Assam University Journal of Science &
Technology : Physical Sciences and Technology
Vol. 5 Number II
21-28, 2010

On Application of Generalized Poisson and Compound Distribution in Control Charts for Number of Defects Per Unit

Kishore K. Das[1] and Dibyojyoti Bhattacharjee[2]
[1]Department of Statistics, Gauhati University, Guwahati, Assam, India.
[2]Department of Business Administration, Assam University, Silchar, Assam, India.
Correspondence; [2]e-mail : dibyojyotibhattacharjee@gmail.com

Abstract

The Poisson distribution, in general remains sensitive to small departure of frequencies especially at the right tail of the distribution. In many situations it may happen that the Generalized Poisson Distribution (GPD) or a compound distribution provides a closer fit to a frequency distribution compared to the usual Poisson distribution. The article examines the possibility of using the GPD and the compound distribution as alternative models to construct the control chart for number of defects per unit (c-chart).

Key words: Control Chart, Poissoness Plot, SQC, Statistical Graphics, Generalized Poisson Distribution, Compound Distribution.

Introduction

The performance of the usual control charts for monitoring number of defects per unit may be greatly affected when the data generated from the stable process does not fit the Poisson distribution well. In many situations such a type of control chart is used when the underlying population distribution may not actually follow the Poisson law and hence, it may not always yield desired result. The departure from the hypothetical model can be due to various factors such as shift in the process average while variance remains constant, dependency amongst the trials, probability of occurrence not constant as assumed, unusual tail behaviour etc. In this article, we explore the possibility of using a generalized Poisson distribution and a compound distribution for the control chart of the number of defects per unit (c-chart), when the underlying distribution does not fit to a Poisson distribution properly or shows some abnormality at the tails. Here by considering a real data set and then a simulated dataset from the Poisson distribution, it is been shown that the frequency distribution of such data gives closer fit to the GPD or compound distribution compared to the Poisson model. We then construct the control charts, OC curves for the GPD case or the compound distribution and compare it with the OC curve under the Poisson model.

The 3-σ Control Chart for Number of Defects Per Unit

Let x_i be the number of defective items in a sample of size n taken from the ith subgroup, where $i = 1, 2,\ldots, k$. Let the population fraction defective be given by p. Now in case the fraction defective is very small and the sample size is very large

	λ known	λ estimated
UCL	$\lambda + 3\sqrt{\lambda}$	$m_1 + 3\sqrt{m_1}$
CL	λ	m_1
LCL	$\lambda - 3\sqrt{\lambda}$	$m_1 - 3\sqrt{m_1}$

In most cases the true value of the parameters are unknown and estimated from actual data generated from the process for setting up control charts. The LCL is taken to be zero in case the value comes out to be negative.

OC Curve for a Control Chart

OC curve depicts how the control charts operates under varying process conditions. The probability of accepting a lot having a fraction defective *p*, is denoted by *L(p)* and is called as the operating characteristic (OC) function and the corresponding curve as the OC curve. If gives the risk of concluding that the process is in control at a specified level when actually the process is operating at a different level.

The Poissonness Plot

For a given set of data related to number of defects per unit, there are different techniques to check if the said data abide by the Poisson law. While the various parametric and non-parametric techniques speaks about the overall goodness of fit but a few graphical tools also exist that study the goodness of fit of discrete data to a given distribution. One advantage of using a graphical technique is that it reveals the goodness of fit point-wise and thus the graphical tools reveal some additional features related to the data. One such graphical technique is the Poissonness plot developed by Hoaglin (1980).

