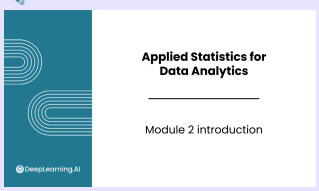
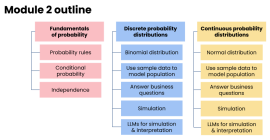
DAG C2M2 scripts

Video title	Isabel	Sean	Slides
L0V1 – Module 2 introduction	✅	✅	✅
L1V1 – Randomness and uncertainty	✅	✅	✅
L1V2 – Probability and the addition rule	✅	✅	✅
L1V3 – The multiplication and complement rules	✅	✅	✅
L1V4 – Conditional probability	✅	✅	✅
L1V5 – Independence	✅	✅	✅
L1V6 – Random variables	✅	✅	✅
L2V1 – Estimation	✅	✅	✅
L2V2 – Sample distributions to population distributions	✅	✅	✅
L2V3 – The Bernoulli distribution	✅	✅	✅
L2V4 – The Binomial distribution	✅	✅	✅
L2V5 – The cumulative distribution function	✅	✅	✅
L2V6 – Random sampling – discrete	✅	✅	✅
L2V7 – Demo: Spreadsheet simulation – discrete	✅	✅	✅
L2V8 – Demo: LLM simulation – discrete	✅	✅	✅
L3V1 – Continuous probability distributions	✅	✅	✅
L3V2 – The normal distribution	✅	✅	✅
L3V3 – The standard normal distribution	✅	✅	✅
L3V4 – Random sampling – normal	✅	✅	✅
L3V5 – Demo: spreadsheet simulation – normal	✅	✅	✅
L3V6 – Demo: LLM simulation – normal	✅	✅	✅
L3V7 – Making decisions with distributions	✅	✅	✅
Coursera dialogue item introduction	✅	✅	✅

Introduction

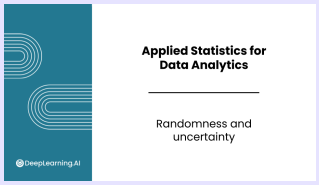
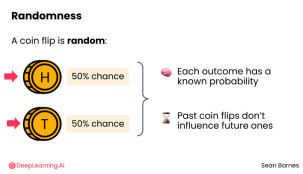
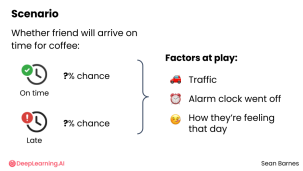
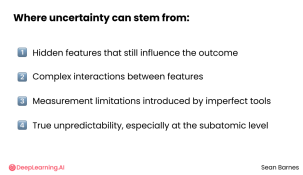
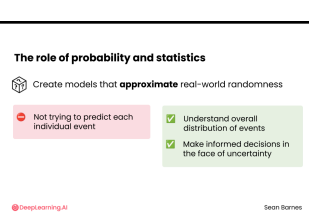
L0V1 – Module 2 introduction

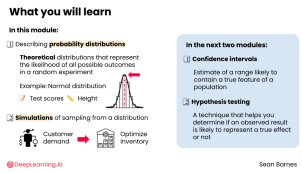
Visual	Script
TH 	Welcome to Module 2: Probability and simulation!
	In this module, you'll start by learning about **[CLICK]** probability – the language used to quantify uncertainty. You'll cover key **[CLICK]** probability rules and concepts like **[CLICK]** conditional probability and **[CLICK]** independence, all with real-world examples you'll encounter as a data analyst. Then you'll explore probability distributions, both **[CLICK]** discrete and **[CLICK]** continuous. You'll learn about common distributions like the **[CLICK]** binomial and **[CLICK]** normal distributions, and how they model real-world phenomena. You'll also see how you can **[CLICK]** use sample data to understand the distribution of your population, and how to **[CLICK]** answer common business questions like how common are certain outcomes or ranges of outcomes? Throughout the lessons on probability distributions, you'll get hands on with **[CLICK]** simulation techniques. You'll see how to generate random data following specific distributions, allowing you to model complex scenarios and inform decision-making. You'll also **[CLICK]** use a large language model to create interfaces for simulation and help you interpret results.
	By the end of this module, you'll have a solid foundation in probability and simulation – crucial tools for any data analyst. These concepts will prepare you for more advanced statistical techniques in future modules, including creating confidence intervals and performing hypothesis tests. Let's get started! Follow me to the next video which is all about randomness and uncertainty 👀

Lesson 1 – Probability

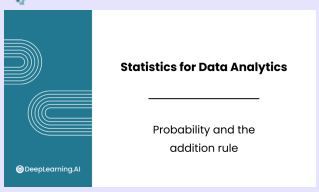
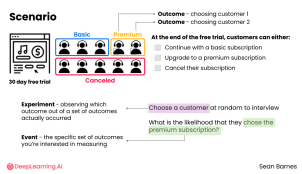
L1v1 – Randomness and uncertainty

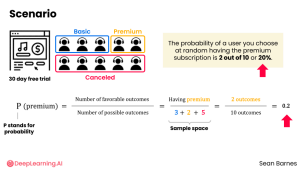
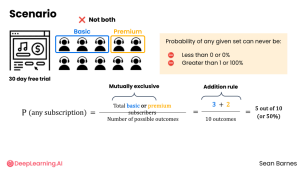
Visual	Script

TH 	Probability is the language of uncertainty. You're waiting for the train during your morning commute – will it come at 7:52 or 7:57? Real world data, like train times, is subject to randomness. Probability provides the tools for you as a data analyst to quantify and reason about this uncertainty.
	Let's talk about randomness for a moment. **[CLICK]** A coin flip is random, a 50/50 chance of **[CLICK]** heads or **[CLICK]** tails. However, coin flips are an example of an experiment where **[CLICK]** each outcome has a known probability (the chance of occurring), and **[CLICK]** past coin flips don't influence future ones. You'll get heads about **[CLICK]** half the time, and **[CLICK]** tails about half the time. Real-world randomness is more complex. I'll give you an example. I'm going to show you some numbers on the next slide, and I'd like you to pick one of them. Don't think too much about it, just pick one.
	**[pause for learner to pick one]** Let me guess… did you pick **[CLICK]** 3? You may be surprised to know that nearly 75% of people do. You can try this on your friends and family. Even for a task as simple as pick one of these four numbers, people don't really pick randomly. What seems like a simple 1 in 4 chance is, behind the scenes, a very complex experiment.
	Here's another example. Say you're trying to predict whether your friend will arrive on time for coffee. There are countless **[CLICK]** factors at play – the **[CLICK]** traffic, whether their **[CLICK]** alarm clock went off, **[CLICK]** how they're feeling that day. Most of these factors are unknown or unmeasurable to you.
	This type of randomness or uncertainty stems from several sources, including • **[CLICK]** Hidden features that you don't know about, but that still influence the outcome • Or, **[CLICK]** complex interactions between features. So, even if you know all the features, they might interact in ways that are difficult to model • You can also have **[CLICK]** measurement limitations introduced by our imperfect tools for observing the world • And, there's also **[CLICK]** **true** unpredictability, especially at the subatomic level. Yes, I'm talking physics. Some things are truly impossible to predict.
	Data analysts use probability and statistics to **[CLICK]** create models that approximate this real-world randomness. You're **[CLICK]** not trying to perfectly predict each individual event, but rather to **[CLICK]** understand the overall distribution of events and **[CLICK]** make informed decisions in the face of uncertainty.

	You've already explored distributions of sample data in the previous module. **[CLICK]** In this module, you'll start by **[CLICK]** describing **probability** distributions, which are **[CLICK]** **theoretical** distributions that represent the likelihood of all possible outcomes in a random experiment. On the other hand, distributions of sample data come from actually going out in the world sampling from a population. One example of a probability distribution is the so-called **[CLICK]** normal distribution, which models the behavior of distributions where **[CLICK]** most values cluster around the mean, like **[CLICK]** test scores or **[CLICK]** heights. It's possible to perform a **[CLICK]** **simulation** of sampling from a distribution, which you'll see throughout the last two lessons. For instance, you might simulate **[CLICK]** customer demand to **[CLICK]** optimize inventory levels. In the **[CLICK]** **next** two modules, you will also explore two statistical tools based on probability theory and distributions: **[CLICK]** **confidence intervals,** an **[CLICK]** estimate of a range likely to contain a true feature of a population, such as the mean, and **[CLICK]** **hypothesis testing,** a **[CLICK]** technique that helps you determine if an observed result is likely to represent a true effect or not. These definitions won't make a ton of sense right now, but you'll become very comfortable with them in the next module.
TH	Now that you've seen the purpose of probability for data analysts – to quantify and model randomness or uncertainty – join me in the next video to learn about the fundamental rules of probability.

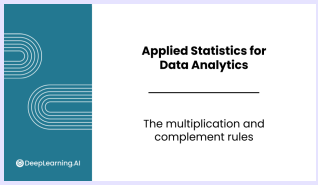
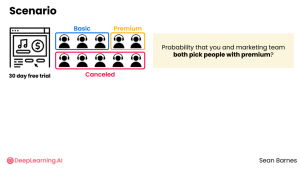
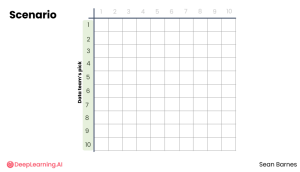
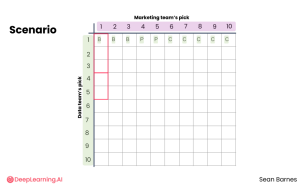
L1v2 - Probability and the addition rule

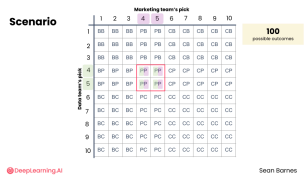
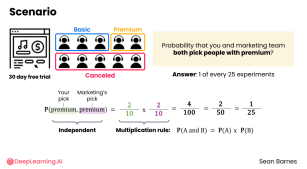
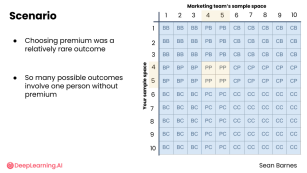
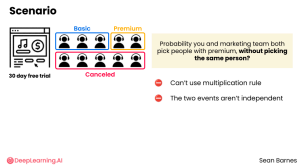
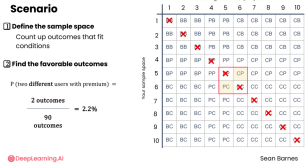
Visual	Script
TH 	Probability is the chance of an event occurring. Whenever the outcome of an event is uncertain, you can use probability to talk about the chance of it occurring. If you flip a coin, the probability it lands on heads is 1 in 2 or 50%. You have an **intuition** about the probabilities associated with this object.
	Let's formalize that intuition about probability using a real world example. Consider a music subscription service that offers a **[CLICK]** 30 day free trial to **[CLICK]** 10 customers. **[CLICK]** At the end of the free trial, customers can either **[CLICK]** continue with a basic subscription, **[CLICK]** upgrade to a premium subscription, or **[CLICK]** cancel their subscription. You found that **[CLICK]** 3 customers chose the basic subscription, **[CLICK]** 2 chose the

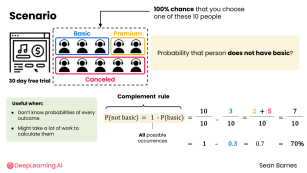
	premium subscription, and **[CLICK]** 5 canceled their subscription. You'll need three terms to talk about probability: experiment, event, and outcome. Say you **[CLICK]** choose a customer at random to interview about their experience. **[CLICK]** What is the likelihood that they chose the **premium** subscription? **[CLICK]** Choosing a customer is called an **[CLICK]** experiment in probability – you are observing which outcome out of a set of outcomes actually occurred. **[CLICK]** A customer having the premium subscription is called the **[CLICK]** "event" – the specific set of outcomes you're interested in measuring. There are **[CLICK]** two different customers with the premium subscription, and choosing each of these customers is called an **[CLICK]** "outcome" of the experiment. For the event "customer has a premium subscription", there are two possible outcomes because you can pick either of the two customers at random.
	Here's how you can write the probability of this event: **[CLICK]** P of **[CLICK]** premium. **[CLICK]** P stands for probability. Now, how do you estimate this probability? The probability of this event – having the premium subscription – is the **[CLICK]** number of favorable outcomes **[CLICK]** over the **[CLICK]** number of possible outcomes. The favorable outcomes are just the ones you're looking for, in this case having the premium subscription. You'll want to add up the number of possible outcomes in the **[CLICK]** denominator – **[CLICK]** 3 people have basic, **[CLICK]** 2 have premium, and **[CLICK]** 5 canceled, **[CLICK]** so in total 10 possible outcomes. This set of outcomes is sometimes called **[CLICK]** the sample space – you're sampling from all these different outcomes. And the event you're interested in is **[CLICK]** having premium, so in the numerator, you have those **[CLICK]** two outcomes that make up this event. So **[CLICK]** the probability of a user you choose at random having the premium subscription is 2 out of 10 or 20%. Probability can be represented either as a proportion – a number between 0 and 1 like 0.2 – or as a percent, like 20%.
	What is the likelihood that you randomly pick a customer to interview with **any** subscription? The **[CLICK]** denominator stays the same, with **[CLICK]** 10 possible outcomes in the sample space, but in this case you are interested in two different events, so **[CLICK]** the total number of basic and premium subscribers. Since these two groups are distinct from each other – either someone chose the **[CLICK]** basic subscription **or** **[CLICK]** the premium one, **[CLICK]** **not** both – you can add them together in the numerator. Being

	separate in this way, where an event can't be both outcomes at the same time is called being **[CLICK]** mutually exclusive. If you got basic, you couldn't have gotten premium. So the outcomes you should include in the numerator are **[CLICK]** the three people with the basic subscription **[CLICK]** plus the two people with premium, or **[CLICK]** 5 out of 10, or 50%. This is called **[CLICK]** the *addition rule*. Notice that the **[CLICK]** probability of any given set of conditions can never be **[CLICK]** less than 0 or 0%, nor can it be **[CLICK]** greater than 1 or 100%. There's nothing more frequent than literally always happening, which is what a probability of 1 means. Nor can there be a negative chance that something will happen.
TH	Probabilities can be combined in many different ways. Follow me to the next video to learn about the multiplication and complement rules.

