

module_prob4_lecture11

Fri, 2/18 12:51PM 15:16

SUMMARY KEYWORDS

normal distribution, sigma, standard deviation, cdf, pdf, probability, variance, random variable, lie, symmetric, remember, suppose, distributed, point, bell curve, left, infinity, minus infinity, observations, contexts

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:04

So now I want to talk about another continuous distribution, perhaps the most important and the most commonly used of them all. And that's called the normal distribution. So this is a distribution, you will find used in economics, in finance, in psychology, in practically any of the sciences or social sciences that you can think of right. And it has been used in contexts like the birth and height of babies of men and women, IQ test scores, exchange rates, stock market in this is realistic prices almost in a wide, one wide variety of contexts. So even if we don't know the name, perhaps normal distribution, perhaps you've seen something like this. So this is exactly the PDF of the normal distribution, but it looks like is a bell to see this. So this mean is the mean, of the of the normal distribution. And this is where most of the values lie, right? So this has this is where the PDF is the highest. And then it slowly tapers off going this way, and slowly tapers off going this way. Right. So this is a bell shaped curve. And sometimes this is also people just call it the bell curve. And there's also a technical another name for it. It's called also the gaussian distribution. So these are the various names that's used for the normal distribution. So the normal distribution has two parameters. One is its mean. And the second is its variance. So the mean is μ , and the variance is σ^2 .



If you



Sumon Majumdar 02:10

say that a random variable is normally distributed, you usually write it like this, that x is normally distributed with mean μ and variance σ^2 . So people instead of the variance, they also write it in terms of the standard deviation. Remember, that's just the square root of the variance. So for example, in Excel, they write the normal distribution in terms of the standard deviation say. So what's the PDF of the normal distribution, it's given by this function here, which is $\frac{1}{\sigma \sqrt{2\pi}}$. And then he remember, e is the base of the

natural logarithm. And this was a number of gels introduced in the previous module, look at when we were looking at the points of distribution. So it's the same e here, e to the power of minus half x minus μ over σ holding squared. And this is in here, the range of x goes from minus infinity to plus infinity.



Now typically, like in



Sumon Majumdar 03:27

our examples that we use, where tie scores, etc, right, none of them go off to minus infinity or to plus infinity. But still the normal distribution is used. And I'll tell you in a short while why it's appropriate to use. But nevertheless, the normal distribution, the range of the normal distribution, technically, this goes from minus infinity to plus infinity.



Now, this



Sumon Majumdar 03:56

is the PDF or the density of the of the normal distribution. And that's the bell curve that, and it's the graph of this that you saw the bell curve that we just shot saw in the previous slide. Now, what about the CDF, now, the CDF you get by integrating the PDF up to a point B , the problem with the normal distribution is this integral, you cannot get a closed form solution of it, there is no specific formula for it. So what you typically use is software or probability tables to figure it out. So in Excel, the command is equal to norm dot Dist. And you put in whatever where you want to find the CDF for re found find the CDF and needed instead of x and put it be there. And then the other parameters of the mean, the standard deviation And I use true true gifts CDF. And if you want to get just the PDF, then you use false. That gives you the PDF. This is a similar sort of commands that we used before to calculate probabilities for the boys on the binomial distribution. Here. It's NORMDIST in Google Sheets is the same formula, but instead of the.is, just nonplussed. And I'll be showing you examples in Excel of actually computing this. Now, what's the shape? And we saw before, right, so this is a symmetric around the mean meal, right? So this middle point is the meal. And it's symmetric around. So what it means is that if you look at here's the mean, and if you look to the left of it, right, so this area, this will be 50%. And if you look to the right, this area, so 50% of the observations will be here. So that's it symmetric, right. So they have the same math on the left as you have on the right. Now, as the mean changes, right, so as new changes, so the PDF either moves to the left or to the right, so here in this diagram, so I started out with new, and suppose the new became bigger, richer increased, the mean, was just that intense is that the PDF will just move to the right. On the other hand, if new became low, right, if you had a lower μ , then this, this would just move to the shape would remain exactly the same. So what about the impact of sigma? Remember, that's the standard deviation. So a sigma changes, what it does is that it the density either flattens out, or it becomes steeper. So in this diagram, what I've looked at is that suppose the

