May 27, 2014

Abstract

Contents

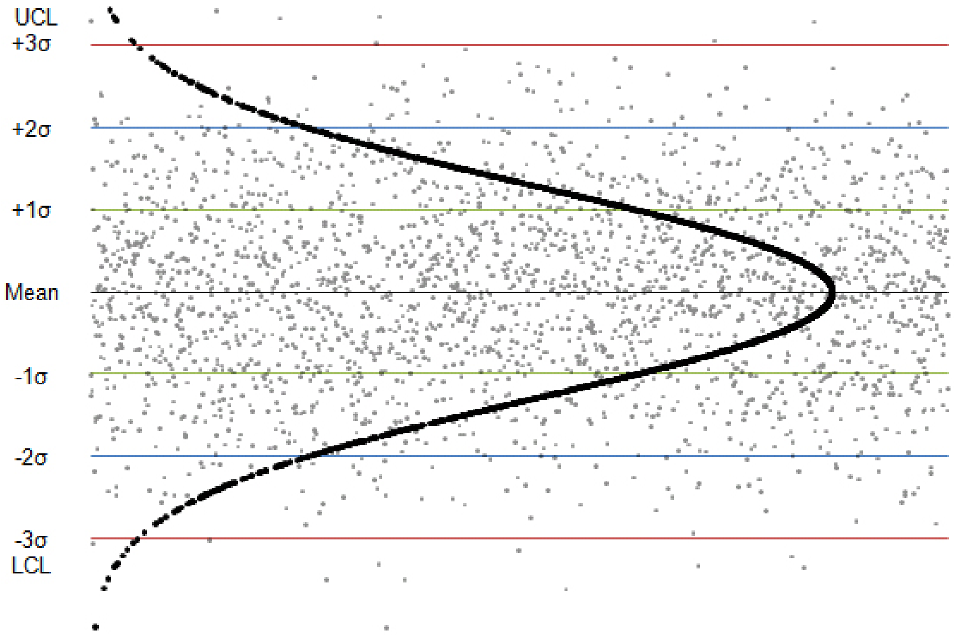
1 Control Charts

1.1 Control Charts

- The control chart is a graph used to study how a process changes over time. Data are plotted in time order. A control chart always has a central line for the average, an upper line for the upper control limit and a lower line for the lower control limit.
- Units are usually the means of small samples.
- These lines are determined from historical data. By comparing current data to these lines, you can draw conclusions about whether the process variation is consistent (in control) or is unpredictable (out of control, affected by special causes of variation).
- There are several types of control chart. In the short term, we will look at the **x-bar** chart (related to mean of values from sample). Other types of chart include R charts and S charts (related to range and standard deviation of the values from each sample)

1.2 Control Limits

- Statistical tables have been developed for various types of distributions that quantify the area under the curve for a given number of standard deviations from the mean (based on the ***normal distribution***).
- Shewhart found that control limits placed at ***three standard deviations from the mean*** in either direction provide an economical tradeoff between the risk of reacting to a false signal and the risk of not reacting to a true signal - regardless the shape of the underlying process distribution.
- If the process has a normal distribution, 99.7% of the population is captured by the curve at three standard deviations from the mean. Stated another way, there is only a 100-99.7%, or 0.3% chance of finding a value beyond 3 standard deviations. Therefore, a measurement value beyond 3 standard deviations indicates that the process has either shifted or become unstable (more variability).


UCL
+3σ
+2σ
+1σ
Mean
-1σ
-2σ
-3σ
LCL

2 Ten R Packages that every Data Scientist Should know

2.1 10 R packages I wish I knew about earlier

- *Following material written by Drew Conway, and was published on the Yhat blog Feb 2013*
- *http://blog.yhathq.com/posts/10-R-packages-I-wish-I-knew-about-earlier.html*

- **qcc** is a library for statistical quality control. Back in the 1950s, the now defunct Western Electric Company was looking for a better way to detect problems with telephone and eletrical lines.

  Remark: *We will discuss these rules shortly*

- They came up with a set of rules to help them identify problematic lines. The rules look at the historical mean of a series of datapoints and based on the standard deviation, the rules help judge whether a new set of points is experiencing a mean shift.

- The classic example is monitoring a machine that produces lug nuts.
  Let's say the machine is supposed to produce 2.5 inch long lug nuts. We measure a series of lug nuts: 2.48, 2.47, 2.51, 2.52, 2.54, 2.42, 2.52, 2.58, 2.51.

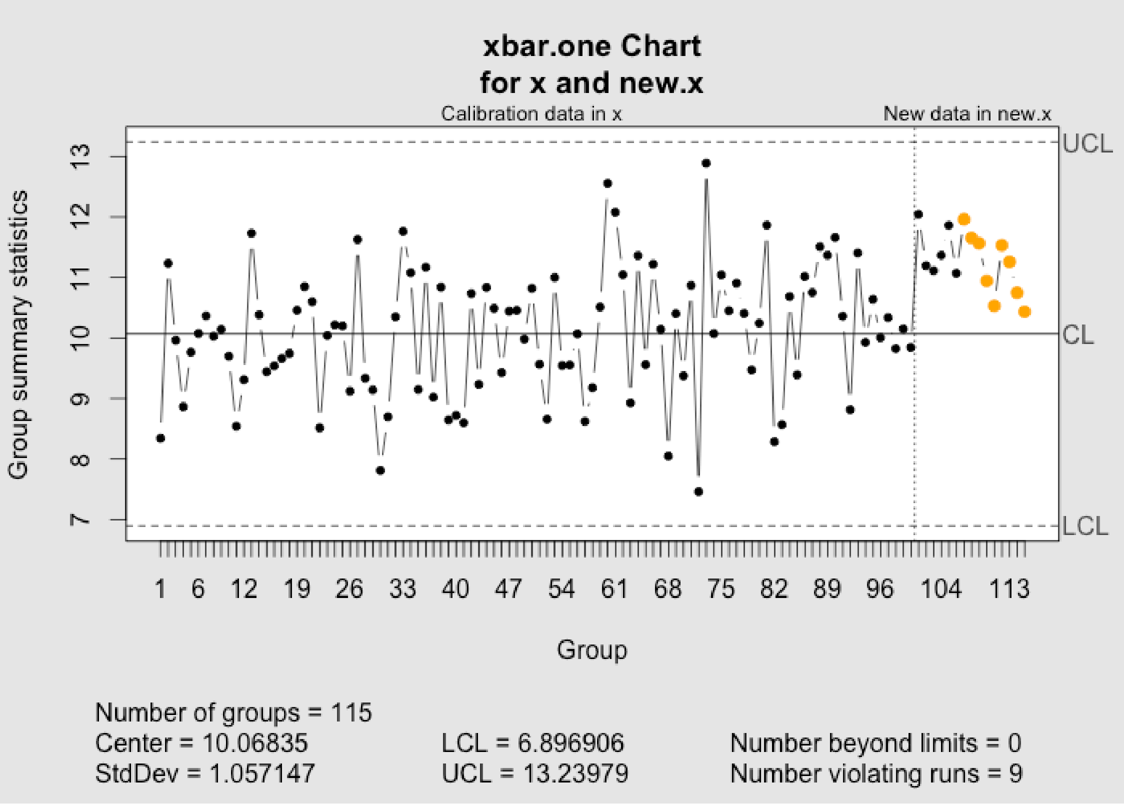
- Is the machine broken? Well it's hard to tell, but the *Western Electric* Rules can help.

- While you might not be monitoring telephone lines, qcc can help you monitor transaction volumes, visitors or logins on your website, database operations, and lots of other processes.

2.1.1 Other Remarks

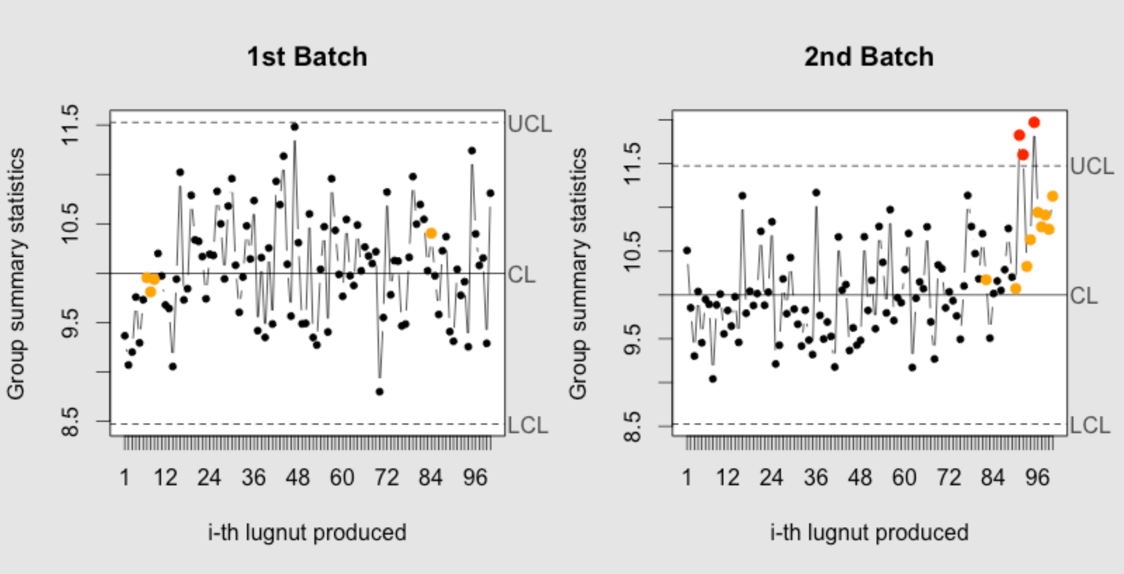
- Quality Control and quality assurance are important functions in most businesses from manufacturing to software development.
- For most, this means that one or more people are meticulously inspecting what's coming out of the factory, looking for imperfections and validating that requirements for products and services produced are satisfied.
- Often times QC and QA are performed manually by a select few specialists, and determining suitable quality can be extremely complex and error-prone.

```
# install.package(qcc)
library(qcc)

# series of value w/ mean of 10 with a little random noise added in
x <- rep(10, 100) + rnorm(100)

# a test series w/ a mean of 11
new.x <- rep(11, 15) + rnorm(15)

# qcc will flag the new points
qcc(x, newdata=new.x, type="xbar.one")
```



Number of groups = 115
Center = 10.06835
StdDev = 1.057147
LCL = 6.896906
UCL = 13.23979
Number beyond limits = 0
Number violating runs = 9

Figure 1:



```
library(qcc)
#make 2 plots in 1 figure
par(mfrow=c(1,2))

#points have base value of 10 w/ normally distributed error
lugnuts <- rep(10, 100) + rnorm(100, mean=0, sd=0.5)
qcc(lugnuts, type="xbar.one", center=10, add.stats=FALSE,
    title="1st Batch",
    xlab="i-th lugnut produced")
```

