Demonstrating Systematic Sampling

Julie W. Pepe, University of Central Florida, Orlando, Florida

Abstract

A real data set involving the number of reference requests at a university library, will be used to present systematic sampling as an alternative to daily collection of information. Excessive collection of data is not only very labor intensive but also unnecessary. Data collected during previous semesters can be used as population information. Since true values are known, systematic samples can be generated and results compared to the population parameters.

Introduction

Systematic sampling is conducted by sampling every k^{th} item in a population after the first item is selected at random from the first k items. If the setting is a manufacturing process, it is easy to instruct someone to pull every 5^{th} item off the line for testing. In marketing, every 10^{th} person could be polled about what product they prefer. It is important to remember that the first item must be randomly selected for the statistical theory to hold true. If there is random ordering in the population of the variable values, then systematic sampling is considered to be equivalent to a random sample.

Library staff members required estimates (for funding reasons) of the number of people the reference librarians helped during each semester. For a past semester "true" numbers were available. Data was collected every hour of every day that the library was open. Could data be collected on only some days or weeks during a semester? The available data provides a unique opportunity to demonstrate systematic sampling.

Methods

The data provided from the library was entered into a SAS® program and then different systematic samples were analyzed for the estimated mean number of references per week. Values of k to be used were picked considering cost and practical considerations of the project. For each of the 3 different k values used, every possible sample for that value of k was calculated. The formula for estimating the mean is: $\sum x / n$. Where x = weekly number of reference requests and n = number of weeks data was collected. This formula is the same formula used in calculating means for simple random sampling.

In order to calculate the true variance of a systematic sample, a measure of correlation between adjacent value pairs must be available. In most cases, population information is not available so variance calculations are usually based on simple random sample variance. As stated by Scheaffer, Mendenhall & Ott (1990), "An unbiased estimate of V(Ysy) cannot be obtained using the data from only one systematic sample." A biased estimator is not a critical problem if the population is random with respect to the variable of interest.

For this example, the population information is available, thus comparisons can be made between the simple random sample variance and the systematic sample variance calculations. Simple random sample variance is calculated as follows.

$$V(Y) = \left(\frac{N-n}{N-1}\right)(\sigma^2 / n) \approx \left(\frac{N-n}{N-1}\right)(s^2 / n)$$

Where s^2 is the variance of the sample, n is the number of weeks, N is the population number of weeks. The systematic sample variance formula is:

$V(Ysy) = (\sigma^2 / n)[1+(n-1)\rho]$

where $\rho = \dfrac{(k-1)nMST - SST}{(n-1)SST}$

ρ = intracluster correlation
MST = mean square total
SST = sums of squares total
k= value of k picked
n= sample size

The values necessary are available from PROC ANOVA or PROC GLM output.

Results

PROC MEANS was used to calculate the means and variances of each systematic sample. Table 1 shows the results of k=4, k=3 and k=2 for samples from the 110 weeks of data available. Simple random sample confidence intervals for the mean were calculated in a data step and plotted in Figure 1. This graph gives the client information on what future sample information would look like. Because complete information was available, the plot shows that all the possible samples captured the true mean value (μ=1493). The true mean value is shown as the horizontal line. The vertical lines are formed by the upper and lower limits with the mean marked as a box. Samples 1 to 4 are for k=4, samples 5 to 7 for k=3 and samples 8 to 9 for k=2. The intervals decrease as n increases (k decreases).

PROC GLM is used to produce values for calculating the systematic variance. Table 2 shows the PROC GLM results. Calculation of ρ = (3*28)67161-30127481/30127481(27). The resulting value of ρ is -0.030. The intracluster correlation is close to zero therefore, the interpretation is the population is random. The resulting variance calculation for systematic samples would then be 1798.95 (references squared). Bound on the error is ± 83.13 references per week. Figure 2 shows the confidence intervals for samples based on population information (specifically intracluster correlation). As in Figure 1, samples 1 to 4 are for k=4, 5 to 7 for k=3 and 8 to 9 for k=2.

