Statistics for Engineering, 4C3/6C3

Written midterm, 16 February 2011

Kevin Dunn, dunnkg@mcmaster.ca **McMaster University**

Note:

- 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 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.
- **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**: 75 marks for 400-level; 85 marks for 600-level students.

---

1 [5 = 3 + 2]

Sulphur dioxide is a byproduct from ore smelting, coal-fired power stations, and other sources.

These 11 samples of sulphur dioxide, SO_2, measured in parts per billion [ppb], were taken from our plant. Environmental regulations require us to report the 90% confidence interval for the mean SO_2 value.

$$180,\ 340,\ 220,\ 410,\ 101,\ 89,\ 210,\ 99,\ 128,\ 113,\ 111$$

1. What is the confidence interval that must be reported, given that the sample average of these 11 points is 181.9 ppb and the sample standard deviation is 106.8 ppb?
2. Why might Environment Canada require you to report the confidence interval instead of the mean?

Solution

1. From the central limit theorem, assuming the 11 values are independent, the mean SO_2 value, $\overline{x} \sim \mathcal{N}\left\{\mu, \sigma^2/n\right\}$, where μ and σ are the distribution from which the raw values come.

   Using an estimate for $\sigma = \hat{s} = 106.8$ we can construct the z-value and confidence interval. z will be t-distributed with $n - 1 = 10$ degrees of freedom, so $c_t = 1.81$ At the 90% confidence level we can then write:

$$\begin{aligned}
-c_t &\leq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leq +c_t \\
\overline{x} - c_t\frac{s}{\sqrt{n}} &\leq \mu \leq \overline{x} + c_t\frac{s}{\sqrt{n}} \\
181.9 - 1.81 \times \frac{106.8}{\sqrt{11}} &\leq \mu \leq 181.9 + 1.81 \times \frac{106.8}{\sqrt{11}} \\
123.6 \text{ ppb} &\leq \mu \leq 240.2 \text{ ppb}
\end{aligned}$$

2. Environment Canada may require the confidence interval since in addition to providing an estimate of the mean (just the midpoint of the CI), it also provides an *estimate of the spread* – variability in your process – if n is known, without requiring access to the raw data.

A wide CI gives an indication that you might in fact be polluting too much on some days, and compensating on others, which is not desirable. The confidence interval's width can also be compared between plants to find the most variable polluters.

2 [6 = 2 + 2 + 2]

Questions related to process monitoring:

1. In which situation would you use a CUSUM chart? Why, in your given situation, would this chart be more advantageous than say a Shewhart chart?
2. How would you select the value of λ used in the EWMA chart?
3. Explain what is meant by common cause variation in process monitoring.

Solution

1.
   - A CUSUM chart is useful in situations where the target must be precisely controlled and there is no room for drift up or down.
   - Shewhart charts are slow to react to small drifts.
   - An example would be drug dosing by an intravenous catheter: too much or too little drug can have negative side effects. The chemical engineering translation of that example is: when too much or too little reactant is added to the reactor it can have negative effects.
2. The general rule is that an EWMA chart behaves:
   - more like a Shewhart chart as $\lambda \to 1$
   - more like a CUSUM chart as $\lambda \to 0$

   Using that, we can select λ based on our desired operation for the chart. In particular, I would always use a testing data set to verify whether known problems are detected, and then adjust λ by trial and error.

   A more sophisticated approach would select λ such that it minimizes the sum of squares of one-step-ahead prediction errors.

   Finally, the rule of thumb for most systems is $\lambda = 0.2 \pm 0.1$ (from the paper by Hunter, and mentioned in class).
3. Common cause variation is the variation present in your process when the process is stable and there are no **special causes** in the data. This is in most cases the product you sell your customers, or send to a downstream operation. Under stable operation you will never get a "0" for the variance.

   One way to describe this is to say that common cause variation is the variation remaining within your lower and upper *control limits* after you have finished phase I of designing your control chart.

3 [7]

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 safety regulations. Assume you can adjust these parameters independently.

Solution

The process capability ratio for an uncentered process, PCR_k, is given by:

$$\text{PCR}_\text{k} = \min\left(\frac{\text{Upper specification limit} - \bar{\bar{x}}}{3\sigma}; \frac{\bar{\bar{x}} - \text{Lower specification limit}}{3\sigma}\right)$$

We know that we must use an uncentered PCR because we operate closer to the lower bound.

