USE OF GINI'S MEAN DIFFERENCE IN ANALYSIS OF MEANS

A. Saghir, A. Saleem[*] and M. S. Anjum[**]

Mirpur University of Science & Technology MUST Mirpur.
[*]Department of Statistics, University of Azad Jammu & Kashmir, Muzaffarabad.
[**]Faculty of Veterinary Sciences, University of Poonch, Rawalakot, Azad Kashmir
Corresponding Author: drazharsaleem@yahoo.com

ABSTRACT: In this study the estimate obtained from Gini's mean difference (G) is used in the construction of decision limits for analysis of means (ANOM). The proposed limits are constant free because the estimate obtained from the G-Chart demands no constant like d_2 and $c_4$used for R and S charts for the unbiased estimation of σ . The comparison of the proposed strategy to existing strategies is made using simulated data from different populations. The comparison reveals that the decision limits constructed by the proposed strategy are least affected by departure from normality.

Key words:Decision limits; Analysis of Mean (ANOM); Simulations; Non-Normality.

INTRODUCTION

Analysis of means (ANOM) is a technique originally developed by Ott (1967) for comparing a group treatment means to see if any one of them differs significantly from the overall mean. It can be thought as an alternative to the analysis of variance (ANOVA) and in fact for only two treatments both the procedures are equivalent. Comparing the sample mean values to the overall grand or target mean value, about which decision lines have been constructed, carries out Ott's procedure. If any sample mean lies outside these decision lines, it is declared significantly different from the target mean level. An ANOM chart, conceptually similar to control chart, portrays decision lines so that statistical significance as well as practical significance of samples may be assessed simultaneously.

The analysis of means has an advantage over analysis of variance, its results can be presented graphically, making the procedure easy to explain and visualize and allowing for an assessment of practical significance as well as statistical significance. Although ANOM has practical advantages over ANOVA but it has no optimal advantage in any mathematical sense because it behaves similarly. Therefore, we can say that two methods are nearly enough equivalent that both will disclose any lack of control among averages and when one say control is good, the other will too. Ott (1967) introduced analysis of means procedure, based on multiple significance tests following the pioneering work of Halperin (1955), for controlling the group of means instead of one mean at a time.

In the analysis of means the estimate of σ is usually calculated from sample range (R) and standard deviation (S). The sample statistic R and S are defined as:

$$R = X_{\max} - X_{\min} \qquad (1)$$

and

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad (2)$$

The sample statistic R is linear function of only two extreme values as obvious from (1). It is widely used to estimate the true process standard deviation σ. It ignores a lot of sample information for larger values of *n* that decreases its efficiency. The sample statistic S is non-linear function of data, as obvious from (2) and is very sensitive to the presence of outlier in the data.

The Gini's mean difference of a set {x_1, x_2, ..., x_n} is defined as:

$$G = 2\sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{i=1}^{n} |x_i - x_j| \Big/ n(n-1) \qquad (3)$$

Jordan (1869) claimed improved precision of G over Gauss's root mean square estimator for normal distribution. According to Von Andrae (1872) the asymptotic efficiency of *G* relative to S is 97.8% for normally distributed data and G, given its much simpler calculation, is a serious competitor to the usually preferred *S*.

Gini (1912) used *G* defined in (3) as an index of variability in a population consisting of x_1, x_2, ..., x_n therefore it is generally known as Gini's mean difference. David (1968) showed that for normally distributed characteristic the sample statistic $\frac{\sqrt{\pi}}{2} \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} |x_i - x_j|}{n(n-1)/2} (= \frac{\sqrt{\pi}}{2} G)$ is an unbiased measure of process variability. Let we name it K , i. e.

$$K = \frac{\sqrt{\pi}}{2} \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} |x_i - x_j|}{n(n-1)/2} \qquad (4)$$

Decision limits in analysis of means: For comparing the individual sample means with the overall mean, Ott (1967) computed the decision lines based on range as an estimate of true process variability as:

$$\left.\begin{aligned} LDL &= \bar{\bar{x}} - H_{(\alpha)}\hat{\sigma}_{\bar{x}} \\ CL &= \bar{\bar{x}} \\ UDL &= \bar{\bar{x}} + H_{(\alpha)}\hat{\sigma}_{\bar{x}} \end{aligned}\right\} \qquad (5)$$

where

$\bar{\bar{x}}$: average of *k* sample means,

$$H_\alpha = \max\left[\frac{\bar{x}_n-\bar{\bar{x}}}{\hat{\sigma}_{\bar{x}}}, \frac{\bar{\bar{x}}-\bar{x}_1}{\hat{\sigma}_{\bar{x}}}\right],$$

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}, \quad \hat{\sigma} = \frac{\bar{R}}{d_2^*},$$

Where,

$\bar{R}$, average of *k* sample ranges,

d_2^*, a factor for estimating σ from $\bar{R}$ and it depends on *k* (whereas the usual d_2 factor in the control chart is independent of *k*), and *n*, sample size.