Summary

Having the population information available, reduced the error, yielding smaller confidence intervals. These confidence intervals would not be available when only sample information is collected. These intervals are presented here for demonstration purposes only. Instead of just assuming population values are random, historical data is used to test the assumption. After calculating the intracluster correlation it was determined weeks had random values for the number of reference requests. Thus, systematic sampling is the perfect tool to use in this situation. It cuts down on the amount of data collection yet is an easy method to utilize in the library setting.

References

SAS Institute Inc. (1990), *SAS Language: Reference, Version 6, First Edition*., Cary NC:SAS Institute Inc.

SAS Institute Inc. (1990), *SAS/STAT Users Guide Vol. 1 and 2, Version 6, Fourth Edition*., Cary NC:SAS Institute Inc.

Scheaffer, Richard L., Mendenhall, William and Ott, Lyman. (1996), *Elementary Survey Sampling*, Fifth Edition. Wadsworth Publishing, Belmont, California.

The author may be contacted at:
University of Central Florida
Department of Statistics
Post Office Box 162370
Orlando, Florida 32816-2370
or pepe@pegasus.cc.ucf.edu

Table 1: Mean and Standard deviations for systematic samples

Population Information N= **110** (number of weeks) Mean=**1493.44** requests

Analysis Variable : VALUE

-------------------------------------- K = 4 sample 1 -------------------------------------

N	Mean	Std Dev	Minimum	Maximum
28	1466.25	543.1177913	440.0000000	2282.00

-------------------------------------- K = 4 sample 2 -------------------------------------

28	1452.75	609.8147032	336.0000000	2305.00

-------------------------------------- K = 4 sample 3 -------------------------------------

27	1564.00	481.1602963	280.0000000	2155.00

-------------------------------------- K = 4 sample 4 -------------------------------------

27	1493.26	476.4300411	500.0000000	2135.00

-------------------------------------- K = 3 sample 1 -------------------------------------

N	Mean	Std Dev	Minimum	Maximum
37	1484.86	538.2227524	336.0000000	2207.00

-------------------------------------- K = 3 sample 2 -------------------------------------

37	1475.84	535.8665108	280.0000000	2305.00

-------------------------------------- K = 3 sample 3 -------------------------------------

36	1520.33	516.0603508	344.0000000	2282.00

-------------------------------------- K = 2 sample 1 -------------------------------------

N	Mean	Std Dev	Minimum	Maximum
55	1514.24	511.2640821	280.0000000	2282.00

-------------------------------------- K = 2 sample 2 -------------------------------------

55	1472.64	543.7315979	336.0000000	2305.00

Table 2: PROC GLM results

General Linear Models Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	201485.36936	67161.78979	0.24	0.8698
Error	106	29925995.68519	282320.71401		
Corrected Total	109	30127481.05455			

R-Square	C.V.	Root MSE	VALUE Mean
0.006688	35.57826	531.33861	1493.4364

Source	DF	Type I SS	Mean Square	F Value	Pr > F
I	3	201485.36936	67161.78979	0.24	0.8698

Source	DF	Type III SS	Mean Square	F Value	Pr > F
I	3	201485.36936	67161.78979	0.24	0.8698

Table 3: Confidence Interval Calculations

	Simple Random Sample Formula				Systematic Formula	
OBS	LOWER	MEAN	UPPER	N	LSYST	USYST
1	1292.56	1466.25	1639.94	28	1383.12	1549.38
2	1257.73	1452.75	1647.77	28	1369.62	1535.88
3	1406.35	1564.00	1721.65	27	1480.87	1647.13
4	1337.15	1493.26	1649.36	27	1410.13	1576.39
5	1343.58	1484.86	1626.15	37	1401.73	1567.99
6	1335.18	1475.84	1616.50	37	1392.71	1558.97
7	1382.06	1520.33	1658.60	36	1437.20	1603.46
8	1418.69	1514.24	1609.78	55	1431.11	1597.37
9	1371.02	1472.64	1574.25	55	1389.51	1555.77

