Note on the hypergeometric distribution

November 3, 2016

Consider an urn with \( N \) balls, of which \( m \) are white, the rest are black. Suppose that \( n \) random balls are removed without replacement, and let \( X \) be the number of white balls drawn. (In the below, we will assume that \( n \leq m, N - m \). Handling the other case, follows similar arguments.

\( X \) is a hypergeometric random variable with parameters \((N, m, n)\). The probability mass function of \( X \) is

\[
Pr(X = i) = \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}.
\]

Observe that

\[
\sum_{i=0}^{n} Pr(X = i) = \sum_{i=0}^{n} \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}} = 1. \tag{0.1}
\]

Expectation of a hypergeometric r.v.

We write \( X \) as a sum of \( n \) random variables \( X_1 + X_2 + \ldots + X_n \) where \( X_k \) is an indicator r.v. which is 1 if the \( k \)-th ball drawn is white and 0 otherwise.

We claim that for each \( k \in [1, n] \),

\[
E(X_k) = \frac{m}{N}, \tag{0.2}
\]

and therefore, by linearity of expectation

\[
E(X) = n \cdot \frac{m}{N}.
\]

We can prove (0.2) several ways.

- Informal proof: If we pick one ball at random out of the urn, the probability it is white is \( m/N \). We claim that this is also true if we consider the fifth ball removed (or any ball). Why? Because consider a sequence of \( n \) balls removed from the urn one at a time. Each permutation of these balls is equally likely. Therefore if the first ball is white with probability \( m/N \), then the \( k \)-th ball is also white with the same probability. Therefore, we have (0.2).
• Formal proof:

\[ E(X_k) = Pr(k\text{-th ball is white}) \]

\[ = \sum_{i=0}^{k-1} Pr(k\text{-th ball is white} | i \text{ of the first } k - 1 \text{ balls white}) Pr(i \text{ of first } k - 1) \]

\[ = \sum_{i=0}^{k-1} \frac{(m-i)}{N-k+1} \cdot \frac{\binom{m}{i} \binom{N-m}{k-1-i}}{\binom{N}{k-1}} \]

and since \( (m-i) \binom{m}{i} = \frac{m!}{i!(m-i)!} = m \binom{m-1}{i} \), and similarly \( (N-k+1) \binom{N}{k-1} = N \binom{N-1}{k-1} \), we have

\[ = \frac{m}{N} \sum_{i=0}^{k-1} \frac{(m-1) \binom{N-m}{k-1-i}}{\binom{N-1}{k-1}} \]

but this sum is 1 by (0.1) applied to a hypergeometric with parameters \( (N-1, m-1, k-1) \), so we get that

\[ E(X_k) = \frac{m}{N} \]