Chemical Engineering: 4C3/6C3
Statistics for Engineering
McMaster University: Final examination

Duration of exam: 3 hours **Instructor: Kevin Dunn**
07 April 2012 **dunnkg@mcmaster.ca**

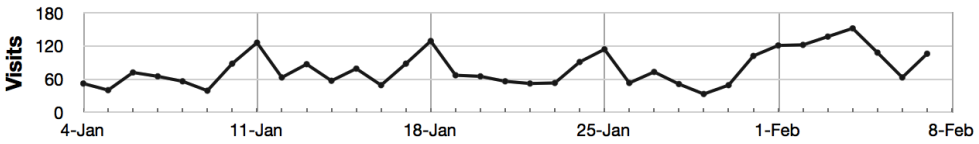
This exam paper has 8 pages and 13 questions. You are responsible for ensuring that your copy of the paper is complete. Please bring any discrepancy to the attention of the invigilator.

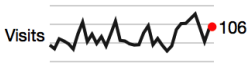
Special instructions

- You may bring any printed materials to the final exam – any textbooks, any papers, *etc*.
- You may use **any calculator** during the exam.
- Please answer the questions in any order in the examination booklet, in pencil or in pen.
- *Time saving tip*: please use bullet points to answer, where appropriate, and **never repeat the question** back in your answer.
- If anything seems unclear, or information appears to be incomplete, please make a *reasonable* assumption and continue with the question.
- **400-level students**: please answer all the questions, except those marked as 600-level questions. You will get credit for answering the 600-level questions though.
- **600-level students** will be held to a higher level of technical accuracy than 400-level students.
- **Total marks**: 100 marks for 400-level; 110 marks for 600-level students.

Question 1 [4 = 2 + 2]

The data shown here are the number of visits to a website:





1. What are the 2 types of plots shown?
2. List some advantages of the plot on the right.

Question 2 [3]

Describe why *and* how real-time control charts can save companies money in processes where there is a long time delay from the point of production until getting the lab results back from quality control testing.

Question 3 [9 = 2 + 2 + 1 + 1 + 3]

A filling line in our company fills bottles by weight. The main quality criterion is to ensure the fill weight is stable over time, around the target of 300 mg. Using data from many days of operation, the standard deviation of fill weights was determined to be 21.7 mg and it remains constant, since the equipment has weekly maintenance.

1. Calculate the Shewhart chart control limits that could be used to monitor the fill weights.
2. A basic Shewhart chart is really not the best choice to monitor the fill weights. Explain why not, and what can be used instead.
3. Calculate the lower and upper control limit for an EWMA chart using $\lambda = 0.1$.
4. Calculate the lower and upper control limit for an EWMA chart using $\lambda = 0.9$.
5. Comment on the Shewhart and two EWMA limits you have calculated in this question, and explain why their numeric values make sense in the context of this monitoring problem.

Question 4 [11 = 2 + 4 + 5]

Using a 2^3 factorial design in 3 variables (**A** = temperature, **B** = pH and **C** = agitation rate), the profit, y, from a chemical reaction was recorded, in standard order.

Experiment	A	B	C	y
1	$-$	$-$	$-$	72
2	$+$	$-$	$-$	73
3	$-$	$+$	$-$	66
4	$+$	$+$	$-$	87
5	$-$	$-$	$+$	70
6	$+$	$-$	$+$	73
7	$-$	$+$	$+$	67
8	$+$	$+$	$+$	87

- **A** = $\dfrac{\text{temperature} - 150°\text{C}}{10°\text{C}}$
- **B** = $\dfrac{\text{pH} - 7.5}{0.5}$
- **C** = $\dfrac{\text{agitation rate} - 50\text{rpm}}{5\text{rpm}}$

1. Show a cube plot for the recorded data.
2. Estimate the main effects and three 2 factor interactions by hand.
3. Interpret all the significant factors you identify in part 2 of this question. Clearly explain what any of the significant 2 factor interactions imply and how it can be used to your advantage to improve the process profitability.

Question 5 [16 = 1 + 2 + 2 + 3 + 4 + 4]

Experiments were conducted by varying the temperature, **T**, and catalyst level, **C**, in order to find conditions that lead to improved conversion, y.

- **T** = 350 K at the low level and 360 K at the high level.
- **C** = 3% catalyst at the low level and 7% catalyst at the high level.

