Simple Random Sampling

Moulinath Banerjee

University of Michigan

September 11, 2012

1 Simple Random Sampling

The goal is to estimate the mean and the variance of a variable of interest in a finite population by collecting a random sample from it. Suppose there are N members of the population, numbered 1 through N and let the values assumed by the variable of interest be $x_1, x_2, \ldots, x_N$. Not all the x_i's are necessarily distinct (for example, if we are interested in estimating the proportion of Democrats in a population of voters, we might assign $x_i = 1$ if the i'th voter is Democrat and 0 if they are Republican. In this case the population proportion of voters is the mean of the x_i's.). We denote the distinct values of the x_i's by $\xi_1, \xi_2, \ldots, \xi_m$ and let n_i denote the frequency of ξ_i in the population. The population mean μ is given by,

$$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{1}{N}\sum_{i=1}^{m} \xi_i\, n_i\,,$$

and the population variance σ^2 is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{m} \xi_i^2\, n_i - \mu^2\,.$$

We denote the relative frequencies of the ξ_i's in the population by $\{p_1, p_2, \ldots, p_m\}$ where $p_i = n_i/N$.

1.1 SRSWR: simple random sampling with replacement

A sample of size n is collected with replacement from the population. Thus, an individual is drawn (randomly), their x value recorded, and the individual is then returned to the population. Now, a second individual is drawn, and the process continues n times. Let

$X_1, X_2, \ldots, X_n$ be the random variables obtained thus. Then, the X_is are an i.i.d. sample from the distribution of a random variable X such that $P(X = \xi_j) = p_j$ for $j = 1, 2, \ldots, m$. Let

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1} E\left(\sum_{i=1}^{n} (X_i - \overline{X})^2 \right) .$$

Then,

$$E(\hat{\mu}) = \mu \quad \text{and} \quad E(\hat{\sigma}^2) = \sigma^2 .$$

This will be proved in class. The standard CLT can be used to construct a C.I. (confidence interval) for μ.

1.2 SRSWOR: simple random sampling without replacement

A sample of size n is collected without replacement from the population. Thus the first member is chosen at random from the population, and once the first member has been chosen, the second member is chosen at random from the remaining $N-1$ members and so on, till there are n members in the sample. A typical sample therefore looks like $(k_1, k_2, \ldots, k_n)$ and the number of all possible ordered samples is easily seen to be $N \times (N-1) \times \ldots (N-n+1)$ and each ordered sample has equal probability, namely $(\Pi_{l=1}^{n} (N-l+1))$ of being selected. Let $X_1, X_2, \ldots X_n$ denote the observed value of the variables for the members in the sample; thus $X_1 = x_{k_1}, \ldots, X_n = x_{k_n}$. The X_i's are random variables and X_1 is easily seen to have marginal distribution given by,

$$P(X_1 = \xi_j) = p_j \, , \; j = 1, 2, \ldots, m \, .$$

In fact each X_i has the same distribution as X_1. To see this note that the event $\{X_i = \xi_1\}$ happens if and only if the i'th member of the sample is one of the members of the population whose variable value equals ξ_1. Without loss of generality assume that the first n_1 members have variable value equal to ξ_1; in other words $x_1 = x_2 = \ldots = x_{n_1} = \xi_1$. Now, the number of ordered samples in which the i'th member is 1, is precisely $\Pi_{l=1}^{n-1} (N-1-l+1) = \Pi_{l=1}^{n-1} (N-l)$. Thus the chance that the i'th member is 1 is precisely $\Pi_{l=1}^{n-1} (N-l)/(\Pi_{l=1}^{n} (N-l+1)) = 1/N$. Similarly, the chance that the i'th member is s where s is between 1 and n_1 is $1/N$. Thus, the chance that $\{X_i = \xi_1\}$ is simply $n_1 \times 1/N = n_1/N = p_1$. It can be argued similarly that $P(X_i = \xi_j) = p_j$ for all j.

We now consider the joint distribution of (X_i, X_j) for $i \neq j$. We will show that the joint distribution of (X_i, X_j) for any $i \neq j$ is the same as that of (X_1, X_2). Now, it is not difficult to see that,

$$P(X_1 = \xi_s, X_2 = \xi_r) = \frac{n_s}{N} \frac{n_r}{N-1} \, , \text{ for } s \neq r \, ,$$

and

$$P(X_1 = \xi_s, X_2 = \xi_s) = \frac{n_s}{N}\,\frac{n_s - 1}{N - 1}\,.$$

Now, consider $P(X_i = \xi_s, X_j = \xi_r)$ for $i \neq j$, and with r and s different. Without loss of generality let $i < j$ and also let $x_1 = x_2 = \ldots = x_{n_s} = \xi_s$ and let $x_{n_s+1} = \ldots = x_{n_s+n_r} = \xi_r$. Thus the event $\{X_i = \xi_s, X_j = \xi_r\}$ is the disjoint union of the events $\{k_i = u, k_j = v\}$ where $1 \leq u \leq n_s$ and $n_s + 1 \leq v \leq n_s + n_r$, and there are $n_s\, n_r$ such pairs. Now, the number of ordered samples that lead to $k_i = u$ and $k_j = v$ is precisely $n_{u,v} = (N-2) \times (N-3) \times \ldots \times (N - 2 - (n-2) + 1)$ (since we are fixing the i'th and the j'th members at u and v respectively and then choosing $n-2$ distinct integers out of the remaining $N-2$ population members. Hence the required probability is $n_{u,v}/N \times (N-1) \times \ldots (N-n+1)$ and this is just $1/N\,(N-1)$. Thus the probability of the event $\{X_i = \xi_s, X_j = \xi_r\}$ is just $n_s\, n_r \times 1/N\,(N-1)$ which is the same as the probability that $\{X_1 = \xi_s, X_2 = \xi_r\}$. The case when $r = s$ can be similarly handled.

