Statistics for Engineers, 4C3/6C3

Assignment 4

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 07 February 2011**

Assignment objectives

- Be comfortable using distributions in R.
- Using unpaired and paired tests via confidence intervals.
- Construct and use Shewhart process monitoring charts

Question 1 [1.5]

In the previous assignment you collected the snowfall and temperature data for the `HAMILTON A` weather station. Here are the data again:

- 1990 to 1999 snowfall: $[131.2, 128.0, 130.7, 190.6, 263.4, 138.0, 207.3, 161.5, 78.8, 166.5]$
- 2000 to 2008 snowfall: $[170.9, 94.1, 138.0, 166.2, 175.8, 218.4, 56.6, 182.4, 243.2]$
- 1990 to 2000 temperature: $[8.6, 8.6, 6.9, 7.1, 7.1, 7.7, 6.9, 7.3, 9.8, 8.8]$
- 2000 to 2008 temperature: $[7.6, 8.8, 8.8, 7.3, 7.7, 8.2, 9.1, 8.2, 7.7]$

1. Use these data to construct a z-value *and* confidence interval on the assumption that the snowfall in the earlier decade (case A) is statistically the same as in 2000 to 2008 (case B).
2. Repeat this analysis for the average temperature values.
3. Do these, admittedly limited, data support the conclusion that many people keep repeating: the amount of snow we receive is less than before and that temperatures have gone up?
4. In the above analysis you had to pool the variances. There is a formal statistical test, described in the course notes, to verify whether the variances could have come from the same population:

$$F_{\alpha/2,\nu_1,\nu_2}\frac{s_2^2}{s_1^2} \quad < \quad \frac{\sigma_2^2}{\sigma_1^2} \quad < \quad F_{1-\alpha/2,\nu_1,\nu_2}\frac{s_2^2}{s_1^2}$$

where we use $F_{\alpha/2,\nu_1,\nu_2}$ to mean the point along the cumulative F-distribution which has area of $\alpha/2$ using ν_1 degrees of freedom for estimating s_1 and ν_2 degrees of freedom for estimating s_2. For example, in R, the value of $F_{0.05/2,10,20}$ can be found from `qf(0.025, 10, 20)` as 0.2925. The point along the cumulative F-distribution which has area of $1-\alpha/2$ is denoted as $F_{1-\alpha/2,\nu_1,\nu_2}$, and α is the level of confidence, usually $\alpha = 0.05$ to denote a 95% confidence level.

Confirm that you can pool the variances in both the snowfall and temperature case by verifying the confidence interval contains a value of 1.0.

> **Note:** The equations in the printed course notes you downloaded, or bought from Titles, are wrong; the above version is correct. Please update your printouts.

Solution (Thanks to Ryan and Stuart)

1 and 2.

Unpaired t-tests were performed on the snowfall datasets from 1990-1999 and 2000-2008 to test the assumption that the amount of snowfall and average temperature has not changed significantly over the last two decades. In order to perform these tests, the following assumptions were made:

- The variances of both samples are comparable
- Independence within each sample and between the sample
- Both samples are normally distributed

The z-value for this test was constructed as follows:

$$z = \frac{(\overline{x}_B - \overline{x}_A) - (\mu_B - \mu_A)}{\sqrt{\sigma^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{(\overline{x}_B - \overline{x}_A)}{\sqrt{\sigma^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Since an external estimate of variance was not available, the estimated variances from each data set were pooled to create an internal estimate using the following formula:

$$s_P^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A - 1 + n_B - 1}$$

Using this internal estimator:

$$z = \frac{(\overline{x}_B - \overline{x}_A) - (\mu_B - \mu_A)}{\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{(\overline{x}_B - \overline{x}_A)}{\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

which follows the t-distribution with $(n_A + n_B - 2) = 17$ degrees of freedom.

Unpacking this z-value, confidence intervals were constructed at the 95% confidence level as follows:

$$\begin{array}{rcccl} c_{t,0.025,17} & \leq & z & \leq & c_{t,0.975,17} \\ (\overline{x}_B - \overline{x}_A) - 2.109816\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} & \leq & \mu_B - \mu_A & \leq & (\overline{x}_B - \overline{x}_A) + 2.109816\sqrt{s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \end{array}$$

The results for the **snowfall** data set:

- $s_P^2 = 2968$
- z-value = 0.0408
- CI is $-51.8 \leq \mu_B - \mu_A \leq 53.8$ in units of centimetres of snow

and for the **temperature** data set:

- $s_P^2 = 0.7211$
- z-value = 0.706
- CI is $-0.55 \leq \mu_B - \mu_A \leq 1.1$ in units of Celcius

3. As the z-values for both t-tests fall within ± 2.11 and the 95% confidence intervals contain zero, the data provided does not support the notion that we receive less snow or that the temperature has risen over the last two decades. i.e. change over the last two decades is not significantly different at the 95% confidence level. In fact, the unpaired t-test on the total snowfall actually indicates that the means of the total snowfall data for the two decades are identical at the 95% confidence level.

