The objectives of this lesson are to describe what the concept of a model is more generally, to explain what the purpose of a model is with respect to a set of data, and last, to describe the process by which a data analyst creates, assesses, and refines a model. In a very general sense, a model is something we construct to help us understand the real world. A common example is the use of an animal which mimics a human disease to help us understand, and hopefully, prevent and/or treat the disease. The same concept applies to a set of data--presumably you are using the data to understand the real world.

In the world of politics a pollster has a dataset on a sample of likely voters and the pollster's job is to use this sample to predict the election outcome. The data analyst uses the polling data to construct a model to predict what will happen on election day. The process of building a model involves imposing a specific structure on the data and creating a summary of the data. In the polling data example, you may have thousands of observations, so the model is a mathematical equation that reflects the shape or pattern of the data, and the equation allows you to summarize the thousands of observations with, for example, one number, which might be the percentage of voters who will vote for your candidate. Right now, these last concepts may be a bit fuzzy, but they will become much clearer as you read on.

A statistical model serves two key purposes in a data analysis, which are to provide a *quantitative summary* of your data and to impose a specific *structure* on the population from which the data were sampled. It's sometimes helpful to understand what a model is and why it can be useful through the illustration of extreme examples. The trivial "model" is simply **no model at all**.

Imagine you wanted to conduct a survey of 20 people to ask them how much they'd be willing to spend on a product you're developing. What is the goal of this survey? Presumably, if you're spending time and money developing a new product, you believe that there is a large *population*of people out there who are willing to buy this product. However, it's far too costly and complicated to ask everyone in that population what they'd be willing to pay. So you take a *sample* from that population to get a sense of what the population would pay.

Recently, I published a book titled *R Programming for Data Science*. Before the book was published, interested readers could submit their name and email address to the book's web site to be notified about the books publication. In addition, there was an option to specify how much they'd be willing to pay for

the book. Below is a random sample of 20 response from people who volunteered this information.

25 20 15 5 30 7 5 10 12 40 30 30 10 25 10 20 10 10 25 5

Now suppose that someone asked you, "What do the data say?" One thing you could do is simply hand over the data---all 20 numbers. Since the dataset is not that big, it's not like this would be a huge burden. Ultimately, the answer to their question is in that dataset, but having all the data isn't a summary of any sort. Having all the data is important, but is often not very useful. This is because the trivial model provides no reduction of the data.

The first key element of a statistical model is *data reduction*. The basic idea is you want to take the original set of numbers consisting of your dataset and transform them into a smaller set of numbers. If you originally started with 20 numbers, your model should produce a summary that is fewer than 20 numbers. The process of data reduction typically ends up with a *statistic*. Generally speaking, a statistic is any summary of the data. The sample mean, or average, is a statistic. So is the median, the standard deviation, the maximum, the minimum, and the range. Some statistics are more or less useful than others but they are all summaries of the data.

Perhaps the simplest data reduction you can produce is the mean, or the simple arithmetic average, of the data, which in this case is $17.2. Going from 20 numbers to 1 number is about as much reduction as you can do in this case, so it definitely satisfies the summary element of a model.

