Statistics for Engineering, 4C3/6C3

Written midterm, 16 February 2012

Kevin Dunn, dunnkg@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.
- You may answer the questions in any order in the answer booklet.
- You may use any table of normal distributions and t-distributions in the midterm; or use the copy that is available in the course notes.
- **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.
- **600-level students** will be held to a higher level of technical accuracy than 400-level students.
- **Total marks**: 70 marks for 400-level students; 75 marks for 600-level students.
- Total time for all levels: 2.5 hours

Question 1 [11 = 2 + 2 + 2 + 2 + 3]

A food production facility fills bags with potato chips with an advertised bag weight of 35.0 grams.

1. The government's *Weights and Measures Act* requires that at most 2.5% of customers may receive a bag containing less than the advertised weight. At what setting should you put the target fill weight to meet this requirement exactly? The check-weigher on the bagging system shows the long-term standard deviation for weight is about 1.5 grams.
2. Out of 100 customers, how many are lucky enough to get 40.0 grams or more of potato chips in their bags?
3. What is the current Cpk of this process?
4. What is your assessment regarding this process's capability? Is it capable?
5. If you wanted to change the Cpk to a value of 1.3, what could you change, and to what new value would you change it?

Solution

1. Target weight calculation:

$$z = \frac{35 - \text{target}}{\sigma} = -1.96$$
$$\text{target} = 35.0 + (1.96)(1.5) = 37.94 \text{ grams}$$

2. Probability of 40.0 grams of more is given by the area above the corresponding z-value:

$$z > \frac{40 - 37.94}{1.50}$$
$$z > 1.37$$

The exact answer is `(1 - pnorm(1.37))*100 = 8.53`, though using crude tables you could use the value corresponding to $z = 1.37$, which is about 92.0%, corresponding to the area below that z-value. So the area above it is 8%, corresponding to 8 people out of 100. Either 8 or 9 people is an acceptable answer, depending on your rounding error in reading the tables.

3. Recall the Cpk is defined relative to the closest specification limit. So in this case it must be due to the lower limit. Cpk = $\frac{\bar{\bar{x}} - \text{LSL}}{3\sigma} = \frac{37.94 - 35.0}{3 \times 1.5} = 0.65$
4. This is a very poor Cpk. Most capable processes have a Cpk of around 1.3 or higher.
5. To obtain Cpk = 1.3 we solve the above equation for $\bar{\bar{x}} = 1.3 \times 3 \times 1.5 + 35.0 = 40.85$ grams for a new target weight, keeping the variance constant.

   Or, we could keep the target weight constant and decrease the process variance (e.g. using better equipment with tighter control). The new $\sigma = \frac{37.94 - 35.0}{3 \times 1.3} = 0.753$ grams, i.e. approximately halve the current standard deviation to double the Cpk value.

   Or, we could perform some combination of both strategies: simultaneously decrease the fill-weight variance and lower the target fill-weight.

   Only one option, either adjusting the mean, or the standard deviation needed to be calculated for full grade.

Question 2 [5 = 2 + 3] (600-level students only)

1. Explain why robust methods are desirable in automatic data analysis systems.
2. What is meant by the break-down point of a robust statistic? Give an example to explain your answer.

Solution

1. Automatic data analysis systems build models from data, or analyze data, without human intervention. Any outliers in the data used can distort the results from the analysis. Robust methods would be insensitive to this distortion. For example using the median instead of the mean would be insensitive, as long as 50% or less of the data are contaminated.
2. The break-down point is the quantity of data in the sample vector that must be discrepant (i.e. outliers) before the robust estimate is affected. More formally, Rousseeuw defines the breakdown point as: "the smallest fraction of the observations that have to be replaced to make the estimator unbounded", where "one can choose which observations are replaced, as well as the magnitude of the outliers, in the least favourable way".

Question 3 [16 = 8 + 8]

A motor company is testing their new automated parallel parking assistant. The aim is to use the automatic system to reduce the time taken to safely parallel park in different lengths of parking spaces (short parking space, or a longer parking space), with different models of their cars.

