Applied Statistics and Data Analysis

Part I

Statistical Process and Quality Control

Introduction

Quality and Process Control

Definition of Quality
- Quality means fitness for use. and
- Quality is inversely proportional to variability.
- My definition: *Is* and *should* are the same.

Statistical Process Control (SPC)
- Statistical process control is, first and foremost, a way of thinking which happens to have some tools attached.

The Magnificent Seven
1. histogram 2. check sheet 3. Pareto chart 4. defect concentration diagram 5. cause-and-effect diagram 6. control chart 7. scatter diagram

Control Charts

The basis of a control chart is a statistical hypothesis test.

Hypothesis test

Question:
is $|\bar{x} - \mu_0|$ significant?
- μ_0 target value
- $\bar{x}$ Arithmetic mean of the measurements

Hypothesis
Two-sided statistical test to check the two alternative hypotheses:

$$\begin{aligned} H_0 :&\mu_0 = \bar{x} \text{ i.e. process is not disturbed} \\ H_1 :&\mu_0 \neq \bar{x} \text{ i.e. process is disturbed} \end{aligned} \tag{1}$$

Statistical test:

Test statistic (z-test, since σ is known)

$$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \tag{2}$$

Critical value i.e. q-quantile of the normal distribution

$$P(|z| \leq z_q) = q \tag{3}$$

with $q = 1 - \frac{\alpha}{2}$, where $\alpha = 0.0027 \to z_q \approx 3$

Statistical conclusion
If $\leq |z|\ z_q \to$ accept null hypothesis, i.e. process is not disturbed.
If $< |z|\ z_q \to$ reject null hypothesis, i.e. process is disturbed.

Reverse it and determine acceptance and rejection limits of the test!

Control limits

$$UCL = \mu_0 + z_q \frac{\sigma}{\sqrt{n}} \qquad LCL = \mu_0 - z_q \frac{\sigma}{\sqrt{n}} \tag{4}$$

Statistical conclusion
If $LCL \leq \bar{x} \leq UCL \to$ process is not disturbed.
If $\bar{x} < LCL or UCL < \bar{x} \to$ process is disturbed.

Problem: In general, process standard deviation is unknown.
- Monitoring the mean and the variation of a process.
- First, monitoring the variation, then (if variation under control) monitoring the mean.

Solution: Control charts! :)

The Control Chart

No.	sample values						mean	sd	range
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	$\bar{x}_1$	s_1	R_1
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2n}	$\bar{x}_2$	s_2	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{in}	$\bar{x}_i$	s_i	R_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_{k1}	x_{k2}	$\cdots$	x_{kj}	$\cdots$	x_{kn}	$\bar{x}_k$	s_k	R_k

Figure 1: Data Set with Mean, Standard Deviation and Range

Mean values

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{5}$$

Standard deviations

$$s_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2} \tag{6}$$

Ranges

$$R_i = max\{x_{ij} | j \in \{1,...,n\}\} - min\{x_{ij} | j \in \{1,...,n\}\} \tag{7}$$

for all $i \in \{1,...,k\}$

Control Chart for $\bar{x}$ and R

R Chart

Centerline

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{8}$$

Control limits

$$UCL = D_4\bar{R}; \quad LCL = D_3\bar{R} \tag{9}$$

$\bar{x}$ Chart based on R chart

Control limits

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}}; \quad LCL = \mu - 3\frac{\sigma}{\sqrt{n}} \tag{10}$$

Problem: μ and σ are in general unknown and must be estimated from the process data.
Two-stage process:
- Make sure that the process standard deviation (R chart) is under statistical control. That is, if some samples are out of bounds, it is recommended to omit these measurements and recalculate the limits.
- Use $\bar{R}$ to estimate the process standard deviation.

Centerline
We know that for an independent sample $x_1, ..., x_n$ from a normal distribution with parameters μ and σ the mean