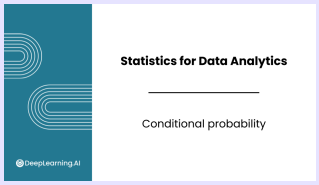
L1V3 – The multiplication and complement rules

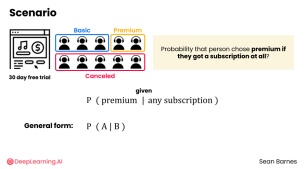
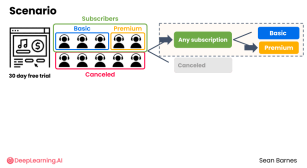
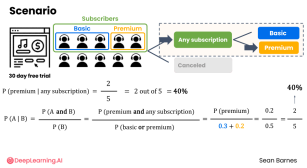
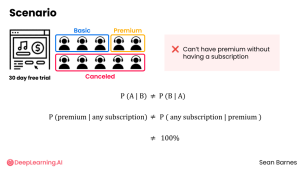
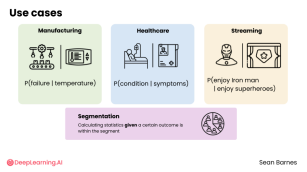
Visual	Script
TH 	How can you determine the likelihood of two events occurring, or of the opposite of an event occurring? Probability rules will help you reason mathematically about these cases.
	Say you're collaborating with the marketing team, and they plan to interview a user as well. What's the **[CLICK]** probability that you and the marketing team, choosing users at random, both pick people with premium?
	You can visualize this experiment in a table. **[CLICK]** In each row you have one of the 10 people you can pick for the data team, so **[CLICK]** 3 basic, **[CLICK]** 2 premium, and **[CLICK]** 5 canceled.
	And in each **[CLICK]** column you have the people the **[CLICK]** marketing team can pick to interview, with the **[CLICK]** same options. So **[CLICK]** BB means both teams picked someone with basic. **[CLICK]** B P means you picked someone with basic and marketing picked someone with premium, and so on.

	This is the sample space, with 10 times 10 or **[CLICK]** 100 possible outcomes. The outcomes you're interested in are where both people that were selected have the premium subscription. Can you spot those outcomes? **[pause for the learner to reflect]** In this case, there are **[CLICK]** four: **[CLICK]** you pick the 4th person and **[CLICK]** marketing picks them too. **[CLICK]** You pick 4, **[CLICK]** marketing picks 5. **[CLICK]** Then 5 **[CLICK]** 4, and finally **[CLICK]** 5 and **[CLICK]** 5.
	To calculate this probability, here's the notation: **[CLICK]** P of premium, premium: the probability that **[CLICK]** you choose a customer with premium and that **[CLICK]** marketing chooses someone with premium. You know that there's a **[CLICK]** 2 in 10 chance that the customer you select chose premium, and there's a **[CLICK]** 2 in 10 chance that the customer selected by the marketing team chose premium, so you can **[CLICK]** multiply these probabilities together. 2/10 * 2/10 **[CLICK]** = 4/100, like you saw in the table, or **[CLICK]** 2/50 or **[CLICK]** 1/25. This is called the **[CLICK]** *multiplication rule*, which applies to estimating the probability of independent events. It's formally written as the **[CLICK]** probability of A and B equals the **[CLICK]** probability of A **[CLICK]** times the **[CLICK]**probability of B. In this scenario, your selection of a person to interview is completely **[CLICK]** independent of the marketing team's selection; therefore, you can multiply the probabilities together to calculate the probability of both outcomes happening together. In the end, only **[CLICK]** 1 of every 25 experiments is likely to include two people who both had the premium subscription.
	That makes sense because **[CLICK]** choosing premium was a relatively rare outcome in the first place. The probability of picking two people who both chose the premium subscription is lower than the chance of just picking one person with the premium subscription, because there are so **[CLICK]** many possible outcomes that involve at least one person without premium.
	Say you and your colleagues on the marketing team want to make sure not to pick the same person. In this case, you **[CLICK]** can't use the multiplication rule, because these **[CLICK]** two events aren't independent – your colleague's choice depends on yours. You'll learn much more about independence later in this lesson.
	For now, you can find the probability using this table. **[CLICK]** First, define the sample space. **[CLICK]** Count up the outcomes that fit the conditions. So, you and your coworker do not choose the same person! That includes all the outcomes except **[CLICK]** these 10 in this diagonal line. So, the **[CLICK]** denominator is 90. Then, **[CLICK]** find the favorable outcomes. Previously you had **[CLICK]** four

	favorable outcomes, but now two of these outcomes have been removed because they involve the same person twice. So, you're left with **[CLICK]** two outcomes. Therefore, the probability of selecting **[CLICK]** two **different** users with premium for the interviews is 2 out of 90 or **[CLICK]** about 2.2%.
	Finally, let's say you want to determine the probability that the person you randomly select to interview does not have the basic subscription. You can use the **[CLICK]** *complement rule*, which is essentially that the **[CLICK]** probability of (not basic) is the same as **[CLICK]** 1 minus the probability of basic. Since 1 represents **[CLICK]** all the possible occurrences, if you subtract out the probability of basic, you're left with everything else. The complement rule is **[CLICK]** useful when **[CLICK]** you don't know the probabilities of every outcome, or **[CLICK]** if it might take a lot of work to calculate them all. To illustrate this example of P(not basic), here are two probabilities. The first is all **[CLICK]** 10 people in the numerator and the **[CLICK]** denominator, which represents the 1 in this equation. There's a **[CLICK]** 100% chance that you choose one of these 10 people. Then, you want to **[CLICK]** subtract the probability that you choose someone with basic. So in the denominator, **again** you have all **[CLICK]** 10 people – that's the sample space. Then in the numerator, the **[CLICK]** three people with basic. And when you subtract these numbers, what do you have left? The **[CLICK]** same denominator, but with **[CLICK]** the 2 people who had premium plus the **[CLICK]** 5 people who canceled in the numerator. So a **[CLICK]** 7 in 10 chance of getting someone without the basic subscription, or 70%. And mathematically you can say **[CLICK]** P(not basic) = 1 minus P(basic), which is **[CLICK]** 1 - 0.3, which is **[CLICK]** 0.7, or a **[CLICK]** 70% chance.
TH	Now that you're familiar with the addition, multiplication, and complement rules, join me in the next video to learn more about conditional probability.

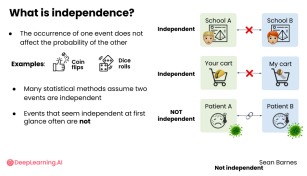
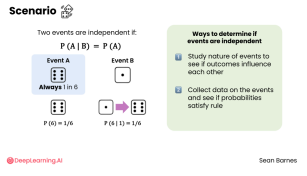
L1V4 – Conditional probability

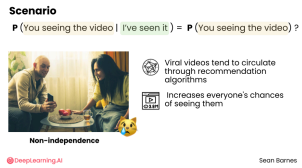
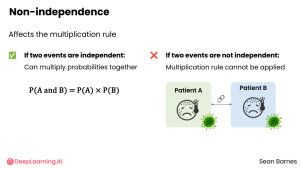
Visual	Script
TH 	Oftentimes in the real world, two events aren't just random, separate events. Events can influence the probabilities of other events, and you can actually calculate the probability of one event happening given that the other already occurred. That's conditional probability.

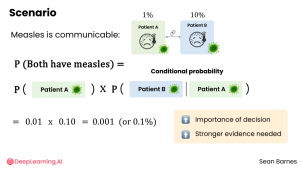
	Let's return to the idea of interviewing subscribers. Say you want to calculate **[CLICK]** the probability that a person chose the premium subscription if they got a subscription at all, so ignoring the people who canceled. Here's how you formalize this concept: you'd say, **[CLICK]** the probability of **[CLICK]** premium **[CLICK]** given **[CLICK]** any subscription. So this vertical line, also called the pipe character, is read as **[CLICK]** "given". More generally, conditional probability is represented as **[CLICK]** P(A given B).
	A flowchart can help you visualize what's happening. There are two branching events here. For the first, either the person **[CLICK]** got a subscription or they canceled. Then, only if they got a subscription, they either got **[CLICK]** basic or premium. What conditional probability does is it just looks at **[CLICK]** this branch of the diagram, and asks: what's the chance that **[CLICK]** someone in this group of subscribers **[CLICK]** got the premium subscription?
	Intuitively, you have narrowed your sample space to just 5 people - the subscribers. **[CLICK]** That's your denominator. And now your numerator is the premium people, of which there are **[CLICK]** two, giving you a probability of premium given any subscription of **[CLICK]** 2 out of 5 or 40%. Let's formalize and generalize that. **[CLICK]** P of A given B is equal to **[CLICK]** P of A and **[CLICK]** B **[CLICK]** over **[CLICK]** P of B. In this case,P of A and B is the **[CLICK]** probability of having premium and having a subscription, so in other words **[CLICK]** the probability of having premium, since you have to have a subscription to have premium. **[CLICK]** So that's 0.2. Then the P(B), **[CLICK]** the probability of having any subscription. This includes the people with basic and premium, so **[CLICK]** 0.3 plus 0.2 equals **[CLICK]** 0.5. That gives you **[CLICK]** 0.2/0.5, which reduces to **[CLICK]** two fifths or **[CLICK]** 40%.
	A quick note that P of A given B is NOT the same as P of B given A. In the example here, P of B given A would be the **[CLICK]** probability of having any subscription given that you have a premium subscription, which is **[CLICK]** 100%. That's because **[CLICK]** you can't have a premium subscription without having a subscription in the first place.
	Conditional probability is widely used in data analytics. You just saw the example with subscribers, but you might also use conditional probability to address business questions like • **[CLICK]** In manufacturing, what's the probability of **[CLICK]** equipment failure given **[CLICK]** a certain temperature in the factory? • **[CLICK]** Or in healthcare, what's the probability **[CLICK]** a patient has a particular condition given **[CLICK]** they exhibit specific symptoms? • **[CLICK]** Or in streaming, what's the probability a given user will enjoy **[CLICK]** the new Iron Man movie if they often **[CLICK]** watch other superhero movies?

	Conditional probability is also related to **[CLICK]** segmentation. You learned previously that segmentation involves calculating your statistics given that a certain outcome is within the segment of data you've defined. In other words, given that a certain condition is true.
TH	Some events in the world depend on each other, so the outcome of one affects the probability of another. But some events are independent – their probabilities are completely separate. Join me in the next video to learn more.

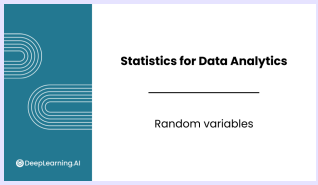
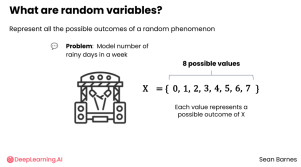
L1V5 – Independence

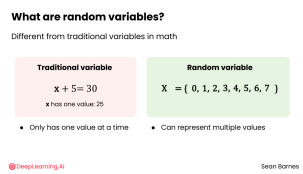
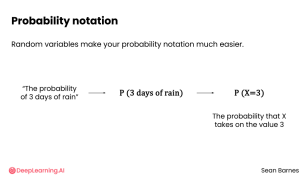
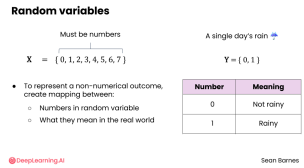
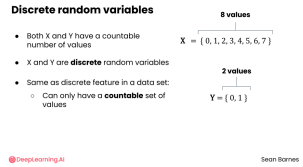
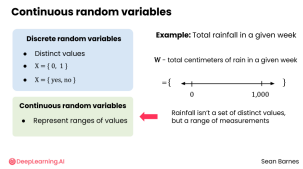
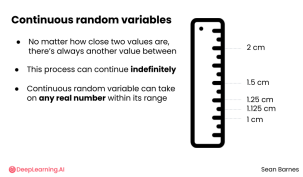
Visual	Script
TH 	You're playing a dice game and you're on a winning streak. Your last four rolls have all been 6s. What's the chance that your next roll is also a 6? It's high right? Well, no. Your chance of rolling a six is 1 in 6, always. That's because the rolls are **independent**. The previous four rolls have no effect whatsoever on the next one.
	Independence in statistics means that the occurrence of one event doesn't affect the probability of the other. **[CLICK]** Coin flips and **[CLICK]** dice rolls are independent. So are **some** real world events. For example, **[CLICK]** two students are taking a geometry test at different high schools. **[CLICK]** Their scores won't affect each other. **[CLICK]** You and I are shopping for groceries in our own cities. What **[CLICK]** you buy and what **[CLICK]** I buy are **[CLICK]** independent; we both made our own decisions, got different coupons, and so on. Whether or not two events are independent is a critical distinction. **[CLICK]** Many statistical methods rely on the assumption that two events are independent, even though subtle connections may exist between the two of them. **[CLICK]** Events that seem independent at first glance often aren't. For example, say you're testing **[CLICK]** two different patients for measles. At first, the tests may seem independent; it's two different people after all. However, measles is a **[CLICK]** highly communicable disease, so if someone at a local clinic tests positive, it's more likely that **[CLICK]** others at the same clinic will have measles too.
	Let's return to dice for a moment; they make things simple. **[CLICK]** Two events are considered independent in probability if they satisfy or follow this rule: **[CLICK]** the probability of A given B equals the probability of A. Here's an example: say that your **[CLICK]** event A is rolling a 6 and **[CLICK]**

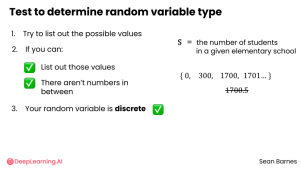
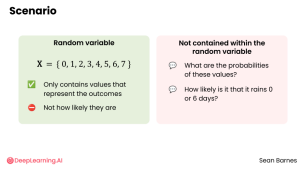
	event B is rolling a 1. **[CLICK]** The probability of rolling a 6, P of A, is 1 in 6. Now say you roll two dice and the first roll is a **[CLICK]** 1. The probability that you roll a **[CLICK]** 6 next given that you just rolled a 1, or p of a given b, is **[CLICK]** **still** 1 in 6. **[CLICK]** No matter what you just rolled, the probability of rolling a 6 is always 1 in 6. So, these events follow the rule above, and are considered independent. **[CLICK]** You can determine if events are independent in a couple of ways. **[CLICK]** First, you can study the nature of the events to see if the outcomes influence each other. In the case of a die, the outcome of one roll doesn't provide any information to help you predict the next one. However, in the real world, proving independence is rarely as easy as it is for dice. Your other option is to test for independence by **[CLICK]** collecting data on the events and seeing if the probabilities satisfy this rule.
	~~Real-world events you'll encounter as a data analyst often don't satisfy the condition of independence. Consider a social media platform with video recommendations. Let's say there's a **[CLICK]** viral cat video taking the internet by storm.~~ ~~Does the **[CLICK]** probability of **[CLICK]** you seeing this video, **[CLICK]** given that I've seen it, equal the probability of **[CLICK]** you seeing it in general? **[pause for thought]** It can be a difficult question to wrap your mind around, but probably not. You're more likely to have seen the video if I've seen it. That's because viral videos tend to circulate through recommendation algorithms, which increases everyone's chances of seeing them.~~ ~~If you've ever **[CLICK]** shown someone a funny video, only to find they've already seen it, you've encountered non-independence. It can be **[CLICK]** disappointing, the videos you were watching really weren't independent of what your friend or family member was watching.~~
	Non-independence affects many statistical calculations, like the multiplication rule you saw earlier. **[CLICK]** If two events are independent, you can **[CLICK]** multiply their probabilities together to get the probability of both events occurring. However, **[CLICK]** if the events are not independent – like **[CLICK]** two measles tests at the same clinic – then the **[CLICK]** multiplication rule cannot be applied.
	In the case of the measles tests, what's the chance of two people testing positive for measles at the same clinic? Imagine the **[CLICK]** probability of the first person having measles is **[CLICK]** 1%. If you assume **[CLICK]** the tests are independent, you'll conclude that the **[CLICK]** probability of the second person having measles is also **[CLICK]** 1%, and **[CLICK]** multiply these: **[CLICK]** 0.01 * 0.01 = 0.0001 or 0.01%. **[CLICK]** One in 10,000.

