CHAPTER 2
ELEMENTS OF THE SAMPLING PROBLEM

2.1 An adequate frame listing individuals in a city is difficult to obtain. For that reason, and because data is desired on a family basis, it would be better to sample dwelling units. An adequate frame for dwelling units is also difficult to obtain, so a cluster sampling approach could be used by sampling city blocks and then measuring water consumption for the families living in the sampled blocks.

2.2 A common way to sample trees is to divide the farm into plots and then randomly or systematically sample plots on which trees would be counted. Unless trees are planted according to a regularly spaced design, it is difficult to use the trees themselves as sampling units.

2.3 The sampling design depends on a careful definition of the population of interest. As it would be almost impossible to get a listing of all cars owned by residents of a city, a better option would be to restrict the population of cars to something like "cars that use city parking lots on a working day" or "cars that belong to people visiting the malls on a weekend." Then, a listing of parking lots or sections of parking lots could serve as frames for collections of cars.

2.4 If the number of plants is not too large, each could serve as a stratum from which employees would be sampled. In this case one would need a list of employees (a frame) for each plant. If the number of plants is large, then a sample of plants (clusters of employees) could be taken and a sample of employees (or all employees) could be interviewed in each sampled plant.

2.5 An area as large as a state is generally broken up into smaller areas, such as counties and farms within counties, for sampling. Each county may contain a number of farms, so there are various sampling options. Counties could be viewed as strata, with farms being sampled from each. If there are many counties, one might sample counties as clusters of farms and then sample farms from each sampled county. In either of these scenarios a list of farms by county would be needed as a frame.

2.6 Most polls of this type are done by telephone using random digit dialing, The state could be stratified by regions, with dialing taking place within each of these regions. Frames may be found for selected populations by using lists of registered voters or lists of property owners, but these frames do not cover the entire population of adults. Personal interviews generally produce the highest response rate, but they are expensive and require a list of individuals from which to sample. Mailed questionnaires also require a list of individual addresses and have the lowest response rate, typically. Telephone interviews are probably the most viable choice for such a survey.

2.7 **(a)** A telephone survey would be the only way to cover the country with a well designed sampling plan in a reasonable time.

(b) If the population is defined as subscribers to the paper, then a mailed questionnaire or interviews could be used. If the population is less well defined to include all readers or potential readers, than a telephone survey with random digit dialing may have to be used.

(c) Homeowners are a well-defined group, and a sample could be contacted through either mailed questionnaires or personal interviews, although the latter would be time consuming. Telephone interviews could also be used, and random digit dialing would not be necessary.

(d) Assuming dogs are registered, it should be relatively easy to sample from the list of registered owners and obtain the survey information by either telephone or mail. If there is no lost of dog owners, this would be a difficult problem probably best solved by random digit dialing.

2.8 "Do you consider yourself a political liberal or conservative?" "Do you favor an increase in the minimum wage?" Once the political label is decided, the options for the second question become more limited. Presented in the reverse order, the respondent has more freedom on the answer to the minimum wage question.

2.9 Closed questions limit options and nuances in answers, but are easier to analyze statistically. An open question could be of the form "What is your opinion on the school tax referendum?" A closed version could be "Do you pan to vote for or against the school tax referendum that is on the ballot in the next election?" This is an extremely closed version, other options could be offered.

2.10 "Do you favor an increase in the minimum wage to keep up with inflation?" "Do you favor an increase in the minimum wage so that many wage earners who are now living below the poverty line can afford to adequately feed and clothe their families?"

2.11 The no-opinion option should be used carefully and sparingly because it gives respondents an easy way out of questions on which they may well have a deeper opinion.

2.12 "I'm sure you are aware that most standardized tests contain multiple choice questions that favor students who memorize lots of facts as compared to those who learn to think deeply. Do you favor or oppose the increased use of standardized test scores to measure academic achievement?"

2.13 After errors of non-observation and errors of observation, the next most common source of errors in surveys is the mishandling of data in the data recording and analysis part of the survey. It is imperative that the data management process contains checks to see that data are recorded correctly, and that all recorded data are part of the analysis.

2.14 The pretest, which need not be on a randomly selected sample, is the best way to discover if a questionnaire contains questions that can be answered in reliable and valid ways. It also helps define issues that should be part of the training of field workers as well as issues of data collection, management and analysis.

