ACCEPTED VERSION

How to compare the means of two samples that include paired observations and independent observations

Abstract

Standard approaches for comparing the means of two samples, comprising both paired observations and independent observations, involve the discarding of valuable information. An alternative test which uses all of the available data, is the partially overlapping samples t-test. Two variations of the test are available, one assuming equal variances, and one assuming separate variances. Issues with standard procedures, and considerations for choosing appropriate tests in the partially overlapping scenario are discussed. An example with details of how to apply the partially overlapping samples t-test is given.

Key Words

Partially overlapping samples; Incomplete observations; Welch's test; Independent samples; Paired samples; Equality of means

Introduction

It is well established that the paired samples t-test can be used for comparing means between two dependent samples (Zimmerman, 1997; Fradette et.al., 2003). The assumptions of the paired samples t-test are that data are randomly sampled from two related populations, and that the differences between the paired observations are approximately normally distributed. It is also well established that the independent samples t-test can be used for comparing means between two independent samples with equal variances (Rasch, Teuscher, and Guiard, 2007; Fradette et.al., 2003). When variances are not equal, the independent samples t-test is not Type I error robust, particularly when the sample sizes are not equal (Ramsey, 1980). When equal variances cannot be assumed, a Type I error robust alternative to the independent samples t-test is Welch's test (Derrick, Toher and White, 2016; Fradette et.al., 2003). For the avoidance of doubt, here the independent samples t-test assuming equal variances is referred to as the independent samples t-test, and the independent samples t-test not assuming equal variances is referred to as Welch's test. The assumptions of the independent samples t-test and Welch's test are that data are randomly sampled from two unrelated populations, which are approximately normally distributed. Welch's test is considered Type I error robust for all but the most extreme deviations from the normality assumption (Rasch, Teuscher, and Guiard, 2007). Extensive testing of these assumptions is not recommended (Rasch, Kubinger, and Moder, 2011; Rochon, Gondan, and Kieser, 2012). A further assumption of these tests is that observations within a sample are independent of each other.

This assumption is critical, violations of the independence of observations assumption make hypothesis testing invalid (Lissitz and Chardos, 1975).

Conventional teaching of statistics usually assumes a perfect world with completely dependent samples or completely independent samples (for example: Magel, 1998). However, a question that is often asked in research is how to compare means between two samples that include both paired observations and unpaired observations. These scenarios are referred to as ‘partially overlapping samples’ (Martinez-Camblor et.al., 2012; Derrick et.al., 2015; Derrick et.al., 2017). Paired samples designs are often advantageous relative to independent samples designs, because paired samples designs allow differences between two samples to be directly compared. However, partially overlapping samples designs are often required due to the limited resource of paired samples, where a number of independent observations are available to compensate. This could also occur in a matched pairs design, when pairing individuals on certain characteristics, there may be some additional independent observations that cannot be reasonably paired on any characteristics. In addition there are occasions where it is desired that observations from a paired samples design, and a separate independent samples design, may be combined, resulting in a partially overlapping samples design.

The approach for analysing partially overlapping samples by design has received relatively little attention within the literature. Consider the scenarios in Figure 1, which demonstrates eight scenarios where there are two samples, each with a different number of paired observations and independent observations.

