African Journal of Mathematics and Computer Science Research Vol. 2(4), pp. 056-062, May, 2009
Available online at http://www.academicjournals.org/AJMCSR

Full Length Research Paper

When does the pooled variance *t*-test fail?

Teh Sin Yin* and Abdul Rahman Othman

School of Distance Education, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia. E-mail: syin.teh@gmail.com or arothman60@yahoo.com.

Accepted 30 April, 2009

The pooled variance *t*-tests used prominently for comparing means between two groups is usually restricted with the assumptions of normality and homogeneity of variances. However, the violation of the assumptions happens in many real world data. In this study, the conditions where the pooled variance *t*-test would fail were investigated. The performance of the *t*-test was evaluated under different conditions. They were sample sizes, type of distributions (normal or non-normal), and unequal group variances. The Type I error rates and power of the pooled variance *t*-tests for different designs were obtained and compared. The results showed that the test failed dramatically when the group sample sizes were small and unequal with slight departure from homogeneity of variances.

Key words: Pooled variance, power, *t*-test, type 1 error.

INTRODUCTION

The *t*-test first proposed in 1908 by William Sealy Gosset, a statistician working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name) (Mankiewicz, 1975; O'Connor and Robertson, 1999-2004; Raju, 2005). Gosset had been hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes (O'Connor and Robertson, 1999-2004). Gosset devised the *t*-test as a method to cheaply monitor the quality of beer and the *t*-test was published in Biometrika in 1908 (Student, 1908). If the two population distributions assumed to have homogenous variances, the pooled variance *t*-test was used for comparison of means. Comparing means is often part of an analysis, for data arising in both experimental and observational studies. In fact, many of the applications of *t*-test have appeared in the text book and literature for the physical sciences and engineering (Dougherty, 1990; Kinney, 2002; Mendenhall and Sincich, 2007; Miller, 1995; Vardernan, 1994).

Although the pooled variance *t*-test statistical method was widely used in physical sciences and engineering, it is usually restricted with the assumptions of normality and homogeneity of variances. No doubt, the violation of the assumptions happens in many real world mdata. Hence, the purpose of this study is to investigate the conditions when the pooled variance *t*-test would fail. The performance of the *t*-test was evaluated under different conditions. They were sample sizes, type of distributions (normal or non-normal), and equal/unequal group variances. The Type I error rates and power of the pooled variance *t*-tests for different designs were obtained and compared.

This paper is organized as follows. In the second part, the pooled variance *t*-test, statistical power and effect sizes are reviewed. In the third part, the design specifications are discussed. The algorithms to obtain the Type I error rates and the power rates of traditional pooled variance *t*-test are given in the fourth part. The fifth part describes the results and discussion. The conclusion and some recommendation for future studies are in the final section.

METHODS

Pooled variance *t*-test

Let $X = X_1, X_2, \ldots, X_n$ and $Y = Y_1, Y_2, \ldots, Y_n$, where X and Y are independent and identically distributed with standard normal distribution. The difference between the two samples means, $\bar{X}$ and $\bar{Y}$ can be tested using the *t*-test. The hypothesis test is stated as:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_a : \mu_x - \mu_y \neq 0$$

If the two population distributions were unknown but

*Corresponding author. E-mail: syin.teh@gmail.com.

assumed to have the same variance, then their standard deviations say s_x and s_y can be pooled together. The pooled variance estimator of σ^2 is given by taking a weighted average of the sample variances, that is:

$$s_p^2 = \frac{s_x^2(n_x - 1) + s_y^2(n_y - 1)}{(n_x - 1) + (n_y - 1)} \quad (1)$$

Where;

s_p^2 = pooled sample variance from population X and Y,

s_x^2 = variance of sample X,

s_y^2 = variance of sample Y,

n_x = size of sample X,

n_y = size of sample Y.

Therefore, the test statistic for pooled variance t-test is defined as:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \quad (2)$$

Where;

μ_x = population mean for X,

μ_y = population mean for Y.

Two sample means need to be estimated before estimating the variances, thus, two degrees of freedom are lost, and the number of degree of freedom is given by

$$df = n_x + n_y - 2. \quad (3)$$

The p-value of one tail t-test is then defined as

$$p = P\left(t > t_0 / t \sim t_{\alpha, n_x + n_y - 2}\right) \quad (4)$$

Where;

t_0 is the calculated t value.

