The SURVEY Program: Description and Exercises

for

Sampling: Design and Analysis

2nd EDITION

Sharon L. Lohr
Arizona State University

Prepared by

Sharon L. Lohr
Arizona State University

BROOKS/COLE
CENGAGE Learning

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

ISBN-13: 978-0-495-12527-3
ISBN-10: 0-495-12527-X

Brooks/Cole
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at:
international.cengage.com/region

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit
academic.cengage.com

Purchase any of our products at your local college store or at our preferred online store
www.ichapters.com

The SURVEY Program

Two thirds of Americans tell researchers they get "most of their information" about the world from television, and the other statistics are so familiar we hardly notice them—more American homes have TVs than plumbing and they're on an average of seven hours a day; children spend more time watching TV than doing anything else save sleeping; on weekday evenings in the winter half the American population is sitting in front of the television; as many as 12 percent of adults (that is, one in eight) feel they are physically addicted to the set, watching an average of fifty-six hours a week; and so on.

–Bill McKibben, *The Age of Missing Information*

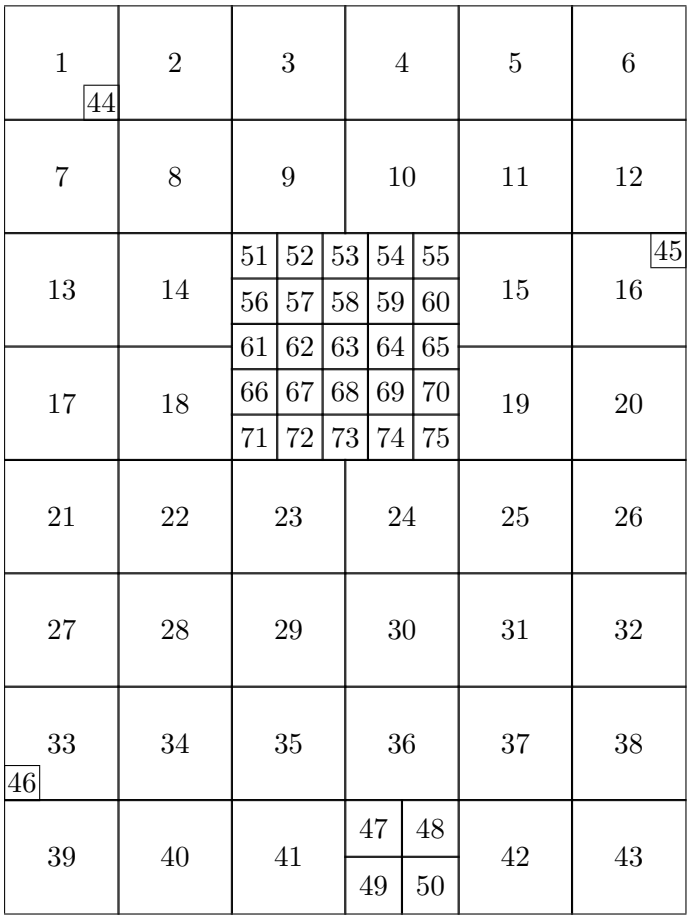
The computer program SURVEY,[1] developed by Theodore C. Chang, simulates the results and costs that might be experienced in actual sample surveys. The exercises using SURVEY are designed to provide a practical illustration of the theoretical aspects of survey design, and to allow comparisons between the different designs discussed in the course. Fortran code and executable files for the program are available on the book website.

Stephens County is a fictitious county in the midwestern part of the United States with a population of approximately 103,000. It has two main cities: Lockhart City, population 57,500, and Eavesville, population 11,700. Both cities are commercial and transportation centers and boast a variety of light industries. Among the county's industrial products are farm chemicals, pet foods, cable and wire, aircraft radios, greeting cards, corrugated paper boxes, industrial gases, and pipe organs. The county has three smaller municipalities: Villegas, Weldon, and Routledge with populations between one and two thousand. These cities are local commercial centers. The surrounding areas are agricultural, although a sizeable number of persons commute to the larger cities. The county's main agricultural products are beef cattle, wheat, sorghum and soybeans.

