Solutions

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 <u>have</u> the same population mean. Roughly, what is the chance of my rejecting the null in an ANOVA? Explain your answer (<u>no marks</u> will be given without an explanation). [1]

This is the situation under the null $H_0: \mu_1 = \mu_2 = \mu_3$
Hence Reject null $\Rightarrow$ Type I error $= 5\%$

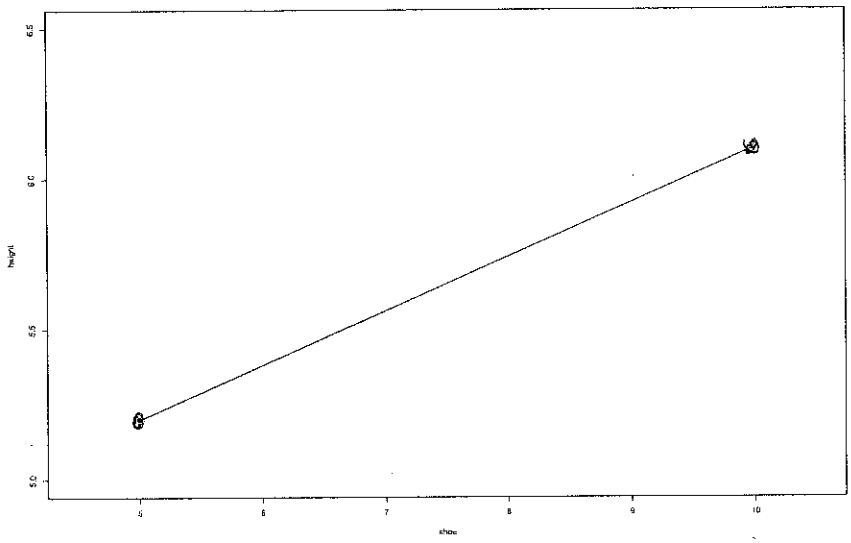
(ii) Under which of the following three scenarios am I most likely to reject the null. (just ring the correct answer - there is <u>no need to justify it</u>). [1]

(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]

Line gives exact fit of points.

Therefore $R^2 = 1$.

(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]

Even though the $R^2 = 1$, there are only two points

Hence the sample size is too small.

Indeed the variance (s.e) of the estimate of the slope, $\hat{\beta}_1$,

will be very large (infinite), which basically tells

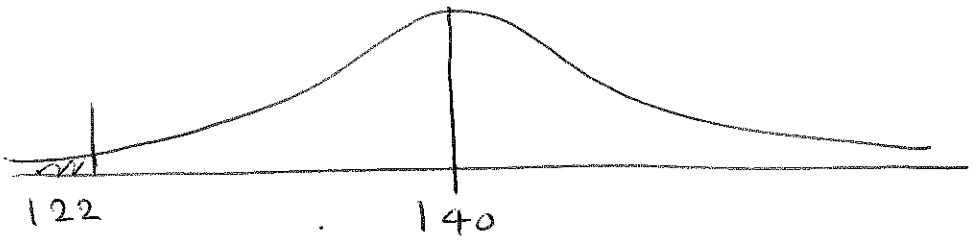
us that no inference can be made.

n.b $s.e(\hat{\beta}_1) = \sqrt{\frac{\sigma^2_\varepsilon}{2-2}}$!!

(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]



$$P(X \leq 122) = P\left(Z < \frac{122-140}{5}\right) = P\left(Z \leq \frac{-18}{5}\right) = P(Z \leq -3.6) = 0.00016$$

(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]

The above result tells us that the chance of observing 122 or less if one is diabetic is 0.00016, which is very small. Therefore at the risk of making a mistake I would not repeat the test - it appears likely that the patient is healthy.

(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]

$$\left[57 \pm 2.04 \times \frac{7}{\sqrt{31}}\right]$$

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

With 95% confidence the true mean lies in the above interval.

(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]

H_0: no dependence between first or second choice H_A: there is a dependence

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

First Choice

Second Choice	Holly	Ind.
Holly	160	50
Ind	40	50

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

method 1 A χ^2-test for independence

method 2 Under the null, it implies that

$P\{\text{Second Holly} \mid \text{First Holly}\} = P\{\text{Second Holly} \mid \text{First Ind}\}$ and

$P\{\text{Second Ind} \mid \text{First Holly}\} = P\{\text{Second Ind} \mid \text{First Ind}\}$

Here we can focus on the subpopulations, those who have taken Holly as first choice and those who taken Ind as first choice and test equality of proportions.

