Comparison of Paired T-Test and Wilcoxon Signed Rank Test for Various Change Measures in Pre-Post Designs: A Simulation Study

Önce-Sonra Denemelerinde Farklı Değişim Ölçümleri İçin Eşleştirilmiş T-Testi ve Wilcoxon İşaretli Sıra Sayıları Testinin Karşılaştırılması: Bir Simülasyon Çalışması

Handan ANKARALI,[a]
Özge YILMAZ,[a]
Duygu AYDIN[a]

[a]Department of Biostatistics, Düzce University Faculty of Medicine, Düzce

Geliş Tarihi/*Received:* 04.02.2012
Kabul Tarihi/*Accepted:* 24.04.2012

Yazışma Adresi/*Correspondence:*
Handan ANKARALI
Düzce University Faculty of Medicine, Department of Biostatistics, Düzce, TÜRKİYE/TURKEY
hankarali@yahoo.com

ABSTRACT Objective: In a pre-post design, most commonly used change measure is simple difference (D), but percent change (PC) and symmetrized percent change (SPC) are not. We aimed to determine whether paired t-test and Wilcoxon signed rank test are suitable for PC and SPC according to type I-error and the power by a simulation. **Material and Methods:** Simulation study was performed by using normally distributed random numbers. Two samples which were called pre and post treatment measures with replacement were taken from a standard normal population which included 10.000 observations. Three different sample size, two different correlation degree, two different standard deviation and two different effect sizes were investigated for the purposes in simulation study. **Results:** The actual type-I error probabilities of paired t-test for PC were higher than nominal level except variances of pre and post measurements were near to homogenous. But they were observed significantly different from its nominal level in all test combinations for SPC. Actual type-I error of WSR for PC and SPC were found different from 5% when sample size is lower than 20. The powers of t-test and WSR test for PC and SPC were also low or moderate level in the simulation conditions. **Conclusion:** PC and SPC measures should only be used for descriptive purposes until appropriate test will found.

Key Words: Symmetrized percent change; percent change score; paired designs; Wilcoxon signed rank test

ÖZET Amaç: Önce-Sonra ölçümlerin alındığı deney tasarımlarında, genellikle işlem etkisini (değişimi) ölçmede yaygın olarak basit fark (D) kullanılır, Bu etkiyi tanımlamak için bazı durumlarda yüzde değişim (PC) veya simetrik yüzde değişim (SPC) de kullanılmaktadır. Bu çalışmada, PC ve SPC değerleri değişim ölçüsü olarak kullanıldığında, eşleştirilmiş t-testi ve Wilcoxon işaretli sıra sayıları testinin I. tip hata ve gücünün ne olacağı araştırılmıştır. **Gereç ve Yöntemler:** Normal dağılım gösteren veriler üretilerek simulasyon çalışması yapılmıştır. Muamele öncesi ve sonrası olarak adlandırılan ve standart normal dağılım gösteren populasyondan her defasında geri iadeli olarak seçilen 10000 gözlem içeren iki örnek dikkate alınarak hesaplamalar yapılmıştır. Amacı gerçekleştirmek için simulasyon çalışmasında, üç farklı örneklem genişliği, iki farklı korelasyon düzeyi, iki farklı standart sapma değeri ve iki farklı etki büyüklüğü değerlendirilmiştir. **Bulgular:** PC değerleri kullanılarak paired t-testi uygulandığında gerçekleşen I. tip hata yapma olasılıkları, once ve sonra ölçümlerinin varyansları homojen iken beklenen seviyede çıkmış diğer koşullarda beklenen seviyeden büyük değer almıştır. Fakat bütün incelenen kombinasyonlarda bu testin SPC ölçüsü için gözlenen I. tip hatası, beklenen düzeyinden oldukça yüksek bulunmuştur. Örneklem genişliği 20'nin altında olduğu durumlarda PC ve SPC ölçüleri için WSR testinin gerçekleşen I. tip hatası %5'ten oldukça farklı bulunmuştur. Çalışmada denenen bütün simulasyon koşullarında, PC ve SPC ölçüleri için paired t-testinin ve WSR testinin gücü düşük vey orta düzeyde bulunmuştur. **Sonuç:** Bu sonuca gore değişimi ifade etmede PC ve SPC ölçülerinin tanımlayıcı amaçlar için kullanılabileceği söylenebilir.

