FAMILY NAME: ....
OTHER NAME(S): ....
STUDENT NUMBER: ....
SIGNATURE: ....

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

November 2011

MATHXXXX
Statistics Sample Exam

(1) TIME ALLOWED – 1.5 hours

(2) TOTAL NUMBER OF QUESTIONS – 3

(3) ANSWER ALL QUESTIONS

(4) THE QUESTIONS ARE OF EQUAL VALUE

(5) THIS PAPER MAY **NOT** BE RETAINED BY THE CANDIDATE

(6) STATISTICAL FORMULAE AND TABLES ARE ATTACHED AT END OF PAPER

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

Answer this question in a separate book marked Question 1

1. [20 marks]

a) Compression strength (in psi) was measured on a sample of 58 specimens of a new alloy being developed for aircraft construction. The data is shown below on a stem-and-leaf plot. In this plot 66|4 denotes 66.4×10^3 psi.

66	4											
67	7											
68	0	0	3	4	6	8	9					
69	0	1	2	3	3	5	5	6	7	8	8	9
70	0	0	1	2	3	3	4	5	6	6	8	9
71	0	1	2	3	3	5	6	6	7	8	8	9
72	1	2	3	4	6	7	9					
73	1	3	5									
74	2	5										
75	3											

i) Comment on the shape of the distribution of the data as displayed in the stem-and-leaf plot.

ii) The mean of the sample is 70.7×10^3 psi and the standard deviation of the sample is 1.78×10^3 psi.

A) Determine a 99% confidence interval for the true mean compression strength of the alloy.

B) State any assumptions you need to make to determine this confidence interval, and explain whether you have sufficient information to justify making these assumptions here.

iii) It is of interest to determine the chance that a specimen of the alloy has a compression strength of less than 70×10^3 psi.

A) What proportion of the sample values are less than 70×10^3 psi?

B) Determine a 95% confidence interval for the true proportion of specimens of this alloy which would have a compression strength of less than 70×10^3 psi.

C) State any assumptions you need to make to determine this confidence interval, and explain whether they seem reasonable assumptions in this situation.

b) There are 15 CDs in a box. On 10 of the CDs there are saved data files, and the other CDs have no data files saved on them.

   i) Suppose that 12 CDs are randomly selected. Determine the probability that exactly 9 of these CDs selected have saved data files.

   ii) Suppose instead that the CDs are randomly selected one at a time (without replacement) and checked to determine whether they have data recorded on them. Determine the probability that the 13th CD to be checked is the 10th CD selected which has saved data files.

Answer this question in a separate book marked Question 2

2. [**20 marks**] An experiment was conducted to compare the degree of soiling for fabric copolymerized with three different mixtures of methacrylic acid. The data is displayed below. We would like to know whether the true mean degree of soiling is the same for all mixtures.

Mixture 1	Mixture 2	Mixture 3
0.56	0.72	0.62
1.12	0.69	1.08
0.90	0.87	1.07
1.07	0.78	0.99
0.94	0.91	0.93
$\bar{x}_1 = 0.918$	$\bar{x}_2 = 0.794$	$\bar{x}_3 = 0.938$

a) What assumptions need to be valid for an Analysis of Variance to be an appropriate analysis here?

Assume from now on that these assumptions are valid.

b) An ANOVA table was partially constructed to summarise the data :

Source	df	SS	MS	F
Treatment	**(1)**	0.0608	**(2)**	0.99
Error	**(3)**	**(4)**	**(5)**	
Total	**(6)**	0.4309		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(6).

c) Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether the mixture significantly influences the degree of soiling. (*You can use the numerical values found in the above table, however you are required to write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language - you may use bounds for the p-value.*)

d) Construct a 95% two-sided confidence interval on the difference between mean soiling degree for Mixture 2 and for Mixture 3, that is, $\mu_2 - \mu_3$. Would you conclude that there is a significant difference between these two means? Explain.

e) The engineers responsible for the study carry out a two-sample t-test to compare the mean soiling degree of Mixture 1 to Mixture 2 and obtain a p-value of 0.14. They then carry out a two-sample t-test to compare Mixture 1 to Mixture 3, and obtain a p-value of 0.86.
Does simultaneously analysing the three pairwise comparisons (these two t-tests and the confidence interval in d)) allow you to come to the same conclusion as the ANOVA F-test in c), at overall level $\alpha = 0.05$? Explain.

