On the Distribution of Hotelling's T^2 Statistic Based on the Successive Differences Covariance Matrix Estimator

JAMES D. WILLIAMS

GE Global Research, Niskayuna, NY 12309

WILLIAM H. WOODALL and JEFFREY B. BIRCH

Virginia Polytechnic Institute & State University, Blacksburg, VA 24061-0439

JOE H. SULLIVAN

Mississippi State University, Mississippi State, MS 39762

In the historical (or retrospective or Phase I) multivariate data analysis, the choice of the estimator for the variance-covariance matrix is crucial to successfully detecting the presence of special causes of variation. For the case of individual multivariate observations, the choice is compounded by the lack of rational subgroups of observations with the same distribution. Other research has shown that the use of the sample covariance matrix, with all of the individual observations pooled, impairs the detection of a sustained step shift in the mean vector. For example, research has shown that, with the use of the sample covariance matrix, the probability of a signal actually *decreases* below the false alarm probability with a sustained step shift near the middle of the data and that the signal probability decreases with the size of the shift. An alternative estimator, based on the successive differences of the individual observations, leads to an increasing signal probability as the size of the step shift increases and has been recommended for use in Phase I analysis. However, the exact distribution for the resulting T^2 chart statistics has not been determined when the successive differences estimator is used. Three approximate distributions have been proposed in the literature. In this paper we demonstrate several useful properties of the T^2 statistics based on the successive differences estimator and give a more accurate approximate distribution for calculating the upper control limit for individual observations in a Phase I analysis.

Key Words: *Approximate Distribution, False Alarm Rate, Multivariate SPC, Sample Size, T^2 Control Chart.*

Dr. Williams is a Statistical Leader in the Applied Statistics Laboratory. He is a Member of ASQ. His email address is james.williams@research.ge.com.

Dr. Woodall is a Professor in the Department of Statistics. He is a Fellow of ASQ. His email address is bwoodall@vt.edu.

Dr. Birch is a Professor in the Department of Statistics. His email address is jbbirch@vt.edu.

Dr. Sullivan is a Professor of Quantitative Analysis. He is a Member of ASQ. His email address is J.Sullivan@msstate.edu.

Introduction

Multivariate statistical process control (SPC) is useful when several measures of a product or process are taken at each sampling stage to assess quality. A common statistical method to simultaneously monitor multiple quality characteristics is use of the Hotelling T^2 control chart. Mason and Young (2002) and Fuchs and Kenett (1998) give expository reviews of the use of the T^2 chart. Multivariate charts are also useful for monitoring quality profiles as discussed by Woodall, Spitzner, Montgomery, and Gupta (2004).

The main objective of a multivariate control chart is to detect the presence of special causes of variation. In particular, for a retrospective Phase I analysis of a historical data set (HDS) the objective is twofold: (1) to identify shifts in the mean vector which might distort the estimation of the in-control mean vector and variance-covariance matrix, and (2) to identify and eliminate multivariate outliers. We seek an in-control subset of the HDS with which we may estimate in-control parameters for use in a Phase II analysis.

The T^2 control chart is a tool to detect multivariate outliers, mean shifts, and other distributional deviations from the in-control distribution. An important aspect of the T^2 control chart for Phase I is how to determine the sample covariance matrix used in the calculation of the chart statistic. When rational subgroups make sense, the implication is that the appearance of a special cause within a subgroup is unlikely, so that all observations within a subgroup share a common distribution. Thus the regular sample covariance matrix is useful, and taking the average over all the subgroups is the common procedure, unless there are special causes that alter the covariance matrix.

With individual observations, taking the sample covariance matrix of the historical data set leads to poor properties in detecting sustained step shifts in the mean vector, as shown in Sullivan and Woodall (1996). Instead, the use of an estimator robust to shifts in the mean vector gives better performance in detecting sustained shifts in the mean vector with a T^2 control chart. A covariance matrix estimator that is robust to a sustained shift in the mean vector is one based on successive differences. Sullivan and Woodall (1996) and Vargas N. (2003) demonstrated that the T^2 statistic based on the successive differences estimator is effective in detecting sustained step and ramp shifts in the mean vector. Sullivan and Woodall (1996) found that the T^2 statistic based on the usual sample variance-covariance matrix estimator was not only less effective in detecting a sustained shift in the mean vector, but, as the magnitude of the shift increased, the power to detect the shift decreased. Vargas N. (2003) also proposed a robust estimator based on the minimum volume ellipsoid and showed that its use led to good properties in detecting shifts in the mean vector. However, the successive differences estimator does not perform very well in detecting outliers, as pointed out by Sullivan and Woodall (1996) and Vargas N. (2003) who also proposed supplementary techniques for the effective detection of outliers. The problem of detecting only outliers has been studied extensively but is not relevant to our research, since the successive differences estimator is not useful for that purpose. Sullivan (2002) proposed an effective way to detect multiple shifts, multiple outliers, and a combination of both.

Mason, Chou, Sullivan, Stoumbos, and Young (2003) noted that the presence of common cause trends, cycles, or autocorrelation can result in extremely large values of the T^2 statistic. Where trends, cycles, or autocorrelation exist in the HDS in the absence of special causes of variation, then other methods must be employed. In this paper we provide a more accurate false alarm probability for the T^2 chart based on the successive differences covariance matrix estimator when the in-control observations are independent and identically distributed (i.i.d.).

Knowledge of the statistical distribution of the control chart statistics is needed to calculate the upper control limit (UCL) of the control chart and estimate control chart performance. If the exact distribution of a control statistic is unknown or intractable, then the UCL can be calculated from either an approximate distribution or from a Monte Carlo simulation.

Unfortunately, the exact small-sample distribution of the T^2 statistic based on the successive differences variance-covariance matrix estimator is unknown. Two approximate distributions have been proposed, one by Sullivan and Woodall (1996) and the other by Mason and Young (2002). Another possible approximate distribution is the asymptotic distribution as the number of samples increases. We give the asymptotic distribution and give recommendations for its use. We show that the distribution of the T^2 statistic depends on its location in the HDS. We also propose an improved small-sample approximation and demonstrate that in many situations our proposed approximation performs better than the other approaches. We discuss some useful properties of the distribution of the T^2 statistic based on the successive differences variance-covariance matrix estimator and compare the performance of the approximate distributions.

Since the case of individual observations (without rational subgroups) is the most challenging situation for estimating the variance-covariance matrix, that is our focus. With rational subgroups of size $n \geq 2$, forming estimates for each subgroup and pooling these estimates is generally recommended.

The T^2 Statistic

In a Phase I analysis, we begin with an HDS consisting of m independent vectors of dimension p observed over time, where p is the number of quality characteristics that are being measured, and $p < m$. We model the case in which the in-control observation vectors or residuals from a time series model, $\mathbf{x}_i, i = 1, \ldots, m$, are i.i.d. multivariate normal random vectors with common mean vector and covariance matrix, i.e.,

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1}$$

It is useful to define the $m \times p$ HDS $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_m' \end{bmatrix}.$$

The Hotelling's T^2 statistic measures the Mahalanobis distance of the corresponding vector from the sample mean vector. The general form of the statistic is

$$T_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \, \mathbf{S}^{-1} \, (\mathbf{x}_i - \bar{\mathbf{x}}) \, ,$$

where $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$ and $\mathbf{S}$ is some estimator of $\mathbf{\Sigma}$.