Anahtar Kelimeler: Simetrik yüzde değişim; yüzde değişim skoru; eşleştirilmiş tasarım; Wilcoxon işaretli sıra sayıları testi

Turkiye Klinikleri J Biostat 2012;4(2):55-9

In some experimental researches, the data are obtained two times when pre and post treatment measures are taken from same subjects. In such designs, pre-treatment measurements are generally accepted as control values and these designs called pre-post design, paired design or self-controlled design. It is widely used in medical researches since this type designs need fewer subjects and eliminate misleading results caused by variations between subjects. However, measures for detecting the treatment effects must be chosen correctly in a pre-post design. These measures are called change measures and simple difference (D) is the most commonly used change measures in statistics. Simple difference is calculated by subtracting pre-treatment measurement values from post-treatment measurement values. To determine treatment effects either paired t-test or Wilcoxon Signed Rank test (WSR) are used depending on shape of the distribution of D in paired designs. In all statistical package programs, simple difference are calculated as change measure and one of above mentioned test statistics are used to analyze the data. These issues were argued in a book.[1]

Various change measures were defined and their usages in different designs were discussed in literature by some authors.[2,3] Another frequently used change measures are Ratio (R), Percent Change (PC) and Symmetrized Percent Chance (SPC). R is calculated by dividing pre-treatment measurement to post-treatment measurement and if pre- and post-treatment values are same, R equal to 1. In this case the mean of R values are tested against "1". To calculate PC measure, difference of pre- and post-treatment measurements is divided by pre-treatment measurements and multiplied by 100. Since the result is "0" when pre- and post-treatment measures are equal in this formula, PC is tested against "0". Another change measure SPC is developed due to PC which has skewed distribution. It is calculated by following formula; [(pre-post)/(pre+post) x 100]. Some authors have been suggested that sometimes these change measures may be able to express the treatment effects better than D.[4,5]

The selection of suitable change measure has become a more important issue when within variance of pre-treatment values is large. For example, we assume that the acne numbers reduced from 20 to 10 and 10 to 0 in a drug research for treatment of acne on the face. Even though there is same amount of change, according to success of the treatment this change may be explained differently. Changing from 10 to 0 may be shown that the treatment is completely effective on the acne but changing from 20 to 10 is not. If PCs are calculated for these subjects, values of 50% and 100% are obtained for the treatment effect, respectively. It is possible to find so many examples in medical researches. Two different change measures are used together by most researchers at their study because they cannot decide which change measure will be more appropriate for the aims and there is no published research on this subject to help researchers. Vickers used PC and some modifications of PC, such as SPC, for descriptive purposes or hypothesis testing.[5] However, some authors claimed that PC and SPC have not normal distribution and thus, using parametric tests is not acceptable and nonparametric alternatives should be used in hypothesis tests.[3,4]

In this study, we aimed to determine suitability of paired t-test and WSR for D, PC and SPC according to actual type-I errors and the power with a simulation study which is constitute for different conditions, such as various sample sizes, effect sizes, correlation degree and variance combination in paired designs.

MATERIAL AND METHODS

In our simulation study, two orjinal samples were taken from a standard normal population which included 10.000 observations. These two samples which are named as populations were called pre and post treatment measures. Resample *with replacement* a bootstrap samples of various sizes (7,10 and 20 in Table 1) were taken 5000 times the original sample. Some numbers in the original sample may be included several times in the bootstrap sample. Others may be excluded.[6]

Three different sample size, two different correlation degree and two different standard deviation combinations were investigated for the purposes in simulation study. In addition two dif-

TABLE 1: Properties of populations and samples in the simulation study.

Correlation Coefficient	n*	Standardized effect sizes		
		$\mu_1:\mu_2$= 0:0 (Δ=0)	$\mu_1:\mu_2$= 0:1 (Δ=1)	$\mu_1:\mu_2$=0:2 (Δ=2)
ρ= 0.10	7	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12
	10			
	20			
ρ= 0.90	7	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12	$\sigma_1:\sigma_2$=1:1 $\sigma_1:\sigma_2$=6:12
	10			
	20			

* Bootstrap sample size.

ferent effect sizes were taken into consideration while evaluating the powers of paired t-test and WSR test. So, we evaluated 36 different combinations in simulation study. Twelve of them were used to calculate actual type-I error and remains were used for the power of tests. Each combination was repeated 5000 times (Table 1). In each sample, taken from population, "D", "PC" and "SPC" were calculated to express changes between pre and post treatment measurements and then test statistics were calculated by using these measures. Mean values of type I errors and power values of the tests which were obtained from 5000 trials were computed as in Table 2 and Table 3. To calculate the powers of tests pre- and post-treatment measures were taken from two different standard normal populations which had different means. At the beginning of the simulation study, nominal type-I error probability was accepted as 5%. FORTRAN POWER STATION DEVELOPER STUDIO IMSL subroutines were used for all calculations.

