

A Correlation Method for Collecting Reference Statistics

Gwenn Lochstet and Donna H. Lehman

While studying a sampling technique for collecting reference statistics, a correlation method for calculating reference statistics using weekly door counts also was tested at the University of South Carolina. Reference statistics and door counts taken on the sample weeks of the test year were correlated, and the resulting correlation coefficient between the two variables was used to calculate weekly reference statistics for the nonsampled weeks. The sum of these calculated weekly values and the actual values of the sampled weeks yielded a yearly total of reference transactions that is comparable to the yearly total determined by using the sampling technique. Thus, the correlation method may offer libraries an accurate and less time-consuming procedure for keeping reference statistics.



The problems and frustrations of trying to keep reference transaction statistics in any busy library are well known. Library staff members simply are not motivated to keep statistics every day because other demands are made on their time to work with patrons and resources. More often than not, the need for patron assistance is deemed more important than the need for accurate counts of reference questions, and it is becoming more difficult for libraries to maintain accurate reference statistics. However, the need to keep and report reference statistics for accreditation and comparison purposes still remains. Therefore, a simpler means of calculating yearly reference transaction totals is needed.

A new Library Statistics Committee was formed at the University of South Carolina's Thomas Cooper Library in the

summer of 1995. The committee was charged with identifying and choosing a method for simplifying the tracking of reference transactions in the public service areas of the library. Until this time, reference areas of the library recorded each reference transaction every hour of every day they were open. The changes in reference service over the past several years have made tracking reference transactions in this manner increasingly unfeasible. Reference librarians no longer spend the majority of their time working at a desk. They constantly move around helping students at computer workstations, fixing printer jams, and explaining the Internet to computer novices. Working away from the desk means working away from the statistics sheets, making it harder to register each question, yet knowing how many patrons are being reached is important. Therefore, the com-

Gwenn Lochstet is the Math-Physics-Astronomy Librarian at the University of Pennsylvania; e-mail: gsl@pobox.upenn.edu. Donna H. Lehman is a Reference Librarian at the University of South Carolina; e-mail: lehmand@tcl.sc.edu.

mittee was looking for a new, less time-consuming method for finding this information.

After reviewing the literature and discussing the findings, the committee found one method that seemed to offer a solution by proposing that statistics be recorded only on certain sample weeks of the year. The number of weeks sampled would be based on the previous year's records. Reference statistics taken during these sample weeks would be used to

The mean and standard deviations were calculated for each group, and after setting a 95 percent confidence limit and an error rate of plus or minus 400, a sample size was set.

calculate a yearly transaction total. This method still required the library to keep statistics fourteen weeks out of every year. The committee continued to search for an even better solution and decided also to use these sample statistics to test a correlation between reference statistics and door counts, which already were being kept in the library. A correlation coefficient and door counts would be used to estimate the number of reference transactions for the weeks when reference questions would not be recorded. It was found that this method worked well, was as accurate as the other method, and could save much time and frustration for public services staff.

Literature Review

Many articles have been written over the years on keeping reference statistics: what, when, and how to measure. Howard D. White discussed these issues in his 1981 article "Measurement at the Reference Desk," published in the *Drexel Library Quarterly*. He described using the Statistical Package for the Social Sciences (SPSS) program to compile reference statistics as well as ways to measure how well certain goals are being achieved in service to the public.¹ Another article, written by Samuel Rothstein and entitled

"The Hidden Agenda in the Measurement and Evaluation of Reference Service, or, How to Make a Case for Yourself," provides examples of ways to use the collection and evaluation of reference statistics as a way to explain reference loads to managers and library administrators. This article also contains examples of surveys and methods for counting reference sources used, as opposed to reference questions asked, for determining workload.² However, few such articles have actually demonstrated practical options for maintaining reference transaction records. Even fewer provide libraries with satisfactory ways to determine accurate counts of reference transactions with proper sampling methods. The statistics committee found three articles that seemed to give some possible options for sampling statistics on a periodic basis.

