Robust attributes control charts (`racc`) in the `rQCC` package

Chanseok Park* and Min Wang†

December 2022

Abstract

The g control charts based on the geometric distribution are widely used in many engineering applications to monitor the number of conforming cases between the two consecutive appearances of nonconformities. However, conventional g charts are based on the maximum likelihood and minimum variance unbiased estimators which are very sensitive to outliers. Thus, they could result in severe bias for obtaining the control limits of the charts. In this note, we provide a brief summary of robust g control charts and a description of how they are constructed using the `racc` function in the R package `rQCC`.

In addition, we also provide

1 Geometric distribution and its parameter estimation

Denote Y_i $(i = 1, 2, \ldots, n)$ to be the number of normal cases (or failures) before observing the first adverse case (or success) in a series of independent Bernoulli trials where its success probability is given by p. Considering the location parameter a, the probability mass function (pmf) of the geometric distribution is given by

$$f(y) = P(Y_i = y) = p(1-p)^{y-a} \tag{1}$$

and its corresponding cumulative distribution function (cdf) is

$$F(y) = P(Y_i \leq y) = 1 - (1-p)^{y+1-a}, \tag{2}$$

where $y = a, a+1, \ldots$. In general, the location shift a is the known minimum possible number of events (usually $a = 0, 1$). Then the mean and variance of Y_i are given by

$$\mu = E(Y_i) = \frac{1-p}{p} + a \quad \text{and} \quad \sigma^2 = \mathrm{Var}(Y_i) = \frac{1-p}{p^2},$$

*Applied Statistics Laboratory, Department of Industrial Engineering, Pusan National University, Busan 46241, Korea. His work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C1091319).

†Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX 78249, USA.

respectively.

In many practical applications, the process parameter p is unknown and needs to be estimated. By using the maximum likelihood (ML) method, we have

$$\hat{p}_{\mathrm{ML}} = \frac{1}{\bar{Y} - a + 1}, \tag{3}$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. Note that the Method-of-Moments (MM) estimator yields the same estimator for p as the ML estimator. One can also use the minimum variance unbiased (MVU) estimator proposed by [1], which is given by

$$\hat{p}_{\mathrm{B}} = \frac{1 - 1/n}{\bar{Y} - a + 1}. \tag{4}$$

However, it is shown to be biased. For more details, see [2]. The correct MVU estimator is given by

$$\hat{p}_{\mathrm{MVU}} = \frac{1 - 1/n}{\bar{Y} - a + 1 - 1/n}. \tag{5}$$

For more details on the *conventional* g control charts based on the above estimators, one can refer to the vignette below.

```
> vignette("acc", package="rQCC")
```

Here we introduce two robust estimators for p developed by [3], which are based on the memoryless property of the geometric distribution and truncation of an empirical distribution.

First, we provide a robust estimator based on the memoryless property. It is immediate from the memoryless property that we have

$$P(X > s + t) = P(X > s) \cdot P(X > t). \tag{6}$$

It should be noted that the above equation works only when X has the pmf of the form $f(x) = p(1-p)^{x-1}$. Care should be taken to use this formula for the geometric random variable Y with location shift a, whose pmf is given by $f(y) = p(1-p)^{y-a}$. Note that $X = Y - a + 1$ is the geometric random variable with the pmf $f(x) = p(1-p)^{x-1}$ regardless of the value of a. Thus, by substituting $X = Y - a + 1$ into (6), we obtain

$$P(Y > s + t + a - 1) = P(Y > s + a - 1) \cdot P(Y > t + a - 1), \tag{7}$$

which works with any location shift a. Simplifying the arguments on the right-hand side of (7) with $s \leftarrow s + a - 1$ and $t \leftarrow t + a - 1$, we have

$$P(Y > s + t - a + 1) = P(Y > s) \cdot P(Y > t). \tag{8}$$

Rewriting (8) using (2), we have

$$1 - F(s + t - a + 1) = \big\{1 - F(s)\big\} \cdot \big\{1 - F(t)\big\},$$

which results in

$$\frac{F(s+t-a+1)-F(t)}{F(s)} = 1 - F(t).$$

Using $F(t) = 1-(1-t)^{t+1-a}$, we have

$$\frac{\hat{F}(s+t-a+1)-\hat{F}(t)}{\hat{F}(s)} = (1-\hat{p})^{t+1-a}. \tag{9}$$

