A Powerful and Robust Nonparametric Statistic for Joint Mean-Variance Quality Control

J.D. Opdyke*

For statistical process control, a number of single charts that jointly monitor both process mean and variability recently have been developed. For quality control-related hypothesis testing, however, there has been little analogous development of joint mean-variance tests: only one two-sample statistic that is not computationally intensive has been designed specifically for the one-sided test of Ho: $\mu_2 \le \mu_1$ and $\sigma_2 \le \sigma_1$ vs. Ha: $\mu_2 > \mu_1$ OR $\sigma_2 > \sigma_1$ (see Opdyke, 2006). Many have proposed other commonly used tests, such as tests of stochastic dominance, exceedance tests, or permutation tests for this joint hypothesis, but the first can exhibit prohibitively poor type I error control, and the latter two can have virtually no power under many conditions. This paper further develops and generalizes the maximum test proposed in Opdyke (2006) and demonstrates via extensive simulations that, for comparing two independent samples under typical quality control conditions, it a) always maintains good type I error control; b) has good power under symmetry and modest power under notable asymmetry; and c) often has dramatically more power *and* much better type I error control than the only other widely endorsed competitor. The statistic (OBMax2) is not computationally intensive, and although initially designed for quality control testing in regulatory telecommunications, its range of application is as broad as the number of quality control settings requiring a one-sided, joint test of both the mean and the variance.

KEYWORDS: CLEC; Location-Scale; Maximum test; Six sigma; Statistical process control; Telecommunications.

* J.D. Opdyke is Managing Director of Quantitative Strategies at DataMineIt, a statistical data mining consultancy specializing in applied statistical, econometric, and algorithmic solutions for the financial and consulting sectors. Clients include multiple Fortune 50 banks and credit card companies, big 4 and economic consulting firms, venture capital firms, and large marketing and advertising firms. e-mail: JDOpdyke@DataMineIt.com web: http:///www.DataMineIt.com

INTRODUCTION

The statistical process control literature recently has seen the development of a number of single control charts that jointly monitor both process mean and variability (see Gan et al., 2004; Costa & Rahim, 2004; Wu et al., 2005; Hawkins & Zamba, 2005; and Reynolds & Stoumbos, 2005). This is an important development since both the location and the spread of data measuring quality are key characteristics, taken together simultaneously, for assessing, quantifying, and monitoring the degree to which the quality goals for a product or service are achieved. However, quality control-related hypothesis testing has seen few developments analogous to those of the statistical process control literature. Only two statistics known to this author (Opdyke, 2006, and Pesarin, 2001, p.325) have been developed specifically to test the one-sided, joint mean-variance hypotheses of:

$$\text{Ho: } \mu_2 \le \mu_1 \text{ and } \sigma_2 \le \sigma_1 \text{ vs. Ha: } \mu_2 > \mu_1 \textbf{ OR } \sigma_2 > \sigma_1 \tag{1}$$

This is simply a test of whether the mean and/or the variance of one population are larger than those of the other population. This is a very simple and important joint hypothesis from a quality control perspective, and yet no easily implemented statistical tests that are robust under real-world data conditions exist to test it. Although many have proposed the use of tests of stochastic dominance when confronted with (1), such as Kolmogorov-Smirnov or similar statistics,[1] these tests, while often powerful, can exhibit prohibitively poor type I error control under the null hypothesis of (1). In other words, while often providing high rates of true positives, these tests also can provide prohibitively high rates of false positives well beyond the acceptable rate (α) set by the researcher (as demonstrated in the simulation study below). Conversely, other tests that have been proposed, such as exceedance tests (e.g. Rosenbaum, 1954) and tests of distributional equality (e.g. permutation tests other than that of Pesarin, 2001, p.325) maintain good type I error control, but can have virtually no power to detect some effects under the alternate hypothesis of (1) (as shown in the simulation study below for Rosenbaum, 1954). The reason that none of these tests work well for (1) is that none are designed specifically for (1): they all are designed to varying degrees to detect distributional characteristics beyond the first two moments[2] – i.e. beyond the mean and the variance.[3] However, for many quality control problems, if the mean and the variance of two samples are essentially equal, then higher moments, such as the kurtosis,[4] are often of far less concern. For example, if two groups of customers are mandated to receive equal quality service, a difference in the kurtosis between the two groups' time-to-service – *if the means and variances are equal* – arguably has a very second-order effect, if any, on the perceived "quality" of service they receive. The tests listed above, however, will either sound false alarms when means and variances are equal but higher moments differ, or fail to detect different means or variances because of a more general statistical design meant to identify differences in higher moments as well. While kurtosis, for example, can be an important characteristic for some types of quality control issues, such as statistics that identify

[1] The one-sided test of whether one sample is "stochastically larger" than another tests the following hypotheses:
Ho: $F_Y(x) = F_X(x)$ for all x, vs. Ha: $F_Y(x) \ge F_X(x)$ for all x and $F_Y(x) > F_X(x)$ for some x (see Gibbons & Chakraborti, 2003, p.232-246). Of course, this is different from the hypotheses of (1), but because tests of stochastic dominance have been proposed for usage with (1), their performance under (1) is examined in this paper.

[2] The term "moment," for all moments higher than the mean, shall herein refer to "moment about the mean."

[3] The mean and the variance, of course, are not the only measures of location and scale, respectively, but they often are the most appropriate for statistical reasons and the most widely used.

[4] While the variance of a distribution measures the degree of dispersion around the mean, the kurtosis measures the degree of dispersion around the "shoulders" – the points one standard deviation on either side of the mean. For many distributions this is very similar to measuring the thickness of the tails of the distribution, and/or the peakedness of its center, which is how most people conceptualize the kurtosis.

individual or small clusters of data outliers, when addressing general issues of quality the mean and the variance typically are the primary concern. The customers in the above example would be most concerned with a slower time-to-service *on average*, and/or a larger variability in their time-to-service, than with a time-to-service distributional peak that was somewhat taller or shorter, *all else equal*. And the former is exactly what is tested by the one-sided, two-sample statistic developed in this study: whether the mean and/or variance of a quality metric of one population or process are larger than those of another similar process.

PREVIOUS AND RELATED WORK

A number of statistics have been developed for the two-sided, location-scale hypotheses of

$$\text{Ho: } F(x) = G(x) \text{ vs. Ha: } F(x) = G\left(\frac{x-\mu}{\sigma}\right), \text{ with } \sigma > 0, \text{ and } \mu \neq 0 \text{ and/or } \sigma \neq 1 \qquad (2)$$

(see O'Brien, 1988; Podgor & Gastwirth, 1994; Buning & Thadewald, 2000; and Manly & Francis, 2002). But from a quality perspective we are more concerned with testing the one-sided hypotheses presented in (1) because the focus is on whether the quality of one population or process is *worse than* (better than) that of the other, not just different from that of the other. One statistic has received widespread attention as a test of (1) in the regulatory telecommunications arena. Seven years' worth of expert testimony, as well as multiple Rulings, Opinions, and Orders handed down by various state and federal regulatory bodies, have supported use of the 'modified' *t* statistic (3) (Brownie et al., 1990) to compare the quality of service provided to two groups of telecommunications customers – competing local exchange carrier (CLEC) customers and incumbent local exchange carrier (ILEC) customers.

$$t_{\text{mod}} = \frac{\left(\bar{X}_2 - \bar{X}_1\right) - \left(\mu_2 - \mu_1\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} \quad \text{with } df = n_1 - 1 \qquad (3)$$

The point of the test is to ensure that the service quality received by CLEC customers is "at least equal" to that received by ILEC customers (see Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996), at S251 (c) (2) (C)), which is necessary to ensure that ILEC customers could and would actually switch to CLEC customers, and that a formerly regulated industry can effectively transition to a fully competitive economic market. The 'modified' *t* statistic will be recognized as the widely used separate-variance *t* statistic (see Appendix) with a slight modification made to the denominator: the study group (subscript 2) variance simply is replaced with the control group (subscript 1) variance.

