Solution to Mike Strub Challenge
Forest Mensurationists Conference 2022

Mauricio Zapata

November 30, 2022

Announcement

You were recently hired by a forest management company to execute a forest inventory, which has a subsequent objective of finding pole trees within a stand. Using high-density lidar point clouds all dominant and codominant trees were identified. A total of 10,000 trees were found, out of which 500 trees were classified as poles using machine learning techniques. Using Ripley's K you have found that the pole trees are randomly distributed within the stand. Your specific task is to estimate the field effort needed to harvest the pole trees. The forest management company told you that harvest constraints limit the size of the harvest to 10 pole trees. Consequently, together with the company, you defined the field effort as the total number of trees to be assessed to achieve the 10 pole trees. You have decided to use a sampling without replacement strategy and visually inspects all the trees within a sample. In summary, your approach to estimate the field effort is to sample without replacement from the 10,000 trees and visually inspect all the trees within a sample until 10 pole trees are found. The location of each sample is chosen randomly. Because you want to impress the hiring company you will estimate not only the field effort (i.e., expectation of the number of inspected trees) but also the variability associated with the effort (i.e., the variance of the inspected number of trees).

1 Solution

Without loss of generality, let's assume that trees are inspected sequentially (sample size one) and randomly selected from the population. The population consists of $N = 10,000$ trees, where $M = 500$ are marked as pole trees); the sampling is performed without replacement, and the sample continues until $r = 10$ pole trees are assessed.

Assessing a pole tree is considered a success, in other cases, the assessment is viewed as a failure. According to the harvest constraints, r, the number of successes in the

sample, is fixed before sampling. Let's say that the number of trees assessed until the rth success appears is n, then $n = x + r$. Where $x = n - r$ is a random variable of counts of failures until the rth success appears. Therefore, n is a random variable, and $E[n] = E[x] + r$.

If the sample schema is with replacement, The random variable x follow a negative binomial distribution with parameters (r, p) (see Lemma 1.1). Since the sample is taken without replacement, then the random variable x follow a negative hypergeometric distribution of parameters (N, r, p), with $r = 1, 2, \ldots, M$, $p = M/N$, and N, M, r are non-negative integers satisfying the condition $m \leq M \leq N$ (see Lemma 1.2)

Thus, the **field effort**, or the expectation of the number of inspected tres until $r = 10$ pole trees are found, is

$$E[n] = E[x] + r$$

Using the results of the Lemma 1.2:

$$\begin{aligned} E[n] &= r\frac{N-M}{M+1} + r \\ &= r\left[\frac{N+1}{M+1}\right] \end{aligned}$$

Using the values provided by the forest management company, the **field effort** will be

$$E[n] = 10\left[\frac{10000+1}{500+1}\right] = 199.6 \simeq 200$$

The variance of this **field effort** is

$$Var[n] = Var[x] + 0$$

Using the results from Lemma 1.3 we have:

$$Var[n] = r\frac{(N+1)(N-M)}{(M+1)(M+2)}\left(1 - \frac{r}{M+1}\right)$$

Using again the values provided by the forest management company the variance of the **field effort** will be:

$$Var[n] = 10\frac{(10000+1)(10000-500)}{(500+1)(500+2)}\left(1 - \frac{10}{500+1}\right) = 3702.3 \simeq 3702$$

Lemma 1.1. *Let X be a random variable representing the counts of success (i.e., a tree is inspected, and it is a pole tree) in a sample taken with replacement at which the rth success occurs. The probability function of X ($P(X = x)$) is*

$$P(X = x|p, r) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

for integer $x \leq r$. Here $0 < p < 1$, and r is a fixed positive integer. X is a negative binomial random variable. If X is defined as the counts of failures (i.e., a tree is inspected and it is not a pole) before the rth success. Then, the probability function of X ($P(X = x)$) is again a negative binomial random variable with function

$$P(X = x|p, r) = \binom{r+x-1}{x} p^r (1-p)^x$$

for $x = 0, 1, 2, \ldots$

The mean

$$E[X] = r(1-p)/p$$

and variance

$$V[X] = r(1-p)/p^2$$

Lemma 1.2. *Let's assume that the sample schema is without replacement of a finite population U whose elements can be partitioned into two groups C (success) and $\bar{C}$ (failure) with cardinality M and $N - M$, respectively. Assume that the group of units belonging to C or $\bar{C}$ is known. A population unit is taken at random and classified in class C or $\bar{C}$; the sampling continues without replacement until the sample contains m units from class C.*

