Systematic Sampling

1. Systematic sampling involves random selection of one element from the first k elements and then selection of every kth element thereafter. k is called the sampling interval and is computed as $k = \mathrm{Int}\left(\frac{N}{n}\right)$ where Int( ) is the integer function returning the integer part of a real number.
2. Reasons for use:

- Usually easier to perform than simple random sampling; costs may be lower per unit to sample; often much easier to train personel in its use; sampling protocol may be more easily followed.
- Can give more information per unit of cost than simple random sampling as the sample is spread out more uniformly over the population. This is often important when sampling in space or time.
- Can be used when the frame is not known prior to sampling. The frame is constructed as the sample is taken.

3. Use the same formulas as for simple random sampling to estimate the population mean, total, and proportion.
4. Variance formulas, however, are problematic. Systematic sampling can be viewed as cluster sampling where a sample of size $m_i = 1$ unit per cluster is taken. From cluster sampling the variance of the systematic sampling scheme can be derived as

$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}[1 + (n-1)\rho]$ where ρ is the intracluster correlation coefficient defined as

$$\rho = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l<j}^{n}(y_{ij}-\bar{y})(y_{il}-\bar{y})}{nk(n-1)\sigma^2}$$ where y_{ij} and y_{il} are from the ith cluster.

5. If the population is randomly ordered then $\rho \approx 0$ and systematic sampling is approximately the same as simple random sampling. This is the typical assumption made.
6. If the elements are ordered in magnitude, then $\rho \le 0$ and so for large N, $\mathrm{Var}(\bar{y}_{sys}) \le \mathrm{Var}(\bar{y}_{SRS})$ and so systematic sampling (sys) is superior to simple random sampling (SRS).
7. If the population is periodic and cycles according to the response, then $\rho > 0$ and so for large N, $\mathrm{Var}(\bar{y}_{sys}) \ge \mathrm{Var}(\bar{y}_{SRS})$ and then systematic sampling is inferior to simple random sampling.

8. In practice, we may not know the ordering of the population, thus we may not know whether or not systematic sampling is approximately the same, better, or much worse than simple random sampling. Cochran also states that the variance of the systematic sample estimator may also increase when a larger sample is taken.
9. Schaeffer et al. (1996) suggest that you plot the data values versus the sample number and study the resulting pattern. If the pattern appears random, then apply the usual simple random sampling formulas. If, however, the pattern appears non-random, then they suggest the use of first differences for estimating the population variance.

   For values from the same population, $E(y_i) = \mu$ so that $E(y_i - y_{i'}) = 0$. Further, $\mathrm{Var}(y_i - y_{i'}) = 2\sigma^2$ if we ignore the dependence among the units in finite population sampling. Thus, use the estimator based upon the $n-1$ first differences $d_i = y_{i+1} - y_i$,

   $\mathrm{var}_d(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_d^2}{n}$ where $s_d^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1} d_i^2$ to provide an alternative estimate of the variance of the mean that is more independent of the trend in the data.

Repeated Systematic Sampling

1. Given a population of size N, we want to take a systematic sample of size n. In the usual procedure we would take a 1 in k sample, where $k = \frac{N}{n}$. In repeated systematic sampling, rather than take 1 systematic sample, we take m systematic samples. To hold n constant, we take m one in M systematic samples, where $M = mk = m\frac{N}{n}$.

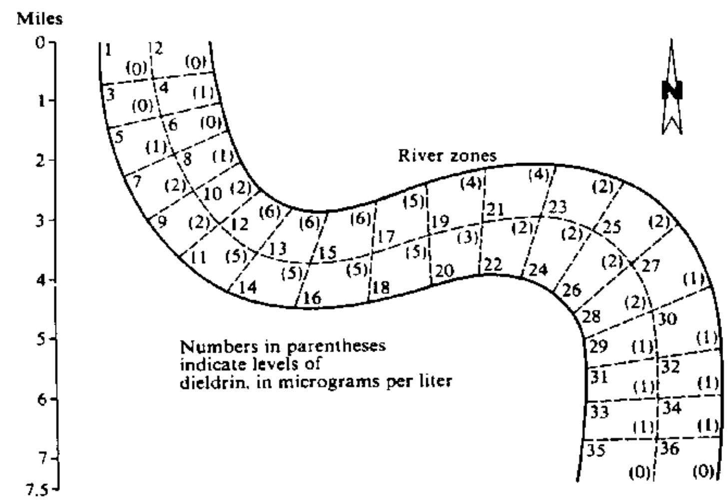
2. Choose m random numbers without replacement between 1 and M, then use the interval M for each systematic sample. This will result in $n' = \frac{N}{M} = \frac{n}{m}$ units for each sample.

