecture on basics of research design.

The next step is for you to carry out the study you have designed by setting up your experiment or making the observations you need to. This is not a trivial task in itself but we will skip forward to the point where you have your data, you have digitized it, and have checked for errors. And now you need to visualize, analyze and interpret the information you have collected such that you can say something about the question you started out with.

The basic tools you have available for this are data visualization and data analysis, and I will admit to you that it is not easy to become proficient in these skills. It is easy enough to follow the instructions in a statistics textbook but unfortunately good analysis and visualization does not result from following a fixed recipe. Rather we have to try and develop an intuition about data and how to treat data in order to extract the information we need.

For this we have to try and get as familiar as possible with numbers and other kinds of data and the ways in which we can work with them. So, to start us off on this journey, let's take a look at the different types of data we might deal with during the course of our research.

(Refer Slide Time: 02:37)



We start with categorical data in which each entity we look at has some state that we can name. So, these are sometimes also called nominal data. For example, this bird might be an Ashy Prinia and the other bird a Jungle Prinia and the third one a Plain Prinia. Or this Ashy Prinia might be male and the other female. Crudely speaking, we can also think of colours as categories like red

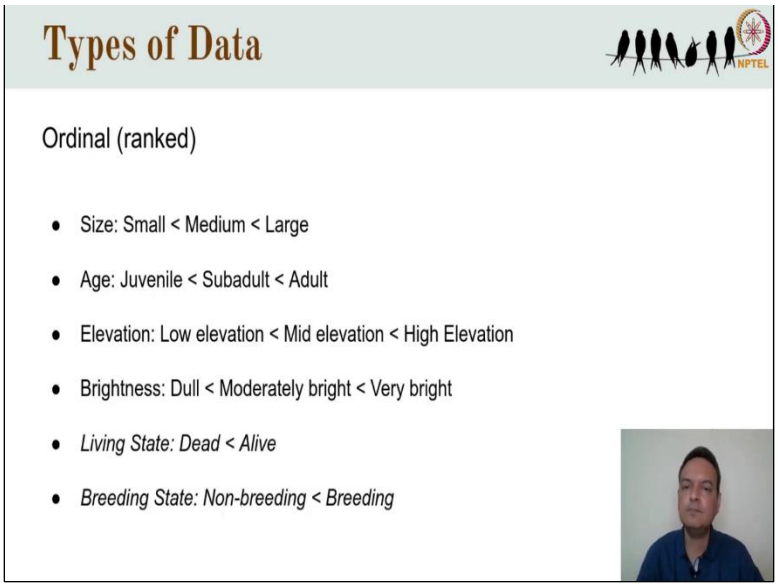
or blue or yellow. Another common category is habitat, like grassland versus wetland versus forest and so on.

Note that the categorical states are distinct entities with no obvious ranking among them. Unless we measure some other attribute of prinias, we cannot say that Ashy is ranked higher than Jungle Prinia (for example). This now changes with ordinal data. Here, the different states can be ranked with respect to each other. Examples include categories of size, like small medium and large; or categories of age, like juveniles, sub-adult and adult.

And here we can actually order the states by the attribute of interest and can say that adults are older than sub-adults which in turn are older than juveniles. This ordering cannot be done for categorical variables. Note that although the different states can be ranked with respect to each other, the interval between ranks need not be uniform and may not be known at all. So, although we know that an adult is older than a sub-adult there is no information inherent there about how different they are in age nor whether the difference is the same as the difference between a sub-adult and juvenile or not.

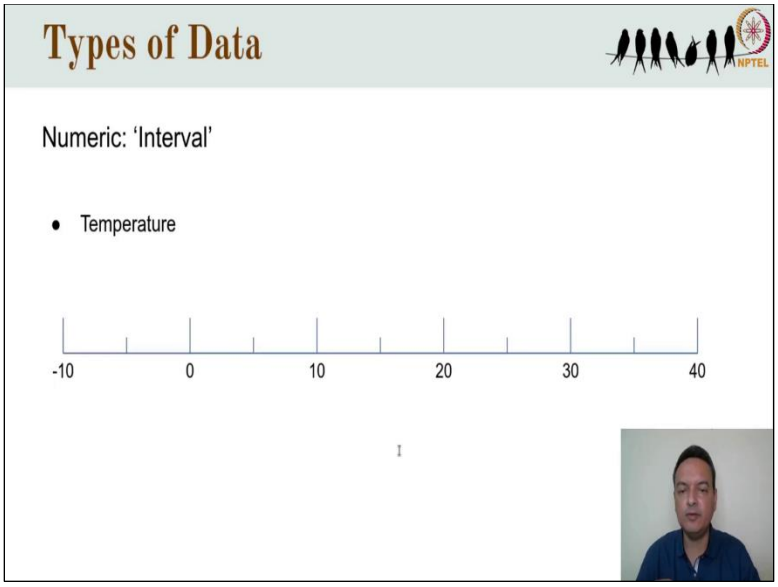
Sometimes people denote ordinal data in numeric form for example labeling juveniles 1, sub adults 2, and adults 3 (for example), but you should be very careful to remember that these data are nevertheless not numeric and do not possess all the properties of numbers that we usually take for granted.

(Refer Slide Time: 04:38)



A particular kind of ordinal data is the kind of dichotomous data shown in the last two examples, although dead can arguably be ranked as less living than alive in practice these kinds of data are usually treated as categorical rather than ordinal.

(Refer Slide Time: 04:58)



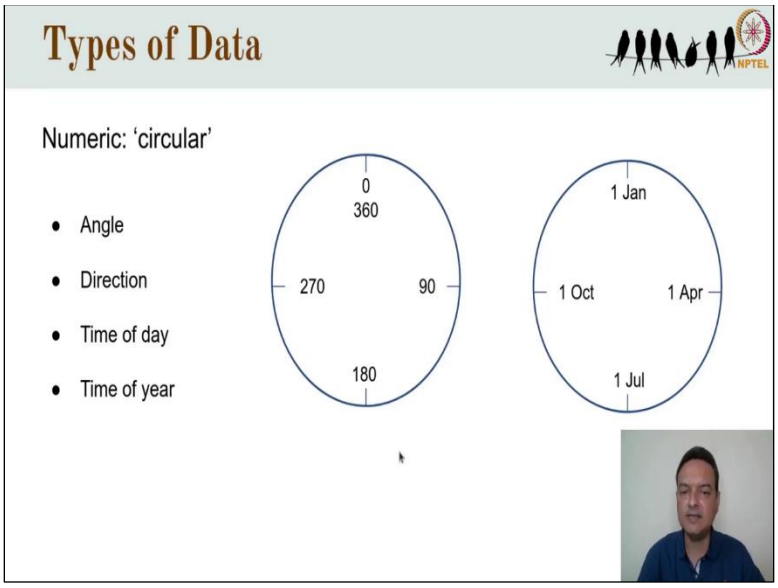
This brings us to numeric data, and in particular interval data, which can not only be ranked from low to high but also have uniform differences or intervals. Take temperature in degrees celsius for example. We know not only that 40 degrees is hotter than 30 degrees Celsius but also that the difference between 30 and 40 degrees is the same as that between 30 and 20 degrees. Further, we can also say that the difference between 40 and 30 degrees is twice as much as the difference between 30 and 25 degrees.

So, the intervals are uniform such that differences are meaningful, and the differences themselves can be compared with each other in various ways. But if you think about it, although you can say that 20 degrees is 10 degrees hotter than 10 degrees you can't say that 20 degrees is twice as hot as 10 degrees. Similarly, you can't say that 30 degrees is 1.5 times as hot as 20 degrees. In other words taking the ratio of two numbers on an interval scale does not make sense. Why is that?

It is because interval scale numbers have no true 0. Zero degrees Celsius is just an arbitrary convention - the freezing point of water. If we used a different convention like assigning 0 as the freezing point of alcohol which is at -114 degrees Celsius then what we now call 20 degrees would actually be 134 degrees and 10 degrees would be 124 degrees. The difference between them remains the same but the ratio completely changes.

Similarly, if we used Fahrenheit as our arbitrary convention instead again there is no true 0. We could measure temperature in Kelvin if we wanted a true 0 but as commonly measured temperature is on the interval scale with no true 0 and therefore ratios of temperatures are not meaningful.

(Refer Slide Time: 06:54)



Now measures that are associated with a true 0 are called ratio scale data because ratios of such numbers make sense, in addition to the intervals being uniform and there being an order among them. Examples include length, area, weight, counts and so on. All of these have true zeroes and

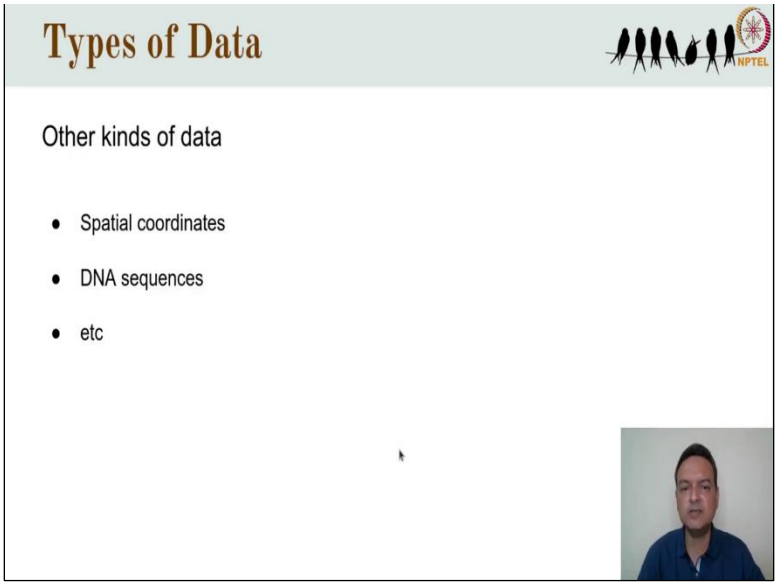
saying that the height of this tree is 1.3 times the height of another or that I have counted twice as many birds in the park compared with at the lake – all of this makes sense.

So, these are the most common types of data that we would usually handle but a quick word about a couple of other data types. Circular data are similar to interval data and that there is no true 0. Examples of circular data are angle, direction, time of day and time of year. In all cases, the so-called starting or 0 point is arbitrary – at the top of the paper for direction; at midnight for time of day; near mid-winter for time of year.

Another key point to remember is that in circular data the numbers wrap around. This means that the difference in angle or direction between 350 degrees and 10 degrees is not 340 degrees rather it is only 20 degrees. So, beware of mechanically performing the usual operations on circular data and treating them as though they are arranged on the standard linear number line.

Now in many situations we may be dealing with only a restricted part of the circle.
For example, the slope of a mountain side might vary only between 0 and 90 degrees -- usually more like 0 and 40 degrees, and within this range we might decide to treat the data as if it were on linear scale. But if you are dealing with data that go all around the circular scale then it is best to look up a special branch of analysis called circular statistics.

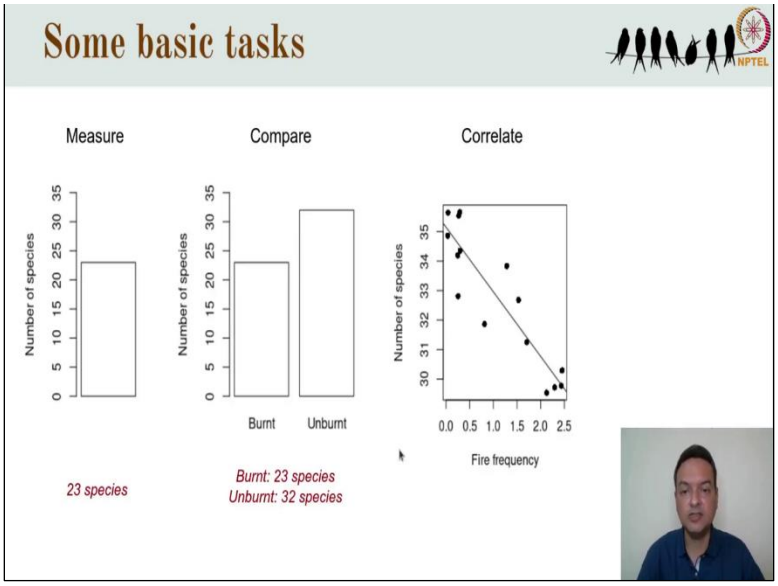
(Refer Slide Time: 08:50)



And there are of course other kinds of data as well. For example, location data are increasingly common in our studies where spatial coordinates like latitude and longitude are key attributes. And if spatial analysis is of a particular interest to you then there is a distinct topic of spatial statistics that you should learn about. In molecular ecology, the raw data might come in DNA sequences or allele frequencies and so on.

Again, there are specialized ways of dealing with these. But for the rest of this video, we will focus on the most common data types in our field. We will talk a bit about categorical data but really the emphasis will be on numerical data, in particular ratio type data for which there is a true 0.

