LECTURE: HYPERGEOMETRIC AND POISSON DISTRIBUTION

1. Hypergeometric Distribution

Here we will look briefly at the hypergeometric distribution, which models sampling without replacement.

Example 1:

You have a bag of 20 marbles, 12 of which are red and 8 of which are black.

(a) You draw a single marble from the bag five times, replacing it before each new draw. What is the probability that 3 of the 5 marbles drawn are red?

This is like a heads-and-tails problem, with probability of success $p = \frac{12}{20} = 0.6$

Let $X \sim \text{Binom}(5, 0.6)$ then

$$P(X = 3) = \binom{5}{3} 0.6^3 0.4^2 \approx 0.346$$

(b) You draw five marbles from the bag <u>without</u> replacement. What is the probability that 3 of the 5 marbles drawn are red?

This is similar to the poker hands. There are $\binom{20}{5}$ possible draws. There are $\binom{12}{3}$ ways to choose the red marbles and $\binom{8}{2}$ ways to choose the black ones

Letting Y = number of red marbles then

$$P(Y = 3) = \frac{\binom{12}{3}\binom{8}{2}}{\binom{20}{5}} \approx 0.397$$

Observation: The probabilities here are quite different. This is because we're taking a relatively large number of marbles compared to the sample size, $\frac{5}{20} = \frac{1}{4}$ What would happen if we took smaller fraction of the total marbles?

Example 2:

Same question, but this time we have 200 marbles, 120 of which are red and 80 of which are black, and you still draw 5 of them.

If X = number of red marbles with replacement then we still have $X \sim \text{Binom}(5, 0.6)$, so $P(X = 3) = \binom{5}{3}0.6^3 0.4^2 \approx 0.346$.

Let Y = number of red marbles without replacement. Then

$$P(Y = 3) = \frac{\binom{120}{3}\binom{80}{2}}{\binom{200}{5}} \approx 0.350$$

In this case, the two probabilities differ by only about 0.5%, which is much smaller.

The idea is that here you're only taking 5 (small) marbles out of 200 (huge), so it doesn't really matter if you replace them or not, the other 195 are still the same

Moral: If we sample a small fraction of the total population, sampling without replacement $\approx$ sampling with replacement, i.e. we can approximate sampling without replacement by a binomial distribution.

The distribution for sampling without replacement in this scenario is known as the **hypergeometric distribution.**

To motivate the notation: Suppose we have a bag of $n = 200$ marbles, $r = 120$ of which are red and the remaining $n - r = 80$ of which are black, and we're taking $m = 5$ marbles from the bag without replacement

Definition:

A discrete random variable Y has a **hypergeometric distribution** if

$$p(y) = \frac{\binom{r}{y}\binom{n-r}{m-y}}{\binom{n}{m}}$$

We write $Y \sim$ Hypergeom (n, r, m)

Here $p(y) = P(Y = y)$ gives you the probability that y of the m marbles are red

The hypergeometric distribution applies whenever we sample without replacement from a population consisting of two distinct groups, such as drawing marbles of two different colors, or polling a population who like either chocolate or vanilla ice cream.

When can we approximate a hypergeometric distribution (sampling without replacement) by a binomial distribution (sampling with replacement)? There is no hard-and-fast rule, but a good guideline is that if the sample size is less than 1/20 of the population size, the binomial distribution is a reasonable approximation.

2. Poisson Distribution

This the final discrete probability distribution we will discuss in this course.

It it used to model the number of events which occur during a fixed time interval under the following two assumptions:

(1) The average rate of occurrence of the events is constant.

(2) The events occur independently from each other.

Often the event in question is relatively rare. Examples of situations in which a Poisson distribution is a good model include:

(1) The number of phone calls received per hour at a call center.

(2) The number of pieces of non-junk mail received per day.

(3) The number of traffic accidents occurring at a particular intersection per week.

(4) The number of customers who enter a restaurant during a 15-minute period (although you might argue here that the average rate of arrival changes depending on the time of day.)

(5) The number of decays per second of a radioactive isotope.

3. Poisson Distribution Construction

Let Y be the number of calls received per hour at a call center, and suppose the average number of calls per hours is λ.

STEP 1: Split our hour up into n subintervals, where each subinterval is so small that at most one phone call can occur per subinterval. Let p be the probability that a phone call occurs in a given subinterval.

STEP 2: This is like a coin-tossing example, where heads means "there is a call in the sub-interval" and tails means "there is no call in the sub-interval."

We can then model Y as $Y \sim \text{Binom}(n, p)$

STEP 3: The average value of Y is $E(Y) = np$

Since the average number of calls per hour is λ, we will let $\lambda = np$ so $p = \frac{\lambda}{n}$

STEP 4: Since $Y \sim$ Binom (n, p) we have

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y}$$

STEP 5: Finally, we will let $n \to \infty$ (think small sub-piece)

$$
\begin{aligned}
\lim_{n\to\infty} p(y) &= \lim_{n\to\infty} \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} \\
&= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-(y-1))}{n^y}\frac{\lambda^y}{y!}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y} \\
&= \frac{\lambda^y}{y!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n \frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-(y-1))}{n}\left(1-\frac{\lambda}{n}\right)^{-y} \\
&= \frac{\lambda^y}{y!}\lim_{n\to\infty}\underbrace{\left(1+\frac{-\lambda}{n}\right)^n}_{\text{this has limit of } e^{-\lambda}} \underbrace{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{y-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-y}}_{\text{these all have limit of 1}} \\
&= e^{-\lambda}\frac{\lambda^y}{y!}
\end{aligned}
$$

The limiting probability $p(y)$ is the pmf for the **Poisson distribution**:

Definition:

A discrete random variable Y has a *Poisson distribution* with parameter $\lambda > 0$ if

$$p(y) = e^{-\lambda}\frac{\lambda^y}{y!} \qquad\qquad y = 0, 1, 2, \ldots$$

We write $Y \sim \text{Poi}(\lambda)$

Note that the Poisson distribution can output a value of 0, which corresponds to no events happening in the fixed span of time.