The null hypothesis is rejected when p-value $\leq \alpha = 0.05$.

The traditional pooled variance t-test is used to assess the fidelity of Type I error rates in experimental designs. In this study, other than the standard normal data, the skewed data was considered by using the property of

$$Y = \sum_{i=1}^{n} Z_i^2 \sim X_n^2$$

Where;

$$Z_1, Z_2, \ldots, Z_n \overset{i.i.d}{\sim} N(0,1).$$

Type I error (α) is defined as probability of rejecting H_0 when null hypothesis is true. In this study, the nominal level is $\alpha = 0.05$. The empirical Type I error rate which were close to nominal level considered as robust. The p-value of two tailed t-test considered in this study is given by

$$p = 2P\left(t > t_0 / t \sim t_{\alpha, n_x + n_y - 2}\right) \quad (5)$$

Statistical power

The definition of power is $1 - \beta$, where β is the probability of Type II error. Based on this definition, the power of a test is given by

Power = 1 – P(Accept H_0 | H_1 is true)
= P(Reject H_0 | H_1 is true). (6)

Essentially, the power of a test depends on three factors, which are the criterion of significance (α), sample sizes (n) and the effect sizes.

Effect size

The effect sizes of the statistic under the design specifications were examined to find the possible choices for μ_x and μ_y. The effect size index (d) for the two group's samples is defined as:

$$d = \frac{|\mu_x - \mu_y|}{\sigma} \quad (7)$$

Where

σ = standard deviation of either population (if they are equal) or the smallest standard deviation (if they are unequal).

According to Cohen (1988), the effect size is considered

Small when $d = 0.2$,
Medium when $d = 0.5$,
Large when $d = 0.8$.

The effect size index is used to choose values of population means for the true alternative hypothesis. Assume the $\mu_x > \mu_y$ so that $\mu_x - \mu_y$ would be positive, the possible choices of μ_x by assuming $\mu_y = 0$, out of the infinite values of $H_a : \mu_x \neq \mu_y$ were given in Table 1.

Table 1. Possible choices of μ_x and μ_y for all of the conditions.

Group sample sizes	Group variances	Pooled variances	Effect sizes, *d*	Range for $\mu_x - \mu_y$
{5, 15}	{1, 1}	1.000	$\frac{\mu_x - \mu_y}{0.5164}$	0.5, 1, 1.5, …
{5, 15}	{1, 2}	1.778	$\frac{\mu_x - \mu_y}{0.6886}$	1.0, 1.5, 2.0, …
{25, 35}	{4, 1}	2.241	$\frac{\mu_x - \mu_y}{0.3920}$	0.5, 1, 1.5, …
{25, 35}	{3, 1}	1.828	$\frac{\mu_x - \mu_y}{0.3540}$	0.5, 1, 1.5, …
{25, 35}	{1, 1}	1.000	$\frac{\mu_x - \mu_y}{0.2619}$	0.5, 1, 1.5, …
{25, 35}	{1, 43}	25.621	$\frac{\mu_x - \mu_y}{1.3255}$	1.0, 1.5, 2.0, …
{25, 35}	{1, 44}	26.207	$\frac{\mu_x - \mu_y}{1.3405}$	1.0, 1.5, 2.0, …
{20, 20}	{1, 1}	1.000	$\frac{\mu_x - \mu_y}{0.3162}$	0.5, 1, 1.5, …
{20, 20}	{1, 11}	6.000	$\frac{\mu_x - \mu_y}{0.7746}$	1.0, 1.5, 2.0, …
{20, 20}	{1, 18}	9.500	$\frac{\mu_x - \mu_y}{0.9747}$	1.0, 1.5, 2.0, …

Design specification

In evaluating the performance of the test procedures, three variables were manipulated. They were: (i) group sizes (ii) type of distribution – normal or non-normal and (iii) pairing of equal/unequal variances. We used small unequal group sizes (5, 15), large unequal group sizes (25, 35), and small equal group sizes (20, 20). For non-normal distributions, we chose the chi-square distributions with three degrees of freedom (χ_3^2) to represent a mild skewed distribution. In terms of variance heterogeneity, different ratios of positive pairing and negative pairing were examined until a break down condition was found for each design.