Stephens County has been organized into 75 districts with the houses within a district numbered consecutively starting with 1. For the purposes of these exercises, you may assume that houses in the same district with close numbers are physically close. The district map of Stephens County is provided in Figure 1. Information about each district is in Figures 2 and 3.

The Stephens County Cablevision Company has been formed to provide cable TV service to Stephens County. It has commissioned this survey to help it with its pricing and programming decisions.

[1]Part of the material in this page previously appeared in Chang et al. (1992), which introduced the SURVEY program.



Area	Districts	No. of houses
Rural areas	1 to 43	7932
Lockhart City	51 to 75	19664
Eavesville	47 to 50	3236
Villegas	44	283
Weldon	45	562
Routledge	46	312

Figure 1: A District Map of Stephens County

Figure 2: STEPHENS COUNTY DISTRICT INFORMATION, part 1

```
Column 1: District number
Column 2: Number of houses
Column 3: Cumulative house count
Column 4: Population
Column 5: Mean assessed house valuation
```

(1)	(2)	(3)	(4)	(5)
1	142	142	526	65248.
2	153	295	624	58759.
3	135	430	508	62319.
4	128	558	560	59416.
5	110	668	455	57202.
6	103	771	404	59290.
7	105	876	421	71122.
8	385	1261	1488	79265.
9	296	1557	1112	75921.
10	287	1844	994	68254.
11	253	2097	929	60660.
12	172	2269	628	53569.
13	198	2467	768	65182.
14	432	2899	1595	77907.
15	248	3147	864	65739.
16	251	3398	915	53771.
17	221	3619	864	68257.
18	297	3916	1099	78449.
19	235	4151	812	70772.
20	171	4322	687	52711.
21	135	4457	525	66739.
22	254	4711	923	66249.
23	203	4914	708	74757.
24	244	5158	825	75766.
25	202	5360	799	68989.
26	103	5463	388	56994.
27	102	5565	398	58940.
28	115	5680	448	60448.
29	180	5860	693	69111.
30	190	6050	766	69685.
31	152	6202	633	70276.
32	141	6343	572	63819.
33	143	6486	610	58636.
34	135	6621	491	55554.
35	178	6799	699	62361.
36	221	7020	811	60052.
37	174	7194	719	55699.
38	101	7295	390	53322.
39	95	7390	312	57174.
40	130	7520	446	55702.

Figure 3: STEPHENS COUNTY DISTRICT INFORMATION, part 2

41	152	7672	533	53285.
42	169	7841	672	56866.
43	91	7932	371	50710.
44	283	8215	1029	60057.
45	562	8777	2079	57233.
46	312	9089	1149	52719.
47	897	9986	3263	62034.
48	734	10720	2623	60764.
49	963	11683	3490	60010.
50	642	12325	2318	54498.
51	525	12850	1825	95123.
52	726	13576	2497	68406.
53	674	14250	1948	53634.
54	585	14835	1219	48643.
55	553	15388	1090	43493.
56	583	15971	1977	95110.
57	911	16882	2691	84394.
58	1051	17933	2663	57657.
59	918	18851	1824	36706.
60	799	19650	1636	44308.
61	545	20195	1853	101906.
62	895	21090	2588	74815.
63	1313	22403	2642	55560.
64	968	23371	2457	62813.
65	717	24088	2203	69846.
66	651	24739	2197	93771.
67	886	25625	2711	82902.
68	912	26537	2750	76832.
69	898	27435	2671	72062.
70	759	28194	2650	79887.
71	722	28916	2568	87383.
72	753	29669	2652	80341.
73	793	30462	2763	79833.
74	725	31187	2560	83354.
75	802	31989	2870	80522.

