May 8 Lec—Binomial Distribution Cont., Hypergeometric, Geometric, Poisson

Midterm: Chapters 2-3

Recall the binomial distribution: for a random variable $X \sim B(N, p)$, we have for $k \in \{0, \ldots, N\}$,

$$p_X(k) = P(X = k) = C_k^N p^k (1-p)^{N-k}$$

Also recall last time's proposition:

Proposition:

1. $\mathbb{E}(X) = Np$
2. $V(X) = Np(1-p)$

Proof (2nd part)

$$V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$
$$= \mathbb{E}(X^2) - (Np)^2$$

The question is, what is $\mathbb{E}(X^2)$? Let's see some motivation and trial-and-error first:

$$\mathbb{E}(X^2) = \sum_{k=0}^{N} k^2 p_X(k)$$

$$= \sum_{k=0}^{N} k^2 C_k^N p^k (1-p)^{N-k}$$

$$= \sum_{k=0}^{N} k \cdot k C_k^N p^k (1-p)^{N-k}$$

We saw last time that $k \cdot C_k^N = N \cdot C_{k-1}^{N-1}$. So if we do some more of the same stuff, but suppose that we have one less k and an extra factor $k-1$.
So what if we can leverage $\mathbb{E}(X(X-1)) = \sum_{k=0}^{N} k(k-1) C_k^N$? We can do that:

$$V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X)^2)$$

Note that $\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X)$.

$$\mathbb{E}(X(X-1)) = \sum_{k=0}^{N} k(k-1)C_k^N p^k (1-p)^{N-k}$$

$$= \sum_{k=2}^{N} k(k-1)C_k^N p^k (1-p)^{N-k}$$

As we mentioned above,
$k(k-1)C_k^N = (k-1)N \cdot C_{k-1}^{N-1} = N \cdot (N-1)C_{k-2}^{N-2}$.
Continuing from above,

$$= N \cdot (N-1) \sum_{k=2}^{N} C_{k-2}^{N-2} p^k (1-p)^{N-k}$$

Making a change of variables $l = k-2$:

$$= N(N-1) \sum_{l=0}^{N-2} C_l^{N-2} p^{l+2} (1-p)^{N-l-2}$$

Exercise

Find $\mathbb{E}(X^3)$ and $\mathbb{E}(X^4)$ for $X \sim B(N,p)$.
Hint: You need to somehow use $\sum k(k-1)(k-2)C_k^N$ instead of $\sum k^3 C_k^N$.

Remark

If $X \sim B(N,p)$, then $Y := N - X$ is also binomial (because it's counting the number of failures) and $Y \sim B(N, 1-p)$.

Example

3.56 and 3.57
Let X be the number of successful explorations that a company makes

to a certain dangerous area. Due to constraints, $P(\text{exploration successful}) = 0.1 =: p$, and we are able to make only $N = 10$ independent trials. (i.e. $X \sim B(10, 0.1)$) We know that $\mathbb{E}(X) = Np = 1$ and $V(X) = Np(1-p) = 0.9$.
Each exploration has a fixed cost $2 \cdot 10^4$, with an additional cost of $3 \cdot 10^4$ if sucesful and $15 \cdot 10^3$ if failure. Find the expected (total) cost.
Let us write (total) cost Y as a discrete random variable in terms of X.

$$Y = 2 \cdot 10^5 + 3 \cdot 10^4 X + 15 \cdot 10^3 (10 - X)$$

$$Y = 15 \cdot 10^3 X + 35 \cdot 10^4$$

($\mathbb{E}(Y) = 365 \cdot 10^3$, $V(Y) = 2025 \cdot 10^5$—I am not a mathematician that counts)

Hypergeometric Distribution

Setting: There's a population of size N (say the total number of students at McGill). There is a subpopulation of size r with a characteristic that you are looking for (say age of at least 21 years). We randomly sample n students. Let X be a random variable representing the number of elements from the subpopulation included into the sample.

Remark:

If we sample *with replacement*, we're done since $X \sim B\left(N, p = \frac{r}{N}\right)$.
But what if we're sampling *without replacement*, i.e. the n elements in the sample are selected simultaneously???

First, what are the possible values of X? Can't be 0 to r always.
$X \leq r$ and $X \leq n$ for sure, so $X \leq \min\{n, r\}$. Since $n - X \geq 0$ and $n - X \leq N - r$, we have $X \geq \max\{0, n + r - N\}$.
So

$$\max\{0, n + r - N\} \leq X \leq \min(n, r)$$

Example

We can also use intuition to figure out the bounds. Suppose we're sampling $n = 5$ balls from an urn of 7 red and 3 blue balls. Let $X = \text{number of red sampled}, Y = \text{number of blue sampled}$. We intuitively think $X \in \{2, \ldots, 5\}$ and $Y \in \{0, \ldots, 3\}$, and we get that indeed with the formulas.

Let's figure out $p_X(x)$ for $\max\{0, n + r - N\} \le x \le \min\{n, r\}$. Using some combinatorics,

$$p_X(x) = \frac{C_x^r \cdot C_{n-x}^{N-r}}{C_n^N}$$

Definition: Hypergeometric distribution

If X has a distribution as above, then X is said to have a hypergeometric distribution with parameter N, r, n.

Proposition:

If X has a hypergeometric distribution with parameters N, r, n:

1. $\mathbb{E}(X) = n \cdot \frac{r}{N}$
2. $V(X) = n \cdot \frac{r}{N} \cdot \left(1 - \frac{r}{N}\right) \cdot \frac{N-n}{N-1}$

Remark

Note that $\frac{N-n}{N-1} \to 1$ as $N \to \infty$, in which case X can be approximated by a binomial random variable with $p = \frac{r}{N}$.

