

## 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 start 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 |
|---|-----------------|-----------------|------------------------------------|---------------------------|--------------------------------|--------|
| <i>Arthurville</i>                                      | 21,500          | 32,841          | 15                                 | 31,863                    | 16.51                          | 1      |
| <i>O'Dell Park</i>                                      | 15,602          | 17,611          | 21                                 | 21,336                    | 21.34                          | 1      |
| <i>Lunnville</i>  | 5282            | 6328            | 16                                 | 53,119                    | 33.11                          | 3      |
| <i>Bandiville</i>                                       | 10,853          | 12,260          | 17                                 | 42,781                    | 26.81                          | 2      |
| <i>Continue with data entry through entire data set</i> |                 |                 |                                    |                           |                                |        |
| <i>Woodville</i>  | 31,338          | 42,132          | 7                                  | 39,388                    | 17.26                          | 4      |
| <i>Lake Grinstead</i>                                   | 18,665          | 21,893          | 61                                 | 41,990                    | 11.59                          | 3      |
| <i>Klepferville</i>                                     | 7033            | 8622            | 5                                  | 39,338                    | 21.53                          | 1      |
| <i>Groves City</i>                                      | 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

$$\mu = \frac{\sum X}{N} \quad \text{Mean of a population}$$

$$\bar{X} = \frac{\sum X}{n} \quad \text{Mean of a sample}$$

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \quad \text{Variance of a population}$$

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1} \quad \text{Variance of a sample}$$

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

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \quad \text{Standard deviation of a sample}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{Standard error of the mean}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \quad \text{Estimate of the standard error of the mean}$$

$$S_p = \sqrt{\frac{P(1 - P)}{n}} \quad \text{Estimate of the standard error of the proportion}$$

$$CI = \bar{X} \pm Z(\sigma_{\bar{X}}) \quad \text{Confidence interval for the mean } (\sigma \text{ known})$$

$$CI = \bar{X} \pm t(s_{\bar{X}}) \quad \text{Confidence interval for the mean } (\sigma \text{ unknown})$$

$$CI = P \pm Z(s_p) \quad \text{Confidence interval for the proportion}$$

$$\bar{D} = \frac{\sum d}{n} \quad \text{Mean difference}$$

$$s_d = \sqrt{\frac{\sum (d - \bar{D})^2}{n - 1}} \quad \text{Standard deviation of the differences}$$

$$s_{\bar{D}} = \frac{s_d}{\sqrt{n}} \quad \text{Estimate of the standard error of the mean difference}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \cdot \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}$$

*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*

$$Z = \frac{X - \bar{X}}{s}$$

*Conversion of a raw score in a sample to a Z score*

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

*Single sample test involving a mean with  $\sigma$  known*

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}}$$

*Single sample test involving a mean with  $\sigma$  unknown*

$$t = \frac{\bar{D}}{S_{\bar{D}}}$$

*Two sample test involving mean difference (matched or related samples)*

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

*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*

$$\chi^2 = \sum \left[ \frac{(f_o - f_e)^2}{f_e} \right]$$

*Chi-Square Test*

$$r = \frac{\sum (Z_X \cdot Z_Y)}{n - 1}$$

*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 Smart

## 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  $\pm .84$

**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.  $\sigma$  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  $\sigma$  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  $\sigma$  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

- Level of significance
- Critical region
- Fail to reject
- .05 and .01
- Region of rejection or critical region; null hypothesis

