DOCUMENT RESUME

ED 129 219 IR 003 995

AUTHOR Regazzi, John J.; Hersberger, Rodney M.

TITLE Library Use and Reference Service: A Regression Analysis.

PUB DATE Jul 76

NOTE 20p.; Paper presented at the American Library Association Annual Conference (95th, Chicago, Illinois, July 18-24, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS *Library Reference Services; Personnel Needs; Statistical Analysis; University Libraries; *Use Studies

IDENTIFIERS ALA 76; Northern Illinois University

ABSTRACT

The hypothesis that there exists a strong linear relation between reference service and library use was strongly supported by data from Northern Illinois University Library. Hourly counts of people using the reference room, total number of direction and location questions asked, and total number of reference questions asked were randomly sampled. Regression analysis tests of the relationships between the total number of questions asked and room use, reference questions and room use, and direction questions and room use all showed significant linear relationships. Library use patterns might, therefore, be useful in planning reference service staffing. (KB)

LIBRARY USE AND REFERENCE SERVICE: A REGRESSION ANALYSIS

by

John J. Regazzi
Information Systems Analyst
The Foundation Center

and

Rodney M. Hersberger
Business & Economics Librarian
Northern Illinois University

The attempt to measure reference service has been a highly debated topic in recent years. This debate has centered more around the issue of whether reference service should or even could be quantified, and less around the actual findings of any quantitative studies. The following study attempts to understand reference services by using strict statistical analyses of the relationship between total library use and these services.

Most of the reported discussion on reference measurement has emphasized the evaluation of past activities and performance. The position generally taken is one of a quantitative defense of the value of a reference service[1] or some other interpretation of past performance. This attitude contrasts markedly with a perceived need reported in RQ under a "Symposium on Measurement of Reference" in which it was found that "the number one need for statistical information centered on information for staffing patterns including peak and idle periods, subject specialization and non-desk time."[2] The literature on reference measurement does not appear to address the crucial question of anticipating or predicting when peak and idle periods occur.

This study seeks to determine if any meaningful variables exist by which reference service patterns can be measured and predicted. Accordingly, it was decided to test the variable which would seem to have the most direct effect on reference service, i.e., library use. The hypothesis was simply: there exists a significant linear regression between reference service and library use. While many reference librarians may intuitively accept this hypothesis, no evidence was found in the literature to suggest anyone has statistically tested such a hypothesis.

METHODOLOGY

Since the summer of 1974, the Northern Illinois University Library has kept records on an hourly basis in the following categories: (a) reference room use measured by counting the number of people using the reference room at a set time each hour; (b) the total number of directional and locational questions asked; and (c) the total number of reference questions asked. The recording technique is a modified version of one employed at Cornell University.[3]

These tabulations were all that was required for this study, but in order to have a more manageable study, these statistics were randomly sampled. This sampling was confined to one day of the week, Monday, in order not to introduce another variable into the study. It may be noted, however, that if a linear regression exists on Mondays, it will likely exist throughout the week. Finally, because the reference room in which the reference desk is located in the Northern Illinois University Library is a major study area, the assumption was made that reference room use was indicative of total library use.

After sampling, the first step was merely to compile the data into a workable format. Calculating averages in each category and each hour based on the random sample accomplished this requirement (see Table 1). The category "Total Number of Questions" was derived by adding "Reference Questions" and "Directional-Locational Questions" and rounding the sum.

RESULTS

The first test undertaken was the relationship between the total number of questions asked and room use. A table calculating the necessary values for the regression equation was compiled (see Table 2).

