Statistics and Application {ISSN 2452-7395(online)}
Vol. 18, No. 1, 2020 (New Series), pp 141-156

Statistical Properties of Sampling Distributions of Different Test Statistics for Different Measures of Change and a New Test: Simulation Study

Handan Ankarali [1], Sengul Cangur [2] and Seyit Ankarali [3]
[1]*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Istanbul Medeniyet University, 34700, Istanbul, Turkey*
[2]*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Duzce University, 81620, Duzce, Turkey*
[3]*Department of Physiology, Faculty of Medicine, Istanbul Medeniyet University, 34700, Istanbul, Turkey*

Received: 04 November 2019; Revised: 18 May 2020; Accepted: 30 May 2020

Abstract

The aim of this study is to analyse the distribution characteristics of four different test statistics, namely the Mean/Standard Error of Mean (Mean/SEmean), Median/Interquartile Range (Median/IQR), Trimmed Mean/Standard Error of Mean (TrMean/SEmean), and Trimmed Mean/Interquartile Range (TrMean/IQR), which can be used to test two measures of change, namely percent change (PC) and modified symmetrised percent change (MSPC). To ensure the selection of suitable test statistics using the two measures of change, the observed type-I errors and powers of the test statistics have been computed. Results demonstrate that the sampling distributions of the four different test statistics by using PC values exhibit skewness. The Mean/SEmean statistic for the MSPC measure exhibits a two-peak value and platykurtic distribution, while the TrMean/SEmean statistic shows a leptokurtic distribution. The Median/IQR test yields robust and powerful results, especially for large sample sizes. This new statistical measure is referred to as the *HS* test.

Key words: Pre-post designs; Paired samples t-test; Change measures; Per cent change; Symmetrised per cent change; Trimmed mean.

1. Introduction

A one-sample paired design is commonly used to assess treatment effects in clinical research. In this design, the measure of the change in continuous data is based on simple difference, relative change or the ratio of post-treatment to baseline. The appropriate measure of change is selected according to the ease of interpretation and ease of data analysis generally. In this process, it is necessary to take into account both in terms of clinical significance and statistical features. The clinical perspective is based on the ease of interpretation, while the statistical perspective comes from the ease of conducting data analysis based on the normal distribution. (Tornqvist *et al.*, 1985; Zhang and Han, 2009; Yamabe *et al.*, 2012). In some cases, it is not easy to decide which measure of change to use, as simple difference may not always identify the change correctly. For example, when the difference observed in an individual with a large initial value is the same as the difference observed in an individual with a small initial value, this difference does not have the same

Corresponding Author: S. Cangur
e-mail: sengulcangur@duzce.edu.tr

meaning biologically. Suppose that a particular drug, which is being evaluated to determine its effectiveness at reducing facial acne, has lowered the number of acne sites from 10 to 0 in one person and from 20 to 10 in another person. If only the difference is taken into consideration, it could be concluded that the drug has the same effect on both persons; however, the clinician would explain that this numerical change does not reflect the actual biological change, and that a more clinically meaningful change has occurred in the first person. Moreover, Waleekhachonloet *et al*. (2007) stated in their study that diet effect is the difference in a person's weight before and after the diet. In the work of these authors, treatment effect is generally expressed as simple difference. This is because the statistical properties of the distribution of simple difference are well known and can be tested using a t-test, which produces strong results. For their part, in a dietary trial with similar content, Kim *et al.* (2009) expressed weight loss as a percent change. In similar clinical studies, researchers argue that percent change better reflects the treatment effect from a clinical perspective compared to simple difference and that its clinical meaning is more accurate. In addition, individuals having large initial values may have large differences, while individuals having very small initial values may have large post/pre ratios. In the literature, various measures of change have been proposed with the aim of eliminating the influence of these differences in individuals' initial values on pre-post change values. There is no in-depth research or consensus on these measures of change (Berry and Ayers, 2006; Wilkinson, 1999). Therefore, it is necessary to examine the characteristics and performances of different measures of change in more detail.

The percent change (PC) and ratio (R) measures are tested with non-parametric tests due to the fact that they often exhibit a highly skewed distribution (Vickers, 2001; Koti, 2001). The other measure of change is symmetrised percent change (SPC) with good statistical properties (Berry and Ayers, 2006; Yamabe *et al*., 2012). However, due to the difficulty associated with interpreting SPC, a very limited amount of research on the matter has been carried out (Yamabe *et al.,* 2012; Berry and Ayers, 2006; Koti, 2001). Another measure of change is modified symmetrised percent change (MSPC) with good statistical properties. MSPC is defined as the mean of two values for a numerator. When we regard the difference in denominator as an index of variation, the measure may be considered as a variation standardised by mean, such as a coefficient of variation (Yamabe *et al.,* 2012). This measure is calculated by eliminating the effect of pre-post mean in the difference between the pre-post measures. This measure, which is similar to the coefficient of variation, is easy to interpret (Yamabe *et al.,* 2012).

The aims of this study are as follows:

a) To test the significance of the two measures of change, PC and MSPC, with four test statistics;
b) To obtain the sampling distributions and 95% confidence limits (critical tail values of the distributions) of these statistics;
c) To calculate the observed type-I errors and power of four tests of each measure of change.

2. Methods

2.1. Measures of Change used in the study

The three most frequently used measures of change in medical research are investigated in this study. The formulas for these measures are provided below.

I. Simple Difference (D) = $(Pre - Post)$

- This measure is commonly used and understood in the literature. It is used in this study for the validation of simulation data, as well as to compare the type I error and power of the t-test calculated for D with other measures' results.