Tables of $H_{(\alpha)}$ for $\alpha = 0.05$ and 0.01have been developed by Ott (1967) for selected values of *k* and selected degree of freedom for error $v = 0.90k(n-1).$

In the situation where number of observations in treatment is constant and is small, Sheesley (1981) used range as an estimate of the within group standard deviation instead of obtaining a pooled estimate of the variance and then taking its square root. The decision limits under the Sheesley's procedure are constructed as:

$$\left.\begin{aligned} LDL &= \bar{\bar{x}} - h_{(\alpha,m,k)}\hat{\sigma}_{\bar{x}} \\ CL &= \bar{\bar{x}} \\ UDL &= \bar{\bar{x}} + h_{(\alpha,m,k)}\hat{\sigma}_{\bar{x}} \end{aligned}\right\} \qquad (6)$$

z $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}, \quad \hat{\sigma} = \frac{\bar{R}}{d_2},$

Where,

$\bar{R}$ is average of *k* sample ranges and $h_{(\alpha,m,k)}$, is critical values depend upon number of sample *k* and degrees of freedom for error *m* are derived by Nelson (1983).

Nelson (1982) obtained the exact critical points for $h_{(\alpha)}$ and used the decision limits as:

$$\left.\begin{aligned} LDL &= \bar{\bar{x}} - h_{(\alpha)}S\sqrt{(k-1)/kn} \\ CL &= \bar{\bar{x}} \\ UDL &= \bar{\bar{x}} + h_{(\alpha)}S\sqrt{(k-1)/kn} \end{aligned}\right\} \qquad (7)$$

where *S*is the pooled standard deviation and $h_{(\alpha)}$ is a critical point, which depends on *k* and *v* (degrees of freedom in S^2). Instead of estimating the pooled standard deviation S we may use $\frac{\bar{S}}{c_4}$ (*c*$_4$ is a constant quantity that depends upon sample size n) as Sheesley (1981) used $\frac{\bar{R}}{d_2}$ as estimate of within group standard deviation.

Proposed decision limits: The process variability control charts are used to monitor process variability e.g. R chart, S chart etc. These charts help in finding the unbiased estimate for process variability e.g. If R chart shows stability, the constant d_2, when divided by $\bar{R}$, provides an unbiased estimate of process variability. Whereas S chart shows stability, when $\bar{S}$ divided by c_4, provides an unbiased estimate of process variability. Riaz and Saghir (2005) developed a design structure of a new Shewhart type control chart namely G -Chart using the estimate K defined in (4) as an estimate of process variability. This chart provides a constant free environment for unbiased estimation of σ (i.e. $\bar{K}$ directly provides unbiased estimate of σ).

In this study we propose a scheme for decision limits based on constant free estimate of σ obtained from G -Char following the idea of Sheesley (1981) as:

$$\left.\begin{aligned} LDL &= \bar{\bar{x}} - h_{(\alpha,m,k)}\hat{\sigma}_{\bar{x}} \\ CL &= \bar{\bar{x}} \\ UDL &= \bar{\bar{x}} + h_{(\alpha,m,k)}\hat{\sigma}_{\bar{x}} \end{aligned}\right\} \qquad (8)$$

where

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}, \quad \hat{\sigma} = \bar{K}$$

$\bar{K}$ is average of *k* sample *K*'s as defined in (4) for an appropriate sample size.

Simulation study: Based on 1,000 random samples of sizes 3 and 10 drawn from normal population $\hat{\sigma}$ is estimated using $\frac{\bar{R}}{d_2}$, $\frac{\bar{S}}{c_4}$ and $\bar{K}$, and the decision limits constructed for ANOM charts using the data sets provided in Appendix. Later, 1,000 random samples of sizes 3 and 10 drawn from some non-normal populations having same standard deviation as the standard deviation of the comparable normal distribution discussed first and the decision limits constructed for ANOM charts using the data sets provided in Appendix.

