Statistics for Engineers, 4C3 / 6C3

Written midterm, 09 February 2015

Kevin Dunn, kevin.dunn@mcmaster.ca **McMaster University**

Note:

- No papers, other than this test and the answer booklet are allowed with you in the midterm. You will be provided with the class-sourced cheat sheet (attached on the last 2 pages).
- You may only use the standard McMaster calculator in 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 suitable and justifiable assumption(s), and continuing to answer the question with that assumption(s)**.
- **Total marks**: 63 marks for 400-level and 74 marks for 600-level. 600-level students have extra questions to complete; 400-level students may attempt certain questions for extra credit, but only where indicated.
- Total time: 2 hours.
- There are 7 pages on the exam, please ensure your copy is complete.

Question 1 [11 (or 15 for 600-level students)]

1. If a random variable has a uniform distribution with mean of 50 and lower bound of 35, and upper bound of 65, what is the probability that an observation has:
   (a) a value less than 45? **[1.5]**
   (b) greater than 120? **[1.5]**
2. For a 95% confidence interval of the mean of n values: (*select the single correct option*) **[1]**
   (a) there is a 95% probability that the given confidence interval contains the true mean
   (b) the n values can all come from different distributions
   (c) the true mean is inside the confidence interval with a 95% probability.
   (d) if the n values are independent, and come from any distribution, then critical values from the t-distribution are used to calculate the confidence interval.
3. What fractional area lies within ± 3 standard deviations of the normal distribution? **[2]**
4. You wish to explain the principle of a box plot and robust statistics to a fellow engineer that has not taken statistics before. So you tell them to imagine a series of numbers from 1 to 1000. Then you proceed to explain the idea of percentiles and quartiles to them. Using that vector of data: **[5 = 1 + 1 + 1 + 1 + 1]**
   (a) what is the approximate first quartile value?
   (b) what is the median?
   (c) what is the approximate IQR?
   (d) what will be the lower whisker (fin) value on the box plot?
   (e) what will the MAD value be that you calculate in R?

5. **600-level students only**: continuing with the prior question: the vector of numbers from 1 to 1000 is randomly contaminated, where 40% of the values are replaced by zero (e.g. for example, someone might have made a copy/paste error in a spreadsheet and overwrote part of the original data).
   (a) Explain, if at all possible, whether you can calculate the median, and give an estimate of what that numeric value will be. **[2]**
   (b) Explain, if at all possible, whether you can calculate the MAD, and give an estimate of what that numeric value will be. **[2]**

Solution

1. (a) The range from 35 to 45, over a total range of 35 to 65 is 33.3%. This is from the uniform distribution, where every value is equally probable.
   (b) Zero probability: uniform distribution ends at 65 in this case.
2. Correct option is (a). Option (c) is wrong.
3. $99.87 - 0.135 = 99.7\%$
4. (a) 250
   (b) 500
   (c) 500
   (d) $250 - 1.5 \times 500 = -500$, but because this is outside the range of the data, it will simple report the smallest value, which is 1.
   (e) MAD = $1.4826\,\text{median}\left\{||x_i - \text{median}\left\{x_i\right\}||\right\} = 1.4826\,\text{median}\left\{||x_i - 500||\right\} = 1.4826 \times 250 = 370.65$
5. Two lines of thinking are possible here. Either interpretation is acceptable in your answers.
   (a) Replace the values with zero, and the zeros are not ignored (they are included in the data). The data will be 40% zeros, followed by 60% non-zeros. The median will be biased downward to some lower value: we cannot tell what exactly.
   (b) Same as part (a).

   Or alternatively:
   (a) Replaced by zeros, but these are excluded/ignored; so the median will still be around 500.
   (b) The MAD will be approximately 370, as shown in part 4.

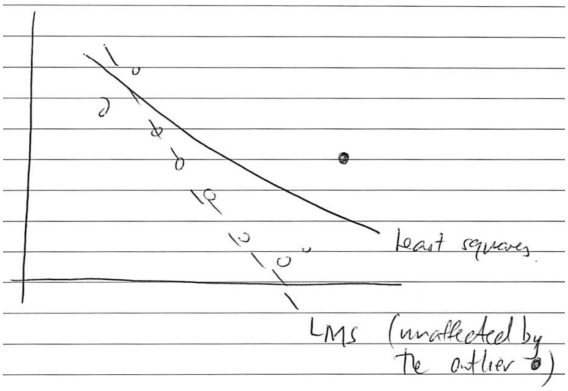
   The key insight is that the statistics are unaffected in this case.

