Probabiltiies and Percentiles: Part 2 - Quantiles

This series probably should have been titled Probabilities and Quantiles - but the alliteration was not as good.

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.*

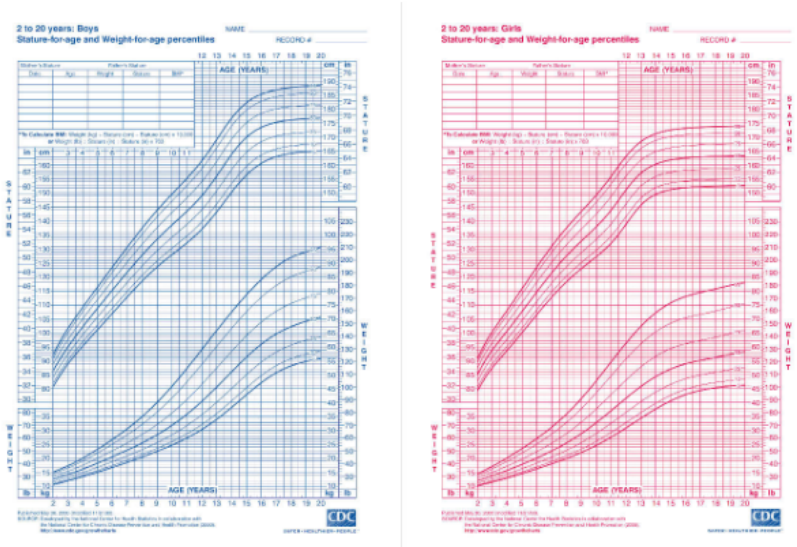
This is the second of three articles, where we get into what quantiles, deciles, and percentiles are. To read about probabilities, check out Part 1. To read about how probabilities and quantiles are different, and when each is useful, read Part 3.

Percentiles (Quantiles)

Quantiles are a way of dividing up a group of data into equal sized and adjacent groups. Equally sized is pretty straight forward to understand; if you have a group of 100 items, and you want to sub-divide that group into five groups (quintiles), each of the sub-groups would be 20 items.

If you have had children (or have been a child) in the United States, you might be familiar with the growth chart percentiles published by the Center for Disease Control (CDC). https://www.cdc.gov/growthcharts/data/set1clinical/cj41c021.pdf

https://www.cdc.gov/growthcharts/data/set1clinical/cj41c022.pdf

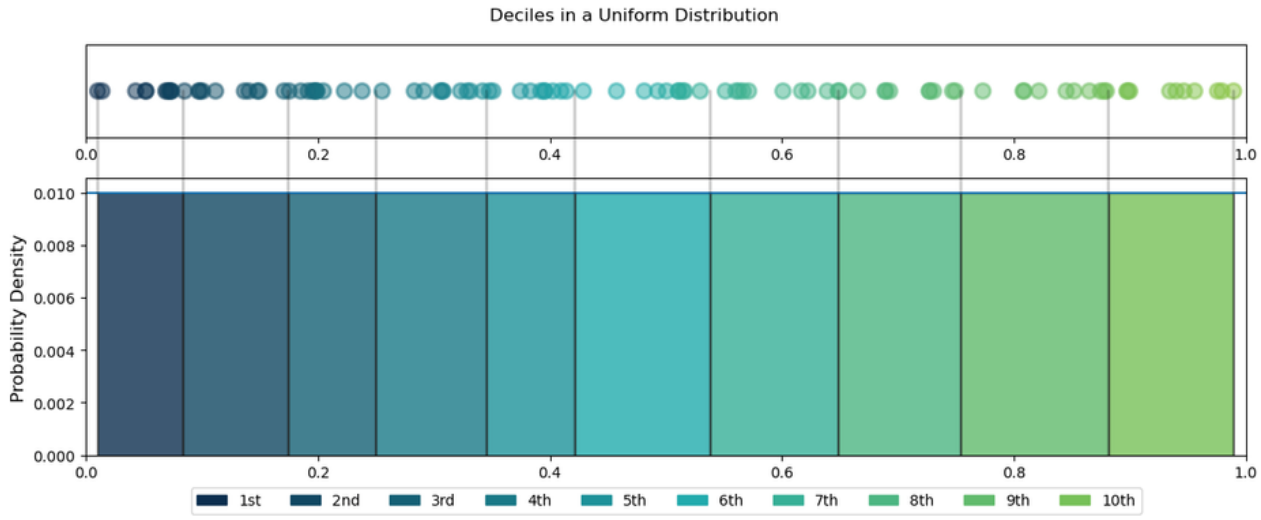


These charts are calculated by taking a large sample of children's' heights and weights at different ages, and then dividing the heights and weights into 100 equally-sized, ordered groups to create a range of values that divide the heights and weights of the population into 100 increments. So if you were told you were in the 50th percentile for height, that meant that 50% of the population was taller, and 50% of the expected population was shorter than you at your age.

