Constructing Multivariate Control Charts with SAS™ Software

Melissa A. Durfee, Wyman-Gordon Company, North Grafton, MA

Abstract

Simultaneous control of two or more related quality characteristics has become an important subject, particularly due to the increased use of automated inspection which may measure many parameters on each unit of manufactured product. Referred to as multivariate quality control, this technique surpasses standard control charts generated on each individual characteristic since the probability of a type I error (i.e., "false alarm") is reduced. Using SAS/STAT™ and SAS/QC™ procedures, a Hotelling T^2 control chart on related characteristics may be constructed. Since the state of the process is characterized by a single number, the value of the T^2 statistic, this method is advantageous when there are two or more quality characteristics of interest. Using a control chart preserves the time sequence of data. Therefore, the identification of runs or other nonrandom patterns is facilitated.

Introduction

The multivariate approach to quality control was first widely publicized in 1947 and 1951 by H. Hotelling who applied this technique in the testing of bombsights. In his procedure, two sights were randomly selected from each lot of 20 sights. Each sight was then tested by dropping four bombs each on two flights. The range error, which is measured in the plane's flight direction, and the deflection error, which is measured perpendicular to the flight path, are correlated. The multivariate approach was used to monitor the quality of each bomb dropped at target.

Hotelling introduced the T^2 control chart as a technique for monitoring the overall quality of a flight, sight, or lot by summing over the appropriate number of bombs involved. Research in this field remained relatively dormant until the 1950's with the increasing availability of computers.

Presently, this technique for controlling multiple variables becomes more pertinent particularly due to the increased application of automated data acquisition which provides a vast amount of manufacturing process data. In process control, many different kinds of measurements are recorded: e.g., temperature, time, tensile strength, hardness, pressure, purity, chemical composition, etc. The amount of information available needs to be reduced to a useful form. Therefore, subsequent efforts may be focused on the key control characteristics which influence the process.

The data records of interest are a set of measurements that describe the properties of a part, lot, or batch or processing parameters directly influencing the process output. Usually, the variables in the record are directly related to each other. For example, in a tensile test, which is a mechanical test where a specimen is pulled apart while subjected to a specified stress and temperature, the strength measurements (yield and ultimate) are inversely related to the ductility properties (% elongation and % reduction of area). Sometimes, each data record is considered individually. Therefore, control chart calculations also consider the case where the subgroup size equals one (n=1).

Objectives

The objectives of multivariate Statistical Process Control (SPC) include the following:

- differentiate between assignable causes and common causes of process variation by detecting data records that are not within the multidimensional normal operating region of the process;
- improve processes by reducing or eliminating assignable causes by identification through root cause analysis;
- reduce the false alarms indicated by a typical Shewhart control chart when correlated or dependent data is present in the process;
- monitor the process effectively with a reduced number of control charts;
- improve the detection of assignable causes by considering the relationships between variables.

Since the process is monitored with fewer false alarms, multivariate SPC promotes true process improvement.

Assumptions

In multivariate control chart analysis, the following are assumed:

- The data follows a multivariate normal distribution.
- An observation outside the control limit indicates an out of control condition at some specified α risk.
- The distribution of the plotted point, T^2, follows a T^2 distribution when the means and variance-covariance matrix are unknown and therefore are estimated from the data.
- The control region is determined by the number of variables in the record. A two-dimensional (bivariate) region is shaped like an ellipse (see Figures 1 and 2). A three-dimensional (or higher) region may look like a blimp or cigar.
- Control regions are chosen such that they encompass almost all the points generated by the process under normal operating conditions.
- Control regions detect assignable causes that change the relationship between the variables.

Control of Means

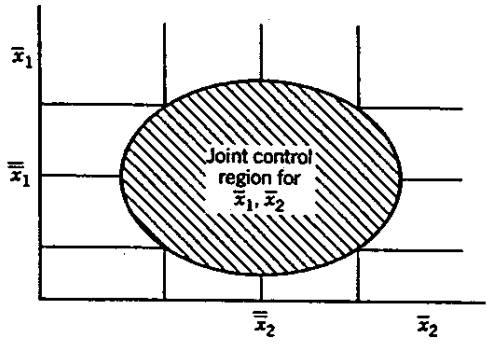
Suppose that two quality characteristics, x_1 and x_2, are jointly distributed according to the bivariate normal distribution. Let $\bar{x}_1$ and $\bar{x}_2$ denote the sample mean values, and s_1 and s_2 represent the standard deviations of the quality characteristics. The covariance between x_1 and x_2 measures the dependence between the two variables and is represented by s_{12}. Assume that s_1, s_2, and s_{12} are estimated from the data, and $\bar{\bar{x}}_1$ and $\bar{\bar{x}}_2$ represent the nominal mean values of the quality characteristics. Then, the following statistic computed from a sample size n using the sample averages, $\bar{x}_1$ and $\bar{x}_2$,:

$$\frac{n}{s_1^2 s_2^2 - s_{12}^2}\left[s_2^2(\bar{x}_1-\bar{\bar{x}}_1)^2 + s_1^2(\bar{x}_2-\bar{\bar{x}}_2)^2 - 2s_{12}(\bar{x}_1-\bar{\bar{x}}_1)(\bar{x}_2-\bar{\bar{x}}_2)\right] \quad (1)$$

is distributed according to a Hotelling's T^2 distribution with 2 and n-1 degrees of freedom. If $T^2 > T^2_{\alpha,2,n-1}$, then at least one of the quality characteristics is out of control, where $T^2_{\alpha,2,n-1}$ is the upper α percentage point of Hotelling's T^2 distribution with 2 and n-1 degrees of freedom.

The control procedure for two variables may be represented graphically. If $\bar{x}_1$ and $\bar{x}_2$ are independent (i.e., $s_{12}=0$), Equation (1) defines an ellipse with the principal axes parallel to the $\bar{x}_1$, $\bar{x}_2$ axes and the center at $(\bar{\bar{x}}_1, \bar{\bar{x}}_2)$, as shown in Figure 1.

Figure 1 - A control ellipse for two independent variables.



Since Equation (1) measures a "distance" from a historical or desired center, a pair of observed sample means $(\bar{x}_1, \bar{x}_2)$ plotting inside the ellipse (i.e., $\leq T^2_{\alpha,2,n-1}$) indicates a state of statistical control. Likewise, a pair of observed means plotting outside the ellipse (i.e., $> T^2_{\alpha,2,n-1}$) indicates the process is out of control. Therefore, the term, control ellipse, is frequently used to describe this region.

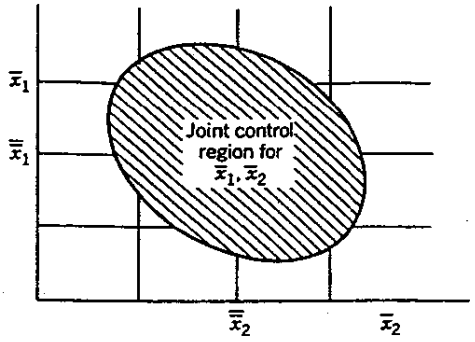
If the two quality characteristics are dependent, then $s_{12} \neq 0$. The corresponding control ellipse is shown in Figure 2. Observe that the principal axes of the ellipse are no longer parallel to the $\bar{x}_1$, $\bar{x}_2$ axes.

However, using the control ellipse results in the following disadvantages:

- not preserving the order of manufacture (i.e., time sequence);
- difficulty in constructing the ellipse for more than two quality characteristics.

Consequently, a technique which maintains the time order of data and may be applied - regardless of the number of characteristics - must be utilized.

Figure 2 - A control ellipse for two dependent variables.



Hotelling T^2 Control Chart

The values of T^2 computed from Equation (1) for each sample may be plotted on a control chart with an upper control limit at $T^2_{\alpha,2,n-1}$ as shown in Figure 3. This control chart may be referred to as either a Hotelling T^2 or a multivariate control chart. Since the manufacturing order is preserved by this chart, runs or other nonrandom patterns of process variation may be investigated. Furthermore, the value of the statistic T^2 characterizes the "state" of the process regardless of how many characteristics are analyzed.