2.15 The response rate is strongly related to the bias in survey results. A low response rate may imply that important segments of the population (such as retired people or single people) are under-represented in the survey data and., hence, in the reported results.

2.16 More people may well beat home at that hour, but they also do not like being interrupted at mealtime. One type of nonresponse is being traded for another, perhaps.

2.18 It is very difficult to get objective information on sensitive issues. The proportion who admitted to cheating is probably well below the actual proportion who cheated. (A technique in Chapter 11 will show how to improve the accuracy of responses in these situations.)

2.19 The results may be a bit biased because students regularly here that mathematics and English are the two subjects in which they need to do well in order to succeed in life.

2.20 The response rate is low and those who are most concerned about being able to pay for a college education are the ones most likely to respond. For others, this is not an important issue and they will tend not to respond. Thus, the result could suffer from a large bias.

2.21 The population being sampled here does not represent the population of the country and the responses are voluntary, not form a randomly selected sample. The question has an inherent bias toward favoring nuclear power plants. All aspects of the survey are directed toward obtaining a highly biased result.

2.22 This is a very low response rate and the GAO was justified in questioning the results. It is quite likely that some of the income groups (especially the low incomes) were greatly underrepresented.

2.23 **(a)** One rating point represents one percent of the viewing households, or

$$95.1 \text{ million} \times 0.01 = 951{,}000 \text{ households}$$

based on the fact that the sampled population is households.

(b) As a percentage, a share is larger than a rating because the denominator of the rating is the total number of sampled households, while the denominator of a share is the total number of sampled households that actually have a TV set turned on (viewing households).

(c) 95.1 million × 0.217 = 20.64 million households could have been viewing this show

(d) Much of the data collected by Nielsen depends upon people in the sampled households either pushing a button on a People Meter or writing in a diary to record what they are watching. This is far from a fool-proof system.

2.25 **(a)**

Target Population	51%	12%	9%
High risk cities	57.9%	33.8%	20.7%
National	58.4%	13.5%	8.3%

In the national survey, the sample percentages are quite close to those reported by the Census. Thus, randomization did a good job.

In the survey of high-risk cities, the black and Hispanic percentages are much higher than those reported for the nation as a whole.

(b) High-risk cities are not the typical cities of the population. One may expect that the randomization actually did a good job here as well.

2.26 **(a)** National: 100 - 84.9 = 15.1 %
High Risk Cities**:** 100 - 80.4 = 19.6 %

(b) National: 0.151 (1000) = 151
High Risk Cities: 0.196 (1000) = 196

2.28

Care about keeping weight down

	NS	EX	FS	CS	Total
A lot	6297	3613	197	2114	12221
Somewhat	2882	1677	90	793	5442
A Little	1441	625	16	354	2436
Don't care	1709	822	22	377	2930
Total	12329	6737	325	3638	23029
% A lot	.51	.54	.61	.58	

(a) It is anticipated that responses will be more honest when questions are asked about peers rather than about the respondent himself or herself.

(b) 12,329,000

(c) 12221 / 23029 = .53

(d) 3,638,000, 2114/3638 = .58

(e) 6297 / 12221 = .51, 2114 / 12221 = .17

(f) No, at least not strongly. As you see from the bottom row of calculated percentages, the percentage of students who care a lot about keeping their weight down is fairly constant across all categories of smoking.

2.29

Care about staying away from marijuana

	NS	EX	FS	CS	Total
A lot	7213	2693	75	857	10838
Somewhat	2482	1861	109	1102	5554
A Little	744	542	27	298	1611
Don't care	1878	1550	119	1312	4859
Total	12317	6646	330	3569	22862
% A lot	.59	.41	.23	.24	

(a) 7213 / 12317 = .59

(b) 857 / 3569 = .24

(c) 7213 / 10838 = .67

(d) 1878 / 4859 =.39

(e) Yes. Non smokers care more about staying away from marijuana than current smokers (59%, 24%, respectively). Also from (a), and (b), among those who care a lot about staying away from marijuana, 59% were non smokers, while 24% was current smokers.

2.30 **(a)** These are conditional proportions, calculated as the percentages of "yes", "no" and "don't know" responses within each smoking category.

(b) Yes. Twelve percent of nonsmokers think smoking help reduce stress, while 46.5% of current smokers believe that.

(c) No. Regardless of their smoking status, teenagers believe that almost all doctors are strongly against smoking.