The two adjustable parameters are $\bar{\bar{x}}$, the process target (operating point) and σ, the process variance. You **cannot** adjust the USL and LSL: these are fixed by customer demands or based on internal specifications.

The current process standard deviation is:

$$1.30 = \frac{64.0 - 56.0}{3\sigma}$$
$$\sigma = \frac{64.0 - 56.0}{3 \times 1.30} = 2.05$$

- Adjusting the *operating point* (we would expect to move the operating point away from the LSL):

$$1.67 = \frac{\bar{\bar{x}} - 56.0}{3 \times 2.05}$$
$$\bar{\bar{x}} = 56.0 + 1.67 \times 3 \times 2.05 = 66.3$$

  So the operating point increases from 64.0 to 66.3 to obtain a higher capability ratio.

- Adjusting the *process variance* (we would expect to have to decrease the process variance, keeping the operating point fixed):

$$1.67 = \frac{64.0 - 56.0}{3 \times \sigma}$$
$$\sigma = \frac{64.0 - 56.0}{3 \times 1.67} = 1.60$$

  Decrease the process standard deviation from 2.05 to 1.60.

4 [25 = 8 + 12 + 3 + 2]

A concrete slump test is used to test for the fluidity, or workability, of concrete. It's a crude, but quick test often used to measure the effect of polymer additives that are mixed with the concrete to improve workability.

The concrete mixture is prepared with a polymer additive. The mixture is placed in a mold and filled to the top. The mold is inverted and removed. The height of the mold minus the height of the remaining concrete pile is called the "slump".

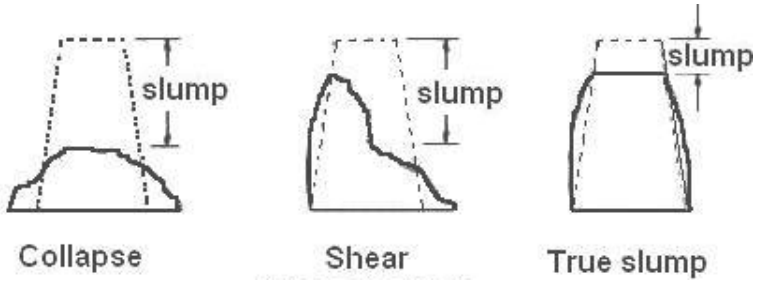


Illustration from Wikipedia

Your company provides the polymer additive, and you are developing an improved polymer formulation, call it B, that hopefully provides the same slump values as your existing polymer, call it A. Formulation B costs less money than A, but you don't want to upset, or loose, customers by varying the slump value too much.

1. You have a single day to run your tests (experiments). Preparation, mixing times, measurement and clean up take 1 hour, only allowing you to run 10 experiments. Describe all precautions, and why you take these precautions, when planning and executing your experiment. Be very specific in your answer (use bullet points).

2. The following slump values were recorded over the course of the day:

Additive	Slump value [cm]
A	5.2
A	3.3
B	5.8
A	4.6
B	6.3
A	5.8
A	4.1
B	6.0
B	5.5
B	4.5

   What is your conclusion on the performance of the new polymer formulation (system B)? Your conclusion must either be "send the polymer engineers back to the lab" or "let's start making formulation B for our customers". Explain your choice clearly.

   To help you, $\overline{x}_A = 4.6$ and $s_A = 0.97$. For system B: $\overline{x}_B = 5.62$ and $s_B = 0.69$.

   Note: In your answer you must be clear on which assumptions you are using and, where necessary, why you need to make those assumptions.

3. Describe the circumstances under which you would rather use a paired test for differences between polymer A and B.

4. What are the advantage(s) of the paired test over the unpaired test?

This question is continued for 600-level students at the end of the exam.

Solution

1. The basic rule is to control what you can and randomize against what you cannot. You should have mentioned some of these items:

- Control: clean equipment thoroughly between runs.
- Control: other factors that might affect the slump: temperature, humidity.
- Control: ensure the same person prepares all mixtures, or randomize the allocation of people if you have to use more than 1 person. Don't let person 1 prepare all the A mixtures and person 2 the B mixtures.
- Control: mixing times and how the mixture is created could have an effect. This should ideally be done by the same person.
- Randomize the order of all the A and B experiments: don't run all the A's, then all the B's, as that will confound with other factors. For example, even though temperature might vary during the day, if we randomize the run order, then we prevent temperature from affecting the results.
- Use raw materials (cement, binder, other ingredients) from all possible suppliers. And the supplier raw materials should be representative.