The control chart calculations may be extended to the case where there are p-related quality characteristics to control jointly. The joint probability distribution of the p quality characteristics is assumed as the p-variate normal distribution. The sample mean for each of the p quality characteristics is computed using a sample size n. This set of quality characteristic means is represented by the p x 1 vector: $\bar{\underline{x}} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$. The statistic plotted on the control chart for each sample is:

$$n(\bar{\underline{x}}-\bar{\bar{\underline{x}}})' s^{-1} (\bar{\underline{x}}-\bar{\bar{\underline{x}}}) \qquad (2)$$

such that $\bar{\bar{\underline{x}}} = [\bar{\bar{x}}_1, \bar{\bar{x}}_2, \ldots, \bar{\bar{x}}_p]$ represents the vector of nominal values for each quality characteristic and s is the variance-covariance matrix of the p quality characteristics. The upper control limit (UCL) is established at $T^2_{\alpha,2,n-1}$. For less than 25 samples, T^2 is calculated using the F distribution:

$$T^2_{\alpha,\,p,\,n-1} = \frac{p(n-1)}{n-p} F_{\alpha,\,p,\,n-p} \qquad (3)$$

For larger samples, the chi-square with p degrees of freedom at the specified α may be utilized. However, this approach implies that $\bar{\bar{\underline{x}}}$ and s are known rather than estimated. Usually, $\bar{\bar{\underline{x}}}$ and s are estimated from the analysis of preliminary samples taken when the process is assumed to be in control.

Subgroup Size n=1

In the instance where each data record is considered individually, Equation (2) for the plotted statistic reduces to:

$$(x-\bar{x})' s^{-1} (x-\bar{x}) \qquad (4)$$

since the subgroup size, n, is set equal to 1 and the subgroup averages are no longer computed. Here, the $\bar{x}$ vector represents the overall process average. Also, the value of n used for the upper control limit calculation in Equation (3) equals the total number of plot points (observations) as opposed to the subgroup size.

If the process is in the start-up phase, Tracy (1992) suggests that the F distribution should not be used for control limit calculations with a subgroup size n=1. The goal of this phase is to establish statistical control (i.e., a "clean" process) and find accurate control limits which allow proceeding to the second phase of process monitor and control. For plotting individual measurements in the start-up phase, the beta distribution should be used in control limit calculations:

$$LCL = \frac{(n-1)^2}{n} B_{1-\alpha/2,\ p/2,\ (n-p-1)/2}$$

$$UCL = \frac{(n-1)^2}{n} B_{\alpha/2,\ p/2,\ (n-p-1)/2} \qquad (5)$$

where $B_{\alpha,\ p/2,\ (n-p-1)/2}$ denotes the $1-\alpha$ percentile of the beta distribution with parameters p/2 and (n-p-1)/2.

Since any shift in the mean results in an increase in T^2, most multivariate control charts establish the lower control limit (LCL) at zero. However, abnormally small T^2 values may result from changes in the variance-covariance matrix. Therefore, Tracy (1992) recommends using a nonzero LCL.

Control Limit Selection Based on α

The upper control limit (UCL) should be established such that almost all plotted points (T^2) are

within the control ellipse when the process is in control. The α risk is the probability of a false alarm occurring. This situation arises when a data point plots outside the control limit and only common causes of process variation are present.

The UCL is established such that the α proportion of the selected distribution (chi-square, beta, or F) exceeds the UCL. Assuming a multivariate normal distribution, the UCL and associated false alarm rate ($1/\alpha$) may be calculated based on various values of α and is indicated in Table 1.

Table 1 - UCL and Associated False Alarm Rate with n=135

α	p	UCL*	False Alarm Rate
.005	2	11.1	1/200
.0027	4	17.6	1/370
.05	5	11.8	1/20
.01	5	16.3	1/100

*calculated using Equation (3)

The false alarm rate reported is equivalent to the average run length (ARL). The ARL equals the average number of plotted points observed before a point indicates an out-of-control condition even if assignable causes are not present. For any Shewhart control chart, the ARL equals $1/\alpha$. With control limits established at $\pm 3\sigma$ (α=.0027), the ARL equals 370 when the process is in control. The probability that an individual point falls outside the control limit equals α.

Multivariate vs. Shewhart

One advantage of multivariate charting of related characteristics becomes evident when considering the following scenario. Suppose that a process is in control and stable and four *independent* variables are plotted on separate Shewhart control charts. The control limits are calculated from the data and established at $\pm 3\sigma$ from the means. Since each variable is plotted separately, the combined probability that all four means will plot within the control limits equals $(.9973)^4 = .9892$. The resulting ARL is reduced to 93 which means that, on average, 1 out of 93 points will generate a false alarm. For *dependent* variables, the actual ARL will differ. In contrast, a multivariate chart with α=.0027 results in a false alarm rate of only 1 out of 370 points.

