CS 547 Lecture 7: Discrete Random Variables

Daniel Myers

The Probability Mass Function

A *discrete* random variable is one that takes on only a countable set of values. A discrete RV is described by its *probability mass function* (pmf),

$$p(a) = P(X = a)$$

The pmf specifies the probability that random variable X takes on the specific value a.

Recall our coin-flipping example. If we flip three coins and count the number of heads that appear, we obtain the following pmf:

$$P(0\ heads) = \frac{1}{8}$$

$$P(1\ head) = \frac{3}{8}$$

$$P(2\ heads) = \frac{3}{8}$$

$$P(3\ heads) = \frac{1}{8}$$

All questions about the behavior of a discrete random variable can be answered using its pmf.

Total Probability

Observe that the probabilities in the number-of-heads pmf add up to 1. Because the random variable must always take on one of its values with non-zero probability, the sum of all its non-zero probabilities must be 1. Mathematically,

$$\sum_{x:p(x)>0} p(x) = 1,$$

where the notation can be read as "sum over the values of x such that $p(x)$ is greater than zero".

Bernoulli Trials

The Bernoulli trial is a simple discrete random variable with only two possible outcomes: 0 and 1. The RV has one parameter, p, and its pmf is

$$P(X = 0) = 1 - p$$
$$P(X = 1) = p$$

You can think of the Bernoulli trial as flipping a weighted coin that comes up heads with probability p and tails with probability $1 - p$. Bernoulli trials are also used to model randomized algorithms that are not guaranteed to return the correct answer to a question.

The Geometric Random Variable

Suppose we perform a series of independent Bernoulli trials, each with parameter p. The geometric random variable describes the number of trials required until we obtain our first success. Its pmf is given by

$$P(X = k) = (1 - p)^{k-1}p$$

That is, the probability that we obtain our first success on the k^{th} trial is the probability of getting $k - 1$ failures (each with probability $1 - p$), and a success on trial k.

Suppose we want to model packet loss in a computer network. If the probability of dropping a packet is p and each packet is independent of the others, then we can model the number of packets sent before a drop as a geometric random variable. For example, if $p = .01$, the probability that we drop the 10^{th} packet is

$$P(X = 10) = (1 - .01)^{10-1}(.01) \approx .009$$

What is the probability that we send more than 10 packets successfully? This can be calculated using total probability. If we send more than 10 packets, then a drop *did not* occur on any of the first 10 trials. The probability is thus

$$P(X > 10) = 1 - \sum_{i=1}^{10} P(X = i) = 1 - \sum_{i=1}^{10}(1 - p)^{i-1}p$$

Expected Value

The *expected value* of a random variable is a weighted average of its possible values.

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

Each value x is weighted by the probability that the random variable X actually takes the value x. Thus, the values that occur most frequently make the greatest contribution to the expected value.

The expected value serves as a measure of *centrality* for a random variable's distribution.

For the number-of-heads example given above, the expected value is

$$E[\textit{number of heads}] = \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 1 + \frac{3}{8} \cdot 2 + \frac{1}{8} \cdot 3 = 1.5$$

Note that the expected value is fractional – the random variable may never actually take on its average value!

Expected Value of a Geometric Random Variable

For the geometric random variable, the expected value calculation is

$$E[X] = \sum_{k=1}^{\infty} k\,P(X = k) = \sum_{k=1}^{\infty} k(1 - p)^{k-1}p$$

Solving this expression requires dealing with the infinite sum. Examining a table of summations shows the following result, which is very close to the correct form:

$$\sum_{i=1}^{\infty} i\,x^i = \frac{x}{(1 - x)^2}$$

Applying a little manipulation brings the expected value's sum into the correct form. We can then use the result from the table of summations and simplify to obtain the final expected value.

$$\begin{aligned}
E[X] &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\
&= \frac{p}{1-p}\sum_{k=1}^{\infty} k(1-p)^{k} \\
&= \frac{p}{1-p}\frac{1-p}{(1-(1-p))^2} \\
&= \frac{1}{p}
\end{aligned}$$

Notation

We'll frequently use $\overline{X}$ in place of $E[X]$, especially in calculations involving other expected values. Many statistics texts use μ to represent the mean of a random variable, but we'll avoid that notation since we've already established a convention of μ as the service rate of a queue.

Properties of the Expected Value

The expected value is a *linear* operation. Scaling or shifting a random random variable simply scales and shifts the expected value by the same amount:

$$E[aX+b] = aE[X]+b$$

The expected value of the sum of two (or more) random variables is simply the sum of their individual expected values:

$$E[X+Y] = E[X]+E[Y]$$

This result holds even when X and Y are not statistically independent.

Variance

The expected value of a random variable provides a description of its average behavior, but it tells us nothing about how much it varies[1]. This is important, because variability is *the* key driver of performance and uncertainty in systems. Even if the system has acceptable average behavior, extreme variability can lead to all kinds of performance problems.

The *variance*, σ^2, of a random variable is defined as the expected variation of a random variable from its own mean,

$$\sigma^2 = E[(X-\overline{X})^2]$$

The square root of the variance, σ, is called the *standard deviation.*

By expanding the squared term and using the linearity properties of the expected value, we can write the variance as,

$$\sigma^2 = E[X^2] - \overline{X}^2$$

That is, the variance can be calculated as the *average of the squares* minus the *square of the average.*

[1] As I once heard an operations research professor quip, "Even MBA students know a mean is useless without a measure of variability."

The term $E[X^2]$ is called the *second moment* of X. The basic expected value $E[X]$ is the first moment. We can define additional *higher moments*: $E[X^3]$ is the third moment, $E[X^4]$ is the fourth moment, and so forth.

Each moment is associated with a different property of the distibution. The first moment is a measure of centrality. The second moment is associated (through the variance) with the spread of the distribution. The third moment is associated with the *skew* of the distribution – whether it is symmetric or asymmetric about the mean. The fourth moment is associated with *kurtosis*, a measure of the "peakedness" of the distribution. Specifying the moments is an alternate way of characterizing the behavior of a random variable. In practice, the first two moments are by far the most important.

Coefficient of Variation

The variance tells us something about the spread of a random variable, but this result has to be interpreted in the context of the variable's overall scale. For example, $\sigma^2 = 1000$ is a huge variance if the mean of the variable is 1, but less significant if the mean is 1000000.

The *squared coefficient of variation* combines the variance and mean into a single number:

$$c_X^2 = \frac{\sigma^2}{\overline{X}^2}$$

Coefficients larger than 1 are associated with high variability. The smaller the coefficient, the less variability in the variable, and a coefficient of 0 indicates a deterministic variable that always takes the same value. The case where $c_X^2 = 1$ is particularly important in queueing systems and is associated with the exponential distribution.