Math 3070, Applied Statistics

Section 1

September 11, 2019

Lecture Outline, 9/11

Section 3.4

- Hypergeometric Random Variable
- Negative Binomial Distribution

Hypergeometric Distribution, Motivation

Suppose a bag contains 5 red balls and 7 green balls. If we draw 4 balls at random, what is the probability that exactly 2 are red?

Solution:

- There are $\binom{12}{4} = 495$ ways to choose 4 balls from the 12 balls in the bag. Each of these 495 outcomes is equally likely.
- Drawing exactly 2 red balls means that the remaining 2 balls drawn are green.
- The number of ways to choose 2 of 5 red balls is $\binom{5}{2} = 10$.
- The number of ways to choose 2 of 7 green balls is $\binom{7}{2} = 21$.
- So the total number of ways to choose 2 red balls and 2 green balls is $10 \cdot 21 = 210$.
- The probability that this occurs is therefore

$$\frac{\binom{5}{2}\binom{7}{2}}{\binom{12}{4}} = \frac{210}{495}$$

.

Hypergeometric Distribution, Definition

In general if we select n individuals at random from a population of size N, where M individuals are of type A, and $N - M$ are of type B, then the number X of selected individuals of type A is a *hypergeometric* random variable:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Example: Suppose that out of a batch of 10 widgets, 7 are defective. If we randomly select 3 of the 10 widgets for inspection, what is the probability that we will find exactly 1 defective?

Here X is hypergeometric with $n = 3$, $N = 10$, $M = 7$, so

$$P(X = 1) = \frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}} = \frac{7 \cdot 3}{120} = 7/40$$

Hypergeometric Distribution, Example

Researchers catch and tag 5 animals of a species thought to be near extinction in a certain region. After the animals have mixed back into the population, 10 animals from the population are randomly selected. Let X be the number of tagged animals out of these 10. If there are actually 25 animals of this type in the region, what is the probability that (a) $X = 2$? (b) $X \leq 2$?

$$P(X = 2) = \frac{\binom{5}{2}\binom{20}{8}}{\binom{25}{10}} \approx .385$$

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{\binom{5}{0}\binom{20}{10}}{\binom{25}{10}} + \frac{\binom{5}{1}\binom{20}{9}}{\binom{25}{10}} + \frac{\binom{5}{2}\binom{20}{8}}{\binom{25}{10}} \approx .699 \end{aligned}$$

Relationship between Binomial and Hypergeometric

If the size of the batch is very large (say, 10000 widgets) and only a few widgets are drawn, then it makes little difference whether we sample with or without replacement, because it is very unlikely that any widget would be chosen more than once anyway. In this case, the hypergeometric and binomial distributions are practically identical.

In mathematical terms, the pmf of a hypergeometric random variable approaches the pmf of a binomial random variable, in the limit as we increase the population size N while keeping the same proportion $p = M/N$.

Binomial as Limit of Hypergeometric

Given a hypergeometric random variable X with $M/N = p$ and n held constant while $N \to \infty$,

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$= \frac{\frac{M(M-1)\cdots(M-x+1)}{x!} \cdot \frac{(N-M)\cdots(N-M-n+x+1)}{(n-x)!}}{\frac{N(N-1)\cdots(N-n+1)}{n!}}$$

$$= \frac{n!}{x!(n-x)!} \frac{M(M-1)\cdots(M-x+1)}{N(N-1)\cdots(N-x+1)} \cdot \frac{(N-M)\cdots(N-M-n+x+1)}{(N-x)\cdots(N-n+1)}$$

$$= \binom{n}{x} \frac{p(p-\frac{1}{N})\cdots(p-\frac{x-1}{N})}{1(1-\frac{1}{N})\cdots(1-\frac{x-1}{N})} \cdot \frac{(1-p)\cdots(1-p-\frac{n-x-1}{N})}{(1-\frac{x}{N})\cdots(1-\frac{n-1}{N})}$$

$$\to \binom{n}{x} p^x (1-p)^{n-x}$$

Binomial as Limit of Hypergeometric, Example

Out of 10,000 widgets, 1,200 of them are defective. Approximate the probability that 4 out of 7 randomly selected widgets are defective.

The population is very large so we can approximate X, the number of defective widgets out of 7, as $X \sim bin(7, 1200/10000)$ even when it accurate to model X as a hypergeometric.

$$P(X = 4) \approx \binom{7}{4}(0.12)^4(1 - 0.12)^3 \approx 0.00494585118$$

Note, "$\approx$" appears where "$=$" normally is after the probability since it's an approximation.

Mean and Variance Hypergeometric, Formula

If X is a Hypergeometric random variable, the number of individuals of type A out of n individuals from a population of size N with M individuals with type A. The PMF is

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

$$E(X) = n\frac{M}{N}$$

and

$$V(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right).$$

Note, this can be calculated from the definition of $E(X)$ and $V(X)$ using the PMF. We'll skip this for now.

