Probabilities and Percentiles: Part 1 - Probabilities

Machine learning is everywhere these days. One of the challenges we run into is understanding how to use the outcomes from the models we create. This problem is amplified by the fact that the terminology data scientists use is often unfamiliar to people without a strong mathematical background.

*Two of the most common ways to talk about the predictions from a machine learning model are the **probability** predicted by the model and the **percentile** of the prediction across the population evaluated by the model. It gets even more confusing since probabilities are often expressed as percentages. In this blog series, we will discuss what each of these outputs tell you and how to use them to solve problems for your business.*

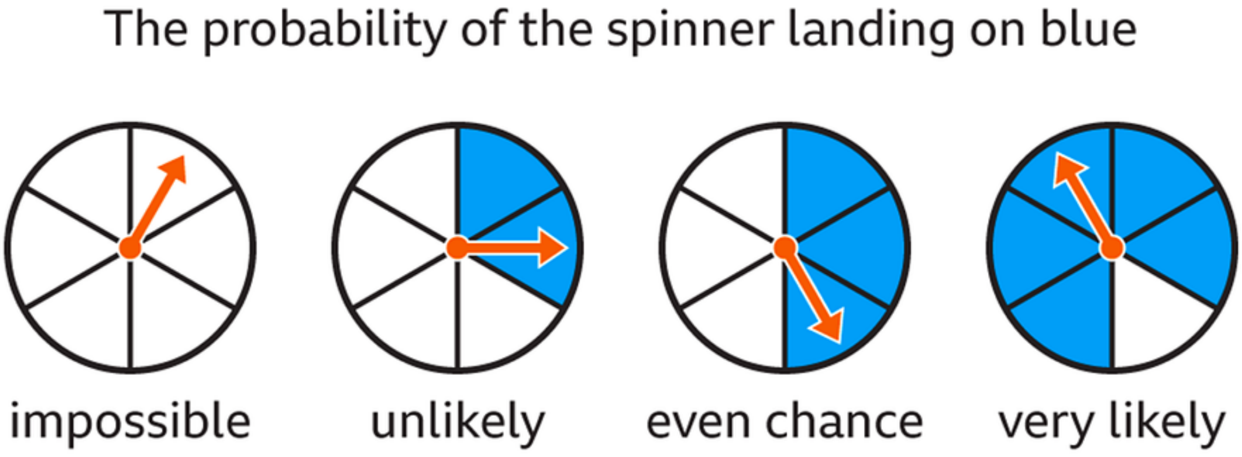
This is the first of three articles, where we get into what a probability represents. If you're more interested in learning about quantiles, read Part 2. If you're ready for the highlights on how probabilities and quantiles are different and when you should use each one, read Part 3.

Probabilities

By default, machine learning models often return a ***probability***. A probability reflects the model's prediction of how likely a record is to belong to a given group. The range of a probability is 0 to 1, including both 0 and 1, where 0 means something is impossible and 1 means that something will happen with certainty.

For example, we might train a model that tells us given the details we have about a customer, how likely are they to buy our product. If the customer buying or not buying our product was equally likely, the probability of both outcomes would be 1/2, or 0.5 (50%). If the customer was more likely to buy our product than not buy our product, the probability would be greater than 0.5, and conversely if they were more likely to not buy than to buy, the probability would be less than 0.5.

In this visualization from the BBC How to describe probabilities and the probability scale , we can see different probabilities visualized as a spinner. On the far left we can see something with a 0 probability, then on the far right we see something with 5/6 (0.83) probability, which is very likely. A scenario that is a certainty would be visualized with all of the wedges on the spinner being blue (probability equal to 1).



What we are able to do by training a machine learning model is use the data we have about a person, place, or thing to adjust the probabilities generated by our model to be more accurate than a random guess between possible outcomes, where all possible outcomes are equally likely. Machine learning models generate a probability how how likely an outcome is for a specific instance of data, given the patterns identified in historical data.

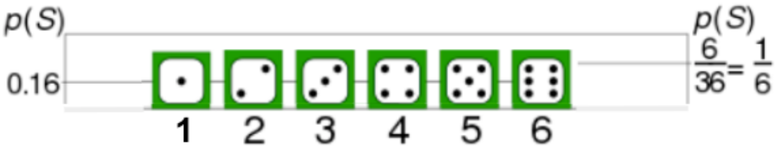
As a side note, we typically treat the “random guess” scenario as a baseline of comparison, to determine if a model has learned something from the data. If a model’s performance on data where we know the outcome is an improvement over randomly guessing at outcome, where all outcomes are equally likely, we know that the model has found a pattern in the data that is helping it make meaningful predictions.