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{11}$$

satisfies

$$E(\bar{x}) = \mu \quad and \quad Var(\bar{x}) = \frac{\sigma^2}{n} \tag{12}$$

The mean is an unbiased estimator with the standard error

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \tag{13}$$

Assumption: R chart is under statistical control.
- The value $\bar{R}$ is a reliable estimate for the mean range.
- The value $\bar{R}$ is a reliable estimate for the process standard deviation

$$\hat{\sigma} = \frac{\bar{R}}{d_2} \tag{14}$$

Any samples excluded for construction of the R chart should also be disregarded for construction of the $\bar{x}$ chart. This results in a sample of $k^\star$ valid samples, (where $k^\star$ denotes the reduced number of samples). Mean values of $\bar{x}_1, ..., \bar{x}_{k^\star}$ provide an estimate of μ, i.e

$$\bar{\bar{x}} = \frac{1}{k^\star}\sum_{i=1}^{k^\star} \bar{x}_i \tag{15}$$

Control limits

$$\begin{aligned} UCL =&\bar{\bar{x}} + 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_2\bar{R} \\ UCL =&\bar{\bar{x}} - 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_2\bar{R} \end{aligned} \tag{16}$$

Control Chart with $\bar{x}$ and s

s Chart

Centerline The centreline of the s chart is denoted by $\bar{s}$ and is calculated from the arithmetic mean of the standard deviations

$$\bar{s} = \frac{1}{k}\sum_{i=1}^{k} s_i \tag{17}$$

Control limits

$$UCL = B_4\bar{s}; \qquad LCL = B_3\bar{s} \tag{18}$$

$\bar{x}$ Chart based on s chart

Using an s chart of a process that is under control, the process standard deviation can be estimated by

$$\hat{\sigma} = \frac{\bar{s}}{c_4} \tag{19}$$

Any samples excluded for construction of the s chart should also be disregarded for construction of the $\bar{x}$ chart. This results in a sample of $k^{\star}$ valid samples, (where $k^{\star}$ denotes the reduced number of samples). Mean values of $\bar{x}_1, ..., \bar{x}_{k^{\star}}$ provide an estimate of μ, i.e

$$\hat{\mu} = \bar{\bar{x}} = \frac{1}{k^{\star}}\sum_{i=1}^{k^{\star}} \bar{x}_i \tag{20}$$

Control limits

$$\begin{aligned} UCL =& \bar{\bar{x}} + 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_3\bar{s} \\ UCL =& \bar{\bar{x}} - 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_3\bar{s} \end{aligned} \tag{21}$$

Individual Control Charts

Individual control charts have exactly one measurement per sample.
Problem: You cannot estimate variability from a single measurement.
Idea: Use variation of two adjacent measurements.

Moving ranges

$$MR_i = |x_{i+1} - x_i| \tag{22}$$

for all $i \in \{1, ..., n-1\}$.

Arithmetic mean of the moving ranges

$$\overline{MR} = \frac{1}{n-1}\sum_{i=1}^{n-1} MR_i \tag{23}$$

Estimated process standard deviation

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{\overline{MR}}{1.128} \tag{24}$$

Since two neighboring measurements were used to calculate the moving ranges we have $d_2 = 1.128$.

Centerline
The centerline for the individuals control chart is the arithmetic mean of the measured values.

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{25}$$

Control limits

$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128}; \qquad LCL = \bar{x} - 3\frac{\overline{MR}}{1.128} \tag{26}$$

Control Charts for Attributes Data – p Chart

Number of defectives under number tested is a discrete random variable.

Given: Random sample of size n, of which D parts are defective We know: The number of defective D under n examined parts follows a binomial distribution with the unknown probability p of success.

Estimated probability

$$\hat{p} = \frac{D}{n} \tag{27}$$

Variance

$$Var(\hat{p}) = \frac{p(1-p)}{n} \tag{28}$$

Given:
- k random samples with $n_1, ..., n_k$ values.
- Each of these samples contains $d_1, ..., d_k$ defective products.

k relative frequencies

$$p_1 = \frac{d_1}{n_1}, \dots, p_k = \frac{d_k}{n_k} \tag{29}$$

Centerline
The centreline and the control limits of a p chart are again determined from a stable trial run with $k^{\star}$ valid samples.
Again $k^{\star} \leq k$ is the reduced number of samples.
Distinguish 2 cases:
1. The sample sizes $n_1, ..., n_k$ are all equal to n.
2. The sample sizes are not all equal.

Case 1

Centerline

$$\bar{p} = \frac{1}{k^{\star}}\sum_{i=1}^{k^{\star}} p_i \tag{30}$$

Control limits

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \qquad LSL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{31}$$

Case 2

Centerline

$$\bar{p} = \frac{d_1 + \cdots + d_{k^{\star}}}{n_1 + \cdots + n_{k^{\star}}} \tag{32}$$

Control limits

$$UCL_i = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}; \qquad LSL_i = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \tag{33}$$

The control limits now depend on the index i.

