The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 69-75

A Simple d_2 Factor (d_2^S) for Control Charts

Jea-Young Lee[1] **and Jae-Woo Lee**[2]

Abstract

A new statistic d_2^S is introduced for constructing control limits. It is easier and more convienient than d_2. We will show the characteristic of d_2^S and evaluate d_2^S through average run length(ARL).

1. Introduction

In statistical quality control, $\bar{x}$ and R control charts(Montgomery, 1996) are widely used for monitoring the process mean and variability. These charts are composed of Center Line that represents the average value of the quality characteristic corresponding to the in-control state, Upper Control Limit(UCL) and Lower Control Limit(LCL). For constructing control limits, we need an estimate of a standard deviation and may use the range of the samples. To do that, Tippett and Lond(1925) proposed d_2 statistic by the mean of relative range(W) and obtained d_2 values for various sample sizes. But, Tippett's d_2 is expressed as a very complex function and it has been calculated by the difficult procedures.

In Chapter 2, we will introduce Tippett's d_2 and a new d_2^S statistic. The comparison between d_2 and d_2^S based on average run length will be appeared in Chapter 3 and show conclusions in Chapter 4.

2. d_2^S factor

Control limits in the $\bar{x}$ and R control charts can be expressed as follows.

1) Associate Professor, Department of Staistics, Yeungnam University, Gyongsan, 712-749, Korea
2) Professional Servicer, SAS Software Korea Ltd., 45-21 Yoido-Dong, Youngdeunpo-Ku, Seoul, 150-010, Korea

$$\text{UCL} = \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$
$$\text{Center Line} = \mu$$
$$\text{UCL} = \mu - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Usually, the grand average, $\bar{\bar{x}}$, is widely used as an estimate of μ and 3 takes place of $Z_{\alpha/2}$. $\hat{\sigma}$ is used with an estimate of σ and it can be expressed by

$$\hat{\sigma} = \frac{\overline{R}}{d_2}$$

where $\overline{R} = \sum_{j=1}^{m} \frac{R_j}{m}$, $R_j = X^j_{\max} - X^j_{\min}$, j : number of subgroups .

The d_2 is defined as expected value of sample relative range of variables that are normally distributed and have the same mean and a standard deviation. The random variable $W = \frac{R}{\sigma}$ is known as relative range and the parameters of the distribution of W are a function of the sample size n. R is the range of the difference between the largest and smallest observations.

The d_2(Tippett and Lond, 1925) is expressed as follows.

$$d_2 = E(W) = E(\frac{R}{\sigma})$$
$$= \int_{-\infty}^{\infty} [1 - (1 - \Phi(x))^n - (\Phi(x))^n] dx,$$

where $\Phi(.)$ is the CDF of standard normal distribution.

If we use $\bar{\bar{x}}$ as an estimator of μ and $\frac{\overline{R}}{d_2}$ as an estimator of σ, then the control limits of the $\bar{x}$ chart are

$$\text{UCL} = \bar{\bar{x}} + \frac{3}{d_2\sqrt{n}} \overline{R} ,$$
$$\text{Center Line} = \bar{\bar{x}} ,$$
$$\text{UCL} = \bar{\bar{x}} - \frac{3}{d_2\sqrt{n}} \overline{R} .$$

Also, the control limits of R chart is as follows.

$$\text{UCL} = \overline{R} + 3d_3 \frac{\overline{R}}{d_2} ,$$
$$\text{Center Line} = \overline{R},$$
$$\text{UCL} = \overline{R} - 3d_3 \frac{\overline{R}}{d_2} ,$$

where d_3 is the standard deviation of W.

By the way, the d_2 of Tippett and Lond (1925) is derived by very complex formular like as above and we may can't obtain d_2 value for n>25 in various statistical packages SAS, SPSS etc. To get rid of these kinds of difficulties, we derived very simple statistics, called d_2^S.

The d_2^S factor is derived as follows:

Let Φ be the cumulative distribution function(CDF) of random variables $X_1, X_2, \cdots, X_n$, $X_{1:n}, \ldots, X_{n:n}$ be the order statistics of $X_1, X_2, \cdots, X_n$ and $\Phi^{-1}(x)$ be the inverse function of cumulative distribution function Φ.