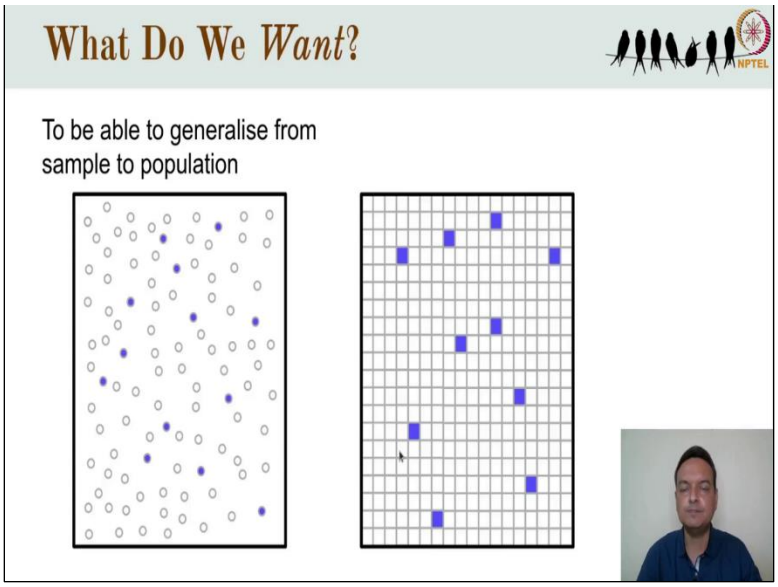
(Refer Slide Time: 09:40)



But before that let's recapitulate what the most common data related tasks are that we would encounter in our research. Sometimes our task is to measure a single thing in a single place, for example the number of bird species in a grassland. More often it is a comparison that is of interest. Here we are comparing the number of species in burnt and unburnt grassland to see which category of grassland has more species and by how many.

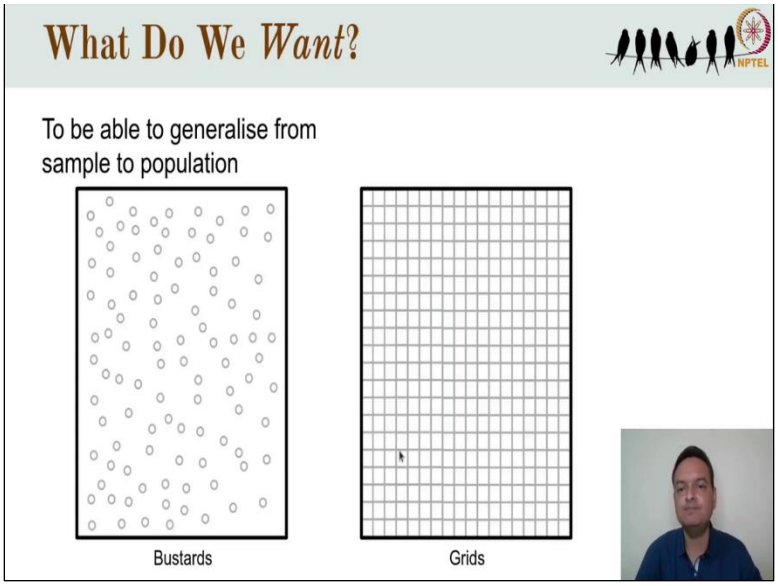
Note there may be more than two categories to compare but we will restrict ourselves to the simple case of two categories. Another common task is to look for associations or correlations between two measures leaving aside the question of whether there is a causal relationship between the two.

(Refer Slide Time: 10:34)



Regardless of the specific task, we need to be very clear on what we are trying to do in a broad sense. In almost all situations we want to understand some larger population through our sample. If you recall, the population is the larger frame to which we want to be able to generalize. Since we typically cannot measure the entire population, we take a sample from the population and we try to design our sampling strategy such that the sample is representative of the population.

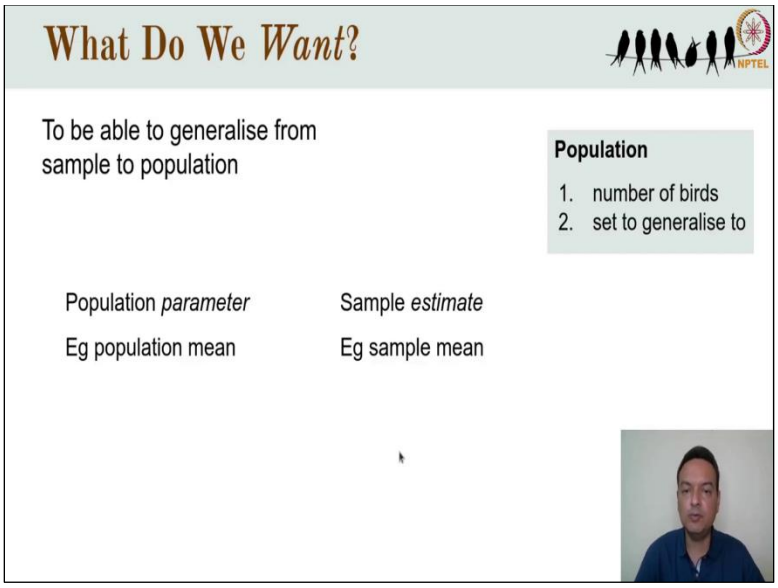
(Refer Slide Time: 11:01)



The word population is used in a statistical sense here. It may refer to an actual population like the population of Great Indian Bustards in Rajasthan or it may refer to a metaphorical population such as the population of all possible one hectare plots in Desert National Park. The whole point of our

study design data collection and analysis is to be able to draw conclusions about the population as a whole from the restricted sample that we are actually able to measure.

(Refer Slide Time: 11:33)

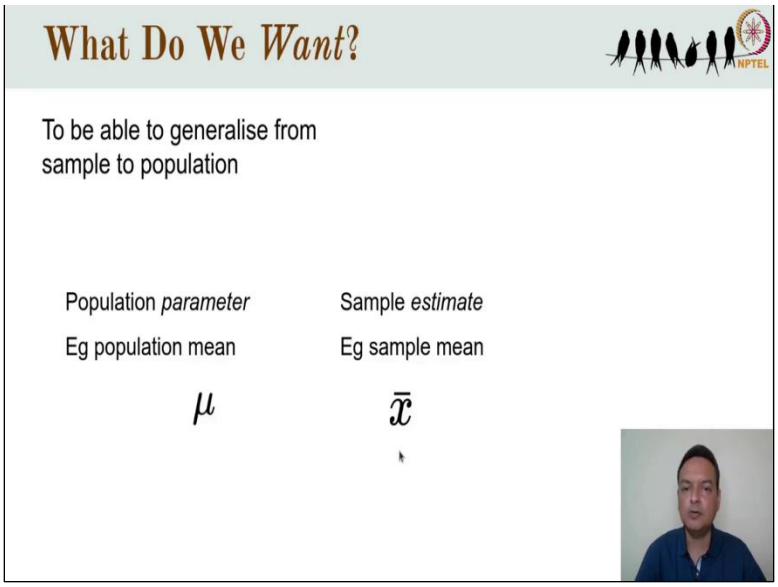


Now some jargon is needed here. That aspect of the population which we are trying to measure is called the parameter. And what we actually measure through our sample is called the estimate. So, for example if we wanted to understand the population density of bustards, we do so by estimating the sample mean density and of course hoping that it is not too far from the true population mean density.

Notice that I use the word population in two senses here -- first to refer to the numbers of bustards and second to refer to the largest set that we want to generalize to. I have done this deliberately not with the intention of confusing you but rather to illustrate that different areas of science use the same word to mean different things and we should guard against the resultant confusion. Here one use of the word population is from the discipline of demographic analysis and the other use is from the discipline of statistics.

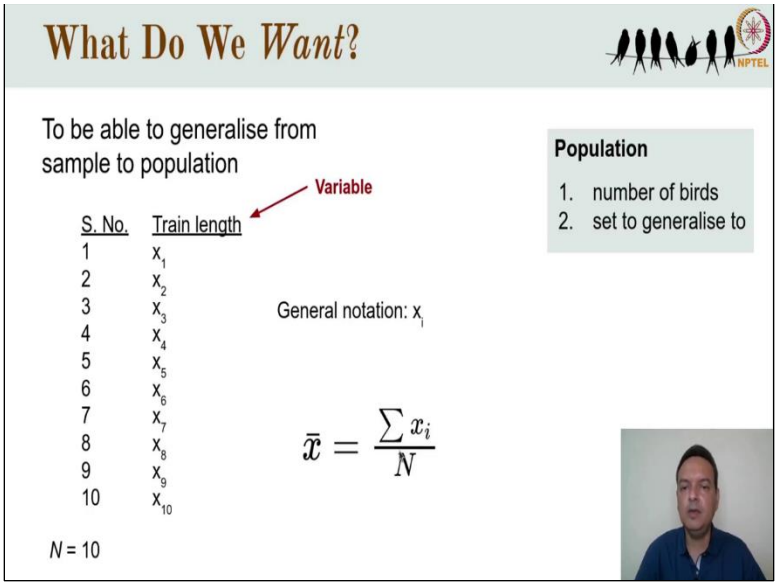
Similar examples of multiple meanings for the same word include the word error and the word hypothesis both will be described a little bit later.

(Refer Slide Time: 12:50)



Another thing that is useful to know is a bit about conventions of statistical notation. One convention is that population parameters are usually written in Greek letters for example the population mean is often written as μ. While on the other hand samples are denoted in Latin letters for example what you measure is usually written as *x* and the convention for the sample mean is to be written in that same letter with a horizontal bar on top, in this case $\bar{x}$.

(Refer Slide Time: 13:21)



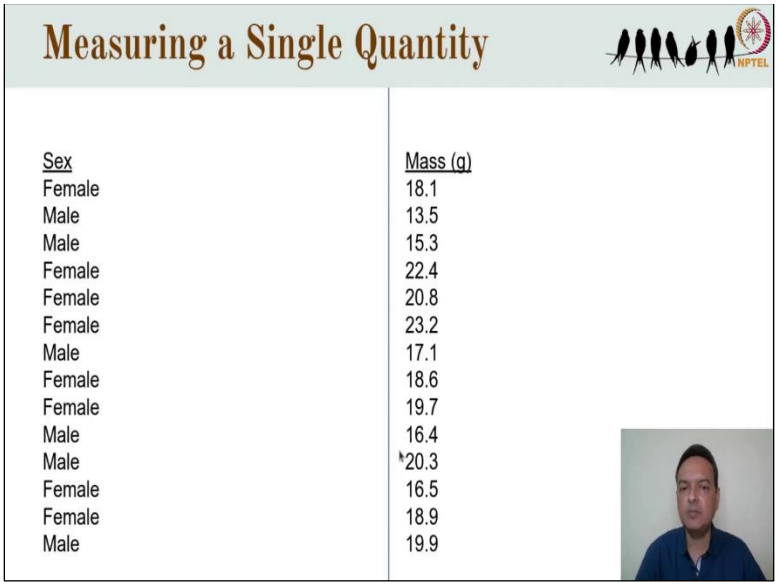
And another convention is to denote specific samples with numbered subscripts. If I measure the trains of 10 peacocks then I could call those 10 measurements x_1, $x_2$and so on until x_{10} – and the general notation would be $\bar{x}$. So, for example the formula for the sample mean would look like this:

$$\bar{x} = \frac{\sum x_i}{N}$$

Another technical word is the term variable which refers to the aspect being measured. Here, Peafowl train length in the notation above we denote the variable by x according to convention but we could just as well denote it by y or z or anything else we wanted.

So, with all this in mind let us start discussing our first possible task which is to measure a single quantity.

(Refer Slide Time: 14:21)

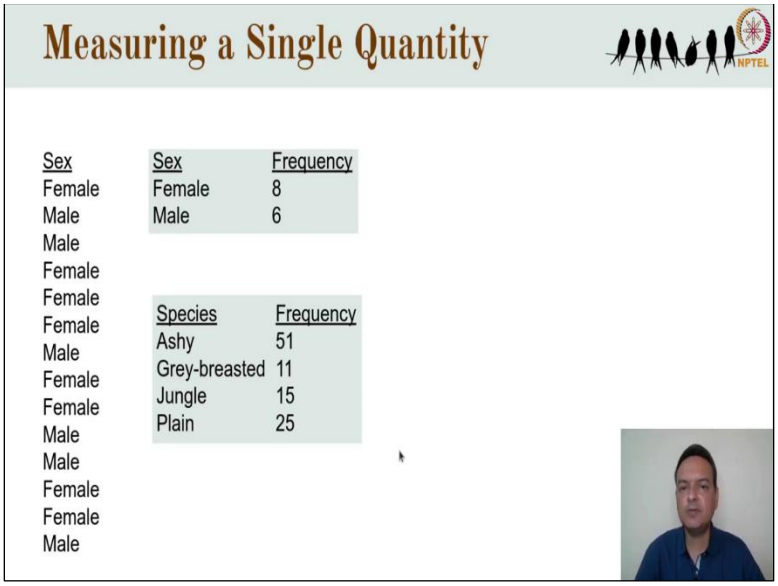


So, a common task is estimating a single specific quantity of interest, whether that is the sex ratio of Great Indian Bustards in desert national park or the density of peafowl in a forest or the average mass of House Crows in a city or the average time spent by Magpie Robins in singing during the breeding season and so on. Remember that through our sample we are estimating what the true quantity might be for the population as a whole that we want to generalize to.

So, let's say we have carefully designed our study as described in the earlier video in this course. and now we have the raw data with us. For a categorical variable such as sex, the raw data might be a series of labels in which each sampled individual is represented by the label male or the label female. For a numeric variable such as body mass, we would have a series of numbers each representing a single individual.

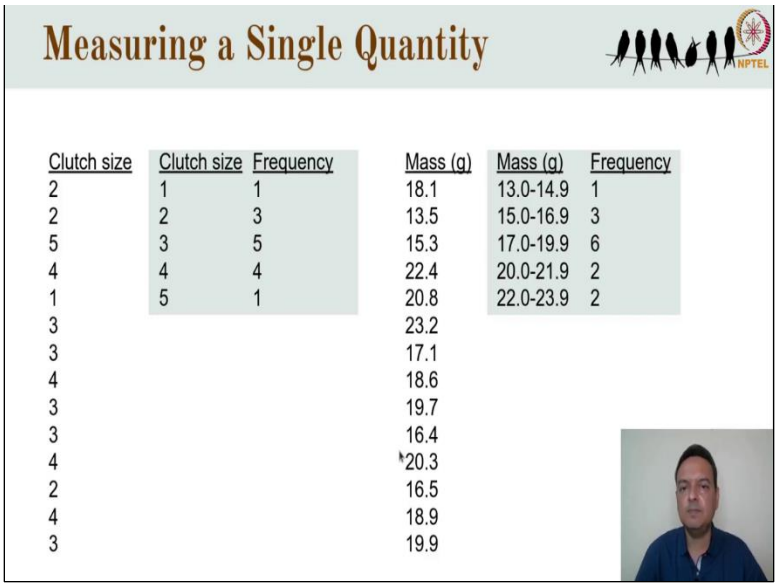
Note that although in these cases each data point represents an individual it could also represent a nest or a transect or a window of observation time and so on. In other words, the raw data depicts some property of the sampling unit, whatever that sampling unit might be. Now we usually cannot make much sense out of a series of labels or numbers. And so, the raw data needs to be summarized in some way.