The following experimental data were collected, using coded form:

Experiment	T	C	y
1	0	0	53
2	−1	−1	36
3	+1	−1	45
4	−1	+1	41
5	+1	+1	60
6	0	0	49
7	1.41	0	52
8	0	1.41	49
9	-1.41	0	41
10	0	-1.41	38
11	0	0	51

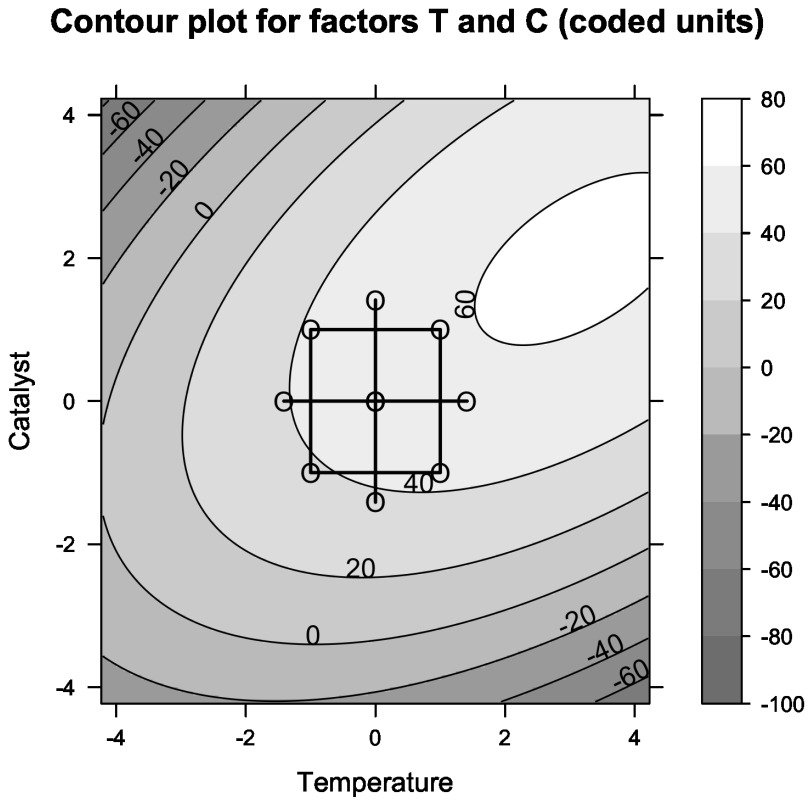
The following output was obtained when building a model in R:

```
Call:
lm(formula = y ~ T + C + T * C + I(T^2) + I(C^2))

Residuals:
      1       2       3       4       5       6       7       8       9
-1.8297  1.2604 -0.7336  2.3564 -2.4565 -1.0423  1.9266  0.5124  3.0021
     10      11
-0.9979 -1.9979

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.9979     1.5518  32.219 5.41e-07 ***
T             5.4550     0.9517   5.732  0.00226 **
C             4.4520     0.9517   4.678  0.00544 **
I(T^2)       -1.6261     1.1356  -1.432  0.21159
I(C^2)       -3.1351     1.1356  -2.761  0.03980 *
T:C           2.5000     1.3439   1.860  0.12193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.688 on ___ degrees of freedom
Multiple R-squared: 0.9298,     Adjusted R-squared: 0.8596
F-statistic: 13.25 on 5 and 5 DF,  p-value: 0.006562
```

1. How many degrees of freedom are available to estimate confidence intervals?
2. Calculate the confidence interval for factor **A** at the 95% level using the above software output.
3. Why might the experimenters have included runs 1, 6 and 11?

4. After finishing experiments 1 to 6, **why** would the experimenters have added points 7 to 10? **What** would they have seen in the data from experiments 1 to 6 that made them add these 4 extra experiments? Provide the necessary calculations to justify your answer.

5. Given the following plot:



at which conditions of **T** and **C**, in real-world units, would you run the next experiment to see if you can achieve maximum conversion?

6. Calculate the predicted conversion at your chosen conditions of **T** and **C**; also give an approximate prediction interval for your prediction.

Question 6 [10 = 3 + 1 + 2 + 4]

1. Give the generators *and* defining relationship, in terms of factors **A**, **B**, **C**, **D**, **E**, and **F**, for a set of fractional factorial experiments using 6 factors, in the fewest number of runs. However, we require that main effects not be confounded with two factor interactions.

2. How many experiments would be required?

3. What is the resolution and projectivity of these experiments?

4. In general, list some advantages of fractional factorial designs and describe how these designs should be used in practice.

Question 7 [9]

Eleven males participated in a exercise and diet program designed to stimulate weight loss. Their weight both before and after participation in the program is shown in the following list. Show whether there is evidence, or not, to support the claim that this particular program is effective in reducing weight. Clearly explain all your calculations.

Individual	Before	After
1	195	187
2	213	195
3	247	221
4	201	190
5	162	175
6	210	197
7	215	199
8	246	221
9	294	278
10	310	285
11	152	178

Question 8 [6 = 2 + 4]

Latent variable methods can solve some shortcomings of *classical* statistical tools. For example, a least squares model built with very highly correlated x-variables will likely have confidence intervals that span zero, for those highly correlated variables.

