An example motivated discourse of the independent samples t-test and the Welch 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— Analyses are given for two example scenarios. Both scenarios comprise two independent samples. One scenario is analysed using the independent samples *t*-test; the other the Welch test. Commentary on the designs and limitations is given through a question and answer process.

Keywords— two-sample t-test, Welch test

I. INTRODUCTION

The independent samples *t*-test is a long-established procedure primarily used to statistically examine whether two means differ based on an assumption of equal variances. A variant of this test, the Welch test (aka Welch-Aspin test, Welch-Aspin-Satterthwaite test), relaxes the assumption of equal variances. Other texts (e.g. [1]) give a good mathematical description of the underpinning mathematics, statistical approximations, and subtleties of these tests.

In essence, 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 i.e., when the message in the data set is double what could reasonably be ascribed to chance. In the context of a two-group problem, the message or signal is how far apart the two means are. In the context of the two-group problem, 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 mean differences*.

The focus of this short note is (a) to give two worked examples which employ these two statistical tests, (b) to discuss emerging issues, and (c) to reflect on what might limit the ability to generalize findings. The motivating examples are described below. The examples will be deconstructed using a series of questions.

II. A MOTIVATING EXAMPLE [EXAMPLE 1]

A dietician wants to evaluate a new low-fat diet she has developed compared to a regular established diet. 60 obese people were selected. 30 of the participants were randomly allocated to the new low-fat diet and the remaining 30 were placed on the regular diet (the regular diet is a managed programme distinct from an *ad lib* diet). At the end of 3 weeks the weight loss (in pounds) of each of the 60 people was measured.

One person on the low-fat diet confessed that they had broken the diet by supplementing the daily intake with curry and beer. The observation for this person was deleted. The resulting data are given in Table 1.

Table 1. Weight lost (in pounds)

Low Fat				Regular		
11.8	10.9	7.7		10.8	5.9	6.6
4.7	11.2	8.9		0.4	9.3	5.2
7.3	11.2	7.6		0.0	4.6	1.5
7.3	8.3	5.7		7.7	9.1	3.0
10.5	6.2	11.7		9.0	9.9	3.6
9.0	10.6	0.7		5.90	4.7	4.4
16.3	16.0	11.9		5.60	10.2	2.8
13.5	4.0			15.4	5.1	5.4
0.2	4.5			0.10	0.3	
14.1	13.8			6.90	1.9	
7.2	8.4			11.3	9.3	

Question 1 Why do you think random allocation was used, and what are the advantages of random allocation?

Answer

The participants would *not* have been *randomly selected*. They are probably volunteers (and they could possibly be a convenience sample). However, consenting participants

could be *randomly allocated* to treatment (either Low Fat, or Regular). This random allocation could be done using a random number table and be done to ensure an equal number in each treatment group.

Undoubtedly, participants will differ from one another in many respects (e.g. different ages, or different number of diets tried in the past, or different motivation to lose weight, or they will differ in initial weight and so on). These individual differences may, or may not, be related to weight loss. These uncontrollable individual differences are often referred to as covariates (i.e. things which might co-vary with outcome). Participants will also differ on things we might not even think about. If we randomly allocate participants to groups, then we would expect, on average, that these individual differences or covariates would balance themselves between the two groups permitting a fair comparison. Of course, there is no guarantee that this will be the case in any one instance.

Relatedly, if the effect of the Low-Fat diet is the same as the effect of the Regular diet (i.e., they are both equally effective, or both equally ineffective) then under random allocation we would anticipate no difference in weight change between the two. On the other hand, if one diet is more efficacious than the other then under random allocation we would anticipate this effect being captured in the sample. Moreover, if covariates are balanced between the two groups then this would rule out the possible explanation that the observed effect is due to the covariates. Hence, random allocation helps with a causality argument (i.e. random allocation helps rule out covariates as being an alternative competing explanation for an observed difference in means).

Additionally, the data under the design will be analysed using statistical methods which are based, and mathematically developed, on assumptions of either random selection or random allocation. Accordingly, the use of random allocation permits a logical justification for analysis using formal statistical methods.

Question 2 What is the research question for this study?

Answer

Undoubtedly, the motivation behind the research is to show that there would be greater weight loss attributable to the new Low-Fat diet. The research question would then be "Is the new low-fat diet better than the regular diet?" This is an example of a superiority study i.e. one treatment arm being superior than the other.