Probability Distributions

The overall behavior of outcomes of a given phenomena can be described by a probability distribution. A probability distribution is a mathematical function that describes the general shape of the probabilities of outcomes across different occurrences.

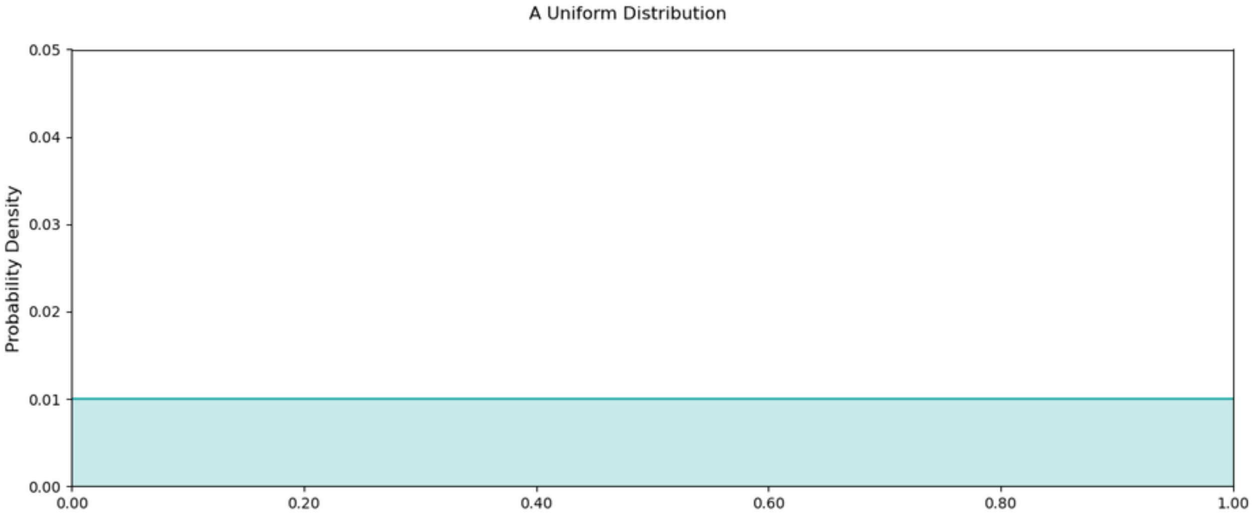
For intuition, we’re going to start by looking at the general shapes of probabilities for possible outcomes of rolling dice (this is what we would call discrete probability distributions, and not a probability distribution, but discrete vs. continuous is a bit of a rabbit hole I’m not going to get into here - but for the record, please note that they are not the same thing).

First, lets imagine we’re rolling just one dice.



Here, because the possible outcomes of rolling a fair die are all equally likely, we see that the probability for each outcome is the same.

