

CHAPTER 3

PRODUCING DATA

SECTION 3.1

OVERVIEW

Chapters 1 and 2 describe methods for exploring data. Such **exploratory data analysis** is used to determine what the data tell us about the variables measured and their relations to each other. Conclusions apply to the data observed and may not generalize beyond these data.

Statistical inference produces answers to specific questions, along with a statement of how confident we are that the answer is correct. Answers are usually intended to apply beyond the data observed. This requires careful **production of data** appropriate for answering the specific questions asked.

Data can be produced in many ways. **Anecdotal data** based on a few isolated cases is usually unreliable. **Available data** collected for other purposes, such as data produced by government agencies, can be helpful but again is not always reliable. **Sampling** selects a part of a population of interest to represent the whole. Done properly, sampling can yield reliable information about a population. **Observational studies** are investigations in which one simply observes the state of some population, usually with data collected by sampling. Even with proper sampling, data from observational studies are generally not appropriate for investigating cause-and-effect relations between variables. **Experiments** are investigations in which data are generated by

2 Chapter 3

active imposition of some treatment on the subjects of the experiment. Properly designed experiments are the best way to investigate cause-and-effect relations between variables.

GUIDED SOLUTIONS

Exercise 3.1

KEY CONCEPTS - anecdotal data

Is this data the result of an experiment? Can there be other explanations for what has occurred?

Exercise 3.5

KEY CONCEPTS - explanatory and response variables, experiments

Remember, experiments are investigations in which data are generated by *active* imposition of some treatment on the subjects. Was that done in this example?

To identify the explanatory and response variables, think about what the experimenter is trying to demonstrate with this study and what is going to be measured.

COMPLETE SOLUTIONS

Exercise 3.1

4 Chapter 3

An isolated, anecdotal case is not a good basis for drawing general conclusions. Although it seems that two acquaintances is a high number to develop brain tumors, we do not have any basis for determining whether this is related to their use of cell phones. Possibly there is another environmental factor related to the incidence of brain tumors that her friends have been exposed to, or there may not be any connection whatsoever between these two incidents of brain tumors.

Exercise 3.5

This is an experiment as a treatment is imposed on the students. The explanatory variable is the teaching method used (standard or computer assisted). The response variable is the increase in knowledge of cell biology as measured by the increase in test score.

SECTION 3.2

OVERVIEW

Experiments are studies in which one or more **treatments** are imposed on experimental **units** or **subjects**. A treatment is a combination of levels of the explanatory variables, called **factors**. The **design** of an experiment is a specification of the treatments to be used and the manner in which units or subjects are assigned to these treatments. The basic features of well-designed experiments are **control**, **randomization**, and **replication**.

Control is used to avoid confounding (mixing up) the effects of treatments with other influences such as lurking variables. One such lurking variable is the **placebo effect**, which is the response of a subject to the fact of receiving any treatment. The simplest form of control is **comparative experimentation** which involve comparisons between two or more treatments. One of these treatments may be a **placebo** (fake treatment), and those subjects receiving the placebo are referred to as a **control group**.

Randomization uses a well-defined chance mechanism to assign subjects to treatments. It is used to create treatment groups which are similar, except for chance variation, prior to application of treatments. Randomized, comparative experiments are used to prevent **bias**, or systematic favoritism of certain outcomes. **Tables of random digits** or computer programs that generate random numbers are well-defined chance mechanisms that are used to carry out randomization. In either case, numerical labels are assigned to experimental units and random numbers from the table or computer software determine which labels (units) are assigned to which treatments.

Replication is the use of many units in an experiment and is used to reduce the effect of any chance variation between treatment groups arising from randomization. Replication increases the sensitivity of an experiment to differences in treatments.

Additional control in an experiment can be achieved by forming experimental units into **blocks** that are similar in some way which is thought to

6 Chapter 3

affect the response. In a **block design**, units are first formed into blocks and then randomization is carried out separately in each block. **Matched pairs** are a simple form of blocking used to compare two treatments. In a matched pairs experiment either the same unit (the block) receives both treatments in a random order or very similar units are matched in pairs (the blocks). In the latter case,

one member of the pair receives one of the treatments and the other member the remaining treatment. Members of a matched pair are assigned to treatments using randomization.

