Statistics and Data Analysis, Fall 2015
Final Exam

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

Note. This exam is in-class and open everything (including all kinds of electronic devices). However, an exam taker is not allowed to communicate with any person during the exam. Cheating will result in severe penalty. All data needed for this exam is contained in the MS Excel file "SDA-Fa15_final_data.xlsx." You do not need to return the problem sheet. The maximum number of points that one may earn is 100.

1. (20 points; 5 points each) The 40 numbers contained in the sheet "P1" form a random sample. The population is known to be normally distributed. The population size is 20000.

   (a) Construct a 95% confidence interval for population mean from the sample data.

   (b) Construct a 90% confidence interval for population mean from the sample data.

   (c) Which of the two intervals above is larger? Intuitively explain why.

   (d) Did you use the z distribution or t distribution in the above two problems? Explain why that distribution is appropriate for this problem.

2. (30 points; 10 points each) Two groups of students took the same course taught by two different instructors. Class 1 had 16 students while class 2 had 25 students. After they took an exam at the same time, their scores were recording in the sheet "P2" in two columns. We consider these scores as sample data (from two populations) in evaluating the teaching effectiveness (in only one aspect, of course) of the two instructors. Let μ_i be the population mean of students' grades of this exam by taking this course from instructor i, $i = 1, 2$. Both populations are assumed to be normally distributed. The population standard deviations for the two classes are known to be 7 and 10.5, respectively.

   (a) For each instructor, we wonder whether the average scores of all her/his students (not limited in the sampled class) are higher than 70. Consider the following hypothesis testing at a 5% significance level:

   $$\begin{aligned} \mathrm{H}_0 :& \mu_i = 70 \\ \mathrm{H}_\mathrm{a} :& \mu_i > 70, \end{aligned}$$

   where $i = 1, 2$. What are the p-values for classes 1 and 2?

   (b) Based on your answers in Part (a), write down the concluding statements for instructors 1 and 2, respectively.

   (c) We wonder whether one instructor has higher teaching effectiveness than the other, i.e., whether $\mu_1 > \mu_2$ or $\mu_2 > \mu_1$. Based on your answers in Parts (a) and (b), would you conclude either $\mu_1 > \mu_2$ or $\mu_2 > \mu_1$? If yes, why and at what significance level? If no, why?

3. (10 points; 5 points each) The sheet "P3" contains the same data set as "P2" in a different format. Each row is the information of one student, including her/his class and scores.

   (a) Construct a regression model to use *class* to explain *scores*. Write down the regression formula, R^2, and the p-value(s) of independent variable(s). Comment on the significance of independent variable(s).

   (b) Answer Problem 2c based on the above regression analysis.

4. (20 points; 5 points each) The sheet "P4" contains the scores of four classes of students taking the same exam. For each student, her/his gender and number of credits in the same semester are also recorded.

   Note. For each of the following problem, if you define your own variables, clearly define them.

(a) Construct a regression model to use *class* to explain *scores*. Let class 1 be the reference level. Write down the regression formula, R^2, and the p-value(s) of independent variable(s). Interpret the coefficients and p-value(s) of independent variable(s).

(b) Repeat Part (a) but with class 2 as the reference level.

(c) Construct a regression model to use $\frac{1}{credits}$ to explain *scores*. Show that this is not a good idea.

(d) Construct a regression model to use *class*, *gender*, and their interaction to explain *scores*. Choose your own reference levels. Comment on the validity of the model. Make one suggestion to refine the model (but do not implement it).

5. (20 points; 2 points each) Answer the following true/false questions. Please do not provide any explanation.

(a) To construct a confidence interval, it is always required for the population to be normally distributed.

(b) To construct a confidence interval, it is always required for the sample size to be at least 30.

(c) When conducting a t test for a population mean, increasing the sample size will make the rejection region smaller.

(d) The p-value of a one-tailed test may be greater than 0.5.

(e) In a regression model, an independent variable cannot be qualitative.

(f) In a regression model, there can be at most 16 independent variables in any software.

(g) For two regression models, the one with the higher adjusted R^2 is typically preferred.

(h) In a regression report, when we read the p-value of one variable, we are testing whether the corresponding coefficient is significant nonzero.

(i) Once we include $x_1 x_2$ in a regression model, it is suggested to also include x_1 and x_2 in the model.

(j) If R^2 of a regression model is less than 50%, the model is considered bad.

6. (Bonus: 10 points; 5 points each) Consider the five points $\{(x_i, y_i)\}_{i=1,...,5}$ contained in the sheet "P6." We try to find β_0 and β_1 so that the equation

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

minimizes the *sum of absolute error*

$$\text{SAE}(\beta_0, \beta_1) = \sum_{i=1}^{5} |\hat{y}_i - y_i|.$$

(a) Find $\text{SAE}(\beta_0, \beta_1)$ for $\beta_0 = -1$ and $\beta_1 = 1$. Show the detailed calculations.

(b) We now consider -1, 0, and 1 are the three possible values of β_0 and 0, 0.5, and 1 are those of β_1. Among the nine combinations of β_0 and β_1, find the one that minimizes $\text{SAE}(\beta_0, \beta_1)$. For the winning combination, show the detailed calculations.