II. Percent Change (PC) = $\frac{(Pre-Post)}{Pre} \times 100$ or PC= $\frac{(Post-Pre)}{Pre} \times 100$

- Although pre- and post- data have the same units, PC values are often unitless or expressed as percentages. This measure means "the proportion of increase (or decrease) for pre-value", and is preferable from a clinical perspective due to its ease of interpretation. However, it exhibits a positive skew when post-data is much larger than pre-data, even if pre- and post-data are normal.

III. Modified Symmetrised Percent Change (MSPC) = $\frac{(Pre-Post)}{\frac{(Pre+Post)}{2}} \times 100$

- Direct interpretation of the SPC value is difficult, but the distribution of SPC maintains symmetry without regard to shape, scale and correlation parameters of the distribution for pre- and post-data (Yamabe *et al.,* 2012). The numerator of the MSPC measure is pre-post difference, while its denominator is a mean of the pre and post values such as a coefficient of variation. This measure is called variability (%) in the bioanalytical field and is used to evaluate the level of reproducibility of assay results using incurred samples (Mario *et al.*, 2007; Douglas *et al.*, 2009). When SPC or MSPC are used to analyse the data, the results can be interpreted after transforming to the robust percent change (RPC) proposed by Berry (1989).

2.2. Test statistics used in the study

Four statistics are investigated in the hypothesis test for the two different measures of change (PC, MSPC) used in the present study. These statistics are as follows:

- Mean/Standard Error of Mean (Mean/SEmean)
- Median/Interquartile Range (Median/IQR)
- Trimmed Mean/Standard Error of Mean (TrMean/SEmean)
- Trimmed Mean/Interquartile Range (TrMean/IQR)

The empirical sampling distributions of these statistics have been obtained and the statistical properties of these distributions determined.

Trimmed mean is average of the remaining values after removing the smallest p% and the largest p% of the values. We have accepted p as equal to 10% in our calculations.

2.3. Simulation study

The present research employs a quasi-Monte Carlo simulation study. The data are generated from a bivariate standard normal distribution. The random variables X_i (i =1,2) denote the response of pre- and post-data following a bivariate standard normal distribution; moreover, the variables satisfy $(X_1, X_2) \sim BN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, where μ_i denotes the location parameters, σ_i the scale parameters, and ρ the correlation parameter between two random variables of pre- and post-data. For the observed type-I error, we generate (X_1, X_2) from$\sim BN(0, 0, 1, 1, \rho)$, while for the power of the tests, we generate (X_1, X_2) from$\sim BN(0, \mu_2[0.5, 1.0, 1.5, 2.0], 1, 1, \rho)$. A total of 18 conditions are evaluated considering six different sample sizes (10, 30, 60, 120, 500, 10,000) and three different correlations between the variates named as pre- and post-measures (0.3, 0.6, 0.9). Each condition is repeated 60,000 times.

In the first step, the results of the Mean/SEmean statistic for D are examined. The purpose of this step is to validate the results produced by the simulation data. The probability of type-I error and power values of the Mean/SEmean statistic are used for this. These values are calculated considering the critical values of the t-distribution at the 5% level (see: Table 3). In the second step, empirical sampling distributions are obtained for the four statistics with regard to PC and MSCP, and the critical tail values of these distributions are determined at the 5% level (see: Figure 2-3 and Table 1-2). In the third step, the observed probability of the type-I error and the power of the Mean/SEmean statistic for D are calculated according to the critical values of the t-distribution in Table 3. In addition, the type-I error and the power values of the Median/IQR and TrMean/IQR statistics selected for MSPC are calculated using the critical values of the empirical sampling distributions in Table 4-5. Subsequently, the results of the selected tests are compared with each other (see: Table 6).

The histograms of the test values from 60,000 trials for each condition have plotted. The lower and upper limit values of the 95% confidence intervals of these distributions are presented in the tables below. In addition, the skewness and kurtosis coefficients are calculated. The dataset used in this phase is referred to as the learning sample. The observed probability of type-I error and the power values of the selected tests are calculated from 40,000 trials by resampling from the bivariate distribution. This dataset is referred to as the test sample. To calculate the power of the tests, the standardised differences between the populations named as pre- and post- are accepted as 0.5SD, 1SD, 2SD and 2.5SD respectively.

The FORTRAN programming language, along with Microsoft Power Station Developer Studio, IMSL Library and the Minitab program (version 16.0) are used in the simulations and to obtain the graphics.

3. Results

3.1. Validation of simulation

In order to carry out the validation of the simulation program, samples (10,000 observations) are taken from a population with a bivariate normal distribution. There is a correlation of 0.60 among them, and differences have been calculated for each sample. The Mean/SEmean test statistic has been used for the differences It has been determined that the distribution of this statistic exhibits a t-distribution; this has been done by calculating the

descriptive statistics and plotting the shape of the sampling distribution following 60,000 trials. The mean of the simple difference is 0.0003±0.9911 and its first, second and third quartiles are -0.6733, -0.0011 and 0.6526 respectively. The mean of the test statistic is close to zero, while the standard deviation is close to one (-0.0042±1.0000), and its first, second and third quartiles are -0.6761, -0.0019 and 0.6753 respectively. Furthermore, when n = 10,000, the 95% confidence limits of this distribution are observed to be ±1.96 (Figure 1). All these results show that the simulation program produces reliable and accurate results.

The PC and MSPC measures have been simultaneously calculated in the validated simulation. In the hypothesis test established for these measures, the sampling distributions at the end of 60,000 trials have been obtained for the four test statistics defined in the methods section.