Let *X* follows Poisson distribution with parameter l and if the total frequency is N, then the expected frequencies of the distribution corresponding to $x = 0, 1, 2, \ldots$ etc. is

$$\eta_x = N \,.\, P(X = x) = N.\ \frac{e^{-\lambda}\lambda^x}{x!}, \qquad (1)$$

which takes a linear form when the logarithm of both sides is considered i.e.,

$$\log\,[\eta_x (x)!] = \log N - \lambda + x \log \lambda \qquad (2)$$

Thus, from (2) it is clear that there is a linear relationship between *x* and $\log(\eta_x (x)!)$ with slope as $\log \lambda$ and intercept as, $\log N - \lambda$. Now, if $\log(o_x (x)!)$ is plotted against *x* then we get the points approximately follow a straight line, provided the data comes from a Poisson distribution. Some of the other references approaching the similar or related issues are Kemsley (1966), Dubey (1966), Ord (1967), Friendly (2003) etc.

The objective of the paper is to detect the departure from Poisson law for a data set especially at the tail area, if any, by drawing the corresponding Poissonness plot with bands around the theoretical straight line. In case such departures are noted, a generalized Poisson Distribution Type-II or compound distribution can be thought of which may provide a better fit. The UCL and the LCL are obtained under the changed circumstances and the corresponding OC curves are also drawn to see the performance of the control chart under the varying process conditions.

Testing a Real Data Set

The data set used here appeared in C. I. Bliss, Biometrics, 1953, 9, p. 176-200. The data set data gives the distribution of the bacterial clumps per field as counted in a milk firm. Here a clump is considered as a defect and each field (A microscopic slide was split into 400 regions of equal area- each region is called a 'field'.) is considered as a unit. Under the assumption that the data comes from the Poisson distribution the maximum likelihood estimator for the parameter of the distribution is computed, and accordingly fitted to a Poisson distribution.

From the data the mle of the parameter (λ) of the Poisson distribution is found to be 2.3985 and accordingly the expected frequencies are computed (See Table 1.). Now before drawing the Poissonness plot to the data, let an accessory be added to it. The accessory is a pair of bands around the theoretical straight line (2). The band would act as a guard line and helps in detecting the observed frequencies that differ significantly from their expected values. The band also eliminates any chance of manipulation of the scales along the y axis, which may make a point look closer or far away from the hypothetical straight line. For a line of the form $y = a + bx$, the 95% confidence interval for y_p for a fixed value of *x* is,

$$y_p \pm t_{0.05,n-2}\hat{\sigma}_{y_p} \qquad (3)$$

where, $\hat{\sigma}_{y_p} = \hat{\sigma}_{y|x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ and $\hat{\sigma}_{y|x} = \left(\frac{\sum (y_i - \bar{y})^2 - b\sum (x_i - \bar{x})(y_i - \bar{y})}{n-2} \right)^{1/2}$

In Table 1, all the computations based on the equations (2) and (3) given provided.

Table 1: The Bliss data and related computations

No. of defects per unit (x)	Observed Frequency (o_x)	Expected Frequency (η_x)	$log\ [o_x (x)!]$	$log\ [\eta_x (x)!]$	Lower band	Upper band
0	56	35.57	4.0254	3.5715	2.4897	4.6532
1	104	86.08	4.6444	4.4552	3.5104	5.4
2	80	104.15	5.0552	5.339	4.5201	6.158
3	62	84.02	5.9189	6.2228	5.5127	6.9328
4	42	50.83	6.9157	7.1065	6.4795	7.7336
5	27	24.6	8.0833	7.9903	7.4092	8.5714
6	9	9.92	8.7765	8.8741	8.2930	9.4552
7	9	3.43	10.7224	9.7578	9.1308	10.3849
8	5	1.04	12.2141	10.6416	9.9316	11.3517
9	3	0.28	13.9004	11.5254	10.7064	12.3443
10	2	0.07	15.7976	12.4091	11.4643	13.3540
11	1	0.02	17.5023	13.2929	12.2111	14.3747

Equation (3) gives the required bands. The Table 1 has sufficient ingredients for drawing the Poissonness plot to Bliss data, and the corresponding plot appears in Fig. 1.