The distributions used for this comparison are:

1. Normal distribution with mean 0 and standard deviation 2.3833 for sample sizes 3.
2. Normal distribution with mean 25 and standard deviation 8 for sample sizes 10.

3. t- distribution with 5 d.f.
4. t- distribution with 10 d.f.
5. t- distribution with 20 d.f.
6. Exponential distribution with 0 as location parameter and 1 as scale parameter.

Based on these 1,000 random numbers drawn from above distribution, process variability is calculated using range, standard deviation and Gini's mean difference and for comparison purposes results are provided here in Table 1.

Table 1. Summary statistics of standard deviation for different distributions.

(i) Normal distribution with $\bar{\bar{X}}$ mean and 2.3833 st.dev for *n* = 3			
Estimate used	N	Mean	Std. Dev.
Using Range	1000	2.3826	0.2795
Using Gini's	1000	2.3835	0.2802
Using Std. Dev.	1000	2.3830	0.2787
(ii) t-distribution with 5 d.f. for *n* = 3.			
Using Range	1000	3.4639	0.4698
Using Gini's	1000	3.4575	0.4718
Using Std. Dev.	1000	3.4712	0.4742
(iii) t-distribution with 10 d.f for *n* = 3.			
Using Range	1000	2.5956	0.3622
Using Gini's	1000	2.5546	0.3669
Using Std. Dev.	1000	2.6023	0.3642
(iv) t-distribution with 20 d.f for *n* = 3.			
Using Range	1000	2.4610	0.3430
Using Gini's	1000	2.4523	0.3452
Using Std. Dev.	1000	2.4759	0.3403
(v) Exponential with 0 as location and 1 as scale parameter for *n* = 3			
Using Range	1000	2.4233	0.4192
Using Gini's	1000	2.4153	0.4294
Using Std. Dev.	1000	2.4554	0.4346
(vi) Normal distribution with $\bar{\bar{X}}$ mean and 8 st.dev for *n* = 10			
Using Range	1000	8.0236	0.4857
Using Gini's	1000	8.0382	0.9489
Using Std. Dev.	1000	8.0269	0.4551
(vii) t-distribution with 5 d.f for *n* = 10.			
Using Range	1000	8.2699	0.7129
Using Gini's	1000	7.9081	1.3440
Using Std. Dev.	1000	7.7857	0.5464
(viii) t-distribution with 10 d.f for *n* = 10.			
Using Range	1000	8.1783	0.5384
Using Gini's	1000	7.9441	1.0123
Using Std. Dev.	1000	8.1893	0.4985
(ix) t-distribution with 20 d.f for *n* = 10.			
Using Range	1000	8.0762	0.5182
Using Gini's	1000	7.9868	0.9494
Using Std. Dev.	1000	8.0969	0.5023
(x) Exponential with 0 as location and 1 as scale parameter for *n* = 10			
Using Range	1000	9.2835	1.3885
Using Gini's	1000	8.7705	2.2182
Using Std. Dev.	1000	9.3795	1.3509

Table1 reveals the following:

1. From Table. 1 (i) and (vi) it is obvious that for the normally distributed data the average value of three estimates gives almost similar estimate of overall variation (i.e. the true process variability in quality terminology). The reason of this result is that as all these estimates are unbiased estimates of the true process variability.
2. From Table. 1 (ii) - (v) and (vii) - (x) it is observed that when data deviates from normality even then the estimates obtained from G chart is best estimate of the true process variability because it is least affected among the three estimates under study.

Now analysis of means procedure is used to explain graphically the proposed schemes of the decision limits for ANOM charts based on the estimates provided in above Table 1 for the data sets provided in Appendix following the idea of Muhammad et .al (1993). The ANOM charts for estimates under study are presented here in fig 1-3 for normal distribution and in fig 4 -9 for other than normal distribution for sample sizes 3, and in fig 10 - 12 for normal distributions and in fig 13 - 18 for other than normal distributions for sample sizes 10.