Second Batch

- First 90 points have mean value of 10 with normally distributed error,
- Last 10 points have mean value of 11 with normally distributed error

```
lugnuts <- c(rep(10, 90), rep(11, 10)) + rnorm(100, mean=0, sd=0.5)
qcc(lugnuts, type="xbar.one", center=10, add.stats=FALSE,
    title="2nd Batch",
    xlab="i-th lugnut produced")
```

```
> set.seed(1234)
> lugnuts <- rep(10, 100) + rnorm(100, mean=0, sd=0.5)
> summary(lugnuts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.827   9.552   9.808   9.922  10.240  11.270
> length(lugnuts)
[1] 100
> newLugnuts <- rep(11, 10) + rnorm(10, mean=0, sd=0.5)
> summary(newLugnuts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.55   10.75   11.00   10.92   11.08   11.21
> length(newLugnuts)
[1] 10
```

```
qcc1 <- qcc(lugnuts, type="xbar.one", center=10, add.stats=TRUE,
    title="1st Batch of 100",
    xlab="i-th lugnut produced")



qcc2 <- qcc(lugnuts, newdata=newLugnuts,
    type="xbar.one", center=10,
    add.stats=TRUE, title="All Lugnuts",
    xlab="i-th lugnut produced")

mode(qcc1)
class(qcc1)
names(qcc1)
```

```
> mode(qcc1)
[1] "list"
> class(qcc1)
[1] "qcc"
> names(qcc1)
 [1] "call"       "type"       "data.name"  "data"       "statistics"
 [6] "sizes"      "center"     "std.dev"    "nsigmas"    "limits"
[11] "violations"
> qcc1$violations
$beyond.limits
integer(0)

$violating.runs
 [1] 13 38 39 40 48 49 50 51 52 53 54 55
```

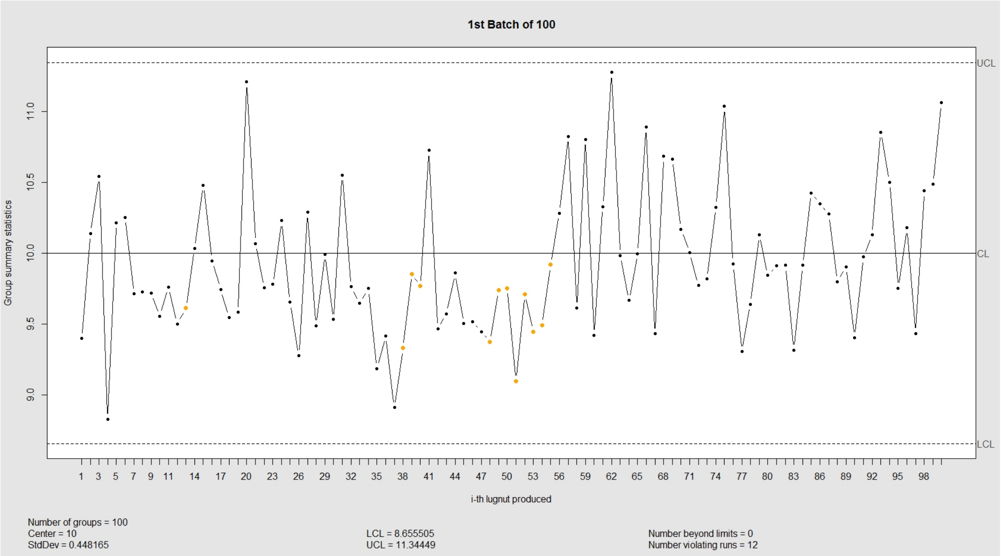


Figure 2:

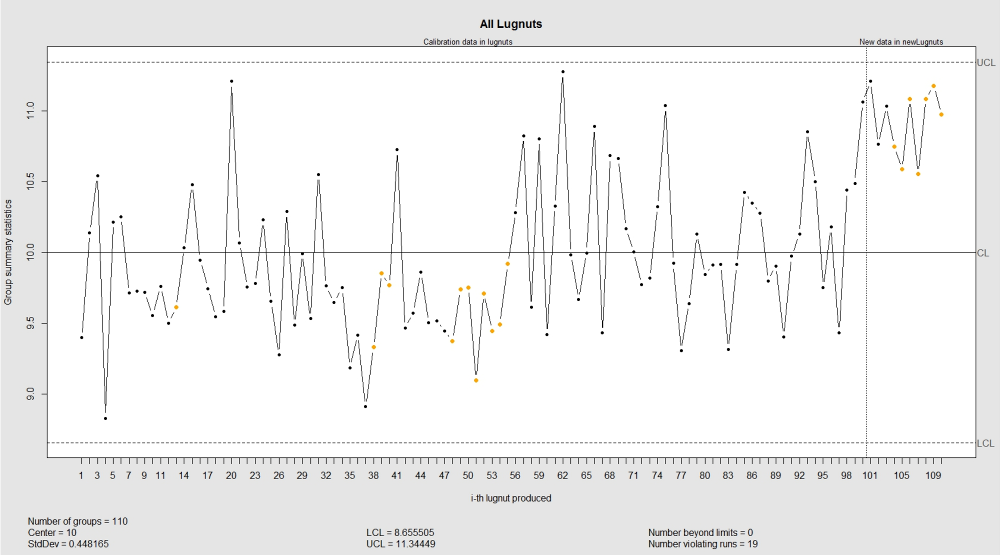


Figure 3:

2.2 Using the summary command

```

Call:
qcc(data = lugnuts, type = "xbar.one", center = 10, add.stats = TRUE,
title = "1st Batch of 100", xlab = "i-th lugnut produced")

xbar.one chart for lugnuts

Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.827   9.552   9.808   9.922  10.240  11.270

Group sample size:  1
Number of groups:  100
Center of group statistics:  10
Standard deviation:  0.448165

Control limits:
      LCL      UCL
 8.655505 11.34449
```

```

Call:
qcc(data = lugnuts, type = "xbar.one", center = 10, newdata = newLugnuts,
add.stats = TRUE, title = "All Lugnuts", xlab = "i-th lugnut produced")

xbar.one chart for lugnuts
.....

Summary of group statistics in newLugnuts:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.55   10.75   11.00   10.92   11.08   11.21

Group sample size:  1
Number of groups:  10

Control limits:
      LCL      UCL
 8.655505 11.34449
```

2.2.1 Example used by Drew Conway

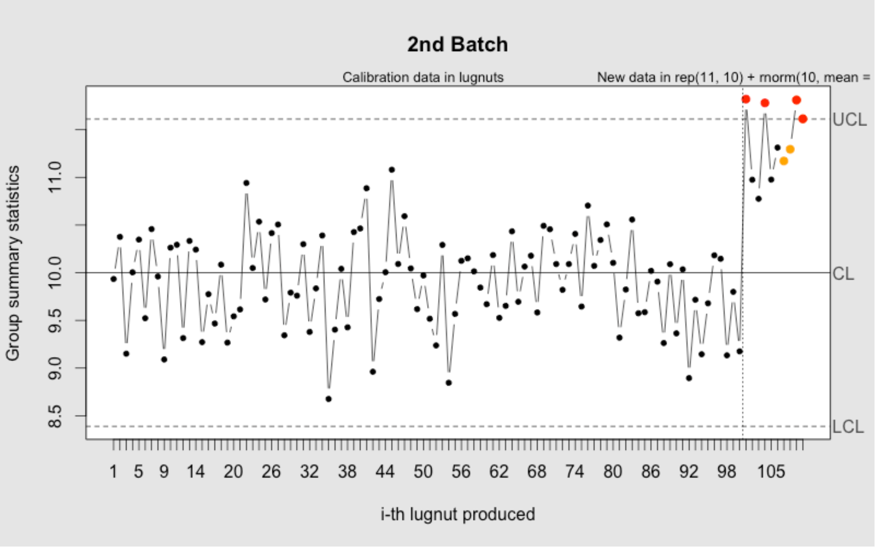


Figure 4:

```
lugnuts <- rep(10, 100) + rnorm(100, mean=0, sd=0.5)
qcc(lugnuts, newdata=rep(11, 10) + rnorm(10, mean=0, sd=0.5),
    type="xbar.one", center=10,
    add.stats=FALSE, title="2nd Batch",
    xlab="i-th lugnut produced")
```

2.2.2 Remarks

- `newdata`
- `add.stats`

2.3 Using R

Advantages of using R statistical program along with the **qcc** package:

- There are several packages for interfacing with databases, RODBC being a common and useful one on MS windows.
- Allows of automation: you can program a regular event loop to check for new data and run a new set of charts and notifications if there is new data.
- The **mail** and **sendmailR** packages were designed to automatically send e-mails with regular reports and warning messages.
- It produces the standard SPC charts, these can go to the screen or a file to be sent out.
- Bespokes tests can be program for out of control signals
- Full programming language with common (and uncommon) statistics so you can preproccess you data in many ways to reduce dimension.
- You can have multiple instances running on multiple or a single computer, each processing for a single department, or you can combine it all into one script to run for all the departments.

3 What is Statistical Process Control

- The concepts of Statistical Process Control (SPC) were initially developed by Dr. Walter Shewhart of Bell Laboratories in the 1920's, and were expanded upon by Dr. W. Edwards Deming, who introduced SPC to Japanese industry after WWII.

- After early successful adoption by Japanese firms, Statistical Process Control has now been incorporated by organizations around the world as a primary tool to improve product quality by reducing process variation.

- Dr. Shewhart identified two sources of process variation:

  Chance variation that is inherent in process, and stable over time,

  Assignable variation , or Uncontrolled variation, which is unstable over time - the result of specific events outside the system.

- Dr. Deming relabeled chance variation as **Common Cause** variation, and assignable variation as **Special Cause** variation.

- Based on experience with many types of process data, and supported by the laws of statistics and probability, Dr. Shewhart devised control charts used to plot data over time and identify both Common Cause variation and Special Cause variation.

3.1 Some Remarks on Multivariate Techniques

- Nowadays, the intensive use of an automatic data acquisition systems and the use of online computers for process monitoring have led to an increased occurrence of industrial processes with two or more correlated quality characteristics, in which the statistical process control and the capability analysis should be performed using multivariate methodologies. *(Edgar Santos-Fernandez)*

3.2 7 Basic Tools of Quality

These are 7 QC tools also known as Ishikawas **7QC** tools

Cause-and-effect diagram : Identifies many possible causes for an effect or problem and sorts ideas into useful categories. (also called *Ishikawa* or *fishbone chart*)

Check sheet : A structured, prepared form for collecting and analyzing data; a generic tool that can be adapted for a wide variety of purposes.

Control charts : Graphs used to study how a process changes over time.

Histogram : The most commonly used graph for showing frequency distributions, or how often each different value in a set of data occurs.

Pareto chart : Shows on a bar graph which factors are more significant.

Scatter diagram : Graphs pairs of numerical data, one variable on each axis, to look for a relationship.

Stratification : A technique that separates data gathered from a variety of sources so that patterns can be seen (some lists replace stratification with flowchart or run chart).

