# Chapter 1: The Where, Why, and How of Data Collection

Section 1.1

1.1. This application is primarily descriptive in nature. The owner wishes to develop a presentation. She will most likely use charts, graphs, tables and numerical measures to describe her data.

1.2. The graph is a bar chart. A bar chart displays values associated with categories. In this case the categories are the departments at the food store. The values are the total monthly sales (in dollars) in each department. A bar chart also typically has gaps between the bars. A histogram has no gaps and the horizontal axis represents the possible values for a numerical variable.

1.3. A bar chart is used whenever you want to display data that has already been categorized while a histogram is used to display data over a range of values for the factor under consideration. Another fundamental difference is that there typically are gaps between the bars on a bar chart but there are no gaps between the bars of a histogram.

1.4. Businesses often make claims about their products that can be tested using hypothesis testing. For example, it is not enough for a pharmaceutical company to claim that its new drug is effective in treating a disease. In order for the drug to be approved by the Food and Drug Administration the company must present sufficient evidence that the drug first does no harm and that it also provides an effective treatment against the disease. The claims that the drug does no harm and is an effective treatment can be tested using hypothesis testing.

1.5. The company could use statistical inference to determine if its parts last longer. Because it is not possible to examine every part that could be produced the company could examine a randomly chosen subset of its parts and compare the average life of the subset to the average life of a randomly chosen subset of the competitor's parts. By using statistical inference procedures the company could reach a conclusion about whether its parts last longer or not.

1.6. Student answers will vary depending on the periodical selected and the periodical's issue date, but should all address the three parts of the question.

1.7.
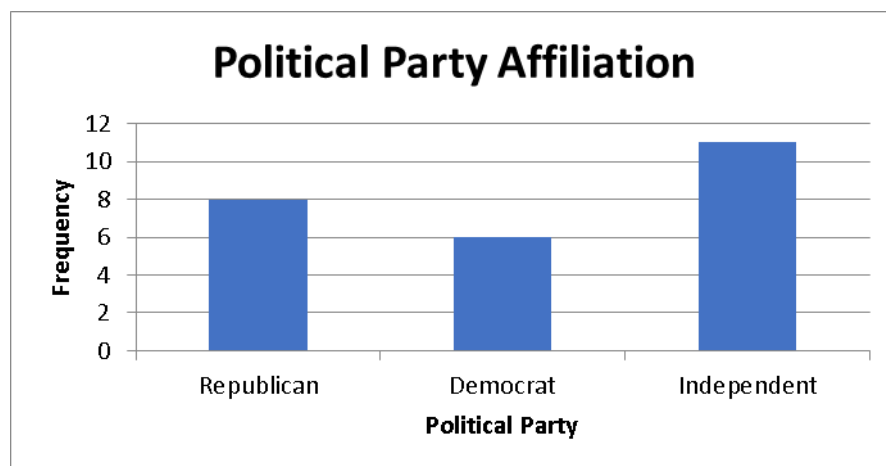
1.8.   Because it would be too costly, too time consuming, or practically impossible to contact every subscriber to ascertain the desired information, the decision makers at *Fortune* might decide to use statistical inference, particularly estimation, to answer its questions. By looking at a subset of the data and using the procedures of estimation it would be possible for the decision makers to arrive at values for average age and average income that are within tolerable limits of the actual values.

1.9.   Student answers will vary depending on the business periodical or newspaper selected and the article referenced. Some representative examples might include estimates of the number of CEO's who will vote for a particular candidate, estimates of the percentage increase in wages for factory workers, estimates of the average dollar advertising expenditures for pharmaceutical companies in a specific year, and the expected increase in R&D expenditures for the coming quarter.

1.10.   Student answers will vary. However, the examples should illustrate how statistics has been used and should clearly indicate the type of statistical analysis employed.

Section 1.2

1.11.   As discussed in this section, the pet store would most likely use a written survey or a telephone survey to collect the customer satisfaction data.

1.12.   A leading question is one that is designed to elicit a specific response, or one that might influence the respondent's answer by its wording. The question is posed so that the respondent believes the researcher has a specific answer in mind when the question is asked, or worded in such a way that the respondent feels obliged to provide an answer consistent with the question. For example, a question such as "Do you agree with the experts who recommend that more tax dollars be given to clean up dangerous and unhealthy pollution?" could cause respondents to provide the answer that they think will be consistent with the "experts" with whom they do not want to disagree. Leading question should be avoided in surveys because they may introduce bias.

