INSE6300 Jia Yuan Yu

2: Control charts

Concordia February 9, 2016

Suppose that you produce integrated circuits (computer chips), voltage measurements can be taken on these chips. Many things can go wrong in the supply chain (silicon wafer, photoresist, etching, etc.). How do we quickly detect that something went wrong somewhere?

The general setting is the following: we take measurements $X_1, X_2, \ldots$ In the beginning, for chips $1, 2, \ldots, \nu$, everything is fine and the measurements are distributed according to a known distribution function F. However, at an unknown time ν, something breaks, and the subsequent chips $\nu+1, \nu+2, \ldots$ generate measurements according to another unknown distribution F'. The unknown ν is called the changepoint. We make the i.i.d. assumption for the subsequences before and after ν.

Control charts are tools for detecting changes in the distribution that characterize a sequence of measurements (random variables).

1 Shewhart $\bar{X}$ control chart

The Shewhart $\bar{X}$ chart is designed to detect changes in distribution that affect the mean.

Suppose that the measurements are $X_1, X_2, \ldots, X_n$ and i.i.d. We assume that the mean μ and the variance σ^2 are given (e.g., estimated using data $\ldots, X_{-2}, X_{-1}, X_0$). The Shewhart $\bar{X}$ chart simply raises an alarm when the empirical average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ falls outside the region

$$\left[\mu - \frac{3\sigma}{\sqrt{n}}, \mu + \frac{3\sigma}{\sqrt{n}}\right]$$

of three times the standard deviation $\sigma/\sqrt{n}$ around the mean μ.

This is repeated for the following batches of measurements:

$$\begin{aligned} X_{n+1}, X_{n+2}, \ldots, X_{2n}, \\ X_{2n+1}, X_{2n+2}, \ldots, X_{3n}, \\ \ldots \end{aligned}$$

1.1 Why does this work?

Suppose that we use data $X_{-n+1}, \ldots, X_{-2}, X_{-1}, X_0$ to estimate μ:

$$\hat{\mu} = \frac{X_{-n+1} + \ldots + X_{-2} + X_{-1} + X_0}{n}.$$

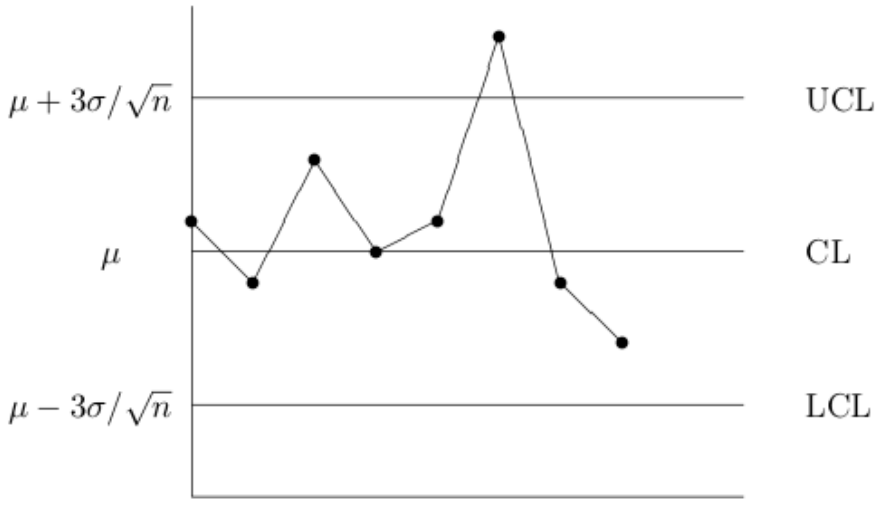

Figure 4.1: Shewhart $\overline{X}$-chart with control lines.

Figure 1: From A. Di Bucchianico, Applied Statistics, Technische Universiteit Eindhoven, 2008.

By the Hoeffding Inequality (Lecture 1), $\hat{\mu}$ is a good estimate of the true mean μ. Similarly,

$$\hat{\mu}' = \frac{X_1 + \ldots + X_n}{n}.$$

is also a good estimate of the true mean μ. Hence, with high probability, we have $|\hat{\mu} - \hat{\mu}'| \leq \varepsilon$.

