An example motivated discourse of the paired samples t-test

Paul White,

Applied Statistics Group,

Faculty of Environment and Technology,

Univeristy of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.white@uwe.ac.uk

Paul Redford,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul2.redford@uwe.ac.uk

James MacDonald,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

james.macdonald@uwe.ac.uk

Abstract— An application of the paired samples t-test is used to discuss the logic underpinning the test and to consider what may be legitimately inferred.

Keywords— paired samples t-test, dependent design

I. INTRODUCTION

The paired samples t-test is a long-established procedure primarily used to statistically examine whether means derived from two dependent samples differ. In this context, the paired samples t-test is logically and numerically equivalent to the one-sample t-test applied to the differences between two dependent samples. Other texts (e.g. [1]) give a good mathematical description of the underpinning mathematics, statistical approximations, and subtleties of these tests.

Broadly speaking, for the two-group problem, the *t*-statistic is a "signal-to-noise" or "message-to-error" ratio. A big value for the *t*-statistic indicates there is a clear message in the data. As a rule of thumb "big values" are values in excess of 2 (in absolute terms) i.e., when the message in the data set is double what could reasonably be ascribed to chance. In the context of two dependent samples, the message or signal is how far apart the two means are. In the context of two dependent samples, the noise or error in the *t*-statistic is how accurately the mean difference is measured and this is referred to as the *standard error of the difference in means.* Mathematically, for dependent samples, the difference between two means is equal to the mean of the differences. Mathematically, for dependent samples, the standard error of the difference in means is equivalent to the standard error of the mean differences. It is for these reasons, that the paired samples *t*-test is logically and numerically equivalent to the one-sample *t*-test applied to the differences between two dependent samples. This will be illustrated using the motivating example.

The focus of this short note is to (a) give a worked example of the paired samples *t*-test (b) to discuss emerging issues, and (c) to reflect on what might limit the ability to generalize findings. The motivating example is given below. The example will be deconstructed using a series of questions.

II. A MOTIVATING EXAMPLE

A quasi-experimental study was carried out to determine whether children exhibit a higher number of aggressive acts after watching a violent television show. The number of aggressive acts for each child before and after the show is given in the table below.

Table 1 Number of aggressive acts before and after watching a violent television show

Child	Before	After
1	4	5
2	6	6
3	3	4
4	2	4
5	4	7
6	1	3
7	0	2
8	0	1
9	5	4
10	2	3

Question 1 What is the research question for this scenario?
Answer

The motivating research question is either "Does exposure to violent materials affect aggressive behavior?" or, the researcher may have a line of reasoning to have a predictive research question such as "Does exposure to violent materials tend to increase aggressive behavior?"

Irrespective, both research questions would be analysed using a two-sided statistical hypothesis (see [2]).

Question 2 What are the scientific hypotheses?
Answer

The scientific hypotheses would be

S_0: Aggression is independent of exposure to violent material

S_1: Aggression is dependent on exposure to violent material

Again, depending on context, S_1 could be predictive (and, as research is not done on a whim, S_1 would most likely be predictive).

Question 3 What is the independent variable in this scenario?
Answer

The independent variable is "Exposure to violent material" which has two levels, "Before" and "After".

Question 4 Do we have dependent or independent samples?
Answer

In this situation, each participant (child) is observed under two different states of nature (prior to exposure, and then after exposure). The same variable (number of aggressive acts) has been measured in each instance. Accordingly, we have dependent (i.e. paired) samples.

Question 5 What is the dependent variable?
Answer

The dependent variable is "number of aggressive acts".

Question 6 What are the statistical hypotheses?
Answer

Suppose watching an aggressive television programme does not affect aggression. If watching an aggressive television programme does not affect aggressive behaviour then on average we would expect no change in the number of aggressive acts. We could call this hypothesis the null hypothesis. What would we expect to observe if the null hypothesis is true? Suppose we consider a single child. For this single child would we expect a zero difference? A zero difference is certainly plausible. If the null hypothesis is true is it possible for a single child to show an increase in the number of aggressive acts or possibly a decrease in the number of aggressive acts? As the number of aggressive acts is not perfectly constant for one time interval to another then it is plausible to observe a non-zero difference for an individual and the null hypothesis to be true. If the null hypothesis is true then we would expect the average difference in the number of aggressive acts to be equal to zero. The last sentence is a bit deceptive; it does not mean the *sample mean* to be exactly equal to zero but rather the mean difference of all children in the population (i.e., the *population mean*).

