Simulation Games and Experiential Exercises in Action, Volume 2, 1975

THE USE OF PROGRAM BAYES IN THE TEACHING OF SAMPLE SIZE DETERMINATION IN SURVEY RESEARCH

James W. Gentry, Kansas State University
Alvin C. Burns, University of West Florida
Edward T. Reutzel, Kansas State University

INTRODUCTION TO THE PROBLEM

The teaching of the typical marketing research course requires discussion of sample size determination. Unfortunately, the treatment of the determination of the necessary sample size is either extremely brief or entirely omitted in the majority of the texts in the area (for example, Buzzell, Cox, and Brown, 1969; Luck, Wales, and Taylor, 1974; and Wentz, 1972).

Authors will argue, justifiably, that the practical difficulty encountered in survey research lies in the need to assure that the sample taken is a representative one, regardless of its size. Consequently, the bulk of discussion on sampling does and should center on sampling plans and procedures rather than sample size per Se. The greater significance of the chosen sample plan to the sample results is not being questioned here. Instead, it is felt that the size of the sample drawn is an important consideration in survey research, meriting more than the cursory treatment it currently receives in most marketing research texts. The need for expanded treatment of sample size determination is evident when one considers the attitudes of a majority of marketing practitioners with respect to large sample sizes. In our dealings with businessmen, we have found strong belief in the law of large numbers. As a result, proposed sample sizes derived from traditional statistical techniques fall far short of their expectations. It may be that these practitioners are aware that the assumptions of random sampling are rather stringent and do not lend themselves easily to the realities of practical application in marketing research. Alternatively, one might be led to the conclusion that they do not have an adequate understanding of the sampling process and/or statistical inference. Whatever the explanation, the point is clear: many businessmen expect to be told that large samples must be taken before any reliable information can be obtained. There appears to be a pervasive belief that the accuracy of sample information increases linearly with the sample size, rather than with the square root of the sample size which is the actual case.

Certainly traditional approaches to the determination of sample size have done little to enhance the student’s comprehension of this decision. Most marketing research text discussions rely on the standard error formula as explained by Ferber (1949). These techniques require that the variance of the population (or the sample proportion) be estimated in order to compute the sample size. This requirement becomes a major criticism, for most students and businessmen consider it counter-intuitive that one must assume what the results of the survey will be in order to develop a sample design to provide those very results. This reaction, coupled with the superficial treatment of the topic in the first place, leads one to search for more attractive alternatives.

One primary weakness of most approaches is that they fail to consider the trade-off between the accuracy of the added information and the cost of obtaining that information. This factor, of course, is a critical consideration In survey sample size determination and must be communicated to students of marketing research. The notion is in no way new, for it was concisely stated by Ferber (1949, p. 184) that "...the primary objective of every sampling survey is to obtain the desired information with maximum precision at a given cost or with a given precision at minimum cost." The Bayesian approach to sampling serves this objective nicely, for it introduces the conditional costs of incorrect sample size decisions directly into the model. Some marketing research texts do devote space to Bayesian analysis (see Cox and Enis, 1972) or to decision theory In general (Boyd and Westfall, 1972); however, only the Green and Tull (1975) text endeavors to apply Bayesian statistics to sample size determination problems. Their discussion concludes with the acknowledgment that the implementation of this procedure is quite laborous. They state that ". . .computer routines can be prepared to deal with optimal size selection in problem classes that appear frequently enough to justify the cost of programming a general computational procedure." (pp. 243-244).

It is clear at this point that there exists a need to utilize an innovative teaching device which explicitly treats precision and cost trade-off s and which simultaneously counters students' bias for large sample sizes. This paper discusses the use of program BAYES, which handles those classes of problems requiring a binomial sampling distribution. Use of BAYES allows marketing research students to learn and apply the Bayesian approach to sample size determination without the necessity of investing hours in tedious hand calculations. The program has considerable flexibility, for not only can students quickly determine the optimal sample size given their initial estimations of the prior probabilities, the cost functions, and the payoff functions, but it also provides them the opportunity to discover the sensitivity of the optimal sample size to changes in these initial conditions.

PROGRAM BAYES

There are, of course, computer programs available which are capable of determining the optimal sample size using the Bayesian approach. The authors are aware, for example, of program OPSAMD developed by Pohi (1974) as well as the algorithm published by

Moskowitz (1973). Whereas OPSAHD is an interactive program, BAYES is not due to system constraints at Kansas State University where BAYES was developed. However, the input-output statements are amenable to simple modification should any user desire to use BAYES on a system that has interactive FORTRAN capability.

