Sampling Biological Populations

Guy L. Steucek

Guy L. Steucek is a professor of biology at Millersville Univ., Millersville, PA, 17551, where he teaches biometry and studies plant growth and development. Presently he is investigating plant architecture at the School of Engineering, Univ. of Massachusetts, and collaborates with biologists, architects and engineers in West Germany.

Perhaps the best way to learn about science is to practice the craft; this is sound pedagogical practice with most disciplines. Students can and should do credible research (Yurkiewicz & Steucek 1970). There is no dearth of problems that beckon our attention. Secondary school and college students represent an enormous resource which could be utilized to conduct research within the confines of science courses. It is no longer possible to teach all that is known about the natural world; emphasis on the process of scientific investigation should not be slighted to squeeze in factual material. On the contrary, the process of scientific investigation should be emphasized, particularly to nonscience majors. Models in biological science have changed dramatically during the past half century, illustrating the limited utility of presenting only factual material.

Since many community problems have a biological basis and since students are able to study biological problems relatively easily, the initiation of scientific research in biology courses is appropriate. Moreover, such projects may satisfy local school district goals of a pedagogical nature and prepare students for effective citizenship; perhaps a government/civics teacher could be involved with the definition of a problem. Local health, environmental, agricultural, industrial and scientific researchers represent good sources of important projects; moreover, collaboration with these scientists could be rewarding for all (Spector & White 1985). The initiation of any research project demands that attention be given to the procedure for sampling the population of interest. I will address this problem.

Biological science is becoming more quantitative; this is evident in journal articles and in laboratory exercises published in manuals. With this shift from a descriptive to a quantitative presentation, students are often introduced to simple statistical techniques, such as Student's t-test and analysis of variance which are often found in appendixes of laboratory manuals. Interestingly, appropriate sampling procedures are often ignored by these same laboratory manuals. These statistical analyses make assumptions about how data were collected and hence place constraints on the sampling procedure.

The use of an appropriate sampling procedure generally takes little time and provides a healthy assurance that the findings will not be discounted on the basis of artifactual bias. This important point should be made to all students of science.

Reasons for Sampling

While all biologists are interested in universal truth manifested in populations, we rarely, if ever, have the opportunity to know population parameters, the indexes that describe a population (Figure 1). We depend on sample statistics for a variety of reasons. It may be physically impossible to observe every individual in a population; time and funds also limit the proportion of a population that may be studied. In addition, if a destructive assay is used to collect information about an organism, we would not wish to observe the entire population because to do so would result in extinction of the species. Often absolute information is not necessary for decisions to be made; a census can be less accurate than a sample if the effort required for completing the census results in sloppy gathering of data. Samples are not only necessary, they are economical.

Types of Samples

A great variety of sampling and survey techniques have been devised (Cochran 1963; Hansen, Hurwitz & Madow 1953). It is not the purpose of this article to evaluate them, but to describe some commonly accepted techniques which can and should be employed by students working on scientific endeavors.

Judgmental Samples:

Judgmental samples are comprised of specimens selected because they are "typical" of the population. This may be a very efficient way to obtain information about a population, but what is "typical" to one person may not be to another. By definition, judgmental samples are biased. Confidence in the results rests entirely on faith in the sampler's judgment, there is plenty of room for artifactual bias. This sampling procedure results in very little error for very small samples, whereas the error for random samples will be high for small samples and less for large samples. Consult Hendricks (1956) for a discussion of this topic. The use of statistical analyses will be meaningless with data from judgmental samples because these techniques assume error to be distributed randomly. This sampling method should be employed only for probing studies during the formative stages in defining a problem. Other sampling procedures should be employed in studies to test hypotheses.

Systematic Samples:

Systematic sampling is the selection of specimens at regular intervals. For example, every tenth plant in a row of corn could be sampled, or soil samples could be taken at the nodes of a grid superimposed on a map. The advantage of this procedure is that it will ensure better coverage of the population than a simple random sample; however, systematic sampling could be less representative than the simple random sample if there is hidden periodicity in the population. For example, if you were to sample a marsh with drainage ditches and your grid was coincident with the ditches, an obvious bias might distort results. In the statistical analysis of data obtained from systematic samples, we assume that the specimens are randomly arranged with reference to space or our grid. The initial specimen to be selected, which establishes the position of the network or grid, should be selected at random as discussed below.