For the sake of brevity, we will only look at a couple of these.

3.3 Multivariate Normal

- The multivariate normal distribution or multivariate Gaussian distribution, is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions.
- One possible definition is that a random vector is said to be k-variate normally distributed if every linear combination of its k components has a univariate normal distribution.
- The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value.

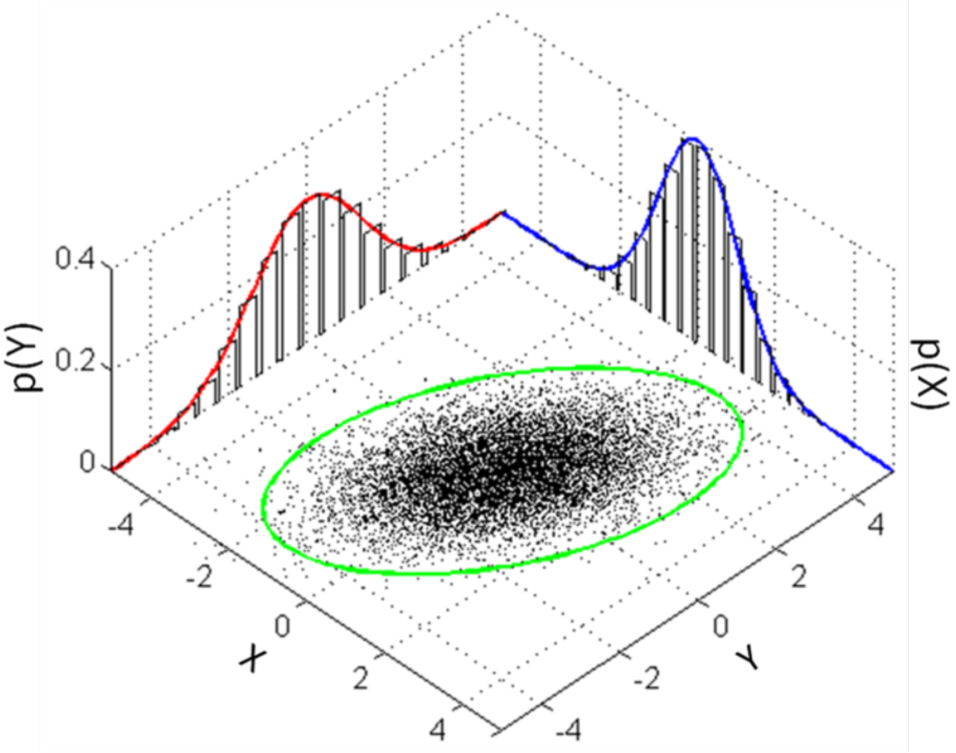


Figure 5:

3.4 Testing for Normality

Graphical Methods

- Histograms
- Normal Probability Plots

Hypothesis Tests for Univariate Data

- Shapiro-Wilk Test (inbuilt with `R`)
- D'Agostino Test (MSQC package)

Hypothesis Tests for Multivariate Data

- Mardia Test (MSQC package)
- Henze and Zirkler (MSQC package)
- Royston Test (MSQC package)

3.4.1 The bimetal data set (MSQC package)

- Bimetal thermostat has innumerable practical uses. These types of thermostats hold a bimetallic strip composed by two strips of different metals that convert the changing of temperature in mechanical displacement due to the difference in thermal expansion.
- Certain type of strip composed of brass and steel is analyzed in a quality laboratory by testing the deflection, curvature, resistivity, and hardness in low and high expansion sides.

```
> tail(bimetal1)
      deflection curvature resistivity Hardness low side Hardness high side
[23,]      20.76     39.98       14.98             22.29              26.03
[24,]      21.00     40.11       15.17             22.04              25.99
[25,]      20.57     39.73       14.35             22.02              25.80
[26,]      20.78     39.83       15.27             21.60              25.89
[27,]      20.96     40.03       15.26             21.98              25.94
[28,]      21.14     39.93       14.98             21.84              25.98
```

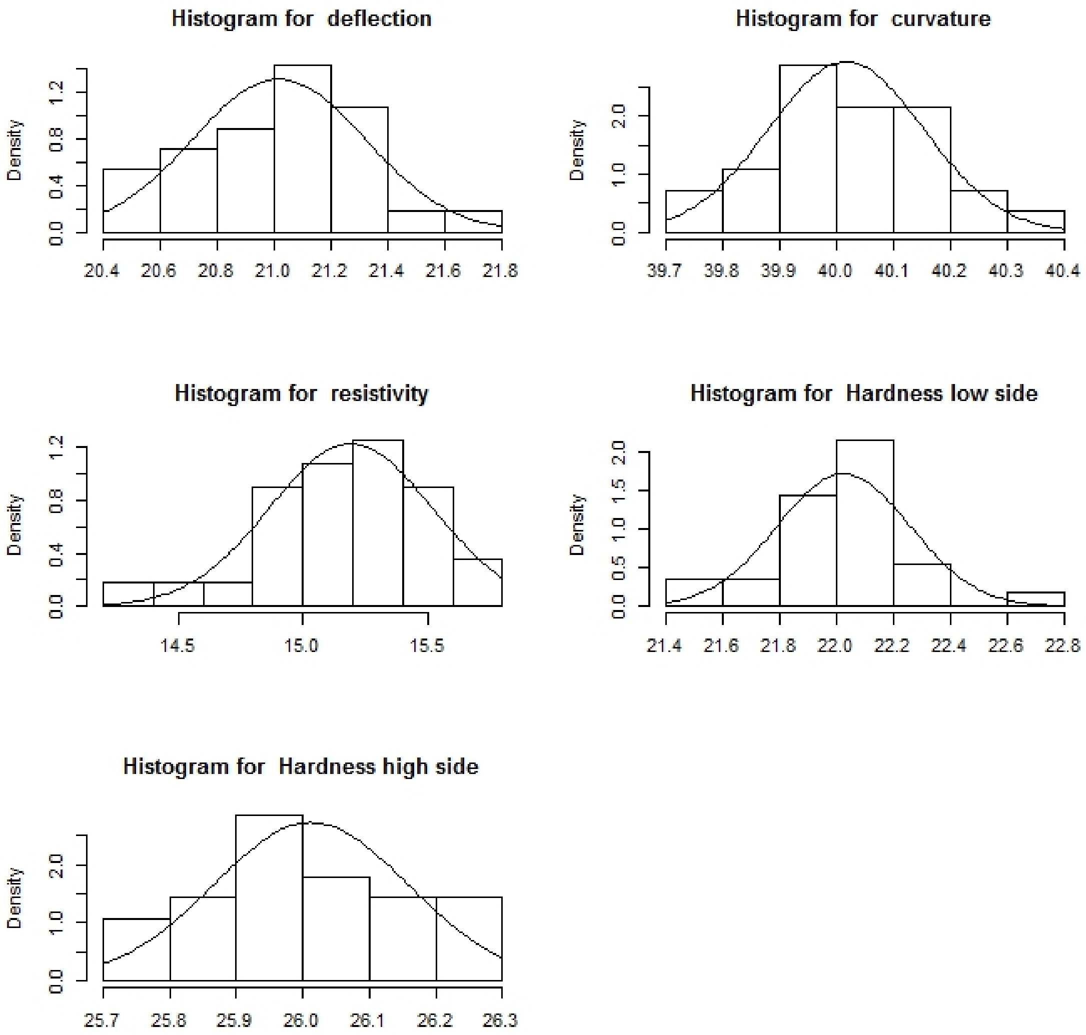

Histogram for deflection
Density
0.0
0.4
0.8
1.2
20.4
20.6
20.8
21.0
21.2
21.4
21.6
21.8
Histogram for curvature
Density
0.0
1.0
2.0
39.7
39.8
39.9
40.0
40.1
40.2
40.3
40.4
Histogram for resistivity
Density
0.0
0.4
0.8
1.2
14.5
15.0
15.5
Histogram for Hardness low side
Density
0.0
0.5
1.0
1.5
2.0
21.4
21.6
21.8
22.0
22.2
22.4
22.6
22.8
Histogram for Hardness high side
Density
0.0
1.0
2.0
25.7
25.8
25.9
26.0
26.1
26.2
26.3

Figure 6:

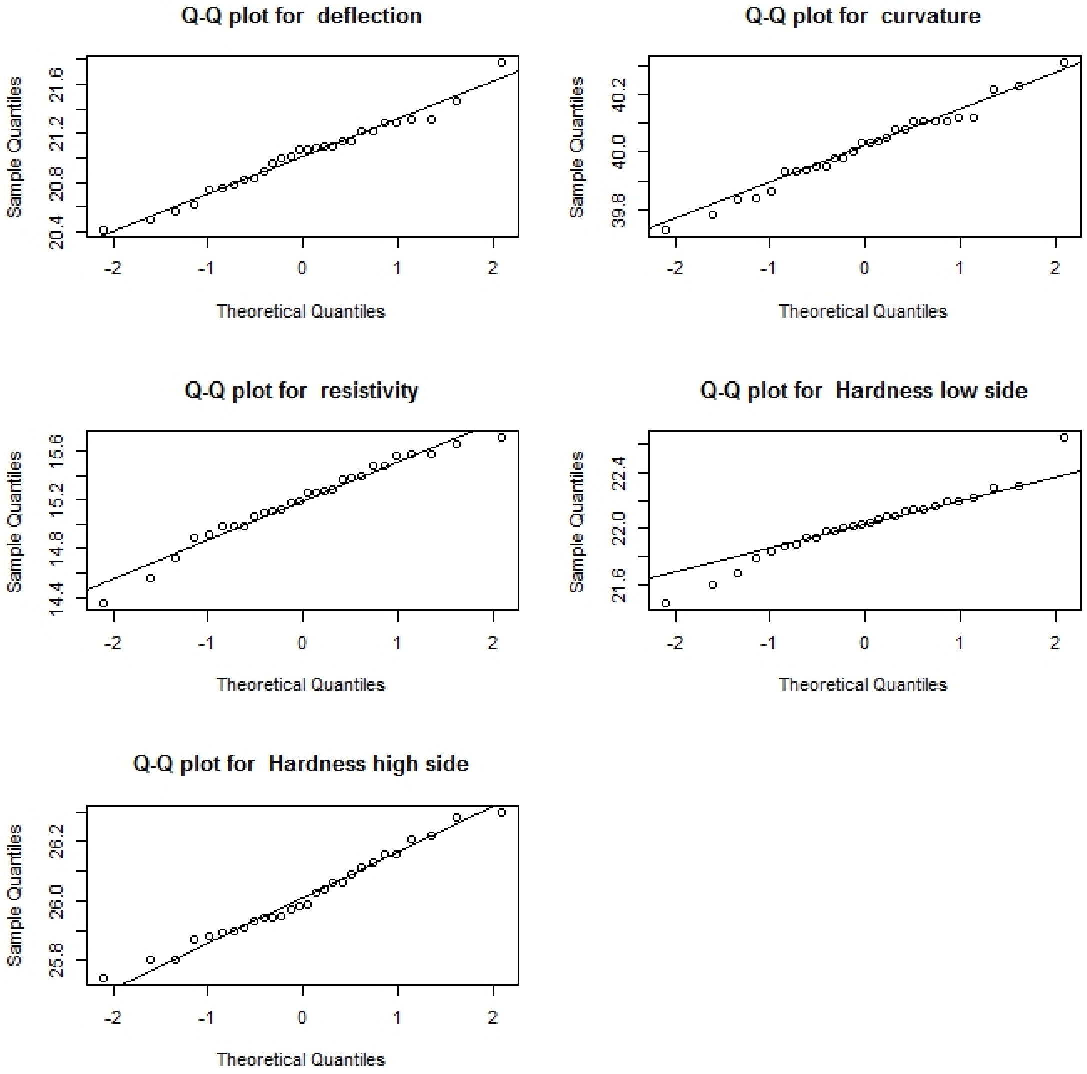

Q-Q plot for deflection
Sample Quantiles
20.4
20.8
21.2
21.6
Theoretical Quantiles
-2
-1
0
1
2
Q-Q plot for curvature
Sample Quantiles
39.8
40.0
40.2
Theoretical Quantiles
-2
-1
0
1
2
Q-Q plot for resistivity
Sample Quantiles
14.4
14.8
15.2
15.6
Theoretical Quantiles
-2
-1
0
1
2
Q-Q plot for Hardness low side
Sample Quantiles
21.6
22.0
22.4
Theoretical Quantiles
-2
-1
0
1
2
Q-Q plot for Hardness high side
Sample Quantiles
25.8
26.0
26.2
Theoretical Quantiles
-2
-1
0
1
2

Figure 7:

3.4.2 D'Agostino Test (MSQC Pacakge)

- Using the bimetal1 data set in MSQC package

```
> for (i in 1 : 5){
+  DAGOSTINO(bimetal1[,i])
+  }
D'Agostino Test
    Skewness
      Skewness coefficient: 0.0831225
      Statistics: 0.2117358
      p-value: 0.8323131
    Kurtosis
      The kurtosis coefficient: 3.0422
      Statistics: 0.591983
      p-value: 0.553862
    Omnibus Test
      Chi-squared: 0.3952759
      Degree of freedom: 2
      p-value: 0.8206669
....
....
D'Agostino Test
    Skewness
      Skewness coefficient: -0.04173762
      Statistics: -0.1063873
      p-value: 0.9152751
    Kurtosis
      The kurtosis coefficient: 4.162062
      Statistics: 1.675258
      p-value: 0.09388364
    Omnibus Test
      Chi-squared: 2.817807
      Degree of freedom: 2
      p-value: 0.2444111
```

3.4.3 Some Multivariate (MSQC Pacakge)

```
> MardiaTest(bimetal1)
$skewness
[1] 6.982112

$p.value
[1] 0.585327

$kurtosis
[1] 33.77373

$p.value
[1] 0.3490892

>
>
>
> HZ.test(bimetal1)
[1] 0.6068650 0.7709586
>
>
> Royston.test(bimetal1)
test.statistic        p.value
     1.1814742      0.9364221
```

3.4.4 Box Cox Transformation

- The Box-Cox transforms nonnormally distributed data to a set of data that has approximately normal distribution.

4 Nelson Rules for Interpreting Control Charts

- The eight tests used in statistical process control were developed by Lloyd S. Nelson, a process control expert. They are based on his determination that the identified patterns are very unlikely to occur in stable processes.
- Thus the existence of any of these patterns in an $\bar{X}$ chart indicates that the process may be unstable, and that one or more assignable causes may exist.
- The table on the next page contains examples of test failure for each of the eight tests, with a description for each graph as to what is required for the illustrated test failure.
- In practice, tests 1,2 and 7 are considered the three most useful.

4.1 Descriptions of Tests

Test 1 - 3 sigma rule Identifies points outside of the control limits

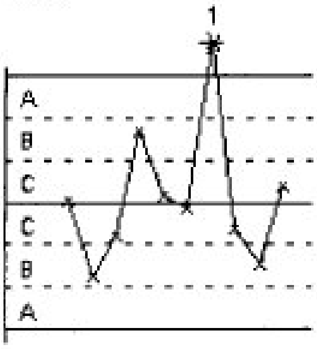
Test 1 identifies points that are more standard deviations from the center line. Test 1 is universally recognized as necessary for detecting out-of-control situations. It has a false alarm rate of only 0.27%.

Test 2 Identifies shifts in the means

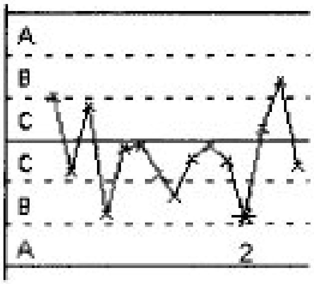
Test 2 signals when 9 points in a row fall on the same side of the center line. The use of Test 2 significantly increases the sensitivity of the chart to detect small shifts in the mean.

When test 1 and test 2 are used together, significantly fewer subgroups are needed to detect a small shift in the mean than are needed when test 1 is used alone. Therefore, adding test 2 helps to detect common out-of-control situations and increases sensitivity enough to warrant a slight increase in the false alarm rate.

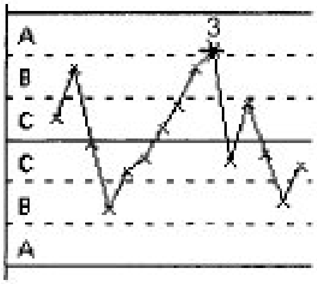
Test 1
One point more than 3 sigmas from center line



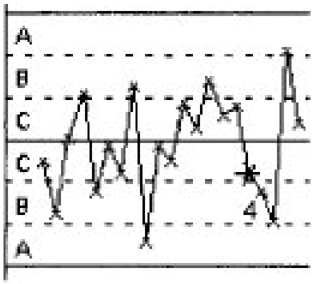
Test 2
Nine points in a row on same side of center line



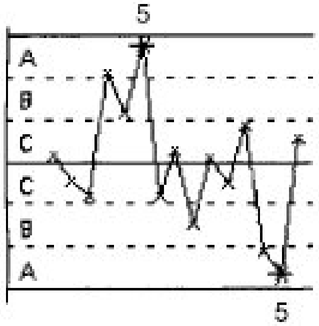
Test 3
Six points in a row, all increasing or all decreasing



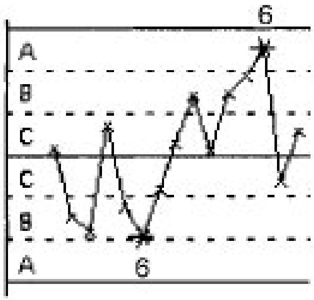
Test 4
Fourteen points in a row, alternating up and down



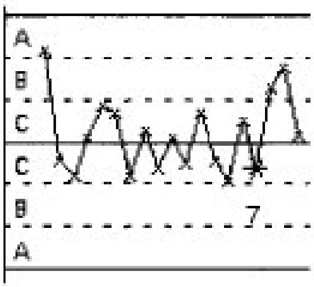
Test 5
Two out of three points in a row more than 2 sigmas from center line (same side)



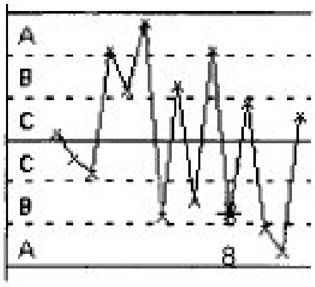
Test 6
Four out of five points in a row more than 1 sigma from center line (same side)



Test 7
Fifteen points in a row within 1 sigma of center line (either side)



Test 8
Eight points in a row more than 1 sigma from center line (either side)



Test 3 k points in a row, all increasing or all decreasing

Test 3 is designed to detect drifts in the process mean.

However, when test 3 is used in addition to test 1 and test 2, it does not significantly increase the sensitivity of the chart to detect drifts in the process mean.

Test 4 k points in a row, alternating up and down

Although this pattern can occur in practice, it is recommended to search for any unusual trends or patterns rather than test for one specific pattern.

Test 5 k out of k=1 points > 2 standard deviations from center line

This test is not quite as informative because it did not uniquely identify special cause situations that are common in practice.

Test 6 k out of k+1 points > 1 standard deviation from the center line

This test is not quite as informative because it did not uniquely identify special cause situations that are common in practice.

Test 7 Identifies control limits that are too wide

Test 7 signals when 12 or 15 points in a row fall within 1 standard deviation of the center line.

Test 7 is used only for the $\bar{X}$ chart when the control limits are estimated from the data. When this test fails, the cause is usually a systemic source of variation (stratification) within a subgroup, which is often the result of not forming rational subgroups.

Test 8 k points in a row > 1 standard deviation from center line (either side)

This test is not quite as informative because it did not uniquely identify special cause situations that are common in practice.

5 Multivariate Control Charts

- With the enhancements in data acquisition systems it is usual to deal with processes with more than one correlated quality characteristic to be monitored.
- A common practice is to control the stability of the process using univariate control charts.
- This practice increases the probability of false alarm of special cause of variation.
- Therefore, the analysis should be performed through a multivariate approach; that is, the variables must be analyzed together, not independently.

5.1 Multivariate Control Charts

- Multivariate control charts monitor multiple process characteristics. Independent variables can be charted individually, but if the variables are correlated, a multivariate chart is needed to determine whether the process is in control.
- Multivariate control charts can detect shifts in the mean or the relationship between several related variables.
- The multivariate control chart plots Hotellings T2 statistic. The calculation for the control limit differs based on whether targets have been specified.