(1)	(2)	(4)	(5)	
1-43	7932	29985	65511	RURAL
44-46	1157	4257	56706	VILLEGAS, WELDON, ROUTLEDGE
1-46	9089	34242	64390	RURAL
47-50	3236	11694	59649	EAVESVILLE
51-75	19664	57505	71117	LOCKHART CITY
1-75	31989	103441	68045	STEPHENS COUNTY

The Interview Questionnaire. The Stephens County Cablevision Company has supplied an interview questionnaire for your use, shown below.

Information from the Interview Questionnaire

I am doing a survey for Stephens County Cablevision, Inc. As you may know, Stephens County will soon have cable service; you can help us make sure that the programming we offer meets your needs by answering the following questions.

1. How many persons aged 12 or older live at this address? Please include any persons you consider to be part of your family; do not include persons renting rooms from you.
2. How many persons aged 11 or younger live at this address?
3. How many television sets are in this household?
4. If cable television service cost $5 per month, would your household subscribe? If it cost $10 per month? $15? $20? $25? (The interviewer records the highest price the respondent would be willing to pay for cable.)
5. How many hours did you, personally, spend watching TV last week, in the period from _____ to _____? Your spouse? Each child? Other persons living in the household? (The interviewer sums these amounts and records the sum. If other persons are available, they are asked directly.)

   For the following types of programming, the total number of hours spent watching the type of programming are recorded.
6. How many hours did you watch news and "public affairs" programming last week? What about other members of the household?
7. Sports
8. Children's programming
9. Movies

In addition, for each surveyed household, the Company has obtained from the county tax assessor the assessed valuation of that household's living quarters. This information is meant to provide a measure of family income (without having to ask about it). Note that Stephens County is somewhat behind the rest of the United States in terms of cable TV and satellite dishes; this may be because the original SURVEY program was written in 1982.

Survey program assumptions. To make as realistic a simulation as possible, certain assumptions have been programmed into SURVEY. These assumptions should be used in efficient design. Assumptions 1 and 2 are obvious; the others seem reasonable.

ASSUMPTIONS FOR SURVEY

1. Each occupied address has at least one adult.
2. Only households with televisions will be willing to subscribe to cable service.
3. All other factors being equal, a household with a higher income will tend to have a more expensive house.
4. Assessed valuation is a reasonably accurate estimate of house price.
5. All other factors being equal, a household with a higher income will tend to be willing to pay more for cable service.
6. All other factors being equal, a household with a higher income will tend to own more television sets. This tendency is much weaker than that of assumption 5 because of the low cost and longevity of most TV sets.
7. Larger families tend to be more willing to subscribe to cable TV.
8. All other factors being equal, a family's willingness to subscribe to cable TV decreases as the other entertainment options available to it increase. These options decrease the further one moves from the population concentrations in Stephens County.
9. Due to zoning and development practices, urban neighborhoods tend to be more homogeneous than rural neighborhoods.

Costs of sampling in Stephens County Of course, one does not obtain information from survey respondents for free. SURVEY has built-in costs for sampling various units:

```
SAMPLING COSTS IN STEPHENS COUNTY
$60 per rural district visited (1-46)
$20 per urban district visited (47-50)
$6 per rural household visited (whether home or not)
$3 per urban household visited (whether home or not)
$10 processing cost per completed interview
```

As an example of the above costs, if the addresses visited and interviewed were 3-47, 3-25, 5-16, 51-25, and 51-36, the sampling cost printed at the end of the output from the program SURVEY would be 2*60 + 1*20 + 3*6 + 2*3 + 5*10 = $214.

Running the SURVEY program. The FORTRAN source code, and the executable files for the IBM PC is provided on the data disk. You are welcome to copy the source code and the executable files, and to use these on your own machine.

To run the SURVEY program on an IBM PC, type `survey.exe` For other operating systems, you need to compile SURVEY, using a FORTRAN compiler available for that system.

