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

Duration of exam: 3 hours **Instructor: Kevin Dunn**
25 April 2013 **kevin.dunn@mcmaster.ca**

This exam paper has 6 pages and 10 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 **ONE** of the 600-level questions (as indicated).
- **600-level students** will be held to a higher level of technical accuracy than 400-level students.
- **Total marks**: 100 marks for 400-level; 122 marks for 600-level students.

Question 1 [20]

1. Name a reason why a company (or yourself) would run a set of saturated fractional factorials. **[2]**
2. Why is the principle of minimizing “data ink” so important in an effective visualization? Give an engineering example of why this important. **[4]**
3. Why are latent variable methods effective for dealing with modern data sets? Your answer must also clearly describe the problem faced with these modern data sets. **[4]**
4. If you are a new employee at a company, e.g. a petrochemical corporation, give two characteristic features than will make you realize an EVOP strategy is being applied on the process. **[3]**
5. Why are robust statistics, such as the median or MAD, important in the analysis of modern data sets? Explain, using an example, if necessary. **[3]**
6. Explain the intention of blocking in experimental designs. **[2]**
7. What is a 6-sigma process? Use the example of a process that is filling cereal boxes with breakfast cereal in your answer, to help explain. **[2]**

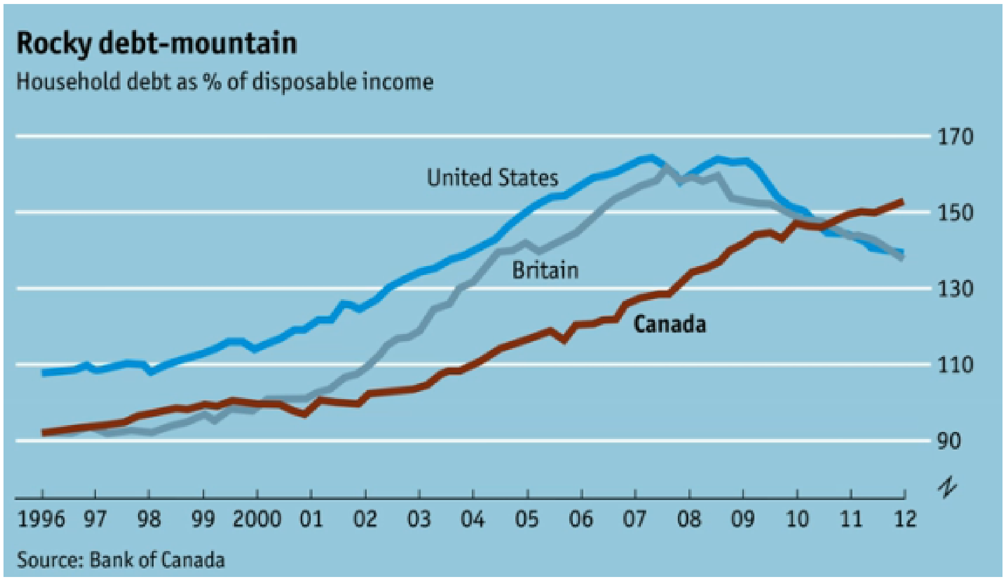
Question 2 [600-level students only (NO extra credit for 400-level students); 8 = 4 + 4]

Your plastic production line (polypropylene) uses the melt flow index as the main measure of quality. A Shewhart chart built on data from the past year of production shows many alarms, periods when unacceptable product is produced, and periods when the production is really stable, with little variability.

1. Now you wish to build a least squares model that uses the reactor temperature as an input variable to predict melt flow index; that is the only purpose of the model. However, you notice that the variance of melt flow index is not constant at all values of temperature. Furthermore, the least squares model errors are not normally distributed. What would you do next if faced with this situation? **[4]**
2. Your manager is requesting you to calculate the C_{pk} value from this same production line. The C_{pk} value is being requested by a potential, new customer. Explain to your manager what you can or cannot provide as a result of this situation. Clearly explain your reasoning **why**. **[4]**

Question 3 [5]

The following data visualization is from *The Economist*.



Since every plot should carry a meaningful message that the author is trying to tell, what is your interpretation of the above figure? In your answer, describe the type of plot being shown, and critique its effectiveness.

Question 4 [6 = 2 + 4]

You have two production lines in your company, producing the same product, which is sold to the same customers. Production line TL-419 has a $C_{pk} = 0.90$ and line TL-417 has $C_p = 1.2$ (notice that one is C_{pk} and the other is C_p).

1. When should one use C_{pk} and when should one use C_p to assess the process capability? **[2]**
2. Write a few bullet points to your manager to explain **which production line** should receive most of the $200,000 annual budget for process improvements. **[4]**

Question 5 [30]

You and your colleagues have been planning a designed experiment for several days to optimize the effect of **A** = pH and **B** = batch duration (time) in a bioreactor. The secretive response variable is just known as y to you; it can be measured with good precision (± 1 unit) . You know that you are near the optimum y, which is a number you are trying to maximize. **A** and **B** are the only factors that can be adjusted.

