UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING

FINAL EXAMINATION, April, 2012

Second Year — Industrial and Mechanical

MIE237H1 — Statistics & Design of Experiments

Calculator Type: 2

Exam Type: B

Examiner: N. Montgomery

a) There are 11 pages including this page.

b) You should also have a 6 page package consisting of two pages of formulae and four pages of tables.

c) You can use the backs of the question pages for rough work.

d) In some questions I ask you to comment or explain something. Better answers get better marks. A formally correct but not very informative answer may not get full marks for such questions.

e) That being said, brevity is the soul of wit. Very lengthy answers are probably wrong and definitely annoying.

f) If the exact degrees of freedom you need does not appear on the tables provided, use the degrees of freedom that is the closest to what you need.

g) There are 50 marks in total. The number of marks available for each part of each question is indicated.

FAMILY NAME:_____

GIVEN NAME:_____

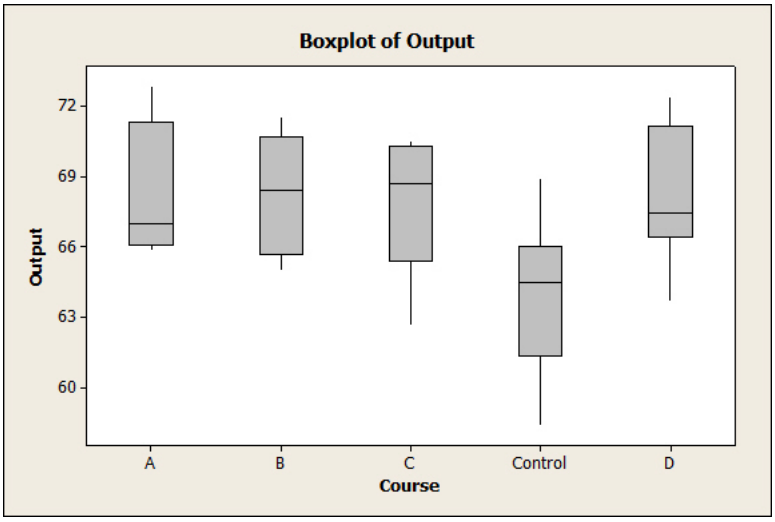
STUDENT NUMBER:_____

1. (15 marks total) An automobile manufacturing company hasn't updated its employee training in many years. They now plan to reevaluate their training program.

A total of $N = 40$ employees will take part in an experiment to compare the value of 4 new training courses, called: A, B, C, and D. The 40 employees will be randomly divided into $k = 5$ groups of $n = 8$ people each. The first group will take course A; the second group will take course B; the third group will take course C; the fourth group will take course D; and the fifth group will take no course at all and will be called the Control group.

After the training courses are complete, each employee's performance is measured over a one month series of shifts, resulting in a score called *Output*, which represents that employee's average production output per shift.

Here are side-by-side boxplots of the results. Minitab puts the groups in alphabetical order, which is why the Control group is between groups C and D, not that it matters.

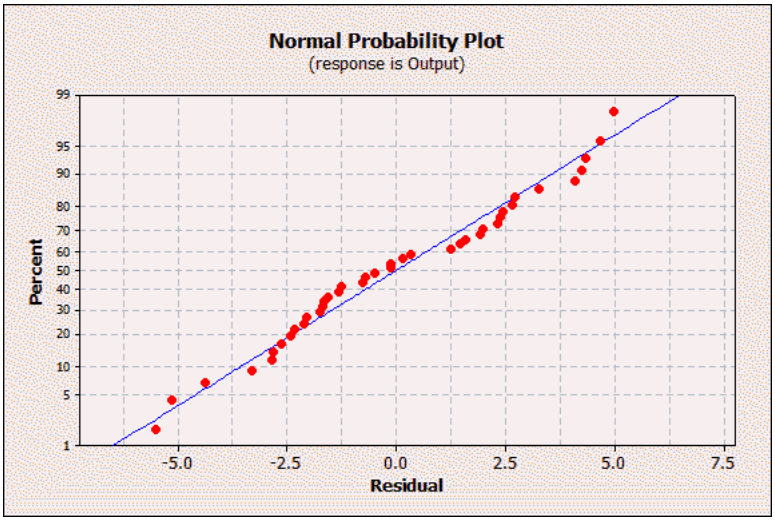


The analysis will be based on the model $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $i \in \{1, 2, 3, 4, 5\}$ and $\varepsilon \sim N(0, \sigma^2)$.

a) **(5 marks)** The total sum of squares is 413.38. The error sum of squares is 301.39. Perform the hypothesis test of $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ versus $H_1 :$ any μ_i different. Use a p-value in your conclusion.

