MARIANNE GOLDSTEIN and JOSEPH SEDRANSK

Using a Sample Technique to Describe Characteristics of a Collection

A sampling procedure is presented which may be employed to identify characteristics of a collection and which then can be used in an evaluative statement and in the description of the scope of the collection. The main results obtained by applying this sampling technique to the Jewish history collections in each of seven university libraries are described in detail. Comparisons among these seven collections relate to the percentage distribution of titles by language and by publication date.

LIBRARIANS WHO ARE INVOLVED in collection building are regularly called upon to make statements about the quality of their collections. Subject librarians seek ways to identify and describe subject strengths. The traditional ways of collection evaluation have included both quantitative and qualitative descriptions of the holdings in subject fields.

The quantitative statement is generally based on one of the following methods of measuring library holdings: (1) measuring linear feet of library materials on shelves, (2) a physical volume count, or (3) use of shelflist measurements, i.e., converting cardholdings in inches or centimeters into number of titles.

For a qualitative evaluation, the librarian attempts to support the quantitative statement by (1) the checking of appropriate bibliographies, (2) the consideration of the levels of programs the collection supports, and (3) the size of student body and faculty that uses it.

Sometimes by the use of formulas[1] a quantitative expression of the quality of the collection is arrived at based on the number of books, periodicals, and documents a specific subject field should have. The results of bibliographic checking are expressed in number or percentage of titles held out of the number of titles in the list. The problem of providing a qualitative evaluation is aptly expressed in the statement that "no easily applicable criteria have been developed for measuring quality in library collections, and this is a subject which should be vigorously pursued."[2]

In this paper we present a technique to identify collection characteristics that can be used in an evaluative statement and in the description of the scope of the collection. Characteristics of books, such as (1) their publication dates; (2) their countries of origin; (3) the languages in which they are written; (4) their publishers (whether private, commercial, or academic); (5) their formats (i.e., book, nonbook, serial, document); and (6) the editions (original,

Marianne Goldstein is reference librarian, and Joseph Sedransk is professor, statistical science and social sciences, in the State University of New York at Buffalo.

reprint, facsimile, etc.) tell the subject specialist something about the nature of the collection. For example, students of history and the humanities generally rely on the availability of library materials with more varied imprint dates than students in the social sciences or natural sciences. In the sciences, recency of publications is usually critical; in history, philosophy, and the humanities, research more often depends on the availability of primary sources for the period or topic under investigation.

The characteristics to be identified are available for each title on the shelflist copy of the catalog card. The catalog card gives in addition to the author and title: the edition, imprint (place, publisher, and date), collation, illustration, the subject tracings (headings), and, often, the format. If the book has been translated this is also indicated.

What characteristics a librarian wishes to identify, using the sample technique presented in this paper, is a decision based on the specific characteristics which enhance that subject area and which, when identified in the particular collection, can lend weight to an evaluative statement. Determination of the number of characteristics to be recorded is based on the size of the sample, the size of the collection, and the total staff time available to record and analyze the information.

What follows is a description of the sample design and estimation methods used in selecting and analyzing a sample of titles from the Jewish history collection in each of seven university libraries. The seven libraries are those at Cornell University, University of Rochester, Syracuse University, and the four university centers of the State University of New York (SUNY)—Albany, Binghamton, Buffalo, and Stony Brook. The sample taken was limited to the shelflist for Library of Congress classification numbers DS 101–151. It did not attempt to include all titles in the collections which deal with the history of the Jewish people (e.g., exclusions include Jews in the U.S. under E184.J4; or World War, 1939–1945—Jews under D810.J4; or Bibliography under Z). Statistics are available for titles held in these areas of Jewish history.[3]

The project was part of a survey conducted in the fall of 1974 to evaluate the Judaic Studies resources of the seven university libraries. While the survey also concerned itself with resources in Jewish literature, Bible studies, and Jewish philosophy and religion, the Jewish history collections were chosen for the example. Because of time limitations, the characteristics sought in the sample were limited to age of publication and language. The original date of publication was used when the book was a reprint.

A systematic sample design was employed in each university's Jewish history collection:

a. The total number of cards in the collection was measured in inches *(X)*.
b. Using the relationship, 100 cards = 1 inch, there are estimated to be $N = 100\,X$ titles in the desired category. For example, if X = 14 inches, $N = 1{,}400$.
c. The sampling interval, i, for the selection of sample cards is defined by

$$i = N/n$$

where n is the desired sample size. That is, after a random start, every *i-th* card is sampled. For each sampled title, the date and country of publication, language, format, etc., are recorded. For example, if we wish to have $n = 200$, then $i = 1400/200 = 7$, and we would select every seventh card. If we desire $n = 300$, $i = 1400/300 = 4.67$, and we would select every fourth card to ensure that our sample size is at *least* 300.