	However, measles is highly communicable. If the first person **[CLICK]** tests positive, the probability of the second person having measles might **[CLICK]** increase to, say, **[CLICK]** 10%. In this case, you would need to use the **[CLICK]** multiplication rule for dependent events. You would calculate: • **[CLICK]** P(Both have measles) = **[CLICK]** P(First person has measles) **[CLICK]** times **[CLICK]** P(Second person has measles | First person has measles) • The last term – the **[CLICK]** conditional probability – quantifies how much the first person having measles affects the chance that others around them have measles too. • This equation comes out to **[CLICK]** 0.01 * 0.10 = 0.001 or 0.1%, so **[CLICK]** one in 1,000 So assuming that the two measles tests were independent gives you an estimate that's off by an order of magnitude, or 10 times different. Remember, **[CLICK]** the more important a decision, **[CLICK]** the stronger the evidence you need, and determining whether two events are independent is one part of establishing rigor.
TH	So that's independence! Join me in the next and final video of this lesson to learn more about representing events using random variables.

L1V6 – Random variables

Visual	Script
TH 	How do you represent all the possible outcomes of an event in a way that allows you to work with those outcomes mathematically? For that purpose, data analysts use random variables. Let's take a look.
	Random variables **[CLICK]** represent all the possible outcomes of a random phenomenon. As an example, say you're working with **[CLICK]** an outdoor event space and they ask you to **[CLICK]** help model the number of rainy days in a week. You could do so with a random variable. Let's call it **[CLICK]** capital X. Random variables are typically represented with capital letters. Weather is unpredictable, so a **random** variable makes sense here. **[CLICK]** What are all the possible outcomes of this random variable? Well, it could rain **[CLICK]** 0 days in a given week, **[CLICK]** 1 day , **[CLICK]** 2 days... all the way up to **[CLICK]** 7 days in that week. So there are **[CLICK]** 8 possible values, 0

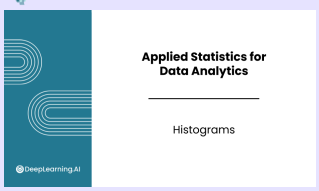
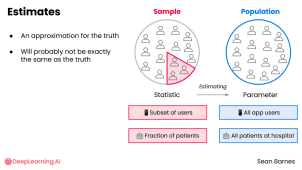
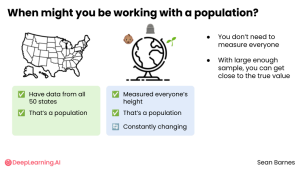
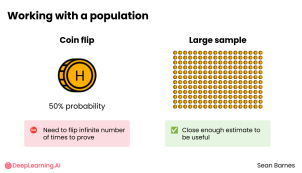
	through 7, and **[CLICK]** each of these values represents a possible outcome of this random variable.
	You can see this type of variable is quite different from a traditional variable in math, which **[CLICK]** only has one value at a time. If you have an equation like **[CLICK]** x + 5 = 30, then **[CLICK]** x only has one value: 25. **[CLICK]** **Random** variables on the other hand **[CLICK]** can represent **multiple** values, each corresponding to a possible outcome.
	Random variables make your probability notation much easier. When you want to say **something like** **[CLICK]** the probability of 3 days of rain, instead of saying **[CLICK]** P of 3 days of rain, you can say **[CLICK]** P of X = 3, or the **[CLICK]** probability that X takes on the value 3.
	The values represented by a random variable **[CLICK]** must be numbers, otherwise you wouldn't be able to work with them mathematically. However, if you want **[CLICK]** to represent a non-numerical outcome, you can create a mapping between the **[CLICK]** numbers in the random variable and **[CLICK]** what they mean in the real world. For example, you could represent a **[CLICK]** single day's rain with a random variable capital **[CLICK]** Y, and have **[CLICK]** 0 represent "not rainy" and **[CLICK]** 1 represent "rainy".
	Notice that **[CLICK]** both X and Y have a countable number of values. **[CLICK]** X has 8, **[CLICK]** Y has 2. Because they each have a set of distinct values, both X and Y are what's called **[CLICK]** discrete random variables. This **[CLICK]** concept of discrete is the same as a discrete feature in a dataset: **[CLICK]** it can only have a countable set of values.
	Since some phenomena in the real world can't be represented with distinct values, you can also have **[CLICK]** continuous random variables, which **[CLICK]** represent ranges of values. So for example, for the same event company, you might be interested in the **[CLICK]** total rainfall in a given week. You could have a random variable **[CLICK]** W represent the total centimeters of rain in a given week. In this case, **[CLICK]** W would add up seven individual daily rainfall measurements. Each day could have **[CLICK]** any amount of rainfall, including **[CLICK]** zero. So W could be any non-negative number, theoretically up to a **[CLICK]** very large amount if you had an extremely rainy week. W is considered a continuous random variable because **[CLICK]** rainfall isn't a set of distinct values, but a range of measurements.
	One way to think about why rainfall is continuous is that **[CLICK]** no matter how close two values are, there's always another value between them. For example, **[CLICK]** take two rainfall amounts: 1 and 2 centimeters. Those are pretty close together. But between those values you could have **[CLICK]** 1.5

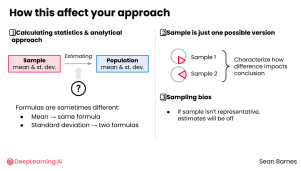
	centimeters. And between 1 and 1.5 centimeters you can have **[CLICK]** 1.25 centimeters. And between those two values, you can have **[CLICK]** 1.125, and so on. **[CLICK]** This process can continue indefinitely, and this is really what is meant by the term "continuous." In practical terms, your measurement precision might be limited by your tools, but in theory, **[CLICK]** a continuous random variable can take on any real number within its range.
	So to figure out if the random variable you're working with is discrete or continuous, what you can do is **[CLICK]** try to list out the possible values it can take on. And **[CLICK]** if you can actually **[CLICK]** list out those values, and **[CLICK]** there aren't numbers in between, **[CLICK]** then your random variable is discrete. Here's an example, here's a random variable **[CLICK]** S, representing the number of students in a given elementary school. Do you think S is a discrete or continuous random variable? **[pause for learner to think]** Well, let's try counting the values. So you could have **[CLICK]** 0 students, **[CLICK]** 300 students, maybe **[CLICK]** 1700, maybe **[CLICK]** 1701. But there couldn't be 1700 **[dramatic emphasis]** **[CLICK]** .5 students. So these are **[CLICK]** separate values you could actually list out, with **[CLICK]** no other numbers in between. That makes S a **[CLICK]** discrete random variable.
	Let's return to the random variable X, the number of days of rainfall in a given week. You saw previously that the values that X can take on are **[CLICK]** 0 1 all the way through 7. **[CLICK]** But what are the probabilities of these values? **[CLICK]** How likely is it in any given week that it rains 0 days or 6 days? That information about the probabilities is **[CLICK]** not contained within the random variable itself. This distinction is a common point of confusion. The random variable **[CLICK]** only contains the values that represent the outcomes, **[CLICK]** not how likely they are. You'll learn more about the probabilities associated with random variables in the next lesson.
TH	That sneak peek takes you to the end of this lesson on probability. You've defined it, learned its core rules, and learned to represent real world phenomena with random variables. Once you finish the practice assessment and practice lab for this lesson, I hope you'll join me in the next one, which is all about estimation and discrete probability distributions. I'll see you there.

Lesson 2 – Discrete probability distributions

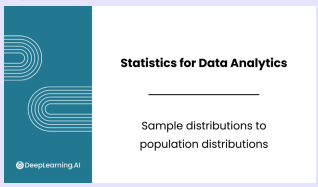
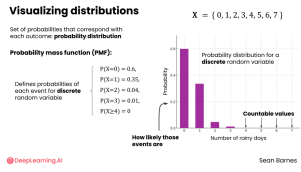
L2V1 – Estimation

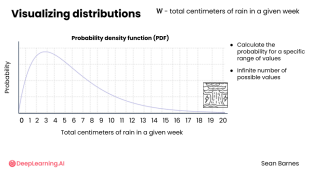
Visual	**Script**

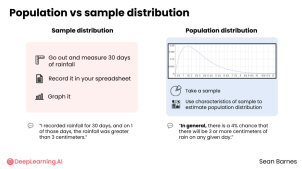
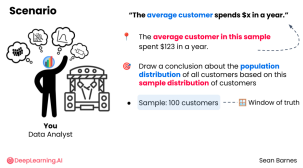
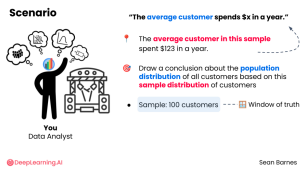
TH 	As you move forward in your statistics journey, it will become much more crucial for you to remember one key thing when working with your data: is this a population or a sample? As you learned in the previous module, it's more than likely going to be a sample. This matters because most statistics have different formulas and interpretations for samples and for the population.
	When you calculate a **[CLICK]** *statistic* from a sample, what you're really doing is **[CLICK]** *estimating* the true population value, which statisticians call a **[CLICK]** *parameter*. The estimate – the statistic – is **[CLICK]** an approximation for the truth – the parameter. You hope that your estimate is accurate, but it **[CLICK]** will probably not be exactly the same as the truth. For example, you might be trying to understand the behavior of **[CLICK]** all app users using just a **[CLICK]** subset of those users, or trying to understand healthcare outcomes for **[CLICK]** all patients at a hospital using just a **[CLICK]** fraction of them. Any statistics you calculate on these samples will be an estimate for the true population parameter.
	When might you be working with a population? Here's an example: if you need information about **[CLICK]** all 50 U.S. states, and you **[CLICK]** have data from all 50, that's a **[CLICK]** population. There's no guesswork about the other states that you did not include in your analysis. But it's usually more nuanced than that. Let's consider a thought experiment. Imagine **[CLICK]** measuring everyone's height on Earth. Put aside the logistics for a second, and say you did it! You **[CLICK]** measured everyone's heights! **[CLICK]** That's a population. BUT even if you could do it, by the time you finished, **[CLICK]** babies would have been born, **[CLICK]** people would have died, **[CLICK]** others would have grown taller. You might know exactly the average height at that moment, but it would be **[CLICK]** constantly changing. The beauty of statistics is that you **[CLICK]** really don't need to measure everyone. With a **[CLICK]** large enough sample, you can get remarkably close to the true value.
	It's the same with **[CLICK]** coin flips. Theoretically, a fair coin has a **[CLICK]** 50% probability of landing on heads, but you'd need to **[CLICK]** flip it an infinite number of times to prove this fact absolutely. In practice, a **[CLICK]** large sample gives you a **[CLICK]** close enough estimate to be useful.

	So how does this difference between populations and samples affect your approach as a data analyst? First, it has implications for the **[CLICK]** statistics you calculate and your analytical approach. When you calculate a **[CLICK]** mean or standard deviation from a sample, know that it's an **[CLICK]** estimate. There's some **[CLICK]** uncertainty about how well it represents the true population parameter. Note that the formulas you learned for mean and standard deviation in the previous module were all for samples. The **[CLICK]** formulas for populations are sometimes different. For example, the population mean and sample **[CLICK]** mean are calculated the same way. But the **[CLICK]** standard deviation has two different formulas. Second, when working with samples, it's important to remember that the data you're analyzing is **[CLICK]** just one possible version of that data. If you were to take **[CLICK]** another sample, you'd get slightly different results. Your job is to **[CLICK]** characterize how that difference impacts your conclusions. Finally, you must be aware of the kinds of **[CLICK]** bias that your sampling method can introduce. **[CLICK]** If your sample isn't representative of the population, your estimates will be off.
TH	As a data analyst, you're often playing detective. You're using the clues from your sample to piece together the bigger picture of the population. Coming up, you'll explore how sample distributions can help you estimate population distributions. **[eyebrows]** Join me in the next video to learn more!

