Statistics for Engineering, 4C3/6C3

Written midterm, 08 February 2010

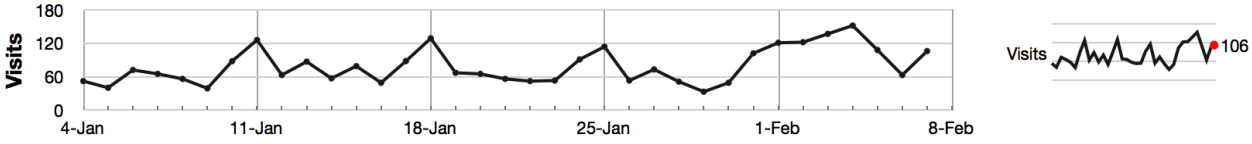
Kevin Dunn, dunnkg@mcmaster.ca **McMaster University**

Note:

- There are 6 pages to this midterm, printed double-sided. Please ensure your copy is complete.
- **Please only write your student number on the answer booklets (do not write your name)**.
- The midterm duration is 2.5 hours for all students.
- 600 level students: should answer **all** questions. Maximum number of points = 50.
- 400 level students: answer the questions marked - please read carefully. Points for a 100% grade = 42.
- 400 level students: you will get extra credit for answering 600-level questions.
- You may use any notes, pages, and books for this midterm exam.
- The only electronic device you should have on your desk is a calculator.
- The use of cellphones, netbooks, laptops, or communication devices of any type are not allowed.
- Please ask the invigilator if you need a copy of the table of normal and t-distributions.
- There are several non-chemical engineering students in this class, so I have included diagrams with some of the questions to help you understand what the question is about.
- For this written exam I will value your ideas and how you think about the problem more than accuracy (I assess accuracy in the take-home midterm). Use approximations when reading the tables of distributions, and state any assumptions, as long as they are reasonable.

1 Visualizing data (400- and 600-level students) [1+2]

The data shown here are the number of visits to the website for this course, http://stats4.eng.mcmaster.ca, since the start of the course on Monday, 04 January 2010. There are 92 students registered for the course, however the site is also publicly available.



1. What are the names (type) of the 2 plots shown?
2. List any 2 interesting features in these data.

Solution

1. The plots are a time-series plot and a sparkline. The sparkline shows exactly the same data, just a more compact form (without the labelling on the axes).
2. Features shown in the data are:
   - a noticeable weekly cycle, peaking with the period just before assignments must be handed in (Sunday and Monday), and people printing class notes
   - after the class on Monday, the web site traffic drops off again
   - there was a high level of traffic during the week of the take-home midterm (2 to 5 February)
   - some days there are more than 92 visits, indicating that students visit the site more than once per day, or it is due to external (non-McMaster) traffic

2 Univariate statistics

2.1 Distributions (400- and 600-level) [1+1]

A food production facility fills bags with potato chips. The advertised bag weight is 35.0 grams. But, the current bagging system is set to fill bags with a mean weight of 37.4 grams, and this done so that only 1% of bags have a weight of 35.0 grams or less.

- Back-calculate the standard deviation of the bag weights, assuming a normal distribution.
- Out of 1000 customers, how many are lucky enough to get 40.0 grams or more of potato crisp goodness in their bags?

Solution

- Calculate the z-value and find which fraction of z falls at or below 1% of the probability area. From the tables this is -2.326.

  Then solve for σ:

$$z = \frac{35 - 37.4}{\sigma} = -2.326$$
$$\sigma = \frac{35 - 37.4}{-2.326} = 1.03 \text{ grams}$$

- Probability of 40.0 grams of more is the area above the corresponding z-value:

$$z > \frac{40 - 37.4}{1.03}$$
$$z > 2.52$$

  The exact answer is `(1 - pnorm(2.52))*1000 = 5.86`, though using tables you could use the value corresponding to $z = 2.5$, which is 99.38%, which is the area below that z-value. The area above it is 0.62%, corresponding to 6.2 people. Either 5, 6 or 7 people is an acceptable answer, depending on your rounding error.

2.2 Tests for differences (400- and 600-level) [8]

You are a new engineer at a pharmaceutical company. One of the steps in the flowsheet is to blend three powders for a tablet: the excipient (an inactive magnesium stearate base), a binder, and the active ingredient. The mixing process is tracked using a wireless near infrared (NIR) probe embedded in a V-blender. The mixer is stopped when the NIR spectra become stable. A new supplier of magnesium stearate is being considered that will save $ 294,000 per year.

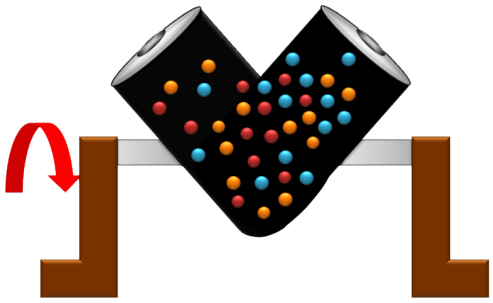


Figure 1: Figure from Wikipedia (http://en.wikipedia.org/wiki/Industrial_mixer)

The 15 most recent runs with the current magnesium stearate supplier had an average mixing time of 2715 seconds, and a standard deviation of 390 seconds. So far you have run 6 batches from the new supplier, and the average mixing time of these runs is 3115 seconds with a standard deviation of 452 seconds. Your manager is not happy with these results so far - this extra mixing time will actually cost you more money via lost production.