4. An F-test was performed on the weather data to test the assumption that the variances for both samples are comparable. The confidence interval for the F-test was constructed on the ratio of sample variances as follows:

$$F_{\alpha/2,\nu_1,\nu_2}\frac{s_2^2}{s_1^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\alpha/2,\nu_1,\nu_2}\frac{s_2^2}{s_1^2}$$

where ν_1 and ν_2 are the degrees of freedom used to compute s_1 and s_2, respectively, i.e. $\nu_1 = n_A - 1$ and $\nu_2 = n_B - 1$.\

Evaluating this expression at the 95% confidence level for both datasets:

Snowfall: $0.3097 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 5.535$

Average temperature: $0.0962 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 1.719$

Assuming that the variances for both samples are from the same population entails that the ratio of population variances is 1. Therefore, as the 95% confidence intervals from the F-test contain a value of 1 for both datasets, the assumption that the variances are comparable is statistically valid. Hence, the F-test supports the pooling of variances for use in the unpaired t-test for both datasets (i.e. it was not incorrect to do so in parts 1 and 2).

```
# Raw data
Snowfall_1 <- c(131.2,128.0,130.7,190.6,263.4,138.0,207.3,161.5,78.8,166.5)
Snowfall_2 <-c(170.9,94.1,138.0,166.2,175.8,218.4,56.6,182.4,243.2)
Temp_1 <- c(8.6,8.6,6.9,7.1,7.1,7.7,6.9,7.3,9.8,8.8)
Temp_2 <- c(7.6,8.8,8.8,7.3,7.7,8.2,9.1,8.2,7.7)

# 95% Confidence interval for unpaired testing
alpha <- 0.05

unpaired_ttest <- function(samp1,samp2,alpha){

        meanA <- mean(samp1)
        meanB <- mean(samp2)

        numA <- length(samp1)
        numB <- length(samp2)

        varA <- var(samp1)
        varB <- var(samp2)

        DOF <- numA-1+numB-1

        Pvar <- ((numA-1)*varA+(numB-1)*varB)/DOF

        z <- (meanB-meanA)/sqrt(Pvar*(1/numA+1/numB))
        if (z>0){prob <- 2*(1-pt(z,df=DOF))} else {prob <- 2*pt(z,df=DOF)}

        LCB <- (meanB-meanA)-qt(1-alpha/2,df=DOF)*sqrt(Pvar*(1/numA+1/numB))
        UCB <- (meanB-meanA)+qt(1-alpha/2,df=DOF)*sqrt(Pvar*(1/numA+1/numB))

        list(numA=numA, MeanA = meanA, varA=varA,
             numB=numB, MeanB = meanB, varB=varB,
             DOF=DOF, PooledVariance=Pvar, zvalue=z,
             prob=prob, LCB=LCB,UCB=UCB)
}

Snowfall_Comparison <- unpaired_ttest(Snowfall_1,Snowfall_2,alpha)
qt(0.025,Snowfall_Comparison$DOF)

Temp_Comparison <- unpaired_ttest(Temp_1,Temp_2,alpha)
```

```
var_ftest <- function(samp1,samp2,alpha){

        numA <- length(samp1)
        numB <- length(samp2)

        varA <- var(samp1)
        varB <- var(samp2)

        dofA <- numA-1
        dofB <- numB-1

        LCB <- qf(alpha/2,dofA,dofB)*varB/varA
        UCB <- qf(1-alpha/2,dofA,dofB)*varB/varA

        list(varA=varA, DOFA = dofA,varB=varB,DOFB=dofB,LCB=LCB,UCB=UCB)
}

Snowfall_varTest <- var_ftest(Snowfall_1,Snowfall_2,alpha)
Temp_varTest <- var_ftest(Temp_1,Temp_2,alpha)
```

Question 2 [1.5]

The percentage yield from a batch reactor, and the purity of the feedstock are available as the Batch yield and purity data set. Assume these data are from phase I operation and calculate the Shewhart chart upper and lower control limits that you would use during phase II. Use a subgroup size of $n = 3$.

1. What is phase I?
2. What is phase II?
3. Show your calculations for the upper and lower control limits for the Shewhart chart on the *yield value*.
4. Show a plot of the Shewhart chart on these phase I data.

Solution (Thanks to Ryan, Stuart and Mudassir)

1. Phase 1 is the period from which historical data is taken that is known to be "in control". From this data, upper and lower control limits can be established for the monitored variable that contain a specified percent of all in control data.
2. Phase 2 is the period during which new, unseen data is collected by process monitoring in real-time. This data can be compared with the limits calculated from the "in control" data.
3. Assuming the dataset was derived from phase I operation, the batch yield data was grouped into subgroups of size 3. However, since the total number of data points (N=241) is not a multiple of three, the data set was truncated to the closest multiple of 3, i.e. $N_{new} = 240$, by removing the last data point. Subsequently, the mean and standard deviation were calculated for each of the 80 subgroups. From this data, the lower and upper control

limits were calculated as follows:

$$\overline{\overline{x}} = \frac{1}{80}\sum_{k=1}^{80}\overline{x}_k = \mathbf{75.3}$$

$$\overline{S} = \frac{1}{80}\sum_{k=1}^{80}s_k = \mathbf{5.32}$$

$$\text{LCL} = \overline{\overline{x}} - 3\cdot\frac{\overline{S}}{a_n\sqrt{n}} = \mathbf{64.9}$$

$$\text{UCL} = \overline{\overline{x}} + 3\cdot\frac{\overline{S}}{a_n\sqrt{n}} = \mathbf{85.7}$$

