UNIVERSITY OF LONDON

Goldsmiths UNIVERSITY OF LONDON

Oct. 22nd 2021

the Hyper-Geometric Distribution

Study Notes | Written by Larry Cui

The hypergeometric distribution is about the unordered sampling without replacement. Here we find its distribution and relevant properties.

1 What does the distribution say?

Hyper-geometric Theorem Suppose an urn contains r red chips and w white chips, where $r + w = N$. If n chips are drawn out at random, without replacement, and if k denotes the number of red chips selected, then

$$P(k \text{ red chips are chosen}) = \frac{\dbinom{r}{k}\dbinom{w}{n-k}}{\dbinom{N}{n}} \quad \text{for} \quad k = 0, 1, 2, ..., n$$

Proof: First of all, it's intuitive to see that if you pick n chips (with k red) from the urn, the combination of the n chips is $\binom{r}{k}\binom{w}{n-k}$. The total amount of the combinations, is a traverse of this format from 0 red chip to n red chips (suppose $r > n$), and must equal to $\binom{N}{n}$.

2 $E(X)$ and Var (X)

2.1 summation method

We use the summation to obtain the expected value,

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \cdot \frac{\dbinom{r}{k}\dbinom{w}{n-k}}{\dbinom{N}{n}} \\
&= \sum_{k=0}^{n} k \cdot \frac{r!}{k!(r-k)!} \cdot \frac{w!}{(n-k)!(w-n+k)!} \Big/ \frac{N!}{n!(N-n)!} \\
&= \sum_{k=0}^{n} n\frac{r}{N} \cdot \frac{(r-1)!}{(k-1)!(r-k)!} \cdot \frac{w!}{(n-k)!(w-n+k)!} \Big/ \frac{(N-1)!}{(n-1)!(N-n)!} \\
&= n\frac{r}{N} \sum_{k=1}^{n} \cdot \frac{\dbinom{r-1}{k-1}\dbinom{w}{n-k}}{\dbinom{N-1}{n-1}} = n\frac{r}{N} \qquad \text{(term } k = 0 \text{ is 0, so the summation starts from } k = 1\text{)}
\end{aligned}$$

Based on this result, we go on to solve for $E(X^2)$:

$$
\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2 \cdot \frac{\binom{r}{k}\binom{w}{n-k}}{\binom{N}{n}} = n\frac{r}{N}\sum_{k=1}^{n} k \cdot \frac{\binom{r-1}{k-1}\binom{w}{n-k}}{\binom{N-1}{n-1}} \\
&= n\frac{r}{N}\sum_{k=1}^{n}((k-1)+1) \cdot \frac{\binom{r-1}{k-1}\binom{w}{n-k}}{\binom{N-1}{n-1}} && \text{the first term in the parenthsis is a format for } E(X-1) \\
&= n\frac{r}{N} \cdot \left[(n-1)\frac{r-1}{N-1} + 1\right] && \text{sum of distribution equals to 1}
\end{aligned}
$$

Now we can use formula $\operatorname{Var}(x) = E(X^2) - E(X)^2$ to get the result, if we have $p = r/N$,

$$
\begin{aligned}
np \cdot \left[(n-1)\frac{r-1}{N-1} + 1\right] - (np)^2 &= \frac{np}{N-1}\left[(n-1)(r-1) + (N-1) - np(N-1)\right] \\
&= \frac{np}{N-1}(nr - r - n + 1 + N - 1 - nr + np) \\
&= \frac{np}{N-1}(N - n + np - Np) \qquad (r = Np) \\
&= np(1-p)\frac{N-n}{N-1}
\end{aligned}
$$

2.2 alternative approach

Covariance Definition Given random variables X and Y with variances, define the *covariance* of X and Y, written Cov(X,Y), as

$$\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

A lemma comes directly from the above definition: when X and Y are independent variables, Cov(X,Y) = 0, since $E(XY) = E(X)E(Y)$.

The following theorem is to find the variance of the sum of two variables.

Theorem 2.2 Suppose X and Y are random variables with finite variances, and a and b are constants. Then

$$\operatorname{Var}(aX + bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\operatorname{Cov}(X, Y)$$

Proof: For convenience, let's denote $E(X)$ by μ_X and $E(Y)$ by μ_Y, then

$$
\begin{aligned}
\operatorname{Var}(aX + bY) &= E[(aX + bY)^2] - (a\mu_X + b\mu_Y)^2 \qquad \text{note: } E(aX) = aE(X) \\
&= E(a^2X^2 + b^2Y^2 + 2abXY) - a^2\mu_X^2 - 2ab\mu_X\mu_Y - b^2\mu_Y^2 \\
&= a^2[E(X^2) - \mu_X^2] + b^2[E(Y^2) - \mu_Y^2] + 2ab[E(XY) - \mu_X\mu_y] \\
&= a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\operatorname{Cov}(X, Y)
\end{aligned}
$$

Lemma If X and Y are independent variables, then

$$\text{Var}\,(aX + bY) = a^2\text{Var}\,(X) + b^2\text{Var}\,(Y)$$

We now revisit the hyper geometric distribution. A random sample of size n is picked without replacement, and the random variable X is defined to the red chip amount of the sample: $X = X_1 + X_2 + \cdots X_n$. If we look each single X_i, we know it's expected value is, no matter of its order:

$$E(X_i) = 1 \cdot \frac{r}{N} + 0 \cdot \frac{N-r}{N} = \frac{r}{N}$$

and $E(X_i^2)$ is same as $E(X_i)$,

$$E(X_i^2) = E(X_i) = \frac{r}{N}$$

again we have $\text{Var}\,(X_i)$,

$$\text{Var}\,(X_i) = E(X_i^2) - E(X_i)^2 = \frac{r}{N} - \left(\frac{r}{N}\right)^2$$

However, for any $j \neq k$, X_j and X_k are not independent. We can calculate their covariance as

$$\begin{aligned}
\text{Cov}\,(X_j, X_k) &= E(X_jX_k) - E(X_j)E(X_k) \\
&= 1 \cdot P(X_j = X_k = 1) - \left(\frac{r}{N}\right)^2 \qquad \text{(in other three scenarios } E(X_jX_k) = 0\text{)} \\
&= \frac{r}{N} \cdot \frac{r-1}{N-1} - \frac{r^2}{N^2} \\
&= -\frac{r}{N} \cdot \frac{N-r}{N} \cdot \frac{1}{N-1}
\end{aligned}$$

Now according to Theorem 2.2, we have the variance of X below (let p denote $\frac{r}{N}$),

$$\begin{aligned}
\text{Var}\,(X) &= \sum_{i=1}^{n} \text{Var}\,(X_i) + 2\sum_{j<k}^{n} \text{Cov}\,(X_j, X_k) \\
&= np(1-p) - 2\binom{n}{2}p(1-p) \cdot \frac{1}{N-1} \\
&= p(1-p)\left[n - \frac{n(n-1)}{N-1}\right] \\
&= np(1-p) \cdot \frac{N-n}{N-1}
\end{aligned}$$