Appendix J

Data Sets and Computer-Based Data Analysis

We'll begin with a definition of a data set, a concept introduced early on in this text. By way of review, a data set may be defined as a bundle or collection of information about one or more variables, typically assembled for the purpose of analysis. It may be that you've collected data on 750 customers, 4500 students, 312 cities, or any other population of interest. Your data set may be very limited (consisting of a small number of cases and variables) or it may be extensive (consisting of hundreds of variables associated with thousands of cases). Despite the variation in size and content, most data sets share certain commonalities in terms of their basic structure. As it turns out, some knowledge about data set structure—even minimal knowledge—can be of substantial benefit to your statistical education in at least two ways.

First, most students who spend any amount of time dealing with statistical applications will eventually find themselves working in a computer-based analytical environment. For example, many statistics courses include an introduction to the use of computers in statistical analysis. Indeed, many statistics courses are structured as two semester courses, with the first semester being devoted to the basics of statistical analysis and the second being devoted to the use of computers in statistical analysis. For other students, the introduction to the use of computers in statistical analysis comes later, maybe in the form of a graduate course or in the world of work. It suffices to say that modern-day statistical analysis is typically undertaken with the assistance of a computer and some rather sophisticated software. In that sense, it's just a practical matter; if you're going to conduct a serious statistical analysis, you'll probably find yourself working with a computer.

The second reason why you should know something about the structure of data sets has to do with the overall learning process. As it turns out, some basic knowledge about the structure of data sets can serve to jump-start your statistical education. At a minimum, it can cause you to starting thinking in terms of cases, variables, and levels of measurement. If you're armed at the outset with a solid understanding of those concepts, your probability of success tends to increase substantially.

With those as reasons enough to take a look at this matter of data set structure, we begin our brief look at the basics of data set structure and the use of computers in statistical analysis. Along the way, you'll discover some good news, along with a few words of caution.

It Usually Starts with Rows and Columns

In the simplest of terms, computer-based data sets are all about rows and columns. As a rule, each case or observation takes one row in the overall format, and different columns are devoted to different variables. There are many different statistical analysis programs on the market, but most share this *row* *and column* format approach. For example, SPSS and SAS are two widely used software programs. While there are differences between them, both rely upon the same general data structure format. Rows are devoted to individual cases or observations; columns are devoted to the different variables. When the underlying software program directs the computer to look at the data, it's directing the system to look at all of the cases or observations in a structured format. What's more, it's telling the system the name that you've assigned to each variable and where each variable will be found. It's really the first step. Specific instructions about what analysis to perform or what type of report to print are matters that come later. What always comes first is the data entry process—the process of entering the information into the computer and letting the computer know the fundamentals about how the data set is structured. To better understand this point, let's take a look at a hypothetical data set.

Imagine for a moment that you had data on 25 different cities. More specifically, let's say that you had the following information on each city:

Name of the city

Total population in 1990

Total population in 2000

Ranking in terms of sales tax revenues collected during 2000 (ranking from 1 to 25)

Median family income in 2000

% of adult population having a college degree

Region of the nation in which the city is located (i.e., north, south, east, or west)

While it's true that a data set with only 25 cases might be small enough to cause you to think about making use of the old fashioned paper/pencil approach, you'd probably want to turn to a computer-based system, at least in the real world. Not only are computer-based systems widely available (and typically at a rather reasonable cost), the matter of data entry is very straightforward. What's more, it's also likely that in the real world you might find yourself working with a much larger number of cases and variables—something that's no problem for the more popular software systems. The SPSS and SAS systems, for example, can easily handle thousands of cases and hundreds of variables.

All of that, though, has to do with capabilities. The issue at hand has to do with how the data set would be structured. Figure J-1 provides an illustration of the general layout that you'd see on a computer terminal screen if you had entered the data described above (i.e., the data on the 25 cities). If you've never really worked with a computer-based statistical analysis program, let me urge you to take a careful look at the illustration. Focus on the overall structure—each row devoted to a single case or observation and each column devoted to a specific variable.

Even if you've taken a close look at Figure J-1, let me ask you to take another look with an eye toward some specific points.

City	1990 Population	2000 Population	Sales Tax Revenue Ranking for 2000	Median Family Income ($)	Adults with College Degree (%)	Region
Arthurville	21,500	32,841	15	31,863	16.51	1
O'Dell Park	15,602	17,611	21	21,336	21.34	1
Lunnville	5282	6328	16	53,119	33.11	3
Bandiville	10,853	12,260	17	42,781	26.81	2
		Continue with data entry through entire data set				
Woodville	31,338	42,132	7	39,388	17.26	4
Lake Grinstead	18,665	21,893	61	41,990	11.59	3
Klepferville	7033	8622	5	39,338	21.53	1
Groves City	24,817	31,992	9	52,167	28.85	4

Figure J-1 Example of Data Set Structure for a Data Set Involving Demographic Data for a Sample of 25 Cities

First, look at the entire illustration with the thought in mind that it could just as easily involve thousands of cases and hundreds of variables (a point that I made earlier in a discussion of the near-limitless capacity of many software packages). Imagine that you could scroll down the terminal screen, with cases appearing, one after another, in a near-endless stream. Similarly, imagine that you're moving across the screen and even more variables begin to appear (e.g., maybe you had 125 different variables in your study). Imagine that additional columns start to appear, again in a near-endless stream.

Secondly, take note of the fact that a row of data constitutes information about a case. For example, the first row of data has information about Arthurville. The second row of data has information about O'Dell Park. And so it goes. One row, one case; another row, another case; all the way through until the last community, Groves City.

