Do Preliminary Tests Validate the Main Tests?

X. Zhu[1], S. Chakraborti[1] and Y. H. Dovoedo[2]
[1]Department of Information Systems, Statistics, and Management Science,
University of Alabama, AL 35487
[2]Department of Mathematics, University of North Alabama, AL 35632

Abstract
In practice, it is common to precede the main test of interest with one or more preliminary tests (pretests). Depending on the results of the preliminary test, one decides what main test to perform. Thus the type I error rate of the main test is actually a "conditional type I error" and there arise questions about whether or not the main test maintains the specified nominal level. In this paper, we revisit the literature on the methodology that is used to estimate the conditional type I error rate. Results are then extended to estimating and deriving the conditional type I error rate in some other situations. The work uses extensive simulations and some exact derivations.

Key Words: Pretest, type I error rate, conditional

1. Introduction

Researchers frequently precede analyses of interest with one or more preliminary tests (pretests), which are used to determine whether assumptions for the main test are met. Based on the results of the preliminary test(s), the main test of interest is performed. The type I error rate of the main test becomes a conditional one and is thus called the conditional type I error rate of the main test. On the contrary, if the main test of interest is performed directly without any pretests, the type I error rate is an unconditional type I error rate.

A lot of researchers have found the conditional type I error rate unacceptably inflated and recommended using any preliminary testing with great caution. Schucany et al. (2006) explored the conditional type I error of the one-sample t-test after a preliminary goodness-of-fit test for normality. Rochon et al. (2012) examined the pooled two-sample t-test instead, after the same preliminary test. Here, we revisit the case studied by Schucany et al. (2006) and apply the methodology to some other problems.

2. Methodology and Example

Schucany et al. (2006) considered testing the hypothesis about the mean $H_0\colon \mu = \mu_0$ against $H_a\colon \mu \neq \mu_0$ based on a random sample$x = \{x_1, x_2, \cdots x_m\}$. One approach is to directly perform the one-sample t-test without doing any preliminary test for normality. The second approach is to use a two-stage procedure, in which we first perform a preliminary test (pretest) of normality (such as the Shapiro-Wilk test) followed by the one-sample t-test if normality is not rejected. The one-sample t-test is the main test of interest in this case, with two different kinds of type I error rates. The Type I error of the one-sample t-test using the second method is a conditional one, as it depends on the outcome of the preliminary test for normality, while that from the first method is an unconditional Type I error, as this test is performed without any preliminary testing. Schucany et al. (2006) varied the underlying distribution from uniform, exponential, Cauchy to normal distribution and estimated both Type I error rates by simulation. They found that the pretest degraded the one-sample t-test and inflated the conditional Type I error rate more than no pretest.

Now this situation is revisited and we first estimate the unconditional type I error rate by generating 10000 random samples of size $n = 5, 10, 20, 30, 50$ from the following distributions: 1) standard normal; 2) t distribution with 5 degrees of freedom (labeled as t5 in Figure 1); 3) Laplace distribution; 4) Uniform [0, 1]; 5) Beta (3, 2); Gamma (3, 1); 7) Chi-square with 2 degrees of freedom (labeled as Chi2 in Figure 1); 8) Exponential (1).

We perform the one-sample t-test on each sample, check if it is significant or not, and find the proportion of significant tests among 10,000 random samples. The unconditional type I error rate is estimated by this proportion. Next, the conditional type I error rate is estimated by generating random samples from each of the distributions as above until 10000 samples pass the preliminary test for normality. For these 10000 samples, the one-sample t-test is performed at the nominal level of $\alpha = 0.05$ and the conditional Type I error rate is estimated by the proportion of significant t-tests (p-value smaller than α) among them. It is observed from the simulation process that even with highly non-normal populations, many samples still pass the screening for normality, which causes more error in the second stage.

Figure 1 and 2 are constructed based on the estimated unconditional and conditional type I error rates, respectively.

