LLO 8200: Summary Statistics

Youtube Generated subtitles

in this video I am going to review what are ultimately going to be the building blocks of every statistical test that we use in data analysis these concepts of central tendency and spread so Central tendencies we call them Central Tendencies because the data tends to center around them okay we have three primary flavors of central tendency the mean the median and the mean the median the mode and the mean sorry have them in a different order okay the score the median is going to be the score that falls in the middle of an ordered list the mode being the most frequent score and the mean being the arithmetic average of the score okay so let's go ahead and unpack each one here so the median is the score that is in the center of an ordered list of all possible values okay the median is such that it splits the upper and the lower half of scores it bifurcates the distribution okay that just means splits in half

it's a very easy way of finding the median okay you can do n plus 1 the quantity n plus one divided by two and that'll tell you what value you need so in the case of n being and the number of observations you have being an odd number the median is going to Simply Be The N plus 1 over 2 term so in this case of 5 in this case of five numbers ten eleven twelve thirteen and fourteen the N is equal to 5. so 5 plus 1 is 6 divided by 2 means the third term so the third term in here 12 splits the distribution in half 10 and 11 being on one side 13 and 14 being on the other okay if n is even then the median is between the two middle scores okay it'll still end up being the N plus one over two term but now when we calculate it it'll be a sort of a half

so we'll so in the case of 10 11 13 and 14 n is now equal to four four plus one is five divided by two is two and a half so the two and a half term will be between the second and third terms so 1 2 11 3 13. so we take 11 plus 13 and we divide by 2 and we get 12 again okay

so the median can be calculated the same way regardless just know that if it ends up being a half number uh that means that we will be taking the average of the two middle numbers so in this case the two and the three okay

the mode is just the most frequent value that exists in a distribution I don't want to go too much into this the mode when we look at the mode in practice it's more going to be just looking to see where the peak of a frequency distribution is or a histogram is okay if you need to find the mode for whatever reason you literally just count all the values you take the one that has the greatest number that's the mode boom okay if there's one we call it unimodal if there's two modes we call it bimodal we'll also call it bimodal if there's close to two modes they don't necessarily have to be the most frequent in the absolute sense we just kind of see like a camel hump we would call it bimodal if there's three it's trimodal and so on and so forth okay but by far the most frequently used um measure of central tendency that we will see and we will use is going to be this mean okay the mean is just the arithmetic average it's the thing you learn in grade school right you add up all the values and you divide by the number of values that you have okay we represented a bunch of different ways it can be represented as X bar with an X over with a line over it also called X bar you can represent it with an M you can represent it with a mu and again the way that we calculate it we just add up all scores and then we divide by the number

of scores that we have so something that we need to remember about the mean is that it's going to be heavily influenced by what we call outliers outliers are those extreme values within a distribution okay the median is the point that balances out the number of observations right it says I have half of the I have half n divided by two of the scores on one side I have n divided by two of the scores on the other side the mean actually sort of Weights the value of the scores okay so having an outlier that's on the Very extreme end what it does is it ends up pulling the mean in One Direction okay um so because the value of that outlier is so much more so we need to be very careful about outliers and being able to identify them appropriately for example when you look at something like mean income mean income will typically be greater than the median income and that has to do with the fact that there are a few people that just make a whole lot of money right substantially more money than what we would typically consider the central or the average person okay so while the median again so while the median balance is the frequency or the counts within the distribution the mean is going to balance out those values so those extreme values have more influence all right something else that I want to note here is that when we have two or more groups and we want to know the average of those groups combined we can do something that's called a weighted average to do that we find the mean of each group we multiply by the group size and then we will sum up uh we'll sum up those products and then divide by the total number of individuals in all groups right we're just rearranging the traditional calculation for the um the traditional calculation for the the mean calculation and we are now Distributing it into like this is one group this is another group okay we can weight them accordingly that's going to allow us a little bit of flexibility when it comes to uh to doing certain things later down the road so if we look at this histogram that I made of IQ scores okay we can calculate the mean of the median you can see them right at their own up on top right median of 99.6 and a mean of 99.8

you can see that in this case the mean and the median are very close right they're within 0.2 units of each other

the mode is somewhere in there as well what this what these numbers aren't able to capture though is not much Beyond where the data happens to be centered around it would be nice if we could say something about how the data are spread okay say something about how far the data points are around this measure of central tendency how much they spread or disperse from The Middle

now we can do this easily by just looking at the minimum and the max scores right that'll give us something that we will call a range but we can do better than that so let's go ahead take this idea of spread and start to explore it a little bit more okay now spread the word that I use spread it's kind of vague um but it's the one that I'm going to use to talk about a bunch of different ways that we measure how variables are distributed across a distribution okay so we have the range that's going to be that Max score by the minimum score that's going to tell you how many possible uh it's going to tell you how many possible values that you can have okay and then we have this thing called quartiles which is breaking up the distribution into 25 units okay so the quartiles will denote the 25th percentile the 50th percentile or the median the 75th percentile and the 100th percentile the maximum okay this is where the use of something called a box plot is particularly helpful if we look at a box plot okay and we uh

we sort of unpack it okay the center line here will Express the median the 50 the bifurcation point the top of the box will represent the upper quartile that is to Say the 25 of data uh between here and here so this is the 75th percent as 75th percentile the lower quartile which is going to be the 25th percentile then we have the minimum and the maximum where we exclude outliers outliers being defined as more than uh more than one and a half times that upper quartile okay so even though the absolute maximum is actually this outlier here okay the maximum for the box plot is actually going to be substance is going to be below that it's going to be this one and a half times uh threshold okay