Models as Expectations

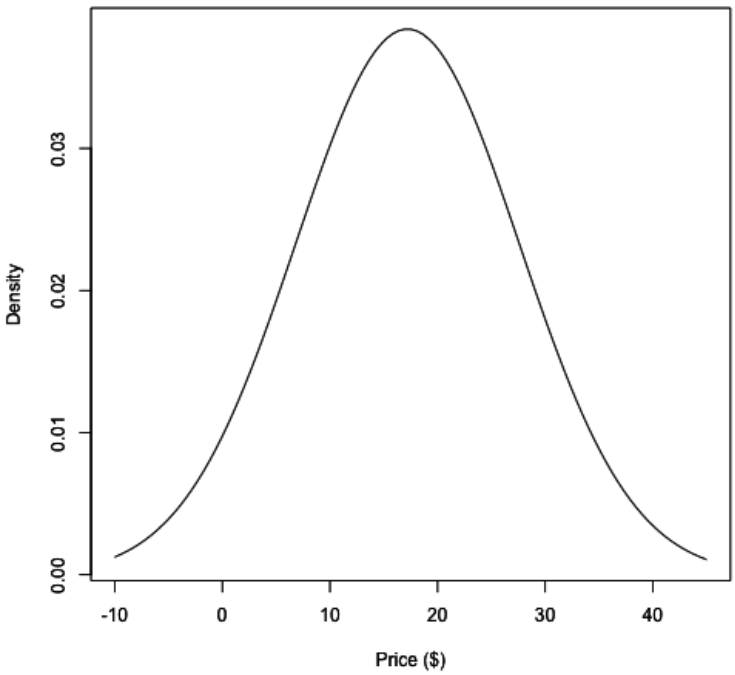
But a simple summary statistic, such as the mean of a set of numbers, is not enough to formulate a model. A statistical model must also impose some structure on the data. At its core, **a statistical model provides a description of how the world works and how the data were generated**. The model is essentially an *expectation* of the relationships between various factors in the real world and in your dataset. What makes a model a *statistical model* is that it allows for some randomness in generating the data.

Applying the normal model

Perhaps the most popular statistical model in the world is the Normal model. This model says that the randomness in a set of data can be explained by the Normal distribution, or a bell-shaped curve. The Normal distribution is fully specified by two parameters---the mean and the standard deviation.

Take the data that we described in the previous section---the amount of money 20 people were willing to pay for a hypothetical new product. The hope is that these 20 people are a representative sample of the entire population of people who might purchase this new product. If that's the case, then the information contained in the dataset can tell you something about everyone in the population.

To apply the Normal model to this dataset, we just need to calculate the mean and standard deviation. In this case, the mean is $17.2 and the standard deviation is $10.39. Given those parameters, our expectation under the Normal model is that the distribution of prices that people are willing to pay looks something like this.



According to the model, about 68% of the population would be willing to pay somewhere between $6.81 and $27.59 for this new product. Whether that is useful information or not depends on the specifics of the situation, which we will gloss over for the moment.

You can use the statistical model to answer more complex questions if you want. For example, suppose you wanted to know "What proportion of the population would be willing to pay

more than $30 for this book?" Using the properties of the Normal distribution (and a little computational help from R), we can easily do this calculation. About 11% of the population would be willing to pay more than $30 for the product. Again, whether this is useful to you depends on your specific goals.

Note that in the picture above there is one crucial thing that is missing---the data! That's not exactly true, because we used the data to draw the picture (to calculate the mean and standard deviation of the Normal distribution), but ultimately the data do not appear directly in the plot. In this case **we are using the Normal distribution to tell us what the population looks like**, not what the data look like.

The key point here is that we used the Normal distribution to setup the shape of the distribution that we *expect* the data to follow. The Normal distribution is our expectation for what the data should look like.