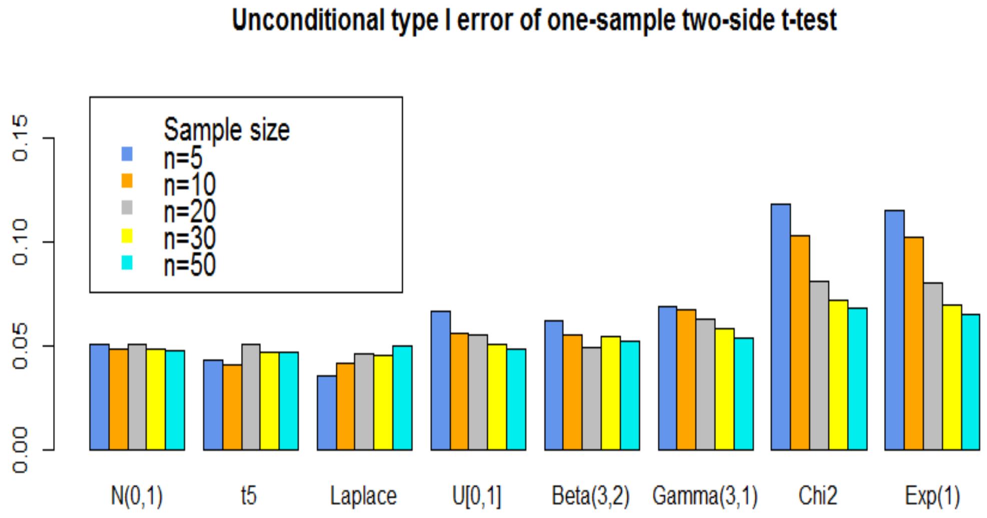

Figure 1: Unconditional type I error of the one-sample t-test

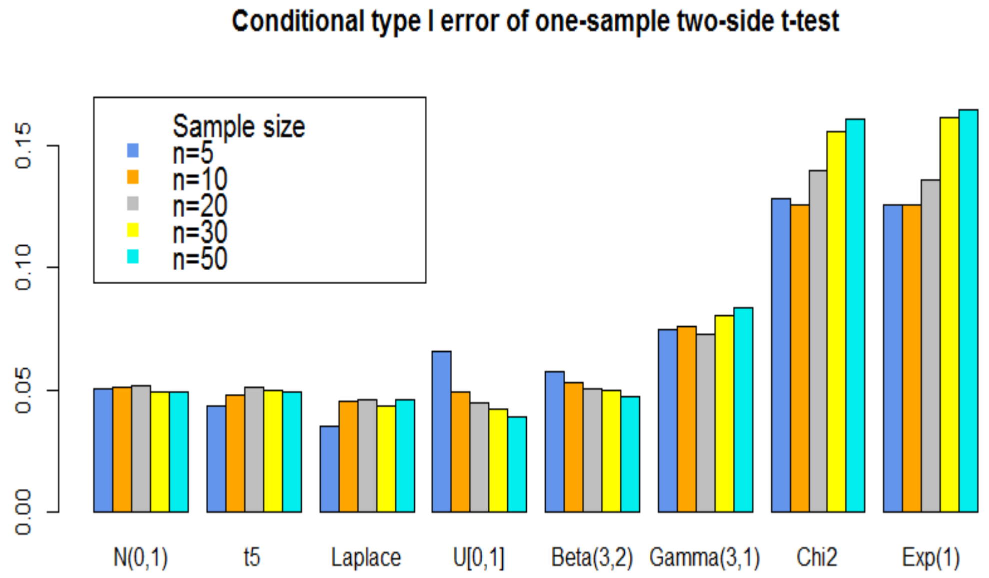

Figure 2: Conditional type I error of the one-sample t-test

We observe from Figure 1 and 2 that for the normal, *t* and Laplace distributions, both the unconditional and conditional type I error rates stay around 0.05, whatever the sample size. For symmetric and roughly bell-shaped distributions, the inflation of the type I error rates is not a problem of concern.

For the uniform and the beta distributions, when the sample size is 5, both the unconditional and conditional type I error rates exceed the nominal level by around 40% and is slightly problematic. When the sample size increases to 30 and then to 50, the two kinds of type I error rates gradually decrease to the nominal level or even a little below. The inflation of the type I error rates is associated with sample size too small. The simulation results fluctuate around the nominal level when the sample size is greater than 10.