Good experiments require attention to details. **Double-blind** experiments are ones in which neither the subject nor the person measuring the response is aware of what treatment is being used. **Lack of realism** in an experiment can prevent us from generalizing the results.

GUIDED SOLUTIONS

Exercise 3.11

KEY CONCEPTS - identifying experimental units or subjects, factors, treatments, and response variables

You need to read the description of the study carefully. To identify the experimental units, ask yourself, exactly on what were the experimental conditions applied?

To identify the factors, ask yourself what question did the experiment wish to answer? What variables does the description say the answer depends on? These are the factors.

To identify the treatments, what combinations of values of the factors were actually used in the experiment? These are the treatments.

To identify the response variables, ask yourself what was measured on the subjects after exposure to the treatments? This is the response variable.

Exercise 3.13

KEY CONCEPTS - design of an experiment, randomization

8 Chapter 3

To begin, identify the subjects, the factors, the treatments, and the response variable. Now outline your design. Be sure to specify

- How many treatments are there, hence how many groups of subjects must you form?

- How will you assign subjects to treatment groups?
- What are the treatments, i.e., what will each subject be required to do?
- What response will you measure and how will you decide if the treatments differ in their effect?

You can outline your design in words or with a picture.

The list of names has been reproduced below. Assign a numerical label to each. Be sure to use the same number of digits for each label.

Acosta	Farouk	Liang	Solomon
Asihiro	Fleming	Maldonado	Trujillo
Bennett	George	Marsden	Tulloch
Bikalis	Han	Montoya	Valasco
Chen	Howard	O'Brian	Vaughn
Clemente	Hruska	Ogle	Wei
Duncan	Imrani	Padilla	Wilder
Durr	James	Plochman	Willis
Edwards	Kaplan	Rosen	Zhang

Now start reading line 130 in Table B. Read across the row in groups of digits equal to the number of digits you used for your labels (for example, if you used two digits for labels, read line 130 in pairs of digits). You will need to keep reading until you have selected all the names for the first treatment. This may require you to continue on to line 131, line 132, and subsequent lines. After you have selected the names for treatment 1, continue in Table B to assign the

10 Chapter 3

nine people to receive treatment 2 and then nine to receive treatment 3. The remaining names are assigned to treatment 4.

Exercise 3.29**KEY CONCEPTS** - matched pairs design, randomization

The first thing you should do is identify the subjects, the factor, the treatments, and the response variable. Next, decide what are the matched pairs in this experiment. How will you use a coin flip to assign members of a pair to the treatments? What will you measure and how will you decide whether the right-hand tends to be stronger in right-handed people?

Exercise 3.32**KEY CONCEPTS** - block designs

a) To assist you, we have arranged the subjects and their excess weight in order of increasing excess weight. Now decide which are the five blocks. (Due to ties in the excess weights, the choice of blocks may not be unique).

Williams 22	Santiago 27	Brunk 30	Jackson 33	Birnbaum 35
Festinger 24	Mann 28	Obrach 30	Stall 33	Tran 35
Hernandez 25	Smith 29	Rodriguez 30	Brown 34	Nevesky 39
Moses 25	Kendall 30	Loren 32	Dixon 34	Wilansky 42

b) Label the names in each block (you should need only the labels 1, 2, 3, 4 in each block since there are only four people in a block), choose a line in Table B, start reading from left to right, assigning each member of a given block to one of the four treatments.

Write your results below, as indicated.

Line in Table B used =

Subjects on regimen A =

Subjects on regimen B =

12 Chapter 3

Subjects on regimen C =

Subjects on regimen D =

Exercise 3.35**KEY CONCEPTS** - properties of random digits

Write your answers (True or False) in the space provided. Remember that a table of random digits is defined to be a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 that has the following properties:

1. The digit in any position in the list has the same chance of being any one of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
2. The digits in different positions are independent in the sense that the value of one has no influence on the value of any other.

