STAT 651 Final Test (2 hours, 15 minutes) **December 14th, 2010**

NAME: **Total number of Marks: /35**

Answer all the questions in the exam (questions are on both sides of the paper). There are 8 questions in this paper. Unless stated otherwise, do all tests at the 5% level.

Advice: Look at the marks allocated for each question and don't spend a disproportionately large amount of time on any one question.

When conducting a test, always state your null and alternative. Also state the distribution and/or the test that you use.

Write your solutions in the question paper.

(1) Suppose that random samples are drawn from three populations (all three populations have variance one). An ANOVA is done at the 5%-level. We denote the true mean of population one as μ_1, the true mean of population two as μ_2 and the true mean of population three as μ_3.

(i) Suppose all three populations have the same population mean. Roughly, what is the chance of my rejecting the null in an ANOVA? Explain your answer (no marks will be given without an explanation). [1]

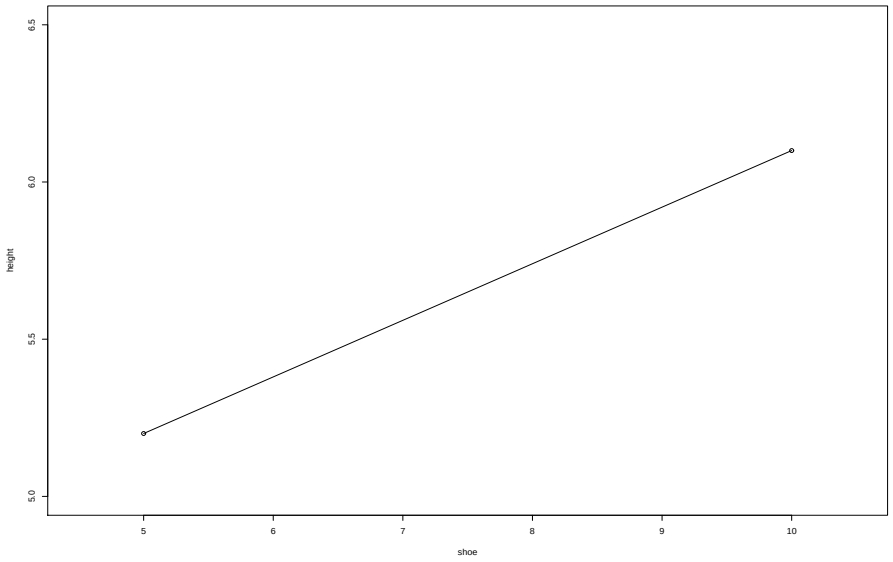
(ii) Under which of the following three scenarios am I most likely to reject the null. (just ring the correct answer - there is no need to justify it). [1]

(A) $\mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 10$, and the size of the samples drawn from all three populations is 50.

(B) $\mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 60$, and the size of the samples drawn from all three populations is 50.

(C) $\mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 12$, and the size of the samples drawn from all three populations is 50.

(2) Doctors want to see whether there is a linear relationship between shoe size and height. A junior doctor collects a sample of <u>two</u> patients, and asks for their shoe size and height. Patient 1 has shoe size 5 and height 5.2, Patient 2 has shoe size 10 and height 6.1. He does a linear regression on these two observations and his results can be found in the plot below.



(i) By studying the above linear regression, what is the R^2? [1]

(ii) Based on the answer in part (i), the junior doctor argues that there is a strong evidence of correlation between shoe size and height. Given the data, do you agree with the junior doctor? [1]

(3) A person needs to take a glucose test to determine whether they are diabetic or not. The results after a test can vary quite a lot. But it is known that glucose levels follow a normal distribution, whose standard deviations is known to be $\sigma = 5$ mg/dL.

A person is said to be diabetic if their mean glucose level is 140 mg/dL or over after a test.

(i) By drawing the appropriate normal distribution calculate the chance of observing a reading of 122 mg/dL or less if that persons mean glucose level is 140. [2]

(ii) Doing multiple glucose tests can be costly. So doctors use the results of the first test to decide whether it is necessary to do another test. In the first glucose test the patient gets a measurement of 122. Based on your answer in part (i), would you repeat the glucose test for that patient or say that they are healthy (give a reason for your answer)? [2]

(4) I want to estimate the mean temperature in College Station during the month of December. Therefore everyday between December 1st to December 31st, I take the average daily temperature reading. I have 31 observations, using this observations I calculate the sample average and the sample standard deviation. The sample average is $\bar{X} = 57^o$F and the sample standard deviation is $s = 7^o$F.