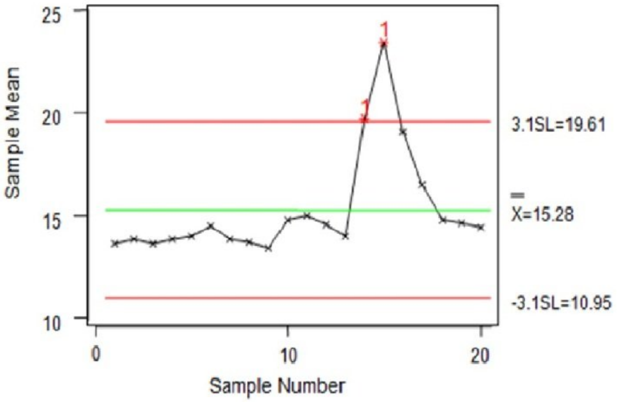

Fig. 1. ANOM chart for data set using range for normal process

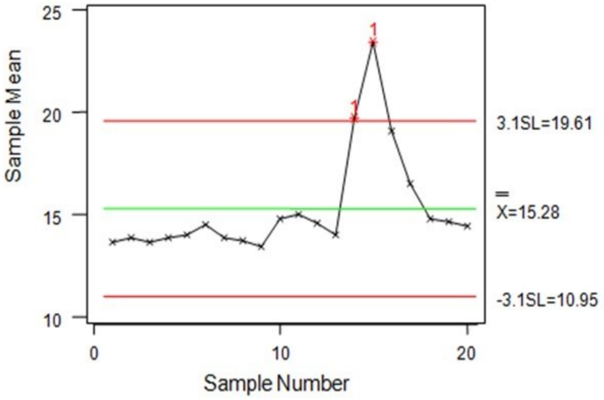

Fig 2. ANOM chart for data set using gini's for normal process

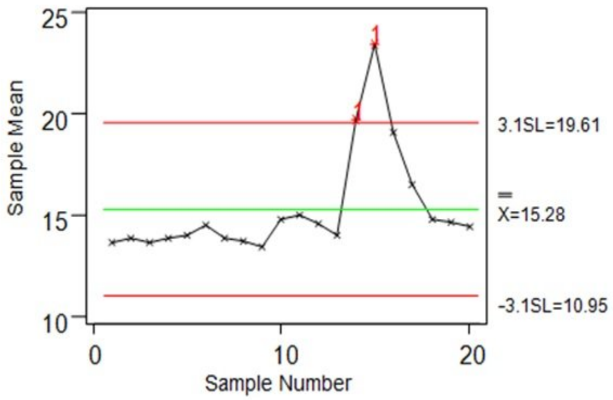

Fig 3. ANOM chart for data set using St.dv. for normal process

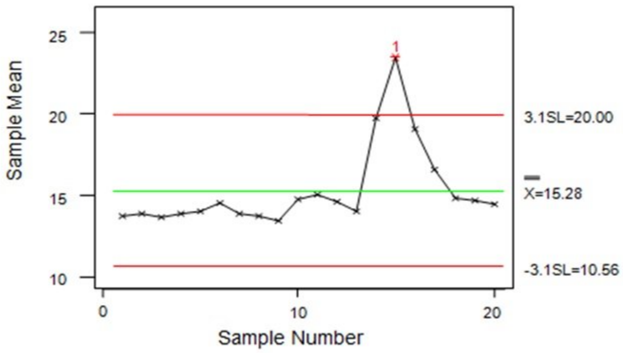

Fig. 4. ANOM chart for data set using range for t 10 d.f process