The manager wants to revert back to the original supplier, but is leaving the decision up to you; what would be your advice? Show all calculations and describe any additional assumptions, if required.

Solution

This question, similar to most real statistical problems, is open-ended. This problem considers whether a significant difference has occurred. And in many cases, even though there is significant difference, it has to be weighed up whether there is a *practical* difference as well, together with the potential of saving money (increased profit).

To get full grade you to have to state any assumptions you make, compute a confidence interval for the difference and interpret it. Additional points were awarded if you considered other factors (see below).

The decision is one of whether the new material leads to a significant difference in the mixing time. It is desirable, from a production point of view, that the new mixing time is shorter, or at least the same. Some notation:

$$\begin{array}{rcl@{\qquad\qquad}rcl} \hat{\mu}_{\text{Before}} = \bar{x}_B &=& 2715 & \hat{\mu}_{\text{After}} = \bar{x}_A &=& 3115 \\ \hat{\sigma}_{\text{Before}} = s_B &=& 390 & \hat{\sigma}_{\text{After}} = s_A &=& 452 \\ n_B &=& 15 & n_A &=& 6 \end{array}$$

Assumptions required to compare the two groups:

- The individual samples within each group were taken independently, so that we can invoke the central limit theorem and assume these means and standard deviation are normal distributed.
- Assume the individual samples within each group are from a normal distribution as well.
- Assume that we can pool the variances, i.e. σ_{Before} and σ_{After} are from comparable distributions.
- Using the pooled variance implies that the z-value follows the t-distribution.
- The mean of each group (before and after) is independent of the other (very likely true).
- No other factors were changed, other than the raw material (we can only hope, though in practice this is often not true, and a paired test would eliminate any differences like this).

Calculating the pooled variance:

$$\begin{aligned} s_P^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A - 1 + n_B - 1} \\ &= \frac{(6-1)452^2 + (15-1)390^2}{6 - 1 + 15 - 1} \\ &= 165837 \end{aligned}$$

Computing the z-value for this difference:

$$\begin{aligned} z &= \frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)}{\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \\ z &= \frac{(2715 - 3115) - (\mu_B - \mu_A)}{\sqrt{165837\left(\frac{1}{6} + \frac{1}{15}\right)}} \\ z &= \frac{-400 - (\mu_B - \mu_A)}{196.7} = -2.03 \qquad \text{on the hypothesis that} \qquad \mu_B = \mu_A \end{aligned}$$

The probability of obtaining this value of z can be found using the t-distribution at 6 + 15 - 2 = 19 degrees of freedom (because the standard deviation is an estimate, not a population value). Using the tables provided in the exam, a value of 0.025, or 2.5% is found (in R, it would be `pt(-2.03, df=19) = 0.0283`, or 2.83%). At this point one can argue either way that the new excipient leads to longer times, though I would be inclined to say that this probability is too small to be due to chance alone. Therefore there is a significant difference, and we should revert back to the previous excipient. Factors such as operators, and other process conditions could have affected the 6 new runs.

Alternatively, and this is the way I prefer to look at these sort of questions, is to create a confidence interval. At the 95% level, the value of c_t in the equation below, using 19 degrees of freedom is `qt(0.975, df=19) = 2.09` (any value close to this from the tables is acceptable):

$$\begin{array}{rcccl}
-c_t & \leq & z & \leq & +c_t \\
(\bar{x}_B - \bar{x}_A) - c_t\sqrt{s_P^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)} & \leq & \mu_B - \mu_A & \leq & (\bar{x}_B - \bar{x}_A) + c_t\sqrt{s_P^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)} \\
-400 - 2.09\sqrt{165837\left(\frac{1}{6}+\frac{1}{15}\right)} & \leq & \mu_B - \mu_A & \leq & -400 + 2.09\sqrt{165837\left(\frac{1}{6}+\frac{1}{15}\right)} \\
-400 - 412 & \leq & \mu_B - \mu_A & \leq & -400 + 412 \\
-812 & \leq & \mu_B - \mu_A & \leq & 12
\end{array}$$

The interpretation of this confidence interval is that there is no difference between the current and new magnesium stearate excipient. The immediate response to your manager could be "*keep using the new excipient*".

However, the confidence interval's asymmetry should give you pause, certainly from a practical point of view (this is why I prefer the confidence interval - you get a better interpretation of the result). The 12 seconds by which it overlaps zero is so short when compared to average mixing times of around 3000 seconds, with standard deviations of 400 seconds. The practical recommendation is that the new excipient has longer mixing times, so "*revert to using the previous excipient*".