However, Opdyke (2004) demonstrated, via both analytic derivation and extensive simulation, that several crucial assumptions made about this statistic are false, making it inappropriate as a test of (1) in any setting. Its asymptotic distribution was shown *not* to be standard normal as previously surmised in submitted expert testimony, but rather, to be normal with a variance that is greater than, less than, or equal to unity depending on the relative sizes of the two population variances, as shown below.

$$t_{\text{mod}} \sim N\left(0,\ \left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)\Big/\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_2}\right)\right) \tag{4}$$

A consequence of this when using standard normal (or student's *t*) critical values, as advised in extensive expert testimony and Brownie et al. (1990), is that it allows a "trade-off" in average service for variability in service, which violates the null hypothesis of (1) with literally *zero* power to detect these violations. This is shown very clearly in Figures 1a and 1b for (very) unbalanced sample sizes, and Figures 2a and 2b for equal sample sizes. Normalizing the 'modified' *t* in an attempt to take care of this problem yields, not surprisingly, the familiar separate-variance *t* statistic.

In addition to this fatal flaw, another problem with using this statistic as a test of (1) is that it has virtually no power to detect differences in variances. For example, under equal means and a study group variance *twice as large* as that of the control group, the asymptotic power of the 'modified' *t*, for $\alpha = 0.05$, is only 0.09 for equal sample sizes, and only 0.12 for very unbalanced ($n_1 / n_2 = 100$) sample sizes, as shown in Figures 3a and 3b, respectively. Although Brownie et al. (1990) originally proposed the 'modified' *t* for use with a different pair of joint hypotheses (5) for which the statistic usually (but not always) has more power than the pooled- and separate-variance *t* tests, it nonetheless remains essentially useless as a test of (1) based on the above findings, extensive expert testimony notwithstanding (see Opdyke, 2004, for extensive citations of expert testimony in regulatory rulings and opinions).

$$\text{Ho: } \mu_2 \le \mu_1 \ \text{ and } \ \sigma_2 \le \sigma_1 \ \text{ vs. Ha: } \mu_2 > \mu_1 \ \textbf{AND} \ \sigma_2 > \sigma_1 \tag{5}$$

As an alternative to the 'modified' *t* statistic, Opdyke (2004) proposed the collective use of several easily implemented conditional statistical procedures. Four tests are proposed by combining O'Brien's (1988) generalized *t* test (OBt) or his generalized rank sum test (OBr) with either of two straightforward tests of variances – Shoemaker's (2003) F_1 test and the 'modified' Levene test (Brown & Forsythe, 1974), which is simply the well-known ANOVA test (see Appendix for corresponding formulae). These easily calculated statistics are combined based on the relative size of the two sample means, as shown in the column headings of Table 1.

Table 1. Conditional Statistical Procedures of Opdyke (2004)

Conditional statistical procedure	if $\bar{X}_2 > \bar{X}_1$, use…	If $\bar{X}_2 \le \bar{X}_1$ or OB fails to reject Ho:, use…
OBtShoe	OBt	Shoemaker's F_1
OBtLev	OBt	'modified' Levene
OBrShoe	OBr	Shoemaker's F_1
OBrLev	OBr	'modified' Levene

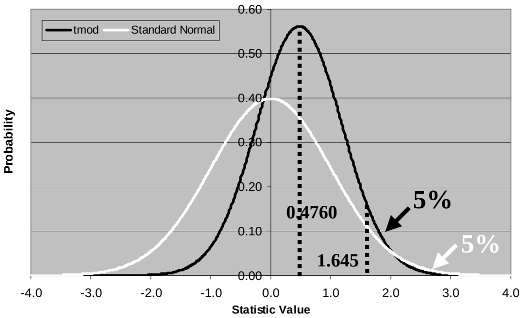

Figure 1a. t_{mod} v. Standard Normal, $\mu_2 > \mu_1$, $\sigma_2/\sigma_1 = 0.5$, $n_1/n_2 = 100$

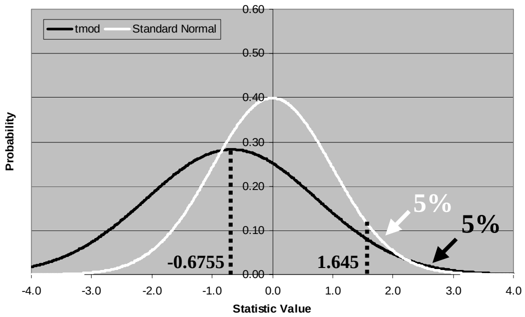

Figure 1b. t_{mod} v. Standard Normal, $\mu_2 < \mu_1$, $\sigma_2/\sigma_1 = 2$, $n_1/n_2 = 100$

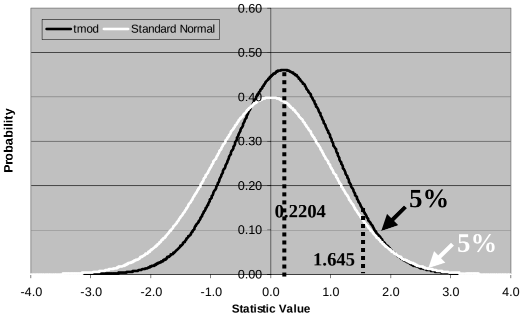

Figure 2a. t_{mod} v. Standard Normal, $\mu_2 > \mu_1$, $\sigma_2/\sigma_1 = 0.5$, $n_1/n_2 = 1$

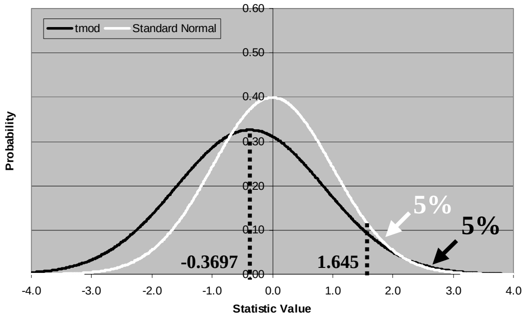

Figure 2b. t_{mod} v. Standard Normal, $\mu_2 < \mu_1$, $\sigma_2/\sigma_1 = 2$, $n_1/n_2 = 1$

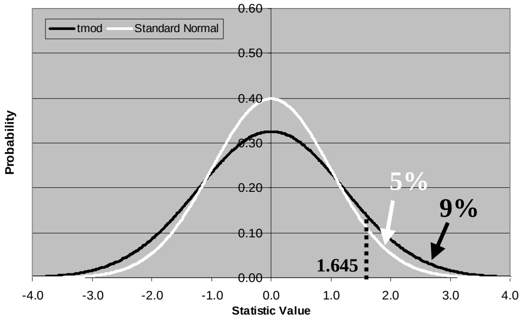

Figure 3a. t_{mod} v. Standard Normal, $\mu_2 = \mu_1$, $\sigma_2/\sigma_1 = 2$, $n_1/n_2 = 1$