How well does this alarm perform? Let's look at the probability of false alarm and the probability of missing a change.

1.2 Probability of false alarm

The probability of false alarm is

$$1 - \mathbb{P}\left(\mu - \frac{3\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{3\sigma}{\sqrt{n}}\right) = \Phi(-3) + (1 - \Phi(2)) = 0.0027.$$

1.3 Probability of missed change

Suppose that there was an incident is the supply chain at time $7n$, such that the measurement $X_{7n+1}, \ldots, X_{8n}$ are distributed according to a new distribution with a new mean γ.

The probability of missed change is

$$\begin{aligned}&\mathbb{P}\left(\mu - \frac{3\sigma}{\sqrt{n}} \leq \bar{X}_{8n} \leq \mu + \frac{3\sigma}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(\mu - \gamma + \gamma - \frac{3\sigma}{\sqrt{n}} \leq \bar{X}_{8n} \leq \mu - \gamma + \gamma + \frac{3\sigma}{\sqrt{n}}\right)\end{aligned}$$

This can be easily approximated with the central limit theorem and integrating (cf. Figure 1.3).

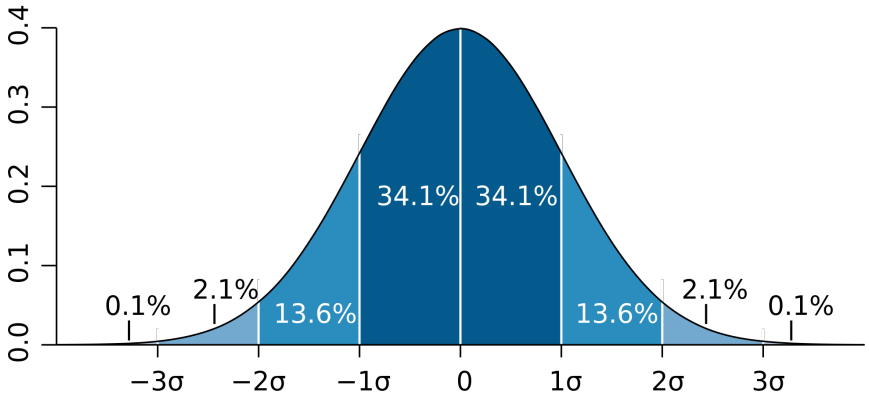

Figure 2: From Wikimedia

1.4 Examples

- Check that $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of $\bar{X}_n$.
- What if instead of 3 standard deviations, we raise an alarm at 2 standard deviations?
- Generate an i.i.d. random sequence in R and test the Shewhard $\bar{X}$ chart.

2 Variance R control chart

What if an incident produces a change in distribution without changing the mean? If the variance changes, then we can use a control chart for variance.

Using sample variance is one approach to detect changes in the variance. However, there is another approach that requires less computation. For $i = \ldots, -1, 0, 1, \ldots$, let $X^i_{(1)}$ and $X^i_{(n)}$ denote the smallest and largest random variables in the dataset $X_{in+1}, \ldots, X_{(i+1)n}$. Let $R^i = X^i_{(n)} - X^i_{(1)}$ denote the range over the ith dataset. We define the interval:

$$[D_3(n)\bar{R}, D_4(n)\bar{R}],$$

where $\bar{R}$ is the true expectation of R^i for $i < 0$, i.e., before any change in distribution, and D_3 and D_4 are functions with values taken from handbooks[1]. This $\bar{R}$ can also be estimated if unknown:

$$\bar{R}_m = \frac{1}{m} \sum_{i=-1}^{-m} R^i.$$

n	c_4	d_2	A_2	D_3	D_4	$D_{.001}$	$D_{.999}$
2	0.7979	1.128	1.880	0.000	3.267		
3	0.8862	1.693	1.023	0.000	2.575		
4	0.9213	2.059	0.729	0.000	2.282	0.199	5.309
5	0.9400	2.326	0.577	0.000	2.115	0.367	5.484
6	0.9515	2.534	0.483	0.000	2.004	0.535	5.619
7	0.9594	2.704	0.419	0.076	1.924	0.691	5.730

Table 4.1: Control chart constants.