Like the Shewhart control chart, the multivariate control chart effectively detects large process shifts. However, neither chart performs well in detecting small shifts in the average. If this aspect is critical, another technique, such as the cumulative sum (cusum) chart, should be utilized.

Constructing the Multivariate Control Chart in SAS

The multivariate control chart is constructed using SAS/STAT and SAS/QC procedures. The PRINCOMP procedure and DATA step are used to compute T^2 for variables X_1 through X_N:

```
PROC PRINCOMP STD OUT=PC;
   VAR X1-XN;
RUN;

DATA PC (DROP=PRIN1-PRINN);
   SET PC;
   TSQ = USS(OF PRIN1-PRINN);
RUN;
```

Next, a control limits data set is created using the SHEWHART procedure with a subgroup size of one:

```
PROC SHEWHART DATA=PC;
   XCHART TSQ*xvar /
      NOCHART
      LIMITN=1
      OUTLIMITS=CLIM
      OUTINDEX='CURRENT';
RUN;
```

The following Screen Control Language (SCL) variables are computed prior to calculating the UCL:

ALPHA - significance level specified (α)
NCHAR - number of characteristics to chart (p)
NCHAR2 - number of characteristics to chart ÷ 2
NOBS - number of observations (n)

A DATA step is used to calculate the T^2 UCL using Equation (3), set the lower control limit (LCL) equal to zero, and store the alpha level:

```
DATA CLIM;
   SET CLIM;
   _UCLX_=FINV(1-&ALPHA,&NCHAR,
   &NOBS-&NCHAR)*(&NCHAR*(&NOBS
   -1))/(&NOBS-&NCHAR);
   _LCLX_=0;
   _ALPHA_=&ALPHA;
RUN;
```

Note that the FINV function must be called using the *upper* α quantile (i.e., 1-α). Finally, the control chart is constructed using the SHEWHART procedure and READLIMITS option:

```
PROC SHEWHART DATA=PC
LIMITS=CLIM;
   XCHART TSQ*xvar='+' /
      READLIMITS;
   LABEL TSQ='T-SQUARED';
RUN;
```

Once the basic chart is constructed, the output may be enhanced using control chart and/or graphics options.

Analysis Aspects

In the utilization of the multivariate control chart as a process analysis technique, the following aspects must be considered:

- Since the control limit calculation in Equation (3) depends on the number of observations, a fixed or historical control limit should be applied to current data with the same number of observations. However, if the chi-square function is used instead, the same α level should be applied again.

- The data must follow a multivariate normal distribution.

- Multivariate outliers in the data may decrease the range of T^2 values, making it difficult to detect extreme ones.

Assessing Multivariate Normality

If T^2 is assumed to follow a chi-square distribution for larger samples, a simple check of multivariate normality may be performed by constructing a Q-Q plot (Figure 4) in the CAPABILITY procedure of SAS/QC. The chi-square distribution is fit using a gamma distribution with parameters: theta=0, sigma=2, and alpha=p/2, where p represents the number of characteristics. Using our prior data set and variables:

```
PROC CAPABILITY DATA=PC;
   QQPLOT TSQ / GAMMA(THETA=0
   SIGMA=2 ALPHA=&NCHAR2);
RUN;
```

A Q-Q plot of T^2 versus the quantiles of the chi-square distribution should plot as a straight line. For further details, refer to the *SAS System for Statistical Graphics, First Edition.*

Detecting Multivariate Outliers

On the chi-square probability plot, outliers appear as points in the upper right that are substantially above the line for the expected quantiles. Unfortunately, like all least squares techniques, the chi-square plot of the T^2 statistic is not resistant to the effect of outliers. Discrepant values not only affect the mean vector, $\bar{x}$, but also inflate the variance-covariance matrix, s. Thus, the effect of a few unusual observations is spread through all the T^2 values. Furthermore, this tends to decrease the range of T^2 values, making it harder to detect extreme ones (Friendly, p. 451).