For the Gamma (3, 1) distribution, which is slightly right skewed, we start to observe more problematic performance on the side of the conditional type I error rate. An increase in the sample size to 50 makes the unconditional type I error rate gradually converge to the nominal level (Figure 1), while at the same time inflates the conditional type I error rate to around 0.08 (Figure 2). If we look at the chi-square and the exponential distributions, this effect is more intense. When the sample size is as small as 5, both the unconditional and conditional type I error rates exceed the nominal level by more than 100%. But when the sample size increases to 50, the unconditional type I error rates have the trend of convergence and decrease to around 0.07 (Figure 1), while the conditional type I error rates have the opposite trend of increase and exceed the nominal level by more than 300% when sample size is 50 (Figure 2).

In practice, we do not know the true underlying distribution. If the sample size is around 10, we should know that both unconditional and conditional type I error rates can be problematic (when the distribution is skewed). When the sample size is greater than 30, the conditional type I error rate can be inflated considerably and thus not acceptable. The unconditional type I error rate, on the other hand, stays around the nominal level.

3. Contributions

The past work on conditional type I error rate has mainly been done for a preliminary goodness-of-fit test. However, other kinds of preliminary tests can also affect the type I error of the main test. We extend this idea and use the methodology to study the conditional type I error rate in the following problems.

(1). Two-sample pooled t-test after the samples pass a preliminary test for equality of variances, assuming normality

Let $x = \{x_1, x_2, \cdots x_m\}$ and $y = \{y_1, y_2, \cdots y_n\}$ be two random samples from two normally distributed populations, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.We want to test whether the two samples have equal means $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$. The unconditional approach is to directly perform the pooled t-test without doing any preliminary test. The conditional approach is to use a two-stage procedure, in which one first performs the preliminary test (pretest) of equality of variances, and then does the pooled t-test if equality of variances is not rejected. The Type I error of the two-sample pooled t-test using the second method is a conditional one, as it depends on the outcome of the preliminary test for variance while that from the first method is an unconditional Type I error, as this test is performed without any preliminary tests. Albers et al. (2000) approximated the conditional type I error rate of the pooled t-test in expectation format using the leading term. They found that it was much larger than the nominal level when the underlying variances are unequal. Here, we derive the exact expression of the conditional type I error rate as well as that of the unconditional one.

A theoretical derivation gives the formula to calculate the unconditional type I error rate

$$\tilde{\alpha} = \int_0^{\infty} \mathrm{f}_{\chi_1^2}(\mathrm{w}) \int_0^{\frac{\mathrm{w}}{\mathrm{c}'}} \int_0^{\frac{\mathrm{s}}{\delta^2}} \mathrm{f}_{\chi_{\mathrm{m}-1}^2}(\mathrm{u}) \mathrm{f}_{\chi_{\mathrm{n}-1}^2}(\mathrm{s} - \delta^2 \mathrm{u})\, \mathrm{du}\, \mathrm{dsdw},$$

where $\mathrm{c}' = \frac{[\mathrm{t}_{\mathrm{m+n-2}}(1-\frac{\alpha}{2})]^2 (\frac{1}{\mathrm{m}}+\frac{1}{\mathrm{n}}) \frac{1}{\mathrm{m+n-2}}}{\frac{\delta^2}{\mathrm{m}}+\frac{1}{\mathrm{n}}}$ and $\delta = \frac{\sigma_1}{\sigma_2}$. The unconditional type I error rates for different values of δ and sample sizes are computed. The results are reported in Table 1 and displayed in Figure 3.

Table 1: Unconditional Type I Error Rate of the Pooled t-test at $\alpha = 0.05$

δ	$n = m = 5$	$n = m = 10$	$n = m = 20$	$n = m = 30$	$n = m = 50$
0.5	0.0585	0.0547	0.0524	0.0516	0.0510
1.0	0.0500	0.0500	0.0500	0.0500	0.0500
1.5	0.0532	0.0519	0.0510	0.0507	0.0504
2.0	0.0585	0.0547	0.0524	0.0516	0.0510

