CHAPTER 1

1.1 (a) The type of beverage sold yields categorical or "qualitative" responses.

(b) The type of beverage sold yields distinct categories in which no ordering is implied.

1.2 Three sizes of U.S. businesses are classified into distinct categories—small, medium, and large—in which order is implied.

1.3 (a) The time it takes to download a video from the Internet is a continuous numerical or "quantitative" variable because time can have any value from 0 to any reasonable unit of time.

(b) The download time is a ratio scaled variable because the true zero point in the measurement is zero units of time.

1.4 (a) The number of cellphones is a numerical variable that is discrete because the outcome is a count. It is ratio scaled because it has a true zero point.

(b) Monthly data usage is a numerical variable that is continuous because any value within a range of values can occur. It is ratio scaled because it has a true zero point.

(c) Number of text messages exchanged per month is a numerical variable that is discrete because the outcome is a count. It is ratio scaled because it has a true zero point.

(d) Voice usage per month is a numerical variable that is continuous because any value within a range of values can occur. It is ratio scaled because it has a true zero point.

(e) Whether a cellphone is used for email is a categorical variable because the answer can be only yes or no. This also makes it a nominal-scaled variable.

1.5 (a) numerical, continuous, ratio scale
(b) numerical, discrete, ratio scale
(c) categorical, nominal scale
(d) categorical, nominal scale

1.6 (a) Categorical, nominal scale.
(b) Numerical, continuous, ratio scale.
(c) Categorical, nominal scale.
(d) Numerical, discrete, ratio scale.
(e) Categorical, nominal scale.

1.7 (a) numerical, continuous, ratio scale *
(b) categorical, nominal scale
(c) categorical, nominal scale
(d) numerical, discrete, ratio scale

*Some researchers consider money as a discrete numerical variable because it can be "counted."

1.8 (a) numerical, continuous, ratio scale *
(b) numerical, discrete, ratio scale
(c) numerical, continuous, ratio scale *
(d) categorical, nominal
*Some researchers consider money as a discrete numerical variable because it can be “counted.”

1.9 (a) Income may be considered discrete if we “count” our money. It may be considered continuous if we “measure” our money; we are only limited by the way a country's monetary system treats its currency.
(b) The first format is preferred because the responses represent data measured on a higher scale.

1.10 The underlying variable, ability of the students, may be continuous, but the measuring device, the test, does not have enough precision to distinguish between the two students.

1.11 (a) The population is “all working women from the metropolitan area.” A systematic or random sample could be taken of women from the metropolitan area. The director might wish to collect both numerical and categorical data.
(b) Three categorical questions might be occupation, marital status, type of clothing. Numerical questions might be age, average monthly hours shopping for clothing, income.

1.12 (a) The American Community Survey is a primary data source since the data are collected by the United States Census Bureau.
(b) The survey is based on a sample of addresses that are randomly selected.

1.13 The answer depends on the specific story.

1.14 The answer depends on the specific story.

1.15 The transportation engineers and planners should use primary data collected through an observational study of the driving characteristics of drivers over the course of a month.

1.16 The information presented there is based mainly on a mixture of secondary data distributed by an organization and data collected by ongoing business activities.

1.17 (a) 001 (b) 040 (c) 902

1.18 Sample without replacement: Read from left to right in 3-digit sequences and continue unfinished sequences from end of row to beginning of next row.
Row 05: 338 505 855 551 438 855 077 186 579 488 767 833 170
Rows 05-06: 897
Row 06: 340 033 648 847 204 334 639 193 639 411 095 924
Rows 06-07: 707
Row 07: 054 329 776 100 871 007 255 980 646 886 823 920 461
Row 08: 893 829 380 900 796 959 453 410 181 277 660 908 887
Rows 08-09: 237
Row 09: 818 721 426 714 050 785 223 801 670 353 362 449
Rows 09-10: 406
Note: All sequences above 902 and duplicates are discarded.

1.19 (a) Row 29: 12 47 83 76 22 99 65 93 10 65 83 61 36 98 89 58 86 92 71
Note: All sequences above 93 and all repeating sequences are discarded.

(b) Row 29: 12 47 83 76 22 99 65 93 10 65 83 61 36 98 89 58 86
Note: All sequences above 93 are discarded. Elements 65 and 83 are repeated.

1.20 A simple random sample would be less practical for personal interviews because of travel costs (unless interviewees are paid to attend a central interviewing location).

1.21 This is a probability sample because the selection is based on chance. It is not a simple random sample because A is more likely to be selected than B or C.

1.22 Here all members of the population are equally likely to be selected and the sample selection mechanism is based on chance. But not every sample of size 2 has the same chance of being selected. For example the sample “B and C” is impossible.

1.23 (a) Since a complete roster of full-time students exists, a simple random sample of 200 students could be taken. If student satisfaction with the quality of campus life randomly fluctuates across the student body, a systematic 1-in-20 sample could also be taken from the population frame. If student satisfaction with the quality of life may differ by gender and by experience/class level, a stratified sample using eight strata, female freshmen through female seniors and male freshmen through male seniors, could be selected. If student satisfaction with the quality of life is thought to fluctuate as much within clusters as between them, a cluster sample could be taken.

