Chemical Engineering: 4C3/6C3

Statistics for Engineering

McMaster University: Final examination

Duration of exam: 3 hours
23 April 2010

Instructor: Kevin Dunn
dunnkg@mcmaster.ca

This exam paper has 8 pages and 15 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 in any printed materials to the final; any textbooks, any papers, etc.
- You may use any calculator during the exam.
- You may not use a cellphone as a calculator. Nor may you use any other communication device (web, email, chat, etc) during the exam.
- You may not use a laptop or netbook.
- You may answer the questions in any order in the examination booklet.
- You may use any table of normal distributions and t-distributions in the exam; or use the copy that was available on the course website, prior the exam.
- The exam covers all material from the course.
- **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.
- **Total marks**: 100 marks for 400-level; 115 marks for 600-level students.

1.0 [5]

Your manager is asking for the average viscosity of a product that you produce in a batch process. Recorded below are the 12 most recent values, taken from consecutive batches. State any assumptions, and clearly show the calculations which are required to estimate a 95% confidence interval for the mean. Interpret that confidence interval for your manager, who is not sure what a confidence interval is.

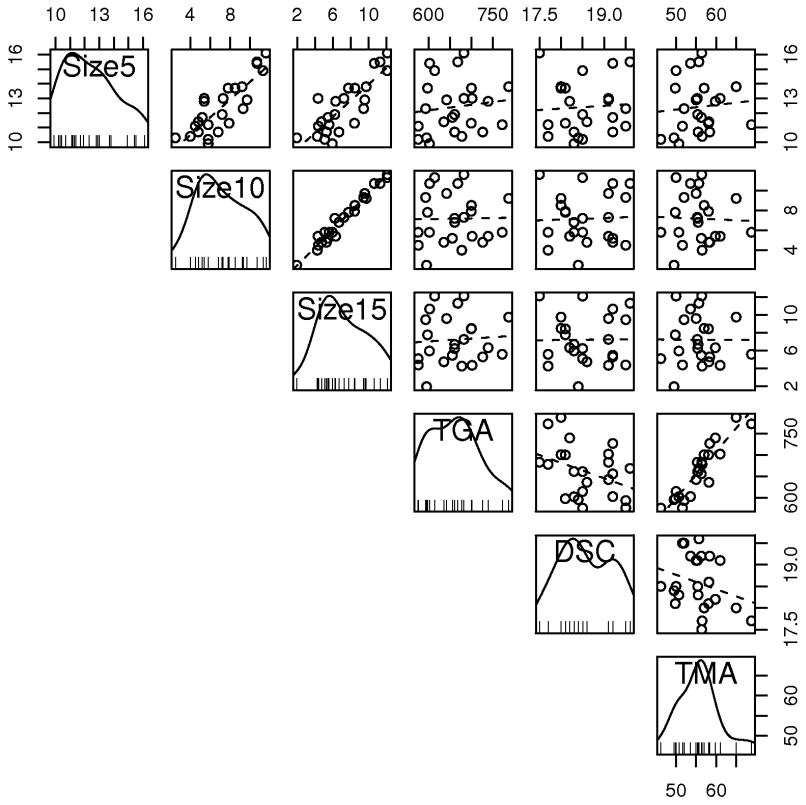
Raw data:	[13.7, 14.9, 15.7, 16.1, 14.7, 15.2, 13.9, 13.9, 15.0, 13.0, 16.7, 13.2]
Mean:	14.67
Standard deviation:	1.16

2.0 [2]

You plan to run a series of 22 experiments to measure the economic advantage, if any, of switching to a corn-based raw material, rather than using your current sugar-based material. You can only run one experiment per day, and there is a high cost to changing the feed pipe to accept the different raw material. Describe two important precautions you would implement when running these experiments, so you can be certain your results will be accurate.

3.0 [5 = 3 + 2]

The following scatterplot matrix shows a multivariate data set. These are the 6 measurements taken on 24 batches of plastic pellets.



