Some Important Distributions

There are four important distributions in probability: binomial, geometric, Poisson and Gaussian. We have covered the first one. We will now cover the second one. Later, we will talk about the Poisson and Gaussian distributions.

Geometric Distribution

Consider the experiment where we flip a coin until we see a head.

\[ \Omega = \{H, TH, TTH, TTTTH, \cdots\} \]

Let \( X \) denote the variable that counts the number of flips until and including the first Heads we obtain. There is a one to one correspondence between the outcomes and the values that the random variable \( X \) takes on.

Assuming the results of the flips \( X_1, X_2, \ldots \) are independent, \( P(X = 1) = P(X_1 = H) = p, \ P(X = 2) = P(X_1 = T, X_2 = H) = (1 - p)p \). More generally,

\[ P(i - 1 \text{ Tails before a Heads}) = (1 - p)^{(i - 1)}p. \]

So, the random variable \( X \) has the following distribution:

\[ P_X(i) = P(X = i) = (1 - p)^{(i - 1)}p \text{ for } i = 1, 2, \cdots \]

A chance of getting a large value with a Geometric random variable is very small (See figure 1).

![Figure 1: The Geometric distribution.](image-url)
**BitTorrent Servers’ Example**

A video is broken down into $n$ chunks. Each server has a random chunk, i.e. one out of $n$ possible choices. We are interested in the number of servers we need to query to have the whole movie (meaning all $n$ chunks). Let $X$ be the number of servers we query before we get the $n$ chunks. How do we analyze this problem? How do we compute $E[X]$? Of course we will need to query at least $n$ servers. But because some servers will store the same chunk, we may need more. The question is, how many more on the average?

There is a natural way of breaking this random variable into a sum of simpler random variables. We make progress whenever we get a **new** chunk. We write $X = Z_1 + \cdots + Z_n$ in order to use linearity of expectation. Let $Z_1$ be the time to get the first new chunk. The first server we query will give a new unobserved chunk no matter what. Therefore $Z_1 = 1$, and $Z_1$ is a deterministic random variable. $Z_2$ is the number of extra servers you query until you see the second new chunk, and similarly $Z_n$ is the number of servers you query until you get the last chunk. The chunks are not ordered here, it does not matter which one comes first as long as we have all $n$ of them at the end.

Let us now determine the distribution of $Z_2$ to compute its expectation. $Z_2 \sim Geom(p)$ where $p$ is the success probability. The probability of success is the probability of getting a new chunk. When you sample the next server, you will find a new chunk with a high probability: $\frac{n-1}{n}$. The expected value of a geometric random variable is $\frac{1}{p}$. Therefore, $E[Z_2] = \frac{n}{n-1}$ which is slightly greater than 1. Similarly, $Z_3 \sim Geom(\frac{n-2}{n})$. In general, $X_i \sim Geom(\frac{n-i+1}{n})$ and $E[Z_i] = \frac{n}{n-i+1}$. In particular, $Z_n \sim Geom(\frac{1}{n})$.

Finally, we can compute the desired expectation

$$E[X] = E[Z_1] + \cdots + E[Z_n] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\left[\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{1}\right].$$

If $n$ is very large, the number in brackets is roughly equal to $\ln n$ (which can be seen by looking at the area under the curve $\frac{1}{x}$). This means we have to query a factor of $\ln n$ more servers than the case where we know exactly where the chunks are (in which case we need only to query $n$ servers.) This is the price of randomness in the protocol.

This is an important application of the geometric distribution. In general, once you figure out the relevant distribution of a random variable in a problem, you just need to determine the parameter of that distribution.