Sampling from Finite Populations

May 5, 1999

1 Sampling

Sampling surveys are used to obtain information about a large population by examining only a small fraction of that population. Examples: Census surveys, agricultural survesy, political surveys. In random sampling each member of the population has a specified probability of being included in the sample and the actual composition of the sample is random.

Advantages:

Guard against biass
Less costly than complete enumeration
May be more accurate than complete count
Can calulate error estimates
Can control size of error by selecting sample size

Population Parameters:

Let x_i be numerical value of interested associated with the ith member of the population $i = 1, \ldots, N$.

Let

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\tau = N\mu = \sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Definition. A choice of a subset of size n from a larger set of size N is called a *simple random sample* if each of the N choose n possible subsets of size n taken without replacement is equally likely to be selected.

Example: Suppose $N = 4$ and $n = 2$ then a simple random sample of size 2 consists of selecting one of the six subsets of size 2: $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, $\{3, 4\}$ where each of the subsets has equal probability of being selected, i.e., $1/6$.

1.1 How to do random sampling

Naive way: List all K subsets of size n from the set of N elelments. Generarte a uniform $[0, 1]$ random number and select the ith sample if $U \in (i - 1/K, i/K]$. This is a lot of work: For $N = 393$ and $n = 16$ we have $K \geq 10^{33}$.

Better way: Let $I_k = 1$ if the kth member is selected and 0 otherwise. Let us compute the conditional distribution of I_k given $I_1, \ldots, I_{k-1}$, First $P(I_1 = 1) = n/N$, now $P(I_2 = 1|I_1 = 0) = n/(N-1)$ and $P(I_2 = 1|I_1 = 1) = (n-1)/(N-1)$. Notice that

$$\begin{aligned} P(I_2 = 1) &= P(I_2 = 1|I_1 = 0)P(I_1 = 0) + P(I_2 = 1|I_1 = 1)P(I_1 = 1) \\ &= \frac{n}{N-1}\frac{N-n}{N} + \frac{n-1}{N-1}\frac{n}{N} \\ &= \frac{n}{N}. \end{aligned}$$

More generally,

$$P(I_j = 1|I_1, \ldots, I_{j-1}) = \frac{n - \sum_{k=1}^{j-1} I_k}{N+1-j},$$

and

$$P(I_j = 1) = \frac{n}{N}.$$

Let $U_1, U_2, \ldots$ be a sequence of independent uniform $[0,1]$ random variables. And let $I_1 = 1$ if $U_1 < n/N$ and in general let $I_j = 1$ if $U_j < \frac{n - \sum_{k=1}^{j-1} I_k}{N+1-j}$ and $I_j = 0$ otherwise. Stop when $I_1 + I_2 + \ldots + I_j = n$.

Example: Let $n = 2$ and $N = 4$. Then $U_1 = 0.0051 < 2/4$ so $I_1 = 1$ and item one is included in the sample. Next $U_2 = 0.3561 > 1/3$ so item 2 is not included in the sample. Next $U_3 = 0.455 \leq 1/2$ so item 3 is included in the sample and we are done.

1.2 Estimating the Population Parameters

Since the composition of the sample is random, the sample mean is random. An analysis of the accuracy with which the sample mean approximates the population mean must therefore be probabilistic in nature.

Assume a simple random sample of size $n < N$ is drawn from the population. We will denote the values of the sample members by $X_1, X_2, \ldots, X_n$. In particular, X_i is not the same as x_i: X_i is the numerical value of the ith member of the sample, which is random and x_i is the ith member of the population which is fixed.

We will consider the *sample mean*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an estimate of the population mean, and

$$T = N\bar{X}$$

as an estimate of the population total.

