ROBUST VARIABLES CONTROL CHARTS BASED ON SAMPLE MEANS

by

John E. Walsh and Grace J. Kelleher

Technical Report No. 33
Department of Statistics THEMIS Contract

May 12, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

ROBUST VARIABLES CONTROL CHARTS BASED ON SAMPLE MEANS

John E. Walsh* and Grace J. Kelleher
Southern Methodist University, Dallas, Texas, U.S.A.
Institute for Defense Analysis, Arlington, Virginia, U.S.A.

ABSTRACT

Consider variables control charts based on sample means (population mean μ and standard deviation σ are known). An assumption that the means of the samples (of size n) are normally distributed is commonly made, on the basis of the Central Limit Theorem. This assumption is often questionable for small n. Fortunately, for the continuous case, the form $\mu \pm k\sigma/\sqrt{n}$ used for limits of two-sided charts ordinarily assures that the in-control probability of a mean falling outside the limits equals the normality value if terms of order 1/n are neglected. This helps explain the successful use of such charts. Also, four types of charts are developed that have accurately determined out-of-control probabilities (the normality values) for the case of in-control if terms of order r^2/n^2 are neglected, with r = 1, 2, 3. These charts have the limits $\mu \pm \sigma\sqrt{3r/n}$ and the advantage of not going far into the tail of the distribution for a sample mean. The first type is of the usual kind, with r = 1, and has an out-of-control probability of about .0833. The sample is taken as two equal-sized subsamples for the second type and r = 2. Both subsample means must fall outside the limits for out-of-control. The probability is about .0069. For the third type, the sample size is n/2, with r = 2, and two successive means must fall outside for out-of-control. The probability that two successive means are outside, and that this occurs first for a given sample (past the second), is about .0064. The sample is taken in three equal subsamples for the fourth type and r = 3. All three means fall outside for out-of-control. The probability is about .00068. Some operating characteristic values and efficiencies are computed for the cases of normality and r = 2, 3 with subsamples.

---

*Research partially supported by Mobil Research and Development Corp., ONR Contract N00014-68-A-0515, and NASA Grant NGR 44-007-028.

INTRODUCTION AND RESULTS

This paper is mainly concerned with the in-control probability properties of a class of variables control charts for continuous-data situations where the customary assumption of sample values from a normal population can be strongly violated. The control charts considered are for surveillance of the population mean. Plotted on a chart (successively) are statistics of independent random samples of size n. The centerline for a chart is μ and the control limits are of the form $\mu \pm k\sigma/\sqrt{n}$. Here, μ is the value "desired" for the population mean and σ is the value for the population standard deviation. The values of μ and σ are known (determined on the basis of past data, management decisions, and perhaps other considerations). The in-control mean of the population sampled is μ and the in-control standard deviation is σ.

Traditionally, a single statistic, its arithmetic mean, is used for each sample and out-of-control occurs if and only if a sample mean is outside the control limits. The value used for k is ordinarily large (say, 3 or 3.09) although much smaller values are sometimes appropriate (for example, see Duncan, 1965). Alternatively, the sample of n could be taken as r independent subsamples, of size n/r, and the mean of each subsample plotted on the chart (with the control limits adjusted accordingly). Out-of-control occurs for a sample if and only if all of the subsample means are outside the control limits. Similar to taking subsamples, the sample size could be reduced to n/r and r times as many samples taken. Then, out-of-control occurs if and only if r successive sample means fall outside the control limits. In this paper the cases of $r = 2, 3$ are considered for subsamples and the case of $r = 2$ for reduced sample size. The case of $r = 1$ is the traditional situation. The value of n/r is, of course, always an integer.

In every case considered, the mean of a sample (or subsample) is plotted on a chart whose control limits are symmetrical about the in-control value μ for the population mean. Direct use of the Edgeworth series expansion (Cramér, 1946) shows that, under fairly general conditions, the probability of a mean falling outside the control limits equals the normality value if terms of order r/n are neglected. This helps explain the successful use of charts of this type. However, even with the more general model, control limits that are as much as three standard deviations (of the observed mean) from the centerline can result in an out-of-control probability that relatively is much different from the normality value. For given n and r, this relative difference, due to inaccuracy of approximation in extreme tails of distributions, tends to decrease as the number of standard deviations used for the control limits decreases. It should nearly always be moderate when the number of standard deviations is less than two.

One way of simultaneously using less than two standard deviations and still having a small in-control probability for out-of-control is to utilize a case with $r > 1$. This results in basing a mean on fewer than n sample values. However, there would ordinarily seem to be an important gain in relative nearness to the normality value, even though the terms neglected are of order r/n rather than order $1/n$.

The material of (Walsh, 1956 and 1958) indicates that the use of control limits that are $\sqrt{3}$ standard deviations from the centerline would result in probabilities for out-of-control that are relatively near the normality values. This situation is examined by use of the Edgeworth series expansion and the normality values are found to be accurate if terms of order r^2/n^2 are neglected.