2. We will initially assume that $\mu_A = \mu_B$, in other words, the outcome is "let's start making formulation B for our customers". We will construct a confidence interval for the difference, $\mu_B - \mu_A$ and interpret that CI.

   - Assume the slump values within each group are independent, which will be true if we take the precautions above. We do this because then we can use the central limit theorem (CLT) to state $\overline{x}_A \sim \mathcal{N}\left(\mu_A, \sigma_A^2/n_A\right)$ and that $\overline{x}_B \sim \mathcal{N}\left(\mu_B, \sigma_B^2/n_B\right)$.
   - Note: we don't require the samples within each group to be normally distributed.
   - Assume the variances are the same: $\sigma_A^2 = \sigma_B^2 = \sigma^2$: this is required to simplify the next step.
   - Assume the $\overline{x}_A$ and $\overline{x}_B$ means are independent. This allows us to calculate a variance value, $\mathcal{V}\left\{\overline{x}_B - \overline{x}_A\right\}$ from which we can create a z-value for $\mu_B - \mu_A$:

   $$z = \frac{(\overline{x}_B - \overline{x}_A) - (\mu_B - \mu_A)}{\sqrt{\mathcal{V}\left\{\overline{x}_B - \overline{x}_A\right\}}}$$

   That denominator variance can be written as:

   $$\begin{aligned}\mathcal{V}\left\{\overline{x}_B - \overline{x}_A\right\} &= \mathcal{V}\left\{\overline{x}_B\right\} + \mathcal{V}\left\{\overline{x}_A\right\} \\ &= \sigma^2\left(\frac{1}{n_B} + \frac{1}{n_A}\right)\end{aligned}$$

   using our previous assumption that the variances are equal. We can verify this with an F-test, but won't do it here.

   Because we do not have an external estimate of the variance, σ^2, available, we must assume a good estimate for it can be found by pooling the estimated variances of the group A and B samples (which requires our equal variance assumption from earlier).

   $$\begin{aligned}s_P^2 &= \frac{4s_A^2 + 4s_B^2}{4+4} \\ s_P^2 &= \frac{4(0.97)^2 + 4(0.69)^2}{4+4} = 0.709\end{aligned}$$

   This pooling also gives us 8 degrees of freedom for the t-distribution, which is how the z-value is distributed.

Using that z-value and filling our assumed difference of zero for the true means, we can construct a 95% confidence interval:

$$\begin{array}{rcccl} -c_t & \leq & z & \leq & +c_t \\ (\overline{x}_B - \overline{x}_A) - c_t\sqrt{s_P^2\left(\frac{1}{n_B}+\frac{1}{n_A}\right)} & \leq & \mu_B - \mu_A & \leq & (\overline{x}_B - \overline{x}_A) + c_t\sqrt{s_P^2\left(\frac{1}{n_B}+\frac{1}{n_A}\right)} \\ 1.02 - 2.3\sqrt{0.709\left(\frac{1}{5}+\frac{1}{5}\right)} & \leq & \mu_B - \mu_A & \leq & 1.02 + 2.3\sqrt{0.709\left(\frac{1}{5}+\frac{1}{5}\right)} \\ -0.21 & \leq & \mu_B - \mu_A & \leq & 2.2 \end{array}$$

The statistical conclusion is that there is **no difference between formulation A and B**, since the CI spans zero. However, the practical interpretation is that the CI only just contains zero, and this should cause us to stop, and really consider the risk of the statistical conclusion.

If one of the data points were in error just slightly, or if we ran a single additional experiment, it is quite possible the CI will *not span zero* anymore. In my mind, this risk is too great, and we risk upsetting the customers.

So my conclusion would be to "send the polymer engineers back to the lab" and have them improve their formulation until that CI spans zero more symmetrically.

3. A paired test should be used when there is something is common *within* pairs of samples in group A and B, but that commonality does not extend between the pairs. Some examples though you could have mentioned:

   Pairing is appropriate: person 1 mixes polymer for test A and B; person 2 mixes polymer for test A and B (but with different time and agitation level that person 2); person 3 mixes ... *etc* Pairing *not* appropriate: person 1 mixes all the polymer A samples; person 2 mixes all the polymer B samples (pairing won't fix this, and even the unpaired results will be inaccurate - see precautions mentioned above). Pairing appropriate: you only have enough cement and raw materials to create the concrete mixture for 2 samples: one for A and one for B. You repeat this 5 times, each time using a different supplier's raw materials.