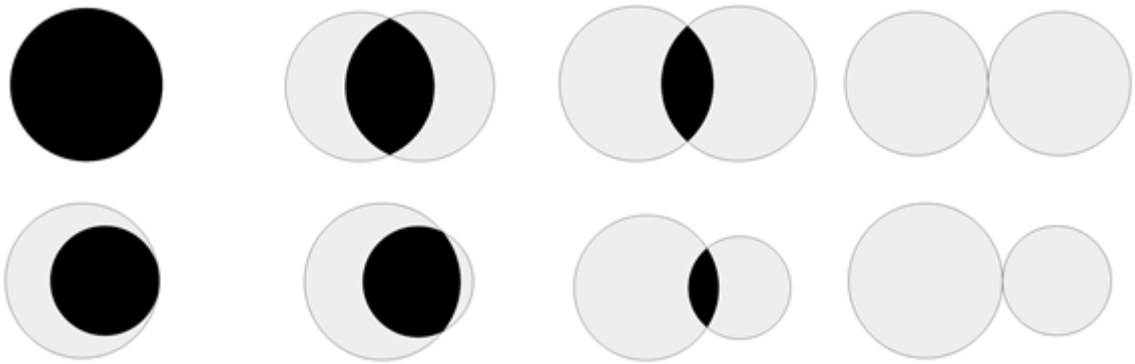

Figure 1. Examples of ‘partially overlapping samples’. In each scenario each of two samples are represented by a circle. The paired observations are represented by the overlap and shaded black. From left to right the graphic shows a decreasing number of paired observations. The relative sample sizes are represented by the size of the circle.

It is not well established how to proceed for the scenarios represented by Figure 1 where there is partial overlap. One ‘standard’ approach if the number of pairs is large, is to perform the paired samples t-test on only the paired observations. Conversely, if the number of independent samples is large a ‘standard’ approach is to perform the independent samples t-test or Welch’s test, on only the independent samples (Looney and Jones, 2003). These standard methods discard data, this adversely impacts the power of the test. Approaches that discard data are likely to maintain adequate power if the number of discarded observations is relatively ‘small’ and the sample sizes are relatively ‘large’. One alternative approach that is commonly applied, is to perform the independent samples t-test on all of the available data. However, this is less powerful than a paired samples approach and ignores the fact that there are matched pairs. Alternative ad hoc approaches using all of the available data, but not mimicking the design structure, will not be considered further in this paper. These alternative approaches emphasise the need for statistically valid tests in the partially overlapping samples case.

A frequent occurrence of partially overlapping samples is a paired samples design with missing observations (Martinez-Camblor et.al., 2012). In this situation partially overlapping samples do not occur by design, and so it is necessary to consider why the samples are incomplete. If data are missing completely at random (MCAR), the reason for missing data is not related to the value of the observation itself, or other variables recorded. An example of data that is MCAR is a question in a survey that is accidentally missed, or data that is accidentally lost. If incomplete observations are MCAR, it is reasonable to discard the corresponding paired observations without causing bias (Donders et.al, 2006). If data are missing at random (MAR), data are missing based on characteristics not directly measured by the missing observation itself. However, the missing data is related to another variable in the dataset. The discarding of information that are MAR is likely to cause bias, therefore the standard approach of pairwise or listwise deletion is not recommended (Schafer, 1997; Donders et.al., 2006). If data are missing not at random (MNAR), the probability of an observation being missing, directly depends on the value of the observation being recorded. When data are MNAR, there is no statistical procedure that can eliminate potential bias (Musil et. al, 2002). This is particularly of concern for analyses with missing data because it is difficult to distinguish between data that is MAR and data that is MNAR. Nevertheless if the amount of missing data is small, the bias is likely to be inconsequential. The literature suggests that up to 5% of observations missing is acceptable (Graham, 2009; Schafer, 1997). There are some that take a more liberal stance in the literature suggesting that up to 20% of data missing may be acceptable (Schlomer, Bauman, and Card, 2010).

For a paired samples design with incomplete observations, researchers often attempt to impute the missing data. Ad hoc basic imputation approaches for imputing missing data are biased solutions

(Schafer, 1997). Mean imputation reduces the variation in the data set. Regression imputation inflates the correlation between variables. More sophisticated techniques, Expected Maximisation and Multiple Imputation, minimise the bias of the parameter estimates (Musil et.al., 2002; Dong and Peng, 2013).