1.13.   An experiment is any process that generates data as its outcome. The plan for performing the experiment in which the variable of interest is defined is referred to as an experimental design. In the experimental design one or more factors are identified to be changed so that the impact on the variable of interest can be observed or measured.

1.14.   There will likely by a high rate of nonresponse bias since many people who work days will not be home during the 9–11 AM time slot. Also, the data collectors need to be careful where they get the phone number list as some people do not have listed phones in phone books and others have no phone or only a cell phone. This may result in selection bias.

1.15.   a.   Observation would be the most likely method. Observers could be located at various bike routes and observe the number of riders with and without helmets. This would likely be better than asking people if they wear a helmet since the popular response might be to say yes even when they don't always do so.

        b.   A telephone survey to gas stations in the state. This could be a cost effective way of getting data from across the state. The respondent would have the information and be able to provide the correct price.

        c.   A written survey of passengers. This could be given out on the plane before the plane lands and passengers could drop the surveys in a box as they de-plane. This method would likely

garner higher response rates compared to sending the survey to passengers' mailing address and asking them to return the completed survey by mail.

1.16. The two types of validity mentioned in the section are internal validity and external validity. For this problem external validity is easiest to address. It simply means the sampling method chosen will be sufficient to insure the results based on the sample will be able to be generalized to the population of all students. Internal validity would involve making sure the data gathering method, for instance a questionnaire, accurately determines the respondent's attitude toward the registration process.

1.17.

Experiments		Direct observations	
Advantages	Disadvantages	Advantages	Disadvantages
The researcher can have control over variables.	The experiment results may apply to only one situation and may be difficult to replicate.	Direct observations may enable access to people in real-life situations.	Direct observations can be viewed as too subjective.

1.18. There are many potential sources of bias associated with data collection. If data is to be collected using personal interviews it will be important that the interviewer be trained so that interviewer bias, arising from the way survey questions are asked, is not injected into the survey. If the survey is conducted using either a mail survey or a telephone survey then it is important to be aware of nonresponse bias from those who do not respond to the mailing or refuse to answer your calls. You must also be careful when selecting your survey subjects so that selection bias is not a problem. In order to have useful, reliable data that is representative of the true student opinions regarding campus food service, it is necessary that the data collection process be conducted in a manner that reduces or eliminates the potential for these and other sources of potential bias.

1.19. For retailers technology that scans the product UPC code at checkout makes the collection of data fast and accurate. Retailers that use such technology can automatically update their inventory records and develop an extensive collection of customer buying habits. By applying advanced statistical techniques to the data the retailer can identify relationships among purchases that might otherwise go unnoticed. Such information could enable retailers to target their advertising or even rearrange the placement of products in the store to increase sales. Manufacturing firms use bar code scanning to collect information concerning product availability and product quality. Credit card purchases are automatically tracked by the retailer and the bankcard company. In this way the credit card company is able to track your purchases and even alert you to potential fraud if purchases on your card appear to be unusual. Finally, some companies are using radio frequency identification (RFID) to track products through their supply chain, so that product delays and inventory problems can be minimized.

1.20. Yes, a survey can be done through the Internet; it is called an Internet survey. An Internet survey reduces the survey cost, the human errors and the time when compared with other methods. It also reaches respondents in different geographic areas effectively.

1.21. Yes, open-ended questions may improve the accuracy of patient's medication because while medical records may not tell doctors what drugs their patients are actually taking, asking open-ended questions can improve the accuracy of medication lists.

1.22.   a.   Direct observation is appropriate for the parents. Parents may personally observe the place or environment where the training takes place, which could help them understand the events and activities conducted in the training.

   b.   An experiment is appropriate for Steven. He may bring along his old bulb and compare the brightness of the bulbs before making his purchase.

   c.   Personal interviews are appropriate for L'Oréal's staff. The staff may investigate the answers provided by the respondents, following which they may ask more questions in their survey to get more accurate answers.

1.23.   We need to be concerned about external validity because we may argue about whether seminar attendance is the factor that improves the performance in exams or whether there are other factors that improve the performance in exams. Even if it is true that the seminar attendance improves the performance in exams for this batch of students, would it have the same result for another batch of students?

