Statistics for Engineering, 4C3/6C3, 2012

Assignment 4

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 06 February 2012, at noon**

Question 1 [1]

Describe what $\overline{S}$ and a_n represent in the derivation of the Shewhart chart control limits.

Solution

The theoretical derivation of the Shewhart chart limits around a mean operating point, $\overline{x}$, is:

$$\overline{x} - c_n\sigma_{\overline{x}} \quad \leq \quad \mu \quad \leq \quad \overline{x} + c_n\sigma_{\overline{x}}$$

The term $\sigma_{\overline{x}}$ is intended to be the standard deviation of the process.

We know, from the Central Limit Theorem, that if we take the average of n data points, the distribution from which that average comes has standard deviation of $\sigma_{\overline{x}} = \sigma/\sqrt{n}$, where σ is the raw data's standard deviation.

In control charts, we use n values in each subgroup, and we call s_k the standard deviation of each subgroup. So $\overline{S}$ is the average of those subgroup standard deviations. However, this average standard deviation is not a perfect estimate of σ, so we correct for that by dividing by a_n.

Notice that as n becomes larger that $a_n \rightarrow 1.0$. This makes intuitive sense, because as our subgroups become larger, the standard deviation, s_k, within the subgroup is a better and better approximation of the average standard deviation, $\overline{S}$.

Question 2 [4; 6 (for 600-level)]

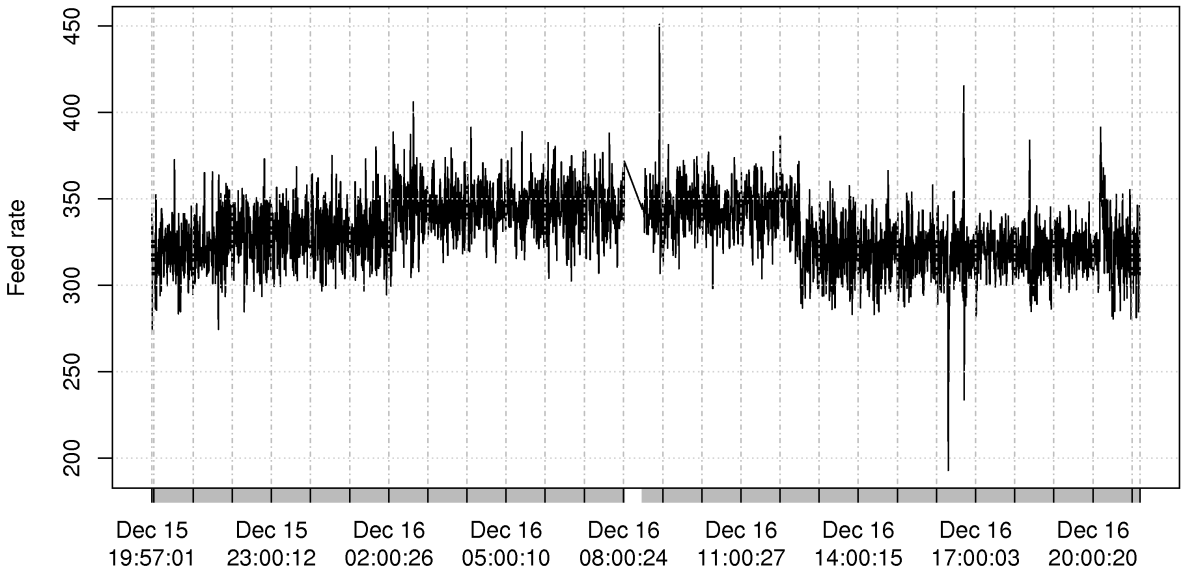
Plant data from a flotation cell are provided on the data set website. One would normally calculate a monitoring chart's limits from a much longer period of data than provided in this data set. We will however use this small data set to illustrate the principle.

Feel free to use *any* software to answer these questions. There is no need to include your code with the solution; just clear explanations of what you are doing, and the corresponding plots.

1. Plot the time-series data for the entire sequence of observations for the `Feed rate` column, which represents the tons of ore fed to the flotation circuit per hour. What do you observe in these data?
2. Use all data on 15 December 2004 (points 1 to 479) from the `Feed rate` variable as your phase 1 data. You have no other process information to go on, so make any assumptions as required. Iteratively prune any outliers to settle on a reasonable set of monitoring parameters. A subgroup size of 4, representing 2 minutes of operation, should be used.
3. Use data from 16 December 2004 (points 480 onwards) from the `Feed rate` variable as your phase 2 (i.e. testing) data. Show the performance of your monitoring parameters calculated in part 1 on these phase 2 data.
4. Explain why the monitoring system works (or doesn't work) as you expect.
5. **600-level students** (extra-credit for 400-level students): implement an alternative monitoring chart using these same data. Describe your calculations.

