Slide 1 – Title Slide

Hello and welcome to Week 7, Part 5 of EGM101: Continuous Probability Distributions. In this lesson, we'll look at probability distributions for continuous variables.

Slide 2 – Probability as Area

I mentioned this briefly in the previous lesson, but we'll look at it in more detail here. Each of the bars on this chart have an area – they're rectangles, so they have an area that is equal to the width of the bar, w, multiplied by the height of the bar, H. And, since the height of the bar is equal to the probability, this means that the area of the bar is the width of the bar, multiplied by the probability.

But, that's not all! For many of the discrete probability distributions that we looked at in the previous lesson, like this one, the width of the bar is equal to 1 – so the area of the bar is actually just the probability.

And, remember that the total probability has to equal 1, which means that the sum of the areas of each of the individual bars has to add up to 1.

So, the area, or probability, of getting a value in some range is just the sum of the bars in that range. For example, we can sum the area of the bars at xi equals 4, 5, and 6, shown in red, to get the probability that we have between 4 and 6 heads in our coin flip experiment. If we do that, we get a value of 0.656.

This all works great for discrete variables, since we have a nicely-defined boundary between values, which means that we have a nicely-defined width. What happens when we have a continuous variable, where we don't have this nicely-defined boundary between values?

Slide 3 – Continuous Probability Distributions

Remember that a continuous variable can take on any value. Looking at a histogram of probability distribution like we saw on the previous slide, we see that this means that the width of the bars gets very, very small – in each panel of this figure, as we go from 10 bars up to 40, the width of the bar for any specific value gets smaller and smaller – which also means that the probability of any specific value gets smaller and smaller.

So, for continuous variables, we can't think about the probability distribution in the same way that we did for discrete variables. Instead, we represent the probability using a function, called the probability density function, or pdf, which is the continuous version of the discrete distributions we looked at previously.

To calculate the area, we can also use something called the cumulative distribution function, or cdf. The value of the cumulative distribution function at a specific value of x tells us the area under the probability density function for all values of x less than that value. As an illustration of this, The gray shaded region represents the area under the probability density function for all values of x less than 4. The value of the cumulative distribution function at x = 4 is equal to the area of the gray shaded region.

If we know the value of the cumulative distribution function for two different values of x, we can calculate the probability that the outcome of our experiment is between those two values.

This also means that we can use the cumulative distribution function to calculate what proportion of values are less than, or greater than, a given value – as we will see, this is extremely useful for hypothesis testing.

Slide 4 – The (Continuous) Uniform Distribution

Just like with discrete probability distributions, there are a number of theoretical continuous probability distributions that are very useful – we will look at a few of the most useful ones, beginning with the uniform distribution, or the continuous uniform distribution to help differentiate it from the discrete version.

In the uniform distribution, all values within some range between values a and b have an equal probability – in fact, the probability is equal to 1 divided by the difference between a and b. It's important to note whether or not the end points a and b are *inclusive* or not – that is, whether they are possible outcome values or not.

For the uniform distribution, outside of the interval between a and b, the probability density function is equal to zero. On the graph here, you can see that for values less than a, the value is zero, and the same for values greater than b. The cumulative distribution function is equal to zero if x is less than a, but if x is greater than b, it is equal to 1; in between a and b, value increases linearly, with a slope equal to 1 over b minus a.

Slide 5 – The Exponential Distribution

The exponential distribution is useful for answering questions like "how long do we have to wait until the next event happens?" For example, if we know that a battery lasts 12 months on average before it needs to be replaced, how likely is it that a new battery lasts longer than 8 months before we have to replace it?

The exponential distribution, shown here, depends on two things: first, the mean "waiting time", denoted using the lowercase Greek letter mu, and something called the "decay parameter", denoted using a lowercase m, which is equal to 1 over mu.

When we use the exponential distribution, we are assuming that the events in question occur continuously and independently at a constant rate. That is, the "mean waiting time", mu, does not change over time. Another way of saying this is that the exponential distribution is "memoryless" – the waiting time does not depend on how much time has passed since the last event.

Slide 6 – The Normal Distribution

We introduced the normal distribution last week, and it is still the most important distribution for statistics and probability. To calculate the probability density function and cumulative distribution function for the normal distribution, we need to know two things: the mean value of the distribution,

denoted using the lowercase Greek letter mu, and the standard deviation of the distribution, denoted using the lowercase Greek letter sigma.

