Data Science 1
Probability
Relative Frequency

Edward L. Boone

Introduction

From the previous videos we understand what a Sample Space $\mathcal{S}$ is and what events are.

Now we want to know how to assign a probability to an event E, denoted $P(E)$.

One way is to use the Empirical approach which looks at the proportion of the time the event was observed.

$$\hat{P}(E) = \frac{n(E)}{n}$$

where n is the number of trials we observed and $n(E)$ is the number of those trials where E occured.

Here we use $\hat{P}$ to show that the probability is estimated from data.

Lionel Messi

Let $\mathcal{S}$ be the sample space of Lionel Messi attempting a goal. Hence, $\mathcal{S} = \{\text{Goal, No Goal}\}$ where the event we are looking for is $E = \{\text{Goal}\}$.

Lionel Messi

Now that we have observed the outcome of $n = 10$ experiments and we observed $n(E) = 3$ we have the empirical probability:

$$\hat{P}(E) = \frac{n(E)}{n} = \frac{3}{10} = 0.3$$

Is his goal completion rate really 0.3?

Lionel Messi

Alternatively, we can go back through history and see how he fared in the Spanish La Liga League.

Season	Goals	Shots
2014/15	43	144
2015/16	26	115
2016/17	37	130
2017/18	33	152
2018/19	36	140
2019/20	19	78
Total	194	759

Data as of 5 May 2020 from
`https://www.infogol.net/en/player/lionel-messi/1529`

Lionel Messi

Using this new information with $n = 759$ experiments and $n(E) = 194$ we have the empirical probability:

$$\hat{P}(E) = \frac{n(E)}{n} = \frac{194}{759} \approx 0.2555$$

Hence there is probability of approximately 0.2555 of Lionel Messi scoring a goal when he takes a shot.

Fishing Example

What is the probability of catching a fish when using net casting?

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

and the Event F is:

$$F = \{\text{Fish}\}$$

Fishing Example

In order to determine $\hat{P}(F)$ we need to see the outcomes from many casts.

Problems

While the Empirical probability also known as *Relative Frequency* approach is intuitive, it is quite expensive as well as other problems.

- It takes time to view each experiment.
- Some experiments are difficult to conduct.
- Rare events take even longer (think volcano eruptions).
- The probability becomes more accurate as the number of experiments increase. Law of Large Numbers:

$$\hat{P}(E) \to P(E) \quad \text{as} \quad n \to \infty$$

Summary

The *Relative Frequency* approach to probability is one way to assign probabilities to events.

- Can be quite expensive.
- Can be difficult.
- May be the only way to estimate the probability of some event.

What if we could come up with a way to estimate a probability with out the need to conduct an experiment?