SURVEY first asks you to enter the desired nonresponse rates. For now, we're assuming that everyone in Stephens County is always at home and cooperative, so enter

`0 0 0 (carriage return)`

Then, when asked, enter the address of each household to be questioned in the form

`district number, house number  (carriage return).`

SURVEY responds "DONE" to each correctly entered address. When you have finished your list of houses enter zero for the district number followed by any house number.

```
SAMPLE RUN
DEMONSTRATION EDUCATIONAL SAMPLE SURVEY PROGRAM
TED CHANG, UNIVERSITY OF KANSAS, SEPT 1986
ENTER FILENAME CONTAINING ADDRESSES--8 OR FEWER  LETTERS
IF ENTERING FROM TERMINAL, TYPE T
t
ENTER FILENAME FOR OUTPUT--8 OR FEWER  LETTERS
myoutput
ENTER DESIRED THREE NONRESPONSE RATES:
NOT-AT-HOMES, REFUSALS, RANDOM ANSWERS
0 0 0
ENTER DISTRICT NUMBER, HOUSE NUMBER
23,45
DONE
ENTER DISTRICT NUMBER, HOUSE NUMBER
22,96
DONE
ENTER DISTRICT NUMBER, HOUSE NUMBER
53,47
DONE
ENTER DISTRICT NUMBER, HOUSE NUMBER
583,22
DISTRICT NUMBERS MUST BE BETWEEN -75 AND 75.
RE-ENTER DISTRICT NUMBER, HOUSE NUMBER
SET DISTRICT NUMBER = 0 TO STOP PROGRAM.
ENTER DISTRICT NUMBER, HOUSE NUMBER
55,9999
IN DISTRICT 55 HOUSE NUMBERS MUST BE BETWEEN 1 AND  553
RE-ENTER DISTRICT NUMBER, HOUSE NUMBER
SET DISTRICT NUMBER = 0 TO STOP PROGRAM.
ENTER DISTRICT NUMBER, HOUSE NUMBER
0,0
```

```
THE COST OF THIS SESSION IS    185 DOLLARS.

This is what the file myoutput looks like:
 ADDRESS   VALUE   1   2   3   4    5    6    7    8    9
 23   45   62673   2   1   3  15  130   11   28    7   20
 22   96   85553   2   2   1  20   86   10   29    7   32
 53   47   83183   2   0   3  10   52    4   39    0    8
 THE COST OF THIS SESSION IS    185 DOLLARS.
```

VALUE = house value, and the numbers in columns labelled 1 through 9 are the household's responses to questions 1 through 9. SURVEY places the answers that each house gives in the file you have specified. You may then edit or print the file using a word processor or spreadsheet.

To analyze the data, you need to use a computer package or program that has subset selection capabilities and allows you to write your own programs. Most programs that only have menus but no programming language are not flexible enough to be useful in survey sampling.

To use any statistical package, you must first use a text editor to remove last line of the output from SURVEY. For some packages, you must remove the first line, with the variable names, as well.

Using R for SURVEY output. In R for windows, you can import the output file into a data frame (here, we call it `survout`) using menu commands. To access any variable, say cable price, use `survout$cable`; the average cable price is `mean(survout$cable)`. Subset selection is simple in R: to print the cable values for the subset of households willing to spend at least $10, type

```
survout$cable[survout$cable >= 10]
```

Using SAS to analyze data from the SURVEY program. If the output is in the file "out.txt" in the C directory, you can read in the data and calculate summary statistics for the variables in SAS with the commands:

```
filename myoutput 'C:\out.txt';
data tv;
   infile myoutput firstobs=2;
   input dist house value over12 under12 numtv cable hourstv news
                          sports child movies;
proc means data=tv;
run;
```

Computer Generation of Random Addresses For any sampling scheme to work effectively, the units must be selected randomly. This is a laborious process and many sample surveys are ruined by attempts to short-cut it.