A common choice for $\mathbf{S}$ is the sample variance-covariance estimator given by

$$\mathbf{S}_1 = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

The T^2 statistics based on $\mathbf{S}_1$ are then

$$T^2_{1,i} = (\mathbf{x}_i - \bar{\mathbf{x}})'\,\mathbf{S}_1^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, \ldots, m.$$

Wilks (1963) and Gnanadesikan and Kettenring (1972) showed that for i.i.d. observations from a multivariate normal distribution, as in Equation (1), the exact distribution of $T^2_{1,i}$ is proportional to a beta distribution, i.e.,

$$T^2_{1,i}\frac{m}{(m-1)^2} \sim BETA\left(\frac{p}{2}, \frac{m-p-1}{2}\right), \quad i = 1, \ldots, m. \tag{2}$$

A detailed proof of Equation (2) is also given in Chou, Mason, and Young (1999).

An alternative choice of $\mathbf{S}$ is one based on *successive differences*, proposed originally by Hawkins and Merriam (1974) and later by Holmes and Mergen (1993). To obtain the estimator, we define $\mathbf{v}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ for $i = 1, \ldots, m-1$ and stack the transpose of these $m-1$ difference vectors into the $(m-1)\times p$ matrix $\mathbf{V}$ as

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \\ \vdots \\ \mathbf{v}_{m-1}' \end{bmatrix}.$$

The estimator of the variance-covariance matrix is then

$$\mathbf{S}_D = \frac{\mathbf{V}'\mathbf{V}}{2(m-1)}. \tag{3}$$

Use of matrix $\mathbf{S}_D$ is analogous to use of the moving range estimate of the variance for a univariate X-chart. Sullivan and Woodall (1996) showed that $\mathbf{S}_D$ is an unbiased estimator of $\mathbf{\Sigma}$ if the observations are i.i.d in Phase I. The resulting T^2 statistics based on $\mathbf{S}_D$ are given by

$$T^2_{D,i} = (\mathbf{x}_i - \bar{\mathbf{x}})'\,\mathbf{S}_D^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, \ldots, m. \tag{4}$$

As noted in Sullivan and Woodall (1996), Holmes and Mergen (1993) incorrectly specify the Phase I UCL of a T^2 chart based on $T^2_{D,i}$ statistics by applying control limits based on a Phase II analysis. Sullivan and Woodall (1996) proposed an approximate distribution for $T^2_{D,i}$ as

$$T^2_{D,i}\frac{m}{(m-1)^2} \sim BETA\left(\frac{p}{2}, \frac{f-p-1}{2}\right), \tag{5}$$

where $f = \frac{2(m-1)^2}{3m-4}$. Mason and Young (2002, pp. 26-27) suggested an adjustment to this approximation, replacing each m in Equation (5) with f, giving

$$T^2_{D,i}\frac{f}{(f-1)^2} \sim BETA\left(\frac{p}{2}, \frac{f-p-1}{2}\right). \tag{6}$$

In this paper we demonstrate that these two approximate distributions may not give accurate UCLs.

Prins and Mader (1997) defined a T^2 statistic based on an alternative successive differences estimator. Their estimator is defined by

$$\mathbf{S}_{PM} = \frac{\mathbf{V}_c'\mathbf{V}_c}{2(m-2)}$$

where $\mathbf{V}_c = (\mathbf{I} - \frac{1}{m-1}\mathbf{J}_{m-1})\mathbf{V}$ and $\mathbf{J}_{m-1}$ is an $(m-1)\times(m-1)$ matrix of ones. However, we do not recommend use of this variance-covariance matrix estimator because it is a biased estimator of $\mathbf{\Sigma}$ for an in-control process. Further, the distribution of T^2 based on $\mathbf{S}_{PM}$ is not known. Prins and Mader (1997) based the UCLs on the $\chi^2(p)$ distribution.

Asymptotic Distribution

Although the exact distribution of $T^2_{D,i}$ is unknown, the asymptotic distribution for large m is $\chi^2(p)$ for each $i = 1, \ldots, m$. A proof can be found in Williams, Woodall, Birch, and Sullivan (2004).

To illustrate, we consider the cases $m = 40$ and 1000 and $p = 4$ and 8. We generated 100,000 $T^2_{D,i}$ values for each case and plotted the statistics against the quantiles of the corresponding $\chi^2(p)$ distribution. We generated a sequence of m $T^2_{D,i}$ values by generating m multivariate normal random variables, $\mathbf{x}_i, i = 1, \ldots, m$ from

Equation (1) and applying Equation (4). This process was repeated 100,000 times to generate an empirical distribution of $T^2_{D,i}$ for each $i = 1, \ldots, m$. Figure 1 gives the Q-Q plots for the four combinations of m and p. The uppermost curve of each plot is for $T^2_{D,1}$ and the lowermost curve is for $T^2_{D,2}$. A straight line is also plotted, indicating how closely the statistics follow the $\chi^2(p)$ distribution.

(*Insert Figure 1 about here.*)

Note that the distributions of the first two $T^2_{D,i}$ statistics are different, and the discrepancy vanishes as each distribution asymptotically approaches $\chi^2(p)$ for $i = 1$ and 2. In a limited simulation study (not reported here) we found the same asymptotic behavior for the values $i = 3, \ldots, m$.

Approximate Distribution

In many applications, only a small sample size (m) is available, so an accurate approximate distribution is needed. To evaluate the approximations of Sullivan and Woodall (1996) and Mason and Young (2002), consider the simple case where the HDS is given by

$$\mathbf{X} = \begin{bmatrix} 0.54 & -1.36 \\ -0.75 & 2.50 \\ 0.51 & 0.37 \\ 0.80 & 0.86 \\ 0.92 & 1.14 \end{bmatrix}.$$

The $T^2_{D,i}$ statistics multiplied by both $\frac{m}{(m-1)^2}$ (Sullivan and Woodall (1996)) and $\frac{f}{(f-1)^2}$ (Mason and Young (2002)), as given in Equations (5) and (6), respectively, are given in Table 1.

(*Insert Table 1 about here.*)

Table 1 shows that the scaling factors proposed are too big to constrain the $T^2_{D,i}$ statistics to be less than one. Based on this example, we conclude that these approximate distributions may not be the best choice in determining an appropriate UCL for the T^2 chart since beta random variables must be between zero and one.

An alternative approach is to divide each $T^2_{D,i}$ statistic by its true maximum value, thus constraining it to be between zero and one so that a beta distribution may be a more accurate approximate distribution. The maximum value of each $T^2_{D,i}$ statistic depends on m and i in the following way

$$MV(m,i) = \frac{2(m-1)}{m}\left(i - \frac{m+1}{2}\right)^2 + \frac{(m-1)^2(m+1)}{6m},\ i = 1, \ldots, m. \qquad (7)$$

A detailed proof can be found in Williams, Woodall, Birch, and Sullivan (2004). Note in Table 1 that the all the $T^2_{D,i}$ statistics divided by $MV(m,i)$ are between zero and one.

It is interesting to note that the maximum value does not depend on p, only on m and i. Also note the symmetry about the center, $\frac{m+1}{2}$. In other words, the $T^2_{D,1}$ and $T^2_{D,m}$ statistics have the same maximum value, $T^2_{D,2}$ and $T^2_{D,m-1}$ have the same maximum value, and so forth. The maximum value is greatest for the cases where $i = 1$ and $i = m$ and is smallest for the center position(s). For the univariate case ($p = 1$), Woodall (1992) demonstrated a similar pattern for the maximum X-chart statistics for individual observations based on the moving range.

To illustrate the variation in the maximum value of $T^2_{D,i}$ for each i, we generated 100,000 $T^2_{D,i}$ statistics, for $i = 1, \ldots, m$ for the cases $p = 2$ and $m = 5$, and plotted the boxplots in Figure 2. The T^2 statistics were generated in the same fashion as described in the previous section.