L2V2 – From sample distributions to population distributions

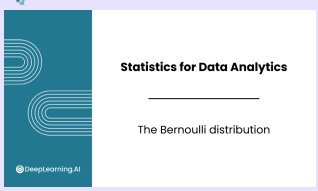
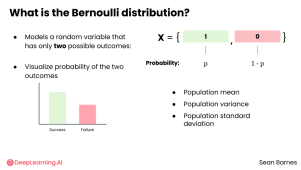
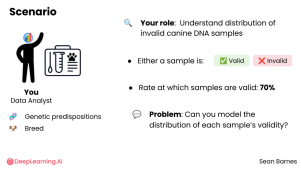
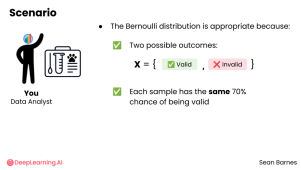
Visual	Script
TH 	In the previous module, you worked extensively with sample distributions, and learned that a distribution tells you how often different values occur in your sample data. But what's the next step? How can you use those sample distributions to draw conclusions about the broader population? Let's take a look, starting with random variables and probability distributions, and working our way to sample and population distributions.
	Say you're working with the random variable **[CLICK]** X from the previous lesson – rainy days per week. Recall that X could take on the values **[CLICK]** 0 all the way through 7. You want to visualize how common it is for a week to have each number of rainy days. **[CLICK]** The set of probabilities that correspond with each outcome in the random variable is called the probability

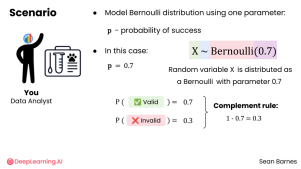
	distribution. If you were to visualize this probability, you would have all the possible values of your random variable, **[CLICK]** 0 through 7 on the x axis and the **[CLICK]** probability of each of these outcomes on the y axis. Say it's most common for there to be zero days of rain, so for the value 0, the probability is **[CLICK]** 0.6. For 1 it's **[CLICK]** 0.35, For 2 it's **[CLICK]** 0.04, for 3 it's **[CLICK]** 0.01, and maybe for 4 and above the probability is **[CLICK]** zero, those events never happen. This function, P of X equals 0 equals 0.6, P of X equals 1 is 0.35, and so on is called the **[CLICK]** probability mass function or PMF. It **[CLICK]** defines the probabilities of each event occurring for a discrete random variable. By the way, this chart looks familiar, right? It's just a column chart with probability on the y axis. This is one example of probability distribution for a discrete random variable. **[CLICK]** You have countable values on the x axis for your random variable X, and **[CLICK]** you're visualizing how likely those events are on the y axis.
	Here is the **[CLICK]** **continuous** random variable you saw in the previous lesson – **measuring** rainfall. You can **[CLICK]** visualize this one too. Add the possible values from the random variable W to the **[CLICK]** x axis, so starting with **[CLICK]** 0, since you can't have negative rainfall, and going up, let's just go up to **[CLICK]** 20 centimeters of rainfall, which is a **[CLICK]** decent amount of rain! Anything beyond that is increasingly unlikely. Can you guess what will be on the y axis? **[pause for learner to think]** ~~That would be roughly the **[CLICK]** probability of each event occurring. You can then **[CLICK]** plot the probabilities across the different rainfall amounts.~~ Analogous to the columns in the discrete case, you can plot a curve, where the points that are higher are more likely to occur. This curve is called the **[CLICK]** probability density function, or PDF. You can use the PDF to **[CLICK]** calculate the probability for a specific range of values. In this case, you can't have individual bars, because as you saw in the previous lesson, there are an **[CLICK]** infinite number of possible values. As a result, you actually use a curve to represent the PDF for a continuous random variable. ~~**[CLICK]** Here's what that curve could look like for different amounts of rainfall.~~
	Now that you've seen these two probability distributions, there's one key distinction here that gets to the heart of statistics. **[enunciate next sentence]** **[CLICK]** These are **distributions**, but these are **not sample** distributions. **[brief pause for effect]**

	You didn't **[CLICK]** go out and measure 30 days of rainfall and **[CLICK]** record it in your spreadsheet and **[CLICK]** graph it. **[CLICK]** This is a model for a population distribution of rainfall, how all days of rainfall behave in a mathematical way. Once **[CLICK]** you've taken a sample, **[CLICK]** you can use the characteristics of that sample distribution to estimate **[CLICK]** this population distribution. It's the difference between saying **[CLICK]** "I recorded rainfall for 30 days, and on 1 of those days, the rainfall was greater than 3 centimeters" and saying **[CLICK]** "In general, there is a 4% chance that there will be 3 or more centimeters of rain on any given day." The first statement is about a **[CLICK]** sample distribution. The second is about a population distribution. You took your sample data and drew a conclusion about the population. It's exciting!! This is the entire goal of statistics, this is really what you've been building up to throughout this course and the previous one.
	Let's look at another concrete example. Say you're working with the same outdoor events company from earlier videos, and they've asked you to **[CLICK]** characterize how much each of its customers spends on tickets in a year. You take a simple random **[CLICK]** sample of 100 customers and **[CLICK]** tally up their spending for the year. You're able to **[CLICK]** characterize your sample distribution with descriptive statistics, so you calculated that the **[CLICK]** mean amount spent is $123 dollars, **[CLICK]** with a standard deviation of $15.40, and **[CLICK]** the median amount spent is $100. Okay. **[pregnant pause]** What now?
	You can use this sample distribution as-is, to communicate insights like **[CLICK]** "The average customer in this sample spent $123 in a year." That's certainly useful, but what you really want to say is something like **[CLICK]** "The average customer spends $x in a year." Period. You want to draw a conclusion about the population distribution of all customers based on **[CLICK]** this sample distribution of customers. **[CLICK]** How does the population of customers behave overall when it comes to spending? **[CLICK]** Do the amounts cluster around some central point? If so, what is that central point? **[CLICK]** Is it a heavily skewed distribution? **[CLICK]** Is it highly variable, or are the amounts quite similar to each other? Remember that your **[CLICK]** sample of 100 customers is your **[CLICK]** window into the truth, and ultimately the point is to see the view, not to see the window. Keep in mind that you can only ever estimate the truth. Your window will always be at least a little dirty.
🎙️ TH	What's interesting about probability distributions is that both discrete and continuous populations often follow distributions with known behavior. Follow

	me to the next video to take a look at one discrete probability distribution.

L2V3 – The Bernoulli distribution

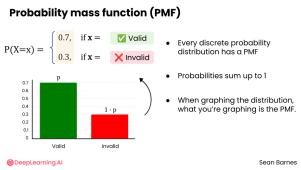
Visual	Script
TH  	In data analytics, your goal is often to predict the behavior of a population of interest. When you suspect your population follows certain patterns, you can use a known probability distribution to model it.
 	The Bernoulli distribution models a **[CLICK]** random variable that has only two possible outcomes: **[CLICK]** success (usually written as 1) with **[CLICK]** probability p, and **[CLICK]** failure (usually written as 0) with **[CLICK]** probability 1 minus p. Notice the complement rule in the definition of the probabilities for the two outcomes. Similar to any other discrete probability distribution, **[CLICK]** you can visualize the probability of the two outcomes using **[CLICK]** a bar or column chart, as well as calculate the **[CLICK]** population mean, **[CLICK]** population variance, and **[CLICK]** population standard deviation.
 	Imagine you're working with **[CLICK]** canine DNA samples collected using home test kits. Companies offer these canine DNA kits to identify **[CLICK]** genetic predispositions or a **[CLICK]** dog's breed. You're working with the test lab to **[CLICK]** understand the distribution of invalid canine DNA samples in the kits. **[CLICK]** Either a sample is **[CLICK]** valid or **[CLICK]** invalid – usually because it was collected incorrectly by the owner. And your partners at the lab tell you that **[CLICK]** the rate at which samples are valid is 70%. The lab is asking you, **[CLICK]** can you model the distribution of each sample's validity? Modeling the probability distribution of valid and invalid samples allows the lab to set realistic expectations for its testing process.
 	You can model this distribution using the Bernoulli distribution. The Bernoulli distribution is **[CLICK]** appropriate for this scenario because • First, **[CLICK]** you have two possible outcomes; each sample is **[CLICK]** either **[CLICK]** valid (a success) or **[CLICK]** invalid (a failure). • And second, **[CLICK]** each sample has the same 70% chance of being valid.



You can model your data using the Bernoulli distribution using only one key parameter: **[CLICK]** p, the probability of success. **[CLICK]** In this case, the probability of success represents the probability of getting a valid sample, and **[CLICK]** is equal to 0.7.

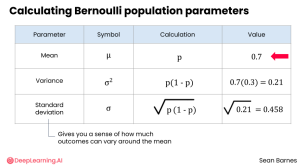
To express this distribution, you'll commonly see **[CLICK]** notation like this: some **[CLICK]** random variable, say X, **[CLICK]** is distributed as **[CLICK]** a Bernoulli **[CLICK]** with parameter 0.7. More generally for any random variable, you'll see the tilde for "is distributed as" followed by the distribution, then the parameters for that distribution. Bernoulli only has one parameter.

Let's look at the probabilities of each outcome in this distribution: **[CLICK]** P of valid is **[CLICK]** 0.7 and, can you guess P of invalid? **[pause for thought]** P of invalid is **[CLICK]** 0.3, **[CLICK]** using the complement rule.



Recall that this set of outcomes and their probabilities is called the probability mass function or PMF. It's not unique to the Bernoulli distribution. **[CLICK]** Every discrete probability distribution has a probability mass function. Notice that the **[CLICK]** probabilities across all of the outcomes sum up to 1.

Can you imagine what a column chart of this distribution would look like? **[pause for thought]** **[CLICK]** Here's a column chart of this Bernoulli distribution. This is a graph of the two outcomes, **[CLICK]** valid (or success) and **[CLICK]** invalid (or failure), with probabilities **[CLICK]** p and **[CLICK]** 1 minus p, respectively. Notice that **[CLICK]** when graphing the distribution, what you're actually graphing is the PMF.



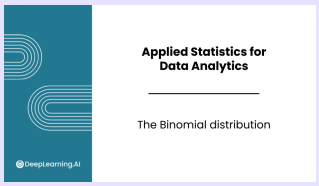
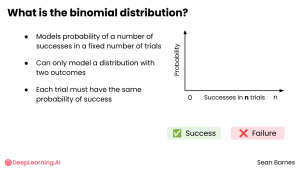
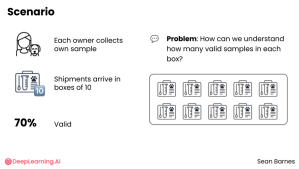
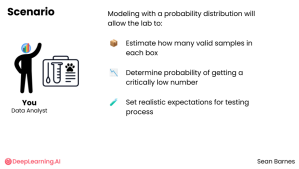
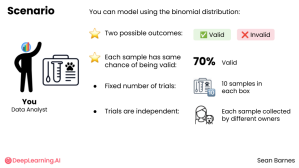
You can also calculate mean, variance, and standard deviation for this distribution, which are all population parameters and have different notation than the sample statistics you calculated previously:

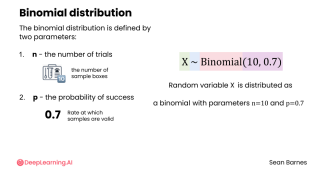
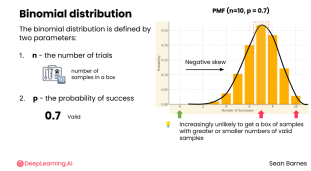
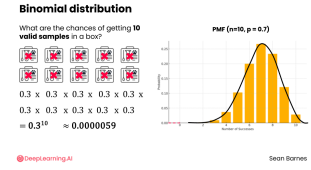
- The mean of the Bernoulli distribution – represented by this symbol **[CLICK]** μ – is equal to **[CLICK]** *p*, the probability of success. Intuitively, thinking about the canine DNA sample validity, if the chance of a sample being valid is **[CLICK]** 0.7, then for many samples over time, the probability of each sample being valid is the same.
- You can also calculate the variance of a Bernoulli distribution ~~denoted by~~ **[CLICK]** σ^2–which is **[CLICK]** p times 1 minus p, or in this case **[CLICK]** 0.7 times 0.3, or 0.21.
- Finally, you previously learned that the standard deviation–represented by **[CLICK]** σ–is the **[CLICK]** square root of the variance, so here **[CLICK]** σ is about 0.458. The standard deviation **[CLICK]** gives you a sense of how much outcomes can vary around the mean. In this case, the outcomes are either 0 or 1, so it makes sense that there is some variation around **[CLICK]** the mean of 0.7.

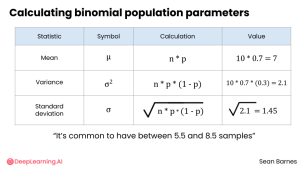
To recap, mu, sigma squared, and sigma are population parameters for this

	distribution, while x-bar, s squared, and s are only for sample distributions.
TH	What's cool about the Bernoulli distribution is that it can be extended into the binomial distribution, which models multiple trials. So, for example, out of 10 random DNA test kits, what is the probability that all 10 are valid samples? Follow me to the next video to take a look.

L2V4 – The Binomial distribution

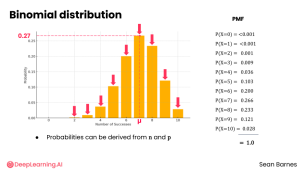
Visual	Script
TH 	In the previous video, you modeled the distribution of sample validity for a single canine DNA test kit. Now, you'll model the distribution of validity for many samples using the binomial distribution.
	The binomial distribution models **[CLICK]** the probability of a **[CLICK]** specific number of successes in a fixed number of independent trials. **[CLICK]** It can only model a distribution with exactly two outcomes, like **[CLICK]** success and **[CLICK]** failure. Bi means "two", after all. And **[CLICK]** each trial must have the same probability of success. The binomial distribution is a discrete probability distribution, and the number of successes can range from **[CLICK]** zero to the **[CLICK]** number of trials.
	Say your partners at the canine DNA lab tell you that **[CLICK]** each individual owner collects their own sample, but **[CLICK]** shipments arrive in boxes of 10. These are the same samples that you previously learned are **[CLICK]** valid at a rate of 70%. The lab is asking you, **[CLICK]** how can we understand how many valid samples are in each box?
	Modeling the DNA test kit scenario with a probability distribution will allow the lab to: • **[CLICK]** Estimate how many valid samples they're likely to get in each box on average • **[CLICK]** Determine the probability of getting a critically low number of valid samples • **[CLICK]** And set realistic expectations for their testing process
	You can **[CLICK]** model this distribution using the binomial distribution. Here's why: • First, you have **[CLICK]** two possible outcomes; each sample is either **[CLICK]** valid (a success) or **[CLICK]** invalid (a failure). • **[CLICK]** Each sample has the same **[CLICK]** 70% chance of being valid. • You have a **[CLICK]** fixed number of trials, since there are **[CLICK]** 10 samples in each box.