Question 2 [600-level students only, other's may attempt it for extra credit: 4]

Redraw the following plot in your answer booklet. On your plot superimpose two lines:

1. A solid single line that shows where the least squares regression line would be. **[1]**
2. A dashed line that indicates where the line that satisfies the least median of squares (LMS) model would be? The LMS objective function is: median $\left\{e_i^2\right\}$. **[1]**
3. Please explain your reasoning and the differences between these two lines. **[2]**

Solution



3. The least squares objective aims to reduce all squared errors, e_i^2, so the line is biased towards these large squared outliers.

   The LMS objective is concerned with minimizing the median squared residual, so it is unaffected by a single (or even several) outliers. It will pass through the majority of the data.

   Your answer must explain both options, and the difference between them must be clear.

Question 3 [10]

A sophisticated chemical company is considering a new reactor modification. Their current plug-flow reactor (a long pipe where the chemical reactions take place inside) is being improved by adding a smaller tubular membrane inside the existing reactor pipe. This membrane has been designed by the engineering team to improve the recovery of the valuable product.

http://www.offshore-mag.com

Test results were collected and the following were recorded:

```
> before.modification <- c(44, 41, 47, 50, 47, 47, 42, 47)
> after.modification  <- c(47, 53, 48, 43, 46, 49, 42)
> mean(before.modification)
[1] 45.625
> mean(after.modification)
[1] 46.85714
> sd(before.modification)
[1] 3.020761
> sd(after.modification)
[1] 3.716117
```

A confidence interval was then calculated using computer software [try it yourself after the midterm: `t.test(before, after, var.equal=TRUE)` in R]:

$$-4.99 \leq \boxed{} \leq 2.52$$

1. If `B` represents "before" and `A` represents "after", is this a confidence interval for $\mu_B - \mu_A$ or for $\mu_A - \mu_B$? **[1]**
2. What is the statistical interpretation of this confidence interval? **[2]**
3. What practical advice do you give your colleagues in the engineering design team, based on this confidence interval result?

   Provide them some guidance based on these numbers, but at the end of your answer your must either recommend the modification, or not. **[4]**
4. A pooled variance was required to calculate this confidence interval; what is that pooled variance value? **[2]**
5. How many degrees of freedom were used on the table of the t-distribution? **[1]**

Solution

1. Since average of the "before" samples is less than the "after" samples, this must be the case for $\mu_B - \mu_A$.
2. Statistically, there was not an improvement in using the tubular membrane.
3. The interval shows a slight negative bias, indicating a small improvement was made. But there is not evidence **in these data** , that the improvement is permanent. It could be pure luck these values were obtained.

   If they are convinced an improvement exists, they need to collect more raw data (tests) on the before and after systems. Also check the data for outliers (though none seem to be present as written here).

   So the conclusion: do not use the modified method.
4. $s_p^2 = 11.3$
5. 13

Question 4 [13 (or 16 for 600-level students)]

1. If overlaying the normal distribution with mean of zero, and standard deviation of 1.0 on top of the t-distribution, please describe how the two distributions are different. How would you be able to tell the two apart? **[2]**
2. Why do we use the t-distribution when calculating the confidence interval for the mean of a data set? **[2]**
3. The following blood sugar values were collected from 17 patients at 07:00. This was before these patients were asked to participate in a drug trial to see if our company's new insulin preparation would assist them with controlling their daily blood sugar levels.

   $x = (6.1, 4.1, 5.4, 5.5, 4.2, 3.1, 5.9, 5.1, 3.5, 4.3, 4.6, 4.9, 1.4, 5.0, 5.2, 3.8, 5.1)$ [mmol.L^{-1}]

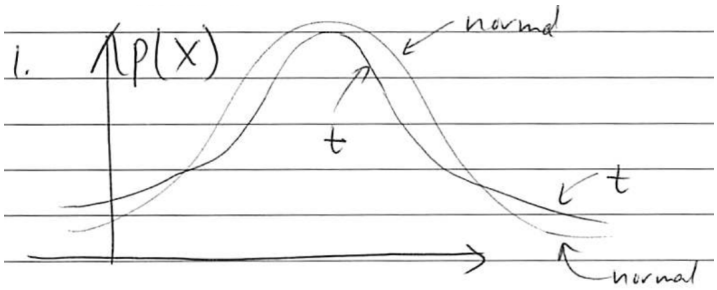
   $\overline{x} = 4.54$ and the standard deviation is $s = 1.15$ [mmol.L^{-1}]

   Use these data to calculate a 95% confidence interval for the blood sugar level of the patients. **[5]**
4. In your calculation above, please explain the assumptions you have had to make. In addition, explain whether those assumptions are accurate for the situation, and/or how you would verify them. **[4]**
5. **600-level students only**: The regulatory requirement for the report asks for you to supply 95% confidence intervals with a width of 1.0 mmol.L^{-1}, or less.