Now take a look at the very top of the illustration—the area that is shaded in the illustration. Those are the names that have been given to the different variables. What you should understand at this point is that most statistical analysis packages have great flexibility in this area. For many of the software packages, for example, you can assign very short names to each variable—short names that you'll use when you issue commands to the system (e.g., when you tell the system to compute the mean and standard deviation for the variable of Pop 90). The real flexibility is found in the fact that many of the packages ultimately allow you to assign very detailed, elongated labels to each variable name. In other words, most programs allow you to expand the short name to provide a far more descriptive name or label. As a rule, the expanded names or labels don't come into play until you actually conduct some sort of analysis. When the analysis is complete and the results appear on the screen or on the printer, you'll see the elongated names or labels appear. For example, you may refer to the variable known as Pop 90 when you're issuing instructions to the system, but the results that appear (after you've completed your analysis) will use the expression Population in 1990 Census (if that's the elongated name or label that you've assigned).

In a sense, the business of assigning elongated names or labels is something that occurs in the background, so to speak. Your real focus is on entering and working with the cases and variables that you see on the screen and doing so on the basis of the short variable names that you assigned at the outset. But that's just one of the background elements. Here are just a few other things that are likely to occur in the background:

Each variable is identified in terms of its level of measurement (e.g., nominal, ordinal, interval/ratio).

Each variable is identified as either numeric or alphanumeric—numeric variables being variables expressed in numbers and alphanumeric variables being variables expressed in numbers, alphabetic characters, or both).

The system has been told how to recognize missing information or deal with cases in which some of the information is incomplete.

The system understands the coding system that you're using (e.g., if you use the letter N to stand for North, the system will understand that and will print out results accordingly).

The list of capabilities could go on and on and on. Suffice it to say that contemporary statistical analysis software packages are extremely sophisticated—so much so that a good amount of time can be spent in exploring the capacities of a single package. What's important at this point, though, is just the basic structure of the data set, and that is something that is fairly uniform across the various packages. Just remember the basic rule of thumb: Cases are in rows; variables are in columns.

If you've never dealt with a data set that was structured for use with a computer program, let me offer the following as a suggested exercise. Simply conjure up a study of some sort—a research project that you might like to conduct. It could be a study of students enrolled at a university, customers at a local store of some sort, prime-time television programs, newspaper editorials, court records, or anything else that might cross your mind. Once you've settled on a topic of interest, imagine that you're going to collect information on, let's say, 20 cases (i.e., 20 students, 20 customers, 20 television programs, etc.). Also imagine that you'll be collecting information on specific variables. For example, maybe your goal is to collect information on the age, sex, place of residence, grade point average, and academic major of each student in your sample. Once you have the basics of your study design in your mind, imagine the way the data would look on a computer screen, assuming that you entered the data into a computer-based data set. If you've had experience working with data sets, the exercise may strike you as rather simplistic. If the world of data sets is something very new to you, though, I suspect you're likely to find the exercise to be a very valuable one.

Assuming that you now have some basic understanding of data-set structure, let me offer a few comments about the day-to-day reliance upon computers for statistical analysis. This is where the mixed message comes into the discussion.

Good News; Words of Caution; It's Up to You

Regardless of when you might get directly involved in computer-based data analysis, my guess is that you'll be a little amazed at the capabilities of most statistical analysis software packages. I've already alluded to the rather extraordinary number of cases and variables that most packages can handle, but that's just the start of it. The truly amazing element is the speed at which the data are manipulated and calculations are performed. Extremely sophisticated analyses can be carried out in split seconds and with the highest levels of accuracy. Just to take one example, imagine that you wanted to calculate a simple average (the mean, as it's referred to in statistical parlance), but you wanted to calculate that average for 127 different variables with a sample involving 38,294 cases. All you have to do is type in a couple of commands, tell the system to go to work, and your results will appear in the blink of an eye.

All of that should be very good news for anyone who's venturing into the world of statistical analysis for the first time. If that's where you are—if you're just beginning your first systematic study of statistical analysis—you might do well to always remember that the sophisticated software is, for the most part, readily available. In doing so, you can take comfort in the fact that you could most likely rely upon some very user-friendly software to do part of the job for you. Consequently, your mind should be freed up a bit for more important matters—important matters such as selecting the appropriate statistical procedures and interpreting the results. Just to set your mind at ease, let me repeat: You can take comfort in the fact that serious statistical analysis is typically done with the assistance of a computer. The days of pencils, paper, and tedious calculations are over. On the other hand, you're never free of the responsibility of knowing how to select and interpret the appropriate statistical procedure.

All of that, of course, returns us to a point raised earlier—namely the importance of developing a solid understanding of the underlying conceptual elements involved in statistical analysis. Statistical calculations represent only one part of the equation, so to speak. The other part—indeed, the most important part—has to do with the logical and conceptual basis of statistical analysis. Simply put, you can always rely upon statistical software to carry out complex calculations, but selecting the appropriate procedure and interpreting the results is something that falls to you.

As to what that means when you're working your way through this or any other text, let me offer the following approach. You should always be careful in your calculations. You should strive for precision. But you should never look at a statistical task with a focus on how long it might take you to work your way through the problem. If there are highly tedious steps involved in a particular procedure, just accept the fact that it's part of the process and there's little you can do except work your way through it. Don't let some temporary frustration about tedious procedures block your understanding of the underlying logic or conceptual basis. In the final analysis, it's your understanding of the underlying logic and conceptual basis that will pay off.