Here $\hat{F}$ is an estimator of F given by

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Y_i \leq t),$$

where $\mathbb{I}(\cdot)$ is an indicator function. Solving (9) for $\hat{p}$, we have

$$\hat{p}_{\text{cdf}} = 1 - \left[\frac{\hat{F}(s+t-a+1)-\hat{F}(t)}{\hat{F}(s)}\right]^{1/(t+1-a)}. \tag{10}$$

The estimator in (10) uses the empirical cdf, which could discard large outliers by selecting appropriate values of t and s. We recommend the choices of t and s such that $\hat{F}(s+t-a+1)$ and $\hat{F}(t)$ approximately cover γ and $\gamma/2$ fractions of the data, respectively. Then we have $t = [q_{\gamma/2}]$ and $s = [q_{\gamma} - q_{\gamma/2} + a - 1]$.

Next, another method for estimating p is based on the truncated geometric distribution with the pmf given by

$$f(y) = \frac{p(1-p)^{y-a}}{1-(1-p)^{d-a+1}},$$

where $y = a, a+1, \ldots, d$. The ML estimator of this truncated distribution is not in a closed-form expression, but it is unique under a certain condition. For more details, see [4]. The closed-form MM estimator, which is quite comparable to the ML estimator, is provided in [5], but it works only for the case of location shift $a = 1$. We can modify this MM estimator so that it works with any location shift and it is given by

$$\hat{p}_{\text{t}} = \frac{(a+d)-2\bar{Y}}{(\bar{Y}-a+1)(d-\bar{Y})-S^2}, \tag{11}$$

where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$. Note that the value of the above MM estimator can be smaller than zero or larger than one. For the case that $\hat{p}_{\text{t}} \geq 1$, we set up $\hat{p}_{\text{t}} = 1$ which implies that Y degenerates at $Y = a$. For the case that $\hat{p}_{\text{t}} = 0$, Y degenerates at $Y = \infty$. Note that Y should always be between a and d with the truncation at d. Thus, we should avoid degenerating at $Y = \infty$ with $\hat{p}_{\text{t}} = 0$. This degeneration occurs if the value of d is too small. Thus, by increasing the value of d we can avoid this case. The condition $(a+d)-2\bar{Y} > 0$ guarantees the existence of the MM estimator over the open interval $(0,1)$. For more details, see [4]. Thus, with $d > 2\bar{Y} - a$, we can avoid degenerating at $Y = \infty$. The minimum positive integer value of d satisfying $d > 2\bar{Y} - a$ is given by $d^* = \lfloor 2\bar{Y} - a \rfloor + 1$, where $\lfloor \cdot \rfloor$ is a floor function.

2 Construction of the robust g control charts

The g chart (about the total number of events) with the sample size n_k has the following control limits

$$\begin{aligned} \mathrm{UCL}(p) &= n_k\left(\frac{1-p}{p}+a\right)+g\sqrt{\frac{n_k(1-p)}{p^2}}, \\ \mathrm{CL}(p) &= n_k\left(\frac{1-p}{p}+a\right), \\ \mathrm{LCL}(p) &= n_k\left(\frac{1-p}{p}+a\right)-g\sqrt{\frac{n_k(1-p)}{p^2}}. \end{aligned} \tag{12}$$

Note that the smallest possible value of the total number of events is $n_k a$. Thus, if $\mathrm{LCL} < n_k a$ in the above limit, we set up $\mathrm{LCL} = n_k a$.

The h chart (about the average number of events) with n_k has the following control limits.

$$\begin{aligned} \mathrm{UCL}(p) &= \frac{1-p}{p}+a+g\sqrt{\frac{1-p}{n_k p^2}}, \\ \mathrm{CL}(p) &= \frac{1-p}{p}+a, \\ \mathrm{LCL}(p) &= \frac{1-p}{p}+a-g\sqrt{\frac{1-p}{n_k p^2}}. \end{aligned} \tag{13}$$

Note that the smallest possible value of the average number of events is a. Thus, if $\mathrm{LCL} < a$ in the above limit, we set up $\mathrm{LCL} = a$.