   In other words, pairing is appropriate when there is something the prevents the $\overline{x}_A$ and $\overline{x}_B$ quantities from being independent.

4. The one advantage of the paired test is that it will cancel out any effect that is common between the pairs (whether that effect actually affects the slump value or not). Pairing is a way to guard against *potential effect*.

   This makes the test more sensitive to the difference actually being tested for (formulation A vs B) and prevents confounding from the effect we are not testing for (suppliers' raw material).

   Unpaired tests, but with randomization will only prevent us from being misled, however that supplier effect is still present in the 10 experimental values. The 5 difference values used in the paired tests will be free from that effect.

5 [23 = 2 + 3 + 2 + 3 + 8 + 5]

A simple linear model relating reactor temperature to polymer viscosity is desirable, because measuring viscosity online, in real time is far too costly, and inaccurate. Temperature, on the other hand, is quick and inexpensive. This is the concept of *soft sensors*, also known as *inferential sensors*.

Data were collected from a rented online viscosity unit and a least squares model build:

$$\hat{v} = 1977 - 3.75T$$

where the viscosity, v, is measured in Pa.s (Pascal seconds) and the temperature is in Kelvin. A reasonably linear trend was observed over the 86 data points collected. Temperature values were taken over the range of normal operation: 430 to 480 K and the raw temperature data had a sample standard deviation of 8.2 K.

The output from a certain commercial software package was:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Model	2	9532.7	4766.35
Error	84	9963.7	118.6
Total	86	19496.4	
Root MSE	XXXXX		
R-Square	XXXXX		

1. Which is the causal direction: does a change in viscosity cause a change in temperature, or does a change in temperature cause a change in viscosity?
2. Calculate the `Root MSE`, what we have called standard error, S_E in this course.
3. What is the R^2 value that would have been reported in the above output?
4. What is the interpretation of the slope coefficient, -3.75, and what are its units?
5. What is the viscosity prediction at 430K? And at 480K?
6. In the future you plan to use this model to adjust temperature, in order to meet a certain viscosity target. To do that you must be sure the change in temperature will lead to the desired change in viscosity.

   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.
7. The standard error features prominently in all derivations related to least squares. Provide an interpretation of it and be specific in any assumption(s) you require to make this interpretation.

Solution

1. The causal direction is that a change in temperature causes a change in viscosity.
2. The `Root MSE` $= S_E = \sqrt{\dfrac{\sum e_i^2}{n-k}} = \sqrt{\dfrac{9963.7}{84}} = \mathbf{10.9}$ Pa.s.
3. $R^2 = \dfrac{\text{RegSS}}{\text{TSS}} = \dfrac{9532.7}{19496.4} = \mathbf{0.49}$
4. The slope coefficient is $-3.75 \frac{\text{Pa.s}}{K}$ and implies that the viscosity is expected to decrease by 3.75 Pa.s for every one degree increase in temperature.
5. The viscosity prediction at 430K is $1977 - 3.75 \times 430 = \mathbf{364.5}$ Pa.s and is $\mathbf{177}$ Pa.s at 480 K.

6. The confidence interval is

$$b_1 \pm c_t S_E(b_1)$$
$$-3.75 \pm 1.98 \frac{S_E^2}{\sum_j (x_j - \overline{x})^2}$$
$$-3.75 \pm 1.98 \frac{10.9}{697}$$
$$-3.75 \pm 0.031$$