(Refer Slide Time: 15:45)



For categorical data the obvious way to summarize is to count up the number of individuals in each category – here there are eight females and six males. I am using here an example with only two categories for simplicity but of course there could be multiple categories. like if you had fifty-one Ashy Prinias, 15 Jungle Prinias, 25 Plain Prinias and 11 Grey Breasted Prinias. So, these two summaries are examples of what is called a frequency table.

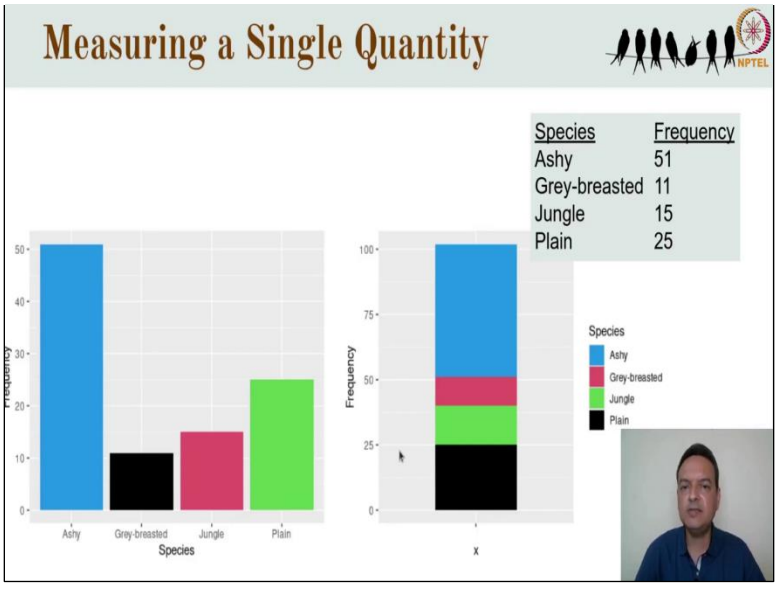
(Refer Slide Time: 16:16)



If the raw data are numeric, the usual next step is to summarize them also in a frequency table. If the numbers can take only integer values over a small range, like the number of eggs laid by a female, then we can just count up how many females laid a certain number of eggs – from one egg to five eggs – here. So, there are two columns in the summary: the number of eggs and the corresponding number of data points in this case females.

And this is also known as the frequency of data points hence the term frequency table. If the range of integer values is large or if the numbers are not integers but rather decimals then we create bins of values and count up how many data points fall into each bin just as before.

(Refer Slide Time: 17:02)



What information can we extract from these frequency tables and how are those related to the population parameter that we are trying to estimate? For categorical data, we often want to estimate the proportion of the population that falls in one category or another and this can easily be calculated from the frequency table. Remember that this is only the sample proportion and we hope (but we do not know for sure) that it is somewhere near the population proportion.

We can visualize the data using a bar graph either side by side or stacked on top of each other. Notice though, that the visual comparison between the two numbers is easier when the bars are side by side. When there are only two levels – here male and female – this visualization does not really add much insight: we can easily just look at the frequency table and understand it. But when there are multiple levels, the visualizations can be easier to understand than the frequency tables themselves.

Here again the bars can be arranged side by side or they can be stacked. The purpose of the visualization is to be able to easily see which categories are more or less frequent and to look at their relative magnitudes. When doing so it is useful to order them in a sensible way.

(Refer Slide Time: 18:21)

If there is no natural way to order them; you could just arrange them in decreasing frequency for a quick comparison. So, this kind of bar graph provides for a very quick and easy comparison in the frequencies of different categories.

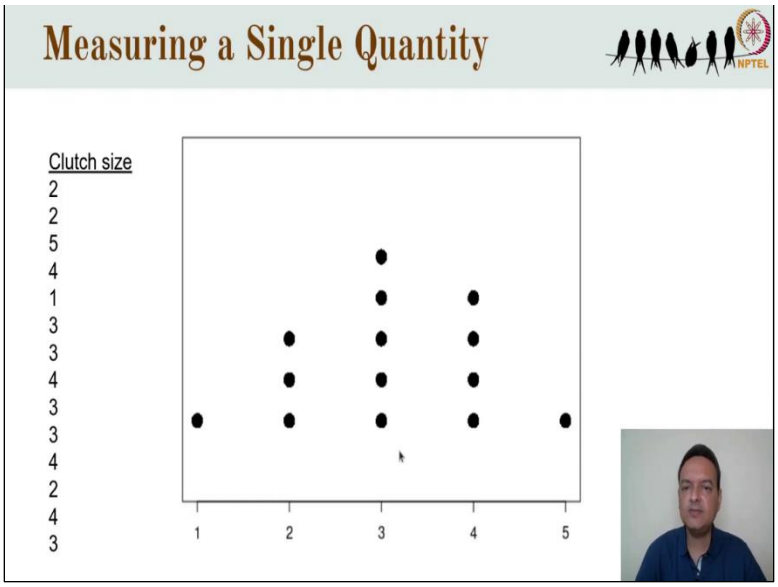
(Refer Slide Time: 18:35)

Now you will often have seen pie charts used to visualize frequency data of this nature but I strongly recommend against doing so for any serious purpose. That is because in bar charts the frequencies are represented by the length or height of the bar and the human eye is very good at estimating relative length. But in pie charts the frequencies are represented by the area of the slice of the pie and the human eye is actually very poor at comparing relative area.

So, do think very carefully before using pie charts in any formal visualization of your data.

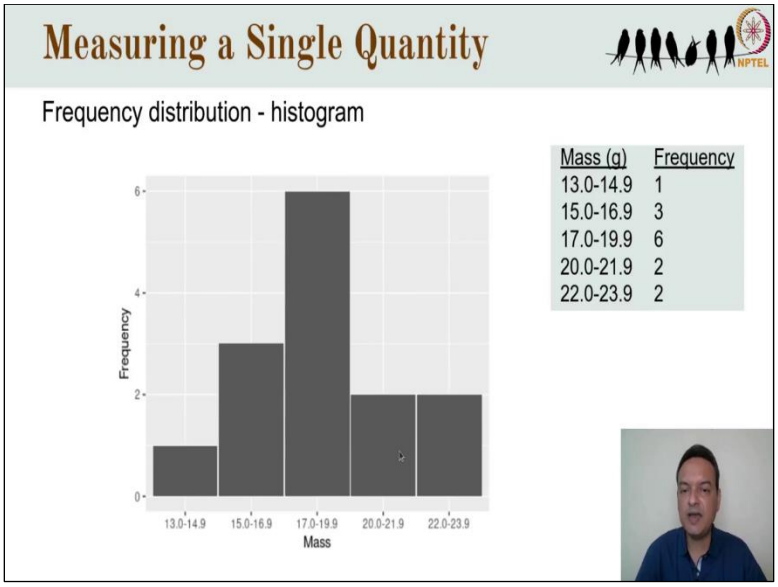
Now let's look at what visualizations we can use for numeric data.

(Refer Slide Time: 19:18)



We can create a dot chart when the data take on integer values, the x-axis here shows the values and the y-axis has one dot for each data point that has that value. We can easily see the minimum and maximum which here are one and five and we can also see that a clutch size of 3 is the most common. Now for numeric data that have decimals let's look at the frequency table.

(Refer Slide Time: 19:46)

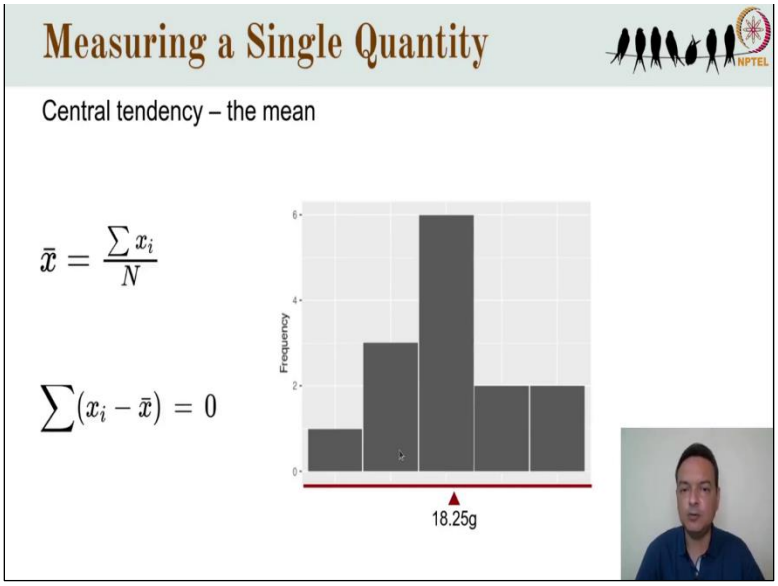


The frequency table also tells us something about the range of the data the max and min and the most frequent bins. But as soon as the table has more than a handful of bins all this is more easily visible by converting the frequency table into a graph. Bar charts of numeric frequency tables are often called frequency distributions or histograms, and from this histogram you can see that the lowest value is somewhere near 13 grams and the highest value is somewhere near 24 grams.

We also see that the most common values are roughly at the midpoint of the lowest and highest values somewhere near 18 grams. Now this is not necessarily true of course – the most common values could easily be away from the midpoint, as we will see shortly. So, two immediate impressions we get from a histogram are what are the what the typical values are and what the span of the data is.

The first is referred to in statistics as the central tendency and the second is the variation. Let's talk about the central tendency some more.

(Refer Slide Time: 20:53)



Very often what we are trying to estimate about a population from our sample is the population's central tendency. The best known such measure is the mean or average; this is a very familiar measure to us all. Perhaps so familiar that we do not usually think about it very much. But what exactly is the mean? Mathematically of course, it is the sum of all observations or numbers divided by the sample size.

$$\bar{x} = \frac{\sum x_i}{N}$$

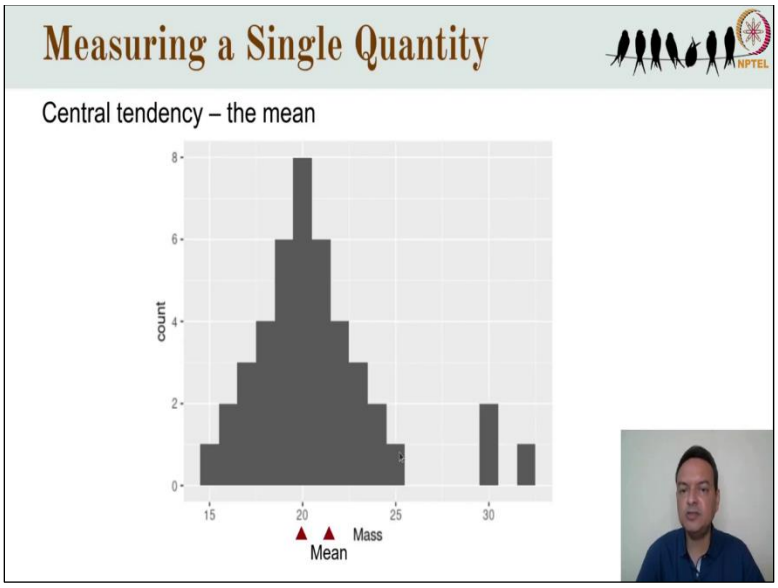
Conceptually, it is the central gravity of a histogram. If the histogram were like a see-saw then the center of gravity would be that point which perfectly balances the two arms. Mathematically, if

you were to take the distance of each point from the mean and take an average, the sum of all the negative values would exactly equal the sum of the positive values. And so, the average deviation from the mean is 0.

$$\sum(x_i - \bar{x}) = 0$$

The implication of this is that the addition of just a few extreme numbers in the data set can affect the mean considerably.

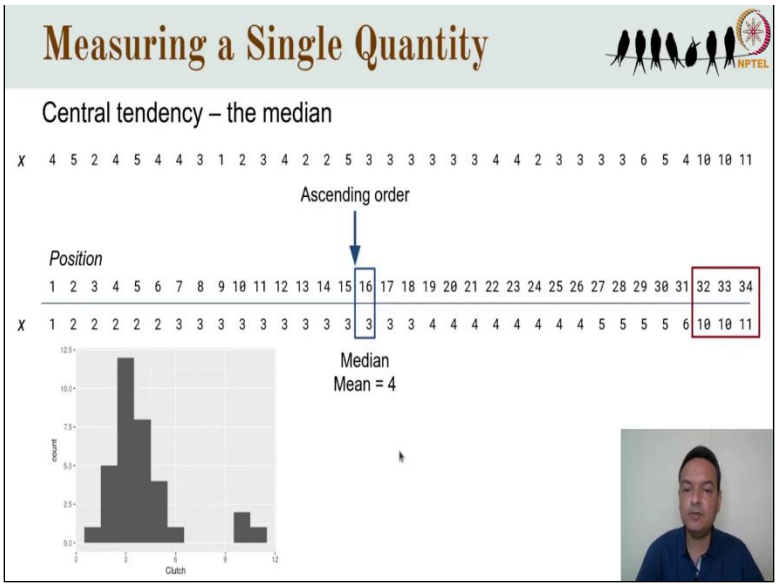
(Refer Slide Time: 21:56)



Let's look at this data set of the mass of 40 individuals. You can see that the distribution is perfectly symmetrical and it has a mean of 20 grams. If we add just three individuals who are unusually heavy then the mean increases by quite a bit changing our estimate of the typical mass of the species. Now the mean is not the only way of measuring central tendency – there is also the mode: the bin or value with the highest frequency.