Mean and Variance Hypergeometric, Binomial Approximation

Consider what happens when $M, N \to \infty$ while $\frac{M}{N} \to p$.

$$E[X] = n\frac{M}{N} \to np$$

$$V(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \to np(1-p)$$

Which matches the same quantites of Binomial Random Variable.

Hypergeometric Distribution, Summary

- $X \sim h(n, N, M)$. If X is the number of individuals of type A out of n individuals from a population of size N with M individuals with type A, then it's PMF is

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

- Not Bernoulli trails since the populations are fixed.
- $$E(X) = n\frac{M}{N}$$
- $$V(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$
- When N is large then X can be approximated as

$$X \sim bin(n, p) \quad p = \frac{M}{N}$$

Negative Binomial, Definition

A **negative binomial** random variable X counts the the number of failures before observing r successes of independent Bernoulli trials with parameter p. The possible values of X are $0, 1, 2, 3, \ldots$ and the PMF is

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x$$

$$X \sim NB(r, p)$$

Note, $X \sim NB(1, p)$ is also called a geometric random variable.

Negative Binomial, Derivation of PMF

Proceed similiarly to the derivation for the binomial.

1. The probability of a distinct outcome is $p^r(1-p)^x$, r successes and x failures.
2. The last trial is always a success. The negative binomial does not care about the order. The remaining $r-1$ can be rearranged among the $x+r-1$ positions to give the same value and probability.

$$P(X=x) = \binom{x+r-1}{r-1} p^r(1-p)^x$$

Negative Binomial, Mean and Variance

A negative binomial random variable X with r successes and trial parameter p has

$$E[X] = \frac{r(1-p)}{p}$$

and

$$Var(X) = \frac{r(1-p)}{p^2}$$

Again, proofs can be done using the definition of $E(X)$ and $Var(X)$ and the pmf, but are easier after discussing independent random variables.

Negative Binomial, Example

Boxes of cereal randomly contain toys. It is known that the expected number of boxes before receiving one with a toy is 2.3. What is the probability that a single box has a toy? What is the probability that your next two boxes both have a toy?

X is the number of box until one toy is found.

$$X \sim NB(1, p)$$

$$E(X) = \frac{1-p}{p} = 2.3 \rightarrow 1 - p = 2.3p \rightarrow p = 1/3.3$$

The probability that a single box has a toy is $1/3.3$.

Negative Binomial, Example

Boxes of cereal randomly contain toys and are independent. It is known that the expected number of boxes before receiving one with a toy is 2.3. What is the probability that a single box has a toy? What is the probability that someone opens three boxes before finding two toys?

Now, we model opening without toys boxes, Y, until 2 toys are found.

$$Y \sim NB(2, 1/3.3)$$

$$\begin{aligned} P(Y = 1) &= \binom{1+2-1}{2-1} \left(\frac{1}{3.3}\right)^2 \left(1 - \frac{1}{3.3}\right) \\ &= \binom{2}{1} \left(\frac{1}{3.3}\right)^2 \left(1 - \frac{1}{3.3}\right) \\ &\approx 0.05565294821 \end{aligned}$$

Negative Binomial, Non-Example

Boxes of cereal randomly contain toys and are independent. It is known that the expected number of boxes before receiving one with a toy is 2.3. What is the probability that someone buys 10 boxes that contain 4 toys?

Now, we are modeling Z boxes with toys out of 10. Each box is a Bernoulli trail with $p = 1/3.3$.

$$Z \sim Bin(10, 1/3.3)$$

$$P(Z = 4) = \binom{10}{4} \left(\frac{1}{3.3}\right)^4 \left(1 - \frac{1}{3.3}\right)^6$$
$$\approx 0.00137113398$$

Notice, the person opens all ten boxes and does not stop as soon as they have 4. The negative binomial stops assumes the 4th toy is found in the tenth box. Which strategy costs less?

Negative Binomial, Summary

- X is the number of failures until r successes are observed of independent Bernoulli trials with parameter p. $X \sim NB(r, p)$

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x$$

- $$E(X) = \frac{r(1-p)}{p} \quad Var(X) = \frac{r(1-p)}{p^2}$$
- Do not confuse with binomial distributions. The negative assumes the last trial is a success and it counts failures.

Midterm September 18, Information

- There is a midterm on September 18th in class. Calculator and notes allowed.
- Study the quizzes, summary slides, and homework. See the Canvas 'Files' tab for information.
- Review on September 16th. Come with questions.
- Reschedule by Wednesday, September 11, if needed. No makeup or late exams.
- One question from this week's material.
- No quiz or homework due on exam weeks. Material is shifted to the later week.