CHAPTER 1 ANSWERS

Exercises 1.1

1.1 (a) The *population* is the collection of all individuals or items under consideration in a statistical study.

(b) A *sample* is that part of the population from which information is obtained.

1.2 The two major types of statistics are descriptive and inferential statistics. Descriptive statistics consists of methods for organizing and summarizing information. Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

1.3 Descriptive methods are used for organizing and summarizing information and include graphs, charts, tables, averages, measures of variation, and percentiles.

1.4 Descriptive statistics are used to organize and summarize information from a sample before conducting an inferential analysis. Preliminary descriptive analysis of a sample may reveal features of the data that lead to the appropriate inferential method.

1.5 (a) An *observational study* is a study in which researchers simply observe characteristics and take measurements.

(b) A *designed experiment* is a study in which researchers impose treatments and controls and *then* observe characteristics and take measurements.

1.6 Observational studies can reveal only association, whereas designed experiments can help establish causation.

1.7 This study is inferential. Data from a sample of Americans are used to make an estimate of (or an inference about) average TV viewing time for all Americans.

1.8 This study is descriptive. It is a summary of the average salaries in professional baseball, basketball, and football for 1995 and 2005.

1.9 This study is descriptive. It is a summary of the assessment results for level of performance of all senior geography majors in 2003 and 2004 at one institution.

1.10 This study is inferential. National samples are used to make estimates of (or inferences about) drug use throughout the entire nation.

1.11 This study is descriptive. It is a summary of the annual final closing values of the Dow Jones Industrial Average at the end of December for the years 2000-2008.

1.12 This study is inferential. Monthly survey data is used to make estimates of the percentages of all music expenditures.

1.13 (a) This study is inferential. It would have been impossible to survey all U.S. adults about their opinions on Darwinism. Therefore, the data must have come from a sample. Then inferences were made about the opinions of all U.S. adults.

(b) The population consists of all U.S. adults. The sample consists only of those U.S. adults who took part in the survey.

1.14 (a) The population consists of all U.S. adults. The sample consists of the more than 500 U.S. adults who were surveyed.

(b) The percentage of 69% is a descriptive statistic since it describes the opinion of the U.S. adults who were surveyed.

1.15 (a) The statement is descriptive since it only tells what was said by the respondents of the survey.

(b) Then the statement would be inferential since the data has been used to provide an estimate of what all Americans believe.

1.16 (a) To change the study to a designed experiment, one would start with a randomly chosen group of men, then randomly divide them into two groups, an experimental group in which all of the men would have vasectomies and a control group in which the men would not have them. This would enable the researcher to make inferences about vasectomies being a cause of prostate cancer.

(b) This experiment is not feasible, since, in the vasectomy group there would be men who did not want one, and in the control group there would be men who did want one. Since no one can be forced to participate in the study, the study could not be done as planned.

1.17 Designed experiment. The researchers did not simply observe the two groups of children, but instead randomly assigned one group to receive the Salk vaccine and the other to get a placebo.

1.18 Observational study. The researchers at Harvard University and the National Institute of Aging simply observed the two groups.

1.19 Observational study. The researchers had no control over who became an elite distance runner and who did not. They simply observed the skinfold thickness of a sample of people in each group.

1.20 Designed experiment. The researchers did not simply observe the two groups of women, but instead randomly assigned one group to receive aspirin and the other to get a placebo.

1.21 Designed experiment. The researchers did not simply observe the three groups of patients, but instead randomly assigned some patients to receive optimal pharmacologic therapy, some to receive optimal pharmacologic therapy and a pacemaker, and some to receive optimal pharmacologic therapy and a pacemaker-defibrillator combination.

1.22 Observational studies. The researchers simply collected available information about the starting salaries of new college graduates.

1.23 (a) This statement is inferential since it is a statement about all Americans based on a poll. We can be reasonably sure that this is the case since the time and cost of questioning every single American on this issue would be prohibitive. Furthermore, by the time everyone could be questioned, many would have changed their minds.

(b) To make it clear that this is a descriptive statement, the new statement could be, "Of 1032 American adults surveyed, 73% favored a law that would require every gun sold in the United States to be test-fired first, so law enforcement would have its fingerprint in case it were ever used in a crime." To rephrase it as an inferential statement, use "Based on a sample of 1032 American adults, it is estimated that 73% of American adults favor a law that would require every gun sold in the United States to be test-fired first, so law enforcement would have its fingerprint in case it were ever used in a crime."

1.24 Descriptive statistics. The U.S. National Center for Health Statistics collects death certificate information from each state, so the rates shown reflect the causes of all deaths reported on death certificates, not just a sample.

1.25 (a) The figure 42800 is an inferential statistic since it is indicated in the statement that it is a projection (probably based on incomplete data for the year 2004). The data may be incomplete in part because in April, 2005, there might still be deaths to occur that are the result of traffic accidents in 2004.

(b) The figure 42643 is a descriptive statistic since it reflects the actual number of traffic deaths for the year 2003.

1.26 (a) The numbers of registered vehicles are descriptive statistics since each state would know exactly how many vehicles were registered.