In this example the bin with the highest frequency remains unchanged with the addition of three unusual individuals, it is still 19.5 to 20.5 grams. So, the mode is unchanged.

(Refer Slide Time: 22:38)



But we will spend a little more time on the *median* which is that number that divides the data set into two equal halves, with 50% of the values lying below the median and 50% above. Here is an example of a series of numbers denoting clutch size -- that is the number of eggs a female has laid in a nest. Now this is the raw data but we arrange the data in ascending order and since we have a total sample size of 31, which is an odd number, the median is the value in the 16th position.

So that there are 15 data points lower than the median and the same number of data points larger than the median. If the total sample size was an even number (say 30) then the simplest way to compute the median would be as the midpoint of the 15th and 16th number. Now this is a relatively symmetrical distribution. So, the median, which is 3, and the mean, which is 3.4, are pretty close to each other.

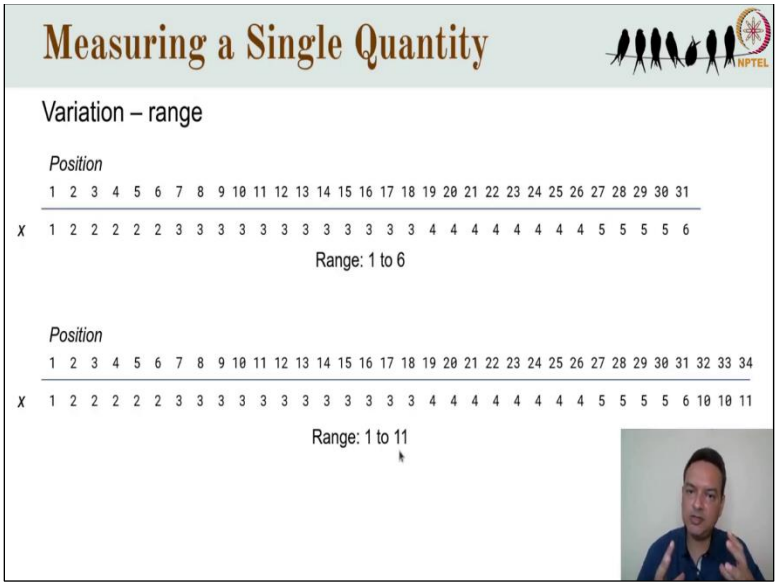
But unlike the mean, the median is not all that sensitive to the presence of small numbers of extreme data. Suppose there were three individuals with an unusually large clutch size say 10, 10 and 11 we do not know whether those are two clutches that is laid by a single female or whether eggs might have been added through intra-specific brood parasitism where another female of the same species laid her egg in the nest.

Nevertheless, we have these extra data -- the extreme values. And the mean then jumps to four eggs but the median stays exactly where it was at 3 eggs, just as before. This property of the median – that it is relatively unmoved by unusual or extreme data is called robustness and we say that the

median is a more robust measure of central tendency than the mean. Now despite this, the vast majority of studies in ornithology and ecology use the mean, and we will largely follow that for the remainder of this video.

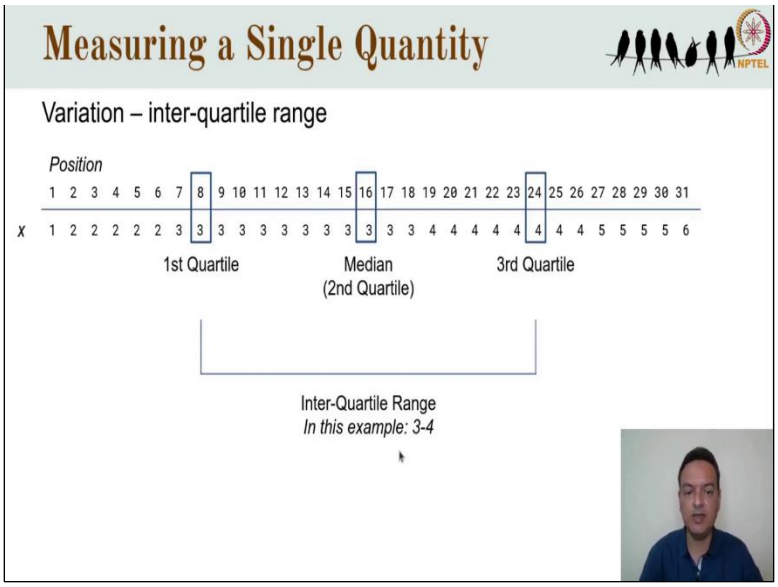
But I would urge you to carefully consider using the median and not just default to using the mean just because that is what others have done.

(Refer Slide Time: 24:53)



Apart from central tendency, the other key thing that we want to understand about a single quantity is its variation -- both because it is interesting in its own right, and also for what it implies about other aspects of data analysis and interpretation. The simplest and most crude measure of variation is the range -- that is, the difference between the smallest and largest values in the data: the minimum and maximum. But the range is unsatisfactory for two reasons. First, it is not a robust measure. And so, a single value can change it considerably, for example taking a data set from earlier the range of these 31 clutch sizes is 1 to 6 but add just those three extra data points and the range is now 1 to 11. So, the range is clearly not robust in the sense that we have used earlier. And a second issue is that most data distributions are clustered around the central tendency and the range does not tell us where the *majority* of data lie which would be something very interesting for us.

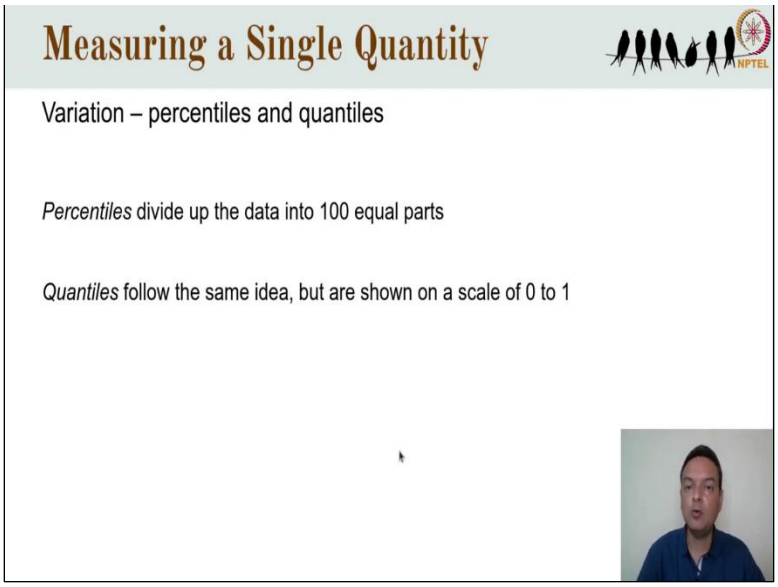
(Refer Slide Time: 25:59)



Let's see if we can learn from the use of the median that we discussed earlier. As you know the median is the value which divides the data set into two equal halves, with half of the data points lying below the median and half falling above. We can do something similar using a different criterion, for example, we can define a point that separates the lower 25% of the data from the upper 75%. Let's call this point a quartile because one quarter of the data lies below.

Its mirror point is that which separates the lower 75% of the data from the upper 25%. Now we take these two quartiles and the median and put them on the sorted data or the frequency distribution of the data and then we have divided the entire data set into four equal parts. One quarter of the data lies below the first quartile, another quarter lies between the first quartile and the median which we can call the second quartile. Yet another quarter lies between the median and the third quartile and the final quarter is above the third quartile. So, now we could use the range between the first and the third quartiles as our measure of variation. It tells us the limits within which *half* of the data lie. But of course, we do not have to restrict ourselves in this way – we can define any other limits.

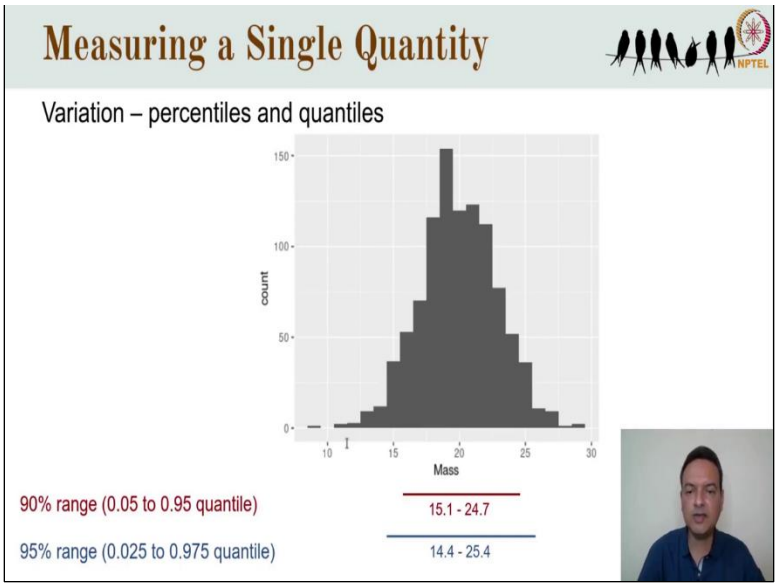
(Refer Slide Time: 27:23)



We could just as well divide the data into say a hundred equal parts, then 1% of the data would lie below the first of the corresponding dividing points another 1% would lie between the first and second dividing points and so on. These dividing points are called percentiles.

Since they cut up the data into 100 equal parts, percentiles are easy to interpret: the 10th percentile is the dividing point between the lowest 10% of values and the upper 90% of values. The 25th percentile is the first quartile; the 50th percentile is the median; and the 75th percentile is the third quartile. You are familiar with this idea from exam scores. If a thousand students write an exam, and your rank in the exam is 10 then you scored in the 99th percentile: only 1% of students scored higher than you. Further when these points are represented on a scale of 0 to 1 rather than 0 to 100, they are referred to as *quantiles*, making the median the 0.5 quantile the first quartile is the 0.25 quantile and so on.

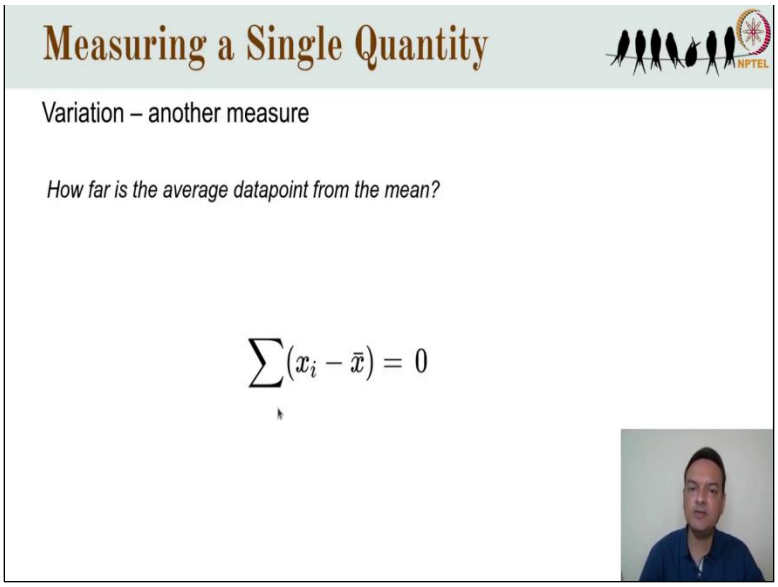
(Refer Slide Time: 28:38)



So, armed with this understanding of quantiles, you will now see that we can choose to represent the variation in our population in any number of ways. For example, here is a distribution of 1000 body masses. We might want to know the points between which 90% of the values lie, in other words which is the 0.05 quantile and the 0.95 quantile. And so, we see that 90% of the body masses lie between 15.1 and 24.7 grams or let's say we want the points between which the central 95% of the values lie -- that is the 0.025 and the 0.975 quantiles. Here we see that 95% of the body masses lie between 14.4 and 25.4 grams. If you are not familiar with quantiles from before all this may be a bit confusing. But I am spending considerable time on them because you will come across them often in your readings and they are fundamental to a further understanding of data and its properties. So, please do watch this section multiple times until it makes sense to you and also play around with numbers on a piece of paper or in a spreadsheet program, sorting them and counting up places until you feel you have gained an intuitive understanding of quantiles.

So, much for quantile-based measures of variation for now. Let's talk about another measure which many of you will be familiar with.

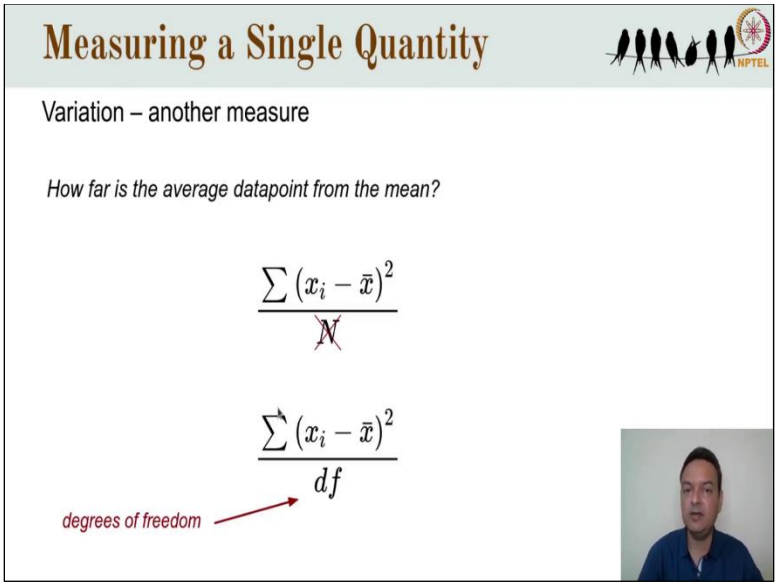
(Refer Slide Time: 30:12)



Let us say we wanted to construct a measure of variation as something that tells us how far the average data point is from the mean. Taking this literally, we can calculate how far each data point is from the mean and then take an average. Unfortunately, as discussed earlier, one of the properties of the mean is that the sum of all deviations from it is 0 with the positive deviations exactly equal to the negative deviations. So, this does not work.