	• And finally, you assume the **[CLICK]** trials are independent because the validity of one sample doesn't appear to affect the others. They are **[CLICK]** each collected by different pet owners. Note that independence is just an assumption, and you can't be 100% certain the samples are truly independent. Do these conditions look familiar? **[pause for thought]** Two of them match the Bernoulli distribution – **[CLICK]** having two possible outcomes, and **[CLICK]** having a fixed chance for success. That's because the Bernoulli distribution is a special case of the binomial distribution where there is only one trial.
 https://i.imgur.com/JWMYd8G.png	The binomial distribution is defined by two parameters: • **[CLICK]** n, the number of trials (in this case, **[CLICK]** the number of samples in a box) • And **[CLICK]** p, the same probability of success from the Bernoulli distribution, **[CLICK]** 0.7 You can write this distribution as **[CLICK]** X **[CLICK]** is distributed as **[CLICK]** a binomial with **[CLICK]** parameters n equals 10 and p equals 0.7.
	These two parameters define the shape of the distribution. Let's visualize the **[CLICK]** probability mass function, or PMF. **[CLICK]** Here's what it looks like for n=10 and p=0.7. Each possible outcome is on the **[CLICK]** x axis, from **[CLICK]** 0 valid samples to **[CLICK]** 10, and on the **[CLICK]** y axis is the probability of any given box containing that number of valid samples. The distribution is **[CLICK]** centered around 7 and forms a **[CLICK]** roughly symmetrical, bell shaped distribution around that center. There is some **[CLICK]** negative skew to the distribution, because you can't get values higher than 10. The rough bell-shaped distribution reflects the idea that **[CLICK]** it's increasingly more unlikely to get a box of samples with increasingly greater or smaller numbers of valid samples.
	What do you think is the chance of getting all 10 valid samples in a box? **[pause for thought]** You can use the multiplication rule to calculate the probability of **[CLICK]** 0.7 times itself 10 times or **[CLICK]** 0.7 raised to the 10, which judging by the **[CLICK]** graph is around **[CLICK]** 0.025.
	What about getting 0 valid samples? **[pause for thought]** Similarly, zero valid samples is **[CLICK]** 0.3 times itself 10 times, or **[CLICK]** 0.3 to the 10, which is a **[CLICK]** vanishingly small number. For cases other than 0 and 10 valid samples, the probability becomes much more complicated since you have to account for all the different ways samples can be configured in a box.



As with any other probability distribution, you can calculate various population parameters such as the mean, variance, and standard deviation:

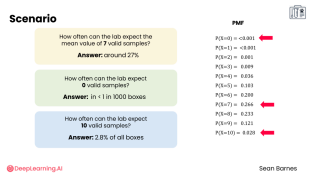
- First, the mean. In the binomial distribution, the mean is **[CLICK]** n * p. In this case, **[CLICK]** 10 samples per box times 0.7, equals 7. You expect 7 valid samples on average.
- Then, the variance, which is calculated as **[CLICK]** n times p times (1 minus p). The variance measures the **spread** of your distribution, in this case you multiply **[CLICK]** 10 times 0.7, so 7 which is the same as the mean, then multiply by 1 - 0.7 or 0.3, which gives you 2.1.
- The standard deviation is calculated from the variance in the same manner as you've seen before, **[CLICK]** just by taking the square root, giving you the **[CLICK]** square root of 2.1, and a standard deviation of 1.45. Since this standard deviation is in the same units as our data (in this case, samples), and the distribution is roughly symmetrical, you can say with relative confidence that **[CLICK]** it's common to have between 5.5 and 8.5 valid samples, or within one standard deviation of the mean.



The binomial distribution also allows you to calculate the probability that a specific number of samples in the box is valid or invalid, which corresponds to **[CLICK]** each of these bars. Approximating from the graph, you can see that **[CLICK]** the mean of 7 is most common at a probability of **[CLICK]** around 0.27.

[CLICK] Here is the full probability mass function for this binomial distribution: **[show values on screen]**. If you add up all these values, they **[CLICK]** sum to 1.

For the binomial distribution, you'll never need to calculate these values by hand, since computers make the process much easier, but just know that all **[CLICK]** these probabilities can be derived just from n and p.



Let's use these probabilities to analyze the DNA testing outcomes. I encourage you to pause the video and try answering the questions.

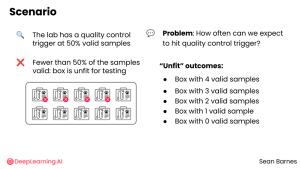
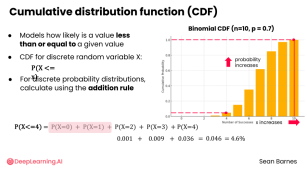
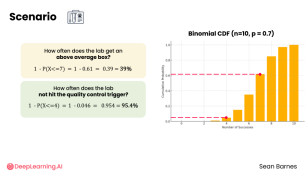
- First, **[CLICK]** how often can the lab expect the mean value of 7 valid samples? **[pause for thought]** **[CLICK]** That would be around 27% of the time.
- **[CLICK]** What about zero valid samples? **[pause for thought]** That outcome is very rare, **[CLICK]** occurring in less than 1 in 1000 boxes of samples.
- **[CLICK]** Having all 10 valid samples is a relatively rare outcome, but not as rare as 0 valid, **[CLICK]** occurring in about 2.8% of all boxes.

TH

You can answer so many interesting questions using the binomial distribution. In the next video, you'll see how to answer even more using the binomial distribution's cumulative distribution function. Follow me to the next video to

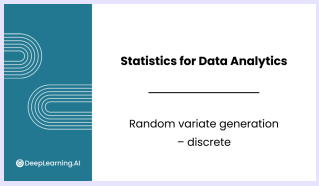
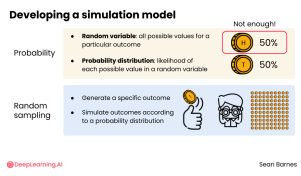
	learn more.

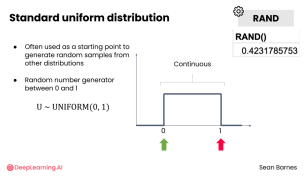
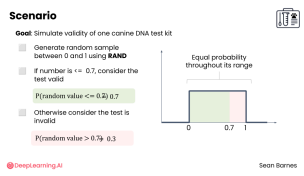
L2V5 – The cumulative distribution function

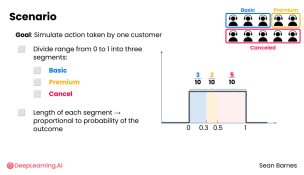
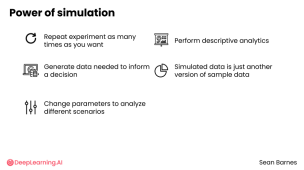
Visual	Script
🎙️ TH	In some cases, you're interested in the probability of a range of outcomes, not just the probability of a single outcome. To calculate these probabilities, you'll need a new tool.
	Imagine that the lab from the previous video has a **[CLICK]** quality control trigger at 50% valid samples. **[CLICK]** If fewer than 50% of the samples are valid, then the box is deemed unfit for testing and customers are reissued a new kit. Your friends at the lab might come to you and say, **[CLICK]** how often can we expect to hit that quality control trigger? In this case, **[CLICK]** you're not just interested in the outcome with 4 valid samples. A **[CLICK]** box with 3 valid samples would also be unfit for testing, as would boxes with **[CLICK]** 2, **[CLICK]** 1, or **[CLICK]** 0 valid samples.
 https://i.imgur.com/M6vA8xj.png	To answer this question, you can use the cumulative distribution function or CDF of the binomial distribution. The CDF **[CLICK]** models how likely it is for a random variable to take on a value less than or equal to a given value. Formally, the **[CLICK]** CDF for a discrete random variable X is defined as **[CLICK]** P of X less than or equal to x. Here's what the **[CLICK]** CDF looks like for the binomial distribution with n = 10 and p = 0.7. It has the same axes as the PMF. You can see that as **[CLICK]** x increases, the **[CLICK]** probability only goes up, ending in **[CLICK]** x = 10 or less with a **[CLICK]** probability of 1. The CDF makes it clear that the probability of getting a box with **[CLICK]** 4 or fewer valid samples is less than 5%. **[CLICK]** For discrete probability distributions, you can calculate the CDF using the addition rule. **[CLICK]** For the probability of getting 4 or fewer valid samples in a box, add the probabilities of each outcome occurring. **[CLICK]** Ignoring the probabilities of 0 and 1 valid samples, which are extremely rare, you can add up P(X=2) which is **[CLICK]** 0.001, P(X=3) at **[CLICK]** 0.009, and P(X=4) which is **[CLICK]** 0.036 and you get about **[CLICK]** 0.046, about 4.6% of the time.
	By summing probabilities, you can also answer questions like, **[CLICK]** how often does the lab get an above average box? What do you think? Pause the video and see if you can estimate it. **[pause for thought]** In that case, you can use the complement rule. **[CLICK]** 1 minus P of X less than or equal to 7, which is around **[CLICK]** 0.61, so you might estimate about **[CLICK]** 0.39, or 39% of the time.

	And similarly, you can apply the complement rule for the question, **[CLICK]** how often does the lab **not** hit the quality control trigger? Pause the video and see if you can calculate it. **[pause for thought]** The answer is **[CLICK]** 1 minus the chance of hitting it, so **[CLICK]** 1 minus 0.046, or **[CLICK]** 0.954. ~~How often do you get average or below boxes? 1 minus 0.382, or 0.618.~~ Since the DNA sample boxes fit the conditions for modeling with the binomial distribution, **[CLICK]** you were able to answer a lot of useful questions, like **[CLICK]** where is the center of this distribution, **[CLICK]** what is the variability, and **[CLICK]** how common are different outcomes or ranges of outcomes?
	Beyond DNA testing kits, in data analytics, the binomial distribution is incredibly useful for **[CLICK]** modeling other scenarios with **[CLICK]** yes/no outcomes, or **[CLICK]** outcomes with success and failure conditions like • **[CLICK]** Customer conversion rates in marketing campaigns • **[CLICK]** Defective product rates in quality control • **[CLICK]** Employee retention in HR analytics • And **[CLICK]** loan default rates in financial services
TH	Both the Bernoulli and binomial distributions are useful discrete probability distributions. How can you determine the many possible outcomes of sampling from these distributions? You'll find out in the next video.

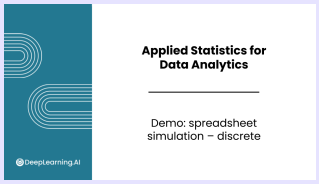
L2V6 – Random sampling – discrete

Visual	Script
TH 	Just like you can sample from a population in the real world, you can also sample from a known probability distribution to generate simulated data. You can then analyze that simulated data to inform business decisions. This process is known as random sampling.
	At the end of the first lesson, you learned the concept of a **[click]** random variable, which represents all of the possible values for a particular outcome of interest. In this lesson, you extended the idea of random variables into **[click]** probability distributions, which represent the **[click]** likelihood of each possible value for a given random variable. In order to develop a simulation model, you need to **[click]** generate a specific outcome that represents the real-world behavior that you are trying to model. For instance, it's not enough to **[click]** know that there's a 50% chance of flipping a coin and landing on heads. You want to **[click]** simulate that coin flip

	and **[click]** observe the actual outcome. Was it heads, or was it tails?And if you simulated a coin flip **[click]** 100 times, how many heads did you get? Without the individual outcomes, you can only analyze what *might* happen, not what *actually happens* in the simulation. ~~It's the difference between imagining that a cup of tea will be delicious and actually taking a sip.~~ **[click]** *Random sampling* allows you to **[click]** simulate outcomes that behave according to a probability distribution. So if you know that a given outcome follows a binomial distribution, you can simulate outcomes according to the given probability of success p.
	The standard uniform distribution is **[click]** often used as a starting point to generate random samples from other distributions. The standard uniform distribution is basically a fancy term for a **[click]** random number generator between 0 and 1. More formally, the standard uniform distribution is **[click]** continuous and it has a minimum of **[click]** 0 and a maximum of **[click]** 1, inclusive. You can represent a uniform random variable as **[click]** U ~ UNIFORM(0, 1). It even has its own function in spreadsheets: **[click]** RAND.
🖥️SC 🔗**Link to final screencast – can we overlay this onto the previous slide while Sean is talking?**	Here's an example of the RAND function in action. You can see that it generates values between 0 and 1. This function simulates one outcome from the standard uniform distribution.
	The standard uniform distribution can be used to generate random samples for many other distributions. For example, to **[CLICK]** simulate the validity of one canine DNA test kit, you could • First **[CLICK]** generate a random sample between 0 and 1 using RAND • Then, **[CLICK]** if the number is less than or equal to 0.7, consider the test valid. • **[CLICK]** Otherwise, the number is greater than 0.7, and the test is invalid. This method works because the uniform distribution has **[CLICK]** equal probability throughout its range of potential values, so the **[CLICK]** probability that a generated random value is less than or equal to 0.7 is …. **[CLICK]** 0.7, which is exactly the probability for a valid test. And similarly, the **[CLICK]** probability that the value is greater than 0.7 is the same as the probability of an invalid test, **[CLICK]** 0.3.

	You can extend this principle to simulate more complex discrete scenarios. For instance, think back to the **[CLICK]** music subscription service that offered its customers a free trial. Recall that customers could either subscribe to the basic plan, upgrade to premium, or cancel their subscription at the end of the trial. To **[CLICK]** simulate the action taken by one customer, you could **[CLICK]** divide the range from 0 to 1 into three segments, each representing one of the three decisions made by the customers: **[CLICK]** basic, **[CLICK]** premium, or **[CLICK]** cancel. The key to making this work is that the **[CLICK]** length of each segment needs to be proportional to the probability of the associated outcome. So the length of the basic segment should be **[CLICK]** 3/10 of the range, the premium segment should be **[CLICK]** 2/10 of the range, and the cancel segment should be the remaining **[CLICK]** 5/10 of the range. Specifically, you could draw from the standard uniform distribution, and if the value is **[CLICK]** less than or equal to 0.3, then you simulate a basic subscription. If the value was between 0.3 and **[CLICK]** 0.5 (inclusive), then you simulate a premium subscription. Otherwise, if the value was **[CLICK]** greater than 0.5, you simulate a cancellation.
	The power in simulation is that you can **[CLICK]** repeat, or replicate, the individual experiment as many times as you want! You can't say the same for the real world, where collecting samples is often costly and time-consuming. This capability to replicate provides you as a data analyst with the opportunity to **[CLICK]** generate the data needed to inform a decision. You can also **[CLICK]** change the simulation's parameters in order to analyze different scenarios. Once you have generated your desired sample size, you can then **[CLICK]** perform any of the descriptive analytics that you learned so far. Your **[CLICK]** simulated data becomes just another version of sample data that you can analyze using all of these techniques.
TH	Let's see this in action. Join me in the next video to create Bernoulli and binomial simulations in a spreadsheet.

L2V7 – Demo: Spreadsheet simulation – discrete

Visual	Script
TH 	Let's see how you can actually run simulations in a spreadsheet using random sampling. There's a lot you can do, and as you get more familiar with tools like Python, you will be able to simulate more and more complex scenarios as well.