5.2 The MSQC package

In his book, Edgar Santos-Fernandez present the multivariate normal distribution, the data structure of the multivariate problems dealt in this book, the mult.chart function that allows the computation in R, and the most used multivariate control charts:

- The control ellipsoid or w2 control chart
- The T2 or Hotelling chart
- The Multivariate Exponentially Weighted Moving Average (MEWMA) chart
- The Multivariate Cumulative Sum (MCUSUM) chart
- The chart based on Principal Components Analysis (PCA)

5.3 The `mult.chart` Function

The performing of the multivariate control chart in R can be carried out with the function mult.chart which is a general function that allows to compute the most accepted and diversified continuous multivariate chart such as

- χ^2
- Hotelling T^2
- MEWMA
- MCUSUM according to Crosier (1988)
- MCUSUM by Pignatiello and Runger (1990)

Finally the function `mult.chart` returns:

- The T2 statistics
- The Upper Control Limit (UCL)
- The sample covariance matrix (S)
- The mean vector (Xmv)
- And if any point falls outside of the UCL and its decomposition

```
mult.chart(dowel1, type = "chi", alpha = 0.05)
```

5.4 T2 control chart

The origin of the T2 control chart dates back to the pioneer works of Harold Hotelling who applied this method to the bombsight problem in Second World War. The Hotelling (1947) procedure has become without doubt the most applied in multivariate process control and it is the multivariate analogous of the Shewhart control chart. For that reason, it is also known as multivariate Shewhart control chart.

```
data("carbon1")
mult.chart(type = "t2", carbon1)
mult.chart(type = "t2", carbon1)$t2
```

5.5 mqcc Example

```
# library(mqcc)
# Ryan (2000, Table 9.2) data with p = 2 variables,
#  m = 20 samples, n = 4 sample size:

X1 = matrix(c(72, 56, 55, 44, 97, 83, 47, 88, 57, 26, 46,
49, 71, 71, 67, 55, 49, 72, 61, 35, 84, 87, 73, 80, 26, 89, 66,
50, 47, 39, 27, 62, 63, 58, 69, 63, 51, 80, 74, 38, 79, 33, 22,
54, 48, 91, 53, 84, 41, 52, 63, 78, 82, 69, 70, 72, 55, 61, 62,
41, 49, 42, 60, 74, 58, 62, 58, 69, 46, 48, 34, 87, 55, 70, 94,
49, 76, 59, 57, 46), ncol = 4)

X2 = matrix(c(23, 14, 13, 9, 36, 30, 12, 31, 14, 7, 10,
11, 22, 21, 18, 15, 13, 22, 19, 10, 30, 31, 22, 28, 10, 35, 18,
11, 10, 11, 8, 20, 16, 19, 19, 16, 14, 28, 20, 11, 28, 8, 6,
15, 14, 36, 14, 30, 8, 35, 19, 27, 31, 17, 18, 20, 16, 18, 16,
13, 10, 9, 16, 25, 15, 18, 16, 19, 10, 30, 9, 31, 15, 20, 35,
12, 26, 17, 14, 16), ncol = 4)

X = list(X1 = X1, X2 = X2)
q = mqcc(X, type = "T2")
summary(q)
```

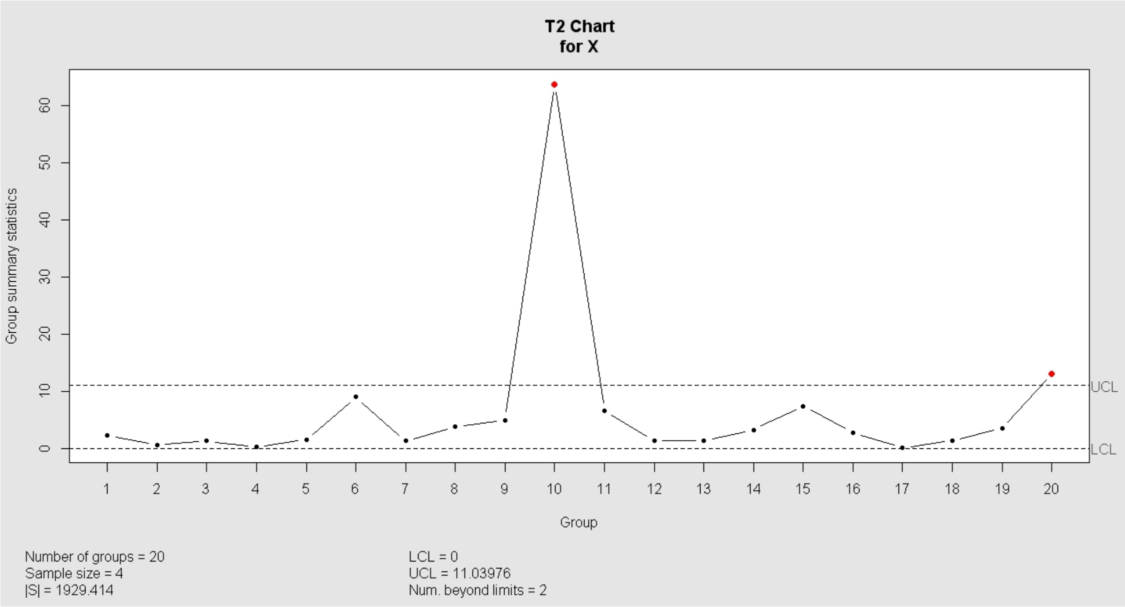


Figure 8:

```
Call:
mqcc(data = X, type = "T2")

T2 chart for X

Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1243  1.3250  2.5030  6.4700  5.3490 63.7600

Number of variables:  2
Number of groups:  20
Group sample size:  4

Center:
     X1      X2
60.3750 18.4875

Covariance matrix:
         X1        X2
X1 222.0333 103.11667
```

```
X2 103.1167  56.57917
|S|:  1929.414

Control limits:
 LCL      UCL
   0 11.03976
```

6 The qcc R package - The 7QC tools revisited

- The **qcc** package was built by Luca Scrucca for nothing but statistical quality control.
- It's extremely easy to use. You provide it with data and it tells you which points are considered to be outliers based on the Shewart Rules.
- It even color codes them based on how irregular each point is.
- Even though statistical quality control an old topic, statistical quality control is still highly relevant. There are probably have lots of jobs, processes, logs, or databse metric tha could be monitored using control charts.

6.1 qcc : Quality Control Charts

Some Remarks

- Shewhart quality control charts for continuous, attribute and count data.
- Cusum and EWMA charts.
- Operating characteristic curves.
- Process capability analysis.
- Pareto chart and cause-and-effect chart.
- Multivariate control charts.

6.2 Types of Control Chart supported by qcc

"xbar" mean - means of a continuous process variable

"R" range ranges of a continuous process variable

"S" standard deviation standard deviations of a continuous variable

"xbar.one" mean one-at-time data of a continuous process variable

"p" proportion proportion of nonconforming units

"np" count number of nonconforming units

"c" count nonconformities per unit

"u" count average nonconformities per unit

"g" count number of non-events between events

6.3 Pareto Chart Analysis.

- A Pareto chart is a barplot where the categories are ordered in non increasing order, and a line is also added to show the cumulative sum.
- Quality problems are rarely spread evenly across the different aspects of the production process or different plants. Rather, a few ”bad apples” often account for the majority of problems.
- This principle has come to be known as the Pareto principle, which basically states that quality losses are mal-distributed in such a way that a small percentage of possible causes are responsible for the majority of the quality problems.
- For example, a relatively small number of ”dirty” cars are probably responsible for the majority of air pollution; the majority of losses in most companies result from the failure of only one or two products. To illustrate the ”bad apples”, one plots the Pareto chart,

6.4 Pareto Analysis (Implementation with qcc package)

```
defect <- c(80, 27, 66, 94, 33)

names(defect) <- c("price code", "schedule date",
 "supplier code", "contact num.", "part num.")

# 1
pareto.chart(defect, ylab = "Error frequency")

#2
pareto.chart(defect, ylab = "Error frequency", xlab = "Error causes", las=1)

#3
pareto.chart(defect, ylab = "Error frequency", col=rainbow(length(defect)))

#4
pareto.chart(defect, cumperc = seq(0, 100, by = 5),
    ylab2 = "A finer tickmarks grid")
```

Output to accompany graphs

```
Pareto chart analysis for defect
                Frequency Cum.Freq. Percentage Cum.Percent.
  contact num.         94        94   31.33333     31.33333
  price code           80       174   26.66667     58.00000
  supplier code        66       240   22.00000     80.00000
  part num.            33       273   11.00000     91.00000
  schedule date        27       300    9.00000    100.00000
```

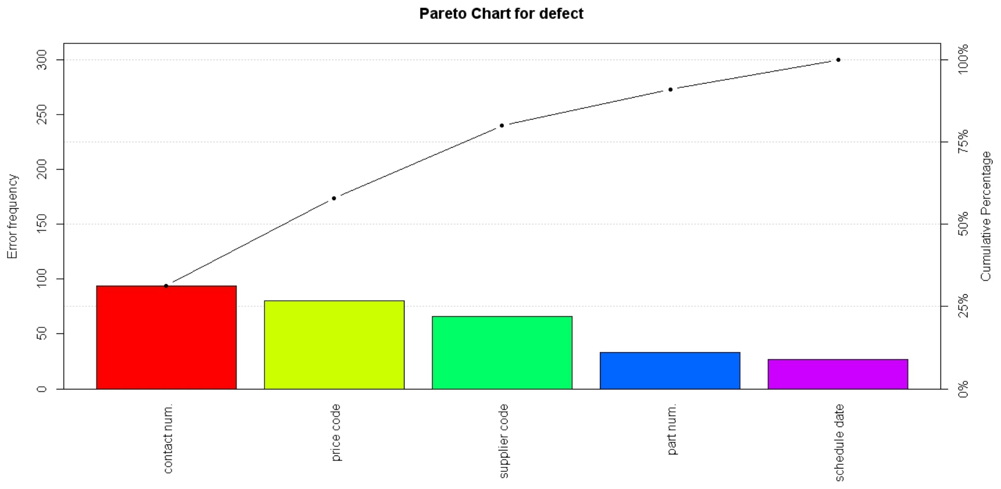


Figure 9: Third Implementation

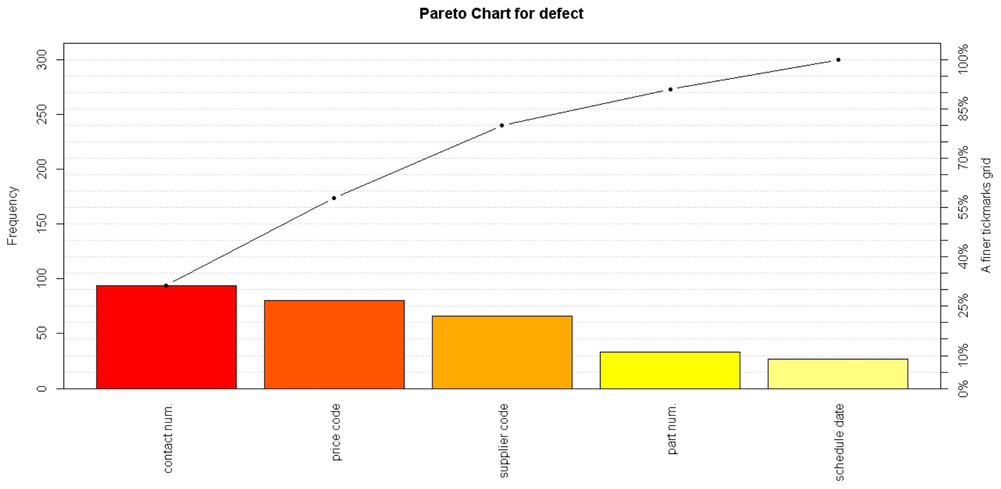


Figure 10: Fourth Implementation

6.5 Cause and Effect Diagrams

The cause and effect diagram is also known as "Ishikawa diagram", and has been widely used in Quality Management. It is one of the Seven Basic Tools of Quality.

```
cause.and.effect(cause=list(
  Measurements=c("Micrometers", "Microscopes", "Inspectors"),
  Materials=c("Alloys", "Lubricants", "Suppliers"),
  Personnel=c("Shofts", "Supervisors", "Training", "Operators"),
  Environment=c("Condensation", "Moisture"),
  Methods=c("Brake", "Engager", "Angle"),
  Machines=c("Speed", "Lathes", "Bits", "Sockets")),
effect="Surface Flaws")
```

