Phase I and Phase II - Control Charts for the Variance and Generalized Variance

R. van Zyl[1], A.J. van der Merwe[2]

[1]Quintiles International, ruaanvz@gmail.com

[2]University of the Free State

Abstract

By extending the results of Human, Chakraborti, and Smit (2010), Phase I control charts are derived for the generalized variance when the mean vector and covariance matrix of multivariate normally distributed data are unknown and estimated from m independent samples, each of size n.

In Phase II predictive distributions based on a Bayesian approach are used to construct Shewart-type control limits for the variance and generalized variance.

The posterior distribution is obtained by combining the likelihood (the observed data in Phase I) and the uncertainty of the unknown parameters via the prior distribution. By using the posterior distribution the unconditional predictive density functions are derived.

Keywords: Shewart-type Control Charts, Variance, Generalized Variance, Phase I, Phase II, Predictive Density

1 Introduction

Quality control is a process which is used to maintain the standards of products produced or services delivered. It is nowadays commonly accepted by most statisticians that statistical processes should be implemented in two phases:

1. Phase I where the primary interest is to assess process stability; and
2. Phase II where online monitoring of the process is done.

Bayarri and Garcia-Donato (2005) gave the following reasons for recommending Bayesian analysis for the determining of control chart limits:

- Control charts are based on future observations and Bayesian methods are very natural for prediction.
- Uncertainty in the estimation of the unknown parameters are adequately handled.
- Implementation with complicated models and in a sequential scenario poses no methodological difficulty, the numerical difficulties are easy handled via Monte Carlo methods.

- Objective Bayesian analysis is possible without introduction of external information other than the model, but any kind of prior information can be incorporated into the analysis if desired.

In this article, control chart limits will be determined for the sample variance, S^2, and the generalized variance $|S|$. Average run-lengths and false alarm rates will also be calculated in the Phase II setting, using a Bayesian predictive distribution.

2 An Example

The data presented in Table 1 represents measurements of inside diameters and represent the number of 0.0001 inches above 0.7500 inches as given in Duncan (1965). The measurements are taken in samples of $j = 1, 2, \ldots, n$ each ($n = 5$) over time. Also shown in Table 1 are the sample variances, S_i^2 for $i = 1, 2, \ldots m$ samples ($m = 10$). These data will be used to construct a Shewart type Phase I upper control chart for the variance, and also to calculate the run-length for future samples of size $n = 5$ taken repeatedly for the process.

From the data in Table 1 the sample variances are calculated by

$$S_i^2 = \frac{1}{n-1} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2 .$$

The pooled sample variance is then determined as

$$S_p^2 = \frac{1}{m} \sum_{i=1}^{m} S_i^2 = 10.72.$$

3 Statistical Calculation of the Upper Control Limit in Phase I

The upper control limit, using the data by Duncan (1965) will be obtained as described by Human, Chakraborti, and Smit (2010).

It is well known that

$$\frac{(n-1)\, S_i^2}{\sigma^2} \sim \chi^2_{n-1}$$

Also, if the underlying distribution is Normal,

$$\frac{m\left(n-1\right)S_p^2}{\sigma^2} \sim \chi^2_{m(n-1)} = \sum_{i=1}^{m} \chi^2_{n-1}.$$

$$\therefore Y_i = \frac{\left(n-1\right)S_i^2/\sigma^2}{m\left(n-1\right)S_p^2/\sigma^2} = \frac{X_i}{\sum_{i=1}^{m} X_i}$$

where $X_i \sim \chi^2_{n-1}$ $(i = 1, 2, \ldots, m)$.

The distribution of $Y_{max} = max\left(Y_1, Y_2, \ldots, Y_m\right)$ obtained from 100000 simulations is illustrated in Figure 1. The value b is then calculated such that the False Alarm Probability (FAP) is at a level of 0.05 (also shown in the figure).

The upper control limit is then determined as:

$$UCL = mbS_p^2 = 10\left(0.3314\right)\left(10.72\right) = 35.526.$$

The data from Duncan (1965) are presented visually in Figure 2. The figure includes the upper control limit as determined above.

4 Upper Control Limit for the Variance in Phase II

In this section, the upper control limit in a Phase II setting will be derived using the Bayesian derived predictive distribution.

Theorem 1. *Assume* $Y_{ij} \sim^{iid} N\left(\mu_i, \sigma^2\right)$ *where* Y_{ij} *denotes the* j^{th} *observation from the* i^{th} *sample where* $i = 1, 2, \ldots, m$ *and* $j = 1, 2, \ldots, n$. *The mean* μ_i *and variance* σ^2 *are unknown.*

Using the Jeffrey's prior $p\left(\mu_1, \mu_2, \ldots, \mu_m, \sigma^2\right) \propto \sigma^{-2}, \sigma^2 > 0, -\infty < \mu_i < \infty, i = 1, 2, \ldots, m$ *it can be proven that the posterior distribution is given by*

$$p\left(\sigma^2 | data\right) = \left(\frac{\tilde{S}}{2}\right)^{\frac{1}{2}k} \frac{1}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}(k+2)} \exp\left(-\frac{\tilde{S}}{2\sigma^2}\right), \sigma^2 > 0 \qquad (1)$$

an Inverse Gamma distribution with $k = m\left(n-1\right)$ *and* $\tilde{S} = m\left(n-1\right)S_p^2$.