RESULTS

Actual type-I error and power of the tests were used to evaluate which test statistic is more suitable for D, PC and SPC measures. Actual type-I error of paired t-test for D preserved its nominal level when the correlation between pre and post treatment is low and sample size is small or moderate (Table 2). When the variances of pre and post treatment measurements were homogenous or near to homogenous, similar results were obtained. However, actual type-I error of paired t-test for PC preserved its nominal

TABLE 2: Mean actual type-I error probabilities of tests for D, PC, SPC when the null hypothesis is true about treatment effect ($\mu_1:\mu_2$= 0:0).

Test	Sample Size (n)	Correlation coefficient	Observed Type I Error Probabilities of Tests (%)					
			ρ= 0.10			ρ= 0.90		
		Standart deviation of pre-post measures	D	PC	SPC	D	PC	SPC
Paired t-test	7	*1:1*	5.40	12.80	2.18	4.23	1.90	1.75
		6:12	5.10	4.17	9.40	5.40	2.40	3.70
	10	*1:1*	4.47	11.83	1.97	5.60	1.97	1.73
		6:12	5.10	5.37	10.77	4.90	1.70	2.90
	20	*1:1*	5.00	16.30	2.12	5.23	2.43	1.80
		6:12	4.40	5.50	11.70	4.50	2.25	3.65
Wilcoxon Sign Rank Test	7	*1:1*	3.40	12.86	2.12	2.60	2.20	1.90
		6:12	3.33	4.30	8.47	4.30	5.00	5.30
	10	*1:1*	4.47	11.83	1.97	3.93	2.10	2.10
		6:12	4.17	4.20	4.73	4.03	2.03	2.47
	20	*1:1*	4.38	4.38	4.20	4.80	7.00	7.00
		6:12	4.40	4.38	4.20	3.86	7.00	7.00

TABLE 3: Mean power of tests for D, PC, SPC when the effect sizes are 1SD and 2SD.

Test	Effect size	Sample Size (n)	Correlation coefficient	Observed Power of Test (%)					
				ρ= 0.10			ρ= 0.90		
			Standart deviation of pre-post measures	D	PC	SPC	D	PC	SPC
Paired t-test	Δ=1.0σ	7	*1:1*	40.3	6.5	4.85	58.83	2.3	2.27
			6:12	5.6	4.2	7.98	5.17	1.93	2.90
		10	*1:1*	61.73	5.57	4.7	77.93	1.97	2.3
			6:12	6.23	3.9	10.67	5.33	1.93	2.67
		20	*1:1*	91.67	7.43	6.1	98.4	2.37	1.97
			6:12	6.43	4.67	10.1	7.93	2.43	3.53
	Δ=2.0σ	7	*1:1*	91.73	3.47	27.6	98.87	2.47	3.53
			6:12	5.7	4.4	9.2	6.37	2.37	2.8
		10	*1:1*	99.3	3.37	25.5	100	2.63	3.1
			6:12	7.2	4.9	9.1	6.7	2.2	3.83
		20	*1:1*	100	4.37	26.2	100	1.97	3.97
			6:12	8.03	4.03	12.63	10.1	2.13	2.96
Wilcoxon Sign Rank Test	Δ=1.0σ	7	*1:1*	29.3	6.7	5.95	44.37	2.3	2.73
			6:12	3.6	4.24	7.86	3.4	2.83	2.8
		10	*1:1*	55.3	3.03	3.1	70.47	2.0	2.23
			6:12	4.73	2.4	5.33	4.03	1.93	2.2
		20	*1:1*	89.3	7.03	7.03	97.67	7.0	7.0
			6:12	5.63	7.1	7.1	6.73	7.03	7.03
	Δ=2.0σ	7	*1:1*	93.43	2.83	32.93	96.47	2.77	4.27
			6:12	3.3	4.57	9.1	4.2	2.63	3.4
		10	*1:1*	98.6	2.2	19.0	99.9	2.2	2.57
			6:12	5.87	3.23	3.7	5.17	2.13	2.17
		20	*1:1*	100	7.0	10.23	100	7.0	7.0
			6:12	7.1	7.03	7.0	9.3	7.0	7.03