1.24.   The results of the survey are based on telephone interviews with 744 adults, aged 18 and older. Students may also answer that the survey could have been conducted using a written survey via mail questionnaire or internet survey. Because telephone interviews were used to collect the survey data nonresponse biases associated with sampled adults who are not at home when phoned, or adults who refuse to participate in the survey. There is also the problem that some adults do not have a landline phone. If written surveys are used to collect the data then it is important to guard against nonresponse bias from those sampled adults who do not complete the survey There is also the problem of selection bias. In phone interviews we may miss the people who work evenings and nights. If written surveys are used we must be careful to select a representative sample of the adult population.

Section 1.3

1.25.   Yes, there is bias in the sampling process because the sample tends to favor the shoppers who are at that particular mall and who are in shops on that particular day.

1.26.   With 7,000/200 = 35, we can perform a 1-in-35 systematic sampling, where we should sample every 35th student until we have reached 200 samples.

$$\text{Part Range} = \frac{\text{Population Size}}{\text{Sample Size}} = \frac{18,000}{100} = 180$$

   Thus, the first person selected will come from employees 1–180. Once that person is randomly selected, the second person will be the one numbered 180 higher than the first, and so on.

1.27.   Whenever a descriptive numerical measure such as an average is calculated from the entire population it is a parameter. The corresponding measure calculated from a subset of the population, that is to say a sample, is a statistic.

1.28.   Statistical sampling techniques consist of those sampling methods that select samples based on chance. Nonstatistical sampling techniques consist of those methods of selecting samples using convenience, judgment, or other nonchance processes. In convenience sampling, samples are chosen because they are easy or convenient to sample. There is no attempt to randomize the selection of the selected items. In convenience sampling not every item in the population has a random chance of being selected. Rather, items are sampled based on their convenience alone. Thus, convenience sampling is not a statistical sampling method.

1.29.   Kris may randomly select any security as her starting point, and then select every (2,531/100 = 25.31) 25th security after that until 100 securities are sampled.

1.30.   A census is an enumeration of the entire set of measurements taken from the population as a whole. While in some cases, the items of interest are obtained from people such as through a survey, in many instances the items of interest come from a product or other inanimate object. For example, a study could be conducted to determine the defect rate for items made on a production line. The census would consist of all items produced on the line in a defined period of time.

1.31.   The customer service manager may face nonresponse bias because the customers who purchased the hairdryer may not return the survey form to him.

1.32.   In stratified random sampling, the population is divided into homogeneous groups called strata. The idea is to make all items in a stratum as much alike as possible with respect to the variable of interest thereby reducing the number of items that will need to be sampled from each stratum. In cluster sampling, the idea is to break the population into heterogeneous groups called clusters (usually on a geographical basis) such that each cluster looks as much like the original population as possible. Then clusters are randomly selected and from the cluster, individual items are selected using a statistical sampling method.

