Definition 1 A sample obtained by randomly selecting one element from the first k elements in the frame and every kth element thereafter is called a 1-in-k systematic sample, with a random start.

1 Estimation of a Population Mean and Total

Estimator of the population mean $\mu$:

$$\hat{\mu} = \bar{y}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$ (1)

where the subscript $sy$ signifies that systematic sampling was used.

Estimated variance of $\bar{y}_{sy}$:

$$V(\bar{y}_{sy}) = \frac{s^2}{n} \left( \frac{N-n}{N} \right)$$ (2)

assuming a randomly ordered population.

The exact variance of $\bar{y}_{sy}$ is given by

$$V(\bar{y}_{sy}) = \frac{\sigma^2}{n} [1 + (n - 1)\rho]$$ (3)

where $\rho$ is a measure of the correlation between pairs of elements within the same systematic sample.

Definition 2 A population is random if the elements of the population are in random order.

Estimator of the population total $\tau$:

$$\hat{\tau} = N \bar{y}_{sy}$$ (4)

Estimated variance of $\tau$:

$$V(N\bar{y}_{sy}) = N^2 V(\bar{y}_{sy}) = N^2 \left( \frac{s^2}{n} \right) \left( \frac{N-n}{N} \right)$$ (5)

assuming a randomly ordered population.

2 Estimation of a Population Proportion

Estimator of the population proportion $p$:

$$\hat{p}_{sy} = \bar{y}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$ (6)

Estimated variance of $\hat{p}_{sy}$:

$$V(\hat{p}_{sy}) = \frac{\hat{p}_{sy} \hat{q}_{sy}}{n-1} \left( \frac{N-n}{N} \right)$$ (7)

where $\hat{q}_{sy} = 1 - \hat{p}_{sy}$ assuming a randomly ordered population.
3 Selecting the Sample Size

Sample size required to estimate $\mu$ with a bound $B$ on the error of estimation:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

(8)

where $D = \frac{\mu^2}{4}$.

4 Repeated Systematic Sampling

Instead of choosing a 1-in-$k$ systematic sample with size $n$, we draw $n_s$ 1-in-$k'$ systematic samples, where $k' = kn_s$.

Estimator of the population mean $\mu$ using $n_s$ 1-in-$k'$ systematic samples:

$$\hat{\mu} = \sum_{i=1}^{n_s} \frac{\bar{y}_i}{n_s}$$

(9)

where $\bar{y}_i$ represents the average of the $i$th systematic sample.

Estimated variance of $\hat{\mu}$:

$$\hat{V}(\hat{\mu}) = \left(\frac{N-n}{N}\right) \frac{\sum_{i=1}^{n_s} (\bar{y}_i - \hat{\mu})^2}{n_s(n_s - 1)}$$

(10)

We can also use repeated systematic sampling to estimate a population total $\tau$, if $N$ is known. The formulas necessary are easy to deduce.

5 Further Discussion of Variance Estimators

Cluster mean square, within-cluster mean square, and total sum of squares are given by

$$\text{MSB} = \frac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2$$

(11)

$$\text{MSW} = \frac{1}{k(n - 1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

(12)

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2$$

(13)

Using these terms,

$$\rho = \frac{(k-1)n\text{MSB} - \text{SST}}{(n-1)\text{SST}}$$

(14)
which, for large $N = nk$, is approximately

$$
\rho \approx \frac{\text{MSB} - \text{MST}}{(n-1)\text{MST}} \tag{15}
$$

where $\text{MST} = \text{SST}/(nk - 1)$.

Suppose $y_1, y_2, \ldots, y_n$ is a random sample with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$. The usual estimator of $\sigma^2$ is based on $\sum (y_i - \bar{y})^2$, but if we knew that $\mu = 0$, the estimator of $\sigma^2$ would be based on $\sum y_i^2$, and $\sum y_i^2/n$ would be an unbiased estimator of $\sigma^2$.