Proof. The proof is provided as part of the appendices. □

The posterior distribution given in Equation 1 is presented in Figure 3.

A predictive distribution derived using a Bayesian approach will be used to obtain the control limits in a Phase II setting. Let S_f^2 be the sample variance of a future sample of n observations from the Normal distribution. Then for a given σ^2 it follows that

$$\frac{(n-1)\,S_f^2}{\sigma^2} = \frac{vS_f^2}{\sigma^2} \sim \chi_v^2$$

which means that

$$f\left(S_f^2|\sigma^2\right) = \left(\frac{v}{2\sigma^2}\right)^{\frac{1}{2}v} \frac{1}{\Gamma\left(\frac{v}{2}\right)} \left(S_f^2\right)^{\frac{1}{2}v-1} \exp\left(-\frac{vS_f^2}{2\sigma^2}\right) \tag{2}$$

Theorem 2. *If S_f^2 is the sample variance of a future sample of n observations from the Normal distribution then the unconditional predictive density of S_f^2 is given by*

$$f\left(S_f^2|data\right) = S_p^2 F_{n-1,m(n-1)} \tag{3}$$

where S_p^2 is the pooled sample variance and $F_{n-1,m(n-1)}$ the F-distribution with degrees of freedom $n-1$ and $m\,(n-1)$.

Proof. The proof is given as part of the appendices. □

The upper control limit in the Phase II setting is then derived as

$$S_p^2 F_{n-1,m(n-1)}\,(\alpha)\,.$$

At $\alpha = 0.0027$ we therefore obtain the upper control limit as

$$S_p^2 F_{n-1,m(n-1)}\,(0.0027) = 10.72 \times 4.8707 = 52.214.$$

The distribution of the predictive density of S_f^2 including the derived upper control limit is presented in Figure 4.

Using the predictive distribution for S_f^2 in Equation 3 the control chart limits are therefore determined as $(52.214, \infty)$.

Assuming that the process remains stable, the predictive distribution for S_f^2 can also be used to derive the distribution of the run-length, that is the number of samples until the control chart signals for the first time.

The resulting rejection region of size α using the predictive distribution for the determination of the run-length is defined as

$$\alpha = \int_{R(\alpha)} f\left(S_f^2|data\right) dS_f^2$$

where

$$R\left(\alpha\right) = (52.214, \infty).$$

Given σ^2 and a stable process, the distribution of the run-length r is Geometric with parameter

$$\psi\left(\sigma^2\right) = \int_{R(\alpha)} f\left(S_f^2|\sigma^2\right) dS_f^2 \tag{4}$$

where $f\left(S_f^2|\sigma^2\right)$ is given in Equation 2.

For a given, unknown $\sigma^2$and a stable process, the uncertainty is described by the posterior distribution defined in Equation 1, denoted by $p\left(\sigma^2|data\right)$.

Theorem 3. *For a given σ^2 the Geometric parameter*

$$\psi\left(\sigma^2\right) = \psi\left(\chi^2_{m(n-1)}\right) \text{ for given } \chi^2_{m(n-1)}$$

which means that it is only dependent on $\chi^2_{m(n-1)}$ and not on σ^2.

Proof. The proof is provided as part of the appendices. □

In Figure 5 the distributions of $f\left(S_f^2|data\right)$ and $f\left(S_f^2|\sigma^2\right)$ for $\sigma^2 = 9$ and $\sigma^2 = 20$ are presented to show the different shapes of the applicable distributions.

In Table 2 results for the run-length at $\alpha = 0.0027$ for $n = 5$ and different values for m are presented. The table present the mean, median, 95% equal tail interval and calculated α value to obtain a run-length of 370 (the expected run length at $\alpha = 0.0027$ is $\frac{1}{0.0027} \approx 370$ if σ^2 is known).

In the case of the diameter example the mean run-length is 29754 and the median run-length 1354. The reason for these large values is the small sample size and number of samples ($n = 5$ and $m = 10$). To get a mean run-length 370 α must be 0.0173 instead of 0.027.

From Table 2 it can be noted that as the number of samples increase (larger m) the mean and median run-lengths converges to the expected run-length of 370.

5 Phase I Control Charts for the Generalized Variance

Assume $\underline{Y}_{ij} \sim^{idd} N(\underline{\mu}_i, \Sigma)$ where $\underline{Y}_{ij}$ $(p \times 1)$ denotes the jth observation vector from the ith sample, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. The mean vector $\underline{\mu}_i$ $(p \times 1)$ and covariance matrix, Σ $(p \times p)$ are unknown.

Define $\underline{\bar{Y}}_i = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$ and $A_i = \sum_{j=1}^{n} \left(\underline{Y}_{ij} - \underline{\bar{Y}}_i\right) \left(\underline{Y}_{ij} - \underline{\bar{Y}}_i\right)'$ $(i = 1, 2, \ldots, m)$.

From this it follows that

$$\underline{\bar{Y}}_i \sim N\left(\underline{\mu}_i, \frac{1}{n}\Sigma\right), (i = 1, 2, \ldots, m),$$

$$A_i = (n-1) S_i \sim W_p(n-1, \Sigma),$$

$$A = \sum_{i=1}^{m} A_i \sim W_p(m(n-1), \Sigma)$$

and

$$S_p = \frac{1}{m(n-1)} A.$$

The generalized variance of the ith sample is defined as the determinant of the sample covariance matrix, i.e., $|S_i|$.