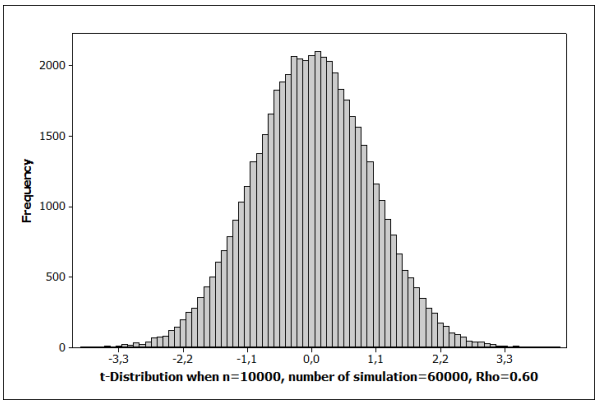

Figure 1: Empirical sampling distributions of Mean/SEmean statistics when n=10,000, number of simulation is 60,000 and Rho=0.60 (Close to standard normal distribution)

3.2. Results of percent change (PC)

The sampling distributions of the four statistics for PC in different conditions are presented in the Annexure A as Figure 2. Sampling distributions of Mean/SEmean statistics are skewed to the left and have two peak values (Figure 2a). The sampling distributions of the TrMean/SEmean statistics are skewed to the left but have only a single peak value (Figure 2b).

The sampling distributions of the Median/IQR and TrMean/IQR statistics have been found to have a single peak and to be approximately symmetrical for all correlation coefficients in large sample sizes. However, the sampling distributions obtained with small sample sizes exhibit a negative skew (Figures 2c and 2d).

The 95% confidence intervals of the sampling distributions of the four test statistics are considered as critical values; these values are presented in Annexure B, Table 1 for all conditions. It has been determined that the sampling distributions of Mean/SEmean and TrMean/SEmean statistics do not have symmetrical distributions under any conditions, while the sampling distributions of the Median/IQR and TrMean/IQR statistics are approximately symmetrical in population (n = 10,000).

3.3. Results of modified percent change (MSPC)

The difference between the pre-post measures has been expressed as MSPC. When the null hypothesis establishing that this difference is not significant has been tested with the

Mean/SEmean statistic, the lower and upper limit values of the test values calculated for the sampling distribution have found to be ±1.887 for n = 10 with 2.5% error probability for each tail (5% in total). These values have been determined to be ±1.710 when n = 10,000. The distribution of this statistic has been found to be symmetrical, but with two peaks. The distribution of the TrMean/SEmean statistic has been found to be symmetrical with a single peak value. The values of this distribution have been calculated as ±1.622 for n = 10 and ±0.081 for n = 10,000 with 2.5% type-I errors.

When compared with the first test statistic, it can be seen that the values of variance have a narrower range; in other words, the variance of the sampling distribution of this test is smaller. The shapes of the distributions of the Median/IQR and TrMean/IQR statistics have been also found to be perfectly symmetrical and single-peaked. It has been determined that as the sample size increases, the variances of the distributions decrease and the values of the ranges become narrower than is the case for the other two statistics. The upper and lower limits of the 95% confidence interval of the sampling distribution have been found to be ±0.430 for n = 10, ±0.015 for n = 10,000 for the Median/IQR statistic and ±0.470 for n = 10 and ±0.029 for n = 10,000 for the TrMean/IQR statistic. All these results demonstrate that the sampling distributions of the four test statistics are not t-distributions (Annexure B, Table 2).

The skewness values of the four sampling distributions in all conditions are close to zero. The kurtosis values of the Mean/SEmean statistic range from –0.5 to –1, depending on the sample size, and it has a platykurtic distribution (Figure 3a). The kurtosis values of the TrMean/SEmean statistic are large and it has positive values between 2 and 6.5 with increasing sample size (Figure 3b). However, the kurtosis values of the Median/IQR and TrMean/IQR statistics are very close to zero. These two distributions are also symmetrical (Figure 3c and 3d).

3.4. Suitable test statistics, observed type-I error and power

Since the Median/IQR and TrMean/IQR test statistics using MSPC exhibit a single peak and a perfectly symmetrical distribution, the critical values reported in Table 2 are used to calculate type-I errors and test powers.

After the critical values have been obtained, 40,000 new samples for each condition in the simulation have produced; these samples are named the test samples. Firstly, a t-test is used for D, and the type-I errors and powers of the test for each simulation condition are calculated as in Annexure B, Table 3. These calculations are made to validate the results of the simulated data. According to Table 3, the probabilities of the observed type-I error of the t-test are around 5%, and the deviation from 5% is not significant. When the sample size is around 10, it is found that the t-test has a low power, and it is also determined that this value increases slightly as the effect size increases. Furthermore, it is determined that the t-test yields fairly strong results in other sample sizes when the effect size is 0.5 standard deviations; moreover, when the effect size increases to 1 standard deviation and above, the power values become 100% or very close to that. Since all these results are known and expected results for the t-test, it can be determined that the simulation program produces valid data.

The majority of the probabilities of making a type-I error for the Median/IQR test statistic calculated for 18 different conditions are found to be very close to 5%, with the lowest being 4.80% and the highest being 5.63% (Annexure A, Table 4). The results of the

TrMean/IQR statistic are also found to be similar: most of the values are close to 5%, with the lowest being 4.63% and the highest being 5.24% (Annexure A, Table 5).

The Median/IQR and TrMean/IQR statistics are found to have high power only for very large sample sizes with an effect size of 0.5SD. While the Median/IQR statistic has sufficiently high power when the effect size is 1SD and $n > 60$, the TrMean/IQR statistic reaches an adequate level of power when the effect size is 1SD and $n \geq 500$. Moreover, the Median/IQR statistic reaches the desired level of power with an effect size of 1.5SD when $n \geq 30$, while the TrMean/IQR statistic reaches the desired level of power with the same effect size when $n \geq 120$. Finally, it is determined that the Median/IQR and TrMean/IQR statistics have the desired level of power with an effect size of 2SD when $n \geq 30$, and that the Median/IQR statistic reaches 100% power level when $n \geq 30$ (see Tables 4 and 5). In light of these results, the two new tests proposed using MSPC can be identified as robust tests like the well-known *t*-test.