Figure 1: Confidence Intervals for Systematic Samples

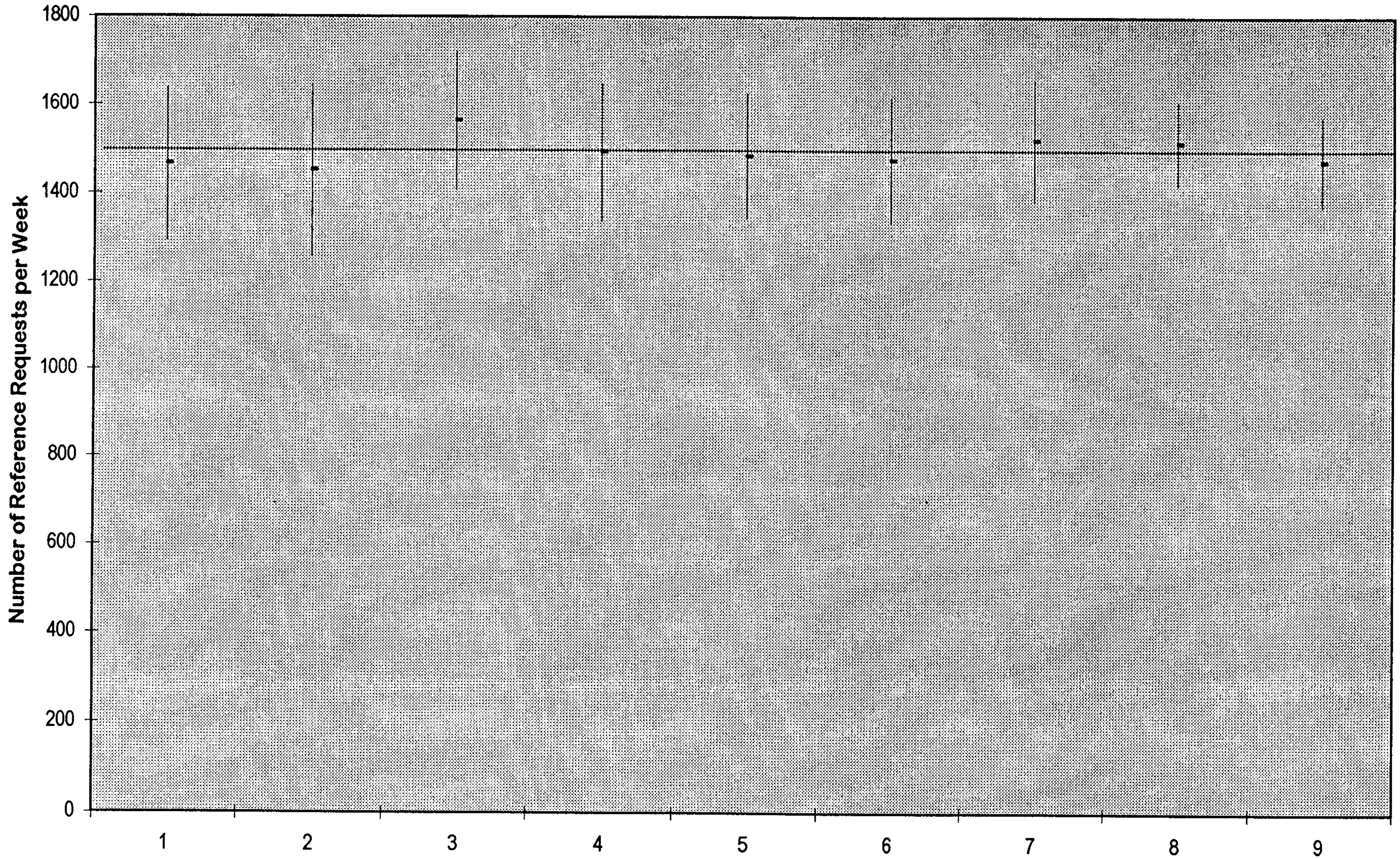

Figure 2: Confidence Intervals Using Population Information