Standard statistical software will perform the paired samples t-test, the independent samples t-test or Welch's test upon command. In SAS the standard 'proc ttest' performs the paired samples t-test, omitting cases pairwise from calculations when any observation from a declared paired variable is missing. Likewise in Unistat, a paired samples t-test is performed, excluding any 'missing values' pairwise. Performing the paired samples t-test in SPSS gives the options of excluding cases pairwise or excluding cases listwise, which are equivalent in the two sample case. In all of these approaches, the unpaired observations are excluded and the analysis is done only on the paired data. Caution should be exercised when using SAS, SPSS or Unistat, because users may be tempted to analyse only the complete pairs when readily presented with the opportunity, and not realise the consequences of not using all of the data. Both Minitab and the standard 't.test' in R present an error message when a paired samples t-test is selected with unequal sample sizes, these software at the very least make users aware there are considerations to take into account with the analysis they are trying to perform.

Derrick et.al., (2017) developed two partially overlapping samples t-tests that make use of all of the available data, that are valid under MCAR and robust under the assumptions of Normality. These test statistics act as a straightforward interpolation between the paired samples t-test, and either the independent samples t-test, or Welch's test. Using these tests for comparing two sample means represents a more powerful alternative to discarding information. In the case of a paired samples design with incomplete observations, these test statistics also represent an alternative to the need to perform complicated imputation techniques.

In this paper, the partially overlapping samples test statistics that make use of all of the available data, accounting for the fact that there is a combination of paired observations and independent observations, are demonstrated by use of example. The paper concludes with a discussion on comparing the use of traditional tests against the partially overlapping samples t-tests.

Worked Example

In this section, an example of the partially overlapping t-test in application is given, with a summary of the calculations and the hypothesis test procedure.

The sleep fragmentation index measures the quality of sleep for an individual over one night. A lower sleep fragmentation score represents less disrupted sleep. The research question is whether the genre of a movie watched before bedtime impacts the quality of sleep.

Study participants are randomly allocated to either a between subjects design (stage 1) or a repeated measures (stage 2) part of the investigation. In the first stage of the study, the sleep fragmentation score is taken over one night, for two groups of individuals. A sample of n_a =8 individuals watch a 'horror' movie before bedtime. A separate sample of n_b =8 individuals watch a 'feel good' movie before bedtime. This first stage is an independent samples design. In a second stage of the study, the sleep fragmentation index is recorded over two separate nights, for a sample of n_c =8 individuals watching a 'feel good' movie and a 'horror' movie on two alternate nights before bedtime (with order counterbalanced). This second stage is a paired samples design. When the two stages of the study are combined, the total number of individuals who watched a 'horror' movie is $n_1 = n_a + n_c = 16$. The total number of individuals who watched a 'feel good' movie is $n_2 = n_b + n_c = 16$. The hypothesis being tested is whether the mean sleep fragmentation scores are the same between individuals watching a 'horror' movie and individuals watching a 'feel good' movie. Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$. The alternative hypothesis, assuming a two-sided test is $H_1 : \mu_1 \neq \mu_2$. The sleep fragmentation scores are given in Table 1.

In this scenario, from a missing data perspective it would be reasonable to assume MCAR. There are no missing data per se; it is the design of the study that results in partially overlapping samples. Therefore standard approaches of discarding either the paired or independent samples are unbiased. However, performing either the paired samples t-test or the independent samples t-test requires discarding exactly half of the observations, and the power of the test is reduced. This therefore is a good example of where a test statistic that makes use of all available data, taking into account both paired and independent observations could be useful.

Table 1: Sleep fragmentation scores obtained for each individual (ID)

Independent Samples (Stage 1)				Paired Samples (Stage 2)		
ID	Horror	ID	Feel good	ID	Horror	Feel good
I1	20	I9	10	P1	14	15
I2	21	I10	16	P2	15	10
I3	16	I11	18	P3	18	15
I4	18	I12	16	P4	20	17
I5	14	I13	15	P5	11	13
I6	12	I14	14	P6	19	19
I7	14	I15	13	P7	14	12
I8	17	I16	10	P8	15	13