(*Insert Figure 2 about here.*)

Note that the maximum values of $T^2_{D,1}$ and $T^2_{D,5}$ are the same and the largest values, whereas the maximum of $T^2_{D,3}$ is the smallest. Also note the symmetry about the center position, $i = 3$. A similar pattern holds for all cases in which $p < m - 1$, where the maximum values are greatest for values of $i = 1$ and m, and smallest in the middle.

As shown in Theorem 2 of Williams et al. (2004), an interesting phenomenon occurs for the case of $p = m - 1$. For this case all of the $T^2_{D,i}$ statistics, $i = 1, \ldots, m$, are necessarily equal to their respective maximum values as given by Equation (7), regardless of the HDS used. This is a degenerate case since each $T^2_{D,i}$ statistic becomes deterministic and equals a constant, as given in Equation (7).

We conjecture that the $T^2_{D,i}$ statistics divided by the true maximum value have an approximate distribution given by

$$T^2_{D,i}\frac{1}{MV(m,i)} \sim BETA(\beta(m,p,i),\gamma(m,p,i)), \quad i=1,\ldots,m, \tag{8}$$

where $\beta(m,p,i)$ and $\gamma(m,p,i)$ are functions of m, p, and i. To obtain the approximate distributions, we generated 100,000 independent $T^2_{D,i}$ statistics for each combination of $p = 2,\ldots,10$, $m = 20, 25, 30,\ldots,70$, and $i = 1,\ldots,m$, assuming, without loss of generality, that $\mathbf{x}_i \sim N_p(\mathbf{0},\mathbf{I})$. We make this assumption based on the fact that the $T^2_{D,i}$ values are invariant to a full-rank linear transformation on the observation vectors, as shown by Sullivan and Woodall (1996). For fixed m and p we then found the maximum likelihood estimates for $\beta(m,p,i)$ and $\gamma(m,p,i)$ for each set of 100,000 $T^2_{D,i}$ statistics, $i = 1,\ldots,m$. Then, using nonlinear regression we fit a parametric function to estimate $\beta(m,p,i)$ and $\gamma(m,p,i)$ as a function of m, p, and i. The functions were chosen based on obvious patterns in the estimated $\beta(m,p,i)$ and $\gamma(m,p,i)$ parameters. In order to adequately capture the form of $\beta(m,p,i)$ and $\gamma(m,p,i)$ as a function of m, p, and i, we fit the functions in the following way:

1. Fit the estimates of $\beta(m,p,i)$ and $\gamma(m,p,i)$ as a function of one of m, p, or i for every combination of the other two.

2. Fit the estimated functional parameters from Step 1 as a function of one of the remaining parameters, holding the third fixed.

3. Finally, fit the estimated functional parameters from Step 2 as a function of the third.

Combining the results of steps 1 – 3, the estimated function for $\beta(m,p,i)$ is

$$\beta(m,p,i) = I_{\{i=1,m\}}\left(\frac{p}{2} - \frac{1}{a_{11}(m-b_{11})}\right) + I_{\{i=2,\ldots,m-1\}}\left(a_{12}p + b_{12}\right), \tag{9}$$

and the estimated function for $\gamma(m,p,i)$ is given by

$$\gamma(m,p,i) = I_{\{i=1,m\}}a_{21} + I_{\{i=2,\ldots,m-1\}}\left[a_{22}\left(i-\frac{m+1}{2}\right)^2 + b_{22}\right], \tag{10}$$

where

$$
\begin{aligned}
I_{\{i=1,m\}} &= \begin{cases} 1 & \text{if } i = 1 \text{ or } i = m \\ 0 & \text{otherwise} \end{cases} \\
I_{\{i=2,\ldots,m-1\}} &= \begin{cases} 1 & \text{if } 2 \leq i \leq m-1 \\ 0 & \text{otherwise} \end{cases} \\
a_{11} &= 6.356e^{-0.825p} + 0.06 \\
b_{11} &= 0.5564p + 0.9723 \\
a_{12} &= 0.54 - 0.25e^{-0.25(m-15)} \\
b_{12} &= -0.085 + 0.2e^{-0.2(m-22)} \\
a_{21} &= (-0.5m + 2)p + \frac{1}{3}(m+3)(m-5) \\
a_{22} &= 0.99 + 0.38e^{0.38(p-13.5)} - \frac{1}{0.25e^{-0.25(p-10)}\left(m - 11 + \frac{(p-7)^2}{3}\right)} \\
b_{22} &= (0.07e^{-0.07(m-42)} - 1.95)p + 0.0833m^2
\end{aligned}
$$

As an example, for the case where $m = 40$, $p = 5$, and $i = 20$, $\beta(40, 5, 20) = 2.618$ and $\gamma(40, 5, 20) = 124.174$. The shape parameters for $i = 1$ are $\beta(40, 5, 1) = 2.330$ and $\gamma(40, 5, 1) = 411.667$. Also note that $\beta(m, p, i) = \beta(m, p, m - i + 1)$ and $\gamma(m, p, i) = \gamma(m, p, m - i + 1)$, demonstrating the symmetry about the center position(s).

One theoretical justification for the form of $\beta(m, p, i)$ and $\gamma(m, p, i)$ is the resemblance to the distribution of $T^2_{1,i}$ given in Equation (2). For example, for large m $\beta(m, p, i)$ is approximately $\frac{p}{2}$, which is the value of the first shape parameter in Equation (2). The form of $\gamma(m, p, i)$ is less intuitive since both $\gamma(m, p, i)$ and the second shape parameter given in Equation (2) are functions of m and p. However, note the similarity of the form of the maximum value of $T^2_{D,i}$ given in Equation (7) and the form of $\gamma(m, p, i)$.

The functions for $\beta(m, p, i)$ and $\gamma(m, p, i)$ have two cases, one case where $i = 1$ or m and the other case where $i = 2, \ldots, m-1$. These cases are treated separately since the distributions of the first and last $T^2_{D,i}$ statistics can be quite different from those of the other $T^2_{D,i}$ statistics. One theoretical justification for this difference can be found by examining the definitions of $\mathbf{S}_D$ and the $T^2_{D,i}$ statistics, given in Equations (3) and (4), respectively. In the calculation of $\mathbf{S}_D$, observation vectors $\mathbf{x}_1$ and $\mathbf{x}_m$ appear only once in the successive differences $\mathbf{v}_i$ whereas all other observation vectors appear twice. Hence it is intuitive that the distributions for $i = 1$ and m should be different from the others. However, the distributions asymptotically approach $\chi^2(p)$, as expected.

Next, we illustrate how well the approximations fit by showing some example Q-Q plots of the (simulated) actual distribution and a suggested approximating distribution. We compare the approximations of Equations (8), (5), and (6) by showing the quantiles from the left hand side of an equation on the vertical axis against the corresponding quantiles from the respective approximation (from the right hand side of the equation) on the horizontal axis. These Q-Q plots are shown in Figure 3, where our proposed approximation from Equation (8) is presented in the left column of plots, the Sullivan and Woodall (1996) approximation (SW), from Equation (5), is in the middle column of plots, and the Mason and Young (2002) approximation (MY), from Equation (6), is in the right column. In these plots, the T^2 statistics are scaled by different factors, as given in Equations (8), (5), and (6).

We show these plots for $p = 4$ and for $m = 30$ and 60 observations. The distributions of the first and last chart statistics are the same, and their common distribution (scaled) is the focus of the first and third rows of the plots. The distribution for the middle chart statistic (scaled) is shown in the second and fourth rows. The first column gives an informal, visual impression of the close relationship of our proposed approximation with the actual distribution, while there is a greater deviation from the theoretical quantiles with the other proposed approximations. There are similar patterns for $p = 8$ and $m = 30$ and 60 observations, and for other cases with small numbers of observations, but these are not shown in the interest of brevity. We generated 10,000 simulated statistics for each plot.