Source	DF	SS	MS	F	P
Course	4	111.99	28.00	3.25	0.023
Error	35	301.39	8.61		
Total	39	413.38			

b) **(2 marks)** Here is a normal quantile plot of the residuals and a summary of the means and standard deviations for each group:



Level	N	Mean	StDev
A	8	68.159	2.847
B	8	68.267	2.737
C	8	67.797	2.845
Control	8	63.896	3.286
D	8	68.021	2.927

Assess whether or not the model assumptions have been satisfied.

Normal plot good.

Balanced experiment and Max variance/min variance $= (3.3/2.7)^2 < 9$.

c) **(3 marks)** Suppose the following statement were true: "Since Course D is the cheapest and fastest to administer, the company was interested in advance of the study to see if there was a significant difference between the control group and Course D". Test this hypothesis H_0*: there is no difference between Course D and the Control* at the $\alpha = 0.05$ level.

$$\omega = \mu_D - \mu_{Co}$$
$$\hat{\omega} = 3.9$$
$$\widehat{\text{Var}}\,(()\,\omega) = 2.153$$
$$\frac{3.9}{\sqrt{2.153}} = 2.66$$
$$p = 2P(t_{35} > 2.66) = 0.01175$$

Reject at 0.05 level.

d) **(3 marks)** Suppose instead that the following statement were true: "The company was surprised to see the observed difference between the control group and Course D, since Course D was so fast and cheap to administer." Test this hypothesis H_0*: there is no difference between Course D and the Control* at the experimentwise error rate of 0.05.

$$q[0.05, 5, 35]\sqrt{\frac{8.611}{8}} = 4.066 \cdot 1.037 = 4.22$$

Do not reject.

e) **(2 marks)** Explain why the conclusions in c) and d) do not contradict each other.

Tukey's test is harder to "pass" due to post-hoc and all-pairwise adjustments etc.

2. **(10 marks total)** A mining company is considering switching to a new brand of oil additive for the diesel engines on its fleet of haul trucks. They are concerned about the amount of calcium contained in the oil additive, since too little can lead to poor oil performance and too much can lead to calcium deposits.

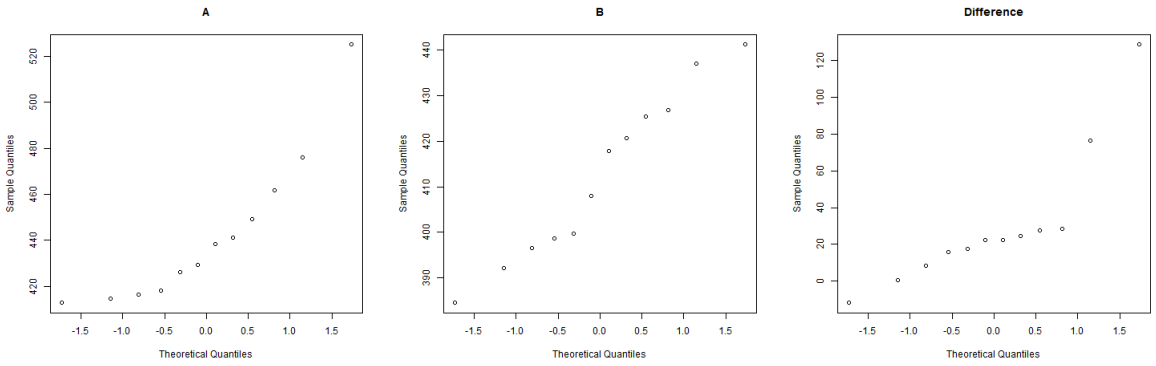
They decide to run an experiment on their 24 haul trucks to see if there is a difference in the average amount of calcium between the old brand and the new brand. The trucks are all of the same model. The trucks are divided at random into two groups of 12 trucks each - group A and group B.

