Robert_S2_L01

Wed, 1/12 12:24PM 13:04

SUMMARY KEYWORDS

values, average, module, stem, observation, number, players, parking, series, histogram, middle, continuous variable, bar graph, represent, toronto maple leafs, descriptive statistics, equal, seesaw, learned, sum

SPEAKERS

Robert McKeown

R Robert McKeown 00:04

Hello, and welcome to our next module in statistics. This one is focused on descriptive statistics. So in the previous module, you learned about how to visualize data, you learned how to create a bar graph with categorical data, you'll learn to how to create a histogram using continuous variables, we're going to pick up and we're going to use what you learned in the previous module and this one, our goal is to understand descriptive statistics. Descriptive Statistics is an attempt to summarize a series of values or a distribution. To summarize it with just one value. And the first concept of summarizing a series of values a series of numbers, sometimes called a distribution is to look at this concept of centerness. And that's what we're going to be doing in this module, we're going to look at centerness, which can be summarized with that concept of average that you're almost certainly familiar with. So I hope your mind's on, you've got a pen piece of paper, you're ready to make some notes as we go. And let's get to it. Now, when we're looking for the middle of a series of values, what we really want to find to represent that middle is an average. And it's a single value that best represents the center of a series of values. Now, the series of values can represent many, many different things, it could represent the players on your favorite basketball team, and each of their shooting percentages, it could be the list of different crimes, and the probability of someone convicted of that crime re offending. The values could be many, many different things. And so you might have a series could be seven, 8, 21, 22, and 23. So there is a series values. Now, how you define the middle depends really on what you're interested in, you might say that the middle is just the middle value might be this 21 right here. Or, you know, none of these numbers are repeating, but maybe, maybe now you want to say, well, the middle is actually out here, because 23 occurs more than once. And it's it's the most likely observation, if each of these is equally weighted. So there are different ways that of find the middle. And we're going to go about doing that right now. You can think of the middle, or an average as being the fulcrum of a seesaw. The point at which a seesaw is perfectly level. Hopefully with with the ground of the ground is also level. But you notice that we have equal weight here. And here. And that's the kind of that's what we're kind of looking at when we're looking at an average. We're gonna look at three different measures of the average. Now formally, averages are known as measures of central tendency. So how does the how do the values in a series? Where do they tend to center and we've got three methods that we're going to discuss. We've got the mean, the median, and the mode.

And they're going to be different variations of looking at the center and thinking about this see saw that we have over here. Our first measure is the mean. And you may be familiar with this. Hopefully, you are but maybe you're not. And we're going to look at it quite formally and look at some of its features as well. So x bar here, this is the mean. And the mean is equal to the sum of all the individual eye observations. So you remember, if you looked at one of our earlier modules on summation notation, you know that the Greek letter Sigma tells you to sum everything to the right of that sigma. And you're going to add up all the values in your series. And each x i represent an individual value in a series of values. The small letter N, or lowercase n is the number of values in a series, and it's often called the sample or population size. We'll look more closely at the difference between a sample and a population in a later module. But just be aware that this n is the number of values in a series. It's also sometimes known as the number of observations. Now here's an example of parking tickets issued by the city of Guelph, Ontario. So we've got five different types of infractions, stopping in a no stopping zone, parking in a no parking zone, staying too long in the parking spot excess at the time limit private parking on someone's private property, and being parked on the side of a highway where you're not allowed to park. And you can see here we've got a bar graph, a horizontal bar graph, but a bar graph nonetheless. And the number of tickets is on the x axis. And you can also see that the actual total has been labeled for us. And so maybe we're interested in knowing the mean number of tickets issued. And if we want to know that, we will add up all the values. And we want to divide by N. What's N, while we've got 1, 2, 3, 4, 5 different types of tickets. And using my calculator, I get 24,502 divided by five, which is equal to 4900.4. Now the mean, we see here that although none of the individual values is 4,400.4, but the mean, the average of all five is 4,400.4. Here we have a histogram. And remember, a histogram is used when we have a continuous variable like salaries, and we've created bins. So we've got zero to $1 million per year. That's this observation here. So we see there eight players on the Toronto Maple Leafs who earned between zero and $1 million per year early that was their stated salary on the where I got the data from. And there were seven players who earn between one and $2 million. There was one player who earned 10 million and there were two players, who earned more than 11 million and less than 12 million. So there's our histogram and how to read it. I'm going to introduce to you something called a stem and leaf plot. It's a way to create a table that will show many values in a succinct fashion. So without with using less space than if we just wrote out each of the values one after the other. Now how do we interpret this stem and leaf plot? Well here this first column is called the stem and the second column is called the leaf. Not to be confused with the the Maple Leafs. Now, the stem is written in millions of US dollars and the leaf are fractions of millions of dollars. So if we focus on let's say this point, this one right here, we've got two values here we've got one player making $3.5 million and another player making $3.8. million. Remember, remember we've got eight players making less than a million. We've got seven in there. We've got so that's 15 players there. We've got two here, two, they're three they're one they're making $7.0 million, one player making 10 point 9,000,002 players making while one player making 11 million the other making $11.6 million a year. What is our n equal to here? Well, we've got 15 19, 15 here for their for 19. Three here for are 22, 23, 24 and 26. So we have 26 players who have played for the Toronto Maple Leafs in the year 2021, 22. Should say the season 21, 22. And if we wanted to calculate the mean, we would sum up each of these values since there's 26. That would take too long in the video. But having done that, with software package that I like to use, I can tell you that the average or like I should say, Be more specific than the average the mean player salary in 21, 22 was $3.2 million US. So I hope you can see how to use the stem and leaf plot. I guess we can think of this line here in this example, it's like a decimal place and the stem here is in millions of dollars. And we've got a fraction of a million dollars or a decimal place or hundreds of $1,000 on the right hand side. Each number represents one value in the series.