CSC 423 – Final Exam

August 11, 2014

Part A. Multiple Choice Questions. 5 pts. each. Answer only 23 out of 25 questions. For each question you may give an optional reason or show your work for possible partial credit. Circle the letter of the best (most correct) answer. Use additional sheets of paper if you need more room to write.
Note: The total number of points for this exam is 167. This will be scaled to be out of 100.

1. The sample median should be used to estimate the center of the population density when that density is
   a. is normal with nonzero μ. b. is skewed. c. is standard normal. d. has thin tails.

2. In the graph of the normal density, the distance between an inflection point and the center of the density is equal to the
   a. center of gravity. b. population standard deviation.
   c. population median. d. critical point of the density.

3. Which of the following correctly states the central limit theorem? If n is the sample size, the larger n is, the closer the distribution of the _____ gets to a normal distribution.
   a. population mean b. sample c. sample mean d. sample variance

4. Which of the following would be a legitimate reason for removing an outlier from a dataset?
   a. The outlier clearly belongs to a different population.
   b. The outlier is the result of natural variability in the measurement of interest.
   c. The outlier is more than two standard deviations from the mean.
   d. The outlier is the only negative number in the dataset.

5. According to Tukey's boxplot method, which points of the following dataset are outliers (mild or extreme)?
   11 38 42 53 57 63 151
   Use the Tukey's Hinges method to compute Q1 and Q3.
   a. 11, 38, 63, and 151 b. 11 and 151 c. 151 only d. There are no outliers.

6. For a dataset with n = 9, what is the third normal score? Use Van der Waerden's method to compute the normal scores? (Using Van der Waerden's method, the normal scores divide the standard normal curve into n + 1 equal areas.)
   a. -0.524 b. -0.431 c. 0.000 d. 0.500

7. Which of these densities generates observations with thick tails?
   a. The normal density N(5, 0.001) b. The normal density N(3, 7000)
   c. The t-distribution with 5 degrees of freedom d. The uniform distribution on the interval [2, 9].

8. Which of the following statistics is unbiased for the population mean?
   a. interquartile range b. sample mean c. sample standard deviation d. sample variance

9. For a sample of 400 blood pressure values, $\bar{x}$ = 120 and s_x = 10. What is the z-score of the blood pressure reading x = 135?
   a. 0.5 b. 0.667 c. 1.0 d. 1.5

10. For the blood pressure sample in Problem 9, if the blood pressure readings are assumed to be normally distributed, which interval is likely to contain almost all (over 99%) of the blood pressure values in the sample?
    a. 119 to 121 b. 110 to 130 c. 100 to 140 d. 90 to 150

11. If n = 9, $\bar{x}$= 7.3, and s_x = 1.8 for a normally distributed dataset, find a 95% confidence interval for μ, using the t-table.
    a. (3.15, 11.5) b. (3.77, 10.83) c. (5.92, 8.68) d. (6.12, 8.48)

12. If for a bivariate dataset, the sample standard deviations are 7 and 12, and the sample correlation is 0.5, what is the sample covariance?
    a. 42 b. 84 c. 162 d. More information is needed to calculate the sample covariance.

13. A one-sample t-test has H_0: μ = 35, n = 7, $\bar{x}$ = 39 and s_x = 4. Is this test significant at the 0.05 level? at the 0.01 level?
    a. no; no b. yes; no c. no; yes d. yes; yes

14. A paired-sample t-test produces the test statistic t = -2.24. Each treatment group has 8 subjects. Is this test significant at the 0.05 level? at the 0.01 level?
    a. no; no b. yes; no c. no; yes d. yes; yes

15. An independent two-sample t-test produces the test statistic t = 2.83. Group 1 has 6 subjects. Group 2 has 9 subjects. Is this test significant at the 0.05 level? at the 0.01 level?
    a. no; no b. yes; no c. no; yes d. yes; yes

16. A regression model that contains as few independent variables as possible is called
    a. consistent b. homoscedastic c. minimally unbiased d. parsimonious

17. What is the definition of a p-value for a statistical test?
    a. The amount of evidence for accepting the null hypothesis.
    b. The probability of obtaining a test statistic value that is as extreme or more extreme than the one actually obtained, given that the null hypothesis is true.
    c. The probability that the null hypothesis is true.
    d. The type 1 error.