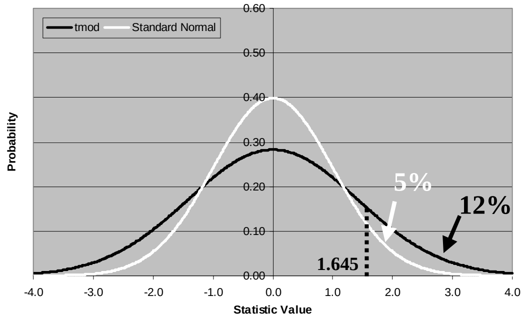

Figure 3a. t_{mod} v. Standard Normal, $\mu_2 = \mu_1$, $\sigma_2/\sigma_1 = 2$, $n_1/n_2 = 100$

Table 2. Implementation of Table 1 Procedures Under Symmetry

	Kurtosis of Distribution	
Sample Sizes	platy- to mesokurtotic (OBt)	leptokurtotic (OBr)
Balanced (Shoemaker's F_1)	OBtShoe	OBrShoe
Unbalanced ('modified' Levene)	OBtLev	OBrLev

For symmetric data, the choice of which of these four tests to use is based on two criteria – whether the data is at least as short-tailed as the normal distribution (platy- to mesokurtotic) vs. long-tailed (leptokurtotic), and whether sample sizes are balanced (or close) vs. at least moderately unbalanced, as shown in Table 2.[5] However, implementing Table 2 by deciding, for example, how unbalanced long-tailed samples must be before using OBrLev rather than OBrShoe requires additional simulations not performed in Opdyke (2004). Subsequently, Opdyke (2006) bypassed this requirement,

[5] Aglina, Olejnik, and Ocanto (1989) make similar distinctions when proposing conditions for the alternate use of two of the scale tests (O'Brien's test (1988) and the Brown-Forsythe test (1974)) used herein to develop the test statistic proposed herein (OBMax2).

combining the Table 1 statistics using a maximum-test approach.

"Maximum tests" – statistics whose scores (*p*-values) are the maximum (minimum) of two or more other statistics – have been devised and studied in a number of settings in the statistics literature with some very favorable results. Neuhäuser et al. (2004) compare a maximum test for the non-parametric two-sample location problem to multiple adaptive tests, finding the former to be most powerful under the widest range of data conditions. Algina, Blair, and Coombs (1995) propose a maximum test for testing variability in the two-sample case,[6] and Blair (2002) constructs a maximum test of location that is shown to be only slightly less powerful than each of its constituent tests under their respective "ideal" data conditions, but notably more powerful than each under their respective "non-ideal" data conditions. These findings demonstrate the general purpose of maximum tests – to often, but not always, trade-off minor power losses under ideal or known data conditions for a more robust statistic with larger power gains across a wider range of possible (and usually unknown) data distributions.

To construct a maximum test for the joint mean-variance hypotheses of (1), it must be recognized that maximum tests are conditional statistical procedures, and the additional variance introduced by such conditioning will inflate the test's size over that of its constituent statistics (and if left unadjusted, probably over the nominal level of the test as shown in Blair, 2002). But the constituent statistics in Table 1 are already conditional statistical procedures, so the *p*-value adjustment used to maintain validity must be large enough to take this "double conditioning" into account (this actually is "triple conditioning" since O'Brien's tests themselves are conditional statistical procedures). The adjustment used in Opdyke (2006) is simply a multiplication of the *p*-values by constant factors (β's), the values of which were determined based on extensive simulations across many distributions. The *p*-value of the maximum test – OBMax – is defined below:

$$p_{OBMax} = \min \begin{pmatrix} p_{OBtShoe} \cdot \beta_{OBtShoe}, \\ p_{OBtLev} \cdot \beta_{OBtLev}, \\ p_{OBrShoe} \cdot \beta_{OBrShoe}, \\ p_{OBrLev} \cdot \beta_{OBrLev}, \\ p_{tsv} \cdot \beta_{tsv}, \\ 1.0 \end{pmatrix} \tag{6}$$

where $\beta_{OBtShoe} = \beta_{OBtLev} = \beta_{OBGShoe} = \beta_{OBGLev} = 2.8$, and $\beta_{tsv} = 1.8$, and p_{tsv} is the *p*-value corresponding to the separate-variance *t* test with Satterthwaite's (1946) degrees of freedom (see Appendix for corresponding formulae).

[6] Algina et al. (1995) report very good power for their test, which combines two of the variance tests used below in the statistic developed herein (OBMax2), but Ramsey & Ramsey (2007) report it is not robust when the nominal level is small (α = 0.01).

While analytic derivation of the asymptotic distribution of OBMax would be preferable to reliance on the simulation-based β's, Yang et al. (2005) show that such derivations for maximum tests are non-trivial, even under much stronger distributional assumptions than can be made with the conditional statistical procedures of Table 1. Babu and Padmanabhan (1996) describe the exact null distribution of their omnibus maximum test as "intractable" and rely on thorough simulation to demonstrate the validity and power of their statistic. Opdyke (2006) takes a similar approach to demonstrate the dramatically greater power of OBMax over the 'modified' t under most alternate hypothesis configurations of (1). However, OBMax has two limitations: it can violate the nominal level (α, the type I error control) when i) $n_2 > n_1$, as well as when, under asymmetry, ii) $\sigma_2 < \sigma_1$ and at least moderately large $n_2 \approx n_1$ (the former condition was not a problem in the setting for which OBMax originally was developed – the regulatory telecommunications arena – since $n_{\text{ILEC}} \geq n_{\text{CLEC}}$ virtually always). This paper eliminates these drawbacks with the development of a more robust statistic – OBMax2 – which maintains validity under asymmetry and any combination of sample sizes, with little loss of power relative to OBMax.

METHODOLOGY

Development of OBMax2

If not stylized for specific asymmetric distributions, most two-sample statistics lose power under asymmetric data, and the constituent tests of OBMax are no exception to this general rule. However, under certain conditions under asymmetry, OBMax fails to maintain validity: if sample sizes are large and equal (or close) and the study group variance is much *smaller* than the control group variance, OBMax (under asymmetry) will often violate the nominal level of the test. This is due to O'Brien's rank sum test (OBr) behaving badly under these conditions – surprisingly, skewed-tail outliers invalidate the Table 1 statistics that use this test under these specific conditions. Although data transformations toward symmetry can alleviate this problem to some degree, there is no guarantee this will fix the problem altogether, if much at all. Instead, Opdyke (2006) proposes the use of another maximum test – OBMax3 – if symmetry cannot be assured. OBMax3 uses only three constituent tests, eliminating the two that use O'Brien's rank sum tests, as shown below:

$$p_{OBMax3} = \min\begin{pmatrix} p_{OBtLev} \cdot \beta_{OBtLev} \;, \\ p_{OBtShoe} \cdot \beta_{OBtShoe} \;, \\ p_{tsv} \quad \cdot \beta_{tsv} \;, \\ 1.0 \end{pmatrix} \tag{7}$$

$$\text{where } \beta_{OBtLev} = \beta_{OBtShoe} = 3.0, \text{ and } \beta_{tsv} = 1.6$$