In particular, it can be concluded that the Median/IQR test yields powerful results when evaluating differences that show deviations of 1SD or larger, while the TrMean/IQR test produces powerful results in cases where the differences are 1.5SD or larger. Since, in many conditions, the Median/IQR test is more powerful than the TrMean/IQR test, it can be concluded that the Median/IQR test could be used as an alternative test to the paired samples *t*-test in some conditions (Annexure B, Table 6). The Median/IQR test statistic can be referred to as the *HS* test, as it has been used for the first time in this study.

4. Discussion

Ankarali and Ankarali (2009) applied only the Mean/SEmean statistic for PC values in their study and found that the test power was very low when they tested the results with *t*-distribution. In light of these authors' simulation results, Mean/SEmean, Median/IQR, TrMean/SEmean and TrMean/IQR test statistics have been investigated for some measures of change.

It is known that the sampling distribution of Mean/SEmean statistic for D is the normal distribution. However, the sampling distributions of this statistic for PC and MSPC are not normal. In addition, it has been observed that the distribution of TrMean/SEmean statistics does not converge to a symmetrical or normal distribution under any conditions. Furthermore, the sampling distributions of the Median/IQR and TrMean/IQR statistics for PC have nearly symmetrical distributions at very large sample sizes, but exhibit a negatively skewed distribution under other conditions. For researchers, if it is more meaningful to use MSCP rather than D when defining a change, the *HS* test proposed in this study can be used when the sample size is large ($n > 100$) and the effect size is higher than 1.5. In other cases, however, Mean/SEmean statistics should be used for D (Table 6). In addition, no suitable statistics have been proposed for PC; this measure of change can be used to interpret the results descriptively.

In many clinical studies, percent change is used to assess treatment efficacy (Curran-Everett and Williams, 2015; Reuter *et al.,* 2012). According to Tornqvist *et al.* (1985), the fact that percent change is independent of the unit of measurement represents an advantage. Unlike simple difference, moreover, percent change is also an effect size that can be easily understood and interpreted by everyone. However, the authors provided no information as to which measure should be used to determine treatment effect. Some researchers would suggest

computing the percent change for each observation, then descriptively reporting the median while reporting no inferential statistics for percent change (Vickers, 2001).

The sampling distributions of the four statistics for the MSPC measure have been found to be symmetrical. However, the Mean/SEmean or TrMean/SEmean statistics are found to be unsuitable for testing the modified symmetrised percent change values; this is because the Mean/SEmean statistic has two peak values and a platykurtic distribution, as in Figure 3a, while the TrMean/SEmean statistic exhibits a leptokurtic distribution, as in Figure 3b. However, it has been concluded that the Median/IQR and Trmean/IQR statistics cannot be tested using a *t*-test, because while their distributions are symmetrical, they are not normal.

When the literature on this subject is examined, it can be seen that the majority of research focuses on how to evaluate the pre-post difference in parallel designs. The most widely preferred model is the ANCOVA model, in which the pre-values are covariate variables and post-values are response variables. However, it is also stated that the power of the test will be low if PC, R and SPC are used as the treatment effect. The aim of such studies is to compare several independent treatment groups in terms of treatment effect (Vickers, 2001; Curran-Everett and Williams, 2015). This study examined how the treatment effect should be expressed in the case of a comparison of dependent measure values obtained before and after treatment in an experimental design with only one group.

5. Conclusion

PC, R, SPC and MSPC measures should be preferred when the variables being studied, such as seizure frequency or laboratory tests, exhibit large intra- and inter-subject variability and a skewed distribution for pre- and post- data. In particular, PC and R can be used for easy interpretation in a clinical context. Statistical analysis based on the parametric tests is not recommended for PC, because PC (or the ratio of two values) will not be normal even if the pre- and post-data are normal (Yamabe *et al.* 2012). For some examples of suitable clinical evaluation, PC can be applied to the treatment evaluation of patients with high-density lipoprotein cholesterol (Adachi *et al.,* 2009), of patients with urge to urinate or urge incontinence based on the number of acraturesis (Homma *et al.* 2003), or of patients with climacteric disorder based on the number of hot flushes (Endrikat *et al.*, 2007).

Statistically, the properties of the SPC and MSPC measures are superior to PC and R. SPC and MSPC do not produce overly large values and show robustness to outliers on the same data analysis. Berry (1989) introduced SPC as the modified percent change with good statistical properties in the medical field. Brouwers and Mohr (1989) argued that the advantage of using SPC over PC is that the transformed variable does not depend on the denominator used in the transformation and the resultant distribution is symmetrical about its mean. However, the interpretation of SPC may not be intuitive for those accustomed to thinking in terms of PC. To improve the interpretability of analysis results, Berry (1989) suggested transforming SPC to the PC scale using the inverse transformation and then the obtained robust percent change (RPC) should be interpreted. SPC is applied to the treatment evaluation of patients with partial epilepsy based on the seizure frequency (Yamauchi *et al.,* 2006) and the evaluation of male patients with osteoporotic fracture based on physical activity (Janney *et al.,* 2010).

According to this study, the statistics TrMean/IQR and Median/IQR (*HS* test) perform well in terms of power for MSCP, especially for higher sample sizes. However, the choice between the change measures D and MSPC will be based on various factors, including the

measurement units, range of data, interpretation of the change measures, etc. The appropriate measure should be selected by striking a balance between both the clinical and statistical points of view. Accordingly, it is safe to conclude that the two test statistics work well using the MSPC measure. When MSPC serves as a more appropriate measure of changes to describe the effect, while the sample size (>120) and effect size (>1.5) are also high, the *HS* test for MSPC should be preffered over a *t*-test for D measure.