Now, suppose $\mu$ is not zero. Choose two sample values, $y_i$ and $y_j$, and construct $d_i = y_i - y_j$. It follows that $E(d_i) = 0$ and $V(d_i) = 2\sigma^2$. If we make up $n_d$ such differences, then $\sum_{i=1}^{n_d} d_i^2 / n_d$ is an estimator of $2\sigma^2$, from which an estimator of $\sigma^2$ is easily obtained. Since we want to estimate a mean $\bar{y}_{sy}$ with a sample of $n$ measurements from a population of $N$ measurements, the estimator based on $d_i$ becomes

$$
\hat{V}_d(\bar{y}_{sy}) = \frac{N - n}{Nn} \frac{1}{2n_d} \sum_{i=1}^{n_d} d_i^2 \tag{16}
$$

Instead of making arbitrary differences, we usually take successive differences of the form

$$
d_i = y_{i+1} - y_i, \quad i = 1, \ldots, (n-1)
$$

6 Exercise

7.1 Systematic sampling is better than random sampling since population is ordered.

7.3 (a)(b)

> population <- c(rep(1, 4), rep(0, 7), rep(1, 3), rep(0, 6), rep(1, 5), rep(0, 3), rep(1, 4), rep(0, 8))

> sam10 <- sys.sample(data = population, k = 10)

> sam10.mean <- apply(sam10, 2, mean)

> var(sam10.mean) * 9/10
[1] 0.1275

> sam5 <- sys.sample(data = population, k = 5)

> sam5.mean <- apply(sam5, 2, mean)
> var(sam5.mean) * 4/5
[1] 0.00875

7.4

> S <- sys.cal(data = c(rep(1, 132), rep(0, 68)), k = 10, N = 2000)

> S$mu
[1] 0.66

> S$mu.B
[1] 0.06243875

7.5

> D <- 0.01^2/4

> n <- (2000 * S$mu * (1 - S$mu)) / (1999 * D + S$mu * (1 - S$mu))

> n
[1] 1635.718

7.6

> Y <- c(12, 11.97, 12.01, 12.03, 12.01, 11.8, 11.91, 11.98, 12.03, 11.98, 12, 11.83, 11.87, 12.01, 11.98, 11.87, 11.9, 11.88, 12.05, 11.87, 11.91, 11.93, 11.94, 11.89, 11.75, 11.93, 11.95, 11.97, 11.93, 12.05, 11.85, 11.98, 11.87, 12.05, 12.02, 12.04)

> S <- sys.cal(data = Y, k = 50, N = 1800, q = 2)

> S$mu
[1] 11.94556

> S$mu.B
[1] 0.02514864

7.7
STAT 631  
Survey Sampling  
Fall 2003

> D <- 0.03^2/4

> n <- 1800 * var(Y)/(1799 * D + var(Y))

> n
[1] 25.46343

7.16

> Y <- c(7030, 6720, 6850, 7210, 7150, 7370, 7000, 6930, 6570, 6910, 7380, 7540, 6720, 6900, 7200, 7100, 6860, 6800, 7050, 7420, 7090)

> S <- sys.cal(data = Y, k = 25, N = 520, q = 2)

> S$mu
[1] 7038.095

> S$mu.B
[1] 108.7363

7.20(b) and 7.27(d)

> X <- seq(1950, 1990, by = 5)

> Y <- c(3642, 4097, 4258, 3760, 3731, 3144, 3612, 3761, 4158)

> Z <- c(24.1, 25, 23.7, 19.4, 18.4, 14.6, 15.9, 15.8, 16.7)

> S = sys.cal(data = Z, k = 5, N = 41)

> S$mu
[1] 19.28889

> S$mu.var
[1] 1.391533

> S$mu.B
Figure 1: Exercise 7.20(b)

[1] 2.312037

> postscript(file = "ex720b.ps")

> plot(x, Z)

> dev.off()

null device

1

> d <- Z[2:length(Z)] - Z[1:(length(Z) - 1)]

> sq <- sum(d^2)/2/length(d)

> mu.var <- sq/length(Z) * (41 - length(Z))/41

> mu.var
[1] 0.2110569

7.21 and 7.27(e)
> Y <- c(385, 377, 393, 479, 708, 1036, 1189, 1190, 1175)

> postscript(file = "ex721.ps")

> plot(X, Y)

> dev.off()
null device
   1

> S <- sys.cal(data = Y, k = 5, N = 41)

> S$tau
[1] 31579.11

> S$tau.var
[1] 20410156

> S$tau.B
[1] 8854.647

> d <- Y[2:length(Y)] - Y[1:(length(Y) - 1)]

> sq <- sum(d^2)/2/length(d)

> mu.var <- sq/length(Y) * (41 - length(Y))/41

> 41^2 * mu.var
[1] 1743648
Figure 2: Exercise 7.21