   What is the fewest number of patients you will you have to sample to obtain an interval that meets their requirements? **[3]**

Solution

1. The t-distribution has broader tails and lower probabilities, $p(x)$ at the center than the normal distribution.



As the degrees of freedom get higher and higher, the t-distribution looks much like the normal distribution.

2. When we do not know the population standard deviation, i.e. when we only know the sample standard deviation.
3. At the 95% confidence level there is 2.5% in each tail, so $c_t \approx 2.12$ from the table. Using this leads to $4.54 \pm 2.12\left(\frac{1.15}{\sqrt{17}}\right)$; or alternatively written: $3.95 \leq \mu \leq 5.13$.
4. Assume the data are independent. This is met because each patient is independent from the other (as long as the blood sugar level analysis was performed in a reproducible way).

   Assume the data are from the normal distribution. This is easy to check with a q-q plot.
5. The confidence interval is to be within a range of ± 1 mmol/L.

$$\text{UB} - \text{LB} = 2c_t\frac{s}{\sqrt{n}}$$

Assuming the $s = 1.15$ will remain constant, even with more samples, then let's try reduce the interval (it current has a width of $5.13 - 3.95 = 1.18$), by taking more than 17 samples.

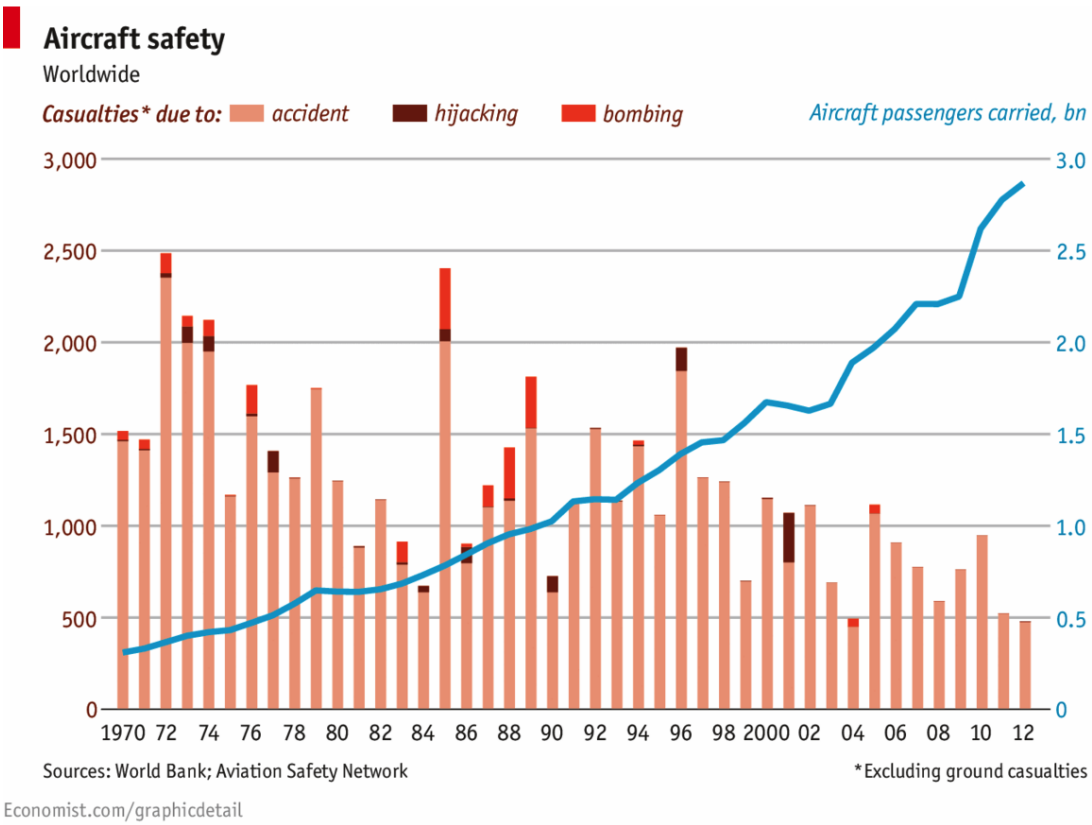
Try 21, so dof=20, then $2c_t\frac{s}{\sqrt{n}} = (2)(2.09)\frac{1.15}{\sqrt{21}} = 1.049$