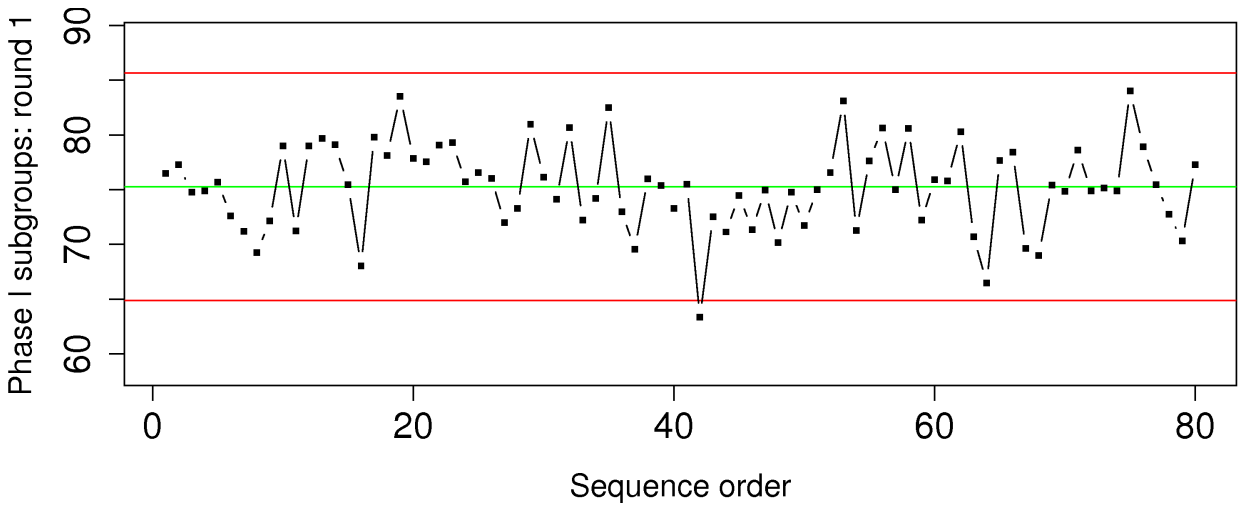
$$\text{using } a_n = 0.886 \qquad \text{for a subgroup size of } 3$$

$$\text{and } \overline{\overline{x}} = 75.3$$

Noticing that the mean for subgroup 42, $\overline{x}_{42} = 63.3$, falls below this LCL, the control limits were recalculated excluding this subgroup from phase I data (see R-code). Following this adjustment, the new control limits were calculated to be:

- LCL = 65.0
- UCL = 85.8

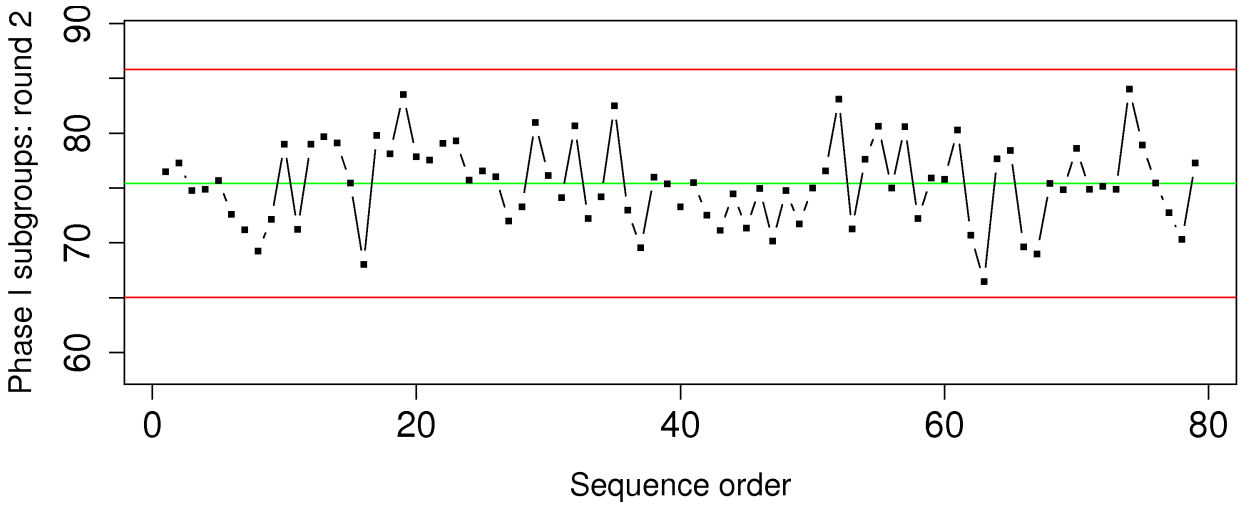
4. Shewhart charts for both rounds of the yield data (before and after removing the outlier):




```
# Thanks to Mudassir for letting me use his recursive source code
# I've made some small updates
# ------------------------------------------------------------------
data <- read.csv('http://datasets.connectmv.com/file/batch-yield-and-purity.csv')
y <- data$yield
variable <- "Yield"
N <- 3

# No further changes required: the code below should work for
# any new data set now
subgroups <- matrix(y, N, length(y)/N)
x.mean <- numeric(length(y)/N)
x.sd <- numeric(length(y)/N)

# Calculate mean and sd of subgroups (see R-tutorial)
x.mean <- apply(subgroups, 2, mean)
```




```
x.sd <- apply(subgroups, 2, sd)
ylim <- range(x.mean) + c(-5, +5)
k <- 1
doloop <- TRUE
# Prevent infinite loops
while (doloop & k < 5){
  # Original definition for a_n: see course notes
  an <- sqrt(2)*gamma(N/2)/(sqrt(N-1)*gamma((N-1)/2))
  S <- mean(x.sd)
  xdb <- mean(x.mean) # x-double bar
  LCL <- xdb - (3*S/(an*sqrt(N)))
  UCL <- xdb + (3*S/(an*sqrt(N)))
  print(c(LCL, UCL))

  # Create a figure on every loop
  bitmap(paste('../images/batch-phaseI-round-', k, '-', variable, '.png',  sep=""),
         type="png256", width=10, height=4, res=300, pointsize=14)
  par(mar=c(4.2, 4.2, 0.5, 0.5))
  par(cex.lab=1.3, cex.main=1.5, cex.sub=1.5, cex.axis=1.5)
  plot(x.mean, type="b", pch=".", cex=5, main="",
       ylab=paste("Phase I subgroups: round", k),
       xlab="Sequence order", ylim=ylim)
  abline(h=UCL, col="red")
  abline(h=LCL, col="red")
  abline(h=xdb, col="green")
  lines(x.mean, type="b", pch=".", cex=5)
  dev.off()

  if (!( any(x.mean<LCL) | any(x.mean>UCL) )){
    # Finally!  No more points to exclude
    doloop <- FALSE
  }
  k <- k + 1

  # Retain in x.sd and x.mean only those entries
  # that are within the control limits
  x.sd <- x.sd[x.mean>=LCL]
  x.mean <- x.mean[x.mean>=LCL]
  x.sd <- x.sd[x.mean<=UCL]
  x.mean <- x.mean[x.mean<=UCL]
} # end: while doloop
```

Question 3 [2]

You want to evaluate a new raw material (B), but the final product's brittleness, the main quality variable, must be the same as achieved with the current raw material. Manpower and physical constraints prevent you from running a randomized test, and you don't have a suitable database of historical reference data either.

One idea you come up with is to use to your advantage the fact that your production line has three parallel reactors, TK104, TK105, and TK107. They were installed at the same time, they have the same geometry, the same instrumentation, *etc*; you have pretty much thought about every factor that might vary between them, and are confident the 3 reactors are identical.

This means that when you do your testing on the new material next week you can run test A using one reactor and test B in another reactor, if you can find the two reactors that have *no statistical difference* in operation.

Normal production splits the same raw material between the 3 reactors. Data on the website contain the brittleness values from the three reactors for the past few runs using the current raw material (A).

Using a series of paired tests, calculate which two reactors you would pick to run your comparative trial on. Be *very specific and clearly substantiate why* you have chosen your 2 reactors.

Solution

Pairing assumes that each reactor was run with the same material, except that the material was split into thirds: one third for each reactor. As described in the section on paired tests we rely on calculating the difference in brittleness, then calculating the z-value of the average difference. Contrast this to the unpaired tests, where we calculated the difference of the averages.

The code below shows how the paired differences are evaluated for each of the 3 combinations. The paired test highlights the similarity between TK105 and TK107, (the same result if you used an unpaired test - you should verify that). However the paired test shows much more clearly how different tanks TK104 and TK105 are, and especially TK104 and TK107.

In the case of TK104 and TK105 the difference might seem surprising - take a look back at the box plots (type `boxplot(brittle)` into R) and how much they overlap. However a paired test cannot be judged by a box plot, because it looks at the case-by-case difference, not the overall between group difference. A better plot with which to confirm the really large z-value for the TK105 and TK107 difference is the plot of the differences.

```
brittle <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')

boxplot(brittle)
# Calculates the paired difference
paired_difference <- function(groupA, groupB, alpha=0.95)
{
    # This function assumes either group has missing data.
    # Find the subset of observations in common.

    groupA.sub <- groupA[!is.na(groupA) & !is.na(groupB)]
    groupB.sub <- groupB[!is.na(groupA) & !is.na(groupB)]

    diffs <- groupB.sub - groupA.sub
    diffs.mean <- mean(diffs)
    diffs.sd <- sd(diffs)
    diffs.N <- length(diffs)

    plot(groupB.sub-groupA.sub, type="b")

    z <- (diffs.mean - 0) / (diffs.sd/sqrt(diffs.N))
    t.critical <- pt(z, df=(diffs.N-1))
    c.t <- qt(1-(1-alpha)/2, df=(diffs.N-1))
```

```
    LB <- diffs.mean  - c.t * diffs.sd / sqrt(diffs.N)
    UB <- diffs.mean  + c.t * diffs.sd / sqrt(diffs.N)

return(list(z, t.critical, diffs.N-1, LB, UB))
}

paired_difference(brittle$TK104, brittle$TK105, alpha=0.95)
# (z=2.64, t.critical=0.991, DOF=17, LB=9.81, UB=88.4)

paired_difference(brittle$TK104, brittle$TK107, alpha=0.95)
# (z=12, t.critical=1, DOF=19, LB=48.3, UB=68.7)

paired_difference(brittle$TK105, brittle$TK107, alpha=0.95)
# (z=-0.33, t.critical=0.37, DOF=20, LB=-46.1, UB=33.5)
```

