# CSC 423 – Final Exam

## August 10, 2016

**Part A. Multiple Choice Questions.** 5 points each. 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.

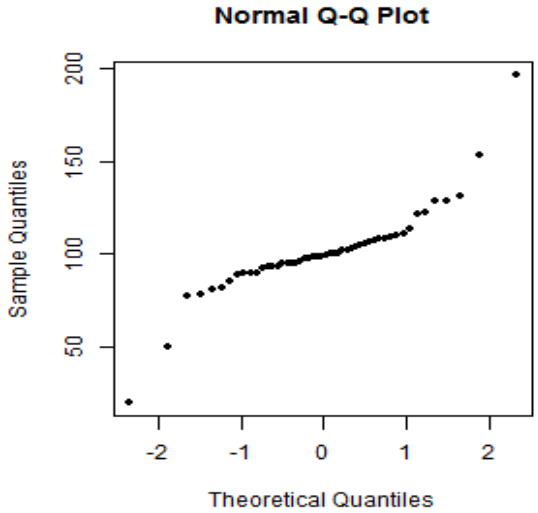
1. What does the greek letter σ denote?
   a. population standard deviation b. population variance c. sample standard deviation d. sample variance

2. Which of the following is a random variable?
   a. population mean b. population standard deviation
   c. sample mean d. degrees of freedom for error

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. If μ represents the population mean and ν represents the population median for a population that is skewed to the right, then
   a. $\mu < \nu$ b. $\mu = \nu$ c. $\mu > \nu$ d. $\mu + \nu = 0$

5. What does this normal plot show about a variable in the dataset? The variable density _____.

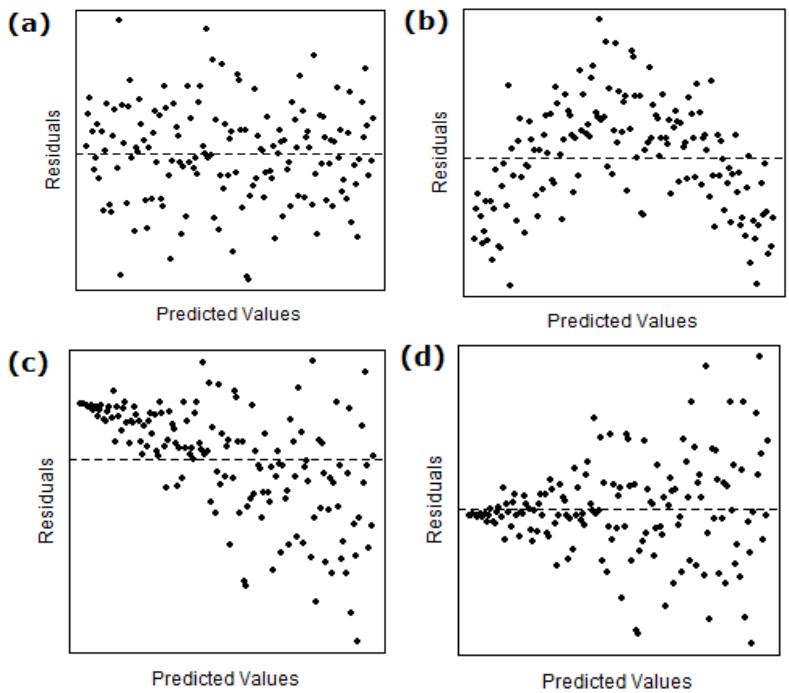



   a. has thick tails b. has thin tails c. is skewed to the left d. is skewed to the right.

6. Which of these assumptions must be true for the independent two-sample t-test to be valid?
   a. the distribution of the underlying population of the two groups must have a t-distribution.
   b. the residuals must be unbiased.
   c. the two samples must be normally distributed.
   d. the combined sample sizes must be smaller than 30.

7. The problem of finding a statistical test for testing the null hypothesis that two samples have the same variance is called the _____ problem.
   a. Behrens-Fisher b. Cauchy-Schwartz c. Kolmogorov-Smirnov d.Welsh-Satterthwaite

8. Which plot shows residuals that are unbiased and homoscedastic?




9. A regression model has multicollinarity if
   a. the dependent variable is highly correlated with at least one of the independent variables.
   b. the model has one or more influence points.
   c. the R-squared value is greater than 0.9.
   d. the variance inflation factor is high.

10. In a logistic regression model, if the odds ratio for an observation is 0.25, what is the probability of success?
    a. 0.2 b. 0.25 c. 0.3333 d. 0.5

**Part B. Short answer questions and problems.** 10 points each. Show your work or explain your answer for each problem.

1. Draw the horizontal box plot of this dataset of hypothetical exam scores:

   8 43 75 78 81 88 95 95 98

   Mark extreme outliers with * and mild outliers with O.

2. Find the expected normal scores for a dataset with n = 3 using Wan der Waerden's method.

3. It is well known that high risk new born babies often have low birth weights. A population of high risk newborns has mean weight 2800 grams. A random sample of 25 high risk newborns is enrolled in a program designed to increase their birth weights. After being in the the program for six weeks, mean of the sample is 3075 grams and its sample standard deviation is 500 grams. Test the null hypothesis that the program does not make a significant difference of the birthweights of newborns enrolled in the program. Use a 95% confidence interval for the test ($\alpha = 0.05$).

4. A random sample of adult female giraffes has mean height 5 meters with standard deviation 1.5. They have a mean weight of 1,000 kilos with standard deviation 75. The correlation between the height and weight is 0.8. Derive a simple regression equation for predicting weight from height for adult female giraffes.

5. For a regression equation, explain the difference between a confidence interval for the expected value of the dependent variable for specified settings of the independent variables and a prediction interval for a new observation for specified settings of the independent variables.

6. For a regression model with 7 independent variables and 24 observations, if the sum of squares for model is 127 and the sum of squares for error is 95, compute the F-value for the overall F test. Is this test significant at the 0.05 level?

7. Compute the R-squared and adjusted R-squared values for the multiple regression model in Problem B6.

8. For the regression model in Problem B6, $\hat{\beta}_1 = 3.5$ and the standard error of $\hat{\beta}_1$ is 1.0. Find a 95% confidence interval for the true regression parameter $\beta_1$.

9. What are influence points? What are some commonly used measures of influence that are used to check regression models?

10. Describe some popular methods of external model validation for a regression model.

11. Explain what logistic regression is and how it differs from ordinary multiple regression.

**Part C. Regression example.** The All Green Franchise Company wants to develop a regression model for predicting X1 (Annual Net Sales in $1000) from the dependent variables : X2 : number of sq. ft./1000, X3 : inventory/$1000, X4 : amount spent on advertizing/$1000, X5 : size of sales district/1000 families, X6 : number of competing stores in district.

Answer these two questions after consulting the SAS and/or R output. The R output does not show the partial residual plots; see the SAS output for them.

1. (30 points) Choose the best regression model out of Model1, Model2, Model3, Model4, or Model 5 for predicting X1 from the independent variables. Justify your choice. For full credit, include the following 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. (10 points) Using your best regression model from Problem C1, what is the predicted annual sales for a store with 2000 square feet of floor space, $450,000 of inventory, $6,000 spent on advertising, 10,000 families in the sales district, and 10 competing stores in the district.