Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2238–2248
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Confidence intervals for the difference between two means

Weiwen Miao[a,*], Paul Chiou[b]

[a]*Department of Mathematics, Haverford College, Haverford, PA 19041, USA*
[b]*Department of Mathematics, Lamar University, Beaumont, TX 77710, USA*

Received 16 May 2007; received in revised form 25 July 2007; accepted 25 July 2007
Available online 7 August 2007

Abstract

This paper compares three confidence intervals for the difference between two means when the distributions are *non-normal* and their variances are *unknown*. The confidence intervals considered are Welch–Satterthwaite confidence interval, the adaptive interval that incorporates a preliminary test (pre-test) of symmetry for the underlying distributions, and the adaptive interval that incorporates the Shapiro–Wilk test for normality as a pre-test. The adaptive confidence intervals use the Welch–Satterthwaite interval if the pre-test fails to reject symmetry (or normality) for both distributions; otherwise, apply the Welch–Satterthwaite confidence interval to the log-transformed data, then transform the interval back. Our study shows that the adaptive interval with pre-test of symmetry has best coverage among the three intervals considered. Simulation studies show that the adaptive interval with pre-test of symmetry performs as well as the Welch–Satterthwaite interval for symmetric distributions. However, for skewed distributions, the adaptive interval with pre-test of symmetry performs better than the Welch–Satterthwaite interval.

Keywords: Behrens–Fisher problem; Coverage probability; Expected length; Preliminary test; Shapiro–Wilk test; Welch–Satterthwaite confidence interval

1. Introduction

The problem of calculating confidence intervals for the difference between the means of two independent normal distributions is covered in numerous elementary statistics text books. The common way is to use the t distribution with pooled sample variances when the population variances *are known* to be equal; otherwise, use the non-pooled Welch–Satterthwaite confidence interval (Welch, 1938; Satterthwaite, 1946) referred as WS interval hereafter. However, in practice, the variances of the populations are usually unknown, and one tends to resolve the uncertainty by using a preliminary test (pre-test) on equality of variances. If the pre-test concludes that the variances are equal, one then pools the sample variances to construct an interval; otherwise, uses the non-pooled WS interval. This may sound like an excellent idea, but recent studies showed that this adaptive procedure is not as good as it sounds (see, e.g., Moser et al., 1989, 1992; Bradley, 1978, 1980a,b). The current practice for the case of *normal* distributions is that one pools the sample variances to construct an interval when the sample sizes are equal; otherwise, uses the non-pooled WS interval.

* Corresponding author.
E-mail addresses: wmiao@haverford.edu (W. Miao), paul.chiou@lamar.edu (P. Chiou).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.07.017

Besides the pooled-variance t confidence interval and the WS non-pooled intervals, several other studies also proposed different forms of confidence intervals for the difference between two *normal* means. For example, Scheffe (1943) proposed an interval with coverage always equal to the nominal level, but he himself later pointed out that the interval is not practical (Scheffe, 1970). Banerjee (1961) proposed another type of t-confidence interval, using two critical values from t distributions. His interval always has coverage equal to or higher than the nominal level. The high coverage probability comes with a price that the interval is always wider than the WS interval. Cochran (1964) proposed a confidence interval inverted from the hypothesis testing. Cochran's interval has the same format as the WS interval and the only difference is the degrees of freedom of the t. Cochran's interval has coverage probability always equal to or higher than the nominal level, but again with the price of being wider. Dalal (1978) proposed a confidence interval whose coverage is always equal to the nominal level, but the interval requires to first obtain a constant, t_α in the paper, that involves solving an equation with the product of two cdfs of t distributions. If the sample sizes are not equal, this constant needs to be found through numerical method. Sprott and Farewell (1993) proposed to provide several confidence intervals based on the plausible range of the variances ratio.

The duality between hypothesis testing and the corresponding confidence interval construction had been well documented. For two independent *normal* distributions, the hypothesis testing on equality of means is the well-known Behrens–Fisher problem. In reality, however, samples may come from non-normal distributions as well as unequal standard deviations, like the distribution of the income data. When both normality and equal variances assumptions are violated, some modifications of the t-test statistics are proposed. For example, instead of using the sample means and variances in the t-test statistic, one can use the symmetrically trimmed sample means and variances (Yuen, 1974; Yuen and Dixon, 1973), the modified maximum likelihood estimators based on symmetrically censored samples (Tiku, 1980), or the asymmetrically trimmed means and variances (Reed and Stark, 1996). Cressie and Whitford (1986) showed that one need only be concerned about the skewness effect, and they proposed a modified t-test to eliminate the bias of skewness. Reed and Stark (2004) did a simulation study on those modified t-tests. They found that when variances are actually equal, the pooled t-test performs well regardless of the underlying distributions. However, when the variance ratio is 1.5, the non-symmetric trimmed procedure performs the best. Gans (1981) also studied the hypothesis testing on equality of means when both normality and equal variances assumptions may be violated. The paper considered normal, uniform and exponential distributions. The three tests studied are the pooled t-test, WS t-test, and the adaptive procedure using the F-test on equality of variances as a pre-test. The conclusion is to use the non-adaptive WS t-test. Stonehouse and Forrester (1998) compared the pooled t-test, WS t-test and the non-parametric Mann–Whitney test. They found that contrary to its popular reputation, the Mann–Whitney U-test showed a dramatic lack of robustness and it is not a proper non-parametric analogue of the t-test.

This paper studies the confidence intervals of the difference between two means when both normality and equal variances assumptions may be violated. We consider three confidence intervals: the WS interval and two adaptive intervals. The WS interval was originally designed for unequal variances situation; however, studies already showed that this interval performs well when the standard deviations are equal (see, e.g., Moser et al., 1989). It motivated us to focus on an adaptive procedure concerning the shape of underlying distributions. On the other hand, the t-test is robust against non-normality, especially when the distributions are symmetric, and our simulation study also shows that the WS interval performs well on symmetric distributions. It ultimately leads us to use a test of symmetry (Miao et al., 2006) as the pre-test. If the pre-test concludes that neither distribution is symmetric, we transform the data into the scale of logarithm, then apply the WS confidence interval to the log-transformed data, and finally adjust the interval back to its original scale. However, if the pre-test indicates otherwise, we use the WS interval.