(*Insert Figure 3 about here.*)

We note that there is a clear improvement in the goodness-of-fit for the proposed approximate distributions of Equation (8) for each case.

Performance Comparison

Control Limits

Once the distribution of $T^2_{D,i}$ is approximated, an approximate upper control limit corresponding to an overall probability of a false alarm may be calculated. Hence, we need the joint distribution of the $T^2_{D,i}$ statistics. However, the $T^2_{D,i}$ values are correlated, since each statistic is based on the same $\bar{\mathbf{x}}$ and $\mathbf{S}_D$, thus making the joint distribution of the $T^2_{D,i}$ values difficult to obtain. As an alternative, Mahmoud and Woodall (2004) suggested using an approximate joint distribution assuming that the $T^2_{D,i}$ statistics are independent. In their simulation study, Mahmoud and Woodall (2004) found that UCLs based on this approach performed well, and we follow their suggestion here. Let α be the probability of a false alarm for any individual $T^2_{D,i}$ statistic. Then the approximate overall probability of a false alarm for a sample of m independent statistics is $\alpha_{overall} = 1 - (1-\alpha)^m$. Thus, for a given overall probability of a false alarm, we use

$$\alpha = 1 - (1 - \alpha_{overall})^{1/m} \tag{11}$$

in calculation of UCLs.

To obtain the UCL that achieves the false alarm rate α, Sullivan and Woodall (1996) recommended

$$UCL_{SW} = \frac{(m-1)^2}{m} BETA_{1-\alpha,p/2,(f-p-1)/2}, \tag{12}$$

where $BETA_{1-\alpha,p/2,(f-p-1)/2}$ is the $1-\alpha$ quantile of a beta distribution with shape parameters $\frac{p}{2}$ and $\frac{f-p-1}{2}$. Mason and Young (2002) suggested the UCL defined by

$$UCL_{MY} = \frac{(f-1)^2}{f} BETA_{1-\alpha,p/2,(f-p-1)/2}. \tag{13}$$

When use of the asymptotic distribution of $T^2_{D,i}$ is justified, then the UCL is given by

$$UCL_{\chi^2} = \chi^2_{1-\alpha,p}, \tag{14}$$

where $\chi^2_{1-\alpha,p}$ is the $1-\alpha$ quantile of a $\chi^2(p)$ distribution.

Because the distribution of $T^2_{D,i}$ varies with i, one UCL is not appropriate for all the $T^2_{D,i}$ statistics. Instead a set of m UCL values is needed, one UCL for each i. We call the set of UCL values the UCL vector, with element i given by

$$UCL_i = MV(m,i)BETA_{1-\alpha,\beta(m,p,i),\gamma(m,p,i)}, \quad i = 1,\ldots,m, \tag{15}$$

and $\mathbf{UCL}_{vec} = (UCL_1, UCL_2, \ldots, UCL_m)$. The T^2 chart based on the UCL vector will signal whenever $T^2_{D,i} > UCL_i$ for any $i = 1,\ldots,m$.

Simulation Study

To compare the performance of the alternative control limits given by UCL_{SW}, UCL_{MY}, UCL_{χ^2}, and $\mathbf{UCL}_{vec}$, we simulated 100,000 Phase I charts for each combination of $m = 20, 25, 30, \ldots, 70$ and $p = 2, 3, \ldots, 10$, recorded the number of signals, and estimated the true false alarm rate as the number of signals divided by 100,000. A Phase I chart can be simulated by generating m i.i.d. multivariate normal random variables according to Equation (1) and applying Equation (4). Then each $T^2_{D,i}$ statistic is compared to the UCL. If one or more of the $T^2_{D,i}$ values exceed the UCL, then the chart produces a signal. The nominal (or target) overall probability of false alarm is set to be $\alpha_{overall} = 0.05$. The large number of simulations provides for an approximate standard error of $\sqrt{\frac{(0.05)(1-0.05)}{100{,}000}} = 0.00068$, so that the estimates are accurate to approximately $\pm 3(0.00068) = \pm 0.002$. In Figure 4 we plot the estimated false alarm rates for the four methods for $p = 2, 3, 4,$ and 5, and in Figure 5 we plot the estimates for $p = 6, 7, 8,$ and 9.

(*Insert Figures 4 and 5 about here.*)

A reference line is plotted at $\alpha_{overall} = 0.05$. It is clear from Figures 4 and 5 that the false alarm rates produced by UCL_{MY} are larger than the nominal value, or higher than intended, for small samples, but the false alarm rate is asymptotically more accurate for large m. The false alarm rates produced by UCL_{SW} are greater than 0.05 for small m and less than 0.05 for large m. Further, as m gets larger, the false alarm rate gets smaller. Thus, even for large values of m, the false alarm rate produced by UCL_{SW} will be smaller than expected. The UCL_{χ^2} performs well for sufficiently large m, with accuracy generally decreasing as m decreases or p increases. For large p and small m, the $\mathbf{UCL}_{vec}$ produces false alarm rates much closer to the nominal level than does UCL_{χ^2}. In a separate simulation study (not reported here) we found similar results using $\alpha_{overall} = 0.01$.

Example

To illustrate use of the four methods, we analyzed the data in Quesenberry (2001). Thirty items were sampled with eleven quality characteristics measured on each item. For the purposes of this example, we consider only the first nine quality characteristics. To calculate $\mathbf{UCL}_{vec}$, we calculate $\beta(m, p, i)$ and $\gamma(m, p, i)$ in Equations (9) and (10) for each observation $i = 1, \ldots, m$ in this dataset. For the first observation, we obtain $\beta(30, 9, 1) = 3.776$ and $\gamma(30, 9, 1) = 158$. We choose an overall probability of a false alarm $\alpha_{overall} = 0.05$, which yields $\alpha = 0.0017$ by Equation (11). Further the maximum value of the $T^2_{D,1}$ statistic is given by $MV(30, 1) = 551.322$ in Equation (7). Then, by Equation (15)

$$UCL_1 = MV(30, 1)BETA_{1-\alpha,\beta(30,11,1),\gamma(30,11,1)} = 39.948.$$

Similarly, $\beta(30, 9, 2) = 4.762$, $\gamma(30, 9, 2) = 223.911$, $MV(30, 2) = 497.189$, producing $UCL_2 = 29.228$, and so forth. The $\mathbf{UCL}_{vec}$ values and the associated $T^2_{D,i}$ values are given in Table 2.

(*Insert Table 2 about here.*)

By Equations (12) – (14), we obtain $UCL_{SW} = 24.828$, $UCL_{MY} = 15.596$, and $UCL_{\chi^2} = 26.474$. In Figure 6 we plot the $T^2_{D,i}$ statistics with the four UCL lines.

(*Insert Figure 6 about here.*)

Observations 2, 7, 13, and 30 exceed the UCL_{MY} control limit, and observation 2 exceeds the UCL_{SW} control limit. No observations exceed the UCL_{χ^2} or $\mathbf{UCL}_{vec}$

control limits. Since these calculations were based on observational rather than simulated data, the true joint distribution of the observations is necessarily unknown.

Discussion

There has been no previous work on the number of samples required for the $T^2_{D,i}$ statistics to approximately follow a $\chi^2(p)$ distribution. However, there have been a number of papers addressing the issue of the number of samples with respect to the use of $T^2_{1,i}$ statistics. For example, Chou, Mason, and Young (2001) made recommendations for the number of samples required for a Phase II analysis with individual observations following a non-normal multivariate distribution. Lowry and Montgomery (1995) considered both Phase I and II, analyzing both individual and subgrouped observations. They used a relative error criteria in comparing the actual UCL and the UCL obtained from assuming that $T^2_{1,i}$ follows a $\chi^2(p)$ distribution to determine the minimum number of samples. Nedumaran and Pignatiello (1999) studied the sample size requirements for using $\chi^2(p)$ for a Phase I analysis with subgrouped observations. From these studies, there is no clear consensus on the minimum number of samples.