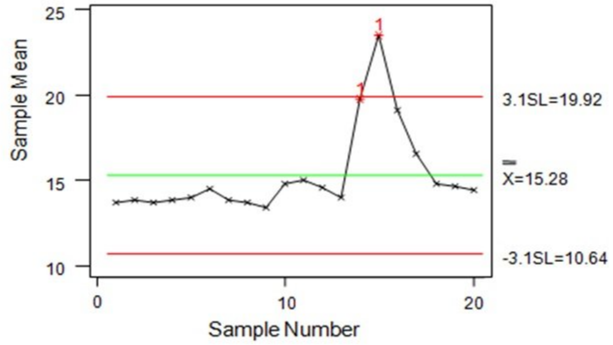

Fig. 5. ANOM chart for data set using Gini's for t 10 d.f process

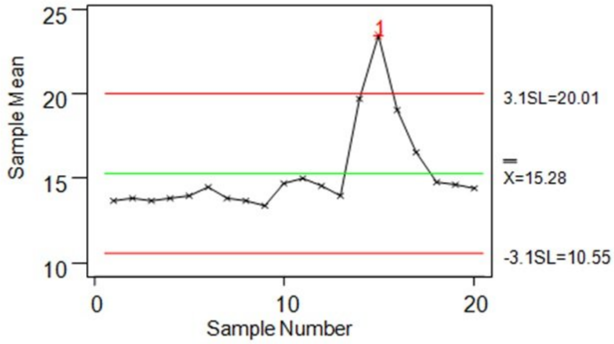

Fig. 6. ANOM chart for data set using St.dv. for t 10 d.f. process

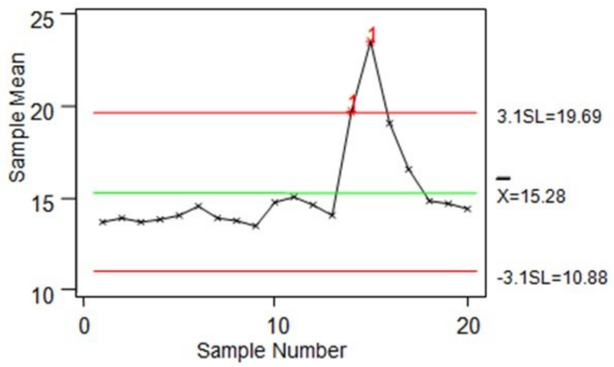

Fig. 7. ANOM chart for data set using range for exponential 1 process.

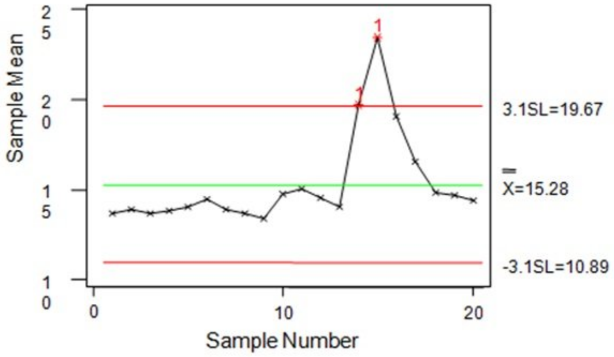

Fig. 8. ANOM chart for data set using gini's for exponential 1 process.