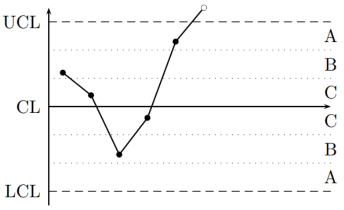
Statistical Properties of Control Charts

Aim of process control using control charts: Keep the process under statistical control. Or, if it is not at the beginning, to put it into statistical control by improving production conditions.

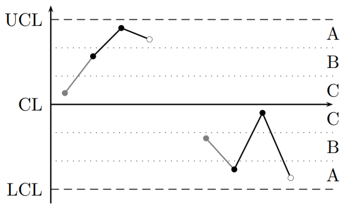
Interpretation of Control Charts

Western Electric Rules

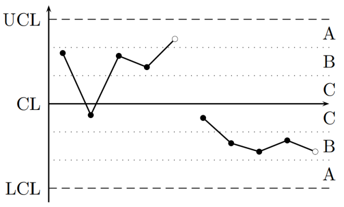
1. Any single data point falls outside the limit defined by UCL and LCL (beyond the 3σ-limit).
2. 2. Two out of three consecutive points fall beyond the limit defined by $\frac{2}{3}$ UCL and $\frac{2}{3}$ LCL on the same side of the centreline (beyond the 2σ-limit).
3. Four out of five consecutive points fall beyond the limit defined by $\frac{1}{3}$UCL and $\frac{1}{3}$ LCL on the same side of the centreline (beyond the 2σ-limit).
4. Nine consecutive points fall on the same side of the centreline (so-called run).



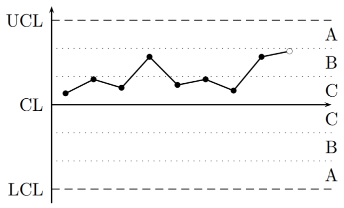
Rule 1: Any point beyond zone A.



Rule 2: Two out of three consecutive points fall on the same side in zone A or beyond.



Rule 3: Four out of five consecutive points fall on the same side in zone B or beyond.



Rule 4: Nine consecutive points fall on the same side of the centreline.

Type I Error and Type II Error

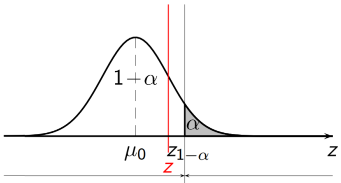
When monitoring a production process with a control chart, as with any statistical test, there are two wrong decisions possible.

$$\begin{aligned} H_0 &: \mu_0 = \mu \text{ i.e. process is not disturbed} \\ H_1 &: \mu_0 \neq \mu \text{ i.e. process is disturbed, } \mu_1 \text{ is true} \end{aligned} \tag{34}$$

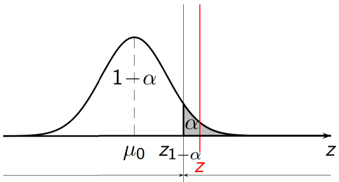
Denoted by:
- μ_0 the target value of the process
- μ the considered statistic, eg. $\mu = \bar{x}$ or $\mu = R$
- μ_1 the true value of the considered statistic.

Two wrong decisions possible:
- If a true null hypothesis H_0 is rejected we make a type I error. An intervention in the process is necessary, because the control limits are exceeded, although the process is not disturbed. This is called a false alarm.
- If a false null hypothesis H_0 is accepted we make a type II error. There is no intervention, since the control limits are not exceeded, although the process is disturbed. This is called an omitted alarm

Type I Error

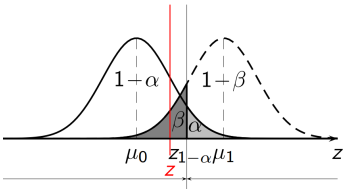


Let h_0 be true: Since $z < z_{1-\alpha}$ the null hypothesis is accepted. This is the right decision, which is made with probability $1 - \alpha$.

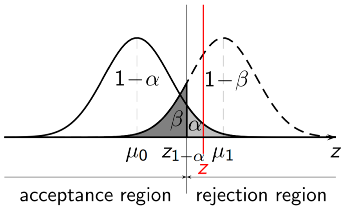


Let h_0 be true: Since $z \geq z_{1-\alpha}$ the null hypothesis is rejected. This is the wrong decision (type I error), which is made with probability α.