Solution

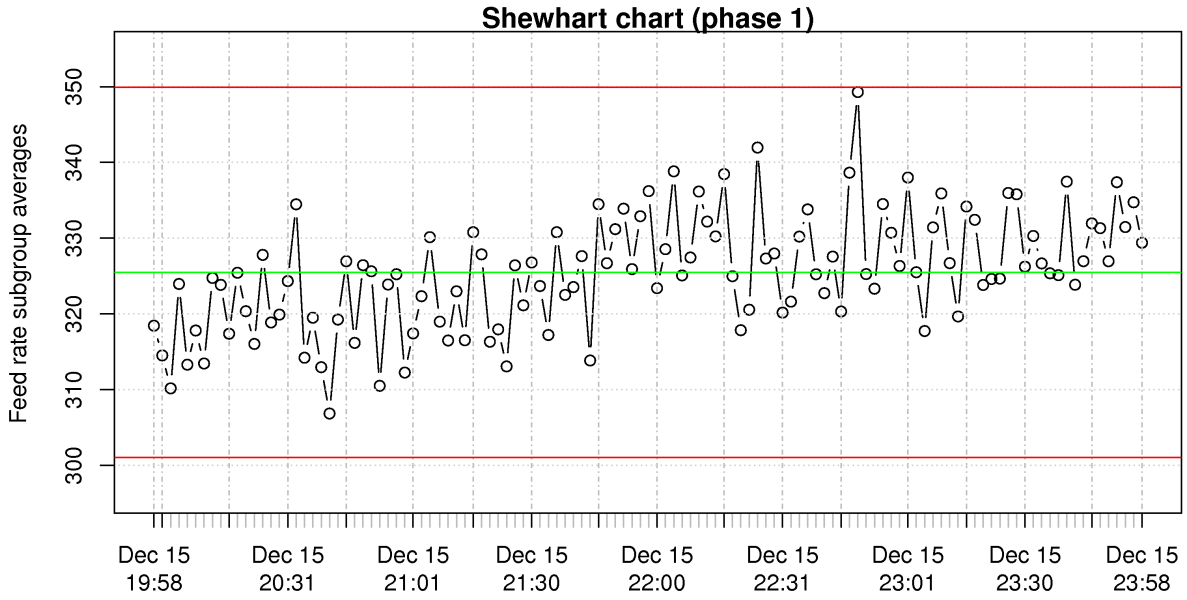
1. A time-series plot, using the `xts` library in R, was generated. Some observations are listed below.



   - The process seems to operate at a certain average level up till 02:00 on 16 December, after which an upward shift occurs.
   - From after 12:00 on the same day it returns back to its original position.
   - There are several large spikes in the data, particularly between 16:00 and 17:00 on 16 December.
   - There is also a very sudden jump in the data, followed by a gradual drift back to stable operation just after 20:00 on 16 December.
   - There is a period of missing data at 08:00 on 16 December.

2. The main assumption used here was that all the data on 15 December were from common-cause operation, i.e. stable operation with no unusual events. This implies that a reasonable guess for the target is the average of the subgroup means, 325.5.

