EXPECTATION, SUMMARY STATISTICS, AND CONDITIONAL PROBABILITIES

1. Probability Density Functions

Discrete Random variables can be characterized by Probability Mass Functions. These functions describe the probabilities of all the values in the sample space of the discrete random variable. For instance, if the discrete random variable in question is the roll of a fair die, which takes values between 1 and 6, we know that

$$P(Y == i) = \frac{1}{6} \forall i \in [1, 6] \tag{1}$$

That is the random variable Y takes an integer value i between 1 and 6, with probability $\frac{1}{6}$ for any given value. For something to be a Probability Mass Function (PMF), the following equation must be true

$$\sum_i P(Y == i) = 1 \tag{2}$$

where the summation is over the sample space of the random variable. We can construct PMFs for all sorts of discrete random variables. For instance, if the random variable represents the outcomes of flipping two coins at random, we can see that there are 3 distinct events in the sample space:(H, H), (H, T), (T, T). That is, getting 2 tails, 2 heads, or 1 head and 1 tail, assuming that the order doesn't matter. The PMF can be expressed as

$$\begin{aligned} P(Y == (HH)) &= \frac{1}{4} \\ P(Y == (TT)) &= \frac{1}{4} \\ P(Y == (HT)) &= \frac{1}{2} \end{aligned} \tag{3}$$

Once again, the summation of the PMF over the sample space is 1. This is the axiom of total probability. The sum of probabilities over the sample space of a random variable has to equal to 1.

In the case of continuous random variables, we don't have a PMF. Instead, we have a continuous function that expresses the probabilities of the various events. This is known as the Probability Density Function or PDF. In the case of a continuous random variable, the number of events in the sample space is infinite and therefore the probability of any distinct event is essentially zero. For instance, if the outside temperature is a continuous random variable that takes values between 0 and 100 degrees Fahrenheit, the probability of any temperature of say 43.25 degrees Fahrenheit is zero. It doesn't make sense to take about individual events but ranges of values. So, the PDF expresses the probability of the continuous random variable being within a range around the value of interest. In other words when we say

$$p(x) = k \tag{4}$$

We are stating that the probability of the random variable Y taking some value in a small neighborhood dx around x is k. That is $p(x - dx/2 \leq x \leq x + dx/2) = kdx$. Similar to the PMF, the summation of the PDF over the sample space is 1. Since these are continuous functions, we need to integrate over the range rather than perform discrete summation.

$$\int_a^b p(x)dx = 1 \tag{5}$$

Where the continuous random variable takes values between a and b, E.g. $a = 0$ and $b = 100$ in the outdoor temperature example we were discussing above. We will encounter integral and differential calculus later in the semester, in case you are not familiar with it already. For now, it suffices to have the intuition that Integration is a summation over the random variable by taking really small intervals (each interval is dx long).

2. Expectation of Random Variables

While PMFs and PDFs are useful to describe random variables, often times we need summary information that captures the essence of the underlying random phenomena. The expected value of random variables (also known as the *mean*) is one such summary statistic that characterizes a random variable. If you perform many trials with the underlying random phenomenon, the expected value states average response you would get. Let's say you experiment with tossing two coins at random many many times. What is the expected value of getting heads? How many heads would you see on average if you toss two coins a large number of times? You can see that this becomes the weighted average of the various events in the sample space. If you toss two coins 1000 times, you'd expect to see only tails about 250 times, two heads about 250 times and 1 head roughly 500 of those trials. The weighted average of seeing heads is therefore $0.25 \times 0 + 0.5 \times 1 + 0.25 \times 2 = 1$. So, the expected value of seeing heads is 1 when you toss two coins at random. You can essentially compute the expected value from the PMF or PDF. It is the weighted average of the PMF.

$$E(x) = \sum_i P(x == i) \times i \tag{6}$$

in the case of a discrete random variable. And for continuous random variables, we get

$$E(x) = \int_a^b xp(x)dx \tag{7}$$

The expected value neatly captures the entire PDF in a single number.

Even if you are not able to compute the expectation using PDF because the form of the PDF is unknown or not analytical, you can always sample from the distribution and estimate the expected value. For instance, the *rand* function in Octave generates uniformly distributed random variables. By uniform we mean that all values between 0 and 1 are equally likely. The expected value of this random variable is 0.5. However, we cannot get a handle on the PDF as implemented by Octave. But we can always sample from this function and use those samples to estimate the expected value. You can easily examine this by getting 1000 random values from *rand* and then computing their mean. You'll see that it is remarkably close to 0.5.

In addition to taking expected values of random variables, we can also computed the expected values of functions of random variables. These are defined the same way the expected value is defined. That is,

$$E(f(x)) = \int_a^b f(x)p(x)dx \tag{8}$$

For instance, we can compute $E(x^n)$, the expected value of the random variable raised to nth power. This is termed the nth moment of x.

3. Variance and Standard Deviation

Knowing the expected value of a random variable is useful. However, it doesn't capture how the PDF is distributed in the sample space. For instance, consider two uniform random variables: x that is uniformly distributed between 0 and 100, and y that is uniformly distributed between 25 and 75. Both x and y have an expected value of 50. But, they have very different behaviors within their ranges. To capture a measure of how "spread" is a random variable around its expected value, we can use *variance* or *standard deviation.*

Variance is defined as follows:

$$Var(x) = E((x - E(x))^2) \tag{9}$$

This can be simplified to

$$Var(x) = E(x^2) - E(x)^2 \tag{10}$$

If you compute the variance of x and y, you'll see that x which ranges from 0 to 100 has a much higher variance (in fact 4 times as much) compared with y.

The standard deviation of a random variable x is defined as the $\sqrt{Var(x)}$. It measures the scale of x and is expressed in the same units as x.

4. Conditional Probabilities

Conditional probability of an event is defined as the probability of an event occurring given the knowledge of some related event occurring. For instance $P(B|A)$ is the probability of event B occurring given the knowledge that an event A has already occurred. When the two events are independent of each other, $P(B|A)$ is simply $P(B)$. That is, knowing that event A has occurred doesn't change the probability of event B occurring. Examples of independent events are the event of getting a H in the first coin toss and getting a H in the second coin toss. Knowing that you already got a H in the first coin toss doesn't affect in any way the probability of getting a H in the second coin toss. On the other hand, if you draw a Spade from a suit of 52 cards, the probability of drawing a second spade given that the first card is a Spade is $\frac{12}{51}$. Whereas, the probability that the second card is a Spade when the first card is not a Spade is $\frac{13}{51}$.

If events A and B are not independent, the probability of the joint event $P(A \cap B)$ is $P(A)P(B|A)$. If A and B are independent, $P(A \cap B) = P(A)P(B)$

Here are some examples of conditional probabilities. Consider that you are taking a census survey. From this, you gather than the population has a uniform distribution of heights between 5 and 6 feet. However, you find that men range from 5.2 to 6 feet in height and women range from 5 to 5.8 feet in height, and men and women are equally probable in the population. Here, we have two conditional PDFs. $P(H|male)$ is a uniform distribution with a mean of 5.6 feet and $P(H|female)$ is another uniform distribution with a mean of 5.4 feet. Similarly, you might find that other factors such as age, income, education level, socio-economic status etc are all inter-related. They are not truly independent of each other and you may be able to compute the expected values and standard deviations of various factors given a conditioning random variable such as socio-economic status, education level, gender etc.