Additional properties are listed in the text below the box containing the definition above.

- a) _____
b) _____
c) _____

COMPLETE SOLUTIONS**Exercise 3.11**

The experimental units are the households that were called.

There are two factors in this study related to the nature of the introductory remarks. One factor is the information provided about the caller (name only, university being represented only, or name and university being represented). The other factor is whether survey results were offered (yes or no).

The treatments are combinations regarding the information about the caller and whether the survey results were offered. Thus there are six treatments ((1) name only and survey results offered, (2) name only and survey results not offered, (3) university represented only and survey results offered, (4) university represented only and survey results not offered, (5) name and university provided and survey results offered, and (6) name and university provided and survey results not offered).

14 Chapter 3

The response variable is whether or not the interview was completed.

Exercise 3.13

In this case, the subjects are the 36 headache sufferers who have agreed to participate in the study. The two factors are antidepressant (placebo or antidepressant given) and stress management training (given or not given). These form the four treatments which we label as:

Treatment 1: Antidepressant and no stress management training.

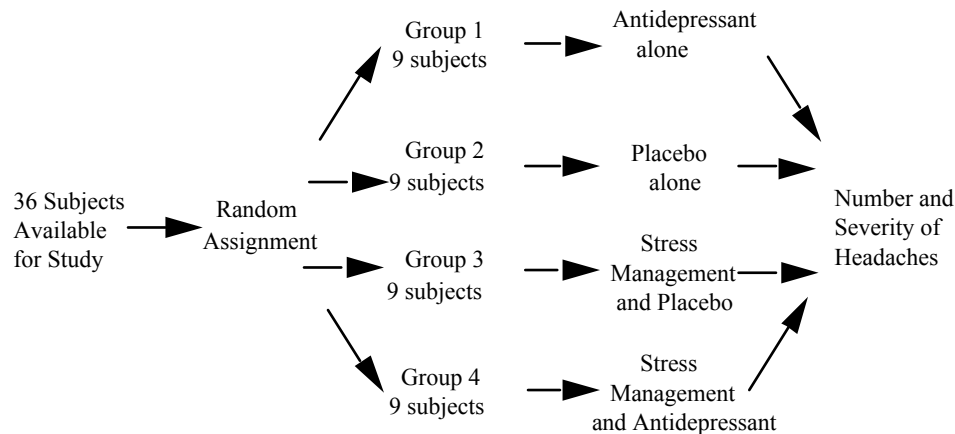
Treatment 2: Placebo (no antidepressant) and no stress management training.

Treatment 3: Placebo (no antidepressant) and stress management training.

Treatment 4: Antidepressant and stress management training.

The response variables are number of headaches over the study period and some measure of the severity of these headaches. The problem does not specify how the severity is to be measured.

The study should be done as follows. Subjects should be randomly assigned to treatments, with nine being assigned to treatment 1, nine being assigned to treatment 2, nine being assigned to treatment 3 and the remainder assigned to treatment 4. Each subject follows their treatment regimen over the course of the study. The average number of headaches for each treatment should be calculated and the results for the four groups compared, as well as a comparison of the severity. A picture which summarizes the experimental design is given below.



Although you can use the applet or Table B, we illustrate the use of Table B to carry out the random assignment of the subjects to the treatments. First label

16 Chapter 3

the 36 names using 2-digit labels. We use the convention of starting with the label 00 and label down the columns. Of course, one could start with another number (such as 01) and label across rows if one wished. The names with labels are

00 Acosta	09 Farouk	18 Liang	27 Solomon
01 Asihiro	10 Fleming	19 Maldonado	28 Trujillo
02 Bennett	11 George	20 Marsden	29 Tullock
03 Bikalis	12 Han	21 Montoya	30 Valasco
04 Chen	13 Howard	22 O'Brian	31 Vaughn
05 Clemente	14 Hruska	23 Ogle	32 Wei
06 Duncan	15 Imrani	24 Padilla	33 Wilder
07 Durr	16 James	25 Plochman	34 Willis
08 Edwards	17 Kaplan	26 Rosen	35 Zhang