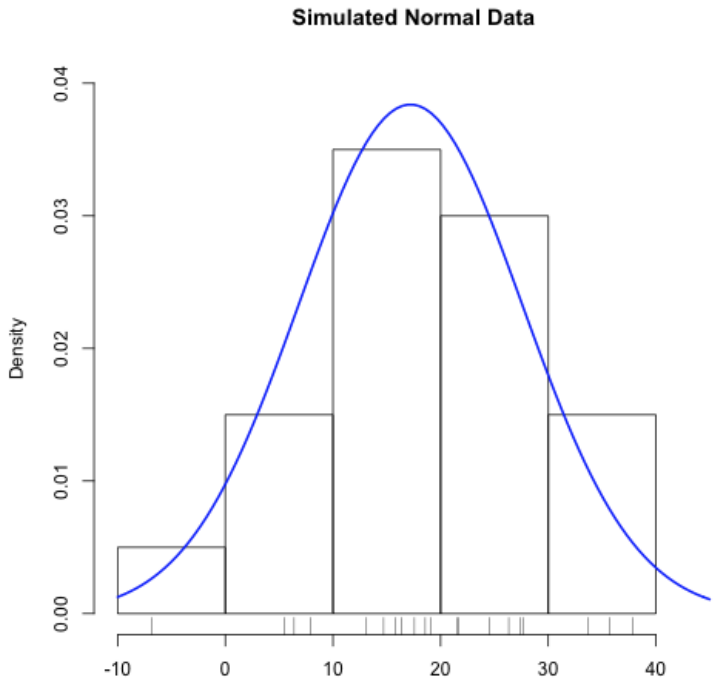
Comparing Model Expectations to Reality

We may be very proud of developing our statistical model, but ultimately its usefulness will depend on how closely it mirrors the data we collect in the real world. How do we know if our expectations match with reality?

Drawing a fake picture

To begin with we can make some pictures, like a histogram of the data. But before we get to the data, let's figure out what we *expect* to see from the data. If the population followed roughly a Normal distribution, and the data were a random sample from that population, then the distribution estimated by the histogram should look like the theoretical model provided by the Normal distribution.

In the picture below, I've simulated 20 data points from a Normal distribution and overlaid the theoretical Normal curve on top of the histogram.



Notice how closely the histogram bars and the blue curve match. This is what we want to see with the data. If we see this, then we might conclude that the Normal distribution is a **good statistical model for the data**.

Simulating data from a hypothesized model, if possible, is a good way to setup expectations *before* you look at the data. Drawing a fake picture (even by hand, if you have to) can be a very useful tool for initiating discussions about the model and what we expect from reality.

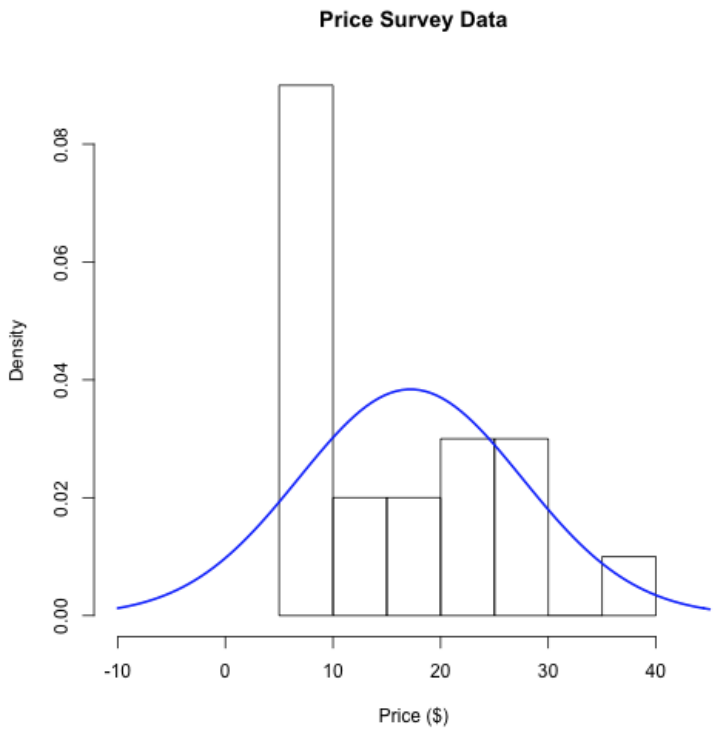
For example, before we even look at the data, we might suspect the Normal model may not provide a perfect representation of the population. In particular, the Normal distribution allows for *negative* values, but we don't really expect that people will say that they'd be willing to pay negative dollars for a book.

So we have some evidence already that the Normal model may not be a *perfect* model, but no model is perfect. The question is does the statistical model provide a reasonable approximation that can be useful in some way?

The real picture

Here is a histogram of the data from the sample of 20 respondents. On top of the histogram, I've overlaid the Normal

curve on top of the histogram of the 20 data points of the amount people say they are willing to pay for the book.



What we would *expect* is that the histogram and the blue line should roughly follow each other. How do the model and reality compare?

