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

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

November 2011

MATHXXXX
Statistics Sample Paper

(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

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

1. **Answer in a separate book marked Question 1**

a) In an air-pollution study, ozone concentrations were taken in a large California city at 5.00 P.M. The eight readings (in parts per million) were

7.9 11.3 6.9 12.7 13.2 8.8 9.3 10.6

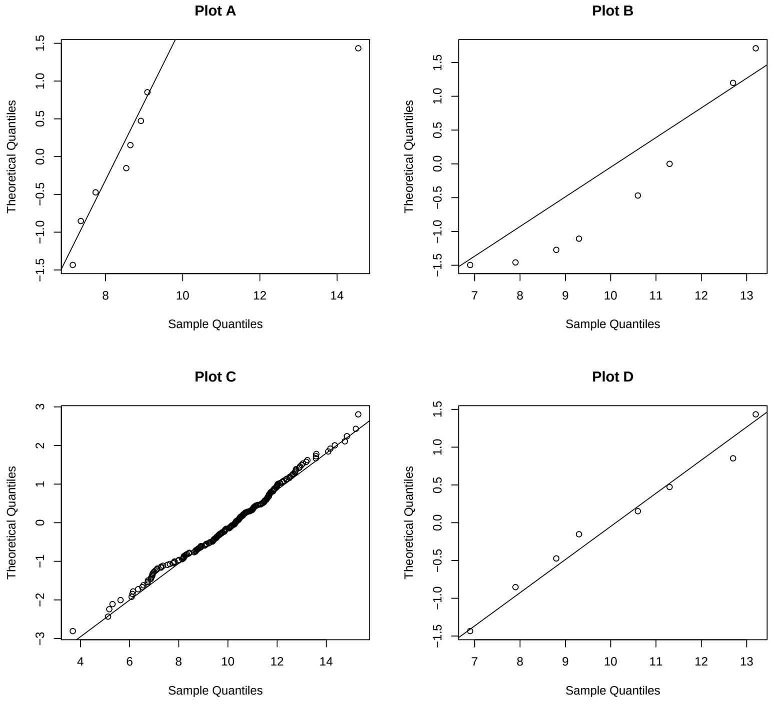
The sample mean and standard deviation are

$$\bar{x} = 10.0875 \text{ ppm} \qquad \text{and} \qquad s = 2.2510 \text{ ppm}$$

i) Determine the five-number summary for this sample.

ii) Draw a boxplot of the data and comment on its features.

iii) A normal quantile plot had been represented for this sample, however it has been mixed up with three other normal quantile plots for other data sets. Which of the 4 quantile plots presented below (A, B, C or D) is the normal quantile plot for this sample? Explain your choice.



iv) From the quantile plot you selected, what is a logical assumption about the underlying distribution of the data? Explain.

v) Based on this sample data, construct a two-sided 95% confidence interval for the true mean ozone concentration in that city.

vi) The mayor claims that mean ozone concentration in his city does not exceed 9 ppm. Can we contradict him? Carry out a one-sided hypothesis test at level $\alpha = 0.05$ and state your conclusion in plain language.

vii) Give bounds on the p-value associated to your test. Interpret these values.

Please see over . . .

b) An article in *The Engineer* reported the results of an investigation into wiring errors on commercial transport aircraft that may produce faulty information to the flight crew. Of 1600 randomly selected aircraft, eight were found to have wiring errors that could display incorrect information to the flight crew.

i) Find an approximate 99% two-sided confidence interval on the proportion of aircraft that have such wiring errors.

ii) How large a sample would be required if we wanted to be at least 99% confident that the observed sample proportion $\hat{p}$ differs from the true proportion p by at most 0.008, regardless of the value p?

2. Answer in a separate book marked Question 2

Suppose that a tensile ring is to be calibrated by measuring the deflection at various loads. In the following table, which gives the results for 12 measurements, the x_i's are the applied load forces in thousands of pounds and the y-values are the corresponding deflections in thousandths of an inch :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	16	35	45	64	86	96	106	124	134	156	164	182

Elementary computations also yield

$$\bar{x} = 6.5 \qquad \text{and} \qquad s_{xx} = \sum_{i=1}^{12}(x_i - \bar{x})^2 = 143$$

The output that results from fitting a simple linear regression model to the data is shown below. The response variable Y is the deflection (in thousandth of an inch) and the predictor variable X is the load (in thousand of pounds). The fitted regression model is given by :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Use the following regression output to answer the questions below.

```
Regression Analysis: Y versus X

The regression equation is Y = 4.35 + 14.8 X


Predictor      Coef   SE Coef     T       P
Constant      4.348    2.244     1.94   0.081
X           14.8182   0.3049    48.60   0.000

S = 3.646   R-Sq = 99.6%   R-Sq(adj) = 99.5%
```

Please see over . . .

a) i) List three essentials assumptions that the error ε in the model must satisfy for the above regression analysis to be valid.

ii) Explain what plots (or other output) you would consider generating to assess how reasonable are these assumptions, and how you would use the output.

Assume from now on that these assumptions are valid.

b) What is the expected change in the deflection for a unit change in the load ?

c) What proportion of variation in the response is explained by the predictor?

d) What is the (sample) correlation between load and deflection? Interpret this value.

e) Give the estimated value of σ, the standard deviation of the error term ε.

f) Carry out a hypothesis test to determine whether the variable X 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 hypothesis, rejection criterion, observed value of the test statistic, p-value, conclusion).

g) Determine a 95% two-sided confidence interval for β_1.

h) The value P associated to the 'Constant' predictor is seen to be equal to 0.081. Interpret this value.

i) Obtain a 95% two-sided prediction interval for the deflection when the load is set to 7.5 thousands of pounds.

3. Answer in a separate book marked Question 3

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5% and 15%. A team of engineers responsible for the study decides to investigate three levels of hardwood concentration : 5%, 10% and 15%. The decide to make up six test specimens at each concentration level. All 18 specimens are tested on a laboratory tensile tester, and the observed tensile strengths (in psi) are shown in the following table :

5%	10%	15%
7	12	14
8	17	18
15	13	19
11	18	17
9	19	16
10	15	18
$\bar{x}_1 = 10$	$\bar{x}_2 = 15.67$	$\bar{x}_3 = 17$
$s_1 = 2.8284$	$s_2 = 2.8048$	$s_3 = 1.7889$

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
Factor	**(1)**	**(2)**	**(3)**	13.04
Error	**(4)**	95.333	**(5)**	
Total	**(6)**	261.111		

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

c) Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether the hardwood concentration significantly influences the tensile strength. You can use the numerical values found in the above table, however you are asked to properly write the detail of the test (null and alternative hypothesis, rejection criterion, observed value of the test statistic, p-value, conclusion - use bounds for the p-value).

d) Construct a 95% two-sided confidence interval on the difference between mean tensile strength at concentration 10% and mean tensile strength at concentration 15%, that is, $\mu_2 - \mu_3$. Would you say that there is a significant difference between these two means? Explain.

e) The engineers responsible for the study also carry out two two-sample t-tests to compare concentration 5% to concentration 10% and concentration 5% to concentration 15%, in terms of the mean tensile strength. They find p-values equal to 0.0059 and 0.0004, respectively. 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. *Hint :* recall the Bonferroni adjustment.