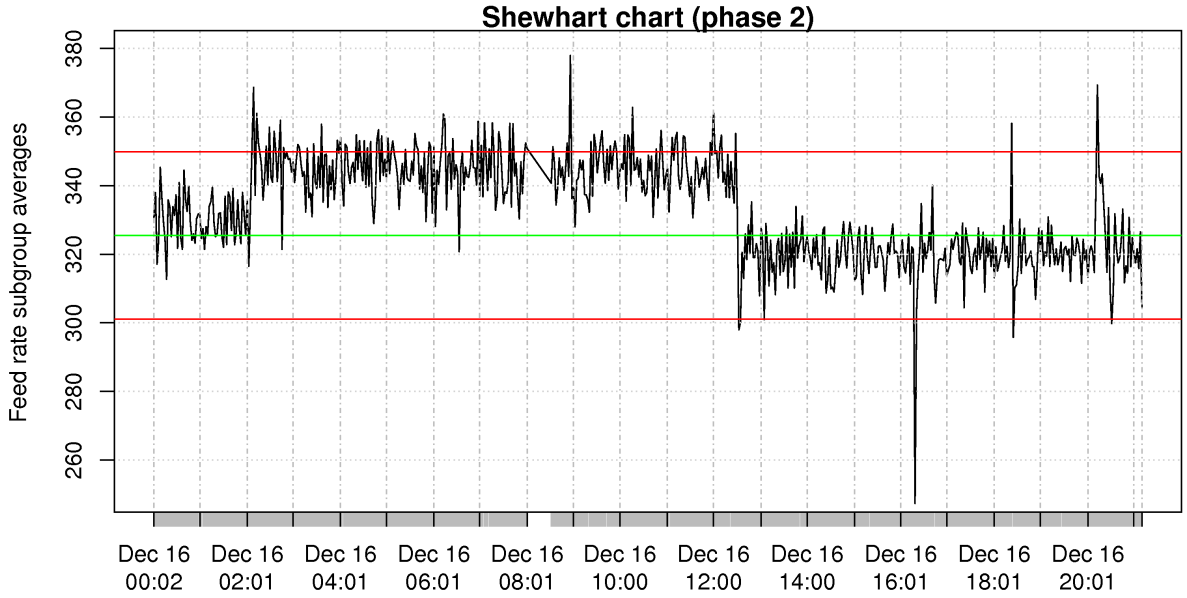
   There were 119 subgroups in the data, and the mean of the subgroup standard deviations is 15.0. This leads to control limits of 301.0 at the lower level and 349.9 at the upper level. A plot of the phase 1 subgroups shows all the subgroups lie within the control limits.



You may have used a sequence (order) based x-axis; that is OK; however a better choice is a time-based x-axis. R code for both options are given below.

My only concern with the phase 1 assumption here is the rising trend in the data during the data: stable operation should how no discernible trends. The Western Electric rules would have triggered alarms with these data.

3. The control chart's performance on the phase 2 data is shown below.



4. The monitoring system generally works as expected.

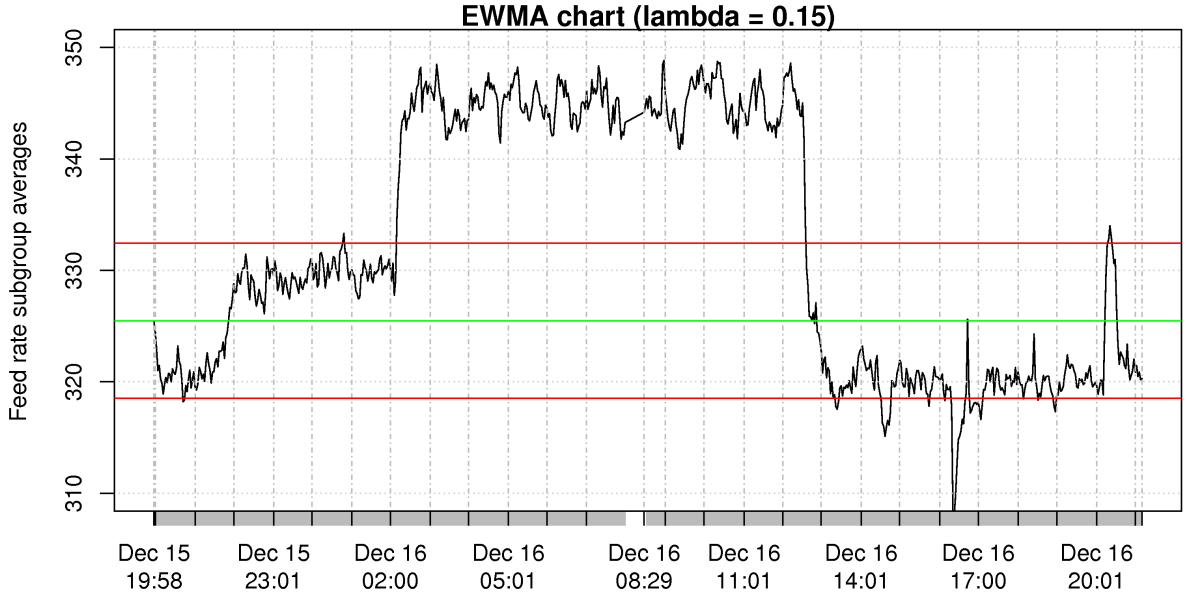
   As hoped for, between 02:00 and 12:00 on 16 December we pick frequent alarms due to the higher feedrate. After 12:30 the process returns to just-below the target. The offset from target is not large enough to raise too many alarms; only the occasional alarm is raised. If Western Electric rules were combined with the Shewhart rules, then this period of time would have raised alarms. The spike just after 16:00 is picked up.