In the calculation of the $T^2_{1,i}$ statistics, we must estimate p parameters in the mean vector and $\frac{p(p+1)}{2}$ variance and covariance parameters, for a total of $p + \frac{p\,(p+1)}{2}$ parameters, as noted by Mason and Young (2002, pp. 40-50). However, it is interesting to note that even with a sample size $m < p + \frac{p\,(p+1)}{2}$, the distribution of the $T^2_{1,i}$ statistics is still the beta distribution as given in Equation (2). Hence, even for small sample sizes, one can still find an appropriate UCL for a Phase I analysis as long as $m \geq p$. However, if $m < p$ then $\mathbf{S}_1$ is a singular matrix, and $\mathbf{S}_1^{-1}$ does not exist.

For the T^2 chart based on $T^2_{D,i}$ statistics, one criteria for specifying the minimum sample size for the asymptotic approximation to be accurate is a comparison between the desired overall probability of a false alarm and the actual probability for a certain sample size. In our simulation study, we estimated the actual probability of a false alarm based on UCL_{χ^2} for $p = 2, \ldots, 9$ and $m = 20$ to $m = 100$. Figures 4 and 5 show that when the number of samples is at least twice the number of parameters estimated $\left(p + \frac{p(p+1)}{2}\right)$, then the UCL_{χ^2} false alarm probability is nearly the nominal value. Hence, we recommend the use of UCL_{χ^2} when $m > 2p + p(p+1) = p^2 + 3p$ and $\mathbf{UCL}_{vec}$ for smaller samples. In a limited simulation study (not reported here) we found that UCL_{χ^2} produces a false alarm rate close to the nominal false alarm rate for a T^2 chart based on $T^2_{D,i}$ statistics when $m > p^2 + 3p$, regardless the value of p. Note, however, that in order to use UCL_{χ^2} a large m is required. Similarly, for the case of the $T^2_{1,i}$ statistic, Hawkins (1974) noted that a very large m is required for use of the $\chi^2(p)$ distribution.

In our determination of the improved approximation given in Equation (8), we limited our scope to $p < 10$. Our proposed approximation is accurate for values of $p < 10$, but for values of $p \geq 10$, the accuracy of our approximation has not been thoroughly studied and remains a topic for future research. A summary of our recommended UCL for a T^2 chart is given in Table 3.

(*Insert Table 3 about here.*)

As is the case with the $T^2_{1,i}$ statistics, the $T^2_{D,i}$ statistics, $i = 1, \ldots, m$, are correlated. The correlations between the $T^2_{D,i}$ statistics, however, are much more complicated than the correlations between the $T^2_{1,i}$ statistics. Where the correlation of the $T^2_{1,i}$ statistics is known to be $\frac{-1}{m-1}$, the correlation of the $T^2_{D,i}$ statistics depends on m, p, and i. Yet, this correlation diminishes in magnitude as m increases relative to p. The importance of the correlation lies in its effect on control chart performance. Figure 3 demonstrates that our method of finding the UCL for the T^2 chart based on the $T^2_{D,i}$ statistic is an improvement over the methods suggested by Sullivan and Woodall (1996) and Mason and Young (2002), and performs better than the chi-square distribution when m is small. Although we motivated the results by ignoring the correlation, our simulations did include the actual correlations. The intermediate properties, such as the α for a single statistic would therefore deviate with small m, but our final result is accurate.

In addition to our proposed approximate distribution, exploring other approximate distributions of the $T^2_{D,i}$ statistics is a topic for further research. A referee suggested that a reasonable approximate distribution is given by

$$\frac{p}{E(T_{D,i}^2)} T_{D,i}^2 \sim \chi^2(p), \quad i = 1, \ldots, m,$$

where $E(T_{D,i}^2)$ is the expected value of $T_{D,i}^2$. In implementing this approximation one would have the challenge of deriving the form of $E(T_{D,i}^2)$, which is a function of m, p, and i. Alternatively, a simulation based approach for computing $E(T_{D,i}^2)$ could also be employed when an analytic solution is infeasible.

Conclusion

We have considered how to accurately determine the upper control limit for a T^2 control chart based on successive differences of multivariate individual observations. We have shown that previously proposed approximations by Sullivan and Woodall (1996) and Mason and Young (2002) give an actual false alarm probability different from what is intended when the sample size is small.

We give an analytical result for the maximum value of the T^2 chart statistics when the successive differences estimator is used, showing that the maximum value is not the same for all observations. We give a formula for how the maximum value depends on the observation number. We divide the T^2 chart statistic by its maximum value to produce a statistic whose distribution is approximately beta, with parameters that we approximate based on simulations.

We then studied the chart performance with our proposed control limit, which varies with the position of the observation. We show that with our proposed limit the actual false alarm probability is much closer to the specified value with small samples.

Acknowledgement

We would like to thank Professor Charles Champ who handled the review process for this article, as well as the two anonymous referees who carefully read the earlier drafts and made many constructive suggestions.

References

Chou, Y. -M.; Mason, R. L.; and Young, J. C. (1999). "Power Comparisons for a Hotelling's T^2 Statistic". *Communications in Statistics, Part B – Simulation and Computation* 28, pp. 1031-1050.

Chou, Y. -M.; Mason, R. L.; and Young, J. C. (2001). "The Control Chart for Individual Observations from a Multivariate Non-normal Distribution". *Communications in Statistics, Part A – Theory and Methods* 30 (8-9), pp. 937-949.

Fuchs, C. and Kenett, R. S. (1998). *Multivariate Quality Control: Theory and Applications.* Marcel Dekker, New York, NY.

Gnanadesikan, R. and Kettenring, J. R. (1972). "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data". *Biometrics* 28, pp. 81-124.

Hawkins, D. M. and Merriam, D. F. (1974). "Zonation of Multivariate Sequences of Digitized Geologic Data". *Mathematical Geology* 6, pp. 263-269.

Hawkins, D. M. (1974). "The Detection of Errors in Multivariate Data Using Principal Components". *Journal of the American Statistical Association* 69, pp. 340-344.

Holmes, D. S and Mergen, A. E. (1993). "Improving the Performance of the T^2 Control Chart". *Quality Engineering* 5, pp. 619-625.

Lowry C. A. and Montgomery D. C. (1995). "A Review of Multivariate Control Charts". *IIE Transactions* 27, pp. 800-810.

Mahmoud, M. A. and Woodall, W. H. (2004). "Phase I Analysis of Linear Profiles with Calibration Applications". *Technometrics* 46, pp. 377-391.

Mason, R. L. and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications.* SIAM, Philadelphia, PA.

Mason, R. L.; Chou, Y. -M.; Sullivan, J. H.; Stoumbos, Z. G.; and Young, J. C. (2003). "Systematic Patterns in T^2 Charts". *Journal of Quality Technology* 35, pp. 47-58.

Nedumaran, G. and Pignatiello, J. J., Jr. (1999). "On Constructing T^2 Control Charts for On-line Process Monitoring". *IIE Transactions* 31, pp. 529-536.

Prins, J. and Mader, D. (1997). "Multivariate Control Charts for Grouped and Individual Observations". *Quality Engineering* 10, pp. 49-57.

Quesenberry, C. P. (2001). "The Multivariate Short-Run Snapshot Q Chart". *Quality Engineering* 13, pp. 679-683.

Rencher, A. C. (2000). *Linear Models in Statistics.* Wiley, New York, NY.