OBMax3 maintains validity under both symmetric and asymmetric data, and under symmetry the largest power losses it suffers relative to OBMax are well under 0.10 (see Opdyke, 2006). However, it unarguably would be preferable to have, rather than two tests, a single test robust to departures from symmetry that also retains most of the power of OBMax. And that is what OBMax2 accomplishes, as defined in (8) below:

$$p_{OBMax2} = p_{OBMax3} \text{ if and only if} \tag{8}$$

a) $s_2^2 \le s_1^2$ *and*

b) $\bar{X}_2 \le \left(\bar{X}_1 + 0.5 s_1\right)$ *and*

c) the null hypothesis of symmetry is rejected for either sample by the test of D'Agostino et al. (1990) at $\alpha = 0.01$ (see Appendix)

$$p_{OBMax2} = p_{OBMax} \quad \textit{otherwise}$$

This conditioning on a), b) and c) in (8) causes minor power losses in OBMax2 ("2" for two maximum tests) compared to OBMax under symmetry, but the worst level violations, even under asymmetry, are small – far smaller than those of the 'modified' *t* and separate-variance *t* statistics, which is a very important finding. Before discussing the simulation study, however, one other adjustment to OBMax2 is presented below.

In the regulatory telecommunications arena for which OBMax originally was developed, the size of the ILEC customer sample (the "control group," subscript 1) almost always dwarfs that of the CLEC customer sample (the "study group," subscript 2), so the behavior of OBMax under $n_2 > n_1$ was not a concern. The present development of OBMax2, however, seeks to generalize its use under the widest range of possible conditions, making it robust and powerful not only under both symmetry and asymmetry, but also under all possible combinations of sample sizes. Since it turns out that, under $n_2 > n_1$, increased variation of OBMax's (and OBMax3's) constituent statistics causes its violation of the nominal level of the test, an additional adjustment is required when $(n_2 / n_1) > 1$ for OBMax2 to maintain validity. This is accomplished simply by increasing the size of the β adjustments as a function of the sample size ratio:

$$\beta_X = \beta_X + \min\left[2.5,\ \max\left(0,\ \log_e\left[n_2/n_1\right]\right)\right] \tag{9}$$

The maximum function in (9) ensures that the β's are increased only if $(n_2 / n_1) > 1$, and the minimum function ensures that the largest adjustment is +2.5, which was shown in simulations of up to $(n_2 / n_1) = (3{,}000 / 30) = 100$ to be adequate. The empirical level and power of OBMax2, as defined in (8) together with (9), are presented in the simulation study results below.[7]

[7] It is important to note that when implementing OBMax2, O'Brien's tests are referenced to the *F* distribution, rather than Blair's (1991) size-correcting critical values, even though doing so would normally violate the nominal level of the test under some conditions, because the *p*-value β adjustments used here explicitly take this size inflation into account, as described above.

Simulation Study

Under a wide range of data distributions, sample size and mean-variance configurations, this study examines the empirical level and power of seven statistics:

(a) OBMax2, as defined in (8) and (9);

(b) OBMax, as defined in (6);

(c) OBMax3, as defined in (7);

(d) the 'modified' *t* statistic (t_{mod}), as defined in (3);

(e) the separate-variance *t* statistic (t_{sv}) with Satterthwaite's (1946) degrees of freedom (see Appendix), to provide a well-known basis for comparison;

(f) Rosenbaum's (1954) exceedance test (Ros), which counts the number of observations in one sample beyond the maximum of the other as a test of Ho: $F(x) = G(x)$ against the general shift alternative;

(g) and the (one-sided) Kolmogorov-Smirnov statistic (K-S) (using Goodman's, 1954, Chi-square approximation – see Siegel & Castellan, 1988, p.148), a widely used test of stochastic dominance whose basic structure, which relies on the difference between the samples' cumulative distribution functions, underlies many such tests.

Although not designed specifically for (1), (f) Ros and (g) K-S are included here because experts have proposed turning to these and similar tests, as well as tests of distributional equality (like permutation tests), when confronted with (1), and it is important to study their behavior under carefully controlled simulations (for example, K-S has been described as being "able to detect not only differences in average but differences in dispersion between the two samples as well." (see Matlack, 1980, p. 359), which easily could be misinterpreted as an endorsement of this statistic for testing (1)).

The simulation study data was generated from highly disparate distributions, including the normal, uniform, double exponential, exponential, and lognormal distributions, for five different pairs of sample sizes ($n_2 = n_1 = 30, 90, \& 300$; $n_2 = 30 \& n_1 = 300$; and $n_2 = 300 \& n_1 = 30$), seven different variance ratios ($[\sigma_2^2/\sigma_1^2] = 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00$) and seven different location shifts ($\mu_2 = \mu_1 - 2\sigma_1$, $\mu_1 - \sigma_1$, $\mu_1 - 0.5\sigma_1$, μ_1, $\mu_1 + 0.5\sigma_1$, $\mu_1 + \sigma_1$, $\mu_1 + 2\sigma_1$), making 1,225 scenarios. N = 10,000 simulations were run for each scenario.

The normal distribution was chosen as a universal basis for comparison; the uniform and double exponential distributions were chosen as examples of very short- and very long-tailed distributions, respectively, to examine the possible effects of kurtosis on the tests; and the exponential and lognormal distributions were chosen to examine the possible effects of extreme skewness on the tests. The simulations do not include cross-distributional comparisons: the two data samples always are generated from the same distribution. While this assumption of same (or very similar) distributions commonly is made by researchers when using two-sample tests, it arguably has stronger justification in the quality control setting since the comparison is of the *quality* of two otherwise similar or identical processes or populations, not of two processes or populations that are potentially completely dissimilar by any of numerous criteria. In the latter case, other statistical tests, or in-depth distributional examinations, are more appropriate.

Sample sizes were chosen to cover a range considered to be fairly small to moderately large, as well as balanced to quite unbalanced, and two common nominal levels were used: $\alpha = 0.05$ and 0.10, bringing the total number of scenarios to 2,450 (600 null hypotheses, and 1,850 alternate hypotheses). [8]

RESULTS

The rejection rates under the null and alternate hypotheses of (1) – empirical level and power, respectively – for OBMax2 are shown in Figure 4 for $\alpha = 0.05$ (results for $\alpha = 0.10$ are similar; complete study results are available from the author upon request). General patterns in power can be observed. Not surprisingly, OBMax2 has more power for detecting mean increases (under equal variances) than for detecting variance increases (under equal means). Also as expected, power increases as sample sizes increase, with slightly greater power under $n_2 = 300$ & $n_1 = 30$ compared to $n_2 = 30$ & $n_1 = 300$ for $\sigma_2 > \sigma_1$, but often vice versa for $\mu_2 > \mu_1$. Power is greatest under short-tailed (uniform) data, decreasing steadily for longer-tailed data as kurtosis increases (to normal and then double exponential data). Power is lowest under skewed data, and the greater the asymmetry, the lower the power (the lognormal samples are more skewed with these mean-variance configurations than the exponential samples).