Type II Error



Let H_0 be false, H_1 true, i.e. the dashed density is true: Since $z < z_1 - \alpha$ the null hypothesis is accepted. This is the wrong decision (type II error), which is made with probability β.



Let H_0 be false, H_1 true, i.e. the dashed density is true: Since $z \leq z_1 - \alpha$ the null hypothesis is rejected. This is the correct decision, which is made with probability $1 - \beta$ (power).

Power Function and Operating Characteristic

The power of a hypothesis test is the probability $1 - \beta$ that the test correctly rejects the null hypothesis when the alternative hypothesis is true, i.e.

$$\text{power} = P(\text{reject } H_0 | H_1 \text{ is true}) = 1 - \beta \tag{35}$$

Power function
Probability to reject the null hypothesis H_0 if μ_1 is true.

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}, \tag{36}$$

or $\mu_1 = \delta\sigma + \mu_0$.
The variable δ is a normalized measure for the deviation of the disturbed from the undisturbed process in units of σ.
In statistical process control the power function is denoted by

$$g(\mu_1) = g(\delta\sigma + \mu_0) = \tilde{g}(\delta). \tag{37}$$

It is a measure for the probability of an intervention in the process.

Undisturbed Process
For an undisturbed process, i.e. $\mu = \mu_0$, we have

$$g(\mu_1) = \tilde{g}(0) = \alpha. \tag{38}$$

Disturbed Process
For a disturbed process we have

$$\tilde{g}(\delta) = \Phi(\delta\sqrt{n} - z_q, 0, 1) + \Phi(-\delta\sqrt{n} - z_q, 0, 1) \tag{39}$$

with Φ being

$$\Phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \tag{40}$$

Average Run Length

If the i_{RL}-th sample is the first to result in an intervention, i.e. this sample is beyond the control limits, then i_{RL} is called the run length of the control chart.
The ARL is the expected value of the probability i.e. the likelihood of exceeding the control limits when performing a test.

$$ARL(\delta) = \frac{1}{p(\mu)} = \frac{1}{g(\mu_1)} = \frac{1}{\tilde{g}(\delta)} \tag{41}$$

Undisturbed Process
If we have an undisturbed process with $\mu_1 = \mu_0$, i.e. with $\delta = 0$, then it follows from equation 38 that

$$ARL(0) = \frac{1}{\alpha}. \tag{42}$$

Disturbed Process
To determine the average run length of a disturbed process with $\mu = \mu_1$, i.e. $\sigma = \frac{\mu_1 - \mu_0}{\sigma}$ we use the power function $\tilde{g}(\delta)$ from equation 39.

Process Capability

The specification limit (SL) is defined by

$$SL = \frac{USL + LSL}{2} \tag{43}$$

This performance is measured with so-called capability process ratios (PCR). The simplest process capability index is

$$C_p = \frac{USL - LSL}{6\sigma} \tag{44}$$

The capability process ratio C_p expresses the ratio of the width of the tolerance range to the width of the process range.

- $C_p = 1$ implies a reject rate of $\alpha \cdot 100\% = 0.27\%$.
- $C_p < 1$ implies a reject rate of more than $\alpha \cdot 100\% = 0.27\%$, i.e. the process capability is not guaranteed.
- $C_p > 1$ implies a reject rate of less than $\alpha \cdot 100\% = 0.27\%$,i.e. the process capability is guaranteed.

Control Charts with Memory

Classical Shewhart control charts
- Decision to interfere with the manufacturing process is based on the result of the current sample.
- No consideration of the development of the manufacturing process in the past (except with western electric rules).

Modern control charts
- have a memory.

Idea
Linear combination of mean values $\bar{x}_j$ of samples from the past

$$y_i = \alpha_i + \sum_{j=1}^{i} \beta_j \bar{x}_j, \tag{45}$$

where α_i and the weights $\beta_1, ..., \beta_i$ can be arbitrary real numbers where the sum of all $\beta s = 1$.
Depending on how the weights are chosen, we get another type of control chart.

CUSUM - Cumulative Sum Control Chart

The CUSUM chart plots the cumulative sums of deviations of measurement values from the target value.

