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

Duration of exam: 3 hours | **Instructor: Kevin Dunn**
18 April 2011 | **dunnkg@mcmaster.ca**

This exam paper has 6 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 do not repeat the question 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 extra credit for answering the 600-level questions though. The distinguishing feature for 600-level students in this exam is that a higher level of technical accuracy is expected.
- **Total marks**: 100 marks for 400-level; 108 marks for 600-level students.

Question 1 [5]

Your new raw material supplier has a C_{pk} value of 1.2 for a critical quality variable, and your previous supplier's C_{pk} is 0.95. Your manager doesn't understand this terminology and wants to understand why you recommended the new supplier, even though their material is more expensive. Give a brief explanation, and an illustration (diagram) to help your manager.

Question 2 [600 level students: 8]

You are investigating four factors, **A**, **B**, **C** and **D** (all of them continuous variables). Time and budget constraints only allow you to run 9 experiments. You must run two experiments per day to finish the experiments within 5 days. Each day there is a different crew of plant operators and staff – they are strongly expected to have an effect on the results.

Write out an experimental table that blocks for the effect of the operators. Your table must show the levels of the 4 factors and have an additional column that indicates which day the experiment should be run (1, 2, 3, 4 or 5). Give bullet point notes that outline the justification for your table.

Hint: blocking can be viewed as adding additional factor(s) to a fractional factorial, with the blocking levels given by the new factor(s).

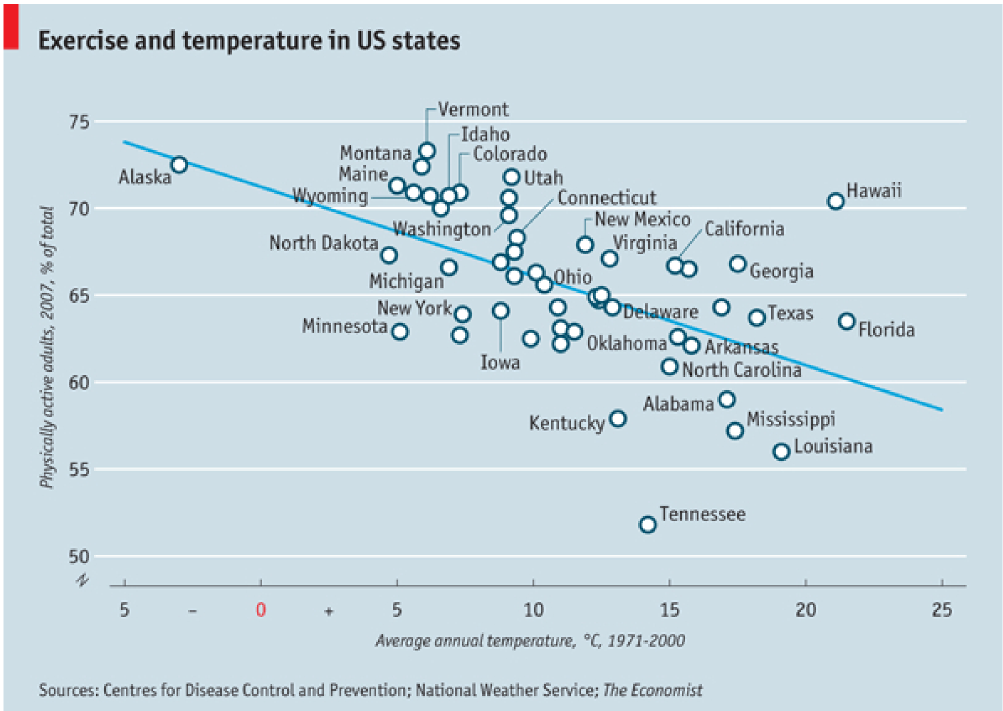
Question 3 [6 = 2 + 4]

Latent variable methods can solve some shortcomings of *classical* statistical tools. For example, in process monitoring we will increase our type II error when monitoring two correlated variables in the usual univariate way.

1. What is a type II error in the context of general process monitoring?
2. List two shortcomings of least squares models that are solved by using latent variable models, such as projection to latent structures (PLS).