Since the parameter p is unknown in practice, we need to estimate it. Suppose that there are m samples (subgroups) from the experiments and the sample size of the ith sample is n_i. Let X_{ij} be the number of independent Bernoulli trials (events) until the appearance of the first nonconforming event in the ith sample, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$. Then X_{ij}'s are independent and identically distributed geometric random variables with location shift a and Bernoulli probability p. Then the estimators, $\hat{p}_{\mathrm{t}}$ and $\hat{p}_{\mathrm{cdf}}$, with these samples are easily obtained from (10) and (11), given by

$$\hat{p}_{\mathrm{cdf}} = 1 - \left[\frac{\hat{F}(s+t-a+1)-\hat{F}(t)}{\hat{F}(s)}\right]^{1/(t+1-a)}$$

and

$$\hat{p}_{\mathrm{t}} = \frac{(a+d)-2\bar{\bar{X}}}{(\bar{\bar{X}}-a+1)(d-\bar{\bar{X}})-S^2},$$

where $\hat{F}(t) = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbb{I}(X_{ij} \le t)$, $\bar{\bar{X}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}/N$ with $N = \sum_{i=1}^{m} n_i$, and $S^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{\bar{X}})^2$. For $\hat{p}_{\mathrm{cdf}}$, $t = [q_{\gamma/2}]$ and $s = [q_\gamma - q_{\gamma/2} + a - 1]$ are obtained from all the m samples.

For a given robust estimator $\hat{p} = \hat{p}_{\text{cdf}}$ or $\hat{p} = \hat{p}_{\text{t}}$ we can obtain the g robust control limits by plugging $\hat{p}$ into the control limits in (12) or (13). As an illustration, the robust control limits of the g or h chart are easily obtained as follows.

```
> library(rQCC)
> x1 = c(11, 2, 8, 2, 4)
> x2 = c(1, 1, 11, 2, 1)
> x3 = c(1, 7, 1)
> x4 = c(5, 1, 3, 6, 5)
> x5 = c(13, 2, 3, 3)
> x6 = c(3, 2, 6, 1, 5)
> x7 = c(2, 2, 8, 3, 1)
> x8 = c(1, 3, 4, 6, 5)
> x9 = c(2, 8, 1, 1, 4)
> data = list(x1, x2, x3, x4, x5, x6, x7, x8, x9)

> result = racc(data, gamma=0.9, type="g", location=1, gEstimator="cdf",
nk=5)
> summary(result)
> plot(result)
```

3 Construction of the robust exponential and Weibull t control charts

3.1 Robust exponential t control chart

The cdf of the exponential distribution is given by

$$F(x) = 1 - e^{-x/\theta},$$

where $\theta > 0$. Let $x_{(i)}$ be the values of the order statistics such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. For notational convenience, we denote $p_i = F(x_{(i)})$. Then the exponential cdf can be linearized as

$$-\log(1 - p_i) \cdot \theta = x_{(i)},$$

where $i = 1, \cdots, n$. We propose a robust estimate of θ as follows:

$$\hat{\theta} = \underset{1 \leq i \leq n}{\text{median}} \left\{ -\frac{x_{(i)}}{\log(1 - p_i)} \right\}. \tag{14}$$

Then, similar to the conventional exponential t chart, its robust version is constructed as follows:

$$\begin{aligned} \text{LCL} &= \{-\log(1 - \alpha/2)\} \cdot \hat{\theta}, \\ \text{CL} &= \{-\log(1/2)\} \cdot \hat{\theta}, \\ \text{UCL} &= \{-\log(\alpha/2)\} \cdot \hat{\theta}, \end{aligned}$$

where $\alpha/2 = \Phi(-g)$ and $\hat{\theta}$ is obtained by (14). For more details on the conventional exponential t control chart, one can refer to `vignette("acc", package="rQCC")`. The control limits of the exponential t chart are obtained as follows.

```
> racc(x, type="t")
```

