## Hypergeometric Random Variables

Another type of random variable is the **hypergeometric** random variable. Following is an example.

**Example:** An office supply store receives a shipment of 20 chairs from a manufacturer. Each chair is in a box containing chair parts which the customer is expected to assemble. The store has had problems before with this manufacturer not including all the parts for each chair. The store manager asks an employee to open up 5 of the boxes to check whether all the parts are there. If two or more boxes are missing parts, the store will send the shipment back. If 6 of the 20 boxes in the shipment are missing parts, what is the probability the store will send the shipment back?

We have seen this type of problem earlier, in Example 3 on page 84, sec 2.4 exercises 9-12 and 18-22.

We let the random variable $X$ be the number of boxes the employee opens which have missing parts. So the problem is asking for $P(X \geq 2)$. The random variable $X$ above is an example of a hypergeometric random variable.

We can find $P(X = 3)$ for example by noting that since there are 6 bad boxes and 14 good boxes in the shipment, there are $C(6,3)$ ways of choosing 3 bad boxes and $C(14,2)$ ways of choosing 2 good boxes. So just as in Example 3 on page 84 we see

$$P(X = 3) = C(6,3) \cdot C(14,2)/C(20,5) \approx .1174$$

To find $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$, it is an easier computation to find $P(X < 2) = P(X = 0) + P(X = 1)$. So we compute

$$P(X = 0) = C(6,0) \cdot C(14,5)/C(20,5) \approx .1291$$

$$P(X = 1) = C(6,1) \cdot C(14,4)/C(20,5) \approx .3874$$

So $P(X \geq 2) = 1 - .1291 - .3874 = .4835$ and the probability the shipment will be returned is .4835.

**Hypergeometric Experiment:** Here is the general form of a hypergeometric random variable. Suppose there is a population of $r$ objects of which $g$ are "good" and $b$ are "bad". (So $r = g + b$.) Suppose you choose $n$ of the objects without replacement. Let the hypergeometric random variable $X$ be the number of objects you chose which were "bad". Then

$$P(X = k) \;=\; C(b,k) \cdot C(g, n-k)/C(r,n)$$

Of course it is not necessary that the objects actually be good or bad, just that they are divided into two types. For example they could be boys and girls, or red marbles and green marbles, or Democrats and Republicans, or whatever. But often hypergeometric random variables occur when you are sampling to find defects so often the two types are literally good and bad.

**Mean and variance of a hypergeometric random variable:** It is possible to calculate the mean and variance of a hypergeometric random variable by the following formulae:

$$\mu = E(X) = nb/r$$

$$\mathrm{Var}(X) = (r-n)nbg/(r^2(r-1))$$

**Bionomial approximation:** It is worth comparing the hypergeometric random variable to a similar binomial random variable. Suppose there is a population of $r$ objects of which $g$ are "good" and $b$ are "bad". (So $r = g + b$.) Suppose you choose $n$ of the objects with replacement. If $Y$ is the number of "bad" objects chosen, then $Y$ is a binomial random variable and $P(Y = k) = C(n,k)p^k q^{n-k}$ where $p = b/r$ and $q = g/r$. The only difference between $Y$ and $X$ is that the hypergeometric random variable $X$ involved choosing without replacement and the binomial $Y$ involved choosing with replacement. As a practical matter, if the population $r$ is much bigger than the sample size $n$ then the binomial random variable $Y$ will be a reasonably good approximation of $X$. This was exploited in some of the problems in section 3.4 of Tan.

**Example:** For example let us look again at Example 5 in section 3.4 of Tan. This involved sampling 20 solar cells from a large production run and finding the probability that 2 or fewer were defective. Since the production run was large, it was reasonable to solve the problem by assuming the number of defectives was a binomial random variable. The answers obtained were slightly inaccurate but the error was insignificant. But if the production run were smaller, say just 100 cells of which 5 were defective it would be better to use a hypergeometric random variable. Redoing this problem for a production run of only 100 cells of which 5 are defective we see that:

$$P(X = 0) = C(5,0) \cdot C(95,5)/C(100,5) \approx .3193$$

$$P(X = 1) = C(5,1) \cdot C(95,4)/C(100,5) \approx .4201$$

$$P(X = 2) = C(5,2) \cdot C(95,3)/C(100,5) \approx .2073$$

and so $P(X \leq 2) \approx .3193 + .4201 + .2073 = 0.9467$. This compares with the answer .9246 obtained in 3.4 of Tan using the binomial approximation.

**Winning at Blackjack:** This difference between the hypergeometric and binomial random variable has been exploited to make a (rather complicated) system for beating the house playing the game of blackjack. The game of blackjack, as with many card games, is a card game where players are dealt cards from a deck. When the deck is exhausted the deck is reshuffled and cards are dealt from the reshuffled deck. Since cards are not replaced after dealing each hand, the dealt cards form a sample without replacement. However, casinos figure their odds for blackjack by using the binomial approximation, erroneously believing that they have set the odds so that they will always make money in the long run. But if a gambler uses a complicated system which exploits the difference between the hypergeometric and binomial random variables, his expected winnings are positive and he can expect to win money in the long run. For example if half the deck is dealt out but only one quarter of the aces have been dealt, the player knows that aces are more likely to be dealt in ensuing hands and can adjust his strategy accordingly. Don't get any ideas of quitting school and making a living playing blackjack though. Your expected winnings are small and if the casino suspects you are using such a system you may be visited by a large person who requests you do your gambling elsewhere.