1. What is the interpretation of a confidence interval calculated from a least squares model, for example $b_{\text{low}} < \beta < b_{\text{high}}$?
2. Describe two shortcomings of other *classical* statistical tools that are addressed by using latent variable models, such as projection to latent structures (PLS) and/or principal components analysis (PCA).

Question 9 [12 = 2 + 2 + 2 + 2 + 2 + 2]

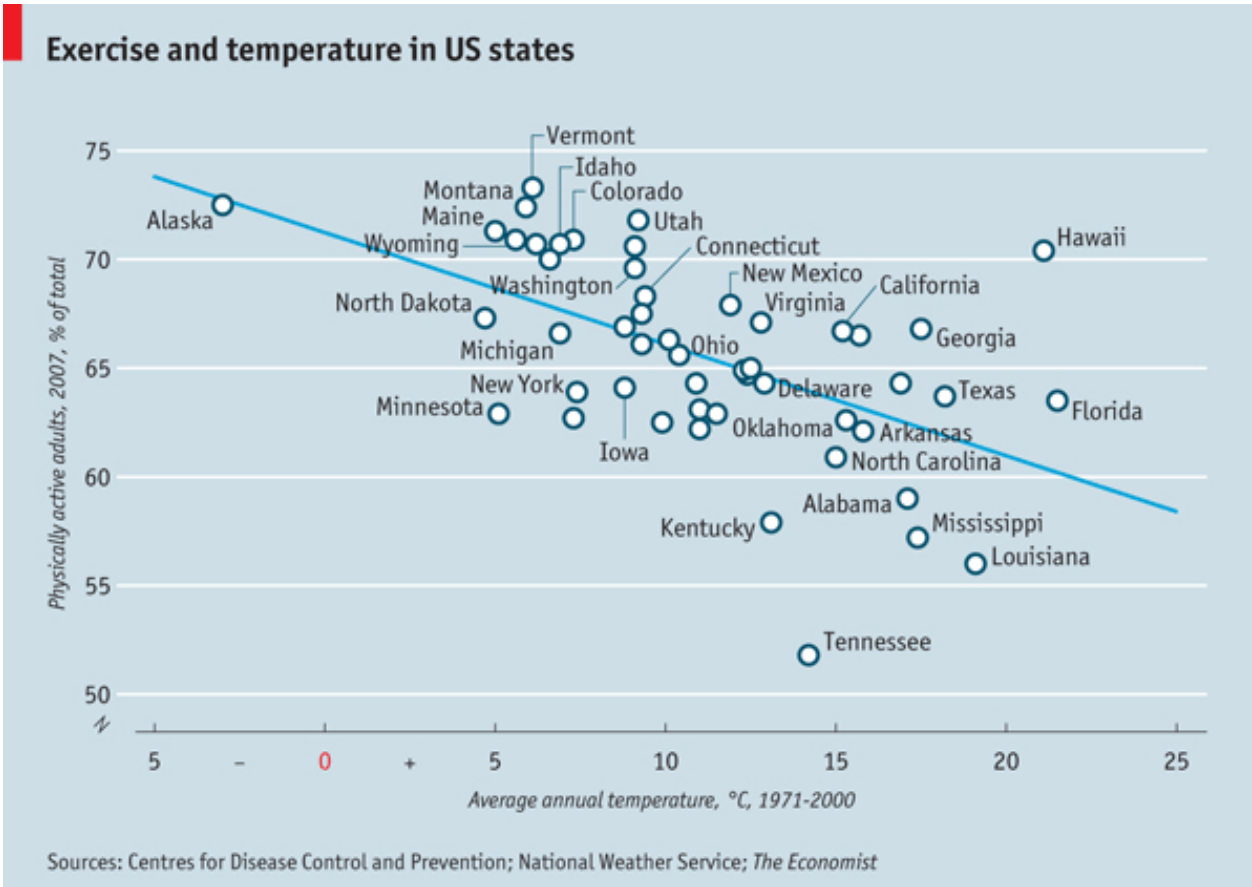
A new bottle filling line is being used for producing a therapeutic nasal spray. The quality assurance (QA) group at your company requires a 95% confidence interval for the average fill weight. This is required to register the manufacturing process with Health Canada, and your company may not sell the product until they get Health Canada approval.

1. What does a confidence interval for the average fill weight tell Health Canada?
2. If the filling standard deviation is known to be 4.4 mg, how many samples must be weighed to ensure the true fill weight is within ± 3.5 mg, i.e. within a range of 7.0 mg?
3. What assumption(s) regarding the sample weights are implicit in the previous question?
4. The current sampling strategy is to run the line for the first 500 bottles and discard them, since the machine and equipment are warming up during this period. Then run another 500 bottles and weigh the last n bottles from those 500 to calculate the confidence interval. Explain why, or why not, this strategy is appropriate.

