

Hypergeometric

The basic outcome is the number of successes in n trials of binary outcome. If sample is selected without replacement this is not binomial because the trials are not independent. When we actually sample we do so without replacement, but since that makes the math more difficult we use nicer math models (e.g, binomial) as convenient approximations. If we *really* wanted the correct answer we'd use hypergeometric.

Let $X_i = \#$ successes on i th trial. (Since these are binary outcomes, the # of successes is either 0 or 1.)

$n =$ sample size

$N =$ population size

$r = \#$ of reds (successes) in population

Define $Y = X_1 + X_2 + \dots + X_n$, that is the number of successes in the sample of size n .

Picking the first choice randomly we have $p(X_1 = 1) = \frac{r}{N}$.

The probability that X_2 , the second randomly chosen choice, is "red" depends on the result of $X_1 \dots$

$$p(X_2 = 1 | X_1 = 1) = \frac{r-1}{N-1}$$

$$p(X_2 = 1 | X_1 = 0) = \frac{r}{N-1}$$

The Law of Total Probability:

$$p(A) = p(A \cap B) + p(A \cap B^C)$$

and by the definition of conditional probability,

$$p(A) = p(A|B) \cdot p(B) + p(A|B^C) \cdot p(B^C)$$

For our choices of "red",

$$p(X_2 = 1) = p(X_2 = 1 | X_1 = 1) \cdot p(X_1 = 1) + p(X_2 = 1 | X_1 = 0) \cdot p(X_1 = 0)$$

$$= \left(\frac{r-1}{N-1}\right) \cdot \frac{r}{N} + \left(\frac{r}{N-1}\right) \left(1 - \frac{r}{N}\right)$$

$$= \left(\frac{r-1}{N-1}\right) \cdot \frac{r}{N} + \left(\frac{r}{N-1}\right) \cdot \left(\frac{N-r}{N}\right)$$

$$= \frac{r^2 - r + rN - r^2}{(N-1)N}$$

$$= \frac{r(N-1)}{(N-1)N}$$

$$= \frac{r}{N}$$

Thus, the sampling without replacement does not violate the constancy of the probability of success; independence, however, is violated.

Let's now find $E(Y)$ and $Var(Y)$.

By virtue of the derivation above, each X_i is identically distributed, and

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

Recall that $Y = X_1 + X_2 + \dots + X_n$.

The expected value is fairly easy to find...

$E(Y) = E(X_1) + E(X_2) + \dots + E(X_n)$ whether or not X_i and X_j are independent.

$$E(Y) = p + p + \dots + p$$

$$= np$$

$$= n(r/N)$$

Finding the $Var(Y)$ is a much more difficult problem....

$Var(Y) = Var(X_1 + X_2 + \dots + X_n)$ when X_i, X_j are not independent is given by:

$Var(Y) = Var(X_1) + Var(X_2) + \dots + Var(X_n) + 2cov(X_1, X_2) + 2cov(X_i, X_j) + \dots + 2cov(X_{n-1}, X_n)$ where $i < j$.

$$\begin{aligned} cov(X_i, X_j) &= E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= E[X_i X_j - \mu_i X_j - \mu_j X_i + \mu_i \mu_j] \\ &= E(X_i X_j) - E(\mu_i X_j) - E(\mu_j X_i) + E(\mu_i \mu_j) \\ &= E(X_i X_j) - \mu_i E(X_j) - \mu_j E(X_i) + \mu_i \mu_j \\ &= E(X_i X_j) - \mu_i \mu_j - \mu_j \mu_i + \mu_i \mu_j \\ &= E(X_i X_j) - \mu_i \mu_j \end{aligned}$$

What, pray, tell, is $E(X_i X_j)$? The product $X_i X_j$ must be either 0 or 1, since X_i and X_j are Bernoulli outcomes.

Then $E(X_i X_j) = 0 \cdot p(0) + 1 \cdot p(1)$

$$= 1 \cdot p(X_i = 1 \text{ and } X_j = 1)$$

(Jon) $\leftarrow = p(X_i = 1 | X_j = 1) \cdot p(X_j = 1)$

$$= \frac{r-1}{N-1} \cdot \frac{r}{N}$$

$$\therefore E(X_i X_j) = \frac{r-1}{N-1} \cdot \frac{r}{N}$$

and

$$cov(X_i X_j) = \frac{r-1}{N-1} \cdot \frac{r}{N} - \frac{r}{N} \cdot \frac{r}{N}$$

There are n $Var(X_i)$ to add, and $\frac{n(n-1)}{2} cov(X_i, X_j)$.

Thus, $Var(Y) = nVar(X_i) + \frac{n(n-1)}{2} [2cov(X_i, X_j)]$

$$= n[p(1-p)] + \frac{n(n-1)}{2} [2(\frac{r-1}{N-1} \cdot \frac{r}{N} - \frac{r}{N} \cdot \frac{r}{N})]$$

$$\dots + \frac{n(n-1)}{2} [2(\frac{r-1}{N-1} - \frac{r}{N}) \frac{r}{N}]$$

$$= np(1-p) + n(n-1)[\frac{r-1}{N-1} - p]p$$

$$= np(1-p) + n(n-1)[\frac{r-1}{N-1} - \frac{p(N-1)}{N-1}]p$$

$$\begin{aligned}
&= np(1-p) + n(n-1)\left[\frac{r-1-pN+p}{N-1}\right]p \\
&= np(1-p) + n(n-1)\left[\frac{r-1-r+p}{N-1}\right]p \\
&= np(1-p) + n(n-1)\left(\frac{1-p}{1-N}\right)p \\
&= np(1-p)\left[1 + \frac{n-1}{1-N}\right] \\
&= np(1-p)\left[1 - \frac{n-1}{N-1}\right]
\end{aligned}$$

finite population correction factor

If we do real sampling but $n \ll N$, $\frac{n-1}{N-1}$ is close to 0 and we forget about it.

Thus, we have at this point shown:

$E(Y)$ is same for binomial and hypergeometric. Also that $Var(Y)$ is close enough to sub and so the binomial can be substituted for hypergeometric.

This means that as a practical matter we can sample