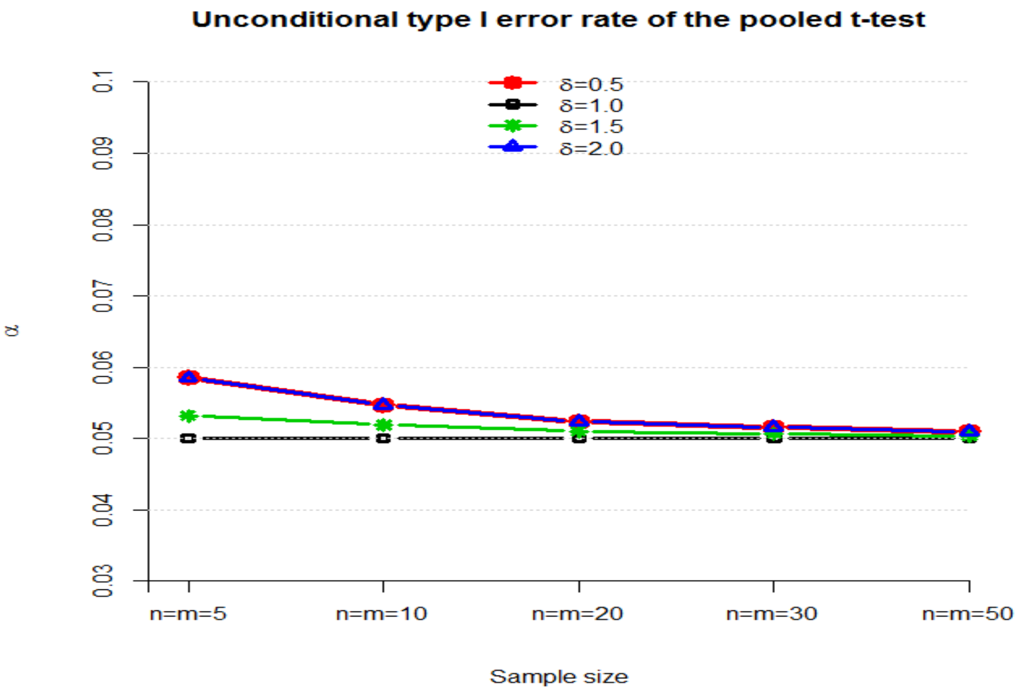

Figure 3: Unconditional Type I Error of the Pooled t-test

When the sample size is smaller than 10, the unconditional Type I error rate has an inflation rate around 15% for $\delta = 0.5,2$. It converges to the nominal level 0.05 regardless of δ when sample size increases to 50.

The conditional type I error rate is given by the formula

$$\widetilde{\alpha} = \int_0^\infty 2[1-\Phi(c'y)] \frac{\int_{\frac{y^2\delta^2}{F_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)\frac{(m-1)}{(n-1)}+1}}^{\frac{y^2\delta^2}{F_{m-1,n-1}\left(\frac{\alpha}{2}\right)\frac{(m-1)}{(n-1)}+1}} f_{\chi^2_{m-1}}\left(y^2-\frac{v}{\delta^2}\right) f_{\chi^2_{n-1}}(v)2ydv}{F_{F_{m-1,n-1}}\left(\frac{F_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)}{\delta^2}\right) - F_{F_{m-1,n-1}}\left(\frac{F_{m-1,n-1}\left(\frac{\alpha}{2}\right)}{\delta^2}\right)} dy,$$

where $c' = \frac{t_{m+n-2}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{1}{m+n-2}}}{\sqrt{\frac{1}{m}+\frac{1}{n\delta^2}}}$.

Some numerical values for the conditional type I error rate are reported in Table 2 and displayed in Figure 4. We compare these numerical values in Table 2 with the unconditional type I error rates.

