Will Monroe
CS 109

Lecture Notes #9
July 15, 2017

More Discrete Distributions

Based on a chapter by Chris Piech

Geometric Distribution

X is a **geometric random variable** ($X \sim \text{Geo}(p)$) if X is number of the independent trials until the first success and p is probability of success on each trial. If $X \sim \text{Geo}(p)$:

$$P(X = n) = (1 - p)^{n-1} p$$
$$E[X] = 1/p$$
$$\text{Var}(X) = (1 - p)/p^2$$

The PMF, $P(X = n)$, can be derived using the independence assumption. Let E_i represent the event that the i-th trial succeeds. Then the probability that X is exactly n is the probability that the first $n - 1$ trials fail, and the n-th succeeds:

$$\begin{aligned} P(X = n) &= P({E_1}^C {E_2}^C \dots {E_{n-1}}^C E_n) \\ &= P({E_1}^C) P({E_2}^C) \dots P({E_{n-1}}^C) P(E_n) \\ &= (1 - p)^{n-1} p \end{aligned}$$

A similar argument can be used to derive the CDF, the probability that $X \le n$. This is equal to $1 - P(X > n)$, and $P(X > n)$ is the probability that at least the first n trials fail:

$$\begin{aligned} P(X \le n) &= 1 - P(X > n) \\ &= 1 - P({E_1}^C {E_2}^C \dots {E_n}^C) \\ &= 1 - P({E_1}^C) P({E_2}^C) \dots P({E_n}^C) \\ &= 1 - (1 - p)^n \end{aligned}$$

Example 1

In the *Pokémon* games, one captures Pokémon by throwing Poké Balls at them. Suppose each ball independently has probability $p = 0.1$ of catching the Pokémon.

Problem: What is the average number of balls required for a successful capture?

Solution: Let X be the number of balls used until (and including) the capture. $X \sim \text{Geo}(p)$, so the average number needed is $E[X] = 1/p = 10$.

Problem: Suppose we want to ensure that the probability of a capture before we run out of Poké Balls is at least 0.99. How many balls do we need to carry?

Solution: We want to know n such that $P(X \le n) \ge 0.99$.

$$\begin{aligned} P(X \le n) = 1 - (1-p)^n &\ge 0.99 \\ (1-p)^n &\le 0.01 \\ \log[(1-p)^n] &\le \log 0.01 \\ n\log(1-p) &\le \log 0.01 \\ n &\ge \frac{\log 0.01}{\log(1-p)} = \frac{\log 0.01}{\log 0.9} \approx 43.7 \end{aligned}$$

So we need 44 Poké Balls. (Note that we flipped the inequality on the last line because we divided both sides by $\log(1-p)$. Since $1 - p < 1$, we know $\log(1-p) < 0$, so we're dividing by a negative number!)

Negative Binomial Distribution

X is a **negative binomial random variable** ($X \sim \text{NegBin}(r, p)$) if X is the number of independent trials until r successes and p is probability of success on each trial. If $X \sim \text{NegBin}(p)$:

$$\begin{aligned} P(X = n) &= \binom{n-1}{r-1} p^r (1-p)^{n-r} \text{ where } r \le n \\ E[X] &= r/p \\ \text{Var}(X) &= r(1-p)/p^2 \end{aligned}$$

Example 2

Problem: A grad student needs 3 published papers to graduate. (Not how it works in real life!) On average, how many papers will the student need to submit to a conference, if the conference accepts each paper randomly and independently with probability $p = 0.25$? (Also not how it works in real life...though the NIPS Experiment[1] suggests there is a grain of truth in this model!)

Solution: Let X be the number of submissions required to get 3 acceptances. $X \sim \text{NegBin}(r = 3, p = 0.25)$. So $E[X] = \frac{r}{p} = \frac{3}{0.25} = 12$.

Hypergeometric Distribution

The remaining three distributions appear occasionally; you don't have to master them for this course, but it can be useful to know they exist.

X is a **hypergeometric random variable** ($X \sim \text{HypG}(n, N, m)$) if X is the number of red balls drawn when n balls are drawn at random, *without replacement*, from an urn with N balls total, m of which are red. If $X \sim \text{HypG}(p)$:

$$\begin{aligned} P(X = k) &= \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \text{ where } 0 \le k \le \min(n, m) \\ E[X] &= n\frac{m}{N} \\ \text{Var}(X) &= \frac{nm(N-n)(N-m)}{N^2(N-1)} \end{aligned}$$

[1] `http://blog.mrtz.org/2014/12/15/the-nips-experiment.html`

Benford Distribution

Benford's law says that "naturally occurring" numbers have an uneven distribution of their *first digits*. This is because, roughly speaking, many collections of numbers are not evenly distributed, but rather their *logs* are evenly distributed. The law says that the fraction of numbers with a first digit of 1 is usually close to $\log_{10}\left(1+\frac{1}{1}\right) \approx 0.301$, the fraction with a first digit of 2 is close to $\log_{10}\left(1+\frac{1}{2}\right) \approx 0.176$, and so on. This forms a probability distribution over the numbers 1 through 9.

More generally, in number base b (for example, in hexadecimal $b = 16$), X is distributed according to Benford's law if:

$$P(X = d) = \log_b\left(1 + \frac{1}{d}\right) \text{ where } 1 \le d < b$$
$$E[X] = (b-1) - \log_b[(b-1)!]$$

Zipf Distribution

X is a **Zipf random variable** ($X \sim \text{Zipf}(s, N)$) if the probability of X obeys an *inverse power law*:

$$P(X = k) = C \cdot \frac{1}{k^s} \text{ where } 1 \le k \le N$$

where C is a normalizing constant (which turns out to be equal to reciprocal of the Nth harmonic number).

In human languages, a Zipf distribution is a good model of the frequency rank index of a randomly chosen word, where N is the number of words in the language, and s also depends on various properties of the language (but is often close to 1). Other processes involving rank-ordering quantities also frequently result in a Zipf distribution, such as the rank of populations of large cities.