Up to this point you would get full grade. One other aspect of this problem that might bother you is the low number of runs (batches) used. Let's take a look at how sensitive the confidence interval is to that. Assume that we perform one extra run with the new excipient ($n_A = 7$ now), and assume the pooled variance, $s_p^2 = 165837$ remains the same with this new run. The new confidence interval is:

$$\begin{array}{rcccl}
(\bar{x}_B - \bar{x}_A) - c_t\sqrt{s_P^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)} & \leq & \mu_B - \mu_A & \leq & (\bar{x}_B - \bar{x}_A) + c_t\sqrt{s_P^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)} \\
(\bar{x}_B - \bar{x}_A) - 2.09\sqrt{165837\left(\frac{1}{7}+\frac{1}{15}\right)} & \leq & \mu_B - \mu_A & \leq & (\bar{x}_B - \bar{x}_A) + 2.09\sqrt{165837\left(\frac{1}{7}+\frac{1}{15}\right)} \\
(\bar{x}_B - \bar{x}_A) - 390 & \leq & \mu_B - \mu_A & \leq & (\bar{x}_B - \bar{x}_A) + 390
\end{array}$$

So comparing this ± 390 with 7 runs, to the ± 412 with 6 runs, shows that the confidence interval shrinks in quite a bit, much more than the 12 second overlap of zero. Of course we don't know what the new $\bar{x}_B - \bar{x}_A$ will be with 7 runs, so my recommendation would be to perform at least one more run with the new excipient, but I suspect that the new run would show there to be a significant difference, and statistically confirm that we should "*revert to using the previous excipient*".

Grading breakdown:

- 2 marks for stating assumptions needed to use the above formula.
- 1 mark for the pooled variance.
- Two options (3 marks for either one):
  - the z-value, and its corresponding probability from the t-distribution with 19 DOF
  - formulate the confidence interval, using the t-distribution with 19 DOF
- 2 marks for interpretation and recommendations from your results, even if your numbers were incorrect, an appropriate recommendation is still graded.

2.3 Paired vs unpaired (400- and 600-level) [2]

List an advantage of using a paired test over an unpaired test. Give an example, not from the class notes, that illustrates your answer.

Solution

One primary advantage of pairing is that any systematic difference between the two groups (A and B) is eliminated. For example, a bias in the measurement will cancel out when calculating the pairs of differences. Any example is suitable as an answer: e.g. laboratory miscalibration; an offset in an on-line sensor, etc.

Other advantages are that the raw data do not need to be normally distributed, only the paired differences.

Another advantage is that randomization of the trials is required in the unpaired case (often a costly extra expense), whereas in the paired case, we only need to be sure the pairs are independent of each other (that's much easier to assume, and often true). For example testing drug A and B on a person, some time apart. The pairs are run on the same person, but each person in the drug trial is independent of the other.

2.4 Paired vs unpaired (600-level only) [4]

An *unpaired* test to distinguish between group A and group B was performed with 18 runs: 9 samples for group A and 9 samples for group B. The pooled variance was 86 units.

Also, a *paired* test on group A and group B was performed with 9 runs. After calculating the paired differences, the variance of these differences was found to be 79 units.

Discuss, in the context of this example, an advantage of paired tests over unpaired tests. Assume 95% confidence intervals, and that the true result was one of "no significant difference between method A and method B". Give numeric values from this example to substantiate your answer.

Solution

One advantage of the paired test is that often a fewer number of samples are required to obtain a more sensitive result than when analyzing the data as from two distinct, unpaired groups.

Construct the confidence interval for both cases, substitute in these values and then compare the confidence intervals. The equations for both confidence intervals are derived directly from the z-value appearing in the class notes.

Unpaired case:

$$
\begin{aligned}
-c_t &\leq \frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)}{\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \leq +c_t \\
(\bar{x}_B - \bar{x}_A) - c_t\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} &\leq \mu_B - \mu_A \leq (\bar{x}_B - \bar{x}_A) + c_t\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \\
(\bar{x}_B - \bar{x}_A) - 2.12 \times \sqrt{86\left(\frac{1}{9} + \frac{1}{9}\right)} &\leq \mu_B - \mu_A \leq (\bar{x}_B - \bar{x}_A) + 2.12 \times \sqrt{86\left(\frac{1}{9} + \frac{1}{9}\right)} \\
(\bar{x}_B - \bar{x}_A) - 9.27 &\leq \mu_B - \mu_A \leq (\bar{x}_B - \bar{x}_A) + 9.27
\end{aligned}
$$

The c_t value for the unpaired case is from the t-distribution with 16 degrees of freedom, a value of around 2.12.