Table 1

Averages for Each Hour

Hour	Number of Reference Questions	Number of Directional-Locational Questions	Total Number of Questions	Individuals in Room
8- 9	.94	2.66	3.68	5.84
9-10	2.63	4.12	6.75	15.48
10-11	2.63	5.80	8.34	23.34
11-12	3.49	6.74	10.02	21.08
12- 1	5.09	7.60	12.62	21.80
1- 2	5.20	8.06	13.14	27.20
2- 3	5.77	7.49	13.14	31.40
3- 4	5.23	7.09	12.42	29.74
4- 5	3.29	6.49	9.31	16.71
5- 6	3.45	4.32	7.85	13.73
6- 7	3.66	5.56	9.05	18.22
7- 8	5.20	6.83	12.00	29.65
8- 9	4.51	5.56	9.68	31.25
9-10	2.17	4.61	6.78	27.64

Table 2

Calculations for Regression Equation

Total Questions x	Use y	xy	x^2	y^2
3.68	5.84	21.49	13.54	31.11
6.75	15.48	104.49	45.56	239.63
8.34	23.34	194.66	69.56	544.76
10.02	21.08	211.22	100.40	444.37
12.62	21.80	275.12	159.26	475.24
13.14	27.20	375.40	172.66	739.84
13.14	31.40	412.60	172.66	985.96
12.42	29.74	369.37	154.26	884.47
9.31	16.70	155.48	86.68	278.89
7.85	13.73	107.78	61.62	188.51
9.05	18.22	164.89	81.90	331.97
12.00	29.65	355.80	1144.00	879.12
9.68	31.25	302.50	97.70	976.56
6.78	22.64	153.50	45.97	512.57
$\Sigma x = 134.78$	$\Sigma y = 308.07$	$\Sigma(xy) = 3204.30$	$\Sigma(x^2) = 1401.77$	$\Sigma(y^2) = 7513.00$
$\bar{x} = 9.63$	$\bar{y} = 22.01$			

The regression equation was then calculated using the following formula:

$$\hat{y} = b_0 + b_1x$$

where

$$b_0 = \bar{y} - b_1\bar{x}$$

and

$$b_1 = \frac{n\Sigma(xy) - \Sigma(x) \cdot \Sigma(y)}{n\,\Sigma(x^2) - (\Sigma x)^2}$$

Then substituting:

$$b_1 = \frac{14 \cdot 3204.3 - 134.78 \cdot 308.07}{14 \cdot 1401.77 - (134.78)^2}$$

and

$$b = 22.01 - 2.29 \cdot 9.63 = -.04$$

thus

$$\hat{y} = (-.04) + 2.29x$$

Having calculated the regression equation as $\hat{y} = 2.29x - .04$, the regression must now be tested to determine if it is significant, and thus a t test was performed. Two hypotheses (H_0 and H_A) were formulated relative to the true slope, B_1: (1) if there is no meaningful linear relationship, then $B_1 = 0$ (null hypothesis, H_0); (2) if there is a linear relationship, then

$B_1 \neq 0$ (alternative hypothesis, H_A). Therefore, in the formula used to test the null hypothesis with a .95 confidence interval and n - 2 degrees of freedom, the null hypothesis can be rejected if $|t| > 2.1788$ (t_{12}, .025), for the following:

$$t_{n-2} = \frac{\frac{b_1 - B_1}{s_{x \cdot y}}}{\sqrt{\Sigma(x^2) - n\bar{x}^2}}$$

where $s_{x \cdot y}$ = the standard deviation for the regression, and where

$$\sqrt{s^2_{x \cdot y}} = \sqrt{\frac{(\Sigma(y^2) - n\bar{y}^2) - b_1(\Sigma(xy) - n\bar{x}\bar{y})}{n - 2}}$$

Substituting

$$= \sqrt{\frac{730.84 - 2.29(236.91)}{12}} = 3.96$$

Therefore

$$t = \frac{\frac{2.29}{-3.96}}{\sqrt{1401.77 - 1298.31}} = 5.87$$

Since t = 5.87 is greater than 2.1788, the null hypothesis can be rejected and a significant linear relationship does exist.