3. Let $\bar{y}_i$ be the mean of the ith systematic sample: $\bar{y}_i = \frac{1}{n'}\sum_{j=1}^{n'} y_{ij}$.

4. $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i$ is an unbiased estimator for μ.

5. $\text{var}(\bar{y}) = \frac{s_m^2}{m}\left(\frac{M-m}{M}\right)$ where $s_m^2 = \frac{\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2}{m-1}$ and $M = \frac{N}{n'}$.

Example 23: (From Problem 4.5, Levy and Lemeshow 1991:95-96) Suppose that a study is planned of the level of the pesticide dieldrin, which is believed to be a carcinogen, in a 7.5 mile stretch of a particular river. To assure representativeness, a map of the river is divided into 36 zones and systematic sampling is to be used. Water samples will be drawn by taking a boat out to the geographic center of the designated zone, and drawing a grab sample of water from a depth of several centimeters below the surface level. The levels of dieldrin, in micrograms per liter, for each of these zones are shown on the map in parentheses.



1. Take a single 1-in-4 systematic sample (for n=9) and describe how you selected the sample, list your sample values, and estimate the mean level of dieldrin in this stretch of the river. Further, give a 95% confidence interval estimate for the mean.
2. Take 3 replications of a systematic sample for a total sample size of n=9 as in (a) above and describe how you selected the 3 samples, list your sample values for each replication, and estimate the mean level of dieldrin in this stretch of the river. Further, give a 95% confidence interval estimate for the mean.