As we shall be selecting thousands of addresses for the simulation studies, a program called ADDGEN has been written by C. G. MacLaren to randomly select addresses from any specified set of districts. ADDGEN will ask the user for a random start. This is any integer between 1 and 1000000 which the program uses as a start point in a long table of random numbers. Given the same start, districts, sample size, and type of computer, ADDGEN will always produce the same sample of addresses. It is extremely important that you record the start in order to repeat a particular sample for further analysis in future assignments. The random start is written on the last line of the output file from ADDGEN.

The program then asks for the districts from which you wish to sample. Any subset of the districts 1 to 75 can be specified. You simply enter the desired district numbers along a line separated by commas. If you want consecutive districts you only need to type the first and last district numbers separated by a - (dash symbol). If you need to continue your list onto a new line simply end the previous line with a $ (dollar symbol), press return, and continue on the next line. Finally the program asks for the number of addresses to be selected from the specified districts.

The program ADDGEN generates an output file named by you in a format suitable for input into the survey program. When running SURVEY, you merely type in the name of the file you created using ADDGEN.

Sample run of ADDGEN. The following is a journal of a sample run which was made using the above procedure. The program ADDGEN was used to create a random sample of size 5 from districts 1-49,60,70. The output file "add.txt" from ADDGEN can be fed to SURVEY.

```
ENTER FILENAME FOR ADDRESS SET--8 OR FEWER LETTERS
add.txt
ENTER RANDOM START--ANY INTEGER BETWEEN 1 AND 1000000
219654
ENTER DISTRICTS FROM WHICH YOU WISH TO SAMPLE
1-49,60,70
     51 DISTRICTS WITH   13241 HOUSEHOLDS HAVE BEEN SPECIFIED
ENTER NUMBER OF ADDRESSES TO BE GENERATED (MAX 1000)
5
DO YOU WANT TO SPECIFY A NEW DISTRICT SET
ANSWER YES OR NO
no
   5 RANDOM ADDRESSES GENERATED WITH RANDOM START  219654

Below are the contents of the file 'address:'
  4    67
  8   246
 18    94
 18   191
 24   244
  0     0                    219654
```

SURVEY exercises

Chapter 2 Exercises

1. Why is the following procedure not suitable for drawing a simple random sample of addresses in Lockhart City?

   (a) Randomly select a district between 51 and 75.

   (b) Randomly select a house from those in the chosen district.

   (c) Reject both district and house selection if the house is already in the sample.

   (d) Repeat a - c until the desired sample size is achieved.

2. No district in Lockhart City has more than 1313 houses. Prove that the following procedure does produce a simple random sample of houses in Lockhart City:

   (a) Randomly select a district between 51 and 75.

   (b) Randomly select a random number (the potential house selection) between 1 and 1313.

   (c) Reject the two random numbers from (a) and (b) if the number in (b) exceeds the number of houses in the district or if the house is already in the sample. Otherwise, add that house to your sample.

   (d) Repeat (a)-(c) until the desired sample size is achieved.

3. Use a random number table to select a simple random sample of size 10 from Lockhart City. Report the list of the random numbers you selected and the addresses to which they correspond. Describe exactly how you converted a random number to an address.

4. Use the SURVEY program to obtain the answers to the questionnaire for your 10 randomly selected addresses. Hand in a printout of the output file. Estimate the following from your sample of 10 households. Give standard errors for your estimates:

   (a) The average number of TVs per household in Lockhart City.

   (b) The average price a household in Lockhart City is willing to pay for cable TV service.

   Actually we only know for each sampled household the price it is willing to pay for service rounded down to the nearest $5. Recognizing this limitation to question 4 of the survey questionnaire, use the answers to that question as the prices that the sampled houses are willing to pay.

5. Use the program ADDGEN to generate 200 random addresses in Lockhart City and then the program SURVEY to obtain the responses of these houses. Estimate

   (a) The average price a household is willing to pay for cable TV.