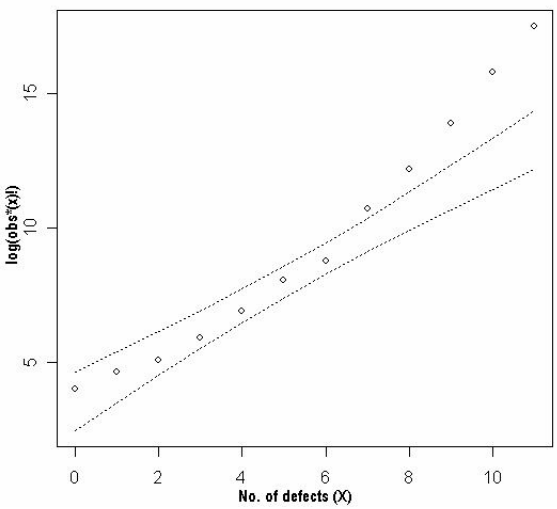

Fig. 1 : Poissonness plot with bands for the Bliss data

From the diagram one thing becomes evident, that the observed frequencies at the tail of the distribution do not follow the Poisson law. The plotted points corresponding to X = 7, 8, 9, 10 and 11 falls outside the band.

Opting for More Appropriate Distributions

Based on the discussion in the previous section, it can be concluded that the tail of the fitted distribution does not tally with the expected frequencies. Thus, a distribution with a better fit especially at the tails must be sought for. There may be several options, initially we think of the compound distribution and then the generalized Poisson distribution.

Compound Distribution

Let us consider the parameter of the distribution $\lambda = \nu\alpha$, where, ν is a constant and α is a random variable. As λ ranges from 0 to ∞ so α also lies in the same range. A probable distribution of α in the range mentioned above, is the gamma distribution with parameter a (say). Accordingly the distribution of X can take the following compound form:

$$P(X = x) = \int_A f(x, \nu\alpha) g(\alpha) d\alpha$$

$$= \int_0^\infty \frac{e^{-\nu\alpha}(\nu\alpha)^x}{x!} \times \frac{1}{\Gamma(a)} e^{-\alpha} \alpha^{a-1} d\alpha = {}^{x+a-1}C_{a-1} \left(\frac{\nu}{\nu+1}\right)^x \left(\frac{1}{\nu+1}\right)^a \quad (4)$$

which follows the negative binomial distribution with parameters a and $p = \frac{1}{\nu+1}$. We know that the mean and variance of the negative binomial distribution is given by $\frac{a(1-p)}{p}$ and $\frac{a(1-p)}{p^2}$.

So we have $\hat{p} = \frac{m_1}{s^2}$ (5)

and $\hat{a} = \frac{\hat{p} m_1}{1 - \hat{p}}$ (6)

where m_1 and s^2 are the sample mean and sample variance respectively.

From the data provided in Table 1 (Observed Frequencies), we compute the values of the parameters as $\hat{p} = 0.5919$ and $\hat{a} = 3.5097 \cong 4$.

Accordingly replacing the estimates of the parameters in (4) we get the probabilities under the compound distribution and accordingly the expected frequencies (See Table. 2).

The Generalized Poisson Distribution

Another suggestive distribution to the data set can be the generalized Poisson distribution type-II (Nandi, Nath and Das, 1999) as an alternative to the Poisson *pmf*. The *pmf* of the GPDII is given by,

$$P(X = x) = \frac{1}{(x)!} \frac{a^2(a+z)}{a(1-z)+z} (a + zx)^{x-2} e^{-(a+zx)}, \; a + xz > 0; \; x = 0, 1, 2, \ldots; a < 0; \; |z| < 1 \quad (7)$$

$= 0$, Otherwise

which will be referred to here as the GPD. The reason for choosing such a distribution is as follows:

(a) The distribution can take different shapes and can provide modeling for different types of data sets especially when the other discrete models provide a poorer fit.

(b) When the probability distribution appears to have a longer tail GPD gives better fit.

(c) When $z = 0$, the GPD tends to the Poisson distribution, so for a data set if, z is not equal to 0 or does not tend to zero, then the GPD may give a better fit.

