▨연구논문

일반화 기하분포를 이용한 ARL의 수정에 관한 연구*

문명상
연세대학교 원주캠퍼스 통계학과

A Study on the Alternative ARL Using Generalized Geometric Distribution

Myung-Sang Moon
Dept. of Statistics, Yonsei University, Wonju

Abstract

In Shewhart control chart, the average run length(ARL) is calculated using the mean of a conventional geometric distribution(CGD) assuming a sequence of identical and independent Bernoulli trials. In this, the success probability of CGD is the probability that any point exceeds the control limits. When the process is in-control state, there is no problem in the above assumption since the probability that any point exceeds the control limits does not change if the in-control state continues. However, if the out-of-control state begins and continues during the process, the probability of exceeding the control limits may take two forms. First, once the out-of-control state begins with exceeding probability p, it continues with the same exceeding probability p. Second, after the out-of-control state begins, the exceeding probabilities may vary according to some pattern. In the first case, ARL is the mean of CGD with success probability p as usual. But in the second case, the assumption of a sequence of identical and independent Bernoulli trials is invalid and we can not use the mean of CGD as ARL. This paper concentrate on that point. By adopting one generalized binomial distribution(GBD) model that allows correlated Bernoulli trials, generalized geometric distribution(GGD) is

* This Research was supported by the 1999 Yonsei Maeji Research Fund.

defined and its mean is derived to find an alternative ARL when the process is in out-of-control state and the exceeding probabilities take the second form mentioned in the above. Small-scale simulation is performed to show how an alternative ARL works.

1. Introduction

The ARL of the control chart is the average number of points that are to be plotted before a point indicates an out-of-control condition, and is a way used to evaluate the decisions regarding sample size and sampling frequency. For any Shewhart control chart, the ARL can be calculated easily as(Montgomery, 1996)

$$\text{ARL} = 1/p, \tag{1.1}$$

from the CGD where p is the probability that any point exceeds the control limit. The CGD assumes a sequence of identical and independent Bernoulli trials. However, if the above assumption is turned out to be an inappropriate one, then the ARL defined in (1.1) should be modified. The purpose of this paper is to derive an alternative ARL when the identical and independent Bernoulli trials assumption is not satisfied.

This paper consists of five sections. The GGD that works even when the identical and independent Bernoulli trials assumption is inappropriate is defined in section 2 using some previous results on GBD. In section 3, an alternative ARL is derived from GGD by finding its mean. Simulation results that compare the usual ARL and an alternative ARL when the identical and independent Bernoulli trials assumption is inappropriate are provided in section 4. Final section is devoted to some concluding remarks.

2. Generalized Geometric Distribution

Consider a sequence of identical and independent Bernoulli trials with probability of success p. If we denote by Y the number of trials required to obtain the first success, then it is well-known that the random variable Y follows CGD with probability mass function(p.m.f.)

$$g(y;p) = pq^{y-1}, \quad y = 1,2,3,\cdots ; \quad q = 1-p. \tag{2.1}$$

However a sequence of identical and independent Bernoulli trials assumption is not satisfied in many cases, and then the equation (2.1) is not an appropriate p.m.f. of Y. Examples are the attendence of a congressman at meetings, a team's probability of winning at successive games, and a probability of survival of a plant in a given area. Hence, a new generalization of the CGD that allows dependence between trials, nonconstant success probabilities from trial to trial, and which contains the CGD as a special case, is necessary. Many results on GBD allowing dependence between trials and nonconstant success probabilities are published already. See Altham(1978), Crowder(1985), Drezner & Farnum(1993), Kupper & Haseman(1978), Madsen(1993), Moore(1987), Ng(1989) and Paul(1985, 1987). Among various models allowing correlated Bernoulli trials, Drezner & Farnum's one which takes account of the previous number of trials is employed in this paper.

Let $P(x,n)$ denote the probability of x successes in n Bernoulli trials, and $S_n(F_n)$ the event of 'success(failure) on the n-th trial'. Let $P(S_n|x, n-1)$ denote the conditional probability of success on the n-th trial after x successes among the previous $n-1$ Bernoulli trials, and let's define $P(F_n|x, n-1)$ similarly. The model allowing correlated Bernoulli trials depends heavily on how we define $P(S_n|x, n-1)$ and $P(F_n|x, n-1)$. Drezner and Farnum defined them as follows so that they take account of the previous number of trials.

$$P(S_n|x, n-1) = (1-\theta)p + \theta\frac{x}{n-1},$$

$$P(F_n|x, n-1) = (1-\theta)(1-p) + \theta\left(1-\frac{x}{n-1}\right), \tag{2.2}$$

where p denotes the probability of success in the first trial, and θ defines the degree of dependence between Bernoulli trials. Note that $P(0,1) = 1-p$ and $P(1,1) = p$.

