LionTutors

www.LionTutors.com

STAT 200 – Final Exam – Practice Exam Solutions

1. A – The college students; Cases are the individual units on which a measurement is taken.

2. D – The amount the student slept the previous night is the explanatory variable because it is explaining the student's performance on the quiz.

3. E – The number of questions answered correctly is the response variable because it is the variable being explained by the amount a student slept the previous night.

4. C – Hours of sleep is quantitative because it is recorded with numerical values that measure the amount of time each student slept the previous night.

5. D – Quiz score is quantitative because it is recorded with numerical value that count the number of correct answers for each student.

6. A – Observational study; There is no treatment applied. The participants are simply observed.

7. A – Sample; The data is being collected from a sample of 200 students and used to represent the population of all Penn State students.

8. B – Population; The data is being collected from <u>all</u> students, or the entire population of interest.

9. C – The population is all students who live in dorm buildings at this university because that is the population from which the sample is randomly selected.

10. C – The population is all people who registered online for the 5K race because that is the population from which the sample is randomly selected.

11. A – For a sample to be considered a simple random sample, each unit in the population must have an equal chance of being selected. There cannot be any specific criteria used for ordering the students (i.e. alphabetical order, class standing, or first returned).

12. B – The results are likely to have response bias because the students are unlikely to tell the professor that they cheated on homework.

13. C – This is a randomized experiment because the students are assigned to use a smart phone or a computer, and with a randomized experiment we are able to claim causation based on the results.
14. B – This is an observational study because the researchers are not influencing the results in any way. They are just observing the people who have already registered for the program. We cannot claim causation based on the results of an observational study.

15. B – How many times have you skipped class this semester?
    B – How long does it take you to walk to your first class?
    A – What is your major?
    A – What is your class year?

16. A – Height is explaining the student's seating preference, and height is a quantitative variable because it is measured with a value.

17. B, C, and D – This is a matched pairs experiment because each student is measured with and without sugar. The identical beverage without sugar is a placebo. Because neither the researcher nor the participants who is assigned to each beverage, this experiment is double-blind.

18. D – Odds = (Number in category) *to* (Number <u>not</u> in category)
    Odds = (Number of Penn State students in "First-Choice" category) *to* (Number of Penn State students in "Second-Choice" and "Neither" categories)
    Odds = 62 *to* (38 + 13)

19. A – Proportion = (Number in category) / (Total)
    Proportion = 51 / 107

20. See numbers below:

    a. 16
    b. 26
    c. 42
    d. 40
    e. 8
    f. 48
    g. 56
    h. 34

21. D – This histogram is right (positively) skewed. In a positively skewed dataset, the mean is greater than the median because the mean is a sensitive statistic that is pulled toward the higher outliers.

22. B – -1.25

    z-score = (Observed Value – Mean) / Standard Deviation
    z-score = (52 – 62) / 8 = -1.25

23. D – The 80th percentile is the value with 80% of the dataset below it.

24. B – The value of 83 MPH would cause the mean to increase and would not affect the median. This would cause the mean and median to become less similar in value.

25. A – 83 MPH is higher than the current mean, so it would increase the value of the mean.

26. A – Since the outlier of 83 MPH is further from the mean than the other values, it would increase the average distance of the values from the mean.

27. B – The IQR is resistant so it would be unaffected by the outlier.

28. A – The minimum is 1 hour.

29. B – 75% of observations fall above the first quartile, 2 hours.

30. B – 25% of observations fall below the first quartile, 2 hours.

31. D – 25% of data falls between each of the numbers in a five-number summary (minimum, Q1, median, Q3, maximum). Therefore, 50% of the data falls between the median and the maximum.

32. D – IQR = Q1 – Q3 = 5 – 2 = 3

33. B – The 95% rule says that 95% of values fall within 2 standard deviations of the mean in either direction.

    95% of resident ages = Mean ± 2 (Standard Deviation)
    95% of resident ages = 46 ± 2 (11.2), or 24.1 to 68.9

34. D – 3 is the first quartile, so 25% of the data would fall below 3. It is false that any value below 3 would be an outlier.