1.3 The distribution of X_i

Denote the distinct values of the population by $y_1, \ldots, y_m$ and denote the number of population members that have the value y_j by $n_j, j = 1, \ldots, m$. Then X_i is a discrete random variable with probability mass function

$$P(X_i = y_j) = n_j/N, \quad j = 1, \ldots, m.$$

Then

$$E[X_i] = \sum_{j=1}^{m} y_j n_j/N$$

$$\begin{aligned} &= \sum_{i=1}^{N} x_i/N \\ &= \mu. \end{aligned}$$

Similarly,

$$\begin{aligned} \text{Var}[X_i] &= E[X_i^2] - E[X_i]^2 \\ &= \sum_{j=1}^{m} y_j^2 n_j/N - \mu^2 \\ &= \sigma^2, \end{aligned}$$

where we have use the fact that $\sum_{i=1}^{N} x_i^2 = \sum_{j=1}^{m} n_j y_j^2$.

We are now ready to compute the mean of $\bar{X}$.

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1} n\mu = \mu.$$

Also,

$$E[N\bar{X}] = N\mu = \tau.$$

Notice, however, that X_i's are NOT independent! Notice that if X_i takes a large value then X_j $j \neq i$ will take on average a smaller number, so X_i and X_j are negatively correlated. Remember we are sampling without replacement.

$$Var[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j)$$

We already know that $Cov(X_i, X_i) = \sigma^2$ and it should be clear that $Cov(X_i, X_j)$ should be a constant, say c independent of i and j. Thus,

$$\text{Var}[\bar{X}] = \frac{1}{n^2}[n\sigma^2 + (n^2 - n)c]$$

It turns out that $c = -\sigma^2/(N-1)$. The derivation of this is fairly tedious so we just state it as a fact, although you may want to prove this for yourself. We now have

$$\text{Var}[\bar{X}] = \frac{1}{n^2}[n - (n^2 - n)/(N-1)]\sigma^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

The second term is less than one, and is known as the finite popilation correction. If we ignore the fpc we are overestimating the variance of $\bar{X}$.

Notice that

$$\text{Var}[T] = N^2 \text{Var}(\bar{X}).$$

1.4 Estimating σ^2

Without the finite population correction, we know that

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estiamtor of σ^2. With finite sampling, however, we obtain

$$E[S^2] = \sigma^2 \frac{N}{N-1},$$

so $\frac{N-1}{N} S^2$ is an unbiased estimator of σ^2.

1.5 Estimating Var$[\bar{X}]$

Well

$$Var[\bar{X}] = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

An an unbiased estimate of $Var[\bar{X}]$ is obtained by simply multiplying $\frac{N-1}{N}S^2$ by $\frac{1}{n}\frac{N-n}{N-1}$ to obtain

$$S^2_{\bar{X}} = \frac{N-n}{N}\frac{S^2}{n}$$

To sumarize, we know that the mean and the variance of $\bar{X}$ and we know how to estimate Var$[\bar{X}]$.

A version of the CLT shows that if n is large, but small relative to N, then $\bar{X}$ is approximatetly normal with mean μ and variance $\frac{\sigma^2}{n}\frac{N-n}{N-1}$, so the probability that μ lies within $\bar{X} \pm z(\alpha/2)\sigma_{\bar{X}}$ is approximately $100(1-\alpha)\%$. Similarly, a $100(1-\alpha)\%$ confidence interval on $T = N\bar{X}$ is given by $N\bar{X} \pm z(\alpha/2)N\sigma_{\bar{X}}$.

1.6 Ratio Estimates

Let N be the population size and let x_j $j = 1, \ldots, N$ and y_j $j = 1, \ldots, N$ denote two measurable characteristics of the population. For example x_j can denote the book value of inventory of item i while y_j may denote the actual value of the inventory of item j. Suppose we know the x_j's but do not know the y_j's.

Suppose we want to estimate the total *value* of all the items $\tau_y = \sum_{j=1}^N y_j$. Without a sample an estimate of τ_y would be the estimate would be $\tau_x = N\mu_x = \sum_{j=1}^N x_j$ the total book value of the items in the population.