As normality is a common hypothesis in many practical situations, our second adaptive interval is to use the Shapiro–Wilk test (Shapiro and Wilk, 1965), the omnibus of testing normality, as the preliminary test. If the pre-test indicates both samples are not normally distributed, one applies the WS interval to log-transformed data, and then adjusts the interval as in the previous case; otherwise, simply uses the WS interval.

The paper is organized as follows. Section 2 proposes the three confidence intervals considered. Section 3 provides the comparison of the coverage probabilities of those three confidence intervals for different types of underlying distributions as well as different standard deviation ratios. The simulation shows that the adaptive interval with pre-test of symmetry has coverage probability close to the nominal level for both symmetric and non-symmetric distributions. The expected lengths of those intervals are compared in Section 4. The result shows that the adaptive interval with pre-test of symmetry is slightly wider than the WS interval. Section 5 provides recommendations for practice.

2. Confidence intervals considered

Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be random samples from two distributions (not necessary normal) with means μ_X, μ_Y and standard deviations σ_X and σ_Y, respectively. Let $\bar{X}$, $\bar{Y}$, S_X^2, S_Y^2 be the sample means and variances for X and Y, respectively. We are interested in the $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

2.1. The non-adaptive WS confidence interval

Let t_k^* be the $(1-\alpha/2)$ quantile of t distribution with degrees of freedom k. When X and Y are normally distributed, it is known that the following WS confidence interval performs well for both equal and unequal variances. The WS interval is defined by

$$I_{\text{ws}} = (\bar{X} - \bar{Y}) \pm t_{df}^* \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\,, \tag{1}$$

where

$$df = \frac{(w_1 + w_2)^2}{w_1^2/(n_X - 1) + w_2^2/(n_Y - 1)}, \quad w_1 = \frac{S_X^2}{n_X}, \quad w_2 = \frac{S_Y^2}{n_Y}.$$

Although the WS interval is designed to take care of non-equal variances situation when both distributions are normal, our simulations show that the WS confidence interval also performs well if the samples come from symmetric distributions. However, the coverage probability for I_{ws} interval can be much lower than its nominal level if the samples come from skewed distributions as well as unequal variances.

2.2. Pre-test of symmetry used in the adaptive interval

As WS interval performs well on symmetric distributions, we were motivated to focus on symmetry of the underlying distributions. To partially resolve the uncertainty whether the underlying distribution is symmetric or not, a pre-test of symmetry can be performed. Miao et al. (2006) has recently proposed a test for symmetry of distributions. Let $X_1, \ldots, X_n$ be a random sample from some distribution. The pre-test is

H_0: the underlying distribution is symmetric
H_a: the underlying distribution is not symmetric.

The test statistic is

$$T = \frac{\bar{X} - M}{J}, \quad J = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \sum_{i=1}^{n} |X_i - M|,$$

where $\bar{X}$ and M are the sample mean and sample median. Note that the numerator of T is the difference between the sample mean and sample median, and the denominator is a robust estimate of standard deviation. This test statistic is asymptotically normally distributed under both the null and alternative hypotheses (see Miao et al., 2006). The test calls to reject the null hypothesis at α' level of significance if

$$|T| \geqslant \frac{z_{\alpha'/2}\sqrt{0.5708}}{\sqrt{n}},$$

where the constant 0.5708 is the asymptotic variance of T when the underlying distribution is normal, and $z_{\alpha'/2}$ is the upper $\alpha'/2$ percentile of a standard normal. It is shown that this test has high power for exponential distributions with small sample sizes.

2.3. Confidence interval when the samples are not symmetric

For skewed distributions, taking the logarithm usually makes the distribution more symmetric. If the preliminary test concludes that both underlying distributions are not symmetric, we apply the WS interval I_{ws} to the log-transformed data. Then the delta method is used to adjust the interval.

First we transform the data X_i to $\log(X_i + c_X)$ and Y_i to $\log(Y_i + c_Y)$, where c_X and c_Y are the constants to make sure $X_i + c_X > 0$ and $Y_i + c_Y > 0$, and then we apply the WS interval to the log-transformed data. Let $[L_{\log}, U_{\log}]$ be the WS confidence interval obtained from $\log(X_1 + c_X), \ldots, \log(X_{n_X} + c_X)$ and $\log(Y_1 + c_Y), \ldots, \log(Y_{n_Y} + c_Y)$. The first-order Taylor expansion for $\log(X + c_X)$ is

$$\log(X + c_X) = \log(\mu_X + c_X) + \frac{1}{\mu_X + c_X}(X - \mu_X) + R,$$

where R is the remainder. Consequently,

$$E[\log(X + c_X)] \approx \log(\mu_X + c_X), \quad E[\log(Y + c_Y)] \approx \log(\mu_Y + c_Y)$$

and

$$E[\log(X + c_X)] - E[\log(Y + c_Y)] \approx \log(\mu_X + c_X) - \log(\mu_Y + c_Y).$$