Unequal group sizes, when paired with unequal variances, can affect Type II error control for tests that compare the typical score across groups (Keselman et al., 1998, 2002; Othman et al., 2004; Syed Yahaya et al. 2006). Therefore, the sample sizes and variances were positively and negatively paired. A positive pairing occurs when the largest group size is associated with the largest group variance, while the smallest group size is associated with the smallest group variance. On the other hand, in a negative pairing, the largest group size is paired with the smallest group variance and the smallest group size is paired with the largest group variance.

This study was based on simulated data. In terms of data generation, the SAS generator RANDGEN (SAS Institute Inc., 2004) was used to obtain pseudo-random standard normal variates (RANDGEN(Y, 'NORMAL')) and to generate the chi-squared variates with three degrees of freedom (RANDGEN(Y, 'CHISQUARE', 3)).

The performances of the pooled variance *t*-test for the specific designs were collected. Here, the group means are set to {0, 0} as to reflect the null hypothesis of equal means ($H_0: \mu_x = \mu_y$). The study conditions that were considered:

Group sample sizes: {5, 15}, {25, 35}, {20, 20}.
Distributions: Normal, Chi-square 3 degrees of freedom (χ_3^2)
Group variances: {1, 1} and various pairings.

Based on these conditions there were 32 Monte Carlo studies. For each condition examined, 1000 data sets were obtained. The nominal level of significance was sets at α = 0.05. When the Type I error rates were robust, the performance of the test was further investigated by looking at the power of the test.

Algorithm for pooled variance t-test

The two algorithms used in this study were the traditional

Table 2. Type I error rates of two sample *t*-test with pooled variances when group sizes {5, 15}.

Distribution	Group Variances				
	Negative pairing		Equal pairing	Positive pairing	
	(3, 1)	(2, 1)	(1, 1)	(1, 2)	(1, 3)
Normal	0.1330	0.0910	.0580	0.0230	0.0210
χ_3^2	0.1150	0.0840	.0560	0.0250	0.0200

Note: Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

pooled variance *t*-test of comparing two means in terms of Type I error rate and power. A slight difference exists between the two algorithms where a modification of the *t*-test program for Type I error rates was made by replacing the null hypothesis with the alternate hypothesis when testing the power.

The algorithm to obtain the Type I an error rate of the traditional pooled variance *t*-test process is as follows:

1) Generate data to reflect the null hypothesis of equal mean is true.
2) Calculate *t*-test statistic based on data generated in Step 1.
3) Determine *p*-value of calculated *t*-test statistic in Step 2.
4) If *p*-value ≤ 0.05, then increase the number of count by one (count = count + 1). Initial count has been set equal to 0.
5) Repeat Step 1 to Step 4 for 1000 times.
6) Obtain the average Type I error rates by dividing count by 1000.
7) Repeat this simulation for 32 different conditions.

RESULTS AND DISCUSSION

The simulation results when the group sizes were small and unequal with various group variances were presented in Table 2. According to Bradley's (1978) liberal criterion of robustness, a test can be considered robust if its empirical rate of Type I error $\hat{\alpha}$, is within the interval $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. Thus, if the nominal level is α = 0.05, the empirical Type I error rate should be within the interval $0.025 \le \hat{\alpha} \le 0.075$. A procedure is considered liberal if its' Type I error rate is greater than the nominal level. Whereas, it is considered conservative if its' Type I error rate is less than the nominal level. Based on this criterion of robustness, the result of the *t*-test indicates failure with slight departure from homogeneity of variances. The negative pairing of group sizes and group variances generated larger Type I error than the positive pairing of group sizes and group variances. Regardless of the distributions, current results also reveal that positive pairings have produced conservative results and negative pairings have produced liberal results. This is in accord with findings in Othman et al. (2004) that positive pairings produced conservative values, while negative pairings generated liberal values. The best values were observed for equal variances which were close to the nominal rate.

Table 3 outlined the power values for the pooled variance *t*-test. For each of the robust Type I error rates in Table 1, there were 15 power values that tally with 15 values of μ_x from 0.5 to 7.5 in increments of 0.5.The highlighted entries are the power rates that higher than the benchmark power rate of 0.80 (Murphy and Myors, 1998). The three robust designs have reached high power at the same rate, that is, when μ_x equal to 2.0. This happened when the effect size is 3.9 for equal variances and for equal pairing, and when the effect size is 3.9 positive pairing of group size (5, 15) and variances (1, 2). It was not surprising as the variances are equal or slightly different between the two groups. This enables the procedure to reach the benchmark power faster.