From the above code we get the 95% confidence intervals as:

$$\begin{aligned} 9.81 &\leq \mu_{105-104} \leq 88.4 \\ 48.3 &\leq \mu_{107-104} \leq 68.7 \\ -46.1 &\leq \mu_{107-105} \leq 33.5 \end{aligned}$$

Now onto the most important part of any statistical analysis: interpreting the results and making a decision.

We can clearly see that TK104 is very different from TK105 and TK107 at the 95% confidence level, because these confidence intervals *do not* include zero (their z-values are very large positive or large negative values).

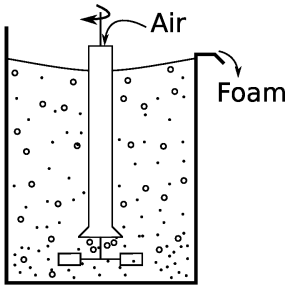
The most similar reactors are TK105 and TK107, because this confidence interval *for the difference* spans zero, and it does this nearly symmetrically, from -46 up to +33, so the risk that this CI was found to span zero due to only a subset of the data is minimal. In fact, a plot of the differences show several large and several small differences.

So you would naturally conclude that the trial should be conducted in reactors TK105 and TK107. However a contrarian point of view holds that you should conduct the trial in TK104 and TK107. This is the confidence interval with the smallest span, i.e. it is the "tightest confidence interval". It interpretation says, well I recognize there is a difference, but I can reliably predict that difference to be only 20 units wide. So I can do my tests in TK104 and TK107, then just subtract off the bias of 20 units. Any tests done in TK105 and TK107 though, should have no *statistically significant* difference, but this confidence limit spans 33+46=79 units, 4 times greater.

My recommendation would be to use TK104 and TK107; however if you answered TK105 and TK107, I will accept that as an answer. But this question should make you realize that most statistical analyses are not clear cut, and you always need to ask what is the engineering significance and implication of your results.

Question 4 [2]

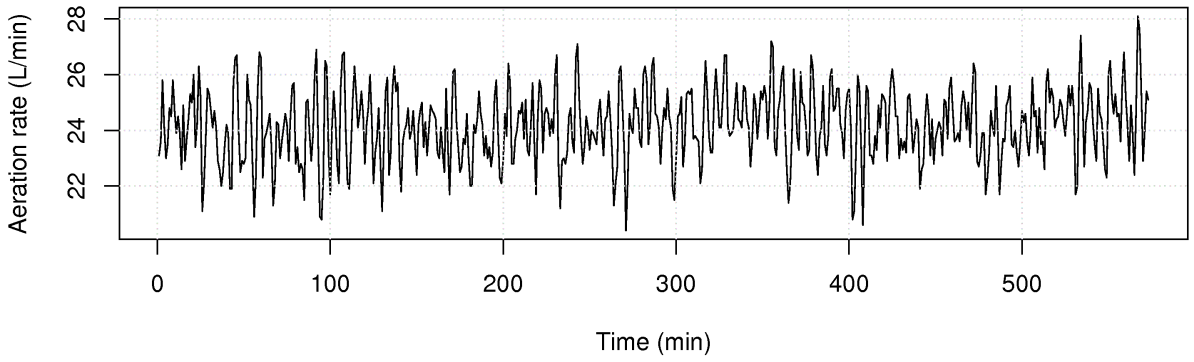
A tank uses small air bubbles to keep solid particles in suspension. If too much air is blown into the tank, then excessive foaming and loss of valuable solid product occurs; if too little air is blown into the tank the particles sink and drop out of suspension.



1. Which monitoring chart would you use to ensure the airflow is always near target?
2. Use the aeration rate dataset from the website and plot the raw data (total litres of air added in a 1 minute period). Are you able to detect any problems?
3. Construct the chart you described in part 1, and show it's performance on all the data. Make any necessary assumptions to construct the chart.
4. At what point in time are you able to detect the problem, using this chart?
5. Construct a Shewhart chart, choosing appropriate data for phase I, and calculate the Shewhart limits. Then use the entire dataset as if it were phase II data.
    - Show this phase II Shewhart chart.
    - Compare the Shewhart chart's performance to the chart in part 3 of this question.