Let $CP_{\log}$ be the coverage probability of $[L_{\log}, U_{\log}]$ for the difference $E[\log(X + c_X)] - E[\log(Y + c_Y)]$, and CP be the coverage of the adjusted interval for $\mu_X - \mu_Y$. We then have

$$\begin{aligned}
CP_{\log} &= P(E[\log(X + c_X)] - E[\log(Y + c_Y)] \in [L_{\log}, U_{\log}]) \\
&\approx P((\log(\mu_X + c_X) - \log(\mu_Y + c_Y)) \in [L_{\log}, U_{\log}]) \\
&= P\left(\frac{\mu_X + c_X}{\mu_Y + c_Y} \in [e^{L_{\log}}, e^{U_{\log}}]\right) \\
&= P(e^{L_{\log}}(\mu_Y + c_Y) \leqslant \mu_X + c_X \leqslant e^{U_{\log}}(\mu_Y + c_Y)) \\
&= P(e^{L_{\log}}(\mu_Y + c_Y) - \mu_Y - c_X \leqslant \mu_X - \mu_Y \leqslant e^{U_{\log}}(\mu_Y + c_Y) - \mu_Y - c_X) \\
&\approx P(\bar{Y}(e^{L_{\log}} - 1) + (c_Y \cdot e^{L_{\log}} - c_X) \leqslant \mu_X - \mu_Y \\
&\leqslant \bar{Y}(e^{U_{\log}} - 1) + (c_Y \cdot e^{U_{\log}} - c_X)) \\
&= CP.
\end{aligned}$$

Hence, the proposed confidence interval for $\mu_X - \mu_Y$ when both distributions are not symmetric is

$$I_{\log} = [\bar{Y}(e^{L_{\log}} - 1) + (c_Y \cdot e^{L_{\log}} - c_X),\ \bar{Y}(e^{U_{\log}} - 1) + (c_Y \cdot e^{U_{\log}} - c_X)]. \tag{2}$$

Note that the adjusted log confidence interval has two approximation steps. The first step is the use of Taylor expansion on the logarithm, and the second one is to use the $\bar{Y}$ to approximate μ_Y. According to the Central Limit Theory, $\bar{Y} \approx N(\mu_Y, \sigma_Y^2/n_Y)$. As $\bar{Y}$ is used to estimate μ_Y, we choose the sample with smaller standard error $(s_Y/\sqrt{n_Y})$ as the Y sample to obtain a better approximation.

2.4. The adaptive intervals

In our proposed adaptive procedure, a preliminary test of symmetry is conducted for both samples. If the pre-test detects that *both* samples are not symmetric, we use the log-adjusted interval $I_{\log}$ proposed in the previous subsection; otherwise, directly use the WS interval I_{ws}. Hence the adaptive confidence interval for $\mu_X - \mu_Y$ incorporating the

pre-test of symmetry is defined by

$$I_{\text{adp\_s}} = \begin{cases} I_{\log} & \text{if the pre-test rejects symmetry for } \textit{both} \text{ samples,} \\ I_{\text{ws}} & \text{otherwise.} \end{cases} \tag{3}$$

As normality is a rather common assumption in statistical inference, and the I_{ws} interval was designed for normal distributions. It motivated us to consider the adaptive confidence interval that incorporates the Shapiro–Wilk test for normality as a pre-test. The adaptive interval with Shapiro–Wilk test is thus defined by

$$I_{\text{adp\_sw}} = \begin{cases} I_{\log} & \text{if the pre-test rejects normality for } \textit{both} \text{ samples,} \\ I_{\text{ws}} & \text{otherwise.} \end{cases} \tag{4}$$

3. Coverage probability

This section provides simulation studies for the coverage probabilities of the three confidence intervals proposed in Section 2. The nominal level of the confidence interval is 95%. For adaptive confidence intervals, the level of the preliminary test is set at 10%. Both symmetric and non-symmetric distributions are considered. For symmetric distributions, we considered normal, 10% contaminated normal ($0.9\text{N}(0, 1) + 0.1\text{N}(0, 3^2)$), double exponential, heavy-tailed t_3 as well as short-tailed uniform distributions. For non-symmetric distributions, slightly skewed chi-squared distribution with degrees of freedom 8 (χ^2_8), heavily skewed lognormal and exponential distributions are considered. Equal sample sizes with $n_X = n_Y = 20$ as well as non-equal sample sizes with $n_X = 40$, $n_Y = 20$ are considered. The ratio of the standard deviations (σ_Y/σ_X) ranges from 0.2 to 5. The results are based on 10^4 simulations. As all the coverage probabilities are higher than 0.8, the error rate for the simulated results is $\sqrt{0.8 * 0.2/10\,000} \approx 0.004$ or less. Table 1 shows the simulation results for coverage probabilities when the two samples are from the same distribution family. Clearly for symmetric distributions, both the symmetry adaptive interval $I_{\text{adp\_s}}$ and the non-adaptive I_{ws} have coverage probabilities very close to the nominal 95% for all sample sizes and all standard deviation ratios combinations considered. This is not surprising as the $I_{\text{adp\_s}}$ uses WS interval for symmetric distributions and the non-adaptive I_{ws} has nice coverage when both distributions are symmetric. For $I_{\text{adp\_sw}}$, the coverage probability is close to the nominal level for normal, contaminated normal, and double exponential distributions. But for heavy-tailed t_3 distribution and the short-tailed uniform distribution, the coverage probability of the $I_{\text{adp\_sw}}$ tends to be slightly higher (about 1–2% higher) than the nominal 95%. This may be due to the fact that the Shapiro–Wilk test has high power for heavy-tailed t_3 and short-tailed uniform distributions, and the log-transformed confidence interval $I_{\log}$does not work as well as the WS confidence interval in those situations.

For slightly skewed χ^2_8, all the three intervals have coverage close to the nominal 95% for all the sample sizes and standard deviation ratios considered. However, for heavily skewed lognormal and exponential distributions, unless the two standard deviations are equal, the coverage for I_{ws} interval was lower than the nominal level. In some situations (lognormal), it is even lower than 90%. On the other hand, the two adaptive intervals have coverage probability much closer to the nominal level. Between the two adaptive intervals, the $I_{\text{adp\_s}}$ has better coverage than the $I_{\text{adp\_sw}}$ for exponential distribution and the lognormal distribution when the sample sizes are large ($n_X = 40$, $n_Y = 20$).