It is suggested that similar studies should be carried out for designs involving more than two repeated measures in future studies.

References

Adachi, H., Imaizumi, T., Murakami, M. and Abe, M. (2009). A phase III, randomized, parallel-group comparative study of Caduet (an amlodipine/atorvastatin combination drug) in patients with concurrent hypertension and hyper-LDL-cholesteremia. *Journal of New Remedies and Clinics*, **58**(**9**), 1496-1510.

Ankarali, H. C. and Ankarali, S. (2009). Which measure should be used for testing in a paired design, Simple difference, percent change, or symmetrized percent change? *Communications in Statistics - Simulation and Computation*, **38**(**1**), 402-415.

Berry, D. A. (1989). *Statistical Methodology in the Pharmaceutical Sciences*. Marcel Dekker, New York.

Berry, D. A. and Ayers, G. D. (2006). Symmetrized percent change for treatment comparisons. *The American Statistician*, **60**(**1**), 27-31.

Brouwers, P. and Mohr, E. (1989). A metric for the evaluation of change in clinical trials. *Clinical Neuropharmacology*, **12**,129-133

Curran-Everett, D. and Williams, C. L. (2015). Explorations in statistics, the analysis of change. *Advances in Physiology Education*, **39**(**2**), 49-54.

Endrikat, J., Graeser T., Mellinger U., Ertan, K. and Holz, C. (2007). A multicenter, prospective, randomized, double-blind, placebo-controlled study to investigate the efficacy of a continuous-combined hormone therapy preparation containing 1mg estradiol valerate/2 mg dienogest on hot flushes in postmenopausal women. *Maturitas*, **58**, 201-207.

Homma, Y., Paick, J. S., Lee J. G. and Kawabe K. On behalf of the Japanese and Korean Tolterodine Study Group (2003). Clinical efficacy and tolerability of extended-release tolterodine and immediate-release oxybutynin in Japanese and Korean patients with an overactive bladder: a randomized, placebo-controlled trial. *BJU International*, **92**(**7**), 741-747.

Janney, C. A., Cauley, J. A., Cawthon, P. M. and Kriska, A. M. (2010). Longitudinal physical activity changes in older men in the osteoporotic fractures in men study. *Journal of the American Geriatrics Society*, **58**(**6**), 1128-1133.

Kim, M. K., Tanaka, K., Kim, M. J., Matuso, T., Endo, T., *et al.* (2009). Comparison of epicardial, abdominal and regional fat compartments in response to weight loss. *Nutrition, Metabolism and Cardiovascular Diseases*, **19**(**11**), 760-766.

Koti, K. M. (2001). On a primary efficacy endpoint. *Drug Information Journal*, **35**(**1**), 157-162.

Rocci, M. L. Jr., Devanarayan, V., Haughey, D. B. and Jardieu, P. (2007). Confirmatory reanalysis of incurred bioanalytical samples. *The American Association of Pharmaceutical Scientists Journal*, **9**(**3**), E336–E343.

Reuter, M., Schmansky, N. J., Rosas, H. D. and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, **61**(**4**), 1402-1418.

Tornqvist, L., Vartia, P. and Vartia, Y. O. (1985). How should relative changes be measured? *The American Statistician*, **39**(**1**), 43-46.

Tsounis, S., Kimiskidis, V. K., Kazis, D., Gkiatas, K., Garganis, K., *et al.* (2011). An open-label, add-on study of pregabalin in patients with partial seizures, a multicenter trial in Greece. *Seizure*, **20**(**9**), 701-705.

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient, a simulation study. *BMC Medical Research Methodology*, **1**, 6.

Waleekhachonloet, O. A., Limwattananon, C., Limwattananon, S. and Gross, C. R. (2007). Group behavior therapy versus individual behavior therapy for healthy dieting and weight control management in overweight and obese women living in rural community. *Obesity Research and Clinical Practice*, **1**(**4**), 223-232.

Wilkinson, L. and Task Force on Statistical Inference, American Psychological Association, Science Directorate (1999). Statistical methods in psychology journals, Guidelines and explanations. *American Psychologist*, **54**(**8**), 594-604.

Yamabe, T., Maruo, K., Shirahata, S. and Goto, M. (2012). Statistical properties of two ratio measures based on pre- and post observed values which are assumed as bivariate power normal distribution. *Ouyou Toukeigaku*, **41**(**1**), 53-68.

Yamauchi, T., Kaneko, S., Yagi, K. and Sase, S. (2006). Treatment of partial seizures with gabapentin: Double-blind, placebo-controlled, parallel-group study. *Psychiatry and Clinical Neurosciences*, **60**, 507-515.

Zhang, L. and Han, K. (2009). How to analyze change from baseline, absolute or percentage change? D-level *Essay in Statistics 2009*. Dalarna University, Borlänge. http,//docplayer.net/20733008-How-to-analyze-change-from-baseline-absolute-or-percentage-change.html. Accessed March 2018.

