hypergeometric variable has smaller variance than does the binomial rv. The correction factor can be written $(1 - n/N)/(1 - 1/N)$, which is approximately 1 when n is small relative to N.

Example 3.36 *(Example 3.35 continued)*

In the animal-tagging example, $n = 10$, $M = 5$, and $N = 25$, so $p = \frac{5}{25} = .2$ and

$$E(X) = 10(.2) = 2$$

$$V(X) = \frac{15}{24}(10)(.2)(.8) = (.625)(1.6) = 1$$

If the sampling was carried out with replacement, $V(X) = 1.6$. Suppose the population size N is not actually known, so the value x is observed and we wish to estimate N. It is reasonable to equate the observed sample proportion of S's, x/n, with the population proportion, M/N, giving the estimate

$$\hat{N} = \frac{M \cdot n}{x}$$

If $M = 100$, $n = 40$, and $x = 16$, then $\hat{N} = 250$. ■

Approximating Hypergeometric Probabilities

Our general rule of thumb in Section 3.4 stated that if sampling was without replacement but n/N was at most .05, then the binomial distribution could be used to compute approximate probabilities involving the number of S's in the sample. A more precise statement is as follows: Let the population size, N, and number of population S's, M, get large with the ratio M/N remaining fixed at p. Then $h(x; n, M, N)$ approaches $b(x; n, p)$, so for n/N small, the two are approximately equal provided that p is not too near either 0 or 1. This is the rationale for our rule of thumb.

The Negative Binomial Distribution

The negative binomial rv and distribution are based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a 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.

The random variable of interest is X = the number of failures that precede the rth success; X is called a **negative binomial random variable** because, in contrast to the binomial rv, the number of successes is fixed and the number of trials is random.

Possible values of X are $0, 1, 2, \ldots$. Let $nb(x; r, p)$ denote the pmf of X. The event $\{X = x\}$ is equivalent to $\{r - 1$ S's in the first $(x + r - 1)$ trials and an S on the $(x + r)$th trial} (e.g., if $r = 5$ and $x = 10$, then there must be four S's in the first 14 trials and the trial 15 must be an S). Since trials are independent,

$$\begin{aligned} nb(x;\ r,\ p) &= P(X = x) \qquad (3.17) \\ &= P(r - 1\ S\text{'s on the first } x + r - 1 \text{ trials}) \cdot P(S) \end{aligned}$$

The first probability on the far right of Expression (3.17) is the binomial probability

$$\binom{x + r - 1}{r - 1} p^{r-1}(1 - p)^x, \qquad \text{while } P(S) = p$$

Proposition

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

$$nb(x;\ r,\ p) = \binom{x + r - 1}{r - 1} p^r(1 - p)^x, \qquad x = 0, 1, 2, \ldots$$

Example 3.37

A pediatrician wishes to recruit 5 couples, each of whom is expecting their first child, to participate in a new natural childbirth regimen. Let $p = P$(a randomly selected couple agrees to participate). If $p = .2$, what is the probability that 15 couples must be asked before 5 are found who agree to participate? That is, with S = {agrees to participate}, what is the probability that 10 F's occur before the fifth S? Substituting $r = 5$, $p = .2$, and $x = 10$ into $nb(x;\ r,\ p)$ gives

$$nb(10;\ 5,\ .2) = \binom{14}{4}(.2)^5(.8)^{10} = .034$$

The probability that at most 10 F's are observed (at most 15 couples are asked) is

$$P(X \le 10) = \sum_{x=0}^{10} nb(x;\ 5,\ .2) = (.2)^5 \sum_{x=0}^{10} \binom{x + 4}{4} (.8)^x = .164$$

In some sources, the negative binomial rv is taken to be the number of trials $X + r$ rather than the number of failures.

In the special case $r = 1$, the pmf is

$$nb(x;\ 1,\ p) = (1 - p)^x p, \qquad x = 0, 1, 2, \ldots \qquad (3.18)$$

In Example 3.10, we derived the pmf for the number of trials necessary to obtain the first S, and the pmf there is similar to Expression (3.18). Both X = number of F's and Y = number of trials ($= 1 + X$) are referred to in the literature as **geometric random variables,** and the pmf (3.18) is called the **geometric distribution.**

In Example 3.18, the expected number of trials until the first S was shown to be $1/p$, so that the expected number of F's until the first S is $(1/p) - 1 = (1 - p)/p$. Intuitively, we would expect to see $r \cdot (1 - p)/p$ F's before the rth S, and this is indeed $E(X)$. There is also a simple formula for $V(X)$.

Proposition

> If X is a negative binomial rv with pmf $nb(x; r, p)$, then
>
> $$E(X) = \frac{r(1-p)}{p}, \qquad V(X) = \frac{r(1-p)}{p^2}$$

Finally, by expanding the binomial coefficient in front of $p^r(1-p)^x$ and doing some cancellation, it can be seen that $nb(x; r, p)$ is well defined even when r is not an integer. This *generalized negative binomial distribution* has been found to fit observed data quite well in a wide variety of applications.

Exercises Section 3.5 (62–72)

62. A shipment of 15 concrete cylinders has been received by a contractor, 5 for a small project and the other 10 for a larger project. Suppose that 6 of the 15 have a crushing strength below the specified minimum. If the 5 for the smaller project are randomly selected from the 15 and X = the number among the 5 that have a below minimum crushing strength, then X has a hypergeometric distribution with parameters $n = 5$, $M = 6$, and $N = 15$. Compute the following:

a. $P(X = 2)$

b. $P(X \le 2)$

c. $P(X \ge 2)$

d. $E(X)$ and $V(X)$

63. Each of 12 refrigerators of a certain type has been returned to a distributor because of the presence of a high-pitched oscillating noise when the refrigerator is running. Suppose that 4 of these 12 have defective compressors and the other 8 have less serious problems. If they are examined in random order, let X = the number among the first 6 examined that have a defective compressor. Compute the following:

a. $P(X = 1)$

b. $P(X \ge 4)$

c. $P(1 \le X \le 3)$

64. A tennis coach has a basket containing 25 balls; 15 of these are Penn balls and the other 10 are Wilsons. Each of four players randomly selects 3 balls for a match.

a. What is the probability that exactly 8 of the balls selected are Penns?

b. What are the mean value and standard deviation of the number of Penn balls selected?

c. What is the probability that the number of Penn balls selected is more than 1 SD away from its mean value?

d. What are the mean value and standard deviation of the number of Penn balls left in the basket?

65. A geologist has collected 10 specimens of basaltic rock and 10 specimens of granite. If the geologist instructs a laboratory assistant to randomly select 15 of the specimens for analysis, what is the pmf of the number of basalt specimens selected for analysis? What is the probability that all specimens of one of the two types of rock are selected for analysis?

66. A personnel director interviewing 11 senior engineers for four job openings has scheduled six interviews for the first day and five for the second day of interviewing. Assume that the candidates are interviewed in random order.

a. What is the probability that x of the top four candidates are interviewed on the first day?

b. How many of the top four candidates can be expected to be interviewed on the first day?

67. Twenty pairs of individuals playing in a bridge tournament have been seeded 1, . . . , 20. In the first part of the tournament, the 20 are randomly divided into 10 east–west pairs and 10 north–south pairs.

a. What is the probability that x of the top 10 pairs end up playing east–west?