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 graduating and working very soon.
- 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]**

Solution

1. Box plots show the variability in a single variable, by plotting a 5 number summary: the median, the first and third quartiles which give an indication of the spread, and the whiskers which give am indication of the extreme values.

2. Any occasion is suitable, but especially when outliers are present in the data, e.g. data taken from an automatic sensor, which might be prone to occasional failure, such as a thermocouple.

3. Quantifying variability is by indicating how spread out the data are. The following are various ways of doing so:

   - use a box plot, which shows the interquartile range and whiskers
   - calculate the interquartile range numerically
   - show the standard deviation (or variance)
   - the median absolute deviation (MAD)
   - show a histogram of the data
   - show a q-q plot (not the most effective way to do this though, as this is not the intention of the plot, but we can see variability by examining the one axis)

4. If we assume the data are normally distributed, then we may use z-values. Further assume that the median as a good estimate of the average, and that the MAD is a good estimate of the standard deviation (*check your software to confirm whether you need to multiply your MAD by a constant to make it match the standard deviation*):

   Then the z-value is:

   $$z = \frac{3.6 - 2.8}{0.45} = 1.77$$

   and from tables of the normal distribution, the probability of observing a z-value great or equal to this is $1 - 0.96 = 0.04$ (values are approximate). So there is a 4% probability.

5. Sparklines are a compact and efficient (from a data-ink perspective) form of representing time-series data. From the course notes: sparklines are small graphics that carry a high density of information.

6. Assume the samples were originally normally distributed and independent of each other:

$$\overline{x} = \frac{14.2 + 0.4}{2} = 7.3$$
$$c_t = 2.13 \quad \text{with 15 degrees of freedom at the 95\% level}$$
$$2c_t \frac{s}{\sqrt{n}} = 14.2 - 0.4 = 13.8$$
$$s = \frac{(13.8)(4)}{(2)(2.13)} = 12.96$$

   (a) Given these values, we can find the 99% confidence interval where $c_t = 2.95$ with 15 degrees of freedom, and 0.5% in each tail.

   Lower bound = $\overline{x} - c_t \frac{s}{\sqrt{n}} = 7.3 - \frac{(2.95)(12.96)}{4} = -2.26$

   Upper bound = $\overline{x} + c_t \frac{s}{\sqrt{n}} = 7.3 + \frac{(2.95)(12.96)}{4} = 16.9$

   So the 99% CI is $-2.26 < \mu < 16.9$

   (b) The interval became wider, and this is expected because a high level of confidence implies the interval has a greater probability of containing the fixed original mean value. In other words, we are more confident this new interval contains the true mean.

7. It passes through $(\overline{x}; \overline{y})$

8. True: the probability is associated with the interval (the true mean is fixed; but our interval changes every time we repeat the experiment). It is **incorrect** to say "there is a 95% chance that the true mean lies within the given interval": that implies the mean can vary for the given interval.

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]**

Solution

1. The usually assumptions apply: assume the water specimens are independent of each other, and their numeric values are normally distributed (checked via a q-q plot).

   The critical t-value from tables, with 0.5% in each tail is $c_t = 2.85$.

   Lower bound = $\overline{x} - c_t \dfrac{s}{\sqrt{n}} = 6.8 - \dfrac{(2.85)(0.9)}{\sqrt{21}} = 6.24$

   Upper bound = $\overline{x} + c_t \dfrac{s}{\sqrt{n}} = 6.8 + \dfrac{(2.85)(0.9)}{\sqrt{21}} = 7.36$

   So the 99% CI is $6.24 < \mu < 7.36$

2. The interval approaches bounds that are infinite, i.e. at 100% confidence we have $-\infty < \mu < \infty$.

3. Yes, since the value of 7.0 is included in the bound, an average pH of 7 units is possible (the average recorded **from this set of data** was 6.8, but a future sample would record a different value).

4. At a given level of confidence, the bounds will become narrower (tighter).

5. One can increase the number of samples, reduce the level of confidence, or more likely in practice, use a combination of the two.