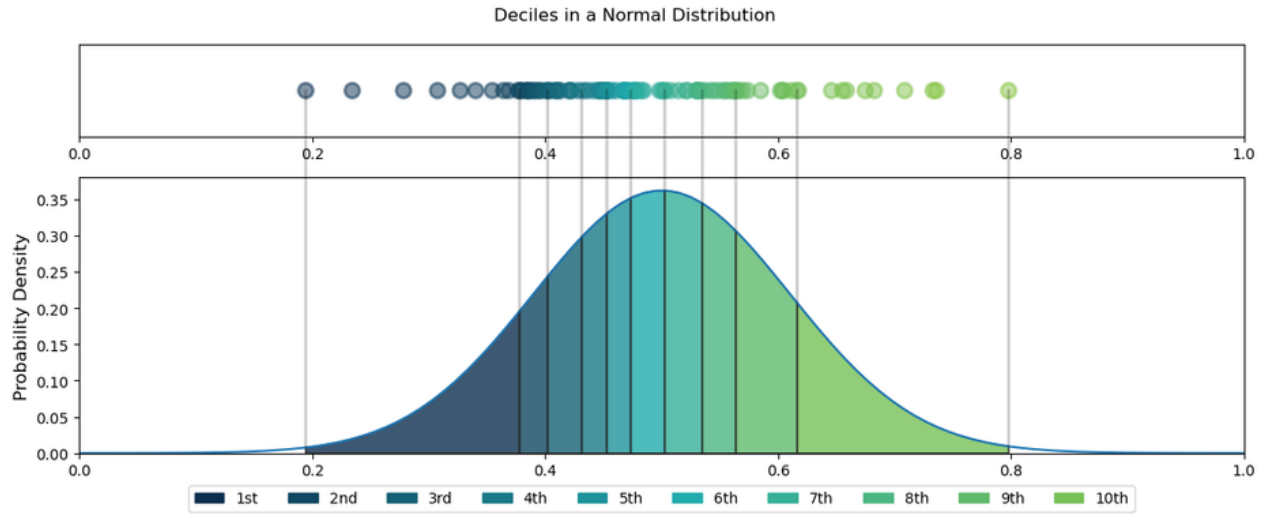
Percentiles and deciles are specific flavors of quantiles. Percentiles are quantiles where the number of groups is set to 100, and deciles are quantiles where the number of groups is set to 10. Percentiles and deciles are directly related to one another that if something is in the 10th percentile it should be in the first decile.

Thinking about our probability distributions, we can visualize how our quantiles split up the individual predictions within the population into equally sized groups. The range of values included in each bin will vary, but the number of records that each bin contains should be the same.

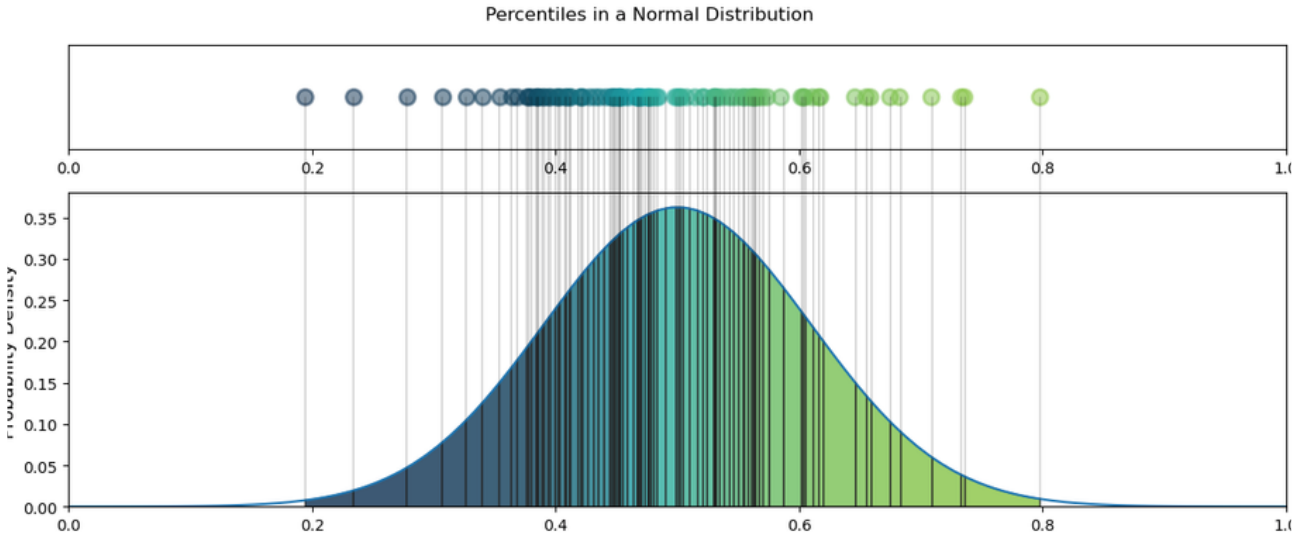
In a uniform distribution, we would expect the range of each of the bins to be roughly equal.



In our normal distribution, we can see that the ranges of the deciles are larger at the beginning and end of our possible probabilities, where instances of the probabilities occur less frequently.



We can see the same visualization with with percentiles instead of deciles. Here, there were 100 samples in the population, so each percentile contains just one sample.



Quantiles like deciles or percentiles are useful for comparing items within a population to one another. This is why quantiles are often used for test rankings or child growth curves. Saying that something is in the 20th percentile, for example, means that 20% of the population is less than that value, and 80% is greater. Regardless of what the value itself represents, a percentile allows us to know where it ranks relative to its peers.

Would You Like to Know More?

Quantiles are an important foundational concept for data analysis. If you'd like to learn more, please read:

W Quantile

Quartiles & Quantiles | Calculation, Definition & Interpretation

To continue your journey of understanding probabilities, quantiles, and how they function as outputs of machine learning models, read part three in the Probabilities and Percentiles blog post series; LINK TO POST 3.