Why do we need Statistics?

Statistics is a collection of tools that you can use to get answers to important questions about data.

You can use descriptive statistical methods to transform raw observations into information that you can understand and share. You can use inferential statistical methods to reason from small samples of data to whole domains.

As a machine learning practitioner, you must have an understanding of statistical methods.

Raw observations alone are data, but they are not information or knowledge. Data raises questions, such as:

What is the most common or expected observation?
What are the limits on the observations?
What does the data look like?
What variables are most relevant?
What is the difference between two experiments?
Are the differences real or the result of noise in the data?

These Questions Are Important…

The results matter to the project, to stakeholders, and to effective decision making. Statistical methods are required to find answers to the questions that we have about data.

We can see that in order to both understand the data used to train a machine learning model and to interpret the results of testing different machine learning models, that statistical methods are required.

This is just the tip of the iceberg as each step in a predictive modeling project will require the use of a statistical method.

Why is Statistics Important to Machine Learning?

…it is needed at each step of a project

It would be fair to say that statistical methods are required to effectively work through a machine learning predictive modeling project.

Below are 10 examples of where statistical methods are used in an applied machine learning project.

Problem Framing: Requires the use of exploratory data analysis and data mining.
Data Understanding: Requires the use of summary statistics and data visualization.
Data Cleaning. Requires the use of outlier detection, imputation and more.
Data Selection. Requires the use of data sampling and feature selection methods.
Data Preparation. Requires the use of data transforms, scaling, encoding and much more.
Model Evaluation. Requires experimental design and resampling methods.
Model Configuration. Requires the use of statistical hypothesis tests and estimation statistics.
Model Selection. Requires the use of statistical hypothesis tests and estimation statistics.
Model Presentation. Requires the use of estimation statistics such as confidence intervals.
Model Predictions. Requires the use of estimation statistics such as prediction intervals.