One method was developed by Michael Halperin and described in a 1974 issue of *RQ*. It required records to be kept on individual days selected randomly after determining the population size, the degree of precision desired, and the standard deviation.³ However, a problem with this method is the keeping of records only on certain days. It is difficult to remember and remind all staff members to keep statistics on several individual and varied days of each month.

A similar problem exists with the method proposed by John M. Maxstadt in his article "A New Approach to Reference Statistics." Writing from Louisiana State University Libraries, Maxstadt describes their method for keeping statistics certain hours of the day on varying days during the year.⁴

Another method, described by Martin Kesselman and Sarah Barbara Watstein in the *Journal of Academic Librarianship* in 1987 and used at New York University's Bobst Library, describes a similar method but calls for the keeping of statistics one week at a time, rather than on single days or hours of the day. The committee felt that this method offered an effective means for keeping reference statistics.⁵

TABLE 1
Reference Statistics July 1994 to June 1995

Week	Total	% of Year	Week	Total	% of Year
1 7/3-7/9	670	1.08%	28 1/8-1/14	393	0.63%
2 7/10-7/16	784	1.27%	29 1/15-1/21	935	1.51%
3 7/17-7/23	930	1.50%	30 1/22-1/28	1,393	2.25%
4 7/24-7/30	827	1.33%	31 1/29-2/4	1,984	3.20%
5 7/31-8/6	735	1.19%	32 2/5-2/11	1,311	2.12%
6 8/7-8/13	534	0.86%	33 2/12-2/18	1,628	2.63%
7 8/14-8/20	279	0.45%	34 2/19-2/25	1,879	3.03%
8 8/21-8/27	588	0.95%	35 2/26-3/4	1,636	2.64%
9 8/28-9/3	1,444	2.33%	36 3/5-3/11	646	1.04%
10 9/4-9/10	1,725	2.78%	37 3/12-3/18	1,390	2.24%
11 9/11-9/17	2,157	3.48%	38 3/19-3/25	1,567	2.53%
12 9/18-9/24	2,073	3.35%	39 3/26-4/1	1,552	2.50%
13 9/25-10/1	2,555	4.12%	40 4/2-4/8	1,650	2.66%
14 10/2-10/8	1,712	2.76%	41 4/9-4/15	1,465	2.36%
15 10/9-10/15	2,770	4.47%	42 4/16-4/22	1,401	2.26%
16 10/16-10/22	1,104	1.78%	43 4/23-4/29	1,625	2.62%
17 10/23-10/29	883	1.42%	44 4/30-5/6	949	1.53%
18 10/30-11/5	1,879	3.03%	45 5/7-5/13	469	0.76%
19 11/6-11/12	2,001	3.23%	46 5/14-5/20	326	0.53%
20 11/13-11/19	2,200	3.55%	47 5/21-5/27	295	0.48%
21 11/20-11/26	1,163	1.88%	48 5/28-6/3	242	0.39%
22 11/27-12/3	1,961	3.16%	49 6/4-6/10	701	1.13%
23 12/4-12/10	1,789	2.89%	50 6/11-6/17	899	1.45%
24 12/11-12/17	755	1.22%	51 6/18-6/24	785	1.27%
25 12/18-12/24	224	0.36%	52 6/25-7/1	653	1.05%
26 12/25-12/31	134	0.22%	Total	61,968	
27 1/1-1/7	318	0.51%	Mean	1,191.692	
			Standard Deviation	658.528	

The Bobst Library Method

The method described by Kesselman and Watstein required the recording of reference statistics during specific weeks of the year. The librarians of the Bobst Library's Statistics Task Force compiled records of weekly reference transactions from the previous year. These totals were recorded, and then the weeks were divided into high, medium, and low groups depending on the total number of questions asked each week. For the Bobst Library, high-use weeks had more than 5,000 transactions, medium-use weeks had between 3,000 and 5,000 transactions, and low-use weeks had fewer than 3,000 transactions. The mean and standard deviations were calculated for each group, and after setting a 95 percent confidence

limit and an error rate of plus or minus 400, a sample size was set. The Bobst Library Statistics Task Force determined that it needed to sample five low-use weeks, seven medium-use weeks, and three high-use weeks. Then the weeks in the academic year to be tested were numbered consecutively and assigned a usage status based on the status of the corresponding week in the previous year. Specific weeks to be sampled were chosen using a table of random numbers.