Paired case:

In this case the vector of differences is w, and by the central limit theorem it is distributed as $w \sim \mathcal{N}\left(\mu_{B-A}, \sigma_w^2/n\right)$, but we use the estimated variance, s_w^2 instead.

$$
\begin{array}{rcccl}
-c_t & \leq & \dfrac{\bar{w} - \mu_{B-A}}{s_w/\sqrt{n}} & \leq & +c_t \\
\\
\bar{w} - c_t \dfrac{s_w}{\sqrt{n}} & \leq & \mu_w & \leq & \bar{w} + c_t \dfrac{s_w}{\sqrt{n}} \\
\bar{w} - 2.3 \dfrac{\sqrt{79}}{\sqrt{9}} & \leq & \mu_w & \leq & \bar{w} + 2.3 \dfrac{\sqrt{79}}{\sqrt{9}} \\
\bar{w} - 6.81 & \leq & \mu_w & \leq & \bar{w} + 6.81
\end{array}
$$

The c_t value for the paired case is from the t-distribution with 8 degrees of freedom, a value of around 2.3.

The key result of this question is that the confidence interval for the paired case is tighter (narrower) than the confidence interval from the unpaired case. Given that the true result was one of no significant difference, it implies that $\mu_A = \mu_B$ and that $\mu_w = 0$. The tighter confidence interval comes purely from the fact that the standard deviation used for the paired case is smaller, $\sqrt{\frac{79}{9}}$ *vs* the $\sqrt{86\left(\frac{1}{9}+\frac{1}{9}\right)}$ from the unpaired case. This is not due to the variances, since $\sqrt{86} \approx \sqrt{79}$, i.e. (9.27 vs 8.88), but rather due to the fact that that unpaired standard deviation is multiplied by $\sqrt{2/9}$, while the paired standard deviation is multiplied by $\sqrt{1/9}$.

So while the c_t value for the paired case is actually larger (widening the confidence interval due to the fewer degrees of freedom), the overall effect is that the paired confidence interval is narrower than the unpaired confidence interval. This result holds for most cases of paired and unpaired studies, though not always.

2.5 Confidence intervals (400- and 600-level) [2+2+1]

You are convinced that a different impeller (mixing blade) shape for your tank will lead to faster, i.e. shorter, mixing times. The choices are either an axial blade or a radial blade.

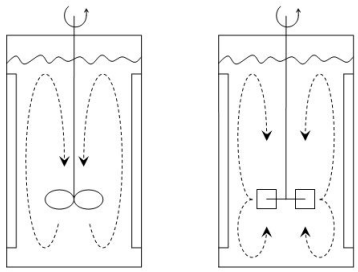


Figure 2: Axial and radial blades; figure from Wikipedia (http://en.wikipedia.org/wiki/Impeller)

Before obtaining approval to run some experiments, your team wants you to explain how you will interpret the experimental data. Your reply is that you will calculate the average mixing time from each blade type and then calculate a confidence interval for the difference. A team member asks you what the following 95% confidence intervals would mean:

1. -453 seconds $\leq \mu_{\text{Axial}} - \mu_{\text{Radial}} \leq 390$ seconds
2. -21 seconds $\leq \mu_{\text{Axial}} - \mu_{\text{Radial}} \leq 187$ seconds

For both cases (a) explain what the confidence interval means in the context of this experiment, and (b) whether the recommendation would be to use radial or axial impellers to get the shortest mixing time.

3. Now assume the result from your experimental test was -21 seconds $\leq \mu_{\text{Axial}} - \mu_{\text{Radial}} \leq 187$ seconds; how can you make the confidence interval narrower?

Solution:

1. This confidence interval spans zero, and nearly symmetrically. This implies the population difference is likely zero, while the symmetry implies their is no preference either way: the difference in mixing times is as low as -453 seconds or as high as 390 seconds. The recommendation is that either the axial or radial impeller could be used, with no expected long-term difference. Use the cheaper impeller; or use the axial impeller if the costs are the same (only because of the very slight imbalance in the CI). Note that there is a 5% chance that the confidence interval does not contain the true difference.
2. This confidence interval also spans zero, so there is **no statistical difference** between the two impellers. However the CI does not span zero symmetrically. The asymmetry of the interval makes me much less comfortable recommending that there is no **practical difference** between the impellers. It often happens in these cases that by removing a single data point that the confidence interval does not span zero anymore. In this case I would recommend either impeller, but if there is no cost difference, I would prefer the radial impeller, as it might have shorter mixing times, especially if the confidence interval quoted here is only due to one observation. A careful review of the raw data would be useful in this case.
3. The confidence interval can be made narrower in 2 ways (as long as the sample mean and sample standard deviation remain stable):
   - Use more data points, n in both groups.
   - Choose a lower degree of confidence, e.g. 90% instead of 95%, which is really just an artificial reduction of the interval.

   One can also reduce the interval by shrinking the standard deviation, but that's usually not a practical possibility. You cannot perform a paired test, as you only have one mixing tank.

Interpreting confidence intervals

Recall the definition of the confidence interval is subtle: it says 95% of the time, the upper and lower bounds of the confidence interval contain the true value of the parameter; it does *not* say there is a 95% probability the true value of the parameter lies inside the bounds. That last part implies the true value of the parameter can vary, which it can't: the true parameter value is fixed, only the bounds change. I wasn't too picky on this point, since a lot depends how you phrase your answer.