Simple Random Sample:

Every possible combination of sampling units has an equal and independent chance of being selected in a random sample. As Folks (1984) states: "The act of physically assigning treatments to experimental units at random 1.) tends to improve the experiment regardless of what the analysis may be, and 2.) justifies an analysis based on probability models"; i.e. accepted statistical procedures can be employed. One function for randomization is to eliminate biases due to judgment or to natural periodicity. Many scientists may claim that randomization is not necessary for their study since they consider their specimens to be arranged in random fashion. Perhaps they are

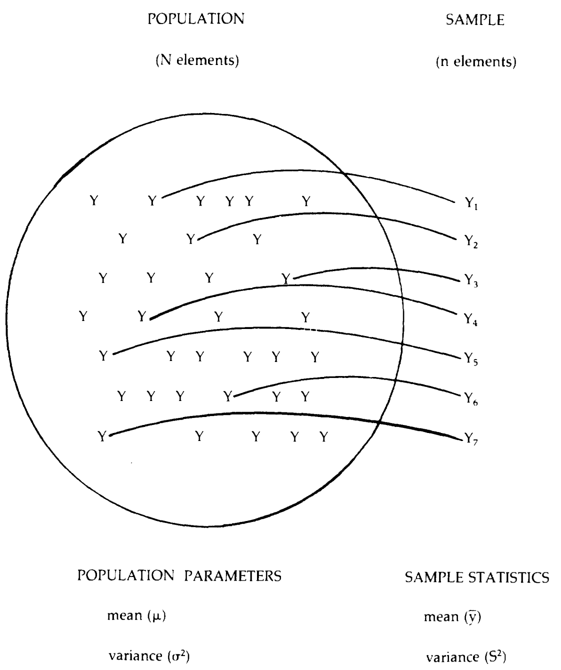

Figure 1. Samples are selected from a population since it is often impossible to observe every element in a population. Sample statistics are estimates of population parameters which are used to describe the nature of the population. The mean ($\bar{y}$) is the average value and variance (s^2) is an index of variability for a sample.

right, but if not, their results could be thoroughly misleading.

While the Peirce-Jastrow experiment reported in 1884 was one of the first to use a mathematically sound scheme of randomization (Stigler 1978), it was not until the 1920s that R.A. Fisher promoted the role of randomization in experimental design. The value of random sampling was then recognized. Today the principle of randomization is commonly accepted in experimentation and is presented matter-of-factly in statistical textbooks. Folks (1984) presents an interesting discussion of randomization.

There is a natural tendency to select a "typical" specimen; it is almost impossible to select a specimen at random while looking at it. Considering a sampling procedure to be random when it is only haphazard is a grave danger. For example, tossing a metal hoop over your shoulder to locate sites for vegetation analysis represents a haphazard sampling procedure for a number of reasons. First, you will be reluctant to toss the hoop where it is difficult to retrieve it. Second, the hoop will bounce off tree stems, hence sites adjacent to trees will be discriminated against. Third, invariably the hoop will roll downhill, resulting in a preponderance of lowland samples. As another example, consider the haphazard sampling of mice from a cage by selecting those you can grab first. The most docile, ill or dead mice will be caught first and hence be assigned to the first sample; the bias is obvious. Students should be asked to think about how haphazard sampling could bias results of an investigation.

The simple random sample uses random numbers to assign treatments to specimens or locate sample sites in the field. These numbers are generated by a random number generator or taken from a table of random numbers. This technique will be described in detail below.

Stratified Random Sample:

On occasion one is able to recognize the presence of sub-populations within the population of interest. For example, a field may be divided into a wet and dry region, or plants could be designated either as flowering or vegetative, or humans could be stratified by the proportion of body fat—lean, normal, fat or obese. Once sub-populations are defined, samples are drawn at random from each sub-population as described above. This procedure permits us to make statements about the sub-populations and increases the precision of estimates about the entire population since some of the sampling error associated with a simple random sample can be described as a sub-population effect. Cochran (1963) discusses ways to allocate samples among sub-populations. Generally one samples a sub-population in proportion to its size relative to the entire population. For example, if one quarter of a field can be designated as wet, then one quarter of the total number of samples should be taken from the wet sub-population. While we will not discuss stratified random sampling in detail, please consider this option when confronted with obvious sub-populations.