The recording procedure is as follows: Once the size of sample is determined, a lined sheet is numbered from 1 to n. A column is drawn for each characteristic to be recorded and given a heading. A sample sheet is shown as Figure 1:

Title	*Call No. (optional)*	*Country*	*Language*	*Date*	*Format*	*Scope, Treatment*
1	DS-	U.S.A.	English	1920	Book	History
2	DS-	Germany	German	1910	Serial	Bibliography

Fig. 1
Sample Sheet for Recording Characteristics

When all the titles in the sample have been recorded, a count is taken of each characteristic (e.g., book with date prior to 1900) of interest. For the given collection the proportion, P, of titles in the entire collection having a specified characteristic is estimated using $\hat{P}$ where $\hat{P}$ equals the proportion of titles in the sample having the specified characteristic. The values of $\hat{P}$ are the primary analytical tool and are presented in Tables 1, 2, and 3. Additional information can be obtained by forming a $100(1-\beta)$ percent confidence interval for P, which is a range of values likely to contain the true, but unknown, value of P. Methods to form a confidence interval for P are presented by Cochran.[4]

The number of cards to be sampled from a given collection is a function of the time available to carry out the sampling (and recording) and the desired precision of estimation of population characteristics. Because the time available to carry out the sampling was unknown initially, the sample size n was arbitrarily set at about 150 for the first two universities visited (Cornell and SUNY at Binghamton). However, after the experience of the trips to Cornell and Binghamton, we were able to determine sample sizes that are feasible in terms of time available to complete the task and which would yield a desired level of precision.

As described above, the proportion, P, of titles with a specified characteristic is estimated using $\hat{P}$. It is desired to select a large enough sample that with high probability the difference between $\hat{P}$ and P will be sufficiently small. More precisely, the investigator specifies two numbers $1-\alpha$ and d. Although the value of P is unknown, the investigator may specify the maximum deviation, d *(e.g.,* $d = .06$*)* between the sample estimate, $\hat{P}$, and P that one would "like" to have. While it cannot be guaranteed in advance of sampling that $\hat{P}$ and P will differ by no more than d units, the investigator may specify the value, $1-\alpha$ *(e.g.,* $1-\alpha = .95$*)*, representing the probability that the maximum deviation will be d units. Then, given values for d and $1-\alpha$, one may find the value of the sample size n required to insure that, with probability $1-\alpha$, $\hat{P}$ and P will differ by no more than d units.

In the Appendix the formulas to determine the required sample sizes are given. In addition, the derivation of the sample sizes used in this investigation is described.

The three tables that follow give the percentages, $\hat{P}$, for the collections for the characteristics outlined above. Comments are provided for each table.

Comments on Table 1

The following percentages represent the largest and smallest sample percentages held in the various languages in the Jewish history collections of the seven university libraries.

	Largest percentage	
In English:	Albany	85%
In German:	Stony Brook	21%
In French:	Buffalo	9%
In Hebrew:	Binghamton	42%

TABLE 1

DS 101-151 JEWISH HISTORY (LANGUAGE DISTRIBUTION)
Percentage of Collection in English and Other Languages

University	N	n	English	German	French	Hebrew	Others	Distribution
Albany	1,489	355	85%	6%	4%	2%	3%	S, R, P, Pol
Binghamton	2,525	151	45	9	3	42	1	A
Buffalo	1,455	269	77	9	9	1	4	S, L, A
Cornell	4,760	158	49	10	4	26	11	.03 R, L; also S, A
Rochester	1,180	264	83	8	5	1	3	R, S, L, I
Stony Brook	1,237	284	70	21	3	1	5	.03 S; also P, R
Syracuse	1,438	214	81	9	1	3	6	A, P, Y, R, L

Abbreviations:
N = Total number of titles
A = Arabic
I = Italian
L = Latin
P = Portuguese
n = number of titles in sample
Pol = Polish
R = Russian
S = Spanish
Y = Serbo-Croatian

Smallest percentage

In English:	Binghamton	45%
In German:	Albany	6%
In French:	Syracuse	1%
In Hebrew:	several	1%

Among the seven universities studied, Binghamton and Cornell have the largest percentages of their titles in Hebrew as would be expected since they were both participants in the Israel PL-480 Program.[5] The percentages of holdings of English-language titles in the Jewish history collections seem larger where there have been no other influencing factors in collection building, i.e., in the Albany, Buffalo, Rochester, Stony Brook, and Syracuse libraries.