Table 2: Conditional Type I error of the Pooled t-test at $\alpha = 0.05$

δ	$n = m = 5$	$n = m = 10$	$n = m = 20$	$n = m = 30$	$n = m = 50$
0.5	0.0665	0.0717	0.0793	0.0842	0.0900
1.0	0.0500	0.0500	0.0500	0.0500	0.0500
1.5	0.0553	0.0558	0.0570	0.0580	0.0596
2.0	0.0665	0.0717	0.0793	0.0842	0.0900

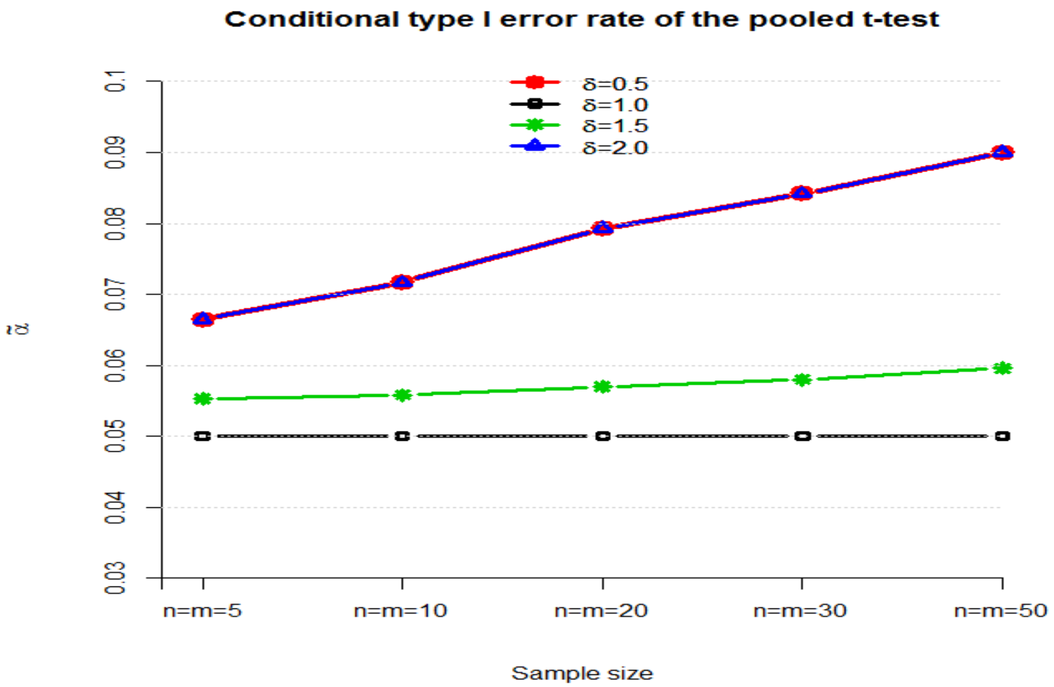

Figure 4: Conditional Type I Error of the Pooled t-test

Conditional Type I error rate has the trend of increase associated with the growth of sample size when δ is extreme. It can exceed the nominal level 0.05 by 80% when the sample size is 50 (Figure 4). We observe that increasing the sample size has opposite effects on the conditional type I error rates and unconditional type I error rates. It helps control the unconditional type I error rates while at the same time inflating the conditional type I error rates.

(2). The rejection of the combined test by Perng et al. (1976), assuming normality

Simultaneous testing the equality of means and that of variances of two populations has been considered by many researchers. Two random samples $x = \{x_1, x_2, \cdots x_m\}$ and $y = \{y_1, y_2, \cdots y_n\}$ are independent of each other from $N(\mu_1, {\sigma_1}^2)$ and $N(\mu_2, {\sigma_2}^2)$ respectively. The practitioner is interested in testing the hypothesis $H_1: \mu_1 = \mu_2 \; and \; {\sigma_1}^2 = {\sigma_2}^2$ against $H_{1a}: \mu_1 \neq \mu_2 \; or \; {\sigma_1}^2 \neq {\sigma_2}^2$. Pearson et al. (1930) considered the likelihood ratio test for this problem and Zhang et al. (2012) derived the exact distribution of the test statistic. Perng et al. (1976) proposed a combination of the pooled t-test statistic and the F-test statistic, which is called the combined test. Zhang et al. (2012) found from simulation that the likelihood ratio test and the combined test perform very similarly. Thus, we use the combined test by Perng et al. (1976) in the following simulations and derivations.