A multivariate trimming procedure is applied to calculate the squared distances that are not affected by potential outliers. This is an iterative process where, on each iteration, a proportion of observations with the largest T^2 values are temporarily set aside. Then, a trimmed mean, $\bar{x}_{(-)}$, and trimmed variance-covariance matrix, $s_{(-)}$, are computed from the remaining observations. Thereafter, new T^2 values are computed using the following formula:

$$(x-\bar{x}_{(-)})' s^{-1} (x-\bar{x}_{(-)}) \qquad (6)$$

Using this trimming approach avoids including observations with large T^2 values in calculations for the remaining observations.

To begin the multivariate trimming, the WEIGHT statement is used in the PRINCOMP procedure. Initially, observations are assigned a weight of 1, and the T^2 values are calculated. Then, an observation with a high chi-square quantile is assigned a weight of 0. The process is repeated, typically until no new observations are trimmed or a specified number of iterations has been conducted. This scheme for outlier detection has been implemented in a general SAS macro, OUTLIER, presented in the *SAS System for Statistical Graphics, First Edition.*

The following arguments describe the OUTLIER macro. PVALUE is the probability such that an observation is trimmed when its T^2 has a probability less than PVALUE. The recommended initial PVALUE should equal the alpha level specified for the UCL calculation. The macro produces an output data set (CHIPLOT) containing the variables TSQ and EXPECTED (the chi-square quantile) in addition to the input variables.

```
%macro OUTLIER(
  data=_LAST_,      /* data set to analyze */
  var=_NUMERIC_,    /* input variables */
  id=,              /* ID variable */
  out=CHIPLOT,      /* data set for plots */
  pvalue=&ALPHA,    /* Prob < pvalue -->
                       weight=0 */
  passes=2,         /* number of passes */
  print=YES);       /* Print OUT= data
                       set? */
```

By default, two passes (PASSES=) are made through the iterative procedure.

Applying the macro OUTLIER to the data set with PVALUE equal to .0027 (the α level from the UCL calculation used for Figure 3) results in the output indicated in Table 2. Four observations have significantly large T^2 values in pass 1. These observations exhibit UCL violations on the multivariate control chart. When these data values are assigned a weight of 0 in the next PRINCOMP step, their T^2 values increase. Although the T^2 values for other observations tend to decrease or remain the same, one new observation, AC8652, is added in pass 2. Since this outlier detection scheme initially identifies observations that exhibit control limit violations, it may be implemented to screen the process data prior to constructing a control chart.

To test the hypothesis of multivariate normality, the output data set, CHIPLOT, is used to plot the T^2 values versus their expected chi-square values as shown in Figure 5. Potential outliers identified with the OUTLIER macro are labeled with the ID variable using the annotate facility in SAS/GRAPH™. The assignable cause for these outliers should be identified and removed from the process.

Conclusion

A multivariate control chart of related quality characteristics surpasses standard Shewhart control charts particularly when considering the reduced false alarm rate and number of charts generated. Shewhart techniques fail to consider the relationships between process variables. With multivariate analysis, processes may be monitored effectively and outliers from multivariate normality may be identified.

References

Friendly, Michael. *SAS System for Statistical Graphics, First Edition.* Cary, NC: SAS Institute Inc., 1991, pp. 447-457.

Hotelling, H. *Techniques of Statistical Analysis.* New York: McGraw-Hill, 1947.

Hotelling, H. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, CA: University of California Press, 1951.

Montgomery, Douglas C. *Introduction to Statistical Quality Control.* New York: John Wiley & Sons, 1985.

Tracy, Nola D., et al. "Multivariate Control Charts for Individual Observations". *Journal of Quality Technology.* Vol. 24, No. 2, April 1992, pp. 88-95.

The Author

Melissa A. Durfee
Wyman-Gordon Company
244 Worcester Street Box 8001
N. Grafton, MA 01536-8001
(508) 839-8083

Table 2 MULTIVARIATE OUTLIER DETECTION

OBS	SN	PASS	CASE	TSQ	PROB
1	AC8654	1	89	21.7114	.00022875
2	AC8658	1	93	22.2732	.00017682
3	AC8664	1	99	34.1861	.00000068
4	AC8666	1	101	18.5146	.00097868
5	AC8652	2	87	21.9979	.00020061
6	AC8654	2	89	27.6026	.00001501
7	AC8658	2	93	39.4177	.00000006
8	AC8664	2	99	60.4184	.00000000
9	AC8666	2	101	21.3610	.00026852