🖥️SC

🔗**Start**
🔗**Solution**

Let's start off with the DNA test kit scenario. If you test a sample's validity, there's a chance that it's destroyed in the process, which is inconvenient and expensive. Since you know the probability of validity, you can instead simulate the outcomes of many boxes mathematically to help your friends at the lab understand the likelihood of hitting their quality control trigger. You are developing a simulation model for this scenario.

Remember, if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab.

DNA test kit simulation – boxes of 10 kits

- Column A
  - Row 2: =RAND()
- Two conditions
  - Success & failure
  - Valid test kit, invalid test kit
- Column B
  - Row 2: `=if(A2<=0.7,"✅ valid","❌ invalid")`
  - Drag the cells until 11th row
- Refreshing
  - Any function without arguments will recalculate on any change to the spreadsheet, `rand` is one of these
  - D2: Insert → checkbox
  - Checking/unchecking will refresh
- Let's see the distribution
  - Select Column B → Insert Chart → Bar chart → Tick Aggregate box → Count
  - Customize → Vertical axis → Min 0 / Max 10
    - Keeps same axis height while refreshing

Recall the music subscription service that offered its users a free trial where users could continue with basic, upgrade to premium, or cancel. Here's one way you could simulate outcomes using a spreadsheet.

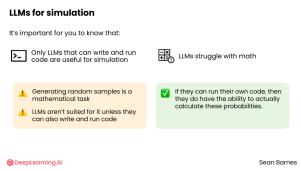
Subscriber simulation

- New sheet
- Column A
  - Row 2: =RAND()
- Three conditions
  - If <= 0.3 that's ⚫ basic, then if <=0.5, 💎 it's premium, otherwise it's a ⛔ cancellation
- Column B
  - Row 2: ~~`=if(A2<=0.3,"⚫ basic",if(A2<=0.5,"💎 premium","⛔ cancellation"))`~~
- If you want to simulate 10 people with a free trial?
  - Drag the cells until 11th row

	• Refreshing
	  ○ D1 header: Refresh button
	  ○ D2: Insert → checkbox
	  ○ Checking/unchecking will refresh
	• Let's see the distribution
	  ○ Select Column B → Insert Chart → Bar chart → Tick Aggregate box → Count
	• Too few samples, still not half and half → Simulate 100 more coin flips
	  ○ Two ways:
	  ■ keep dragging cells (tedious!)
	  ■ Use RANDARRAY function together with ARRAYFORMULA
	  • Col A: delete any extra cells below A2 to avoid overwrite error
	  • A2: `=RANDARRAY(100)`
	  • Col B: `=arrayformula(if(A2:A101<=0.3,"⚫ basic",if(A2:A101<=0.5,"💎 premium","⛔ cancellation")))`
	• Update the chart range. This is much better!

🎙️ TH	Great work simulating those outcomes! You can see how it's much more convenient than going out and testing dozens of boxes in real life. In the next video, you'll see how to use a large language model for simulation via random sampling. I'll see you there.

L2V8 – Demo: LLM simulation – discrete

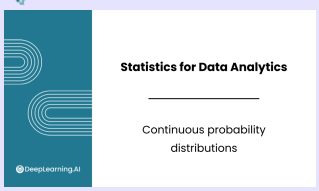
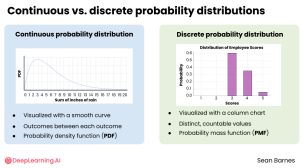
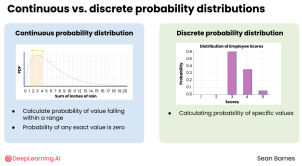
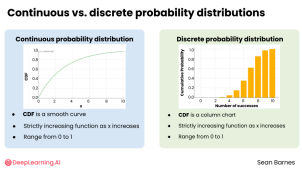
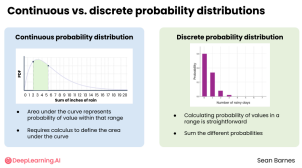
Visual	Script
🎙️ TH	So how can you use large language models for sampling? They're an incredibly powerful tool with some key limitations.
	Before getting into using LLMs for simulation, it's important for you to know that **[CLICK]** only LLMs that can write and run code are useful for simulation. You'll learn more about the nature of this limitation and see it in action in the next lesson. In short, remember that **[CLICK]** LLMs struggle with math. Because **[CLICK]** generating random samples is a mathematical task, **[CLICK]** LLMs aren't suited for it unless they can also write and run code. **[CLICK]** If they can run their own code, then they do have the ability to actually calculate these probabilities.
🖥️ SC 🔗 **Claude**	Okay, so let's take this over to Claude, which is capable of running some really cool simulations. Claude has this feature called Artifacts that allows it to write and run code. Let's see what it can do. To start your simulation, you can give it this prompt, Create an interface for sampling boxes of 10 canine DNA test kits. Tell it that the probability that any given kit is valid is 70%.

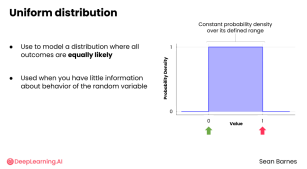
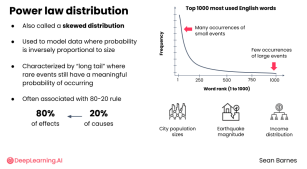
	PROMPT: *Create an interface for sampling boxes of 10 canine DNA test kits. The probability that any given kit is valid is 70%. In the interface, when I click a button, have a visual of the 10 samples appear, with valid samples in green and invalid samples in red. In the top right corner, keep track of the values of valid and invalid kits sampled so far, as well as the percent of all kits that have been valid up to this point.* And then tell it what kind of interface you want. In the interface, when you click a button, have a visual of the 10 samples appear, with valid samples in green and invalid samples in red. In the top right corner, keep track of the values of valid and invalid kits sampled so far, as well as the percent of all kits that have been valid up to this point. You're going to see it generating all this code which essentially creates a website inside a website. The fact that Claude is writing code tells you that it does have a way to actually sample randomly from a distribution compared to an LLM without this ability, which can only guess what word comes next with no guarantee that the guess is truly random. If you try generating a new sample, it does exactly what you asked it to do. It has this image of different samples in red and green. And it shows you how many valid kits there were, and how many invalid kits, and the percent of valid samples. So 80% valid, you and I know that the true rate is 70%, so you can just keep going. Same sample again. Now you have another sample with five invalid kits. And you can keep going and just generate a ton of new samples. And you'll see that that valid percentage hovers right around 70 percent. This kind of interface can help you visualize what these different scenarios would look like. You can see that it's relatively common to get kits with many invalid tests.
🎙️ TH	Great work creating a simulation with the large language model. That takes you to the end of this lesson, and there's only one more to go! Once you've completed the practice assessment and practice lab for this lesson, follow me to the next one to learn about continuous probability distributions.

Lesson 3 – Continuous probability distributions

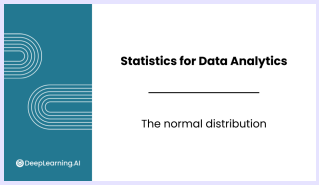
L3V1 – Continuous probability distributions

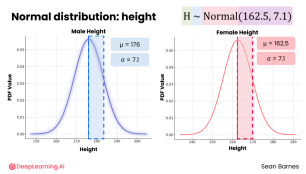
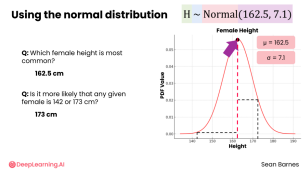
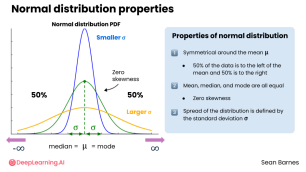
Visual	Script

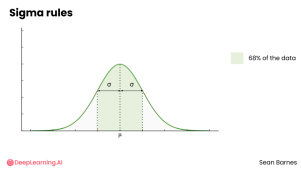
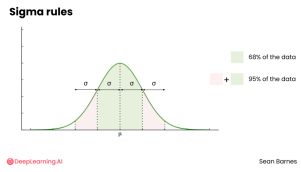
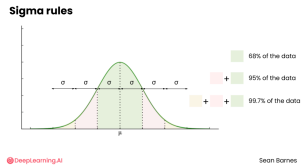
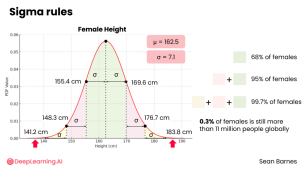
TH 	Similar to the Bernoulli and binomial distributions you saw in the previous lesson, there are known probability distributions for continuous random variables as well. The most important continuous probability distribution in data analytics is the normal distribution, and you'll also learn more about the uniform and power law distributions.
	The major differences between continuous and discrete probability distributions are related to how the probabilities of different outcomes are represented and how the statistics are calculated. • First, you saw previously that continuous distributions don't have countable values. Because of this difference, continuous distributions are typically **[CLICK]** visualized with a smooth curve rather than **[CLICK]** a column chart. The smooth curve reflects that there are **[CLICK]** outcomes in between each outcome, whereas the column chart for discrete distributions shows that there are **[CLICK]** distinct, countable values. This curve is called **[CLICK]** a probability density function or PDF, which is analogous to the **[CLICK]** probability mass function or PMF for a discrete probability distribution.
	• Rather than **[CLICK]** calculating the probability of specific values like you would in a discrete probability distribution, for a continuous distribution, you **[CLICK]** calculate the probability of a value falling within a range. Since there are infinitely many values in between each value for a continuous probability distribution, **[CLICK]** the probability of any exact value is zero.
	• The cumulative distribution function or CDF for a continuous probability distribution is also a **[CLICK]** smooth curve, compared with the **[CLICK]** column chart you saw for a discrete CDF. However, the two are similar in that they are both **[CLICK]** strictly increasing functions as x increases, and **[CLICK]** they range from 0 to 1.
	• Because the PDF is a smooth curve, **[CLICK]** the area under that curve between two points represents the probability of a value falling within that range. **[CLICK]** Calculating the probability of values falling with a range was straightforward with a discrete probability distribution, since you can just **[CLICK]** sum the different probabilities. For a continuous random variable, the probability is more complex to calculate, **[CLICK]** requiring calculus to define the area under the curve. You won't see the calculus here, since you will rarely if ever need to calculate these values by hand. However you will learn the intuition for these calculations in a moment.

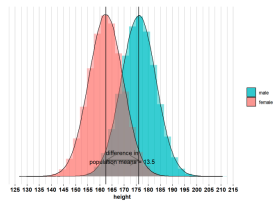
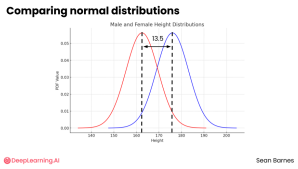
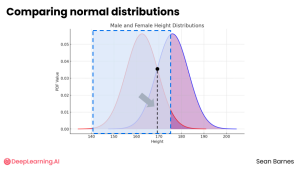
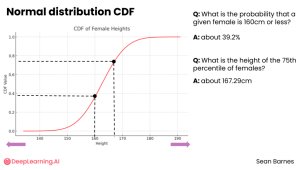
	Let's start with the uniform distribution, which you can **[CLICK]** use to model a distribution where all outcomes within a specified range are equally likely to occur. You used the uniform distribution in the previous lesson for generating random samples. It has a **[CLICK]** constant probability density over its defined range. The uniform distribution is **[CLICK]** often used when you have little information about the behavior of the random variable other than an estimated **[CLICK]** minimum or **[CLICK]** maximum value. Otherwise, it's useful for simulation, as you saw earlier.
	In most scenarios, the distribution of outcomes is not uniform. For example, take a look at **[CLICK]** this distribution of the frequencies of the top 1000 most used words in English. This is a fairly typical power law distribution, sometimes **[CLICK]** also called a skewed distribution. This distribution can be **[CLICK]** used to model data where the probability of an outcome is inversely proportional to its size. In other words, there are **[CLICK]** many occurrences of small events and **[CLICK]** few occurrences of large events. The power law distribution is **[CLICK]** characterized by its "long tail," where rare events still have a meaningful probability of occurring. It's quite common in nature, and can be **used to model** not just word frequencies, but **[CLICK]** city population sizes, **[CLICK]** earthquake magnitudes, **[CLICK]** income distribution, and more. The power law is **[CLICK]** often associated with the 80-20 rule, where **[CLICK]** 80% of effects come from **[CLICK]** 20% of causes. For example, a small number of words are used very frequently, while most words are used rarely.
TH	Great work examining those distributions. Follow me to the next video to learn about the most important continuous distribution in data analytics: the normal distribution.

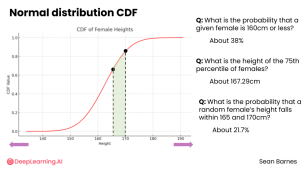
L3V2 – The normal distribution

Visual	Script
TH  https://miro.medium.com/v2/resize:fit:1131/1*SD4MtJcsheHu5uqH	It turns out that many real world phenomena follow a distribution that looks roughly like this **[normal distribution appears on screen]** – the values are clustered symmetrically around the mean. Outcomes on both sides of the mean become increasingly more rare. This is the normal distribution.