Line 130 from Table B is reproduced below. We should read line 130 in pairs of digits from left to right. We have placed vertical bars between consecutive pairs to indicate how we have read the table. We underline those pairs that correspond to labels in our list and that have not been previously selected. On line 130 we have

69|05|1 6|48|17| 87|17|4 0|95|17| 84|53|4 0|64|89| 87|20|1 9|72|45

We only find 5 of our labels so we need to continue reading on line 131.

05|00|7 1|66|32| 81|19|4 1|48|73| 04|19|7 8|55|76| 45|19|5 9|65|65

We now have eight labels and continuing to line 132 gives us the 9th label for the group assigned to treatment 1.

68|73|2 5|52|59| 84|29|2 0|87|96| 43|16|5 9|37|39| 31|68|5 9|71|50

The treatment 1 group consists of Clemente, James, Kaplan, Marsden, Maldonado, Acosta, Wei, Chen, and Plochman. Continuing in line 132 (and ignoring pairs corresponding to previously selected subjects) we assign the next nine subjects to treatment 2,

|52|59| 84|29|2 0|87|96| 43|16|5 9|37|39| 31|68|5 9|71|50

45|74|0 4|18|07| 65|56|1 3|33|02| 07|05|1 9|36|23| 18|13|2 0|95|47

27|

giving the treatment 2 group as Tullock, Vaughn, Liang, Durr, Howard, Wilder, Bennett, Ogle, and Solomon. Continuing in line 134, the treatment 3 group is

|81|6 7|84|16| 18|32|9 2|13|37| 35|21|3 3|77|41| 04|31|2 6|85|08

66|92|5 5|56|58| 39|10|0 7|84|58| 11|20|6 1|98|76| 87|15|1 3|12|60

08|42|1 4|

Zhang, Montoya, Rosen, Edwards, Fleming, George, Imrani, Han, and Hruska assigned to treatment 3. The nine remaining subjects, Ashiro, Bikalis, Duncan, Farouk, O'Brian, Padilla, Trujillo, Valasco, and Willis are assigned to treatment 4. Using Table B can become tedious with a large number of subjects and it is best to leave such calculations to a computer.

Exercise 3.29

We have 15 subjects available. The factor is “hand” and there are two treatments. Treatment 1 is squeezing with the right hand and treatment 2 is squeezing with the left hand. The response is the force exerted as indicated by the reading on the scale.

To do the experiment, we use a matched pairs design. Each subject is a block and each subject uses each hand (the matched pairs are the two hands of a particular subject). We should randomly decide which hand to use first, perhaps by flipping a coin. We measure the response for each hand and then compare the forces for the left and right hands over all subjects to see if there is a systematic difference between the two hands.

If we use a coin to do the randomization, we might decide that if we get heads the right-hand goes first. Tails means the left-hand goes first. For 15 flips of a coin we got

HTTHTTHTHHHTTTT

This tells us that subject 1 uses the right-hand first, subject 2 the left hand first, subject 3 the left-hand first, etc. Notice with this scheme the number of times the right-hand goes first is 6 and the number of times the left-hand goes first is 9.

Exercise 3.32

a) The subjects and their excess weights, rearranged in increasing order of excess weight, are listed below. The columns are the five blocks. We have labeled the subjects in each block from 1 to 4.

Block 1	Block 2	Block 3	Block 4	Block 5
1 Williams 22	1 Santiago 27	1 Brunk 30	1 Jackson 33	1 Birnbaum 35
2 Festinger 24	2 Mann 28	2 Obrach 30	2 Stall 33	2 Tran 35
3 Hernandez 25	3 Smith 29	3 Rodriguez 30	3 Brown 34	3 Nevesky 39
4 Moses 25	4 Kendall 30	4 Loren 32	4 Dixon 34	4 Wilansky 42

b) We used lines 130 and 131 in Table B, which are given below.