Stony Brook reflects to a noticeable extent the impact of faculty and research interests in German Judaica. With the exception of Stony Brook, the percentages of titles in German held in the university libraries are similar enough to suggest the holdings of many German titles in common.[6] In each collection the sample percentage of French titles is no larger than that of German titles. From Table 1, it may be seen that Albany, Buffalo, and Rochester have similar distributions of titles among the various languages, offset only by Buffalo's larger percentage of French and German titles.

COMMENTS ON TABLE 2

Pre-1900

Cornell, Rochester, and Syracuse have significant special collections and, generally, each has acquired more pre-1900 publications than the other universities. In particular, Syracuse has acquired the collection of the nineteenth-century

TABLE 2

DS 101-151 JEWISH HISTORY (CHRONOLOGIC DISTRIBUTION)
Percentage of Collection in Publication Periods Given

University	N	n	Pre-1900	1901-1950	1951-1960	1961-1974
Albany	1,489	355	6%	26%	15%	53%
Binghamton	2,525	151	2	17	9	72
Buffalo	1,455	269	6	24	16	54
Cornell	4,760	158	11	18	11	60
Rochester	1,180	264	9	36	15	40
Stony Brook	1,237	284	7	22	10	61
Syracuse	1,438	214	8	27	17	48

See Table 1 for explanation of abbreviations.

German historian Leopold von Ranke.

1901–1950

Rochester with 36 percent of its collection dated 1901–50 has the largest percentage in this publication period.

1951–1960

The holdings of titles with 1951–60 publication dates range from 9 to 17 percent. These percentages are substantially less than those for the 1961–74 period.

1961–1974

Each library has the largest percentage of its imprints in this period, accounting for 40 percent or more of the titles in each library's Jewish history collection. Binghamton, Stony Brook, and Cornell have at least 60 percent of their titles bearing 1961–74 publication dates, indicating sizeable acquisitions in these years. The reasons for this are: (1) the general publication explosion, (2) relative affluency, (3) similar patterns of acquisitions, e.g., approval plans, (4) impact of the Israel PL-480 Program in the cases of Cornell and Binghamton.

Similarities

From Table 2 it is seen that Cornell and Stony Brook have similar percentage distributions (over the four time periods). The similarity with Cornell may reflect Stony Brook's apparently successful acquisition of a balanced collection for the study of Jewish history. This is surprising, considering the recent development of Stony Brook's collection.

While Albany, Buffalo, and Syracuse may be seen to have similar percentage distributions, they differ from Cornell and Stony Brook in their pattern of acquisition.

Dissimilarities

From Table 2 it is clear that Binghamton and Rochester have quite dissimilar percentage distributions. Binghamton has an unusually high percentage (72 percent) of 1961–74 publications, and Rochester has an unusually low percentage (40 percent) of 1961–74 publications. Further, Rochester has a significantly higher percentage of 1901–60 publications (51 percent total) when compared with the other six university libraries. Rochester's distribution suggests a selective acquisition policy and the acquisition of titles with pre-1961 imprints through gifts or special collections. Binghamton experienced very little growth until 1961–74.

Comments on Table 3

For the years 1961–74 publications in English make up the largest part of each collection (except for Binghamton). In particular, Albany and Stony Brook have the largest percentages corresponding to English titles. German

TABLE 3

DS101-151 Jewish History
Percentage of Collection in Various Languages in Years 1961–1974

University	N	n	English	German	French	Hebrew	Others Dist.	Total Percentage
Albany	1,489	355	46%	3%	2%	1%	1%–S	53%
Binghamton	2,525	151	33	1	1	36	1 –	72
Buffalo	1,455	269	37	7	7	1	2 –A, I, S	54
Cornell	4,760	158	27	3	3	21	6 –A, R, S	60
Rochester	1,180	264	33	3	2	1	1 –R	40
Stony Brook	1,237	284	43	12	2	1	3 –P, R, S	61
Syracuse	1,438	214	37	3	1	1	6 –A, I, L, S, R, Y	48

See Table 1 for explanation of abbreviations.

and French titles are approximately equal in number, except at Stony Brook which shows strength in German Judaica. Buffalo's relatively large percentages of German and French titles reflect an acquisition policy based on recognized research interests. At both Binghamton and Cornell there are large percentages of titles in Hebrew. These reflect the impact of participation in the Israel Public Law-480 Program. Note that Cornell has a more widespread distribution of titles in various languages than does Binghamton which has concentrated primarily on English and Hebrew titles.