**Application Questions/Problems:**  
**Hypothesis Involving a Single Sample Mean With  $\sigma$  Known**

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

**Application Questions/Problems:**  
**Hypothesis Test Involving a Single Sample Mean With  $\sigma$  Unknown**

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

**CHAPTER 8**

**General Thought Questions**

- True
- False

- Mean differences
- The difference between means

**Application Questions/Problems:**  
**Matched/Related Samples Design**

- a.  $H_0: \mu_D = 0$ ; b.  $t = 2.81$ ; c. Critical value = 2.145; d. Reject the null at .05 level.
- a.  $H_0: \mu_D = 0$ ; b.  $t = 1.54$ ; c. Critical value = 2.064; d. Fail to reject the null at .05 level.
- a.  $H_0: \mu_D = 0$ ; b.  $t = 2.57$ ; c. Critical value = 2.045; d. Reject the null at .05 level.
- a.  $H_0: \mu_D = 0$ ; b.  $t = 2.28$ ; c. Critical value = 2.048; d. Reject the null at .05 level.
- a.  $H_0: \mu_D = 0$ ; b.  $t = 1.70$ ; c. Critical value = 2.045; d. Fail to reject the null at .05 level.

**Application Questions/Problems:**  
**Independent Samples Design**

- a.  $H_0: \mu_1 - \mu_2 = 0$ ; b.  $t = 2.97$ ;  
c. Critical value = 2.042; d. Reject the null at .05 level.
- 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.
- 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.
- 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.
- 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**

- Alternative or research
- Directional hypothesis
- Two-tailed
- One-tailed
- Rejecting; true
- 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  $\times$  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  $\pm 1$ ,  $\pm 2$ , and  $\pm 3$  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 ( $\mu$ ) and a standard deviation ( $\sigma$ ), the sampling distribution of sample means will have a mean equal to mu ( $\mu$ ) 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 ( $\mu$  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 ( $\mu$ )** 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, symmetrical 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

## Numbers

- 1-2-3 Rule, 66, 75
- regression analysis and, 301
- usefulness of, 76

## A

- Absolute values, 33
- Accuracy, vs. precision, 121
- Alpha errors (Type I errors), 165–167, 171, 213
- Alpha level, 159, 166. *See also* Level of significance
- Alternative hypothesis, 204–206
- Analysis of variance (ANOVA), 221–254
  - application, 233–242
  - benefits over *t* tests, 222
  - components, 242
  - estimates of variance and, 237, 242
  - F* statistic interpretation, 243–244
  - logic of, 223–230
  - means, 226–227
  - null hypothesis and, 230
  - one-way versus two-way, 250
  - post hoc testing, 244–248, 250
  - procedure for reviewed, 237, 248
  - questions raised by results, 244
  - uses, 222
- ANOVA. *See* Analysis of Variance (ANOVA)
- Area under the normal curve, 55, 65
  - table of, 79–85
  - t* distributions and, 128–131
- Arithmetic mean. *See* Mean

## Associations

- benefits over *t* tests, 222
- between variables, 257, 259–262, 270, 274–283, 286, 290–298, 302
- causation and, 261, 281–283
- curvilinear, 279
- direct (positive), 278, 281
- direction of, 277
- inverse (negative), 278, 281
- involving two interval/ratio level variables, 274–304
- linear, 277
- measures of, 271
- negative. *See* Associations, inverse
- non-existent, 279–280
- perfect, 279
- positive. *See* Associations, direct
- one-way vs. two-way, 249
- reciprocal, 282
- significant, 261
- strength of, 270, 277–281, 289
- types of, 277–280
- a* term (Y-intercept), 297, 299
- Average deviation, 33

## B

- Bar graphs, 54
- Baseline, 54, 88, 157
  - standardized normal curve, 76–82
  - Z* values, 79–80
- Beta errors (Type II errors), 167, 213–218
- Between-groups degrees of freedom, 239
- Between-groups estimate of variance, 238