2.6 Robust methods (600-level only) [2]

The enrichment paper by Rousseeuw discusses the breakdown point of a statistic. Describe what the breakdown point is, and give two examples: one with a low breakdown point, and one with a high breakdown point. Use a vector of numbers to help illustrate your answer.

Solution

PJ Rousseeuw (Tutorial to Robust Statistics, *Journal of Chemometrics*, **5**, 1-20, 1991, link to the paper) defines the breakdown point on page 3 of his paper as "... 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".

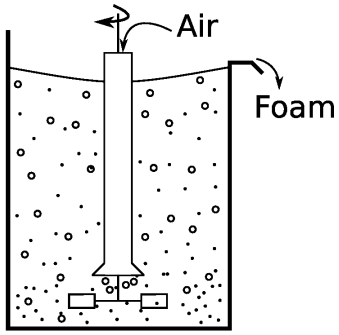
A statistic with a low breakdown point is the mean, of the n values used to calculate the mean, only 1 needs to be replaced to make the estimator unbounded; i.e. its breakdown point is $1/n$. The median though has a breakdown point of 50%, as one would have to replace 50% of the n data points in the vector before the estimator becomes unbounded.

Use this vector of data as an example: $[2, 6, 1, 9151616, -4, 2]$. The mean is 1525270, while the median is 2.

3 Process monitoring

3.1 Monitoring charts (400- and 600-level) [1+1+2]

1. Which type of monitoring chart would be appropriate to detect unusual spikes (outliers) in your production process?
2. A tank uses small air bubbles to keep solid particles in suspension. If too much air is blown into the tank, then excessive foaming occurs; if too little air is blown into the tank the particles sink and drop out of suspension. Which monitoring chart would you use to ensure the airflow is always near target?



3. Answer **only one** of the following questions, whichever you prefer:
   - Do you think a Shewhart chart would be suitable for monitoring the closing price of a stock on the stock market? Please explain your answer if you agree, or describe an alternative if you disagree.
   - Describe how a control chart could be used to prevent over-control of a batch-to-batch process. (A batch-to-batch process is one where a batch of materials is processed, followed by another batch, and so on).

Solution

1. A Shewhart chart has no memory, and is suited to detecting unusual spikes in your production. CUSUM and EWMA charts have memory, and while they would pick up this spike, they would also create a long duration of false alarms after that. So those charts are much less appropriate.
2. A CUSUM chart would be a suitable chart to monitor that the airflow is near target. While a Shewhart chart is also intended to monitor the location of a variable, it has a much larger run length for detecting small shifts. An EWMA chart with small λ (long memory) would approximate a CUSUM chart, and so would also be suitable.
3. Answer either question:
   - No, a Shewhart chart is not suitable for monitoring stock prices. Stock prices are volatile variables (not stable), so there is no sense in monitoring their location. Hopefully the stock is moving up, which it should on average, but the point is that stock prices are not stable. Nor are stock prices independent day-to-day.

     So what aspect of a stock price is stable? The difference between the opening and closing price of a stock is remarkably stationary. Monitoring the day-to-day change in a stock price would work. Since you aren't expected to know this fact, any reasonable answer that attempts to monitor a *stable* substitute for the price will be accepted. E.g. another alternative is to remove the linear up or down trend from a stock price and monitor the residuals.

     There are many alternatives; if this sort of thing interests you, you might find the area called technical analysis worth investigating. An EWMA chart is widely used in this sort of analysis, and most of you recommended this in your answer.
   - Over-control of any process takes place when too much corrective action is applied. Using the language of feedback control, your gain is the right sign, but the magnitude is too large. Batch processes are often subject to this phenomenon: e.g. the operator reduces the set-point temperature for the next batch, because

the current batch produced product with a viscosity that was too high. But then the next batch has a viscosity that is too low, so the operator increases the temperature set-point for the following batch. This constant switching is known as over-control (the operator is the feedback controller and his/her gain is too high, i.e. they are over-reacting).

A control chart such as a Shewhart chart would help the operator: if the previous batch was within the limits, then s/he should not take any corrective action. Only take action when the viscosity value is outside the limits. An EWMA chart would additionally provide a one-step ahead prediction, which is an advantage.

3.2 Shewhart charts (400- and 600-level) [1+2+2]

You need to construct a Shewhart chart. You go to your company's database and extract data from 10 periods of time lasting 6 hours each. Each time period is taken approximately 1 month apart so that you get a representative data set that covers roughly 1 year of process operation. You choose these time periods so that you are confident each one was from in control operation. Putting these 10 periods of data together, you get one long vector that now represents your phase I data.

- There are 8900 samples of data in this phase I data vector.
- You form subgroups: there are 4 samples per subgroup and 2225 subgroups.
- You calculate the mean within each subgroup (i.e. 2225 means). The mean of those 2225 means is 714.
- The standard deviation within each subgroup is calculated; the mean of those 2225 standard deviations is 98.