Try 25, so dof=24, then $2c_t\frac{s}{\sqrt{n}} = (2)(2.065)\frac{1.15}{\sqrt{25}} = 0.95$

So somewhere between 21 and 25 samples would be required.

Question 5 [7]

The following data visualization is from The Economist and the 3 arrows added manually.



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 plots being shown, and critique their effectiveness. Use bullet points in your answer.

Solution

- Accidents are by far the greatest source of aircraft-related deaths.
- This is shown in the stacked bar plot.
- These types of deaths show a drop-off over time.
- High-jackings and bombings are almost non-existent now.
- Bar plot can be improved by a better colour choice (doesn't shown up well in grey).
- The time-series plot (superimposed) shows increasing passengers over time.
- That emphasizes the message even more: casualties decreased even while air traffic increases (so percentage-wise this is even better).
- Might be initially unclear that the time-series y-axis is on the right, but not hard to figure that out.
- The plot has good time-based context: all the way back to 1970.

Question 6 [9 = 3 + 3 + 3]

You want to speed up your testing on a new polymer formulation that biodegrades faster than your competitor's product. We learned that statistical tests should keep things as constant as possible, except for the thing being varied.

But you do not have time to wait! You need to get your new product on the market as soon as possible. Your company has 3 reactors: J, K and L. You realize that you just need to use only 2 of these reactors to complete your test work in time.

So you look back at historic data and discover data when the same raw material was fed to 2 of the reactors, and the output from the reactor was measured. You calculate these 95% confidence intervals from the historic data (over 100 different unique batches of material was used in your calculation):

- $-31.4 \leq \mu_K - \mu_L \leq 134$
- $-21.4 \leq \mu_K - \mu_J \leq 41.2$
- $-81.8 \leq \mu_L - \mu_J \leq 77.6$

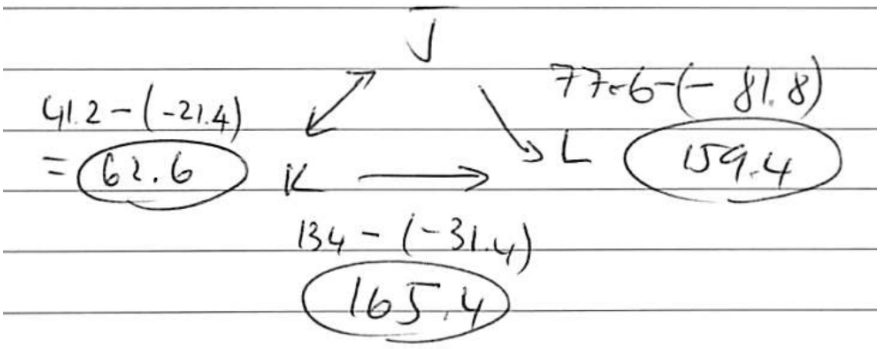
1. Which two reactors would you pick to do your future tests in, so that you use two reactors that are the most consistent and similar in operation? Explain your reasoning **[3]**
2. But then your colleague from Guelph tells you that you should have done a paired test instead. Should you have? Is your colleague wrong? Explain whether a paired test would be more appropriate in this case. **[3]**
3. Just to verify things though, you go to your raw data and, fortunately, in this situation, the way the experimental work was run allows the data to be analyzed as paired results (because detailed raw records were kept).

   You calculate these confidence intervals for the paired difference:

   - $9.81 \leq \mu_{L-K} \leq 88.4$
   - $48.3 \leq \mu_{J-K} \leq 68.7$
   - $-46.1 \leq \mu_{J-L} \leq 33.5$

   Which two reactors should you use based on these results? **[3]**

Solution



1. If aiming for consistency you would like an interval that is as small as possible. If aiming for similarity you would like an interval to span zero, and symmetrically so.