Now we could potentially get rid of the negative sign by taking the absolute deviation but instead we will use another method of removing the negative sign which is by squaring.

(Refer Slide Time: 30:52)

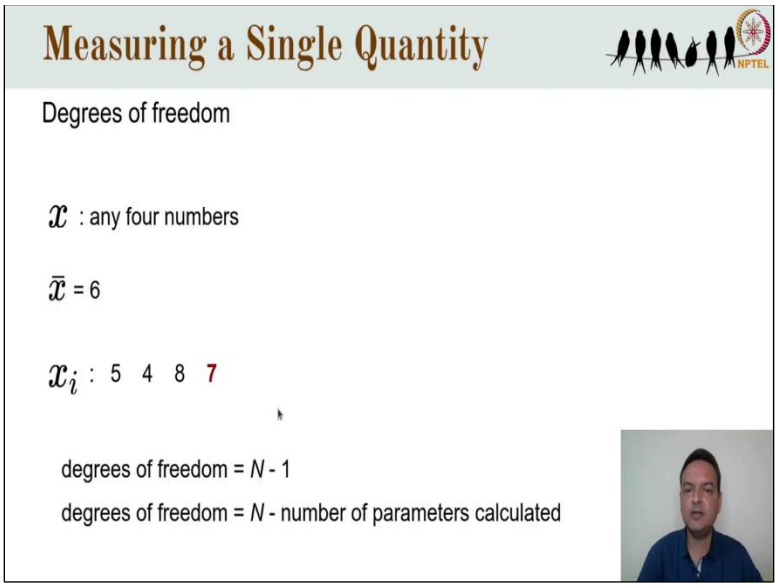


We square each deviation and add all the squared deviations up and take the mean. Now, one further complication, when taking the mean, we divide the sum of squared deviations not by the sample size (that is how many numbers we have) but rather by what is called the degrees of freedom (*df*)

$$\frac{\sum(x_i - \bar{x})^2}{df}$$

So, let us make a quick detour to understand what degrees of freedom means.

(Refer Slide Time: 31:17)

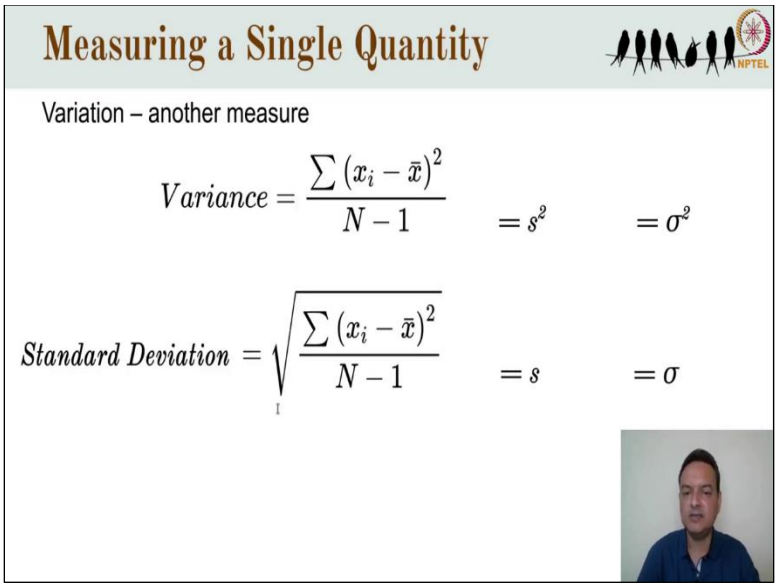


Suppose I tell you that I have a series of four numbers without any more information you are free to imagine absolutely any four numbers and we say that the degrees of freedom in this case is four. Now imagine that I also tell you that the mean of these four numbers is 6. Now, a constraint has been added which means that you cannot imagine absolutely any four numbers anymore. I can pull out 1, 2, 3 numbers out of a hat but once I have reached three numbers the fourth one has to be a particular number such that the overall mean is what I told you.

Let's try this out. I can choose anything for the first number say 5, then I can choose anything for the second number say 4 and I can choose anything for the third number say 8. Now, once you have chosen three of the four numbers the fourth number can be only one thing. If the mean of the series is set at six, the fourth number *has* to be seven. No other number in the fourth place would give a mean of six.

So, I had the freedom to choose any three numbers but once those three numbers are chosen the fourth was fixed. We say in this situation that the degrees of freedom is (*N-1).* In general, the degrees of freedom to be used in a calculation is the total number of data points minus the number of parameters previously estimated from the data. In this example, we have calculated one parameter – the mean -- so the degrees of freedom is (4 – 3). In fact, in many statistical applications, it is more appropriate to use the degrees of freedom rather than the overall sample size. So, with that in mind let us get back to our measure of variation.

(Refer Slide Time: 33:05)



We divide by the degrees of freedom rather than the sample size. And so now this quantity is the average square deviation from mean. This is a standard statistical measure called *variance* and since the numerator has square deviations, the variance is in square units. If x is centimeters of peacock's tail then the mean may be 130 centimeters but the variance is say 25 centimeters squared. To bring the variance back to the original units, we have to take the square root so, 5 centimeters and this is called the *standard deviation.*

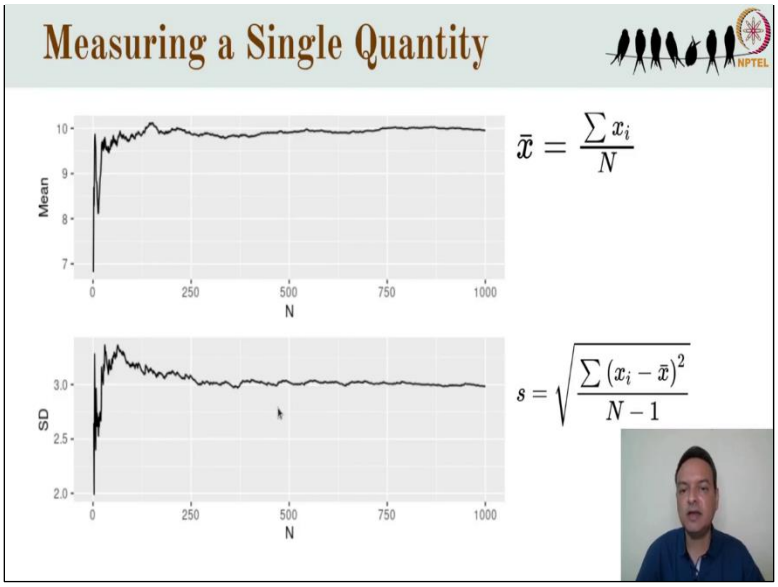
Recall also that these calculations are being performed on a sample from the population you want to generalize to. So, what we have is the sample standard deviation, ‘s’, which you hope is not very far from the population standard deviation, sigma. Again, Latin letter for the sample estimate and Greek letter for the population parameter. Variance sometimes is not given its own symbol

but rather is represented as s square or sigma square depending on whether we mean the sample variance or the population variance.

Now intuitively you can think of the standard deviation as the average distance of data points from the mean but in actual fact it is a bit larger than the average distance because squaring the distances before averaging gives extra weightage to more extreme data points, something that is generally worth remembering. By the way you might also see a formula for standard deviation that has N in the denominator rather than N – 1. And the difference is this: if you have the full population that you are interested in, then you would use N; if you have a sample from the population then the degrees of freedom N - 1 is what is used. In nearly all research, we are sampling from a larger population and so we use the degrees of freedom rather than N.

Now one point that people often get confused about is what happens to the standard deviation when sample size increases. Let's look at this by first examining what happens to the mean.

(Refer Slide Time: 35:23)

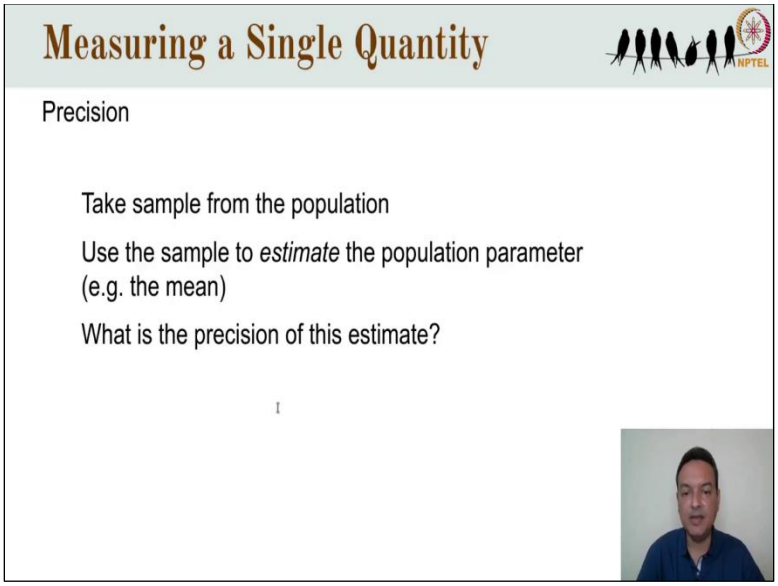


Let's say, we have a series of numbers which is our sample from the population. We start with the first number in our sample and then the second and the third and as we proceed with our sampling, we go from left to right in this graph adding numbers and calculating the running mean with each additional data point. As you can see the mean bounces around at low sample sizes and then gradually settles down, converging on what we hope eventually is the true population mean.

The running standard deviation of the same sample shows a similar pattern -- fluctuating at first, and then settling down near what we hope is the true population standard deviation. Now some people look at the N in the denominator of the formula for standard deviation and therefore expect the standard deviation to decrease with sample size. But that does not happen because for each additional sample (each additional N) an additional deviation is being added in the numerator as well.

So, this far we have talked about how to describe our data in terms of different measures of central tendency and variation, but in the context of research design and analysis our goal is most often to be able to say something about the unknown population parameter from our sample estimate. Let's explore that next.

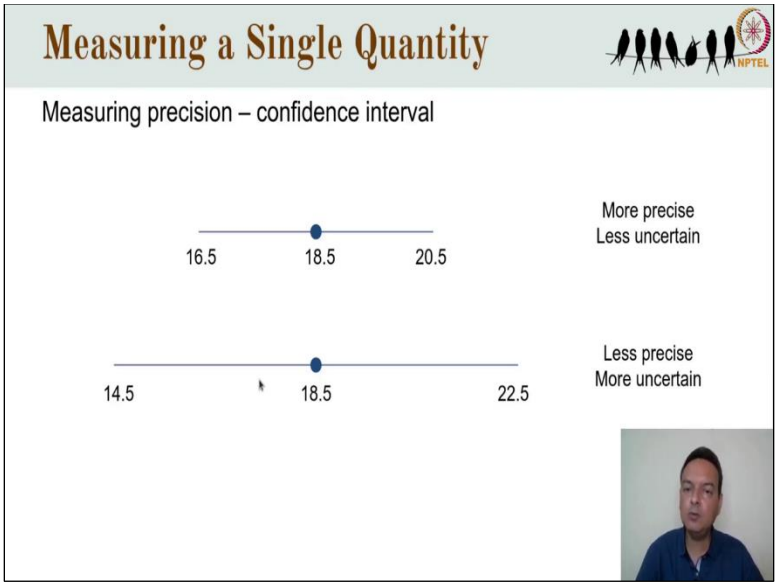
(Refer Slide Time: 36:43)



Say our research task is to find out the density of peafowl in Tadoba sanctuary or to measure the mass of 10 day old bustard chicks or to say something about the proportion of time that male koels spend singing in the breeding season. For all these we take some sort of sample -- whether of transects, or of broods, or of individual birds; and from the mean of these samples, we want to say something about the population.

In other words, we hope to get close to the population parameter – in this case the mean – through our sample estimate. So, we go out, following all principles of good study design as discussed in the previous video, and come up with some estimate of the population mean. Now how do we know the *precision* of this estimate or its inverse the *uncertainty* in the estimate. The precision or uncertainty is a crucial component of our result, so, that we can conclude something about the population parameter. Our best guess for the parameter is of course the sample mean, but the question is it a good guess -- a precise guess – or is it a poor guess -- an imprecise guess?