Table 2 presents the coverage probabilities when two distributions are not from the same family. The simulation shows that when one sample is from normal, the other is 10% contaminated normal, the coverage for all three intervals are about the same: they are all slightly higher than the nominal 95% except when the two standard deviations are equal. When one sample is normal, and the other is slightly skewed χ^2_8, all three intervals have coverage close to the nominal 95%. When one sample is χ^2_8, and the other is very skewed lognormal distribution with larger standard deviations, none of the three intervals has good coverage. When the χ^2_8 has larger standard deviations than the lognormal, the non-adaptive I_{ws} has coverage close to the nominal 95%, while the adaptive intervals $I_{\text{adp\_s}}$ and $I_{\text{adp\_sw}}$ have coverage 2% and 4%, respectively, lower than the nominal level. This might be due to the fact that adaptive procedure requires that both distributions are not symmetric (or normal) in order to perform the log-transformation, while the symmetry test (or Shapiro–Wilk test) has low power to detect asymmetry for slightly skewed χ^2_8 distribution. Hence the log-transformation is not really kicked in to improve the performance. When both distributions are very skewed, i.e., one is lognormal and the other is exponential, the non-adaptive $I_{\text{adp\_s}}$ has coverage below 90% when the standard deviation

Table 1
Coverage probability for samples from the same family[a]

	Standard deviation ratio (σ_Y/σ_X)								
	0.2	0.25	1/3	0.5	1	2	3	4	5
Normal distribution									
I_{ws}	0.9480	0.9498	0.9479	0.9521	0.9499	0.9446	0.9466	0.9487	0.9503
I_{adp_sw}	0.9487	0.9498	0.9484	0.9527	0.9502	0.9449	0.9471	0.9491	0.9510
I_{adp_s}	0.9486	0.9498	0.9481	0.9525	0.9503	0.9448	0.9468	0.9490	0.9505
I_{ws}	0.9491	0.9484	0.9481	0.9501	0.9495	0.9485	0.9515	0.9487	0.9501
I_{adp_sw}	0.9499	0.9489	0.9488	0.9507	0.9502	0.9489	0.9520	0.9496	0.9504
I_{adp_s}	0.9494	0.9486	0.9486	0.9502	0.9497	0.9489	0.9517	0.9489	0.9502
Contaminated normal distribution									
I_{ws}	0.9562	0.9565	0.9545	0.9564	0.9534	0.9577	0.9550	0.9550	0.9500
I_{adp_sw}	0.9625	0.9605	0.9601	0.9613	0.9590	0.9624	0.9593	0.9598	0.9538
I_{adp_s}	0.9568	0.9572	0.9555	0.9574	0.9540	0.9584	0.9559	0.9555	0.9507
I_{ws}	0.9518	0.9497	0.9482	0.9485	0.9505	0.9543	0.9520	0.9542	0.9549
I_{adp_sw}	0.9603	0.9590	0.9582	0.9582	0.9576	0.9622	0.9591	0.9618	0.9627
I_{adp_s}	0.9525	0.9505	0.9490	0.9493	0.9511	0.9551	0.9523	0.9545	0.9552
Double exponential distribution									
I_{ws}	0.9541	0.9569	0.9508	0.9513	0.9469	0.9522	0.9547	0.9540	0.9506
I_{adp_sw}	0.9601	0.9623	0.9564	0.9583	0.9532	0.9582	0.9597	0.9601	0.9570
I_{adp_s}	0.9551	0.9580	0.9519	0.9528	0.9493	0.9543	0.9565	0.9549	0.9517
I_{ws}	0.9512	0.9498	0.9506	0.9509	0.9539	0.9527	0.9527	0.9539	0.9569
I_{adp_sw}	0.9617	0.9603	0.9602	0.9601	0.9631	0.9623	0.9612	0.9621	0.9648
I_{adp_s}	0.9525	0.9523	0.9533	0.9529	0.9549	0.9535	0.9533	0.9542	0.9572
t_3 Distribution									
I_{ws}	0.9579	0.9510	0.9525	0.9533	0.9592	0.9555	0.9579	0.9618	0.9563
I_{adp_sw}	0.9645	0.9581	0.9601	0.9601	0.9661	0.9615	0.9639	0.9685	0.9627
I_{adp_s}	0.9587	0.9527	0.9541	0.9540	0.9609	0.9570	0.9589	0.9633	0.9577
I_{ws}	0.9535	0.9562	0.9533	0.9531	0.9588	0.9551	0.9549	0.9561	0.9553
I_{adp_sw}	0.9655	0.9691	0.9652	0.9632	0.9694	0.9639	0.9635	0.9651	0.9657
I_{adp_s}	0.9557	0.9595	0.9548	0.9550	0.9599	0.9559	0.9555	0.9568	0.9557
Uniform distribution									
I_{ws}	0.9491	0.9468	0.9508	0.9504	0.9488	0.9465	0.9518	0.9495	0.9506
I_{adp_sw}	0.9573	0.9567	0.9601	0.9587	0.9595	0.9552	0.9601	0.9593	0.9614
I_{adp_s}	0.9512	0.9495	0.9524	0.9517	0.9510	0.9496	0.9531	0.9521	0.9529
I_{ws}	0.9505	0.9501	0.9505	0.9496	0.9509	0.9472	0.9490	0.9457	0.9497
I_{adp_sw}	0.9647	0.9638	0.9663	0.9649	0.9683	0.9664	0.9690	0.9647	0.9685
I_{adp_s}	0.9524	0.9524	0.9539	0.9520	0.9527	0.9484	0.9503	0.9465	0.9504
χ^2_8 Distribution									
I_{ws}	0.9415	0.9418	0.9435	0.9457	0.9512	0.9453	0.9433	0.9439	0.9471
I_{adp_sw}	0.9469	0.9478	0.9483	0.9484	0.9484	0.9486	0.9479	0.9497	0.9523
I_{adp_s}	0.9429	0.9435	0.9450	0.9460	0.9505	0.9460	0.9446	0.9469	0.9487
I_{ws}	0.9481	0.9463	0.9475	0.9478	0.9494	0.9441	0.9398	0.9424	0.9424
I_{adp_sw}	0.9593	0.9576	0.9563	0.9507	0.9471	0.9487	0.9431	0.9486	0.9477
I_{adp_s}	0.9517	0.9496	0.9501	0.9491	0.9481	0.9448	0.9409	0.9433	0.9443
Lognormal distribution									
I_{ws}	0.8806	0.8864	0.8944	0.9258	0.9629	0.9233	0.8970	0.8903	0.8836
I_{adp_sw}	0.9596	0.9478	0.9351	0.9130	0.9490	0.9126	0.9352	0.9511	0.9562
I_{adp_s}	0.9287	0.9199	0.9162	0.9118	0.9497	0.9107	0.9165	0.9254	0.9241
I_{ws}	0.9111	0.9184	0.9318	0.9470	0.9513	0.9035	0.8874	0.8820	0.8770
I_{adp_sw}	0.9728	0.9603	0.9414	0.9175	0.9473	0.9318	0.9493	0.9546	0.9545
I_{adp_s}	0.9602	0.9526	0.9375	0.9199	0.9433	0.9281	0.9416	0.9455	0.9419
Exponential distribution									
I_{ws}	0.9250	0.9223	0.9294	0.9408	0.9568	0.9395	0.9338	0.9210	0.9235
I_{adp_sw}	0.9677	0.9653	0.9640	0.9525	0.9524	0.9507	0.9667	0.9649	0.9690
I_{adp_s}	0.9425	0.9419	0.9424	0.9429	0.9515	0.9422	0.9457	0.9413	0.9435