Example

An urn contains 7 red balls and 3 blue balls. A random sample of 5 balls is selected. The payoff is +$2 for each red ball included in the sample and -$3 for each blue ball included in the sample. Let Y be the total payoff. Find $\mathbb{E}(Y)$ and $V(Y)$.

$$Y = 2X - 3(5-X)$$
$$= 5X - 15$$

So we have

$$\mathbb{E}(Y) = 5\mathbb{E}(X) - 15$$
$$= 5 \cdot 5 \cdot \frac{7}{10} - 15 = 2.5$$
$$V(Y) = 5^2 V(X)$$
$$= 25 \cdot 5 \cdot \frac{7}{10} \cdot \frac{3}{10} \cdot \frac{5}{9} \approx 14.583$$

Geometric distribution

Quick refresher

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \quad |x| < 1$$

$$\sum_{k=1}^{\infty} x^k = \frac{1}{1-x} - 1 = \frac{x}{1-x}, \quad |x| < 1$$

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}, \quad x \neq 1$$

(if $x = 1$, of course we have

$$\sum_{k=0}^{n} 1 = n + 1$$

)

Now we describe the geometric distribution.
An experiment leads to success or failure. Different trials are independent.

Let X be the number of trials needed to reach the first success.
$X \in \{1, 2, 3, \ldots\} = \mathbb{N}^*$.
We now derive the probability function.

$$p_X(k) = P(X = k) = P(\underbrace{FFF\ldots F}_{k-1}S)$$

$$= (1-p)^{k-1} \cdot p$$

Definition: Geometric distribution

If X follows the above distribution, X is said to have a geometric distribution with parameter p. We write $X \sim \text{Geometric}(p)$.

Scribe's Note

Again, for brevity I will write $X \sim G(p)$.

Remark

Note that

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} p(1-p)^{k-1}$$

$$= p \cdot \sum_{k=1}^{\infty} (1-p)^{k-1}$$

$$= \frac{p}{1-p} \cdot \sum_{k=1}^{\infty} (1-p)^{k}$$

$$= \frac{p}{1-p} \cdot \frac{1-p}{1-(1-p)} = 1$$

Proposition

For $X \sim G(p)$, we have

1. $\mathbb{E}(X) = \frac{1}{p}$

2. $V(X) = \frac{1}{p}\left(\frac{1}{p} - 1\right)$

Proof

1.
$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=1}^{\infty} k p_X(k) \\ &= \sum_{k=1}^{\infty} k p (1-p)^{k-1} \\ &= p \cdot \left(\sum_{k=1}^{\infty} k (1-p)^{k-1}\right)\end{aligned}$$

From calculus, analysis, or an oracle:
for $|x| < 1$ we have

$$\sum_{k=1}^{\infty} k x^{k-1} = \sum_{k=1}^{\infty} \frac{d}{dx} x^k = \frac{d}{dx}\left(\sum_{k=1}^{\infty} x^k\right) = \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}$$

Similarly
(well you should be able to prove this on your own, can't LaTeX quickly enough in class)

Example

Refer to 3.70.
An oil inspector digs a succession of holes in a given area to find a productive well. $P(\text{successful trial}) = 0.2 =: p$. Let X be the number of wells that the inspector has to dig to find the first productive well. So $X \sim G(0.2)$. What is the probability that the third hole will be the first productive hole?
$P(X = 3) = 0.8^2 \cdot 0.2 = 0.128$
Now what if he can only afford to dig 10 holes?
Note that this is *not* binomial!!!

$$\begin{aligned}P(X \leq 10) &= 1 - P((X \leq 10)^c) \\ &= 1 - P(X > 10)\end{aligned}$$

$$= 1 - \sum_{k=11}^{\infty} 0.8^{k-1} \cdot 0.2$$

$$= 1 - 0.2 \cdot \sum_{l=10}^{\infty} 0.8^{l}$$

$$= 1 - 0.2 \cdot \sum_{i=0}^{\infty} 0.8^{i+10}$$

$$= 1 - 0.2 \cdot 0.8^{10} \sum_{i=0}^{\infty} 0.8^{i}$$

$$= 1 - 0.2 \cdot 0.8^{10} \cdot \frac{1}{0.2}$$

$$= 1 - 0.8^{10}$$

Poisson Distribution

Another quick refresher

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x \quad \forall x \in \mathbb{R}$$

There is no known real-life experiment that has exactly a Poisson distribution. You will always be told to assume that a certain experiment follows a Poisson distribution.

Definition: Poisson distribution

Given $\lambda > 0$ we have

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$$\implies 1 = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!}$$

Therefore the function

$$p(x) = \begin{cases} 0 & x \notin \mathbb{N} \\ e^{-\lambda} \cdot \frac{\lambda^x}{x!} & x \in \mathbb{N} \end{cases}$$

is a probability function. The random variable X that has this probability function is called the **Poisson random variable**. A discrete random variable X is said to have a **Poisson distribution** with parameter (or *mean*—we will justify this terminology later) $\lambda > 0$ if $X \in \mathbb{N}$ and $P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}\ \forall k \in \mathbb{N}$.
We denote this as $X \sim \text{Poisson}(\lambda)$.

Scribe's note

For brevity I will write $X \sim P(\lambda)$.