DOCUMENT RESUME

ED 086 218 IR 000 068

AUTHOR	Lied, Terry R.; Tolliver, Don L.
TITLE	A General Statistical Model for Increasing Efficiency and Confidence in Manual Data Collection Systems Through Sampling.
INSTITUTION	Purdue Univ., Lafayette, Ind. Instructional Media Research Unit.
REPORT NO	IMRU-12-73
PUB DATE	Nov 73
NOTE	14p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Confidence Testing; Data Collection; *Library Research; Library Surveys; Models; Research Methodology; *Sampling; *Statistical Analysis; Statistical Surveys; University Libraries; *Use Studies

ABSTRACT

Through utilization of effective sampling procedures, libraries may obtain substantial savings in terms of data collection costs. A theoretical statistical sampling model is presented and two types of random sampling techniques are empirically compared as to their effectiveness in estimating a library usage parameter. Implications are drawn for the possible use of these techniques in library setting. (Author)

A General Statistical Model for Increasing Efficiency and Confidence in Manual Data Collection Systems Through Sampling

By

Terry R. Lied

and

Don L. Tolliver

November 1973

IMRU-12-73

ABSTRACT:

Through utilization of effective sampling procedures, libraries may obtain substantial savings in terms of data collection costs. A theoretical statistical sampling model is presented and two types of random sampling techniques are empirically compared as to their effectiveness in estimating a library usage parameter. Implications are drawn for the possible use of these techniques in a library setting.

Introduction

Without question, current information on the operations of a large university library system is essential for its proper management and administration. Increasingly, managers of libraries are faced with the need for more data to better monitor the library system. Added data becomes necessary to complete internal comparisons, to observe a library sub-system over time, to compare one library with others, and to satisfy external requests for varied and more detailed data. It is likely that this continued pressure for additional data will eventually overload manual collection routines.

This overload may cause administrators to examine the various continuous counting procedures that have become established daily library routines. They begin to search for more efficient data gathering methods to replace traditional procedures. Often seemingly "straight forward" sampling techniques are instituted with an intent to efficiently meet the requirements for data gathering. Yet, these techniques may or may not be effective in providing the required data.

The main objective of this study was to compare two accepted sampling techniques and determine which method would provide the best estimate of a library usage parameter. The first sampling technique examined was a pure random sampling method, and the second was a stratified random sampling technique.

The Theoretical Sampling Model

One of the sampling techniques selected to estimate the library usage parameter was the pure random sampling method (Dixon & Massey, 1969). As applied to this problem, the technique was one in which the particular semester days selected for estimating the parameter would be chosen at random and without replacement. This particular sampling technique was chosen for examination because it is simple to emoloy, it is free of bias (when properly used), and it is a widely used sampling method. The pure random sampling method is based on a theoretical model which requires some elaboration.

Assume that a number of equal-sized samples of semester days is drawn (without replacement) from the population of calendar days in one semester. For each of the days selected in each sample, a number is obtained corresponding to the total number of patrons utilizing the library for that particular day. The distribution of the means of each of these samples is assumed to be normally distributed and has a standard deviation. This standard deviation is known as the standard error of the mean and is represented by the following equation.

$$\text{S.D. of } \bar{x} = \sqrt{\frac{\sigma^2}{N}\left(\frac{N_p - N}{N_p - 1}\right)}$$

Where S.D. of $\bar{x}$ = standard error of the mean.

σ^2 = the population variance; in this case, the variance of the daily number of patrons utilizing the library for one semester.

N_p = the size of the population; in this case the total number of days the library is open during the semester.

N = size of the sample; in this case, the total number of days chosen for sampling the number of patrons utilizing the library.

Random sampling can be effectively used in conjunction with the above mathematical relationship to provide estimates of library usage parameters as well as confidence regions around those estimated parameters. It should be noted that the sampling procedure itself would yield the estimates of the parameters, whereas, the above mathematical relationship could be used to provide the confidence regions around these estimated parameters.