Define

$$T_i = \frac{|A_i|}{|\sum_{i=1}^{m} A_i|} = \frac{|A_i^*|}{|\sum_{i=1}^{m} A_i^*|}$$

where $A_i^* \sim W_p\left(n-1, I_p\right)$.

Also

$$T = max\left(T_1, T_2, \ldots, T_m\right) = max\left(T_i\right), i = 1, 2, \ldots, m$$

Now

$$T_i = \frac{|A_i|}{|\sum_{i=1}^{m} A_i|} = \frac{|S_i|}{m^p\,|S_p|}$$

Therefore a $(1-\alpha)\,100\%$ upper control limit for $|S_i|$ $(i = 1, 2, \ldots, m)$ is $m^p\,|S_p|\,T_{1-\alpha}$

Figure 6 presents a histogram of 100,000 simulated values of $max\left(T_i\right)$ for the two dimensional case (p=2,m=10 and n=6). The upper control limit as presented in Table 3 is presented on the figure. Table 3 also presents the upper control limit for the one dimensional (p=1) and the three dimensional (p=3) situations.

By using a Bayesian procedure a predictive distribution will be derived to obtain control chart limits in Phase II.

Using the Jeffrey's prior

$$p\left(\underline{\mu}, \Sigma\right) \propto |\Sigma|^{-\frac{1}{2}(p+1)} \quad -\infty < \underline{\mu} < \infty, \Sigma > 0$$

the posterior distribution of Σ is derived as

$$|\Sigma|\,|data \sim |A| \prod_{i=1}^{p} \left(\frac{1}{\chi^2_{m(n-1)+1-i}}\right) \tag{5}$$

and the predictive distribution of a future sample generalized variance $|S_f|$ given Σ as

$$|S_f|\,|\Sigma \sim \left|\frac{1}{n-1}\Sigma\right| \prod_{i=1}^{p} \chi^2_{n-i} \tag{6}$$

By combining Equation 5 and Equation 6 the unconditional predictive distribution is given by

$$\left|S_f^2\right|\,|data \sim \left(\frac{1}{n-1}\right)^p |A| \left(\prod_{i=1}^{p} \frac{n-i}{m\,(n+1)+1-i}\right) F^* \tag{7}$$

where

$$F^* = \prod_{i=1}^{p} F_{n-i,m(n-1)+i-i}.$$

Equation 7 can be used to obtain the control chart limits.

Similarly for the variance, the rejection region of size α is defined as

$$\alpha = \int_{R(\alpha)} f\left(|S_f|\,|data\right) d\,|S_f|\,.$$

Given Σ and a stable process, the distribution of the run-length r is Geometric with parameter

$$\psi\left(|\Sigma|\right) = \int_{R(\alpha)} f\left(|S_f|\,|\Sigma\right)$$

where $f\left(\left|S_{_f}\right|\,|\Sigma\right)$ is given in Equation 6.

Theorem 4. *For an unknown value of* $|\Sigma|$ *and uncertainty described by Equation 5 the Geometric parameter can be shown to be*

$$\psi\left(|\Sigma|\right) = P\left\{\prod_{i=1}^{p} \chi^2_{n-i} \geq \left(\prod_{i=1}^{p} \chi^2_{m(n-1)+1-i}\right)\left(\prod_{i=1}^{p} \frac{n-i}{m\,(n-1)+1-i}\right) F^*_\alpha\right\}$$

for a given $\prod_{i=1}^{p} \chi^2_{m(n-1)+1-i}$.

Proof. The proof is provided as part of the appendices. □

For further details see Menzefricke (2002, 2007, 2010a,b).

Mean and median run-length results at $\alpha = 0.0027$ for $n = 50$, $m = 50$ and 100 for the one, two and three dimensional cases are presented in Table 4.

6 Conclusion

Phase I and Phase II control chart limits have been constructed using Bayesian methodology. In this article we have seen that due to Monte Carlo simulation the construction of control chart limits using the Bayesian paradigm are handled with ease. Bayesian methods allow the use of any prior to construct control limits without any difficulty. It has been shown that the uncertainty in unknown parameters are handled with ease in using the predictive distribution in the determination of control chart limits. It has also been shown that an increase in number of samples m and the sample size n leads to a convergence in the run-length towards the expected value of 370 at $\alpha = 0.0027$.

References

Bayarri, M., Garcia-Donato, G., 2005. A bayesian sequential look at u-control charts. Technometrics 47 (2), 142–151.

Duncan, A., 1965. Quality Control and Industrial Statistics. Vol. 3rd. Richard D. Irwin, Inc.

Human, S., Chakraborti, S., Smit, C., 2010. Shewart-type control charts for variation in phase i data analysis. Computational Statistics and Data Analysis 54, 863–874.

Menzefricke, U., 2002. On the evaluation of control chart limits based on predictive distributions. Communications in Statistics - Theory and Methods 31(8), 1423–1440.

Menzefricke, U., 2007. Control chart for the generalized variance based on its predictive distribution. Communications in Statistics - Theory and Methods 36(5), 1031–1038.

Menzefricke, U., 2010a. Control chart for the variance and the coefficient of variation based on their predictive distribution. Communications in Statistics - Theory and Methods 39(16), 2930–2941.

Menzefricke, U., 2010b. Multivariate exponentially weighted moving average chart for a mean based on its predictive distribution. Communications in Statistics - Theory and Methods 39(16), 2942–2960.