Question 4 [11 = 2 + 2 + 2 + 2 + 2 + 1]

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


Exercise and temperature in US states

Sources: Centres for Disease Control and Prevention; National Weather Service; *The Economist*

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. What interesting aspect(s) do you see in this plot that you would want to investigate, and why?
4. The author has shown a linear regression line. Is the intercept term meaningful in this case; please explain.
5. Calculate an estimate of the linear model's slope, and give an interpretation for it.
6. Is the Tennessee point a high leverage, high discrepancy and/or highly influential point in the least squares model?

Question 5 [6 = 2 + 2 + 2]

Engineering systems record and archive a variety of data – e.g. plant instrumentation, sensors on bridges, cars, roads and buildings, digital cameras, NIR spectra, *etc*. What major objectives might these data be used for? Briefly describe any 3 major objectives and give a concise example of each one.

Question 6 [10 = 4 + 6]

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, not from McMaster, 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**.

That engineer mistakenly thought if he runs the experiments on alternating days that he would be able to get an unbiased result of the cheaper catalyst's effect on conversion. Given below are the conversion data for March 2011:

Date	Catalyst used	Conversion [%]	Date	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 (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. Also describe how *you* would have run the experiments if you only had those 8 values to analyze.
2. Fortunately, your knowledge from this course can be used to rescue the experiment. Give bullet point notes and an appropriate rough illustration to describe how you will analyze *all the data* available to you (much more than shown in this table). However, please use the data in the above table in your answer, showing one or two example calculations.

Question 7 [5]

In this course we learned mainly about *classical* statistical tools, such as least squares regression, process monitoring, visualizing data using scatter plots, and design of experiments. But we also discussed at the end of the course how each of these is really just a form of modelling. For example, least squares is just building an approximate model $\mathbf{y} = f(\mathbf{x}, \mathbf{b})$, where $\mathbf{x}$ and $\mathbf{y}$ are the data used to build the model and $\mathbf{b}$ are the model parameters.

Describe what the "model parameters", $\mathbf{b}$, are in a generic monitoring chart, and describe the objective of this model.

Question 8 [14 = 2 + 1 + 4 + 7]

One of the experiment projects investigated by a current 4C3/6C3 student was understanding effects related to the preparation of uncooked, breaded chicken strips.

The student investigated these 3 factors in a full factorial design *:

- **D** = duration: low level at 15 minutes; and high level = 22 minutes.
- **L** = level of oven rack: low level = use middle rack; high level = use low oven rack (this coding was used because the lower rack applies more heat to the food).
- **P** = preheated oven or not: low level = short preheat (30 seconds); high level = complete preheating.

* The student actually investigated 4 factors, but found the effect of oven temperature to be negligible!

The response variable was y = taste, the average of several tasters, with higher values being more desirable.

Experiment	**D**	**L**	**P**	**Taste**
1	$-$	$-$	$-$	3
2	$+$	$-$	$-$	9
3	$-$	$+$	$-$	3
4	$+$	$+$	$-$	7
5	$-$	$-$	$+$	3
6	$+$	$-$	$+$	10
7	$-$	$+$	$+$	4
8	$+$	$+$	$+$	7

The full factorial model that can be calculated from these 8 experiments is:

$$y = 5.75 + 2.5x_{\text{D}} - 0.5x_{\text{L}} + 0.25x_{\text{P}} - 0.75x_{\text{D}}x_{\text{L}} - 0.0x_{\text{D}}x_{\text{P}} - 0.0x_{\text{L}}x_{\text{P}} - 0.25x_{\text{D}}x_{\text{L}}x_{\text{P}}$$

1. What is the interpretation of the $+2.5x_{\text{D}}$ term in the model?
2. From the above table, at what conditions should you run the system to get the highest taste level?
3. Does your previous answer match the above model equation? Explain, in particular, how the non-zero two factor interaction term affects taste, and whether the interaction term reinforces the taste response variable, or counteracts it.
4. If you decided to investigate this system, but only had time to run 4 experiments, write out the fractional factorial table that would use factors **D** and **L** as your main effects and confound factor **P** on the **DL** interaction.

   Now add to your table the response column for taste, extracting the relevant experiments from the above table.