Following the above reasoning, let μ_1 denote the theoretical mean number of aggressive acts before exposure and let μ_2 denote the theoretical mean number of aggressive acts after exposure. The statistical hypotheses would be

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Alternatively, let μ_D denote the theoretical change in mean number of aggressive acts. The statistical hypotheses would be

$$H_0: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

Note that even if the scientific rationale is predictive (i.e. even if the researcher has reason to believe that exposure will tend to increase the number of aggressive acts) this does **not** translate into one-sided hypotheses or a one-tailed test. One-sided hypotheses should only be used if one possibility can be logically discounted (in this case we cannot, pre-study, logically discount the possibility that exposure will reduce the number of aggressive acts), or if decision making (e.g. an interim analysis in a clinical trial might use a one-sided test for progression of a clinical trial). Essentially, we would nearly always consider two-sided hypotheses unless there was a compelling argument otherwise.

Question 7 What can be deduced from the following output?

Table 1 Descriptive Statistics for number of aggressive acts before and after

	N	Min	Max	Mean	Std. Dev
Before	10	0	6	2.70	2.058
After	10	1	7	3.90	1.792
Difference	10	-1.0	3.0	1.20	1.135

Table 2 Summary data from paired samples t-test

	Mean	Std Dev	Std Error Mean	*t*	*df*	*p*
[Before] [After]	-1.20	1.135	.359	-3.343	9	.009

Answer

Inspection of the data indicates that the number of aggressive acts increase after exposure for 8 out of the 10 children (with one decreasing, and one remaining the same). Before exposure, the mean number of aggressive acts was 2.7 per child. This mean rose by 1.2 aggressive acts per child on average post intervention to 3.90 aggressive acts on average.

The mean difference in the number of aggressive acts is 1.2 (this is the "signal"). We know that this sample mean difference is an estimate of the true intervention effect. We know that this sample mean will be in error (i.e. different samples of size 12 would give different estimates for the true difference). This noise or error referred to is the standard error of the mean and, in this case, has the value 0.358. Hence the signal-to-noise ratio, in absolute terms, is 1.2 ÷0.359 = 3.343.

In this case the signal-to-noise ratio, or more formally the sample *t*-value is 3.34 and is in excess of 2 given by the introductory rule of thumb. In fact the *p*-value (see [6]) is given as 0.009 indicating a statistically significant effect.
We would summarise this result as

"Analysis of the data using the paired samples *t*-test indicates that the mean number of aggressive acts has increased in the sample, from 2.70 pre exposure to 3.90 post exposure and this increase is a statistically significant increase (t = 3.343, df = 9, p = .009, two-sided)".

Note that this is a statistical conclusion; it does not attribute a causal change in the sample, or beyond the sample.

Question 8 What can be deduced from the following output given in Table 3 and Table 4?

Answer

Inspection of the data indicates that the number of aggressive acts increase after exposure for 8 out of the 10 children (with one decreasing, and one remaining the same). The mean number of aggressive acts rose by 1.2 aggressive acts per child. In this case the sample *t*-value is 3.34 and the *p*-value is given as 0.009 indicating a statistically significant effect. We would summarise this result precisely as before i.e.

"Analysis of the data using the paired samples *t*-test indicates that the mean number of aggressive acts has increased in the sample, from 2.70 pre exposure to 3.90 post exposure and this increase is a statistically significant increase (t = 3.343, df = 9, p = .009, two-sided)".

Note that this is a statistical conclusion; it does not attribute a causal change in the sample, or beyond the sample.

Table 3 Listing of differences (changes)

Child	Before	After	Difference
1	4	5	1
2	6	6	0
3	3	4	1
4	2	4	2
5	4	7	3
6	1	3	2
7	0	2	2
8	0	1	1
9	5	4	-1
10	2	3	1

Table 4 Output from the one sample t-test on differences

One-Sample Test

	Test Value = 0			
	T	df	Sig. (2-tailed)	Mean Difference
Difference	3.343	9	.009	1.200

III. A FORMAL STATISTICAL VIEW

Mathematicians and statisticians would probably lay out the thought procedures behind the test as given below. This layout makes reference to, (a) the mathematical formula for the t-test and the (see [1]), and the Central Limit Theorem (see [3])

* **Hypotheses**

$H_0: \mu_D = 0$

$H_1: \mu_D \neq 0$

* **Significance level**, $\alpha = 0.05$
* **Test Statistic**

$$t_{n-1} = \frac{\bar{x} - \mu_D}{s/\sqrt{n}}$$

$$t_{n-1} = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{\bar{x}}{s/\sqrt{n}}$$

(assuming the null hypothesis to be true)

* **Null distribution**

If the null hypothesis is true, and if the differences have been sampled from a normal distribution, then the test statistic will have a *t*-distribution with 9 degrees of freedom (*n*-1 = 10-1 = 9). Also, note, that if the data has *not* been sampled from a normal distribution but reliance can be placed on the results of the Central Limit Theorem then the test statistic will have a distribution which can reasonably be approximated by a *t*-distribution with 9 degrees of freedom.