(b) The vehicle miles traveled are inferential statistics since there is no way to know exactly how many miles each car traveled during a given year.

(c) While the NHTSA may have taken a sample of drivers (or vehicles) and estimated the total miles based on the sample, this estimate could be based in part on the number of vehicles registered, federal gas taxes collected, average miles per gallon estimates, the ages and types of vehicles registered, and estimates of the percentages of highway and city driving.

(d) The highway fatality rates are inferential statistics. While the number of fatalities can be known exactly and is a descriptive statistic, the number of vehicle miles is an estimate (inferential statistic), and therefore the fatality rate (which is a ratio of the two numbers) is also an estimate resulting from the inferred vehicle miles.

Exercises 1.2

1.27 A census is generally time consuming, costly, frequently impractical, and sometimes impossible.

1.28 Sampling and experimentation are two alternative ways to obtain information without conducting a complete census.

1.29 The sample should be representative so that it reflects as closely as possible the relevant characteristics of the population under consideration.

1.30 The online poll clearly has a built-in non-response bias. Since it was taken over the Memorial Day weekend, most of those who responded were people who stayed at home and had access to their computers. Most people vacationing outdoors over the weekend would not have carried their computers with them and would not have been able to respond.

1.31 Dentists form a high-income group whose incomes are not representative of the incomes of Seattle residents in general.

1.32 There are many possible answers. Surveying people regarding political candidates as they enter or leave an upscale business location, surveying the readers of a particular publication to get information about the population in general, polling college students who live in dormitories to obtain information of interest to all students are all likely to produce samples unrepresentative of the population under consideration.

1.33 (a) Probability sampling consists of using a randomizing device such as tossing a coin or consulting a random number table to decide which members of the population will constitute the sample.

(b) No. It is possible for the randomizing device to randomly produce a sample that is not representative.

(c) Probability sampling eliminates unintentional selection bias, permits the researcher to control the chance of obtaining a non-representative sample, and guarantees that the techniques of inferential statistics can be applied.

1.34 (a) Simple random sampling is a procedure for which each possible sample of a given size is equally likely to be the one obtained.

(b) A simple random sample is one that was obtained by simple random sampling.

(c) Random sampling may be done with or without replacement. In sampling with replacement, it is possible for a member of the population to be chosen more than once, i.e., members are eligible for re-selection after they have been chosen once. In sampling without replacement, population members can be selected at most once.

1.35 Simple random sampling.

1.36 One method would be to place the names of all members of the population under consideration on individual slips of paper, place the slips in a container large enough to allow them to be thoroughly shuffled by shaking or spinning, and then draw out the desired number of slips for the sample while blindfolded. A second method, which is much more practical when the population size is large, is to assign a number to each member of the population, and then use a random number table, random number generating device, or computer program to determine the numbers of those members of the population who are chosen.

1.37 (a) GLS, GLA, GLT, GSA, GST, GAT, LSA, LST, LAT, SAT.

(b) There are 10 samples, each of size three. Each sample has a one in 10 chance of being selected. Thus, the probability that a sample of three officials is the first sample on the list presented in part (a) is 1/10. The same is true for the second sample and for the tenth sample.

1.38 (a) E,M E,A M,L P,L L,A

E,P E,B M,A P,A L,B

E,L M,P M,B P,B A,B

(b) One procedure for taking a random sample of two representatives from the six is to write the initials of the representatives on six separate pieces of paper, place the six slips of paper into a box, and then, while blindfolded, pick two of the slips of paper. Or, number the representatives 1-6, and use a table of random numbers or a random-number generator to select two different numbers between 1 and 6.

(c) 1/15; 1/15

1.39 (a) E,M,P,L E,M,L,B E,P,A,B M,P,A,B

E,M,P,A E,M,A,B E,L,A,B M,L,A,B

E,M,P,B E,P,L,A M,P,L,A P,L,A,B

E,M,L,A E,P,L,B M,P,L,B

(b) One procedure for taking a random sample of four representatives from the six is to write the initials of the representatives on six separate pieces of paper, place the six slips of paper into a box, and then, while blindfolded, pick four of the slips of paper. Or, number the representatives 1-6, and use a table of random numbers or a random-number generator to select four different numbers between 1 and 6.

(c) 1/15; 1/15

1.40 (a) E,M,P E,P,A M,P,L M,A,B

E,M,L E,P,B M,P,A P,L,A

E,M,A E,L,A M,P,B P,L,B

E,M,B E,L,B M,L,A P,A,B

E,P,L E,A,B M,L,B L,A,B

(b) One procedure for taking a random sample of three representatives from the six is to write the initials of the representatives on six separate pieces of paper, place the six slips of paper into a box, and then, while blindfolded, pick three of the slips of paper. Or, number the representatives 1-6, and use a table of random numbers or a random-

number generator to select three different numbers between 1 and 6.