Table 1 *(Continued)*

	Standard deviation ratio (σ_Y/σ_X)								
	0.2	0.25	1/3	0.5	1	2	3	4	5
I_{ws}	0.9345	0.9389	0.9407	0.9540	0.9490	0.9293	0.9177	0.9207	0.9184
I_{adp_sw}	0.9833	0.9793	0.9720	0.9528	0.9475	0.9587	0.9581	0.9595	0.9613
I_{adp_s}	0.9617	0.9602	0.9589	0.9523	0.9425	0.9459	0.9401	0.9439	0.9461

[a]For each distribution, the first three rows are for sample sizes $n_X = n_Y = 20$, the fourth to sixth rows are for sample sizes $n_X = 40$ and $n_Y = 20$. The simulation was done by: first draw two random samples from the same distribution, then multiply the second sample by the corresponding standard deviation ratio.

Table 2
Coverage probability when two distributions are from different families[a]

	Standard deviation ratio (σ_Y/σ_X)								
	1	2	3	4	5				
$n_X = n_Y = 20$	X is normal, Y is contaminated normal								
I_{ws}	0.9500	0.9615	0.9695	0.9787	0.9777				
I_{adp_sw}	0.9508	0.9634	0.9712	0.9797	0.9790				
I_{adp_s}	0.9502	0.9627	0.9704	0.9795	0.9783				
$n_X = 40, n_Y = 20$									
I_{ws}	0.9544	0.9604	0.9703	0.9779	0.9811				
I_{adp_sw}	0.9549	0.9623	0.9720	0.9791	0.9825				
I_{adp_s}	0.9547	0.9604	0.9704	0.9781	0.9813				
$n_X = 20, n_Y = 40$									
I_{ws}	0.9496	0.9533	0.9603	0.9653	0.9650				
I_{adp_sw}	0.9500	0.9569	0.9643	0.9687	0.9682				
I_{adp_s}	0.9498	0.9556	0.9649	0.9705	0.9700				
	0.2	0.25	1/3	0.5	1	2	3	4	5
X is normal, Y is χ^2_8									
I_{ws}	0.9504	0.9492	0.9492	0.9471	0.9479	0.9448	0.9452	0.9371	0.9408
I_{adp_sw}	0.9521	0.9517	0.9512	0.9497	0.9492	0.9469	0.9475	0.9385	0.9423
I_{adp_s}	0.9514	0.9502	0.9500	0.9479	0.9486	0.9456	0.9464	0.9379	0.9413
I_{ws}	0.9492	0.9528	0.9469	0.9528	0.9420	0.9432	0.9451	0.9418	0.9440
I_{adp_sw}	0.9516	0.9543	0.9476	0.9539	0.9435	0.9445	0.9467	0.9433	0.9454
I_{adp_s}	0.9503	0.9539	0.9473	0.9531	0.9423	0.9435	0.9455	0.9423	0.9441
X is χ^2_8*, Y is lognormal*									
I_{ws}	0.9445	0.9470	0.9501	0.9472	0.9546	0.9232	0.8973	0.8820	0.8788
I_{adp_sw}	0.9132	0.9222	0.9188	0.9177	0.9142	0.8479	0.8243	0.8127	0.8164
I_{adp_s}	0.9273	0.9294	0.9325	0.9292	0.9312	0.8846	0.8596	0.8455	0.8448
I_{ws}	0.9487	0.9501	0.9491	0.9524	0.9417	0.9051	0.8763	0.8740	0.8776
I_{adp_sw}	0.9120	0.9127	0.9068	0.9181	0.8748	0.7692	0.7447	0.7445	0.7503
I_{adp_s}	0.9281	0.9249	0.9227	0.9339	0.9256	0.8742	0.8490	0.8458	0.8497
X is lognormal, Y is exponential									
I_{ws}	0.8858	0.8840	0.8998	0.9213	0.9572	0.9393	0.9299	0.9228	0.9235
I_{adp_sw}	0.9624	0.9564	0.9435	0.9249	0.9336	0.9390	0.9554	0.9557	0.9601
I_{adp_s}	0.9237	0.9202	0.9208	0.9185	0.9365	0.9344	0.9418	0.9388	0.9418
I_{ws}	0.9079	0.9158	0.9296	0.9490	0.9485	0.9341	0.9251	0.9219	0.9248
I_{adp_sw}	0.9518	0.9685	0.9478	0.9455	0.9386	0.9511	0.9563	0.9534	0.9506
I_{adp_s}	0.9533	0.9463	0.9381	0.9435	0.9389	0.9480	0.9486	0.9480	0.9446