Perform the requested analysis using SURVEYSELECT and SURVEYMEANS

```
*****************************************************************;
* systematic.sas - Selection of Systematic Sample from River   *;
* zones for determination of Dieldrin levels. Problem 4.5 in   *;
* Levy and Lemeshow 1991:95-96.                                 *;
*****************************************************************;
Options LS=78 PS=60 PageNo=1 NoDate NoCenter
        FORMCHAR='|----|+|---+=|-/\<>*';
Title1 "Systematic Sampling of River Zones";
Data River;
 Input Zone Dieldrin @@;
 Label Zone="River Zone" Dieldrin="Dieldrin ug/l";
Datalines;
1 0 2 0 3 0 4 1 5 1 6 0 7 2 8 1 9 2 10 2
11 5 12 6 13 6 14 5 15 6 16 5 17 5 18 5 19 4 20 3
21 4 22 2 24 2 23 2 26 2 25 2 27 1 28 2 29 1
30 1 31 1 32 1 33 1 34 1 35 0 36 0
;

Proc Sort Data=River;
 By Zone;
Run;
*Proc Print Data=River NoObs;
* Var Zone Dieldrin;
*Run;

Title2 "Single Systematic Sample";
Proc SurveySelect Data=River Out=Sample1
     Seed=25645834 Method=SYS Rate=0.25 Stats;
Run;


Proc Print Data=Sample1;
 Var Zone Dieldrin;
Run;

Title3 "Summary Statistics";
Proc Means Data=Sample1 N Mean Var StdErr;
 Var Dieldrin;
Run;

Title3 "Estimation Under Assumption of Approximate SRS";
Proc SurveyMeans Data=Sample1 Rate=0.25 Mean CLM CV;
 Var Dieldrin;
 Weight SamplingWeight;
Run;
```



```
*****************************************************************;
* Now use repeated systematic sampling: For m=3 replications   *;
* with a total sample size of n=9 from a population of N=36    *;
* zones, we will take m=3 1-in-k=m(N/n)=3(36/9)=12 systematic  *;
* samples. Thus RATE=1/12=0.0833333.                            *;
*****************************************************************;
Title2 "Repeated Systematic Samples";
Proc SurveySelect Data=River Out=Sample2 Method=SYS
     Seed=25645834 Rep=3 Rate=0.083333333 Stats;
Run;

Data Sample2;
 Set Sample2;
 SamplingWeight=SamplingWeight/3;
Run;

Proc Print Data=Sample2;
 Var Replicate Zone SamplingWeight Dieldrin;
Run;

Title3 "Summary Statistics of Individual Replicates";
Proc Means Data=Sample2 N Mean Var StdErr;
 Class Replicate;
 Var Dieldrin;
 Output Out=Summary N=N Mean=Dieldrin;
Run;

Title3 "Summary Statistics of Replication Means";
Proc Means Data=Summary N Mean Var StdErr;
 Where (_TYPE_=1);
 Var Dieldrin;
Run;

Title3 "Estimation Under Repeated Systematic Samples";
Proc SurveyMeans Data=Sample2 Rate=0.25 Mean CLM;
 Cluster Replicate;
 Var Dieldrin;
 Weight SamplingWeight;
Run;

Title3 "Variance Components for Intracluster Correlation";
Proc Mixed Data=Sample2 CovTest;
 Class Replicate;
 Model Dieldrin=;
 Random Replicate;
 Parms / Nobound;
Run;
```

```
Systematic Sampling of River Zones                                        1
Single Systematic Sample
The SURVEYSELECT Procedure

Selection Method    Systematic Random Sampling
Input Data Set               RIVER
Random Number Seed        25645834
Sampling Rate                 0.25
Sample Size                      9
Selection Probability         0.25
Sampling Weight                  4
Output Data Set            SAMPLE1

Systematic Sampling of River Zones                                        2
Single Systematic Sample

Obs    Zone    Dieldrin
 1       2         0
 2       6         0
 3      10         2
 4      14         5
 5      18         5
 6      22         2
 7      26         2
 8      30         1
 9      34         1

Systematic Sampling of River Zones                                        3
Single Systematic Sample
Summary Statistics
The MEANS Procedure
    Analysis Variable : Dieldrin Dieldrin ug/l

N            Mean        Variance       Std Error
-------------------------------------------------
9       2.0000000       3.5000000       0.6236096
-------------------------------------------------

Systematic Sampling of River Zones                                        4
Single Systematic Sample
Estimation Under Assumption of Approximate SRS

The SURVEYMEANS Procedure

            Data Summary
Number of Observations             9
Sum of Weights                    36

                              Statistics
                         Std Error     Lower 95%     Upper 95%      Coeff of
Variable          Mean     of Mean    CL for Mean   CL for Mean     Variation
-----------------------------------------------------------------------------
Dieldrin      2.000000    0.540062      0.754615      3.245385      0.270031
-----------------------------------------------------------------------------
```



```
Systematic Sampling of River Zones                                        5
Repeated Systematic Samples
The SURVEYSELECT Procedure

Selection Method    Systematic Random Sampling
Input Data Set               RIVER
Random Number Seed        25645834
Sampling Rate           0.08333333
Selection Probability     0.083333
Sampling Weight                 12
Number of Replicates             3
Total Sample Size                9
Output Data Set            SAMPLE2

Systematic Sampling of River Zones                                        6
Repeated Systematic Samples
                            Sampling
Obs    Replicate    Zone     Weight     Dieldrin

 1         1          5      4.00000        1
 2         1         17      4.00000        5
 3         1         29      4.00000        1
 4         2          3      4.00000        0
 5         2         15      4.00000        6
 6         2         27      4.00000        1
 7         3         10      4.00000        2
 8         3         22      4.00000        2
 9         3         34      4.00000        1

Systematic Sampling of River Zones                                        7
Repeated Systematic Samples
Summary Statistics of Individual Replicates
The MEANS Procedure
               Analysis Variable : Dieldrin Dieldrin ug/l

      Sample
   Replicate      N
      Number    Obs    N            Mean        Variance       Std Error
-------------------------------------------------------------------------
           1      3    3       2.3333333       5.3333333       1.3333333
           2      3    3       2.3333333      10.3333333       1.8559215
           3      3    3       1.6666667       0.3333333       0.3333333
-------------------------------------------------------------------------

Systematic Sampling of River Zones                                        8
Repeated Systematic Samples
Summary Statistics of Replication Means
The MEANS Procedure

    Analysis Variable : Dieldrin Dieldrin ug/l
N            Mean        Variance       Std Error
-------------------------------------------------
3       2.1111111       0.1481481       0.2222222
-------------------------------------------------
```

```
Systematic Sampling of River Zones                                        9
Repeated Systematic Samples
Estimation Under Repeated Systematic Samples

The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                3
Number of Observations            9
Sum of Weights           36.0000001


                               Statistics

                               Std Error       Lower 95%       Upper 95%
Variable            Mean         of Mean     CL for Mean     CL for Mean
------------------------------------------------------------------------
Dieldrin        2.111111        0.192450        1.283065        2.939157
------------------------------------------------------------------------
```

```
Systematic Sampling of River Zones                                       10
Repeated Systematic Samples
Variance Components for Intracluster Correlation

The Mixed Procedure

             Covariance Parameter Estimates

                          Standard         Z
Cov Parm      Estimate       Error     Value        Pr Z

Replicate      -1.6296      1.0370     -1.57      0.1161
Residual        5.3333      3.0792      1.73      0.0416


              Fit Statistics

Res Log Likelihood                  -16.7
Akaike's Information Criterion      -18.7
Schwarz's Bayesian Criterion        -17.8
-2 Res Log Likelihood                33.3


  PARMS Model Likelihood Ratio Test

    DF    Chi-Square      Pr > ChiSq

     1          2.89          0.0893
```