The estimates of the parameters of the GPD viz. a and z, obtained via method of moments are as follows :

$$\hat{a} = \sqrt{\frac{m_1(m_2 - m_1)}{m_2 - m_1^2}} \quad (8)$$

$$\hat{z} = 1 - (\hat{a} + 1)\frac{m_1}{m_2} \quad (9)$$

where, $m_i = i^{th}$ sample raw moment about zero.

For the data provided in Table 1 (Observed Frequencies), we compute the values of the parameters as $\hat{a} = 2.1104$ and $\hat{z} = 0.2431$ and accordingly the expected frequencies under GPD can be computed. Table 2, shows that the observed and expected frequencies of the variate under different models.

Table 2 : Observed Frequencies and Expected Frequencies from Different Models

No. of defects per unit (x)	Observed Frequencies	Expected Frequency		
		Compound Distribution	GPD	Poisson Distribution
0	56	63.49	61.99	35.57
1	104	90.94	91.98	86.08
2	80	83.68	84.89	104.15
3	62	62.72	63.01	84.02
4	42	41.66	41.35	50.83
5	27	25.54	25.10	24.6
6	9	14.78	14.47	9.92
7	9	8.19	8.04	3.43
8	5	4.39	4.36	1.04
9	3	2.29	2.31	0.28
10	2	1.17	1.21	0.07
11	1	1.15	1.29	0.01
Total	400	400	400	400
χ^2		5.7526	5.1491	30.8452
d.f.		6	6	5

It can be noted that both GPD and compound distribution gives a better fit to the observed frequencies compared to the Poisson model. Also the GPD fits the data even better than the compound distribution as evident from Table 2. This may also be verified by calculating the chi-square statistic for the goodness of fit for all the models- in this case the GPD gives the minimum chi-square value.

Control Limits and OC Curves

The control limits under the compound distribution are given by:

	a, ν known	a, ν estimated
UCL	$av + 3\sqrt{av(v+1)}$	$m_1 + 3s$
CL	av	m_1
LCL	$av - 3\sqrt{av(v+1)}$	$m_1 - 3s$

and the control limits under GPD II are given by:

	a, z known	a, z estimated
UCL	$\frac{a^2}{a(1-z)+z} + 3\sqrt{\frac{a^2(a+z)}{[a(1-z)+z]^2(1-z)}}$	$m_1 + 3\sqrt{m_2 - m_1^2}$
CL	$\frac{a^2}{a(1-z)+z}$	m_1
LCL	$\frac{a^2}{a(1-z)+z} - 3\sqrt{\frac{a^2(a+z)}{[a(1-z)+z]^2(1-z)}}$	$m_1 - 3\sqrt{m_2 - m_1^2}$

For the given data, $m_1 = 2.42$, $m_2 = 9.945$ and $s = 2.022$, so the control limits under compound distribution and GPD are as follows:
UCL = 8.486, CL = 2.42 and LCL = 0 and
UCL = 8.486, CL = 2.42 and LCL = 0 respectively.
Thus, both the distribution gives the same control limits and hence any one of them can be preferred over the Poisson distribution.

While for the same data the control limits under Poisson distribution is given by:
UCL = 7.0869, CL = 2.42 and LCL = 0

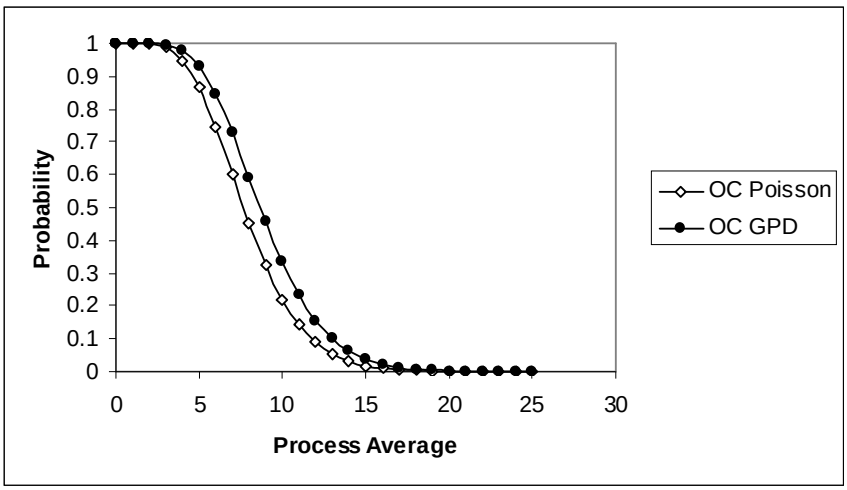

Fig. 2: Probability of a point falling within control limits (OC curves)