Methods

The library usage parameter (the mean number of patrons utilizing a university library daily during one semester) estimated by these sampling techniques is mathematically expressed as follows:

$$\mu = \frac{\sum X}{N_p}$$

Where μ = the mean number of patrons utilizing the library daily during one semester.

X = the number of patrons utilizing the library for any given day during the semester. This term is then summed over all days of the semester.

N_p = the total number of days during the semester (N_p = 112).

Data were collected on the actual number of patrons utilizing the library each day for one semester. This count was made on a continuous basis by personnel assigned to library exits who had been instructed to record, with a counting device, the number of patrons exiting the library. The <u>mean</u> number of daily patrons utilizing the library during the semester was found to be 1416; the standard deviation was found to be 739.

In summary, then, the parameter or population to be estimated by the two sampling techniques was the mean number of daily patrons utilizing the library during the semester, and as stated above, this value was computed beforehand and was found to be 1416.

The distribution for the theoretical sampling model is presented in Table 1. It contains the expected error specifying the confidence regions for various sample sizes at the 68% and 95% confidence levels. The mean number of daily patrons (1416) utilizing the library during the semester as well as the standard deviation (739), the population size (112), and the sample size (30) were used to derive these estimates, i.e., the appropriate values were substituted into the equation explained above. This provided the expected error for confidence intervals of 68% and 95% for each of the sample sizes listed in Table 1. For example, if the sample size were 35, we would expect that 68 times out of 100 the true value of the estimated parameter would fall within ± 104 units of the estimated value of the parameter; also, 95 times out of 100 the true value of the estimated parameter would fall within ± 208 units of the estimated value of the parameter.

Thus, if one knew the population variance, the population size (e.g., number of semester days) and chose a particular sample size, then confidence regions around the parameter (estimated by the random sampling method) could be obtained. In practice, the only variable that would be left unspecified after one sample of 35 (or any other sample size that might be selected) had been taken and the parameter estimated would be the population variance. However, if the sample size is 30 or more, the variance of the elements of the sample would closely approximate the population variance. On the other hand, if the sample size is substantially less than 30, the population variance could be estimated by using previously collected data if it were available. For example, if one wanted to estimate the previously referred to parameter using a sample size of much less than 30, it might be necessary to obtain the variance of the number of patrons utilizing the library during some previous semester to estimate the population variance. Of course, this is predicated upon the assumption that the variance would not differ substantially from semester to semester. In many instances, this could be a tenuous assumption. At any rate, a decision would have to be made which would involve weighing the accuracy of a small sample size versus the biased estimate that could occur by using the variance of a previous semester.

It was determined that 28 days would constitute a satisfactory sample from the total number of days the library was open during the semester. Once this decision was reached and the procedures described earlier had been arranged, then much of the ground work had been established for instituting a procedure which compared the effectiveness of two sampling techniques.

Figure 1 contains the sampling distribution of sample means (of number of patrons utilizing the library daily for one semester) that were obtained empirically. This was accomplished by drawing without replacement, 30 independent random samples of 28 semester days from the population of 112 semester days. The ordinate of Figure 1 contains the actual frequency with which each of the sample means fell within the specified interval of values indicated along the abscissa. The size of the interval was 50 units. This interval width was selected for it was felt that it would yield the most accurate visual representation of the data. It should be noted that the empirical distribution approximates a normal distribution. Also, the values tend to distribute themselves about the true value of the parameter. The mean of the empirical distribution was

1380 and the standard deviation was 148. The expected mean of this distribution would be 1416 while the expected standard deviation would be 121 (see Table 1). Thus, this empirically derived sampling distribution provided an accurate representation of a sampling distribution that might be obtained by using 30 such random samplings with each sample constituting 28 semester days; the values of the mean and standard deviation of this empirical distribution conformed rather well to the theoretical distribution.