[As an aside, it is conceded that sometimes we might want to consider whether a new treatment is equivalent to, or alternatively, not inferior to an existing treatment. For these situations there are other branches of statistics which deal with *equivalence* and *non-inferiority*.]

Question 3 What are the scientific hypotheses for this study?

Answer

S_0: The new low-fat diet and the regular diet are equally good.

S_1: The new-low fat diet is better than the regular diet.

Question 4 What could be the statistical hypotheses for this study?

Answer

Let μ_1 denote the theoretical mean weight loss under the Low Fat diet and let μ_2 denote the theoretical mean weight loss under the Regular diet. The statistical hypotheses would be

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

Note that even though the scientific rationale is predictive (i.e. the researcher has reason to believe that Low-Fat will be better than Regular) 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 Regular diet may be better than the Low-Fat diet), 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 and unless there was a compelling argument otherwise.

Question 5 What is the independent variable in this study? How many levels does it have?

Answer

The independent variable is Diet. Diet has two levels (Low-Fat, Regular)

Question 6 What is the dependent variable?

Answer

The dependent variable is weight loss (in pounds) over the three-week trial duration.

Question 7 Do we have independent or dependent samples?

Answer

This is an independent design (aka a between-subjects design). Participants are in one group and one group only; there is no logical mechanism of matching any one person in Low-Fat with another in Regular; we would further assume weight loss in one person does not affect weight loss in another.

Question 8 The independent samples *t*-test has been used to analyse the data. What can be concluded from the following computer output?

Answer

From Table 2, it can be seen there is a greater average weight loss in the Low-Fat group (M = 9.0 pounds) than the average weight loss in the Regular group (M = 5.9 pounds). Not all participants lost the same amount of weight. The standard deviation, which approximates, on average how far participants deviate from the average, is marginally higher in the Low-Fat group (SD = 4.0) than the Regular group (SD = 3.8).

Table 2 Group Statistics

	Diet	N	Mean	Std. Dev
Weight Loss (Pounds)	Low-Fat	29	9.007	4.034
	Regular	30	5.863	3.811

Analysis using the independent samples *t*-test indicates that the differences between sample means cannot easily be explained as a chance outcome (p = .003).

Table 3 Independent Samples Test

	t-test for Equality of Means			
	t	df	Sig. (2 tailed)	Mean Difference
Weight Loss (Pounds)	3.078	57	.003	3.143

In Table 3 it is noted that the t-statistic, i.e. the signal to noise ratio, is 3.078 and the associated *p*-value (.003) is smaller than the usual threshold of 0.05 used for statistical significance.

A summary for a results section of a report could be along the following lines: "Analysis of the data using the independent samples *t*-test indicates that mean weight loss in the Low-Fat group (M = 9.007, SD = 4.034) is significantly greater than mean weight loss in the Regular group (M = 5.863, SD = 3.811) (t = 3.078, df = 57, p = .003, two-sided).

Note

(a) It is not sufficient to say that the means significantly differ; the direction of effect should be given.

(b) We would also give the effect size (see [3]).

(c) The above is a statistical conclusion and should not be confused with a scientific conclusion. Scientific conclusions can only be considered after considering, any, and all limitations.

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

There are a number of positives with this design, most notable the use of random allocation and a dependent variable which should not suffer from measurement error.

The first question we should ask is whether we can really argue, that *for the sample participants*, can we really attribute the increased weight loss in the Low Fat group to the diet? Clearly, the participants in the study would not be naïve to its purpose and may modify their behavior in different ways e.g. they might take up more exercise. Accordingly, the weight loss itself may be a function of both diet and other factors. However, it is arguable that these "other factors" would apply to both diets.

We should also recall that an observation was deleted because the participant did not adhere to the low-fat diet regime. Is the deletion of this observation justifiable? In an Intention to Treat analysis (ITT, or Intent to Treat) the observation would not be deleted. For instance, suppose the low-fat diet had a very high drop-out rate because participants found it difficult to adhere to the regime. In this situation, ignoring the drop-out cases, and only reporting on those that adhered to the diet (i.e. reporting possibly only on those with a high degree of motivation to lose weight), would seemingly give results that would put the low-fat diet in a favourable light. The deletion of the data for the non-adherent participant weakens the internal validity. Of course, there may be others who broke the diet and the investigator is unware of these violations. Treatment *fidelity* should be considered.