   The number of samples would need to be quadrupled to reduce the bound by half (approximately). Strictly speaking, as n increases, the c_t value also changes, but this change is rather small and can readily be ignored. For example $c_t = 2.85$ for 20 degrees of freedom at the 95% level, while it becomes $c_t = 2.64$ for 80 degrees of freedom. So degrees of freedom would have to be slightly more than quadrupled to reduce the CI by 50%.

   Using a lower level of confidence, one that halves the c_t value is required: $0.5 \times 2.85 = 1.425$, which is approximately found when there is 8% in each tail. So in this case, a $100 - 2 \times 8 = 84\%$ confidence interval be roughly half the width.

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

$$-7.5 < \mu_{\mathtt{C}} - \mu_{\mathtt{A}} < 1.2 \text{ months}$$

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]**

Solution

1. *Explanation*: since the bounds span zero, it indicates Chem Eng students statistically remain unemployed as long as any other engineers, however the slight asymmetry in the bounds indicates that, at least with this sample, they remain unemployed shorter than other engineers.

   Note: the bounds have a 95% chance of containing the true difference $\mu_{\mathrm{C}} - \mu_{\mathrm{A}}$.
2. Paired tests are used when there is a common element between the two sample groups (i.e. the samples being treated or compared), but that commonality is not of interest to be tested, however it might affect the test's outcome. By pairing, we cancel out the commonality, and eliminate the potential for bias.
3. A paired test is not possible in this case. The two groups being compared are "chemical engineering students" and "engineering students". Since the former is by definition a member of the latter, the experiment cannot be run in a paired manner.

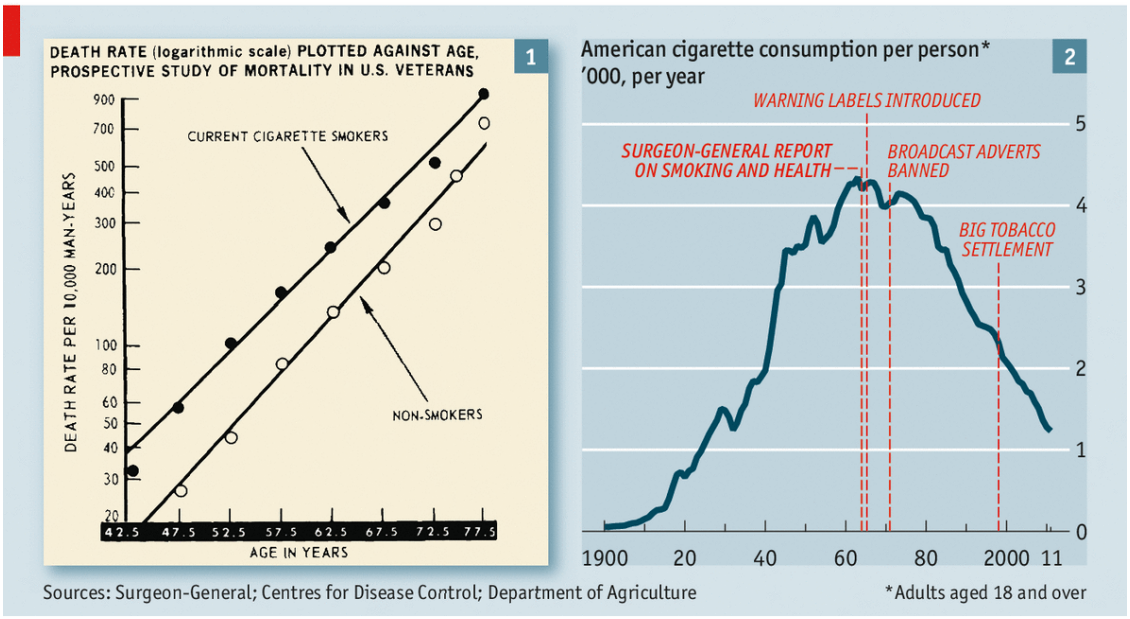
Question 4 [400-level: 7; 600-level: 9]

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)]**

Solution

1. Scatter plot
2. Time-series plot



3. The conclusions is that smoking leads to a higher death rate when compared to a non-smoker, at all ages. The fact that the two lines approach each other indicates the *proportional difference* decreases at higher ages. The fact that the plot applies to females and males, indicates that smoking has the same effect, no matter what the subject's sex.