	72344	
	14850	
	88727	
	79042	
	31796 ↑	
94719	34656	80018
23388 ←	(22463)	→ 65774
15446	11034 ↓	98143
45412	43556	27518
32818	62953	78831
	73799	
	40829	
	96395	
	67803	
	31093	
	35611	
	69930	
	80651	
	11796	
	55355	

Figure 2. A portion of the table of random numbers. The first number selected is in the box; one may proceed up or down columns or across rows either to the left or to the right as the arrows illustrate.

Selection of Random Numbers from a Table

The generation of random numbers is far more difficult than it may seem. An interesting discussion of the futility in defining randomness is presented by Kac (1983,1984). Fortunately tables of random numbers have been published, and one is reproduced in Table I (Sokal & Rohlf 1969). A table of random digits consists of a series of digits 0, 1, 2, 3, . . ., 9 and each digit occurs with the same relative frequency. The numbers are generated in such a manner that all digits occur with equal probability for any consistent path through the table, but no pattern or order is evident in their sequence. These tables have been tested for randomness and contain no detectable bias. Consult Rand (1966) should you need a very large table of random numbers. How does one select random numbers from this table?

Procedure

Since the selection of random numbers of size 10 or less is rather simple and the selection of random numbers size 11 or greater is a bit more involved, we will make this distinction in the description of the procedure and with the use of two examples. Teachers/researchers may wish to use a one-digit example with a class initially and proceed to the two- or more-digit situation when working on a problem.

a. Define the number of random numbers to be selected. Also, define the maximum size of a random number to be selected.

These numbers would be the same if all specimens are to be used in your investigation. For a one-digit example, suppose we wish to select six random numbers from a total of six numbers. For a two-digit sample we wish to select five random numbers from a total of 14, i.e. nine specimens would not be observed. In this sample the maximum size of a random number is 14.

b. Locate the initial random number.

This is done simply by placing your finger on the table of random numbers while looking elsewhere. For example, I hit the number 22463 which is located in row 27 and column 7 of Table I (see also Figure 2). For the one-digit example we will use the last digit (3). When two digits are required, as in our second example, use the last two digits (63) of this number; however, we could use any of these digits. The initial random digit for the first sample will be 3.

TABLE I. Table of Random Digits organized in 10 columns and 50 rows. This table was taken from Sokal and Rohlf (1969).