BAYES also shares a limitation with most other available programs in that It requires a binomial sampling distribution. This limitation is no problem when the purpose of the survey is to obtain the proportion of people answering “yes” to a particular question or the proportion that buys brand A as opposed to buying brand B (or all other brands grouped together). An application of BAYES in such a case will be documented later. The limitation renders the program of little use when the purpose of the survey is to estimate a mean (a continuous distribution such as the normal distribution is required) or when the proportion of people classified in k different groups (a multinomial distribution is required).

INPUTS TO BAYES

The use of BAYES in student exercises guarantees that the learning experience will involve a myriad of considerations which have impact on the sample size. Students are required to consider action alternatives, possible outcomes, associated payoffs, and research costs[1] The following list of inputs should indicate the nature of a priori decisions the student must utilize.

1. The number of alternative courses of action available to the manager and the label for each one.

2. The number of states of the world and a proportion which describes each one. For example, the typical Bayesian sampling example involves sampling from a production process which may have either 1% defective items or 10% defective items. An analogous marketing research situation might be one in which the decision maker is interested in whether the firm’s market share was 10%, 20%, or 30%.

3. The payoffs accruing to the alternative course of action contingent upon the possible states of the world. These may be input either in functional form or as individual amounts. The functional form has the advantage that new payoffs do not have to be input when states of the world are added or deleted.

4. The prior probabilities for the states of the world.

5. The cost of sampling, expressed in the form

$$CSa+bN+cNd \tag{1}$$

Ferber (1949) recommends the particular form

$$CS = a + bN - cN^2 \qquad (2)$$

which accounts for the fixed costs for overhead and a variable cost that gradually decreases per respondent as learning takes place.

6. The form of analysis sought.

ALTERNATIVE OUTPUT FROM BAYES

1. Regardless of the option chosen by the student, the program outputs the prior analysis (determination of the most profitable action before obtaining any sample information) and the expected value of perfect information.

2. The first option is the calculation of the expected value of sample information (EVSI), the cost of sampling (CS), and the expected net gain from sampling (ENGS) for a particular sample size n. If desired, students can ask for the complete printing of all the probability revisions involved in the calcu- lation of EVSI. For example, it is possible for them to see how the probability of each state of the world is changed when r people out of n answer the question positively (r will vary from 0 to n). The amount of printout required to output the n + 1 possible sample outcomes becomes very costly for large n's. However, students have found that the extensive output (for small n's) greatly helps them to understand the procedure being performed by the computer.

3. The second option is for an enumeration procedure, starting at n_0 (which is supplied by the student) and terminating shortly after the optimal n has been determined. A plot of EVSI, CS, and ENGS is included in the output, as well as the listing of the needed number of respondents answering the question positively (or negatively, depending on the prior action) before one should change actions. The primary drawback to the enumeration procedure is the amount of computer time required. This was especially true for those cases in which the specified starting sauiple size is far from the optimal n.

4. The third option provides the optimal sample size through the use of the algorithm outlined by Moskowitz (1973). There is no simple empirical solution to the complex maximization problem involving the ENGS function (Raiffa and Schlaifer, 1961). Consequently, Moskowitz's search technique, which quickly determines the switchover point (the point where the optimal action changes) for the various n's without performing the full analysis, appears to be the quickest approach available. Whether or not the approach is the fastest technique is a moot question; it is clear that the approach is quicker than explicit enumeration. To solve the simple two-action, two state-of-the-world problem discussed in Moskowitz (1973) using explicit enumerations, the program takes .183 minute when the enumeration stops shortly after

the optimal size is found and .231 minute when the enumeration is allowed to proceed to

$$\text{NMAX} = \frac{\text{EVPI - Fixed Cost of Sampling}}{\text{Variable Cost of Sampling}} \tag{3}$$

Using the search algorithm the program takes .179 minute to derive the same results. (All runs were made on an IBM 370/158 computer.) The amount of time saved through the use of Moskowitz' algorithm increases greatly when the optimal n is large. (The optimal sample size in the trial problem was 10.)

APPLICATION OF BAYES IN A CASE

BAYES has been used as a student experience in conjunction with the "Transit Radio, Inc. (A)" case from Boyd and Westfall (1972, pp. 384-385). The following excerpts from the case develop its scenario.