Assuming normality and MCAR, the partially overlapping samples t-test is a Type I error robust method for comparing means between the two samples (Derrick et.al., 2017). To calculate elements for the partially overlapping samples t-test let; $\bar{x}_1$ = mean of all observations in Sample 1 (i.e. the mean for the n_1 observations for individuals watching a 'horror' movie), $\bar{x}_2$ = mean of all observations in Sample 2 (i.e. the mean for the n_2 observations for individuals watching a 'feel good' movie), s_1 = standard deviation of all observations in Sample 1, s_2 = standard deviation of all observations in Sample 2, and *r* = Pearson's correlation coefficient for the paired observations only (i.e. in n_c). There are two forms of the partially overlapping samples t-test; t_1 for when equal variances between the two samples can be assumed, and t_2 for when equal variances between the two samples cannot be assumed.

The partially overlapping samples t-test assuming equal variances acts as an interpolation between the independent samples t-test and the paired samples t-test, and is defined by Derrick et.al. (2017) as:

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}-2r\left(\frac{n_c}{n_1 n_2}\right)}} \text{ where } s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$$

If the null hypothesis is true, the test statistic t_1 follows a t-distribution with approximate degrees of freedom given as:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c} \right)(n_a + n_b).$$

If equal variances cannot be assumed, the partially overlapping samples t-test which acts as an interpolation between Welch's test and the paired samples t-test is defined by Derrick et.al. (2017) as:

$$t_2 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2r\left(\frac{s_1 s_2 n_c}{n_1 n_2} \right)}}$$

If the null hypothesis is true, the test statistic t_2 follows a t-distribution with degrees of freedom approximated by:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c} \right)(n_a + n_b) \text{ and where } \gamma = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

These test statistics can be viewed as a generalised form of the two sample t-tests. When there are no independent observations, t_1 and t_2 default to the paired samples t-test. When there are no paired observations, t_1 defaults to the independent samples t-test, and t_2 defaults to Welch's test.

For either version of the partially overlapping samples t-test, if $\mu_1 > \mu_2$ (i.e. the population mean score for 'horror' movie is greater than the population mean score for 'feel good' movie), then it is anticipated that this will be reflected in the sample values above, and the expectation is to observe a large positive value of the test statistic. Conversely if $\mu_1 < \mu_2$, the expectation would be for a large but negative value of the test statistic to be observed. In absolute terms it is anticipated that large values of the test statistic will be observed if the null hypothesis is not true. The null hypothesis is rejected if the observed value of the test statistic is greater than the critical value from a t-distribution with the degrees of freedom as defined by v_1 or v_2.

The elements of the calculation of the test statistics are[1]: n_1=16, n_2=16, n_a=8, n_b=8, n_c=8, $\bar{x}_1$=16.125, $\bar{x}_2$=14.125, s_1=2.986, s_2=2.778, r=0.687, s_p=2.884, γ=29.845, t_1=2.421, t_2=2.419, v_1=18.500, v_2=18.422.

[1] Unrounded values are used in each part of the calculation, each element displayed to 3 decimal places.

The calculated value of the test statistic t_1 is 2.421. The calculated value of the test statistic t_2 is 2.419. Using the degrees of freedom v_1=18.500 or v_2=18.422, from the t-distribution at the 5% significance level the critical value is 2.097. The calculated value of the test statistic is greater than the critical value, therefore the null hypothesis is rejected (p=0.026)**.**

Instead of performing the above calculations manually, the partially overlapping samples t-tests can be easily performed in R, using the package 'Partiallyoverlapping' (Derrick, 2017). In the following, let 'a' represent 'horror' movie and 'b' represent 'feel good' movie. Example R code to enter the data and perform the analyses assuming equal variances is given below:

```
install.packages(Partiallyoverlapping)
library(Partiallyoverlapping)
a.unpaired<-c(20,21,16,18,14,12,14,17)
b.unpaired<-c(10,16,18,16,15,14,13,10)
a.paired<-c(14,15,18,20,11,19,14,15)
b.paired<-c(15,10,15,17,13,19,12,13)
Partover.test(a.unpaired,b.unpaired,a.paired,b.paired,var.equal=TRUE)
#Output: statistic=2.421, parameter=18.500, p.value=0.026.
```

Alternatively, to perform the test when equal variances are not assumed, the 'var.equal=TRUE' command can be dropped. The results from either test performed replicate their respective manual calculation and show that the samples from group 'a' (Horror movie) and group 'b' (Feel good movie) have significantly different means at the 5% significance level.