Sullivan, J. H. and Woodall, W. H. (1996). "A Comparison of Multivariate Control Charts for Individual Observations". *Journal of Quality Technology* 28, pp. 398-408.

Sullivan, J. H. (2002). "Detection of Multiple Change Points from Clustering Individual Observations". *Journal of Quality Technology* 34, pp. 374-383.

Vargas N., J. A. (2003). "Robust Estimation in Multivariate Control Charts for Individual Observations". *Journal of Quality Technology* 35, pp. 367-376.

Wilks, S. S. (1963). "Multivariate Statistical Outliers". *Sankhyā A* 25, pp. 407-426.

Williams, J. D.; Woodall, W. H.; Birch, J. B.; and Sullivan, J. H. (2004). "Distributional Properties of the Multivariate T^2 Statistic Based on the Successive Differences Covariance Matrix Estimator". Technical Report No. 04-5, Department of Statistics, Virginia Polytechnic Institute and State University.

Woodall, W. H. (1992). "A Note on Maximum Z-scores for Control Charts for Individuals". *Communications in Statistics, Part A – Theory and Methods* 21, pp. 3211-3217.

Woodall, W. H.; Spitzner, D. J.; Montgomery, D. C.; and Gupta, S. (2004). "Using Control Charts to Monitor Process and Product Quality Profiles". *Journal of Quality Technology* 36, pp. 309-320.

Table 1: The $T^2_{D,i}$ statistics scaled according to Sullivan and Woodall (1996), Mason and Young (2002), and Equation (7) for a data set.

i	1	2	3	4	5
$\frac{m}{(m-1)^2}T^2_{D,i}$	2.572	1.499	0.016	1.017	2.294
$\frac{f}{(f-1)^2}T^2_{D,i}$	6.578	3.828	0.041	2.599	5.860
$\frac{1}{MV(m,i)}T^2_{D,i}$	0.857	0.999	0.016	0.678	0.765

Table 2: The $T^2_{D,i}$ statistics and $\mathbf{UCL}_{vec}$ values based on the Quesenberry (2001) data.

i	$T^2_{D,i}$	$\mathbf{UCL}_{vec}$	i	$T^2_{D,i}$	$\mathbf{UCL}_{vec}$	i	$T^2_{D,i}$	$\mathbf{UCL}_{vec}$
1	6.418	39.948	11	7.220	29.232	21	12.734	29.235
2	26.400	29.228	12	12.538	29.229	22	6.390	29.236
3	7.880	29.230	13	19.228	29.225	23	3.183	29.236
4	7.498	29.232	14	9.820	29.222	24	6.755	29.236
5	12.233	29.233	15	9.780	29.219	25	10.585	29.235
6	5.656	29.235	16	15.377	29.219	26	13.100	29.233
7	21.705	29.236	17	13.979	29.222	27	4.395	29.232
8	4.124	29.236	18	6.375	29.225	28	7.008	29.230
9	6.578	29.236	19	5.280	29.229	29	12.765	29.228
10	6.835	29.235	20	15.234	29.232	30	20.193	39.948

Table 3: Recommended UCL for the T^2 chart based on the $T^2_{D,i}$ statistic.

	$p < 10$	$p \geq 10$
$m > p^2 + 3p$	UCL_{χ^2}	UCL_{χ^2}
$m \leq p^2 + 3p$	$\mathbf{UCL}_{vec}$	No recommendation

Figure 1: Q-Q plots of empirical quantiles of the $T^2_{D,i}$ statistic ($i = 1$ and 2) versus a $\chi^2(p)$ distribution. The uppermost curve is for $T^2_{D,1}$ and the lowermost curve is for $T^2_{D,2}$.

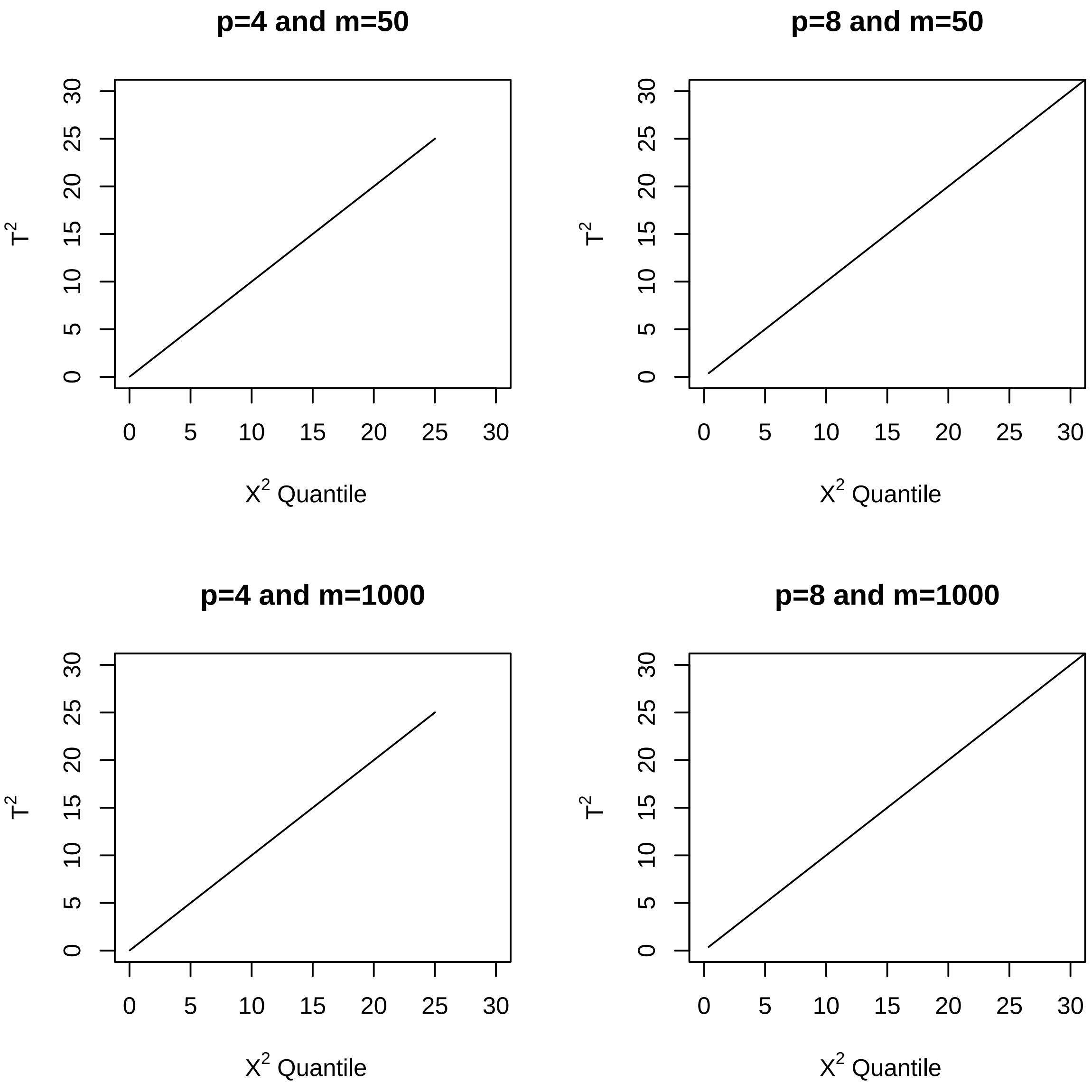

Figure 2: Boxplots of $T^2_{D,i}$ for $p = 2$ and $m = 5$.