1. Give an unbiased estimate of the process standard deviation?
2. Calculate lower and upper control limits for operation at ± 3 of these standard deviations from target. These are called the action limits.
3. Operators like warning limits on their charts, so they don't have to wait until an action limit alarm occurs. Discussions with the operators indicate that lines at 590 and 820 might be good warning limits. What percentage of in control operation will lie inside the proposed warning limit region?

Solution

1. An unbiased estimate of the process standard deviation is $\hat{\sigma} = \dfrac{\bar{S}}{a_n} = \dfrac{98}{0.921} = 106.4$, since the subgroup size is $n = 4$.
2. Using the data provided in the question:

$$\text{UCL} = \bar{\bar{x}} + 3\frac{\bar{S}}{a_n\sqrt{n}} = 714 + 3 \times \frac{98}{0.921 \times 2} = 874$$
$$\text{LCL} = \bar{\bar{x}} - 3\frac{\bar{S}}{a_n\sqrt{n}} = 714 - 3 \times \frac{98}{0.921 \times 2} = 554$$

3. Since Shewhart charts assume a normal distribution in their derivation, we can use the same principle to calculate a z-value, and the fraction of the area under the distribution. But you have to be careful here: which standard deviation do you use to calculate the z-value? You should use the subgroup's standard deviation, not the process standard deviation. The Shewhart chart shows the subgroup averages, so the values of 590 and 820 refer to the subgroup values.

   If that explanation doesn't make sense, think of the central limit theorem: the mean of a group of samples, $\bar{x} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, where σ^2 is the process variance, and σ^2/n is the subgroup variance of $\bar{x}$.

$$z_{\text{low}} = \frac{x_{\text{low}} - \bar{\bar{x}}}{\hat{\sigma}/\sqrt{n}} = \frac{590 - 714}{106.4/\sqrt{4}} = -2.33$$
$$z_{\text{high}} = \frac{x_{\text{high}} - \bar{\bar{x}}}{\hat{\sigma}/\sqrt{n}} = \frac{820 - 714}{106.4/\sqrt{4}} = +2.00$$

The area below -2.33 is `pnorm(-2.33) = 0.009903076`, though I will accept any value around 1%, eyeballed from the printed tables. The area below +2.00 is 97.73%, which was on the tables already. So the total amount of normal operation within the warning limits is 97.73-1.00 = **96.7%**.

The asymmetry in their chosen warning limits might be because a violation of the lower bound is more serious than the upper bound.

3.3 Process capability (400- and 600-level)

- 400 level students: only answer the potato chip question, or the plastics question **[2]**
- 600 level students: answer both questions **[4]**

A. Potato chip question [0.5 + 0.5 + 1]

Back to the potato chips example. The bagging system fills the bags with a target weight of 37.4 grams and the lower specification limit is 35.0 grams. Assume the bagging system fills the bags with a standard deviation of 0.8 grams (note, this is **not** the answer to a previous question):

1. What is the current Cpk of the process?
2. To what target weight would you have to set the bagging system to obtain Cpk=1.3?
3. How can you adjust the Cpk to 1.3 without adjusting the target weight (i.e. keep the target weight at 37.4 grams)?

Solution

1. Recall the Cpk is defined relative to the closest specification limit. So in this case it must be due to the lower limit. Cpk = $\dfrac{\bar{\bar{x}} - LSL}{3\sigma} = \dfrac{37.4 - 35.0}{3 \times 0.8} = 1.0$
2. To obtain Cpk = 1.3 we solve the above equation for $\bar{\bar{x}} = 1.3 \times 3 \times 0.8 + 35.0 = 38.12$ grams.
3. Changing the lower specification limit is not an option to raise Cpk, because the bags are sold as containing 35.0 grams of snackfood. Changing the specification limit is in general an artificial way of changing Cpk. The only practical way to improve Cpk is to decrease the process variance (e.g. using better equipment with tighter control). The new $\sigma = \dfrac{37.4 - 35.0}{3 \times 1.3} = 0.615$ grams.

B. Plastic sheet question [0.5 + 0.5 + 0.5 + 0.5]

Plastic sheets are manufactured on your blown film line. The Cp value is 1.7. You sell the plastic sheets to your customers with specification of 2 mm $\pm$ 0.4 mm.