Group A trucks (with identification numbers A01, A02, up to A12) use the old brand of oil additive. Group B trucks (with identification numbers B01, B02, up to B12) use the new brand of oil additive. The trucks then all operate in the same mine for the next 500 operating hours (about 30 days) as usual. An oil sample is then taken from each truck and the amount of calcium in parts per million is determined by a laboratory.

A summer student took the data and made the following spreadsheet with it. The first row of actual data is from group A. The second row is from group B. The third row is the difference between the number in the first row and the number in the second row. At the end of each row are the observed sample averages and the observed sample standard deviations for the numbers in that row.

Sample ID	01	02	03	04	05	06	07	08	09	10	11	12	Average	SD
A	441	416	476	462	426	413	415	429	449	525	438	418	442	33
B	425	408	400	437	399	385	392	441	427	396	421	418	412	20
Difference	16	8	76	25	27	28	23	-12	22	129	17	0	30	38

Here are the normal quantile plots for all three rows of data:



Use the following page to provide an analysis of the data that answers the question *is there a difference in the average amounts of calcium between the old brand and the new brand.* Include the following:

- specify an appropriate model;
- perform the hypothesis test using a p-value in your conclusion;
- comment on whether or not the model assumptions have been satisfied, and if they haven't

been satisfied, whether the violation casts doubt on the validity of your conclusion.

Two sample t-test. Model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$s_p^2 = 744.5$$

$$30/s_p\sqrt{1/12 + 1/12} = 2.69$$

$$p = 2P(t_{22} > 2.69) = 0.0196$$

Evidence against H_0.

Equal variance OK. Normality violated - probably OK via CLT (but suspicion OK too.)

(Incorrect paired model result: $p = 2P(t_{11} > 30/(38/\sqrt{12})) = 0.0181$)

3. **(10 marks total)** A dataset with sample size $n = 63$ consists of an output variable y and five (possible) input variables x_1, x_2, x_3, x_4, x_5. There are 32 possible linear regression models of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$. Excluding an x_i from a model is done by setting its β_i to 0.

Here are the sample correlation coefficients for all pairs of variables, both input and output:

	y	x1	x2	x3	x4
x1	0.156				
x2	0.798	0.023			
x3	0.292	-0.118	0.346		
x4	0.794	0.251	0.596	0.143	
x5	0.674	-0.143	0.888	0.249	0.579

a) **(3 marks)** The final model that you would select using a general sequential method (possibly with forward and backward steps, if necessary) would definitely consist of one of three models summarized here:

Model	Output	R^2
1	Predictor Coef SE Coef T P Constant 0.86870 0.09879 8.79 0.000 x2 0.61177 0.08863 6.90 0.000 x4 0.60958 0.09010 6.77 0.000	79.4%
2	Predictor Coef SE Coef T P Constant 0.9252 0.1301 7.11 0.000 x2 1.1474 0.2036 5.64 0.000 x5 -0.1050 0.1066 -0.99 0.328	66.3%
3	Predictor Coef SE Coef T P Constant 0.88132 0.09562 9.22 0.000 x2 0.9028 0.1531 5.90 0.000 x4 0.63633 0.08784 7.24 0.000 x5 -0.18098 0.07892 -2.29 0.025	81.1%

Give a brief possible explanation of the behaviour of the p-value associated with β_5 in these three models.

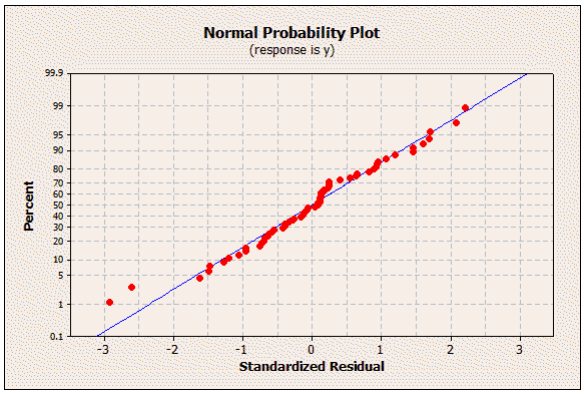
b) **(2 marks)** Which of the three models would you choose, and why?