(c) 1/20; 1/20

1.41 (a) C,W,H C,W,V C,W,A C,H,V C,H,A
C,V,A W,H,V W,H,A W,V,A H,V,A

(b) 1/10; 1/10

1.42 (a) 1/45

(b) 1/252

(c) First assign the digits 0 though 9 to the ten cities as listed in the exercise. Select a random starting point in Table I of Appendix A and read in a pre-selected direction until you have encountered 5 different digits. For example, if we start at the top of the fifth column of digits and read down, we encounter the digits 4,1,5,2,5,6. We ignore the second '5'. Thus our sample of five cities consists of Osaka, Tokyo, Miami, San Francisco, and New York. Your answer may be different from this one.

(d) We can use Excel to generate random numbers. If we enter the expression =RAND() in a cell, Excel returns a random number between 0 and 1. If we multiply this number by 10, we will get a number between 0 and 10, and if we take the integer part of that result, we will get one of the digits 0 through 9. For example in cell A1, we enter the expression =INT(10*RAND()). One of digits 0 through 9 will result. We copy the content of A1 to the clipboard and paste it into cells A2 through A10. Then we choose the first 5 unique digits for our sample. Our result is 3, 8, 6, 3, 5, 3, 9, 6, 5, 2. The first 5 unique digits are 3, 8, 6, 5, 9. Thus our sample of 5 cities is Los Angeles, Manila, New York, Miami, and London. Your result may be different from ours.

1.43 (a) I am using Table I to obtain a list of 10 random numbers between 1 and 500 as follows.

I start at the three digit number in line number 14 and column numbers 10-12, which is the number 452.

I now go down the table and record the three-digit numbers appearing directly beneath 452. Since I want numbers between 1 and 500 only, I throw out numbers between 501 and 999, inclusive. I also discard the number 000.

After 452, I skip 667, 964, 593, 534, and record 016.

Now that I've reached the bottom of the table, I move directly rightward to the adjacent column of three-digit numbers and go up.

I record 343, 242, skip 748, 755, record 428, skip 852, 794, 596, record 378, skip 890, record 163, skip 892, 847, 815, 729, 911, 745, record 182, 293, and 422.

I've finished recording the 10 random numbers. In summary, these are:

452 016 343 242 428
378 163 182 293 422

(b) We can use Excel to generate random numbers. If we enter the expression =RAND() in a cell, Excel returns a random number between 0 and 1. If we multiply this number by 500, we will get a number between 0 and 500, and if we take the integer part of that result, we will get one of the integers 0 through 499. If we add 1 to this result, we will get a random number between 1 and 500. For example, in cell A1, we enter the expression =INT(500*RAND())+1. One of integers 1 through 500 will result. We copy the content of A1 to the clipboard and paste it into cells A2 through A30. Then we choose the first 10 unique three-digit numbers for our sample. Our result is 489, 451, 61, 114, 389, 381, 364, 166, 221, 437, 266, 46, 422, 388, 401, 387, 276, 248, 21,

198. The first 10 unique three-digit numbers are 489, 451, 61, 114, 389, 381, 364, 166, 221, and 437. Your result may be different from ours. [Note: Each time the ENTER key is depressed, the random numbers will be recalculated, so make sure that you do not press the ENTER key until you have recorded all of your random numbers.]

1.44 (a) I am using Table I to obtain a list of 20 different random numbers between 1 and 80 as follows.

I start at the two digit number in line number 5 and column numbers 31-32, which is the number 86. Since I want numbers between 1 and 80 only, I throw out numbers between 81 and 99, inclusive. I also discard the number 00.

I now go down the table and record the two-digit numbers appearing directly beneath 86.

After skipping 86, I record 39, 03, skip 97, record 28, 58, 59, skip 81, record 09, 36, skip 81, record 52, skip 94, record 24 and 78.

Now that I've reached the bottom of the table, I move directly rightward to the adjacent column of two-digit numbers and go up.

I skip 84, record 57, 40, skip 89, record 69, 25, skip 95, record 51, 20, 42, 77, skip 89, skip 40(duplicate), record 14, and 34.

I've finished recording the 20 random numbers. In summary, these are

39	03	28	58	59
09	36	52	24	78
57	40	69	25	51
20	42	77	14	34

(b) We can use Excel to generate random numbers. If we enter the expression =RAND() in a cell, Excel returns a random number between 0 and 1. If we multiply this number by 80, we will get a number between 0 and 80, and if we take the integer part of that result, we will get one of the integers 0 through 79. If we add 1 to this result, we will get a random number between 1 and 80. For example, in cell A1, we enter the expression =INT(80*RAND())+1. One of integers 1 through 80 will result. We copy the content of A1 to the clipboard and paste it into cells A2 through A30. Then we choose the first 20 unique two-digits numbers for our sample. Our result is 55, 47, 66, 66, 2, 72, 56, 10, 31, 5, 55, 19, 39, 57, 44, 57, 60, 23, 34, 55, 43, 9, 49, 62, 47, 15, 32, 38, 74, 10. The first 20 unique two-digit numbers are 55, 47, 66, 2, 72, 56, 10, 31, 5, 19, 39, 57, 44, 60, 23, 34, 43, 9, 49, and 62. Your result may be different from ours. [Note: Each time the ENTER key is depressed, the random numbers will be recalculated, so make sure that you do not press the ENTER key until you have recorded all of your random numbers.]