[a]For last three sets of families, the first three rows are for sample sizes $n_X = n_Y = 20$, the fourth to sixth rows are for sample sizes $n_X = 40$ and $n_Y = 20$. The simulation was carried out in the following way to guarantee that the standard deviation ratio is at the given level: for the normal and contaminated normal case, choose the variance of the contamination; for the normal and χ^2_8 case, choose the variance of the normal; for the χ^2_8 and lognormal case, choose the mean of the lognormal; for the lognormal and exponential case, choose the rate of the exponential distribution.

ratio is less than $\frac{1}{3}$ for small sample size situations. However, both adaptive intervals perform well, and their coverage probabilities are close to the nominal 95%.

The following list summarizes the main results of the coverage probabilities:

- When the two distributions are either symmetric or only slightly skewed, no matter whether they are from the same family or not, all the three intervals have coverage probabilities close to the nominal level. Among the three, the non-adaptive I_{ws} and the symmetry adaptive interval $I_{\text{adp\_s}}$ have better coverage than the $I_{\text{adp\_sw}}$.
- When one distribution is slightly skewed and the other is very skewed, none of the interval has acceptable coverage. Further research is needed in this situation.
- When both distributions are very skewed, I_{ws} is not acceptable as its coverage may drop below 90% in some situations, but both adaptive intervals have acceptable coverage probabilities. The $I_{\text{adp\_s}}$ has the best overall coverage.

4. Expected length

Besides coverage probability, length is an important factor as well in judging the performance of a confidence interval. Let L_{ws}, $L_{\text{adp\_sw}}$ and $L_{\text{adp\_s}}$ denote the length of I_{ws}, $I_{\text{adp\_sw}}$ and $I_{\text{adp\_s}}$, respectively. Table 3 provides simulation results for the ratio of expected length of $I_{\text{adp\_sw}}$ and $I_{\text{adp\_s}}$ to that of I_{ws}, i.e., $E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$ and $E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$ for sample sizes $n_X = n_Y = 20$ and $n_X = 40$, $n_Y = 20$ when two samples are from the same family.

Clearly, almost all the ratios are bigger than 1 which indicates that the I_{ws} has the shortest expected length among the three intervals considered. Notice that the length of the Shapiro–Wilk adaptive interval $I_{\text{adp\_sw}}$ sometimes can be 2 or 3 times wider than the other two intervals. On the other hand, its coverage is not significantly higher than that of the symmetry adaptive interval $I_{\text{adp\_s}}$, hence using the interval $I_{\text{adp\_sw}}$ is not recommended. From now on, we only compare the non-adaptive I_{ws} to the symmetry adaptive interval $I_{\text{adp\_s}}$.

For normal and χ^2_8 distributions, when sample sizes are equal or the larger sample is associated with smaller variances, the adaptive interval $I_{\text{adp\_s}}$ has about the same width as the I_{ws}; when the larger sample is associated with larger variances, the adaptive interval $I_{\text{adp\_s}}$ is about 4–7% wider than the I_{ws}. For contaminated normal, double exponential, and uniform distributions, when the larger sample is associated with smaller variances, the adaptive interval $I_{\text{adp\_s}}$ is about 2–4% wider than the I_{ws}; when the larger sample is associated with larger variances or when the sample sizes are equal, the adaptive interval $I_{\text{adp\_s}}$ is about 10–15% wider than the I_{ws}. However, for heavy-tailed t_3 distribution, the $I_{\text{adp\_s}}$ can be about 20–30% wider than the I_{ws}. For very skewed lognormal and exponential distributions, when the sample sizes are equal, the adaptive $I_{\text{adp\_s}}$ is about 10–20% (20–40% for exponential) wider than the non-adaptive I_{ws}; when the larger sample is associated with smaller variances, the adaptive interval $I_{\text{adp\_s}}$ is about 7% (or 40% for exponential) wider than the $I_{\text{adp\_s}}$; when the larger sample is associated with larger variances, the adaptive interval $I_{\text{adp\_s}}$ is about 30% (or 80% for exponential) wider than the I_{ws}. Notice that for lognormal distributions, when the standard deviations are actually equal, the adaptive $I_{\text{adp\_s}}$ is shorter than the non-adaptive I_{ws}. Its coverage is also very close to the nominal level. This is not surprising as the symmetry test has high power on the very skewed lognormal distribution, and the log-transformed data are normal, hence the adaptive interval $I_{\text{adp\_s}}$ works better than the non-adaptive I_{ws}.

Table 4 provides the ratio of expected length when two samples come from different families. Note that this table only reports the ratio of $E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$ as the other interval, $I_{\text{adp\_sw}}$, is not recommended. For normal and contaminated normal case, when the larger sample is associated with smaller variances, the $L_{\text{adp\_s}}$ is about 7% wider than the L_{ws}; when the larger sample is associated with larger variances, the $L_{\text{adp\_s}}$ can be twice as wide as the L_{ws}; when the samples sizes are equal, the adaptive $L_{\text{adp\_s}}$ is about 20–40% wider than the L_{ws} unless the variances are equal. For normal and χ^2_8, when the normal has larger standard deviations, the $L_{\text{adp\_s}}$ is about 5% wider than the L_{ws}; when the χ^2_8 has larger standard deviations and sample sizes are small, the $L_{\text{adp\_s}}$ is about 10–15% wider than the L_{ws}; when the χ^2_8 has larger standard deviations and sample sizes are large, the $L_{\text{adp\_s}}$ has about the same size as the L_{ws}. For χ^2_8 and lognormal case, when the skewed lognormal distribution has larger variances, the $L_{\text{adp\_s}}$ is actually narrower than the L_{ws}; when the lognormal has smaller variances, the $L_{\text{adp\_s}}$ can be twice as wide as the L_{ws}. For lognormal and exponential case, the $L_{\text{adp\_s}}$ is about 30% wider than the L_{ws} when sample sizes are equal or the exponential distribution has smaller variance; however, when the exponential distribution has larger variances, the $L_{\text{adp\_s}}$ is about 60% wider ($n_X = 40, n_Y = 20$) than the L_{ws}.