Between-groups sum of squares ( $SS_B$ ),  
235–237

Between-groups variation, 229, 232

Between the means, 224

Bimodal distributions, 27, 60

$b$  term (slope -of the line), 297–299

## C

Calculated test statistic, 186, 196

$F$  ratio and, 242

standard deviation known, 160–164

standard deviation unknown, 170

Cases, 6–8

Categorical data, 255

Categories, 8

Causation, 261, 281–283

Center, of a distribution, 20

Central Limit Theorem, 100–104, 120

populations and, 155

relevance of, 117

Central tendency, 20–28

Chance

chi-square test of independence  
and, 261

statement of, null hypothesis and, 151

Chi-square test of independence,

256–269

application, 262–265

calculation of, 267

degrees of freedom, 269

departure from chance, 261–262

formula for, 265–268

interpretation, 268–269

limitations of, 270

logic of, 257–261

null hypothesis, 262

significant association, 261

variables and, 256–263

CI. *See* Confidence interval

Coefficient of determination ( $r^2$ ), 289

Collectively exhausted categories, 9

Comparisons

different scales of measurements,  
283–284

two distributions, 28, 35–36

Confidence intervals, 108–147

examples of, 111, 114–117, 132–134,  
137–139

formulas for, 112, 114, 118–120, 138

for the mean, 109–136

for proportions, 136–144

with standard deviation known, 109–123

with standard deviation unknown, 123–136

Confidence level. *See* Level of confidence

Constants, 297–298

Contingency tables, 257

pattern of chance and, 261

three-by-three, 262

Continuous distributions, 69

Correlation analysis, 280–293

application, 287–288

degrees of freedom, 291–292

interpretation, 292–293

formula for, 284–287

logic of, 283

null hypothesis, 291–292

used for varied scales of measurement,  
283–284

Correlation coefficient ( $r$ )

formula for, 285, 287

positive/negative, 280–281, 289

Critical region, 162

Critical value, 160–162, 269

Cross products of  $Z$  scores, 284–288

Curves. *See* Distributions; Normal curves

Curvilinear association, 279

## D

Data, 6–8. *See also* Variables

ascending/descending order and, 24

categorical, 255

types of, 5

Data distributions. *See* Distributions

Data points, 6

Data presentation,

graphs and curves, 53–68

scatter plots, 275–280

Data sets, 6

Degrees of freedom ( $df$ ), 128

analysis of variance and, 239, 242

between-groups estimate of variance,  
239–240

chi-square test of independence and, 269

correlation analysis, 291–292

single sample test with  $\sigma$  unknown, 170

$t$  distributions, 128–130

$t$  test for difference of means of  
independent variables, 186

$t$  test for the mean difference of related  
samples, 196

within-groups estimate of variance,  
239–240



Departure from chance, 261  
 Dependent variable, 282  
 Descriptive statistics, 13  
 Deviations from the mean, 29–32,  
     73–76, 79  
 Difference of means test, 179, 196  
 Direct (positive) association, 278, 281  
 Directional hypothesis, 205, 209–213  
 Direction of association, 277  
 Discrete distributions, 68  
 Dispersion, 28–47  
 Distributions, 6, 19–51, 60–67  
     bimodal, 27, 60  
     continuous, 69  
     discrete, 68  
     frequency, 54–56  
     normal, 90  
     point of inflection of, 62  
     shapes of, 52–70  
     skewed, 58  
     symmetrical, 58  
     tail of, 59  
     unimodal, 27, 60

## E

Effect, 216  
 Equal intervals, 9  
 Equality, null hypothesis and, 151  
 Errors. *See also entries at* Standard error  
     sampling error and, 94, 97–99  
     Type I/Type II, 165–167, 213–218  
 Expected frequencies, 264–268  
 Extreme values  
     effect on the mean, 23–24  
     graphing a distribution, 57–58  
     occurring with low probability, 161  
     in one-tailed test scenario, 211  
     standardized normal distribution, 85  
     statistician's interest in, 85, 89  
     in two-tailed test scenario, 210

## F

Fail to reject the null hypothesis, 160–161  
 Family of  $t$  distributions, 126–132  
     table for, 128–132  
     vs.  $Z$  distributions, 135  
 Formulas  
     chi-square test of independence,  
         265–268  
     confidence intervals, 112, 114,  
         117–120, 138

$F$  ratio, 241  
     the Mean, 21  
     the Median, 24  
     positional, 24  
      $r$  (correlation coefficient), 285, 287  
     regression analysis, 297–300  
     standard deviation, 39, 41, 132  
     standard deviation of the differences, 182  
     standard error of the difference  
         of means, 194  
     standard error of the estimate ( $s_e$ ), 300  
     standard error of the mean, 169  
     standard error of the mean difference, 185  
     standard error of the mean estimates, 124  
     standard error of the proportion  
         estimates, 138  
      $t$  ratio, 186  
     variance, 39, 41  
      $Z$  scores, 283