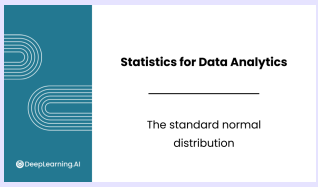
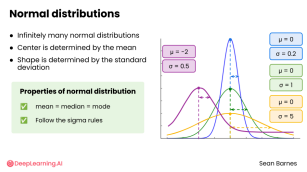
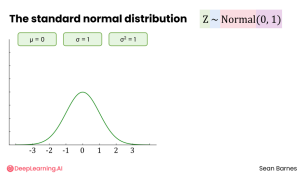
[10eyjw.png](#)	
	One real-world phenomenon that follows the normal distribution is human height. Here are two curves, both graphs of the normal distribution probability density function. The one on the left shows **[click]** male heights. On the x axis you have the height and on the y axis, the probability density function value. The male height distribution has a **[click]** mu of 176 centimeters, which is about 5 foot 9 inches for all you Imperial users out there, and a **[click]** sigma of 7.1 centimeters, about 2 and a half inches. Notice that these are population parameters, not sample statistics. Now take a look at the **[click]** female heights, same axes, with a **[click]** mu of 162.5 centimeters, about 5 foot 4 inches, and the **[click]** same standard deviation as the distribution for male heights. The distribution for female heights, just as an example, can be written as **[click]** H, for heights, **[click]** is distributed as **[click]** Normal with **[click]** parameters mu equals 162.5 and sigma equals 7.1. Notice how they're symmetrical **[click]** **[click]** and centered around their respective means. Heights, blood pressure readings, and test scores are often normally distributed. Beyond natural phenomena, though, many statistical methods assume that data follows a normal distribution.
	The normal distribution can also help you answer questions like: • **[click]** Which female height is most common? Can you guess? **[pause for thought]** That would be the mode, which in the case of a normal distribution also happens to equal the mean and median at **[click]** 162.5 cm. • **[click]** Is it more likely that any given female is 142 or 173cm? 142cm is **[click]** 20.5cm away from the mean, while 173cm is **[click]** 10.5cm away from the mean, so **[click]** 173cm is more likely as it is closer to the mean.
	Let's break down the key features of a normal distribution: • First, it's **[click]** symmetrical around the mean (μ). The left side is a mirror image of the right side. **[click]** 50% of the data is to the **[click]** left of the mean and 50% is to the **[click]** right. • Second, **[click]** the mean, median, and mode are all equal and located at the center of the distribution, which means there is **[click]** zero skewness. • Finally, the **[click]** spread of the distribution is defined by the standard deviation sigma. The tails of the distribution technically extend out to

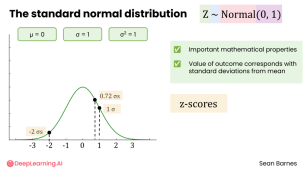
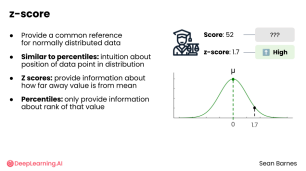
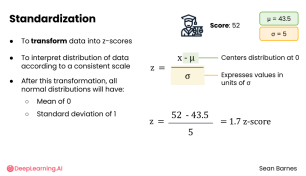
	[click] positive and **[click]** negative infinity on either side. The mean determines where the peak of the curve is located, while the standard deviation determines how spread out the data is. A larger standard deviation means a flatter, more spread-out curve like **[click]** this one, while a smaller standard deviation results in a taller, narrower curve, like **[click]** this one. So while this bell-shaped curve may seem more “normal” than this flatter curve, these are all normal distributions ~~because they meet the three criteria listed here.~~ This curve you’ve been looking at is the **[click]** probability density function or PDF of the normal distribution. As a reminder, for continuous distributions, the height of the curve in the PDF shows you the relative likelihood of any given range of values.
	One useful property of the normal distribution is the set of sigma rules, sometimes called the empirical rule or the 68-95-99.7 rule. This rule tells you that • About 68% of the data falls within one standard deviation of the mean. This is the one sigma rule
	• About 95% falls within two standard deviations – the two sigma rule.
	• And about 99.7% falls within three standard deviations – the three sigma rule. Together, these are called the sigma rules
	This rule can help you understand probabilities in a normal distribution. Here is the distribution for female heights again. The **[click]** mu is 162.5 cm and the **[click]** sigma is 7.1 cm. The height one sigma **[click]** below the mean is **[click]** 155.4 centimeters, while the height one sigma **[click]** above the mean is **[click]** 169.6 cm. So using the one sigma rule, you know that **[click]** 68% of females have a height between those two values. **[click]** Two standard deviations below the mean is **[click]** 148.3 cm, while two above is **[click]** 176.7 cm, so, by the two sigma rule, you know that **[click]** 95% of females are between those two heights. And according to the three sigma rule **[click]** 99.7% of females will be **[click]** between **[click]** 141.2 and **[click]** 183.8 centimeters tall, so between about 4 foot 8 and slightly above 6 feet. That said, **[click]** 0.3% of females is still more than 11 million people globally. While heights above or below these values are rare, millions of people will fall

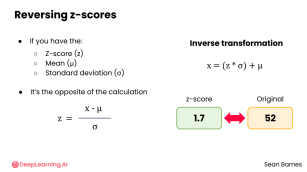
	within **[click]** these tails of more unusual values.
 Don't use this image specifically, but it should look roughly like this 	You can also use the PDF to compare between two distributions. For example, here are the distributions of male and female heights plotted together. This likely matches your intuition, but you can see that male heights are centered at a higher mean. There are a few more interesting insights here: • First, the difference in mean is about **[click]** 13.5, which means the average male is about that much taller than the average female.
	• Second, you can see there's **[click]** significant overlap between these two distributions. About **[click]** 51.5% of male heights fall within the central 95% of female heights. Can you spot the point where any given person of that height is equally likely to be male or female? **[pause for thought]** That would be **[click]** just below 170 cm.
	Here is the cumulative distribution function or CDF for female heights. It looks like an s-shaped curve. Remember that the CDF shows the probability that a random variable will take on a value less than or equal to a given value. On the x axis you have the values of your random variable, in this case heights, with the mean of 162.5 at the center, and on the y axis you have the cumulative probability of obtaining at most that value. Imagine the x axis extending from **[click]** negative to positive infinity beyond this image. The cumulative probability never quite hits zero on the left, nor does it quite hit one on the right. There's always a non-zero chance of some slightly more extreme value occurring. In the real world, there are constraints on human height – for example, you can't be negative centimeters tall – but the normal distribution includes infinitely long tails. The CDF helps you answer questions like: • **[click]** What is the probability that a given female is 160cm or less? Can you approximate it using the CDF? **[pause for thought]** This probability corresponds to this **[click]** height of the CDF at 160cm, which is about **[click]** 39.2%. • **[click]** What is the height of the 75th percentile of females? **[click]** This line represents the 75th percentile of females, about **[click]** 167.29 cm

	• What is the probability that a random female's height falls within 165 and 170cm? That corresponds to **[click]** P of (X < or equal to 170) minus P of (X < or equal to 165), which is about **[click]** 21.7%.
TH	Great work exploring heights and the properties of the normal distribution. Follow me to the next video to explore a special case of the normal distribution.

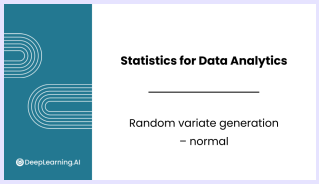
L3V3 – The standard normal distribution

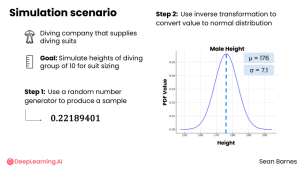
Visual	Script
TH 	The standard normal distribution. It might seem redundant at first: standard... normal... but let's take a look at why the standard normal distribution is so important.
	There are **[CLICK]** infinitely many normal distributions, as the **[CLICK]** center is determined by the mean parameter and the **[CLICK]** shape is determined by the standard deviation. Suppose you have a normal distribution with a **[CLICK]** mean of 0 and a **[CLICK]** standard deviation of 1, represented by the green probability density function curve in this chart. Now suppose you have a **[CLICK]** mean of 0 and a **[CLICK]** standard deviation of 2, your normal distribution is wider and flatter, and looks like the yellow curve. And if instead you have a **[CLICK]** mean of 0 and a **[CLICK]** standard deviation of 0.5, your normal distribution is taller and has less variability, like the blue curve in this chart. You can also have a normal distribution with a different mean that is shifted from the previous examples you've seen, like the **[CLICK]** purple curve with a **[CLICK]** mean of -2 and a **[CLICK]** standard deviation of 1. All of these normal distributions have **[CLICK]** mean equals median equals mode, and all of these **[CLICK]** follow the set of sigma rules, where 68% of the data is within 1 sigma of the mean, 95% is within 2 sigmas, and so on.
	But which normal distribution is the most normal-est of them all? It turns out that the **[click]** normal distribution with a **[click]** mu of 0 and a **[click]** standard deviation of 1 is called the **standard** normal distribution. You can represent this distribution using the notation the **[click]** random variable Z **[click]** is distributed as **[click]** normal **[click]** with a mu of 0 and a sigma of 1. Now remember from the previous module that the variance is the square of

	the standard deviation sigma, so the standard normal distribution also has a **[click]** variance of 1. Despite its humble name, the standard normal distribution is actually quite special. All these distributions you just saw are normal distributions, it's just that the **standard** normal distribution has some
	important mathematical properties that can help you in a variety of ways. One interesting property of the standard normal distribution is that the **[click]** value of any outcome in this distribution also corresponds to the number of standard deviations it is away from the mean. Let's break that down visually on this graph of the standard normal distribution. For example, the **[click]** value 1 in the distribution falls here, and represents **[click]** 1 standard deviation above the mean of 0. If you have a value of **[click]** -2, that value falls here, and represents **[click]** 2 sigmas below the mean. A value of **[click]** 0.72 falls here on the distribution, and represents **[click]** 0.72 sigmas above the mean. Because the standard normal distribution is often represented using the random variable capital Z, this value is also commonly referred to as a **[click]** "z score." **[pause to absorb]**
	Z scores **[CLICK]** provide a common reference for normally distributed data. Say you have some data you suspect is normally distributed, but it's really difficult to evaluate how big or small any given value is. For example, suppose your **[CLICK]** friend in law school comes to you and says they scored a **[CLICK]** 52 on their recent midterm. Is that **[CLICK]** high? **[CLICK]** Low? **[CLICK]** It's hard to tell. I certainly wouldn't know. However, if you hear that your friend's score has a **[CLICK]** z score of 1.7, well then you know it's **[CLICK]** 1.7 sigmas above the mean and that they **[CLICK]** did well above average! Z scores are similar to percentiles in this way – they both give you an **[CLICK]** intuition about the position of a data point relative to the rest of the distribution. If a test score is in the 95th percentile, then it's higher than 95% of the data. However, the advantage of Z scores is that they **[CLICK]** provide more information about how far away a given value is from the mean, whereas percentiles **[CLICK]** only provide you with information about the rank of that value.
	Now that you understand how z scores can help you interpret data points, let's see how to **[CLICK]** transform your original data into z scores, which is called standardization. You can perform standardization when you are looking for a way **[CLICK]**to interpret the distribution of your data according to a consistent scale.

	To calculate the z score for a given value, take your observed value **[CLICK]** x, subtract mu, and divide by sigma. This process has a two different effects: • First, by subtracting the mean, it **[CLICK]** centers the distribution at 0 • Then, by dividing by sigma, it **[CLICK]** expresses all the values in units of standard deviations **[CLICK]** After this transformation, all normal distributions will have a **[CLICK]** mean of 0 and a **[CLICK]** standard deviation of 1, regardless of their original parameters. This standardization process is crucial in many statistical analyses and allows you to compare values from different normal distributions on a common scale. For example, to find the z score for the test score **[CLICK]** from a moment ago, assume the scores are distributed as normal with a mean of **[CLICK]** 43.5 and a standard deviation of **[CLICK]** 5. So to calculate the z score, take your friend's **[CLICK]** score of 52 and **[CLICK]** subtract the mean of 43.5, then **[CLICK]** divide by the standard deviation of 5. This calculation produces a **[CLICK]** z-score of 1.7, which again represents 1.7 standard deviations above the mean.
	Z-score transformations are also reversible. **[CLICK]** If you know the **[CLICK]** z-score, **[CLICK]** mean, and **[CLICK]** standard deviation, you can calculate the original value using **[CLICK]** this equation: x = (z * σ) + μ. **[CLICK]** It's the opposite of the calculation **[CLICK]** you just did, and it's called an **[CLICK]** inverse transformation. You can always transform your data between **[CLICK]** z scores and the **[CLICK]** original scale.
TH	It's okay not to remember all of the details for now, but know that you'll use Z scores in the next two modules to construct confidence intervals and conduct hypothesis testing. For now, Just know that z scores represent the number of standard deviations from the mean in the standard normal distribution. Follow me to the next video to see how you can generate random samples from the normal distribution.

L3V4 – Random sampling – normal

Visual	Script
TH 	Let's take a look at sampling from any normal distribution, which starts from the standard normal distribution, then uses the inverse z-score transformation you learned about in the previous video.

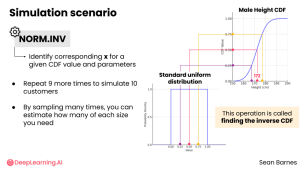


Say you're working for a **[click]** diving company that supplies diving suits for customers. You want to **[click]** simulate the heights of customers in a diving group of 10 for suit sizing.

You want to randomly generate heights, but the problem is heights aren't evenly distributed. So, unlike a uniform distribution where every value has the same chance of being generated, you need a more complicated function to map the random number from the uniform distribution onto probabilities of the normal distribution.

Here's one way to sample from the normal distribution.

1. First, **[click]** use a random number generator to produce a sample from a standard uniform distribution, just as you did for the discrete examples in the previous lesson. For example, you might get 0.2218.
2. Then, use an **[click]** inverse transformation to convert the value from Step 1 to the normal distribution value of interest. In the case of **[click]** male heights, you've already seen that it has a **[click]** mean of 172 cm and a **[click]** standard deviation of 7.1 cm. Let's see how this step works.



Here's the **[click]** cumulative density function, or CDF, for male height. Now, remember, the CDF ranges from 0 to 1, representing the cumulative probability as you move from -infinity to +infinity. Hmm, that's interesting, the **[click]** standard uniform distribution also ranges from 0 to 1. Suppose you were to generate a random value from the standard uniform distribution, let's say it's **[click]** 0.5. If you were to place this random sample on the y-axis of this CDF it would land **[click]** right in the middle.

Now imagine drawing a line across to the **[click]** CDF curve and then down to the **[click]** x-axis. What height would it correspond to? **[PAUSE for effect]** **[click]** The mean height, 172 cm! The probability of observing a value less than or equal to 172 is 0.5, since the distribution is symmetric about the mean.

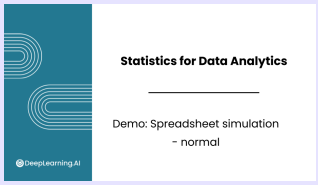
As you generate random values via the **[click]** uniform distribution, they will project to values on the **[click]** x-axis, which correspond to different sampled heights.

[click] This operation is called finding the inverse CDF, and you can use the spreadsheet function **[click]** NORM.INV **[click]** to identify the corresponding x for a given CDF value and set of normal distribution parameters (mu, sigma). You'll use it in a moment.

Now, you'll repeat this process **[click]** 9 more times to simulate the distribution of the 10 customers on the tour. **[click]** By sampling many times, you can estimate how many of each suit size you typically need.