We can record a sample as $s = \{i_1, i_2, \ldots, i_n\}$ where n refers to the sample size. Let s_c denote the set of sample units from class C with cardinality m and $s_{\bar{c}}$ denote the set of the remaining $n - m$ sample units from class $\bar{C}$. Thus the sample is $s = s_c \cup s_{\bar{c}}$ where $s_c \cap s_{\bar{c}} = \emptyset$. If r (the number of successes in the random sample without replacement) is fixed, with $r = 1, 2, \ldots, m$ and $p = m/n$, the random variable x (number of failures in the sample until the rth success appears) follow a negative hypergeometric distribution with probability function

$$P(X = x) = \frac{\binom{x+r-1}{x}\binom{N-x-r}{M-r}}{\binom{N}{M}}$$

for integer $x = 0, 1, 2, \ldots, N - M$; $0 < p < 1$

Proof

The event $(X = x)$ has a probability that is the product of the two following probabilities:

- *$EI = \{$the sample contains $r - 1$successes and x failures$\}$*

- $EII = \{it\ is\ obtained\ the\ r th\ success\ in\ the\ last\ draw \mid EI\}$

The probability of one particular sample with $r-1$ successes and x failures is

$$\begin{aligned} P(r-1 \text{ successes, } x \text{ failures}) &= \frac{\frac{M!}{(M-(r-1))!}\frac{(N-M)!}{(N-M-x)!}}{\frac{N!}{(N-r-x+1)!}} \\ &= \frac{\frac{(N-x-r+1)!}{(M-r+1)!(N-M-x)!}}{\frac{N!}{M!(N-M)!}} \\ &= \frac{\binom{N-x-r+1}{M-r+1}}{\binom{N}{M}} \end{aligned}$$

The different possible orders in which $r-1$ successes and x failures can appear are:

$$\binom{x+r-1}{x}$$

Therefore the probability of event EI is:

$$P(EI) = \binom{x+r-1}{x}\frac{\binom{N-x-r+1}{M-r+1}}{\binom{N}{M}}$$

The probability of event EII is

$$P(EII) = \frac{M-(r-1)}{N-x-(r-1)}$$

Thus

$$\begin{aligned} P(X=x) &= P(EI) \times P(EII) \\ &= \binom{x+r-1}{x}\frac{\binom{N-x-r+1}{M-r+1}}{\binom{N}{M}} \times \frac{M-(r-1)}{N-x-(r-1)} \\ &= \frac{\binom{x+r-1}{x}\binom{N-x-r}{M-r}}{\binom{N}{M}} \end{aligned}$$

Lemma 1.3. *The expected value of X is*

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{N-M} xP(X=x) \\
&= \sum_{x=0}^{N-M} x \frac{\binom{x+r-1}{x}\binom{N-x-r}{M-r}}{\binom{N}{M}}
\end{aligned}
$$

Khan (1994) and Espejo et al. (2008) showed that for the negative hypergeometric distribution, its expected value is:

$$E[x] = r\frac{N-M}{M+1}$$

And the variance can be derived as follow

$$V[X] = E[X^2] - (E[X])^2 = r\frac{(N+1)(N-M)}{(M+1)(M+2)}\left(1 - \frac{r}{M+1}\right)$$

2 References

- Khan, R.A. (1994). A note on the generating function of a negative hypergeometric distribution. Sankhya : The Indian Journal of Statistics B, 56(3), 309-313.
- Espejo, R.M., H.P. Singh, and S. Saxema. (2008). On inverse sampling without replacement. Statistical Papers 49, 133-137.