After recording reference statistics on all the sample weeks, means were calculated for each usage group that were then multiplied by the total number of weeks in each group for the year. For example, the mean for the high-use group was multiplied by twenty, the total number of

weeks in that group. The products of the means and total number of weeks from each usage group were added to obtain the number of total yearly transactions.

Sampling Methodology

Applying the Bobst Library method, the Library Statistics Committee determined a set of weeks for sampling reference statistics at the Thomas Cooper Library. The weekly nontelephone reference statistics of the main reference department of Thomas Cooper Library for fiscal year 1994–1995 were used as the prestudy data to determine high-, medium-, and low-use weeks. These statistics were used because they were the most complete set of weekly statistics available for a whole fiscal year. Daily reference statistics for the various departments of the library are normally added together and kept as monthly totals. Only main reference retained its daily statistics for fiscal year 1994–1995. Thus, its daily statistics were compiled into weekly totals, except for the month of October 1994. The daily statistics for this month were missing, so daily and weekly totals were extrapolated from random statistical methods using the month total for the department. Only the nontelephone questions for 1994–1995 were counted for the prestudy data because it was thought initially that telephone questions would not be recorded during the study. This assumption proved to be false, and they were later recorded. Despite these difficulties with the prestudy data, they still represented the majority of the reference questions asked in fiscal year 1994–1995. As such, they can offer an indication of the trends of reference activity during that fiscal year.

Table 1 shows the weeks in fiscal year 1994–1995, the weekly nontelephone main reference statistics, each week's percentage of the year's total of nontelephone main reference statistics, the mean of weekly totals, and the standard deviation. Weekly totals were divided into high-, medium-, and low-use periods. It was decided that weeks with more than 1,700 questions would be con-

sidered high weeks, weeks with 800 to 1,700 questions medium weeks, and weeks with fewer than 800 questions low weeks.

The mean and standard deviations for each usage period were calculated. These values were used to determine the necessary sample size for each usage period of the study period. A confidence limit of 95 percent with an error of plus or minus 250 was chosen to ensure that 95 percent of the time the means of randomly selected sample weeks for each usage period would fall within plus or minus 250 of the means from the 1994–1995 prestudy data. The following formula was used to determine the sample size for each usage period:

$$n = \left(\frac{zS}{e} \right)^2$$

where n = sample size, z = accuracy based on the confidence limit in standard deviation units (1.96 for 95% confidence limit), s = standard deviation, and e = error.

After the sample size for each usage period was ascertained, the actual weeks for the study sample were selected. Following the Bobst Library method, each week in 1994–1995 was numbered consecutively, and the corresponding week in the study period was given the same number and usage period designation. A random table was then used to pick the necessary number of weeks for each usage period.⁶ When the initial investigations had been completed in the fall of 1995, calendar year 1996 was chosen as the study period because the Reference Statistics Task Force did not want to wait another semester before implementing the sampling method. The weeks still corresponded, but the first week of the study period, December 31, 1995 to January 6, 1996, was numbered week 27 to parallel the week of January 1, 1995 to January 7, 1995 in the 1994–1995 data. If most fiscal years behave similarly, the use of a calendar year should not skew the data significantly.

At the conclusion of the study period, after reference statistics had been col-

lected by the various departments of the library for each of the sample weeks of 1996, a yearly total was calculated using the following formula:

$$Y = m_l w_l + m_m w_m + m_h w_h$$

where Y = year total, m_l = the mean of the low-sample weeks, w_l = the number of low weeks in 1996, m_m = the mean of the medium-sample weeks, w_m = the number of medium weeks in 1996, m_h = the mean of the high-sample weeks, and w_h = the number of high weeks in 1996.