MULTIVARIATE CHART: 4 PROPERTIES WG 15909 INT & STAT HOT TENSILE #1200

Yield Stress (in KSI)/Ultimate Strength (in KSI)/% Elongation/% Reduction of Area

PAGE 2 OF 3

CONTROL LIMIT STATUS: CURRENT (0.0027 CONFIDENCE)

08/11/93

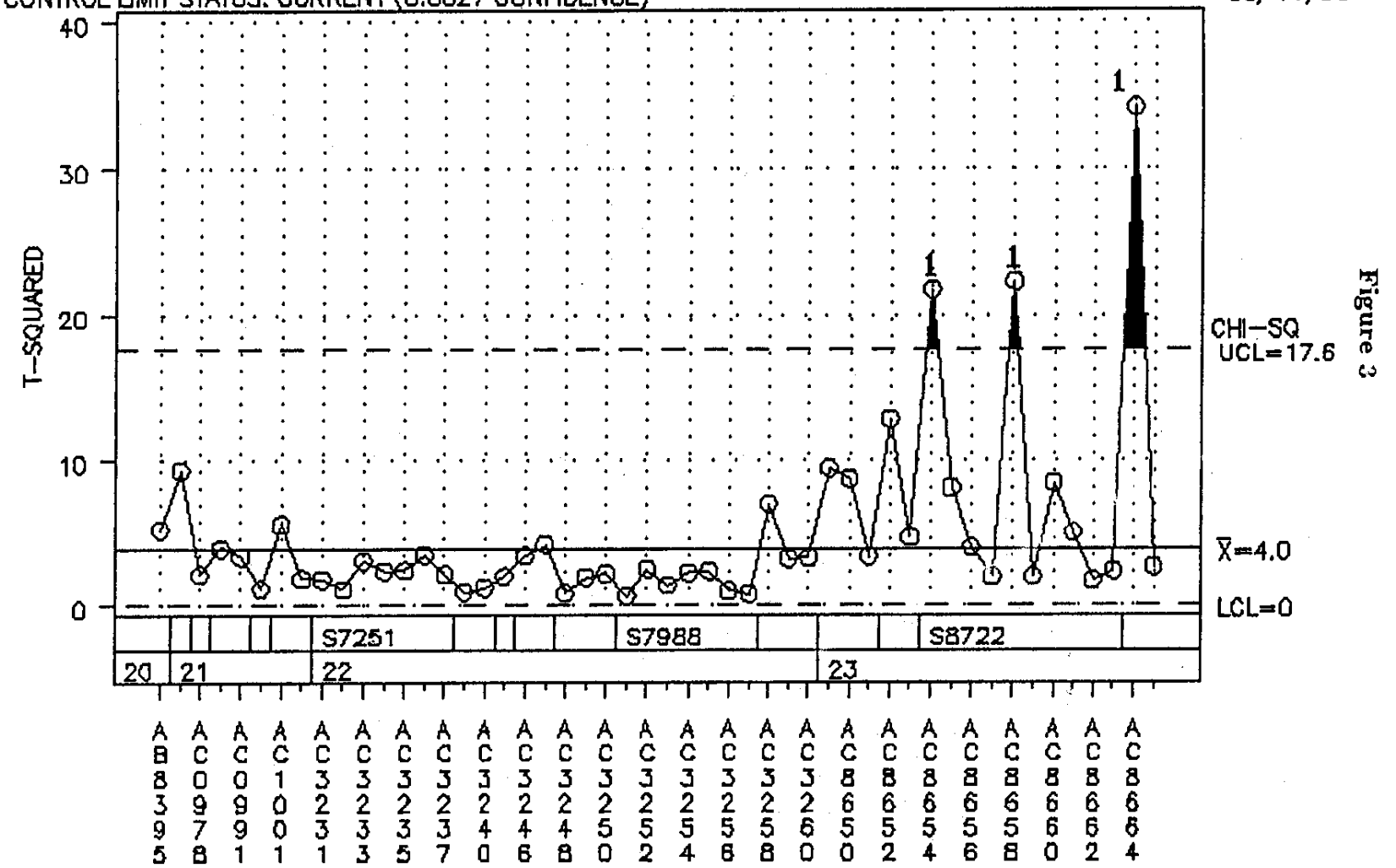


Figure 3

