Statistics for Engineering, 4C3/6C3, 2012

Assignment 5

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 13 February 2012, at noon** (no late handins)

Covers paired tests, process capability and least squares modelling. This assignment looks longer than it is, and it is good preparation for the midterm. Most questions are from previous midterms and exams.

Question 1 [1]

List an advantage of using a paired test over an unpaired test. Give an example, not from the 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.

Question 2 [3] (600-level students)

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.

Unpaired case:

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

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.

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

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, $\sqrt{\dfrac{79}{9}}$ *vs* the $\sqrt{86\left(\dfrac{1}{9} + \dfrac{1}{9}\right)}$ 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.

Question 3 [2]

A bagging system fills 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:

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 = $\frac{\overline{\overline{x}}-LSL}{3\sigma} = \frac{37.4-35.0}{3\times 0.8} = 1.0$
2. To obtain Cpk = 1.3 we solve the above equation for $\overline{\overline{x}} = 1.3 \times 3 \times 0.8 + 35.0 = 38.12$ 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 $\sigma = \frac{37.4-35.0}{3\times 1.3} = 0.615$ grams.

Question 4 [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:

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

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:

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

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: $\hat{m} = 42.0 - 0.0772 \times T$

1. 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).

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

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

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

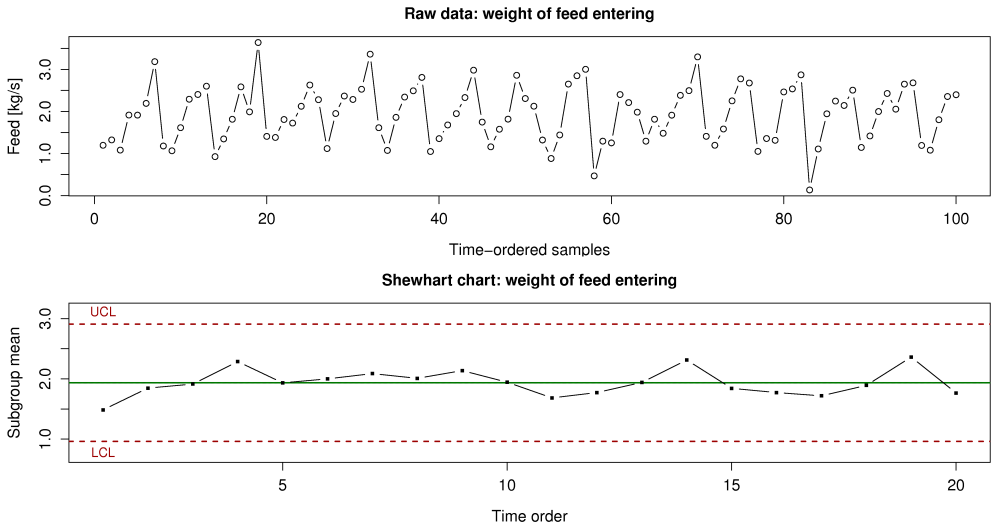
You may also have chosen to answer at the 99% confidence level:

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

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.

Question 5 [2]

The following charts show the weight of feed entering your reactor. The variation in product quality leaving the reactor was unacceptably high during this period of time.



1. What can your group of process engineers learn about the problem, using the time-series plot (100 consecutive measurements, taken 1 minute apart).
2. Why is this variability not seen in the Shewhart chart?
3. Using concepts described in this course, why might this sort of input to the reactor have an effect on the quality of the product leaving the reactor?

Solution

1. The time-series plot shows a cyclical, almost saw-tooth, pattern in the weight of feed entering. I would investigate the feeding equipment to see what is leading to these fluctuations in the feed weight. Perhaps some rotary device is responsible for the periodic variation.
2. The variability is not seen in the Shewhart monitoring chart. The Shewhart chart used subgroups of size 5 (20 Shewhart samples for 100 time-series samples). These fluctuations obviously cancel out when calculating the Shewhart subgroups (a limitation of the Shewhart chart).
3. As engineers we are aiming for stability in our processes; stability in the raw material characteristics, stability in how we operate the process over time and minimizing as many disturbances as possible. If we can do this, it will lead to greatly improved consistency in our products (low output variability). Having this sort of input to the reactor means we have to provide apply (feedback) control to counteract it. In this case the feedback control may not have been effective to eliminate the feed variation, or the feedback control itself caused other disruptions to the process quality.