Figure 5 graphs a histogram of the differences in power between OBMax2 and the 'modified' *t* (1,106 of the 1,850 simulations had a non-zero power difference; the large mass point of zero difference is excluded from all histograms). OBMax2 completely dominates the 'modified' *t* whenever the study group variance is larger and the study group mean is equal or smaller (which is most of the 79% of alternate hypothesis cases where there is a difference in power). This is simply a demonstration of the modified *t*'s inability to detect differences in variances, as seen above in Figures 3a and 3b. However, when the study group mean is slightly larger – i.e. under slight location shifts – the 'modified' *t* does have a slight power advantage over OBMax2 (most of the 21% of alternate hypothesis cases where there is a difference in power). This is the price paid by OBMax2 for its ability to serve simultaneously as a test of differences in either means and/or variances. However, this modest power advantage disappears as samples increase to moderate sizes, as seen in Figures 6-8.

[8] A permutation test was not included among the competing statistics in this simulation study because, for N=10,000 simulations per scenario, 2,450 scenarios, and assuming only half a second of runtime to conduct each computationally intensive permutation test (a very conservative estimate for the larger samples in the simulation study), the number of runtime seconds = 2,450 x 10,000 x 0.5 = 12,250,000 = 141.8 days. However, what can be said about OBMax2 relative to permutation tests, Persarin's (2001) in particular, is that all else equal, the former would be preferred because it is not computationally intensive.

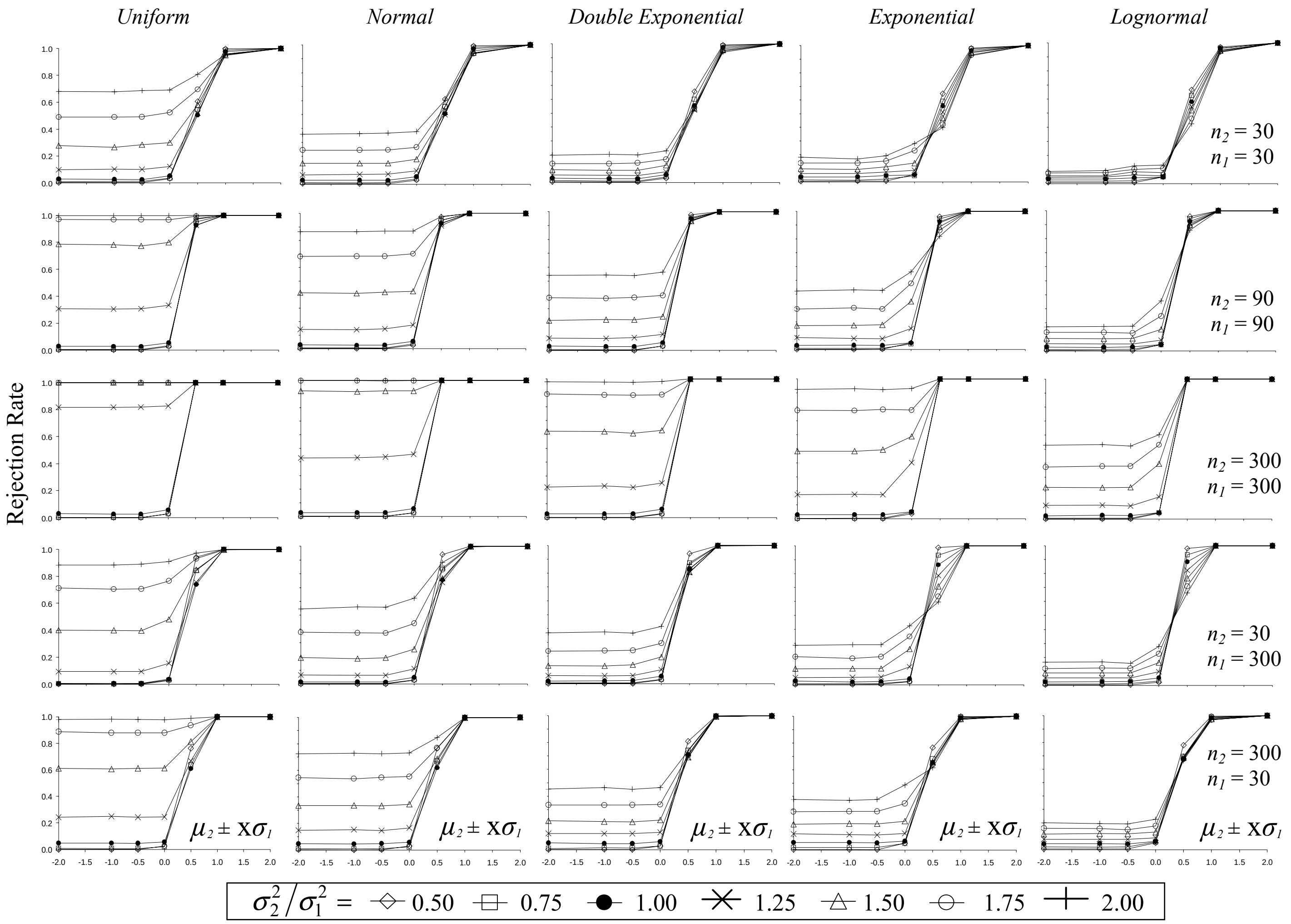

Figure 4. OBMax2 Rejection Rate (Empirical Level & Power) by Distribution by Sample Size by Mean Shift by Variance Ratio (α = 0.05)