1.45 (a) The possible samples of size one are G L S A T

(b) There is no difference between obtaining a sample of size one and selecting one official at random.

1.46 (a) The only possible sample of size five is GLSAT.

(b) There is no difference between obtaining a sample of size five and taking a census.

1.47 No. Only adults with access to the Internet would have been able to respond to the survey. Thus, not every adult had an equal chance of being chosen for the sample.

1.48 (a) In Exercise 1.43(b), we wanted 10 random numbers between 1 and 500 inclusive, so we take m = 1 and n = 500. Using a random number generator to generate a random number r between 0 and 1, we then

calculate the expression 1 + (500 − 1 + 1)r or 1 + 500r and round it down to the nearest integer to get a random number between 1 and 500. Then repeat this process until 10 different random numbers have been generated.

(b) In Exercise 1.44(b), we wanted 20 different random numbers between 1 and 80 inclusive, so we take m = 1 and n = 80. Using a random number generator to generate a random number r between 0 and 1, we then calculate the expression 1 + (80 − 1 + 1)r or 1 + 80r and round it down to the nearest integer to get a random number between 1 and 80. Then repeat this process until 20 different random numbers have been generated.

Exercises 1.3

1.49 (a) Answers will vary, but here is the procedure: (1) Divide the population size, 500, by the sample size, 10, and round down to the nearest whole number if necessary; this gives 50. (2) Use a table of random numbers (or a similar device) to select a number between 1 and 50, call it k. (3) List every 50th number, starting with k, until 10 numbers are obtained; thus, the first number on the required list of 10 numbers is k, the second is k+50, the third is k+100, and so forth (e.g., if k=6, then the numbers on the list are 6, 56, 106, ...).

(b) Systematic random sampling is easier.

(c) The answer depends on the purpose of the sampling. If the purpose of sampling is not related to the size of the sales outside the U.S., systematic sampling will work. However, since the listing is a ranking by amount of sales, if k is low (say 2), then the sample will contain firms that, on the average, have higher sales outside the U.S. than the population as a whole. If the k is high, (say 49) then the sample will contain firms that, on the average, have lower sales than the population as a whole. In either of those cases, the sample would not be representative of the population in regard to the amount of sales outside the U.S.

1.50 (a) Answers will vary, but here is the procedure: (1) Divide the population size, 80, by the sample size, 20, and round down to the nearest whole number if necessary; this gives 4. (2) Use a table of random numbers (or a similar device) to select a number between 1 and 4, call it k. (3) List every 4th number, starting with k, until 20 numbers are obtained; thus the first number on the required list of 20 numbers is k, the second is k+4, the third is k+8, and so forth (e.g., if k=3, then the numbers on the list are 3, 7, 11, 15, ...).

(b) Systematic random sampling is easier.

(c) No. In Keno, you want every set of 20 balls to have the same chance of being chosen. Systematic sampling would give each of 4 sets of balls [(1, 5, 9,...,77), (2, 6, 10,...,78), (3, 7, 11,...,79) and (4, 8, 12,...,80)], a 1/4 chance of occurring, while all of the other possible sets of balls would have no chance of occurring.

1.51 (a) Number the suites from 1 to 48, use a table of random numbers to randomly select three of the 48 suites, and take as the sample the 24 dormitory residents living in the three suites obtained.

(b) Probably not, since friends are more likely to have similar opinions than are strangers.

(c) There are 384 students in total. Freshmen make up 1/3 of them. Sophomores make up 7/24 of them, Juniors 1/4, and Seniors 1/8. Multiplying each of these fractions by 24 yields the proportional allocation, which dictates that the number of freshmen, sophomores, juniors, and seniors selected should be, respectively, 8, 7, 6, and 3.

Thus a stratified sample of 24 dormitory residents can be obtained as follows: Number the freshmen dormitory residents from 1 to 128 and use a table of random numbers to randomly select 8 of the 128 freshman dormitory residents; number the sophomore dormitory residents from 1 to 112 and use a table of random numbers to randomly select 7 of the 112 sophomore dormitory residents; and so forth.

1.52 (a) Each category of "Percent free lunch" should be represented in the sample in the same proportion that it is present in the population of top 100 ranked high schools. Thus 50/100 of the sample of 25 schools should be from the 0 to under 10% free lunch category, 18/100 from the second category, 11/100 from the third, 8/100 from the fourth, and 13/100 from the last. Multiplying each of these fractions by 25 gives us the sample sizes from each category. These sample sizes will not necessarily be integers, so we will need to make some minor adjustments of the results. The first category should have (50/100)(25) = 12.5. The second should have (18/100)(25) = 4.5. Similarly, the third, fourth, and fifth categories should have 2.75, 2, and 3.25 for their sample sizes. We round the third and fifth sample sizes each to 3. After flipping a coin, we round the first two categories to 12 and 5. Thus the sample sizes for the five Percent free lunch categories should be 12, 5, 3, 2, and 3 respectively. We would now use a random number generator to select 12 out of the 50 in the first category, 5 out of the 18 in the second, 3 out of the 11 in the third, 2 of the 8 in the fourth, and 3 of the 13 in the last category.