(b) A simple random sample is one of the simplest to select. The population frame is the registrar’s file of 4,000 student names.

(c) A systematic sample is easier to select by hand from the registrar’s records than a simple random sample, since an initial person at random is selected and then every 20th person thereafter would be sampled. The systematic sample would have the additional benefit that the alphabetic distribution of sampled students’ names would be more comparable to the alphabetic distribution of student names in the campus population.

(d) If rosters by gender and class designations are readily available, a stratified sample should be taken. Since student satisfaction with the quality of life may indeed differ by gender and class level, the use of a stratified sampling design will not only ensure all strata are represented in the sample, it will also generate a more representative sample and produce estimates of the population parameter that have greater precision.

(e) If all 4,000 full-time students reside in one of 10 on-campus residence halls which fully integrate students by gender and by class, a cluster sample should be taken. A cluster could be defined as an entire residence hall, and the students of a single randomly selected residence hall could be sampled. Since each dormitory has 400 students, a systematic sample of 200 students can then be selected from the chosen cluster of 400 students. Alternately, a cluster could be defined as a floor of one of the 10 dormitories. Suppose there are four floors in each dormitory with 100 students on each floor. Two floors could be randomly sampled to produce the required 200 student sample. Selection of an entire dormitory may make distribution and collection of the survey easier to accomplish. In contrast, if there is some variable other than gender or class that differs across dormitories, sampling by floor may produce a more representative sample.

1.24 (a) Row 16: 2323 6737 5131 8888 1718 0654 6832 4647 6510 4877
Row 17: 4579 4269 2615 1308 2455 7830 5550 5852 5514 7182
Row 18: 0989 3205 0514 2256 8514 4642 7567 8896 2977 8822
Row 19: 5438 2745 9891 4991 4523 6847 9276 8646 1628 3554
Row 20: 9475 0899 2337 0892 0048 8033 6945 9826 9403 6858
Row 21: 7029 7341 3553 1403 3340 4205 0823 4144 1048 2949
Row 22: 8515 7479 5432 9792 6575 5760 0408 8112 2507 3742
Row 23: 1110 0023 4012 8607 4697 9664 4894 3928 7072 5815
Row 24: 3687 1507 7530 5925 7143 1738 1688 5625 8533 5041
Row 25: 2391 3483 5763 3081 6090 5169 0546
Note: All sequences above 5000 are discarded. There were no repeating sequences.

(b) 089 189 289 389 489 589 689 789 889 989
1089 1189 1289 1389 1489 1589 1689 1789 1889 1989
2089 2189 2289 2389 2489 2589 2689 2789 2889 2989
3089 3189 3289 3389 3489 3589 3689 3789 3889 3989
4089 4189 4289 4389 4489 4589 4689 4789 4889 4989

(c) With the single exception of invoice #0989, the invoices selected in the simple random sample are not the same as those selected in the systematic sample. It would be highly unlikely that a random process would select the same units as a systematic process.

1.25 (a) A stratified sample should be taken so that each of the four strata will be proportionately represented.

(b) Since the stratum may differ in the invoice amount, it may be more important to sample a larger percentage of invoices in stratum 1 and stratum 2, and smaller percentages in stratum 3 and stratum 4. For example, 50/5000 = 1% so 1% of 500 = 5 invoices should be selected from stratum 1; similarly 10% = 50 should be selected from stratum 2, 20% = 100 from stratum 3, and 69% = 345 from stratum 4.

(c) It is not simple random sampling because, unlike the simple random sampling, it ensures proportionate representation across the entire population.

1.26 (a) For the third value, Apple is spelled incorrectly. The twelfth value should be Blackberry not Blueberry. The fifteenth value, APPLE, may lead to an irregularity. The eighteenth value should be Samsung not Samsun.

(b) The eighth value is a missing value.

1.27 The second value “2.7MB” and the eighth value “1,079” are potential irregularities.

1.28 (a) The times for each of the hotels would be arranged in separate columns.

(b) The hotel names would be in one column and the times would be in a second column.

1.29 There will be ten records (rows) for each day that can be organized into the following variables (columns): (1) Hotel name, (2) Number of budget-priced rooms occupied, (2) Number of moderate-priced rooms occupied, (3) Number of deluxe—priced rooms occupied, and (4) Total number of rooms occupied.

1.30 Before accepting the results of a survey of college students, you might want to know, for example:
Who funded the survey? Why was it conducted? What was the population from which the sample was selected? What sampling design was used? What mode of response was used: a personal interview, a telephone interview, or a mail survey? Were interviewers trained? Were survey questions field-tested? What questions were asked? Were they clear, accurate, unbiased, valid? What operational definition of "vast majority" was used? What was the response rate? What was the sample size?

1.31 (a) Possible coverage error: Only employees in a specific division of the company were sampled.
(b) Possible nonresponse error: No attempt is made to contact nonrespondents to urge them to complete the evaluation of job satisfaction.
(c) Possible sampling error: The sample statistics obtained from the sample will not be equal to the parameters of interest in the population.
(d) Possible measurement error: Ambiguous wording in questions asked on the questionnaire.