Figure 3: From A. Di Bucchianico, Applied Statistics, Technische Universiteit Eindhoven, 2008.

The R control chart simply raises an alarm when R^i, for $i \geq 0$, falls outside the interval $[D_3(n)\bar{R}, D_4(n)\bar{R}]$.

3 CUSUM

Suppose that the distribution before and after the changepoint are functions of a parameter θ, e.g.,

$$\begin{aligned} f_{\theta_0}, &\quad \theta = \theta_0 \quad \text{for } X_1, \ldots, X_\nu, \\ f_{\theta_1}, &\quad \theta = \theta_1 \quad \text{for } X_{\nu+1}, X_{\nu+2} \ldots. \end{aligned}$$

This parameter can be the mean, the variance, or any other parameter of the distribution function. In this setting, we can use the cumulative sum control chart to detect the changepoint.

Let

$$Z_i = \log\left(\frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)}\right), \quad \text{for } i = 1, 2, \ldots$$

and

$$\begin{aligned} C_1 &= 0, \\ C_i &= \max(0, C_{i-1} + Z_i), \quad \text{for } i \geq 2. \end{aligned}$$

[1]Montgomery, Douglas (2005). Introduction to Statistical Quality Control. Hoboken, New Jersey: John Wiley & Sons.

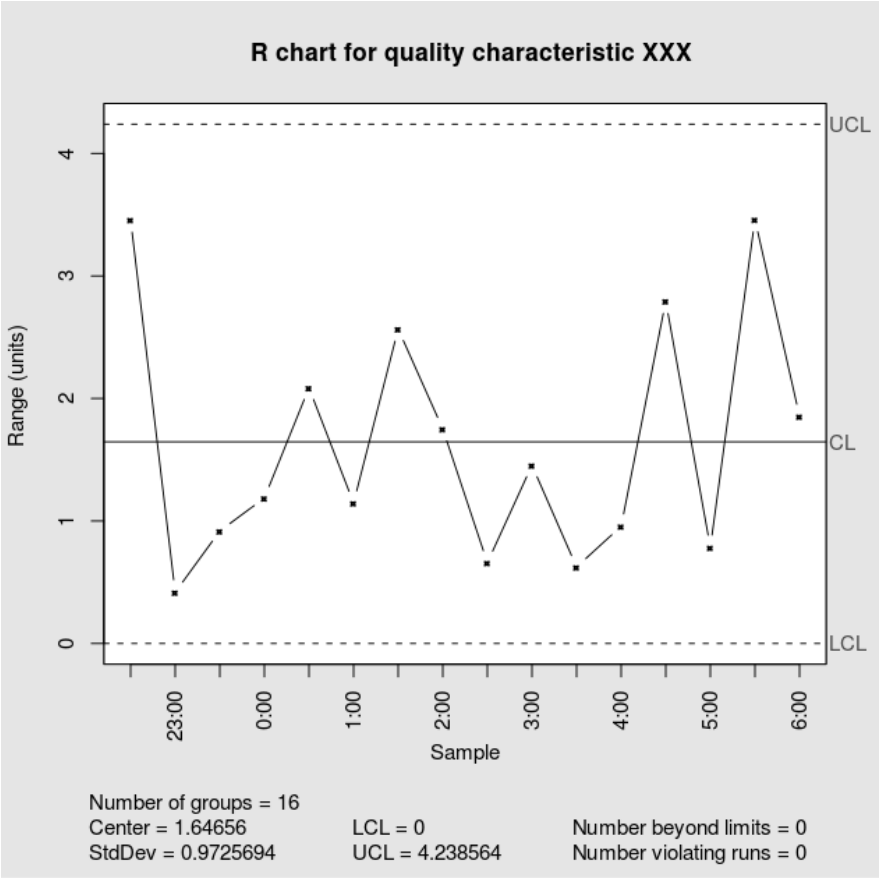

Figure 4: From Wikimedia

For a fixed constant $b > 0$, the CUSUM control chart raises an alarm at the first time i when $C_i \geq b$. The value of b is chosen to trade-off false detections and delays in detection.

4 References

- A. Di Bucchianico, Applied Statistics, Technische Universiteit Eindhoven, 2008.