Discrete Probability Notes

CMS 380 Simulation and Stochastic Modeling

Elements of Probability

This section is primarily for general background. Read Timo Seppäläinen's notes for a more thorough mathematical overview.

The goal of probability theory is to reason about the outcomes of experiments. Here, *experiment* is a general term that covers pretty much any observation we'd like to make about the world, ranging from very simple contexts like dice rolls and coin flips to challenging measurements of complex physical systems. Probability theory provides tools to reason about the range of possible outcomes for an experiment, including the relative likelihood of its possible outcomes.

First, some basic terminology. The result of an experimental observation is called an *outcome*. The set of all possible outcomes for a particular experiment is called the *sample space*, usually denoted by Ω. A single outcome ω is therefore an element of the sample space, $\omega \in \Omega$.

For example, suppose we flip two coins. The sample space consists of four possible outcomes: $\{HH, HT, TH, TT\}$.

If we roll a standard six-sided die, the sample space consists of the six values, $\{1, 2, 3, 4, 5, 6\}$. If we rolled two six-sided dice, the sample space would have 36 possible outcomes (six possible values for the first die times six for the second die).

It's also possible for the sample space to be infinite, although this necessarily makes things more abstract. Consider, for example, the experiment of tossing an *infinite* number of coins. The outcome of the experiment could be thought of as a vector with infinitely many elements,

$$(x_1, x_2, x_3, \ldots)$$

where each x_i is the outcome of one coin toss. A little later, we'll talk about queueing systems, where the number of customers waiting in line is (theoretically) unbounded and can take on any value $0, 1, 2, 3, \ldots$

An *event* is a set of possible outcomes that share some quality of interest. In set notation, event $E \subseteq \Omega$. For example,

- If we flip three coins, the event *Exactly one coin is heads* corresponds to three outcomes, $\{TTH, THT, HTT\}$.
- If we roll a six-sided die, the event *The result is even* corresponds to three outcomes, $\{2, 4, 6\}$.
- The event *The roll is greater than 4* corresponds to $\{5, 6\}$.

Intuitively, the probability of an event E, denoted $P(E)$, is the fraction of time event E would occur if we performed an infinitely large number of experiments. Letting N_E represent the number of observations that result in E,

$$P(E) = \lim_{N \to \infty} \frac{N_E}{N}$$

This definition is intuitively reasonable, but makes several technical assumptions about the existence of the limit and the fact that the ratio converges to a single value for all possible sequences of experimental outcomes. Modern theorists use a set of *axioms of probability* as the building blocks of probability theory. We're not going to discuss the axioms here, since we don't need them for anything else in the class, but you can read about them in the more technical notes if you're interested.

Random Variables

Very often, we're less interested in the outcome of an experiment than we are in the value of some function calculated from the outcome.

- For example, roll two dice and add their faces. The actual experimental outcome is result of the roll, but the value of interest is the sum calculated from the two dice.

- Flip 10 coins and count the number that come up heads. The outcome is the 10-element vector describing each coin's value, but the quantity of interest is the number of heads. Different experimental outcomes will yield the same number of heads.

- Flip a coin repeatedly and report the number of trials required until the first head comes up.

- Measure the latency of a disk request in a database system. This depends in a complex way on the state of the system while the request is being processed, but the quantity of interest is ultimately a single number.

A *random variable* is a function that associates a number with the outcome of an experiment. More formally, it is a real-valued function defined on the sample space. In most cases, you can reason about a random variable of interest without worrying about the technical details of the underlying sample space.

Continuous random variables can take real numbers as values and are useful when modeling physical quantities like time or distance. *Discrete* random variables take values from a finite or countably infinite set. *The rest of this note deals with discrete variables*; we'll talk about continuous random variables in the next unit.

Probability Mass Functions

Recall that we wanted to develop the theory of probability to reason about the outcomes of experiments. Because the value taken by a random variable depends on the outcome of an experiment, *we can also use probability to reason about random variables*.

The set of probabilities associated with the values of a random variable is called the *distribution* of the variable. Describing the distribution of a RV totally summarizes its behavior and allows us to draw inferences about it. Let X be a discrete random variable (it's common to use capital letters to represent random variables). Define the *probability mass function* of X to be a function p

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

That is, the pmf p(a) specifies the probability that random variable X takes on value a.

Consider the example of flipping three coins and counting the number of heads. The pmf of this random variable is

$$p(0\ heads) = 1/8$$

$$p(1\ head) = 3/8$$

$$p(2\ heads) = 3/8$$

$$p(3\ heads) = 1/8$$

These probabilities come directly from the sample space of the experiment: one of the eight possible outcomes yields zero heads, three of the eight yield one head, and so forth.

For simple experiments like this one, it may be possible to write down the entire pmf. In other cases, it's more convenient to express the pmf as a formula.