   (b) The average number of TV's in a household in Lockhart City.

   (c) The proportion of houses willing to pay at least $10 for cable service. This really means, of course, at least $10.

   Be sure to give standard errors for all estimates. (Use the fpc, even though it may not be strictly necessary.) Make sure you save the sample you obtained for this exercise—you will use it again in later chapters.

6. Using your sample of size 200, estimate the average assessed valuation in Lockhart City. Does a 95% confidence interval include the known value of $71117? Estimating a known quantity is often used to check the representativeness of a sample.

7. Draw a histogram or stem-and-leaf diagram of the responses to question 8 of the survey (number of hours watching children's TV) using the sample you drew in Exercise 5. Does the distribution of number of hours spent watching children's TV for households in Lockhart City appear normal? Find an approximate 95% confidence interval for the mean number of hours spent watching children's TV. Based on your histogram, is constructing a confidence interval an appropriate thing to do? Why or why not? (Hint: Do you think that the sampling distribution of the mean viewing time for children's TV could be normal?)

   Chapter 3 Exercises.

8. In the quest to estimate the average price a household in Stephens County is willing to pay for cable TV service, we are fortunate to know a great deal about some demographic aspects of the county, as given in the district map and tables in Appendix A. According to the SURVEY assumptions, what information might be used to stratify Stephens County in order to improve the precision of estimates? Are any other reasons for stratification relevant to Stephens County?

9. Use any considerations you like to divide Stephens County into strata. Your stratification should divide Lockhart City into approximately five strata. Why did you choose your stratification variable? Count the total number of households in each of your strata. (You may use the program ADDGEN to do this.)

   The remainder of these exercises concern Lockhart City ONLY.

10. Using ADDGEN generate a stratified random sample of size 200 from Lockhart City with your stratification in the previous question and proportional

allocation. Find the responses using the program SURVEY. Estimate the average price a household in Lockhart City is willing to pay for cable service and the average number of TV's per household in Lockhart City. How do these estimates compare with those obtained with simple random sampling and sample mean and ratio estimates? Which estimates are the most precise?

11. Pilot studies are often used to estimate S_h. In this case we are fortunate to have a very large pilot study from the sample of size 200 used in Chapter 2. Divide your sample from Chapter 2 into the strata you chose above and thus obtain estimates of the variances S_h^2 in each of the strata for the average price a household is willing to pay for cable TV service.

12. The sampling costs for Stephens County are given in Appendix A. Using your estimates of S_h, optimally allocate a sample of size 200 to estimate the average price a household in Lockhart City is willing to pay for cable TV service. Using that allocation take a stratified random sample of Lockhart City and estimate the average price a household is willing to pay for cable TV service and the average number of TV's per household.

13. Under what conditions can optimal allocation be expected to perform much better than proportional allocation. Do these conditions occur in Lockhart City? Comment on the relative performance that you observed between these two allocations.

14. Using the variances estimated in question 5 of Chapter 2, what size sample would be needed with simple random sampling to achieve the same precision in estimating the average price a household is willing to pay as a stratified sample of size 200 using the strata you have designed and optimal allocation? proportional allocation?

15. Are there any deficiencies in your design? How would you correct them if you were to do this exercise a second time?

Chapter 4 Exercises.

16. Using the same sample of size 200, repeat problem 5 of Chapter 2 using a ratio estimate with assessed value of the house as the auxiliary variable. Which estimate of the mean gives greater precision? How are your results related to the SURVEY program assumptions? Be sure to include an appropriate plot of the data.

17. Using your sample of size 200, estimate the average number of adults per household in Lockhart City households willing to pay at least \$10 for cable service. Give the standard error and the estimated coefficient of variation of your estimate.

18. Using your sample of size 200, estimate the total number of adults in Lockhart City who live in households willing to pay at least \$10 for cable service. Give the standard error and the estimated coefficient of variation of your estimate.