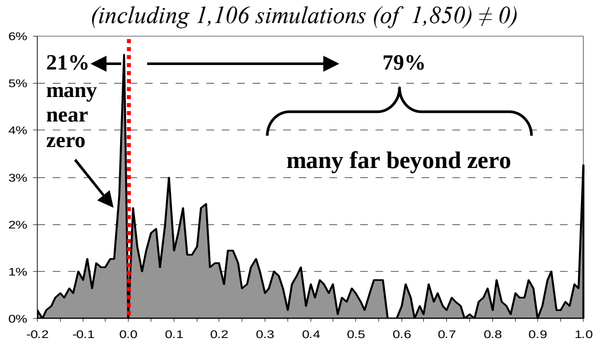

Figure 5. OBMax2 power minus t_{mod} power, all sample sizes

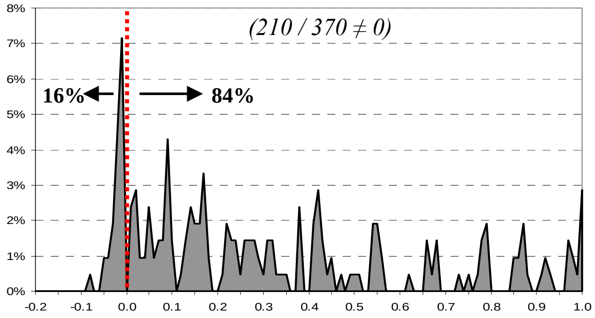

Figure 7. OBMax2 power minus t_{mod} power, $n_1 = n_2 = 90$

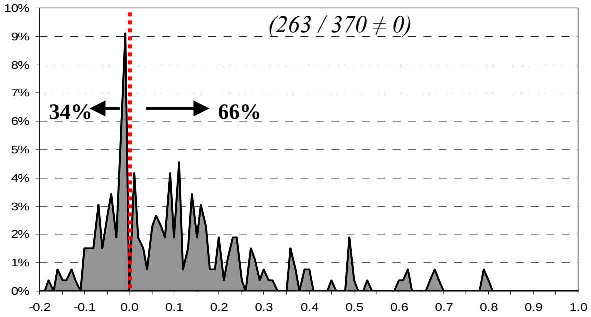

Figure 6. OBMax2 power minus t_{mod} power, $n_1 = n_2 = 30$

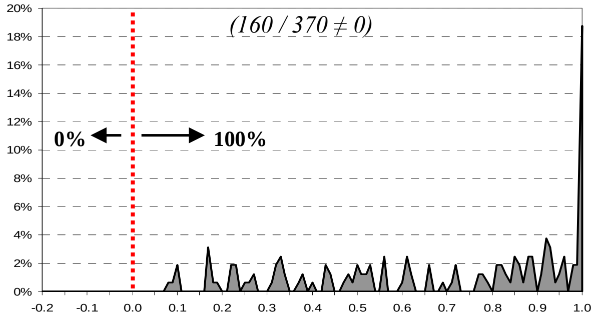

Figure 8. OBMax2 power minus t_{mod} power, $n_1 = n_2 = 300$

When there is a difference in power between OBMax2 and OBMax3, the former almost always wins, with the largest power gains under asymmetry and $\sigma_2 > \sigma_1$ (0.23 and 0.18 for α = 0.05 and 0.10, respectively). OBMax3 only has more power when $n_2 > n_1$, but then it often substantially violates the nominal level. Compared to OBMax, the largest power loss of OBMax2 approaches 0.13 under symmetry, which is the price paid for OBMax2's very good level control, even under asymmetry, as seen in Figure 9 compared to the 'modified' *t*. Both the 'modified' *t* and separate-variance *t* statistics often substantially violate the nominal level of the test under skewed data (this is a well known result for the one-sample *t*-test – see Chaffin & Rhiel, 1993, Boos & Hughes-Oliver, 2000, and Zhou & Gao, 2000). The worst level violations of OBMax2, on the other hand, are what most researchers would consider reasonable, if not quite good: between 0.06 and 0.07 when α = 0.05, and just over 0.11 when α = 0.10. Regarding the other tests, although the Rosenbaum exceedance statistic always maintains validity, it often has dramatically less power than OBMax2, especially if the study group mean is smaller than the control group mean, when it often has absolutely no power to detect a larger study group variance (which actually is consistent with its design). This latter finding also is true of the K-S statistic which, although sometimes more powerful than OBMax2 under small location shifts, often severely violates the nominal level when means are identical but the study group variance is *smaller*. This is because the smaller variance causes the distance between the empirical cumulative distribution functions of the two samples, which determines the significance of K-S, to be sizeable on either side of the common mean (see Figures 10 and 11). This causes high rejection rates for one-sided tests simultaneously in *both* directions. In this study, for example, when $\mu_1 = \mu_2$, $\sigma_2 / \sigma_1 = 0.50$, and $n_1 = n_2 = 300$, the K-S rejection rate, when α = 0.05, is over 0.95 for the uniform, exponential, and lognormal distributions, and very large for the double exponential (0.35) and normal (0.57) distributions as well. This makes it and similarly structured tests of stochastic dominance that rely on the difference between cumulative distribution functions unusable for the joint hypotheses of (1).

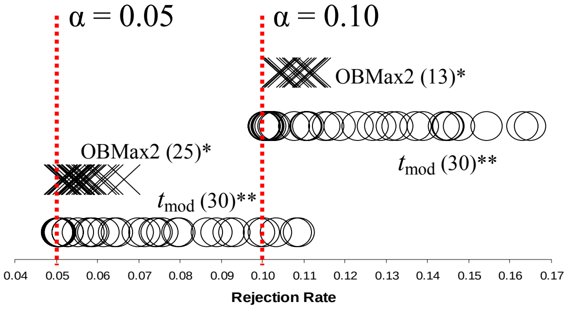

Figure 9. Level Violations: t_{mod} vs. OBMax2 (600 null hypotheses)

NOTE: *15/38 from $n_2 = 300$ & $n_1 = 30$, **47/60 from skewed data

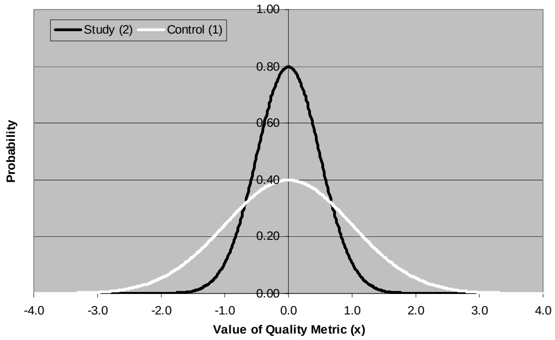

Figure 10. PDFs: K-S Violates Nominal Level when $\mu_1=\mu_2$, $\sigma_2<\sigma_1$

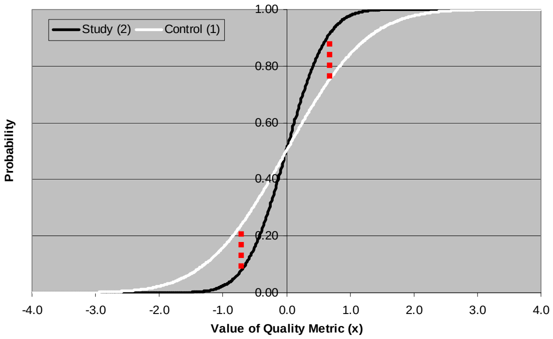

Figure 11. CDFs: K-S Violates Nominal Level when $\mu_1=\mu_2$, $\sigma_2<\sigma_1$

DISCUSSION AND CONCLUSIONS

In quality control settings requiring a one-sided test of the joint mean-variance hypotheses of (1), many researchers and investigators to date have suggested using inappropriate statistics. Although supported by extensive expert testimony, the ‘modified’ *t* was shown here and in Opdyke (2004, 2006) to be completely powerless under many, if not the majority, of alternate hypothesis scenarios under (1). Tests of stochastic dominance, too, are inappropriate. Arguably the most commonly used of these – the Kolmogorov-Smirnov statistic – was shown in this simulation study not only to have no power to detect a larger study group variance when the study group mean is much smaller (which is consistent with its design), but also to severely violate the nominal level of the test when means are equal and the study group variance is *smaller*. The first of these two drawbacks also was shown to be true for a common exceedance test. The one type of test not examined here that researchers and quality investigators sometimes turn to when confronted with (1) are permutation tests. Permutation tests are computationally intensive, nonparametric, rank-based statistics that test whether the entire distributions of two populations are equal: Ho: $F(x) = G(x)$ vs. Ha: $F(x) \neq G(x)$. As such, they are not particularly powerful for detecting differences in variances alone (see Good, 2000, pp.32-33), and only one permutation test has been designed specifically as a joint mean-variance test of (1). Pesarin (2001) develops such a statistic by combining two permutation tests – one of scale and one of location – using any of several *p*-value combining functions (pp.147-148, 325). However, his test has one, and potentially two drawbacks relative to OBMax2: first, it is computationally intensive which, all else equal, is a limitation, and even runtime prohibitive for larger samples. Secondly, the combining function it relies upon must be chosen with care since many in common usage (e.g. Fisher and Liptak) would decrease the power of the test under many situations due to their “trade-off” nature: a small and significant *p*-value from one of the constituent tests can be “undone” by a large *p*-value from the other test when the two are combined, in which case the overall test will lose power under (1). While the Tippet function, which is itself a maximum-test approach, appears to avoid this problem, OBMax2 never suffers from a "trade-off" problem and thus, may be "safer" (i.e. more powerful, all else equal) if the Tippet function cannot be used with Pesarin's test for some reason (see Pesarin, 2001, pp.147-149, for further details).