***Question**. Write down the complete pmf for the random variable representing the sum of two six-sided dice.*

Total Probability

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

$$\sum_x p(x) = 1$$

Total probability can often be used to simplify calculations. For example, what's the probability that the sum of two dice returns a value greater than 3?

A straightforward way to approach this question is to use the pmf to add up the probabilities for all values greater than 3:

$$P(sum\ greater\ than\ 3) = p(4) + p(5) + \ldots + p(12)$$

This is fine, but can become cumbersome in other problems if the sum contains too many terms (or is infinite). Use total probability to rewrite the expression as $1 - P(sum\ is\ less\ than\ or\ equal\ to\ 3)$.

$$1 - (P(2) + P(3)) = 1 - 1/36 - 2/36 = 33/36$$

You'll see more examples using this strategy later.

Bernoulli Trials

The *Bernoulli trial* is a simple discrete random variable with only two possible outcomes: 0 and 1. It is used in modeling situations where there are two outcomes of interest, such as success or failure in a game, or randomized algorithms that are not guaranteed to return the correct answer every time. It is also the building block of more complex random variables, like the geometric and binomial distributions described below.

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 $1 - p$.

Expected Values and Variance

The *expected value* of a random variable is a weighted average of its possible values. It is an important summary of the variable's behavior, because it represents the “average” or “most likely” value. You can think of it as the random variable’s equivalent of the traditional arithmetic mean that we’ve already used to summarize quantitative data.

The formula for the expected value of a random variable X is

$$E[X] = \sum_x xp(x)$$

Thus, each possible value x is weighted by the probability that $X = x$. Values that occur most frequently tend to have a strong influence on the expected value, but a very large x can still influence the calculation even if its associated $p(x)$ is small.

What is the expected number of heads obtained by flipping 3 coins?

$$E[X] = \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 1 + \frac{3}{8} \cdot 2 + \frac{1}{8} \cdot 3 = \frac{12}{8} = 1.5$$

The expected number of heads is 1.5. Notice that the expected value can be fractional even though the RV can only take whole number values.

What is the expected value of rolling a six-sided die?

$$E[X] = \sum_{x=1}^{6} \frac{1}{6}x = \frac{21}{6} = 3.5$$

Again, the expected value can be fractional.

What is the expected value of the Bernoulli trial with parameter p?

$$E[X] = 0 \cdot (1 - p) + 1 \cdot p = p$$

When asked to calculate an expected value, always start by writing down the definition, then substituting in the value of the pmf, $p(x)$. If the domain of the RV is finite, you can usually calculate $E[X]$ directly. Working with infinite sums is more challenging; you often have to do some manipulation to bring the sum into a form that has a known result. Take a look at the next section for some examples.

There are two common ways to notate the expected value of a random variable X: $E[X]$ and $\overline{X}$. You may have seen notation similar to the latter in statistics books, where the bar is used to indicate an average.

Properties of the Expected Value

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

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

Question. *Prove the linearity of the expected value operation. Hint: start by writing down the definition, using* $ax + b$ *for the value.*

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.

Question. *What is the expected value of the sum of two six-sided dice?*

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,

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

Expanding the inner term and applying the linearity of the expected value

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

yields a convenient formula:

$$\begin{aligned}\sigma^2 &= E[X^2 - 2X\overline{X} + \overline{X}^2] \\ &= E[X^2] - 2\overline{X}E[X] + E[\overline{X}^2] \\ &= E[X^2] - 2\overline{X}^2 + \overline{X}^2 \\ &= E[X^2] - \overline{X}^2\end{aligned}$$

Note: the move from line 2 to line 3 uses the fact that $\overline{X}^2$ is a scalar value and that $\overline{X}$ and $E[X]$ are alternate names for the same value.

The convenient rule for calculating the variance is therefore

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

The variance is the *average of the squares* minus the *square of the average*.

***Question**. Calculate the variance of the Bernoulli trial with parameter p. Use the definition of the expected value to calculate $E[X^2]$.*

A point of notation: if we're dealing with only one RV, it's fine to just write σ^2. If you have multiple RVs in one problem, you can use subscripts to distinguish them: σ^2_X, σ^2_Y, etc.

Higher Moments

$E[X^2]$ has a special name: the *second moment* of X. The regular expected value, $E[X]$, is the *first moment*. It's also possible to define 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 distribution. 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.

Higher moments come up in some applications, because you can completely characterize the behavior of a random variable by specifying its moments.

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 10 million.

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 relative to the mean. 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 (it has no variance).

$c_X^2 = 1$ is a special case: it's associated with the exponential distribution, which is very important for analyzing queueing systems. We'll learn more about the exponential in the next unit on continuous probability.