In short, it's probably a good idea to remind yourself every now and then that you could, if push came to shove, rely upon computer-based data analysis for almost any sort of statistical analysis. In doing so, you're apt to lower your stress level, at least to some degree. But when you do that, you'd be well served to keep your mind focused on the more important issues—logic and concepts.

Appendix K

Some of the More Common Formulas Used in the Text

$\mu = \dfrac{\sum X}{N}$ *Mean of a population*

$\overline{X} = \dfrac{\sum X}{n}$ *Mean of a sample*

$\sigma^2 = \dfrac{\sum (X - \mu)^2}{N}$ *Variance of a population*

$s^2 = \dfrac{\sum (X - \overline{X})^2}{n - 1}$ *Variance of a sample*

$\sigma = \sqrt{\dfrac{\sum (X - \mu)^2}{N}}$ *Standard deviation of a population*

$s = \sqrt{\dfrac{\sum (X - \overline{X})^2}{n - 1}}$ *Standard deviation of a sample*

$\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ *Standard error of the mean*

$s_{\overline{X}} = \dfrac{s}{\sqrt{n}}$ *Estimate of the standard error of the mean*

$S_p = \sqrt{\dfrac{P(1 - P)}{n}}$ *Estimate of the standard error of the proportion*

$\mathrm{CI} = \overline{X} \pm Z(\sigma_{\overline{X}})$ *Confidence interval for the mean (σ known)*

$\mathrm{CI} = \overline{X} \pm t(s_{\overline{x}})$ *Confidence interval for the mean (σ unknown)*

$\mathrm{CI} = P \pm Z(s_P)$ *Confidence interval for the proportion*

$\overline{D} = \dfrac{\sum d}{n}$ *Mean difference*

$s_d = \sqrt{\dfrac{\sum (d - \overline{D})^2}{n - 1}}$ *Standard deviation of the differences*

$s_{\overline{D}} = \dfrac{s_d}{\sqrt{n}}$ *Estimate of the standard error of the mean difference*

$$s_{\overline{X}1-\overline{X}2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \bullet \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$ *Estimate of the standard error of the difference between means*

$$Z = \frac{X - \mu}{\sigma}$$ *Conversion of a raw score in a population to a Z score*

$$Z = \frac{X - \overline{X}}{s}$$ *Conversion of a raw score in a sample to a Z score*

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$ *Single sample test involving a mean with σ known*

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$ *Single sample test involving a mean with σ unknown*

$$t = \frac{\overline{D}}{S_{\overline{D}}}$$ *Two sample test involving mean difference (matched or related samples)*

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}1-\overline{X}2}}$$ *Two sample test involving difference between means (independent samples)*

$$F = \frac{MS_B}{MS_W}$$ *F ratio for Analysis of Variance*

$$MS_B = \frac{SS_B}{df_B}$$ *Mean square between*

$$MS_W = \frac{SS_W}{df_W}$$ *Mean square within*

$$\chi^2 = \sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$$ *Chi-Square Test*

$$r = \frac{\sum(Z_X \bullet Z_Y)}{n - 1}$$ *Correlation Coefficient*

Answers to Chapter Problems

CHAPTER 1

1. Academic major; test performance
2. Gender; attitude toward abortion
3. Nominal
4. Ordinal
5. Interval or interval/ratio
6. Nominal
7. Ratio
8. Ordinal
9. 500; 23,419
10. Sample; population
11. Statistics; parameters
12. Descriptive; inferential

CHAPTER 2

General Thought Questions

1. Mean, median, and mode
2. Mean
3. Mode
4. Median
5. Range; dispersion
6. 82
7. Bi-modal distribution; the modes are 18 and 21
8. Mean deviation or average deviation
9. 0
10. 0; squaring
11. Square root
12. $n - 1$; n

Application Questions/Problems

1. **a.** 4.2; **b.** 3rd score; **c.** 4; **d.** 1.84; **e.** 5.70; **f.** 2.39
2. **a.** 15.88; **b.** 4.5th score; **c.** 15.5; **d.** 12; **e.** 2.88; **f.** 12.41; **g.** 3.52
3. **a.** 3.89; **b.** 5th score; **c.** 4; **d.** 1 and 4; **e.** 1.90; **f.** 6.61; **g.** 2.57
4. **a.** 4.67; **b.** 4.50; **c.** 7; **d.** 2.35
5. 3
6. **a.** 3.20; **b.** 2.10
7. **a.** 2; **b.** .50
8. 73
9. 170

CHAPTER 3

General Thought Questions

1. Symmetrical
2. Right; left
3. Inflection
4. 1
5. Coincide (or are equal)

Application Questions/Problems

1. 68%
2. 95%
3. 99%
4. 50%; 50%
5. 2
6. 78
7. 80
8. 140
9. 950

CHAPTER 4

General Thought Questions

1. Infinite
2. 0; 1

Application Questions/Problems

1. 44.84%
2. 49.06%
3. 38.88%
4. 2.5%
5. 2.5%
6. .5%
7. .5%
8. 6.30%
9. 13.14%
10. approximately −.84
11. approximately .39
12. approximately ±.84

CHAPTER 5

General Thought Questions

1. Equal
2. Does not
3. All
4. Sampling frame
5. Sample
6. Sampling error
7. Error
8. Means
9. Mean
10. Standard error of the mean
11. Standard deviation; square root
12. Normal

