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

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

November 2011

STATISTICS SAMPLE EXAM

(1) TIME ALLOWED – 3 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) **ONLY** CALCULATORS WITH AN AFFIXED "UNSW APPROVED" STICKER MAY BE USED

(7) STATISTICAL FORMULAE ARE AT THE START OF PART B
STATISTICAL TABLES ARE ATTACHED AT END OF PAPER

Part A – consists of questions 1 – 3

Part B – Statistics consists of questions 4 – 6

Both parts must be answered

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

Part B – Statistics

1. Answer in a separate book marked Question 1

a) Oil pipes are internally coated to prevent corrosion. A study was undertaken of the surface roughness (in microns) for 20 sampled pipe-sections from pipes used in oil fields. The data has been sorted and listed below:

$$1.06, 1.09, 1.19, 1.26, 1.27, 1.40, 1.51, 1.72, 1.95, 2.03,$$

$$2.05, 2.13, 2.13, 2.16, 2.24, 2.31, 2.41, 2.50, 2.57, 2.64.$$

The sample has mean of $\bar{x} = 1.8810$ and standard deviation of $s = 0.5239$.

i) Determine the 5 number summary for the data.

ii) Determine whether there are any outliers (using the $1.5 \times IQR$ criterion).

iii) Construct a boxplot for the data and comment on it.

iv) Determine a 95% confidence interval for the mean surface roughness of coated oil pipes.

v) State what assumptions you made to determine the confidence interval for the mean surface roughness. For each assumption comment on whether you have the information to check whether it is valid in this instance.

b) A *k-out-of-n* system is one that will function if and only if at least k out of the n individual components in the system function. Assume that individual components function independently of each other. Assume also each individual component functions with probability 0.9.

Determine the long-run proportion of 3-*out-of*-5 systems that will function.

2. Answer in a separate book marked Question 2

a) During periods of high electricity demand power output from a gas turbine can drop dramatically. One way to counter this is by cooling the air coming in to the engine. A study looked at the heat rates (in kJ/kWhour) from three types of gas turbines (*traditional, advanced and autoderivative)*. The following ANOVA table summarises the data, but it has one missing entry indicated by *.

Source	df	SS	MS	F
Factors	2	55360022	27680011	15.74
Error	64	112537186	1758394	
Total	*	167897208		

i) Calculate the value represented by *. What does this number represent in the study?

ii) Set up suitable notation, and state the null and alternative hypotheses for the ANOVA F-test. What assumptions need to be made for an Analysis of Variance to be an appropriate analysis here?

iii) Using a significance level of $\alpha = 0.01$, (and making the necessary assumptions) complete out the ANOVA F-test. Include in your answer: the observed value and distribution of the test statistic; a mathematical expression for the p-value; the range of values within which the p-value falls (and a statement of how these are obtained); and the conclusion of the test stated in plain language.

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 π by at most 0.008, regardless of the value $\hat{p}$?

c) Suppose that X and Y are independent standard normal variables:

$$X \sim N(0,1), \qquad Y \sim N(0,1).$$

i) What is the distribution of $X + Y$?

ii) Calculate $P(X + Y < 1)$.

3. Answer in a separate book marked Question 3

Concrete is vulnerable to shock vibrations, which may cause hidden damage to the material. In a study of vibration phenomena, an experiment is carried out and data is reported: including the variables `ppv` — peak particle velocity (mm/sec), and `Ratio` — ratio of ultrasonic pulse velocity after impact to that before impact in concrete prisms. Investigators fit the simple linear linear regression model:

$$\texttt{Ratio} = \beta_0 + \beta_1 \times \texttt{ppv} + \epsilon. \qquad (\star)$$

Using the regression analysis output (which includes residual plots and a fitted line plot) at the end of the question, answer the following questions.

a) i) Write the null and alternative hypotheses to test whether the variable `ppv` is significant in predicting the variable `Ratio`.

   ii) Carry out the test at the 1% significance level.

b) i) Determine the value of R^2;

   ii) hence determine the (sample) correlation between the variables `Ratio` and `ppv`.