5. Our company aims for this line to be a 6-sigma process. Given that the target fill weight is 70 mg, with symmetrical specification limits of ± 10 mg around the target, what is our current process capability?
6. If we are to remain symmetrically around the target, within the given specification limits, to what standard deviation must we control the fill weight to be a 6-sigma process?

Question 10 [8 = 2 + 2 + 2 + 2]

The following figure, taken from The Economist shows the percentage of physically active adults against the average annual temperature, broken down by geographical regions, according to the USA state.



1. Since visualization plots can often stand alone without accompanying text, what is the plot's author asking you to infer from this visualization?
2. Is there a causal relationship in the data? Explain your answer.
3. The author has shown a linear regression line. Is the intercept term meaningful in this case; please explain.
4. Calculate an estimate of the linear model's slope, and give an interpretation for it.

Question 11 [8 = 4 + 4]

In a chemical plant there are two "identical" reactors that can be run in parallel. Raw materials from a common source are fed continuously to the reactors, but the composition and purity of the raw materials are known to vary or drift with time. The standard operating temperature of these reactors is 120°C, but there is a proposal to use 140°C. It is claimed that this increased temperature should improve the yield of the process.

1. Set up an experimental design over a six-day period to test the hypothesis that the increased temperature will improve the yield. Be specific and concise in your instructions to the plant operators. They will run the experiments exactly as instructed. Both reactors can be operated each day but each can only be operated at one temperature on a given day.
2. Show how you would analyze the results from your experimental program.

Question 12 [4 = 2 + 2]

A company has been producing a polymer for the past 5 years. There are plenty of historical data available (two values per day) that give the conversion of the raw material monomer to the final product, the polymer. Another engineer has just finished a sequence of $N = 8$ experiments (13 to 20 in the table below), to test whether a cheaper catalyst, **B**, has any effect on product conversion when compared to the existing catalyst, **A**.

The engineers mistakenly thought that by running the experiments on alternating days they would be able to get an unbiased result of the cheaper catalyst's effect on conversion. Given below are the conversion data:

Experiment	Catalyst used	Conversion [%]	Experiment	Catalyst used	Conversion [%]
1	**A**	85	11	**A**	84
2	**A**	78	12	**A**	80
3	**A**	81	13	**B**	76
4	**A**	79	14	**A**	79
5	**A**	97	15	**B**	71
6	**A**	70	16	**A**	76
7	**A**	87	17	**B**	86
8	**A**	74	18	**A**	75
9	**A**	89	19	**B**	92
10	**A**	77	20	**A**	83

1. Describe why the 8 experiments, numbered 13 to 20, could show a misleading result when trying to test the difference between catalysts **A** and **B** using only those 8 data points.
2. Draw a table in your answer booklet to show how *you* would have allocated the choice of catalyst **A** or **B** on days 13 to 20.

Question 13 [600-level student question: 10 = 7 + 3]

Biological drugs are rapidly growing in importance in the treatment of certain diseases, such as cancers and arthritis, since they are designed to target very specific sites in the human body. This can result in treating diseases with minimal side effects. Such drugs differ from traditional drugs in the way they are

manufactured – they are produced during the complex reactions that take place in live cell culture. The cells are grown in lab-scale bioreactors, harvested, purified and packaged.

These processes are plagued by low yields which makes these treatments very costly. Your group has run some experiments to learn more about the system and find better operating conditions to boost the yield. The following factors were adjusted in the usual factorial manner:

- **G** = glucose substrate choice: a binary factor, either $\mathbf{G}-$ at the low level code or $\mathbf{G}+$ at the high level.
- **A** = agitation level: low level = 15 rpm and high level = 25 rpm, but can only be set at integer values.
- **T** = growth temperature: 30°C at the low level, or 35°C at the high level, and can only be set at integer values in the future.
- **C** = starting culture concentration: low level = 1100 and high level = 1500, and can only be adjusted in multiples of 50 units.

A fractional factorial in 8 runs, created by aliasing **C = GAT**, has given the following 8 model coefficients, when **C**, **G**, **A** and **T** are centered and scaled (coded) in the usual way:

- **I + GATC** $= 24$
- **G + ATC** $= +3.5$
- **A + GTC** $= -1.5$
- **T + GAC** $= +4.0$
- **C + GAT** $= +3.5$
- **GA + TC** $= -0.18$
- **GT + AC** $= -0.09$
- **GC + AT** $= +0.20$

The aim is to find the next experiment that will improve the yield, measured in milligrams, the most. Since your manager has seen that temperature has a strong effect, she has requested the next experiment be run at 40°C, which also happens to be the highest level you can adjust the bioreactor to.

1. Give the experimental conditions for all 4 factors for the next experiment. The conditions are to be reported in both real-world units, as well as in the usual coded units of the experiment, presented in a table.
2. What is the expected yield at your proposed experimental conditions?

The end.