Appendix D: Solutions to Selected Exercises

D

Chapter 1

1.2 The change was made to protect the privacy of the adolescent in answering sensitive questions. The estimate of the proportion increased slightly immediately after the change, suggesting the earlier values were probably underestimated.

1.3 No, the difference in the infant mortality between Pennsylvania and Louisiana may be due to the difference in the racial/ethnic composition of the two states. The race-specific rates were indeed lower in Louisiana than in Pennsylvania. The proportion of blacks in Louisiana was sufficiently greater than that in Pennsylvania to make the overall rate higher than the overall rate in Pennsylvania.

Chapter 2

2.2 Not necessarily, as the choice of scale is dependent on the intended use of the variable. For example, we know that those completing high school have more economic opportunities than those that didn't and the same is true for those completing college. Hence, there is a greater difference between 11 and 12 years of education than between 10 and 11 years, and the same is true for the difference between 15 and 16 years compared to 13 and 14 or 14 and 15.

2.4 Counting the beats for 60 seconds may be considered too time-consuming. On the other hand, counting for 20 seconds or 15 seconds and multiplying by 3 or 4 may be unreliable. Counting for 30 seconds and multiplying by 2 may be a good compromise.

2.7 Age recorded in census is considered to be more accurate than that reported in death certificate which was reported by grieving relatives and other informants. In order to alleviate some of these disagreements, the age-specific death rates are usually calculated by five-year age groups.

Chapter 3

3.2 The actual expenditures increased, whereas the inflation-adjusted expenditures decreased. The trend in the inflated-adjusted expenditures would provide a more

realistic assessment of the food stamp program because it takes into account the decrease in the purchasing power of the dollar.

3.6 b.

A/C	1	0	Total
1	8	20	28
0	8	10	18
Total	16	30	46

c. A (28) B (28) C (16)

3.9 Since the total number of hospitals by type is not available, it is not possible to calculate the mean occupancy rate.

3.12 a. Mean = 747,482,813.3, CV = 344.1 percent

b. Median = 10^5, geometric mean = 541,170

c. The geometric mean, 5.4×10^5, seems to capture the sense of the data better than the mean or median.

3.14 b. The adjusted correlation between the new variables (protein/calories and total fat/calories) is 0.094; the adjusted correlation better characterizes the strength of the relationship; the unadjusted correlation of 0.648 is due to the fact that both protein and total fat are related to calories.

Chapter 4

4.2 a. 0.685

b. 0.524

c. $(0.426) * (0.524) = 0.223$

d. $(0.372 - 0.223)/(1 - 0.524) = 0.313$

4.5 $1 - \{1 - (1 - 0.99) * (0.2)\}^{120} = 0.214$

4.8 a. 82,607/97,196 = 0.850

b. (91,188 − 82,607)/98,672 = 0.087

Chapter 5

5.1 a. $(1 - 0.8593) = 0.1407$

b. At least 10 persons or less than 2 with $p = 0.0388$

c. Virtually zero

5.3 0.0146; 0.6057 (= 0.7073 − 0.1016)

5.6 Probability is 0.0116 (= 1 − 0.9884); would investigate further.

5.10 $z = -1.3441$; Pr $(x < 7) = 0.0895$; yes, we believe the data are normally distributed; can be verified by a normal probability plot.

Chapter 6

6.2 Read 25 four-digit random numbers and, if any random numbers are 2000 or greater, subtract a multiple of 2000 to obtain numbers less than 2000. Eliminate duplicates and draw additional random numbers to replace the number eliminated.

6.5 a. The population consists of all the pages in the book; the pages can be randomly sampled and number of words counted on the selected pages would constitute the data.

b. All moving passenger cars during the one-week period can be considered as the population. The population can be framed in two dimensions: time and

space. Passing cars can be observed at randomly selected locations at randomly selected times and the total number of cars and the number with only the driver can be observed.

c. The population consists of all the dogs in the county. Households in the county can be sampled in three stages: census tracts, blocks, and households. The number of the dogs found in the sample households and the number of dogs that have been vaccinated against rabies can then be recorded.

6.8 a. Some people have unlisted telephone numbers and others do not have telephones. People who have recently moved into the community are also not listed. Thus, these groups are unrepresented in the sample. The advantage is that the frame, although incomplete, is already compiled.

6.11 a. 30 classes can be randomly allocated to two curricula.

b. A simple random allocation of six teachers to two curricula may not be appropriate; instead, teachers can be matched based on teaching experience before randomly allocating one member of each pair to the new curriculum and the other member to the old curriculum.

6.14 a. Fewer subjects would be needed compared with the two-group comparison design.

b. The random assignment of subjects to the initial diet presumably balanced the sequencing effect but it might not be adequate because of the small sample size.

c. The carry-over effect is ineffectively controlled by not allowing a wash-out period and the granting of a leave to some subjects.

6.16 a. Randomized block design

b. The effect of organizational and leadership types is not controlled effectively, although the matching may have reduced the effect of this confounder.