69051 64817 87174 09517 84534 06489 87201 97245

05007 16632 81194 148731 04197 85576 45195 96565

For block 1, we read these from left to right, one digit at a time. The first label we encounter is assigned to regimen A, the next to regimen B, the next to

regimen C, and the remaining label is then automatically assigned to regimen D. We underline those that are one of our labels, skipping repeats. The vertical lines indicates when we have completed a block. We summarize our results on the next page.

Regimen A = Williams, Kendall, Obrach, Brown, Birnbaum

Regimen B = Moses, Mann, Loren, Stall, Wilansky

Regimen C = Hernandez, Santiago, Brunk, Jackson, Nevesky

Regimen D = Festinger, Smith, Rodriguez, Dixon, Tran

Exercise 3.35

a) This is false. Randomness does not mean each digit appears the exact same number of times in each row. For example, look at line 150 in Table B. There are only two 0s in this row.

b) True. Randomness means each digit has an equal chance ($1/10$) of being a 0, each pair an equal chance ($1/100$) of being 00, each triple an equal chance ($1/1000$) of being 000, etc. This point is made in the section "how to randomize" in the text a few paragraphs below the box containing the definition of a table of random digits.

c) False. Following the logic in (b), any set of four digits has an equal chance ($1/10000$) of being 0000. Thus the digits 0000 can appear, but the chance of any four digits being this sequence is quite small.

SECTION 3.3

OVERVIEW

The **population** is the entire group of individuals or objects about which we want information. The information collected is contained in a **sample** which is the part of the population we actually get to observe. How the sample is chosen, that is, the **design**, has a large impact on the usefulness of the data. A useful sample will be representative of the population and will help answer our questions. "Good" methods of collecting a sample include the following:

probability samples

simple random samples, also called **SRS**

stratified random samples

multistage samples

All these sampling methods involve some aspect of randomness through the use of a formal chance mechanism. Random selection is just one precaution that a person can take to reduce **bias**, the systematic favoring of a certain outcome.

22 Chapter 3

Samples we select using our own judgment, because they are convenient, or "without forethought" (mistaking this for randomness) are usually biased in some way. This is why we use computers or a tool like a **table of random digits** to help us select a sample.

A **voluntary response sample** includes people who choose to be in the sample by responding to a general appeal. They tend to be biased, as the

sample is overrepresented by individuals with strong opinions, which are often negative.

Other kinds of bias to be on the lookout for include:

nonresponse bias which occurs when individuals who are selected do not participate or cannot be contacted,

undercoverage which occurs when some group in the population is given either no chance or a much smaller chance than other groups to be in the sample, and

response bias which occurs when individuals do participate but are not responding truthfully or accurately due to the way the question is worded, the presence of an observer, fear of a negative reaction from the interviewer, or any other such source.

These types of bias can occur even in a randomly chosen sample and we need to try to reduce their impact as much as possible.

GUIDED SOLUTIONS

Exercise 3.39

KEY CONCEPTS - populations and sources of bias

What variable was measured and what was the sample? Now, try and identify the population as exactly as possible. Where the information is not complete, you may need to make assumptions to try to describe the population in a reasonable way. Make sure not to confuse the population of interest with the population actually sampled. When they don't coincide there is always a strong potential for bias. What are some possible sources of bias in this example?

Exercise 3.41

KEY CONCEPTS - selecting a SRS with a table of random numbers

The table of random numbers can be used to select a SRS of numbers - in order to use it to sample from the students in the statistics course, the individuals in the course need to be assigned numbers. So that everyone does the problem the "same" way, we have first numbered the students according to alphabetical order in the list.