Application Questions/Problems

1. 24.12; .40
2. 30; .40
3. 120; 3
4. 615; 4.50
5. 55; 1.70

CHAPTER 6

General Thought Questions

1. Mean
2. Estimate; sample
3. Mean
4. Decreases
5. Inverse
6. Width
7. Increase; decrease
8. σ divided by the square root of n
9. s divided by the square root of n

Application Questions/Problems: Confidence Interval for the Mean With σ Known

1. **a.** 12.50; **b.** 14.14; **c.** 8.77; 15.00
2. **a.** 145.30–154.70;
   b. 143.81–156.19
3. **a.** 51.73–56.27; **b.** 51.01–56.99
4. **a.** 74.02–75.98; **b.** 73.71–76.29
5. 76.24–79.76
6. 488.20–507.80
7. 515.47–528.53
8. 513.41–530.59
9. 108.06–111.94

Application Questions/Problems: Confidence Interval for the Mean With σ Unknown

1. **a.** 1.25; **b.** 2.50; **c.** 2.58; **d.** 6.48
2. **a.** 24.14–27.86; **b.** 23.49–28.51
3. **a.** 360.96–443.04;
   b. 346.69–457.31
4. **a.** 73.81–86.19;
   b. 71.61–88.39
5. 3.68–5.12
6. $35.52–$41.98
7. 81.05–90.95
8. 94.40–107.60
9. 5.83 ounces–6.89 ounces

Application Questions/Problems: Confidence Interval for the Proportion

1. 33.22%–46.78%
2. 9.68%–30.32%
3. 25.51%–38.49%
4. 7.47%–22.53%
5. 8.67%–17.33%
6. 57.51%–64.49%
7. 54.96%–67.04%
8. 71.06%–76.04%
9. 70.39%–76.71%

CHAPTER 7

General Thought Questions

1. Type I
2. Type I
3. Type II
4. Null hypothesis

5. Level of significance
6. Critical region
7. Fail to reject
8. .05 and .01
9. Region of rejection or critical region; null hypothesis

Application Questions/Problems: Hypothesis Involving a Single Sample Mean With σ Known

1. H_0: $\mu = 6.88$; b; $Z = 4.03$; Reject the null at the .05 level.
2. H_0: $\mu = 72.55$; b; $Z = 2.77$; Reject the null at the .05 level.
3. H_0: $\mu = 61$; b; $Z = -2.94$; Reject the null at the .05 level.
4. H_0: $\mu = 10.45$; b; $Z = 1.67$; Fail to reject the null at the .05 level.
5. H_0: $\mu = 155$; b; $Z = 3.33$; Reject the null at the .05 level.
6. H_0: $\mu = 75$; b; $Z = 2.00$; Reject the null at the .05 level.

Application Questions/Problems: Hypothesis Test Involving a Single Sample Mean With σ Unknown

1. H_0: $\mu = 8.45$; b; $t = -3.53$; Critical value = 2.045; Reject the null at the .05 level.
2. H_0: $\mu = 8.25$; b; $t = -.99$; Critical value = 2.160; Fail to reject the null at the .05 level.
3. H_0: $\mu = 15.23$; b; $t = -5.78$; Critical value = 2.064; Reject the null hypothesis at the .05 level.
4. H_0: $\mu = 10.65$; b; $t = 2.66$; Critical value = 2.042; Reject the null hypothesis at the .05 level.
5. H_0: $\mu = 12.16$; b; $t = -1.48$; Critical value = 2.064; Fail to reject the null hypothesis at the .05 level.
6. H_0: $\mu = 12.56$; b; $t = -1.90$; Critical value = 2.045; Fail to reject the null hypothesis at the .05 level.

CHAPTER 8

General Thought Questions

1. True
2. False
3. Mean differences
4. The difference between means

Application Questions/Problems: Matched/Related Samples Design

1. **a.** H_0: $\mu_{\overline{D}} = 0$; **b.** $t = 2.81$; **c.** Critical value = 2.145; **d.** Reject the null at .05 level.
2. **a.** H_0: $\mu_{\overline{D}} = 0$; **b.** $t = 1.54$; **c.** Critical value = 2.064; **d.** Fail to reject the null at .05 level.
3. **a.** H_0: $\mu_{\overline{D}} = 0$; **b.** $t = 2.57$; **c.** Critical value = 2.045; **d.** Reject the null at .05 level.
4. **a.** H_0: $\mu_{\overline{D}} = 0$; **b.** $t = 2.28$; **c.** Critical value = 2.048; **d.** Reject the null at .05 level.
5. **a.** H_0: $\mu_{\overline{D}} = 0$; **b.** $t = 1.70$; **c.** Critical value = 2.045; **d.** Fail to reject the null at .05 level.

Application Questions/Problems: Independent Samples Design

1. **a.** H_0: $\mu_1 - \mu_2 = 0$; **b.** $t = 2.97$; **c.** Critical value = 2.042; **d.** Reject the null at .05 level.
2. **a.** H_0: $\mu_1 - \mu_2 = 0$; **b.** $t = 1.82$; **c.** Critical value = 2.009; **d.** Fail to reject the null at .05 level.
3. **a.** H_0: $\mu_1 - \mu_2 = 0$; **b.** $t = -1.47$; **c.** Critical value = 2.056; **d.** Fail to reject the null at .05 level.
4. **a.** H_0: $\mu_1 - \mu_2 = 0$; **b.** $t = -1.58$; **c.** Critical value = 2.048; **d.** Fail to reject the null at .05 level.
5. **a.** H_0: $\mu_1 - \mu_2 = 0$; **b.** $t = -2.93$; **c.** Critical value = 2.042 (use critical value for 30 degrees of freedom); **d.** Reject the null at .05 level.