1.32 The results are based on an online survey. If the frame is supposed to be smartphone and tablet users, how is the population defined? This is a self-selecting sample of people who responded online, so there is an undefined nonresponse error. Sampling error cannot be determined since this is not a probability sample.

1.33 Before accepting the results of the survey, you might want to know, for example: Who funded the study? Why was it conducted? What was the population of industries from which the sample was selected? What sampling design was used? What mode of response was used: a personal interview, a telephone interview, or a mail survey? Were interviewers trained? Were survey questions field-tested? What other questions were asked? Were the questions clear, accurate, unbiased, and valid? What was the response rate? What was the margin of error? What was the sample size? What frame was used?

1.34 Before accepting the results of the survey, you might want to know, for example: Who funded the study? Why was it conducted? What was the population of automobile executives from which the sample was selected? What sampling design was used? What mode of response was used: a personal interview, a telephone interview, or a mail survey? Were interviewers trained? Were survey questions field-tested? What other questions were asked? Were the questions clear, accurate, unbiased, and valid? What was the response rate? What was the margin of error? What was the sample size? What frame was used?

1.35 A population contains all the items of interest whereas a sample contains only a portion of the items in the population.

1.36 A statistic is a summary measure describing a sample whereas a parameter is a summary measure describing an entire population.

1.37 Categorical random variables yield categorical responses such as yes or no answers. Numerical random variables yield numerical responses such as your height in inches.

1.38 Discrete random variables produce numerical responses that arise from a counting process. Continuous random variables produce numerical responses that arise from a measuring process.

1.39 Both nominal scaled and ordinal scaled variables are categorical variables but no ranking is implied in nominal scaled variable such as male or female while ranking is implied in ordinal scaled variable such as a student's grade of A, B, C, D and F.

1.40 Both interval scaled and ratio scaled variables are numerical variables in which the difference between measurements is meaningful but an interval scaled variable does not involve a true zero such as standardized exam scores while a ratio scaled variable involves a true zero such as height.

1.41 Items or individuals in a probability sampling are selected based on known probabilities while items or individuals in a nonprobability samplings are selected without knowing their probabilities of selection.

1.42 Missing values most frequently arise when you record a nonresponse to a survey question, but also occur with other data sources for a variety of reasons. Outliers are values that seem excessively different from most of the other values.

1.43 In unstacked arrangement, you create separate numerical variables for each of the multiple groups. In stacked arrangement, you pair the single numerical variable with a second categorical variable that contains multiple categories.

1.44 Microsoft Excel could be used to perform various statistical computations that were possible only with a slide-rule or hand-held calculator in the old days.

1.45 (a) The population of interest is all customers who use mobile shopping.
(b) The sample is the 1,600 U.S. adults who have made purchases via their mobile device and decided to complete the online survey.
(c) One possible answer: A parameter of interest is the proportion of customers who use mobile shopping and spend more time in the app.
(d) A statistic used to estimate the parameter of interest in (c) is the proportion of customers who use mobile shopping and will spend more time in the app out of the 1,600 U.S. adults who complete the online survey.

1.46 The answers to this question depend on which article and its corresponding data set is being selected.

1.47 (a) The population of interest is the CEOs in a wide range of industries representing a mix of company sizes from across three global regions: Asia, Europe, and the Americas.
(b) The sample was the 1,322 CEOs in a wide range of industries representing a mix of company sizes from across three global regions: Asia, Europe, and the Americas surveyed by PwC.
(c) One possible answer: A parameter of interest is the proportion of CEOs in the population who see data mining and analysis as strategically important for their organization.
(d) A statistic used to estimate the parameter of interest in (c) is the proportion of CEOs in the sample who see data mining and analysis as strategically important for their organization.

1.48 Answers will vary.
(a) Place of Birth for the Foreign-Born Population.
(b) Categorical variable.
(c) Not relevant.

1.49 Answers will vary.

(a) How attentive was the salesperson to your needs?

(b) For how many years have you been working in the FIELD? Please count the years from your first job to now.

1.50 (a) The population of interest was the collection of all the 10,000 benefitted employees at the University of Utah when the study was conducted.

(b) The sample consisted of the 3,095 benefitted employees participated in the study.

(c) gender: categorical; age: numerical; education level: numerical; marital status: categorical; household income: numerical; employment category: categorical

1.51 (a) (i) key social media platforms used: categorical
(ii) frequency of social media usage: numerical, discrete
(iii) demographics of key social media platform users: most likely categorical

(b) Answers will vary.
(i) What is your gender?
(ii) What is your ethnic group?
(iii) What is the key social media platform that you use?
(iv) Do you own any Apple products?
(v) Do you consider yourself a savvy internet user?

(c) Answers will vary.
(i) How many hours on average do you spend on the social media platforms?
(ii) How many members are there in your immediate family?
(iii) How many computers (include desktop, laptop and tablets) do you own?
(iv) What is your internet speed (in Mbps) at home?
(v) How many cell phones do you own?