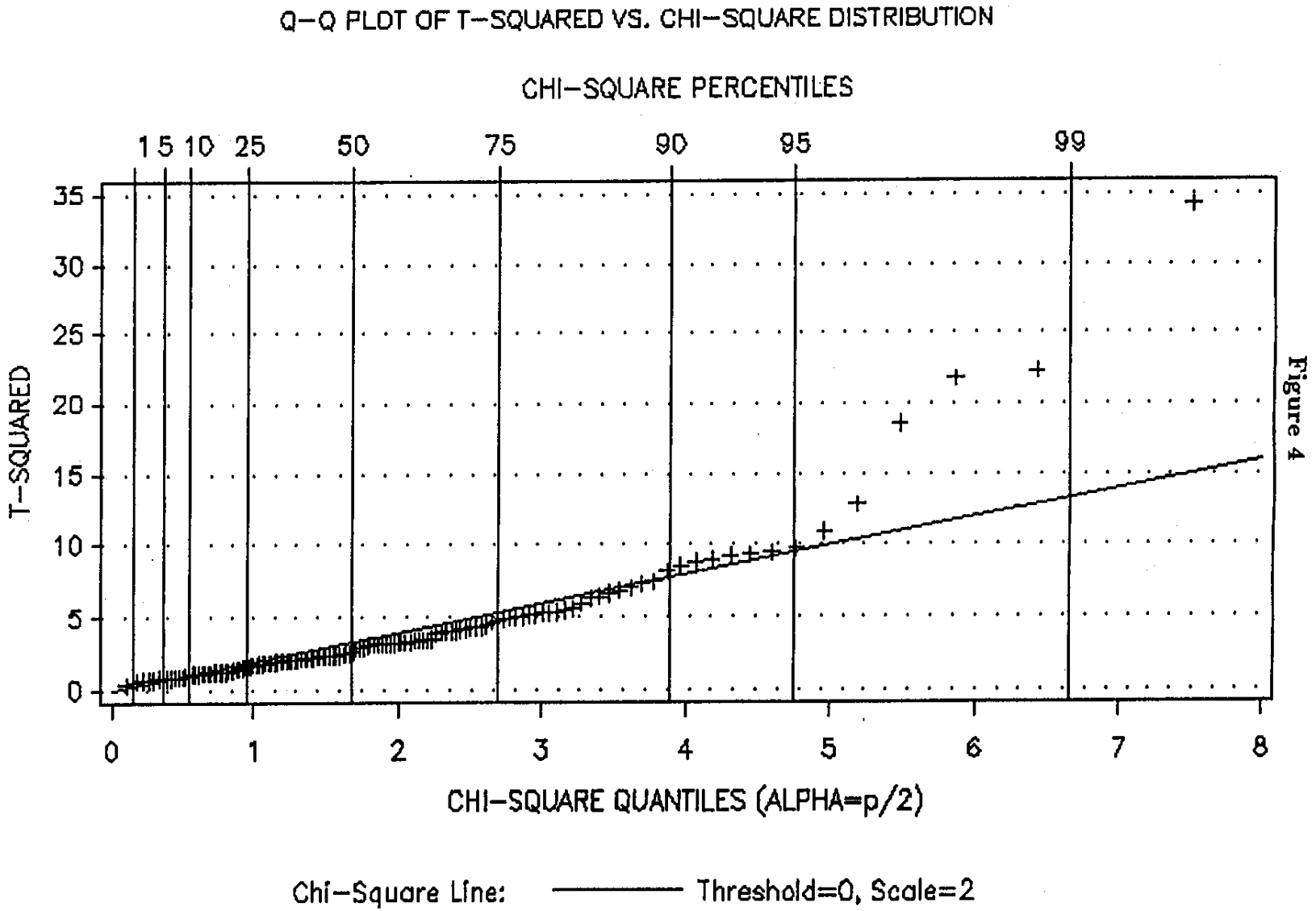

Q–Q PLOT OF T–SQUARED VS. CHI–SQUARE DISTRIBUTION
CHI–SQUARE PERCENTILES
1 5 10 25 50 75 90 95 99
T–SQUARED
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8
CHI–SQUARE QUANTILES (ALPHA=p/2)
Chi–Square Line: Threshold=0, Scale=2