(i) Construct a 95% CI for the true mean temperature. [1]

(ii) What does this CI tell us about the true mean? [1]

(5) Netflicks (an online movie rental company) are trying to understand the behaviour of its customers. They want to know whether the first film a customer has rented has any influence on the second film the customer rents. They divide the films into two categories, Hollywood and independent films, and random select 300 customers who have rented two films in the past two weeks. They observed the following:

Choice	numbers
First movie Hollywood and second movie Hollywood	160
First movie Hollywood and second movie Independent	40
First movie Independent and second movie Independent	50
First movie Independent and second movie Hollywood	50

(i) Based on this data Netflick wants to conduct a statistical test, what would the null and the alternative be? [1]

(ii) Represent the data in an easy to read table which can easily be interpretated. [1]

(iii) Suggest two methods for testing the hypothesis in part (i) [2]

(iv) Do one of the tests suggested in part (iii) (at the 5% level). What is the conclusion of the test. [3]

(6) The National Bureau of Statistics (NBS) wants to determine whether there is a dependence between the income of a man (over the age of 40) and the number of children he has. The NBS sample 50 men (over the age of 40), and ask their income and the number of children. In this sample the incomes varied from 25K - 120K and the number of children from none to four.

Two NBS statisticians Jay and May are wondering how to determine whether there is a relationship. They can not decide on the appropriate test to use. Jay did a linear regression and May did an ANOVA. The JMP output for both tests is given in Figure 1 and 2.

(i) State the hypothesis that should be tested? [1]

(ii) By looking at the data and understanding what we want to test, which method (linear regression or ANOVA) do you think is the most appropriate. Give reasons for your answer (no marks will be given without a reason). [2]

(iii) By selecting the appropriate output, what are the conclusions of the test (no marks will be given without a reason - ring the appropriate parts of the output)? [2].

(iv) Give one of the main assumptions of the chosen test and explain how to check that this assumption is satisfied. [1].

(7) The amount spent on alcohol and tobacco in two different populations is being analysed. 20 individuals are drawn from each population. They are (separately) plotted on Figures 3 and 4. The x-axis corresponds to the amount spent on tobacco and the y-axis the amount spend on alcohol. The results for the linear regression of both samples are given in Figures 3 and 4. Let β_{P1} denote the true slope of population one and β_{P2} denote the true slope of population two.

(i) In the parameter estimates table, state the hypothesis that is being tested for the row denoted 'column 2'. [1]

(ii) By studying both the plots and the outputs explain why the p-value for the slope β_1 is smaller for the sample taken from population 1 than population 2. [2]

(iii) Using the output (Figures 3 and 4), construct confidence intervals for the slopes for both populations (in other words 95% confidence intervals for β_{P1} and β_{P2}). Based on these confidence intervals what would your conclusion be of testing whether they both had the same slope, ie. $H_0 : \beta_{P1} - \beta_{P2} = 0$ against $H_A : \beta_{P1} - \beta_{P2} \neq 0$ (do not do a test, but base your conclusions on the confidence intervals). [2]

(8) Doctors want to understand whether 'tall' male toddlers (children below the age of 3) tend to be taller than 'tall' female toddlers. Let μ_b denote the mean height of the tall boys and μ_g denote the mean height of the tall girl.

(i) State the hypothesis the scientists want to test. [1]

(ii) To statistically test their hypothesis they sampled 30 daycare centers and measured the tallest male and female toddler in each daycare center. Therefore they had 30 pairs of heights (measured in inches). It is unclear to the scientists whether there is dependence between the pairs. Hence they plot the paired data (see Figure 5). They also do both an independent sample t-test and a paired t-test.

(a) Based on the plot in Figure 5, which do you think is the most appropriate test to do? [1]

(b) Using the output and your answer to part (a), what are the results of the hypothesis test in part (i) (give a reason for your answer).

[1]

(c) Construct a 99% CI for the mean difference.

[1]

(iii) Using the output, test the research hypothesis that there is correlation between the tall girl and tall boy heights at daycare centers. Do your results support the test that you chose in part (a)?

[3]