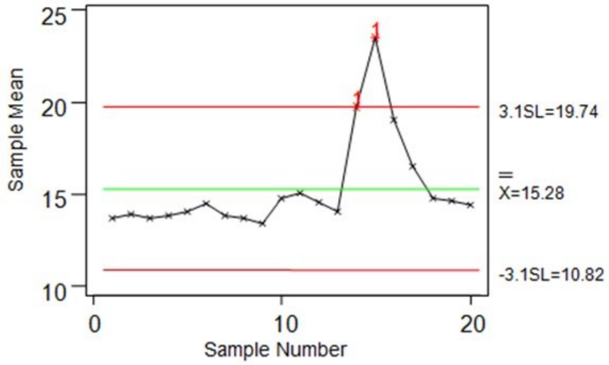

Fig. 9. ANOM chart for data set using St.dv for exponential 1 process

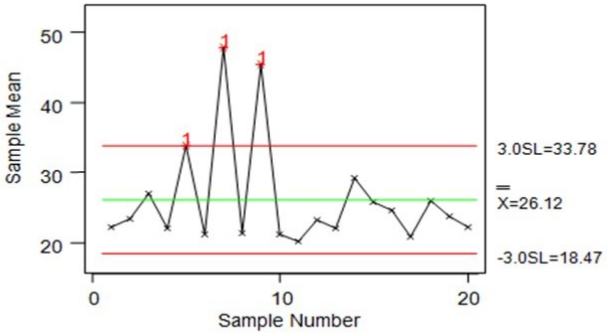

Fig. 10. ANOM chart for data set using range for normal process

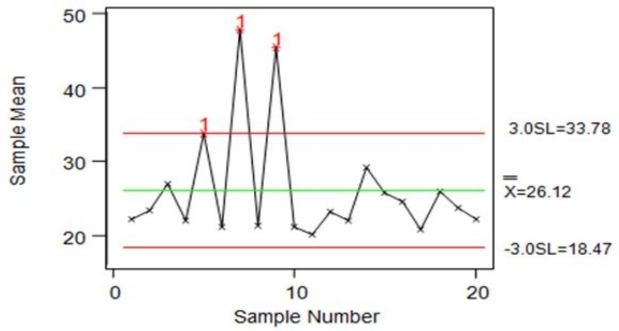

Fig. 11. ANOM chart for data set using Gini's for normal process

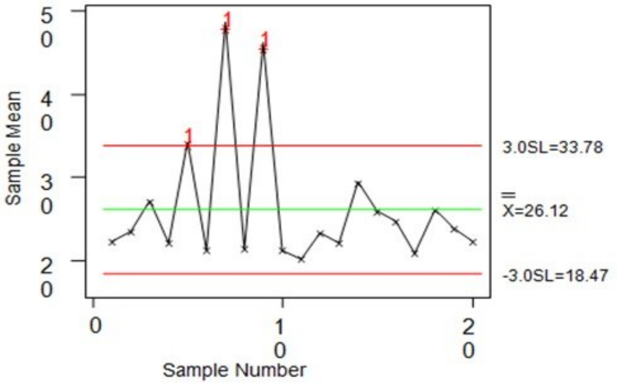

Fig.12. ANOM chart for data set using St..dv. for normal process

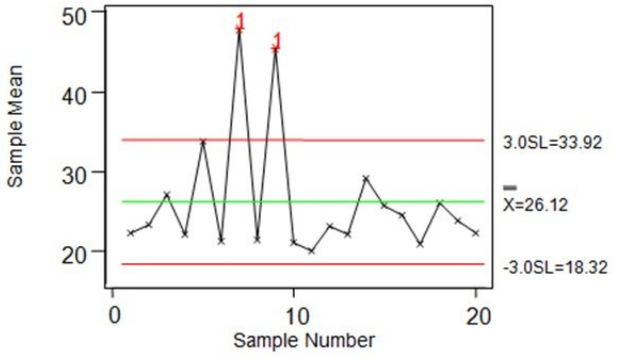

Fig 13 ANOM chart for data set using range for t with 10 d.f normal process

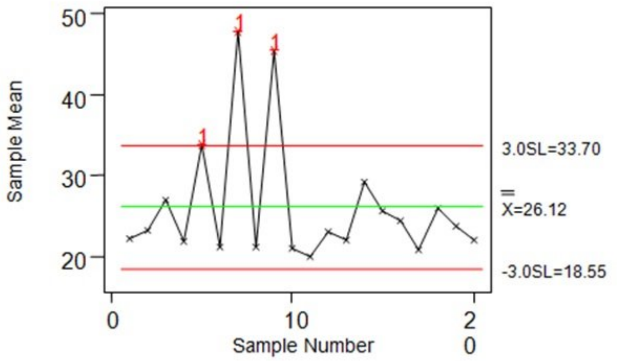

Fig 14 ANOM chart for data set using gini's for t with 10 d.f normal process

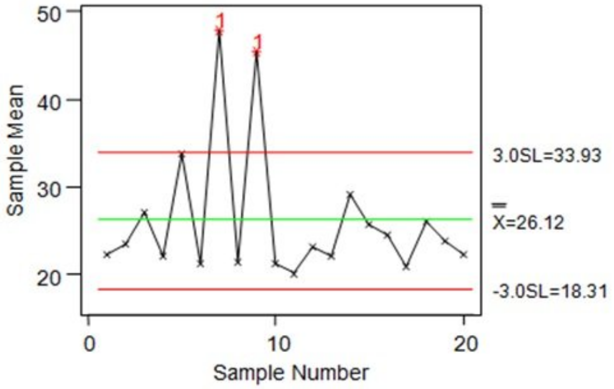

Fig.15. ANOM chart for data set using St.dv for t with 10 d.f Process

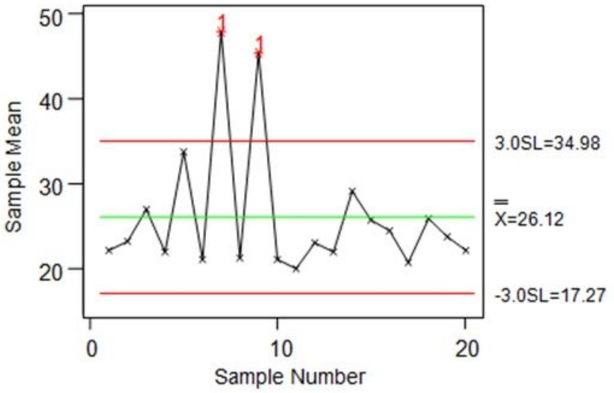

Fig.16. ANOM chart for data set using range for exponential 1 Process

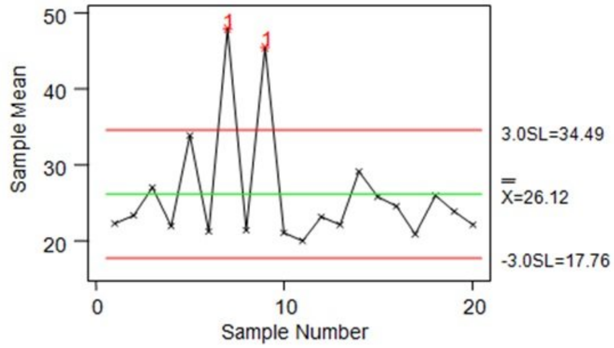

Fig.17. ANOM chart for data set using Gini's for exponential 1 process

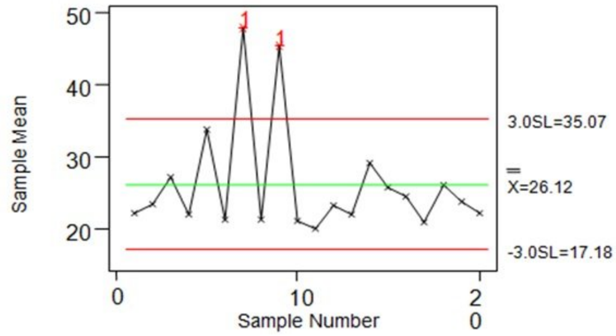

Fig 18 ANOM chart for data set using St.dv for exponential 1 process

In the above figures the symbol ‘1’ represent the point which is out of control or not consistent with the overall mean. The decision limits of the above graphs are based on:

$\bar{\bar{X}} = 15.28$ for data set 1, m =36, n =3, k=20, $h(0.05, m, k) = 3.148$ for the figures 1-9 and $\bar{\bar{X}} = 26.12$ for data set 2, m =162, n =10, k=20, $h(0.05, m, k) = 3.016$ for the figures 10-18 and $\hat{\sigma}$ is obtained from the above Table 1.