At first glance, it looks like the histogram and the Normal distribution don't match very well. The histogram has a large spike around $10, a feature that is not present with the blue curve. Also, the Normal distribution allows for negative values on the left-hand side of the plot, but there are no data points in that region of the plot.

So far the data suggest that the Normal model isn't really a very good representation of the population, given the data that we sampled from the population. It seems that the 20 people surveyed have strong preference for paying a price in the neighborhood of $10, while there are a few people willing to pay more than that. These features of the data are not well characterized by a Normal distribution.

Reacting to Data: Refining Our Expectations

Okay, so the model and the data don't match very well, as was indicated by the histogram above. So what do do? Well, we can either

1. Get a different model; or
2. Get different data

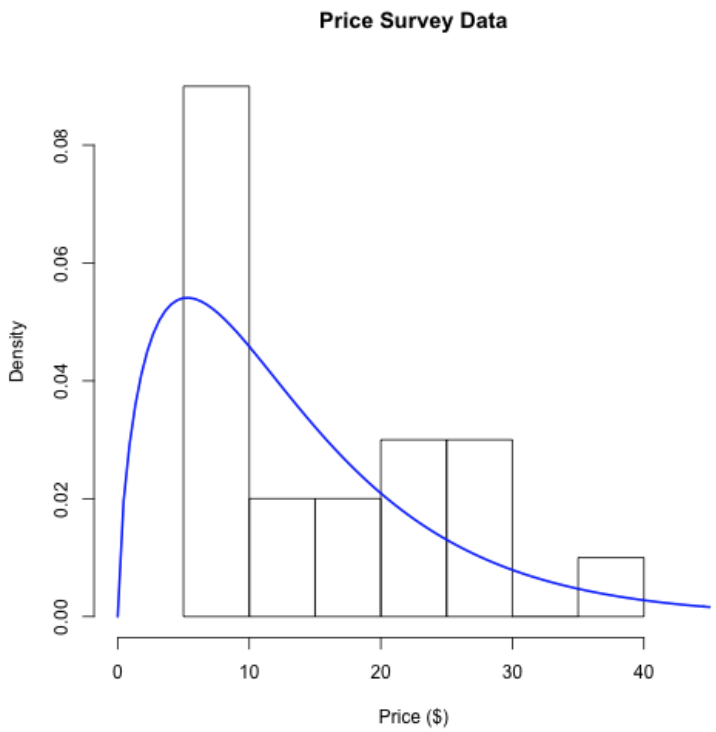
Or we could do both. What we do in response depends a little on our beliefs about the model and our understanding of the data collection process. If we felt strongly that the population of prices people would be willing to pay should follow a Normal distribution, then we might be less likely to make major modifications to the model. We might examine the data collection process to see if it perhaps led to some bias in the data. However, if the data collection process is sound, then we might be forced to re-examine our model for the population and see what could be changed. In this case, it's likely that our model is inappropriate, especially given that it's difficult to imagine a valid data collection process that might lead to negative values in the data (as the Normal distribution allows).

To close the loop here, we will choose a different statistical model to represent the population, the *Gamma distribution*. This distribution has the feature that it only allows positive values, so it eliminates the problem we had with negative values with the Normal distribution.

Now, we should go back to the top of our iteration and do the following:

1. Develop expectations: Draw a fake picture---what do we expect to see before looking at the data?
2. Compare our expectations to the data
3. Refine our expectations, given what the data show

For your reference, here is a histgram of the same data with the Gamma distribution (estimated using the data) overlaid.



How do the data match your expectations now?

You might ask what difference does it make which model I use to represent the population from which the data were generated? Well, for starters it might affect the kinds of predictions that you might make using the model. For example, recall before that were interested in what proportion of the population might be willing to pay at least $30 dollars for the book. Our new model says that only about 7% of the population would be willing to pay at least this amount (the Normal model claimed 11% would pay $30 or more). So different models can yield different predictions based on the same data, which may impact decisions made down the road.

Note: Parts of this reading were taken from *The Art of Data Science* by Peng & Matsui.

Complete