   The fact that death rate is higher at higher ages is not a surprising conclusion: it is a natural consequence of aging (NR=not relevant).

4. The human eye and brain are quick to decode a visual plot. This question is forcing you to think what specifically it is about the plot that makes it so effective (many of you rushed your answer without explaining why and how the message is effectively conveyed):

   - The use of **two data sets** with open and closed circles, but plotted on the same axes, emphasizes the difference between smokers and non-smokers.
   - The plot shows 3 variables on a 2D plot: death rate, age, and the effect of smoking as compared to the baseline of non-smoking.
   - Normally cause-and-effect is shown on the two axes of the scatter plot. But, in this case neither the horizontal nor vertical axes are particularly insightful (they simply show higher death rates at a higher age). The insight and effectiveness comes from the *difference* between the two lines.
   - This effectiveness is due to the horizontal (or vertical) offset between the two curves, which emphasizes the cause-and effect relationship between smoking and earlier probability of death. As described below, the offset is essentially a controlled experiment where only a single variable is changed: smoking vs non-smoking.

     Bear in mind the plot is a summary of over 7,000 prior studies, which would imply that people from all sorts of educational backgrounds, ethnicities, socioeconomic standings, *etc* would have participated. Therefore the only difference between these two diagonal lines is smoking.

The horizontal offset indicates the reduced life-expectancy: a 55-year old non-smoker has the same life expectancy of a 47.5 year old smoker (a difference of 7 to 8 years).

The vertical offset indicates the increased probability of death within a cohort of the **same age** (i.e. keeping age constant), e.g. a group of people aged 67.5 years will have a death rate of 200 per 10,000 man-years lived for non-smokers, while smokers have an elevated death rate of 350 per 10,000 man-years lived. This interpretation can be seen as a visual paired test: the age is in common, the only difference is smoking vs non-smoking. There are 7 pairs of data at the same age level.

The death rate approximately **doubles** for smokers vs non-smokers, an extremely significant and very large factor: to double the probability of death.

- I would argue that the use of log-axes *diminishes* the severity of the message, however, I must admit that log axes make the plot look more aesthetically appealing, by being linear. The lack of outliers also helps to emphasize the message.

Note: the above answer is far more comprehensive than expected for full grade.

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	...
H_2S	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
H2S           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]**

Solution

1. The intercept is -9.8 and indicates the taste level the cheese would have if there were no H_2S content. It is not meaningful, as there didn't appear to be any cheeses without H_2S. Furthermore, a negative taste does not seem possible in the raw data.
2. The slope coefficient is 5.8 (having units of taste divided by units of H_2S content), and implies the taste improves by 5.8 units *on average*, for every one unit increase in the level of H_2S.
3. The standard error is $S_E = \sqrt{\dfrac{\text{RSS}}{n-k}} = \sqrt{\dfrac{3286}{30-2}} = 10.8$ units of taste.
4. The $R^2 = \dfrac{\text{RegSS}}{\text{TSS}} = \dfrac{4377}{7663} = 0.571$, or 57%.
5. $\hat{y} = -9.76 + 5.78(5.0) = 19.1$ units of taste, would be expected.

   There are raw data points around 5.0 units of H_2S, for example, at 5.04, which has a taste of 20.9 units. This prediction is comparable to that. Consider also that the standard error

is $S_E = 10.8$ units, so this is good, especially when contrasted against the range of the raw data for `Taste`.

6. The true slope coefficient, β_1, has a confidence interval:

$$-c_t < \frac{b_1 - \beta_1}{S_E(b_1)} < c_t$$

where $c_t = 2.05$ using 28 degrees of freedom at the 95% confidence level, since there are 30 data points and 2 parameter estimates. The standard error for b_1 is 0.946, given in the software output. Using this, the confidence interval can be found as:

$$\begin{aligned} 5.78 - 2.05 \times 0.946 &< \beta_1 < 5.78 + 2.05 \times 0.946 \\ 3.84 &< \beta_1 < 7.72 \end{aligned}$$

This indicates bounds within which we expect to find the true slope coefficient, β_1, at the 95% confidence level. From this we conclude the effect of H2S on taste is that it leads to a statistically significant increase.

The end.