Question 6 [3]

For a distillation column, it is well known that the column temperature directly influences the purity of the product, and this is used in fact for feedback control, to achieve the desired product purity. Use the distillation data set , and build a least squares model that predicts `VapourPressure` from the temperature measurement, `TempC2`. Report the following values:

1. the slope coefficient, and describe what it means in terms of your objective to control the process with a feedback loop
2. the interquartile range and median of the model's residuals
3. the model's standard error
4. a confidence interval for the slope coefficient, and its interpretation.

You may use any computer package to build the model and read these values off the computer output.

Solution

The solution to this question can be almost entirely solved using R, though any other language could be used. These commands, with the output that follows, were used:

```
> distillation <- read.csv('http://datasets.connectmv.com/file/distillation-tower.csv')
> model <- lm(distillation$VapourPressure ~ distillation$TempC2)
> summary(model)

Call:
lm(formula = distillation$VapourPressure ~ distillation$TempC2)

Residuals:
     Min       1Q   Median       3Q      Max
-5.59621 -2.37597  0.06674  2.00212 14.18660

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         195.96141    4.87669   40.18   <2e-16 ***
distillation$TempC2  -0.33133    0.01013  -32.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.989 on 251 degrees of freedom
Multiple R-squared: 0.8098,     Adjusted R-squared: 0.8091
F-statistic:  1069 on 1 and 251 DF,  p-value: < 2.2e-16
```

1. This predictive model allows us to achieve better control of the vapour pressure, because we can predict it from temperature (measured in real-time), rather than wait several hours for the laboratory vapour pressure value. The slope coefficient is -0.331, and since no units were given, I can't expect any in your solution; however one should report the units, which is this case would be units of pressure divided by units temperature (e.g. psi/K). What this means, in terms of feedback control of the vapour pressure is that we must decrease the temperature to raise the vapour pressure. This is important when tuning the feedback control loop in 2 ways: (a) firstly, the the sign of the gain in the feedback controller (i.e. negative gain) must be the same as the process gain to achieve a stable feedback loop, (b) the magnitude of the slope provides an estimate of how sensitive the vapour pressure is to temperature. For example: do we have to add a large amount of energy into the distillation column to achieve a smallish reduction in vapour pressure? The answer depends heavily on the units, which I omitted to provide.
2. These are reported in the above software output: (a) the residual IQR is 2.00 - (-2.38) = 4.38 units of vapour pressure, while (b) the median residual is close to zero, as expected.
3. The model's standard error is 2.989 in the output, or around 3.00 units of vapour pressure.
4. The slope coefficient's confidence interval can be calculated from its z-value = $\dfrac{b_1 - \beta_1}{S_E(b_1)}$; but we require the standard error of the slope coefficient, which is $S_E(b_1) = 0.01013$ from the software output. The value for $c_t = 1.969$ from the t-distribution at the 95% confidence level, with $n - k = 253 - 2 = 251$ degrees of freedom (a normal distribution would work equally well in this case).

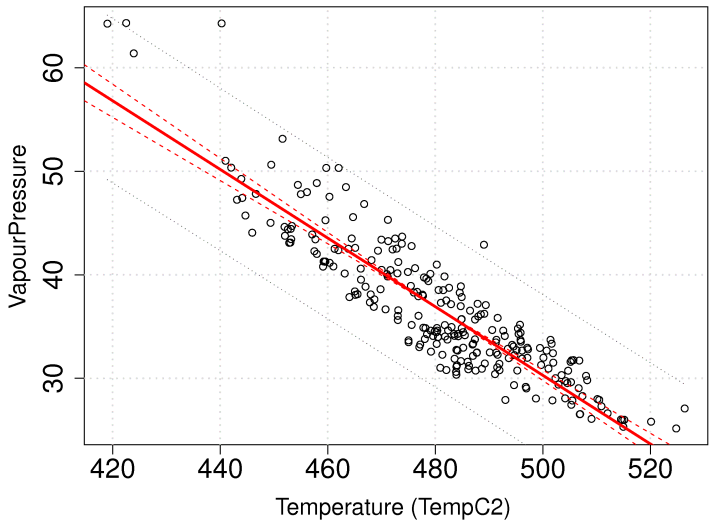
$$\begin{aligned} -c_t &\leq \frac{b_1 - \beta_1}{S_E(b_1)} \leq +c_t \\ -0.33133 - 1.969 \times 0.01013 &\leq \beta_1 \leq -0.33133 + 1.969 \times 0.01013 \\ -0.35 &\leq \beta_1 \leq -0.31 \end{aligned}$$

