Lecture 13

The hypergeometric distribution

Suppose we have N balls; B of them are black and the remaining $N - B$ are white. We sample n balls at random without replacement (assuming that $n \leq B$). Let X denote the number of black balls drawn. What is the distribution of X? If the sampling were done with replacement then we know the answer is "Binomial$(n\,,p)$," where $p = B/N$. But sampling without replacement changes that answer a little. Indeed, it is not hard to check that

$$P\{X = k\} = \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}} \qquad \text{for } k = 0, \ldots, n.$$

This is called the *hypergeometric distribution* with parameters $(N\,, B\,, n)$.

Example 1 (Mean of a hypergeometric)**.** What is EX? One representation is, of course, the following:

$$EX = \sum_{k=0}^{n} k \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}}.$$

Although this can be simplified directly, the direct method is arduous. Instead we use the method of indicator variables: We can write $X = I_{A_1} + \cdots + I_{A_n}$, where A_j denotes the event that the jth draw is a black ball. The addition rule for expectation tells us that

$$EX = P(A_1) + \cdots + P(A_n) = \frac{nB}{N}.$$

Example 2 (SD of a hypergeometric)**.** What is SDX? Again we use the method of indicator variables; namely, we write $X = I_{A_1} + \cdots + I_{A_n}$, where A_j denotes the event that the jth draw is a black ball. But now note that the

A_j's are *not* independent. Therefore, we need to be more careful. First, note that for every two random variables J_1 and J_2 that have finite second moments,

$$E\left[(J_1 + J_2)^2\right] = E(J_1^2) + E(J_2^2) + 2E(J_1 J_2).$$

This and induction together yield the following: For all random variables $J_1, \ldots, J_n$ that have finite second moments,

$$E\left[(J_1 + \cdots + J_n)^2\right] = \sum_{i=1}^n E(J_i^2) + 2 \sum_{1 \le i < j \le n} E(J_i J_j).$$

We apply this with $J_i := I_{A_i}$ to find that

$$\begin{aligned} E(X^2) &= \sum_{i=1}^n E(I_{A_i}^2) + 2 \sum_{1 \le i < j \le n} E(I_{A_i} I_{A_j}) \\ &= \sum_{i=1}^n E(I_{A_i}) + 2 \sum_{1 \le i < j \le n} E(I_{A_i} I_{A_j}) \\ &= \sum_{i=1}^n P(A_i) + 2 \sum_{1 \le i < j \le n} P(A_i \cap A_j). \end{aligned}$$

On one hand, $P(A_i) = B/N$; therefore $\sum_i P(A_i) = nB/N$. On the other hand, if $i < j$ then "by symmetry,"

$$P(A_i \cap A_j) = P(A_1 \cap A_2) = P(A_2 \mid A_1) P(A_1) = \frac{B-1}{N-1} \frac{B}{N} = \frac{B(B-1)}{N(N-1)}.$$

Therefore,

$$\begin{aligned} E(X^2) &= \frac{nB}{N} + 2 \sum_{1 \le i < j \le n} \frac{B(B-1)}{N(N-1)} \\ &= \frac{nB}{N} + 2 \binom{n}{2} \frac{B(B-1)}{N(N-1)} \\ &= \frac{nB}{N} + \frac{n(n-1)B(B-1)}{N(N-1)}. \end{aligned}$$

It follows that

$$E(X) = \frac{nB}{N} \quad \text{and} \quad \mathrm{Var}(X) = \frac{nB}{N} + \frac{n(n-1)B(B-1)}{N(N-1)} - \frac{n^2 B^2}{N^2}.$$

We simplify the variance further as follows: Let $p := B/N$ denote the proportion of black balls. Then,

$$
\begin{aligned}
\mathrm{Var}(X) &= np + np\frac{(n-1)(B-1)}{N-1} - n^2p^2 \\
&= np\left[1 + \frac{(n-1)(B-1)}{N-1} - np\right] \\
&= \frac{np}{N-1}\left[N - 1 + (n-1)(B-1) - n(N-1)p\right] \\
&= \frac{np}{N-1}\left[N - 1 + (n-1)(Np-1) - n(N-1)p\right] \\
&= \frac{np}{N-1}\left[N - n - Np + np\right] = \frac{np}{N-1}\left[(N-n) - p(N-n)\right] \\
&= npq\frac{N-n}{N-1},
\end{aligned}
$$

with $q := 1 - p =$ proportion of white balls. Therefore,

$$
\mathrm{SD}(X) = \sqrt{npq} \cdot \sqrt{\frac{N-n}{N-1}}.
$$

If the sample size $n \ll N$, then $(N-n)/(N-1) \approx 1$. Therefore $\mathrm{Var}(X) \approx \sqrt{npq}$; i.e., there isn't much difference between with and without replacement sampling when the sample size n is much smaller than the population size!

It turns out that there is also a central limit theorem [for X standardized; that is, for all $-\infty \le a \le b \le \infty$ and B and n fixed,

$$
P\left\{a \le \frac{X - np}{\sqrt{npq} \cdot \sqrt{\frac{N-n}{N-1}}} \le b\right\} \approx \Phi(b) - \Phi(a) \quad \text{as } N \to \infty.
$$