Chapter 7

7.1 a. Sample mean = 11.94; sample standard error = 1.75; the 95 percent confidence interval = (8.42, 15.46).

c. Sample median = 8; the 95 percent confidence interval = {3 (19th observation), 12 (32nd observation)}.

d. The 95 percent tolerance interval to cover 90 percent of observation, based on normal distribution = $11.94 \pm 1.992(12.5) = (0, 36.84)$; based on distribution-free method, the interval (0, 45) gives 96.9 percent confidence level to cover 90 percent of observations; the latter method is more appropriate, since the data are not distributed normally (distribution is skewed to the right).

7.4 Would expect a negative correlation because those states that have the higher workplace safety score should have the lower fatality rates. $r = -0.435$. Since the data are based on population values, there is no need to calculate a confidence interval. However, if we viewed these data as a sample in time, then the formation of a confidence interval is appropriate. The 95 percent confidence interval = (−0.636, −0.178); a significant negative correlation exists, since the confidence interval does not include zero.

7.7 Correlation = 0.145; these data may be viewed as a sample in time; the 95 percent confidence interval = (−0.136, 0.404); no significant linear relation exists, since the confidence interval includes zero; region of the country, perhaps reflecting the unemployment levels, may play a role.

7.10 a. (0.052, 0.206)
b. Difference (1990 − 1983) = −0.06; confidence interval = (−0.202, 0.082); no difference, since the confidence interval includes zero.

7.13 Difference = −0.261; 99 percent confidence interval = (−0.532, 0.010); no difference, since the confidence interval includes zero, although a 95 percent confidence interval would not include zero.

Chapter 8

8.2 The decision rule is to reject the null hypothesis when the number of pairs favoring diet 1 is 14 to 20 with $\sigma = 0.0577$ and $\beta = 0.0867$.

8.6 The percent predicted FVC is used, since it is adjusted for age, height, sex, and race. H_0: $\mu_1 = \mu_2$; H_a: $\mu_1 > \mu_2$. One-sided test is used to reflect the expected effect of asbestos on pulmonary function; assuming unknown and unequal population variances, $t' = 30.27$ with df = 126.5; p-value is virtually zero; reject the null hypothesis, suggesting that those with less than 20 year exposure have significantly larger forced vital capacity than those with 20 or more years of exposure.

8.9 H_0: $\mu_d = 0$, H_a: $\mu_d < 0$; $t_d = -10.03$, which is smaller than $t_{23,0.01} = -2.50$; reject the null hypothesis, suggesting that the weight reduction program worked.

8.12 H_0: $\pi = 0.06$; H_a: $\pi < 0.06$ (one-sided test); $z = 1.745$, which is larger than $z_{0.95} = 1.645$; reject the null hypothesis, suggesting that there is evidence for the community's attainment of the goal.

8.14 $r = -0.243$; H_0: $\rho = 0$; H_a: $\rho \neq 0$; $\lambda = -0.8224$, which is not smaller than $z_{0.05} = -1.645$; fail to reject the null hypothesis, no evidence for nonzero correlation; $p = 0.21$.

8.16 H_0: $\mu = 190$; H_a: $\mu \neq 190$; $t = 3.039$, which is larger than $t_{14,0.01} = 2.6245$; reject H_0.

Chapter 9

9.1 Medians are 12.25 for group 1, 7.75 for group 2 and 5.80 for group 3; average ranks are 36.5 for group 1, 23.3 for group 2 and 16.9 for group 3; the Kruskal-Wallis test is appropriate to use; the test statistics H = 15.3, df = 2 and $p = 0.001$, indicating the medians are significantly different.

9.4 Divide into three groups based on the toilet rate (1 to 61), (133 to 276) and (385 to 749), with 9 observations in group 1, 6 in group 2, and 6 in group 3; since H = 6.67 with $p = 0.036$, we reject H_0.

9.7 The results by the Wilcoxon signed rank test are consistent with that obtained by the sign test in Exercise 9.6, although the p-value is slightly smaller with the Wilcoxon signed rank test than with the sign test.

Chapter 10

10.3 Note that there are 2 out of 10 cells with expected counts less than 5, but the smallest expected count (3.18) is greater than 1 [= 5 ∗ (2/10)] and the chi-square test is valid; $X^2 = 6.66$, df = 4, $p = 0.1423$, we fail to reject the null hypothesis at the 0.05 significance level; This is a test of independence because it appears

that the subjects were selected at random, not by degree of infiltration. By assigning scores of –1, 0, 1, 2, and 3, we calculate $X^2 = 6.67$, df = 1, $p = 0.0098$; we reject the null hypothesis of no trend; by assigning scores of –1, 0, 0.5, 1, and 1.5, we calculate $X^2 = 6.36$, df = 1, $p = 0.0117$; we again reject the null hypothesis of no trend.