1.33.   Using Excel, choose the Data tab, select Data Analysis from the Analysis Group , then Random Number Generation—shown as follows:
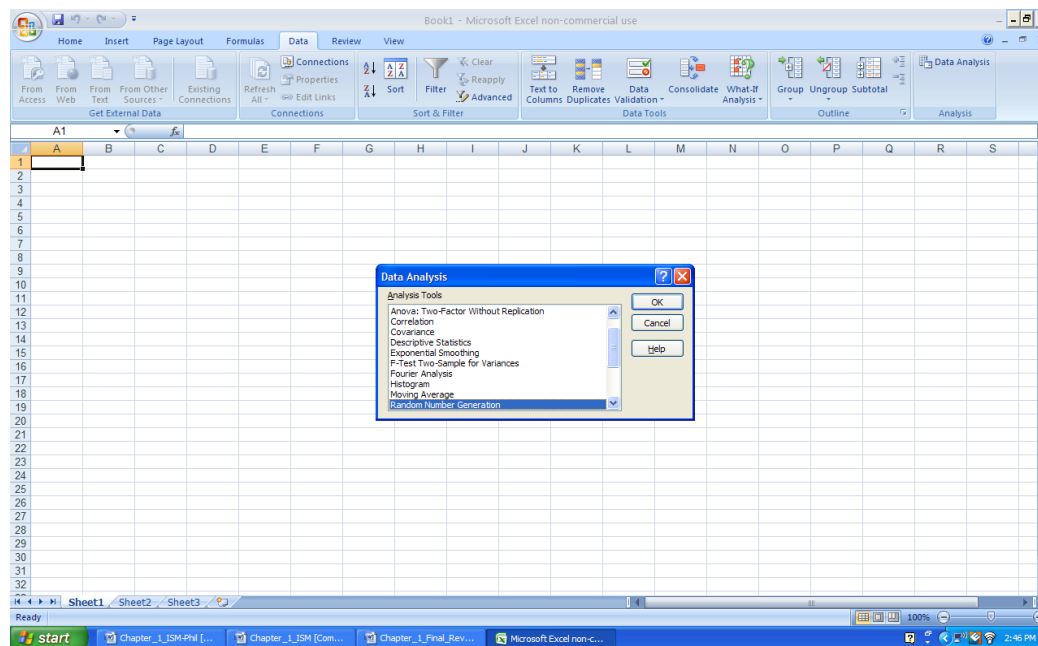
The next step is to complete the random number generation dialog as follows:
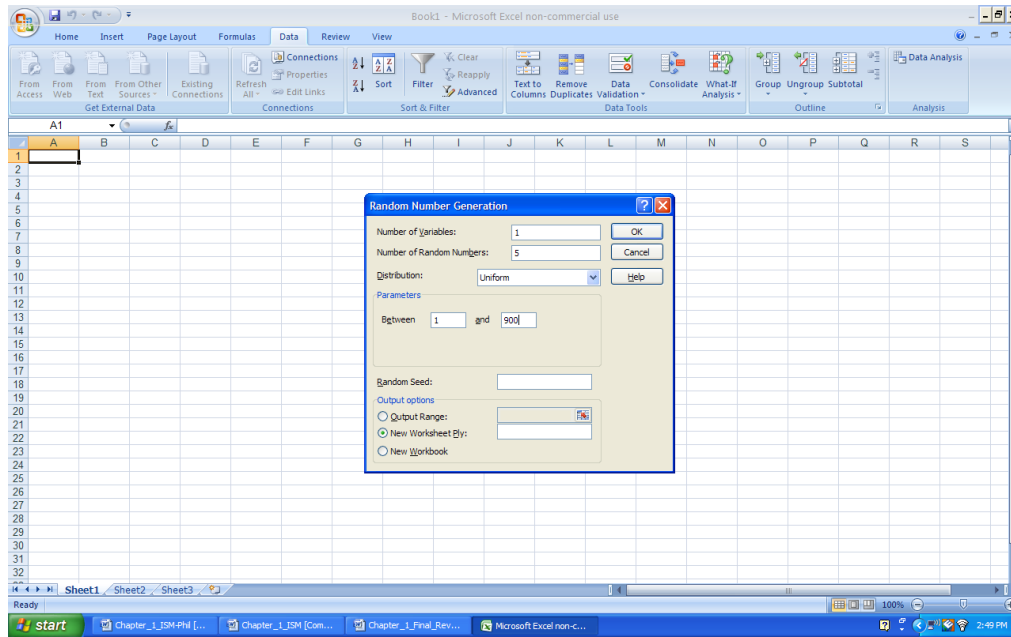
The resulting random numbers generated are:

344.4182
91.51183
537.2394
809.2961
796.264

Note, the students' answers may differ since Excel generates different streams of random numbers each time it is used. Also, if the application requires integer numbers, the Decrease Decimal option can be used.

1.34. The manager used the stratified random sampling method because the entire university was divided into the subgroups of administration, teaching staff and non-teaching staff.

1.35. This is a statistic. A poll would be a sample of eligible voters rather than all eligible voters.

1.36. Solution
   a. Stratified random sampling
   b. Simple random sampling or possibly cluster random sampling
   c. Systematic random sampling
   d. Stratified random sampling

1.37. This is a statistical sample. Every employee has an equal chance of being selected using this method. In fact, this is an example of a simple random sample because every possible sample of size 50 has an equal chance of being selected.

1.38. a. Student answers will vary
   b. Cluster sampling could be used to ensure that you get all types of cereal. Make each cluster the area where certain cereals are located (i.e., isle, row, shelf, etc.)
   c. Cluster sampling would give you a better idea of the inventory of all types of cereal. Simple random sampling could possibly end up with only looking at 2 or 3 cereal types.

1.39. Students should choose the Data tab, select Data Analysis from the Analysis group—Random Number Generation process. Students' answers will differ since Excel generates different streams of random numbers each time it is used, but 40 random numbers should be generated from a uniform distribution with values ranging from 1 to 578. Since the application requires integer numbers, the Decrease Decimal option should be used.