3.1 [3]

How would you describe the distribution of each variable, as well as the relationships between the variables. Only describe the relationships that appear interesting.

3.2 [2]

Describe the limitations, if any, of this visualization technique for multivariate data sets.

4.0 [4]

If an exponentially weighted moving average (EWMA) chart can be made to approximate either a CUSUM or a Shewhart chart by adjusting the value of λ, what is an advantage of the EWMA chart over the other two? Describe a specific situation where you can benefit from this.

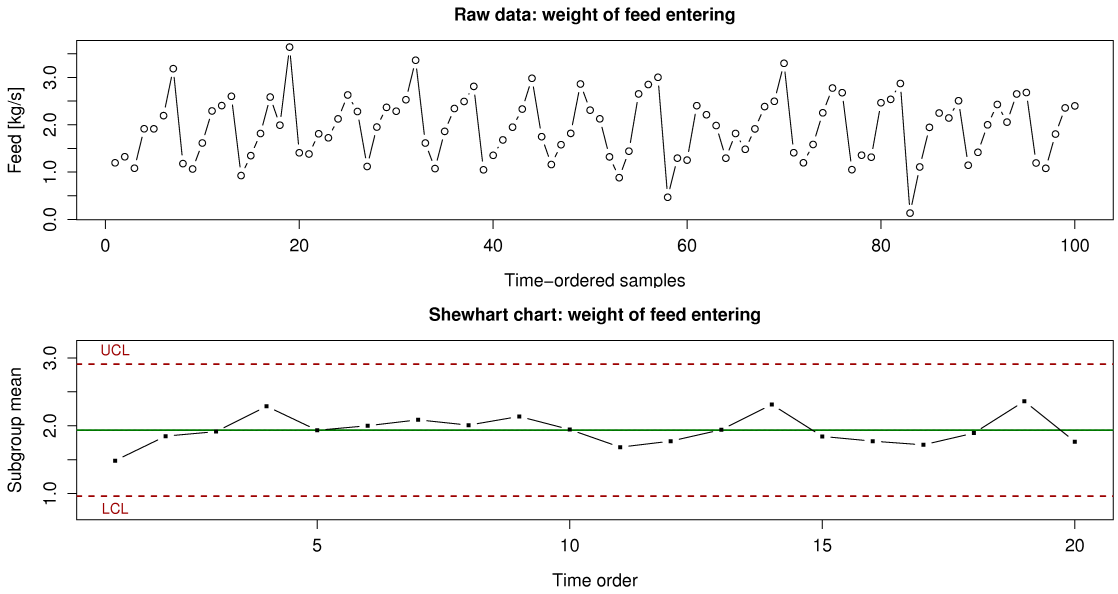
5.0 [6]

The most recent estimate of the process capability ratio for a key quality variable was 1.30, and the average quality value was 64.0. Your process operates closer to the lower specification limit of 56.0. The upper specification limit is 93.0.

What are the two parameters of the system you could adjust, and by how much, to achieve a capability ratio of 1.67, required by recent safety regulations. Assume you can adjust these parameters independently.

6.0 [4 = 2 + 2]

The following charts show the weight of feed entering your reactor. The variation in product quality leaving the reactor was unacceptably high during this period of time.



6.1 [2]

What can your group of process engineers learn about the problem, using the time-series plot (100 consecutive measurements, taken 1 minute apart).

6.2 [2]

Using concepts learned in this course, why might this sort of input to the reactor have an effect on the quality of the product leaving the reactor?

7.0 [6 = 1+2+3]

We have performed a fractionated design, 2^{5-2}. Let the five factors be **A**, **B**, **C**, **D** and **E**.

- What is the resolution of this design?
- How would you set the levels of **D** and **E**?
- What is the projectivity of the design? Explain the significance of what this projectivity means if factors **B**, **C** and **E** are shown to be unimportant.

8.0 [10]

In your start-up company you are investigating treatment options for reducing the contamination level of soil that has been soaked with hydrocarbon products. You have two different heaps of contaminated soil from two different sites. You expect your treatment method to work on any soil type though.