18. Which two-sample t-test is used when the variances of the treatment groups is not assumed to be equal?
    a. Cauchy-Schwartz b. Kolmogorov-Smirnov c. Schroeder-Bernstein d. Welch-Satterthwaite

For Problems 18 and 19, $\bar{x} = 43$, $s_x = 3.4$, $\bar{y} = 197$, $s_y = 23.0$ $r_{xy} = 0.8$, for the simple regression equation $y = a + bx$,

19. What is the estimated slope?
    a. 0.12 b. 5.41 c. 6.76 d. 12.58

20. What is the estimated intercept?
    a. -35.7 b. -93.9 c. -154.0 d. 429.7

21. For a regression model with 3 regression parameters, n = 8, SSM = 21.4, and SSE = 5.78. Use the overall F-test to check if any of the independent variables are significant. Are any of them significant at the 0.05 level? At the 0.01 level?
    a. no; no b. yes; no c. no; yes d. yes; yes

22. For a regression model with 8 independent variables and 100 observations, if the R-squared value is 0.857, what is the adjusted R-squared value?
    a. 0.16 b. 0.38 c. 0.48 d. 0.84

23. Where is a dummy variable used? As a(n) _____ in a regression model
    a. dependent variable b. estimated parameter c. estimated residual d. independent variable

For problems 24 and 25, a regression model has 4 independent variables and if MSE = 0.057,

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 16 & 8 & 8 & 8 & 8 \\ 8 & 8 & 4 & 4 & 4 \\ 8 & 4 & 8 & 4 & 4 \\ 8 & 4 & 4 & 8 & 4 \\ 8 & 4 & 4 & 4 & 8 \end{pmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} 279.4 \\ 159.5 \\ 151.0 \\ 155.7 \\ 144.2 \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.3125 & -0.125 & -0.125 & -0.125 & -0.125 \\ -0.125 & 0.250 & 0 & 0 & 0 \\ -0.125 & 0 & 0.250 & 0 & 0 \\ -0.125 & 0 & 0 & 0.250 & 0 \\ -0.125 & 0 & 0 & 0 & 0.250 \end{pmatrix}$$

Note: the first column of X consists of a column of ones for the intercept. For problems 24 and 25, don't perform the entire matrix multiplication, just perform enough of the multiplication to obtain the information about the estimated intercept parameter.

24. Compute estimated the intercept parameter.
    a. 4.1 b. 7.9 c. 11.0 d. 16.0

25. Compute standard error of intercept parameter.
    a. 0.13 b. 0.37 c. 1.14 d. 3.66

Part B. Regression example. A company wants to compare the overall HR rating of supervisors with four employee questionaire questions. Each question is answered with a rating from 0 to 100.

Question 1: "Handles employee complaints." Question 2: "Does not allow special privileges."

Question 3: "Opportunity to learn new things." Question 4: "Rate of advancing to better jobs."

The independent variables are the ratings for each question (0 to 100): x1, x2, x3, and x4.

The dependent variable y is the HR rating of the supervisor (also 0 to 100).

Answer these two questions after consulting the SAS or R output.

1. (30 points) Choose the best regression model out of MODEL1, MODEL2, MODEL3, or MODEL4 for predicting y from the independent variables. Justify your choice. For full credit, include the following concepts in your discussion: R-squared value, adjusted R-squared value, parsimonious, overall F-test, t-values and p-values for the coefficients corresponding to individual independent variables, residual plot, normal plot, influence points, multicollinarity.

2. (7 points) Using your best regression model, what is the predicted value of the overall supervisor rating for a questionnaire that has the following employee ratings:
   x1 = 68 x2 = 55 x3 = 60 x4 = 40

Part C: Short Essay Questions. 15 pts. each. For full credit write in compete sentences and paragraphs. Answer only 2 out of 5 questions. Approximately one third to one half page per question. Write your answers on separate sheets of there is not enough room on the exam sheets.

1. Choose one of the statistical packages SAS or R. Your company is thinking of using this package. Write a memo to the purchasing department describing the good and bad points of the statistical package. (Don't compare SAS and R, just discuss one of the packages.)

2. Explain what a random variable is. Why random are variables are important for statistics in general, and for regression analysis in particular?

3. Explain what influence points are. Give at least two popular measures of influence.

4. Explain what multicollinarity is and how to remedy it.

5. Why is the normal distribution so important in statistics? How do you check if a sample comes from a normal distribution?