Correlation Methodology

The random sample of weeks derived from the preceding sampling methodology was also used to test a new way of "collecting" reference statistics. Door counts are taken daily at Thomas Cooper Library. If a significant positive correlation between reference statistics and door counts existed, the weekly door count values could be used with a correlation coefficient to calculate statistically accurate weekly values for reference statistics.⁷ After the reference sampling was completed for 1996, Quattro Pro 6 was used to calculate the Pearson Correlation Coefficient between the door count values and the reference statistics values for the sample weeks. Then, a one-tail t test was employed at the 0.01 significance level to test whether the correlation was significant or due to chance. A t distribution table was consulted to determine a critical value for significance,⁸ and the following formula was used to calculate the t value for the test sample:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

where r = the Pearson Correlation Coefficient, and n = the number of weeks sampled. If the calculated t is larger than the critical value, then the correlation is significant.

If a significant positive correlation could be established, the following linear regression formula would be used to calculate a weekly reference statistics value

based on a weekly door count value:

$$Y = r_{xy} \left(\frac{S_y}{S_x} \right) (X - X_m) + Y_m$$

where Y = the weekly reference statistics, r_{xy} = the Pearson Correlation coefficient between weekly door counts and weekly reference statistics, s_y = the standard deviation of weekly reference statistics from the sample weeks, s_x = the standard deviation of weekly door count values from the sample weeks, X = the weekly door counts, X_m = the mean of weekly door counts from the sample weeks, and Y_m = the mean of weekly reference statistics from the sample weeks.

To measure the reliability of the predicted weekly reference statistics value, the standard error would be calculated. The standard error is the standard deviation of all the weekly reference statistics values that would correspond to any one door count value based on the data from the sample weeks. The following formula is used to calculate the standard error:

$$E = S_y \sqrt{1 - r_{xy}^2}$$

where E = the standard error, s_y = the standard deviation of weekly reference statistics from the sample weeks, and r_{xy} = the Pearson Correlation coefficient between weekly door counts and weekly reference statistics.

To determine a yearly total of reference statistics, a weekly reference statistics value would have to be calculated for each nonsampled week of 1996. Then these calculated weekly values and the actual weekly values for the sampled weeks would be added together for a yearly sum.

Results

Table 2 summarizes the results of the sampling method. The sample sizes were three for low weeks, five for medium weeks, and six for high weeks. The mean reference statistics for low weeks was 1,156.333. The mean for medium weeks was 4,284.600, and the mean for high weeks was 4,240.167. The calculation for

TABLE 2
Reference Statistics Calculations for 1996 based on Sample Data

Low Weeks	Questions	Medium Weeks	Questions	High Weeks	Questions
5/5-5/11	1,306	1/21-1/27	4,139	9/1-9/7	3,427
6/16-6/22	2,002	3/17-3/23	4,221	9/15-9/21	4,621
12/22-12/28	161	3/31-4/6	4,406	10/27-11/2	4,464
		4/7-4/13	4,428	11/10-11/16	4,988
		10/20-10/26	4,229	11/24-11/30	3,083
				12/1-12/7	4,858
Total	3,469		21,423		25,441
Mean	1,156.333		4,284.6		4,240.167
Yearly Total:	162,784				

total yearly reference statistics based on the sampling method was 162,784.

Table 3 summarizes the results of the correlation method. The Pearson Correlation Coefficient between door counts and reference statistics for the sample weeks was 0.957. The highest possible coefficient is 1, which indicates that there is a linear relationship between the vari-

ables. A linear correlation would allow exact predictions of the value of one variable based on the known value of the other. The critical value for the *t* test is 2.681. The calculated *t* value from the test was 11.428. The standard error was plus or minus 423.023.

Table 4 shows the recorded door count for every week of 1996 and the predicted weekly reference statistics value based on the door counts. The sum of the calculated weekly reference statistics values and the actual weekly reference statistics values produces a yearly reference statistics total of 162,878.