(b) From part (a), two schools would be selected from the strata with a percent free lunch value of 30-under 40.

1.53 Stratified Sampling. The entire population is naturally divided into subpopulations, one from each lake, and random sampling is done from each lake. The stratified sampling is not with proportional allocation since that would require knowing how many fish were in each lake.

1.54 (a) In probability sampling, a random number device is used to determine which members of the population will constitute the sample from the population. Since this poll was taken from an online sample, the only people who could respond to the survey were those with Internet access. Others could not be chosen at all. More likely, the responders were also self-chosen, not randomly selected from all Internet users.

(b) It is not systematic random sampling. That would require an ordered list of some type (alphabetical, ID number, etc.) and then only those chosen would receive a notice to go online to complete a survey form. Even then, there is no assurance that the person receiving the notice would be the one to complete the survey. It is not cluster sampling for similar reasons. The Harris Poll would need to have information enabling them determine to what cluster every Internet user belonged and then they would have to sample everyone from that cluster. Similarly for stratified sampling and multistage sampling. They would need to have information about each user to enable them to determine what stratum each user was in or some way to know whom to sample at the next stage. Since often there are multiple users of a single computer, there could be no assurance that the poll was actually sampling the people who were targeted to be sampled.

1.55 (a) This is a poll taken by calling randomly selected U.S. adults. Thus, the sampling design appears to be simple random sampling, although it is possible that a more complex design was used to ensure that various political, religious, educational, or other types of groups were proportionately represented in the sample.

(b) The sample size for the second question was 78% of 1010 or 788.

(c) The sample size for the third question was 28% of 788 or 221.

1.56 No. In your text, Example 1.9, only 48 different samples are possible. A

sample containing students 5,6, and 7 is not possible at all. While the 48 possible samples are equally likely, there are other samples that could be obtained through simple random sampling that are not possible at all in systematic sampling. Thus not all possible samples are equally likely. Nevertheless, if there is no pattern or cycle to the data, this method will tend to give about the same results as simple random sampling.

1.57 (a) It is also true for systematic random sampling if the population size divided by the sample size results in an integer for m. The chance for each member to be selected is then still equal to the sample size divided by the population size. For example, suppose the population size is N=10 and the sample size is n=2. The chance that each member in simple random sampling to be selected is 2/10 = 1/5. In systematic random sampling for the same example, m=5. The possible samples of size two are 1 and 6, 2 and 7, 3 and 8, 4 and 9, and 5 and 10. Therefore, the chance that a member is selected is equal to the chance of one of those five samples being selected, which is the same as simple random sampling of 1/5.

(b) It is not true for systematic random sampling if the population size divided by the sample size does not result in an integer for m. For example, suppose the population size is N=15 and the sample size is n=2. After dividing the population size by the sample size and rounding down to the nearest whole number, we get m=7. You would select every 7^{th} member after a random starting place k, between 1 and 7, is determined. If k=1, you would select the first and eighth member. If k=7, you would select the seventh and fourteenth member. In this situation, the last member (fifteenth) can never be selected. Therefore, the last member of the sample does not have the same chance of being selected as any other member in the population.

1.58 Refer to example 1.11. If we approached this problem as a simple random sample each member would have a chance of being selected equal to the sample size divided by the population size: 20/250, or 2/25.

If we approached this same example as a stratified sample with proportional allocation, we would select 2 out of 25 households in the upper income group, 14 out of the 175 households in the middle income group, and 4 out of 50 households in the lower income group. Thus the chance that an upper income household is selected is 2/25. The chance that a middle income household is selected is 14/175 = 2/25. Finally, the chance that a lower income household is selected is 4/50 = 2/25. Thus, the chance that each member is selected is the same as a simple random sample.

1.59 From the information about the sample, we can conclude that the population of interest consists of all adults in the continental U.S. The sample size was 2010 except that for questions about politics, only registered voters were considered part of the sample. The sample size for those questions was 1637.

The overall procedure for drawing the sample was multistage (actually, three stages were used) sampling: the first stage was to randomly select 520 geographic points in the continental U.S.; then proportional sampling was used to randomly sample a number of households with telephones from each of the 520 regions in proportion to its population; finally, once each household was selected, a randomizing procedure was used to ensure that the correct numbers of adult male and female respondents were included in the sample.

The last paragraph indicates the confidence that the poll-takers had in the results of the survey, that is, that there is a 95% chance that the sample results will not differ by more than 2.2 percentage points in either direction from the true percentage that would have been obtained by surveying all adults in the actual population, or by more than 2.5 percentage points in either direction from the true percentage that would have been obtained by surveying all registered voters in the population.

The last sentence says that smaller samples have a larger margin of error,@ an explanation for the difference in the maximum percentage points of error for all adults and for registered voters.

Exercises 1.4

1.60 The three basic principles of experimental design are control, randomization, and replication.

Control: Two or more treatments should be compared.

Randomization: The experimental units should be randomly divided into groups to avoid unintentional selection bias in constituting the groups.