Let $U_i = \Phi(X_i)$ and $U_{1:n}, \ldots, U_{n:n}$ be the order statistics of $U_1, \ldots, U_n$, then $U_{i:n} = \Phi(X_{i:n})$, and we know already $U_{(i:n)}$ is a BETA(i, n-i+1) distribution. Therefore, we have a probability density function (pdf), $b_i(u_{(i:n)})$, of $U_{(i:n)}$ as

$$b_i(u_{i:n}) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} (u_{i:n})^{i-1}(1-u_{i:n})^{n-i}$$

We can then calculate the expected value of $x_{(i:n)}$, $E[x_{(i:n)}]$, by using transformation technique, from the pdf $h_i(x_{(i:n)}) = b_i(\Phi(x_{(i:n)})) \cdot |\Phi'(x_{(i:n)})|$ of $x_{(i)}$, that is,

$$X_{i:n} = \Phi^{-1}(U_{i:n})$$

$$h_i(x_{i:n}) = b_i[\Phi(x_{i:n})] \cdot |\Phi'(x_{i:n})|$$

Using previous equations, we obtain the following a proposition.

Proposition 2.1 (Lee and Rhee, 1997)

Assume Φ is CDF of random variables $X_1, X_2, \cdots, X_n$ and $X_{1:n}, \ldots, X_{n:n}$ are the order statistics of $X_1, X_2, \cdots, X_n$, $\Phi^{-1}(x)$ is the inverse function of CDF. And let $b_i(u_{i:n})$ be pdf of $U_{i:n}$, $h_i(x_{i:n})$ be pdf of $x_{i:n}$. Then $E[X_{i:n}] \cong \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right)$, c∈[0, 1).

proof)

$$\begin{aligned} E[X_{i:n}] &= \int_{-\infty}^{\infty} x_{i:n}\ h_i(x_{i:n})\ dx_{i:n} \\ &= \int_{-\infty}^{\infty} x_{i:n}\ b_i[\Phi(x_{i:n})] \cdot |\Phi'(x_{i:n})|\ dx_{i:n} \\ &= \int_0^1 \Phi^{-1}(u_{i:n})\ b_i(u_{i:n})\ du_{i:n} \\ &\cong \int_0^1 \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right) b_i(u_{i:n})\ du_{i:n} \\ &= \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right) \int_0^1 b_i(u_{i:n})\ du_{i:n} \\ &= \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right), \end{aligned}$$

where $c\in[0,\ 1)$. Therefore, we have $E(X_{i:n})\cong\Phi^{-1}\left(\frac{n-c}{n-2c+1}\right)$.

Proposition 2.2

$$E(W)\cong 2\times\Phi^{-1}\left(\frac{n-c}{n-2c+1}\right), \qquad c\in[0,1).$$

proof) The relative range(W) can be expressed as follows.

$$W=\frac{R}{\sigma}=\frac{X_{n:n}-X_{1:n}}{\sigma},$$

where n is a sample size of subgroup. The expected value of W is

$$\begin{aligned}E(W) &= E\left(\frac{X_{n:n}-X_{1:n}}{\sigma}\right)\\ &= E\left(\frac{X_{n:n}-\mu-X_{1:n}+\mu}{\sigma}\right)\\ &= E\left(\frac{X_{n:n}-\mu}{\sigma}\right)-E\left(\frac{X_{1:n}-\mu}{\sigma}\right)\\ &= E\left(X'_{n:n}\right)-E\left(X'_{1:n}\right)\\ &\cong \Phi^{-1}\left(\frac{n-c}{n-2c+1}\right)-\Phi^{-1}\left(\frac{1-c}{n-2c+1}\right)\\ &= 2\times\Phi^{-1}\left(\frac{n-c}{n-2c+1}\right),\end{aligned}$$

where $c\in[0,\ 1)$.

By the above propositions, we can define d_2^S as the expected value of relative range.

$$d_2^S\equiv 2\times\Phi^{-1}\left(\frac{n-c}{n-2c+1}\right).$$

Based on c=1/3, 3/8, 1/2 values, d_2^S values have been obtained for n=1,2, ...,25 and n>25 (Table 1) and we may hard to find the most appropriate d_2^S value based on c values with d_2 of Tippett and Lond(1925). Futhermore, what we want thing to evaluate the decisions between d_2 and d_2^S regarding sample size and sampling frequency is through the average run length (ARL) of the control chart. We will do the comparison between d_2 and d_2^S based on ARL.

3. Comparison between d_2 and d_2^S by ARL

The ARL equation is well known as evaluating the performance of the control limits.

Essentially, the ARL is the average number of points that must be plotted until a point indicates an out-of-control condition. For any Shewhart control chart(Montgomery, 1996), the ARL can be calculated easily from

$$ARL = \frac{1}{p} ,$$

where p= P [any point exceeds the control limits].