Chapter 5 Exercises.

19. We would like to see if a cluster sample from the rural areas of Stephens County can improve on the precision of a simple random sample of size 100 while costing the same. To do this, we need to know the cost of sampling 100 houses randomly in districts 1 through 43. Use ADDGEN to generate ten different simple random samples of size 100 from the rural districts; calculate how much each of those different samples would cost, and average the costs to get an estimate of the cost of sampling 100 houses randomly in districts 1 through 43.

20. Design a two-stage cluster sampling scheme for the rural areas (districts 1-43) of Stephens County. Your design should choose between 25% and 50% of the districts (clusters) with equal probability, should subsample within each chosen district with sample size proportional to district size (number of houses), and should cost about the same amount as a simple random sample of size 100.

21. Using your sample from the previous exercise, estimate the average price a rural household is willing to pay for cable TV using both an unbiased estimate and a ratio estimate. Be sure to give standard errors, and to plot the data appropriately.

Chapter 6 Exercises.

22. Design a self-weighted sampling scheme for districts 1-43, with districts (clusters) chosen with probability proportional to size and with replacement. Your design should have the same number of clusters and cost about the same amount as the sample in Chapter 5. For this sample, estimate the average price a rural household is willing to pay for cable TV, along with the standard error of your estimate.

23. Comment on the relative performance of the three estimates:

    (a) Cluster sample with equal probabilities, unbiased estimate

    (b) Cluster sample with equal probabilities, ratio estimate

    (c) Cluster sample with probabilities proportional to district size

    Which estimate is most precise? Explain.

Chapter 7 Exercises.

24. Design a stratified cluster survey for Stephens county. Stratify on two variables: urban/rural and assessed valuation. Then within each stratum, select two districts with probability proportional to population, and sample an equal number of households within each district selected. Construct the sampling weights for each household in your sample.

25. Execute the sample and estimate the average price a household is willing to pay for cable TV. Be sure to give standard errors.

26. Compare your results with those from an SRS with the same number of households. What is the estimated design effect for your survey? How do the costs compare?

Chapter 8 exercises

When running SURVEY, you may have noticed the prompt

```
ENTER DESIRED THREE NONRESPONSE RATES:
NOT-AT-HOMES, REFUSALS, RANDOM ANSWERS
```

If you enter

.3 0 0

in response, about 30% of the households in Stephens County will "not be home." If you enter

0 .3 0

about 30% of the households in Stephens County will refuse to say how much they would be willing to pay to subscribe to cable TV. If you enter

0 0 .3

about 30% of the households in Stephens County will give random answers to certain questions.

27. Generate 200 random addresses for a simple random sample of the households in Stephens County. You will use this same list of addresses for all of the problems in this chapter. Draw the full sample of size 200 specified by those addresses with no nonresponse. This sample gives the values you would have if all households responded. Estimate the means for the assessed value of the house, and for each of questions 1 through 9.

28. Using the list of addresses from Exercise 27, draw a simple random sample of size 200 with 30% unit nonresponse rate. You will find that about 30% of the households have the information on district, household number, and assessed value, but the words "NOT AT HOME" instead of answers to questions 1 through 9. Find the means for the assessed value of the house, and for questions 1 through 9 for just the responding households. How do these compare with the results from the full simple random sample? Is there evidence of nonresponse bias?

29. Apply two-phase sampling to the nonrespondents, taking a random subsample of 30% of the nonrespondents. (Assume that all households respond to the second call.) Now estimate for the price a household is willing to pay for cable TV and the number of TV's, along with their standard errors. How do these estimates compare with those in Exercise 28?

30. Poststratify your sample from Exercise 28 using the strata you constructed in Chapter 4. Now calculate the poststratified estimates for the price a household is willing to pay for cable TV and the number of TV's. Are these closer to the values from Exercise 27? What are you assuming about the nature of the nonresponse when you use this weighting scheme? Do you think these assumptions are justified?