The table below lists the cars and conditions under which the tests were performed. The experiments were performed in totally randomized order to those listed below. The manual parking time is the median time of 12 representative drivers, of different levels of skill.

Car	Parking length	Automatic time [s]	Manual parking time [s]	Manual − Automatic time [s]
A	Short	35	44	9
B	Short	47	49	2
C	Short	19	39	20
D	Short	22	41	19
E	Short	33	44	11
A	Long	24	35	11
B	Long	35	39	4
C	Long	18	26	8
D	Long	22	35	13
E	Long	28	39	11

The following summary calculations have been made for your reference, though you might not need all this information:

- All automatic parkings: mean = 28 s, standard deviation = 9.1 s
- All manual parkings: mean = 39 s, standard deviation = 6.3 s
- All short space parkings: mean = 37 s, standard deviation = 10 s
- All long space parkings: mean = 30 s, standard deviation = 7.5 s
- Differences between manual and automatic times: mean = 11 s, standard deviation = 5.7 s

If required, you may assume that variances can be pooled. All calculations should be at the 95% confidence level. Always state the degrees of freedom you use when reading from statistical tables.

1. Is parking in a short parking space significantly extended in duration than parking in a long parking space? Show all calculations and list all assumptions. When giving any assumption, either explain why it is reasonable, or explain how you might test whether it is reasonable.
2. Does parking with the automatic system reduce the time when compared to regular, manual drivers? You must clearly justify your choice of statistical test and list all assumptions in your answer.

Solution

Most people lost grades for not giving a full set of assumptions required to use the statistical test, *and* then justifying the assumption as being either reasonable, or explaining how they would verify that assumption as being reasonable. Calculations count for a portion of your grade in this course, but I'm more concerned that (a) you explain why you picked a certain calculation type, and (b) explain your results from the equations.

1. Combining all data from both manual and automatic parkings, we can test if parking in a short space is significantly extended in duration than parking in a long parking space. We will use a test of

differences between two groups with $n_{\text{short}} = 10$ and $n_{\text{long}} = 10$, so we get the most degrees of freedom possible.

Assume (4 points):

- Short parking space values are from a normal distribution, long parking space values from a normal distribution (tested via a q-q plot)
- $\overline{x}_S$ and $\overline{x}_L$ are independent: reasonable especially for automatic portion (computer based parking); there might be some relationship within the manual portion of short and long parkings, because a driver might "learn" how to better park the car the next time (e.g. park car A in short parking, then again in long parking). However if experiments are randomized, then this learning effect disappears.
- If the previous assumption is true, then can consider pooling the variances if we additionally assume equal variances in both short and long distributions (this can be tested by an F-test, not required, as per the question). Pooled variance is = $\frac{9}{18} \cdot (10^2) + \frac{9}{18} \cdot (7.5^2) = 8.83^2 = 78.1\text{seconds}^2$, from the 18 degrees of freedom $(10 - 1 + 10 - 1)$.
- Assume experiments within each group are independent: reasonable because the for the automatic parkings, the vehicles are different. For the manual parkings, the drivers might get better each time, but that's why the experiments are randomized.

The average time for all short space parkings is 37 seconds, from 10 experiments. The average time for all long space parkings is 30 seconds, from 10 experiments. Constructing the confidence for the difference between these two means:

$$\overline{x}_S - \overline{x}_L \pm c_t \sqrt{8.83^2 \left(\frac{1}{10} + \frac{1}{10}\right)}$$

$$c_t = \pm 2.1 \text{ with 18 degrees of freedom}$$

$$(37 - 30) \pm (2.1)(3.95)$$

$$7.0 \pm 8.29$$

The confidence interval is: $-1.29 \leq \mu_S - \mu_L \leq 15.3$

This interval spans zeros, so statistically there is no difference between short and long parkings in terms of time duration. However, the interval only just spans zero, so there is still some chance that short parking bays have longer parking times on average.