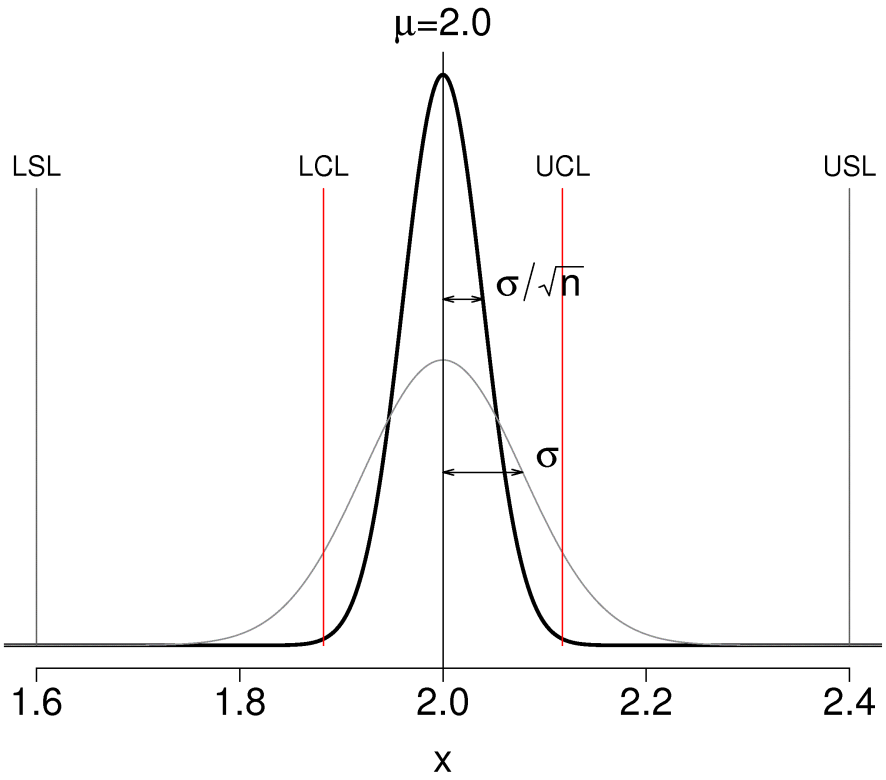
1. List three important assumptions you must make to interpret the Cp value.
2. What is the theoretical process standard deviation, σ?
3. What would be the Shewhart chart limits for this system using subgroups of size $n = 4$?
4. Illustrate your answer from part 2 and 3 of this question on a diagram of the normal distribution.

Solution

1. The notes show that Cp values require us to assume that (a) the process values follow a normal distribution, the process was centered when the data were collected, and (c) that the process was stable (use a monitoring chart to verify this last assumption).
2. The range from the lower to the upper specification limit is 0.8 mm, which spans 6 standard deviations. Given the Cp value of 1.7, the process standard deviation must have been $\sigma = \dfrac{0.8}{1.7 \times 6} = 0.0784$ mm.

3. This time we have the process standard deviation, so there is no need to estimate it from historical phase I data (remember the assumption that Cp and Cpk value are calculated from stable process operation?). The Shewhart control limits would be: $\bar{\bar{x}} \pm 3 \times \frac{\sigma}{\sqrt{n}} = 2 \pm 3 \times 0.0784/2$. The LCL = 1.88 mm and the UCL = 2.12 mm.

4. An illustration is shown here with the USL, LSL, LCL and UCL, and target values. This question merely required you to show the LCL and UCL within the LSL and USL, on any normal distribution curve. However, for illustration, I have added to the diagram the distribution for the Shewhart chart (thicker line) and distribution for the raw process data (thinner line).



4 Variance and covariance (400- and 600-level students) [3]

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

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

Prove this statement, and clearly explain all steps in your proof.

Solution

I don't normally concentrate on proofs in the course, unless they show something interesting, or are used over and over. This short mathematical statement fits both criteria.

The important point with this proof is that $\bar{x}_A$ and $\bar{x}_B$ are the variables, not x. These variables come from a normal distribution (Central limit theorem), as long as we assume independent sampling: $\bar{x}_A \sim \mathcal{N}\left(\mu; \sigma^2/n_A\right)$, and similarly for $\bar{x}_B$.

$$\begin{aligned}
\mathcal{V}\{\bar{x}_B - \bar{x}_A\} &= \mathcal{V}\{\bar{x}_B + (-\bar{x}_A)\} \\
&= \mathcal{V}\{\bar{x}_B\} + 2\text{Cov}\{\bar{x}_B, (-\bar{x}_A)\} + \mathcal{V}\{-\bar{x}_A\} \\
&= \mathcal{V}\{\bar{x}_B\} + 0 + (-1)^2 \mathcal{V}\{\bar{x}_A\} \\
&= \mathcal{V}\{\bar{x}_B\} + \mathcal{V}\{\bar{x}_A\}
\end{aligned}$$

The second line is a result from the course notes. The third line requires that we assume the between-group means $\bar{x}_B$ and $\bar{x}_A$ are independent, and so they are uncorrelated (their covariance is zero). This was one of the key assumptions when we studied between-group differences; and is one assumption that is often true in many real cases.

5 Least squares modelling (400- and 600-level) [3+2+3]

The production of low density polyethylene is carried out in long, thin pipes at high temperature and pressure (1.5 kilometres long, 50mm in diameter, 500 K, 2500 atmospheres). One quality measurement of the LDPE is its melt index. Laboratory measurements of the melt index can take between 2 to 4 hours. Being able to predict this melt index, in real time, allows for faster adjustment to process upsets, reducing the product's variability. There are many variables that are predictive of the melt index, but in this example we only use a temperature measurement that is measured along the reactor's length.