	The beauty of this method is that it mimics real-world variability. Just like not everyone has exactly the average height, these samples will vary around the mean, giving you a realistic picture of your potential customers and helping you order the right inventory.
🎙️ TH	I bet you're interested in seeing this process in action. Let's take it over to a spreadsheet to simulate the heights of many random samples of males.

L3V5 – Demo: Spreadsheet simulation – normal

Visual	Script
🎙️ TH 	Let's see how you can generate random samples following the normal distribution using a spreadsheet.

🖥️ SC
🔗 Start
🔗 Solution
	Recall that male heights are roughly normally distributed. Let's see how to randomly sample heights from that distribution. This allows you to see the distribution of heights, for example to order the right numbers of different diving suit sizes depending on the number of individuals on a given trip
	
	And don't forget, if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab.
	
	Male heights from normal distribution
	• Define parameters
	  ○ E2: 172
	  ○ F2: 7.1
	• Create samples from uniform distribution
	  ○ A2: `=randarray(100)`
	• Sample from normal distribution
	  ○ B2: `=NORMINV(A2,F2,G2)`
	• Now crate an array option
	  ○ B2: `=arrayformula(norminv(A2:A101, E2, F2))`
	• Refreshing
	  ○ D1 header: Refresh button
	  ○ D2: Insert → checkbox
	  ○ Checking/unchecking will refresh
	• Visualize
	  ○ Insert chart → histogram
	• Show sample mean and stdev
	  ○ E3: sample mean
	  ○ E4: `=average(B2:B101)`

	○ F3: Sample standard deviation ○ F4: `=stdev(B2:B101)`
🎙️ TH	It's cool, right? I encourage you to try out this simulation on your own. You can also try changing the parameters to simulate female heights. Follow me to the next video to learn more about LLM simulations.

L3V6 – Demo: LLM simulation – normal

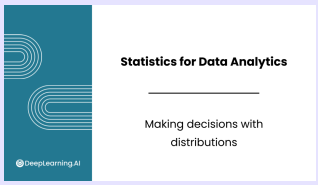
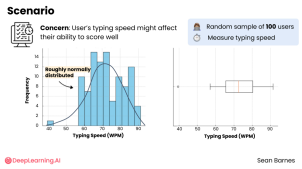
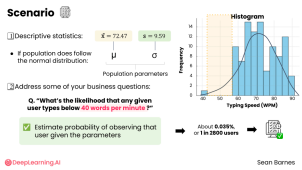
Visual	Script
🎙️ TH	LLMs are great thought partners for your work in data analytics. Earlier, you saw how to use them to create simulations. Let's take a closer look at their limitations as well as how to use them to generate random samples from the normal distribution.
🖥️ SC 🔗 **ChatGPT – logged out** 🔗 **Old screencast**	Recall that you can't use LLMs for sampling unless they can write and run code. Let's investigate that claim more closely. Let's start with ChatGPT 4oh-mini. This is a logged out account, not the paid version. It can't run code and it's very, very similar to what you'll work with in the Coursera labs that you'll see in this course. Say you want to simulate 100 samples from the standard normal distribution. You can ask ChatGPT to do that for you. ***PROMPT:*** *Simulate 100 samples from a standard normal distribution. Print them in a single comma separated list.* And it gives you 100 samples. Considering a mu of 0 and a sigma of 1, these all look, you know, pretty standard and normal. What you want to do is actually test whether these come from a normal distribution. If you copy these data points and come over to the paid version of ChatGPT, this is ChatGPT 4-oh, you can ask it to graph this data in a histogram. ***PROMPT:*** *Graph this data in a histogram: LIST OF NUMBERS THAT WAS OUTPUT* Since this is the paid version of ChatGPT, it has the advanced data analytics feature, and it can actually write and run code. You can see it's writing some

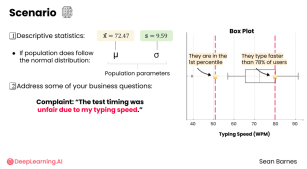
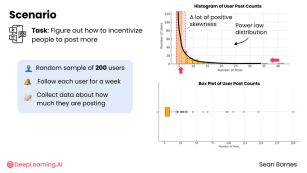
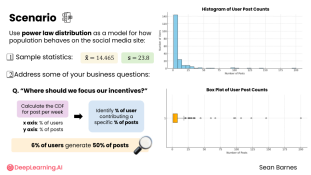
	code to plot a histogram. And there's your histogram. Let me ask you, does this look normally distributed to you? **[pause for thought]** Well, looks can be deceiving. This is beyond the scope of this course, but you can actually statistically test how likely it is that this data did in fact come from a normal distribution. I'll ask ChatGPT to do that. ***PROMPT:*** *What's the chance that this data comes from a normal distribution? Provide the test statistic, plot a Q-Q plot and interpret the results.* And it's going to provide you with a little information. Don't worry too much about this plot. But essentially this dark orange curve is what one half of the normal distribution looks like on its side, and this lighter yellow color is what the data looked like. But just focus on the conclusion here. "This test strongly suggests that the data does not follow a normal distribution."
🖥️SC 🔗**ChatGPT – paid**	One thing that LLMs on their own can help you with is to help set up a simulation or to help interpret your results. For example, say you want to create a simulation of stock prices. I will sample randomly from the normal distribution to simulate each stock. Please critique this approach and make suggestions to improve it. ***PROMPT:*** *I want to create a simulation of stock prices. I will sample randomly from the normal distribution to simulate each stock. Please critique this approach and make suggestions to improve it.* Okay. So it says using a normal distribution to simulate stock prices is a common approach, but it does have limitations. Stock returns often have fat tails and skewness, meaning that extreme prices are more common than in a normal distribution. It suggests some different distributions stocks might follow. Plus maybe you should include external factors like announcements, economic news, market shocks, and so on. It gives you a lot of really advanced suggestions, but the top one of using a different distribution or incorporating real world events could be something that you might be able to do.
🖥️SC 🔗**Claude**	Okay, so let's take this over to Claude, which is capable of running some really cool simulations. Remember that Claude can write and run code using its

	Artifacts feature. ***PROMPT:*** *Create an interface for sampling from a normal distribution of female heights with a mean of 162.5 centimeters, and a standard deviation of 7.1. When I click the sample button, create a histogram from the sampled heights, and calculate the sample mean x bar and the sample standard deviation s. Display these next to the mu and sigma.* Say you wanted to create an interface for sampling from a normal distribution of female heights with a mean of 162.5 centimeters, and a standard deviation of 7.1. When you click the sample button, you want it to create a histogram from the sampled heights, and calculate x bar and s, and display these next to the mu and sigma. So mu and sigma are the values you already gave it. This simulation can help you see all the different possible scenarios you might get if you randomly sampled one hundred females and measured their height. Okay, let's go ahead and generate a new sample. So you have your heights on the x axis, you have the frequency on the y axis, and on the left you have the population parameters, and on the right you have the sample statistics. So even with just a hundred samples, the sample mean is super, super close to mu. The distribution of heights changes, but they do appear to be roughly normally distributed. **[showing specific histogram]** It's possible to get this type of outcome with a simple random sample where you really don't have that many people at 160. Or at 152.
🎙️ TH	Great work with the LLM – you'll get to practice with it in a moment. Any LLM can help you devise a simulation and communicate results, but you should only run a simulation using an LLM that can write and run code. Follow me to the next video to put it all together and learn how distributions can be used to make decisions.

L3V7 – Making decisions with distributions

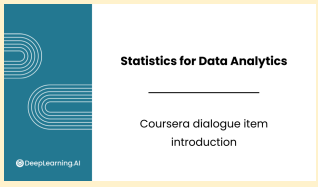
Visual	**Script**
🎙️ TH	Let's put this all together and see how you can use distribution techniques to inform different business decisions in two different scenarios.

	
 https://i.imgur.com/eEkXFIc.png	Imagine that you're working with a **[CLICK]** digital testing company that provides timed standardized tests that users can take in their browsers. **[CLICK]** One concern the testing company may have is that a user's typing speed might affect their ability to score well on the test even if they know the material well. You decide to take a simple **[CLICK]** random sample of 100 users and **[CLICK]** measure their typing speed. Your first step is to visualize the distribution, so you create **[CLICK]** a histogram and **[CLICK]** a box plot. What kind of distribution do you see? **[pause for thought]** It's not 100% clear, but the data does look **[CLICK]** roughly normally distributed. The sample size is only 100, so there is some variability from the typical bell curve. But you have some evidence based on the visual distribution of the data that the population may follow a roughly normal distribution.
	Recognizing that you've made an assumption, you can proceed with creating **[CLICK]** descriptive statistics for your data and try to **[CLICK]** address some of your business questions. First, calculate your sample mean **[CLICK]** x-bar and standard deviation **[CLICK]** s. They are **[CLICK]** 72.47 and **[CLICK]** 9.59 words per minute respectively. **[CLICK]** If your population does indeed follow the normal distribution, you can use your x-bar and s as approximations for **[CLICK]** mu and **[CLICK]** sigma, your **[CLICK]** population parameters. The normal distribution provides assumed probabilities for the big gap in your data between **[CLICK]** 42 and 55 or so. In reality, you might also expect more extreme values than what your sample produced in the population. Using your population model, you can answer questions like: • **[CLICK]** What's the likelihood that any given user types below 40 words per minute, which is the **[CLICK]** critical threshold for being able to complete the test on time? Even though you didn't sample anyone with a typing speed that low, the normal distribution can help you **[CLICK]** estimate the probability of observing that user given the parameters you estimated. You calculate that the probability that any given user types below 40 wpm is about **[CLICK]** 0.035%, or 1 in 2800 users. That probability can inform whether you feel the test is **[CLICK]** valid given the current time limit.

	Say a user files a complaint with the testing company, saying **[CLICK]** the test timing was unfair due to their typing speed. If they say their typing speed is **[CLICK]** 80 words per minute, you can find the percentile of this individual's typing speed using the normal distribution. You calculate that **[CLICK]** this user falls in the 78th percentile, meaning **[CLICK]** they type faster than 78 percent of users. This may be grounds for rejecting this individual's complaint. By contrast, if another user complains and their typing speed is tested at **[CLICK]** 51 words per minute, you can estimate that they are in **[CLICK]** the 1st percentile of typers – quite slow, which may qualify them for extended time.
 https://i.imgur.com/FnGRyMk.png https://i.imgur.com/bcqovqc.png	Say you're working with a **[CLICK]** social media company and you're tasked with **[CLICK]** figuring out how to incentivize people to post more often. You generate a simple **[CLICK]** random sample of 200 users, **[CLICK]** follow each user for one week, and **[CLICK]** collect sample data about how much they are posting. Your first step is visualizing the distribution, so you create **[CLICK]** a histogram and **[CLICK]** box plot of the data. What kind of distribution do you see? **[pause for thought]** It looks like it might follow a **[CLICK]** power law distribution, since you have a **[CLICK]** large concentration of users near 0 with a **[CLICK]** long tail in the positive direction, indicating **[CLICK]** a lot of positive skewness.
	You can **[CLICK]** use the power law distribution as a model for how the population behaves on the social media site. There are some further steps needed to validate this hypothesis, but just imagine you're able to confirm that this sample follows this distribution. First, **[CLICK]** you can calculate your sample statistics. The power law distribution actually has different parameters from the normal distribution, but you can still calculate the **[CLICK]** mean x-bar, which is 14.465, and your **[CLICK]** sample standard deviation, which is 23.8. Just note that because this distribution isn't normal, you can't apply the set of sigma rules you learned previously. Using this probability distribution as a model for your population, you can start to **[CLICK]** address some of your business questions, keeping in mind that you're operating under the assumption that your population behaves in this way. • ~~For example, you might wonder, **[CLICK]** where should we focus our incentives? Based on the power law distribution, you can **[CLICK]** calculate the CDF for posts per week, with the **[CLICK]** percent of users on the x-axis and the **[CLICK]** percent of posts on the y-axis.~~ • ~~The CDF allows you to **[CLICK]** identify the percentage of users contributing a specific percentage of posts. So you can calculate that **[CLICK]** 6% of users generate 50% of posts, meaning that you may~~

	~~want to [CLICK] focus on incentives for those 6% of power users.~~ Say the marketing team lets you know that **[CLICK]** a particular incentive causes users in the bottom 50% of activity to create one more post per week. Using your **[CLICK]** model of the power law distribution, you note that **[CLICK]** the median number of posts would increase by 33% from 3 to 4. The mean number of posts would increase by 0.5, since you're adding 1 post to half of all users. So it would be up about **[CLICK]** 5% from 10.47 to 10.97. This difference makes sense since the mean is heavily influenced by values in the tail of the distribution compared with the median. And **[CLICK]** the total number of posts would also increase by the same percentage as the mean. This shift helps you characterize the effect this incentive might have if it was implemented for the entire population.
🎙️ TH	Great work using distributions and descriptive statistics to make decisions! That brings you to the end of Module 2 of this course. You've learned a ton about probability, distributions, and simulation. You're practically a pro at this! Coming up next are the graded assessment and lab for this module. In the lab, you'll extend your analysis of the forest fire prevention dataset to help estimate wildfire locations based on the sample data you have. Once you've completed the lab and assessment, I'll see you in Module 3: confidence intervals.

Coursera dialogue item introduction

Visual	Script
🖥️ SC 🔗 **Example screencast** 	Coming up, you'll complete your first Coursera dialogue item. The Coursera dialogue is driven by a large language model and is a really great opportunity for you to talk through your thinking in natural language. It can help you both check and refine your understanding of the course material. Keep in mind this is a new and evolving item, so you may see a slightly different interface or have a slightly different experience than the one in this video. The goal of this item is to have a conversation with the Coursera coach to review the course material thus far. At the start of the dialogue, the coach will explain the upcoming task and give you a few checkpoints to guide the conversation. As you talk with the coach, it will assess your understanding of the material against these checkpoints so that you can meet each one.

	You'll reply to the coach just as you would with a large language model. So for example, it's asking, are you ready to begin? So you can just say “yes” and it will share the task with you. You can respond to the task by typing directly into the chat. At any point, You can ask the coach how you're doing and it will evaluate your responses against the checkpoints and give you some suggestions or tell you you're on the right track. If you're stuck, you can click the “I’m stuck” button to get some friendly hints. And you can also create a new chat if you want to start the conversation over. Once you’ve completed all three checkpoints, the Coach will give you an overall assessment of your understanding. From there, you can either continue your conversation or move on to the next item! This is a great opportunity for you to review, ask questions, and test yourself. So once you’ve completed the dialogue item, I’ll see you in the next video.