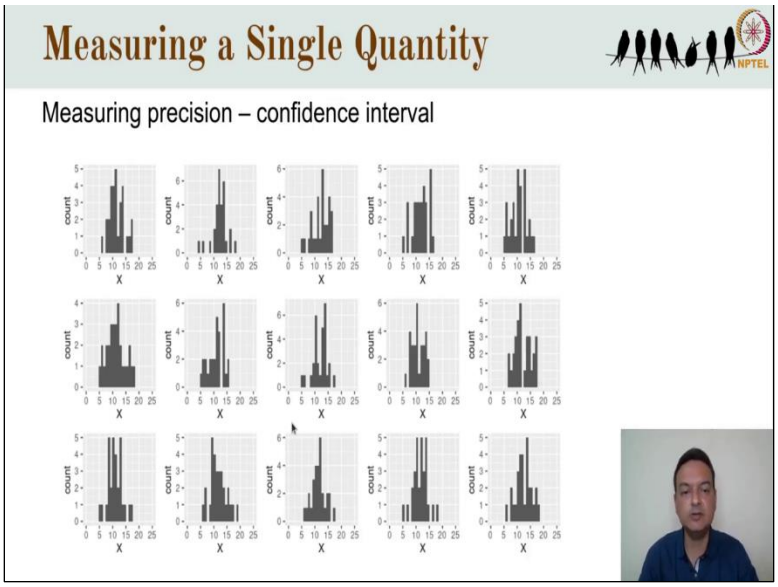
(Refer Slide Time: 37:45)



We do this by constructing what is called a confidence interval, a range within which the true population mean is likely to fall. So we might say that our best guess or estimate for the population mean body mass is 18.5 grams and we are fairly sure it is somewhere between 16.5 and 20.5 grams (if our guess is somewhat precise) or we might calculate that it is somewhere between 14.5 and 22.5 grams (if our guess is not so precise).

So, how do we calculate the confidence interval? How do we calculate the degree of precision in our guess?

(Refer Slide Time: 38:29)



We begin by recognizing that when we take a sample of, say, 30 data points from a population, this is only *one* of many *many* possible samples of 30 we could have got by chance. We just happened to choose these specific 30 sampling units from which we measured these 30 numbers and calculated a mean. If we had run our random number generator again and sampled another 30 we could have chosen a different set with potentially different measurements and a different mean.

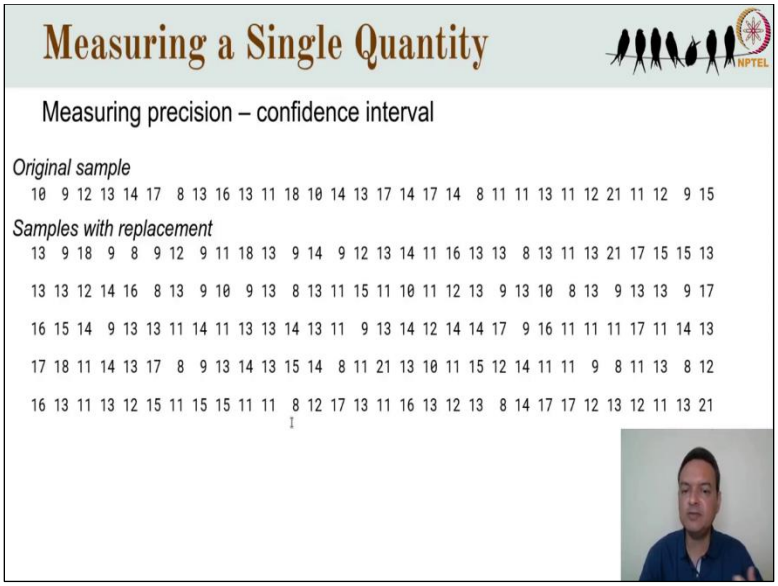
And here is another set of 30 numbers we could have sampled again, with a different mean, all differing by chance. So, the 30 numbers we happen to measure are one set out of potentially thousands or millions of possible sets we might have drawn by chance. One aspect of major interest is to ask ourselves, if I had drawn another set of 30 from the population or another or another how different would these have been from the actual set of 30 that I did get?

If all sets give very similar sample means that means that chance had a relatively small effect here and no matter what set of 30 I happen to get, the sample mean is likely to be close to the population mean. On the other hand, if the possible sampling sets are very different in their means then chance looks to have a very large role to play and some sample means could be quite far from the true population mean; and perhaps the sample mean that I got in my data is in fact quite far from the true population mean.

Now if only we could construct a frequency distribution of sample means from lots of possible sets that we could have drawn from the population. If we had such a frequency distribution, we could see whether it is a tight distribution with very little variation or a broad distribution with lots of variation. In the first case, we will have more confidence that the population mean is close to the sample mean that that we got

than the confidence we would have in the second case where it is a broad distribution of sample means. Now we will not be able to *actually* sample the population many times to create this frequency distribution, but we can do the next best thing.

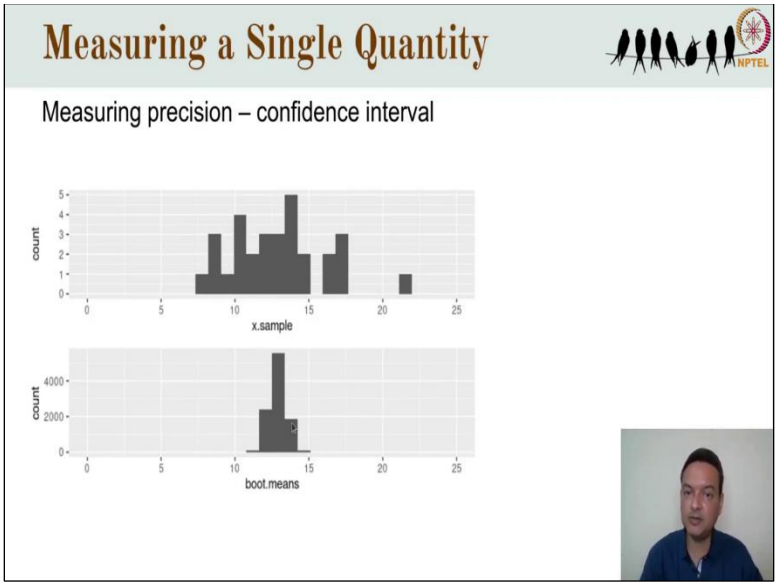
(Refer Slide Time: 42:12)



We take our sample data set as our best representation of what the population is like. It is really our *only* representation of what the population is like. So, imagining that these numbers look like the population as a whole, we convert this sample into an imaginary population by repeating the numbers many times. And then we draw a random set of numbers with the same sample size as the original from that imaginary population, and we calculate the mean of that sample. And we repeat this, drawing repeatedly a sample of size N (our original sample size) from the imaginary population, each time calculating the sample mean and writing it down. By the way, in practice, this is called *sampling with replacement.* It is the same if you write down your N numbers on small pieces of paper which you all put in a single basket, and then for the first sample you randomly choose one paper, note the number written down, put it back, choose another paper note, number down put it back and so on until you have N numbers.

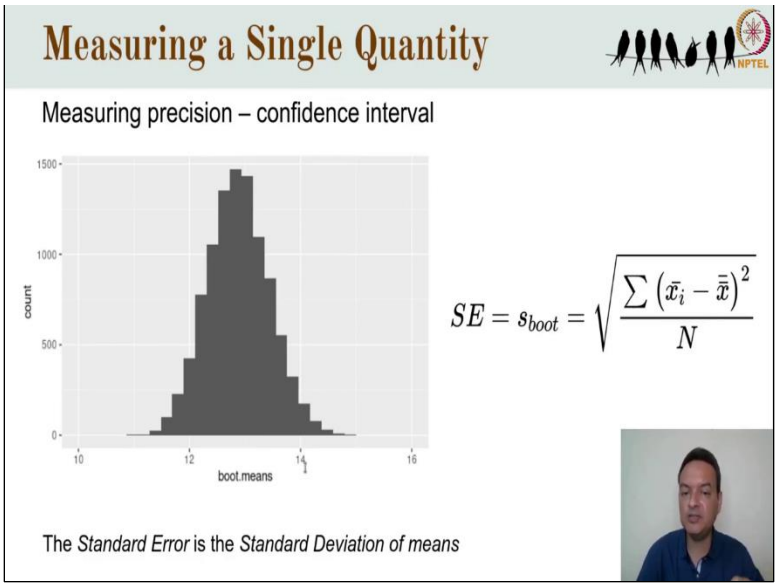
So, that is sampling with replacement using pieces of paper in a basket. This method of sampling repeatedly with replacement from a set of numbers is called *bootstrap* and you carry out this bootstrap procedure many times, each time calculating the mean of the bootstrap sample. Let's say you do this a thousand times -- then you can draw up a frequency distribution of the thousand bootstrap means.

(Refer Slide Time: 42:10)



When compared with the original distribution of the raw data, the distribution of bootstrap means is much narrower, which makes sense, since although the data might be widely distributed, the distribution of means will be narrower because the mean is, after all, a measure of central tendency. And as the sample size of the original sample increases the distribution of bootstrap means becomes narrower and narrower.

(Refer Slide Time: 42:38)



So, how does this distribution of bootstrap means help us in determining the precision of our original estimate -- that is how closely the sample mean is likely to be to the population mean? Well, it depends on the width of the distribution of bootstrap means. So, then how can we characterize the width of this distribution? We can do so exactly the way we did so before when we were talking about the distribution of the raw data rather than the distribution of bootstrap means.

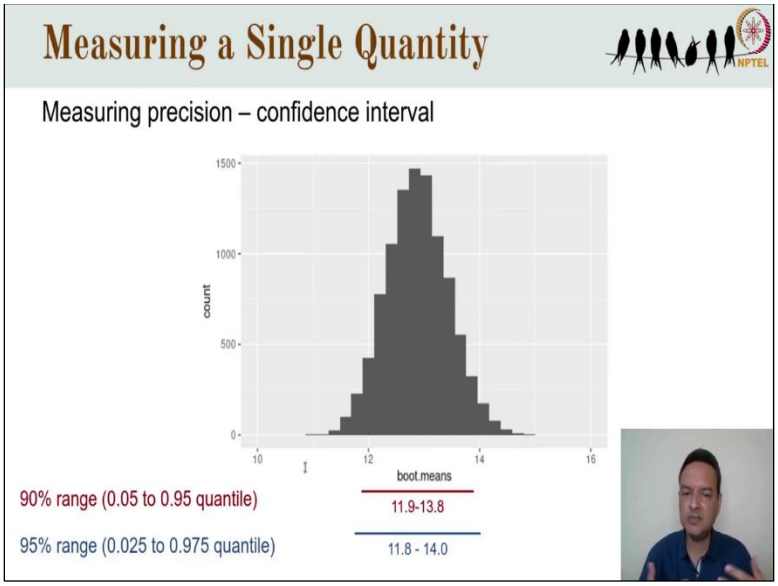
We can calculate the root average square deviation from the mean otherwise known as the standard deviation. In this case, the overall mean is the mean of bootstrap means and we are looking at the deviation of each bootstrap mean from the mean of bootstrap means. Remember that the bootstrap means represent different possible outcomes we could have got if we had sampled another random set of data. And the bootstrap frequency distribution represents the range of those possible outcomes.

Now I apologize but it is time to introduce another bit of jargon. While the standard deviation of a set of numbers is simply called the standard deviation, the standard deviation of the means we spoke about is called the standard error. Do not get thrown by the use of the word error here. In this statistical context, the word error means deviation or difference it does *not* mean mistake.

So, how do we interpret the standard error? The standard error is a measure of the *precision* in your original sample estimate of the unknown population mean. A large standard error tells you that the distribution of bootstrap means is large or broad and therefore your precision is low -- the true mean could be anywhere within a large range of values. By contrast a small standard error tells you that the distribution of bootstrap means is small or tight and therefore your precision is high – the true mean is likely to lie between a small range of values.

But say that is not good enough. We want to be able to describe an actual range of values between which we think – based on the data – that the true population mean lies. For example, we want to say that our best guess is the population mean lies between 12 and 14 or 11 and 15. How can we get those numbers?

(Refer Slide Time: 44:59)

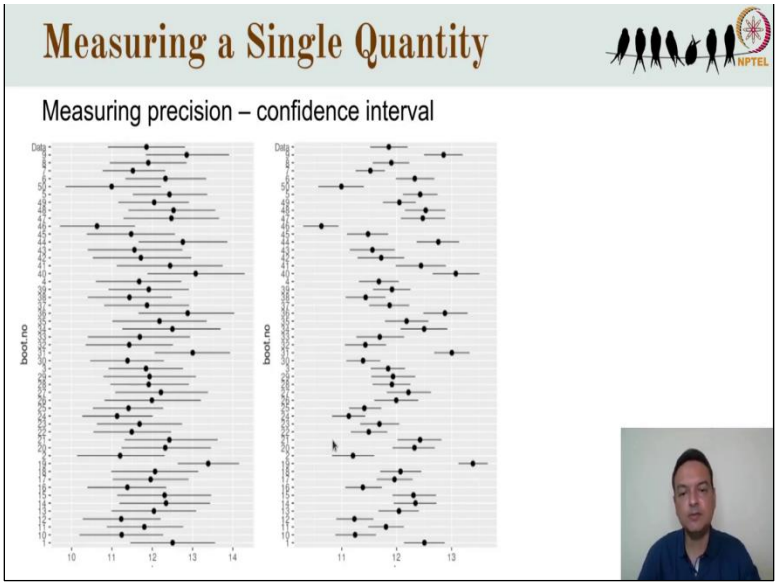


Well, one way to do this is to go back to the distribution of bootstrap means and now we take advantage of what we have learned about percentiles and quantiles. We could ask what value corresponds to the 5th percentile or the .05 quantile and what value corresponds to the 95th percentile or the 0.95 quantile. If you are following carefully, you will realize that 90% of the distribution lies between between the 5th and the 95th percentile. If you have not understood this it might be good to go back and review the section on quantiles.

So, we call these two values the confidence limits and the range between them the confidence interval what we have done now is to identify the 90% confidence interval within which 90% of the bootstrap means lie. More conventionally, you look up the 2.5th percentile and the 97.5th percentile and through that, identify the 95% confidence interval, within which 95% of the bootstrap means lie.

Some people advocate using the 50% confidence interval instead and of course the nature of frequency distributions and quantiles is such that the 50% confidence interval is narrower than the 90% confidence interval which in turn is a bit narrower than the 95% confidence interval.