Your limited line of credit allows only 9 experiments, even though you have identified at least 6 factors which you expect to have an effect on the treatment. Write out the set of experiments that you believe will allow you to learn the most relevant information, given your limited budget. Explain your thinking, and present your answer with 7 columns: 6 columns showing the settings for the 6 factors and one column for the heap from which the test sample should be taken. There should be 9 rows in your table. What is the projectivity and resolution of your design?

9.0 [11 = 1 + 2 + 2 + 2 + 2 + 2]

A least squares model was built using data recorded during January to March 2010, from a system with two nearly identical reactors.

$$y = 45.1 + b_{\text{flow}}x_{\text{flow}} + b_{\text{reactor}}x_{\text{reactor}}$$

The variable y is a response variable, measured in units of cP, and is a function of the feed flow rate, x_{flow}, and the reactor, x_{reactor}. The feed flow rate varies between 0.5 and 1.3 kg/s under normal operation. The x_{reactor} variable is 0.0 for reactor TK102 and is 1.0 for reactor TK103; $b_{\text{flow}} = 7.6\dfrac{\text{cP}}{\text{kg/s}}$ and $b_{\text{reactor}} = -4.5$cP.

9.1 [1]

What is the predicted value of the response variable when the flow rate into TK103 is 0.9 kg/s?

9.2 [2]

Does this model mean that flow rate has a different effect on the response variable, depending on whether TK102 or TK103 is being used? Please explain why.

9.3 [2]

Using bullet points, what do you understand from the:

- $b_{\text{flow}} = 7.6\dfrac{\text{cP}}{\text{kg/s}}$ term in the model
- $b_{\text{reactor}} = -4.5$cP term in the model

9.4 [2]

How does your description in the previous question change, if at all, given the standard error is $S_E = 4.3$?

9.5 [2]

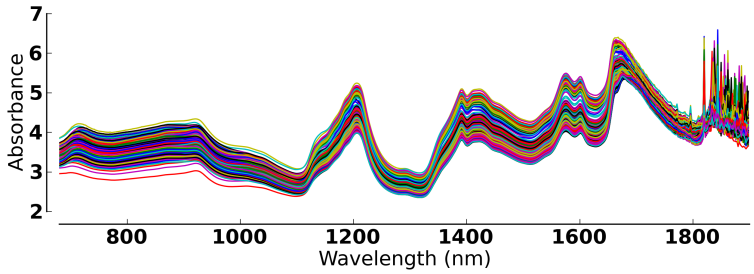
What is your interpretation of $S_E = 4.3$? The measurement error for y has standard deviation of ± 3.9 cP.

9.6 [2]

Using data from April 2010, this predictive model had an RMSEP (root mean square error of prediction) value of 4.52 cP. What is your opinion of this model's predictive ability?

10.0 [5]

The following graph shows data measured using a near infrared probe. These probes are now commonly installed on piping and reactors to quantify and monitor material flowing past. The figure shows some data from the probe; this particular probe records absorbance values at 650 wavelengths every second.



A monitoring system that uses these spectral measurements is required, but it is not possible to monitor each of the 650 wavelengths separately. Describe, for this system, the process monitoring plot that you would use in phase II to ensure the absorbance spectra are consistent with the operation that was present in phase I. How would you react to an alarm from this monitoring plot?

11.0 [24 = 2 + 2 + 4 + 3 + 1 + 6 + 6]

A factorial experiment was run to investigate the settings that minimize the production of an unwanted side product. The two factors being investigated are called **A** and **B** for simplicity, but are:

- **A** = reaction temperature: low level was 420 K, and high level was 440 K
- **B** = amount of surfactant: low level was 10 kg, high level was 12 kg

A full factorial experiment was run, randomly, on the same batch of raw materials, in the same reactor. The system was run on two different days though, and the operator on day 2 was a different person. The recorded amount, in grams, of the side product was:

Experiment	Run order	Day	A	B	Side product formed
1	2	1	420 K	10 kg	89 g
2	4	2	440 K	10 kg	268 g
3	5	2	420 K	12 kg	179 g
4	3	1	440 K	12 kg	448 g
5	1	1	430 K	11 kg	196 g
6	6	2	430 K	11 kg	215 g

11.1 [2]

What might have been the reason(s) for including experiments 5 and 6?