## F ratio

analysis of variance and, 225  
 formula for, 241  
 null hypothesis and, 231

## Frequency ( $f$ )

expected frequencies and, 264–268  
 frequency distributions and, 7, 54–56  
 observed frequencies and, 263–268

## G

Grand mean, 227, 242  
 Graphs, 53–68  
     scatter plots, 275–280  
 Group mean, 228, 242

## H

Harmonic mean, 246  
 Honestly Significant Difference (HSD),  
     245–248  
 Hypothesis, 150–152  
 Hypothesis testing  
     directional hypothesis testing and,  
         209–213  
     examples of, 187, 196  
     non-directional hypothesis testing and,  
         207–209  
     with one sample, 148–177  
     phrasing conclusions about, 168  
     procedures for summarized, 223  
     process summarized, 161t  
     with standard deviation known, 149,  
         152–168

Hypothesis testing (*continued*)  
 with standard deviation unknown, 149,  
 168–172  
 with two samples, 178–202

**I**

Independent samples, 188–197  
 Independent variable, 282  
 Inferential statistics, 14, 15, 93, 94, 109  
 degrees of freedom and, 128  
 four fundamental concepts of, 93–107  
 Inflection point, 64  
 Influence, vs. causation, 282  
 Information. *See* Data  
 Interval/ratio level of measurement, 9, 10  
 Intervals, 9, 110  
 Interval width, 120–123  
 Inverse (negative) association, 278, 281

**K**

## Key terms

analysis of variance, 250  
 chi-square test of independence, 271  
 confidence intervals, 144  
 correlation analysis, 304  
 data, 48  
 distributions, 48, 69  
 hypothesis testing, 173, 199  
 inferential statistics, 106  
 normal curves, 91  
 null hypothesis, 219  
 regression analysis, 304  
 statistics, 17

**L**

Least squares line, 296  
 Level of confidence, 110, 119, 121,  
 131–135, 140  
 Level of significance, 130  
 chi-square test of independence and, 269  
 correlation analysis and, 291  
 standard deviation known, 159–162  
 standard deviation unknown, 169–172  
 Type I errors and, 166  
 Levels of measurement, 8–10  
 Linear associations, 277  
 Line of best fit, 295–297

**M**

Marginal totals, 257, 261  
 Margin of error, 94, 139

Matched samples, 179–188  
 the Mean, 20–24

Mean difference, 183–187

## Means

deviations from, 29–34, 73–76, 79  
 formula for, 21  
 grand mean and, 227  
 group mean and, 228

Mean square between ( $MS_B$ ), 237

Mean square within ( $MS_W$ ), 237

Measurement. *See* Levels of measurement;

Scales of measurement

the Median, 24–26

the Mode, 26–28

Mu,  $\mu$  (population mean), 23, 42, 99

Mutually exclusive categories, 9

**N**

$N$  (symbol), 21

median of a population, 24

$n$  (symbol), 22

median of a sample, 24

Negative association. *See* Inverse (negative)  
 association

Negative (inverse) association, 278

Negative skew, 59

No difference, null hypothesis and, 151

Nominal level of measurement, 8

Non-directional hypothesis, 205, 207–209

No relationship, null hypothesis and, 151

Normal curves, 71–92

area under. *See* Area under the normal  
 curve

examples of, 73–76

mean of, 78

standard deviation of, 61–67

theoretical, 76–79

Normal distributions. *See* Normal curves

Null hypothesis, 150, 153, 203–220

vs. alternative hypothesis, 204–206

analysis of variance and, 230

chi-square test of independence and, 262

correlation analysis and, 291

independent samples and, 192

level of significance and, 159–162

phrasing conclusions about, 168

rejection of, 165, 222

related samples and, 184

symbol for, 184

$n$  versus  $n - 1$ , 44–47

**O**

Observations, 6  
 Observed frequencies, 263–268  
 One-tailed tests, 206–213  
 One-way ANOVA (analysis of variance), 221, 249  
 Opinion polls, 136–137, 140  
 Order, ascending or descending, 24, 25  
 Ordinal level of measurement, 8, 9