01- Agarwal	08 - Dewald	15 - Huang	22 - Puri
02 - Alfonseca	09 - Fleming	16 - Kim	23 - Richards

24 Chapter 3

03 - Baxter	10 - Fonseca	17 - Lujan	24 - Rodriguez
04 - Bowman	11 - Gates	18 - Mourning	25 - Santiago
05 - Brown	12 - Goel	19 - Nunez	26 - Shen
06 - Cortez	13 - Gomez	20 - Peters	27 - Vega
07 - Cross	14 - Hernandez	21 - Pliego	28 - Watanabe

If you go to line 139 in the table and start selecting two digit numbers, then you should get the same answer as given in the complete solution.

Exercise 3.46

KEY CONCEPTS - systematic sampling

a) This is like the example except there are now 200 addresses instead of 100, and the sample size is now 5 instead of 4. With these two changes, you need to think about how many different systematic samples there are. Two different systematic samples are:

systematic sample 1 =	01, 41, 81, 121, 161
systematic sample 2 =	02, 42, 82, 122, 162

How many systematic samples are there altogether? Choosing one of these systematic samples at random is equivalent to choosing the first address in the sample. The remaining four addresses follow automatically by adding 40. Carry this out using line 120 in the table.

b) Why are all addresses equally likely to be selected? First, how many systematic samples contain each address? The chance of selecting an address is the same as the chance of selecting the systematic sample that contains it. With this in mind, what is the chance of any address being chosen? By the definition of a SRS, all samples of 5 addresses are equally likely to be selected. In a systematic sample, are all samples of 5 addresses even possible?

Exercise 3.51

KEY CONCEPTS - sampling frame, undercoverage

a) Which households wouldn't be in the sampling frame? Make some educated guesses as to how these households might differ from those in the sampling frame (other than the fact that they don't have a phone number in the directory).

b) Random digit dialing makes the sampling frame larger - which households are added to it?

Exercise 3.55**KEY CONCEPTS** - wording of questions

Questions can be worded in such a way that makes it seem as though any reasonable person should agree (disagree) with the statement. Which questions are slanted towards a desired response? Are all the questions clear?

COMPLETE SOLUTIONS**Exercise 3.39**

The variable being measured is approval of the president's overall job performance which is recorded as approve or don't approve. The sample is the 1210 adults that were actually interviewed. The population of interest is probably all adult citizens of the U.S. or possibly just registered voters.

There are several possible sources of bias in the study. The states of Alaska and Hawaii were omitted and there is no reason to believe that the adult residents of these states were not intended to be part of the population (they may not have been included in the sample due to the higher cost of calling residents of these states). Any systematic differences in the opinions of the adults in Alaska and Hawaii and the remaining states will bias the results. Also, only residents with phones could be contacted and if the phone numbers were selected from phone books then residents with unlisted numbers could not be in the sample. This is another possible source of bias, which is just any systematic error in the way the sample represents the population. Finally, there may be bias due to nonresponse, as all adults contacted by phone may not have been willing to give their opinion.

Exercise 3.41

To choose a SRS of 6 students to be interviewed, first label the members of the population by associating a 2 digit number with each.

01- Agarwal	08 - Dewald	15 - Huang	22 - Puri
02 - Alfonseca	09 - Fleming	16 - Kim	23 - Richards
03 - Baxter	10 - Fonseca	17 - Lujan	24 - Rodriguez

28 Chapter 3

04 - Bowman	11 - Gates	18 - Mourning	25 - Santiago
05 - Brown	12 - Goel	19 - Nunez	26 - Shen
06 - Cortez	13 - Gomez	20 - Peters	27 - Vega
07 - Cross	14 - Hernandez	21 - Pliego	28 - Watanabe

Now enter Table B and read two-digit groups until 6 students are chosen. Starting at line 139

55588 99404 70708 41098 43563 56934 48394 51719
12975 13258 13048

The selected sample is 04 - Bowman, 10 - Fonseca, 17 - Lujan, 19 - Nunez, 12 - Goel, and 13 - Gomez.

