Math 342 —Hypergeometric distribution

Problem 1. Suppose in a population of 1000 deer at a national park, 200 are captured by the wildlife service. These deer are tagged, say with some identifying mark, and are released back into the park with the rest of the population. After a period of time, a new sample of 50 deer are re-captured. Let X count the number of tagged deer in this group of 50.

a. Does X have the hypergeometric distribution? If so, what are the parameters?

b. Find the probability that there are exactly 10 tagged deer in our re-capture sample.

c. Find the probability that there are at least 10 tagged deer in our re-capture sample.

d. Find the probability that there are between 5 and 15 tagged deer in our re-capture sample.

e. Find the expected number of tagged deer in our re-capture sample.

Problem 2. Suppose that a batch of 100 items contains 6 that are defective and 94 that are not defective. If X is the number of defective items in a randomly drawn sample of 10 items from the batch, find

a. $P(X = 0)$

b. $P(X > 2)$

c. $E[X]$

Problem 3. In a population of 10,000 people, 8000 are in favor of a proposal for a new law. Suppose that a poll of 25 people is taken. Find the probability that between 15 and 20 people in the poll are in favor of the proposal, assuming (a) sampling is done with replacement and (b) sampling is done without replacement. What do you notice? *This problem is demonstrating the following idea: if $r \to \infty$ and $N \to \infty$ in such a way that r/N converges to a value $p \in (0, 1)$, a random variable which is hypergeometrically distributed with parameters r, N, n can be approximated by a random variable that is binomially distributed with parameters n and p. In other words, in the large population size limit, sampling with or without replacement are essentially the same.*

Problem 4 (Challenge Problem)**.** We have previously learned the following formula for the variance of a sum of two random variables: $V(X + Y) = V(X) + V(Y) + 2\,\mathrm{Cov}(X, Y)$ where $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$. Note that this can also be written as

$$V(X + Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(Y, Y) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X).$$

This can be generalized to the sum of n random variables:

$$\begin{aligned} V(X_1 + \cdots + X_n) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) \\ &= V(X_1) + \cdots + V(X_n) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j). \end{aligned}$$

Use this to derive the formula for the variance of a hypergeometrically distributed random variable $X \sim \mathrm{HyperGeo}(r, N, n)$:

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

Aside from the last factor, notice the similarity with the formula for the variance of the binomial distribution $np(1 - p)$.