   Next, write out the model equation and estimate the 4 model parameters from your reduced set of experiments. Compare and comment on your model coefficients, relative to the full model equation from all 8 experiments.

Question 9 [15 = 2 + 4 + 1 + 2 + 6]

Your company is developing a microgel-hydrogel composite, used for controlled drug delivery with a magnetic field. A previous employee, J.E.J., did the experimental work but she has since left the company. You have been asked to analyze the existing experimental data.

- Response variable: y = sodium fluorescein (SF) released [mg], per gram of gel
- The data collected, in the original units:

Experiment	Order	A = microgel weight [%]	B = hydrogel weight [%]	y
1	4	4	10	1.19
2	1	8	10	0.926
3	6	4	16	1.54
4	3	8	16	0.89
5	2	6	13	0.85
6	5	6	13	0.88
7	9	3.2	13	1.25
8	7	8.8	13	1.11
9	10	6	17.2	1.36
10	8	6	8.8	0.979

1. What was likely the reason the experimenter added experiments 5 and 6?
2. Why might the experimenter have added experiments 7, 8, 9 and 10 after the first six? Provide the necessary calculations to justify your answer.
3. What is the name of the type of experimental design chosen by J.E.J. for all experiments?
4. Using these data, you wish to estimate a nonlinear approximation of the response surface. Write out the equation of such a model (*also read the next question*).
5. Write out both the $\mathbf{X}$ matrix, and the corresponding symbolic entries in $\mathbf{b}$ that you would use to solve the equation $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ to obtain the parameter estimates of the model you proposed in the previous part.

Question 10 [12 = 2 + 4 + 2 + 4]

Some data were collected from tests where the compressive strength, x, used to form concrete was measured, as well as the intrinsic permeability of the product, y. There were 16 data points collected. The mean x-value was $\overline{x} = 3.1$ and the variance of the x-values was 1.52. The average y-value was 40.9. The estimated covariance between x and y was -5.5.

The least squares estimate of the slope and intercept was: $y = 52.1 - 3.6x$.

1. What is the expected permeability when the compressive strength is at 5.8 units?
2. Calculate the 95% confidence interval for the slope if the standard error from the model was 4.5 units. Is the slope coefficient statistically significant?
3. Provide a rough estimate of the 95% prediction interval when the permeability is at 5.8 units.
4. Now provide a more accurate, calculated 95% prediction confidence interval for the permeability at 5.8 units.

Question 11 [11 = 8 + 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 **Gm** at the low level code or **Gp** at the high level.
- **A** = agitation level: low level = 10 rpm and high level = 20 rpm, but can only be set at integer values.
- **T** = growth temperature: 30°C at the low level, or 36°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 = 1400, 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 model:

$$y = 24 + 3x_{\text{G}} - 1.0x_{\text{A}} + 4.0x_{\text{T}} - 0.2x_{\text{G}}x_{\text{A}} - 0.79x_{\text{G}}x_{\text{T}} - 0.25x_{\text{A}}x_{\text{T}} + 3.5x_{\text{G}}x_{\text{A}}x_{\text{T}}$$

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, he has requested the next experiment be run at 40°C, which is also the highest level you can adjust the bioreactor to.

1. Give the experimental conditions for all 4 factors for the next experiment. Give the conditions in both the real-world units above, as well as in the usual coded units of the experiment, presented in a table.
2. Report the expected yield at your proposed experimental conditions.

Question 12 [5]

A high-volume food production facility fills bags with potato chips. The advertised bag weight is 35.0 grams. But, the current bagging system is set to fill bags with a mean weight of 37.4 grams, and this is done so that only 2.5% of bags have a weight of 35.0 grams or less.

Out of 1000 customers, how many are lucky enough to get 40.0 grams or more of potato chips in their bags?

Question 13 (optional)

Should you have time, I'm curious as to your opinion on allowing future students to use tablet computers, such as iPads or even cell phones, in engineering midterms and final exams. Especially in courses, such as this one, where open book exams are used. We expect university textbooks to be available almost exclusively in electronic form, in the very near future. And since the internet and wi-fi capability on these devices cannot not be turned off (or at least enforced), does this affect your opinion given?

The end.