The simulation results when the group sizes were large and unequal with various group variances were presented in Table 4. Regardless of the distributions, equal group variances satisfy the Bradley's robustness interval. However, large unequal group sizes fail faster with departure from homogeneity of variances for negative pairing. While, positive pairing of group sizes and group variances seems robust. The results were unaffected by distributions (normal or non-normal). Again, the results indicate that positive pairings have conservative Type I error rates and negative pairings have liberal Type I error rates.

The power values for the robust Type I error rates in Table 4 were presented in Table 5a (for normal distribution) and Table 5b (for non-normal distribution). For normal distribution in Table 5a, the design with homogeneous variance reached 0.8 fastest. This was followed by designs with negative pairings with group variances (3, 1) and (4, 1), respectively, which achieve power of 0.8 at the same time, that is, when μ_x equal to 1.5. That is, when their effect sizes equal to 3.8 and 4.2, respectively. When the different of group variance increase for designs with positive pairing, the power achieve the benchmark slowest before, For example, the design with group variances (1, 43) achieved the power of 0.8 at μ_x equal to 3.5 and effect size equal to 2.6. The comparison in terms of distributions (Table 5a and 5b), shows that the normal distribution has a slight edge over the χ_3^2 distribution in achieving the benchmark power level.

The simulation results when the two group sizes were small and equal with various group variances were presented in Table 6. The pooled variance *t*-test for normal distribution with equal sample sizes were not affected by differences in the population variances.

Table 3. Power of two sample *t*-test with pooled variances when group sizes {5, 15}.

Distribution					
Normal		Chi-Square (3), χ^2_3			
Group variance		Group variance			
(1, 1)		(1, 1)		(1, 2)	
d	Power	*d*	Power	*d*	Power
1.0	0.1640	1.0	0.1690	0.7	0.1080
1.9	0.4420	1.9	0.4940	1.5	0.3370
2.9	0.7810	2.9	0.7730	2.2	0.5790
3.9	0.9580	3.9	0.9410	2.9	0.8140
4.8	0.9970	4.8	0.9890	3.6	0.9150
5.8	1.0000	5.8	1.0000	4.4	0.9740
6.8	1.0000	6.8	0.9990	5.1	0.9940
7.7	1.0000	7.7	1.0000	5.8	0.9980
8.7	1.0000	8.7	1.0000	6.5	1.0000
9.7	1.0000	9.7	1.0000	7.3	1.0000
10.7	1.0000	10.7	1.0000	8.0	1.0000
11.6	1.0000	11.6	1.0000	8.7	1.0000
12.6	1.0000	12.6	1.0000	9.4	1.0000
13.6	1.0000	13.6	1.0000	10.2	1.0000
14.5	1.0000	14.5	1.0000	10.9	1.0000

Note: Bold values indicate power rate ≥ 0.80.

Table 4. Type I error rates of two sample *t*-test with pooled variances when group sizes {25, 35}.

Distribution	Group Variances					
	Negative pairing			Equal pairing	Positive pairing	
	(5, 1)	(4, 1)	(3, 1)	(1, 1)	(1, 43)	(1, 44)
Normal	0.0890	0.0710	0.0700	0.0520	0.0250	0.0190
χ^2_3	0.0980	0.0920	0.0690	0.0430	0.0340	0.0470

Note: Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

However, when the distribution was non-normal, Type I error fail when one of the variances is larger than the other by around 10 units.

The corresponding power values for the designs that produced robust Type I error rates in Table 6 were presented in Table 7. The comparison in terms of distributions shows that when group variances were equal, they achieved high high power at the same rate, that is, when effect size equals to 0.3.

Conclusion

This study showed that unbalanced designs with small sample sizes and slight departure from variance homogeneity produced Type I error rates that are liberal or conservative. In other words, the assumption of homogeneity of variances for the *t*-test is important when the group sample sizes were small regardless of the distributions. The power values indicate that the three robust designs arrived have same speed to achieve at high power at the same rate.

The Type I error rates also showed that unbalanced design with large sample sizes fails faster with departure from the equal variance case towards negative pairing, regardless of distributions. Therefore, even though the central limit theorem assured sample means to be distributed normally, testing of the homogeneity of variances is still needed to ensure that the variances are within the reasonable limits.