We observed that for normally distributed data (for data set.1 of sample sizes 3) all the estimates produced same decision limits and as a result the state of control (whether in control or out of control) remains same in all the three cases as is obvious from figures 1 - 3. For non-normally distribute data the decision limits based on for the estimates obtained from G chart is least effected as compared to the estimates obtained R and S charts as is clear from 5-9. The same is observed for the data set.2 of sample sizes 10 as its obvious from figures 10-18. Consequently we claim that the discriminatory power of the ANOM chart is least affected by departure from normality when the decision limits are constructed using the estimates obtained from G chart. The claim is apparently supported in three figures 4-6 where point 14 is actually out of control and the affected decision limits based on the estimates obtained from G chart is showing point 14 out of control (fig 4) while is not the case if the estimates are obtained from R and S charts.

APPENDIX

Data Set 1.

An experiment was conducted, Muhammad et al. (1993), for food consumed (mg) per INSTAR LARVA under different treatments and was measured as:

Treatment	R_1	R_2	R_3
1	14.30	11.82	14.88
2	17.65	7.98	15.97
3	11.82	12.81	16.35
4	13.02	12.49	15.97
5	12.74	13.40	15.89
6	14.21	13.80	15.46
7	14.64	13.05	13.88
8	13.64	12.52	14.92
9	16.97	10.72	12.54
10	15.05	13.71	15.52
11	17.00	12.75	15.34
12	15.35	14.38	13.97
13	12.71	15.35	13.97
14	19.03	22.68	17.54
15	25.86	24.04	20.66
16	17.39	19.11	20.76
17	16.62	13.72	19.30
18	13.84	14.92	15.64
19	14.51	13.65	15.80
20	15.98	13.83	13.42

Data Set 2.

A hypothetical data set generated from normal process (row wise) with mean **25** and st.dev **8** (with row **5, 7** and **9** disturbed by adding a constant amount **7, 15** and **10** respectively).