> Transit Radio, Inc. was negotiating for the franchise to install FM radio loudspeakers on buses in Eastville, an eastern city of 95,000. If it got the franchise, Transit Radio would broadcast music over the loudspeakers and would sell time to advertisers for commercials. The amount of income it might obtain from such advertising would determine how much it could afford to offer for the franchise.
>
> Officials of Transit Radio believed they would have to pay approximately $50,000 for the franchise, but they were not sure this was a price they could afford to pay. The equipment to be installed on each bus would cost about $200. Since there were 48 buses on the city line, the total investment in equipment would be about $9,600.
>
> To help in arriving at an estimate of the potential advertising income, the company executives decided to make a survey among Eastville residents of 15 years of age or more. They engaged a consultant to plan the details of the research. He developed a questionnaire which included the following question:
>
> 1. Have you ridden on an Eastville city bus in the last seven days?...
>
> Pretesting of the questionnaire and discussions with bus line officials led the research consultant to the following estimates of the answers that would be obtained to these questions:
>
> Q#1 70 percent will have ridden a bus in the Last seven days...
>
> 1. How large a sample should the research consultant draw?...

The case was designed to be an application of the standard error formula, but it can be enlarged by adding information on costs so that the Bayesian approach can also be used. The alternative actions are: 1) Put radios on the buses, and 2) do nothing.

The states of the world used in this example were .10 (10% of people rode the bus), .30, .50, .70, and .90, and the initial estimates of the prior probabilities were .05, .2, .4, .25, and .1. One of the points of the modified case is for the students to think about the reasonableness of the given priors and for them to test the sensitivity of the final sample size when the priors are changed. For the purposes of this example, the payoffs are given as

		STATES OF THE WORLD				
		10%	30%	50%	70%	90%
ACTIONS	Put radios on the bus	-15000	-5000	10,000	20,000	40,000
	Do nothing	0	0	0	0	0

In actual usage, the students are required to compute the annual cost of the preliminary investment and to estimate operating and maintenance costs for the system (data regarding these costs are supplied). The students are told to assume that there is a functional relationship between the proportion of people riding the bus and the number of advertisers who will buy time. Also the students are told to estimate the cost of sampling as a function of n. This requires them to make estimates of the fixed costs (computer time, questionnaire construction, training interviewers---assuming that the available people will all be used regardless of the sample size selected), the linear variable costs (per interview rate, postage), and the non-linear variable costs in which learning occurs (coding time, travel costs). The function used for this example was:

$$CS = \$200 + \$3N - \$.002N^2 \quad (4)$$

Table 1 is an example of the output from program BAYES given the initial conditions. Page limit constraints prevent the presentation of the entire output.

In the example discussed above, the I~Opti~1ut solution from the search routine (N = 79 with ENGS of $1,198.56) differs from the “optimal” solution from the enumeration procedure (N 81 with ENGS of $1,200.21). In fact, explicit enumeration from 1 to NMAX finds that the true optimal solution is N=88 with ENGS of $1,200.61. The different solutions result from the use of Noskowitz’s decision rule for finding the region of the global optimum. This decision rule is valid when the cost of sampling function Is linear, which was not the case in this example. Thus explicit enumeration is required to find the true optimum. However, this is an extremely expensive approach. The program was allowed to run 5 minutes and it had progressed from an N of 1 to an N of 183. It would have taken hours to have reached NMAX (567 in this situation). The search procedure took .425 minute and the enuiner- ation procedure that uses Moskowitz’s stopping rule took .660 minute. For the initial conditions given in this case, all sample sizes between 53 and 135 provide values of ENGS greater than

$1,150 and less than $1,201. This information gives the manager much flexibility in fitting his sampling design into his budget.

CONCLUSIONS

As a learning experience, program BAYES greatly facilitates the student's comprehension of the Bayesian approach to sample size determination. Whereas most marketing research texts have abbreviated discussions of standard error methods and those that do treat the Bayesian approach make disparaging comments about computational constraints, program BAYES provides the student with a flexible tool that he can apply to practically any problem involving a binomial sampling distribution.

REFERENCES

1. Buzzell, Robert D., Cox, Donald F., and Brown, Rex V. Marketing Research and Information Systems, New York: McGraw- Hill, 1969.

2. Cox, Keith K. and Enis, Ben M. The Marketing Research Process, Pacific Palisades, Cal: Goodyear Publishing Co., 1972.

3. Ferber, Robert. Market Research, New York: McGraw-Hill, 1949.

4. Green, Paul E., and Tull, Donald S. Research for Marketing Decisions, Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

5. Luck, David J., Wales, Hugh G. and Taylor, Donald A. Marketing Research, Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

6. Moskowitz, Herbert. "An Algorithm for Determining Bayesian Attribute Single Sampling Acceptance Plans," Krannert Paper, No. 427, Purdue, October 1973.