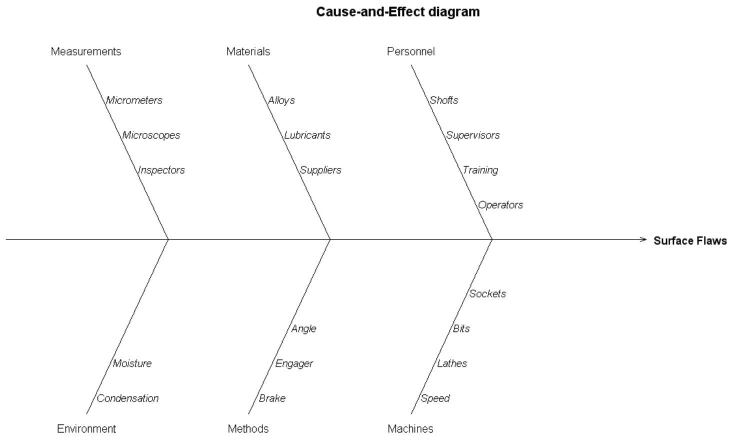


Figure 11:

6.5.1 Implementation with Six Sigma Package

```
effect <- "Flight Time"
causes.gr <- c("Operator", "Environment", "Tools", "Design",
"Raw.Material", "Measure.Tool")
causes <- vector(mode = "list", length = length(causes.gr))
causes[1] <- list(c("operator #1", "operator #2", "operator #3"))
causes[2] <- list(c("height", "cleaning"))
causes[3] <- list(c("scissors", "tape"))
causes[4] <- list(c("rotor.length", "rotor.width2", "paperclip"))
causes[5] <- list(c("thickness", "marks"))
causes[6] <- list(c("calibrate", "model"))
ss.ceDiag(effect, causes.gr, causes, sub = "Paper Helicopter Project")
```

6.6 Constructing Process Maps

```
inputs.overall<-c("operators", "tools", "raw material", "facilities")
outputs.overall<-c("helicopter")
steps<-c("INSPECTION", "ASSEMBLY", "TEST", "LABELING")
```

```
#Inputs of process "i" are inputs of process "i+1"
input.output<-vector(mode="list",length=length(steps))
input.output[1]<-list(c("sheets", "..."))
input.output[2]<-list(c("sheets"))
input.output[3]<-list(c("helicopter"))
input.output[4]<-list(c("helicopter"))
```

Parameters of each process

```
x.parameters<-vector(mode="list",length=length(steps))
```

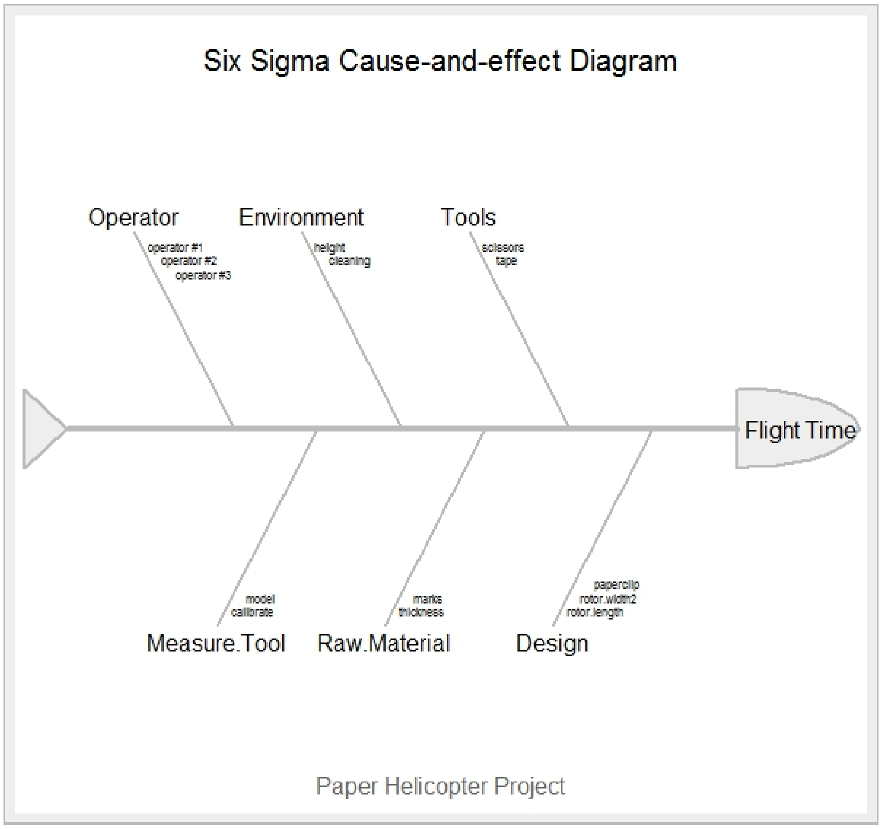


Figure 12:

```
x.parameters[1]<-list(c(list(c("width", "NC")),list(c("operator", "C")),
                        list(c("Measure pattern", "P")), list(c("discard", "P"))))
x.parameters[2]<-list(c(list(c("operator", "C")),list(c("cut", "P")),
                        list(c("fix", "P")), list(c("rotor.width", "C")),
                        list(c("rotor.length",
                                                                     list(c("paperclip", "C"))
x.parameters[3]<-list(c(list(c("operator", "C")),
list(c("throw", "P")),
                        list(c("discard", "P")),
                        list(c("environment", "N"))))
x.parameters[4]<-list(c(list(c("operator", "C")),
list(c("label", "P"))))
```

```
x.parameters
```

```
#Features of each process
y.features<-vector(mode="list",length=length(steps))
y.features[1]<-list(c(list(c("ok", "Cr"))))
y.features[2]<-list(c(list(c("weight", "Cr"))))
y.features[3]<-list(c(list(c("time", "Cr"))))
y.features[4]<-list(c(list(c("label", "Cr"))))
y.features
ss.pMap(steps, inputs.overall, outputs.overall,
        input.output, x.parameters, y.features,
        sub="Paper Helicopter Project")
```

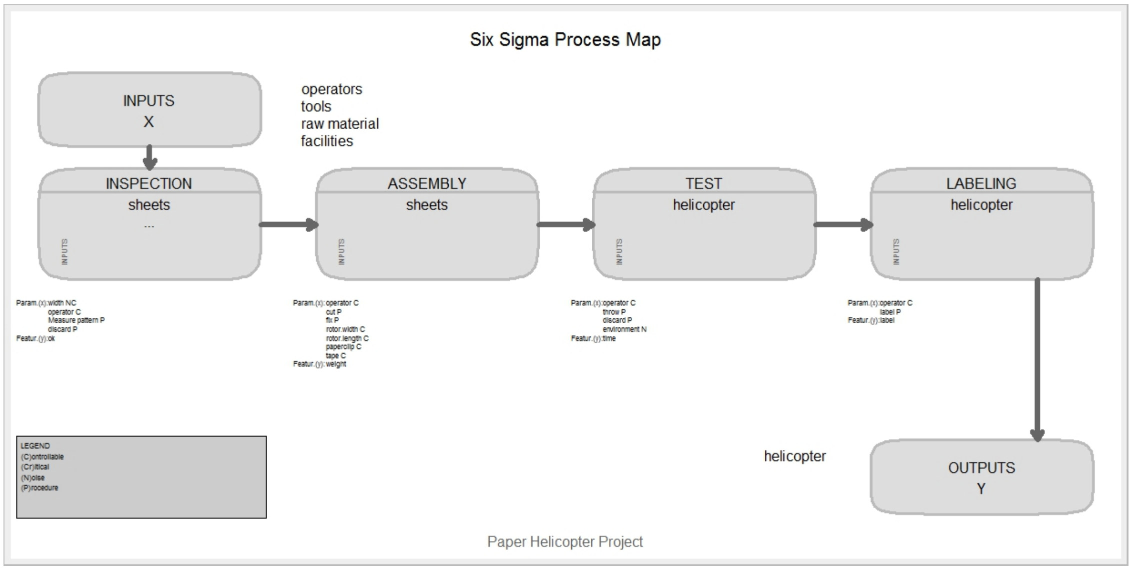


Figure 13:

7 More on Control Charts

7.1 Control Chart Selection

- Correct control chart selection is a critical part of creating a control chart. If the wrong control chart is selected, the control limits will not be correct for the data.

- The type of control chart required is determined by the type of data to be plotted and the format in which it is collected.

- Data collected is either in variables or attributes format, and the amount of data contained in each sample (subgroup) collected is specified.

- **Variables data** is defined as a measurement such as height, weight, time, or length. Monetary values are also variables data.

  * Generally, a measuring device such as a weighing scale, vernier, or clock produces this data.
  * Another characteristic of variables data is that it can contain decimal places e.g. 3.4, 8.2.

- **Attributes data** is defined as a count such as the number of employees, the number of errors, the number of defective products, or the number of phone calls. A standard is set, and then an assessment is made to establish if the standard has been met.

  * The number of times the standard is either met or not is the count. Attributes data never contains decimal places when it is collected, it is always whole numbers, e.g. 2, 15.

7.2 Attribute Control Charts

- The Shewhart control chart plots quality characteristics that can be measured and expressed numerically. We measure weight, height, position, thickness, etc. If we cannot represent a particular quality characteristic numerically, or if it is impractical to do so, we then often resort to using a quality characteristic to sort or classify an item that is inspected into one of two "buckets".
- An example of a common quality characteristic classification would be designating units as "conforming units" or "nonconforming units".
- Another quality characteristic criteria would be sorting units into "non defective" and "defective" categories. Quality characteristics of that type are called ***attributes***.
- *Note that there is a difference between "nonconforming to an engineering specification" and "defective" – a nonconforming unit may function just fine and be, in fact, not defective at all, while a part can be "in spec" and not fucntion as desired (i.e., be defective).*
- Examples of quality characteristics that are attributes are the number of failures in a production run, the proportion of malfunctioning wafers in a lot, the number of people eating in the cafeteria on a given day, etc.

7.3 Types of Attributes Control Charts

- Control charts dealing with the proportion or fraction of defective product are called ***p-charts*** (for proportion).
- Control charts dealing with the number of defective product are called ***np-charts***.
- Control charts dealing with the number of defects or nonconformities are called ***c-charts*** (for count).
- There is another chart which handles defects per unit, called the ***u-chart*** (for unit). This applies when we wish to work with the average number of nonconformities per unit of product.