What about the external validity i.e. the ability to generalize? The main problem is sample size. Firstly, n = 30 per group is a small number to argue for "representative" data. Secondly, small data sets can give variable results. At best we can say there is prima facie evidence of greater weight loss on average in the Low-Fat diet group compared to Regular but this is very much subject to confirmation using much larger samples and using different populations.

III. A MOTIVATING EXAMPLE [EXAMPLE 2]

The following data (Table 4) relate to reaction times (in milliseconds) in a sample of n = 9 male lecturers between the age of 50 and 55, and a sample of n = 13 male students between the ages of 20 and 25. All n = 22 attended the same university.

Table 4 Reaction times (in milliseconds)

Age 50 to 55		*Age 20 to 25*	
134	216	126	276
139	216	131	282
142	227	139	289
164	289	160	301
189		238	301
		251	335
		251	

Question 10 Give the research question and scientific hypotheses for this study.

Answer

Presumably the research question is "Are reaction times age dependent?". It could be the case that the researcher has a rationale to suggest reaction times increase with age (i.e. people become slower with increasing age) and in this case the research question might be "Do reaction times decrease with age?". Irrespective, the scientific hypotheses could be

S_0: Reaction times are not age dependent

S_1: Reaction times are age dependent

Question 11 Would it have been possible to use random allocation in this study? What about random selection? What are the populations of interest? Is there confounding?

Answer

Random allocation cannot be performed. You cannot randomly allocate a participant to an age group; their age is their age.

Technically random selection could be done. This would require two lists. One list being a list of all male students who meet eligibility criteria (aged 20 to 25) and another list of a male lecturers aged 50 to 55. Accessing these lists might be a major barrier. If the lists could be accessed, then certainly random selection could be performed. In practice though, it is probably the case that participants are recruited using a convenience sample or through a volunteer sample, but these approaches might create bias.

The populations of interest, would presumably be (from the researcher's perspective), males aged 20 to 25 and males aged 50 to 55. However, the participants are all from the same university and arguably the populations are males aged 20 to 25 at that particular university, and males aged 50 to 55 at that particular university. However, all those aged 20 to 25 are students and all those aged 50 to 55 are lecturers. Age and occupation are completely confounded; we cannot separate out their effects. Are we testing for an effect due to Age, or an effect due to Occupation? This could be a major limitation on the conclusions which could be drawn.

Question 12 What is the independent variable in this study? How many levels does it have? What is the dependent variable? Do we have independent or dependent samples?

Answer

The independent variable is age group which has two levels (20 to 25, 50 to 55). Of course, it could be argued that the independent variable is occupation with two levels (student, lecturer). Or you could argue that the IV has two-levels comprising male students aged 20-25 and male lecturers aged 50-55.

The dependent variable is reaction time.

This an independent samples design (a between-subjects design) as each participant is in one group, and there is no logical way to match participants between the two groups.

Question 13 What could be the statistical hypotheses for this study?

Answer

Let μ_1 denote the theoretical mean reaction time for those aged 50 to 55 and let μ_2 denote the mean reaction time for those aged 20 to 25. The statistical hypotheses would be

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

Question 14 Does it matter that we have unequal sample sizes?

Answer

Having unequal sample sizes does not invalidate the statistical analysis. However, if you were to design a two-group independent study would you, intuitively, aim to have equal or unequal numbers in each group? Most would say "equal". In fact, the power of a two-group study is dictated by the smaller sample size. As a rule, the smaller the sample size the smaller the power. (Power being the probability that the correct statistical decision is made.) For fixed resources the largest smallest sample size would be when the two samples have equal sample sizes.

Question 15 The data has been analysed using Welch's test with relevant output given below. What can you conclude from the output?

Answer

In the sample, the mean reaction time for the 50 - 55 age group (M = 237ms) is higher than the sample mean reaction time for the 20 – 25 age group (M = 191ms). However, in

both groups there is extensive variation (*SD* = 51.5ms in the 20-25 age group, and *SD* = 72.8ms in the 50 – 55 age group).