**P**

Parameters, 14  
 Pearson's  $r$ . See Correlation analysis  
 Percentages  
   confidence intervals for, 136–141  
   converting proportions to, 82  
   used for confidence intervals, 114  
 Perfect association, 279  
 Point of inflection, 62  
 Population parameters, 14  
 Populations, 10–13, 23  
   calculating standard variation of, 40  
   calculating variance of, 35  
   Central Limit Theorem and, 100–104, 155  
   confidence intervals and, 109–144  
   distributions and, 90  
   inferential statistics and, 109  
   sampling error and, 97–99  
 Positional formula, 24  
 Positive (direct) association, 278  
 Positive skew, 59  
 Post hoc testing, 244–248  
 Power, 213–217  
 Precision, vs. accuracy, 121  
 Predictions  
   line of best fit and, 295–297  
   regression analysis and, 299  
 Probabilistic distribution, 151  
 Probability, 151  
 Proportions  
   confidence intervals for, 136–144  
   converting to percentages, 81

**Q**

Q statistic, 245–247  
 Questions to answer  
   alternative hypothesis, 205  
   analysis of variance, 222, 225–228  
   associated variables, 262

association between variables, 277, 278, 279  
 basics of statistics, 6–15  
 between-groups sum of squares ( $SS_B$ ), 236  
 categorical data, 256  
 central limit theorem, 117  
 chi-square test of independence, 257, 263, 265  
 coefficient of determination, 290  
 confidence intervals, 110–114, 117, 122–124, 130, 132, 137–140  
 contingency tables, 259  
 correlation analysis, 281  
 critical region, 162  
 critical value, 162  
 cross products, 286  
 degrees of freedom, 240, 292  
 dependent/independent variables, 282  
 directional/non-directional hypothesis, 205  
 distributions, 56, 59, 62, 63, 66  
 $d$  values, 183  
 effect, 216  
 $F$  ratio, 242  
 hypothesis testing, 151, 157, 159, 169  
 independent samples, 189  
 line of best fit, 297  
 mean square between ( $MS_B$ ), 237, 239  
 mean square within ( $MS_W$ ), 237, 241  
 measures, 22–31, 34, 37–39, 44, 45  
 normal curves, 76, 77  
 null hypothesis, 291  
 one-tailed tests, 212  
 power, 216  
 probability, 152  
 reciprocal associations, 283  
 related samples, 180  
 scatter plots, 277, 280  
 significant difference, 153  
 standard deviation, 118  
 standard error, 118  
 standard error of the mean, 126  
 standard error of the mean difference, 185  
 $t$  ratio, 186  
 two-tailed tests, 208  
 Type I errors, 165, 166  
 Type II errors, 167  
 within-groups sum of squares ( $SS_W$ ), 233

**R**

- $r$  (correlation coefficient)
  - formula for, 285, 287
  - positive/negative, 280–281, 289
- $r^2$  (coefficient of determination), 289
- Random sampling, 94–97
- the Range, 28
- Ratio level of measurement, 8, 9
- Ratios
  - calculating, 43
  - $F$ , 225, 231, 232, 237, 241–244
  - $t$  test for difference of means of independent samples, 195
  - $t$  test for the mean difference of related samples, 185–186
- Raw scores, 86–88
- Reciprocal associations, 282
- Region of rejection, 162
- Regression analysis, 293–302
  - equation for, 297–300
  - prediction and, 293, 295–302
- Regression line, 295–297
- Rejection of null hypothesis, 158–159.
  - See also Null hypothesis
  - failing to reject null hypothesis, 160–161
  - at .01 level of significance, 161
  - at .05 level of significance, 160–161
  - in Type I errors, 165
- Related samples, 179–188
- Research hypothesis, 204–206
- Research questions. See Hypothesis testing
- Resources for further reading, 17, 143