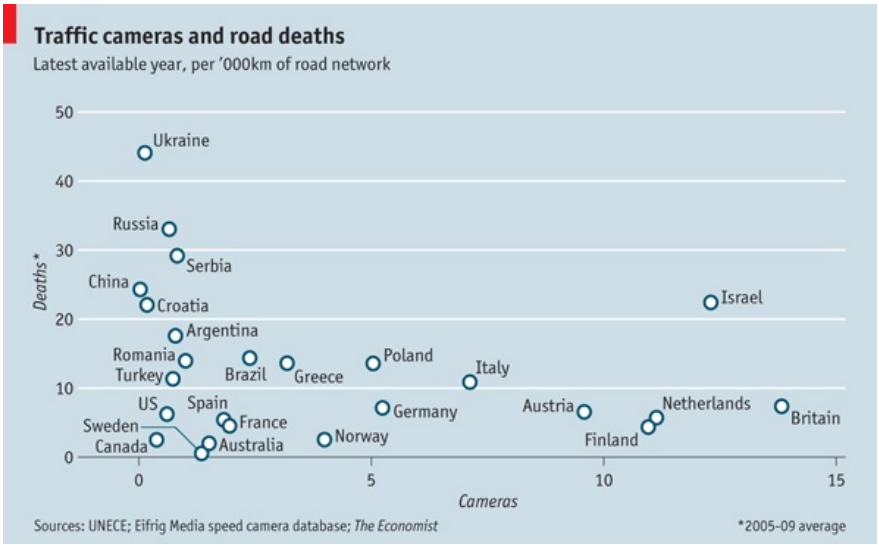
where $\frac{(x_j - \overline{x})^2}{n-1} = 8.2$, so one can solve for $(x_j - \overline{x})^2$ (though any reasonable value/attempt to get this value should be acceptable) and $c_t = 1.98$, using $n - k$ degrees of freedom at 95% confidence.

Interpretation: this interval is extremely narrow, i.e. our slope estimate is precise. We can be sure that any change made to the temperature in our system will have the desired effect on viscosity in the feedback control system.

7. The standard error, $S_E = 10.9$ Pa.s is interpreted as the amount of spread in the residuals. In addition, if we assume the residuals to be normally distributed (easily confirmed with a q-q plot) and independent. If that is true, then S_E is the one-sigma standard deviation for the residuals and we can say 95% of the residuals are expected within a range of $\pm 2S_E$.

6 [9 = 1 + 1 + 1 + 2 + 2 + 2]

Traffic cameras have their proponents (it improves road safety) and opponents (it's just a money grab). The plot below shows the number of cameras per 1000km of roadway and number of traffic deaths per 1000km of roadway. It is from the The Economist website.



1. What type of plot is this?
2. If you had to describe this relationship to a colleague, what would you say?
3. Identify and describe anything interesting in this plot that would lead you to search for more information.
4. What is the causal direction (line of reasoning) that the plot's author is wanting you to follow?

5. Which region of the plot would a linear regression model do an adequate job of describing? Feel free to answer with an illustration of your own.

6. An alternative model is possible to describe this relationship. Describe that model, perhaps providing an illustration of it, and be specific on how would use that model on new data points.

You will find all the user comments and criticism for this article quite informative (http://www.economist.com/node/21015161).

Solution

1. A scatter plot.

2. There is a negative (cor)relation between number of cameras installed and road deaths, when accounting for the distance of paved roadway (1000's of km). This is particularly true across "developed" countries in Europe.

   Note: it is a fact that there is negative correlation; the cause behind the correlation is obviously disputable.

3. I have summarized several interesting points noted by you:

   - Why is Israel an outlier (shorter road network? more drivers per km? poor driver training?)
   - Canada's location on the chart is interesting, especially when contrasted to "similar" countries such as Britain and Finland. Is it perhaps that Canada has a longer length of the road network per person, a factor that has not been accounted for in the plot. But then how we account for Sweden, and its relationship to Finland?
   - Why do countries such as Croatia, Russia, Serbia and Ukraine have such a high death rate on their roads?
   - How would the plot change if traffic *density* was taken into account, not just road network length.
   - There are diminishing returns in road safety from about 10 cameras/1000 km.
   - Would road conditions, weather, road design, presence of police officers and their effectiveness (accepting bribes?) change the plot?
   - Are the reduced deaths in European and North American countries perhaps due to safer cars, better roads, police presence, stricter alcohol laws and driver education, such as the graded G-licensing system? Cameras might have no effect at all.
   - At a low number of cameras the variation in deaths across the different countries is too wide. Surely cameras are not the only factor.

   As many realized, there are obviously many other factors at work. Someone suggested doing an experiment: that is exactly the only way we can be certain about the effect of cameras. The cost of such an experiment is prohibitive, let alone a political and logistical ordeal.

4. The author's clear intention with this plot is to *initially* ask you to believe that a higher number of traffic cameras is causally related to a lower number of road deaths. However, the author has also provided enough data points here so that you can start to question this relationship, as noted in the previous part to this question.

5. A linear regression might do a reasonable job of describing the lower part of the plot: road deaths of 20/1000km or lower, if we remove the Israel outlier. However this will be a weak relationship. Notice that these data points are mostly from European countries.

   Another region that would have a steeper negative slope would be the left side: traffic cameras of 4/1000km or lower, over the entire range of deaths, again a weak relationship, dominated mostly by Ukraine and Russia.