As the compound distribution and the GPD gives the same control limits and provides better fit to the data compared to the Poisson distribution so the control chart constructed using the compound distribution/GPD is better. Using the Poisson distribution would lead to the under estimation of the UCL and accordingly in such cases there is a chance of declaring a process out of control when the process is actually within control.

The Simulated Data Set

Next data of size 150 is simulated from the Poisson distribution with mean 2.5. Under the assumption that the data comes from the said distribution the maximum likelihood estimator for the parameter of the Poisson distribution is computed, and accordingly fitted to a Poisson distribution. From the data the estimate of the parameter (λ) of the Poisson distribution is found to be 2.5867 and based on that value of λ the expected frequencies are computed.

Table 3 : Calculation based on the simulated data

No. of defects per unit	Observed Frequency	Rounded Expected Frequency
0	14	11
1	32	29
2	29	38
3	40	33
4	17	21
5	4	11
6	9	5
7	4	2
8	1	0
Total	150	150

The corresponding Poissonness plot with bands is as follows:

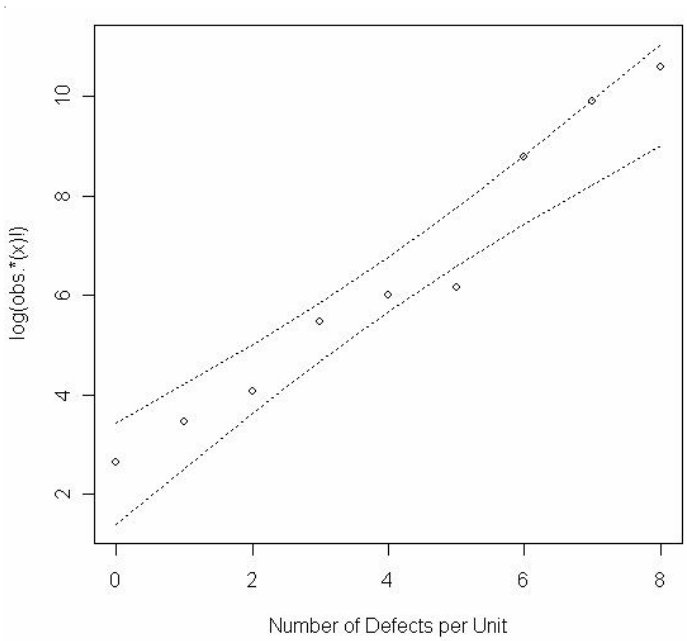

Fig. 3 : Poissonness plot with bands for the simulated data

From the diagram one thing becomes evident, that the observed frequencies at the tail of the distribution are not in close proximity to the Poisson law. The plotted points corresponding to X = 5, falls outside the band while the points corresponding to X = 6 and 7 falls just on the band.

From the data provided in Table 3 (Observed Frequencies), we compute the values of the parameters as $\hat{p} = 0.8427$ and $\hat{a} = 13.8574 \cong 14$ which are used as parameters of the compound distribution. The estimates of parameters for GPD II are $\hat{a} = 2.45878$ and $\hat{z} = 0.08332$. Table 4, shows that the observed and expected frequencies of the variate under different models computed from the simulated data.