(Refer Slide Time: 46:33)



Now how does the confidence interval help us and what does it mean? The confidence interval represents a range of values in which the unknown population mean is likely to fall. Formally speaking, if we collect a sample of data and construct a 95% confidence interval as just described and do this repeatedly, each time constructing a 95% confidence interval, then the true population mean will fall within those intervals in 95% of cases, and will fall outside those intervals in 5% of cases. In the simulation shown here, the true population mean is 12 and you can see that in 3 out of 50 confidence intervals (marked with red arrows) the mean (which is 12) is not included in those three confidence intervals. So, in 47 out of 50 confidence intervals (or 94% of cases) the confidence intervals contain the true mean. Similarly, if we construct a 50% confidence interval for a sample of data and go out and do this again and again each time constructing a 50% confidence interval

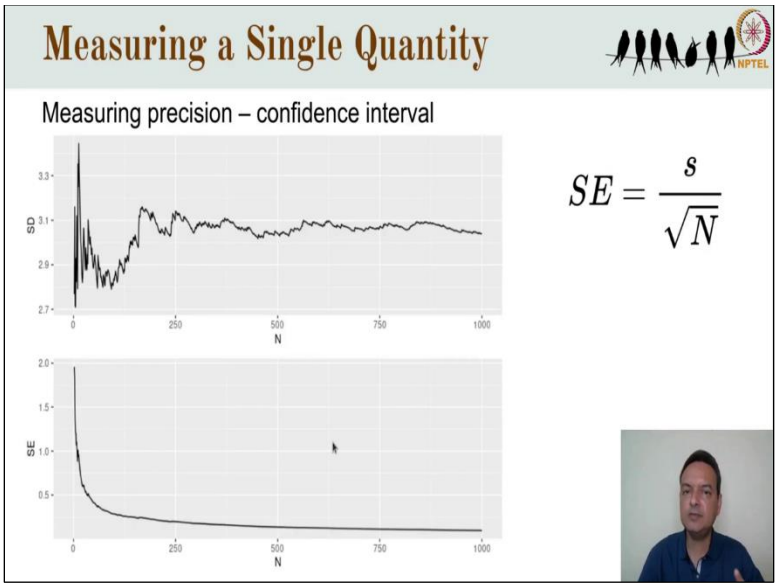
then the true population mean will fall within those intervals in 50% of cases and will fall outside those intervals in 50% of cases.

So, to summarize, your 95% confidence interval will contain the true mean in 95% of cases. Your 50% confidence interval will contain the true mean in 50% of cases and so on. And this is of course assuming that your sample is unbiased and randomly chosen from the population and all that good stuff. Statistical analysis can't make up for flaws in study design, unfortunately.

So, now to recap we have calculated a measure of central tendency – say the mean – and we have a measure of precision – the confidence interval. The narrower the confidence interval, the greater the precision. That is all we need in order to say something about our measurement of a single quantity. But here is a question: when constructing confidence intervals is there a shortcut we can take rather than going through all this fuss about sampling and resampling and calculating all those bootstrap means and so on. Well, yes, it turns out that there is a shortcut.

And in fact, it is something that was used almost exclusively until personal computers became widely available and is still often used today.

(Refer Slide Time: 49:04)



It turns out that under some circumstances, we do not need to do the bootstrap process to get a distribution of bootstrap means and then take the standard deviation of bootstrap means in order

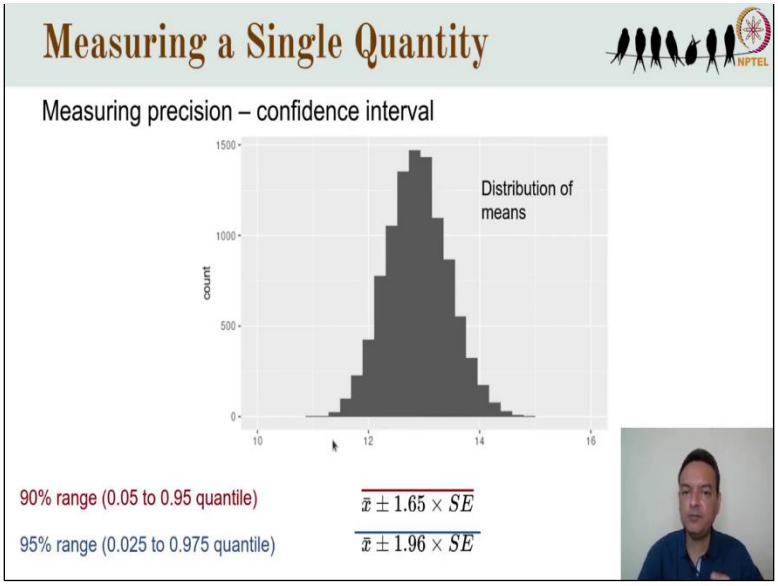
to calculate the standard error. Instead, we can use this simple formula to calculate the standard error (SE)

$$SE = \frac{S}{\sqrt{N}}$$

Note that because of the extra N in the denominator, the standard error does get smaller as N gets larger, unlike the standard deviation which may fluctuate but does not decrease as more samples are added.

Now, we have the standard error. How do we get to the quantiles?

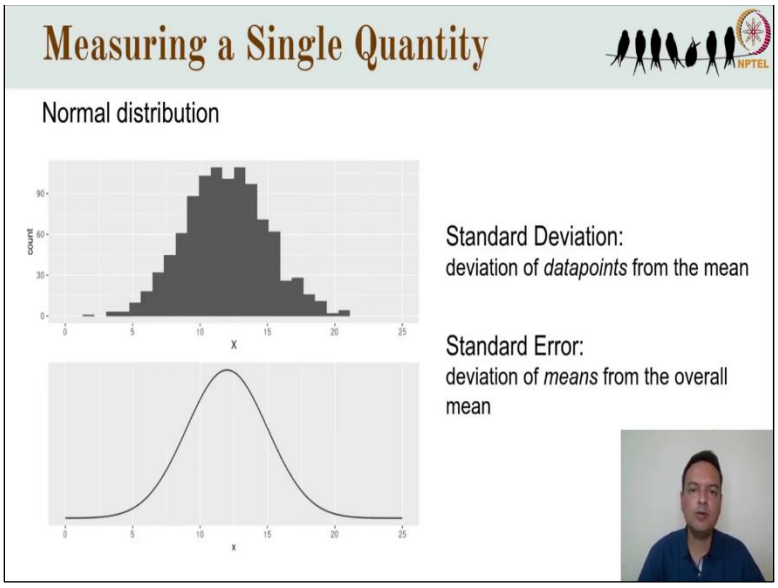
(Refer Slide Time: 49:41)



It turns out that under some specific conditions, the standard error has magical and very useful properties. For example, in a particular kind of frequency distribution of means the standard error has the property that if you take the mean of the means and look at the range encompassed by the mean + or - 1.65 times the standard error then that range covers 90% of the distribution or if you look at the range encompassed by the mean ±1.96 times the standard error then that range covers 95% of the distribution.

Now where does 1.65 and 1.96 come from; where do these multipliers come from? They come from the properties of a particular, specific, distribution called the *normal distribution* and it is only for a normal distribution that these properties hold.

(Refer Slide Time: 50:36)

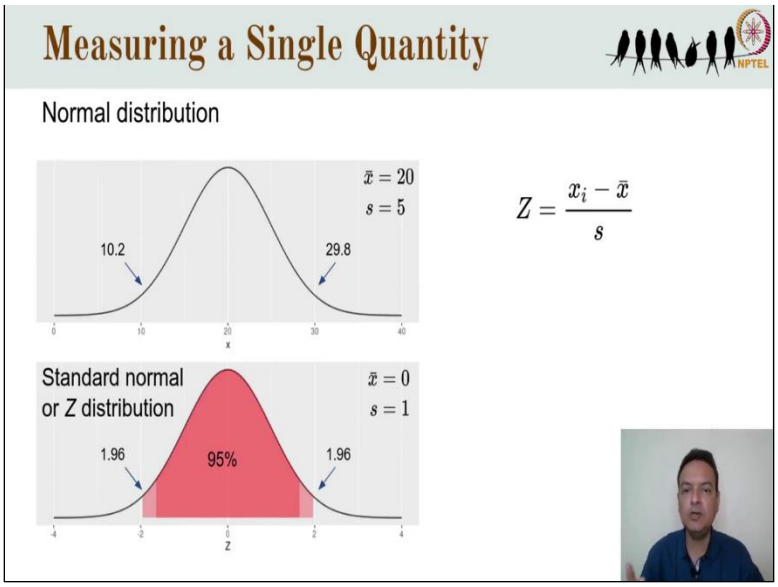


The normal distribution is a symmetrical distribution characterized by two quantities - the mean in the middle, and the standard deviation. What you see here is a histogram but one can also draw a distribution as a line rather than in steps. Roughly speaking, the height of the line for any x value represents how common that value is in the distribution. For a normal distribution, the mean lies at the middle, which is also the highest point on the curve. Changing the mean shifts the curve left or right.

The standard deviation specifies how wide the curve is. The smaller the standard deviation the narrower the curve with extreme observations being less frequent. And the larger the standard deviation, the broader the distribution, and extreme observations are more frequent. You will need to be alert if you do not want to get confused. Remember that we use the term standard deviation to refer to the root average square deviation of raw data and standard error to refer to the root average square deviation of means.

But the term standard deviation is a generic term for any distribution and you will have to be alert to know when it is the variation in raw numbers being talked about, or when it is the variation in means being talked about.

(Refer Slide Time: 52:00)



So, for any particular value of mean and standard deviation, the normal distribution has a fixed shape. Here is a normal distribution with a mean of 20 and a standard deviation of 5. With this information we can find out what proportion of the distribution is covered within any particular range of x. In fact, it is most convenient to measure the distance from the mean in terms of not x but standard deviations of x.

So, we can do a small transformation, subtract each x (each data point) from the mean ($\bar{x}$) and divide by the standard deviation (s).

$$Z = \frac{x_i - \bar{x}}{s}$$

The resulting numbers have a mean of 0 (because now positive and negative x is balanced out) and a standard deviation of 1. This new distribution is called the z-distribution or standard normal distribution and we have converted the original data into what is called z-scores. The z-scores are the deviations of the data point from the mean, but measured in units of standard deviation.

So, in the original normal distribution, if a data point has a value of 10 and we know that the standard deviation is 5, that means that 10 lies two standard deviations below the mean and therefore it has a z score of -2 and is placed at -2 in the z-distribution. Conversely, if the z-score is 1 then we know that the data point lies one standard deviation above the mean and we can calculate

(knowing the mean and standard deviation) that this is at a value of 25 on the original scale of the data. So, you can convert between the original data and the z scores in either direction.

Finally, from the properties of the z distribution we know that the rang -1.65 to +1.65 covers the central 90% of the distribution. This corresponds to ± 1.65 standard deviations from the mean in the original data, which is ± 1.65 times 5 which is ±8.25. So, in the original distribution 90% of the data lie between 11.75 and 28.25.

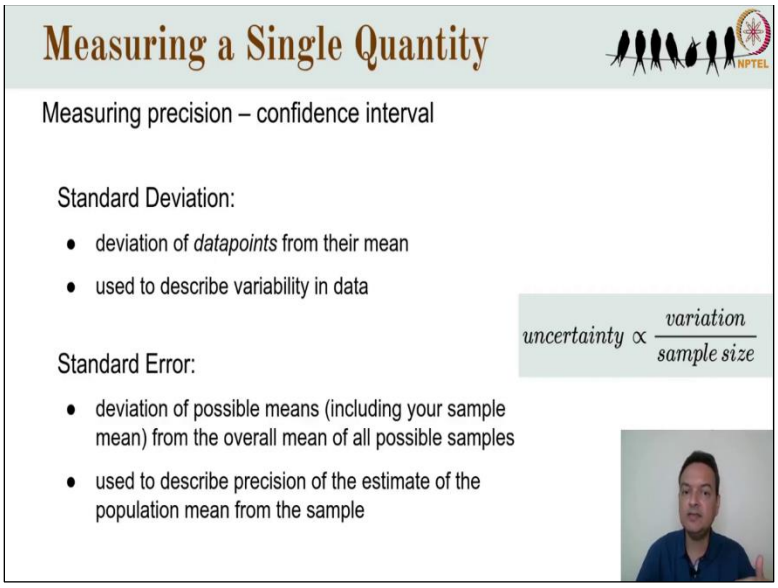
Similarly, in the z distribution, the range -1.96 to +1.96 covers the central 95% of the distribution which corresponds to 10.2 to 29.8 in the original data. Please work it out for yourself and see if you get the same numbers that I did.

(Refer Slide Time: 54:42)

So, now we can take full advantage of the normal and the z distributions. If we have our sample mean and sample standard deviation, we can calculate the standard error and then we know from the properties of these distributions that the mean ±1.96 times the standard error contains 95% of the distribution of possible means we could have got had we sampled repeatedly. And so, the 95% confidence interval is mean ±1.96 times the standard error.

And if we wanted a different coverage say of 90% or 50% for the confidence intervals, we would just need to look up a z-table to see what the multiplying factor would be.