When using the partially overlapping samples t-test at the 5% significance level, there is a statistically significant difference in the mean sleep fragmentation index between individuals watching a 'horror' movie prior to bedtime, and individuals watching a 'feel good' movie prior to bedtime. The results suggest that individuals watching a 'feel good' movie before bedtime, have less disrupted sleep compared to individuals watching a 'horror' movie before bedtime.

Discussion

Further consideration is given to the choice between traditional tests that discard information, and the partially overlapping samples t-tests. Table 2 gives a summary of results obtained from the example in Table 1. This shows results when performing 'standard' tests and results from performing the partially overlapping samples t-tests, with their respective statistical decisions at the 5% significance level.

Table 2. Summary of results for the worked example, including the calculated value of each test statistic (t), the degrees of freedom (df), the p-value (p) and the statistical decision.

Test	t	df	p	Decision
Paired samples t-test	1.821	7.000	0.111	Fail to reject H_0
Independent samples t-test	1.667	14.000	0.118	Fail to reject H_0
Welch's test	1.667	13.912	0.118	Fail to reject H_0
Partially overlapping samples t-test (t_1)	2.421	18.500	0.026	Reject H_0
Partially overlapping samples t-test with Welch's df (t_2)	2.419	18.422	0.026	Reject H_0

It can be seen from Table 2 that the choice of test to apply is important because the statistical decision is not the same. This example emphasises the lower power for the traditional approaches. In general, the more observations used in the calculation of a test statistic, the greater the power of the test will be. However, rare situations may arise where the independent observations and the paired observations have mean differences in opposing directions. In these situations the partially overlapping samples t-test may cancel out these differences, but to ignore either the paired observations or independent observations could create bias.

In the worked example, the two samples are partially overlapping by design. It is also possible to encounter a partially overlapping samples design, with incomplete observations. In these situations, the partially overlapping samples t-test can similarly be performed on all available observations, when the missing observations are MCAR. To demonstrate this, consider the situation where there are occasional errors with the machine recording sleep fragmentation. As a result of errors, let the 'horror' observations for individuals 'I1' and 'P1' be missing. There is now one missing independent 'horror' observation and one missing paired observation. The resulting reduction in sample size is further to the detriment of the paired samples t-test, the independent samples t-test and Welch's test. Using the

partially overlapping samples t-test, the 'feel good' observation for individual 'P1' is not discarded, it is treated as an independent observation. Revised elements of the partially overlapping samples t-test are; n_1=14, n_2=16, n_a=7, n_b=9, n_c=7, $\bar{x}_1$=16.000, $\bar{x}_2$=14.125, s_1=2.961, s_2=2.778, r=0.736, s_p=2.864, γ=26.903, t_1=2.208, t_2=2.194, v_1=17.733, v_2=17.148. Assuming equal variances and using the test statistic t_1, the p-value is 0.041. For completion, using the test statistic t_2, the p-value is 0.042. The null hypothesis is rejected at the 5% significance level and the statistical conclusions are as before.

The assumptions of the partially overlapping samples t-test match the assumptions of the independent samples t-test. The assumptions are that observations within a sample are independent of each other, observations are sampled from normally distributed populations and equal variances between the two groups. The assumptions of the partially overlapping samples t-test with Welch's degrees of freedom, match the assumptions of Welch's test. This assumes that observations within a sample are independent of each other and observations are sampled from normally distributed populations. Similarly as stated for the standard tests that discard data, extensive testing of these assumptions is not recommended. The partially overlapping samples t-test with Welch's degrees of freedom is Type I error robust with equal and unequal variances, and the power difference relative to the independent samples t-test is negligible. Many authors advocate the routine use of Welch's test in the two independent samples case, (for example, Ruxton, 2006; Rasch, Kubinger and Moder, 2011). Therefore, if in doubt and normality and MCAR can be assumed, the partially overlapping samples t-test with Welch's degrees of freedom can be used routinely in the two partially overlapping samples case.