	1	2	3	4	5	6	7	8	9	10	
1	48461	14952	72619	73689	52059	37086	60050	86192	67049	64739	1
2	76534	38149	49692	31366	52093	15422	20498	33901	10319	43397	2
3	70437	25861	38504	14752	23757	59660	67844	78815	23758	86814	3
4	59584	03370	42806	11393	71722	93804	09095	07856	55589	46020	4
5	04285	58554	16085	51555	27501	73883	33427	33343	45507	50063	5
6	77340	10412	69189	85171	29082	44785	83638	02583	96483	76553	6
7	59183	62687	91778	80354	23512	97219	65921	02035	59847	91403	7
8	91800	04281	39979	03927	82564	28777	59049	97532	54540	79472	8
9	12066	24817	81099	48940	69554	55925	48379	12866	51232	21580	9
10	69907	91751	53512	23748	65906	91385	84983	27915	48491	91068	10
11	80467	04873	54053	25955	48518	13815	37707	68687	15570	08890	11
12	78057	67835	28302	45048	56761	97725	58438	91528	24645	18544	12
13	05648	39387	78191	88415	60269	94880	58812	42931	71898	61534	13
14	22304	39246	01350	99451	61862	78688	30339	60222	74052	25740	14
15	61346	50269	67005	40442	33100	16742	61640	21046	31909	72641	15
16	66793	37696	27965	30459	91011	51426	31006	77468	61029	57108	16
17	86411	48809	36698	42453	83061	43769	39948	87031	30767	13953	17
18	62098	12825	81744	28882	27369	88183	65846	92545	09065	22655	18
19	68775	06261	54265	16203	23340	84750	16317	88686	86842	00879	19
20	52679	19595	13687	74872	89181	01939	18447	10787	76246	80072	20
21	84096	87152	20719	25215	04349	54434	72344	93008	83282	31670	21
22	63964	55937	21417	49944	38356	98404	14850	17994	17161	98981	22
23	31191	75131	72386	11689	95727	05414	88727	45583	22568	77700	23
24	30545	68523	29850	67833	05622	89975	79042	27142	99257	32349	24
25	52573	91001	52315	26430	54175	30122	31796	98842	37600	26025	25
26	16586	81842	01076	99414	31574	94719	34656	80018	86988	79234	26
27	81841	88481	61191	25013	30272	23388	22463	65774	10029	58376	27
28	43563	66829	72838	08074	57080	15446	11034	98143	74989	26885	28
29	19945	84193	57581	77252	85604	45412	43556	27518	90572	00563	29
30	79374	23796	16919	99691	80276	32818	62953	78831	54395	30705	30
31	48503	26615	43980	09810	38289	66679	73799	48418	12647	40044	31
32	32049	65541	37937	41105	70106	89706	40829	40789	59547	00783	32
33	18547	71562	95493	34112	76895	46766	96395	31718	48302	45893	33
34	03180	96742	61486	43305	34183	99605	67803	13491	09243	29557	34
35	94822	24738	67749	83748	59799	25210	31093	62925	72061	69991	35
36	34330	60599	85828	19152	68499	27977	35611	96240	62747	89529	36
37	43770	81537	59527	95674	76692	86420	69930	10020	72881	12532	37
38	56908	77192	50623	41215	14311	42834	80651	93750	59957	31211	38
39	32787	07189	80539	75927	75475	73965	11796	72140	48944	74156	39
40	52441	78392	11733	57703	29133	71164	55355	31006	25526	55790	40
41	22377	54723	18227	28449	04570	18882	00023	67101	06895	08915	41
42	18376	73460	88841	39602	34049	20589	05701	08249	74213	25220	42
43	53201	28610	87957	21497	64729	64983	71551	99016	87903	63875	43
44	34919	78901	59710	27396	02593	05665	11964	44134	00273	76358	44
45	33617	92159	21971	16901	57383	34262	41744	60891	57624	06962	45
46	70010	40964	98780	72418	52571	18415	64362	90636	38034	04909	46
47	19282	68447	35665	31530	59832	49181	21914	65742	89815	39231	47
48	91429	73328	13266	54898	68795	40948	80808	63887	89939	47938	48
49	97637	78393	33021	05867	86520	45363	43066	00988	64040	09803	49
50	95150	07625	05255	83254	93943	52325	93230	62668	79529	65964	50

In the second example, since 63 is larger than 14, you might wish to ignore this random number and proceed to the next random number. By continuing this practice you would have to ignore 86 percent of the numbers (i.e. those greater than 14) in the table, this would be tedious to say the least and you might require a table larger than that presented in Table 1. To alleviate this problem, simply divide the two-digit number (63) by the total number (14) and use the remainder (7) as the random number; i.e. the first random number is 7. A remainder of zero (0) designates the number 14 as the random number.

Caution: since we wish every number to have an equal and independent chance of being selected, we should eliminate some two-digit numbers from consideration. The two-digit number 00 is taken to mean 100, since we begin counting with the number 1. Since 14 goes into 99 seven times with a remainder of 1 and 14 goes into 100 (the random number 00) seven times with a remainder of 2, the numbers 1 and 2 will be encountered more frequently (7 times versus 6 times) than 3, 4, 5, . . ., and 14. Therefore, the random numbers 99 amd 00 should not be used when proceeding as outlined above. This consideration must be acknowledged when using remainders as a source of random numbers.

c. Proceed to the next random number.

You may move up or down a column of numbers or from left to right or from right to left (Figure 2). We will move down the column of random numbers. The next two-digit number is 34, located in row 28 and column 7 of Table 1. Should you encounter a random digit, or generate one as a remainder, that has been used previously, skip it and proceed to the next number in the table. For the first sample, the second random number is 4. Moving down one more row of random numbers, the third random number is 6. Since the next tabulated number is 3 and we have used it, skip it, and proceed to the next. The fourth random number is 5, after skipping two 9s since they are out of the range. The fifth random number is 1 after skipping two 3s. And the sixth or last random number is 2 by default. Therefore, the sequence of six random numbers for the first sample is 3, 4, 6, 5, 1 and 2.

Considering the second sample, 34 divided by 14 has a remainder of 6 which is the second random number (refer to Figure 2). Proceeding to the next two-digit number, 56 divided by 14 has a remainder of 0, therefore 14 is the third random number. The two-digit number 53 provides the fourth random number: 11. The next two-digit number from the table is 99; this should be skipped as described above. The fifth random number is 1 which was obtained using the two-digit number 29. Therefore, the five random numbers for the second sample are 7, 6, 14, 11 and 1.