**S**

- $s$  (standard deviation of a sample), 37–39
- Sample, 10–13, 178–202
  - analysis of variance and, 222
  - calculating standard variation of, 46
  - independent, 188–197
  - related, 179–188
- Sample means
  - analysis of variance and, 221, 227–231, 242
  - HSD procedure and, 245
  - symbol for, 23
- Sample size
  - analysis of variance and, 231
  - precision of estimates and, 122
  - selecting, 45
- Sample statistics, 14
- Sampling distribution, 100
- Sampling distribution of sample means, 99–100
- Sampling error, 94, 97–99
- Sampling frames, 96
- Scales of measurement, 10, 37, 68
- Scatter plots, 275–280
- Sigma, lowercase ( $\sigma$ ), 37–39
  - calculation, 39–40
  - Central Limit Theorem, 103
- Sigma, uppercase ( $\Sigma$ ), 21
- Significant association, 261
- Significant difference, 152–156, 164
- Skew, positive/negative, 59
- Skewed distributions, 58
- Slope of the regression line, 297
  - calculation, 298–299
- Standard deviation, 37–47
  - area under the normal curve and, 66
  - formulas for, 39, 41, 132
  - hypothesis testing and, 149, 152–172
  - of normal curve, 61–67
  - of sampling distribution of sample means, 102
- Standard deviation of the differences, 182–187
- Standard error, formula for, 117
- Standard error of the difference of means
  - estimate of, 192–195
  - formula for, 194
- Standard error of the estimate ( $s_e$ ), 300–302
- Standard error of the mean, 102
  - confidence intervals and, 111, 114–117, 123–126
  - estimating, 123–126
  - formula for, 169
  - hypothesis testing and, 149, 156–158, 169–172
- Standard error of the mean difference, 185, 192
- Standard error of the proportion,
  - formula for estimating, 138
- Standardized normal curves. See Normal curves
- Standard variation
  - for populations, 40
  - for samples, 46
- Statement of chance, null hypothesis and, 151

Statistical analysis, 13–16

Statistics, 4–18  
 definitions of, 14  
 inferential. *See* Inferential statistics  
 vs. parameters, 14  
 reasons for studying, 5

Strength of association, 278

Sum of squares, 232, 242

Symmetrical distributions, 58

Symmetry, 58

**T**

Table of Areas Under the Normal Curve, 79–85, 131

Table of random numbers, 96

Tail of the distribution, 59

*t* distributions, 126–132  
 table for, 128–132  
 vs. *Z* distributions, 135, 143

Test scores, 85–89

*t* ratio, formula for, 186

treatment groups, analysis of variance and, 222

Tukey's Honestly Significant Difference (HSD), 245–248

Two-tailed tests, 206, 212

Two-way ANOVA (analysis of variance), 249

Type I errors (alpha errors), 165–167, 171, 213

Type II errors (beta errors), 167, 213–218

**U**

Unimodal distributions, 27, 60

Universe. *See* Populations

**V**

Variability, 28–47

Variables, 6  
 associations between, 257, 259–262, 270, 274–283, 286, 290–298, 302  
 chi-square test of independence and, 256–263  
 dependent/independent, 282  
 X and Y, 281–283

Variance, 34–37  
 estimating, 237, 242  
 formula for, 39, 41  
 for populations, 35

Variation, types of, 228–230

**W**

Within-groups degrees of freedom ( $df_w$ ), 240

Within-groups estimate of variance, 238

Within-groups sum of squares ( $SS_w$ ), 232–234, 237

Within-groups variation, 229, 232

**X**

X-axis, 54, 275, 295

X variable, 281–283

**Y**

Y-axis, 55, 275, 297, 299

Y-intercept (a term), 297, 299

Y-prime, 297

Y variable, 281–283

**Z**

*Z* distributions, vs. *t* distributions, 135, 143

Zero point, of ratio level of measurement, 9

*Z* ratio, 87