   This shows, at the 95% confidence level, the range within which we can expect to find the true slope coefficient. This range is remarkably narrow; i.e. our feedback controller gain is unlikely to change on either extreme. So we can likely design our control loop at the center point, and be sure it will work over the entire range of

expected operation. Please also cross reference the solutions to question 2.4 in the written midterm to correctly understand what a confidence interval is.

If you used 99% confidence levels, the answer should be: $-0.358 \leq \beta_1 \leq -0.305$.

We have illustrated the actual slope (thick, solid line) at the upper and lower bounds of the slope coefficient (thin, dashed lines) in the accompanying figure. Not required for this question, but added nevertheless, are the prediction intervals for $\hat{y}_i$.



I recommended that you reproduce R's output yourself. The code below calculates these same values.

```
# Calcuations with R

distillation <- read.csv('http://datasets.connectmv.com/file/distillation-tower.csv')
model <- lm(distillation$VapourPressure ~ distillation$TempC2)
summary(model)

# Calculations by hand
# ------------------------

# Confidence level
alpha = 0.99

# Raw data
x <- distillation$TempC2
y <- distillation$VapourPressure
n = length(x)

# Some intermediate values
x.bar = mean(x)
y.bar = mean(y)
num <- sum((x - x.bar) * (y - y.bar))
den <- sum((x - x.bar) * (x - x.bar))

# Model coefficients
b1 <- num/den
b0 <- y.bar - b1 * x.bar
c(b0, b1)

# Model predictions and residuals, with their summary (IQR and median)
predictions <- b0 + x*b1
residuals <- y - predictions
```

```
summary(residuals)

# Calculate the 3 standard errors
SE <- sqrt(sum(residuals^2) / (n-2))
SE.b1 <- sqrt(SE^2 / den)
SE.b0 <- sqrt(SE^2 *(1/n + (x.bar^2)/den))
c(SE, SE.b0, SE.b1)

# Confidence intervals for the least squares parameters
z.b0 = b0/SE.b0
z.b1 = b1/SE.b1
c(z.b0, z.b1)
t.critical = qt(1-(1-alpha)/2, df=(n-2))
t.critical
b0.LB <- b0 - t.critical*SE.b0
b0.UB <- b0 + t.critical*SE.b0
b1.LB <- b1 - t.critical*SE.b1
b1.UB <- b1 + t.critical*SE.b1
c(b0.LB, b0.UB)
c(b1.LB, b1.UB)

# R2, TSS, RegSS, RSS, Adjusted R2
TSS <- sum((y - y.bar)^2)
RegSS <- sum((predictions-y.bar)^2)
RSS <- sum(residuals^2)
R2 <- RegSS/TSS
RSS.adj <- 1- (RSS/(n-2)) / (TSS/(n-1))
c(TSS, RegSS, RSS, R2, RSS.adj)

# Error bounds for y-hat
x.new = seq(min(x), max(x), diff(range(x))/100)
y.new = b0 + x.new*b1
error.delta <- t.critical*SE*sqrt(1+ 1/n + ((x.new-x.bar)^2)/den)

# Plot of the raw data, least squares line, prediction interval for yhat,
# slope coefficient confidence interval range
plot(x,y, cex.lab=1.5, cex.main=1.8, cex.sub=1.8, cex.axis=1.8, main="",
          xlab="Temperature (TempC2)", ylab="VapourPressure")
grid(lwd=2)
points(x, y)
lines(x.new, y.new + error.delta, col="gray40", lty=3)
lines(x.new, y.new - error.delta, col="gray40", lty=3)

abline(a=b0, b=b1, col="red", lty=1, lwd=3)

# One extreme of the beta_1 slope CI
abline(a=(y.bar-b1.UB*x.bar), b=b1.UB, col="red", lty=2, lwd=1)
# Other extreme of the beta_1 slope CI
abline(a=(y.bar-b1.LB*x.bar), b=b1.LB, col="red", lty=2, lwd=1)
```

END