We are now in a position to study the properties of the sample-based estimates of μ and σ^2. We estimate μ by $\hat{\mu} = (X_1 + X_2 + \ldots + X_n)/n = \overline{X}$ and the sample variance by $\hat{\sigma}^2 = (1/(n-1)) \sum_{i=1}^{n} (X_i - \overline{X})^2$. In the sampling with replacement case, we have seen that $\hat{\mu}$ and $\hat{\sigma}^2$ are unbiased estimates of μ and σ^2 respectively. In the SRSWOR case $X_1, X_2, \ldots, X_n$ are identically distributed as X_1 and it is easy to check that $E(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Now,

$$E(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n\,\mu = \mu\,.$$

Thus as in SRSWR, $\overline{X}$ is an unbiased estimator of μ. In SRSWR, $\mathrm{Var}(\overline{X}) = \sigma^2/n$; however this is not the case with SRSWOR because the X_i's are not independent and the correlation factor consequently needs to be taken into account, while computing the variance of $\overline{X}$. To compute the variance of $\overline{X}$ we proceed as follows.

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \frac{1}{n^2} \sum_{i,j} \mathrm{Cov}\,(X_i, X_j) \\
&= \frac{1}{n^2} \left(\sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \right) \\
&= \frac{1}{n}\,\sigma^2 + \frac{n-1}{n}\,\mathrm{Cov}(X_1, X_2)\,.
\end{aligned}$$

Now,

$$\mathrm{Cov}(X_1, X_2) = E(X_1\, X_2) - \mu^2\,.$$

Also, letting p_{ij} denote the probability that $X_1 = \xi_i$ and $X_2 = \xi_j$, we have

$$\begin{aligned}
E(X_1\,X_2) &= \sum_{i,j} \xi_1\,\xi_j\,p_{ij} \\
&= \sum_{i=1}^{m} \xi_i\,p_i \sum_{j=1}^{m} \xi_j\,\frac{p_{ij}}{p_i} \\
&= \sum_{i=1}^{m} \xi_i\,p_i \left(\sum_{j=1}^{m} \xi_j\,\frac{n_j}{N-1} - \frac{\xi_i}{N-1} \right) \\
&= \sum_{i=1}^{m} \xi_i\,p_i \left(\sum_{j=1}^{m} \xi_j\,\frac{N}{N-1}\,p_j - \frac{\xi_i}{N-1} \right) \\
&= -\sum_{i=1}^{m} \frac{1}{N-1}\,\xi_i^2\,p_i + \frac{N}{N-1} \left(\sum_{i=1}^{m} \xi_i\,p_i \right)^2 \\
&= -\frac{1}{N-1}(\sigma^2 + \mu^2) + \frac{N}{N-1}\,\mu^2 \\
&= -\frac{1}{N-1}(\sigma^2) + \mu^2\,.
\end{aligned}$$

Thus,

$$\mathrm{Cov}(X_1, X_2) = -\frac{1}{N-1}(\sigma^2) + \mu^2 - \mu^2 = -\frac{1}{N-1}(\sigma^2)\,.$$

Now, plugging the above into the expression for $\mathrm{Var}(\overline{X})$, we get,

$$\mathrm{Var}(\overline{X}) = \frac{1}{n}\,\sigma^2\,(1 - \frac{n-1}{N-1})\,.$$

Thus, we get what is called " a finite population correction factor" for the variance.

Now, if we try to estimate σ^2 as before, by s^2, where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\,,$$

the estimate is no longer unbiased (in contrast to what happens with SRSWR). Rather,

$$
\begin{aligned}
E(s^2) &= \frac{1}{n-1}\, E\left(\sum_{i=1}^{n} (X_i - \overline{X})^2\right) \\
&= \frac{1}{n-1}\, E\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \\
&= \frac{1}{n-1}\, \left(\sum_{i=1}^{n} E(X_i^2) - n\, E(\overline{X}^2)\right) \\
&= \frac{1}{n-1}\, \left(n\,\sigma^2 + n\,\mu^2 - n\, \left(\mathrm{Var}(\overline{X}) + \mu^2\right)\right) \\
&= \frac{1}{n-1}\, \left(n\,\sigma^2 - n\,\frac{1}{n}\,\frac{N-n}{N-1}\,\sigma^2\right) \\
&= \frac{1}{n-1}\,\sigma^2\,\frac{nN - n - N + n}{N-1} \\
&= \frac{1}{n-1}\,\sigma^2\,(n-1)\,\frac{N}{N-1} \\
&= \sigma^2\,\frac{N}{N-1}\,.
\end{aligned}
$$