or at least the last value within that one and a half times threshold so here's an example where we look at that distribution of IQ scores and I've given it an outlier of 160. you see that the vast majority of the data are going to be in this Center chunk here but we have this one outlier this is one value that's more than three and a half times this value sort of floating off here in la la land we can also think of instead of the range and the quartiles we can think about the deviations okay or how much an individual score deviates from the mean okay we express that as the x minus M so the value of the point minus the mean in this in this example here we have uh participant one um has a score of two the mean is four so the deviation score is minus two now that's cool because what we can do is we can look at these individual scores and we can say okay this individual is two points below the average this person is one point below the average this person's at the average this person's one above this person's two above okay so these deviation scores are super helpful when we want to talk about where individuals are relative to the average but if I want to say something about how the entire group of participants are relative to the average talk about the typical spread um it's not helpful because when we add up these deviations remember what I said the mean is balancing the values so that means if one participant is a couple value points below there's another participant that's a couple value points above to balance it out so when we sum up these raw deviations we're always going to get an answer of 0. so that's not going to be super helpful for us trying to express what's going on at a group level so we want to try to create some kind of mean deviation so let's think about the deviation the mean deviation we know that they're equal to zero one thing that we can do is the problem here is that these deviations have a sign so let's not look at deviations instead let's look at the squared deviations right because if we square a number we'll get rid of the sign so let's do that let's go ahead we'll square those deviations and then we'll get rid of the sign for us

now when we take all of those squared deviations and we add them up and then divide by anything the number is not going to be equal to zero there's one exception but we already know that when we have that exception we don't have a variable right the only way that this value will be equal to zero is if the sum is equal to zero the only way that that is the case is if all the values are exactly equal to the mean right so the mean squared deviation will not equal zero so we have a measure of spread here by taking the squares of these raw deviations adding them all up and then taking the average of them

okay dividing by the number that we have but this is in squared units right we've taken whatever unit we have here we've squared it and then divided by a constant so we still have

squared units that's kind of like talking about apples and apples squared so we can just take the square root of it if we take the square root of this squared deviation the sum of squared deviations that we calculated we end up with this thing that we call the standard deviation we call it standard deviation because saying the mean deviation is not accurate but it's kind of what it is right this is like on average this is how much the scores are spread around our measure of central tendency okay it is this distance squared divided by the number of observations and then we take the square root of that so it's actually the area the average area and the square root of that is the average distance okay so this is the average distance of a score from the mean we express it with a sigma for the population and we express it with an S for the sample okay we will largely be dealing with samples in this class so we are going to be dividing that by n minus 1 and not by n Okay the reason that we divide by n minus 1 is for inferential reasons we are estimating a parameter of the population okay so just standard deviation boom now

that thing that I called the mean squared deviation that also has a name which is the variance okay I'm expressing the population as Sigma squared and where the sample as s squared and you see that it's the same as the standard deviation we just now haven't taken the square root

okay so when we talk about the variance what we're talking about is we're talking about area okay because a distance squared is an area okay now the numerator of this term is known as the sums of squares because it's the sum of the squared deviations okay so another way for us to write this is for us to write it as the sum of squares divided by n minus 1.

all we are doing here is we are coming up with an average or typical measure of spread of distance from the mean that's our standard deviation now there are going to be times where we play around with variance instead where we're actually looking for sort of the overlap in area and not this average distance so we do need to have an understanding of what the relationship between standard deviation and variance are we also need to know that the relationship between variance and standard deviation is just or sorry between variance and sums of squares is that the sums of squares it's just been divided by the number of observations okay so these are proportional to each other

there are a couple other measures of spread that we use skew and kurtosis skew is going to be the big one for us when we are looking at distributions and looking at whether or not they are appropriate and this has to do with issues of uh normality assumption okay the three types of skew are going to be no skew that is when the mean and the median are basically in the same spot the distribution I showed you before with IQ scores there was basically no skew okay you could argue there was a slight skew but for all intents and purposes they were the same okay you have a positive skew when your mean is to the right of the median okay because the positive end of the distribution is being pulled in One Way by an outlier okay and the negative will be when the mean is to the left of the median and that is because the negative end of the distribution the Left End is being pulled is being pulled by an outlier okay so when the negative side of the distribution is elongated it is a negative skew when the positive end is elongated it is a positive skew okay our statistics are able to handle a small amount of skew but when we have distributions that are heavily

skewed we're going to have to rely on things like Transformations or perhaps even the median instead of the mean in order to make accommodations okay the other type of spread that is typically covered is this idea of kurtosis you have no kurtosis or mesokurtic and then you have high positive kurtosis which is leptokurtic because it is tall and thin it done leapt up and then you have platycurtic which is the short and flat or negative kurtosis it's because it's shaped like a plateau or a platypus bill okay these are things that are just calculated they're not a thing that I'm going to ask you to calculate but you do need to understand that for the most part we're not going to be concerned with kurtosis okay particularly leptokurtic distributions we're a little bit more concerned about platy critic distributions because when they are exceptionally flat and short the distribution is no longer normal and it actually comes closer to what we call a uniform distribution so we do need to kind of look at it but we're a little bit softer about this criteria than we are some of the others okay

so the biggest take-homes from this for us okay you need to be able to identify where your data is clustered that is to say the central tendency okay you need to understand how it varies throughout the distribution this idea of variance and standard deviation and skew and as we continue on we will bring up some issues that relate to things like spread and central tendency like for example there's a thing called restriction of range okay or we will discuss something called partitioning of variance okay you need to understand what these terms mean so that we can sort of do a thorough job unpacking those issues and those Concepts a little bit later down the road okay it'll make the experience a lot easier if you sort of know these um so just kind of spend a few minutes reviewing them but again it shouldn't be anything too terribly surprising for anyone who has had uh applied stats