Recursive procedure
Using two statistics C^+, resp. C^- the CUSUM chart sums up deviations above, resp. below the target value

$$\begin{aligned} C_i^+ &= max\{0, \bar{x}_i - (\mu_0 + K) + C_{i-1}^+\}, \\ C_i^- &= max\{0, (\mu_0 - K) - \bar{x}_i + C_{i-1}^+\}. \end{aligned} \tag{46}$$

C^+ and C^- only sum up deviations from the target value, which are greater than the reference value K. The starting values of the recursion are $C^+ = 0$ and $C^- = 0$.

If a shift of Δ is to be detected then set

$$K = \frac{\Delta}{2}. \tag{47}$$

The constant K is called reference value.

If the process is under controll the expected values of the statistic are both 0.
If the process is not under control, then the statistic sums up the deviations. If the sum of de deviations (C^+ and C^-) exceed the decision intercal H, then we should stop the process and look for the cause.

Rule of thumb for choosing the constants K and H: Let $\hat{\sigma}$ be an estimate for the process standard deviation. - reference value $= K\frac{\hat{\sigma}}{2}$ - decision interval $H = 5\hat{\sigma}$

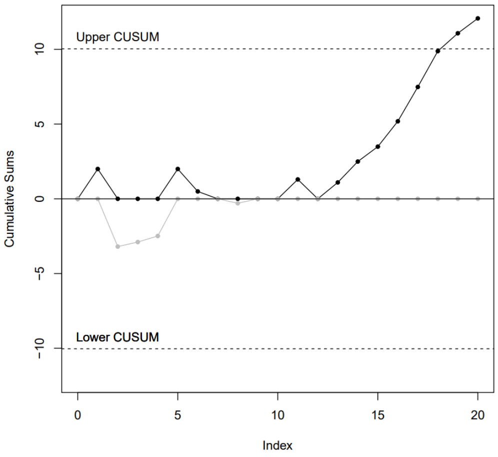


Figure 2: CUSUM-Chart

EWMA - Exponentially Weighted Moving Average

Idea
- Monitoring means.
- Wheights β_j decay exponentially.

Smoothing parameter λ
Lambda lays between 0 and 1. The smoothing parameter λ determines the influence of the previous sample mean $\bar{x}_j$ on the statistic. The smaller λ, the more values $\bar{x}_j$ are used for the decision. For $\lambda = 1$ only one sample is used and we get the well-known Shewhart $\bar{x}$ chart.

Weights

$$\begin{aligned} \alpha_i &= (1-\lambda)^1 \mu_0 \\ \beta_j &= \lambda(1-\lambda)^{i-j} \text{with } j \in \{1, 2, ..., i\}. \end{aligned} \tag{48}$$

Statistics

$$y_i = (1-\lambda)^i \mu_0 + \lambda \sum_{j=1}^{i} (1-\lambda)^{i-j} \bar{x}_j \tag{49}$$

Start: $y_0 = \mu_0$

The same with recursion

$$y_i = (1-\lambda) y_{i-1} + \lambda \bar{x}_j \tag{50}$$

Assumptions
If the process is under control, then $\bar{x}_i$ comes from a normal distribution with the expected value μ_0 and the standard deviation $\frac{\sigma}{\sqrt{n}}$.
The standard deviation is either known or can be estimated from data. The statistic y_i is then also normally distributed with $E(y_i) = \mu_0$ and

$$Var(y_i) = \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\frac{\sigma^2}{n} \tag{51}$$

Control limits
These assumptions lead to the 3σ control limits:

$$\begin{aligned} UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}} \\ LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{52}$$

The asymtotic control limits are:

$$\begin{aligned} UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \\ LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{53}$$

Estimate of process standard error

$$\hat{\sigma} = \frac{\bar{s}}{c_4} \tag{54}$$

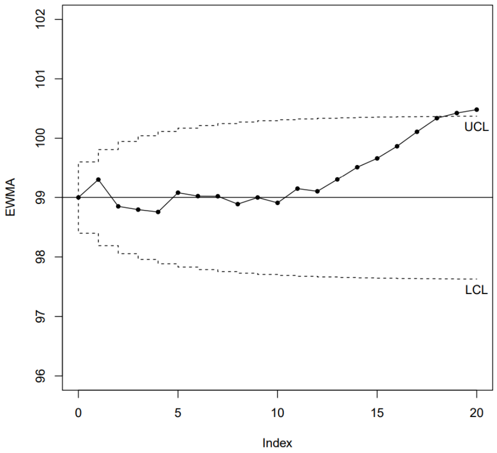


Figure 3: EWMA-Chart

Part II
Multiple Regression

Part III
Design of Experiment