ANNEXURE A

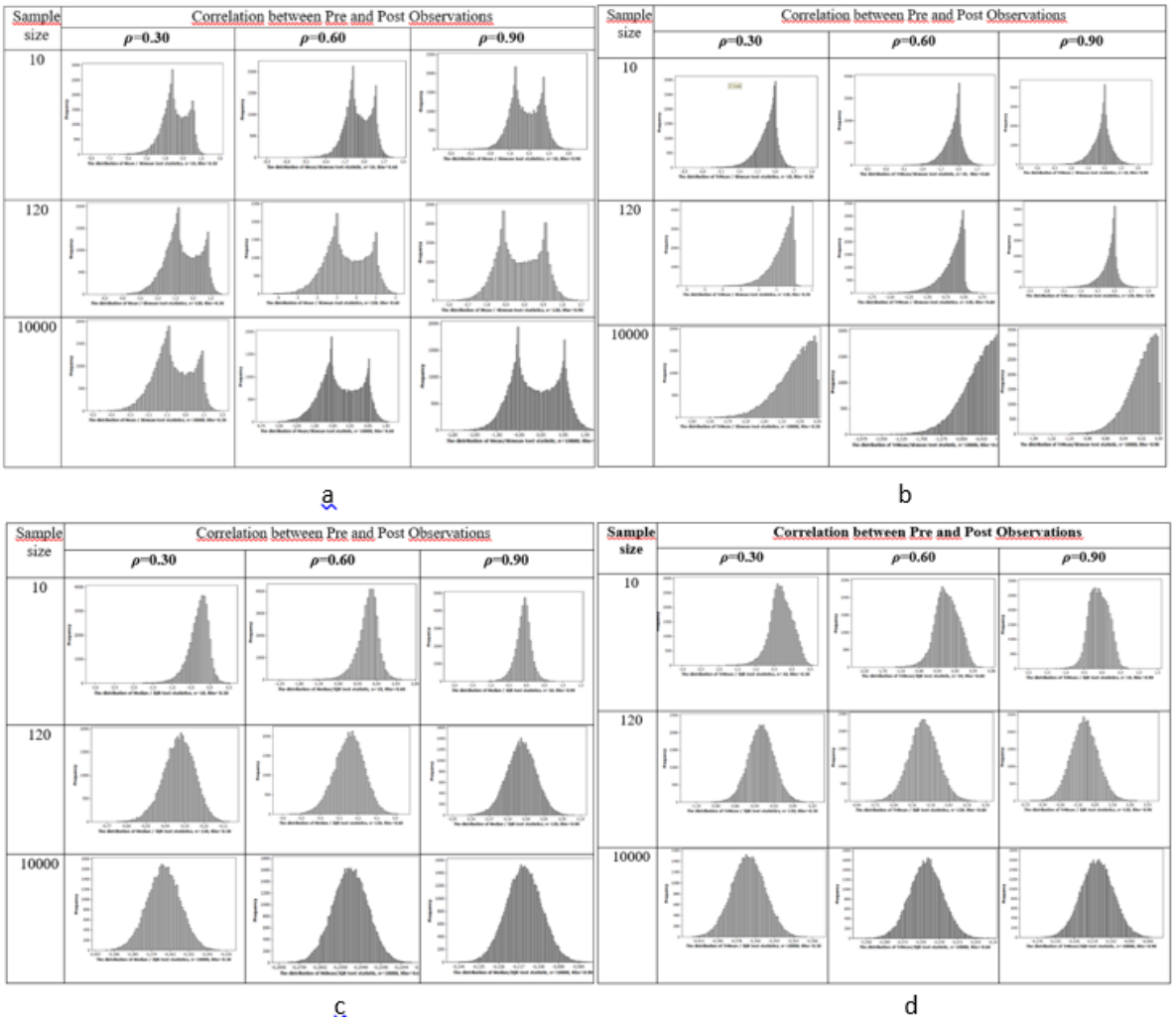

Figure 2: Empirical sampling distributions of a) Mean/SEmean statistics, b) TrMean/SEmean statistics, c) Median/IQR statistics and d) TrMean/IQR statistics for using PC