TABLE 3
Sample Reference and Door Count Correlation for 1996*
(0.01 significance level, one tail t-test, 12 degrees of freedom)

Week	Questions—Y	Door Count—X
1/21-1/27	4,139	13,027
3/17-3/23	4,221	15,665
3/31-4/6	4,406	15,602
4/7-4/13	4,428	14,162
5/5-5/11	1,306	4,896
6/16-6/22	2,002	7,239
9/1-9/7	3,427	12,339
9/15-9/21	4,621	19,643
10/20-10/26	4,229	19,476
10/27-11/2	4,464	19,075
11/10-11/16	4,988	19,043
11/24-11/30	3,083	11,275
12/1-12/7	4,858	20,518
12/22-12/28	161	577
Total	50,333	192,537
Mean	3,595.214	13,752.643
S.D.	1,458.258	6,075.881

* Pearson *r* = 0.957; Critical Value = 2.681; *t* Test = 11.428; and Standard Error = 423.023

Discussion

Although the estimated 1996 reference statistics total of 162,784, using the sampling method, had over 10,000 questions more than the total of the 1994-1995 year (150,450), it is a feasible number considering the many changes and additions of technology the library experienced in 1996. It is difficult to comment on the accuracy of this value because a number of factors that were not individually accounted for during the sampling could have influenced it (e.g., during some sample weeks, individual library staff members forgot to record reference transactions at the proper time and values had to be estimated later). However, the fact that the yearly total from the correlation method is so similar to the yearly total from the sampling method would seem to indi-

TABLE 4
Weekly Reference Statistics Based on Sample Correlation for 1996

Week	Door Count	Reference Questions	Week	Door Count	Reference Questions
1/1-1/6	3,589	1,261	6/30-7/6	5,817	1,772
1/7-1/13	8,861	2,472	7/7-7/13	6,186	1,857
1/14-1/20	12,750	3,365	7/14-7/20	7,186	2,087
1/21-1/27	13,027	4,139	7/21-7/27	6,886	2,018
1/28-2/3	14,459	3,757	7/28-8/3	6,976	2,039
2/4-2/10	14,537	3,775	8/4-8/10	6,478	1,924
2/11-2/17	14,946	3,869	8/11-8/17	3,543	1,250
2/18-2/24	14,609	3,792	8/18-8/24	7,277	2,108
2/25-3/2	13,460	3,528	8/25-8/31	12,153	3,228
3/3-3/9	5,022	1,590	9/1-9/7	12,339	3,427
3/10-3/16	14,376	3,738	9/8-9/14	19,197	4,846
3/17-3/23	15,665	4,221	9/15-9/21	19,643	4,621
3/24-3/30	14,878	3,854	9/22-9/28	20,840	5,223
3/31-4/6	15,602	4,406	9/29-10/5	19,287	4,866
4/7-4/13	14,162	4,428	10/6-10/12	17,543	4,466
4/14-4/20	17,283	4,406	10/13-10/19	14,014	3,655
4/21-4/27	14,968	3,874	10/20-10/26	19,476	4,229
4/28-5/4	7,488	2,156	10/27-11/2	19,075	4,464
5/5-5/11	4,896	1,306	11/3-11/9	17,831	4,532
5/12-5/18	5,294	1,652	11/10-11/16	19,043	4,988
5/19-5/25	4,594	1,492	11/17-11/23	21,496	5,374
5/26-6/1	3,016	1,129	11/24-11/30	11,275	3,083
6/2-6/8	6,295	1,882	12/1-12/7	20,517	4,858
6/9-6/15	7,278	2,108	12/8-12/14	16,372	4,197
6/16-6/22	7,239	2,002	12/15-12/21	3,432	1,225
6/23-6/29	7,581	2,178	12/22-12/28	577	161
			Sum Total	610,334	162,878

cate that the correlation method is at least as accurate as the sampling method. The difference is that the correlation method does not require any record of reference statistics as long as door counts are recorded and the correlation coefficient between the two variables remains fairly constant.

The high Pearson Correlation Coefficient of 0.957 indicates that there is an extremely strong, almost linear, correlation between weekly door counts and weekly reference statistics values. Furthermore, because the calculated t value of the one-tail t test was so much larger than the critical value, one can be quite confident that there is a strong positive correlation between the two variables that is not due to chance.