(Refer Slide Time: 55:21)

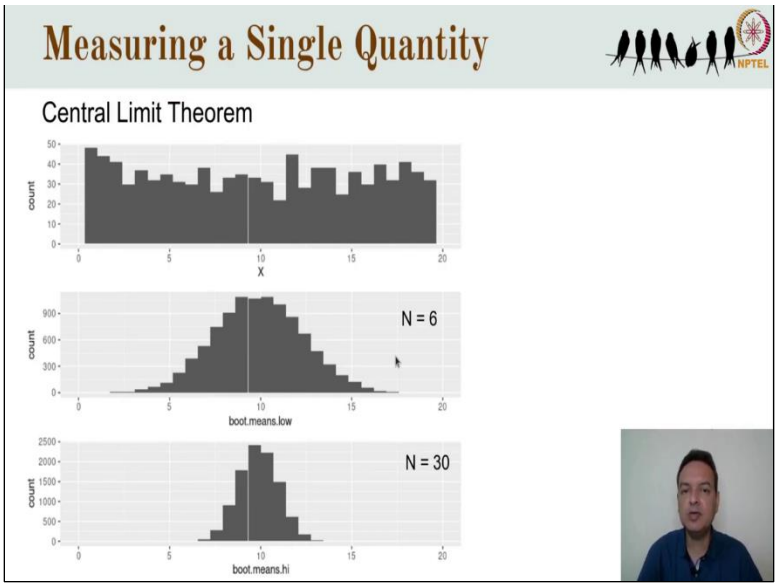


Please carefully understand once again the difference between standard deviation and standard error. Standard deviation describes the variation in the actual data points. The mean ±1.96 times the standard deviation would encompass 95% of data points if the data follows a normal distribution. On the other hand, the mean ±1.96 standard error would encompass 95% of the imaginary means from imaginary multiple sets of samples.

In other words, standard deviation describes variation in the data and standard error describes variation in the means. You use standard deviation when you want to describe how variable your data are. You use standard error when you want to describe how precise (or the converse, how uncertain) your sample estimate of the population mean is.

Now, as discussed in the earlier lecture on research design, the uncertainty of your sampling estimate is directly proportional to the variation in the population and inversely proportional to the sample size. More variation in the population leads to more variation in your sample which is reflected in the standard deviation of the sample. Higher sample size brings the means of multiple separate samples close together reducing the standard error and this is also reflected in N being in the denominator in the formula for standard error.

(Refer Slide Time: 56:48)



Now, I said earlier that the standard error has these magical properties under *certain conditions*. The specific condition is that the distribution of sampling means (of which our observed mean is one) follows the *normal distribution*. If it does not follow the normal distribution then we cannot use the multipliers 1.96 and 1.65 and so on and life gets more complicated.

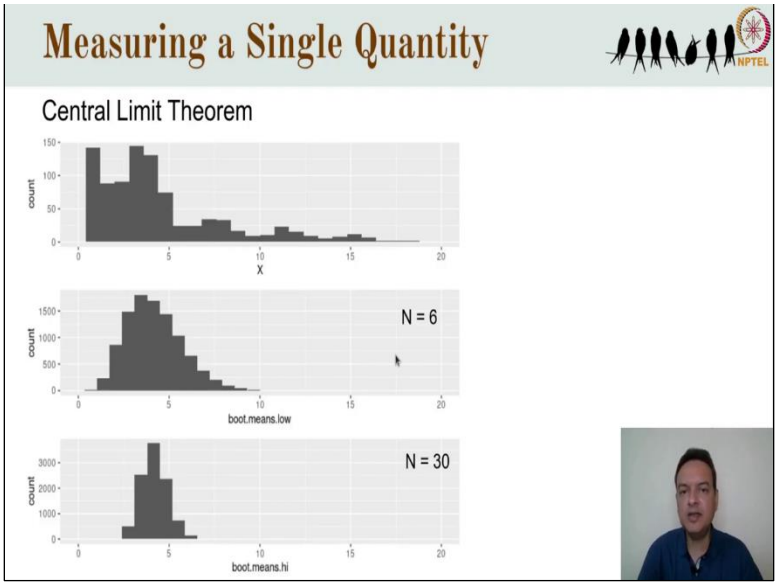
Now, we often see that people just assume that the conditions of a normal distribution apply. So, the question is: how reasonable is this assumption? Well, it turns out that under a surprisingly broad range of conditions it is a fairly reasonable assumption. We find that even if the underlying data look very different from a normal distribution, the distribution of means from multiple sets of samples from the same data looks quite symmetrical and normal. In fact, this observation is codified in what is called the central limit theorem which says that the distribution of sample means approximates a normal distribution as sample size gets larger,, and especially if the underlying distribution is not too asymmetrical.

Here are some examples. Here we have a distribution of the underlying data that is very different from normal – something like a uniform distribution – and the most common values are not even at the center. But when you take repeated samples from these exact data and plot the distribution of means, you see that the distribution of means is quite symmetrical and almost like a normal distribution.

The most common value of the mean is towards the center, and that is because the act of taking the mean tends to shrink values from the extremes of the distribution towards the center. It is worth pondering about this a little bit to see if this makes sense to you. Now, this distribution is what you get when you take samples of size six at a time. So, sample size is six; and repeat this over and over calculating the mean each time.

If you were to take larger samples -- of size 30 – instead, then the distribution of sample means would look even more like a normal distribution and would also be narrower. I hope by now you understand why.

(Refer Slide Time: 58:59)



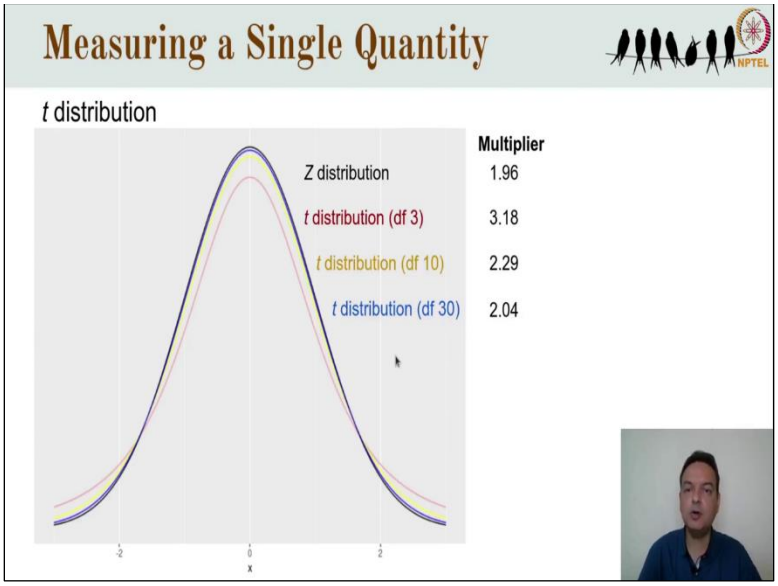
If we started with a very different shaped distribution of the data, this one is skewed towards the right with a small number of extremely large values. Then, if we drew samples of size 6, the distribution of resultant means looks a bit more symmetrical but it is still skewed somewhat to the right. So, not quite symmetrical but if we took samples of size 30 each then the distribution of resultant means is very close to normal.

So, these two examples demonstrate the central limit theorem, which holds that the distribution of means is approximately normal, especially when the underlying distribution is not too skewed and when the sample size is large. So, in many cases when we are dealing with numeric data, it can be

reasonable to assume that the distribution of possible sampling means is normal and we can then use the properties of the normal distribution to calculate the precision of our estimates.

But beware of variables that have lower or upper limits like 0 or 1 for these the assumption may not hold particularly at small sample size.

(Refer Slide Time: 1:00:11)

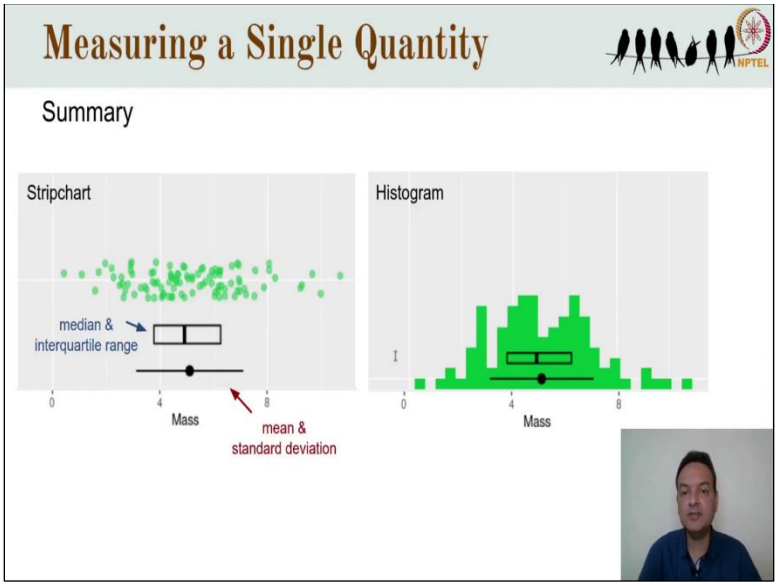


And one further point if your sample size is less than 30, the distribution of means is likely to be a bit different from a standard normal distribution. In this situation the distribution is likely to follow what's called a t distribution which is like a normal distribution but with a higher frequency of extreme values. In other words, with greater area under the two tails – the extreme values to the left and right of the mean.

So, when the sample size is less than 30, one must use a multiplier from the t distribution, not from the standard normal distribution. You know that for 95% coverage in the normal distribution, you have to look at ±1.96 standard deviations from the mean; and this multiplier does not change with sample size. By contrast, the t distribution changes shape as sample size increases and so the appropriate multiplier changes as well as the sample size gets larger and larger the t distribution shrinks towards the normal distribution.

For example, for a sample size of 4 (usually corresponding to a degrees of freedom 3) the multiplier is as high as 3.18 but as the sample size increases the t distribution shrinks towards the normal distribution and our multiplier gets closer and closer to 1.96. So, because the t distribution reduces to the normal distribution at large sample sizes. In practice most statistical software programs always use the t distribution to calculate the multiplier, since it takes care of matters regardless of sample size.

(Refer Slide Time: 1:01:48)



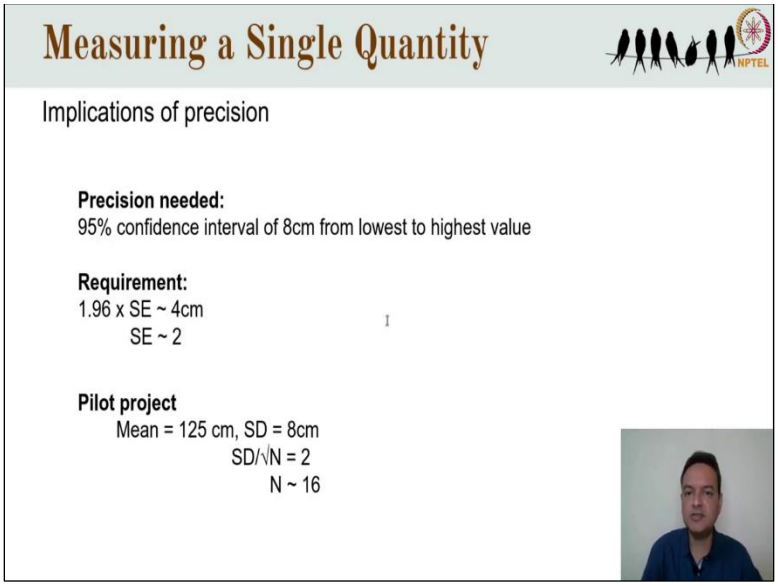
So, that brings us to the end of our description of what is involved in measuring a single quantity, we are typically interested in some measure of central tendency -- most often the mean – but given the mean's lack of robustness, perhaps we should use the median more often.

Also, if you want to understand the variation in our sample then we can visualize it using a strip chart or a histogram. Note in the strip chart, the points are jittered, that is, each point is moved a small distance at random along the vertical axis, so that points with exactly the same values are visible and do not obscure each other. We can summarize the variation by calculating the interquartile range or the standard deviation or for that matter the 2.5th and 97.5th percentiles and we can overly overlay these on the graph. Now, this is for describing the variation in the population in the sample. If on the other hand, we are trying to make inferences about the parameters like the population mean we also need a measure of precision of the estimate from our sample, And we would calculate something like the 95% confidence interval around the sampling mean using a

bootstrap method. Or if it is justified, we can take the shortcut of calculating the standard error and use the properties of the normal distribution to find out the limits of the 95% confidence interval. And if this is a primary interest to us then we would show the confidence interval in the strip chart or histogram rather than the standard deviation.

When you are making graphs like this please ensure that you label your graphs carefully to tell the viewer what measure of central tendency you are presenting – whether it is mean or median – and which measure of variation you are showing -- the standard deviation or interquartile range, for example. Or if you are presenting precision then are you showing the standard error or the conference interval and if you're showing the confidence interval then what percent cover are you showing -- a 90% confidence interval or 95% confidence interval and so on.

(Refer Slide Time: 1:03:54)



While estimating a population parameter like the mean, remember also that if the confidence interval is very large then we may actually have learned nothing much from our study. So, it is useful to work backwards in specifying what degree of precision you are aiming for, and *then* calculating what sample size you might need given the variation in the population. For example, let's say I want to estimate mean Peafowl train length with such precision that I get a 95% confidence interval that spans a range of only four centimeters.

So, what we need is for 1.96 times the standard error of our study to be around 2 centimeters since we calculate the mean plus as well as minus that number and that will give us an overall confidence range of 4 centimeters. So, we require a standard error of roughly one. Now, if I do a small pilot project measuring the train lengths of 10 peafowl and the mean of these 10 is I find to be 125 centimeters with a standard deviation of 8.

Then, assuming that this sample standard deviation of 8 is somewhat close to the population standard deviation, I know that the standard deviation (s) divided by the square root of N will be my standard error

$$SE = \frac{s}{\sqrt{N}}$$

which we know needs to be around 1. So, therefore the square root of N needs to be around 8, which means that the sample size in my final data collection should be roughly 64 individual peacocks. By contrast, if I were fine with a 95% confidence interval spanning a range of 8 centimeters then my sample size would only need to be around 16.

So, now you know the answer to that eternal question – what sample size should I have? The answer is that it *depends*!! Specifically, it depends on the variation in your population and it depends on how much precision you need in order to address the research question you are tackling. The first you can measure in a pilot study or possibly by getting a clue from the literature. The second is a decision you have to make – often a subjective decision, and a decision you still have to make even if you are uncomfortable with the subjectiveness of it.