35. B – It is likely that a person's weight will decrease as the number of hours spent exercising increases, which would make the variable negatively associated.

36. B – If the correlation is 0, there is no association between the two variables – negative or positive. That means that the best line through the data would be horizontal.

37. D – Correlation is affected by sample size.

38. Top left – A
Top right – D
Bottom left – B
Bottom right – C

39. B – The earlier years tend to have higher numbers of injuries.

40. You know than the correlation for #3 will be closer to 0 than the correlation for #1 because they have the same pattern; however, #3 has more observations. When the pattern stays constant and the number of observations increases, correlation will move closer to 0.

#1 = -.75
#2 = -.02
#3 = -.73
#4 = -.99

41. B – The correlation is very close to 0, so we suspect there is no linear relationship between the two variables. Correlation only describes linear relationships, so it is incorrect to say that we suspect there is no relationship in general between the variables. Since the data appears to be curvilinear, it cannot be well described using correlation.

42. C < D < B < A

43. B – This is an observational study because the company is just recording the salary for males and females. They are not applying a treatment or manipulating the results in any way. With an observational study, we cannot claim causation.

44. D – Side-by-side boxplots are the best visual display when we are comparing a quantitative variable (salary) for multiple groups (males and females).

45. C – We would compare the average salary for males and females using the difference in means.

46. D – The sample proportion, p-hat = (number in category) / (sample size) = 400 / 500 = 0.80

47. D – p
B – $\hat{p}$
A – μ
F - $\bar{x}$
C - $\mu_1 - \mu_2$
E - $\hat{p}_1 - \hat{p}_2$

48. B – We are <u>95% confident</u> that the true <u>population parameter</u> is within a 95% confidence interval. The other three choices are common misinterpretations of confidence intervals.

49. B, C – Any value within a confidence interval is a plausible value for the population parameter.

50. B – We expect 95% of all 95% confidence intervals to capture the true population parameter. That means that we expect the other 5% to <u>fail to capture</u> the population parameter.

    .05 (500) = 25

51. A – The population proportion p should be the center of the sampling distribution, 0.33.
52. B – 0.60 is near the edge of the sampling distribution, so it is unlikely to occur. However, it is still in the range of possible sample proportions so it may occur occasionally.

53. C – Increasing the sample size will not affect the center of the distribution, but it will decrease the spread of the distribution.

54. C – The widest confidence interval goes with the highest confidence level.

55. A – Confidence interval = Sample estimate ± Margin of error

    Note that we are given the margin of error, 6%, so all we have to do is add and subtract it from the sample statistic, 18%.

56. C – 87.8
    A – 81.5
    B – 75.9
    B – 85.0

57. B – A larger sample size will result in a narrower confidence interval, all else held constant.

58. B – The range of the confidence interval will be twice the margin of error because of the calculation below:

    Confidence interval = Sample estimate ± Margin of error

    (130 – 90) / 2 = 20

59. D – A bootstrap distribution is found by taking repeated samples of the original sample size (n=16) from the original sample, with replacement. To find 500 bootstrap statistics, we would do this 500 times.

60. B – 1,000
    C – 90%
    A – 9.24
    E – 3.18

61. B - 95% Confidence Interval = Sample estimate ± 2 x SE

    We use the original sample statistic as the sample estimate and the standard error of the bootstrap distribution to calculate the confidence interval.

62. B, C, and D – Any value within the range of the bootstrap distribution is a plausible value for the population mean.

63. C – Each dot in the bootstrap distribution represents the bootstrap statistic for one bootstrap sample. Bootstrap samples are found by sampling with replacement from the original sample.
64. D – We cannot determine an exact value for the population mean, we can only use the bootstrap distribution to estimate a range for the population parameter.

65. A – The sample statistic is the correlation, r, from the original sample found in the right corner of the output.

66. B – Since there are 100 bootstrap statistics, a 92% interval would include the middle 92 dots in the distribution. That means that there are 8 dots outside of the interval, or 4 on each end. There are 4 dots below 0.191 and 4 dots above 0.678, so those are the boundaries of our 92% confidence interval.