Appendices

A Proofs

A.1 Proof of Theorem 1

The likelihood function, i.e., the distribution of the data is

$$L\left(\mu_1,\mu_2,\ldots,\mu_m,\sigma^2|data\right)=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}mn}\prod_{i=1}^{m}\prod_{j=1}^{n}\exp\left\{-\frac{1}{2}\left(y_{ij}-\mu_i\right)^2/\sigma^2\right\}.$$

Deriving the posterior distribution as Poster $\propto$Likelihood $\times$Prior, and using the Jeffrey's prior it follows that

$$\mu_i|\sigma^2,data\sim N\left(\bar{y}_i,\frac{\sigma^2}{n}\right),i=1,2,\ldots,m$$

and

$$p\left(\sigma^2|data\right)=\left(\frac{\tilde{S}}{2}\right)^{\frac{1}{2}k}\frac{1}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}(k+2)}\exp\left(-\frac{\tilde{S}}{2\sigma^2}\right),\sigma^2>0$$

an Inverse Gamma distribution with $k=m\left(n-1\right)$ and $\tilde{S}=m\left(n-1\right)S_p^2$.

A.2 Proof of Theorem 2

For a given σ^2 it follows that

$$\frac{\left(n-1\right)S_f^2}{\sigma^2}=\frac{vS_f^2}{\sigma^2}\sim\chi_v^2,$$

which means that

$$f\left(S_f^2|\sigma^2\right) = \left(\frac{v}{2\sigma^2}\right)^{\frac{1}{2}v} \frac{1}{\Gamma\left(\frac{v}{2}\right)} \left(S_f^2\right)^{\frac{1}{2}v-1} \exp\left(-\frac{vS_f^2}{2\sigma^2}\right)$$

where $v = n-1$ and $S_f^2 > 0$.

The unconditional predictive density of S_f^2 is given by

$$\begin{aligned} f\left(S_f^2|data\right) &= \int_0^\infty f\left(S_f^2|\sigma^2\right) p\left(\sigma^2|data\right) d\sigma^2 \\ &= \frac{(v)^{\frac{1}{2}v}\left(\tilde{S}\right)^{\frac{1}{2}k}\left(S_f^2\right)^{\frac{1}{2}v-1}\Gamma\left(\frac{v+k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{v}{2}\right)\left(\tilde{S}+vS_f^2\right)^{\frac{1}{2}(v+k)}} \qquad S_f^2 > 0 \end{aligned}$$

where $v = n-1$, $k = m\left(n-1\right)$ and $\tilde{S} = kS_p^2 = m\left(n-1\right)S_p^2$.

$$\therefore f\left(S_f^2|data\right) = S_p^2 F_{n-1,m(n-1)}$$

A.3 Proof of Theorem 3

For a given σ^2

$$\begin{aligned} \psi\left(\sigma^2\right) &= P\left(S_f^2 > S_p^2 F_{n-1,m(n-1)}\left(\alpha\right)\right) \\ &= P\left(\frac{\sigma^2\chi_{n-1}^2}{n-1} > S_p^2 F_{n-1,m(n-1)}\left(\alpha\right)\right) \qquad \text{for given } \sigma^2 \\ &= P\left(\frac{m(n-1)S_p^2}{\chi_{m(n-1)}^2}\frac{\chi_{n-1}^2}{n-1} > S_p^2 F_{n-1,m(n-1)}\left(\alpha\right)\right) \quad \text{for given } \chi_{m(n-1)}^2 \\ &= P\left(\chi_{n-1}^2 > \tfrac{1}{m}\chi_{m(n-1)}^2 F_{n-1,m(n-1)}\left(\alpha\right)\right) \quad \text{for given } \chi_{m(n-1)}^2 \\ &= \psi\left(\chi_{m(n-1)}^2\right) \qquad \text{for given } \chi_{m(n-1)}^2 \end{aligned}$$

A.4 Proof of Theorem 4

For a given $|\Sigma|$

$$\begin{aligned}
\psi\left(|\Sigma|\right) &= P\left\{|S_f| > \left(\tfrac{1}{n-1}\right)^p |A| \left(\prod_{i=1}^p \tfrac{n-i}{m(n-1)+1-i} F_\alpha^*\right)\right\} \\
&= P\left\{\left|\tfrac{1}{n-1}\Sigma\right| \prod_{i=1}^p \chi^2_{n-i} \geq \left(\tfrac{n}{n-1}\right)^p |A| \left(\prod_{i=1}^p \tfrac{n-i}{m(n-1)+1-i}\right) F_\alpha^*\right\} \\
&= P\left\{|A| \prod_{i=1}^p \left(\tfrac{1}{\chi^2_{m(n-1)+1-i}}\right) \prod_{i=1}^p \chi^2_{n-i} \geq |A| \left(\prod_{i=1}^p \tfrac{n-i}{m(n-1)+1-i}\right) F_\alpha^*\right\} \\
&= P\left\{\prod_{i=1}^p \chi^2_{n-i} \geq \left(\prod_{i=1}^p \chi^2_{m(n-1)+1-i}\right) \left(\prod_{i=1}^p \tfrac{n-i}{m(n-1)+1-i}\right) F_\alpha^*\right\}
\end{aligned}$$

for a given $\prod_{i=1}^p \chi^2_{m(n-1)+1-i}$.