Zhang et al. (2012) suggested a three stage procedure to compare the means and the variances of the two populations. The three-stage procedure begins with the likelihood ratio test or the combined test to test the null hypothesis H_1. According to this three-stage procedure, if one fails to reject H_1, the conclusion is that there is no evidence that the two normal populations have different means or different variances. If, on the other hand, H_1 is rejected, one needs to check which caused H_1 to be rejected, unequal variances or unequal means. Then comes the second stage of testing for equal variances: $H_2\colon {\sigma_1}^2 = {\sigma_2}^2$ against $H_{2a}\colon\ {\sigma_1}^2 \neq {\sigma_2}^2$using the F-test. If the null hypothesis H_2 is not rejected, one concludes there is no evidence of unequal variances. One goes to the third stage and uses the pooled t-test to check: $H_3\colon \mu_1 = \mu_2$versus $H_{3a}\colon\ \mu_1 \neq \mu_2$. A flow chart of the procedure is in Figure 5.

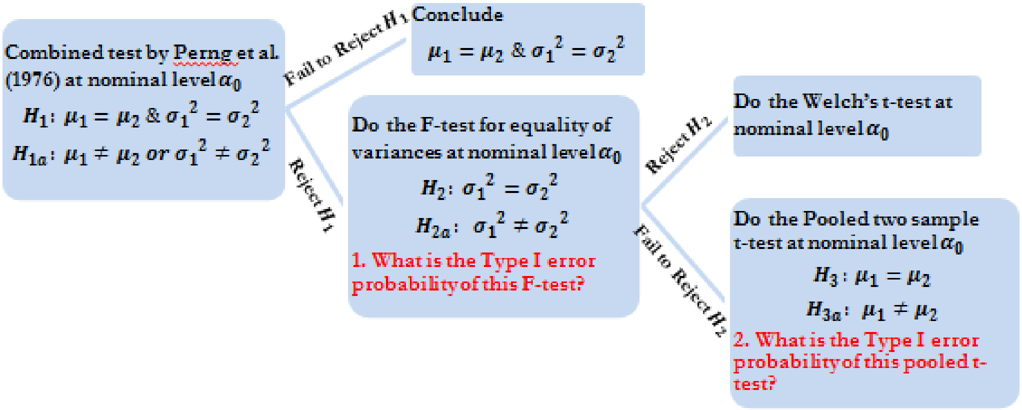
Combined test by Perng et al. (1976) at nominal level α_0
$H_1\colon \mu_1 = \mu_2$ & ${\sigma_1}^2 = {\sigma_2}^2$
$H_{1a}\colon \mu_1 \neq \mu_2$ or ${\sigma_1}^2 \neq {\sigma_2}^2$
Fail to Reject H_1
Conclude
$\mu_1 = \mu_2$ & ${\sigma_1}^2 = {\sigma_2}^2$
Reject H_1
Do the F-test for equality of variances at nominal level α_0
$H_2\colon {\sigma_1}^2 = {\sigma_2}^2$
$H_{2a}\colon\ {\sigma_1}^2 \neq {\sigma_2}^2$
1. What is the Type I error probability of this F-test?
Reject H_2
Do the Welch's t-test at nominal level α_0
Fail to Reject H_2
Do the Pooled two sample t-test at nominal level α_0
$H_3\colon \mu_1 = \mu_2$
$H_{3a}\colon\ \mu_1 \neq \mu_2$
2. What is the Type I error probability of this pooled t-test?

Figure 5: Flow chart of the three-stage procedure by Zhang et al. (2012)

The type I error of the F test in the second stage is a conditional one as H_1 has already been rejected. It stays at 0.478 for any sample sizes as long as m and n are equal. The derivation is provided in Appendix A. Compared to the nominal level of 0.05, the conditional Type I error rate has inflated by 860%. Furthermore, an increase in sample size does not help improve the conditional Type I error rate.

Now despite the fact that there is inflated Type I error rate in stage two, we move to the third stage and study the type I error rate of the pooled t-test. If the practitioner rejects H_1 and fails to reject H_2, he should perform a pooled t-test on the same pair of samples to check for equality of means. The type I error of the pooled t-test at the third stage is a conditional one based on the fact that H_1 is rejected and H_2 is not rejected.

A theoretical derivation of the conditional type I error rate is provided in Appendix B. It stays at 0.821 regardless of the sample size, while the unconditional type I error rate of the t-test being 0.05. The conditional Type I error rate had inflated by more than 1500%. Furthermore, an increase in sample size does not help control the conditional Type I error rate.

