Hypergeometric Distribution

Assume we are drawing cards from a deck of well-shulffed cards with replacement, one card per each draw. We do this 5 times and record whether the outcome is ♠ or not. Then this is a binomial experiment.

If we do the same thing without replacement, then it is NO LONGER a binomial experiment.

However, if we are drawing from 100 decks of cards without replacement and record only the first 5 outcomes, then it is approximately a binomial experiment.

What is the exact model for drawing cards without replacement?

Hypergeometric Distribution

1. The population or set to be sampled consists of N individuals, objects, or elements (a *finite* population).
2. Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

Definition

For any experiment which satisfies the above 3 conditions, let $X =$ the number of S's in the sample. Then X is a **hypergeometric random variable** and we use $h(x; n, M, N)$ to denote the pmf $p(x) = P(X = x)$.

Hypergeometric Distribution

Examples:
In the second cards drawing example (without replacement and totally 52 cards), if we let $X =$ the number of ♠'s in the first 5 draws, then X is a hypergeometric random variable with $n = 5$, $M = 13$ and $N = 52$.
For the pmf, the probability for getting exactly x ($x = 0, 1, 2, 3, 4,$ or 5) ♠'s is calculated as following:

$$p(x) = P(X = x) = \frac{\binom{13}{x} \cdot \binom{39}{5-x}}{\binom{52}{5}}$$

where $\binom{13}{x}$ is the number of choices for getting x ♠'s, $\binom{39}{5-x}$ is the number of choices for getting the remaining $5 - x$ non-♠ cards and $\binom{52}{5}$ is the total number of choices for selecting 5 cards from 52 cards.

Hypergeometric Distribution

Examples:
For the same experiment (without replacement and totally 52 cards), if we let $X =$ the number of ♠'s in the first 20 draws, then X is still a hypergeometric random variable, but with $n = 20$, $M = 13$ and $N = 52$. However, in this case, all the possible values for X is $0, 1, 2, \ldots, 13$ and the pmf is

$$p(x) = P(X = x) = \frac{\binom{13}{x} \cdot \binom{39}{20-x}}{\binom{52}{20}}$$

where $0 \leq x \leq 13$.

Hypergeometric Distribution

Proposition

If X is the number of S's in a completely random sample of size n drawn from a population consisting of M S's and $(N-M)$ F's, then the probability distribution of X, called the **hypergeometric distribution**, *is given by*

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

for x an integer satisfying $\max(0, n-N+M) \le x \le \min(n, M)$.

Remark:
If $n < M$, then the largest x is n. However, if $n > M$, then the largest x is M. Therefore we require $x \le \min(n, M)$.
Similarly, if $n < N-M$, then the smallest x is 0. However, if $n > N-M$, then the smallest x is $n-(N-M)$. Thus $x \ge \min(0, n-N+M)$.

Hypergeometric Distribution

Example: (Problem 70)
An instructor who taught two sections of engineering statistics last term, the first with 20 students and the second with 30, decided to assign a term project. After all projects had been turned in, the instructor randomly ordered them before grading. Consider the first 15 graded projects.
a. What is the probability that exactly 10 of these are from the second section?
b. What is the probability that at least 10 of these are from the second section?
c. What is the probability that at least 10 of these are from the same section?

Hypergeometric Distribution

Proposition

The mean and variance of the hypergeometric rv X having pmf $h(x; n, M, N)$ are

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

Remark:
The ratio $\frac{M}{N}$ is the proportion of S's in the population. If we replace $\frac{M}{N}$ by p, then we get $E(X) = np$ and $V(X) = \left(\frac{N-n}{N-1}\right) \cdot np(1-p)$.
Recall the mean and variance for a binomial rv is np and $np(1-p)$.
We see that the mean for binomial and hypergeometric rv's are equal, while the variances differ by the factor $(N-n)/(N-1)$.

Hypergeometric Distribution

Example (Problem 70) continued:
An instructor who taught two sections of engineering statistics last term, the first with 20 students and the second with 30, decided to assign a term project. After all projects had been turned in, the instructor randomly ordered them before grading. Consider the first 15 graded projects.
d. What are the mean value and standard deviation of the number of projects among these 15 that are from the second section?
e. What are the mean value and standard deviation of the number of projects not among these 15 that are from the second section?

Negative Binomial Distribution

Consider the card drawing example again. This time, we still draw cards from a deck of well-shulffed cards with replacement, one card per each draw. However, we keep drawing until we get 5 ♠'s. Let X = the number of draws which do not give us a ♠, then X is NO LONGER a binomial random variable, but a **negative binomial random variable**.

Negative Binomial Distribution

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either s success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so $P(S \text{ on trial } i) = p$ for $i = 1, 2, 3, \ldots$.
4. The experiment continues (trials are performed) until a total of r successes have been observed, where r is a specified positive integer.

Definition

For any experiment which satisfies the above 4 conditions, let $X =$ the number of failures that precede thr r^{th} success. Then X is a **negative binomial random variable** and we use $nb(x; r, p)$ to denote the pmf $p(x) = P(X = x)$.

Negative Binomial Distribution

Remark:

1. In some sources, the negative binomial rv is taken to be the number of trials $X + r$ rather than the number of failures.
2. If $r = 1$, we call X a **geometric random variable**. The pmf for X is then the familiar one

$$nb(x; 1, p) = (1 - p)^x p \qquad x = 0, 1, 2, \ldots$$

Negative Binomial Distribution

Proposition

The pmf of the negative binomial rv X with parameters r = number of S's and $p = P(S)$ is

$$nb(x; r, p) = \binom{x + r - 1}{r - 1} \cdot p^r (1 - p)^x$$

Then mean and variance for X are

$$E(X) = \frac{r(1-p)}{p} \quad \text{and} \quad V(X) = \frac{r(1-p)}{p^2},$$

respectively

Negative Binomial Distribution

Example: (Problem 78)
Individual A has a red die and B has a green die (both fair). If they each roll until they obtain five "doubles" $(1-1, 2-2, \ldots, 6-6)$*, what is the pmf of* $X =$ *the total number of times a die is rolled? What are* $E(X)$ *and* $V(X)$*?*