On the other hand, if a sample of size n is taken an estimate of τ_y that ignores book values is

$$N\bar{Y} = N\frac{1}{n}\sum_{i=1}^n Y_i$$

where Y_i is the actual value of the ith item in the sample.

It may be that the x_j's are highly correlated with the y_j's. In this case we may want to use the information on book values to estimate τ_y. One way to do this is by using the estimate

$$\frac{\mu_x}{\bar{X}}N\bar{Y} = N\mu_x\frac{\bar{Y}}{\bar{X}} = N\mu_x R,$$

where $R = \bar{Y}/\bar{X}$.

You can think of $N\mu_x R = \tau_x R$ in two ways. First, you may think of R as providing you with an estimate of the actual to book value ratio of the inventory, and correcting the estimate τ_x by the ratio R. Alternatively, you can think of $\frac{\mu_x}{\bar{X}}$ as a correction to the estimate $N\bar{Y}$. Notice that if the sample happens to contain low values items then $N\bar{Y}$ will underestimate $\sum_j y_j$, but the factor $\mu_x/\bar{X} > 1$ will partially correct this. On the other hand, if the sample happens to contain large values items then $N\bar{Y}$ will overestimate $\sum_j y_j$, but the factor $\mu_x/\bar{X} < 1$ will partially correct this.

Question: How does the two estimators $N\mu_x R$ and $N\bar{Y}$ of $\sum_j y_j$ compare?

Most people judge estimators by their mean square errors. If the estimator Z is used to estimate z then the mean square error is $E[(Z-z)^2]$. If $E[Z] = z$ we say that the estimator is unbiased and the mean square error reduces to the variance of Z. More generally, if $E[Z] = \mu$ and $Var[Z] = \sigma^2$ we have

$$E[(Z-z)^2] = \sigma^2 + (\mu - z)^2.$$

The second term $\mu - z$ is called the bias of the estimator.

It turns out that if $\bar{X}$ and $\bar{Y}$ are highly correlated the MSE of $N\mu_x R$ can be much smaller than the MSE of $N\bar{Y}$ although $\tau_x R$ is typically a biased estimator of τ_y.

Let $r = \frac{\mu_y}{\mu_x}$ and

$$R = \frac{\bar{Y}}{\bar{X}}.$$

We need to compute the mean and the variance of R. The approximate variance of R is given by

$$Var[R] \simeq \frac{1}{n}\frac{N-n}{N-1}\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right).$$

Unfortunately, we may not know the values of σ_x^2, σ_Y^2 and σ_{XY}. In this case, we may estimate them, respectively, by

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})^2,$$

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2,$$

and

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}),$$

obtaining an estimate of Var$[R]$:

$$s_R^2 = \frac{1}{n}\frac{N-n}{N-1}\frac{1}{\bar{X}^2}(R^2 s_X^2 + s_Y^2 - 2Rs_{XY}).$$

An approximate $100(1-\alpha)\%$ confidence interval for r is

$$R \pm z(\alpha/2)s_R.$$

Going back to our case, where we do know μ_x, we can use instead

$$s_R^2 = \frac{1}{n}\frac{N-n}{N-1}\frac{1}{\mu_x^2}(R^2 s_X^2 + s_Y^2 - 2Rs_{XY}).$$

Then an approximate $100(1-\alpha)\%$ confidence interval for μ_y is

$$\mu_x R \pm z(\alpha/2)\mu_x s_R.$$

Finally, an approximate $100(1-\alpha)\%$ confidence interval for τ_y is

$$\tau_x R \pm z(\alpha/2)\tau_x s_R.$$

For the hospital data discussed in class, the 95% confidence interval for the total number of discharges based on $N\bar{Y}$ was $(302752, 430146)$ and had a width equal to 127,394. On the other hand, the 95% confidence interval based on $\tau_x R$ was $(293681, 335790)$ and had a widt of 42,108.