Figure 2 contains the sampling distribution of means that were obtained using the same data, the same sample size, the same number of samples, but a slight modification of the previous sampling method. This modification took into account the effects that various days of the week might have on patron utilization of the library. (It should be obvious that other variables also might significantly affect sampling results.) The mean of this distribution was 1392 and the standard deviation was 94. It can be seen that the sample estimates, in general, more nearly approximated the true value of the parameter in this case than by the method shown in Figure 1. The mean error (obtained by summing the absolute values of each of the sample errors and dividing by the number of samples) in estimating the parameter by this method was 77. The mean error over all samplings in estimating the parameter by the previous method was 114. The difference between these two values was statistically significant ($p < .01$). More importantly, the reduction in the mean error was 32.5%. Therefore, some elaboration of this modified version of the previous sampling technique is in order.

The sampling technique that was employed in Figure 2 is known as stratified random sampling. For the semester used in this example it was observed that the number of patrons utilizing the library for certain weekdays (e.g., Saturdays and Sundays) was at great variance from the number of patrons utilizing the library for other weekdays. Therefore, the sample was selected in such a way that each day of the week was included four times in each of the 28 day samples. Thus, stratification insured that each day of the week was represented an equal number of times in the sample although each day included in the sample was selected randomly for each of the seven strata.

Conclusions

A theoretical model has been presented that makes it possible to estimate confidence regions. This model was based on a random sampling method which did not involve stratification, however, the identical procedure for estimating confidence regions can be used to estimate confidence regions for the stratified random sampling method.

The confidence regions that would result by using this procedure if the stratified random sampling design were employed would be expected to be somewhat wider than the actual confidence regions; i.e., the parameter estimates would be more precise than the width of the estimated confidence region would indicate. This presents no major problem in that interpretations of parameter estimates based on wider confidence regions would consequently tend to be more cautious than interpretations based on narrower confidence regions. The fact remains that such parameter estimates are generally more precise if a stratified random design is properly used instead of the purely random design.

The brief examination and comparison of two sampling techniques demonstrates the increased precision that can accure from careful consideration of the characteristics of the elements that are being measured. For example, the observation that great variability existed in the daily patron usage of the library suggested that the sample should be stratified and that a fixed portion of the sample should be taken from each of the strata. This procedure ensured that the proportion of the sample in each of the seven strata was the same as the proportion of the population in each of the seven strata. If the stratified random sampling design is to be used correctly, it is necessary that the proportion of the sample in each of the strata be the same as the proportion of the population in that strata.

Careful consideration must be given to the idiosyncrasies, habits, and makeup of the elements of the population that is sampled. Such consideration should improve the utilization of sampling procedures, thereby yielding more precise estimates of library parameters.

References

1. Dixon, W. J. and F. J. Massey, Introduction to Statistical Analysis, Third Ed. McGraw-Hill, New York, 1969.

2. Drott, M. Carl, "Random Sampling: a Tool for Library Research," College and Research Libraries, 30 (No. 2): 119-125 (March 1969).

3. Tolliver, Don L., "Study of Purdue Libraries Statistics," IMRU-01-72, Instructional Media Research Unit, Purdue University Libraries and Audio-Visual Center, July 1972.

4. Winer, B. J., Statistical Principles in Experimental Design, Second Ed. McGraw-Hill, New York, 1971.

TABLE 1. Expected Error in Patron Count for Confidence Intervals of 68% and 95%

N	Expected Error for 68% Confidence Interval	Percent Error* for 68% Confidence Interval	Expected Error for 95% Confidence Interval	Percent Error* for 95% Confidence Interval
7	±272	19.20	±543	38.40
14	±186	13.10	±372	26.30
21	±146	10.3	±292	20.6
28	±121	8.5	±243	17.1
35	±104	7.3	±208	14.7
42	± 91	6.4	±181	12.8
49	± 80	5.6	±159	11.3
56	± 70	4.9	±140	9.9
63	± 62	4.4	±124	8.7
70	± 54	3.8	±109	7.6
77	± 47	3.3	± 95	6.6
84	± 40	2.8	± 81	5.6
91	± 34	2.4	± 67	4.8
98	± 27	1.9	± 53	3.8
105	± 18	1.3	± 36	2.5
112	± 0	0	± 0	0

*Note: Percent error is equal to $\frac{\text{Expected Error}}{\text{Mean Daily Patrons}} \times 100$ (In this case, mean daily patrons equals 1416.)