The last finding showed that pooled variance *t*-test for normal distribution with equal sample sizes were not affected by differences in the population variances. However, when the distribution was non-normal, Type I error fails when one of the variances is larger than the

Table 5. Power of two sample *t*-test with pooled variances when group sizes {25, 35}.

(a) Normal Distribution							
Group variance							
(4, 1)		(3, 1)		(1, 1)		(1, 43)	
d	**Power**	***d***	**Power**	***d***	**Power**	***d***	**Power**
1.3	0.2440	1.4	0.2960	1.9	0.4700	0.4	0.0460
2.6	0.6730	2.8	0.7680	3.8	0.9580	0.8	0.0800
3.8	0.9500	4.2	0.9760	5.7	1.0000	1.1	0.1590
5.1	0.9960	5.6	0.9990	7.6	1.0000	1.5	0.3060
6.4	1.0000	7.1	1.0000	9.5	1.0000	1.9	0.4470
7.7	1.0000	8.5	1.0000	11.5	1.0000	2.3	0.6530
8.9	1.0000	9.9	1.0000	13.4	1.0000	2.6	0.8020
10.2	1.0000	11.3	1.0000	15.3	1.0000	3.0	0.8920
11.5	1.0000	12.7	1.0000	17.2	1.0000	3.4	0.9520
12.8	1.0000	14.1	1.0000	19.1	1.0000	3.8	0.9770
14.0	1.0000	15.5	1.0000	21.0	1.0000	4.1	0.9940
15.3	1.0000	16.9	1.0000	22.9	1.0000	4.5	0.9980
16.6	1.0000	18.4	1.0000	24.8	1.0000	4.9	1.0000
17.9	1.0000	19.8	1.0000	26.7	1.0000	5.3	1.0000
19.1	1.0000	21.2	1.0000	28.6	1.0000	5.7	1.0000

(b) Chi-Square 3 df distribution, χ_3^2							
Group variance							
(4, 1)		(3, 1)		(1, 1)		(1, 43)	
d	**Power**	***d***	**Power**	***d***	**Power**	***d***	**Power**
1.3	0.2920	1.4	0.1690	1.9	0.0630	0.4	0.0670
2.6	0.7940	2.8	0.4390	3.8	0.1460	0.8	0.1570
3.8	0.9920	4.2	0.7860	5.7	0.2290	1.1	0.2470
5.1	0.9990	5.6	0.9460	7.6	0.3490	1.5	0.3600
6.4	1.0000	7.1	0.9860	9.5	0.4870	1.9	0.4880
7.7	1.0000	8.5	0.9980	11.5	0.6190	2.3	0.6360
8.9	1.0000	9.9	1.0000	13.4	0.7370	2.6	0.7130
10.2	1.0000	11.3	1.0000	15.3	0.8270	3.0	0.8300
11.5	1.0000	12.7	1.0000	17.2	0.9100	3.4	0.9060
12.8	1.0000	14.1	1.0000	19.1	0.9230	3.8	0.9190
14.0	1.0000	15.5	1.0000	21.0	0.9650	4.1	0.9670
15.3	1.0000	16.9	1.0000	22.9	0.9890	4.5	0.9740
16.6	1.0000	18.4	1.0000	24.8	0.9960	4.9	0.9910
17.9	1.0000	19.8	1.0000	26.7	0.9920	5.3	0.9940
19.1	1.0000	21.2	1.0000	28.6	0.9960	5.7	0.9950

Note: Bold values indicate power rate ≥ 0.80.

Table 6. Type I error rates of two sample *t*-test with pooled variances when group sizes {20, 20}.

Distribution	Group variances		
	(1, 1)	**(1, 11)**	**(1, 18)**
Normal	0.0660	0.0380	0.0510
χ_3^2	0.0590	0.0870	0.0880

Note: Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

Table 7. Power of two sample *t*-test with pooled variances when group sizes {20, 20}.

