Data Science 1
Probability 1
Beginnings

Edward L. Boone

Introduction

In order to understand Data Science we need to understand how data arises.

Typically data arises from an *Experiment* which elicits a response from the system we are considering.

- The *Experiment* could be in a controlled environment where the researcher applies a treatment to a subject to elicit a response. This is considered an Experimental Study.
- The *Experiment* could be in an uncontrolled environment where the researcher simply observes behavior and records the response without any intervention. This is considered an Observational Study.

From a statistics standpoint we want to know whether or not the response or outcome we observe is rare or common.

Lionel Messi

Consider this video where Lionel Messi is attempting to score.

Lionel Messi

Now consider this video where Lionel Messi is attempting to score.

Since we have not yet seen the outcome of the goal attempt we do not know what happened.

Hence we are uncertain about the outcome.

Probability

Probability attempts to understand the uncertainty associated with an Experiment.

Probability does not just consider one outcome of the Experiment, instead it considers all possible outcomes of the experiment.

It then assigns how likely each of the possible outcomes are to occur.

Since it considers all outcomes many people consider it to be abstract and theoretical. And it is!

Sample Space

In the example of Messi trying to score what are the possible outcomes?

- Makes the goal.
- Misses the goal.

The collection of all possible outcomes from the experiment is called the *Sample Space* and is denoted by $\mathcal{S}$. The Sample Space for this example is:

$$\mathcal{S} = \{\text{Goal, No Goal}\}$$

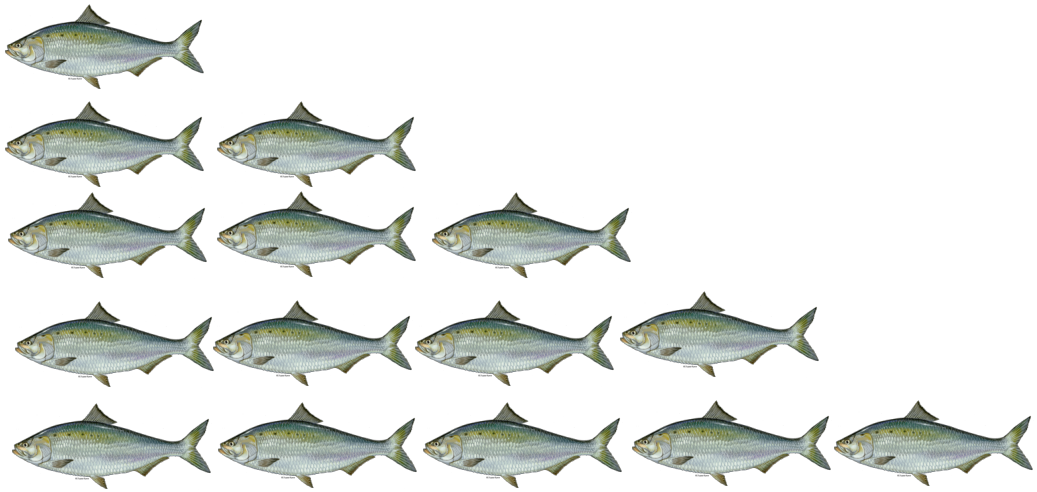
There are no other possibilities that we are interested in considering.

Fishing

Consider the net fisherman here.

Fishing

What are the possible outcomes?



Here it is difficult to write out $\mathcal{S}$ directly.

We need a easier way to think about $\mathcal{S}$.

Random Variables

A *Random Variable* is a function, X, that assigns outcomes A in $\mathcal{S}$ to real valued numbers.

Think of the Random Variable as the scale that measures the outcome.

Some random variables are natural and easy to understand.

If Lionel Messi makes the goal then he scores 1 point other wise he scores 0 points.

$$X = \begin{cases} 1 & \text{if goal} \\ 0 & \text{otherwise} \end{cases}$$

Random Variables

Other random variables are a bit more difficult to define.

Consider the fishing example.

- X could be the number of fish he caught.
- X could be the weight of the fish he caught.
- X could be the value of the fish at market. This may or may not be directly related to the above due to price negotiation.

Random Variables

Other random variables are a bit more difficult to define.

Consider the fishing example.

- X could be the number of fish he caught.
- X could be the weight of the fish he caught.
- X could be the value of the fish at market. This may or may not be directly related to the above due to price negotiation.

Summary

Now that we understand what an experiment is and the idea of a random variable to measure the outcomes in a sample space we can continue on and look at each one more in depth as there are some key differences between the types of random variables.