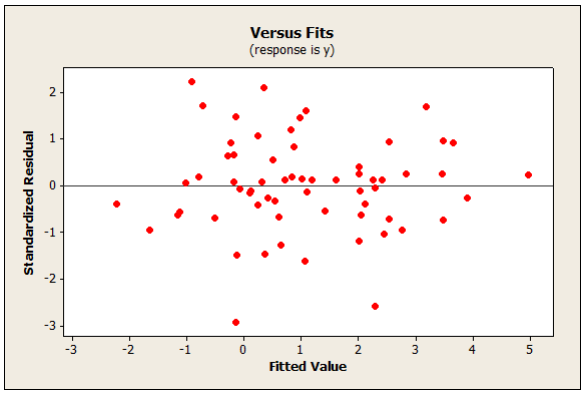
Model 3 - all small p-values.

c) **(3 marks)** The total sum of squares is 176.3. For the model you chose in b), perform the hypothesis test using an F distribution that answers the question "is there any significant linear relationship between the inputs and the response".

SSR=142.96, SSE=33.32, MSR=47.55, MSE=0.559, F=85.3

d) **(2 marks)** The usual residual plots for the three models look almost the same and very similar to the following:





Determine if the model assumptions have been satisfied in this case.

4. (10 marks total) A chemistry lab gathered $n = 12$ observations on the yield y (as a percentage, in other words, as a fraction of 100) of a chemical reaction taken at various temperatures x (in degrees Celsius). A simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ will be considered.

Here is the computer output with some entries covered up with ****. You might also need the fact that $\overline{x} = 225$.

```
The regression equation is
y = **** + ***** x

Predictor      Coef   SE Coef      T      P
Constant     63.540     1.289  49.28  0.000
x          ********  0.005561  *****  *****

S = 1.07691   R-Sq = 97.3%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression      **  423.47  ******  ******  *****
Residual Error  **   11.60  ******
Total           **  435.07
```

a) **(2 marks)** What is the regression equation?

63.54 + 0.106 x

b) **(4 marks)** The dataset consists of 12 rows of numbers in two columns y and x. Compute the observed sample variance of the y column and the observed sample variance of the x column.

var(x)=3410, var(y)=39.55

c) **(4 marks)** Assume that the normal error and equal variance assumptions are satisfied. Compute an interval with 95% confidence for the percentage yield that you would get by performing the chemical reaction at $x = 400$ degrees Celsius. Your interval will not be of any practical use. Explain the likely cause of the problem. (A diagram might help but is not essential.)

$$106.04 \pm 2.228 \cdot 1.077 \sqrt{1 + \frac{1}{12} + \frac{(400 - 225)^2}{37510}}$$

$$(102.7, 109.3)$$

5. **(3 marks total)** This question is based on a true story.

Two weeks ago I visited a company that performs experiments on a particular component of some system, to see how long it takes the component to fail.

After a set of experiments is performed (typically, 80 to 120 of them), the resulting dataset of component failure times is entered in a software package and analyzed as follows. A few details are missing because they did not actually have a manual for the software they were using.

The software package computes the sample mean $\overline{y}$ and the sample standard deviation s of the data, and identifies any points that are further than $k \cdot s$ units away from $\overline{y}$ as "outliers". (They couldn't tell me what k was exactly.) The outliers are discarded, and a new $\overline{y}$ and s are computed. This procedure continues until no outliers remain.

A normal quantile plot along with a sample mean and sample standard deviation is produced using this reduced dataset. The company's conclusions are based on these final results (which may include the construction of a confidence interval or a hypothesis test, but that doesn't matter here.)

Comment on the data analysis procedure I've just described.

6. **(2 marks total)** One of the tables in the package of formulae and tables, with the title "$\alpha = 0.05$ *critical values of the studentized range distribution*", is used for Tukey's all-pairwise-comparison procedure. Explain why it makes intuitive sense that the numbers along each row always increase.