Distribution							
Normal						Chi-Square (3), χ_3^2	
Group variance						Group variance	
(1, 1)		(1, 11)		(1, 18)		(1, 1)	
d	**Power**	***d***	**Power**	***d***	**Power**	***d***	**Power**
0.2	0.3680	0.6	0.1000	0.5	0.0860	0.2	0.3730
0.3	0.8760	1.3	0.2440	1.0	0.1860	0.3	0.8740
0.5	0.9890	1.9	0.4700	1.5	0.3400	0.5	0.9940
0.7	1.0000	2.6	0.7240	2.1	0.5150	0.7	1.0000
0.8	1.0000	3.2	0.8770	2.6	0.6960	0.8	1.0000
1.0	1.0000	3.9	0.9700	3.1	0.8460	1.0	1.0000
1.2	1.0000	4.5	0.9900	3.6	0.9250	1.2	1.0000
1.3	1.0000	5.2	0.9980	4.1	0.9760	1.3	1.0000
1.5	1.0000	5.8	0.9990	4.6	0.9930	1.5	1.0000
1.7	1.0000	6.5	1.0000	5.1	0.9990	1.7	1.0000
1.8	1.0000	7.1	1.0000	5.6	1.0000	1.8	1.0000
2.0	1.0000	7.7	1.0000	6.2	1.0000	2.0	1.0000
2.2	1.0000	8.4	1.0000	6.7	1.0000	2.2	1.0000
2.3	1.0000	9.0	1.0000	7.2	1.0000	2.3	1.0000
2.5	1.0000	9.7	1.0000	7.7	1.0000	2.5	1.0000

Note: Bold values indicate power rate ≥ 0.80.

other by around 10 units.

On the other hand, the findings revealed that the Type I error rate was positively related to power with the alternative hypothesis $\mu_x > \mu_y$. The power values approached the benchmark level faster as the Type I error rates get closer to the nominal level. The larger the group sizes, the faster the design achieved the benchmark power level. This indicates the sample sizes are positively related to the power. The magnitude of the effect under alternate hypothesis, i.e. the effect sizes, also affects the power values. They too have a positive relationship.

This study can be extended to the performance assessment of any existing commonly used procedure in terms of Type I error rates and power rates.

ACKNOWLEDGMENT

The authors would like to acknowledge the work that led to this paper publication funded by the Fundamental Research Grant Scheme of the Malaysian Ministry of Higher Education, and supported by the Universiti Sains Malaysia Fellowship.

REFERENCES

Bradley JV (1978). Robustness?, British J. Math. Statist. Psych. 31: 321-339.

Cohen J (1988). Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York.

Dougherty ER (1990). Probability and Statistics for the Engineering, Computing, and Physical Sciences, Prentice Hall.

Keselman HJ, Wilcox RR, Algina J, Othman AR, Fradette K (2002). Trimming, transforming statistics and bootstrapping: Circumventing the biasing effects of heteroscedasticity and non-normality, J. Mod. Appl. Stat. Methods 1(2): 288-309.

Keselman HJ, Huberty CJ, Lix LM, Oleijnik S, Cribbie RA, Donahue B, Kowalchuk RK, Lowman LL, Petoskey MD, Keselman JC, Levin JR (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses, Rev. Edu. Res. 68(3): 350-386.

Kinney JJ (2002). Statistics for Science and Engineering, Addison Wesley, Boston.

Mankiewicz R (1975). The Story of Mathematics.Princeton, University Press.

Mendenhall W, Sincich T (2007). Statistics for Engineering and the Sciences, 5th ed., Pearson/Prentice Hall, Upper Saddle River, New Jersey.

Miller I (1995). Statistical Methods for Quality: With Applications to Engineering and Management, Irwin Prentice Hall, Miller.

Murphy KR, Myors B (1998). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Lawrence Erlbaum, Mahwah, New Jersey.

O'Connor JJ, Robertson EF (1999-2004). Student's *t*-test, MacTutor, History of Mathematics archive.

Othman AR, Keselman HJ, Padmanabhan AR, Wilcox RR, Algina J, Fradette K (2004). Comparing measures of the “typical” score across treatment groups, Br. J. of Math. Stat. Psychol. 57(2): 215-234.

Raju TN (2005). William Sealy Gosset and William A. Silverman: Two "students" of science, Pediatrics 116(3): 732-735.

SAS Institute Inc (2004). SAS OnlineDoc® 9.1.2., SAS Institute Inc., Cary, NC.

Student (1908). The probable error of a mean, Biometrika 6:1-25.

Syed Yahaya SS, Othman AR, Keselman HJ (2006). Comparing the "typical score" across independent groups based on different criteria for trimming, Metodološki Zvezki-Advances in Methodology and Statistics, 57(2): 49-62.

Vardernan SB (1994). Statistics for Engineering Problem Solving, PWS, Boston.