Table 1. Values of d_2 and d_2^S for various sample sizes

n	d_2	d_2^S		
	Tippett	c=1/3	c=3/8	c=1/2
2	1.12838	1.13190	1.17891	1.34898
3	2.05875	1.68324	1.73885	1.93484
4	2.32593	2.04015	2.09826	2.30070
5	2.53441	2.30070	2.35952	2.56310
6	2.70436	2.50424	2.56310	2.76599
7	3.33598	2.67036	2.72898	2.93047
8	2.84720	2.81014	2.86840	3.06824
9	2.97003	2.93047	2.98831	3.18644
10	3.07751	3.03586	3.09327	3.28971
11	3.17287	3.12945	3.18644	3.38124
12	3.25846	3.21351	3.27008	3.46333
13	3.33598	3.28971	3.34587	3.53765
14	3.40676	3.35932	3.41511	3.60549
15	3.47183	3.42335	3.47877	3.66783
16	3.53198	3.48258	3.53765	3.72546
17	3.58788	3.53765	3.59239	3.77902
18	3.64006	3.58908	3.64350	3.82901
19	3.68896	3.63729	3.69141	3.87586
20	3.73495	3.68265	3.73648	3.91993
21	3.77834	3.72546	3.77902	3.96150
22	3.81938	3.76598	3.81928	4.00085
23	3.85832	3.80443	3.85748	4.03817
24	3.89535	3.84100	3.89381	4.07367
25	3.93063	3.87586	3.92843	4.10750
26	.	3.90916	3.96150	4.13980
27	.	3.94101	3.99315	4.17071
28	.	3.97154	4.02347	4.20033
29	.	4.00085	4.05258	4.22876
30	.	4.02902	4.08056	4.25609
35	.	4.15542	4.20611	4.37870
40	.	4.26276	4.31271	4.48281
45	.	4.35585	4.40516	4.57610
50	.	4.43790	4.48666	4.65270

To illustrate, for the $\bar{x}$ chart with 3-sigma limits, p=0.0027 is the probability that a single point falls outside the limits when the process is in control. Therefore, the average run length of the $\bar{x}$ chart when the process is in control (called ARL_0) is $ARL_0 = \frac{1}{p} = \frac{1}{0.0027} \simeq 370$. That is, even if the process remains in control, an out-of-control signal will be generated every 370 samples, on the average.

In here, we want to compare d_2 and d_2^s in terms of ARL by using Monte Carlo simulations of 2,000,000 times based on each sample size. We first generate 2,000,000 random samples for each sample size from normal distribution and then we can obtain 2,000,000 means of subgroups from generated samples. After calculating 2,000,000 sample means for each subgroup, $\bar{x}$ control limits are calculated based on c=1/3, 3/8, 1/2 and Tippett values. Values of c(1/3, 3/8, 1/2) are well known factor that is used for normality test. If each mean of subgroups falls outside of these control limits, then it is considered out-of-control. Finally, we calculate average run length(ARL).

Table 2. ARL comparison when the process is in-control

n	ARL			
	Tippett	c=1/3	c=3/8	c=0.5
2	377.358	350.939	239.406	82.163
3	393.701	387.447	286.205	115.393
4	361.664	397.693	307.267	137.108
5	353.774	415.800	323.520	156.678
6	379.747	425.260	336.304	167.673
10	375.049	426.712	357.782	202.593
15	377.858	434.028	369.413	222.891
20	376.719	434.972	375.164	237.727
25	368.528	423.998	370.576	242.043
30	·	432.994	379.723	254.550
35	·	415.714	371.885	257.169
40	·	425.532	379.651	261.814
45	·	419.551	375.164	261.917
50	·	427.533	382.117	266.809

From the Table 2, we find that d_2^s's ARL values are varied based on c values 1/3, 3/8 and 1/2. For n>15 and c=3/8, the ARL has very stablized value with 370s. Of course, Tippett's that is also very consistant. But, Tippett's values for n>25 are hard to get general statistical software like as SAS, SPSS etc. Therefore, we have the following conclusions for d_2^s statistic with c=3/8.

4. Conclusions

The Tippett's d_2 value is obtained using quadratures and filled in by interpolation, using first Lagrangian formulae, and a difference formula. Therefore, these procedures are very complex and need deep calculation. In fact, in most statistical computer package like as SAS, SPSS and BMDP, we can not obtain d_2 value for n>25. On the other hand, a new statistic d_2^S, provided the expected values of sample range, is very easier and simpler than Tippett's d_2. Futhermore, for n>10 the ARL of d_2^S(with c=3/8) is very stable and similar to that of d_2, and for n>25, we can use the simple d_2^S(c=3/8) statistic for the Shewhart control limits.

5. References

[1] Lee, J.-Y. and Rhee, S.-W. (1997). 특정분포에 따른 확률 Plot들의 정규성과 Bimodality비교, *한국통계학회논문집*, 제 4권 1호, 243-254.

[2] Montgomery, D. C. (1996). *Introduction to statistical quality control*, 3rd ed., John Wiley, New York.

[3] Tippett, L. H. C. and Lond, B. SC. (1925) On the extreme individuals and the range of samples taken from a normal population, *Biometrika*, VOL. 17, 364-387.