Table 3
Ratio of expected length for samples from the same family[a]

	Standard deviation ratio (σ_Y/σ_X)								
	0.2	0.25	1/3	0.5	1	2	3	4	5
Normal									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.06	1.05	1.06	1.06	1.05	1.07	1.08	1.07	1.08
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.04	1.03	1.03	1.03	1.03	1.04	1.04	1.04	1.05
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.08	1.10	1.11	1.08	1.03	1.05	1.03	1.04	1.04
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.06	1.06	1.07	1.06	1.02	1.01	1.01	1.02	1.01
Contaminated normal									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	2.18	2.32	2.23	2.25	2.27	2.41	2.49	2.62	2.56
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.08	1.11	1.12	1.10	1.11	1.11	1.14	1.11	1.15
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	3.85	3.82	3.84	3.55	2.15	2.18	2.23	2.28	2.29
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.17	1.18	1.17	1.15	1.05	1.04	1.04	1.04	1.04
Double exponential									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.96	1.98	2.03	2.05	2.04	2.18	2.24	2.33	2.30
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.07	1.09	1.08	1.08	1.08	1.09	1.09	1.11	1.12
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	3.30	3.36	3.41	3.12	1.98	1.99	2.09	2.12	2.09
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.14	1.15	1.14	1.12	1.04	1.03	1.03	1.03	1.03
t_3									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	2.60	2.81	2.82	2.91	2.94	3.07	3.33	3.25	3.26
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.18	1.20	1.21	1.22	1.22	1.21	1.29	1.30	1.24
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	4.84	4.97	4.85	4.44	2.86	2.83	2.74	2.87	2.87
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.38	1.37	1.33	1.35	1.12	1.10	1.09	1.11	1.11
Uniform									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.49	1.51	1.54	1.55	1.42	1.63	1.66	1.67	1.69
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.06	1.08	1.07	1.07	1.05	1.07	1.08	1.08	1.09
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	2.54	2.59	2.57	2.48	1.54	1.68	1.76	1.79	1.77
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.11	1.12	1.12	1.11	1.02	1.03	1.03	1.03	1.04
χ^2_8									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.07	1.06	1.06	1.05	1.00	1.05	1.06	1.07	1.07
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.02	1.02	1.02	1.01	1.00	1.02	1.02	1.02	1.02
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.22	1.21	1.19	1.11	1.00	1.04	1.05	1.05	1.06
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.07	1.07	1.06	1.04	1.00	1.01	1.01	1.02	1.02
Lognormal									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.24	1.22	1.16	1.04	0.89	1.06	1.16	1.21	1.24
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.17	1.16	1.11	1.01	0.90	1.03	1.11	1.14	1.16
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.43	1.38	1.27	1.06	0.87	1.00	1.06	1.08	1.08
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.38	1.34	1.24	1.05	0.89	1.00	1.05	1.07	1.07
Exponential									
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	1.86	1.86	1.81	1.69	1.43	1.70	1.81	1.84	1.86
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.43	1.43	1.40	1.34	1.20	1.35	1.40	1.42	1.42
$E[L_{\text{adp\_sw}}]/E[L_{\text{ws}}]$	2.35	2.33	2.19	1.85	1.39	1.56	1.59	1.60	1.61
$E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$	1.88	1.87	1.76	1.54	1.23	1.36	1.39	1.40	1.40

[a]For each distribution, the first two rows are for $n_X = n_Y = 20$, the third and fourth rows are for $n_X = 40$ and $n_Y = 20$. The simulation was done by: first draw two random samples from the same distribution, then multiply the second sample by the corresponding standard deviation ratio.

The following list summarizes the main findings for expected length:

- The Shapiro–Wilk adaptive interval $L_{\text{adp\_sw}}$ is sometimes 2–3 times wider than the other intervals, hence using this interval is not recommended.
- The non-adaptive L_{ws} has the smallest length in most cases.
- The ratio $E[L_{\text{adp\_s}}]/E[L_{\text{ws}}]$ is closer to 1 when larger sample is associated with smaller variance.

Table 4
The $E[L_{adp\_s}]/E[L_{ws}]$ when two distributions are from different families[a]

	Standard deviation ratio (σ_Y/σ_X)								
	1	2	3	4	5				
X is normal, Y is contaminated normal									
$n_X = n_Y = 20$	1.04	1.20	1.30	1.41	1.47				
$n_X = 40, n_Y = 20$	1.01	1.03	1.07	1.07	1.06				
$n_X = 20, n_Y = 40$	1.02	1.32	1.78	2.22	2.55				
	0.2	0.25	1/3	0.5	1	2	3	4	5
(Normal, χ^2_8)	1.05	1.05	1.04	1.04	1.12	1.15	1.14	1.13	1.15
	1.05	1.05	1.05	1.04	1.01	1.01	1.02	1.01	1.01
(χ^2_8, Lognormal)	1.43	1.45	1.43	1.36	1.14	0.95	0.93	0.93	0.93
	2.13	2.13	2.04	1.81	1.17	0.95	0.95	0.95	0.95
(Lognormal, Exp.)	1.37	1.36	1.31	1.20	1.07	1.20	1.27	1.30	1.31
	1.66	1.62	1.50	1.25	1.13	1.25	1.29	1.30	1.30

[a]For last three sets of families, the first row is for sample sizes $n_X = n_Y = 20$, the second row is for sample sizes $n_X = 40$ and $n_Y = 20$. The simulation was carried out in the following way to guarantee that the standard deviation ratio is at the given level: for the normal and contaminated normal case, choose the variance of the contamination; for the normal and χ^2_8 case, choose the variance of the normal; for the χ^2_8 and lognormal case, choose the mean of the lognormal; for the lognormal and exponential case, choose the rate of the exponential distribution.