4. Discussion

In this paper, we approach the conditional type I error of some tests by simulation and theoretical derivation. The inflation in the conditional type I error rate is considerably large. Increasing the sample size does not help control the conditional type I error and sometimes inflate it even more. We do not recommend preliminary tests to the practitioners.

If we know that the main test statistic is robust to the violation of the assumption, we can perform the test of interest directly. For example, we know that the t-test statistic is robust to non-normal underlying distributions. We can directly perform the t-test on the sample(s). When unsure about meeting the underlying assumptions of the main test, the practitioner should look for an alternative test (may be a nonparametric test). For example, the F-test statistic is very sensitive to non-normality. If we are not sure that the data come from normal distributions, we suggest that the practitioner perform a nonparametric two-sample Wald test of equality of variances.

Further investigations are underway.

References

Albers, W., Boon, P. C. and Kallenberg, W. C. M. (2000), "The asymptotic behavior of tests for normal means based on a variance pre-test", Journal of Statistical Planning and Inference, 88, 47-57
Perng, S. K. and Littell, R. C. (1976), "A Test of Equality of Two Normal Population Means and Variances," Journal of the American Statistical Association, Vol.71, No.356, pp.968-971
Pearson, E. S. and Neyman, J. (1930), "On the Problem of Two Samples," Joint Statistical Papers (1967), eds. J. Neyman and E. S. Pearson, Cambridge: Cambridge University Press, pp. 99–115.
Rochon, J., Gondan, M. and Keiser, M. (2012), " To test or not to test: Preliminary assessment of normality when comparing two independent samples", BMC Medical Research Methodology, 12:81Schucany, W. R. and Ng, H. K. T. (2006), "Preliminary Goodness-of-Fit Tests for Normality do not Validate the One-Sample Student t", Communications in Statistics - Theory and Methods, 35: 12, 2275—2286
Zhang, L., Xu, X. and Chen, G. (2012), "The Exact Likelihood Ratio Test for Equality of Two Normal Populations", The American Statistician, 66:3, 180-184

Appendix

Appendix A

We derive the conditional Type I error of the F-test in the second stage theoretically. If the nominal level is α, then the conditional Type I error is given by the formula:

$$\text{Conditional Type I error of the F} - \text{test} = \frac{\text{P(reject the combined test \&reject the F} - \text{test}|\text{H}_1\text{is true)}}{\alpha}.$$

$$\text{P(reject the combined test \&reject the F} - \text{test)} = \alpha - \int_0^{\chi_4^2(\alpha)+2\log\alpha} f_X(x) \int_{-2\log\alpha}^{\chi_4^2(\alpha)-x} f_Y(y)dydx,$$

where X and Y both follow χ_2^2.
Now if we let $\alpha = 0.05$, $\text{P(reject the combined test \&reject the F} - \text{test)} = 0.026078$ and thus the conditional Type I error is

$$\text{Conditional type I error} = \frac{0.026078}{0.05} = 0.478. = 0.478.$$

Appendix B

The conditional Type I error of the pooled t-test in the third stage is given as:

$$\text{Conditional Type I error of the pooled t} - \text{test} = \frac{\text{P(reject the combined test \&pass the F} - \text{test \&reject the t} - \text{test}|\text{H}_1 \text{ is true)}}{\text{P(reject the combined test \&pass the F} - \text{test}|\text{H}_1 \text{ is true)}}.$$

$$\begin{aligned}\text{P(reject combined test \&pass F test \&reject t test)} \\ = 1 - \alpha - \int_0^{-2\log\alpha} f_Y(y) \int_0^{\chi_4^2(\alpha)-y} f_X(x)\,dx\,dy \\ - \int_{\chi_4^2(\alpha)+2\log\alpha}^{-2\log\alpha} \int_{\chi_4^2(\alpha)-y}^{-2\log\alpha} f_X(x)\, f_Y(y)\,dxdy,\end{aligned}$$

where X and Y both follow χ_2^2.

Now if we let $\alpha = 0.05$, the conditional Type I error is 0.821.