Thus, only OBMax2 remains as a generally powerful, valid, and easily implemented test of (1) – whether the mean and/or variance of one process or population are larger than (worse than) those of another. For many quality control problems, like the equal service requirement in the regulatory telecommunications arena, this is the specific hypothesis that matters. The simulation study presented above demonstrates OBMax2's notable power under symmetry, and its much more modest power under asymmetry, which unfortunately is common among two-sample statistics and is the price paid for OBMax2's good level control under all conditions. Still, improving its power under skewed data is a worthy objective, and remains the subject of continuing research, as is the derivation of its asymptotic distribution.

To end on a cautionary note, it should be (re)emphasized that OBMax2 was developed specifically for hypothesis testing, for which it was shown to have good level control for (1) at commonly used nominal levels. In other words, its *p*-values appear to be uniformly distributed (at or under the appropriate "height") in the range that matters for most hypothesis tests. It would *not* be advisable, however, to use OBMax2's *p*-values for other purposes requiring a presumption of uniformity across its entire domain of zero to one. As one might expect, "β-inflating" the *p*-values of OBMax2's constituent statistics to maintain its overall validity has the consequence of "bunching up" its otherwise uniform *p*-value distribution at 1.0, as seen in Figure 12. One positive note from this finding, however, is that an examination of all of the 600 null distributions of OBMax2's *p*-values that were generated in this study reveals good level control for even larger nominal levels, such as $\alpha = 0.15$ and 0.20. And power given robust level control is all that matters for a hypothesis test, and this is the criterion by which OBMax2 far surpasses any of its competitors.

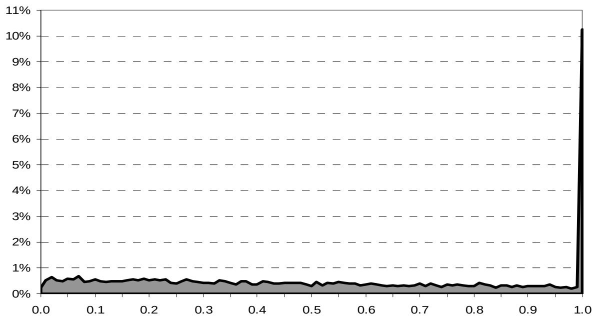

Figure 12. OBMax2 p-values:
Normal Data, μ_1=μ_2, σ_1=σ_2, n_1=n_2=30, N simulations=10,000

APPENDIX: Statistical Formulae

OBt and OBr: O'Brien's OBt test involves running the following ordinary least squares regression on pooled data including both samples:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i , \tag{10}$$

where y is a dummy variable indicating inclusion in the study group sample, and x is the performance metric variable. If the parameter on the quadratic term (β_2) is (positively) statistically significant at the 0.25 level, use the critical value of the overall equation (an F test of β_1=β_2=0) to reject or fail to reject the null hypothesis; if it is not, use the critical value of the overall equation of the following ordinary least squares regression instead:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{11}$$

(Of course, this latter test of β_1=0 is equivalent to a two-sample t test.) O'Brien's OBr test is identical to the OBt test except that the pooled-sample ranks of x are used in the regressions instead of the x data values themselves.

'modified' Levene test: The 'modified' Levene test requires a simple data transformation: take the absolute value of each data point's deviation from its respective sample median (as per Brown & Forsythe (1974)), and then calculate the usual one-way ANOVA statistic using these transformed values (as per Levene (1960)). The resulting statistic (12) is referenced to the F distribution as usual.

Let $z_{ij} = \left| x_{ij} - \tilde{x}_i \right|$ where $\tilde{x}_i$ is sample i's median

$$W_o = \frac{\sum_i n_i \left(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot} \right)^2 \Big/ (g-1)}{\sum_i \sum_j \left(z_{ij} - \bar{z}_{i\cdot} \right)^2 \Big/ \sum_i (n_i - 1)} \sim F_{(g-1), \sum_i (n_i - 1)} \tag{12}$$

where $\bar{z}_i = \sum z_{ij} \big/ n_i$ and $\bar{z}_{\cdot\cdot} = \sum \sum z_{ij} \big/ n_i$

However, because this test is designed as a two-tailed test, and the hypotheses being tested in (1) are one-tailed, the p-value resulting from this test, when used conditionally with O'Brien's tests as in Table 1, must be subtracted from 1.0 if the study group sample variance is less than the control group sample variance.

Shoemaker's F_1 test: Shoemaker's F_1 test is simply the usual ratio of sample variances referenced to the F distribution, but using different degrees of freedom:

$$s_2^2 \big/ s_1^2 \sim F_{df_2, df_1} \quad \text{where} \quad df_i = 2n_i \Big/ \left(\frac{\hat{\mu}_4}{\hat{\sigma}^4} - \frac{n_i - 1}{n_i - 3} \right), \tag{13}$$

i = 1, 2 corresponds to the two samples, and μ_4 and σ^4 are estimated from the two samples when pooled:

$$\hat{\mu}_4 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^4 \Big/ \left(n_1 + n_2 \right) \tag{14}$$

$$\hat{\sigma}^4 = \left[\left((n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right) \big/ \left(n_1 + n_2 \right) \right]^2 \tag{15}$$

Shoemaker (2003) notes that the biased estimate for σ^4 is used for improved accuracy.

separate-variance t test: The separate-variance t test, also known as the Welch or Behrens-Fisher t test, is below:

$$t_{sv} = \frac{\left(\bar{X}_2 - \bar{X}_1 \right) - \left(\mu_2 - \mu_1 \right)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \tag{16}$$

where $s_1^2 = \dfrac{\sum_{i=1}^{n_1} \left(X_{1_i} - \bar{X}_1 \right)^2}{(n_1 - 1)}$, $s_2^2 = \dfrac{\sum_{i=1}^{n_2} \left(X_{2_i} - \bar{X}_2 \right)^2}{(n_2 - 1)}$,

$\bar{X}_1 = \dfrac{\sum_{i=1}^{n_1} X_i}{n_1}$, and $\bar{X}_2 = \dfrac{\sum_{i=1}^{n_2} X_i}{n_2}$

Satterthwaite's (1946) degrees of freedom for t_{sv} is:

$$df = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2} \right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1} \right)^2}{(n_1 - 1)} + \dfrac{\left(\dfrac{s_2^2}{n_2} \right)^2}{(n_2 - 1)}} \tag{17}$$

If df is not an integer, it should be rounded down to the next smallest integer (see Zar (1999), p.129)

test of D'Agostino et al. (1990): The test of D'Agostino et al. (1990) is calculated as follows:

$$g_1 = \frac{k_3}{s^3} = \frac{\dfrac{n \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^3}{(n-1)(n-2)}}{\sqrt{\left(s^2 \right)^3}}, \qquad \sqrt{b_1} = \frac{(n-2) g_1}{\sqrt{n(n-1)}} \tag{18}$$

$$A = \sqrt{b_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}, \quad B = \frac{3\left(n^2 + 27n - 70 \right)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C = \sqrt{2(B-1)} - 1 \ , \quad D = \sqrt{C} \ , \quad E = \frac{1}{\sqrt{\ln D}}, \quad F = \frac{A}{\sqrt{\dfrac{2}{C-1}}},$$

$$Z_{g_1} = E \ln \left(F + \sqrt{F^2 + 1} \right) \sim \phi(0,1)$$

For one-tailed testing of skewness to the left, check $\Pr\left(Z \le Z_{g_1} \right)$; for skewness to the right, check $\Pr\left(Z \ge Z_{g_1} \right)$. See Zar (1999), pp.115-116, for further details.