- When both distributions are symmetric or only slightly skewed, the $E[L_{adp\_s}]$ is only slightly larger than $E[L_{ws}]$. If one distribution or both distributions are very skewed, the $E[L_{adp\_s}]$ is much larger than $E[L_{ws}]$.

5. Conclusion and discussions

This paper studies three confidence intervals for the difference between two means when both normality and homogeneity of variance assumptions are violated. The paper shows that using the Shapiro–Wilk test of normality as the pre-test, whether the homogeneity of variance assumption is satisfied or not, provides counter productive results. This is consistent with the results in Schucany and Ng (2006) for one-sample *t*-test. They concluded that graphic diagnostics are better practice than a formal pre-test. Furthermore, the paper shows that the WS interval works well for symmetric non-normal distributions regardless of the standard deviation ratio. In other words, the more important feature is the symmetry of the underlying distributions, not the normality. This may be due to the fact that the Central Limit Theory supports the Welch–Satterthwaite *t*-approximation when the underlying distributions are symmetric.

This paper shows that if both the underlying distributions are skewed and the homogeneity of variances assumption is also violated, the Welch–Satterthwaite interval has much lower coverage probability than the nominal level. The paper thus proposes an adaptive interval incorporating a pre-test of symmetry for underlying distributions. If the pre-test concludes neither distribution is symmetric, we propose to use an interval that first transforms the data into logarithm, then applies the WS interval to the log-transformed data and finally adjusts log-interval back to the original scale. Our simulation study shows that this adaptive interval performs as well as the WS interval for symmetric distributions, while it has much better coverage probability than the WS interval for skewed distributions. Therefore, the use of the adaptive interval I_{adp_s} when the underlying distributions are generally unknown is recommended.

Acknowledgment

We would like to thank Professor Joseph Gastwirth for his suggestions.

References

Banerjee, S., 1961. On confidence interval for two-mean problem based on separate estimates of variances and tabulated values of *t*-table. Sankhya 23, 359–378.

Bradley, J.V., 1978. Robustness? British J. Math. Statist. Psych. 31, 144–152.

Bradley, J.V., 1980a. Nonrobustness in classical tests on means and variances: a large-scale sampling study. Bull. Psychonomic Soc. 15, 275–278.
Bradley, J.V., 1980b. Nonrobustness in *Z*, *t* and *F* tests at large sample sizes. Bull. Psychonomic Soc. 16, 333–336.
Cochran, W.G., 1964. Approximation significance level of the Behrens–Fisher test. Biometrics 20, 191–195.
Cressie, N.A.C., Whitford, H.J., 1986. How to use the two-sample *t*-test. Biometrical J. 2, 131–148.
Dalal, S.C., 1978. Simultaneous confidence procedures for univariate and multivariate Behrens–Fisher type problems. Biometrika 65, 221–225.
Gans, D.J., 1981. Use of a preliminary test in comparing two sample means. Comm. Statist. Simulation Computat. 10, 163–174.
Miao, W., Gel, Y.L., Gastwirth, J.L., 2006. A new test of symmetry about an unknown median. In: Hsiung, A.C., Ying, Z., Zhang, C.H. (Eds.), Random Walk, Sequential Analysis and Related Topics—A Festschrift in Honor of Yuan-Shih Chow. World Scientific Publisher, Singapore, pp. 199–214.
Moser, B.K., Stevens, G.R., Watts, C.L., 1989. The two-sample *t* test versus Satterthwaite's approximation *F* test. Comm. Statist. Theory Methods 18, 3963–3975.
Moser, B.K., Stevens, G.R., Watts, C.L., 1992. Homogeneity of variance in the two-sample means test. Amer. Statist. 46, 19–21.
Reed III, J.E., Stark, D.B., 1996. Hinge estimators of location: robust to asymmetry. Comput. Methods Programs in Biomedicine 49, 11–17.
Reed III, J.E., Stark, D.B., 2004. Robust two-sample statistics for testing equality of means: a simulation study. J. Appl. Statist. 31, 831–854.
Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometrics Bull. 2, 110–114.
Scheffe, H., 1943. On solutions of the Behrens–Fisher problem, based on the *t*-distribution. Ann. Math. Statist. 14, 35–44.
Scheffe, H., 1970. Practical solution of the Behrens–Fisher problem. J. Amer. Statist. Assoc. 65, 1501–1508.
Schucany, W., Ng, H.K., 2006. Preliminary goodness-of-fit tests for normality do not validate the one-sample Student *t*. Comm. Statist. Theory Methods 35, 2275–2286.
Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrica 52, 591–611.
Sprott, D.A., Farewell, V.T., 1993. The difference between two normal means. Amer. Statist. 47, 126–128.
Stonehouse, J.M., Forrester, G.J., 1998. Robustness of the *t* and *U* tests under combined assumption violations. J. Appl. Statist. 25, 63–74.
Tiku, M.L., 1980. Robustness of MML estimators based on censored samples and robust test statistics. J. Statist. Plann. Inference 4, 123–143.
Welch, B.L., 1938. The significance of the difference between two means when the population variances are unequal. Biometrika 29, 350–362.
Yuen, K.K., 1974. The two-sample trimmed *t* for unequal population variances. Biometrika 61, 165–170.
Yuen, K.K., Dixon, W.J., 1973. The approximate behavior and performance of the two-sample trimmed *t*. Biometrika 60, 369–374.