7.4 p-charts

- The p-chart is a type of control chart used to monitor the **proportion of nonconforming units** in a sample, where the sample proportion nonconforming is defined as the ratio of the number of nonconforming units to the sample size, n.
- The p-chart only accommodates dichotomous PASS/FAIL-type inspection as determined by a series of tests, effectively applying the specifications to the data before they are plotted on the chart.
- Other types of control charts display the magnitude of the quality characteristic under study, making troubleshooting possible directly from those charts.
- A p-chart is an attributes control chart used with data collected in subgroups of varying sizes. Because the subgroup size can vary, it shows a proportion on nonconforming items rather than the actual count.
- *p-charts show how the process changes over time. The process attribute (or characteristic) is always described in a yes/no, pass/fail, go/no go form.*
- Example: use a p-chart to plot the proportion of incomplete insurance claim forms received weekly. The subgroup would vary, depending on the total number of claims each week.

7.5 np-charts

The np-chart is a type of control chart used to monitor the number of nonconforming units in a sample. An np-chart is an *attributes* control chart used with data collected in subgroups that are the **same size**.

It is an adaptation of the p-chart and used in situations where personnel find it easier to interpret process performance in terms of concrete numbers of units rather than the somewhat more abstract proportion.

The np-chart differs from the p-chart in only the three following aspects:

- The control limits are
  $$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}$$
  , where n is the sample size and $\bar{p}$ is the estimate of the long-term process mean established during control-chart setup.
- The number nonconforming (np), rather than the fraction nonconforming (p), is plotted against the control limits.
- The sample size, n, is constant.

7.6 The c-chart

- In this chart, we plot the number of defectives (per batch, per day, per machine, per 100 feet of pipe, etc.).
- This chart assumes that defects of the quality attribute are rare, and the control limits in this chart are computed based on the Poisson distribution (distribution of rare events).
- The c-chart is a type of control chart used to monitor "count"-type data, typically total number of nonconformities per unit.It is also occasionally used to monitor the total number of events occurring in a given unit of time.
- The c-chart differs from the p-chart in that it accounts for the possibility of more than one nonconformity per inspection unit, and that (unlike the p-chart and u-chart) it requires a fixed sample size.
- The p-chart models "pass"/"fail"-type inspection only, while the c-chart (and u-chart) give the ability to distinguish between (for example) 2 items which fail inspection because of one fault each and the same two items failing inspection with 5 faults each; in the former case, the p-chart will show two non-conformant items, while the c-chart will show 10 faults.
- The Poisson distribution is the basis for the chart and requires the following assumptions:
  - ∗ The number of opportunities or potential locations for nonconformities is very large
  - ∗ The probability of nonconformity at any location is small and constant
  - ∗ The inspection procedure is same for each sample and is carried out consistently from sample to sample

7.7 The u-chart

- In this chart we plot the rate of defectives, that is, the number of defectives divided by the number of units inspected (the n; e.g., feet of pipe, number of batches).
- Unlike the C chart, this chart does not require a constant number of units, and it can be used, for example, when the batches (samples) are of different sizes.
- The u-chart is a type of control chart used to monitor "count"-type data where the sample size is greater than one, typically the average number of nonconformities per unit.
- The u-chart differs from the c-chart in that it accounts for the possibility that the number or size of inspection units for which nonconformities are to be counted may vary. Larger samples may be an economic necessity or may be necessary to increase the area of opportunity in order to track very low nonconformity levels.

8 The qcc R package - Other Types of Graph

8.1 Operating Characteristic (OC) Curves

- A common supplementary plot to standard quality control charts is the so-called operating characteristic or OC curve (see example below). One question that comes to mind when using standard variable or attribute charts is how sensitive is the current quality control procedure? Put in more specific terms, how likely is it that you will not find a sample (e.g., mean in an X-bar chart) outside the control limits (i.e., accept the production process as "in control"), when, in fact, it has shifted by a certain amount?
- This probability is usually referred to as the (beta) error probability, that is, the probability of erroneously accepting a process (mean, mean proportion, mean rate defectives, etc.) as being "in control."
- Note that operating characteristic curves pertain to the false-acceptance probability using the sample-outside-of- control-limits criterion only, and not the runs tests described earlier.
- Operating characteristic curves are extremely useful for exploring the power of our quality control procedure. The actual decision concerning sample sizes should depend not only on the cost of implementing the plan (e.g., cost per item sampled), but also on the costs resulting from not detecting quality problems. The OC curve allows the engineer to estimate the probabilities of not detecting shifts of certain sizes in the production quality.

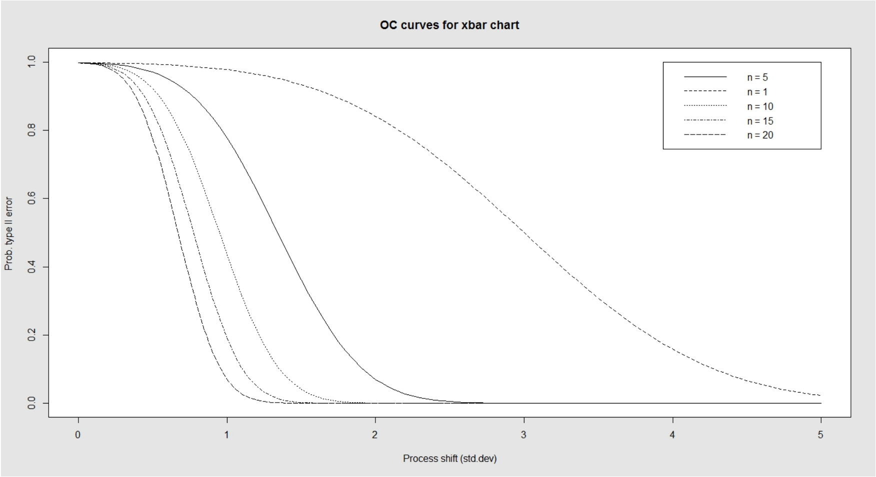


Figure 14:

8.1.1 pistonrings Data

```
data(pistonrings); attach(pistonrings);

diameter <- qcc.groups(diameter, sample)
beta <- oc.curves.xbar(qcc(diameter, type="xbar", nsigmas=3, plot=FALSE))
print(round(beta, digits=4))

# or to identify points on the plot use
## Not run: oc.curves.xbar(qcc(diameter,
    type="xbar", nsigmas=3, plot=FALSE), identify=TRUE)

detach(pistonrings)
```

8.2 Moving Average (MA) Chart

- To return to the piston ring example, suppose we are mostly interested in detecting small trends across successive sample means.
- For example, we may be particularly concerned about machine wear, leading to a slow but constant deterioration of quality (i.e., deviation from specification).
- Another way is to use some weighting scheme that summarizes the means of several successive samples; moving such a weighted mean across the samples will produce a moving average chart (as shown in the following graph).

8.2.1 GExponentially-weighted MA (EWMA) Chart

```
data(pistonrings)
attach(pistonrings)
diameter <- qcc.groups(diameter, sample)
q <- ewma(diameter[1:25,], lambda=0.2, nsigmas=3)
summary(q)

q <- ewma(diameter[1:25,], lambda=0.2, nsigmas=2.7,
newdata=diameter[26:40,], plot = FALSE)
summary(q)

plot(q)
```

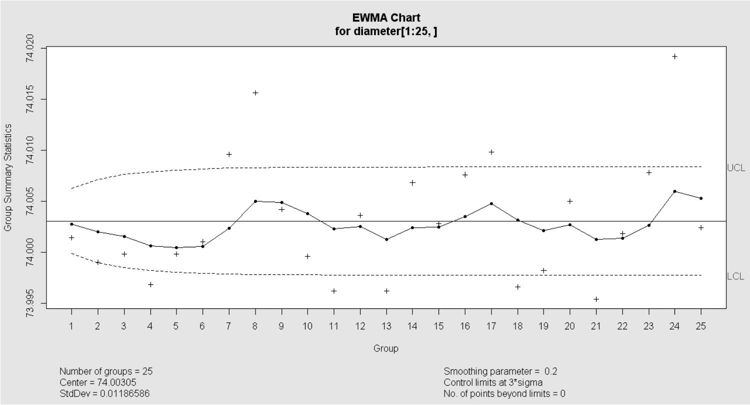


Figure 15:

```
> summary(q)

Call:
ewma(data = diameter[1:25, ], lambda = 0.2, nsigmas = 2.7,
newdata = diameter[26:40,     ], plot = FALSE)

ewma chart for diameter[1:25, ]
```

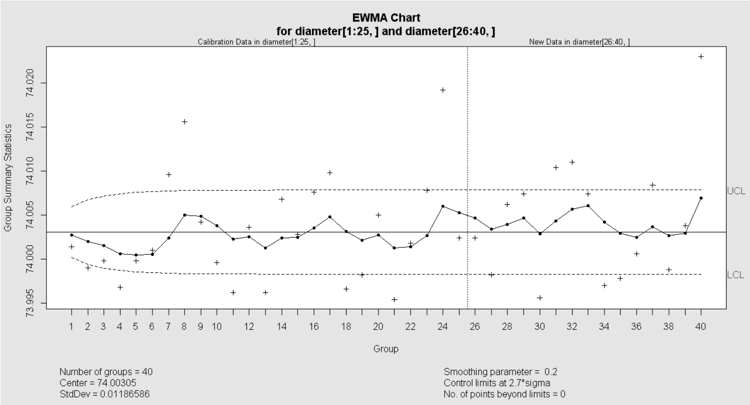


Figure 16:

```
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  74.00   74.00   74.00   74.00   74.01   74.02

Group sample size:  5
Number of groups:  25
Center of group statistics:  74.00305
Standard deviation:  0.01186586

Summary of group statistics in diameter[26:40, ]:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  74.00   74.00   74.00   74.00   74.01   74.02

Group sample size:  5
Number of groups:  15

Smoothing parameter: 0.2
Control limits:
         LCL      UCL
1   74.00018 74.00591
2   73.99938 74.00672
...
40  73.99827 74.00782
```

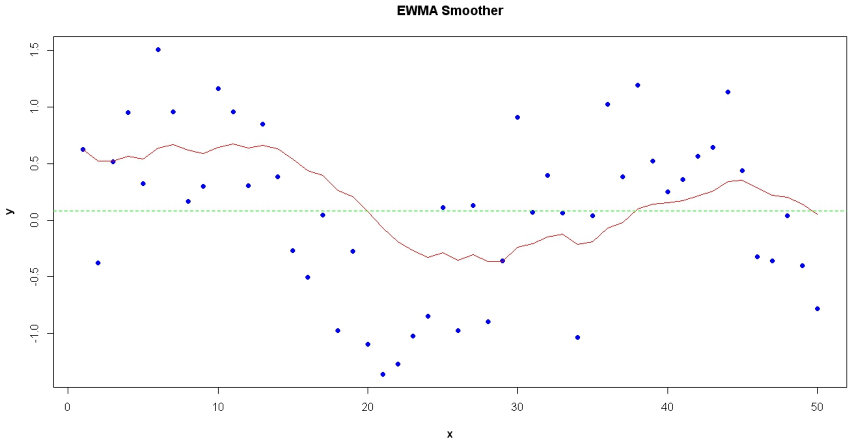


Figure 17:

8.2.2 EWMA : Individual observations

```
x <- c(33.75, 33.05, 34, 33.81, 33.46, 34.02, 33.68, 33.27, 33.49, 33.20,
33.62, 33.00, 33.54, 33.12, 33.84) # viscosity data (Montgomery, pag. 242)
q <- ewma(x, lambda=0.2, nsigmas=2.7)
summary(q)
```

```
x <- 1:50
y <- rnorm(50, sin(x/5), 0.5)
plot(x,y,pch=16,col="blue",font.lab=2)
lines(ewmaSmooth(x,y,lambda=0.1), col="red")
abline(h=mean(y),col="green",lty=2)
title("EWMA Smoother")
```

8.3 CUSUM charts

- CUSUM charts, while not as intuitive and simple to operate as Shewhart charts, have been shown to be more efficient in detecting small shifts in the mean of a process.
- In particular, analyzing ARL's for CUSUM control charts shows that they are better than Shewhart control charts when it is desired to detect shifts in the mean that are 2 sigma or less.