To compare the distribution of titles by language for two periods, pre-1961 and post-1961, two new tables may be constructed. For example, a table for Albany for the post-1961 period would show the following:

English	87 percent
German	5
French	4
Hebrew	2
Others	2
TOTAL	100

This information is derived from Table 3, where 87 percent *(= .46/.53)* is the percentage of titles in English in the post-1961 period among all titles in that period.

We have constructed the aforementioned tables but include only the following comparisons of holdings with pre-1961 and post-1961 publication dates: Albany, Rochester, and Stony Brook show very little alteration in distribution. Binghamton's distribution has changed from (pre-1961) one having extensive representation for both English and German titles to (post-1961) one with about equal percentages of English and Hebrew titles. Cornell exhibits a similar shift from English and German to English and Hebrew, but at Cornell there is, in each period, a moderate representation of titles in the "other" languages.

For Buffalo the sample percentage of titles in each of German and French changes substantially from 5 percent of the collection in pre-1961 publications to 12 percent in post-1961. As a corollary of this, the percentage of titles in English is 86 percent in pre-1961 and 70 percent in post-1961. At Syracuse there are some changes in distribution; a smaller representation for English and a larger representation for "other" languages in the post-1961 period.

Conclusions

In the university libraries at Albany, Buffalo, Rochester, Stony Brook, and Syracuse the preponderance of titles is in English with German and French titles ranking second and third. By contrast, both Binghamton and Cornell have substantial percentages of titles in both English and Hebrew. Of particular note at Stony Brook is the high percentage of German titles in relation to its rather small collection. This indicates specialized interest concerning the history of German Jewry in the nineteenth and twentieth centuries.

The percentage distributions of titles by language are quite similar for Albany, Buffalo, and Rochester. However, each of these distributions is substantially different from those at Cornell and Binghamton where there are large percentages of titles in Hebrew.

Books with pre-1900 imprints are found more extensively at Cornell, Rochester, and Syracuse. It is likely that many of these holdings were acquired by gift or by purchase of scholarly collections. In addition, Rochester has a larger relative percentage of titles with 1901–60 imprints than the other six libraries. Thus, Rochester's distribution suggests a more gradual acquisition of selected titles over a considerable time period.

One may note Stony Brook's similarity to Cornell in the percentage distribu-

tion of titles over the time periods shown, this despite the fact that Stony Brook is the youngest of the university libraries. Strong similarities in distribution of titles by publication date appear for Albany, Buffalo, and Syracuse. Distinct dissimilarities in distribution are observed between Rochester and Binghamton, which are not surprising since most of Binghamton's growth has occurred since 1950. Binghamton's pre-1961 holdings are relatively weak.

The heaviest acquisition period for all seven university libraries was 1961–74. Except at Binghamton and Cornell, English titles were acquired primarily, with German- and French-language titles ranking next in the number of acquisitions. The relatively large percentage of German-language titles acquired at Stony Brook in relation to its small collection is unusual. At Binghamton, Hebrew titles predominate with English second, while at Cornell, English and Hebrew rank first and second respectively. The importance of Hebrew titles at Cornell and Binghamton is, of course, the result of participation in the Israel PL-480 Program which operated between 1964 and 1973.

Finally, the study indicates that strengths of collections, special interests, periods of heavy acquisitions and/or publishing, and book selection policies can be identified by sampling a library's collections. The sample technique used in this study would be particularly useful in a comparative evaluation of the holdings in one subject area at a number of similar libraries.

Appendix

Sample Size Determination

It is assumed that it is desired to use $\hat{P}$ to estimate P so that, with probability $(1-\alpha)$, the difference between $\hat{P}$ and P will be less than d units. The formulas[7] for the required sample size n are shown as formulas A and B:

A. $$n_o = \left\{ z^2_{(1-\frac{\alpha}{2})} \right\} P(1-P)/d^2$$

B. $$n = \frac{n_o}{(1+\frac{n_o}{N})}$$

In these formulas, P is the proportion of titles in the given collection with the specified characteristic; d is the margin of error (specified by the investigator), N is the number of titles in the entire collection, and $z_{(1-\frac{\alpha}{2})}$ is a number completely determined by a specification of the value of the probability, $(1-\alpha)$. The value of $z_{(1-\frac{\alpha}{2})}$ can be read from tables of the normal probability distribution. For example, for $\alpha = 0.05$, $z_{(1-\frac{\alpha}{2})} = 1.96$ while for $\alpha = 0.10$, $z_{(1-\frac{\alpha}{2})} = 1.65$.