When $\sqrt{3}$ standard deviations are used for control limits, these limits are $\mu \pm \sigma\sqrt{3r/n}$. For r = 1, the out-of-control probability is about .0833 for the situation of in-control. The value .0833 is small enough for some cases and in other cases $\mu \pm \sigma\sqrt{3/n}$ could be used as warning limits (for example, see Duncan, 1965). For r = 2 and subsamples, the probability is about .0069. For r = 3 and subsamples, the probability is about .00068. Finally, for r = 2 and reduced samples, the probability is about .0064.

The question arises as to whether use of r = 2 or 3 results in substantially less favorable operating characteristics as compared to use of r = 1 and limits that result in the same out-of-control probability for the situation of in-control. An investigation of operating characteristic (OC) function values occurs for the situation of normality and subsamples with r = 2, 3. These OC values are compared with those for the corresponding cases of r = 1. The value of σ is fixed (at the in-control value) for these computations. It is found that the OC values for r = 2 are definitely inferior to those for r = 1, and that those for r = 3 are very much inferior to those for r = 1. Approximate efficiencies are 68 percent for r = 2 and 57 percent for r = 3. Thus a gain in robustness leads to a loss with respect to OC values. However, the loss may be much smaller when the population sampled is strongly nonnormal.

Finally, it should be noticed that the in-control "robustness" from use of $\sqrt{3}$ standard deviations arises from an approximately constant value for the sum of the probability of being above the upper control limit and the probability of being below the lower limit. These two probabilities can be greatly different. They are, of course, equal for the normality situation.

The next section contains the basis for the in-control results that have been stated. The final section contains the material on OC values and efficiencies.

BASIS FOR IN-CONTROL PROBABILITIES

Let $\bar{X}$ be the mean of a random sample of size n/r from a continuous population with mean μ, variance σ^2, and 2v-th moment about the mean μ_{2v}, (v = 2, 3,...). The probability density function of the continuous random variable $\bar{X}$ is required to exist and to be continuous. Let

$$Y = \sqrt{n}\, U(\bar{X} - \mu)/\sigma \sqrt{r},$$

where the random variable U is independent of $\bar{X}$, and takes the value 1 with probability 1/2 and the value -1 with probability 1/2. It is easily seen that the random variable Y has a probability density function f(y) that is continuous and symmetrical. The expected value of Y is zero and its variance is unity. The fourth moment about the mean is easily shown (for example, see Cramér, 1946) to be

$$E(Y^4) = 3 + r(\mu_4/\sigma^4 - 3)/n.$$

Thus, since all the off moments about the mean are zero, the sum of the first seven terms of the Edgeworth series (Cramér, 1946) for f(y) is

$$\varphi(y) + (r/24n)(\mu_4/\sigma^4 - 3)\varphi^{(4)}(y),$$

where

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad \varphi^{(4)}(y) = d^4\varphi(y)/dy^4.$$

Examination of $E(Y^{4+2v})$, $(v = 1, 2, \ldots)$, and substitution shows that the remaining terms of the Edgeworth series for f(y) are $O(r^2/n^2)$. Thus,

$$f(y) = \varphi(y) + (r/24n)(\mu_4/\sigma^4 - 3)\varphi^{(4)}(y) + O(r^2/n^2).$$

However, for any $K > 0$,

$$\begin{aligned} 1 - P(-K \le Y \le K) &\equiv 1 - P[-K \le \sqrt{n}(\overline{X} - \mu)/\sigma\sqrt{r} \le K] \\ &\equiv 1 - P[\mu - K\sigma\sqrt{r/n} \le \overline{X} \le \mu + K\sigma\sqrt{r/n}], \end{aligned}$$

where the last term is the probability of $\overline{X}$ falling outside the limits $\mu \pm K\sigma\sqrt{r/n}$. Thus, for control limits symmetrical about μ, the probability of $\overline{X}$ being outside the limits equals the normality value plus a term of order r/n.

When $K = \sqrt{3}$, evaluation of $1 - P(-\sqrt{3} \le Y \le \sqrt{3})$ shows that the coefficient of $(\mu_4/\sigma^4 - 3)$ becomes zero. Hence, the probability of $\overline{X}$ being outside $\mu \pm \sigma\sqrt{3r/n}$ limits equals the normality value plus terms of order r^2/n^2. The approximate in-control probabilities stated for use of control limits $\mu \pm \sigma\sqrt{3r/n}$ are based on the assumption that terms of order r^2/n^2 are unimportant.