The peak of the probability density function is at mu, and because the distribution is symmetrical, 50% of the area under the curve is at values less than the mean, and 50% of the area is at values greater than the mean. That is, the value of the cumulative distribution function at mu is equal to 0.5.

We can also see that as we move away from the mean in either direction, the probability decreases quickly – this is something that we'll come back to shortly.

Slide 7 – Z-scores

The issue with using the normal distribution is that it can be difficult to compare the distribution for variables that have different ranges of values. To help with this, we can define a statistic, z (or zed), which tells us how many standard deviations away from the mean any value of x is.

This has the effect of standardizing the normal distribution – in fact, the distribution of z scores is called the "standard normal distribution", and it's a special case of the normal distribution where the mean is equal to 0, and the standard deviation is equal to 1. We can estimate the probability of a value using its z-score – and, as we will soon see, this makes it very easy to compare data that have different scales – by converting to z-scores, we can easily tell how far away from the mean of the distribution a particular value is.

In ancient times, people used statistical tables to look up values of the cumulative distribution function based on the value of z – in fact, if you look in any statistics textbook, you may still see these relics of a bygone age.

Now, though, in the year of our Lord two thousand and twenty-two, we have computers that we can use for this.

Slide 8 – The Empirical Rule

If our data are normally distributed, we can calculate the probability that any value randomly selected from the dataset is close to the mean, using the standard normal distribution.

The probability that a randomly selected value, X, is within one standard deviation of the mean is equal to 68.27%; within two standard deviations, it's 95.45%; and within three standard deviations, it's 99.73%.

Another way of saying this is that 68.27% of all values of x fall within one standard deviation (or one-sigma) of the mean; 95.45% fall within two standard deviations of the mean; and 99.73% of all values fall within three standard deviations of the mean. On the graph here, the blue shaded region is equal to 68.27% of the area under the curve; the blue and the red area combined are 95.45% of the area, and the blue, red, and black areas are 99.73% of the area.

This is known as the “empirical rule” – we can use it to help remember how what percentage of our data lies within some interval from the mean value. You may also see this under the even more helpful name of the “68-95-99.7” rule, where the values are actually included in the name of the rule.

Slide 9 – Example: The Fish Question

At the beginning of the week, I asked a question that we’re now finally ready to answer. Let’s say that we know that on average, salmon in our favorite stream are 70 cm long, with a standard deviation of 15 cm. If we go fishing in our favorite stream, what is the likelihood of catching a fish 100 cm or longer?

To answer this, we have to calculate the z-score for 100 cm. So, we start with the formula for the z-score, plug in the different numbest – 100 for x, 70 for mu, 15 for sigma, and we end up with a z-score of 2. That means that a fish 100 cm long is 2 standard deviations larger than the average length for fish in this stream.

Now, remember that probabilities have to add up to 1. So, the probability of catching a fish less than or equal to a certain length and the probability of catching a fish greater than a certain length add up to 1. This means that the probability of catching a fish greater than a certain length has to be equal to 1 minus the probability of catching a fish less than that length.

this means that we can use the cumulative distribution function to calculate the probability of catching a fish greater than 100 cm, using the mean and standard deviation that we know from earlier.

Alternatively, we can use the standard normal distribution and the z-score. The probability of getting a z-score greater than 2 is equal to 1 minus the cumulative distribution function value for a z-score of 2, which works out to be about 0.0227. In other words, we have about a 2% chance of catching a fish that big, given what we know about the size distribution of fish in our stream.

We can also plot this – the black dashed line is the probability density function, and the blue shaded region represents the area under the curve greater than 100 cm (or, greater than two standard deviations away from the mean value). This is something that we’ll see a lot more of next week when we learn about hypothesis testing.

Slide 10 – Summary

In this lesson, we’ve discussed how with continuous variables, we have to think about the probability of a range of values, rather than individual values.

We’ve also seen a number of examples of commonly-used continuous distributions, like the uniform distribution, which we use when all values have an equal likelihood; the exponential distribution, which we can use for estimating the amount of time that takes place between different events; and the normal distribution, which many people wind up using for, well… everything.

With the normal distribution, we also learned about z-scores and the standard normal distribution, which we can use for comparing distributions of values that have different scales.

Slide 11 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapters 5 and 6; Caswell, Chapters 8.9 and 13.5 to 13.9; and Weiss, Chapter 6.

I've also included links to three articles about some of the topics covered here – about the Uniform and Exponential Distributions, and about the Empirical Rule. And, finally, there's a link to a video from Khan Academy that goes into more detail about continuous probability distributions.

That's all for this lesson – I hope you found it interesting, and if you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!