The sample size n given by formula B will never be larger than n_o. Thus, if the sample size n is chosen as

$$n = n_o = \left\{ z^2_{(1-\frac{\alpha}{2})} \right\} P(1-P)/d^2$$

the selected sample will certainly be large enough to achieve, with probability $1-\alpha$, the specified margin of error, d. When planning a study, this is often a useful procedure since use of *both* formulas A and B to determine n requires knowledge of N, the total number of titles in the collection.

Suppose that it is desired to have $\alpha = 0.05$ and $d = 0.06$. Then,

$$n_o = \frac{(1.96)^2 P(1-P)}{(.06)^2} .$$

Now note the relationship of $P(1-P)$ with P, as shown in Figure 2.

P	.1	.2	.3	.4	.5	.6	.7	.8	.9
P(1 – P)	.09	.16	.21	.24	.25	.24	.21	.16	.09

Fig. 2
Relationship of P(1 – P) with P

Thus, $P(1 - P)$ assumes its largest value when $P = 0.5$. Taking $P(1 - P) = (0.5)(0.5) = 0.25$, the sample size

$$n = n_o = \frac{(1.96)^2 (0.25)}{(0.06)^2} = 267$$

will be sufficient to ensure with probability 0.95 a margin of error not larger than 0.06 irrespective of the proportion, P, being estimated.

The sample size calculated in this manner may, however, be *larger* than necessary because the proportion, P, for the characteristic of interest may differ from 0.5; and because formula B has not been used to determine n. To illustrate the latter point assume $\alpha = 0.05$, $d = 0.06$, $P = 0.5$, and $N = 1400$. Then $n_o = 267$ and, using B, $n = 224$. Thus, if it were known prior to sampling that $N = 1400$ for a specific collection, a sample of size 224 rather than one of size 267 would be selected. Since a sample of size 224 is all that is needed, there is a reduction in sample size of 267 – 224 = 43 titles because of knowing the value of N ($N = 1400$, here).

Calculations such as those made above indicated that, for most collections, and for $d = 0.06$, $\alpha = 0.05$, a sample of about 250 titles would be adequate. The actual sample sizes differ from 250 because (1) there were differential amounts of time available for sampling and (2) there was rounding error. The latter point can easily be demonstrated by considering a collection with $N = 1400$ titles and a desired sample size $n = 250$. Then i (the sampling interval) = $1400/250 = 5.6$. If $i = 5$, the actual sample size will be $1400/5 = 280$ titles, while if $i = 6$, the actual sample size will be $1400/6 = 233$ titles.

References

1. Formulas referred to are the following: Verner W. Clapp and Robert T. Jordan, "Quantitative Criteria for Adequacy of Academic Library Collections," *College & Research Libraries* 26:371–80 (Sept. 1965). Interinstitutional Committee of Business Officers, University of Washington, *A Model Budget Analysis System for Program 05 Libraries* (Washington State Univ., March 1970). "Standards for College Libraries," *College & Research Libraries News* 36:277–79, 290–301 (Oct. 1975).
2. *Report of the Advisory Committee on Planning for the Academic Libraries of New York State* (Albany: The University of the State of New York, State Education Department, 1973), p.vii.
3. Marianne Goldstein. *A Survey of Library Resources in Judaic Studies in the FAUL and SUNY Center Libraries, With Recommendations Toward Formulating Plans for Possible Areas of Cooperative Collection Development* (Buffalo: SUNY at Buffalo, Lockwood Reference Dept., 1976). Available as an ERIC publication (ED 125651).
4. W. G. Cochran, *Sampling Techniques,* 2d ed. (New York: Wiley, 1963), Section 3.6.
5. Israel Public Law-480 Program. "Within the framework of what is commonly referred to as the Public Law-480 Program, the United States Government supplied some 25 American Research Libraries with a copy of virtually every monograph, book and periodical then published in Israel that was, or might eventually be, of research value. From 1964–1973, approximately 1,665,000 items were supplied, with an average of 65,000 for each full participant." See Charles Berlin, "Library Resources for Jewish Studies in the United States," *American Jewish Yearbook* 75:10 (1974/75).
6. In the sampling process, books in German on Jewish history listed in major bibliographies were recognized. The *Survey* mentioned above in reference 3 also included some bibliographic checking in Jewish history bibliographies. Moreover, most libraries had some approval plan arrangements with the German book firm, Harrassowitz.
7. Note that formulas A and B presume the use of simple random sampling. While we have used systematic sampling, the two sampling methods should be essentially the same for the populations being sampled. (See Cochran, *Sampling Techniques,* Section 8.5, p.214.) Further, the "normal approximation" used to derive formulas A and B should be appropriate for most cases since the sample sizes are large.