All numbers are larger than 5, hence we can do either one of these tests.

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

<u>Method 1</u> χ^2-test indep.

<u>Observed</u>

	Holly	Ind	
Holly	160	50	210
Ind	40	50	90
	200	100	300

<u>Expected</u>

	Holly	Ind	
Holly	140	70	210
Ind	60	30	
		100	

What we expect to see if independent

χ^2 T = $\frac{(160-140)^2}{140} + \frac{(50-70)^2}{70} + \frac{(40-60)^2}{60} + \frac{(30-50)^2}{30} = 28$.

The rejection level is 3.841. Since $28 > 3.841$, There is evidence to reject the null.

<u>Method 2</u> : Consider the proportion estimates ~~made~~ of the subsamples:

$\hat{\pi}_{H|H} = \frac{160}{200} = 0.8$ $\quad \hat{\pi}_{H|I} = \frac{50}{100} = 0.5$

$H_0: \pi_{H|H} - \pi_{H|I} = 0$

$H_A: \pi_{H|H} - \pi_{H|I} \neq 0$

<u>S.E</u> $= \sqrt{\frac{0.8 \times 0.2}{200} + \frac{0.5 \times 0.5}{100}} = 0.057$.

Hence the non-rejection region is $[0 \pm 1.96 \times 0.057] = [-0.11, 0.11]$.

Since $\hat{\pi}_{H|H} - \hat{\pi}_{H|I} = 0.3$ does not lie in this region evidence to reject null

(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]

To look for dependence we focus on the mean (but it can be trickier than this).

H_0: mean salary for all categories same H_A: mean salary different.

(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]

(a) Number of children is numerical discrete.

(b) The question asks for <u>dependence</u> <u>not</u> <u>linear dependence</u>. Hence we are trying to detect changes in the mean (b) is most important. Here we choose ANOVA.

(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].

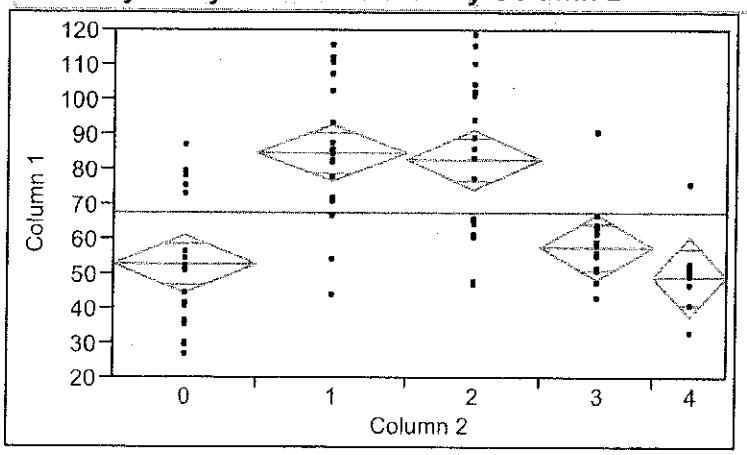
Based on the ANOVA output we see that the p-value is less than 0.001, thus we reject the null.

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

The residuals should be close to normal. Thus calculate the residuals $\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{\cdot j}$ and make a QQplot of residuals.

Oneway Analysis of Column 1 By Column 2

Figure 1 - ANOVA



Missing Rows 85

Oneway Anova

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	4	19038.932	4759.73	14.3727	<.0001*
Error	77	25499.621	331.16		
C. Total	81	44538.553			

$F_{4,77}$

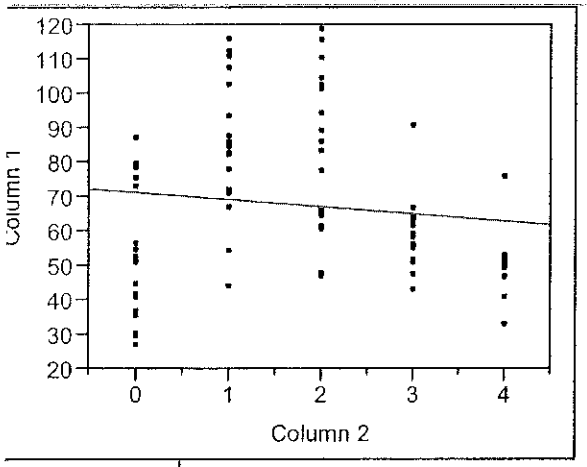
Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
0	19	52.6968	4.1749	44.384	61.010
1	20	84.5680	4.0692	76.465	92.671
2	18	82.5861	4.2893	74.045	91.127
3	15	57.4773	4.6987	48.121	66.834
4	10	48.9370	5.7547	37.478	60.396

Std Error uses a pooled estimate of error variance

Bivariate Fit of Column 1 By Column 2

Figure 2 - linear regression.