c) i) Determine a 95% confidence interval for β_1;

   ii) hence determine a 95% confidence interval for the change in the mean of `Ratio` for an increase of 100 mm/sec in `ppv`.

d) In the raw data, what (approximately) were the range of values for the peak particle velocity?

e) For the regression analysis to be valid, the error ϵ in Model ($\star$) must be independent, and normally distributed $N(0, \sigma)$ for some constant σ.
This part of this question is about the assumption that the standard deviation σ is constant:

   i) There are four residual plots. Which of these can be used to verify whether the constant σ assumption is plausible? Explain whether this assumptions is supported, with specific reference to the regression analysis output.

   ii) What is the estimate of σ from the regression analysis of the data?

f) Determine a 95% confidence interval for β_0.

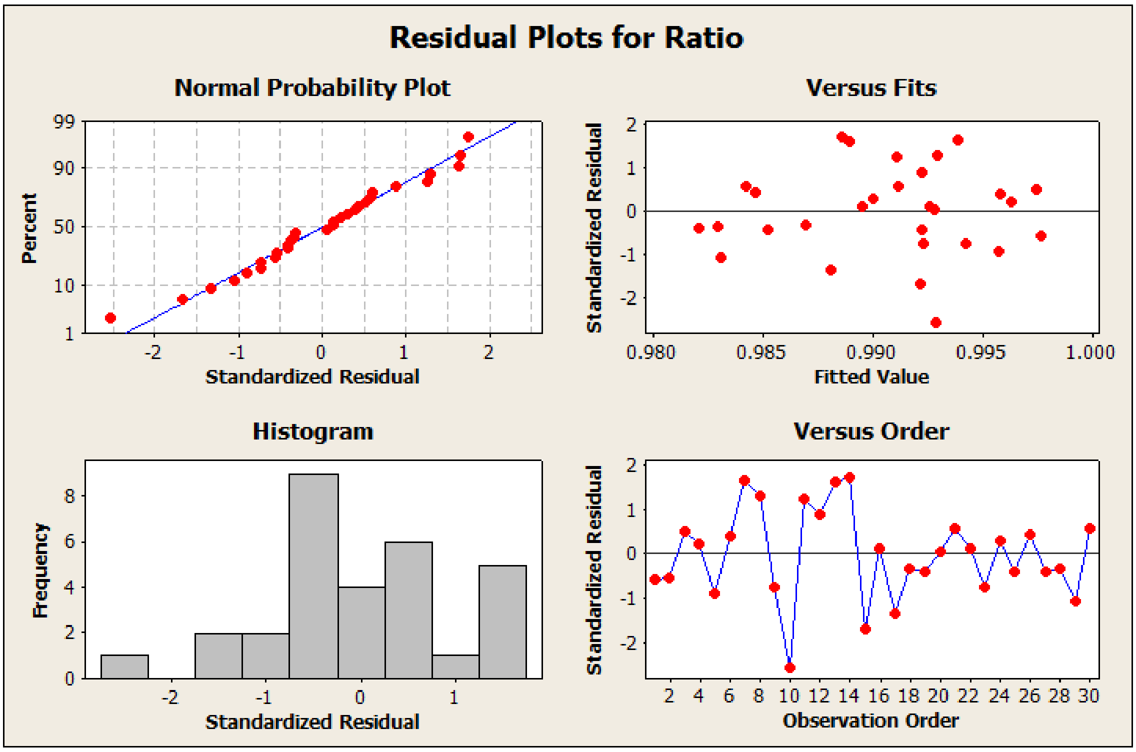
Please see over ...