Finally, consider approximate evaluation of the probability that the first out-of-control is observed at a specified one of the successive samples for the case of $r = 2$ and reduced sample size. Out-of-control cannot occur prior to the second sample. For the second sample it is approximately $(.0833)^2$, the approximate in-control probability that the mean of a sample of size n/2 falls outside the limits $\mu \pm \sigma\sqrt{6/n}$. For the third and further samples, the nearly equal values

$$(.0833)^2/(1 + .0833), \qquad (.0833)^2/[1 + .0833 - (.0833)^2]$$

approximately bound the probability for out-of-control (approximate bounds because .0833 is approximate).

OC VALUES AND EFFICIENCIES

OC values and approximate efficiencies are determined for $r = 2$, 3 and subsamples. The data are assumed to be a random sample from a normal distribution and computations are limited to the situation where σ remains fixed at its in-control value.

Let μ' be the true value of the population mean. The OC value is, of course, the probability that out-of-control does not occur as a function of μ' (n, μ, and σ have known values). Actually, the OC values are expressed as a function of

$$\psi = \sqrt{n}(\mu - \mu')/\sigma$$

which is more convenient.

The control limits are $\mu \pm \sigma\sqrt{6/n}$ for the case of $r = 2$ and subsample size n/2. Corresponding control limits for $r = 1$ and sample size n are $\mu \pm 2.70\sigma/\sqrt{n}$. The OC value for $r = 2$ can be expressed as

$$1 - [1 - P(-1.732 + .707\psi \le Z \le 1.732 + .707\psi)]^2,$$

where Z denotes a random variable whose distribution is normal with zero mean and unit standard deviation. The OC value for the corresponding case with r = 1 is

$$P(-2.70 + \psi \le Z \le 2.70 + \psi).$$

The values of these expressions are unchanged if ψ is replaced by $-\psi$, so that only nonnegative values of ψ need to be considered.

OC values are computed for ψ = 2.0, 3.0, 4.0, 4.5, 5.0, 5.5. The results are given in the following table.

OC Values for r = 2 and Corresponding r = 1 Cases

ψ =	2.0	3.0	4.0	4.5	5.0	5.5
r = 2 Case	.863	.576	.220	.144	.061	.032
r = 1 Case	.758	.382	.097	.036	.011	.003

Smaller values are, of course, preferable. To obtain an indication of the relative efficiency of the r = 2 case (compared to the corresponding r = 1 case) the value of n for the r = 2 case is increased by 48 percent. Then, the OC value is, with the original n (not increased) used in ψ,

$$1 - [1 - P(-1.732 + .86\psi \le Z \le 1.732 + .86\psi)]^2.$$

An OC value of .032 occurs for ψ = 4.5, an OC value of .010 for ψ = 5.0, and an OC value of .003 for ψ = 5.5. This suggests that the efficiency for the r = 2 case is about 68 percent.

Finally, consider the case of r = 3 and subsample size n/3, with control limits $\mu \pm 3\sigma/\sqrt{n}$. Corresponding control limits for r = 1 and sample size n are $\mu \pm 3.40\sigma/\sqrt{n}$. The OC value for r = 3 is

$$1 - [1 - P(-1.732 + .578\psi \le Z \le 1.732 + .578\psi)]^3.$$

The OC value for the corresponding case with $r = 1$ is

$$P(-3.40 + \psi \le Z \le 3.40 + \psi).$$

Again, a value is unchanged if ψ is replaced by $-\psi$.

OC values are computed for $\psi = 3.0, 4.0, 5.0, 6.0, 6.5$. The results are given in the following table

OC Values for r = 3 and Corresponding r = 1 Cases

$\psi =$	3.0	4.0	5.0	6.0	6.5
r = 3 Case	.875	.629	.324	.130	.066
r = 1 Case	.655	.274	.055	.005	.001

To obtain an idea of the relative efficiency of the $r = 3$ case, the value of n for the $r = 3$ case is increased by 76 percent. Then, with the original n in ψ, the OC value is

$$1 - [1 - P(-1.732 + .765\psi \le Z \le 1.732 + .765\psi)]^3.$$

An OC value of .054 occurs for $\psi = 5.0$. An OC value of .006 occurs for $\psi = 6.0$, and an OC value of .002 for $\psi = 6.5$. This suggests that the efficiency for the $r = 3$ case is about 57 percent.

REFERENCES

Cramér, Harald, Mathematical Methods of Statistics, Princeton University Press, 1946

Duncan, Acheson J., Quality Control and Industrial Statistics (3rd Ed.), R. D. Irwin, Inc., 1965

Walsh, John E., "Validity of approximate normality values for $\mu \pm k\sigma$ areas of practical type continuous population," Annals of the

Institute of Statistical Mathematics, Vol. 8 (1956), pp. 79-86

Walsh, John E., "Further consideration of normality values for $\mu \pm k\sigma$ areas of continuous populations," *Annals of the Institute of Statistical Mathematics*, Vol. 19 (1958), pp. 127-29