CHAPTER 9

General Thought Questions

1. Alternative or research
2. Directional hypothesis
3. Two-tailed
4. One-tailed
5. Rejecting; true
6. Failing to reject; false

Application Questions/Problems: Alternative or Research Hypotheses

1. **a.** H_0: There is no significant difference between on-campus and commuter students with respect to grade point average.
   b. H_1: There is a significant difference between on-campus and commuter students with respect to grade point average.
   c. H_2: On-campus students have a significantly higher grade point average than commuter students.
   d. H_3: Commuter students have a significantly higher grade point average than on-campus students.
2. **a.** H_0: There is no significant difference in length of sentences handed out to white and non-white defendants in first-offense drug trafficking cases.
   b. H_1: In first-offense drug trafficking cases, the length of sentence handed out to non-white defendants is significantly different than the length of sentence handed out to white defendants.
   c. H_2: In first-offense drug trafficking cases, the length of sentence handed out to non-white defendants is significantly higher than the length of sentence handed out to white defendants.
   d. H_3: In first-offense drug trafficking cases, the length of sentence handed out to white defendants is significantly higher than the length of sentence handed out to non-white defendants.
3. **a.** H_0: There is no significant difference between rural and urban areas in terms of levels of voter participation.
   b. H_1: There is a significant difference between rural and urban areas in terms of levels of voter participation.
   c. H_2: The level of voter participation is significantly higher in rural areas than it is in urban areas.
   d. H_3: The level of voter participation is significantly higher in urban areas than it is in rural areas.
4. **a.** H_0: There is no significant difference between the levels of water pollution in creeks in the southern part of the state and levels of water pollution in creeks in the northern part of the state.
   b. H_1: There is a significant difference between the levels of water pollution in creeks in the southern part of the state and levels of water pollution in creeks in the northern part of the state.
   c. H_2: Levels of water pollution in creeks in the southern part of the state are significantly higher than levels of water pollution in creeks in the northern part of the state.
   d. H_3: Levels of water pollution in creeks in the northern part of the state are significantly higher than levels of water pollution in creeks in the southern part of the state.

Application Questions/Problems: One-tailed and Two-tailed Critical Values

1. **a.** 1.96; **b.** 1.64; **c.** 2.58; **d.** 2.33
2. **a.** 2.131; **b.** 1.721; **c.** 1.74; **d.** 1.734

CHAPTER 10

General Thought Questions

1. F
2. Between; within
3. Find the difference or deviation between each score and the mean of each category; square the deviations; add the squared deviations; sum the squared deviations across all categories.
4. Find the difference or deviation between each category mean and the grand mean; square the deviations; multiply the squared deviations in each category by the number of cases in the category; sum across all categories.
5. Degrees of freedom
6. Degrees of freedom
7. $n - k$
8. $k - 1$
9. Mean square between
10. Mean square within
11. $\mu_1 = \mu_2 = \mu_3$

Application Questions/Problems

1. **a.** 5; **b.** 30
2. **a.** 3.59; **b.** 3.01; **c.** Reject the null at the .05 level.

3. **a.** 2.69; **b.** 2.98; **c.** Fail to reject the null at the .05 level.
4. **a.** 6.77; **b.** 3.40; **c.** Reject the null at the .05 level.
5. **a.** $\mu_1 = \mu_2 = \mu_3$; **b.** Sample 1 = 8.00, Sample 2 = 6.00, Sample 3 = 9.00; **c.** 7.79; **d.** 40.64; **e.** 170.00; **f.** 2; **g.** 26; **h.** 6.54; **i.** 20.32; **j.** 3.11; **k.** Fail to reject the null at the .05 level.
6. **a.** $\mu_1 = \mu_2 = \mu_3 = \mu_4$; **b.** Northern = 4.00, Southern = 4.00, Eastern = 6.00, Western = 5.00; **c.** 4.74; **d.** 25.43; **e.** 122.00; **f.** 3; **g.** 35; **h.** 3.49; **i.** 8.48; **j.** 2.43; **k.** Fail to reject the null at the .05 level.
7. **a.** $\mu_1 = \mu_2 = \mu_3$; **b.** Day Shift = 4.40, Afternoon Shift = 4.75, Night Shift = 4.60; **c.** 4.57; **d.** .27; **e.** 19.14; **f.** 2; **g.** 11; **h.** 1.74; **i.** .14; **j.** .08; **k.** Fail to reject the null at the .05 level.
8. **a.** $\mu_1 = \mu_2 = \mu_3$; **b.** Male = 3.17, Female = 7.17, Mixed Gender = 5.17; **c.** 5.17; **d.** 48; **e.** 34.52; **f.** 2; **g.** 15; **h.** 2.30; **i.** 24; **j.** 10.43; **k.** Reject the null at the .05 level.

CHAPTER 11

General Thought Questions

1. Contingency
2. Categorical
3. Observed
4. Expected
5. (row total × column total)/*n*
6. $(r - 1) \times (c - 1)$
7. 4
8. 12
9. 15
10. 8

Application Questions/Problems

1. 9.488; Fail to reject the null hypothesis at the .05 level.
2. 21.026; Reject the null hypothesis at the .05 level.
3. 3.841; Reject the null hypothesis at the .05 level.
4. 9.488; Reject the null hypothesis at the .05 level.
5. 21.026; Fail to reject the null hypothesis at the .05 level.
6. **a.** 3 degrees of freedom; **b.** .10; Fail to reject the null at the .05 level.
7. **a.** 4 degrees of freedom;
   b. 18.35; Reject the null at the .05 level.
8. **a.** 2 degrees of freedom;
   b. .27; Fail to reject the null at the .05 level.