Replication: A sufficient number of experimental units should be used to ensure that randomization creates groups that resemble each other closely and to increase the chances of detecting differences among the treatments.

1.61 (a) Experimental units are the individuals or items on which the experiment is performed.

(b) When the experimental units are humans, we call them subjects.

1.62 (a) The treatment group consisted of the 2444 patients who took Prozac.

(b) The control group consisted of the 1331 patients who received a placebo.

(c) The treatments were administering Prozac and administering the placebo.

1.63 (a) There were three treatments.

(b) The first group, the one receiving only the pharmacologic therapy, would be considered the control group.

(c) There were three treatment groups. The first received only pharmacologic therapy, the second received pharmacologic therapy plus a pacemaker, and the third received pharmacologic therapy plus a pacemaker-defibrillator combination.

(d) The first group (control) contained 1/5 of the 1520 patients or 304. The other two groups each contained 2/5 of the 1520 patients or 608.

(e) Each patient could be randomly assigned a number from 1 to 1520. Any patient assigned a number between 1 and 304 would be assigned to the control group; any patient assigned to the next 608 numbers (305 to 912) would be assigned to receive the pharmacologic therapy plus a pacemaker; and any patient assigned a number between 913 and 1520 would receive pharmacologic therapy plus a pacemaker-defibrillator combination. Each random number would be used only once to ensure that the resulting treatment groups were of the intended sizes.

1.64 (a) Experimental units: the perishable items in the study

(b) Response variable: a measure of the deterioration of the items

(c) Factors: two factors - storage time and storage temperature

(d) Levels of each factor: three storage temperatures and five storage times

(e) Treatments: the fifteen different combinations of storage temperature and storage time resulting from testing deterioration at each of the three storage temperature for each of the five storage times

1.65 (a) Experimental units: batches of the product being sold

(b) Response variable: the number of units of the product sold

(c) Factors: two factors - display type and pricing scheme

(d) Levels of each factor: three types of display of the product and three pricing schemes

(e) Treatments: the nine different combinations of display type and price resulting from testing each of the three pricing schemes with each of the three display types

1.66 (a) Experimental units: fields of oats

(b) Response variable: crop yield of the oats per acre

(c) Factors: variety of oats and concentration of manure on the fields

(d) Levels of each factor: three varieties of oats and four concentrations of manure

(e) Treatments: the twelve combinations of oat variety and manure concentration resulting from testing each of the three oat varieties with each of the four concentration levels of the manure

1.67 (a) Experimental units: female lions

(b) Response variable: whether or not the female lion approached the male lion dummy

(c) Factors: length and color of the mane on the male lion dummy

(d) Levels of each factor: two different mane lengths and two different mane colors

(e) Treatments: the four combinations of mane length and color

1.68 (a) This is a completely randomized design since the flashlights were randomly assigned to the different battery brands.

(b) This is a randomized block design since the four different battery brands would be randomly assigned within each set of four flashlights from each of the five flashlight brands.

1.69 Double-blinding guards against bias, both in the evaluations and in the responses. In the Salk vaccine experiment, double-blinding prevented a doctor's evaluation from being influenced by knowing which treatment (vaccine or placebo) a patient received; it also prevented a patient's response to the treatment from being influenced by knowing which treatment he or she received.

1.70 (a) Simple random sampling corresponds to completely randomized designs since selection is randomly made from the entire population.

(b) Stratified sampling corresponds to randomized block designs since selection is randomly made from within each strata.

Review Problems for Chapter 1

1. Student exercise.

2. Descriptive statistics are used to display and summarize the data to be used in an inferential study. Preliminary descriptive analysis of a sample often reveals features of the data that lead to the choice or reconsideration of the choice of the appropriate inferential analysis procedure.

3. Descriptive study. The scores are merely reported.

4. Descriptive study. The paragraph describes the results of 1,020 respondents surveyed.

5. Inferential study. The results of a sample are used to make inferences about the age distribution of all British backpackers in South Africa.

6. (a) Since 18% reported that they had abused Vicodin, this figure applies to the sample and is therefore descriptive.

(b) The 4.3 million youths abusing Vicodin is clearly inferential since it applies to the entire population, not just the sample.

7. (a) An *observational study* is a study in which researchers simply observe

characteristics and take measurements.

(b) A *designed experiment* is a study in which researchers impose treatments and controls and *then* observe characteristics and take measurements.

8. This is an observational study. To be a designed experiment, the researchers would have to have the ability to assign some children at random to live in persistent poverty during the first 5 years of life or to not suffer any poverty during that period. Clearly that is not possible.

9. This is a designed experiment since the researcher is imposing a treatment and then observing the results.

10. A literature search should be made before planning and conducting a study.

11. (a) A representative sample is one that reflects as closely as possible the relevant characteristics of the population under consideration.

(b) Probability sampling involves the use of a randomizing device such as tossing a coin or die, using a random number table, or using computer software that generates random numbers to determine which members of the population will make up the sample.

(c) A sample is a simple random sample if all possible samples of a given size are equally likely to be the actual sample selected.

