

Statistics for Engineers, 4C3 / 6C3

Written midterm, 12 February 2014

Kevin Dunn, kevin.dunn@mcmaster.ca

McMaster University

Note:

- You may bring in any printed materials to the midterm; any textbooks, any papers, *etc.*
- You may use any calculator during the midterm.
- **To help us with grading, please start each question on a new page, but use both sides of each page in your booklet.**
- You may answer the questions in any order on all pages of the answer booklet.
- This exam requires that you apply the material you have learned here in 4C3/6C3 to new, unfamiliar situations, which is the level of thinking we require from students that will be
- Any ambiguity or lack of clarity in a question may be resolved by making a suitable and justifiable assumption, and continuing to answer the question with that assumption(s).
- **Total marks:** 54 marks for 400-level and 68 marks for 600-level, 12% of course grade. 600-level students have extra questions to complete; 400-level students may attempt these questions for extra credit, where indicated
- Total time: 2 hours (nominally), though you have “infinite” time to complete it. There are 4 pages on the exam, please ensure your copy is complete.

Question 1 [20 = 2 + 1 + 3 + 3 + 1 + 5 + 1 + 3 + 1]

1. Name one purpose of a box plot and state how it achieves that purpose. [2]
2. Describe a case where the median is a more suitable measure of location than the mean. [1]
3. We said that a major aim of learning about statistics is to quantify variability in a data set. State a couple of ways you would go about doing that for a single column of data from a new data set? [3]
4. If breaking strength (a quality property) data from 65 plastic samples show a median value of 2.8, and a MAD of 0.45 units, estimate the probability of observing a value that is 3.6 units or higher. Be clear in all assumptions you make in arriving at your answer. [3]
5. Complete the sentence with a statement that is factually true: sparklines are _____. [1]
6. If the 95% confidence interval for the amount of impurity in a new catalyst is 0.4 to 14.2. The interval was originally based on 16 samples. [5]
 - (a) Find the 99% confidence interval now.
 - (b) Clearly explain why the interval in the prior part changed in the way it did, and why this is expected.
7. The least squares model curve (line) always passes through a particular set of x and y data points. Which ones? [1]

8. What is an outlier in the context of linear regression? Draw an illustration to substantiate your answer. [3]
9. True or False? For a 95% confidence interval, there is a 95% probability that the given interval contains the true mean. [1]

Question 2 [400-level: 8; 600-level: 13]

The pH of water from a treatment facility is measured daily. Water specimens collected over 21 days yield a sample mean value of 6.8 and a sample standard deviation of 0.9. The measured pH values are assumed to be normally distributed.

1. Calculate the 99% confidence interval for the mean pH. [4]
2. What happens to the confidence interval as the degree of confidence approaches 100%? [1]
3. The process settings require the average pH of the water to be 7.0. Does the evidence shown above support this requirement? [2]
4. You have a large budget; what happens to the confidence interval as you take more and more samples? [1]
5. Describe two ways by which the length of the confidence interval in the first part of the question could be reduced by 50%. Give quantitative answers. [600-level: 5; extra credit for 400-level]

Question 3 [400-level: 4; 600-level: 11]

A survey of alumni that graduated from a particular university's chemical engineering department in 2000 to 2005 had 124 bachelors students that participated. The survey asked for the number of months the students were unemployed during the period from 2000 to 2005.

An independent survey of engineering students graduating from Ontario universities was used as a reference. That survey had similar data available on the duration of unemployment, but was for a period from 1999 to 2005.

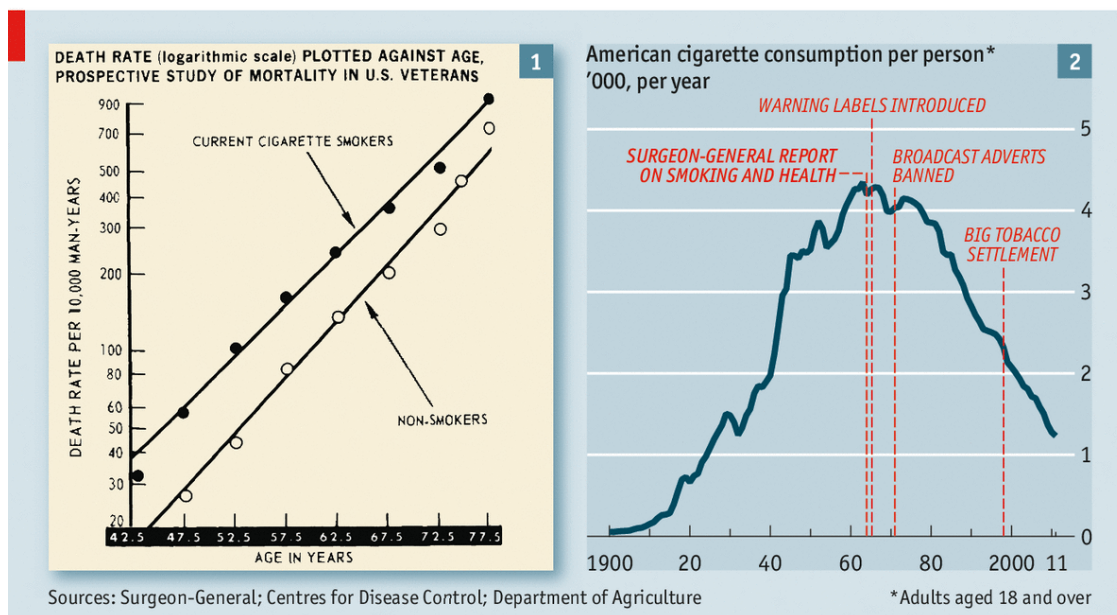
Let C refer to data from the chemical engineering students, and let A refer to data from all engineering students. The following 95% confidence interval was constructed:

$$-7.5 < \mu_C - \mu_A < 1.2 \text{ months}$$

1. Give a clear interpretation of the above confidence interval, one that is clear enough so that the chair of the chemical engineering department, who doesn't understand statistics, can understand. (*Disclaimer:* this question is obviously not referring to McMaster University). [4]
2. In which circumstances would you use a paired test for differences? [600-level only: 4]
3. If possible, briefly describe how you would set up a paired test for the above situation. If a paired test is not possible for this situation, please describe why. [600-level only: 3]

Question 4 [400-level: 7; 600-level: 9]

The plot on the left appeared in a report on 11 January 1964 by the US Surgeon-General “Smoking and Health”. Ten scientists (all men; half smokers) analyzed 7,000 studies to assess the effects of tobacco on the human body.



1. The plot on the left is a _____. [1]
2. The plot on the right is a _____. [1]
3. The plot on the left was for males. The report by the Surgeon-General concluded by saying “The data for women point [to the same conclusion]”. What conclusion(s) do the data ask you to make? [3]
4. It is hard to believe that people doubted (and some still do) a cause-effect relationship between these two variables. However describe **specifically** what it is about the left hand side plot that makes it an effective message. **600-level** students should be especially clear in their answers, illustrating their answer with quantitative values from the plot. [400-level: 2; 600-level: 4 (extra credit for 400-level)]

Question 5 [15]

A small data set is available that uses the average taste of mature cheddar cheese determined by several judges; it relates the taste to several explanatory variables, one of which is level of H_2S in the cheese. H_2S is the gas responsible for the bad smell of sewers and swamps. Higher taste values indicate a better tasting cheese. The aim of the model is to understand the nature of the relationship between the variables, and potentially build a predictive model for taste.

There are 30 data points in the original dataset, and 10 pairs of data are randomly selected and shown below, so you can get a feel for the raw data.

Taste	12.3	20.9	25.9	37.3	5.5	16.8	38.9	54.9	57.2	6.4	...
H₂S	3.14	5.04	7.60	8.72	4.79	3.66	9.06	6.75	7.91	4.70	...

Other information is that the average Taste was 24.5 units, the variance of Taste was 264, the average H₂S content was 5.94 units, and the variance of H₂S was 4.52.

The following output is from a particular software package, but many packages, such as R, Excel, SAS, JMP, Minitab, and others will produce a similar table. One of the goals of this course is that you are comfortable interpreting the statistical output from any software.

Residuals:

Min	1Q	Median	3Q	Max
-15.427	-7.611	-3.493	6.421	25.686

Coefficients:

	Estimate	Std. Error
(Intercept)	-9.7884	5.958
H ₂ S	5.7764	0.946

Residual standard error: _____ on _____ degrees of freedom

Multiple R-squared: _____, Adjusted R-squared: 0.5559

1. What is the intercept in the least squares model? Give its value as well as interpretation for it. [2]
2. Give a clear interpretation for the slope coefficient of 5.8 in this model. [2]
3. An excerpt from the Analysis of Variance table is provided below

Analysis of Variance			
Source	DF	Sum of Squares	Mean Square
Model	_____	4377	_____
Error	_____	3286	_____
Total	_____	7663	
Root MSE	_____		
R-Square	_____		

Calculate the Root MSE value, or in other words, what we have called standard error, S_E , in this course. [2]

4. What is the R^2 value that would have been reported in the above output? [2]
5. What is the prediction of taste at an H₂S concentration of 5 units? Contrast it to the sample of raw data provided. [3]
6. What is the 95% confidence interval for the slope coefficient, *and interpret* this confidence interval in the context of how you plan to use this model. [4]

The end.