Assume a sequence of correlated Bernoulli trials satisfying (2.2) and let X be the random variable denoting the number of trials required to obtain the first success. Then, the p.m.f. of X can be written as follows:

$$f(x;\theta) = P(0,x-1)\,P(S_x|0,x-1), \quad x=1,2,3,\cdots, \tag{2.3}$$

or

$$f(x;\theta) = \begin{cases} p, & \text{if } x=1, \\ p(1-p)(1-\theta)\{(1-\theta)(1-p)+\theta\}^{x-2}, & \text{if } x=2,3,4,\cdots. \end{cases} \tag{2.4}$$

Define the distribution given in (2.3) and (2.4) as GGD. The resulting class of GGD includes the CGD(when $\theta=0$) as a special case. Although (2.4) is easier to handle, (2.3) will be adopted in this paper since it is expected that we can extend the present GGD result to that of generalized negative binomial distribution (GNBD) using (2.3)-type p.m.f. Furthermore, the recursive formula for the k-th moment of GGD can be obtained more easily if we use (2.3). The following theorem shows that $f(x;\theta)$ defined in (2.3) satisfies one condition of p.m.f.

Theorem 1. With $f(x;\theta)$ given in (2.3), it follows that $\sum_{x=1}^{\infty} f(x;\theta) = 1$.

Proof. Let $A = \sum_{x=1}^{\infty} f(x;\theta) = \sum_{x=1}^{\infty} P(0,x-1)\,P(S_x|0,x-1)$. Then,

$$\begin{aligned} A &= p + \sum_{x=2}^{\infty} P(0,x-1)\,P(S_x|0,x-1) \\ &= p + \sum_{y=1}^{\infty} P(0,y)\,P(S_{y+1}|0,y), \qquad \text{where } x-1=y, \\ &= p + \{(1-\theta)(1-p)+\theta\} \sum_{y=1}^{\infty} P(0,y-1)\,P(S_{y+1}|0,y) - p^2\theta(1-\theta) \\ &= p + \{(1-\theta)(1-p)+\theta\}\{A+(1-\theta)p-p\} - p^2\theta(1-\theta). \end{aligned}$$

Rearranging terms on A yields

$$p(1-\theta)A = p - (p\theta)\{(1-\theta)(1-p)+\theta\} - p^2\theta(1-\theta),$$

and we have $A=1$. □

The other condition of p.m.f. is $0 \le f(x;\theta) \le 1$ for all x, and it is related to the range of θ in GGD.

Theorem 2. For (2.3) to be a proper GGD p.m.f., θ should satisfy the following:

$$1-1/p \le \theta \le 1.$$

Proof. By Theorem 1, it is sufficient to ensure that $f(x;\theta) \ge 0$. It requires that $P(0, x-1) \ge 0$ and $P(S_x | 0, x-1) \ge 0$. We get $\theta \ge 1-1/p$ from the first one since $P(F_{x-1} | 0, x-2)$ is included in it. $\theta \le 1$ is easily derived from the second one. Hence the result follows. □

From Theorem 2, it is seen that negative θ is allowed.

3. An Alternative ARL

In this section, the recursive formula for the k-th moment of GGD is derived and based on that result, an alternative ARL is suggested.

Theorem 3. Let the random variable X follow GGD with initial success probability p and correlation related parameter θ. Then, the recursive formula for the k-th moment of X is given as follows:

$$E(\mathrm{X}^k) = \frac{\{(1-\theta)(1-p)+\theta\}}{p(1-\theta)} \sum_{i=0}^{k-1} \binom{k}{i} E(\mathrm{X}^i) - \left(\frac{2^k\theta - 1}{1-\theta} \right), \qquad k = 1, 2, 3, \cdots.$$

Proof. By definition,

$$\begin{aligned} E(\mathrm{X}^k) &= \sum_{x=1}^{\infty} x^k P(0, x-1)\, P(S_x | 0, x-1) \\ &= \sum_{y=0}^{\infty} (y+1)^k P(0, y)\, P(S_{y+1} | 0, y), \quad \text{where } x-1 = y, \\ &= \sum_{y=0}^{\infty} \left\{ \sum_{i=0}^{k} \binom{k}{i} y^i \right\} P(0, y)\, P(S_{y+1} | 0, y) \end{aligned}$$

$$= \sum_{i=0}^{k} \binom{k}{i} \left\{ \sum_{y=2}^{\infty} y^i P(0,y) P(S_{y+1} | 0,y) + p(1-p)(1-\theta) \right\} + p$$

$$= \sum_{i=0}^{k} \binom{k}{i} [\{(1-\theta)(1-p)+\theta\} \cdot$$

$$\sum_{y=2}^{\infty} y^i P(0,y-1) P(S_{y+1} | 0,y) + p(1-p)(1-\theta)] + p$$

$$= \{(1-\theta)(1-p)+\theta\} \sum_{i=0}^{k} \binom{k}{i} E(\mathrm{X}^i) - 2^k p\theta + p.$$