The previous tests demonstrate that a linear regression is statistically meaningful; however, to provide an indication of the strength of the linear regression, a Pearson Coefficient of Correlation (r) was calculated as follows:

$$r = \frac{n\Sigma(xy) - \Sigma x \cdot \Sigma y}{\sqrt{[n\Sigma(x^2) - (\Sigma x)^2] \cdot [n\Sigma(y^2) - (\Sigma y)^2]}}$$

Substituting

$$= \frac{14 \cdot 3204.3 - (134.78 \cdot 308.07)}{\sqrt{[14 \cdot 1401.77 - (134.78)^2] \cdot [14 \cdot 7513 - (308.07)^2]}}$$

$$= .86$$

This calculation illustrates that a fairly strong linear regression does exist; moreover, the coefficient of determination (r^2) which describes the percent of variation in observed y values explained by the regression on x, is also calculated. The coefficient of determination (r^2) is strong as well, accounting for almost 75 percent of the variation.

$$100 \cdot r^2 = \frac{\text{explained variation}}{\text{total variation}} = 73.96\%$$

In order to complete the analysis, the standard error of the estimate (s_e) was calculated. The s_e gives the size of the interval in which predictions based on the linear regression will fall. The formula is as follows:

$$s_e = \pm\sqrt{\frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{n - 2}}$$

where y' is the estimated value for y using the regression equation.

A table was constructed for this test (see Table 3), and from this table we could substitute for the above equation:

$$s_e = \pm\sqrt{\frac{326.86}{14.2}} = \pm\sqrt{27.24} = \pm 5.22$$

The same series of analyses was made for the relationship between (a) "Reference Questions" and "Room Use," and (b) "Directional-Locational Questions" and "Room Use." In each analysis there was a significant linear relationship. The Reference Question and Room Use regression showed a coefficient of correlation of $r = .83$ and a coefficient of determination of $100 \cdot r^2 = 68.32\%$. The Directional-Locational Questions and Room Use regression had a coefficient of correlation of $r = .74$ and a coefficient of determination of $100 \cdot r^2 = 54.76\%$. For complete calculations see Appendix 1 and Appendix 2. It is worth noting that it appears that room use has a somewhat more dramatic influence on actual reference questions than on directional-locational questions.

Table 3

Values for Standard Error of the Estimate Formula

y	y'	$y - y'$	$(y - y')^2$
5.84	8.39	-2.55	6.50
15.48	15.42	.06	.00
23.34	19.06	4.28	18.32
21.08	22.91	-1.83	3.35
21.80	28.86	-7.06	49.84
27.20	30.05	-2.85	8.12
31.40	30.05	1.35	1.82
29.74	28.40	1.34	1.80
16.70	21.28	-4.58	20.98
13.73	17.94	-4.21	17.72
18.22	20.68	-2.46	6.05
29.65	27.44	2.21	4.88
31.25	22.13	9.12	83.17
22.64	15.49	7.15	51.12

$$\sum_{i=1}^{n} (y_i - y'_i)^2 = 326.86$$

DISCUSSION OF RESULTS

The study clearly illustrates that a strong linear regression exists between library use and reference service. Thus, one can safely assume that as library use decreases, demand for reference service will also decrease. As noted above, many reference librarians may intuitively recognize this relationship, however, the strength of this relationship heretofore has been without statistical analysis.

Once the strength of the regression has been established the possible applications are numerous. The implications for staffing and scheduling are most predominant. For example, reference service at Northern Illinois University ceases at 10 p.m. every night, Sunday through Thursday. As in most libraries, the feeling is that little if any reference service is required after that time. However, it was determined that the average Room Use count for 10 to 11 p.m. was some 20.2 persons. By substituting once again into the regression equation, $y = b_0 + b_1\hat{x}$, one can predict the number of reference questions which would be asked within the range of the standard error of the estimate, i.e., $x \pm |s'_e|$. Thus

$$20.2 = (-.04) + 2.29\,\hat{x}$$

$$20.2 = 2.29\,\hat{x}$$

$$\hat{x} = 8.84$$

Since the standard error of the estimate $|s_e|$ for this particular regression is 5.22, one can then predict that between 3.62 and 14.06 questions would be asked during the period 10 p.m. to 11 p.m. If these figures are compared with those in Table 1, it will be noted that (1) even using the lower limit of the above range, at least as many questions will be asked as for the hour 8 a.m. to 9 a.m. for which the library does staff the reference desk; (2) the point estimate $(\hat{x})$ of 8.84 exceeds the demand for an additional four hours of staffed desk service; and (3) the upper limit of 14.06 exceeds all of the remaining hours.