Table 1: Data for Constructing a Shewart-Type Phase I Upper Control Chart for the Variance

Sample Number/ Time (i)	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	s_i^2
1	15	11	8	15	6	16.5
2	14	16	11	14	7	12.3
3	13	6	9	5	10	10.3
4	15	15	9	15	7	15.2
5	11	14	11	12	5	11.3
6	13	12	9	6	10	7.5
7	10	15	12	4	6	19.8
8	9	12	9	8	8	2.7
9	8	12	14	9	10	5.8
10	10	10	9	14	14	5.8

Table 2: Mean and Median Run-length at $\alpha = 0.0027$ for $n = 5$ and Different Values of m

m	n	Mean	Median	95% Equal Tail Interval	Calculated α for Mean Run Length of 370
10	5	29 754	1 354	(54;117 180)	0.0173
50	5	654	470	(121;2 314)	0.0044
100	5	482	411	(156;1 204)	0.0035
200	5	422	391	(197;829)	0.0031
500	5	389	379	(244;596)	0.0028
1 000	5	379	374	(274;517)	0.0028
5 000	5	371	370	(322;428)	0.0027
10 000	5	370	370	(335;410)	0.0027

Table 3: Upper 95% Control Limit, $T_{0.95}$ for $T = max\,(T_i)$ for the Generalized Variance in Phase I for $m = 10$, $n = 6$ and $p = 1, 2$ and 3

p	m	n	$T_{0.95}$
1	10	6	0.3035
2	10	6	0.04429
3	10	6	0.00445

Figure 1: Distribution of $Y_{max} = max\,(Y_1, Y_2, \ldots, Y_m)$ (100,000 simulations)

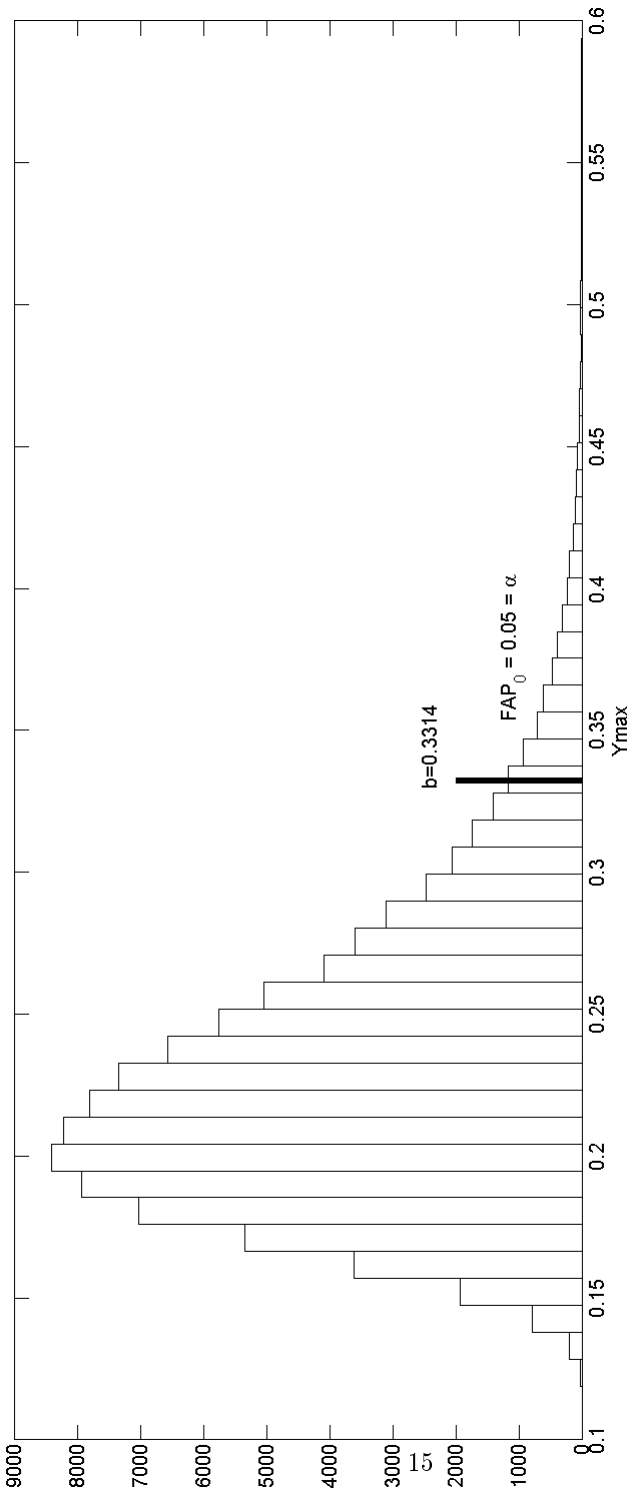

Figure 2: Shewart-type Phase I Upper Control Chart for the Variance - $FAP_0 = 0.05$