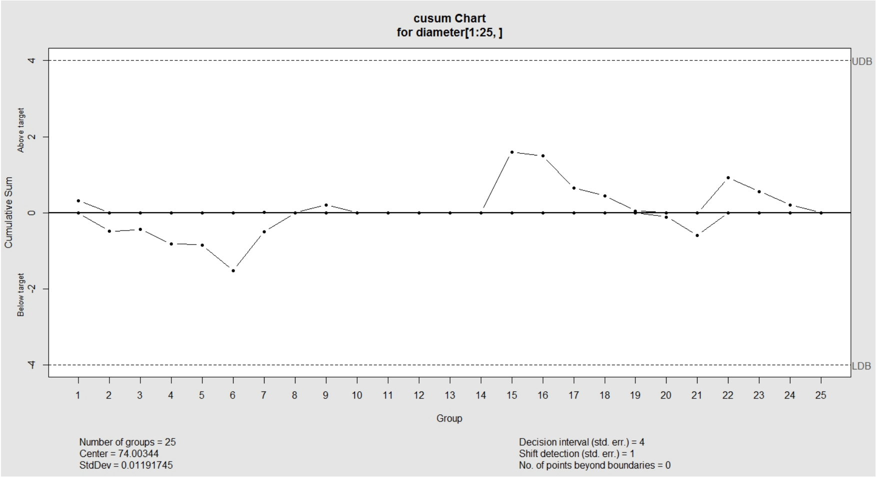


Figure 18:

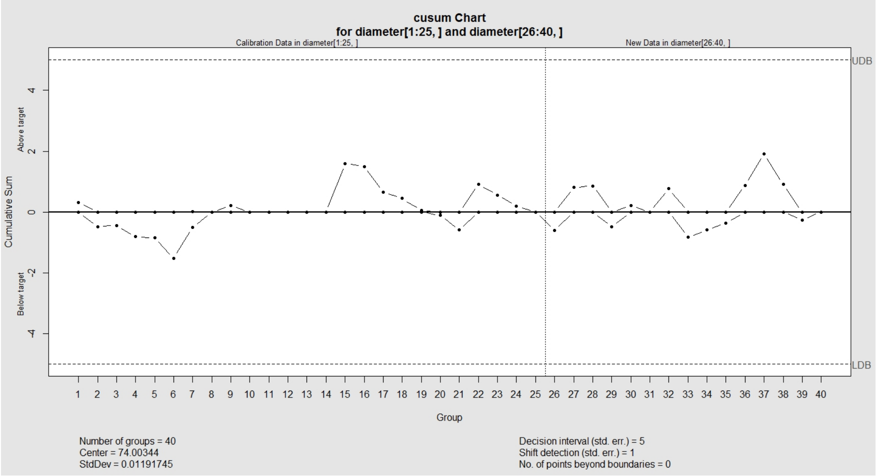


Figure 19:

8.4 Individual/Moving-Range chart

- The individual/moving-range chart is a type of control chart used to monitor variables data from a business or industrial process for which it is impractical to use rational subgroups.
- The chart is necessary in the following situations:
  - ∗ Where automation allows inspection of each unit, so rational subgrouping has less benefit.
  - ∗ Where production is slow so that waiting for enough samples to make a rational subgroup unacceptably delays monitoring
  - ∗ For processes that produce homogeneous batches (e.g., chemical) where repeat measurements vary primarily because of measurement error
- The ”chart” actually consists of a pair of charts: one, the individuals chart, displays the individual measured values; the other, the moving range chart, displays the difference from one point to the next.
- As with other control charts, these two charts enable the user to monitor a process for shifts in the process that alter the mean or variance of the measured statistic.

9 Process Capability

Process capability is the measure of process performance. Capability refers to the ability of a process to make parts that are well within engineering specifications. A capability study is done to answer the questions, *"Does the process need to be improved?"* and *"How much does the process need to be improved?"*

To define the study of process capability from another perspective, a capability study is a technique for analyzing the random variability found in a production process. In every manufacturing process there is variability. This variability may be large or small, but it is always present. It can be divided into two types:

- Variability due to common (random) causes
- Variability due to assignable (special) causes

The first type of variability can be expected to occur naturally within a process. It is attributed to common causes that behave like a constant system of chances. These chances form a unique and describable distribution. This variability can never be completely eliminated from a process. Variability due to assignable causes, on the other hand, refers to the variation that can be linked to specific or special causes. If these causes, or factors, are modified or controlled properly, the process variability associated with them can be eliminated. Assignable causes cannot be described by a single distribution.

9.1 Capability Study

- A capability study measures the performance potential of a process when no assignable causes are present (when it is in statistical control). Since the inherent variability of the process can be described by a unique distribution, usually a normal distribution, capability can be evaluated by utilizing this distributions properties.
- Simply put, capability is expressed as the proportion of in-specification process output to total process input.
- Capability calculations allow predictions to be made regarding quality, enabling manufacturers to take a preventive approach to defects. This statistical approach contrasts to the traditional approach to manufacturing, which is a two-step process: production personnel make the product, and quality control personnel inspect and eliminate those products that do not meet specifications.
- This is wasteful and expensive, since it allows time and materials to be invested in products that are not always usable. It is also unreliable, since even 100% inspection would fail to catch all defective products.
- Control Limits are Not an Indication of Capability
- Those new to SPC often believe they dont need capability indices. They think they can compare the control limits to the specification limits instead.
- This is not true, because control limits look at the distribution of averages and capability indices look at the distribution of individuals. The distribution of individuals will always spread out further than the distribution of averages.

9.2 What is Process Capability?

Distribution of averages compared to distribution of individuals, for the same sample data. Control limits (based on averages) would probably be inside specification limits, even though many parts are out of specification. This shows why you should not compare control limits to specification limits.

Therefore, the control limits are often within the specification limits, but the ± 3 Sigma distribution of parts is not. Subgroup averages follow more closely a normal distribution. This is why we can create control charts for processes that are not normally distributed. But averages cannot be used for capability calculations, because capability concerns itself with individual parts, or samples from a process. After all, parts, not averages, get shipped.

9.3 Capability Indices

Capability The uniformity of product which a process is capable of producing. Can be expressed numerically using CP, CR, CpK, and Zmax/3 when the data is normally distributed.

CP For process capability studies: CP is a capability index defined by the formula. CP shows the process capability potential but does not consider how centered the process is. CP may range in value from 0 to infinity, with a large value indicating greater potential capability. A value of 1.33 or greater is usually desired.

CR For process capability studies: the inverse of CP, CR can range from 0 to infinity in value, with a smaller value indicating a more capable process.

CpK For process capability studies: an index combining CP and K to indicate whether the process will produce units within the tolerance limits. CpK has a value equal to CP if the process is centered on the nominal; if CpK is negative, the process mean is outside the specification limits; if CpK is between 0 and 1, then some of the 6 sigma spread falls outside the tolerance limits. If CpK is larger than 1, the 6 sigma spread is completely within the tolerance limits. A value of 1.33 or greater is usually desired.

9.4 Interpreting Capability Indices

- The greater the CpK value, the better. A CpK greater than 1.0 means that the $6\sigma(\pm 3\sigma)$ spread of the data falls completely within the specification limits. A CpK of 1.0 means that one end of the 6σ spread falls on a specification limit. A CpK between 0 and 1 means that part of the 6σ spread falls outside the specification limits. A negative CpK indicates that the mean of the data is not between the specification limits.

- Since a CpK of 1.0 indicates that 99.73% of the parts produced are within specification limits, in this process it is likely that only about 3 out of 1,000 need to be scrapped or rejected. Why bother to improve the process beyond this point, since it will produce no reduction in scrap or reject costs? Improvement beyond just meeting specification may greatly improve product performance, cut warranty costs, or avoid assembly problems.

- Many companies are demanding CpK indexes of 1.33 or 2.0 of their suppliers products. A CpK of 1.33 means that the difference between the mean and specification limit is 4σ (since 1.33 is 4/3). With a CpK of 1.33, 99.994% of the product is within specification. Similarly a CpK of 2.0 is 6σ between the mean and specification limit (since 2.0 is 6/3).

- This improvement from 1.33 to 2.0 or better is sometimes justified to produce more product near the optimal target. Depending on the process or part, this may improve product performance, product life, customer satisfaction, or reduce warranty costs or assembly problems.

- Continually higher CpK indexes for every part or process is not the goal, since that is almost never economically justifiable. A cost/benefit analysis that includes customer satisfaction and other true costs of quality is recommended to determine which processes should be improved and how much improvement is economically attractive.

9.5 Process Capability Analysis

- Process capability compares the output of an in-control process to the specification limits by using capability indices.

- The comparison is made by forming the ratio of the spread between the process specifications (the specification ”width”) to the spread of the process values, as measured by 6 process standard deviation units (the process ”width”).

9.6 Intepreting Process Capability Indices

- **CP**
  Historically, this is one of the first capability indexes used. The "natural tolerance" of the process is computed as 6s . The index simply makes a direct comparison of the process natural tolerance to the engineering requirements. Assuming the process distribution is normal and the process average is exactly centered between the engineering requirements, a CP index of 1 would give a "capable process." However, to allow a bit of room for process drift, the generally accepted minimum value for CP is 1.33. In general, the larger CP is, the better. The CP index has two major shortcomings. First, it cannot be used unless there are both upper and lower specifications. Second, it does not account for process centering. If the process average is not exactly centered relative to the engineering requirements, the CP index will give misleading results. In recent years, the CP index has largely been replaced by CPK (see below).

- **CPM**
  A CPM of at least 1 is required, and 1.33 is preferred. CPM is closely related to CP. The difference represents the potential gain to be obtained by moving the process mean closer to the target. Unlike CPK, the target need not be the center of the specification range.