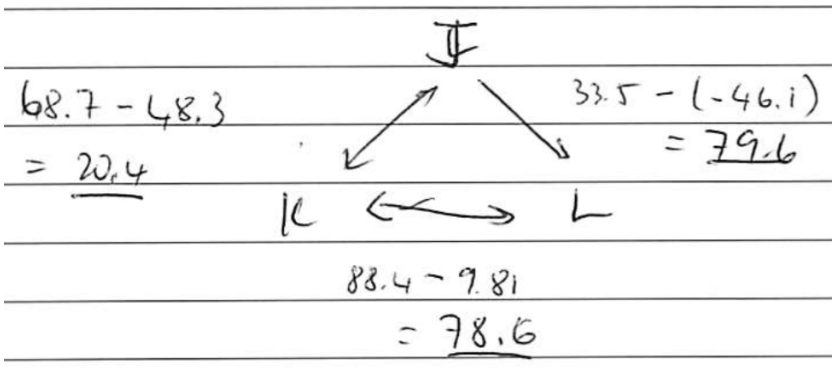
   This rules out K and L.

   You can argue for (K vs J) or for (J vs L) on either grounds. Statistics seldom makes a decision for you: it provides you the basis to help numerically guide your decision.

   Personally, I would use K and J: I would value consistency (narrower bounds, less variance) over similarity. I would rather have consistency: a constant bias can be added or subtracted.

2. Your colleague is wrong. There is no consistent parameter in common to remove. You are running the experiments in two reactors in the future (equivalent of cooking in the kitchen with two mixing bowls).

   Pairing is done when there is a common element (potential bias) that is not being tested, but that should be cancelled out.

3. 

   J and L would be preferred if from a direct reading of the paired test result, since it spans zero symmetrically (i.e. no difference between the reactors).

   J and K are the most consistent, but don't span zero. One can argue that it is worthwhile, because even if different, it is consistently difference, at a value of about 58 (the midpoint), which could be added or subtracted to align the results from the two reactors.

Question 7 [13 = 2 + 3 + 2 + 2 + 2 + 2]

A data set containing temperature data as the input variable and melt index data (a valuable quality property) as the output data has been collected. The company has used this to predict melt index in real-time, on-line, so they do not have to take lab samples.

Instead of presenting the raw data here, the following summary data is presented:

- 17 data pairs collected
- $\sum_i (x_i - \overline{x})^2 = 26,090$
- $\sum_i (y_i - \overline{y})^2 = 2,502,528$
- $\sum_i (x_i - \overline{x})(y_i - \overline{y}) = 159,242$
- $\overline{x} = 450$
- $\overline{y} = 1340$

1. If you had to plot these data, would you observe mostly a positive, or negative, or mostly no correlation? Explain your answer. **[2]**
2. What would be a suitable estimate of the population value for the intercept, b_0, in the linear model $y = \beta_0 + \beta_1 x + \epsilon$? **[3]**
3. The R^2 was calculated from the model to be $R^2 = 0.68$. What would be the residual sum of squares values, RSS, in the ANOVA table? **[2]**
4. Provide an estimate for the standard error. **[2]**
5. What is the interpretation of the standard error in this example, and explain how you would have verified your explanation if you had software available to you. **[2]**
6. What is the best prediction of melt index you can currently provide for your colleague when the temperature is 500 K? **[2]**

Solution

1. A positive correlation, since $r = \dfrac{\text{cov}(x,y)}{\sqrt{\mathcal{V}(x)\mathcal{V}(y)}}$, and the numerator is $159242/n > 0$, so the correlation is positive.

2. $b_1 = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \dfrac{159242}{26909} = 6.10$, which also confirms the positive correlation.

   $b_0 = \overline{y} - b_1 \overline{x} = 1340 - (6.10)(450) = -1407$

3. $R^2 = \dfrac{\text{RegSS}}{\text{TSS}} = 0.68$; TSS $= 2502528$, so RegSS $= 1701719$, then RSS $= 800809$.

4. $S_E = \sqrt{\dfrac{\text{RSS}}{n-k}} = \sqrt{\dfrac{800809}{17-2}} = 231.1$.

5. The S_E is interpreted as the standard deviation of the residuals and gives a measure of spread of the residuals, provided they are normally distributed. We can easily verify that with a q-q plot.

6. At $T = 500\,\text{K}$, we have $\hat{y} = (6.10)(500) - 1406.6 = 1643.4$.

The end.