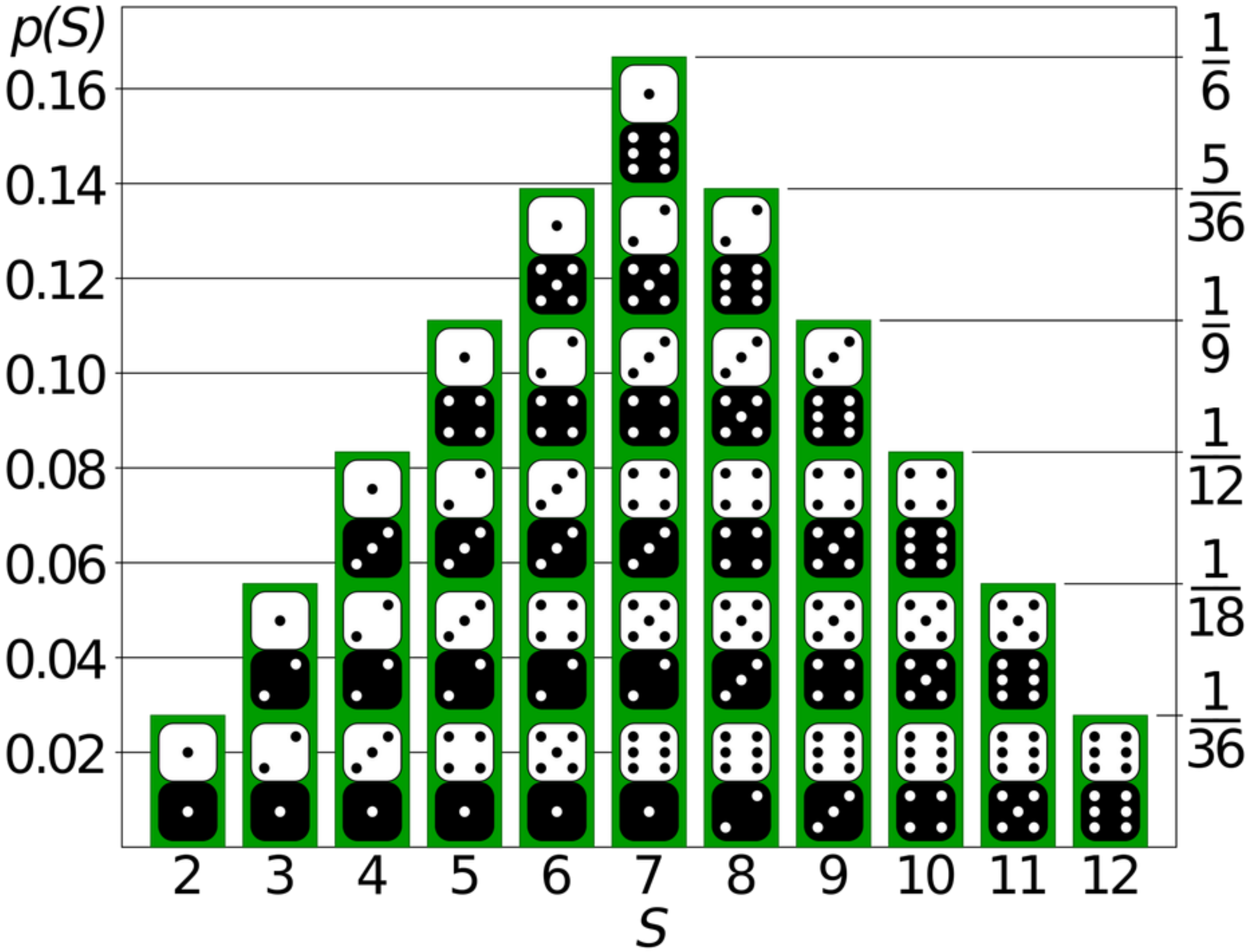
If we are working with something where it is possible for the values to be anything in-between a given range (like probability), the single dice example would be the equivalent of a uniform distribution, where all possible outcomes are equally likely.



The shape of a uniform probability distribution is a horizontal line where the value of the horizontal line is the 1 divided by the number of possible outcomes.

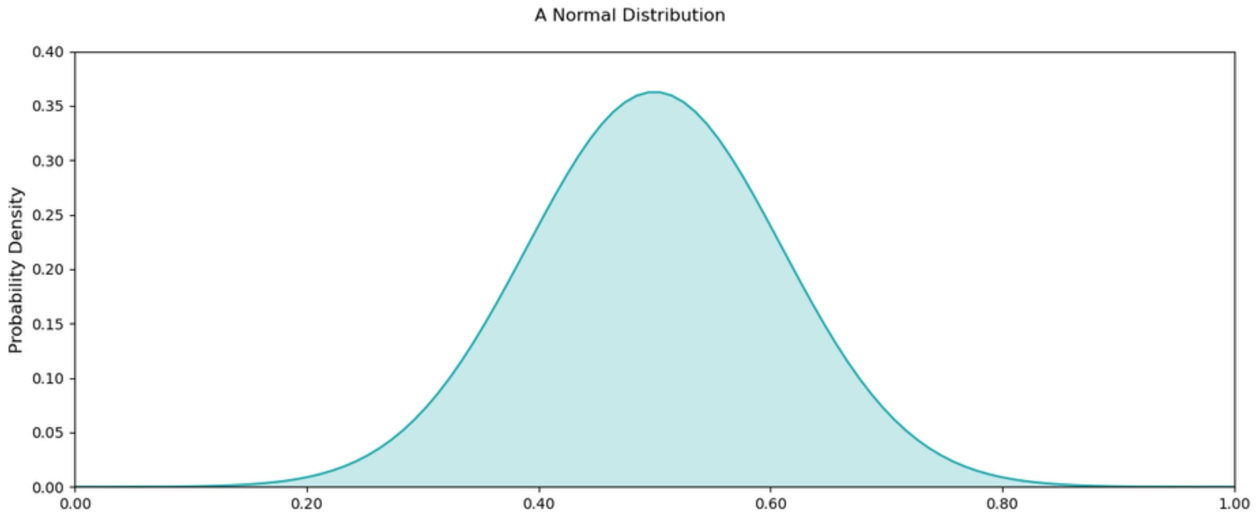
Next, let’s imagine we’re rolling two dice, which is the setup you’ll see for games like Settlers of Catan or Craps.