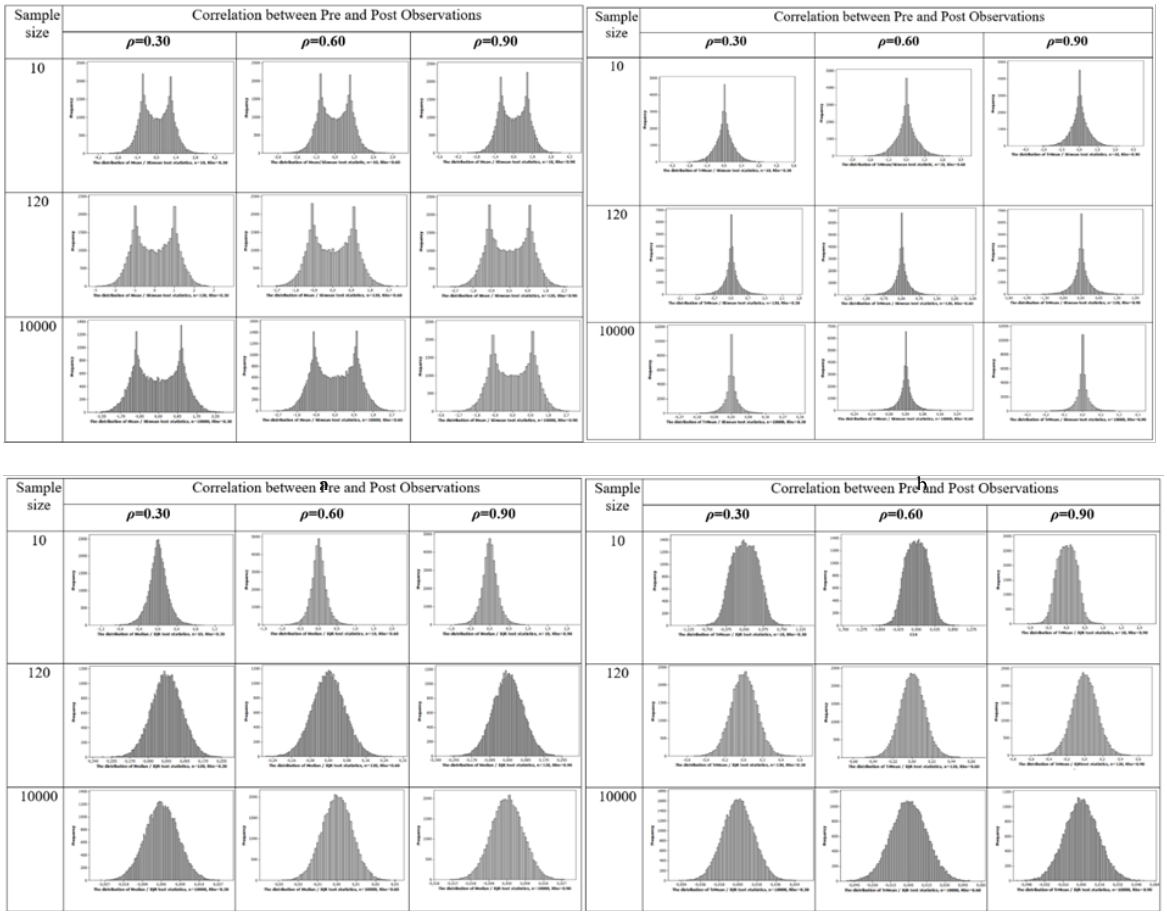

Figure 3: Empirical sampling distribution of (a) Mean/SEmean statistics, (b) TrMean/SEmean statistics, c) Median/IQR statistics and d) TrMean/IQR statistics for MSPC

ANNEXURE B

Table 1: Critical values for 95% confidence level of sampling distributions of four test statistics for PC

n	ρ	Percent Change ($PC = \left(\frac{Post-Pre}{Pre}\right)x100$)							
		Mean/SEmean		TrMean/SEmean		Median/IQR		TrMean/IQR	
		$S1L_\alpha$	$S1R_\alpha$	$S2L_\alpha$	$S2R_\alpha$	$S3L_\alpha$	$S3R_\alpha$	$S4L_\alpha$	$S4R_\alpha$
10	0.30	-3.376	1.207	-3.238	0.691	-0.952	0.112	-0.959	0.257
	0.60	-2.815	1.344	-2.657	0.911	-0.770	0.201	-0.784	0.309
	0.90	-2.232	1.593	-2.018	1.242	-0.568	0.317	-0.593	0.386
20	0.30	-3.166	1.181	-2.981	0.596	-0.820	0.002	-1.176	0.455
	0.60	-2.615	1.314	-2.376	0.788	-0.653	0.092	-1.042	0.547
	0.90	-2.131	1.515	-1.844	1.086	-0.469	0.208	-0.892	0.666
60	0.30	-3.011	1.168	-2.580	0.067	-0.622	-0.135	-0.788	0.063
	0.60	-2.517	1.287	-1.988	0.262	-0.479	-0.041	-0.653	0.167
	0.90	-2.050	1.480	-1.412	0.586	-0.320	0.081	-0.512	0.291
120	0.30	-2.981	1.147	-2.396	-0.014	-0.548	-0.201	-0.655	-0.071
	0.60	-2.527	1.279	-1.816	0.033	-0.410	-0.101	-0.532	0.031
	0.90	-2.031	1.484	-1.173	0.335	-0.260	0.025	-0.388	0.162
500	0.30	-2.960	1.146	-2.158	-0.028	-0.453	-0.284	-0.505	-0.226
	0.60	-2.495	1.289	-1.532	-0.017	-0.327	-0.174	-0.384	-0.113
	0.90	-2.034	1.489	-0.853	0.024	-0.185	-0.044	-0.246	0.019
10000	0.30	-2.952	1.145	-2.068	-0.030	-0.386	-0.348	-0.398	-0.336
	0.60	-2.488	1.272	-1.421	-0.0183	-0.267	-0.233	-0.280	-0.220
	0.90	-2.015	1.473	-0.659	-0.0107	-0.130	-0.099	-0.144	-0.086

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, $S1L_\alpha$ and $S1R_\alpha$: Left and Right tail critical values for Mean/SEmean; $S2L_\alpha$ and $S2R_\alpha$: Left and Right tail critical tables value for TrMean/SEmean; $S3L_\alpha$ and $S3R_\alpha$: Left and Right tail critical tables value for Median/IQR; $S4L_\alpha$ and $S4R_\alpha$: Left and Right tail critical table values for TrMean/IQR

Table 2: Critical values for 95% confidence level of sampling distributions of four test statistics for MSPC

n	*ρ*	Modified Percent Change ($MSPC = \left(\frac{Post-Pre}{\frac{Pre+Post}{2}}\right) x100$)							
		Mean/SEmean		TrMean/SEmean		Median/IQR		TrMean/IQR	
		$S1L_\alpha$	$S1R_\alpha$	$S2L_\alpha$	$S2R_\alpha$	$S3L_\alpha$	$S3R_\alpha$	$S4L_\alpha$	$S4R_\alpha$
10	0.30	–1.887	1.877	–1.622	1.622	–0.430	0.430	–0.470	0.470
	0.60	–1.886	1.884	–1.609	1.602	–0.434	0.434	–0.473	0.473
	0.90	–1.884	1.864	–1.600	1.600	–0.435	0.435	–0.476	0.476
30	0.30	–1.765	1.765	–1.180	1.180	–0.274	0.274	–0.470	0.470
	0.60	–1.760	1.760	–1.158	1.158	–0.273	0.273	–0.480	0.480
	0.90	–1.758	1.758	–1.170	1.170	–0.272	0.272	-0.482	0.482
60	0.30	–1.737	1.736	–0.968	0.954	–0.195	0.195	–0.395	0.395
	0.60	–1.740	1.726	–0.978	0.961	–0.196	0.196	–0.398	0.398
	0.90	–1.733	1.725	–0.969	0.973	–0.196	0.196	–0.400	0.400
120	0.30	–1.730	1.736	–0.698	0.693	–0.138	0.139	–0.271	0.271
	0.60	–1.736	1.728	–0.721	0.721	–0.139	0.139	–0.276	0.276
	0.90	–1.721	1.717	–0.715	0.717	–0.140	0.140	–0.277	0.277
500	0.30	–1.704	1.735	–0.358	0.363	–0.068	0.068	–0.130	0.130
	0.60	–1.724	1.710	–0.355	0.357	–0.068	0.068	–0.129	0.129
	0.90	–1.727	1.728	–0.351	0.354	–0.068	0.068	–0.131	0.131
10000	0.30	–1.710	1.711	–0.081	0.081	–0.015	0.015	–0.029	0.029
	0.60	–1.715	1.731	–0.081	0.081	–0.015	0.015	–0.029	0.029
	0.90	–1.718	1.701	–0.082	0.082	–0.015	0.015	–0.029	0.029