You could also have done this question using a paired test: 5 differences of long minus short under automatic parking, and another 5 from manual parking. Use these 10 differences and calculate the mean and standard deviation, and calculate a confidence interval. The conclusion shows that parking in long spaces is of shorter duration.

2. The most appropriate statistical test is a paired test. There is a commonality between the experimental pairs in each row (the same car and same parking conditions) that can be removed by subtraction. This commonality does not exist between the rows. If we don't subtract, and do an ordinary test of differences, the effect of long and short parking bays and different car types will contaminate the experimental values, since we didn't control for these effects. This subtraction and subsequent analysis will be more sensitive than a regular statistical tests of differences.

   We assume these 10 difference values are independent, which is likely to be, for the same reason given in the first part of the question.

We also assume these 10 differences are normally distributed as $\overline{w} \sim \mathcal{N}(\mu_w, \sigma_w/n)$, which can be checked with a q-q plot.

$$\overline{w} \pm c_t \frac{s_{\overline{w}}}{\sqrt{n}}$$
$$c_t = \pm 2.3 \text{ with 9 degrees of freedom}$$
$$11 \pm 2.3 \cdot \frac{5.7}{\sqrt{10}}$$
$$11 \pm 2.3 \cdot \frac{5.7}{\sqrt{10}}$$
$$11 \pm 4.15$$

So a confidence interval for the average difference is $6.85 \leq \mu_w \leq 15.1$ seconds. So parking with the automatic system results in reduced times when compared to regular, manual drivers.

Question 4 [11 = 2 + 2 + 3 + 2 + 2]

One criterion for a "clean room" is to maintain a positive differential pressure; i.e. the pressure in the room must exceed the pressure on the other side of the door. This is to prevent contaminants from entering the room via any gaps around the door. Differential pressure gauges are commercially available for this purpose and can be made to record their data automatically.

1. Which process monitoring chart would you choose to monitor that the differential pressure remains at the target value of +30 Pascal. Explain your choice.
2. Give an example of a type II error in the context of this example.
3. In general, in the context of process monitoring, what can you do if too many type II errors occur on a particular monitoring chart?
4. Explain what is meant by a process being "in a state of statistical control".
5. Explain what a "special cause" is in the context of the above example.

Solution

1. CUSUM or EWMA chart with small λ. A Shewhart chart alone will not be too helpful, because some drifts may not be picked. A Shewhart chart with the Western Electric rules may work.
2. A change in the room differential pressure occurs (e.g. the door is propped open, or the incoming supply air rate is increases, raising the pressure in the room), but this shift does not show up outside the alarm limits.
3. You can make the control limits narrower or tighter, but this is at the expense of increased type I errors. Alternatively you could use another monitoring chart.
4. A state of statistical control indicates that no special causes are present: the monitoring chart should remain within limits.
5. A special cause is a destabilizing, or unusual event. For example, the door is propped open and the differential pressure will drop from target. Someone opens the damper and too much air is charged to the room, raising the differential pressure above target.

Question 5 [10 = 2 + 2 + 2 + 2 + 2]

The following confidence interval is reported by our company for the amount of sulphur dioxide measured in parts per billion (ppb) that we send into the atmosphere.

$$123.6 \text{ ppb} \leq \mu \leq 240.2 \text{ ppb}$$

Only $n = 21$ raw data points (one data point measured per day) were used to calculate that 90% confidence interval. A z-value would have been calculated as an intermediate step to get the final confidence interval, where $z = \dfrac{\overline{x} - \mu}{s/\sqrt{n}}$.

1. What assumptions were made about those 21 raw data points to compute the above confidence interval?
2. Which lower and upper critical values would have been used for z? That is, which critical values are used before unpacking the final confidence interval as shown above.
3. What is the standard deviation, s, of the raw data?
4. Today's sulphur dioxide reading is 460 ppb and your manager wants to know what's going on; you can quickly calculate the probability of seeing a value of 460 ppb, or greater, to help judge the severity of the pollution. How many days in a 365 calendar-day year are expected to show a sulphur dioxide value of 460 ppb or higher?
5. Explain clearly why a wide confidence interval is not desirable, from an environmental perspective.