When we look at the possible outcomes of rolling two dice, we see that because there are the most ways to “make” a 7, it ends up having the highest probability of outcomes, 1/6. As we look at values further away from 7, they become less likely. This is why you end up with so many sheep when you have a settlement on a sheep with a 6. Its also why the robber, or the house, wins on a 7 role.



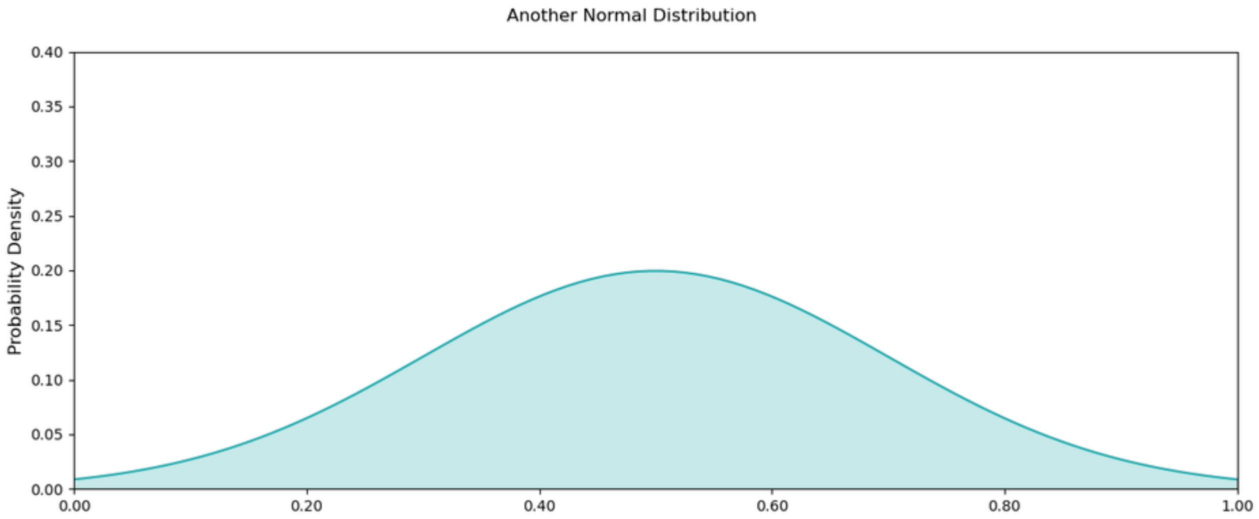
W File:Dice Distribution (bar).svg

If it were possible to have values in between the dice role outcomes (which we can approximate by increasing the number of dice being rolled for each trial), this would approximately equate to a normal distribution, also known as the bell curve, where the outcomes towards the middle (average) values of a population are more likely than the values at either extreme of a population.