67. A – Since the entire confidence interval for the population correlation is above 0, we can conclude that there is a positive correlation between these two variables.

68. C – The smaller the p-value, the stronger the evidence against the null hypothesis.

69. C – Sampling distribution (centered at the true population parameter)
A – Sample of n=30 (the original sample will have the biggest spread because it represents individual values rather than possible sample means)
B – Bootstrap distribution (centered near the original sample statistic)

70. B – The sampling distribution would be centered at the true population parameter rather than the sample statistic, but the standard error (spread) would be approximately the same.

71. C – Ho: $p = 0.55$, Ha: $p > 0.55$

Since we are dealing with one categorical variable, our parameter is p. We use a greater-than alternative becase we want to test whether the new method is <u>more</u> effective, or if the success rate is higher.

72. B – Ho: $\rho = 0$, Ha: $\rho < 0$

Since we are dealing with two quantitative variables, our parameter is the population correlation ρ (rho). We use a less-than alternative because we want to test whether there is a negative association between the two variables.

73. H_o: $p = 0.5$
H_a: $p < 0.5$

74. We <u>fail to reject</u> the null hypothesis (because the p-value is greater than or equal to 0.05). We <u>do not have</u> evidence that less than half of students feel sufficiently prepared when entering exams at the testing center.

75. B – Researcher 1 gives stronger evidence against the null hypothesis because he/she found a lower p-value. Researcher 1 should reject the null because the p-value is less than 0.05, but researcher 2 should fail to reject the null because the p-value is greater than or equal to 0.05.
76. B – Researcher 2 will have the smallest p-value because the statistic gives the most support for the alternative hypothesis, Ha: mu1 – mu 2 ≠ 0.

77. B – The p-value is the probability of getting a sample statistic as extreme or more extreme than our sample statistic in the direction of the alternative hypothesis.

78. A – The distribution should be centered at the null value, 0. We should shade the area more extreme than our test statistic, 0.12, in the direction of the alternative hypothesis. Since we have a two-sided alternative, we shade the area above 0.12 and below -0.12.

79. A result is considered to be statistically significant if the observed relationship in the sample is large enough that it is unlikely to have occurred by random chance if the null hypothesis is true, meaning that the explanatory variable is important in the population.

80. C – 1/100 bootstrap statistics is above the sample statistic of 0.70, so the p-value would be 1/100 or 0.01.

81. B - $\hat{p} = 0.22$
    A - $\hat{p} = 0.24$
    A - $\hat{p} = 0.26$
    C - $\hat{p} = 0.32$

82. B – False; The randomization distribution is created under the assumption that the null hypothesis is true, or that there is no difference between the two methods.

83. A – True; The randomization distribution is centered at the null value.

84. B – False; If the difference in sample means is 0.99, the manager will not find statistical significance because the p-value would be much greater than 0.05.

85. The probability of seeing a sample proportion of 0.546 or any value larger is 0.024, assuming the population proportion is equal to 0.5.

86. D – A researcher should attempt to replicate results before reporting them because there is a chance that an error was made in the first hypothesis test.

87. C – The significance level is the probability of committing a type 1 error.

    If 40 hypothesis tests are performed using a 0.05 level of significance, 5% of them would be expected to commit a type 1 error.

    5% * 40 = 2 type 1 errors

88. C – Researcher 3 uses the largest sample size. With all else held constant, a larger sample size will provide more evidence against the null and in favor of the alternative.

89. A – Any value outside of the 95% confidence interval can be rejected at the 0.05 level of significance. Since 0.50 is not within the confidence interval, Ho: p = 0.50 can be rejected. In choices B and C, the null value is within the confidence interval and therefore cannot be rejected as possible population proportions.

    Statistical significance has been found when the null is rejected in favor of the alternative.

90. E – This situation is dealing with the difference between means for two independent groups.

91. B – We use the t-distribution when dealing with means. The degrees of freedom (df) is the smaller of $n_1 - 1$ and $n_2 - 1$.

    df = 10 – 1 = 9

92. C – The z* for a 90% confidence interval is the value of z* where 90% of the standard normal distribution is between –z* and +z*

93. B – These are paired samples because we are looking at the difference in weight loss for each set of identical twins.