Linear Fit

Column 1 = 71.035682 - 2.0868506*Column 2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	629.502	629.502	1.1469
Error	80	43909.051	548.863	Prob > F
C. Total	81	44538.553		0.2874

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	71.035682	4.233245	16.78	<.0001*
Column 2	-2.086851	1.94861	-1.07	0.2874

(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]

The default is: $H_0: \beta = 0$ again $H_0: \beta \neq 0$ This can be change to $H_0: \beta = \beta_0$ again $H_A: \beta \neq \beta_0$.

(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]

The variation about the slope of plot 2 is greater than the variation about the slope of plot 1. Thus standard error for slope plot 1 is smaller than s.e of plot 1. Thus p-value is smaller

(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]

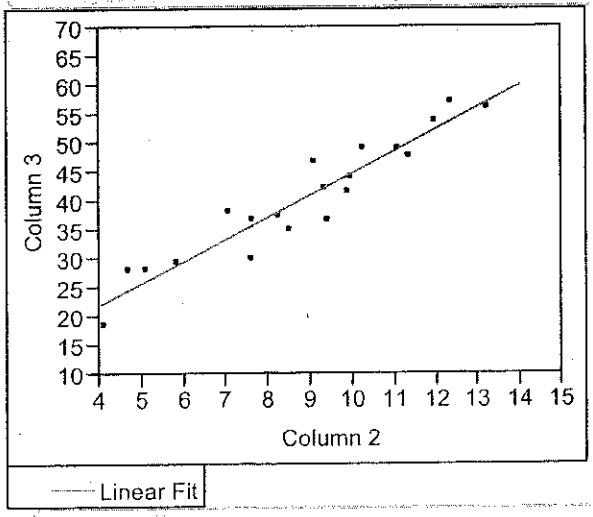
CI for β_{P1} is $[3.21 \pm t_{18}(0.025) \times 0.30]$

$= [3.21 \pm 2.1 \times 0.30] \equiv [3.17, 4.44]$

CI for β_{P2} is $[4.22 \pm 2.1 \times 0.96] = [2.42, 6.45]$.

As both intervals intersect, there is no suggestion that they have different slopes.

Bivariate Fit of Column 3 By Column 2



Linear Fit

Column 3 = 6.4762776 + 3.8075861*Column 2

Summary of Fit

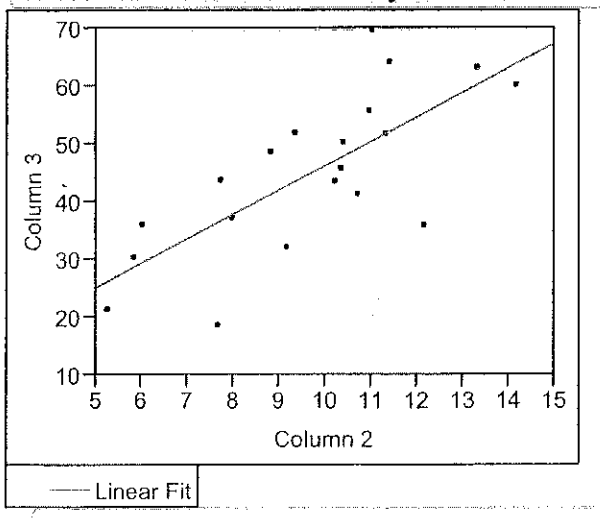
RSquare	0.899778
RSquare Adj	0.89421
Root Mean Square Error	3.371534
Mean of Response	40.0807
Observations (or Sum Wgts)	20

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.4762776	2.748868	2.36	0.0300*
Column 2	3.8075861	0.299521	12.71	<.0001*

Figure 3 - Samples from population 1

Bivariate Fit of Column 3 By Column 2



Linear Fit

Column 3 = 3.7810457 + 4.2256539*Column 2

Summary of Fit

RSquare	0.536593
RSquare Adj	0.510849
Root Mean Square Error	9.716032
Mean of Response	44.757
Observations (or Sum Wgts)	20

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.7810457	9.234549	0.41	0.6870
Column 2	4.2256539	0.925584	4.57	0.0002*

Figure 4 - Samples from population 2

↓

Small variation

↓

large variation

(2) The larger the variation, the larger the variation of the residuals and the larger the s.e and p-value.