***Question**. What is the squared coefficient of variation of the Bernoulli trial as a function of p?*

The Geometric Distribution

Consider a series of independent Bernoulli trials, each with parameter p. The *geometric random variable* describes the number of trials required to obtain the first success. Its pmf is

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

The first trial succeeds with probability p, so

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

. If $X = 2$, then the first trial must have failed and the second trial succeeded,

$$P(X = 2) = (1 - p)\, p$$

Similarly, if $X = 3$, then the first two trials must have failed before the third succeeded, so

$$P(X = 3) = (1 - p)^2\, p$$

In general, obtaining success on the kth trial requires that the first $k-1$ trials be failures.

Practice Problems

Packet Loss*. 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, then the number of packets sent before a drop is a geometric random variable. What is the probability of dropping the 10th packet as a function of p?*

This is a direct application of the pmf.

$$P(X = 10) = (1 - p)^{10-1}(p)$$

Here's a trickier problem: what is the probability that the first drop occurs any time after sending five packets successfully?

If at least five packets have been sent successfully, then a drop *did not* occur while sending the first five packets. Using total probability:

$$P(X > 5) = 1 - \sum_{k=1}^{5} (1 - p)^{k-1} p$$

You can use this expression to calculate the exact result for any value of p.

Your First Urn Problem*. Filling urns with balls is a common hobby among probability theorists. An urn contains R red balls and B black balls. What is the probability of drawing a red ball at random from the urn?*

The probability of getting an red ball is the fraction of red balls:

$$p = \frac{R}{R + B}$$

Suppose you draw balls from the urn ***with replacement****, returning each ball after it is drawn. What is the probability that you draw the first red ball on your third try?*

This is again a straightforward calculation from the pmf, with the slight twist of having the parameter p be calculated in terms of R and B:

$$P(X = 3) = \left(1 - \frac{R}{R+B}\right)^2 \frac{R}{R+B}$$

What is the probability that you need more than three tries to get the first red ball? Use total probability, as in the previous example.

***Disks**. Suppose that a datacenter has n disks. During a random month, each disk has an independent 5% chance of failing. What is the probability that a disk survives for exactly a year and then fails? What is the probability that a disk survives for more than three months?*

Expected Value

Calculating the expected value of the geometric RV is a good example of working with more complex summations. Using the pmf and the definition of the expected value:

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

What should we do with that infinite sum?

First rule of everything: don't panic when you see something that looks difficult. There are techniques for simplifying summations, but lots of useful results have already been collected into tables of summations. Examining a table of sums will turn up the following result, which is almost what we need:

$$\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}$$

Let $x = 1 - p$. The only issue is the exponent, which is $k - 1$ in the expected value formula, but needs to be k for the table summation to work. Divide by

$1 - p$ to bring the sums into agreement.

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

The final result turns out to be simple:

$$E[X] = \frac{1}{p}$$

For example, if the probability of success is $p = .5$, we'd expect to obtain success on the second trial. If $p = .10$, we'd expect to obtain the first success on the tenth trial.

The Binomial Distribution

Consider the experiment of performing n independent Bernoulli trials, each with parameter p. The *binomial distribution* describes the probability of obtaining k successes out of the n trials. Its pmf is

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

The pmf uses the binomial coefficient: the number of ways to select k items from a fixed set of n elements.

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

Here's how to interpret the pmf. If k of the n trials result in successes, there must be k successes and $n - k$ failures. Successes occur with probability p and failures occur with probability $1 - p$, so the combined probability must be

$$(1-p)^{n-k}p^k$$

The binomial coefficient accounts for the fact that there may be multiple ways to select the k successes out of the n trials—the binomial distribution doesn't care about *which* of the n trials yielded the k successes, only that a total of k successes occurred.

Practice Problems

Failure Analysis. *A communication system has 5 independent components, each of which is working correctly each day with probability 99%. The system can function as long as 3 or more components work correctly. What is the probability that the system fails on a randomly chosen day?*

This kind of problem is a common application of the binomial distribution. First, note that the quantity we want is

$$P(X = 3) + P(X = 4) + P(X = 5)$$

because the system works if three or more components function. Calculating the terms is a straightforward application of the pmf:

$$P(X = 3) = \binom{5}{3}(.01)^{5-3}(.99)^{3} \approx .00097$$

The other calculations are similar. Adding up all three probabilities gives a result greater than 99.99%.

Screws. *A company makes screws that are defective with probability .01, independent of one another. The company sells these screws in packs of 10, with an offer to replace any pack that has more than 1 defective screw. Derive an expression for the proportion of packs the company must replace.*

Let us note, in passing, that a failure rate of .01 for manufacturing screws is *insanely* high. Like, complete lunacy.

This is a case where using total probability is effective. The pack must be replaced if there is more than one defective, so

$$P(\text{replacement}) = 1 - P(0 \text{ defects}) - P(1 \text{ defect})$$

The only tricky thing about this problem is deciding what counts a a “success”. If we take each screw to be a Bernoulli trial with $p = .01$,

$$P(X = 1) = \binom{10}{1}(.99)^{10-1}(.01)^{1}$$

This formulation treats a defective screw as a “success” for the purposes of making the Binomial calculation.