These are the data of temperature (K) and melt index (units of melt index are "grams per 10 minutes").

Temperature = T [Kelvin]	441	453	461	470	478	481	483	485	499	500	506	516
Melt index = m [g per 10 mins]	9.3	6.6	6.6	7.0	6.1	3.5	2.2	3.6	2.9	3.6	4.2	3.5

The following calculations have already been performed for you:

- Number of samples, $n = 12$
- Average temperature = $\bar{T} = 481$ K
- Average melt index, $\bar{m} = 4.925$ g per 10 minutes.
- The summed product, $\sum_i \left(T_i - \bar{T}\right)\left(m_i - \bar{m}\right) = -422.1$
- The sum of squares, $\sum_i \left(T_i - \bar{T}\right)^2 = 5469.0$

1. Use this information to build a predictive linear model for melt index from the reactor temperature.
2. What is the model's standard error and how do you interpret it in the context of this model? You might find the following software software output helpful, but it is not required to answer the question.

```
Call:
lm(formula = Melt.Index ~ Temperature)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5771 -0.7372  0.1300  1.2035  1.2811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) --------    8.60936   4.885 0.000637
Temperature --------    0.01788  -4.317 0.001519

Residual standard error: 1.322 on 10 degrees of freedom
Multiple R-squared: 0.6508,     Adjusted R-squared: 0.6159
F-statistic: 18.64 on 1 and 10 DF,  p-value: 0.001519
```

3. Quote a confidence interval for the slope coefficient in the model and describe what it means. Again, you may use the above software output to help answer your question.

Solution

1. The simplest linear predictive model possible is $m = \beta_0 + \beta_1 T + \varepsilon$, predicting the melt index from temperature. Once we find estimates for these coefficients we write: $m = b_0 + b_1 T + e$. And one way to calculate these coefficients is by least squares. In the class notes we showed that for a variable x used to predict a variable y

that:

$$b_0 = \bar{y} - b_1 \bar{x}$$
$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Using the pre-calculated values, and that in our case $T = x$, and that $m = y$

$$b_1 = \frac{-422.1}{5469.0} = -0.0772 \frac{\text{g per 10 minutes}}{K}$$
$$b_0 = 4.925 + 0.0772 \times 481 = 42.0 \text{g per 10 minutes}$$

A predictive model of melt flow is: $\hat{m} = 42.0 - 0.0772 \times T$

2. The standard error, S_E can be read directly from the software output as 1.322 g per 10 minutes. If you like, you could also have calculated it by hand, using the above predictive model, calculating residuals ($e_i = m_i - \hat{m}_i$), from which the standard error is $\sqrt{\frac{\sum_i^n e_i^2}{n-k}}$, where $n = 12$ and $k = 2$ (there are 2 parameters in the model). However I recommend you always use the software output and avoid these tedious hand calculations.

   The interpretation of the standard error for this model is that the approximate prediction error of melt index has a standard deviation of 1.322 grams per 10 minutes (if the residuals are normally distributed).

3. The slope coefficient estimate, b_1 has standard error of 0.01788 (from the software output), or it could be calculated as $S_E^2(b_1) = \frac{S_E^2}{\sum_j (T_j - \bar{T})^2} = \frac{1.322^2}{5469.0} = 0.01788^2 = 3.19 \times 10^{-4}$.

   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 $n - k = 12 - 2 = 10$.

$$\begin{array}{rcccl} -c_t & \leq & \frac{b_1 - \beta_1}{S_E(b_1)} & \leq & +c_t \\ b_1 - c_t S_E(b_1) & \leq & \beta_1 & \leq & b_1 + c_t S_E(b_1) \\ -0.0772 - 2.23 \times 0.01788 & \leq & \beta_1 & \leq & -0.0772 + 2.23 \times 0.01788 \\ -0.117 & \leq & \beta_1 & \leq & -0.037 \end{array}$$

   You may also have chosen to answer at the 99% confidence level:

$$\begin{array}{rcccl} b_1 - c_t S_E(b_1) & \leq & \beta_1 & \leq & b_1 + c_t S_E(b_1) \\ -0.0772 - 3.17 \times 0.01788 & \leq & \beta_1 & \leq & -0.0772 + 3.17 \times 0.01788 \\ -0.134 & \leq & \beta_1 & \leq & -0.0205 \end{array}$$

   This shows, at which ever confidence level (95% or 99%), the range within which we can expect to find the true slope coefficient. This slope represents the magnitude by which the melt index changes, on average, for a one degree change in temperature. If we plan to manipulate the melt index using temperature, then this range will help us estimate an upper and lower bound for the effort required to adjust the melt index.

6 Final note

I know it has been pretty busy course so far with 4C3/6C3. I have been really impressed by the level of enthusiasm and high quality of work received in the assignments and take-home midterm.

The next half of this course will definitely have a lighter assignment load. Anyway, enjoy the midterm break - you won't have to think about this course again until 22 February.

END