11.2 [2]

Was the blocking for a potential day-to-day effect implemented correctly in the design? Please show your calculations.

11.3 [4]

Write out a model that will predict the amount of side product formed. The model should use coded values of **A** and **B**. Also write out the $\mathbf{X}$ matrix and $\mathbf{y}$ vector that can be used to estimate the model coefficients using the equation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

11.4 [3]

Solve for the coefficients of your linear model, either by using $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ directly, or by some other method.

11.5 [1]

Assuming the blocking for the day-to-day effect was implemented correctly, does your model show whether this was an important effect on the response or not? Explain your answer.

11.6 [6]

You have permission to run two further experiments to find an operating point that reduces the unwanted side product. Where would you place your next two runs, and show how you select these values. Please give your answer in the original units of **A** and **B**.

11.7 [6]

As you move along the response surface, performing new experiments to approach the optimum, how do you know when you are reaching an optimum? How does your experimental strategy change? Please give specific details, and any model equations that might help illustrate your answer.

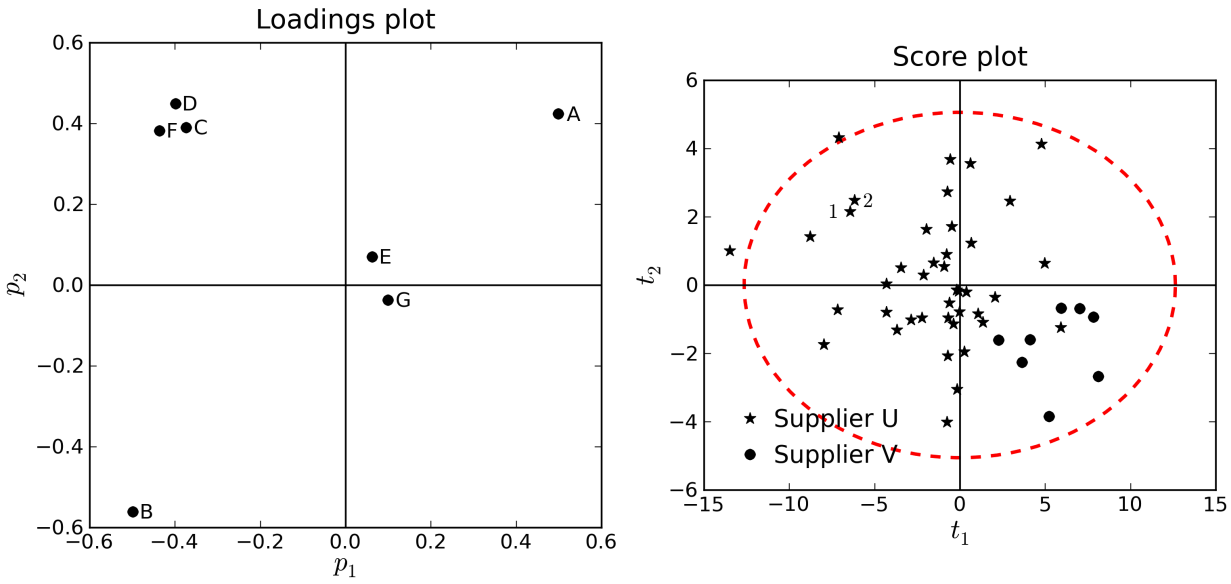
12.0 [4]

Using the article by George Box on "*Quality Improvement – An expanding domain for the application of scientific method*", describe two simple tools that can be implemented to improve quality control.

13.0 [14 = 5 + 6 + 3]

The following considers a principal component analysis (PCA) model, built from a matrix **X**, containing 7 columns of laboratory measurements that were taken on several batches of raw material. These 7 measurements, called **A**, **B**, **C** ... up to **G**, are thought to completely characterize the raw material's properties. In particular, variables **C**, **D** and **F** are related to the energy costs required to process the material.