The only noticeable problem with the correlation method is the large standard error. Because most of the predicted weekly reference statistics values were in the 10^4 order of magnitude, the standard error of 423.023 was comparatively large. Actually, for the week of December 22, 1997 to December 28, 1997, the predicted value of 161 for weekly reference statistics is actually less than the standard error. Perhaps the sizable standard error was due to the small sample size of reference statistic weeks. A larger sample size of more actual data for weekly door counts and weekly reference statistics will probably decrease the standard deviation of weekly reference statistics (s_y). According to the standard error formula, if the correlation between the two variables re-

mains high, a smaller standard deviation for weekly reference statistics will decrease the standard error.

Conclusion

The correlation method appears to be a viable option for collecting reference statistics, but it will need more testing before it can be confirmed as a functional alternative to recording daily reference transactions. Because of the inconsistencies in the prestudy data for this study, it is recommended that daily statistics be taken for one or even two years to provide a substantial set of weekly reference statistics that could be correlated to the corresponding weekly door counts. It may be difficult to persuade a library to do this because one of the main purposes of trying to determine an alternative method is to reduce the time spent on taking reference statistics. However, the correlation results from this set of data could possibly be used for many years to calculate accurate reference statistics with only a very minimal amount of actual recording of weekly statistics. After an accurate correlation coefficient is obtained, it should only be necessary to record actual reference statistics on two or three randomly sampled weeks a year in order to verify that the correlation coefficient has not changed significantly. (A benchmark value such as 10% should be determined beforehand.) The initial work of recording daily statistics could, therefore, be a worthy investment to save much time in the future without significantly sacrificing accuracy.

It is important that the prestudy be collected carefully and accurately. Although door counts were used in this study to correlate to reference statistics, it is not the only variable that can be used.

Mathematically speaking, any two distinct variables can be tested for correlation, but realistically, it is known that

According to the standard error formula, if the correlation between the two variables remains high, a smaller standard deviation for weekly reference statistics will decrease the standard error.

some library statistics are more directly related to reference question counts than others. It would be wise to choose a variable that is most likely to have a very strong correlation with reference statistics. The closer the correlation coefficient is to one, the more accurate the calculated weekly reference transaction values will be.

If door counts are used, telephone and other remote questions should be eliminated from the correlation because it is unlikely that the patrons asking these questions would be included in door counts. Also, if reference service points are located in different buildings, door counts for each building should be included in the correlation calculations. Unfortunately, this was not done for the two remote reference service areas of the Thomas Cooper Library because the weekly door counts for these areas were not available.

Clearly, much planning and commitment are involved in the initial stages of the correlation method for collecting reference statistics, but once in place, this method may enable libraries to acquire the necessary information from reference statistics without requiring reference librarians to take time away from their growing responsibilities to record every transaction.

Notes

1. Howard D. White, "Measurement at the Reference Desk," *Drexel Library Quarterly* 17 (winter 1981): 3-35.
2. Samuel Rothstein, "The Hidden Agenda in the Measurement and Evaluation of Reference Service, or, How to Make a Case for Yourself," *Reference Librarian* 11 (fall-winter 1984): 45-52.

3. Michael Halperin, "Reference Question Sampling," *Reference Quarterly* 14 (fall 1974): 20–23.
4. John M. Maxstadt, "A New Approach to Reference Statistics," *College & Research Libraries News* 49 (Feb. 1988): 85–87.
5. Martin Kesselman and Sarah Barbara Watstein, "The Measurement of Reference and Information Services," *Journal of Academic Librarianship* 13 (Mar. 1987): 24–30.
6. *CRC Standard Probability and Statistics Tables and Formulae* (Boca Raton, Fla.: CRC Pr., 1991), 367–71.
7. Correlation and linear regression formulae and techniques were taken from Susan H. Hill, *No-Frills Statistics: A Guide for the First-Year Student* (Totowa, N.J.: Rowman & Allanheld Publishers, 1984), 149–77.
8. *Ibid.*, 119.