The result follows by rearranging the terms on $E(\mathrm{X}^k)$. □

Result 1. Let X be the GGD random variable defined in Theorem 3. The expectation and variance of X is obtained as follows using Theorem 3.

$$E(\mathrm{X}) = \frac{1-p\theta}{p(1-\theta)}, \quad Var(\mathrm{X}) = \frac{(1-p)(1+p\theta)}{p^2(1-\theta)^2}.$$ □

Final theorem providing an alternative ARL when the process is in out-of-control is given in Theorem 4 using (2.2), (2.3) and Result 1.

Theorem 4. In Shewhart control chart, when the process is in-control with probability of exceeding control limits p_0, ARL(in-control)= $1/p_0$. When the out-of-control state begins with exceeding probability p and if it varies according to the pattern given in (2.2), then ARL(out-of-control) = $\frac{(1-p\theta)}{p(1-\theta)}$.

Proof. When the process is in-control state with probability of exceeding control limits p_0, the CGD is the distribution of run length since identical and independent Bernoulli trials assumption is satisfied. If the out-of-control state begins with probability of exceeding control limits p, and if it varies according to (2.2), then the distribution of run length is GGD defined in (2.3) and the result follows from Result 1. □

4. Simulation Results

When the process is in out-of-control state and probabilities of exceeding control limits vary according to some pattern, ARL should not be defined as the mean of CGD as usual. Particularly, if the probabilities of exceeding control limits vary according to (2.2), then an alternative ARL presented in the previous section should be used. An example of a process with varying probabilities of exceeding control limits when it is in out-of-control state(that is, a process satisfying the correlated Bernoulli trials assumption), is given below.

Example: In the manufacture of automotive engine piston rings, a critical quality characteristic is the inside diameter of the ring. Suppose that the process is in-control state if mean inside diameter($= \mu$) is 75mm. After maintaining that mean value for some period, suppose that a problem outbreaks to a machine related to the production of piston ring, and the out-of-control state begins at that moment with $\mu = 75.5$mm. If this mean value(=75.5) is maintained afterwards, then the usual identical and independent Bernoulli trials assumption is appropriate in calculating ARL(out-of-control) and CGD should be used in finding it. However, once a problem outbreaks to the machine, it would be more reasonable to assume that the condition of the machine getting worse. So, the value of μ may vary according to some fashion(for instance, $75.5 \rightarrow 75.6 \rightarrow 75.8 \rightarrow \cdots$, or according to (2.2)). In this case, the usual identical and independent Bernoulli trials assumption is not satisfied in finding ARL. Instead, the correlated Bernoulli trials assumption is to be employed and ARL should be obtained from GGD.

In this section, small-scale simulation results are provided to compare the usual ARL and an alternative ARL. The following values of parameters are used in simulation.

$$p = 0.001,\ 0.005,\ 0.010,\ 0.020,\ 0.050,\ 0.100,\ \text{and}$$
$$\theta = 0.010,\ 0.050,\ 0.100,\ 0.150,\ 0.200,\ 0.300.$$

For each combination of p and θ, 100 GGD random samples are generated as follows:

Step 1. Generate Bernoulli random sample with initial success probability p using IMSL subroutine RNBIN.

Step 2. If the first sample is turned out to be 'success', stop and set X(=GGD random variable)=1.

Step 3. If not, generate the second one with success probability $P(S_2|0,1)$ using IMSL subroutine RNBIN again. If it is 'success', then stop and set X=2, otherwise generate the third one with success probability $P(S_3|0,2)$.

Step 4. Repeat Step 3-like procedure for the k-th($k=4,5,\cdots$) sample until the first 'success' is obtained. For each one the success probability is $P(S_k|0,k-1)$. If the first 'success' is obtained in the k-th sample, stop and set X= k.

After 100 GGD random samples are obtained according to the above steps, usual sample mean of them is calculated. The above simulation procedure is replicated 500 times.

Three ARL values are provided in <Table 1>. ARL(simu.) is the mean of 500 simulated ARL sample means, ARL(GGD) is obtained from Theorem 4, and ARL(CGD) is the reciprocal of initial exceeding probability when the out-of-control state begins. The estimate of θ, $\hat{\theta}$, is also included in the table. It is a method of moment estimator of θ and its formula is,

$$\hat{\theta} = \frac{p\overline{X}-1}{p(\overline{X}-1)}.$$

From the simulation results given in <Table 1>, the following conclusions can be made.

(1) ARL(simu.) and ARL(GGD) take simular values regardless of p and θ, as can be expected.