   My concern with this monitoring chart is that is does not pick up the offset from target quite as clearly as hoped for from 12:30 onwards on 16 December.

5. The monitoring system may be improved by using an EWMA chart. It was found, by trying various values of λ, that $\lambda = 0.15$ gives a slightly clearer signal that the process was not stable from about 13:00 onwards on 16

December. Lower values of λ given even better signals, but they also have too-narrow phase 1 control limits.

There is also an interesting oscillatory behaviour between 02:00 and 12:00 on 16 December that should be investigated. This was not visible in the Shewhart chart. Overall, the EWMA chart gives a smoother, more readable, monitoring chart, in my opinion.




```
data <- read.csv('http://datasets.connectmv.com/file/flotation-cell.csv')
summary(data)

# Time-series data for the entire sequence of observations for the
# ``Feed rate`` column.  Use the xts library for better plots; search the
# software tutorial for "xts" to see how.
library(xts)
date.order <- as.POSIXct(data$Date.and.time, format="%d/%m/%Y %H:%M:%S")
Feed.rate <- xts(data$Feed.rate, order.by=date.order)
bitmap('flotation-feedrate.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(3.0, 4, 1, 0.2))
plot(Feed.rate, ylab="Feed rate", main="")
dev.off()

# Use all data on 15 December 2004 (points 1 to 479) from the ``Feed rate``
# variable as your phase 1 data. You have no other process information to go
# on, so make any assumptions as required. Iteratively prune any outliers to
# settle on a reasonable set of monitoring parameters. A subgroup size of 4,
# representing 2 minutes of operation, should be used.
phase1.point <- 479
phase1 <- data$Feed.rate[1:phase1.point]

N.raw = length(phase1)
N.sub = 4                                              # subgroup size
subgroup.1 <- matrix(phase1, N.sub, N.raw/N.sub)
N.groups <- ncol(subgroup.1)
dim(subgroup.1)                                          # 4 by 119 matrix

subgroup.1.sd <- apply(subgroup.1, 2, sd)
subgroup.1.xbar <- apply(subgroup.1, 2, mean)

# Take a look at what these numbers mean
plot(subgroup.1.sd, type="b",  ylab="Subgroup spread")
# there's evidence process really isn't stable
```

```
plot(subgroup.1.xbar, type="b", ylab="Subgroup average")

# Report your target value, lower control limit and upper control limit,
target <- mean(subgroup.1.xbar)
Sbar <- mean(subgroup.1.sd)
an <- sqrt(2) * gamma(N.sub/2) / (sqrt(N.sub-1) * gamma(N.sub/2 - 0.5))
sigma.estimate <- Sbar / an  # a_n value is from the table when subgroup size = 5
LCL <-  target - 3 * sigma.estimate/sqrt(N.sub)
UCL <- target + 3 * sigma.estimate/sqrt(N.sub)
c(LCL, target, UCL)

# Sequence order Shewhart-chart
bitmap('flotation-feedrate-phase1.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(2, 4, 1, 0.2))
plot(subgroup.1.xbar, ylab="Feed rate subgroup averages",
      main="Shewhart chart (phase 1)", ylim=c(LCL-5, UCL+5), type="b")
abline(h=target, col="green")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
dev.off()

# Improved Shewhart chart: uses time on the x-axis
Feed.rate.subgroup.1 <- xts(subgroup.1.xbar,
           order.by=date.order[seq(N.sub, phase1.point, by=N.sub)])
bitmap('flotation-feedrate-phase1-time.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(3, 4, 1, 0.2))
plot(Feed.rate.subgroup.1, ylab="Feed rate subgroup averages",
      main="Shewhart chart (phase 1)", ylim=c(LCL-5, UCL+5), type="b")
abline(h=target, col="green")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
dev.off()

#.        Use data from 16 December 2004 (points 480 onwards) from the ''Feed rate''
# variable as your phase 2 (i.e. testing) data. Show the performance of
# your monitoring parameters calculated in part 1 on these phase 2 data.
phase2.point <- length(data$Feed.rate)
phase2 <- data$Feed.rate[480:phase2.point]
N.raw = length(phase2)
N.sub = 4                                                  # subgroup size
subgroup.2 <- matrix(phase2, N.sub, N.raw/N.sub)
N.groups <- ncol(subgroup.2)
dim(subgroup.2)                                            # 4 by 610 matrix

subgroup.2.sd <- apply(subgroup.2, 2, sd)
subgroup.2.xbar <- apply(subgroup.2, 2, mean)

# Take a look at what these numbers mean
plot(subgroup.2.sd, type="b",  ylab="Subgroup spread")
# there's evidence process really isn't stable
plot(subgroup.2.xbar, type="b", ylab="Subgroup average")

# Ordinary phase 2 Shewhart chart
bitmap('flotation-feedrate-phase2.png', type="png256", width=10, height=7, res=300, pointsize=14)
par(mar=c(4.2, 4.2, 1.2, 0.2))  # (B, L, T, R); defaults are par(mar=c(5, 4, 4, 2) + 0.1)
plot(subgroup.2.xbar, ylab="Subgroup means",
     main="Shewhart chart (phase 2)", ylim=c(250, 380), type="b")
abline(h=target, col="green")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
```

```
dev.off()

# Improved Shewhart chart: uses time on the x-axis
Feed.rate.subgroup.2 <- xts(subgroup.2.xbar,
     order.by=date.order[seq(phase1.point+N.sub, phase2.point, by=N.sub)])
bitmap('flotation-feedrate-phase2-time.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(3, 4, 1, 0.2))
plot(Feed.rate.subgroup.2, ylab="Feed rate subgroup averages",
      main="Shewhart chart (phase 2)", ylim=c(250, 380), type="l")
abline(h=target, col="green")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
dev.off()

# Implement an alternative monitoring chart using these same data: EWMA chart.
# EWMA function below is directly from the course notes
ewma <- function(x, lambda, target=x[1]){
    N <- length(x)
    y <- numeric(N)
    y[1] = target
    for (k in 2:N){
        error = x[k-1] - y[k-1]
        y[k] = y[k-1] + lambda*error
    }
return(y)
}
# Use all data in the EWMA chart
N.raw = length(data$Feed.rate)
N.sub = 4                                                    # subgroup size
subgroup <- matrix(data$Feed.rate, N.sub, N.raw/N.sub)
N.groups <- ncol(subgroup)
subgroup.xbar <- apply(subgroup, 2, mean)
lambda <- 0.15

# Calculate the EWMA values around the phase 1 target
ewma.values <- ewma(subgroup.xbar, lambda, target=mean(subgroup.1.xbar))
Feed.rate.ewma <- xts(ewma.values, order.by=date.order[seq(N.sub, N.raw, by=N.sub)])
LCL <-  target - 3 * sigma.estimate/sqrt(N.sub) * sqrt(lambda/(2-lambda))
UCL <- target + 3 * sigma.estimate/sqrt(N.sub) * sqrt(lambda/(2-lambda))

bitmap('flotation-feedrate-EWMA-time.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(3, 4, 1, 0.2))
plot(Feed.rate.ewma, ylab="Feed rate subgroup averages",
      main="EWMA chart (lambda = 0.15)", ylim=c(310, 350), type="l")
abline(h=target, col="green")
abline(h=UCL, col="red")
abline(h=LCL, col="red")
dev.off()
```