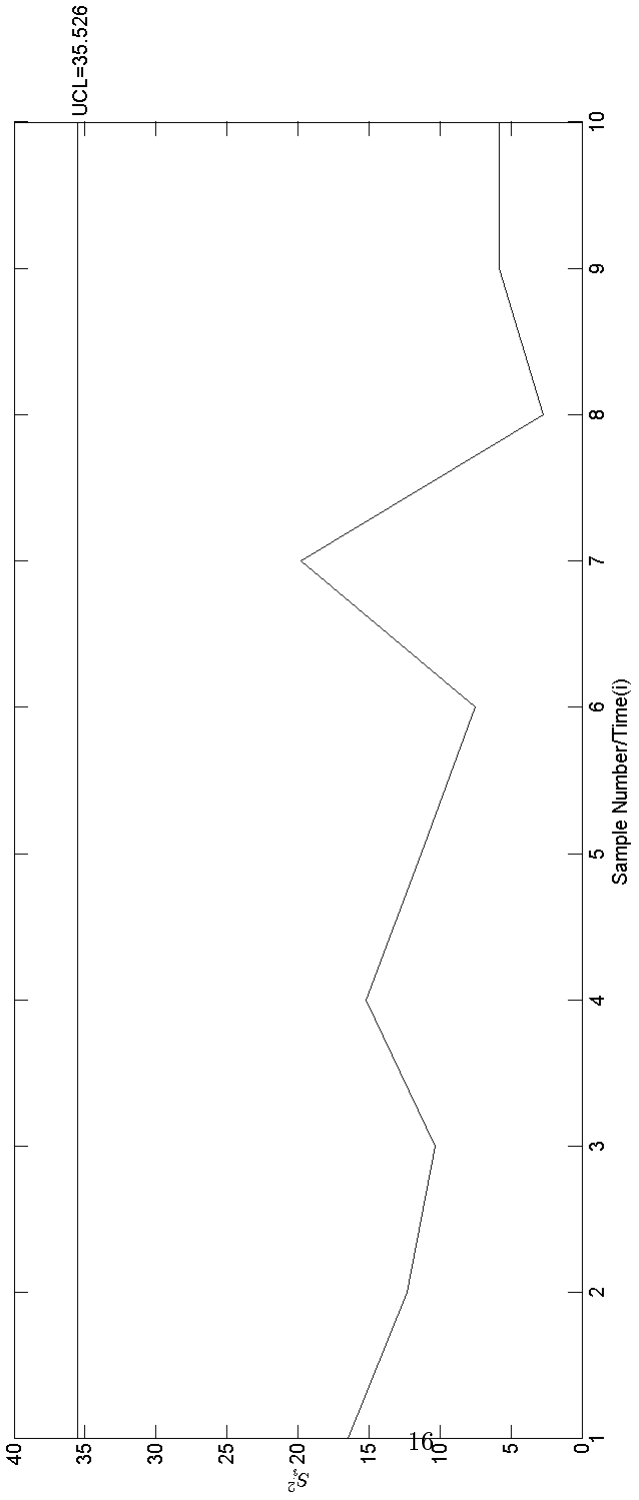

Figure 3: Distribution of $p\left(\sigma^2|data\right)$-Simulated Values

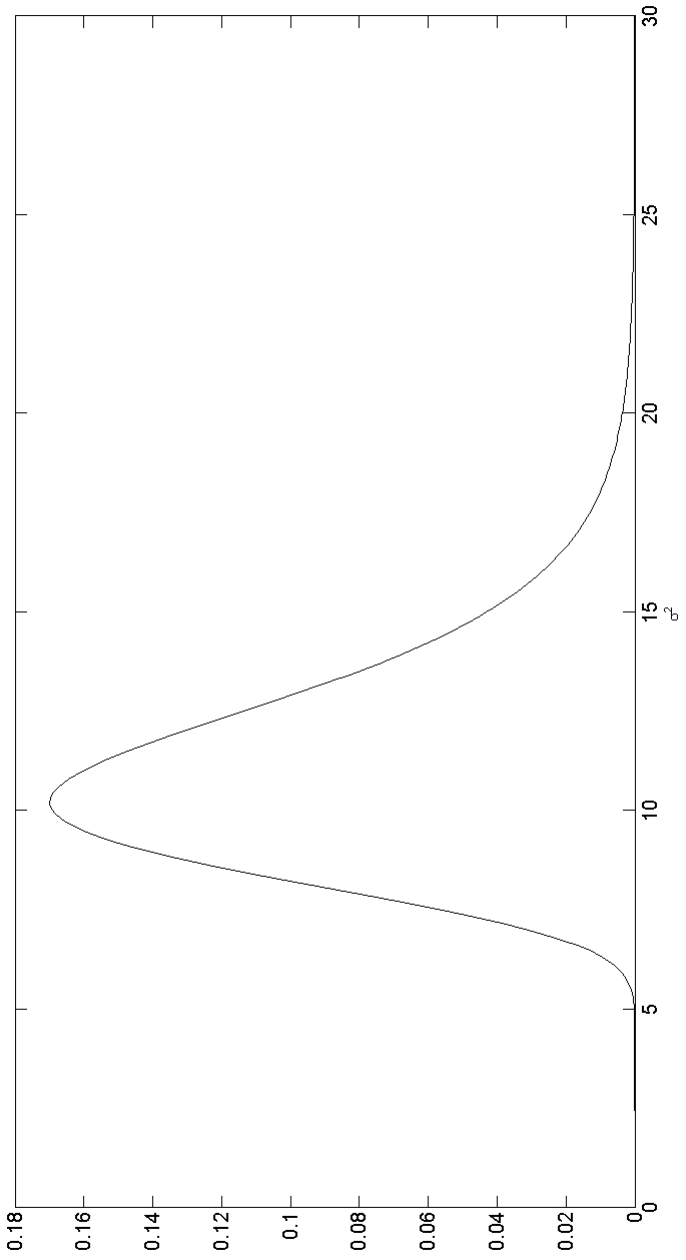

Figure 4: Distribution of $f\left(S_f^2|data\right)$

