Discrete Probability Distributions

CMS 380 Simulation and Stochastic Modeling

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 $1 - p$.

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.

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.

The Poisson Distribution

The Poisson distribution is frequently used to model the occurrence of random events over time. These events could be things like arrivals to a queueing system, requests submitted to a web server, occurrences of natural disasters, or particles emitted by radioactive isotopes. The Poisson distribution is discrete and it has a single parameter $\lambda > 0$, which is called the **rate** of the distribution.

Suppose we're interested in the probability that k events occur in a period of length t. If the events occur at **constant rate**, the distribution of the number of events that occur during the period is Poisson. The probability of k events occurring in time t is given by the Poisson pmf:

$$P(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

In some applications, the length of the period t is considered to be fixed and the pmf can be simplified into

$$P(X = k) = \frac{e^{-\lambda}(\lambda)^k}{k!}$$

The pmf appears to be quite a bit more complicated than the geometric and binomial mass functions, and it doesn't lend itself to an easy interpretation. The short description of the derivation of the Poisson distribution is that it's actually related to the binomial distribution.

It can be shown that a binomial distribution with a large number of trials n and a small probability of success p is approximately equal to a Poisson

distribution with rate $\lambda = np$. The Poisson pmf comes from setting

$$p = \frac{\lambda}{n}$$

in the binomial pmf and taking the limit as $n \to \infty$.

Expected Value

The expected number of events in a period of length t is λt. If the period length t is fixed, the expected value of the Poisson distribution is simply λ.

This makes sense if we interpret λ as the rate of events:

- *Customers arrive to a queueing system at a constant rate of 2 per minute. How many would we expect to arrive in an hour?* If the rate of arrivals is constant, we'd expect 120 customers to arrive in an hour.
- *I receive an average of eight e-mails per hour. How many e-mails would I expect to receive in 15 minutes?* The rate is $\lambda = 8$ per hour and the interval is .25 hours: two e-mails in 15 minutes.

Constant Rate

A key assumption of the Poisson distribution is that events occur at **constant rate** λ. Understanding the concept of "constant rate" turns out to be a little tricky, but is important for reasoning about other applications of the Poisson distribution to arrival streams.

First, "constant rate" **does not** mean that the time between events is constant! The time between Poisson events is a random variable: there can be shorter or longer interevent times. However, the mix of shorter and longer interevent times balances out on average. Any long gap between events must eventually, in the long run, be counterbalanced by a sequence of bursty events with short interevent times so that the overall rate remains λ.

if events occur at constant rate λ, then the expected number of events that occur in a period of length t is always λt, no matter how long or short t is. Further, all non-overlapping periods of length t are statistically equivalent to

each other, so the process does not change over time and the expected value relationship holds for all periods of length t.

More E-mails. *Suppose that I receive an average of 12 e-mails per hour. What is the expected time between e-mails?*

If events occur at constant rate λ, the expected time between events is $1/\lambda$. Therefore, I expect to receive one e-mail every six minutes on average.

Practice Problems

Floods *A certain region experiences flooding at a rate of one major flood every 100 years. What is the probability of going 100 years without a major flood if the number of floods per century is Poisson distributed?*

We're interested in the case of zero floods when floods occur at a rate of one per century. This is a case where the interval t is fixed at one century, so we don't need to consider it. Applying the Poisson pmf yields

$$P(X = 0) = \frac{e^{-1}(1)^0}{0!} \approx .3678$$

There's about a 37% probability of no major flood in 100 years.

What's the probability of having more than one flood event in 100 years?

Use total probability:

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

Performing the calculations will show that $P(X = 0)$ and $P(X = 1)$ are both approximately 37%, so the probability of more than one flood event is about 26%.

Stop Sending Me E-mails *I receive e-mails at a constant rate of 10 per hour. What's the probability that I make it half an hour without getting any e-mails?*

This is equivalent to the probability of receiving no e-mails in a period of length $t = .5$ hours with $\lambda = 10$ per hour. Applying the Poisson pmf:

$$P(X = 0) = \frac{e^{-10 \cdot .5}(10 \cdot .5)^0}{0!} \approx .0067$$

Alas, there is only a .6% chance I can go an entire half hour without receiving an e-mail.