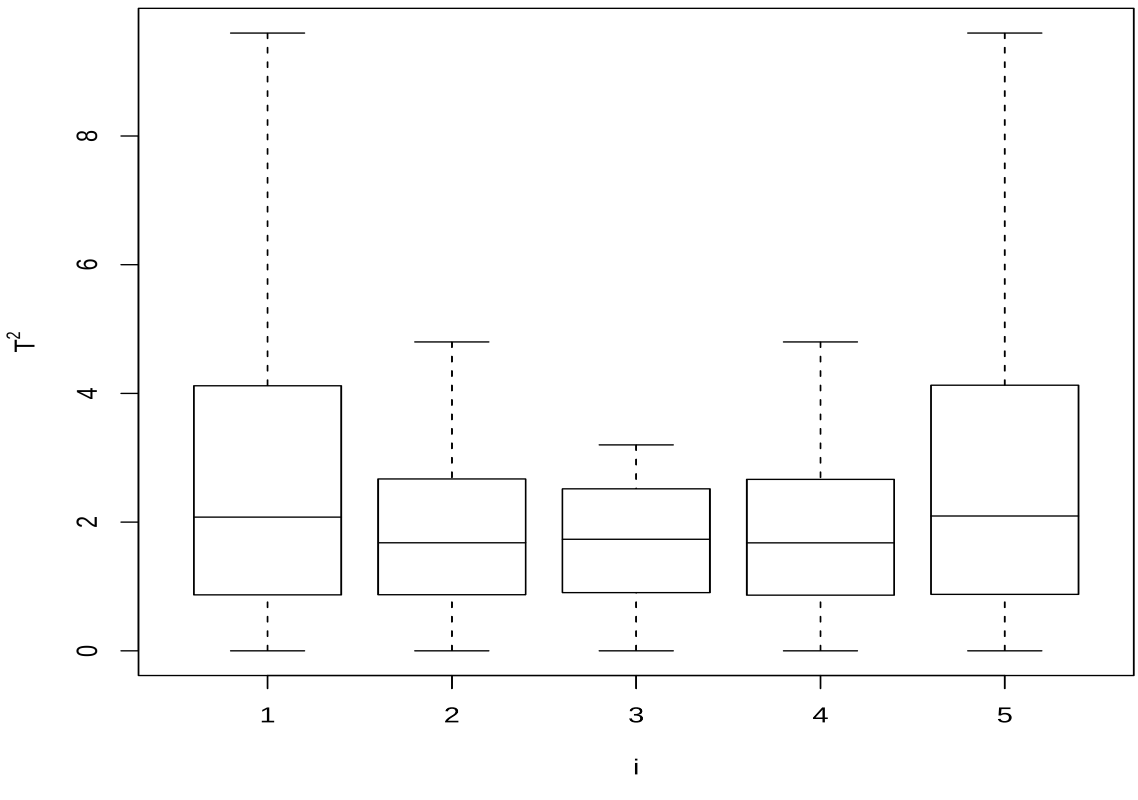

Figure 3: Q-Q plots of empirical versus the theoretical quantiles. The first column is for the proposed approximation, the second column is for the Sullivan and Woodall (1996) approximation, and the last column is for the Mason and Young (2002) approximation.

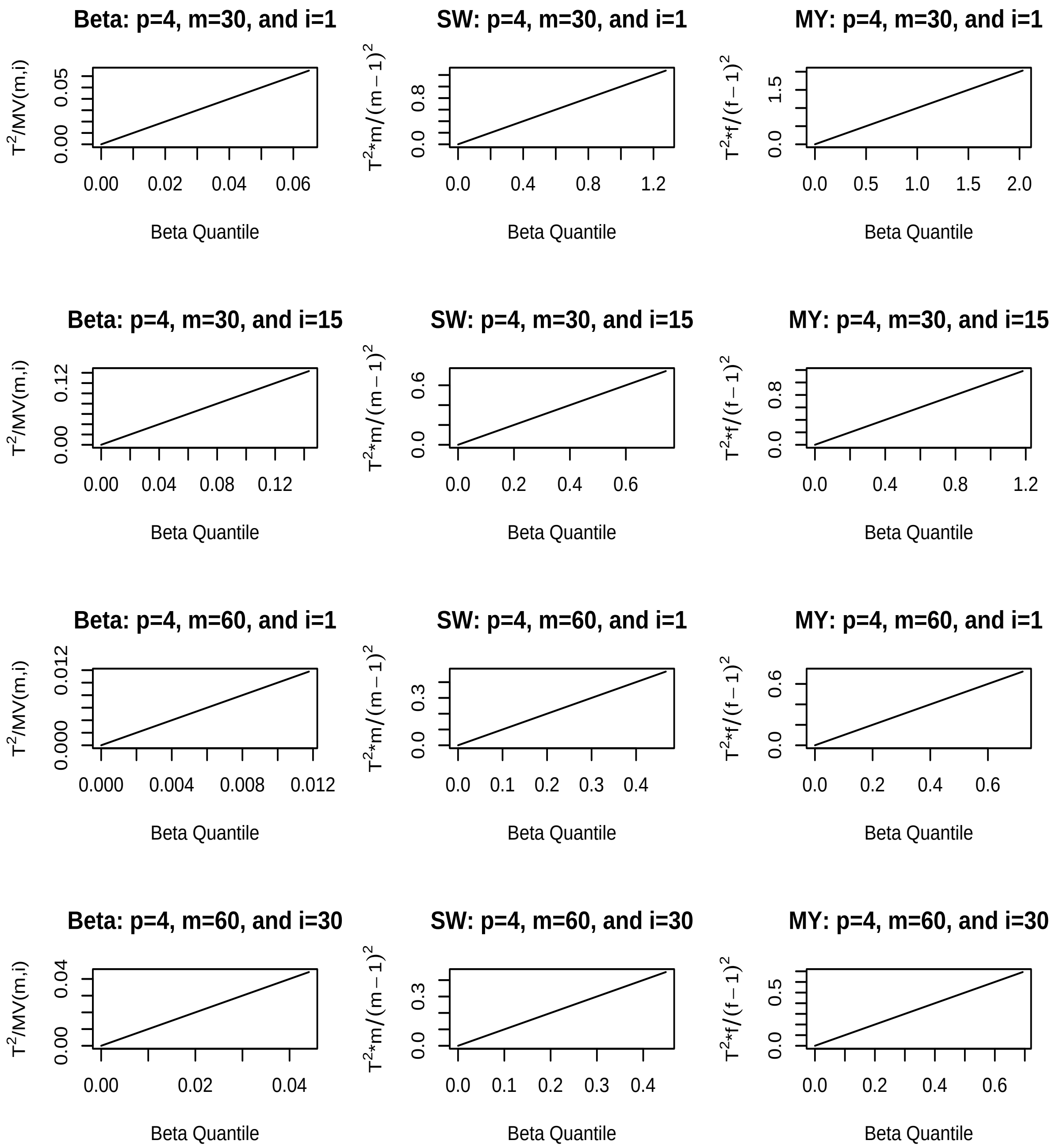

Figure 4: Overall Probability of a False Alarm for $p = 2, 3, 4, 5$. The intended false alarm rate is 0.05. Key: Proposed approximation (solid), S&W (dashed), M&Y (dotted), and $\chi^2(p)$ (dash-dotted).