12. Because Yale is a very expensive school, incomes of parents of Yale students will not be representative of the incomes of all college students' parents.

13. (a) This method does not involve probability sampling. No randomizing device is being used and people who do not visit the campus cafeteria have no chance of being included in the sample.

(b) The dart throwing is a randomizing device that makes all samples of size 20 equally likely. This is probability sampling.

14. (a) H,P,S H,P,A H,P,E H,S,A H,S,E

H,A,E P,S,A P,S,E P,A,E S,A,E

(b) Since each of the 10 samples of size three is equally likely, there is a 1/10 chance that the sample chosen is the first sample in the list, 1/10 chance that it is the second sample in the list, and 1/10 chance that it is the tenth sample in the list.

(c) (i) Make five slips of paper with each airline on one slip. Draw three slips at random. (ii) Make 10 slips of paper, each having one of the combinations in part (a). Draw one slip at random. (iii) Number the five airlines from 1 to 5. Use a random number table or random number generator to obtain three distinct random numbers between 1 and 5 inclusive.

(d) Your method and result may differ from ours. We rolled a die (ignoring 6's and duplicates) and got 2, 5, 2, 6, 4. So our sample consists of Pinnacle Airlines, Atlantic Southeast Airlines, and Alaska Airlines.

15. (a) Table I can be employed to obtain a sample of 15 random numbers between 1 and 100 as follows. First, I pick a random starting point by closing my eyes and putting my finger down on the table.

My finger falls on three digits located at the intersection of a line with three columns. (Notice that the first column of digits is labeled "00" rather than "01".) This is my starting point.

I now go down the table and record all three-digit numbers appearing directly beneath the first three-digit number that are between 001 and 100 inclusive. I throw out numbers between 101 and 999, inclusive. I also discard the number 0000. When the bottom of the column is reached, I move over to the next sequence of three digits and work my way back up the table. Continue in this manner. When 10 distinct

three-digit numbers have been recorded, the sample is complete.

(b) Starting in row 10, columns 7-9, we skip 484, 797, record 082, skip 586, 653, 452, 552, 155, record 008, skip 765, move to the right and record 016, skip 534, 593, 964, 667, 452, 432, 594, 950, 670, record 001, skip 581, 577, 408, 948, 807, 862, 407, record 047, skip 977, move to the right, skip 422 and all of the rest of the numbers in that column, move to the right, skip 732, 192, record 094, skip 615 and all of the rest of the numbers in that column, move to the right, record 097, skip 673, record 074, skip 469, 822, record 052, skip 397, 468, 741, 566, 470, record 076, 098, skip 883, 378, 154, 102, record 003, skip 802, 841, move to the right, skip 243, 198, 411, record 089, skip 701, 305, 638, 654, record 041, skip 753, 790, record 063.

The final list of numbers is 082, 008, 016, 001, 047, 094, 097, 074, 052, 076, 098, 003, 089, 041, 063.

(c) Using Excel, we enter the expression =INT(100*RAND())+1 in cell A1, copy the content of A1 to the clipboard and paste it into cells A2 to A20. Then we use the first 15 unique numbers as our random sample. Our results were the numbers 46, 99, 90, 31, 75, 98, 79, 14, 44, 13, 66, 49, 37, 87, 73, 26, 61, 71, 72, 2. Thus our sample consists of the first 15 numbers 46, 99, 90, 31, 75, 98, 79, 14, 44, 13, 66, 49, 37, 87, 73. Your sample may be different.

16. (a) Systematic random sampling is done by first dividing the population size by the sample size and rounding the result down to the next integer, say m. Then we select one random number, say k, between 1 and m inclusive. That number will be the first member of the sample. The remaining members of sample will be those numbered k+m, k+2m, k+3m, ... until a sample of size n has been chosen. Systematic sampling will yield results similar to simple random sampling as long as there is nothing systematic about the way the members of the population were assigned their numbers.

(b) In cluster sampling, clusters of the population (such as blocks, precincts, wards, etc.) are chosen at random from all such possible clusters. Then every member of the population lying within the chosen clusters is sampled. This method of sampling is particularly convenient when members of the population are widely scattered and is most appropriate when the members of each cluster are representative of the entire population. Cluster sampling can save both time and expense in doing the survey, but can yield misleading results if individual clusters are made up of subjects with very similar views on the topic being surveyed.

(c) In stratified random sampling with proportional allocation, the population is first divided into subpopulations, called strata, and simple random sampling is done within each stratum. Proportional allocation means that the size of the sample from each stratum is proportional to the size of the population in that stratum. This type of sampling may improve the accuracy of the survey by ensuring that those in each stratum are more proportionately represented than would be the case with cluster sampling or even simple random sampling. Ideally, the members of each stratum should be homogeneous relative to the characteristic under consideration. If they are not homogeneous within each stratum, simple random sampling would work just as well.

17. (a) Answers will vary, but here is the procedure: (1) Divide the population size, 100, by the sample size 15, and round down to the nearest whole number; this gives 6. (2) Use a table of random numbers (or a similar device) to select a number between 1 and 6, call it k. (3) List every 6th number, starting with k, until 15 numbers are obtained; thus the first number on the required list of 15 numbers is k, the second is k+6, the third is k+12, and so forth (e.g., if k=4,

then the numbers on the list are 4, 10, 16, ...).