CHAPTER 12

General Thought Questions

1. −1.00
2. +1.00
3. −1.00 to +1.00
4. No association
5. Scatter plot
6. Correlation
7. Determination
8. Line of best fit; least squares line
9. $a + bx$
10. Y'
11. a
12. b

Application Questions/Problems

1. **a.** 6.38; **b.** 2.96; **c.** 6; **d.** 2.55; **e.** 11.11; **f.** .93
2. **a.** 10.00; **b.** 1.56; **c.** 56.00; **d.** 15.60; **e.** 3.53; **f.** .39
3. **a.** −.89; This is a strong, negative relationship; **b.** .79; 79% of the variation in Y is attributable to variation in X; **c.** $r = 0$; Reject the null at the .05 level
4. **a.** .39; This is a weak, positive relationship; **b.** .10; 10% of the variation in Y is attributable to variation in X; **c.** $r = 0$; Reject the null at the .05 level
5. **a.** .90; This is a strong, positive relationship; **b.** .81; 81% of the variation in Y is attributable to variation in X; **c.** $r = 0$; Reject the null at the .05 level
6. $a = 29.91$ and $b = -.23$
7. $a = 17$ and $b = 4.15$
8. **a.** −.998; This is a strong, negative relationship; **b.** .996; 99.6% of the variation in Y is attributable to variation in X; **c.** $a = 4.12$ and $b = -.04$
9. **a.** .83; This is a strong, positive relationship; **b.** .69; 69% of the variation in Y is attributable to variation in X; **c.** $a = -\$32.01$ and $b = 3.67$
10. $41.39

Glossary

1-2-3 Rule A statement of how much area under the normal curve is found between ±1, ±2, and ±3 standard deviations from the mean.

***a* term in the regression equation ($Y' = a + bX$)** The *Y*-intercept; the point at which the regression line crosses the *Y*-axis.

alternative hypothesis A hypothesis that stands in opposition to the null hypothesis. It may be directional or nondirectional.

ANOVA (analysis of variance, one-way) A test to determine if there is a significant difference among three or more groups or samples.

average deviation See *mean deviation.*

***b* term in the regression equation ($Y' = a + bX$)** The slope of the regression line; the change in *Y* that accompanies a unit change in *X*.

between-groups degrees of freedom The number of degrees of freedom associated with the estimate of between-groups variance; equivalent to the number of groups minus 1.

between-groups estimate of variance See *mean square between.*

between-groups sum of squares The sum of the squared deviation of each sample mean from the grand mean, weighted by the number of cases in each sample, and summed across all samples.

bimodal distribution A distribution with two modes.

calculated test statistic The result of a hypothesis-testing procedure; the value that is compared to a critical value when testing the null hypothesis.

categorical data Information obtained on variables measured at the nominal or ordinal level; responses that can be classified into categories.

Central Limit Theorem A statement about the relationship between a population and a sampling distribution based on that population. The Central Limit Theorem is stated as follows:

> If repeated random samples of size *n* are taken from a population with a mean or mu (μ) and a standard deviation (σ), the sampling distribution of sample means will have a mean equal to mu (μ) and a standard error equal to $\frac{\sigma}{\sqrt{n}}$. Moreover, as *n* increases the sampling distribution will approach a normal distribution.

central tendency The center or typicality of a distribution. The three most common measures of central tendency are the *mean, median,* and *mode.*

chi-square test of independence A test to determine whether there is an association between two categorical variables.

coefficient of determination The value of r^2; a measure of the amount of variation in Y that is attributable to variation in X.

confidence interval for a proportion A statement of two values (or an interval) within which you believe the true proportion of the population is found.

confidence interval for the mean A statement of two values (or an interval) within which you believe the true mean of the population (μ or mu) is found.

contingency table A classification tool that reveals the various possibilities (contingencies) in the comparison of variables; a table that presents data in terms of all combinations of two or more variables.

correlation A procedure designed to determine the strength and direction of an association between two interval/ratio level variables. Also known as Pearson's *r*.

correlation coefficient The value of *r*; a measure of the strength and direction of an association between two interval/ratio level variables. The value of *r* can range from –1.0 to +1.0.

critical region The portion of a sampling distribution that contains all the values that allow you to reject the null hypothesis. If the calculated test statistic (e.g., *Z* or *t*) falls within the critical region, the null can be rejected.

critical value The point on a sampling distribution that marks the beginning of the critical region; the value that is used as a point of comparison when making a decision about a null hypothesis. If the calculated test statistic (e.g., *Z* or *t*) meets or exceeds the critical value, the null hypothesis can be rejected.

curvilinear association An association between two variables that would, if represented in a scatter plot, conform to a general pattern of a curved line.

data Information.

data distribution A listing of the values or responses associated with a particular variable in a data set.

data point The individual pieces of information in a data set.

data set The collection or bundle of information relative to specific variables.

dependent variable The variable that's presumed to be influenced by another variable.

descriptive statistics Statistical procedures used to summarize or describe data.

directional hypothesis An alternative or research hypothesis that specifies the nature or direction of a hypothesized difference. It asserts that there will be a difference or a change in a particular direction (increase or decrease).

dispersion (variability) The extent to which the scores in a distribution are spread around the mean value or throughout the distribution. The two most commonly used measures of dispersion are the *variance* and the *standard deviation*.

effect The change in a measurement that is attributable to a treatment condition or stimulus of some sort.

estimate of the standard error of the mean An estimate of the standard deviation of the sampling distribution of sample means; a function of the standard deviation of a sample.

expected frequency The frequency that would be expected to occur in a particular cell, given the marginal distributions and the total number of cases in the table.