(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]

$H_0: \mu_b \leq \mu_g \qquad H_A: \mu_b > \mu_g.$

(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]

Independent sample t-test (Fust plot does not show dependency)

(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).

Looking at the output of an independent sample t-test, we reject the null (p-value is less than 0.0001) [1]

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

$$\left[3.69 \pm t_{59}(0.005) \times 0.36\right]$$

[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)?

$H_0: \rho = 0$

$H_A: \rho \neq 0.$

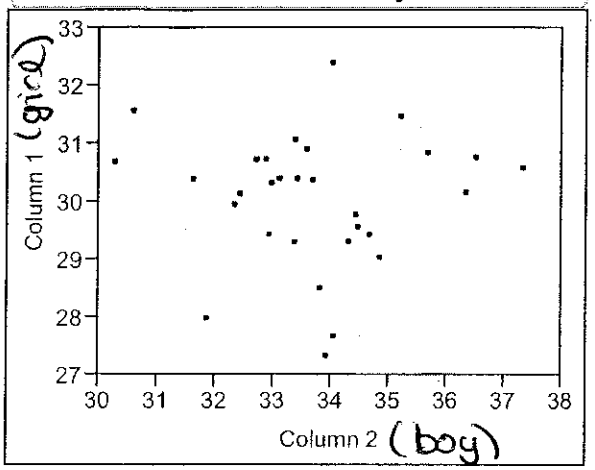
The correlation (estimated) coefficient is $\hat{r} = -0.0175$.

The standard error of ρ is $\sqrt{\frac{1-\hat{r}^2}{30-2}} = 0.179$ [3]

Based on this the non-rejection region is

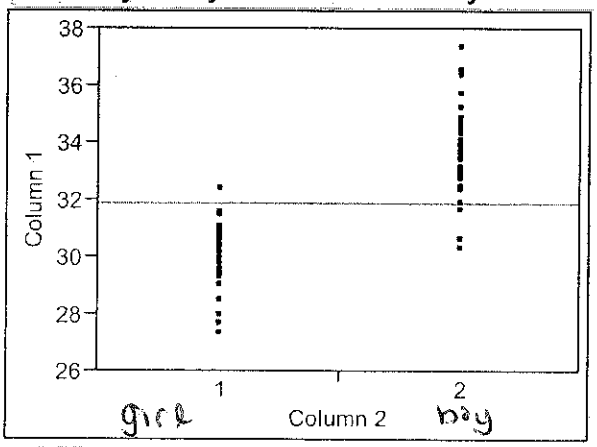
$[-0.5 \times 7, 0.5 \times 7]$. Since -0.0175 lies outside inside this region there is not enough evidence of dependence. Though with appreciably weak.

Bivariate Fit of Column 1 By Column 2



Oneway Analysis of Column 1 By Column 2

Independent Sample t-test.



Difference = average boy − average girl

= 3.68

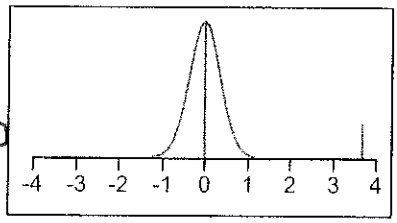
t Test

2-1

Assuming unequal variances

Difference	3.68867	t Ratio	10.28368
Std Err Dif	0.35869	DF	58
Upper CL Dif	4.40829	Prob > |t|	<.0001*
Lower CL Dif	2.96905	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

p-value = less than 0.0001.



Enough evidence to reject the hypothesis in part (a).

Paired t-test

Matched Pairs

Difference: Column 2-Column 1

Column 2	33.6941	t-Ratio	10.19948
Column 1	30.0054	DF	29
Mean Difference	3.68867	Prob > |t|	<.0001*
Std Error	0.36165	Prob > t	<.0001*
Upper 95%	4.42833	Prob < t	1.0000
Lower 95%	2.949		
N	30		
Correlation	-0.0175		