When selecting a larger number of random numbers you will undoubtedly come to the edge of the table. If when moving across rows/columns you encounter the edge of the table, simply move up, down, or across to the next row/column. When locating a new set of rows/columns you need only move one digit over, even when using two, three or more digit numbers. The direction and pattern of movement should always be the same, a consistent path should be used.

The use of random numbers to assign treatments to specimens or to locate sites for samples in the field will be discussed in detail later.

Random Number Generators

Numerous random number generators are available for use on calculators and computers; they compute pseudo-random numbers in a completely deterministic way. For example, the RANDOMIZE statement and the RND function can be used with IBM BASIC to obtain random numbers. While most of these programs are relatively sound, some are rather unreliable after many iterations. I have found some programs to degenerate after 15 iterations. Methods of generating random numbers are discussed in Abramowitz & Stegun (1964) and by Jansson (1960). Table or random numbers have been tested for their validity and are convenient to use in the field, hence they may be preferable to a particular random number generator.

Assigning Treatments to Specimens

Assigning treatments to specimens with a biased or haphazard method can be responsible for erroneous results. While you may think the treatment was a causative agent, in fact, the "treatment" response may have been an artifactual bias. To guard against this possibility, treatments should be assigned to specimens at random; the procedure follows.

Procedure

a. Determine the number of specimens to be used for all treatments.

For example, suppose we wish to study the influence of vitamins on growth of rat pups; three pups could be fed lab-chow and serve as controls and three treated pups could be fed lab-chow laced with vitamins. Six rat pups are needed.

b. Separate out unusual specimens prior to selecting a sample.

In order to reduce the variability in results, it is best not to include the unusual specimens that are available. Make this distinction prior to taking samples. In our example, one possible animal may be nearly as mature as an adult, hence would grow little no matter what it was fed; obviously this specimen should not be included, nor should the pup that lost an ear in a fight the day before.

In many cases, such as the characterization of a natural population, unusual individuals should not be disregarded.

c. Number the specimens that are available to be used in the experiment.

After separating out the atypical specimens we label each rat pup with a number, leaving, say, a total of nine specimens.

d. Select the number of random numbers necessary to assign individuals' treatments.

Use the procedure describing the selection of random numbers. The first three random numbers selected will assign rat pups to a diet of lab chow, and the next three random numbers will designate rat pups to be fed the vitamin enriched diet.

The procedure will work with any number of treatments and will provide assurance that the responses observed are not artifacts.

In selecting specimens to characterize a natural population, assign a number to every specimen and do not disregard what appear to be unusual individuals. Make up a sample by selecting random numbers and observing the corresponding specimens.

This procedure can also be used to assign specimens to locations in animal cages or on a greenhouse bench. If specimens representing different treatments are to be cultured, it is best not to place all specimens with the same treatment in a group since slight differences in the environment between groups could bias results. The arrangement of specimens should be random. To do this, number the specimens and assign them to locations using random numbers. The procedure is similar to that discussed above.

Random Selection of Sample Sites

When sampling a region in which there may be an infinite number of sample sites, it is not possible or even practical to number every site. It makes little difference if the region under consideration is a microscopic specimen or a continent, the problem is the same. To avoid artifacts that are associated with judgmental and haphazard sampling, sample sites should be located using a random process; the procedure follows.

Procedure

a. Define the area to be sampled with a map and place the coordinate system on the map.

See Figure 3. A genuine example of a plot to be sampled is illustrated with the aerial photograph of a small wooded plot in Lancaster County, Pennsylvania, in Figure 4. This could be an excellent opportunity to use stratified random sampling when geographical regions can be recognized.

b. State the number of samples to be taken from the region. Note the limits of the coordinate system which overlays the region.

For example, all points within the region depicted in Figure 3 can be located using coordinates with units between 0 and 20 on the north/south axis and between 0 and 30 on the east/west axis.

c. Select pairs of random numbers to define coordinates that will locate sample sites.

Since a region to be sampled may be irregular in shape, it is possible that some sets of coordinates will locate a sample site outside of the region; ignore these coordinates and proceed to obtain the desired number of samples within the region. With the example presented in Figure 3, random numbers for the north/south coordinate would be between 1 and 20 and between 1 and 30 for the east/west coordinate.