CONCLUSION

In order to develop a consistent staffing plan in the above instance, for example, some consideration should be given to either staffing between 10 p.m. and 11 p.m., or not staffing between 8 a.m. and 9 a.m., for at the very least the demand will be equal. This same application could easily apply to the other days of the week, thus giving librarians administrative information about reference demand without staffing or even observing the reference desk.

The regression will undoubtedly vary among libraries; however, the application of such a procedure may provide librarians with an effective tool for determining consistent and cost-effective staffing patterns for library reference desks.

NOTES

[1] Manual D. Lopez, "Academic Reference Service: Measurement, Cost, and Value," RQ 12, No. 3: 234-242 (Spring 1973).

[2] Symposium on Measurement of Reference," RQ 14, No. 1: 8 (Fall 1974).

[3] Caroline T. Spicer, "Measuring Reference Service: A Look at the Cornell University Libraries Reference Question Recording System," Bookmark 31, No. 3: 79-81 (January-February 1972).

APPENDIX 1

Regression Analysis for Reference Questions and Room Use

Reference Questions x	Room Use y	xy	x^2	y^2
.94	5.84	5.49	.88	31.11
2.63	15.48	40.71	6.92	239.63
2.63	23.34	61.38	6.92	544.76
3.49	21.08	73.57	12.18	444.37
5.09	21.80	110.96	25.91	475.24
5.20	27.20	141.44	27.04	739.84
5.77	31.40	181.18	33.29	985.96
5.23	29.74	155.54	27.35	884.47
3.29	16.70	54.94	10.82	278.89
3.45	13.73	47.37	11.90	188.51
3.66	18.22	66.69	13.40	331.87
5.20	29.65	154.18	27.04	879.12
4.51	31.25	140.94	20.34	976.56
2.57	22.64	58.18	6.60	512.57
$\Sigma x = 53.66$	$\Sigma y = 308.07$	$\Sigma(xy) = 1292.57$	$\Sigma(x^2) = 230.59$	$\Sigma(y^2) = 7513.00$
$\bar{x} = 3.83$	$\bar{y} = 22.01$			

Appendix 1 (Continued)

Regression equation:

$$\hat{y} = b_0 + b_1x$$

$$b_0 = y - b_1\bar{x}$$

$$b_1 = \frac{n\Sigma(xy) - \Sigma(x) \cdot \Sigma(y)}{n\Sigma(x^2) - (\Sigma x)^2}$$

substituting

$$= \frac{14 \cdot 1292.57 - 53 \cdot 66 \cdot 308.07}{14 \cdot 230.59 - (53.66)^2}$$

$$= \frac{1564.94}{348.86} = 4.49$$

and

$$b_0 = 22.01 - 4.49 \cdot 3.83 = 4.81$$

therefore

$$\hat{y} = 4.81 + 4.49x$$

Testing the regression

$H_0 : B_1 = 0$ (not meaningful)

$H_A : B_1 \neq 0$ (meaningful)

(for confidence interval of .95)

Reject if $|t| > 2.1788$ $(t_{12}, .025)$

Appendix 1 (Continued)

$$t_{n-2} = \frac{b_1 - B_1}{\frac{S_{x \cdot y}}{\sqrt{\Sigma(x^2) - n\bar{x}^2}}}$$

$$s_{x \cdot y} = \sqrt{s^2_{x \cdot y}} = \sqrt{\frac{(\Sigma(y^2) - n\bar{y}^2) - b_1(\Sigma(xy) - n\bar{x}\bar{y})}{n - 2}}$$

$$= \sqrt{\frac{(7513.0 - 14 \cdot 22.01^2) - 4.49(1292.57) - 14 \cdot 3.83 \cdot 22.01)}{12}}$$

$$= \sqrt{\frac{730.84 - 504.68}{12}} = \sqrt{18.15} - 4.34$$

$$\therefore \quad t = \frac{\frac{4.49}{4.34}}{\sqrt{25.72}} = 5.2 \qquad \therefore \quad \text{Reject } H_0$$