1.40. a. The population should be all users of cross-country ski lots and trailheads in Colorado.
    b. Several sampling techniques could be selected. Be sure that some method of ensuring randomness is discussed. In addition, some students might give greater weight to frequent users of the lots. In which case the population would really be user days rather than individual users.
    c. Students using Excel should choose the Data tab, select Data Analysis from the Analysis group—Random Number Generation process. Students' answers may differ since Excel generates different streams of random numbers each time it is used. Since the application requires integer numbers, the Decrease Decimal option should be used.

1.41. a. Since there are 4,000 patient files we could give each file a unique identification number consisting of 4 digits. The first file would be given the identification number "0001." The last file would be given the identification number of "4000." By assigning each patient a number and randomly selecting the 100 numbers allows each possible sample of 100 an equal chance of being selected.
    b. Either use a random number table (randomly select the starting row and column), or use a computer program, such as Microsoft Excel, which has a random number generator.
    c. Since each patient is assigned a 4-digit identification number, we would need a 4-digit random number for each random number selected.
    d. Answers will vary.

Section 1.4

1.42. a. Time-series
    b. Cross-sectional
    c. Time-series
    d. Cross-sectional

1.43. Qualitative data are categories or numerical values that represent categories. Quantitative data is data that is purely numerical.

1.44. a. Ordinal—categories with defined order
    b. Nominal—categories with no defined order
    c. Ratio
    d. Nominal—categories with no defined order

1.45. Nominal data involves placing observations in separate categories according to some measurable characteristic. Ordinal data also involves placing observations into separate categories, but the categories can be rank-ordered.

1.46. Since the circles involve a ranking from best to worst, this would be ordinal data.

1.47. a. The population of interest is the entire company of 500 employees.

    b. The sample is the 50 selected employees obtained by choosing every 10th employee from the alphabetically ordered list of employees.

    c. The systematic sampling method is being used.

1.48. a. Nominal Data
      b. Ratio Data
      c. Ratio Data
      d. Ratio Data
      e. Nominal

1.49. a. Cross-sectional
      b. Time-series
      c. Cross-sectional
      d. Cross-sectional
      e. Time-Series

1.50. Columns A–G are nominal—they are all codes
      Columns H–L are ratio level.

End of Chapter Exercises

1.51. a. Ratio
      b. Nominal
      c. Ratio
      d. Ordinal
      e. Interval
      f. Ratio
      g. Nominal.

1.52. Nominal data or ordinal data.

1.53. Interval or ratio data.

1.54. Ratings are typical uses of ordinal scale data. And since ratings are based on personal opinion, even though people are using the same scale, a direct comparison between the two ratings is not possible. This is a common problem when people are asked to rate an object using an ordinal scale.

1.55. Answers will vary with the student. But a good discussion should include the following factors:

      Sampling techniques and possible problems selecting a representative sample.

      Determining how to measure confidence.

      Structuring questions to avoid bias.

      The measurement scale associated with the questions.

      The fact this poll is specifically intended to develop time-series data.

1.56. Answers will vary with the student.

1.57. Answers will vary with the student.

1.58. a. It is cross-sectional data because it was collected on a specific day, which is at the same point of time.

b. Franchise code: qualitative

Most frequently purchased item: qualitative

Number of items purchased on last visit: quantitative

Total spent on last visit ($): quantitative

Satisfaction on purchased item: qualitative

c. Franchise code: nominal

Most frequently purchased item: nominal

Number of items purchased on last visit: interval

Total spent on last visit ($): interval

Satisfaction on purchased item: ordinal

1.59. The most common and basic form of probability sampling is simple random sampling. To conduct the sampling, we must first assign a number to each of the 400 employees in the company, and then use a random numbers table or computer software to select 15 random employees from the entire company to attend the training program.

1.60. a. Student answers will vary but one method would be personal observation at grocery stores or another method would be to simply look at their sales. Are buyers of the energy drinks purchasing bottles or cans?

b. If using personal observation just have people at grocery stores observe people over a specified period of time and note which are selecting cans and which are selecting bottles and look at the percentages of each.

c. You would be looking at ratio data because you could have a true 0 if, for example, no one purchased bottles.

d. Depends on the way the data are collected. Sales data would be quantitative.

1.61. a. Simple random sampling is the fairest sampling method to use to select the grand door prize.

b. The host of the evening dinner may mix all the tickets in a box and select the ticket for the grand door prize.

1.62. The appropriate design would be a stratified random sampling method. Start by dividing the students into class standing (Freshman, Sophomore, Junior, and Senior). Then randomly select students from each strata.