This method can be expanded into a two-step process (Peterson & Calvin 1965). In the initial step, the general sample location is found using a coarse scale, then the second set of coordinates is used in conjunction with a fine scale to identify the specific site to sample. For example, the coarse scale may be in kilometers and the fine scale in meters.

Summary

Students should be encouraged to participate in

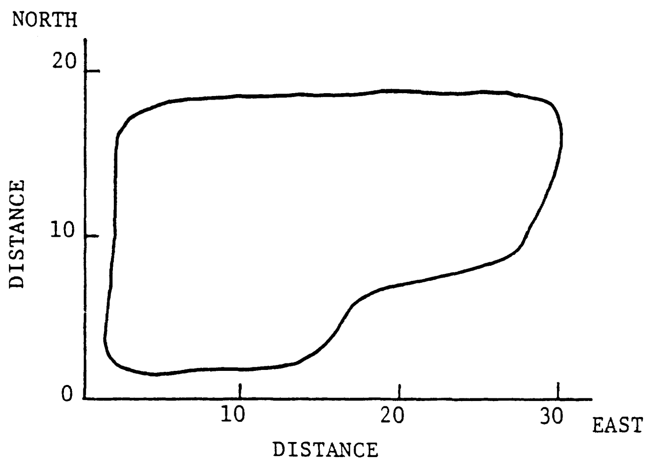

Figure 3. The region to be sampled is presented with two axes, one running north/south and the other east/west. All points within the region can be located using coordinates with units between 0 and 20 on the north/south axis and between 0 and 30 on the east/west axis. Pairs of random numbers are used as coordinates to locate sample sites within the region.

Figure 4. Aerial photograph of a small wooded plot in Lancaster County, Pennsylvania, with two axes superimposed on the photograph to illustrate how random samples could be located in this wooded area.

scientific research which requires that specimens be allocated to treatments or that sample sites be selected without bias. Perhaps the only way to avoid artificial bias is through the use of a randomization process. Proper use of random number table can be mastered easily by students in secondary school and college, and doing so provides an introduction to principles underlying statistical analysis.

Acknowledgments

I am grateful for the valuable suggestions offered by Jean Labriola, Larry Reinking and Spencer Stover; and thank Scott Kriner for technical assistance and Ken Miller and Don Rudy for the aerial photograph. In addition, I thank R.R. Sokal, F.J. Rohlf and W.H. Freeman and Company for the use of Table 1.

References

Abramowitz, M. & Stegun, I.A. (Eds.). (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Washington, DC: National Bureau of Standards Applied Mathematics Series 55. U.S. Government Printing Office. p 949–953.

Cochran, W.G. (1963). *Sampling techniques.* New York: John Wiley & Sons.

Folks, J.L. (1984). Use of randomization in experimental research. In K. Hinkelmann (Ed.), *Design, Statistical Models, and Genetic Statistics.* New York: Marcel Deckker.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). *Sample survey methods and theory: Volume I, methods and applications.* New York: John Wiley & Sons.

Hendricks, W.A. (1956). *The mathematical theory of sampling.* New Brunswick, NJ: The Scarecrow Press.

Jansson, B. (1966). *Random number generators.* Stockholm, Sweden: Victor Pettersons.

Kac, M. (1983). Marginalia: What is random? *American Scientist, 71,* 405–406.

Kac, M. (1984). Marginalia: More on randomness. *American Scientist, 72,* 282–283.

Petersen, R.G. & Calvin, L.D. (1965). Sampling. In C.A.B. Black (Ed.), *Methods of Soil Analysis: Part 1.* Madison, WI: American Society of Agronomy.

Rand Corporation. (1966). *A million random digits with 100,000 normal deviates.* New York: The Free Press.

Sokal, R.R. & Rohlf, F.J. (1969). *Introduction to biostatistics.* San Francisco: W.H. Freeman.

Spector, B.S. & White, A.Q. (1985). Teachers and researchers: Potential symbionts. *American Biology Teacher, 47*(2), 92–93.

Stigler, S.M. (1978). Mathematical statistics in the early stages. *Annals of Statistics, 6,* 239–265.

Yurkiewicz, W.J. & Steucek, G.L. (1969). Undergraduate Research. *Association of the Pennsylvania State College and University Faculty Journal, 23*(3), 3–4.