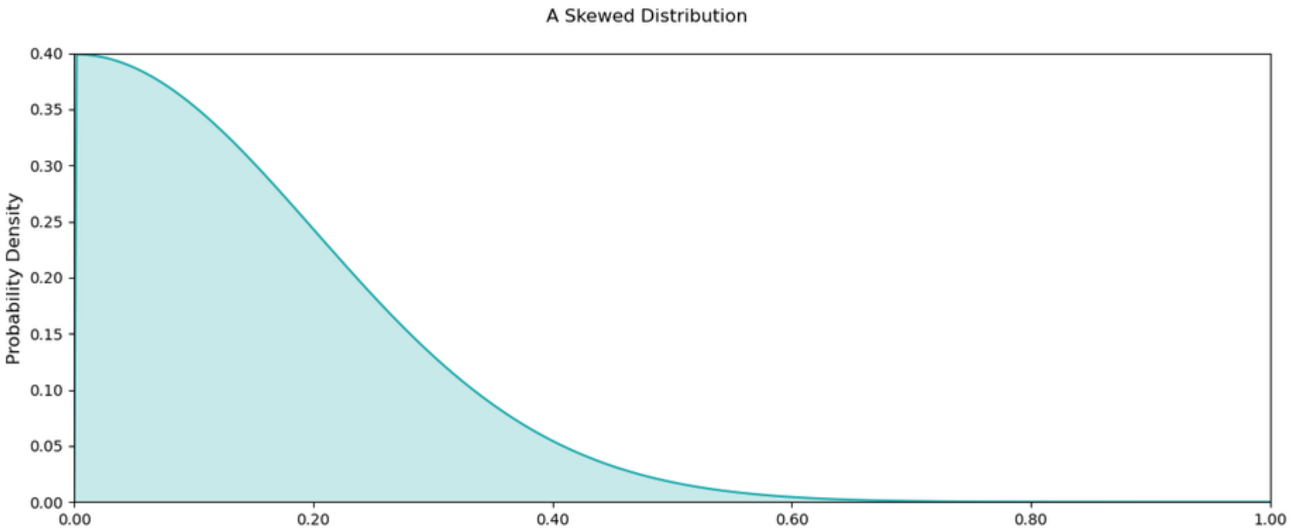


In this normal probability distribution, we see that the most common outcome is a 0.5, with just over 30% of the values in the population occurring at 0. In both directions, as you move away from 0.5, a smaller proportion of the outcomes occur.

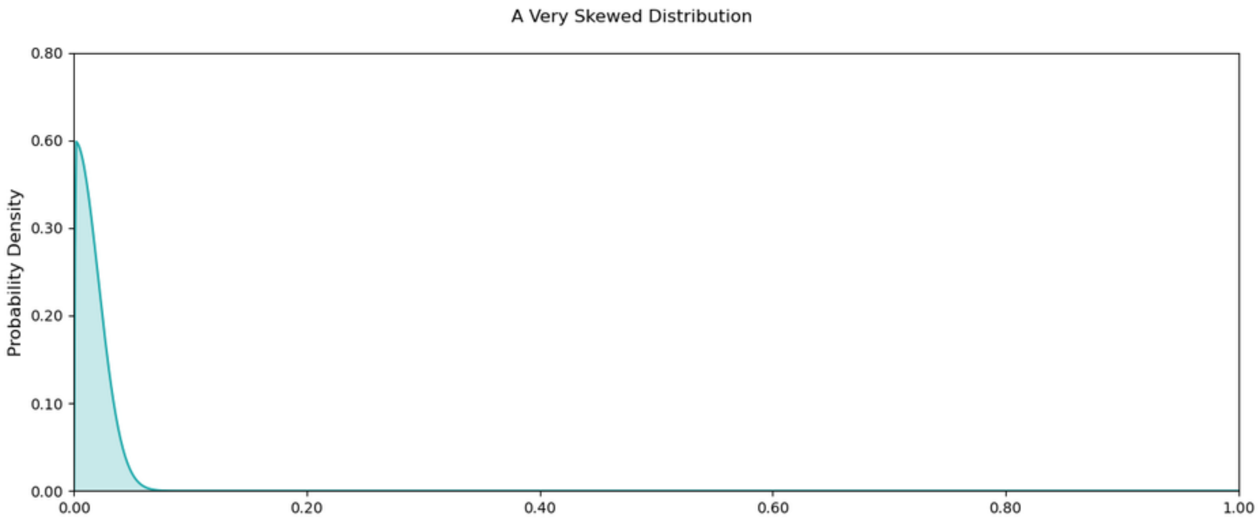
Probability distributions can take on many different shapes. Here we have another normal distribution centered at 0.5, but as you can see the peak of the distribution is lower and wider. Just under 20 percent of the instances occur at 0.5.



Or we can have a skewed distribution, which means that the population is not symmetrically distributed around the most common outcome. Here, we can see a skewed distribution, while the average value of the population is close to 0.



We can see that effectively no records have a probability greater than 0.7 in this population, and most of the probabilities are very low. This probability distribution might represent something that is somewhat infrequent or unlikely to happen - like a person's chances of finding a dollar bill on the sidewalk on a given day. A more extreme example of this would be an individual's probability of getting struck by lightning.



My imagined probability distribution for lifetime lightning strikes.

What is important to take out of this is that without knowing the probability distribution of a given problem, it can be difficult to make statements on where a specific probability returned by a model falls in a population relative to other probabilities beyond "more likely" or "less likely".

For example, in a probability distribution where the average probability of a positive outcome is 0.25, a 0.5 probability might be one of the highest probability events the model predicts, but in a probability distribution where the average is 0.8, that same probability might be one of the less likely predictions a model makes. In both cases, the individual probability means the same thing, that a given event has a 50/50 (or 1 in 2) chance of occurring, but relative to the overall population of events, this might be one of the highest probabilities returned, or one of the lowest.

Would You Like to Know More?

Probabilities and probability distributions are an important fundamental concept in machine learning. For additional reading, check out:

Probability: the basics (article) | Khan Academy

Probability

Basic Concepts of Probability

To continue reading this series and learn about Quantiles, read LINK TO POST 2.