***F* ratio** The ratio of the between-groups estimate of variance to the within-groups estimate of variance. The *F* ratio is frequently referred to as the ratio of the mean square between to the mean square within.

family of *t* distributions A series of sampling distributions (of the *t* statistic) developed by Gossett. The shape of any one distribution is a function of sample size (or degrees of freedom, equal to $n - 1$).

frequency distribution A table or graph that indicates how many times a value or score appears in a set of values or scores.

grand (overall) mean The mean that would result if the values of all cases in an ANOVA application were added and the sum divided by the total number of cases.

group (sample) mean The mean of an individual sample in an ANOVA application.

hypothesis A statement of expectations. See also *null hypothesis* and *alternative hypothesis.*

independent samples Samples selected in such a manner that the selection of any case in no way affects the selection of any other case.

independent variable The variable that's presumed to influence another variable.

inferential statistics Statistical procedures used to make statements or inferences about a population, based on sample statistics.

interval level of measurement A system of measurement based on an underlying scale of equal intervals. See also *interval/ratio level of measurement* and *ratio level of measurement.*

interval/ratio level of measurement Since there is no practical difference between the interval and ratio levels of measurement when it comes to statistical analysis, the terms are often combined to refer to any scale of measurement that is either interval or ratio.

least squares line See *line of best fit.*

level of confidence The amount of confidence that can be placed in an estimate derived from the construction of a confidence interval. Level of confidence is mathematically defined as 1 minus the level of significance. The level of confidence is a statement of the percentage of times (99%, 95%, etc.) one would obtain a correct confidence interval if one repeatedly constructed confidence intervals for repeated samples from the same population.

level of significance The probability of making a Type I error.

linear association An association between two variables that would, if represented in a scatter plot, conform to a general pattern of a straight line.

line of best fit The line that passes through a scatter plot in such a way that the square of the distance from each point in the plot to the line is at a minimum. Also known as the *regression line* or the *least squares line.*

margin of error A term used to express the width of a confidence interval for a proportion.

marginal totals The row and column totals that are presented in the margins of a table.

matched or related samples Samples selected in such a manner that cases included in one sample are somehow related or matched to cases in another sample. In some instances, the matching is achieved by using the same subjects tested in two situations (for example, in a before/after test situation). In other instances, the matching is achieved by matching subjects or cases on the basis of relevant criteria.

mean The most widely used measure of central tendency. The mean is calculated by summing all the scores in a distribution and dividing the sum by the total number of cases in the distribution.

mean deviation An infrequently used measure of dispersion based, in part, on the absolute deviations from the mean of the distribution. Also known as the *average deviation.*

mean square between The between-groups estimate of variance; calculated by dividing the between-groups sum of squares by the between-groups degrees of freedom.

mean square within The within-groups estimate of variance; calculated by dividing the within-groups sum of squares by the within-groups degrees of freedom.

median The score that divides a distribution in half; the midpoint of a distribution, or the point above and below which one-half of the scores or values are located. The formula for the median is a positional formula; it will tell you the position of the median in the distribution, not its value.

mode The response or value that appears most frequently in a distribution. The mode is the only measure of central tendency that is appropriate for nominal level data.

mu (μ) The mean of a population.

negative (inverse) association A pattern of association in which the variables track in opposite directions; as one variable increases in value, the other variable decreases in value.

negative skew The shape of a distribution that includes some extremely low scores or values. A distribution is said to have a negative skew if the tail of the distribution points toward the left.

nominal level of measurement The simplest level of measurement; a system of measurement based on categories that are mutually exclusive and collectively exhaustive.

non-directional hypothesis An alternative or research hypothesis that does not specify the nature or direction of a hypothesized difference. It simply asserts that a difference will be present.

normal curve A unimodal, symmetrial curve that is mathematically defined on the basis of the mean and standard deviation of an underlying distribution.

null hypothesis A statement of equality; a statement of no difference; a statement of chance. In the case of a hypothesis test involving a single sample mean (that is compared to a known population mean), the null is typically a statement of the value of the population mean.

observed frequency The result or frequency presented in each cell of a contingency table.

one-tailed test situation A research situation in which the researcher is looking for an extreme difference that is located on only one side of the distribution.

ordinal level of measurement A level of measurement that presumes the notion of order (greater than and lesser than).

parameter A characteristic of a population. Compare *statistic.*

Pearson's *r* See *correlation.*

perfect association A pattern of association between variables in which there is perfect predictability; knowledge of the value of one variable allows a precise prediction of the value of the other variable.

point of inflection The point at which a normal curve begins to change direction. It is one standard deviation above or below the mean of the underlying distribution.

population All possible cases; sometimes referred to as the *universe.* It is often thought of as the total collection of cases that you're interested in.

positive (direct) association A pattern of association in which the variables track in the same direction; as one variable increases in value, the other variable increases in value.

positive skew The shape of a distribution that includes some extremely high scores or values. A distribution is said to have a positive skew if the tail of the distribution points toward the right.

power The ability of a test to reject a false null hypothesis.

random sample A sample selected in such a way that every unit has an equal chance of being selected, and the selection of any one unit in no way affects the selection of any other unit. In a random sample, all combinations are possible.