Solution

1. The 21 data points are independent and come from *any distribution* of finite variance. The values are **not** required to come from a normal distribution.
2. From the t-distribution at 20 degrees of freedom, with 5% in each tail: 1.72. The t-distribution is used because the standard deviation is estimated, rather than being a population deviation.
3. The standard deviation may be calculated from:

$$\begin{aligned} UB - LB = 240.2 - 123.6 = 2 \times c_t \frac{s}{\sqrt{n}} &= (2)(1.72)\frac{s}{\sqrt{n}} \\ s &= \frac{(116)(\sqrt{n})}{(2)(1.72)} \\ s &= 154.5 \end{aligned}$$

4. The probability calculation requires a mean value. Our best guess for the mean is the midpoint of the confidence interval, which is always symmetric about the estimated process mean, $\overline{x} = \dfrac{240.2 - 123.6}{2} + 123.6 = 181.9$. Note that this is not the value for μ, since μ is unknown.

$$z = \frac{460 - 181.9}{154.5} = 1.80$$

Probability is $1 - 0.9641 = 0.0359$, or about $0.0359 \times 365 = 13.1$, or about 13 days in the year.

5. A wide confidence interval implies that our sulphur dioxide emissions are extremely variable (the confidence interval bounds are a strong function of the process standard deviation). Some days we are putting more pollution up into the air and balancing it out with lower pollution on other days. Those days with high pollution are more environmentally detrimental.

Question 6 [15 = 2 + 2 + 2 + 2 + 2 + 3 + 2]

The mass of steam required to heat a building can be related to the average ambient temperature. Being able to predict the mass of steam required, s, when given the ambient temperature, T, can help in energy planning, and ultimately lead to energy reduction.

The table below lists the mass of steam produced [ton] in the past when the average temperature over a 2 hour period, recorded in K, was observed outside.

Temperature = T [Kelvin]	267	268	272	273	278	281	283	288	289	293	296
Steam produced = s [tons]	220	251	211	210	155	152	122	157	100	64	58

The following calculations have already been performed for you:

- Number of samples, $n = 11$
- Temperature: mean = $\overline{T} = 281$ K and standard deviation is 10.0 K
- Average steam produced, $\overline{s} = 155$ tons, and standard deviation is 64.4 tons.

The modified output from a certain statistical package is:

```
Coefficients:
                          Standard
                Value     Error
-----------------------------------
(Intercept)   1871.6936   183.2183
T               -6.1168     0.6523

Residual standard error: ____ on ____ degrees of freedom
Multiple R-squared: ____
```

A portion of the analysis of variance table is given below:

```
     Analysis of Variance
-----------------------------
                       Sum of
Source                Squares
-----------------------------
Due to the model        37572
Due to error             3845
-----------------------------
Total                   41417
```

1. What is the interpretation of the intercept? Is it a useful piece of knowledge derived from the model?
2. What is the interpretation of the slope coefficient? Is it a useful piece of knowledge derived from the model?
3. What is the multiple R^2 value that would have been calculated for this model?

4. What is the standard error, S_E, value that would have been calculated for this model?
5. How would you interpret your calculated standard error value, and what assumptions are required to match your interpretation?
6. Give a confidence interval for the slope coefficient and interpret what it means.
7. Which other input variable might be added to the linear model to help improve the model's prediction ability?

Solution

1. Intercept = 1872 tons, the amount of steam that is expected to be produced, when the ambient temperature in 0 K. This is clearly not a useful piece of information, there are no data around this region to support this interpretation.
2. Slope = -6.1 tones/K: indicates we expect to produce 6 tons less of steam, for every 1 K rise in ambient temperature.
3. $R^2 = 0.9072 = \dfrac{37572}{41417} = 1 - \dfrac{3845}{41417}$
4. $S_E = \sqrt{\dfrac{3845}{11-2}} = 20.7$ tons.
5. Assuming the residuals are normally distributed (easily checked with a q-q plot), the standard error of 20.7 tones is the one-sigma standard deviation for error. So we expect about 70% of our residuals to lie within a range of ± 20.7 tons, and 95% of residuals to lie within a range of ± 41.4 tons.
6. The slope coefficient estimate has standard error of 0.6523 (from the software output), or it could be calculated as $S_E^2(b_1) = \dfrac{S_E^2}{\sum_j \left(T_j - \overline{T}\right)^2} = \dfrac{20.7^2}{(10.0^2 \times 10)} \approx 0.65$. The $\times 10$ term is to because the reported standard deviation for temperature is divided by $n-1$, but we require only the numerator portion.