**This is only the beginning:** It is worth pointing out a bit of unreality in the problems we present here. The problems all assume the number $b$ of bad objects is known. In real life, of course, you would not actually know this number $b$ unless you sampled all $r$ objects in the population, which is usually not feasible to do. For example, the customers of the office supply store might prefer to purchase an unopened box because they might suspect an opened box was returned for some reason. Also the employees have more productive things to do than opening every single box of every single shipment and checking to see if it is okay.

In reality you most likely want to take the results of your small sample and use them to estimate the value of $b$. This is in effect what the store manager was doing. She figured that if there were 2 or more defectives in the sample of 5, then there were probably lots of defectives in the whole batch, enough to make it worthwhile to refuse the shipment. Using statistics it is possible to answer such questions as given that $X = 2$, what is the probability $b \geq 5$? This is beyond the scope of this course, but the ideas we present here are a prerequisite for finding the answer.

**Self-Check problems:**

1. A bag with 15 jellybeans contains 3 licorice flavored jelly beans. You pick 5 jellybeans from the bag at random and eat them.
   a) What is the expected number of licorice jellybeans you eat?
   b) What is the probability you eat less than 2 licorice jellybeans?

2. A shipment of 100 scooters contains 10 defective scooters. You test six scooters. let $X$ be the number of defective scooters you find. Find the mean, variance and standard deviation of $X$.

## Exercises

1. A box of a dozen eggs contains two rotten eggs. You make a cake using three eggs from the box. Let $X$ be the number of rotten eggs you use in your cake. Find the probability distribution for $X$ and draw a histogram for $X$. Find the expected value and standard deviation of $X$.

2. Redo problems 36, 37, and 38 of section 3.4 using hypergeometric random variables. In 36a, 37, and 38 assume 2 of the 20 are defective, in 36b assume 1 of the 20 is defective.

3. Eighteen uncounted punch card ballots from a recent election are discovered under an election official's couch. Six of them have a dimpled chad. A reporter gets to sneak a peek at three of the ballots. Let $X$ be the number of ballots with dimpled chads the reporter sees. Compute the probability distribution of $X$. What is the expected value of $X$? What is the standard deviation of $X$?

4. A six person committee is chosen at random from a class with 12 boys and 9 girls. What is the probability that there are 2 boys and 4 girls on the committee?

5. A bunch of grapes contains 60 grapes, of which 4 are rotten. You eat 6 of the grapes at random. Let $X$ be the number of rotten grapes you eat.

a) Find $P(X > 0)$ and $P(X \leq 4)$.

b) Find the expected value, variance, and standard deviation of $X$.

6. The border patrol of a certain country randomly searches cars coming into the country for smuggled goods.

a) In one hour, fifty cars pass a border checkpoint, ten of these cars are smuggling goods and the border patrol selects five cars for a thorough search. What is the probability that two or more smuggler's cars are searched?

b) The head border patrol officer at one checkpoint thinks for some reason that smugglers tend to drive flashy red convertibles. So he orders his officers to only select flashy red convertibles for a search. In one hour, fifty cars pass the checkpoint, ten of these cars are smugglers, twelve are flashy red convertibles, and six smugglers are not driving flashy red convertibles. The border patrol selects five flashy red convertibles for a thorough search. What is the probability that two or more smuggler's cars are searched?

7. You are dealt 5 cards from a standard 52 card deck with 4 aces. Let $X$ be the number of aces in your hand.

a) What is $P(X \leq 2)$?

b) What is the probability your hand contains two or fewer aces?

c) What is the expected number of aces in your hand?

d) Suppose you know that four other players have no aces in the hands they were dealt. What is the probability your hand contains two or fewer aces?

**Answers to Self-Check problems:**

1. Let the random variable $X$ be the number of licorice jellybeans you eat. Then $X$ is a hypergeometric random variable and

$$E(X) = 5 \cdot 3/15 = 1$$

so you expect to eat an average of one licorice jellybean.

$$\begin{aligned}
P(X \leq 2) &= P(X = 0) + P(X = 1) \\
&= C(3,0) \cdot C(12,5)/C(15,5) + C(3,1) \cdot C(12,4)/C(15,5) \\
&= \frac{3!}{0! \cdot 3!} \cdot \frac{12!}{5! \cdot 7!} \cdot \frac{5! \cdot 10!}{15!} + \frac{3!}{1! \cdot 2!} \cdot \frac{12!}{4! \cdot 8!} \cdot \frac{5! \cdot 10!}{15!} \\
&= \frac{3!}{3!} \cdot \frac{5!}{5!} \cdot \frac{10!}{7!} \cdot \frac{12!}{15!} + \frac{3!}{2!} \cdot \frac{5!}{4!} \cdot \frac{10!}{8!} \cdot \frac{12!}{15!} \\
&= \frac{10 \cdot 9 \cdot 8}{15 \cdot 14 \cdot 13} + \frac{3 \cdot 5 \cdot 10 \cdot 9}{15 \cdot 14 \cdot 13} \\
&= \frac{720}{2730} + \frac{1350}{2730} = \frac{2070}{2730} \approx .7582
\end{aligned}$$

2.

$$E(X) = 6 \cdot 10/100 = .6$$

$$\mathrm{Var}(X) = (100 - 6) \cdot 6 \cdot 10 \cdot 90/(100 \cdot 100 \cdot 99) \approx .5127$$

$$(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{.5127} \approx .7160$$