Day 1: Statistics and Machine Learning

Hi, in this lesson, you will discover the 5 reasons why a machine learning practitioner should deepen their understanding of statistics.  
  
**1. Statistics in Data Preparation**  
Statistical methods are required in the preparation of train and test data for your machine learning model.  
  
This includes techniques for:

* Outlier detection.
* Missing value imputation.
* Data sampling.
* Data scaling.
* Variable encoding.
* And much more.

A basic understanding of data distributions, descriptive statistics, and data visualization is required to help you identify the methods to choose when performing these tasks.  
  
**2. Statistics in Model Evaluation**  
Statistical methods are required when evaluating the skill of a machine learning model on data not seen during training.  
  
This includes techniques for:

* Data sampling.
* Data resampling.
* Experimental design.

Resampling techniques such as k-fold cross-validation are often well understood by machine learning practitioners, but the rationale for why this method is required is not.  
  
**3. Statistics in Model Selection**  
Statistical methods are required when selecting a final model or model configuration to use for a predictive modeling problem.  
  
These include techniques for:

* Checking for a significant difference between results.
* Quantifying the size of the difference between results.

This might include the use of statistical hypothesis tests.​  
  
**4. Statistics in Model Presentation**  
Statistical methods are required when presenting the skill of a final model to stakeholders.  
  
This includes techniques for:

* Summarizing the expected skill of the model on average.
* Quantifying the expected variability of the skill of the model in practice.

This might include estimation statistics such as confidence intervals.​  
  
**5. Statistics in Prediction**  
Statistical methods are required when making a prediction with a finalized model on new data.  
  
This includes techniques for:

* Quantifying the expected variability for the prediction.

This might include estimation statistics such as prediction intervals.​  
  
**Your Task**  
For this lesson, you must list 3 reasons why you personally want to learn statistics.  
  
Post your answer online. I would love to see what you discover.  
  
In the next lesson, you will discover a concise definition of statistics.  
  
Jason.