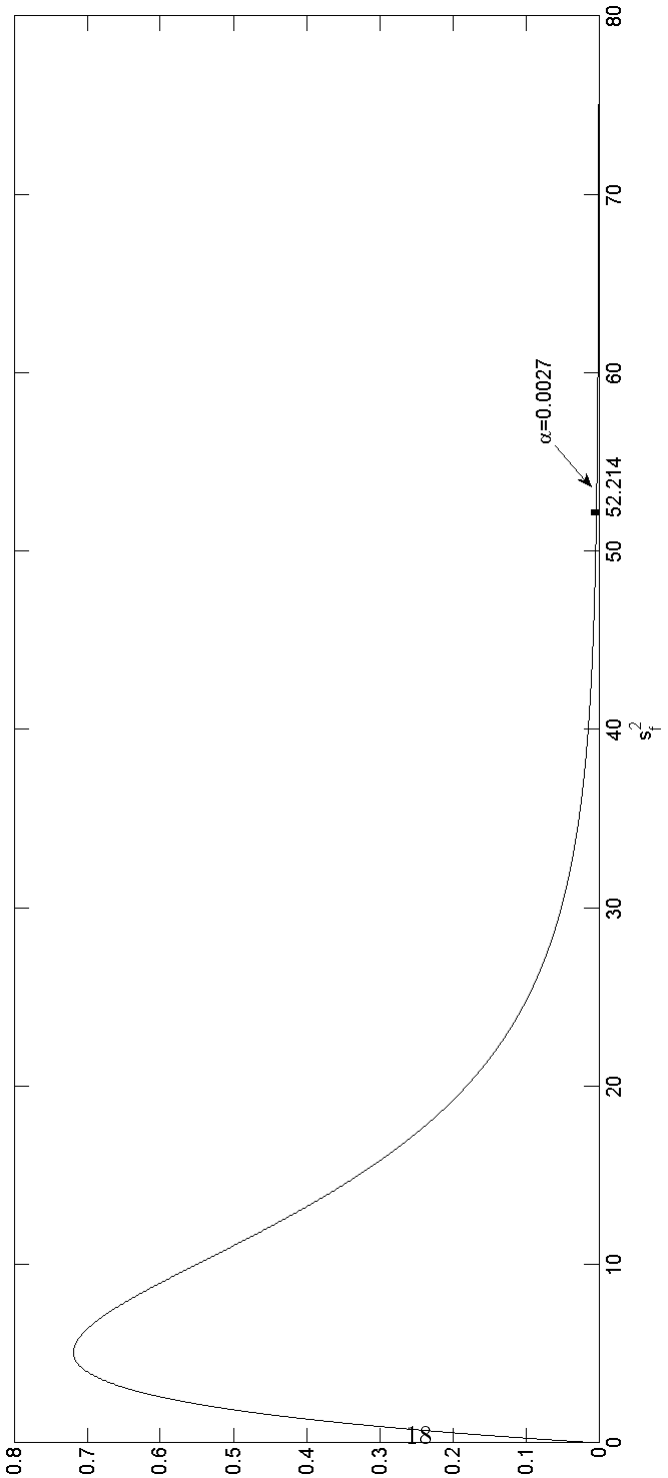

Figure 5: Distributions of $f\left(S_f^2|\sigma^2\right)$ and $f\left(S_f^2|data\right)$ showing $\psi\left(\sigma_1^2\right)$

