5.3 The Hypergeometric Distributions

The hypergeometric random variable is an analogue of the binomial random variable but differs with respect to the sampling experiment underlying the random variable. A binomial random variable $X \sim \text{Binom}(n, p)$ can be viewed as a sum of $n$ iid Bernoulli random variables where $p = \Pr(X_i = 1)$. From a sampling context, the generating experiment consists of $n$ trials that draw randomly and with replacement from a population consisting of $N$ units. Each population unit can be classified according to a binary variable (having two levels). $A$ units possess the attribute value of interest, $N - A = B$ do not, and $p = A/N$. The random variable $X$ counts the number of units in the sample with the attribute value of interest.

In the case of the hypergeometric random variable, the population again consists of $N$ units wherein $A$ possess the attribute value of interest and $B = N - A$ do not, but now sampling is random but without replacement. Since the probability of sampling a unit with the attribute value changes on each draw, the number of sampled units with the attribute value cannot be cast as a sum of independent Bernoulli random variables.

Suppose that $X$ counts the number of units among a sample of $n \leq N$ that possess the attribute value. The probability of the event $\{X = x\}$ will be non-zero provided that $x \leq A$ and $x \leq n$, i.e.,

$$0 \leq x \leq \min\{n, A\}.$$  

Similarly, the number of units in the sample without the attribute value must be no greater than $B \Rightarrow n - x \leq B \Rightarrow n - B \leq x$. Also, $0 \leq x$, and $\max\{0, n - B\} \leq x$.

The probability of interest, $\Pr(X = x)$, can be determined by elementary counting methods: there are $\binom{A}{x}$ ways of drawing $x$ items from among $A$ items and $\binom{B}{n-x}$ ways of drawing the remaining $n-x$ sample items from among $B$ items. The number of possible samples is $\binom{N}{n}$; hence,

$$\Pr(X = x|A, B, n) = \begin{cases} \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{N}{n}}, & \max\{0, n - B\} \leq x \leq \min\{n, A\}, \\ 0, & \text{otherwise}. \end{cases} \quad (1)$$  

Three parameters are used $(A, B, n)$. The probability density function is $f(x|A, B, n) = \Pr(X = x|A, B, n)$ as defined by equation (1). R functions to compute hypergeometric probabilities are `dhyper`, `phyper`, etc.
The mean and variance of a hypergeometric random variable

Suppose that \( A \) items are red and \( B \) items are blue. Let

\[
X_i = \begin{cases} 
1, & \text{if the attribute of the } i\text{th item is red,} \\
0 & \text{otherwise.}
\end{cases}
\]

The number of red items in the sample is

\[
X = \sum_{i=1}^{n} X_i,
\]

and the distribution of \( X \) is hypergeometric with parameters \( n \), \( A \) and \( B \).

With no other information besides \( A \) items are red and \( B \) items are blue, and sampling is random, the probability that the \( i \)th draw produces a red is \( \Pr(X_i = 1) = A/(A + B) \).

The expected number of red items is

\[
E(X) = E\left( \sum_{i=1}^{n} X_i \right) = \frac{nA}{A + B}.
\]

Each \( X_i \) is Bernoulli in distribution (i.e., \( X_i \sim \text{Binom}(1, A/(A + B)) \)), though \( X \) is not binomial because \( X_1, \ldots, X_n \) are not iid. Regardless, the \( X_i \)'s can be used to determine variance of \( X \). First,

\[
E(X_i) = \frac{A}{A + B}
\]

\[
E(X_i^2) = E(X_i) = \frac{A}{A + B}
\]

\[
\Rightarrow \text{Var}(X_i) = \frac{AB}{(A + B)^2}.
\]

The covariance of \( X_i \) and \( X_j, i \neq j \) is

\[
\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)
\]

\[
= \Pr(X_i = 1, X_j = 1) - \Pr(X_i = 1)\Pr(X_j = 1)
\]

\[
= \Pr(X_i = 1|X_j = 1)\Pr(X_j = 1) - \Pr(X_i = 1)^2
\]

\[
= \frac{A(A - 1)}{(A + B)(A + B - 1)} - \left( \frac{A}{A + B} \right)^2
\]

\[
= \frac{AB}{(A + B)^2(A + B - 1)}
\]
Finally, returning to the hypergeometric random variable \( X = \sum_i X_i \),

\[
\text{Var}(X) = \sum_i \text{Var}(X_i) + 2 \sum_i \sum_{j > i} \text{Cov}(X_i, X_j)
\]

\[
= \frac{nAB}{(A + B)^2} - \frac{n(n-1)AB}{(A + B)^2(A + B - 1)}
\]

\[
= \frac{nAB}{(A + B)^2} \frac{A + B - n}{A + B - 1}.
\]

The difference between the hypergeometric sampling experiment and the binomial sampling experiment is that the hypergeometric experiment samples without replacement (versus with replacement for the binomial). A comparison of the two random variables is based on defining the binomial probability as the fraction of items possessing the attribute of interest (red in color). Specifically, let \( T = A + B \), then

\[
p = \frac{A}{A + B} = \frac{A}{T}.
\]

For \( Y \sim \text{Binom}(n, p) \),

\[
E(Y) = np = \frac{A}{T}
\]

\[
\text{Var}(Y) = np(1 - p) = \frac{nAB}{T^2}
\]

In comparison, for \( X \) with a hypergeometric distribution obtained from sampling \( n \) items without replacement from \( T = A + B \) items,

\[
E(X) = np = \frac{A}{T}
\]

\[
\text{Var}(Y) = \frac{nABT - n}{T^2} \frac{T - n}{T - 1}
\]

\[
= np(1 - p) \frac{T - n}{T - 1}.
\]

The factor \( \alpha = (T - n)/(T - 1) \), known as the finite population correction factor in sampling\(^1\), represents the difference between sampling with and without replacement. Notice that if \( T >> n \), then there is little or no difference between the binomial and hypergeometric random variable. Consequently, if sampling is without replacement and the population size \( T \) is much larger than \( n \), statisticians will analyze \( Y \) by treating \( Y \) was if it were binomial in distribution (less difficult). These usually minor differences are also ignored in the teaching of elementary statistics.

Theorem 5.3.4. provides a formal justification of the approximation of a hypergeometric

\(^1\)A branch of statistics focusing on sampling designs or protocols and the analysis of data obtained from sampling designs.
random variable by an appropriate binomial random variable.

The usual statistical context in which the hypergeometric random variable arises is in the estimation of the proportion of units in a population possessing the attribute value of interest. In this situation, $A$ and $B$ are not known and the objective is to estimate $p = A/T$. The best estimator\(^2\) is the sample proportion $Y/n$, where $Y$ is the number of observations in a sample of $n$ that possess the attribute value of interest.

**Question 1**: Compute the expected value and variance of the estimator $Y/n$. 

\[^2\text{(dependent on a measure of quality)}\]