Table 4. Observed Frequencies and Expected Frequencies from Different Models

Counts (x)	Observed Frequencies	Expected Frequency (Rounded to integral values)		
		Compound Distribution	GPD	Poisson Distribution
0	14	14	14	11
1	32	30	31	29
2	29	35	36	38
3	40	30	30	33
4	17	20	20	21
5	4	11	11	11
6	9	6	5	5
7	4	3	2	2
8	1	1	1	0
Total	150	150	150	150
χ^2		10.9997	14.1313	16.9614
d.f.		4	4	5

It can be noted that both GPD and compound distribution gives a better fit to the observed frequencies compared to the Poisson model. Also the compound distribution fits the data even better than the GPD as evident from Table 4.

For the given data, $m_1 = 2.5866$, $m_2 = 9.76$ and $s = 1.7519$, so the control limits under GPD, compound distribution and Poisson distribution are: UCL = 7.8426, CL = 2.5866 and LCL = 0, UCL = 8.1123, CL = 2.5866 and LCL = 0 and UCL = 6.41147, CL = 2.5866 and LCL = 0 respectively.

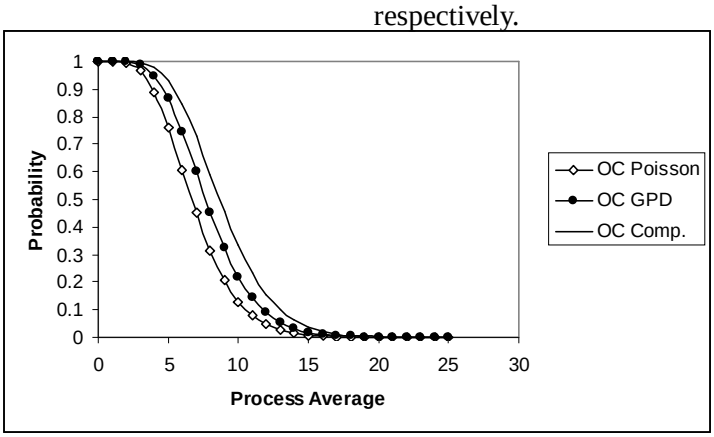

Fig. 4: Probability of a point falling within control limits (OC curves)

As the compound distribution gives the best fit to the data compared to the Poisson distribution and the GPD so the control chart constructed using the compound distribution is the true one. Using the Poisson distribution or the GPD would lead to the under estimation of the UCL and accordingly in such cases there is a chance of declaring a process out of control when the process is actually in control.

Conclusion

Statistical process control when applied correctly can be very effective. From the above study it can be observed that the application of control charts without perfect knowledge about the shape of the underlying distribution may result in wrong signals about the status of the manufacturing process. It is therefore important to check whether or not the data generated from a given process follows a given type of probability distribution and satisfies its chief characteristics before choosing a particular distribution as the underlying model. Here for the data set under consideration we have seen that the compound distribution is a better alternative to the usual Poisson distribution, but there may be cases when the GPD may perform better than the compound distribution.

As an extension to this work one may use this technique for drawing an appropriate p chart, where the under lying distribution of fraction defective is usually the binomial distribution. In such a case the departure of the observed frequencies from those expected from the binomial law, if any, can be scrutinized using the binomialness plot (See Hoaglin (1980)). If any such violation is noted then one may fit a GPD or an appropriate compound distribution or whatever is suitable.

Reference

Dubey, S. D. (1966). Graphical tests for discrete distributions. *Amer. Statistician.* 20 (3): 23-24

Friendly, M. (2003). Categorical data analysis with graphics. http://www.math.yorku.ca/SCS/Courses/grcat/grcfoils1_4.pdf

Hoaglin, D.C. (1980). A Poissonness plot. *The American Statistician.* 34 (3): 146-149

Kemsley, W. F. F. (1966). Discussion on the paper by Messrs Chatfield. Ehrenberg and Goodhardt. *Journal of the Royal Statistical Society A.*129: 360-362

Nandi, S. B.; Nath, D.C.; Das, K. K. (1999). A class of generalized Poisson distributions. *Statistica. anno.* LIX (4) : 487-498

Ord, J. K. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society A.* 130: 232–238