3.2 Robust Weibull t control chart

The cdf of the Weibull distribution is given by

$$F(x) = 1 - \exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\},$$

where $\theta > 0$ and $\beta > 0$ represent the scale and shape parameters, respectively. Let $x_{(i)}$ be the values of the order statistics such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. For notational convenience, we denote $p_i = F(x_{(i)})$. Then p_i can be easily estimated by using the plotting position, an increasing step function jumping at $x_{(i)}$. In this `rQCC` package, we use the `ppoints()` function to estimate $p_i = F(x_{(i)})$, which is based on Blom [6]. Then it is given by

$$p_i = \begin{cases} \dfrac{j - 3/8}{n + 1/4} & \text{for} \quad n \leq 10 \\ \dfrac{j - 1/2}{n} & \text{for} \quad n \geq 11 \end{cases}, \ . \tag{15}$$

The Weibull cdf can be linearized as

$$\log\left(-\log(1 - p_i)\right) = -\beta \log\theta + \beta \log x_{(i)}, \tag{16}$$

where $i = 1, \cdots, n$. By denoting $y_i^* = \log\left(-\log(1 - p_i)\right)$, $x_i^* = \log x_{(i)}$, $\beta_0^* = -\beta\log\theta$, and $\beta_1^* = \beta$, we can rewrite (16) as

$$y_i^* = \beta_0^* + \beta_1^* x_i^*, \quad i = 1, \cdots, n.$$

Then, based on observations $\{(x_1^*, y_2^*), \cdots (x_n^*, y_n^*)\}$, we can easily calculate the estimate of β_1^*, denoted by $\hat{\beta}_1^*$, by using the repeated median estimate [7], which is given by

$$\hat{\beta}_1^* = \underset{1 \leq i \leq n}{\text{median}} \ \underset{j \neq i}{\text{median}} \ \frac{y_i^* - y_j^*}{x_i^* - x_j^*}.$$

After $\hat{\beta}_1^*$ is obtained, we can estimate $\hat{\beta}_0^*$ easily using

$$\hat{\beta}_0^* = \underset{1 \leq i \leq n}{\text{median}} \left(y_i^* - \hat{\beta}_1^* x_i^*\right).$$

After $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are obtained, we obtain the original parameter estimates by reparametrizing as

$$\hat{\beta} = \hat{\beta}_1^* \ \text{ and } \ \hat{\theta} = e^{-\hat{\beta}_0^*/\hat{\beta}_1^*}. \tag{17}$$

Then, similar to the conventional Weibull t chart, its robust version is constructed as follows:

$$\begin{aligned} \text{LCL} &= \{-\log(1-\alpha/2)\}^{1/\hat{\beta}} \cdot \hat{\theta}, \\ \text{CL} &= \{-\log(1/2)\}^{1/\hat{\beta}} \cdot \hat{\theta}, \\ \text{UCL} &= \{-\log(\alpha/2)\}^{1/\hat{\beta}} \cdot \hat{\theta}, \end{aligned}$$

where $\hat{\beta}$ and $\hat{\theta}$ are from (17). For more details on the conventional Weibull t control chart, one can refer to `vignette("acc", package="rQCC")`. The control limits of the Weibull t chart are obtained as follows.

```
> racc(x, type="t", tModel="W")
```

References

[1] J. C. Benneyan. Performance of number-between g-type statistical control charts for monitoring adverse events. *Health Care Management Science*, 4:319–336, 2001.

[2] C. Park and M. Wang. A study on the g and h control charts. *Communication in Statistics – Theory and Methods*, To appear, 2023.

[3] C. Park, L. Ouyang, and M. Wang. Robust g-type quality control charts for monitoring nonconformities. *Computers & Industrial Engineering*, 162:107765, 2021.

[4] C. Park, K. Gou, and M. Wang. A study on estimating the parameter of the truncated geometric distribution. *The American Statistician*, 76(3):257–261, 2022.

[5] C. Kapadia and R. Thomasson. On estimating the parameter of a truncated geometric distribution by the method of moments. *Annals of the Institute of Statistical Mathematics*, 27:269–272, 1975.

[6] G. Blom. *Statistical Estimates and Transformed Beta Variates*. Wiley, New York, 1958.

[7] A. F. Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982.