(b) Yes, unless for some reason there is some kind of trend or a cyclical pattern in the listing of the athletes.

18. (a) The number of full professors should be (205/820) x 40 = 10. Similarly, proportional allocation dictates that 16 associate professors, 12 assistant professors, and 2 instructors be selected.

(b) The procedure is as follows: Number the full professors from 1 to 205, and use Table I to randomly select 10 of the 205 full professors; number the associate professors from 1 to 328, and use Table I to randomly select 16 of the 328 associate professors; and so on.

19. The statement under the vote is a disclaimer as to the validity of the survey. Since the vote reflects only the responses of volunteers who chose to vote, it can not be regarded as representative of the public in general, some of whom do not use the Internet, nor as representative of Internet users since the sample was not chosen at random from either group.

20. (a) This is a designed experiment.

(b) The treatment group consists of the 158 patients who took AVONEX. The control group consists of the 143 patients who were given a placebo. The treatments were the AVONEX and the placebo.

21. The three basic principles of experimental design are control, randomization, and replication. Control refers to methods for controlling factors other than those of primary interest. Randomization means randomly dividing the subjects into groups in order to avoid unintentional selection bias in constituting the groups. Replication means using enough experimental units or subjects so that groups resemble each other closely and so that there is a good chance of detecting differences among the treatments when such differences actually exist.

22. (a) Experimental units: tomato plants

(b) Response variable: yield of tomatoes

(c) Factor(s): tomato variety and density of plants

(d) Levels of each factor: These are not given, but tomato varieties tested would be the levels of variety and the different densities of plants would be the levels of density.

(e) Treatments: Each treatment would be one of the combinations of a variety planted at a given plant density.

23. (a) Experimental Units: The children

(b) Response variable: Whether or not the child was able to open the bottle

(c) Factors: The container designs

(d) Levels of each factor: Three (types of containers)

(e) Treatments: The container designs

24. This is a completely randomized design. All of the experimental units (batches of doughnuts) were assigned at random to the four treatments (four different fats).

25. (a) This is a completely randomized design since the 24 cars were randomly assigned to the 4 brands of gasoline.

(b) This is a randomized block design. The four different gasoline brands are randomly assigned to the four cars in each of the six car model groups. The blocks are the six groups of four identical cars each.

(c) If the purpose is to learn about the mileage rating of one particular

car model with each of the four gasoline brands, then the completely randomized design is appropriate. But if the purpose is to learn about the performance of the gasoline across a variety of cars (and this seems more reasonable), then the randomized block design is more appropriate and will allow the researcher to determine the effect of car model as well as of gasoline type on the mileage obtained.

26. The explanation informs the reader that only random sampling of a population can be used to draw valid inferences about the population as a whole.

27. The data in this study were clearly not collected via a controlled experiment in which some participants were forced to do crossword puzzles, practice musical instruments, play board games, or read while others were not allowed to do any of those activities. Therefore, any data relative to these activities and dementia arose as a result of observing whether or not the subjects in the study carried out any of those activities and whether or no they had some form of dementia. Since this would be an observational study, no statement of cause and effect can rightfully be made. It cannot be claimed as result of the study that "Crosswords Reduce Risk of Dementia."

28. The researchers did not impose or manipulate any of the conditions of this study. They didn't decide who had cancer, who didn't have cancer, who had hepatitis B, or who had hepatitis C. This study was an observational study and not a controlled experiment. Observational studies can only reveal an association, not causation. Therefore, the statement in quotes is valid. If the researchers wanted to establish causation, they would need a designed experiment.

29. From the information about the sample, we can conclude that the population of interest consists of all adults in the continental U.S. The sample size was 895.

The overall procedure for drawing the sample was multistage. The first stage was to randomly select telephone exchanges (area codes) from a population of 69,000 possible telephone exchanges. Proportional sampling was used to randomly sample these exchanges from regions proportional to the population; then, randomization was used to complete the phone numbers and select a person from each household; finally, randomization was used to call cell phone numbers as well.

The next paragraph indicates the confidence that the poll-takers had in the results of the survey, that is, that there is a 95% chance that the sample results will not differ by more than 3 percentage points in either direction from the true percentage that would have been obtained by surveying all adults in the actual population.

Case Study: Greatest American Screen Legends

(a) The population of interest in the AFI survey is from the American film community.

(b) The sample is the 1800 leaders from the American film community **interviewed**.

(c) No. The population of all American moviegoers includes many people who are not in the American film community. Furthermore, the American film community has a very specialized interest and possibly a different viewpoint as to what constitutes a great actor or actress than many others in the American movie-going population.

(d) Descriptive. It merely describes the opinion of those in the sample without trying to draw an inference about the opinions of all moviegoers.

(e) Inferential. This statement would be an attempt to draw an inference about the opinion of all artists, historians, critics, and other cultural dignitaries based on the opinions of those 1800 people who were interviewed.