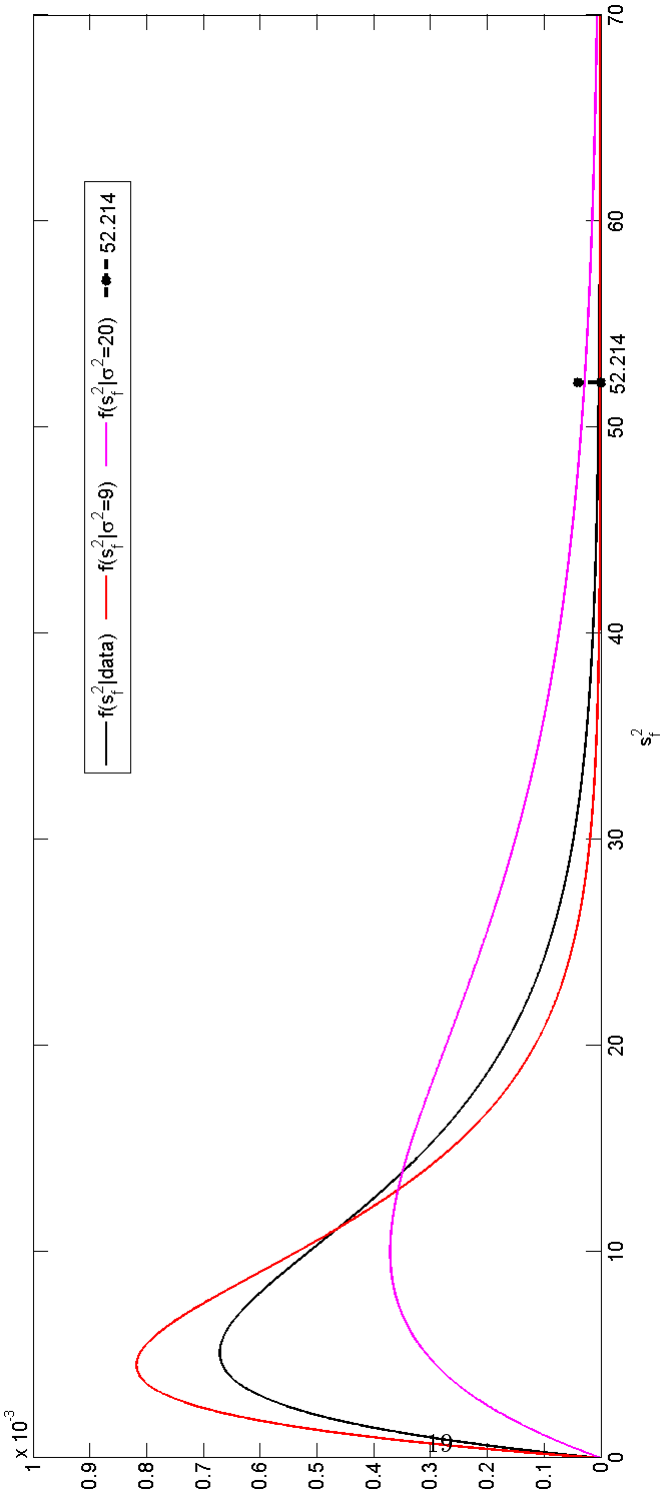

Figure 6: Histogram of $max\,(T_i)$-100,000 Simulations

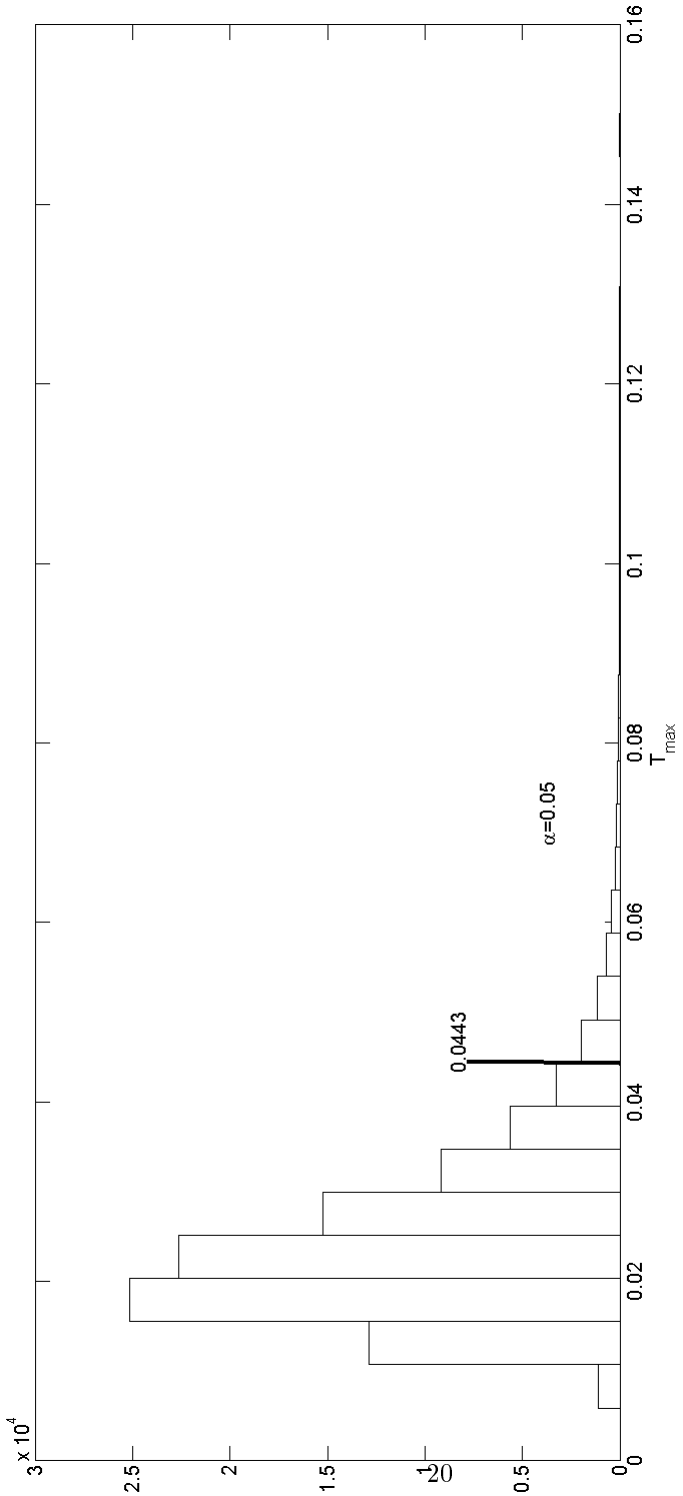

Table 4: Mean and Median Run-length at $\alpha = 0.0027$ for $n = 50$, $m = 50$ and 100 and $p = 1, 2$ and 3

p	m	n	Mean	Median	95% Equal Tail Interval
1	50	50	482	414	(185;1 198)
1	100	50	431	402	(197;841)
2	50	50	466	404	(162;1 128)
2	100	50	423	396	(205;819)
3	50	50	461	407	(165;1 063)
3	100	50	424	399	(209;786)