REFERENCES

Algina, J., Blair, R.C., & Coombs, W.T. (1995), “A Maximum Test fo Scale: Type I Error Rates and Power,” *Journal of Educational and Behavioral Statistics*, 20, 27-39.

Algina, J., Olejnik, S. F., & Ocanto, R. (1989), “Error Rates and Power Estimates for Selected Two-sample Tests of Scale,” *Journal of Educational Statistics*, 14, 373–384.

Babu, G. & Padmanabhan, A.R. (1996), “A Robust Test for Omnibus Alternatives,” *Research Developments in Probability and Statistics*, E. Brunner and M. Denker (eds.), VSP, Utrecht, 319-327.

Blair, R.C. (2002), “Combining Two Nonparametric Tests of Location,” *Journal of Modern Applied Statistical Methods*, 1(1), 13-18.

Blair, R.C. (1991), “New Critical Values for the Generalized *t* and Generalized Rank-sum Procedures,” *Communications in Statistics*, 20, 981-994.

Boos, D., & Hughes-Oliver, J. (2000), “How Large Does n Have to be for Z and t Intervals?” *The American Statistician*, 54(2), 121-128.

Brown, M. & Forsythe, A. (1974), “Robust Tests for the Equality of Variances,” *Journal of the American Statistical Association*, 69, 364-367.

Brownie, C., Boos, D. D. & Hughes-Oliver, J. (1990), “Modifying the *t* and ANOVA *F* Tests When Treatment is Expected to Increase Variability Relative to Controls,” *Biometrics*, 46, 259-266.

Buning, H., & Thadewald, T. (2000), “An Adaptive Two-sample Location-scale Test of Lepage Type for Symmetric Distributions,” *Journal of Statistical Computation and Simulation*, 65(4), 287-310.

Chaffin, W., & Rhiel, G. (1993), “The Effect of Skewness and Kurtosis on the One-Sample T-Test and the Impact of Knowledge of the Population Standard Deviation,” *Journal of Computation and Simulation*, 46, 79=90.

Costa, A.F.B., & Rahim, M.A. (2004), “Monitoring Process Mean and Variability with One Non-central Chi-square Chart,” *Journal of Applied Statistics*, 31(10), 1171-1183.

D’Agostino, R.B., A. Belanger, & R.B. D’Agostino, Jr. (1990), “A Suggestion for Using Powerful and Informative Tests of Normality,” *The American Statistician*, 44, 316-321.

Gan, F.F., Ting, K.W., & Chang, T.C. (2004), “Interval Charting Schemes for Joint Monitoring of Process Mean and Variance,” *Quality and Reliability Engineering International*, 20(4), 291-303.

Gibbons, J., Chakraborti, S. (2003), Nonparametric Statistical Inference, 4th ed., Marcel Dekker, Inc., New York.

Good, P., (2000), *Permutation Tests*, 2nd ed. New York: Springer-Verlag New York, Inc.

Goodman, L.A. (1954), “Kolmogorov-Smirnov Tests for Psychological Research,” *Psychological Bulletin*, 51, 160-168.

Hawkins, D. M. & Zamba, K.D. (2005), “Statistical Process Control for Shifts in Mean or Variance Using a Changepoint Formulation,” *Technometrics*, 47(2), 164-173.

Levene, H. (1960), “Robust Tests for Equality of Variances,” in *Contribution to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al., eds., Stanford University Press, 278-292.

Manly, B., & Francis, C. (2002), “Testing for Mean and Variance Differences with Samples from Distributions that may be Non-normal with Unequal Variances,” *Journal of Statistical Computation and Simulation*, 72(8), 633-646.

Matlack, W.F., (1980), *Statistics for Public Policy and Management*, Belmont, CA: Duxbury Press.

Neuhäuser, M. Büning, H., & Hothorn, L. A. (2004), “Maximum Test Versus Adaptive Tests for the Two-sample Location Problem,” *Journal of Applied Statistics*, 31(2), 215-227.

O’Brien, P. (1988), “Comparing Two Samples: Extensions of the *t*, Rank-sum, and Log-rank Tests,” *Journal of the American Statistical Association*, 83, 52-61.

Opdyke, J.D. (2004), “Misuse of the ‘modified’ *t* Statistic in Regulatory Telecommunications,” *Telecommunications Policy*, 28, 821-866.

Opdyke, J.D. (2006), “A Single, Powerful, Nonparametric Statistic for Continuous-data Telecommunications ‘Parity Testing’,” *Journal of Modern Applied Statistical Methods*, 4(4), 372-393.

Pesarin, F. (2001), *Multivariate Permutation Tests with Applications in Biostatistics*, John Wiley & Sons, Ltd., New York.

Podgor, M.J., & Gastwirth, J.L. (1994), “On Non-parametric and Generalized Tests for the Two-sample Problem with Location and Scale Change Alternatives,” *Statistics in Medicine*, 13(5-7), 747-758.

Ramsey, P.H. & Ramsey, P.P. (2007), “Testing Variability in the Two-Sample Case,” Communications in Statistics – Simulation and Computation, Vol. 36 (2), 233-248.

Reynolds, M.R., & Stoumbos, Z.G. (2005), “Should Exponentially Weighted Moving Average and Cumulative Sum Charts be Used with Shewhart Limits?” *Technometrics*, 47(4), 409-424.

Rosenbaum, S. (1954), “Tables for a Nonparametric Test of Location,” *Annals of Mathematical Statistics*, 25, 146-150.

Satterthwaite, F. W. (1946), “An Approximate Distribution of Estimates of Variance Components, *Biometrics Bulletin*, 2, 110-114.

Siegel, S. & Castellan, N. John, (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., New York: McGraw-Hill.

Shoemaker, L. H. (2003), “Fixing the *F* test for Equal Variances,” *The American Statistician*, 57, 105-114.

Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996). [available at http://www.fcc.gov/telecom.html]

Wu, Zhang, Tian, Y., & Zhang, Sheng (2005), “Adjusted-loss-function Charts with Variable Sample Sizes and Sampling Intervals,” *Journal of Applied Statistics*, 32(3), 221-242.

Yang, Song, Li Hsu, & Lueping Zhao (2005), “Combining Asymptotically Normal Tests: Case Studies in Comparison of Two Groups,” *Journal of Statistical Planning and Inference*, 133(1), 139-158.

Zar, J.H., (1999), *Biostatistical Analysis*, 4th ed., Upper Saddle River, NJ: Prentice-Hall.

Zhou, X., & Gao, S., (2000), “One-Sided Confidence Intervals for Means of Positively Skewed Distributions,” The American Statistician, 54(2), 100-104.