Conclusion

A common issue in psychology is a paired samples design with incomplete observations, or a study that otherwise results in both paired observations and independent observations being observed. These scenarios are referred to in the literature as partially overlapping samples.

In these scenarios, the discarding of observations is common practice. However, discarding observations may cause bias, and has a substantial impact on power when sample sizes are small and/or if the number of discarded observations is large. The partially overlapping samples approach uses all available data and has appeal when the assumption of normality has not been grossly violated, and the MCAR assumption is reasonable. These solutions do not detract from other analytical strategies but do provide a simple generalisation of the standard two sample t-tests.

References

Derrick, B. (2017). Partiallyoverlapping: Partially Overlapping Samples t-Tests. R package version 1.0.

Derrick, B., Dobson-McKittrick, A., Toher, D. & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. Journal of Applied Quantitative Methods, 10(3), 1-14.

Derrick, B., Russ, B., Toher, D. & White P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. Journal of Modern Applied Statistical Methods, 16(1).

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. The Quantitative Methods for Psychology, 12(1), 30-38. doi:10.20982/tqmp.12.1.p030

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087-1091. doi:10.1016/j.jclinepi.2006.01.014

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 1-17. doi:10.1186/2193-1801-2-222

Fradette, K., Keselman, H.J., Lix, L., Algina, J., & Wilcox, R. (2003). Conventional and Robust Paired and Independent Samples t-tests: Type I Error and Power Rates, Journal of Modern Applied Statistical Methods, Vol 2, Iss 2, 481-496. doi:10.22237/jmasm/1067646120

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual Review of Psychology, 60, 549-576. doi:10.1146/annurev.psych.58.110405.085530

Lissitz, W., & Chardos, S. (1975). A study of the effect of the violations of the assumption of independent sampling upon the type one error rate of the two-sample t-test. Educational and Psychological Measurement, 35, 353-359. doi:10.1177/001316447503500213

Looney, S., & Jones, P. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data, Statistics in Medicine, Vol 22, 1601-1610. doi:10.1002/sim.1514

Magel, R. C. (1998). Testing for Differences between two brands of Cookies. Teaching Statistics, 20(3), 81-83. doi:10.1111/j.1467-9639.1998.tb00775.x

Martinez-Camblor, P., Corral, N., & de la Hera, J.M. (2012). Hypothesis test for paired samples in the presence of missing data, Journal of Applied Statistics, Vol 40, No.1, 76-87. doi: 10.1080/02664763.2012.734795

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. Western Journal of Nursing Research, 24(7), 815-829. doi:10.1177/019394502762477004.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's t test with unequal variances. Journal of Educational and Behavioral Statistics, 5(4), 337-349. doi:10.2307/1164906

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. Statistical papers, 52(1), 219-231. doi:10.1007/s00362-009-0224-x

Rasch, D., Teuscher, F., & Guiard, V. (2007). How robust are tests for two independent samples?. Journal of Statistical Planning and Inference, 137(8), 2706-2720. doi:10.1016/j.jspi.2006.04.011

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. BMC Medical Research Methodology, 12(1), 1. doi:10.1186/1471-2288-12-81

Ruxton, G. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology. 17(4), 688-690. doi:10.1093/beheco/ark016

Schafer, J. L. (1997). Analysis of incomplete multivariate data. CRC press.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. Journal of Counseling Psychology, 57(1), 1. doi:10.1037/a0018082

Zimmerman D. (1997). A note on the interpretation of the paired samples t-test, Journal of educational and behavioral statistics,22,3, 349 – 360. doi:10.3102/10769986022003349