94. C – One population mean because we are looking at one quantitative variable (average rent); Hypothesis test because we want to know **if** the average rent is greater than $3,000.

95. A – Difference in two population proportions because we are comparing a categorical variable (opinion) for two groups (Republicans and Democrats); Confidence interval because we want to determine **how different** the proportions are (we want to estimate the value of the difference)

96. C – Difference in two population means because we are comparing a quantitative variable (LDL cholesterol level) for two groups (people who eat potato chips and people who don't); Hypothesis test because we want to determine **if** the level is higher for those who eat potato chips

97. B – A null hypothesis says that the explanatory variable will have no effect on the response variable.

98. A – A chi-square test is used with two categorical variables.

99. C – There is no relationship between illegal drug consumption and gender.

    It is incorrect to use = or ≠ symbols because illegal drug consumption and gender are not values.

100. D – Expected count = (Row total x Column total) / Overall total for the table

Expected count = (295)(209)/629

101. A – Degrees of freedom (df) takes the number of row categories minus one times the number of column categories minus one. You do not include totals or headings, just the number of categories.

df = (r – 1)(c – 1)
df = (2 – 1)(2 – 1)
df = 1

102. A – Reject the null because the p-value is less than 0.05. Conclude the alternative is true, or that there is a relationship between gender and illegal drug consumption in the population.

103. A – Since the pattern of OS used is the same for students who live on campus and students who live off campus, there is no support of a relationship between the two variables.

104. B – For each one-unit increase in the x-variable, y-hat changes by the slope of 2.

105. B – If the x-variable increases by 11, y-hat increases by 11(2), or 22.

You can also plug in numbers for the x-variable that are 11 inches apart:

Weight = 40 + 2(Height)

Weight for someone who is 60 inches tall:
Weight = 40 + 2(60)
Weight = 160

Weight for someone who is 71 inches tall:
Weight = 40 + 2(71)
Weight = 182

Difference:
182 – 160 = 22

106. A – Scatterplots <u>cannot</u> prove causation.

107. B – To find correlation (r), we can take the square-root of squared-correlation (R^2). We <u>do not</u> use R^2 adjusted.

r = sqrt(0.439)

The square root could be negative or positive; however, we know that the slope of the regression line is positive (b_1 = 0.4692), so we know that the correlation must be positive.

108. A – Exam 2 = 45.1 + 0.469(Exam 1)

y-hat = b_0 + b_1x
b_0 – The y-intercept or "constant coefficient"
b_1 – The slope or "Exam 1 coefficient"

109. C – R^2 is the proportion of variation in y explained by x.

110. B - Ho: $B_1 = 0$, Ha: $B_1 \neq 0$; We use B_1, the population slope, because the hypotheses are always about population parameters. The null has to have an equal sign, and the alternative is that the slope is not 0, or that there is a linear relationship in the population.

111. E – We use the T and P values from the "Exam 1" row because we are testing Exam 1 (the explanatory variable) as a linear predictor of Exam 2 (the response variable).

112. C – The multiple linear regression equation uses the constant plus the coefficient times each explanatory variable.

113. A – The predicted final score will increase by the slope for each 1 point increase in the midterm score, assuming all other variables are held constant.

114. B – A dummy variable will either be equal to 1 or 0. Since the slope is -1.158, the predicted final score will decrease by 1.158 when the variable is present.

115. B – Quiz average is a significant linear predictor of final exam score because it has a p-value less than 0.05. This only holds true when all explanatory variables are present in the model.

116. D – DummyGender has a p-value greater than 0.05, so it is not a significant linear predictor when combined with the other explanatory variables in this model.

117. C – In the ANOVA table, we see that the p-value is less than 0.05. That tells us that at least one of the slopes in the model differs from 0, or is a significant linear predictor of final exam score. It **does not** tell us that all of the slopes differ from 0.

118. C – R-squared = 38.5%