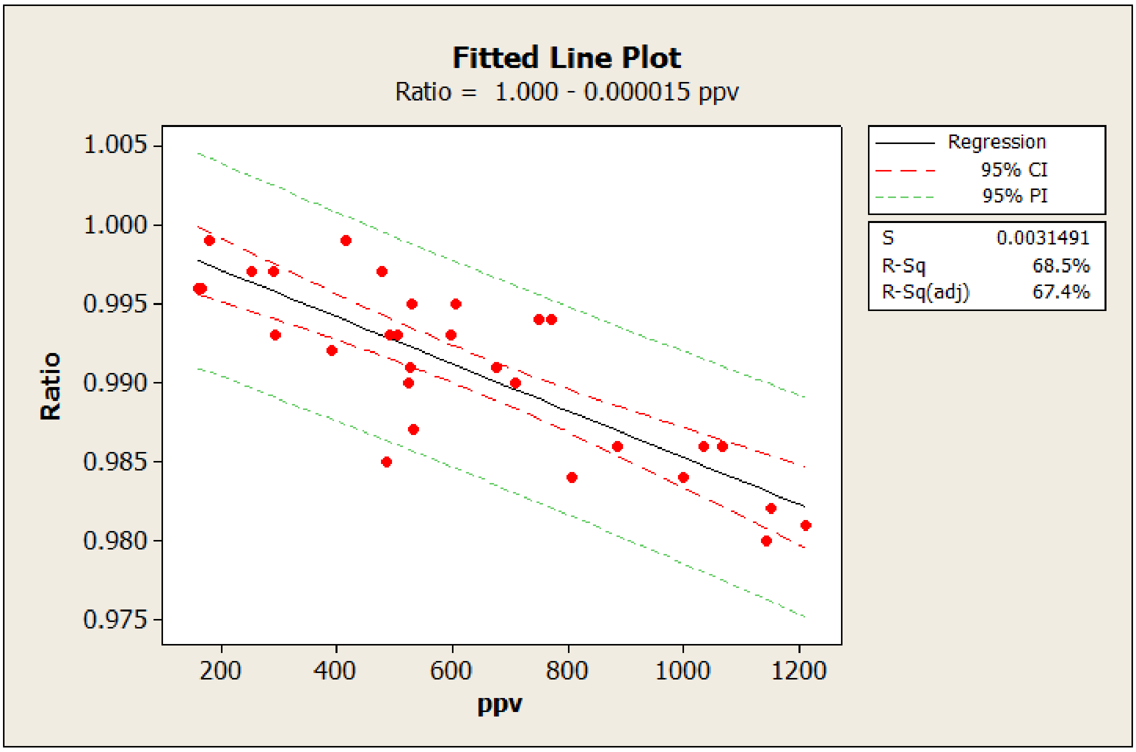
Regression analysis output for Question 3

```
Regression Analysis: Ratio versus ppv

The regression equation is
Ratio = 1.00 - 0.000015 ppv


Predictor          Coef     SE Coef       T      P
Constant        1.00007     0.00131  761.91  0.000
ppv         -0.00001484  0.00000190   -7.80  0.000


S = 0.00314906   R-Sq = 68.5%   R-Sq(adj) = 67.4%


Analysis of Variance

Source          DF          SS          MS      F      P
Regression       1  0.00060380  0.00060380  60.89  0.000
Residual Error  28  0.00027766  0.00000992
Total           29  0.00088147


Unusual Observations

Obs  ppv     Ratio       Fit    SE Fit   Residual  St Resid
 10  486  0.985000  0.992864  0.000629  -0.007864     -2.55R

R denotes an observation with a large standardized residual.

Predicted Values for New Observations

New
Obs       Fit    SE Fit          95% CI                  95% PI
  1  0.988947  0.000625  (0.987666, 0.990228)  (0.982370, 0.995523)

Values of Predictors for New Observations

New
Obs  ppv
  1  750
```

Note: Regression analysis continues on the next page with residual plots and a fitted line plot.

Please see over ...


Residual Plots for Ratio
Normal Probability Plot
Percent
Standardized Residual
Versus Fits
Standardized Residual
Fitted Value
Histogram
Frequency
Standardized Residual
Versus Order
Standardized Residual
Observation Order


Fitted Line Plot
Ratio = 1.000 - 0.000015 ppv
Regression
95% CI
95% PI
S 0.0031491
R-Sq 68.5%
R-Sq(adj) 67.4%
Ratio
ppv