Answer this question in a separate book marked Question 3

3. [**20 marks**] The relationship between applied stress (variable X, in kg/mm^2) and time to failure (variable Y, in hours) for '18-8 stainless steel' under uniaxial tensile stress is important for many applications.

a) In one study ten different settings of applied stress were used under fixed conditions of temperature and alkalinity.
The resulting data values are given in the table below giving y_i, recorded observations of failure times, for different applied stress levels x_i. A scatterplot is given in the left subplot of Figure 1.

i	1	2	3	4	5	6	7	8	9	10
x_i	2.5	5	10	15	17.5	20	25	30	35	40
y_i	63	58	55	61	62	37	38	45	46	19

The regression model to be fitted is given by

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon.$$

A scatterplot of the residuals versus fitted values is provided in the right subplot of Figure 1.

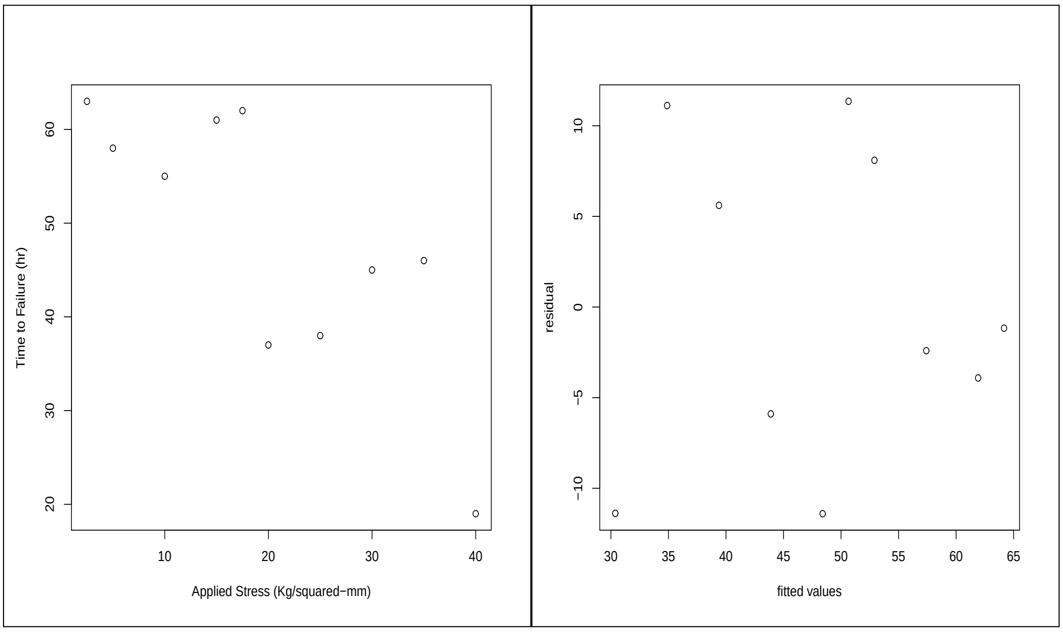

Figure 1: Left Subplot: Plot of Failure Times (h) vs Applied Stress (kg/mm^2). Right Subplot: Plot of regression residuals vs fitted values.

Some regression output is given here for use in your answer to this part of this question.

```
Regression Analysis: Y versus X

The regression equation is Y = 66.4177  -0.9009 X

Predictor   Coef     SE Coef      T       P
Constant    66.4177   5.6481      11.759  2.5e-06
X           -0.9009   0.2428      -3.711  0.00595

S =  9.124   R-squared: 0.6325   Adjusted R-squared: 0.5866
```

i) A) List three essential assumptions that the residuals in the model must satisfy for the above regression analysis to be valid.

B) Using the plots provided in Figure 1 comment on the suitability of these statistical assumptions for this model, where applicable. One or two qualitative statements per plot will suffice to answer this question.

ii) What proportion of variation in the failure time Y is explained by the amount of applied stress X?

iii) Carry out a hypothesis test to determine whether the applied stress is significant in this model, at significance level $\alpha = 0.05$. (*You can use the numerical values found in the above output, however you are asked to properly write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language.*)

iv) Using the regression output above, determine a 95% confidence interval for β_1.

b) Now suppose that based on a different study with many more observations, it is determined that that the relationship between applied stress X (in kg/mm^2) and time to failure Y (h) is described by a simple linear regression model with $\mathbb{E}\big[Y|X = x\big] = 65 - 1.2x$ and $\mathbb{V}\text{ar}\big[Y|X = x\big] = 8$.

i) Suppose that Y_1 (in hours) denotes an observation on time-to-failure made with $x_1 = 25$ (in kg/mm^2) and Y_2 (in hours) denotes an independent observation made with $x_2 = 24$ (in kg/mm^2).
State the distribution of $Y_1 - Y_2$, and state the mean and variance.

ii) A) Find the probability that the time to failure Y_1 (hrs) at the larger applied stress $x_1 = 25$ (in kg/mm^2) exceeds the time to failure at the lower applied stress level $x_2 = 24$ (in kg/mm^2).

B) Comment on this probability and what it says about the effect of changing applied stress.