1. How do we know we are near an optimum when implementing the response surface methodology on a system? **[3]**
2. Response surface methods follow a predictable algorithm when starting off, usually when we are far from the optimum. But describe one way in which the response surface strategy (algorithm) changes when approaching an optimum. **[2]**

You specified some experiments to your trusted operator to run while you were on vacation. The list was provided to him in **standard order**, but you asked him to run them randomly. When you returned back from vacation you had the following waiting in your email:

Experiments, in the order they were run	**A**	**B**	y
1	9.6	400	230
2	6.0	500	97
3	7.5	400	240
4	5.4	400	140
5	8.4	500	210
6	9.0	500	215
7	9.0	300	250
8	7.5	541	153
9	6.0	300	241
10	7.5	259	280

3. Why must experiments be run in random order? **[2]**
4. Draw a cube plot of the experiments, and superimpose approximate contour lines. **[4]**
5. What is the name of this set of experiments? **[2]**
6. Your operator is not currently available to talk with. What do you suspect happened with experiments 5 and 6? **[2]**
7. In general terms (i.e. do not provide exact values), where would you choose to operate the process next to maximize the y value. Give your answer in real-world units for factors **A** and **B**. **[3]**
8. Write out the

   (a) the equation for converting real-world units of **A** and **B** to coded units,

   (b) the model equation that you would use to fully exploit the potential of the data acquired,

   (c) the $\mathbf{X}$ matrix for this model

   (d) the $\mathbf{y}$ vector for this model

   (e) and state how many degrees of freedom you will have to estimate confidence limits with.

   You must write this matrix and vector so that you can estimate the coefficients using the least squares formula $\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. **[12]**

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 set of experiments using these 6 factors. Each experiment is extremely expensive on this process, so you must minimize the total experimental cost.
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 [600-level students only (extra credit will be given to 400-level students); 14]

Your group is developing a new product, but have been struggling to get the product's stability, measured in days, to the level required. You are aiming for a stability value of 50 days or more. Four factors have been considered:

- **A** = monomer concentration: 30% or 50%
- **B** = acid concentration: low or high
- **C** = catalyst level: 2% or 3%
- **D** = temperature: 393K or 423K

These eight experiments have been run so far:

Experiment	Order	A	B	C	D	Stability
1	5	$-$	$-$	$-$	$-$	40
2	6	$+$	$-$	$-$	$+$	27
3	1	$-$	$+$	$-$	$+$	35
4	4	$+$	$+$	$-$	$-$	21
5	2	$-$	$-$	$+$	$+$	39
6	7	$+$	$-$	$+$	$-$	27
7	3	$-$	$+$	$+$	$-$	27
8	8	$+$	$+$	$+$	$+$	20

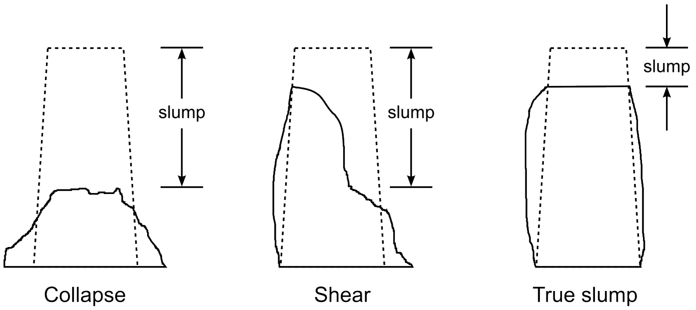
The generator for this design was factor **D = ABC**, and the resulting least squares model:
$y = 29.5 - 5.75x_A - 3.75x_B - 1.25x_C + 0.75x_D + 0.50x_Ax_B + 1.0x_Ax_C - 1.0x_Ax_D$

1. What is the interpretation of the "-3.75" coefficient for x_B in the above model? **[2]**
2. What will the two factor interaction, **AC**, be confounded with? **[2]**
3. Where would you run the next experiment to try get the stability above 50 or greater? Take a step of -1 in coded units for factor **C**. Your answer must end with a table, that gives the values of **A**, **B**, **C**, and **D** in real world units. **[8]**
4. What is the predicted stability value at the conditions specified in your previous answer? **[2]**

Question 8 [20]

A concrete slump test is used to test for the fluidity, or workability, of concrete. It's a crude, but quick test often used to measure the effect of polymer additives that are mixed with the concrete to improve workability.

The concrete mixture is prepared with a polymer additive. The mixture is placed in a mold and filled to the top. The mold is inverted and removed. The height of the mold minus the height of the remaining concrete pile is called the "slump".



Your company provides the polymer additive, and you are developing an improved polymer formulation, call it **B**, that hopefully provides the same slump values as your existing polymer, call it **A**. Formulation **B** costs less money than **A**, but you don't want to upset, or lose, customers by varying the slump value too much.

1. You have a single day to run your tests (experiments). Preparation, mixing times, measurement and clean up take 1 hour, only allowing you to run 10 experiments. Describe the precautions, and why you take these precautions, when planning and executing your experiment. Be very specific in your answer (use bullet points). **[4]**

2. The following slump values were recorded over the course of the day:

Additive	**Slump value [cm]**
A	5.2
A	3.3
B	5.8
A	4.6
B	6.3
A	5.8
A	4.1
B	6.0
B	5.5
B	4.5

What is your conclusion on the performance of the new polymer formulation (system **B**)? Your conclusion must either be "send the polymer engineers back to the lab to make a better polymer **B**" or "let's start making formulation **B** for our customers". Explain your choice clearly.

To help you, $\overline{x}_A = 4.6$ and $s_A = 0.97$. For system **B**: $\overline{x}_B = 5.62$ and $s_B = 0.69$.

Note: In your answer you must be clear on which assumptions you are using and, where necessary, why you need to make those assumptions. **[10]**

3. Describe the circumstances under which you would rather use a paired test for differences between polymer A and B. **[3]**
4. What are the advantage(s) of the paired test over the unpaired test? **[3]**

Question 9 [6]

At the end of the course we learned that data can be used for five major purposes, whether this is a for a company, or even for use in your own life (e.g. your hobbies, *etc*.)

List 3 of the purposes, and describe a corresponding tool that we learned about in the 4C3/6C3 course that helps achieve each that purpose. Be clear in your answer as to how that tool matches the purpose.

Question 10 [3]

What is the single most interesting thing you learned in 4C3/6C3?

The end.