   From this we can construct the confidence interval for the actual slope coefficient, β_1. I have used the 95% confidence level, but you could use any level you prefer. The degrees of freedom to use for the t-distribution are $n - k = 11 - 2 = 9$.

$$\begin{aligned} -c_t &\leq \frac{b_1 - \beta_1}{S_E(b_1)} \leq +c_t \\ b_1 - c_t S_E(b_1) &\leq \beta_1 \leq b_1 + c_t S_E(b_1) \\ -6.1168 - 2.26 \times 0.6523 &\leq \beta_1 \leq -6.1168 + 2.26 \times 0.6523 \\ -7.59 &\leq \beta_1 \leq -4.64 \end{aligned}$$

   This shows that at the 95% confidence level, the range within which we can expect to find the true slope coefficient. This can help guide us in the sensitivity.
7. Wind speed will have an effect on the building and the heat losses. Another example might be humidity, or the average number of people in the building.

Question 7 [7 = 2 + 3 + 2]

In the course notes on the section on comparing differences between two groups we used, without proof, the fact that:

$$\mathcal{V}\{\bar{x}_B - \bar{x}_A\} = \mathcal{V}\{\bar{x}_B\} + \mathcal{V}\{\bar{x}_A\}$$

Using the fact that $\mathcal{V}\{cx\} = c^2\mathcal{V}\{x\}$, you can show that:

$$\mathcal{V}\{\bar{x}_B + \bar{x}_A\} = \mathcal{V}\{\bar{x}_B\} + \mathcal{V}\{\bar{x}_A\}$$

1. The first equation is only correct when an important assumption is true; what is that assumption?
2. *Based on an actual industrial problem*: A filling machine doses a drug to a canister. The patient will inhale the drug (imagine an asthma pump). The weight of the drug in the canister must be added as precisely and accurately as possible, to avoid patient over- or under-dosing.

   The weight filled will fluctuate with temperature in the building and is theoretically calculated as having a standard deviation of 32mg due to typical temperature variations. The filling line has 6 machines that fill the canisters and the variability from machine-to-machine is 40mg. The operators calibrate the machines at the start of each shift, and their estimated calibration accuracy is estimated at 15mg. The wear and tear on the machine parts over the year is estimated to only add an extra 10mg of variation.

   What is the expected long-term standard deviation of fill weights recorded from this process? What assumption(s) do you have to make to calculate this?
3. Continuing the above question, assume the long-term standard deviation of fill weights was 40 mg (not the correct answer for the previous question), what is the process capability ratio if the filler operates midway between the upper and lower specification limits, where USL = 1200 mg and LSL = 800 mg?

Solution

1. Assume the operation in system A is independent of system B's operation. This implies $\overline{x}_A$ is independent of $\overline{x}_B$.
2. Using a similar concept to the above, the variation from multiple sources can be added up, as long the variation from each source is independent of the other. In this case, it would require the variance due to temperature is unrelated to the machine-to-machine variance. We'd have to check every pair of combinations to ensure they are independent.

   If this assumption is true, then the total variance in the process is $\sigma^2_{\text{total}} = 32^2 + 40^2 + 15^2 + 10^2 = 54.3^2$, or $\sigma_{\text{total}} = 54.3$ mg.
3. PCR = $\dfrac{\text{USL} - \text{LSL}}{6\sigma} = \dfrac{1200 - 800}{(6)(40)} = 1.67$, which is a reasonable capability ratio.

The end.