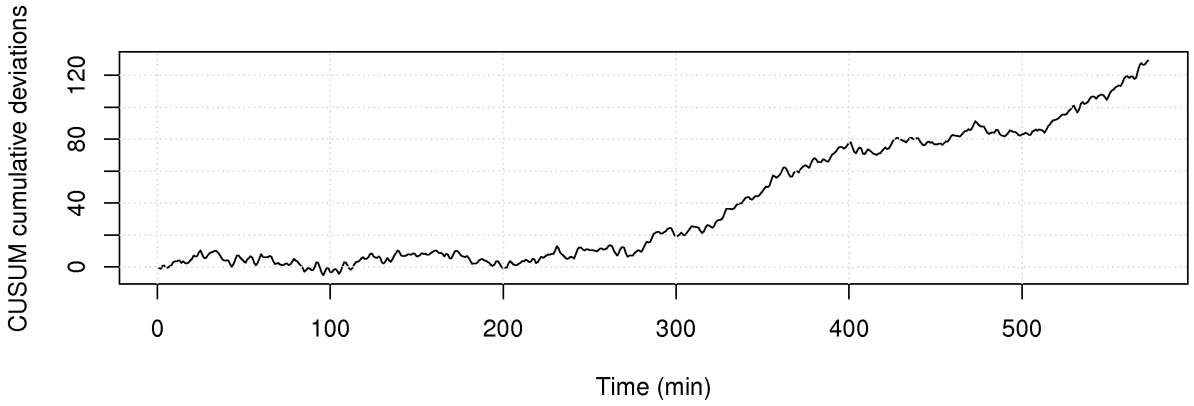
Solution (thanks to Ryan and Stuart)

1. A CUSUM chart would be the most appropriate monitoring chart to ensure the airflow is always near the intended target. A EWMA chart could also be used for the same purpose, but the value of λ would have to be set fairly low (i.e. long memory) such that the EWMA would approximate the CUSUM.
2. The aeration rate dataset is depicted below:



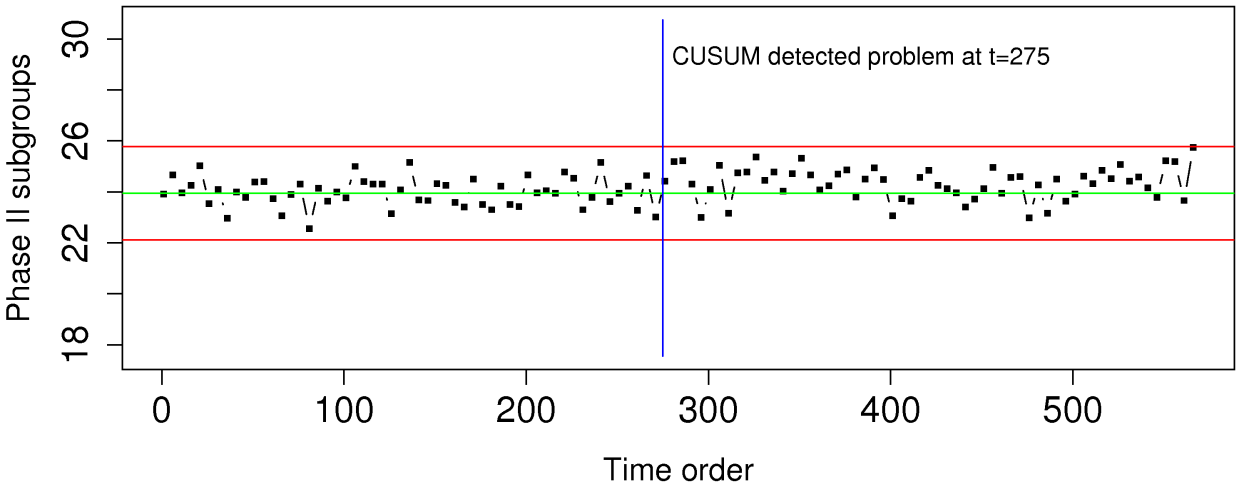
   It is very difficult to assess problems from the raw data plot. There might be a slight upward shift around 300 and 500 minutes.
3. Assumptions for the CUSUM chart:
    - We will plot the CUSUM chart on raw data, though you could use subgroups if you wanted to.
    - The target value can be the mean (24.17) of all the data, or more robustly, use the median (24.1), especially if we expect problems with the raw data (true of almost every real data set).

   The CUSUM chart, using the median as target value showed a problem starting to occur around $t = 300$. So we recalculated the median, using only data from 0 to $t = 200$, to avoid biasing the target value. Using this median instead, 23.95, we get the following CUSUM chart:
4. The revised CUSUM chart suggests that the error occurs around 275 min, as evidenced by the steep positive slope thereafter. It should be noted that the CUSUM chart begins to bear a positive slope around 200 min, but this initial increase in the cumulative error would likely not be diagnosable (i.e. using a V-mask).
5. Using the iterative Shewhart code from the previous question, we used



- Phase I was taken far enough away from the suspected error: 0 - 200 min
- Subgroup size of $n = 5$
- $\bar{\bar{x}} = 23.9$
- $\bar{S} = 1.28$
- $a_n = 0.940$
- LCL = $23.9 - 3 \cdot \frac{1.28}{0.940\sqrt{5}} = 22.1$
- UCL = $23.9 + 3 \cdot \frac{1.28}{0.940\sqrt{5}} = 25.8$

The Shewhart chart applied to the entire dataset is shown below. In contrast to the CUSUM chart, the Shewhart chart is unable to detect the problem in the aeration rate. Unlike the CUSUM chart, which has infinite memory, the Shewhart chart has no memory and cannot adequately assess the location of the monitored variable in relation to its specified target. Instead, the Shewhart chart merely monitors aeration rate with respect to the control limits for the process. Since the aeration rate does not exceed the control limits for the process (i.e. process remains in control), the Shewhart chart does not detect any abnormalities.