14.9926	17.8493	26.4133	15.0199	32.5929	23.8213	18.9816	33.0385	22.3844	16.8063
21.6634	27.1479	15.1405	22.1316	19.9353	30.4281	16.7610	14.7840	39.5966	25.8214
21.4046	24.0037	42.2537	21.1991	38.7195	26.5414	30.8343	11.8979	26.4504	27.2104
16.1441	34.7498	22.3463	13.1158	19.8447	25.4109	26.2692	13.1091	32.6411	16.3399
35.9986	43.9226	13.7717	39.8997	36.8978	31.7777	37.8997	37.9920	32.9964	26.8796
33.2620	22.4642	26.3375	16.0709	10.2923	17.2225	14.5295	16.8935	20.3843	34.2682
48.4405	53.3277	35.4845	50.1250	55.5317	56.2211	44.6285	49.1427	42.4264	42.3570
17.8394	20.0611	13.3025	12.2820	34.2306	21.3968	17.5088	34.3816	20.9972	21.1166
43.1083	46.2056	49.2902	57.2600	49.7069	43.0985	38.6308	43.5406	45.5045	37.1890
27.6154	31.9127	23.4333	22.5963	7.4311	13.6870	25.6735	19.8647	25.9113	12.9879
2.0732	26.7588	23.8237	29.9192	15.0837	13.8754	12.9054	28.4594	22.6949	25.1610
31.5406	18.0311	32.4378	16.9898	21.7639	31.0040	20.0494	25.8890	20.2270	13.3578
10.5338	10.5150	27.5925	40.2056	26.0264	28.8224	24.4927	16.1548	19.4376	16.4114
33.3330	27.1122	35.7043	21.7985	30.5135	17.3169	29.8923	31.8018	33.7078	30.4746
23.3671	30.6085	21.2627	13.6770	30.9617	23.6628	35.0568	24.1758	20.5567	33.9112
12.1388	30.6350	23.3612	28.4510	28.6651	29.3254	18.4560	14.4898	39.8276	19.9882
15.0065	34.0063	13.4529	26.7821	22.1141	12.2504	19.6651	24.0522	19.9415	20.9734
34.6125	32.5291	29.8270	18.4894	29.6426	22.7328	27.4935	21.8881	30.5271	12.1292
33.2991	23.6034	19.5604	24.2852	35.5245	7.5181	24.4244	21.9147	16.4995	31.0257
38.9386	16.1893	12.5455	10.2326	35.5856	6.8294	31.0576	21.7895	33.4454	15.0012

Conclusion: The estimate for within variation (i.e. the true process variability in quality terminology) obtained from G-chart is the best estimate of the true process variability because it is constant free (i.e. it demands no constant like d_2 and c_4 used for R and S charts for an unbiased estimation of true process variability) and least affected among the three estimates under study. The decision limits for ANOM chart using the estimate obtained from G chart are least affected by departure from normality. Also for the larger sample sizes the affect of non-normality is more for the decision limits of ANOM chart when the estimate of σ is obtained from R and S charts as compared to G-Chart. Consequently we claim that the discriminatory power of the ANOM chart is least affected by departure from normality when the decision limits are constructed using the estimates obtained from G chart.

REFERENCES

David, H. A. Gini's Mean difference rediscovered. Biometrika 55: 573-575, (1968).

Gini, C. Variabilita e mutabilita, contributo allo studio delle distribuzioni e delle relazoine statistiche. Studi Fconomico- Giuredice dell' Universita di Cagliari 3 (part 2) i-iii, 3-159, (1912).

Halperin, M., S. W. Greenhouse, J. Cornfield and J. Zalokar. Tables of percentage points for the studentized maximum absolute deviate in normal samples. Journal of American Statistical Association 50: 185-195, (1995).

Jordan W. Ueber die Bestimmung der Genauigkeit mehrfach wiederholter Beobachtungen einer Unbekannten. Astronom, Nachr. 74: 209-226, (1869).

Muhammad, F., S. Ahmad and M. Abiodullah. Use of Probability Weighted Moments in the Analysis of Means. Biometrical Journal., 35(3): 371-378, (1993).

Nelson, P. R. Multivariate normal and t- distribution with communication in statistics, part B. Simulation and Computation. 11, 239-248, (1982).

Nelson, P. R. Additional uses of the Analysis of Means and Extended Tables of Critical Values. Technometrics., 35(1): 61-71, (1983).

Ott, E. R. Analysis of Means: A graphical procedure. Industrial Quality Control., 24: 101-109, (1967).

Riaz, M and A. Saghir, Monitoring Process Variability Using Gini's Mean Difference. Preprint, (2005).

Sheesley, J. H.Simplified factors for analysis of means when the standard deviation is estimated from the range. Journal of Quality Technology., 13: 184-185, (1981).

Von, A. C. G. Ueber die Bestimmung des wahrscheinlichen Fehlers durch die gegebenen Differenzen von m gleich genauen Beobachtungen einer Unbekannten. Astronomy. Nachr. 79, 257-272, (1872).