Probability: Random Variables

Expected value versus Sample mean

Consider the following question:
In what situations should we calculate the average using expected value versus using the sample mean?

The expected value of discrete random variable is specified by its values $x_1, x_2, , x_k$ and associated probabilities $P(X = x_1), P(X = x_2), , P(X = x_k)$. The formula for expected value is:

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k) = \sum_{i=1}^{i=k} x_i P(X = x_i)$$

On the other hand when we observe data from the distribution of X then we have observations $y_1, y_2, , y_n$ (so we have a total of n observations). Then we calculate the sample average by usual formula

$$\bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n}$$

Suppose that a fair six sided die is rolled 5 times, and we get 3, 6, 5, 5 and 4. Then the sample average is

$$\bar{y} = \frac{3 + 6 + 5 + 5 + 4}{5} = 4.6$$

These values are sampled from the probability distribution of X (here X is a value of the die). The probability mass function of X is plotted below:

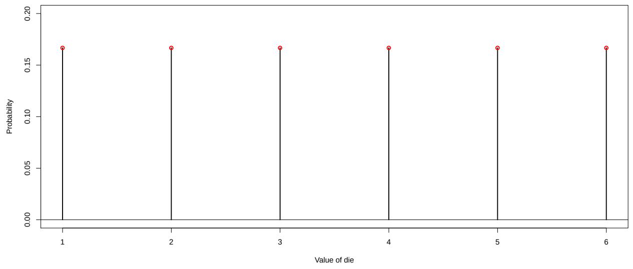

Figure 1: Probability Mass Function of fair dice

Each probability is equal to 1/6. We know that

$$E(X) = 3.5$$

Note that sample mean is different from the expected value. Since sample mean is calculated from random sample, we expect sample means to vary from sample to sample.
We also expect that as sample size increases, the sample mean will approach expected value of X, this important result is called **The Law of Large Numbers**.
To see how this law works, we simulated rolling a fair die different number of times (from 1 up to 100) each time calculating the average. That is what we get:

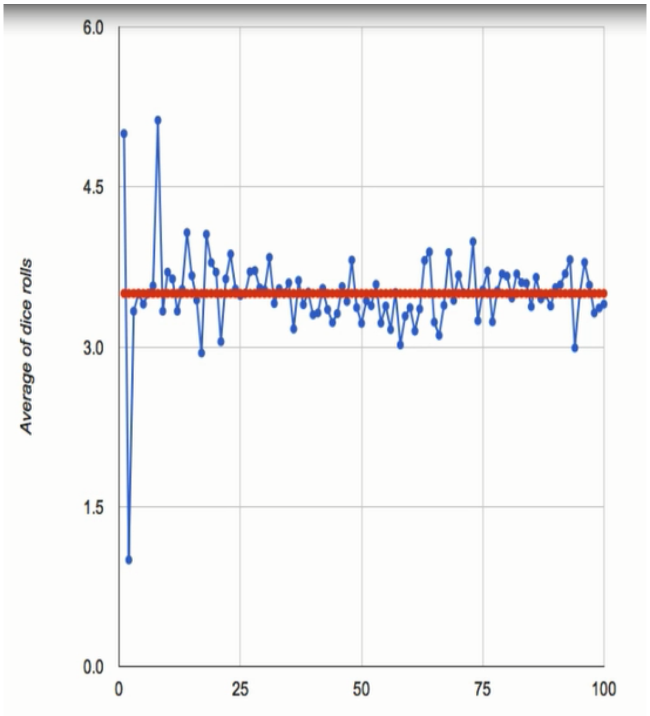

Figure 2: Law of Large Numbers for the dice example

In the x-axis we have number of rolls and y-axis shows average mean, the red line is our expectation which is 3.5.
From this plot we see that as number of rolls increases (as it becomes closer to 100), the average means get closer to the expected value.
This shows how sample mean is related to the expected value $E(X)$.