If you used the Western Electric rules, in addition to the Shewhart chart limits, you would have picked up a consecutive sequence of 8 points on one side of the target around $t = 350$.

Question 5 [1.5]

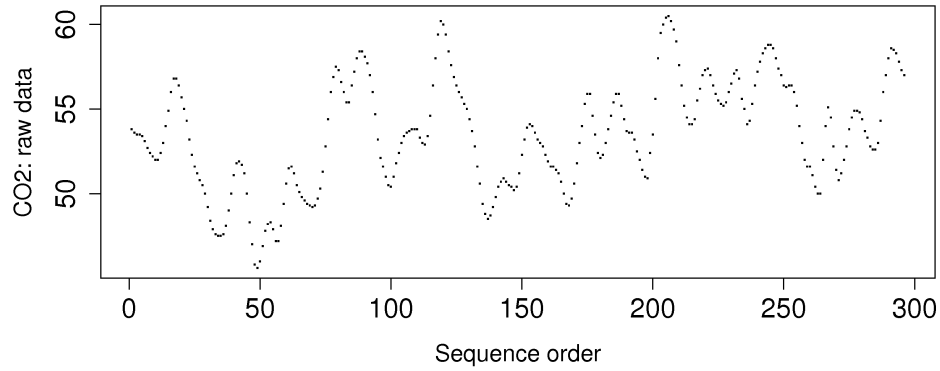
Note: For 600-level students

The carbon dioxide measurement is available from a gas-fired furnace. These data are from phase I operation.

1. Calculate the Shewhart chart upper and lower control limits that you would use during phase II with a subgroup size of $n = 6$.
2. Is this a useful monitoring chart? What is going in this data?
3. How can you fix the problem?

Solution (thanks to Ryan and Stuart)

First a plot of the raw data will be useful:



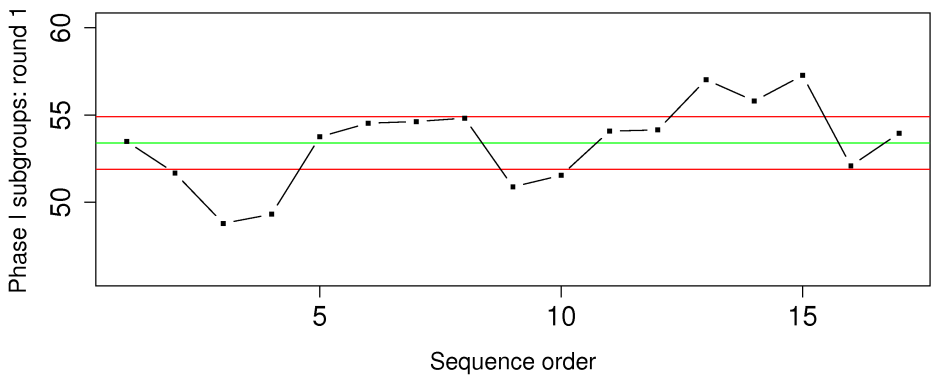
1. Assuming that the CO_2 data set is from phase I operation, the control limits were calculated as follows:
   - Assume subgroups are independent
   - $\bar{\bar{x}} = \frac{1}{K}\sum_{k=1}^{K} \bar{x}_k = 53.5$
   - $\bar{S} = \frac{1}{K}\sum_{k=1}^{K} s_k = 1.10$
   - $a_n = 0.952$
   - LCL = $53.5 - 3 \cdot \frac{1.10}{0.952\sqrt{6}} = 53.5$
   - UCL = $53.5 + 3 \cdot \frac{1.10}{0.952\sqrt{6}} = 54.0$
2. The Shewhart chart, with a subgroup of 6, is not a useful monitoring chart. There are too many false alarms, which will cause the operators to just ignore the chart.

   The problem is that the first assumption of independence is not correct. As shown in the previous assumption

   > The raw data show that the subgroups will be related; in this case $n = 6$ and any 6 points side-by-side will still have a relationship between them.

3. One approach to fixing the problem is to subsample the data, i.e. only use every k^{th} data point as the raw data, e.g. $k = 10$, and then form subgroups from that sampled data.

   Another is to use a larger subgroup size. We will introduce a method later on that can be used to verify the degree of relationship: the autocorrelation function, and the corresponding `acf(...)` function in R. Using



this function we can see the raw data are unrelated average the 17th lag, so we could subgroups of that size. However, even then we see the Shewhart chart showing frequent violation, though fewer than before.

Yet another alternative is to use an EWMA chart, which takes the autocorrelation into account. However, the EWMA chart limits are found from the assumption that the subgroup means (or raw data, if subgroup size is 1), are independent.