31. For the respondents, fit the linear regression model $y = a+bx$, where $y =$ price household is willing to pay for cable, and $x =$ assessed value of the house. Now, for the nonrespondents, impute the predicted value from this regression model for the missing y values, and use the "completed" data set to estimate the average price a household is willing to pay for cable. Compare this estimate to the previous one, and to the estimate from the full data set. Is the standard error given by your statistical package correct here? Why or why not?

32. Generate another set of data from the same address list, this time with a 30% item nonresponse rate. (The nonresponse parameters are 0, .3, 0.) What is the average price the respondents are willing to pay for cable? Using the respondents, develop a regression model for cable price based on the other variables. Impute the predicted values from this model into your missing observations and recalculate your estimate.

33. Perform another imputation on the data, this time using a sequential hot-deck procedure. Impute the value of the household immediately preceding the one with the missing item (if that one also has missing data, move up through the previous households until you find one that has the data and then impute that value). How does the value using this imputation scheme differ from the estimate in the previous Exercise?

Chapter 9 Exercises.

34. Draw a simple random sample of size 200 from Lockhart City. You may use the sample from Chapter 2 if you wish. We want to estimate $B = \bar{y}_U/\bar{x}_U$, the ratio of the price a household is willing to pay for cable TV (y) to the assessed value of the house (x). Use the linearization method to estimate the variance of $b = \bar{y}/\bar{x}$.

35. Randomly divide your sample into 10 different subsamples, each of size 20. This can be done in SAS by creating a new variable SCRAMBLE which has 200 uniform random numbers between 0 and 1. Sort the data by the variable SCRAMBLE; then assign the first 20 observations to group 1, the second 20 to group 2, etc. The group means can be easily calculated by doing a one-way

analysis of variance on the data. Now find the ratio $b_i = \bar{y}_i/\bar{x}_i$ for each group, and use the random group method to estimate the variance of b.

36. Calculate 200 different estimates $b_{(j)}$ of B, each using all but one of the 200 data points. Calculate the jackknife estimate of the variance of $b = \bar{y}/\bar{x}$.

37. How do your variance estimates from 34, 35, and 36 compare?

Chapter 10 Exercises.

38. Take a SRS of 400 households in Stephens County. Cross-classify the sample on two variables: whether the household has any children under age 12, and the number of televisions in the household (1, 2, or more than 2). Test the null hypothesis that the two variables are not associated.

39. Use your sample from the SURVEY Exercises in Chapter 5. Test the association between number of televisions (1, 2, 3 or more) and the price a household is willing to pay for cable TV (less than \$10, \$10 or more). What method did you use to account for the survey design?

Chapter 11 Exercises.

40. Use your stratified sample with optimal allocation from Exercise 12, and fit a regression model predicting the amount a household is willing to spend for cable TV from the assessed value of the house. As part of your analysis, of course, you should plot the data. Give standard errors for your parameter estimates. Does it make a difference for the parameter estimates whether you include the weights or not? Should you consider different regression models for the different strata?

41. Repeat Exercise 40, using the cluster sample from Exercise 22. What effect does the clustering have on the regression coefficients and their standard errors?

Chapter 12 Exercises.

42. Draw a simple random sample of size 500 from Lockhart City. But pretend that you do not see the price a household is willing to pay for cable TV; you only see the assessed value of the house. Use the assessed value to divide the Phase I sample into five strata of approximately equal size.

43. Draw a stratified Phase II sample with proportional allocation, of 100 observations. Estimate the average amount that households are willing to spend on cable, along with the standard error. How does the precision of this estimate compare with that of a simple random sample with the same overall cost?

44. Repeat Exercise 43, after determining the optimal allocation using results Section 12.5. Use information from the samples drawn in Chapter 3 to estimate the S_h^2, or postulate a model for them.