Only two components were found to be significant, explaining all useful variation in **X**. There were no large outliers off the model plane.



13.1 [5]

Batches marked as 1 and 2 in the score plot are from the same supplier and they arrived at your site on the same day. Give an explanation, together with a supporting equation, to explain why these points are close together on the score plot.

13.2 [6]

Using the above plots, describe:

- The relationship you expect to see in the raw data between measurements **A** and **B**.
- The relationships you expect to see in the raw data between variables **C**, **D** and **F**.
- The characteristics of variables **E** and **G** as they relate to the raw material properties.

13.3 [3]

Material from supplier U costs $31/kg, while it costs $42/kg from supplier V. Explain, using any concepts learned in this course, which other costs you, as the recipient of material U, might incur. In other words, which costs do you need to take into account when making a decision to choose between supplier U or V?

14.0 [600-level question: 10 = 1 + 1 + 2 + 2 + 2 + 2]

In your company, one of the final quality parameters of a tablet (drug) is its stability. The current laboratory test requires that it be dissolved for 40 minutes in a liquid at a fixed temperature. An investigation into using **A** = a lower temperature, and **B** = a shorter dissolution time, is underway. The objective is to ensure that the new settings of **A** and **B** have at least the same level of accuracy as the current test.

All tablets used in the experiments came from a sample where the stability was known to be 65 ± 5.

Temperature	Dissolution time (minutes)		
	20	30	40
302 K	52	58	60
310 K	60	68	63

This table contains the stability values measured from the laboratory tests, which were run independently and in a random order. The baseline settings are at **A** = temperature = 310K, and **B** = time = 40 minutes.

14.1 [1]

Draw a cube plot that shows the results from the test.

14.2 [1]

Give a model for this system that has both first-order and second-order (quadratic) terms for **A** and **B**.

14.3 [2]

Write down the $\mathbf{X}$ matrix and the $\mathbf{y}$ vector that you would use to estimate your model parameters when using computer software to implement: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

14.4 [2]

Explain why you are unable to solve for $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

14.5 [2]

Which additional experiments would you perform (give their specific settings in actual units) in order to resolve the problem in the previous question.

14.6 [2]

Assume now that you have the solution vector, $\mathbf{b}$ for the full quadratic model. Describe how you would use this model to determine the effect of the temperature and dissolution time on the accuracy of the test. What results do you expect, based on a visual inspection of the raw data.

15.0 [600-level question: 5]

The following PCA model has been calculated from a data set with 6 variables.

$$\text{Loadings} = \begin{bmatrix} -0.52 & 0.24 \\ -0.52 & 0.32 \\ -0.52 & 0.13 \\ -0.28 & 0.58 \\ +0.13 & 0.37 \\ -0.25 & 0.58 \end{bmatrix}$$

$$\text{Mean of each column in the original data} = \begin{bmatrix} 12, & 7, & 29, & 660, & 19, & 55 \end{bmatrix}$$

$$\text{Standard deviation of each column in the original data} = \begin{bmatrix} 1.8, & 2.5, & 6.3, & 59.7, & 0.6, & 5.0 \end{bmatrix}$$

$$\text{Variance of } \mathbf{t_1} = 2.67$$

$$\text{Variance of } \mathbf{t_2} = 2.03$$

Calculate the following from the new observation vector given below:

- the t_1 score value,
- Hotellings T^2, assuming that t_2 = 0.35
- the squared prediction error, assuming that $\widehat{x}_{\text{new},a=2} = \begin{bmatrix} 1.85 & 1.88 & 1.81 & 1.15 & -0.31 & 1.05 \end{bmatrix}$

The raw data are written here in the same order as used in the PCA model:

$$\mathbf{x}_{\text{new, raw}} = \begin{bmatrix} 16 & 11 & 39 & 682 & 17 & 56 \end{bmatrix}$$

The end.