Robert_S2_L12

Wed, 1/12 12:25PM 16:29

SUMMARY KEYWORDS

cumulative distribution function, interval, frequency, probability, cumulative probability, inches, data, graph, histogram, x axis, probability density function, y axis, equal, probability distribution, households, sampled, create, cumulative frequency, function, observations

SPEAKERS

Robert McKeown

R Robert McKeown 00:05

Hello everyone, and welcome back. In this video, we're going to be looking at how data can be represented as a distribution, there could be two distributions, we're going to look at a probability density function and a cumulative density function. This is similar to turning frequencies into probabilities. And I'm going to illustrate this to you using examples and graphs. So let's get to it. As you learned for visualizing data, data can be represented according to frequency. So you could sort it into a frequency interval and then create a frequency distribution. So let's refresh your memory. We'll start with something you know. And then we'll add on something you don't know. Now, when it comes to creating histograms and working with data, this is something we commonly do with continuous data. And since you know, each observation with enough decimal places tends to be unique. And so there's no mode. With continuous data, there's often no mode, and many different variables. So we're going to want to group the data into class intervals, for example, from four to six centimeters. Each of these intervals, such as four to six centimeters is a class interval, and encompasses a range of values. We saw this on all the histograms that we've seen, here's one we worked with earlier, where we're looking at soccer player heights and inches upon close inspection of the X axis here, that each then is one inch interval, one inch interval. And whether you're 60.2 inches or 60.8 inches, you're going to be in this interval right here, regardless, regardless of the fact that the numbers are not identical. Also, notice that the y axis is frequency. And so upon close inspection, it looks like we walk across here, there are 15 values that fall between 64 and 65 inches. Now here, we have two different histograms. We've got number of children on the X axis of both, and this is from a sample of us families. And we can see that the maximum number of children that each family had was eight. And we can see the frequency on this diagram here. So for example, more than 300 of the households sampled had no children. The graph here, the bar graph, or the histogram over here, it looks very similar, but the y axis is changed and the y axis is actually a percentage. That's useful, because now you can see clearly that out of all the households sampled about looks like 24%. So one, almost one in four households have no children. And this is what we can call a probability. density function. And we are using the frequencies to estimate the probability density function or the probability function. Here we are using the frequencies to try to come up with a measure of the probabilities, the probability known here as the probability density function. Now we're looking

at our PDF here our probability density function. This is the same graph that we had in the previous slide on the right hand side. We've got percentage on the Y axis, and we've got number of children on the X axis. It might be interesting to know something about the cumulative distribution function. So distribute the cumulative distribution for example, how many households have three children or less, how many households have three children or less well looks like something like maybe, let's say 85% of households in the survey have three children or less. The cumulative distribution function adds the probability distribution functions as we go from lowest to highest. So we've got say, say we have 24% here, and let's call that 17%. Here, then this point right there, where we're looking at households that have zero, or one, children, how many households have zero or one children? Well, it looks like 41%. How many households have two children or less. That would be right there. And let's call this 25%. And now we're going to have 66% of households have two children or less have zero children, one children or two children. And so this cumulative distribution function is another way to demonstrate and to visualize data also to quickly calculate the probabilities. The probabilities of what the probabilities of having less than a certain amount of whatever variable is on the x axis, in this case, number of children. As an exercise, let's take a look at the histogram that we have down here histogram, it has four intervals. So we've got 60, to 64, 64, to 68, 68, to 72, and 72 to 76. So there's four inches in each interval. And we've got labels here that I hope you can read, clearly, we've got just in case you can't, we've got 9, 55, 29, and 8, we want to use the frequencies to create the probability density function and the cumulative distribution function. Rather than graph it, we will put it into this table right here. Now we have four ranges. So let's start with the first one, we had 60 to 64. What's the frequency? How many units are in the 60 to 64? band while there's nine? What's the cumulative frequency? Well, it's the first band the most leftward interval. And so we're going to just keep that as nine nine plus zero. And what's the probability? That or what fraction, really the probability is saying what fraction of players are between 60 and 64 inches tall? To answer this, we're going to need to know how many observations we have. And we have nine plus 55 plus 29, plus eight, which is equal to 101, which is the same number of observations we had before. Now if we want to know the frequency, we're going to take nine Sure, you can see, we're going to take nine divided by 101. And that's going to give us well, something very close to 9%. I will give us 8.9%. And what is the cumulative probability? Well, it's going to be 8.9%, because this is the most leftward interval, and there are no intervals before. Now let's move to the next interval 64 to 68. There are 55 players that fall into this interval. What's the cumulative frequency? Well, it's going to be 55 plus nine, which equals 64. And what's The probability that a player or what fraction of players are between 64 and 68 inches tall? Well, that's going to be 55 divided by 101. And we'll call that 54.5%. And what's the cumulative probability? Well, it's going to be 8.9 plus 54.5, which is going to be equal to 63.4. We're now ready to move on to the next interval. We've got 29 players here in this interval. What's the cumulative frequency? Well, it's going to be 64 plus 29, which gives us 7793. And now we're ready to calculate the probability or the the fraction of players that are between 68 and 72 inches tall, that's going to be 29 divided by 101, which is equal to 28.7. What's our cumulative probability? Well, we've got 63.4 plus 28.7%.

That's going to give us

R Robert McKeown 11:47

92.1%. Now we're ready for our last interval.

We've got 7276.

Robert McKeown 12:00

What's the frequency in here? It's eight. What's the cumulative frequency is 93.8, which equals 101. What is the fraction what fraction of players are between 72 and 76 inches tall? Well, eight divided by 101 is going to be very close to 8% 7.9%. And we're ready for our cumulative probability, that's going to be 92.1% plus 7.9%, which is equal to 100%. We've completed the table took a fair bit of work. But you can see we've got our cumulative probabilities, which we can graph we can draw our cumulative distribution function. So if we start at the beginning, we've got our intervals. Maybe we could have

64 here 68, 72 and 76.

Robert McKeown 13:26

That means our inches

inches in height.

Robert McKeown 13:38

And we're starting at zero. But at zero, if we have a 60. Here, we want to start with those 60 inch players. And so we're going to go up to 8.9%. Draw a horizontal line until we get to 64% which point the probability shoots up too far shoots up to 63.4%. Then we get to 68% shoots up

to 90 to

Robert McKeown 14:26

1%. Then when we get to 72% all the way to 100%. So there's our cumulative distribution function. For soccer player heights when we use an interval of four inches. Why did we use an interval of four inches? Well, I didn't want us to have too many calculations. It took us quite a

bit of time to make those calculations when we only had four intervals to consider. What would this look like if we used a smaller interval? Well, here we have the cumulative distribution function, where we have the smallest intervals possible, believe these intervals are half an inch. Although it looks like some, there might not even they might even be smaller than that. And you can see our cumulative distribution function has many more steps. So more intervals leads to more steps. And I wouldn't want to have to calculate this by hand or even draw this by hand. You had a chance to create a probability distribution function by hand, you created a cumulative distribution function by hand. These are useful things to do now, because when you move further into your study of statistics, there are going to be assumptions and things you need to understand about the underlying probability distribution. Having a intuition about what a PDF and a CDF are is going to be helpful for you