Table 5 Group Statistics

	Age group	N	Mean	Std. Deviation
Reaction time (msecs)	20 – 25	9	190.67	51.522
	50 – 55	13	236.92	72.760

In Table 6 the *t*-statistic has an absolute value of 1.746 (p = .096) indicating that the difference in means (46.3ms) is not sufficiently large enough to cast doubt on this difference being anything other than chance or natural sampling variation.

Table 6 Independent Samples Test

	t-test for Equality of Means			
	t	df	Sig. (2-tailed)	Mean Diff
Reaction time (msecs)	-1.746	19.97	.096	-46.26

Analysis of the data using the Welch t-test indicates that the difference in mean reaction times between the two groups is not a statistically significant difference (t = 1.746, df = 19.97, p = .096, two-sided).

Question 16 What, if anything, can we scientifically conclude from this study?

Answer

Nothing. The difference between means does not achieve statistical significance; this does not translate to equal means, as espoused by the mantra "*absence of evidence is not evidence of absence*" [4]. Age and occupation are *confounded* and the sample is from one university; it would therefore seem that the groups are not 20-25 versus 50-55 but are male students aged 20-25 at a particular university versus male lecturers aged 50 – 55 at the same particular university. As this is a volunteer sample we cannot have any certainty that there are no other biases and/or whether the samples are representative of those in the university (which is highly unlikely given the small sample sizes).

IV. Discussion

The two examples in this paper have been discussed without mentioning the normal distribution. It is often stated that a necessary assumption for a valid application of the two versions of the *t*-test is for the "data to be normally distributed". This is not true.

It is certainly true that the initial mathematical derivation of the sampling distribution for the independent samples *t*-test was based on an assumption of normality. However, mathematically the requirement is for the *means* to be normally distributed and not necessarily the data. It is also true that if the means are approximately normally distributed then the *t*-test works well.

There is another theorem in statistics, the Central Limit Theorem (CLT), which indicates that *means* will have a distribution which approximates the normal distribution to a greater or lesser degree. The quality of this approximation is dependent on sample size (the bigger the sample sizes the better the approximation). The quality of this approximation is also dependent on the degree of skewness in the originating population (large skew works against having a good approximation). With very little skew reliance can be placed on the results of the Central Limit Theorem for sample size of approximately 15 or larger; for moderate skew samples of size n = 30 or larger would be needed, and for higher degrees of skewness sample sizes of up to n = 60 might be needed for means to be approximately normally distributed. If the originating data was from a normal distribution, or an approximate normal distribution, then the *t*-test is valid for any sample sizes.

In the first example the independent samples *t*-test was used (assuming equal variances) and in the second the Welch test was used. Why? In the first example (the diet example) participants were randomly allocated and in the absence of an effect the random allocation would ensure equality of means and equality of variances. In the second example (reaction times) the two groups are pre-existing groups which might differ in means. If you are open to the possibility that the means might differ then you should be open to the possibility that the variances might differ.

If you use the independent samples *t*-test when population variances are not equal then this really becomes problematic if sample sizes differ. If you use the Welch test when population variances are equal then this is not problematic providing the originating distribution is approximately normally distributed. For these reasons some mathematical statisticians have made a clarion call to always using the Welch test [5] but also see [6].

Some may hold the view that the sample data should be formally tested for normality and the assumption of equal variances formally tested using Levene's test. However, it is well known that preliminary testing to choose a statistical test has ramifications on the rates of false positive and false negative findings. So tread with caution.

In this brief note we have not considered "effect size". A statistically significant finding might not necessarily reflect a finding of substance (i.e. it might not be practically

meaningful or clinically meaningful). Effect size for the two-group problem is covered in [2].

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) Cohen's *d* and the *t*-test, *Qualitative and Quantitative Research Methods Project*, Univeristy of the West of England, 1 –***

[3] Cook C (2013) Clinimetrics Corner: Use of Effect Sizes in Describing Data, *The Journal of Manual & Manipulative Therapy*, Vol 15, No 3, 54 – 57

[4] Altman, DG; Bland JM (1995), Absence of evidence is not evidence of absence, *British Medical Journal*. Vol 311, 485.

[5] Ruxton, G. D. (2006), The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, *Behavioral Ecology*, Vol 17, 688–690

[6] Derrick, B; Toher, D; White, P (2016), Why Welch's test is Type I error robust, *The Quantitative Methods for Psychology*, Vol 12. No 1, 30–38.