6. Some alternatives are:

   - Use a non-parametric smoother, such as the loess smoother. Once build, we use it by knowing the number of cameras installed for a new country (on the x-axis), and then read across to the y-axis and predict the expected number of road deaths per 1000km.
   - A nonlinear function could be fit to the data: such as a hyperbolic function. This would be used in the same way as the loess smoother.
   - Another option is a sequence of boxplots in categories: 0 to 3 cameras (showing very wide variation on the y-axis), 3 to 5, 5 to 7, *etc*. Again this be used for a new data point by knowing the number of cameras and then predicting the median value from the boxplot. We automatically get an indication of the variability in our estimate.
   - Others suggested a table; which is great for presenting the given data, but not usable to look up a new country that we didn't have in the existing dataset.

Questions for 600-level students only

7 **[6 = 3 + 3]** for 600-level students (400 level students may attempt this question for extra credit)

This question is a continuation of question 4. Please refer back to that question for context.

1. Clearly explain which assumptions are used for paired tests, and why they are likely to be true in this case?
2. The slump tests were actually performed in a paired manner, where pairing was performed based on the cement supplier. Five different cement suppliers were used:

Supplier	Slump value [cm] from A	Slump value [cm] from B
1	5.2	5.8
2	3.3	4.5
3	4.6	6.0
4	5.8	5.5
5	4.1	6.2

   Use these data, and provide, if necessary, an updated recommendation to your manager.

Solution

1. Pairing requires/assumes that the paired objects have something in common (e.g. a common bias due to the cement raw material). This common bias will be cancelled out once we calculate the difference in measurements.

   - The difference values calculated, w_i, are assumed to be independent. This is likely true in this case because each raw material supplier is different (unrelated) to the other.

- If the differences are independent, then the central limit theorem can be safely assumed so that the average of these differences, $\overline{w} \sim \mathcal{N}\left(\mu_w, \sigma_w^2/n\right)$.

2. The 5 difference values are $w_i = [0.6,\ 1.2,\ 1.4,\ -0.3,\ 2.1]$ and the average difference value is $\overline{w} = 1$ and its estimated variance is $s_w^2 = 0.815$.

   Create the z-value against the t-distribution with 4 degrees of freedom ($c_t = 2.78$), at the 95% confidence level, and unpack it into a confidence interval.

$$\begin{aligned} -c_t &\leq z \leq +c_t \\ \overline{w} - c_t\sqrt{\tfrac{s^2}{n}} &\leq \mu_w \leq \overline{w} + c_t\sqrt{\tfrac{s^2}{n}} \\ 1 - 2.78\sqrt{\tfrac{0.815}{4}} &\leq \mu_w \leq 1 + 2.78\sqrt{\tfrac{0.815}{4}} \\ -0.12 &\leq \mu_w \leq 2.12 \end{aligned}$$

   The interpretation is that the true difference in slump, μ_w, when accounting for variation from the cement raw material, is again not statistically significant, at the 95% confidence level.

   Practically though, there is a bit of a risk, due to the imbalance (asymmetry) in the confidence interval. It would be reluctant to hinge my company's profitability on this result, especially with the fact that there are only 4 experiments. So my personal conclusion would be to still "send the polymer engineers back to the lab".

8 **[4 = 2 + 2]** for 600-level students (400 level students may attempt this question for extra credit)

1. Describe what is meant by the *breakdown point* of a statistic, such as the standard deviation, or its robust counterpart, the median absolute deviation.

2. What is an advantage of using robust methods over their "classical" counterparts?

Solution

1. The breakdown point, as described in "Tutorial to Robust Statistics" by PJ Rousseeuw is: "the smallest fraction of the observations that have to be replaced to make the estimator unbounded. In this definition one can choose which observations are replaced, as well as the magnitude of the outliers, in the least favourable way."

   The mean and the standard deviation have a breakdown point of $1/n$ meaning that only a single data point (outlier) can make them unbounded.

2. Several advantages:

   - Robust methods are insensitive to outliers, which is useful when we need a measure of location or spread that is calculated in an automated way. It is increasingly prevalent to skip out the "human" step that might have detected the outlier, but our datasets are getting so large that we can't possibly visualize or look for outliers manually anymore.

   - As described in the above paper by Rousseeuw, robust methods also emphasize outliers. Their "lack of sensitivity to outliers" can also be considered an advantage.

END