Figure 4

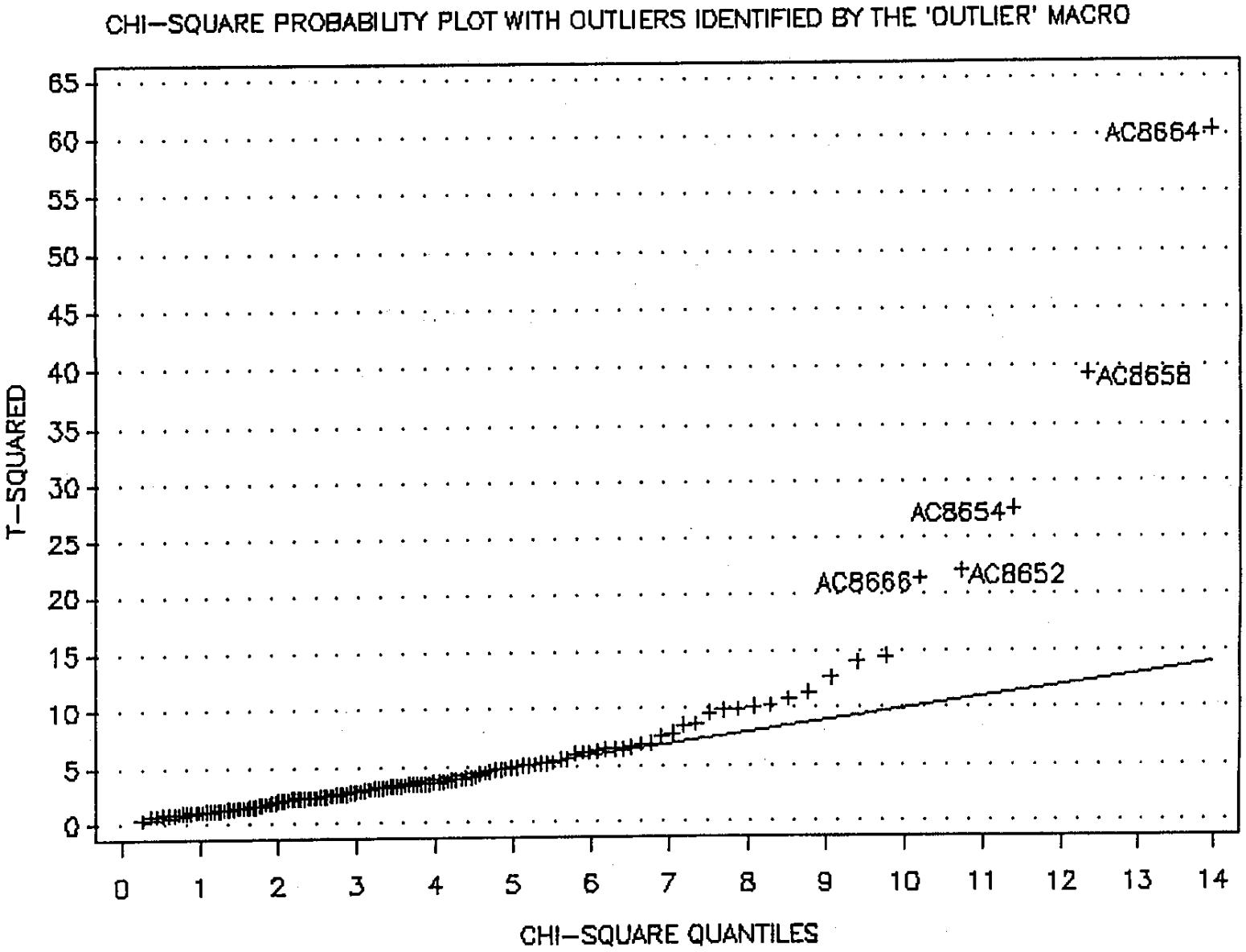

CHI–SQUARE PROBABILITY PLOT WITH OUTLIERS IDENTIFIED BY THE 'OUTLIER' MACRO
T–SQUARED
CHI–SQUARE QUANTILES
AC8664
AC8658
AC8654
AC8666
AC8652
0
5
10
15
20
25
30
35
40
45
50
55
60
65
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Figure 5