FIGURE 1

Empirical Frequency Distribution of Means

Using Random Sampling Design

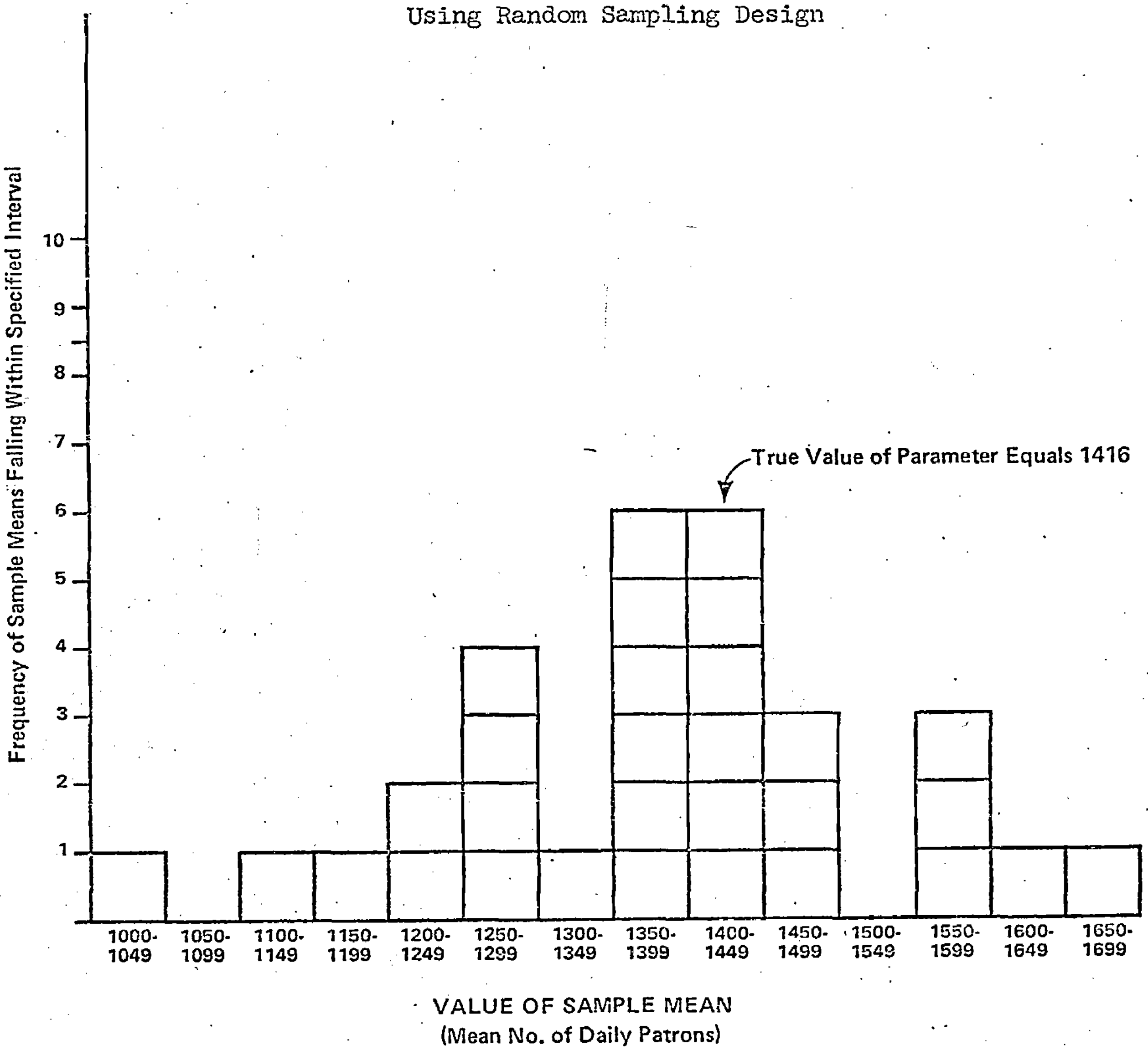

FIGURE 2

Empirical Frequency Distribution of Means

Using Stratified Random Sampling Design