range A statement of the difference between the highest and lowest scores or values in a distribution. As a measure of dispersion or variability, the range is simple to calculate, but it doesn't say much about the distribution.

ratio level of measurement A level of measurement that has all the properties of the interval level of measurement, plus the presence (or possibility) of a true or legitimate zero (0) point. See *interval/ratio level of measurement.*

region of rejection See *critical region.*

regression analysis A technique that allows the use of existing data to predict future values.

regression equation The equation that describes the path of the line of best fit. The regression equation is used to predict a value of Y (referred to as Y' or Y-prime) on the basis of an X value ($Y' = a + bX$).

regression line See *line of best fit.*

research hypothesis See *alternative hypothesis.*

sample A portion of a population.

sampling distribution of sample means The result you would get if you took repeated samples from a given population, calculated the mean for each sample, and plotted the sample means.

sampling error The difference between a sample statistic and a population parameter that is due to chance.

sampling frame A physical representation of the population; a listing of all the elements in a population.

scatter plot A visual representation of the values of two variables on a case-by-case basis.

skewed distribution A distribution that departs from symmetry, in the sense that most of the cases are concentrated at one end of the distribution.

standard deviation A widely used measure of dispersion or variability. The standard deviation is the square root of the variance.

standard error of the difference of means The standard deviation of a sampling distribution of the difference between two sample means. The sampling distribution, in this case, would be the result of repeated sampling—each time taking two samples, calculating the mean of each sample, calculating the difference between the means, and recording/plotting the differences. The standard error would be the standard deviation of the sampling distribution.

standard error of the estimate An overall measure of the difference between actual and predicted values of *Y*.

standard error of the mean The standard deviation of a sampling distribution of sample means.

standard error of the mean difference The standard deviation of a sampling distribution of mean differences between scores reflected in two samples. The sampling distribution, in this case, would be the result of repeated sampling—each time looking at two related samples, and focusing on the difference between the individual scores in each sample. The individual differences would be treated as forming a distribution, and that distribution has a mean. The repeated samplings would result in repeated mean differences. The recording/plotting of those mean differences would constitute the sampling distribution. The standard error would be the standard deviation of the sampling distribution.

standardized normal curve A unimodal, symmetrical, theoretical distribution based on an infinite number of cases, having a mean of 0 and a standard deviation of 1.

statistic A characteristic of a sample. Compare *parameter.*

strength of association The extent to which the value of one variable can be predicted on the basis of the value of another variable.

symmetrical distribution A distribution in which the two halves are mirror images of each other.

table of areas under the normal curve A table of values that tell you what proportion of the area under the normal curve is found between the mean and any *Z* value.

tail of the distribution In a skewed distribution, the elongated portion of the curve.

two-tailed test scenario A research situation in which the researcher is looking for an extreme difference that could be located at either end of the distribution.

Type I error Rejection of the null hypothesis when the null is true.

Type II error Failure to reject the null hypothesis when the null is false.

unimodal distribution A distribution with only one mode.

universe See *population.*

variable Anything that can take on different quantities or qualities; anything that can vary.

variance A widely used measure of dispersion or variability. The variance is equal to the standard deviation squared.

within-groups degrees of freedom The number of degrees of freedom associated with the within-groups estimate of variance; equivalent to the number of cases minus the number of groups.

within-groups estimate of variance See *mean square within.*

within-groups sum of squares The sum of the squared deviations of each score from its sample mean, summed across all samples.

***Y* prime (*Y′*)** The *Y* value that you are attempting to predict, based on a given value for *X* and the regression equation.

***Z* (*Z* score)** A point along the baseline of a standardized normal curve.

***Z* ratio** The result of finding the difference between a raw score and a mean, and dividing the difference by the standard deviation. This procedure converts a raw score into a *Z* score.

References

Cuzzort, R. P., & Vrettos, J. S. (1996). *The elementary forms of statistical reason*. New York: St. Martin's.

Dunn, D. S. (2001). *Statistics and data analysis for the behavioral sciences*. New York: McGraw-Hill.

Elifson, K. W., Runyon, R. P., & Haber, A. (1990). *Fundamentals of social statistics* (2nd ed.). New York: McGraw-Hill.

Gravetter, F. J., & Wallnau, L. B. (1999). *Essentials of statistics for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth.

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of statistics for the behavioral sciences* (4th ed.). Pacific Grove, CA: Wadsworth.

Healy, J. F. (2002). *Statistics: A tool for social research* (6th ed.). Belmont, CA: Wadsworth.

Howell, D. C. (1995). *Fundamental statistics for the behavioral sciences* (3rd ed.). Belmont, CA: Duxbury.

Hurlburt, R. T. (1998). *Comprehending behavioral statistics* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York: Radius.

Moore, D. S. (2000). *The basic practice of statistics* (2nd ed.). New York: W. H. Freeman.

Pagano, R. R. (2001). *Understanding statistics in the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth.

Popper, K. R. (1961). *The logic of scientific discovery*. New York: Science Editions.

Pryczak, F. (1995). *Making sense of statistics: A conceptual overview*. Los Angeles, CA: Pryczak Publishing.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Russell, B. (1955). *Nightmares of eminent persons, and other stories*. New York: Simon & Schuster.

Salkind, N. J. (2000). *Statistics for people who think they hate statistics*. Thousand Oaks, CA: Sage.

Utts, J. M., & Heckard, R. F. (2002). *Mind on statistics*. Pacific Grove, CA: Duxbury.

Index

R

S