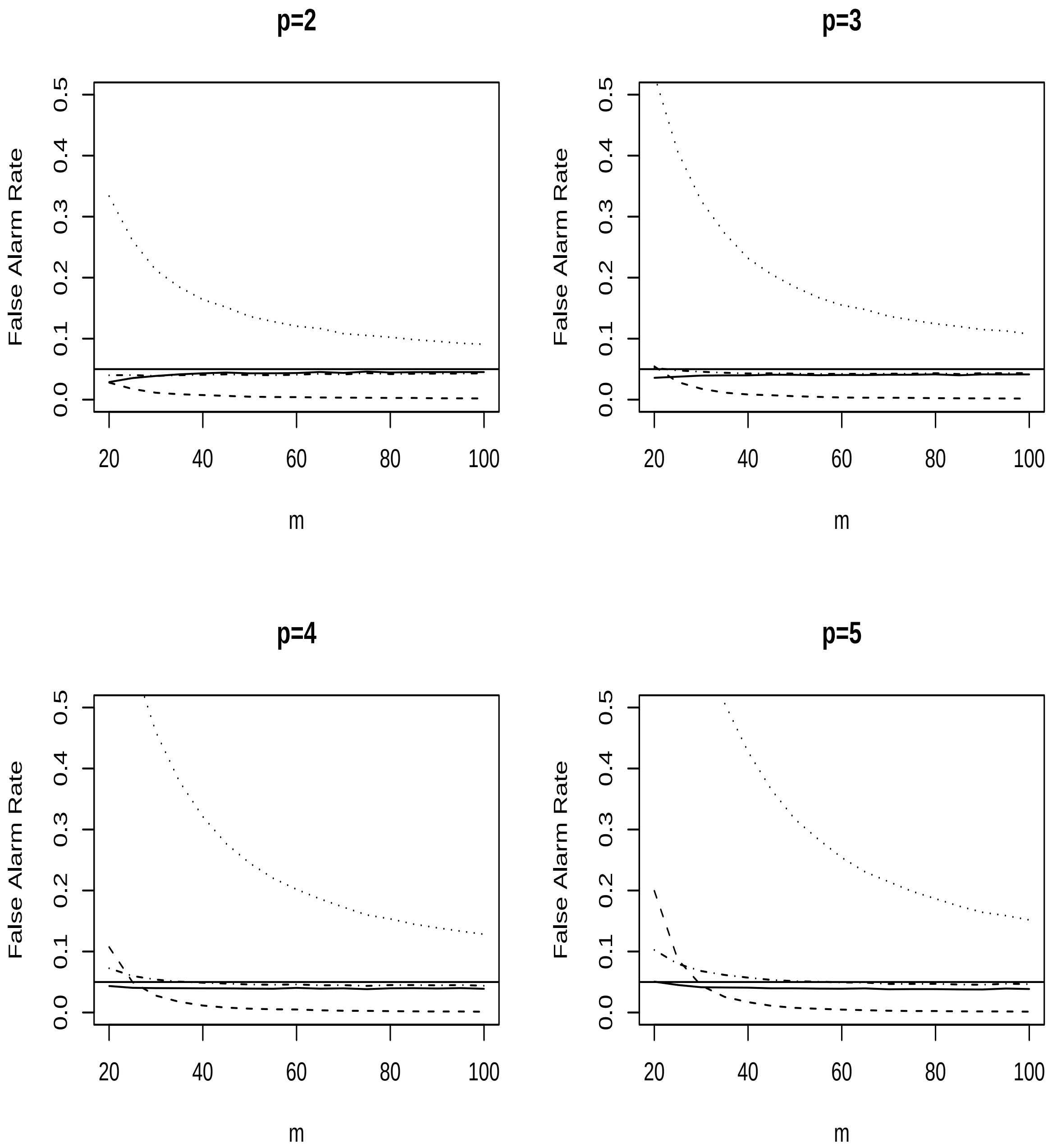

Figure 5: Overall Probability of a False Alarm for $p = 6, 7, 8, 9$. The intended false alarm rate is 0.05. Key: Proposed approximation (solid), S&W (dashed), M&Y (dotted), and $\chi^2(p)$ (dash-dotted).

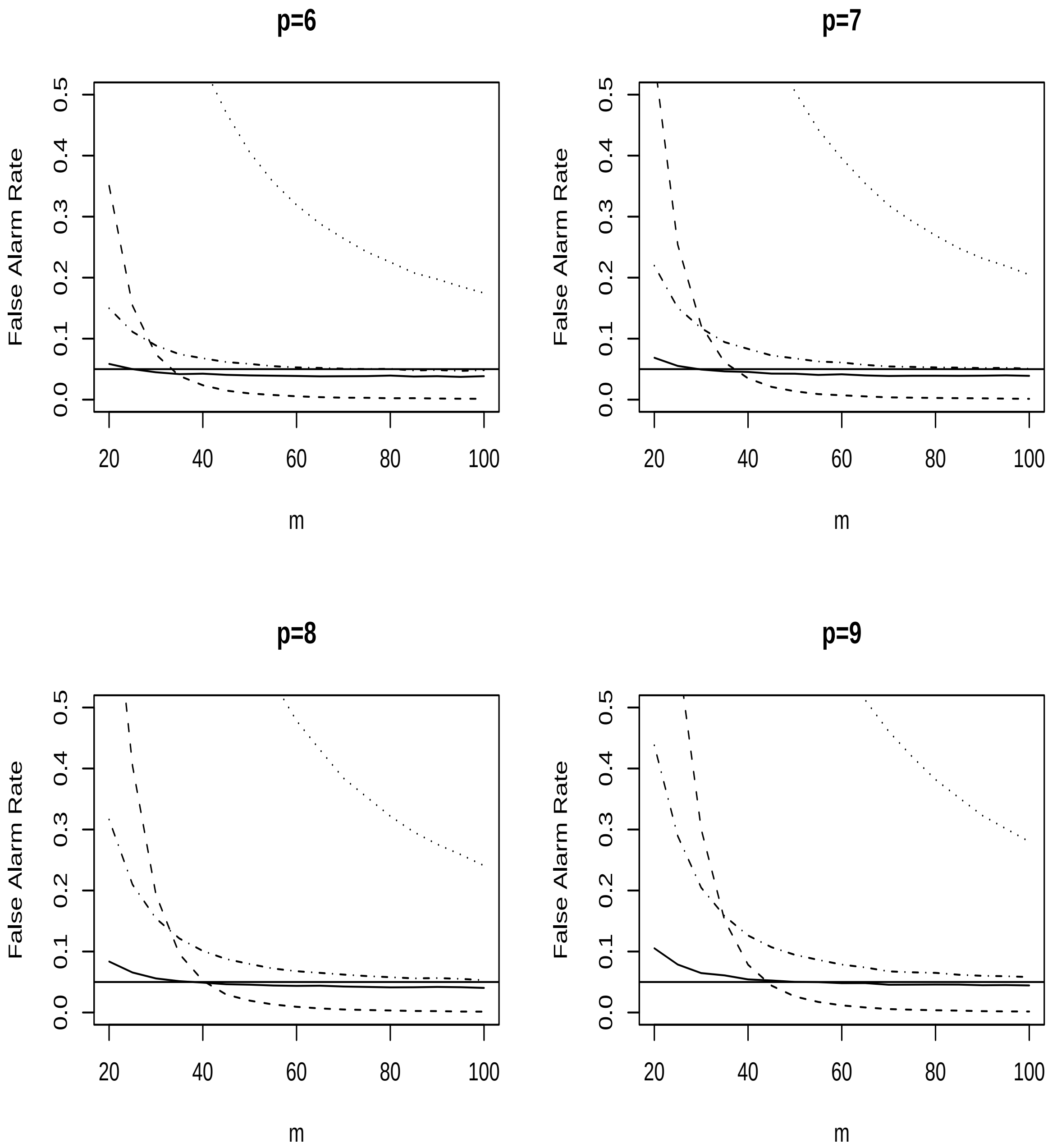

Figure 6: $T^2_{D,i}$ statistics and UCL values for the Quesenberry (2001) data. Key: Proposed approximation (solid), S&W (dashed), M&Y (dotted), and $\chi^2(p)$ (dash-dotted).