Exercise 3.46

a) We want to select 5 addresses out of 200, so we think of the 200 addresses as forty lists, each containing 5 addresses. We choose one address from the first 40, and then every 40th address after that. The first step is to go to Table B, line 120 and choose the first two digit random number you encounter that is one of the numbers 01, ..., 40.

35476

The selected number is 35, so the sample includes addresses numbered 35, 75, 115, 155, and 195.

b) Each individual is in exactly one systematic sample, and the systematic samples are equally likely to be chosen. In our previous example, there were 40 systematic samples, each containing 5 addresses. The chance of selecting any address is the chance of picking the systematic sample that contains it, which is 1 in 40.

A simple random sample of size n would allow every set of n individuals an equal chance of being selected. Thus, in this exercise, when using a SRS the sample consisting of the addresses numbered 1, 2, 3, 4, and 5 would have the same probability of being selected as any other set of 5 addresses. For a systematically selected sample, all samples of size n do not have the same probability of being selected. In our exercise the sample consisting of the addresses numbered 1, 2, 3, 4, and 5 would have zero chance of being selected since the numbers of the addresses do not all differ by 40. The sample we selected in (a), 35, 75, 115, 155, and 195 had a 1 in 40 chance of being selected, so all samples of five addresses are not equally likely.

Exercise 3.51

a) Households omitted from the frame are those which do not have a telephone number listed in the telephone directory. The types of people who might be

underrepresented are poorer (including homeless) people who cannot afford to have a phone, and the group of people who have unlisted numbers. It is harder to characterize this second group. As a group they would tend to have more money as you need to pay to have your phone number unlisted or it might include more single women who do not want their phone numbers available and

possibly people whose jobs put them in contact with large groups of people who might harass them if their phone number was easily accessible.

b) People with unlisted numbers will be included in the sampling frame. The sampling frame would now include any household with a phone. One interesting point is that all households will not have the same probability of getting in the sample, as some households have multiple phone lines and will be more likely to get in the sample. So, strictly speaking, random digit dialing will not actually provide a SRS of households with phones. Just a SRS of phone numbers!

Exercise 3.55

a) The beginning of the question suggests that cell phone use is associated with brain cancer. This initial suggestion and the wording "the danger of using cell phones" would lead most reasonable people to be in favor of including a warning label. The question is slanted in favor of this response.

b) The question is clear but is slanted in favor of national health insurance. The reason for agreeing with a question should not be contained within the question.

c) The question is slanted as it contains reasons why you should support recycling. As a question, the wording is a little technical for the general population and a simpler version such as "Do you favor economic incentives to promote recycling?" would be better.

SECTION 3.4

OVERVIEW

Statistical inference is the technique which allows us to use the information in a sample to draw conclusions about the population. To understand the idea of statistical inference, it is important to understand the distinction between **parameters** and **statistics**. A **statistic** is a number we calculate based on a sample from the population - its value can be computed once we have taken the sample, but its value varies from sample to sample. A statistic is generally used to estimate a population **parameter** which is a fixed but unknown number that describes the population.

The variation in a statistic from sample to sample is called **sampling variability**. It can be described through the **sampling distribution** of the statistic which is the distribution of values taken by the statistic in all possible samples of the same size from the population. The sampling distribution can be

described in the same way as the distributions we encountered in Chapter 1. Three important features are:

- a measure of center
- a measure of spread
- a description of the shape of the distribution

The properties and usefulness of a statistic can be determined by considering its sampling distribution. If the sampling distribution of a statistic is centered (has its mean) at the value of the population parameter, then the statistic is **unbiased** for this parameter. This means that the statistic tends to neither overestimate nor underestimate the parameter.

Another important feature of the sampling distribution is its spread. If the statistic is unbiased and the sampling distribution has little spread or variability, then the statistic will tend to be close to the parameter it is estimating for most samples. The variability of a statistic is related to both the sampling design and the sample size n . Larger sample sizes give smaller spread (better estimates) for any sampling design. An important feature of the spread is that as long as the population is much larger than the sample (at least 100 times), the spread of the sampling distribution will depend primarily on the sample size, not the population size.