n: Sample size; *ρ*: Correlation coefficient between pre-values and post-values, PC: Percent Change, R: Ratio, MSPC: Modified Percent Change, SD: Standard Deviation, SEmean: Standard Error of Mean, TrMean: Trimmed Mean, IQR: Interquartile Range, $S1L_\alpha$ and $S1R_\alpha$: Left and Right tail critical values for Mean/SEmean; $S2L_\alpha$ and $S2R_\alpha$: Left and Right tail critical tables value for TrMean/SEmean; $S3L_\alpha$ and $S3R_\alpha$: Left and Right tail critical tables value for Median/IQR; $S4L_\alpha$ and $S4R_\alpha$: Left and Right tail critical table values for TrMean/IQR

Table 3: Observed type-I errors and power of *t*-test for difference in paired samples

n	*ρ*	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10 (t_α=±2.262)	0.30	5.13	22.23	66.43	94.48	99.74
	0.60	4.99	35.10	87.78	99.75	100
	0.90	4.81	87.86	100	100	100
30 (t_α=±2.045)	0.30	5.34	61.10	100	100	100
	0.60	4.88	84.00	100	100	100
	0.90	4.96	99.99	100	100	100
60 (t_α=±2.000)	0.30	4.99	89.73	100	100	100
	0.60	5.05	98.96	100	100	100
	0.90	5.02	100	100	100	100
120 (t_α=±1.980)	0.30	5.02	100	100	100	100
	0.60	5.13	100	100	100	100
	0.90	4.73	100	100	100	100
500 (t_α=±1.965)	0.30	5.12	100	100	100	100
	0.60	5.05	100	100	100	100
	0.90	5.23	100	100	100	100
10000 (t_α=±1.960)	0.30	5.06	100	100	100	100
	0.60	4.90	100	100	100	100
	0.90	4.81	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 4: Observed type-I errors and power of Median/IQR statistic (*HS* test) for MSPC

n	ρ	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10	0.30	5.01	5.71	12.93	32.66	58.88
	0.60	4.99	5.95	14.39	34.17	57.51
	0.90	4.95	7.19	19.68	37.84	56.82
30	0.30	5.05	7.12	31.10	81.31	100
	0.60	5.12	7.76	36.61	89.50	100
	0.90	4.82	12.5	47.86	80.53	100
60	0.30	4.91	9.56	56.77	97.99	100
	0.60	4.80	10.86	64.72	97.90	100
	0.90	4.95	2130	82.82	98.53	100
120	0.30	5.24	13.74	86.26	100	100
	0.60	4.97	17.36	92.06	100	100
	0.90	4.90	47.86	99.46	100	100
500	0.30	5.28	43.70	100	100	100
	0.60	5.26	55.67	100	100	100
	0.90	5.31	93.80	100	100	100
10000	0.30	5.50	100	100	100	100
	0.60	5.63	100	100	100	100
	0.90	5.55	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 5: Observed type-I errors and power of TrMean/IQR test statistic for MSPC

n	ρ	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10	0.30	5.15	5.86	14.56	38.59	68.42
	0.60	5.01	6.08	16.29	39.40	64.55
	0.90	4.96	7.23	21.16	40.32	60.06
30	0.30	5.19	6.13	16.06	49.34	84.92
	0.60	4.99	5.78	16.85	49.10	82.10
	0.90	4.96	6.62	20.36	51.18	80.01
60	0.30	5.24	6.16	20.30	67.44	94.32
	0.60	4.82	6.26	21.91	63.72	92.10
	0.90	4.75	6.85	22.27	62.11	90.30
120	0.30	4.99	7.23	36.98	95.60	100
	0.60	4.70	7.64	37.30	96.81	100
	0.90	4.63	20.36	34.48	98.80	100
500	0.30	5.23	15.53	90.24	100	100
	0.60	5.12	18.23	90.50	100	100
	0.90	4.83	27.77	89.67	100	100
10000	0.30	5.09	100	100	100	100
	0.60	5.02	100	100	100	100
	0.90	4.92	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 6: Which test should be preferred?

<table>
<tr><th rowspan="2">n</th><th rowspan="2">ρ</th><th rowspan="2">Type-I error (Observed alpha)</th><th colspan="4">Power of Tests
Standardized effects size</th></tr>
<tr><th>Δ = 0.5</th><th>Δ = 1.0</th><th>Δ = 1.5</th><th>Δ = 2.0</th></tr>
<tr><td>10</td><td>0.30</td><td rowspan="19">Paired Samples t-test, Median/IQR statistic and TrMean/IQR statistic have same results in terms of observed type-I error</td><td colspan="4" rowspan="19">In this conditions, Paired Samples t-test should be used for simple difference

In these conditions, as the statistical properties of the t-test and HS test are similar, HS test (Median/IQR) used for MSPC is recommended.
This is because MSPC considers the differences between individuals in pre-treatment when computing the treatment effect. This makes it more accurate measure of change.</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>30</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>60</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>120</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>500</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>10000</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
</table>

n: Sample size; *ρ*: Correlation coefficient between pre-values and post-values