standard deviation of the variance increases. So remember, variance gives you the dispersion of the random variable. So what it means is that the CDF, now it spreads out, right, it moves to the left and to the right, and it comes down around one, the total area under the PDF, right, this has to be equal to one. So if you want to still keep the area one and spread it out, the only way this can happen is that this has to be coming down. Right. So this comes down and flattened. So that's what's happening to the PDF, right. So this is the PDF. And this is, again, is the PDF for a higher standard deviation or higher variance. So there are some other more interesting properties as well. So first, remember I said it's symmetric around the mean. So if you look to the left, there's 50%. And if you look to the right, it's 50%. So the probability that the random variable is less than equal to its mean, that probability is point five, the probability that the random variable is greater than the mean. That's also point five. Now so here's some interesting, another interesting property. And this is sometimes called the 68-95-99.7 rule. So what is this rule? So if you look at the mean, right, and you look at one standard deviation to the left of it, so you look at $\mu - \sigma$, and one standard deviation to the right of it, so that's $\mu + \sigma$. So if you look at that area in between 68% of the observations lie in that. So technically, if you look at the random variable X , and you'll see what's the probability that it lies between $\mu - \sigma$ and $\mu + \sigma$, that's going to be point six, eight. So that's the 68-95-99.7 rule.



What if, what if you look at



Sumon Majumdar 10:11

two sigma away, right? So $\mu + 2\sigma$ and $\mu - 2\sigma$. So what it does is that this interval now expands, right. And if you look at that, then you find that 95% of your observation lie between $\mu - 2\sigma$ and $\mu + 2\sigma$. So, between $\mu - 2\sigma$ and $\mu + 2\sigma$, if you look at what's the probability that the random variable external the lies between those two, that's point nine, five. And so this is the starting about the 68-95-99.7 rule, right? So this is the second tip. That is if you look at two sigma away from the from the mean, right, minus two sigma and plus two sigma away from the mean, then 95% of the observations will lie there. If you go a step further, right, and you look at $\mu + 3\sigma$ and $\mu - 3\sigma$, so, three sigma away from the mean 99.7% of the observations lie within that, right? So, if you look at the random variable X , and you look at what's the probability that it lies between $\mu + 3\sigma$ and $\mu - 3\sigma$, that probability is point 997. So, in a sense, right, if you look at the mean, and you go $\mu + 3\sigma$ and $\mu - 3\sigma$, so, 99.7% Nearly every, all of these observations, right? lie within that, right? So, it's almost certain, right, it's 99.7% probability that the random variable will lie within data. Right? So that's why this is called the 68, 95, 99.7. So, right, so what does this use? So this is used in things like quality control, suppose you know that a light bulbs life is normally distributed with a mean of 100 and sigma equal to 50. So what it means is that if you look at 1000, right? Remember I said, if you look at 1000, plus three sigma, so that's three times 50. So this is what is this 1150, right, and 1000 minus three sigma. So that would be 1000, minus three times 50. So that's 850. So 99.7% of the boards will have a life between 850 to 1150 hours. So when when you see advertisements that this ball will almost surely have a life between this and this, right? This is the type of calculations. And as you can see that as this sigma, the variance becomes smaller and smaller, right? This thing, this interval, within which 99.7% of the bulbs will like that will become tighter and tight. So quality control when they're

doing checks, what they are trying to do is reduce the sigma so that they are more and more sure that most of the bulbs will have like a lifetime, very close around \$1,000. So let's, so I want to do another example along these lines. So this is an example of suppose a door dashes drivers delivery time is normally distributed with mean new of 20 minutes, and sigma equal to four. Right? So remember, this is the standard deviation. And what I would like to calculate is what's the probability that it lies within one standard deviation within two standard deviation within three standard deviation. So what I'm going to do is I'm going to stop the clip here. And I'm going to take this out exactly in the next clip. So if If you wish to try go ahead and try this out on your own please do so and that will give you some practice about the 6895 99.7