If the parameter p is the proportion of the population with a particular characteristic, then the statistic \hat{p} , the proportion in the sample with this characteristic, is an unbiased estimator. Provided the samples are selected at random, **probability** theory can be used to tell us about the distribution of a statistic.

GUIDED SOLUTIONS

Exercise 3.59

KEY CONCEPTS - statistics and parameters

In deciding whether a number represents a parameter or a statistic, you need to think about whether it is a numerical characteristic of the population of interest or whether it is a numerical characteristic of the particular sample that was selected. Statistics vary from sample to sample; parameters are fixed numerical characteristics of the population.

Exercise 3.65

KEY CONCEPTS - variability of the sample proportion.

a) "As long as the population is much larger than the sample (say, at least 100 times as large), the spread of the sampling distribution for a sample of fixed size

34 Chapter 3

n is approximately the same for any population size." You need to think about how this rule applies to this example.

b) Is the rule given in part (a) applicable here? Read it carefully.

Exercise 3.69**KEY CONCEPTS** - sampling distributions

a) The table of random numbers contains the 10 digits, 0, 1, 2, ..., 9, which are "equally likely" to occur in any position selected at random from the table. If we want an egg mass to be present 20% of the time, then two digits correspond to the presence of an egg mass and the remaining eight digits correspond to the absence of an egg mass. Does it matter which two digits correspond to the presence of an egg mass?

While any two digits could be used, so that everyone does the same thing, let the occurrence of the digits 0 or 1 correspond to the presence of an egg mass, and the remaining digits correspond to the absence. Also, let's start on line 128 of the table.

15689 14227

These 10 random digits correspond to our 10 sample areas. There are two sample areas with egg masses (correspond to a digit of 0 or 1), so that $\hat{p} = .2$ for this sample.

b) For this part of the problem, everyone will be taking their 20 samples from different parts of the random number table. Some of you may know how to get random samples from your computer software. Work with your own samples here. Your answers will not agree exactly with that given in the complete solution - the general pattern should be similar. If everyone took 2000 samples instead of 20, would the sampling distributions from person to person show more or less agreement?

COMPLETE SOLUTIONS**Exercise 3.59**

2.503 cm. is a property (the mean) of the carload (population) of ball bearings and is the value of a parameter. **2.515** cm. is a property of the sample of 100 bearings inspected. It is the value of a statistic.

Exercise 3.65

a) The population is at least 100 times the sample size $n = 2000$ for each of the states. So the variability in the sample proportion based on $n = 2000$ will be approximately the same for the population size of any state.

b) The problem switches here. The rule applies to a given sample size - the variability of the sample proportion based on a fixed number of observations is approximately the same for any population size. Now the sample size will vary from state to state. For Wyoming, 1/10 of 1% of the population is a sample size of about $n = 494$ and 1/10 of 1% of the population of California is a sample size of about $n = 34000$. Since larger sample sizes give smaller spread, there will be differences in the variability of the sample proportion from state to state. California's sample proportion will be much less variable than the sample proportion from Wyoming.

Exercise 3.69

a) Done in the guided solution

b) These are the values of \hat{p} in the 20 samples we obtained using the computer to generate random digits, followed by the stem and leaf plot.

sample	\hat{p}
1	0.1
2	0.1
3	0.0
4	0.1
5	0.0
6	0.4
7	0.0
8	0.3
9	0.2
10	0.0
11	0.2
12	0.0
13	0.4
14	0.2
15	0.4
16	0.2
17	0.2
18	0.1
19	0.3
20	0.2

```

0.0 | 00000
0.1 | 0000
0.2 | 0000000
0.3 | 00
0.4 | 000

```

The mean of the distribution is 0.17. The shape looks fairly symmetric with a center near 0.2. Your stem and leaf plot may look quite different from this - with only 20 samples the distributions may vary quite a bit from person to person. If everyone took 2000 samples, which would require the sampling be

done using a computer, then the shapes of the distributions would be quite similar from person to person.