So we are finally left with the conclusion that perhaps there data really are not from in control operation, or, if they are, we must manually adjust the limits to be wider.

```
data <- read.csv('http://datasets.connectmv.com/file/gas-furnace.csv')
CO2 <- data$CO2
N.raw <- length(CO2)
N.sub <- 6   # change this value to 10, 15, 17, 20, etc

# Plot all the data
bitmap('../images/CO2-raw-data-assign4.png', type="png256",
       width=10, height=4, res=300, pointsize=14)
par(mar=c(4.2, 4.2, 0.5, 0.5))
par(cex.lab=1.3, cex.main=1.5, cex.sub=1.5, cex.axis=1.5)
plot(CO2, type="p", pch=".", cex=2, main="",
     ylab="CO2: raw data", xlab="Sequence order")
dev.off()

# Create the subgroups on ALL the raw data.  Form a matrix with
# 'N.subgroup' rows by placing the vector of data down each row,
# then going across to form the columns.

# Calculate the mean and standard deviation within each subgroup
# (columns of the matrix)
subgroups <- matrix(CO2, N.sub, N.raw/N.sub)
subgroups.S <- apply(subgroups, 2, sd)
subgroups.xbar <- apply(subgroups, 2, mean)
ylim <- range(subgroups.xbar) + c(-3, +3)

# Keep adjusting N.sub until you don't see any autocorrelation
# between subgroups
acf(subgroups.xbar)

# Create a function to calculate Shewhart chart limits
shewhart_limits <- function(xbar, S, sub.size, N.stdev=3){
    # Give the xbar and S vector containing the subgroup means
    # and standard deviations.  Also give the subgroup size used.
    # Returns the lower and upper control limits for the Shewhart
    # chart (UCL and LCL) which are N.stdev away from the target.
```

```
    x.double.bar <- mean(xbar)
    s.bar <- mean(S)
    an <- sqrt(2)*gamma(sub.size/2)/(sqrt(sub.size-1)*gamma((sub.size-1)/2))
    LCL <- x.double.bar - 3*s.bar/an/sqrt(sub.size)
    UCL <- x.double.bar + 3*s.bar/an/sqrt(sub.size)
return(list(LCL, x.double.bar, UCL))
}
limits <- shewhart_limits(subgroups.xbar, subgroups.S, N.sub)
LCL <- limits[1]
xdb <- limits[2]
UCL <- limits[3]
c(LCL, xdb, UCL)

# Any points outside these limits?
bitmap('../images/CO2-phaseI-first-round-assign4.png', type="png256",
       width=10, height=4, res=300, pointsize=14)
par(mar=c(4.2, 4.2, 0.5, 0.5))
par(cex.lab=1.3, cex.main=1.5, cex.sub=1.5, cex.axis=1.5)
plot(subgroups.xbar, type="b", pch=".", cex=5, main="", ylim=ylim,
    ylab="Phase I subgroups: round 1",  xlab="Sequence order")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
abline(h=xdb, col="green")
lines(subgroups.xbar, type="b", pch=".", cex=5)
dev.off()
```

Something to think about

Being RRSP season, it is tempting to start buy and selling stock, mutual funds and exchange traded funds (ETFs). One issue faced by any investor is when is a good time to buy or to sell.

Using the tools of process monitoring, think about you can use control limits to decide when to sell a poorly performing stock (going below the LCL?) and when to buy a weak stock that is strengthening (going up, over the UCL). One issue with stock prices is of course the lack of independence between the daily prices (form subgroups!).

But using the concepts of process monitoring you can devise a trading strategy that prevents trading too frequently (rapid buying and selling), as well as from selling/buying when there is just "common cause" fluctuations in the data. The control limits are set based on your personal level of risk.

You should always verify your trading strategies with historical data, and Yahoo Finance provides CSV data dumps for free. There are also *many* R packages that automatically get the data for you from Yahoo. Calculate your control limits and simulate your buying/selling strategy using data, for example from 2004 to 2006, then test your strategy on phase II data, from 2007 onwards. You could conceivably write it as a nonlinear optimization problem that calculates limits to maximize your profit. Its expected that different stock sectors will have different limits (e.g. compare slower moving financial stocks to high-tech stocks), because their variability is different.

If this sort of concept seems interesting, take a further look at technical analysis.

Question (not for credit)

> **Note:** This question should take you some time to complete and is open-ended.

A common unit operation in the pharmaceutical area is to uniformly blend powders for tablets. In this question we consider blending an 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 stablize. A new supplier of magnesium stearate is being considered that will save $ 294,000 per year.

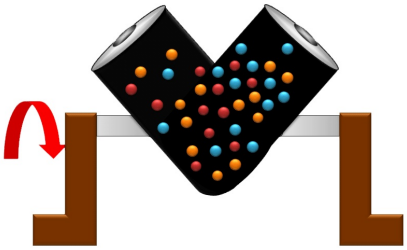


Illustration from Wikipedia

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).

You should always state any assumptions you make, compute a confidence interval for the difference and interpret it.

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:

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

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:

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

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 tables, 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):

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

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*".

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:

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

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 $\overline{x}_B - \overline{x}_A$ 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*".

END