(2) Three ARL values included in the table are similar for all p when θ is relatively small(say, less than 0.010).

(3) For moderate and large θ, ARL(CGD) values are quite different from two other ARL's. However, for a given θ, the ratio ARL(simu.)/ARL(CGD) ($\approx$ ARL(GGD)/ARL(CGD)) takes similar vales regardless of p. Hence, whether to use ARL(GGD) or ARL(CGD) depends heavily on θ, but not on p.

< Table 1 > ARL values for various combinations of p and θ

θ \ p		0.001	0.005	0.010	0.020	0.050	0.100
0.010	ARL(simu.)	1,005.813	200.427	100.712	50.430	20.243	10.074
	ARL(GGD)	1,010.090	202.010	101.000	50.495	20.192	10.091
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	-3.825E-3	-7.327E-3	-2.494E-3	-1.001E-3	1.996E-3	-3.500E-3
0.050	ARL(simu.)	1,055.696	210.121	105.375	52.463	21.096	10.476
	ARL(GGD)	1,052.579	210.474	105.211	52.579	21.000	10.474
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	4.293E-2	3.872E-2	4.180E-2	3.738E-2	4.576E-2	4.012E-2
0.100	ARL(simu.)	1,106.132	222.416	110.905	55.683	22.039	10.988
	ARL(GGD)	1,111.000	222.111	111.000	55.444	22.111	11.000
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	8.664E-2	9.283E-2	8.953E-2	9.446E-2	8.737E-2	8.917E-2
0.150	ARL(simu.)	1,176.455	236.149	117.010	58.501	23.366	11.530
	ARL(GGD)	1,176.294	235.118	117.471	58.647	23.353	11.588
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	1.421E-1	1.457E-1	1.381E-1	1.383E-1	1.417E-1	1.364E-1
0.200	ARL(simu.)	1,250.028	248.957	124.071	61.937	24.680	12.279
	ARL(GGD)	1,249.750	249.750	124.750	62.250	24.750	12.250
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	1.920E-1	1.896E-1	1.874E-1	1.885E-1	1.881E-1	1.928E-1
0.300	ARL(simu.)	1,435.420	285.020	141.827	70.632	28.127	13.849
	ARL(GGD)	1,428.143	285.286	142.429	71.000	28.143	13.857
	ARL(CGD)	1,000	200	100	50	20	10
	$\hat{\theta}$	2.966E-1	2.921E-1	2.902E-1	2.887E-1	2.926E-1	2.917E-1

5. Concluding Remarks

The GGD model that allows correlated Bernoulli trials is defined in this paper. This model can be used in finding ARL in Shewhart control chart when the process is in out-of-control state since the probabilities of exceeding control limits

may vary according to some fashion. The recursive formula for the k-th moment of GGD is derived, and its mean is introduced as an alternative ARL when the process is in out-of-control state. Simulation results show that an alternative ARL is more effective than usual ARL when θ is relatively large. However, it turns out that the initial exceeding probability p does not play an important role in comparing them.

References

[1] Altham, P. M. E.(1978), "Two generalizations of the binomial distribution," *Applied Statistics,* Vol. 27, pp. 162-167.

[2] Crowder, M.(1985), "Gaussian estimation for correlated binomial data," *Journal of the Royal Statistical Society, Series B,* Vol. 47, pp. 229-237.

[3] Drezner, Z. and Farnum, N,(1993). "A generalized binomial distribution," *Communications in Statistics-Theory and Methods,* Vol. 22, pp. 3051-3063.

[4] IMSL User's Manual(1989), IMSL Inc., Houston, TX.

[5] Kupper, L. L. and Haseman, J. K.(1978), "The use of a correlated binomial model for the analysis of certain toxicological experiments," *Biometrics,* Vol. 34, pp. 69-76.

[6] Madsen, R.(1993), "Generalized binomial distributions," *Communications in Statistics-Theory and Methods,* Vol. 22, pp. 3065-3086.

[7] Montgomery, D. C.(1996), Introduction to Statistical Quality Control, 3rd Edition, John Wiley & Sons, New York.

[8] Moore, D. F.(1987), "Modeling extraneous variance," *Applied Statistics,* Vol. 36, pp. 8-14.

[9] Ng, T. H.(1989), "A new class of modified binomial distributions with applications to certain toxicological experiments," *Communications in Statistics-Theory and Methods,* Vol. 18, pp. 3477-3492.

[10] Paul, S. R.(1985), "A three-parameter generalization of the binomial distribution," *Communications in Statistics-Theory and Methods,* Vol. 14, pp. 1497-1506.

[11] Paul, S. R.(1987), "On the beta-correlated binomial distribution-A three parameter generalization of the binomial distribution," *Communications in Statistics-Theory and Methods,* Vol. 16, pp. 1473-1478.