Question 3 [1]

Your process with Cpk of 2.0 experiences a drift of 1.2σ away from the current process operating point towards the closest specification limit. What is the new Cpk value; how many defects per million items did you have before the drift? And after the drift?

Solution

The new Cpk value is 1.5. The number of defects per million items at Cpk = 2.0 is 0.00098 (essentially no defects), while at Cpk = 1.5 it is 3.4 defects per million items. You only have to consider one-side of the distribution, since Cpk is by definition for an uncentered process, and deals with the side closest to the specification limits.

```
Cpk <- 1.5
n.sigma.distance <- 3 * Cpk
defects.per.million <- pnorm(-n.sigma.distance, mean=0, sd=1) * 1E6
```

Question 4 [2]

From the 2010 midterm. Show full calculations please.

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 1 data.

- There are 8900 samples of data in this phase 1 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 $\hat{\sigma} = \frac{\overline{S}}{a_n} = \frac{98}{0.921} = 106.4$, since the subgroup size is $n = 4$.
2. Using the data provided in the question:

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

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, $\overline{x} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, where σ^2 is the process variance, and σ^2/n is the subgroup variance of $\overline{x}$.

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

The area below -2.33 is `pnorm(-2.33) = 0.009903076`, though any value around 1%, eyeballed from the printed tables, is acceptable. 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.

END