*** Alternative distribution**.

If the true population mean difference is greater than 0 then we would anticipate that this will be reflected in the sample and in this case we would expect to see large positive values of the test statistic.

If the true population mean difference is less than 0 then we would anticipate that this will be reflected in the sample and in this case we would expect to see "large but negative" values of the test statistic.

In absolute terms we anticipate large values of the test statistic if the null hypothesis is not true.

*** Critical values** ± 2.262 (see statistical tables)

* **Decision Rule**

Reject the null hypothesis if the observed value of the test statistic is greater than +2.262 or if less than -2.262; otherwise fail to reject the null hypothesis.

* **Calculation.**

The sample mean is equal to 1.2, the sample standard deviation is equal to 1.135, the sample size is equal to 10 and hence

$$\frac{1.2}{1.135/\sqrt{10}} = 3.34$$

* **Statistical Decision.**

The calculated value for the test statistic does fall into the critical region and therefore we would reject the null hypothesis at the 5% significance level.

* **Statistical Conclusion**.

At the $\alpha = 0.05$ significance level there is sufficient statistical evidence to reject the null hypothesis that the population mean is equal to 0. In other words, the sample mean for the differences differs from a hypothesised population mean of 0 and this observed difference cannot readily be explained as a chance effect attributable to random variation through sampling. We conclude that there is statistical evidence that the mean number of violent acts has increased in the sample ($t = 3.34$, $df = 9$, $p = 0.009$, two-tailed).

Question 9 What, if anything, can we scientifically conclude from this study?

Answer

Not a lot! In the sample the mean has changed by a statistically significant amount; however we cannot really say, "for this sample there has been a causal effect attributable to the one violent tv show" ; for instance there is no control group (perhaps a control with non-violent material), or we could not rule out other competing explanations e.g. the children might be hungrier after watching the show and this might make them "hangry". Essentially, the internal validity is compromised.

Even if we could argue for good internal validity then there is no way we could argue that $n = 9$ is representative of a wider population, or that this one specific tv show was representative of all violent shows.

Scientifically, we cannot conclude anything other than arguing there is prima facie evidence to conduct more and better research into this phenomenon.

IV. Discussion

This note has covered a single application of the paired samples *t*-test. This paired sample *t*-test can be used in a two level matched design, or a two level repeated measures design (see [4]).

In the classical derivation of the test there is an assumption that the paired *differences* have been sampled from a normal distribution. In practice perfect normality will not exist but the test does work well if there are minor departures from normality. In fact the assumption is whether the *mean* difference *is normally distributed*. *If* the differences are normally distributed *then* the mean difference will also be normally distributed. That said, the mean difference could be approximately normally distributed by virtue of the Central Limit Theorem even if differences are not normally distributed (see [3]). By way of example, the paired *t*-test could be used with Likert-like data which is clearly discrete and clearly non-normal (see for instance, [5]). Indeed, in the give example, the number of aggressive acts, or the difference in aggressive acts, cannot be normally distributed, due to the fact that this count data is (a) discrete with (b) a limited range whereas the normal distribution is (a) continuous with (b) an infinite range.

It is widely recognized that statistical significance is, in itself, not the complete story and effect size should also be reported. Effect size for the two-group problem is covered in [6].

Acknowledgments

The authors gratefully acknowledge the support of the Learning and Teaching Initiative in the Faculty of Health and Applied Sciences, University of the West of England, Bristol in supporting the wider Qualitative and Quantitative Methods teaching programme.

SERIES BIBLIOGRAPHY

White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –6

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5

White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5.

White P, Redford PC, and Macdonald J (2019) That assumption of normality, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –11

White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent samples, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

REFERENCES

[1] Howell DC (2019) Statistical Methods for Psychology, Wadsworth: USA

[2] White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, Qualitative and Quantitative Research Methods Project, University of the West of England, 1 –4

[3] White P, Redford PC, and Macdonald J (2019) That assumption of normality, Qualitative and Quantitative Research Methods Project, University of the West of England, 1 –11

[4] White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, Qualitative and Quantitative Research Methods Project, University of the West of England, 1 –4

[5] Derrick, B. and White, P. (2017) Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics*, 18 (3).

[6] White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent samples, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4