7. Pohl, Norval F. "On Teaching the Economic Value of Sample Information (The Discrete Case): An Interactive Computer Approach," Proceedings, Sixth Annual Meeting of AIDS, Atlanta, 1974, p. 40.

8. Raiffa, Howard, and Schlaiffer, Robert. Applied Statistical Decision Theory, Cambridge, Mass: MIT Press, 1961.

9. Wentz, Walter B. Marketing Research: Management and Methods New York: Harper and Row, 1972.

10. Winkler, Robert L. Introduction to Bayesian Inference and Decision, New York: Holt, Rinehart, and Winston, 1972.

TABLE 1

Sample Output Showing Results of Explicit Enumeration (Using Moskowitz's Stopping Rule)

N	CS	EVSI	ENGS	C*
1	203.0	0.0	-203.0	-1.
2	206.0	0.0	-206.0	-1.
3	209.0	250.7	41.8	0.
4	212.0	441.3	229.3	0.
5	215.0	473.7	258.8	0.
6	217.9	592.1	374.2	1.
7	220.9	698.2	477.3	1.
8	223.5	718.0	494.7	1.
9	226.8	792.3	565.4	2.
10	229.8	853.4	623.6	2.
11	232.8	861.8	629.1	2.
12	235.7	922.8	687.1	3.
13	238.7	957.2	718.5	3.
14	241.6	959.8	718.2	4.
15	244.6	1015.7	771.1	4.
16	247.5	1032.2	784.7	4.
17	250.4	1053.1	802.7	5.
18	253.4	1085.7	832.4	5.
19	256.3	1089.7	833.5	5.
20	259.2	1124.3	865.1	6.
21	262.1	1140.8	878.7	6.
22	265.0	1151.8	886.7	7.
23	267.9	1180.4	912.5	7.
24	270.8	1195.4	914.5	7.
25	273.8	1210.6	936.8	8.
26	276.6	1225.7	949.0	8.
27	279.5	1232.8	953.3	9.
28	282.4	1257.8	975.4	9.
29	285.3	1262.8	977.5	9.
30	288.2	1282.0	994.4	10.
31	291.1	1296.2	1005.1	10.
32	294.0	1301.1	1007.2	11.
33	296.8	1323.0	1026.2	11.
34	299.7	1327.9	1028.2	11.
35	302.6	1343.8	1041.2	12.
36	305.4	1356.0	1050.6	12.
37	308.3	1359.3	1051.0	13.
38	311.1	1378.5	1067.4	13.
39	314.0	1383.2	1069.2	13.
40	316.8	1396.0	1079.2	14.
41	319.6	1407.0	1087.4	14.
42	322.5	1409.1	1086.6	15.
43	325.3	1426.1	1100.9	15.
44	328.1	1430.5	1102.4	15.
45	331.0	1440.9	1110.0	16.
46	333.8	1450.8	1117.0	16.
47	336.6	1452.0	1115.4	17.
48	339.4	1467.0	1127.6	17.
49	342.2	1471.2	1129.0	17.
50	345.0	1479.6	1134.6	18.
51	347.8	1488.5	1140.7	18.
52	350.6	1489.0	1138.4	19.
53	353.4	1502.3	1149.0	19.
54	356.2	1506.3	1150.1	19.
55	359.0	1513.1	1154.1	20.
56	361.7	1521.1	1159.4	20.
57	364.5	1521.1	1156.6	20.
58	367.3	1532.9	1165.7	21.
59	370.0	1536.7	1166.4	21.
60	372.8	1542.1	1169.3	22.
61	375.6	1549.3	1171.8	22.
62	378.3	1549.6	1171.3	22.
63	381.1	1559.5	1178.4	23.
64	383.8	1563.0	1179.2	23.
65	386.6	1567.3	1180.8	24.
66	389.3	1573.9	1184.6	24.
67	392.0	1574.3	1182.3	24.
68	394.8	1582.6	1187.9	25.
69	397.5	1585.8	1188.3	25.
70	400.2	1589.3	1189.1	26.
71	402.9	1595.2	1192.3	26.
72	405.6	1595.8	1190.1	26.
73	408.3	1602.7	1194.4	27.
74	411.0	1605.7	1194.7	27.
75	413.8	1608.5	1194.7	28.
76	416.4	1613.8	1197.4	28.
77	419.1	1614.5	1195.3	28.
78	421.8	1620.3	1198.5	29.
79	424.5	1623.1	1198.6	29.
80	427.2	1625.3	1198.1	30.
81	429.9	1630.1	1200.2	30.
82	432.6	1630.8	1198.2	30.

OPTIMAL N IS 81 WITH ENGS OF 1200.2