10.5 $X^2 = 20.41$, df = 1, $p < 0.0001$, we reject the null hypothesis; the proportion of violation is nearly three times higher for the nonattendees (73.5 percent) than the attendees (24.3 percent). Without more information, we cannot draw any conclusion about the effect of attending the course. Our interpretation depends on whether the course was attended before or after the violation was found.

10.8 $X^2 = 103.3$, df = 1, $p < 0.0001$, ignoring the radio variable, significant; $X^2 = 24.65$, df = 1, $p < 0.0001$, ignoring the newspaper variable, significant. The newspaper variable seems to have the stronger association. However, it is difficult to recommend one media over the other, since these two media variables, in combination, appear to be related with the knowledge of cancer. Additionally, since people were not randomly assigned to the four levels of media, to use these results about knowledge of cancer, we must assume that the people in each of the four levels of the media initially had the same knowledge of cancer. Without such an assumption, it is difficult to attribute the status of cancer knowledge to the different media.

Chapter 11

11.2 a. For the group with serum creatinine concentration 2.00–2.49 mg/dL, the five-year survival probability is 0.731 with standard error of 0.050; for the group with serum creatinine concentration 2.5 mg/dL or more, the five-year survival probability is 0.583 with a standard error of 0.058.

b. Despite the considerable difference in the five-year survival probabilities, the two survival distributions are not significantly different at the 0.01 level, with $X^2_{CMH} = 3.73$ and $p = 0.0535$, reflecting the small sample size.

11.6 Median for the fee-for-service group = 28.8 month; median for HMO = 29.5; the two survival distributions are not significantly different.

Chapter 12

12.2 $F = 0.51$, $p = 0.479$, no significant difference; the test results are the same as those obtained using the *t*-test, with the same *p*-value; $F_{1,\, n-1,1-\alpha} = t^2_{n-1,1-\alpha/2}$.

12.5 Degrees of freedoms are 2, 2, 4 and 18 for smoking status, lighting conditions, interaction, and error, respectively; $F = 0.213$ for interaction, which is not significant; $F = 12.896$ for smoking status, significant; $F = 45.276$ for lighting conditions, significant.

Chapter 13

13.1 The zero value for the degree of stenosis in the 10th observation is suspicious, which appears to be a missing value rather than 0 percent of stenosis; the zero value for the number of reactive nuclei at initial survey in the 12th observation is also suspicious, but it may well be a reasonable value, because there are other

smaller numbers such as 1 and 2; the scatter plot seems to suggest that there is a very weak linear relationship; a regression analysis yields $\hat{\beta}_0 = 22.2$, $\hat{\beta}_1 = 2.90$, $F = 10.04$, $p = 0.007$; the 10th observation had the largest standardized residual and the 6th observation had the greatest leverage almost three times greater than the average leverage; eliminating the 6th observation, $\hat{\beta}_0 = 21.2$ and $\hat{\beta}_1 = 3.05$.

13.4 Eliminating the two largest blood pressure values (14th and 50th observations) and the two smallest values (22nd and 27th observations), $\hat{\beta}_0 = 63.1$, $\hat{\beta}_1 = 0.726$ and $R^2 = 23.4$ percent.

13.5 r = −0.138; $\hat{\beta}_0 = 8.53$, $\hat{\beta}_1 = -0.063$, $R^2 = 1.9$ percent, $F = 0.19$ and $p = 0.669$; by adding the new variable, $\hat{\beta}_0 = 8.13$, $\hat{\beta}_1 = -0.067$, $\hat{\beta}_2 = 0.110$ (new variable), $R^2 = 37.8$ percent, $F = 2.73$ and $p = 0.118$; the new variable captured the nonlinear effect of BMI on serum cholesterol.

Chapter 14

14.1 The exponential of 0.392 is 1.48 (odds ratio), suggesting that those with the extensive operation have the greater proportion of surviving less than 10 years; this result is consistent with the data in the table, which shows that 51.4 percent [= 129/(129 + 122)] of patients with the extensive operation survived less than 10 years, whereas 41.7 percent [= 20/(20 + 28)] of patients with the not extensive operation survived less than 10 years.

14.3 Coding female = 1 and male = 0 for the sex variable, and died = 0 and survived = 1 for the survival status, the fitted logistic regression model yields: logit $(\hat{\pi})$ = 1.57 − 0.07 (age) + 1.33 (sex). Exp (1.33) = 3.78 indicates a woman's odds of survival is nearly 4 times the odds of survival for a man holding age constant. Exp [−0.07 * (40 − 15)] = 0.17 suggests that a 45-year-old person's odds of survival is about 1/5 of the odds of 15-year-old person of the same sex.

Chapter 15

15.1 The weight is the number of adults in the households with one phone and for the households with two phones is ½ of the number of adults; the weighted percent (35.8) means that 35.8 percent of adults in the community are smokers; the unweighted percent (30.0) means that 30 percent of telephone locations in the community have at least one smoker.

15.2 The standard errors for the prevalence rate and the odds ratio are 0.59 and 0.090, respectively; the standard errors based on the range are 0.61 and 0.096, respectively.