level only when variances of pre and post measurements were near to homogenous. In all other condition, actual type-I error probabilities were higher than nominal level (Table 2). Actual type-I error probabilities of paired t-test for SPC were observed significantly different from its nominal level in all test combinations. Actual type-I error of WSR for D preserved its nominal level, especially when the sample size higher than 10. Although, its actual levels for PC and SPC were found different from 5% when sample size is lower than 20 (Table 2).

Actual type-I error of paired t-test for D was preserved at level of 5%, when the correlation level between pre and post treatment was high. However actual level for PC and SPC in the same condition was too lower than nominal level. The

results of actual type-I error of WSR for D, PC and SPC were similar to above mentioned results when the correlation level is high (Table 2).

In addition, two standardized effect sizes were used for calculation of the tests power and then the powers were compared with each other (Table 3).

When the effect size is 1 standard deviation and variances are homogenous, the power of paired t-test for D was moderate levels and it was increasing as parallel with correlation degree. In other word, when the variances of pre and post measurements are different from each other, the power of t-test for D was quite low. However, the powers of t-test for PC and SPC were also low and they were independent from sample size and homogeneity of variance for above mentioned conditions. In the same conditions, the power of WSR was very low especially for PC and SPC. When the effect size is two standard deviation and variances are homogenous, the power of t-test and WSR for D were obtained larger than small effect sizes' condition. In addition, the powers of t-test and WSR for SPC at low correlation levels were little increased, but they were not changed in other conditions (Table 3).

DISCUSSION

Simple difference is most frequently used change measure for prediction of treatment effect in simple paired designs, because existing statistical tests are developed for the simple difference. However, some researchers claim that sometimes PC or its modifications can be more suitable change measures than D for prediction of treatment effect.[4,7] If the between subjects variation in pre treatment measurements is large, PC measure may be more suitable for treatment effect than D. In this case, either change measures are used together in statistical analysis or PC, R or SPC are only computed for descriptive purposes. Parametric tests were used to analyze PC or R.[5] Because the change measures values, except D, have not got normal distribution, some other researchers suggested that nonparametric tests should be used for analyzing these type measures.[3,4,8] However, detailed information has not found in literature about which statistical test is suitable for change measures which have non-normal distributions. Choosing a statistical test to analyze various change measures data is one of the most difficult decisions that a researcher makes in the experimental process because choosing the wrong test may lead to the wrong conclusion. Sometimes PC and SPC are superior to describe treatment effect than D. In this study, we evaluated whether statistical tests which are used for D are also suitable for PC and SPC regarding type-I error and power or not. According to our results both paired t-test and WSR give similar and good results for D and these findings are consistent with a previous study performed.[7] However we observed that actual type-I error and power of these two tests for PC and SPC are significantly different from nominal level.

CONCLUSION

PC and SPC give more information than D about treatment effect, paired t-test and Wilcoxon signed rank tests are not suitable for PC and SPC. We conclude that new test statistic should be developed for PC and SPC and these measures should only be used for descriptive purposes until suitable test will found.

REFERENCES

1. Bonate LP. Analysis of Pretest-Posttest Designs. 1st ed. New York: Chapman & Hall/CRC; 2000. p.1-203.
2. Dimitrov DM, Rumrill PD Jr. Pretest-posttest designs and measurement of change. Work 2003;20(2):159-65.
3. Ercan I, Yazici B, Yang Y, Ozkaya G, Cangur S, Ediz B, et al. Misusage of statistics inmedical research. Eur J Intern Med 2007;4(3):128-34.
4. Berry DA, Ayers GD. Symmetrized percent change for treatment comparisons. Am Stat 2006;60(1):27-31.
5. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
6. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. 1st ed. New York: Chapman & Hall/CRC; 1994. p.1-456.
7. Blair RC, Higgins JJ. Comparison of the power of the paired samples t-test to that of Wilcoxon's signed-ranks test under various population shapes. Psychol Bull 1985;97(1): 119-28.
8. Kaiser L. Adjusting for baseline: Change or percentage change? Stat Med 1989;8(10): 1183-90.