Testing the strength of the linear regression using the Pearson Coefficient of Correlation (r)

$$r = \frac{n\Sigma(xy) - \Sigma x \cdot \Sigma y}{\sqrt{[n\,\Sigma(x^2) - (\Sigma x)^2] \cdot [n\Sigma(y^2) - (\Sigma y)^2]}}$$

$$= \frac{14 \cdot 1292.57 - 308.07 \cdot 53.66}{\sqrt{[14 \cdot 230.59 - (53.66)^2][14 \cdot 7513 - (308.07)^2]}}$$

$$= \frac{1564.94}{1893.29} = .83$$

Coefficient of determination (r^2)

$100 \cdot r^2$

$100 \cdot (.83)^2 = 68.32\%$

APPENDIX 2

Regression Analysis for Directional-Locational Questions and Room Use

Directional Questions x	Room Use y	xy	x^2	y^2
2.66	5.84	15.53	7.08	31.11
4.12	15.48	63.78	16.97	239.63
5.80	23.34	135.37	33.64	544.76
6.74	21.08	142.08	45.43	444.37
7.60	21.80	165.68	57.76	475.24
8.06	27.20	219.23	64.96	739.84
7.49	31.40	235.19	56.10	985.96
7.09	29.74	210.86	50.27	884.47
6.49	16.70	108.38	42.12	278.89
4.32	13.73	59.31	18.66	188.51
5.56	18.22	101.30	30.91	331.97
6.83	29.65	202.51	46.65	879.12
5.56	31.25	173.75	30.91	976.56
4.61	22.64	104.37	21.25	512.57
$\Sigma x = 82.92$	$\Sigma y = 308.07$	$\Sigma(xy) = 1937.34$	$\Sigma(x^2) = 522.71$	$\Sigma(y^2) = 7513.00$
$\bar{x} = 5.92$	$\bar{y} = 22.01$			

Appendix 2 (Continued)

$$t_{n-2} = \frac{b_1 - B_1}{\dfrac{s_{x \cdot y}}{\sqrt{\Sigma(x^2) - n\bar{x}^2}}}$$

$$s_{x \cdot y} = \sqrt{s^2_{x \cdot y}} = \sqrt{\frac{(\Sigma(x^2) - n\bar{y}^2) - b_1(\Sigma(xy) - n\bar{x}\bar{y})}{n - 2}}$$

$$= \sqrt{\frac{(7513 - 14 \cdot 22.01) - 3.57(1937.34 - 14 \cdot 5.92 \cdot 22.01)}{12}}$$

$$= 23.81$$

$$\therefore \quad t = \frac{3.45}{\dfrac{23.81}{\sqrt{522.71 - 14 \cdot 5.92^2}}} = 4.64 \quad \therefore \text{ Reject } H_0$$

Testing the strength of the linear regression using the Pearson Coefficient of Correlation (r).

$$r = \frac{n\Sigma(xy) - \Sigma x \cdot \Sigma y}{\sqrt{[n \cdot \Sigma(x^2) - (\Sigma x)^2] \cdot [n \cdot \Sigma(y^2) - (\Sigma y)^2]}}$$

$$= \frac{14 \cdot 1937.34 - 82.93 \cdot 308.07}{\sqrt{[14 \cdot 522.71 - (82.93)^2][14 \cdot 7513 - (308.07)^2]}} = .74$$

Coefficient of determination (r^2)

$$100 \cdot r^2$$

$$100 \cdot (.74)^2 = 54.76\%$$