Parametric hypothesis tests for the difference between two population means

L N Dzikiti, MSc; **B V Girdler-Brown**, FCPHM, FFPH, BCom Hons (Econ)

School of Health Systems and Public Health, Faculty of Health Sciences, University of Pretoria, South Africa

Corresponding author: *B V Girdler-Brown (brendangirdlerbrown@gmail.com)*

When one wishes to perform a statistical-hypothesis test, the first important step is to select the correct, most appropriate, test to perform. This article aims, firstly, to outline the test-selection criteria when one wishes to obtain statistical evidence about the equality of population means. Thereafter, Stata statistical software (StataCorp, USA) commands will be given for the various tests.
If the population means of interest are numerical and have known probability-sampling distributions, then the standard recommended statistical-hypothesis test, with data from samples, is either the classic *t*-test, or a variant of it.
In this article, the first part concentrates on the selection of the appropriate test to perform when using sample data to determine whether two population means are likely to differ. The selection criteria/steps are illustrated by a diagram (Fig. 1).
The second part describes how to perform the tests using Stata. In that section, some Stata output is also presented, and the interpretation of the output is explained.

Strengthen Health Syst 2017;2(2):40-46. DOI:10.7196/SHS.2017.v2.21.60

When one wishes to perform a statistical-hypothesis test, the first important step is to select the correct, most appropriate, test to perform. This article does not attempt to explain the underlying theory of the statistical tests it describes. Readers are urged to acquaint themselves with the theory by reading, or dipping into, a good textbook on the subject. There are many good textbooks available, but the one we recommend is by Pagano and Gauvreau.[1] Their explanations are clear, up to date and easy to understand.

Furthermore, the focus in this article is on the various two-sample *t*-tests and how to perform them using Stata (StataCorp, USA). Therefore, manual calculations will not be described. In addition, details about how to perform single-sample tests, analysis of variance (ANOVA) (for when one wishes to compare more than two independent-sample means) and non-parametric tests, as important as these topics are, will not be covered here. The authors have assumed a prior basic understanding of the principles of hypothesis testing, including the following concepts: the difference between population parameters and sample statistics; sampling error; the null hypothesis and the null value; alpha and the *p*-value; beta and and the power of a test; the 95% confidence interval (CI) for the difference between two means; type I and type II errors; and single-tailed v. two-tailed tests.

Selecting the appropriate test

Parametric or non-parametric tests?

Hypothesis tests of the difference between two population means are performed using data from two samples that are assumed to have been selected in a random or probabilistic way. The null hypothesis is of the form:

H_0: $\mu 1 - \mu 2 = 0$

Two population means may be compared using either parametric or non-parametric ('distribution-free') hypothesis tests. Parametric tests are performed when one knows the sampling probability distribution for the difference between the two means, and the assumptions for the parametric tests have been met. Non-parametric tests are used when the nature of this sampling distribution is not known and cannot be surmised, or when the assumptions for a valid performance of the parametric test have not all been met. Although non-parametric tests have fewer restrictions than parametric tests, one should be aware that they also have conditions for their appropriate performance and these should always be checked for before embarking on a non-parametric test.

Whereas parametric tests will result in a *p*-value, as well as a 95% CI, for the difference between the two means, the non-parametric tests will only produce a *p*-value. The null hypothesis for a non-

parametric test may not be that the two population means are equal. In addition, in general (certainly not invariably, however), non-parametric tests are less powerful than parametric tests. This is because non-parametric tests generally make use of less of the information contained in the sample.

For all these reasons, when performing a test to establish whether there is likely to be a difference between two population means, one should usually use a parametric test in preference to a non-parametric option, if conditions for the parametric test have been met.

Independent samples or paired samples?

The parametric tests for comparing means from independent samples are the *t*-test and the Welch test. The paired *t*-test is a suitable parametric test for comparing means from paired samples. Measurements are said to be paired when they are taken on the same unit of study (e.g. the same person or the same facility, depending on the unit of analysis). In some cases, where two groups of study participants are very closely matched, it may also be acceptable to treat the two groups as 'paired'. When pairing is present, the paired *t*-test is in general more powerful than treating the two sets of readings as independent samples and then performing a *t*-test or Welch test. Therefore, if pairing is present, rather perform a paired *t*-test.

Some examples of paired data sets would include the weights of a group of people who were measured before the participants started a diet and exercise programme, and then measured again after a suitable interval. Another example might be sets of anatomical measurements of distances between surface landmarks on the left- and right-hand sides of the body.

Conditions for the *t*-test and the Welch test

For the independent samples situation, valid use of either of the parametric hypothesis tests mentioned (the *t*-test and the Welch test) requires that for both samples being compared, the data are drawn at random from a population of data that have a normal, or bell-shaped, frequency distribution. If this assumption is not satisfied, or if there is uncertainty as to whether or not it is satisfied, then distribution-free methods should be considered, unless the sample sizes are large.

In practice, if the frequency distributions of the two samples appear approximately bell-shaped (unimodal and not too skewed) then it is safe to consider using these two parametric tests. If sample sizes are small, however, say <30, then one may wish to perform, first, a hypothesis test, such as a Shapiro-Wilk test, to assess whether each sample is likely to have been drawn at random from a normally distributed population of data.

For large samples (>30) it is unecessary to perform tests of normality, as the *t*-test and the Welch test are robust against departures from normality when samples are large. As a result, with large samples, there is also usually no need to first transform the data (for example, using log transformations with positively skewed data). One has to be clear, however, that for skewed data, even if the samples are large, the mean may not be an appropriate measure of central tendency, so one must first satisfy oneself that comparison of the differences between two means is a useful exercise.

Should one use the *t*-test or the Welch test?

These two tests are used to compare means from two independent samples, as in the situation where, for example, the mean birth weight of babies born to smoking mothers is compared to the mean birth weight of babies born to non-smoking mothers. However, there is a specific independent-samples *t*-test assumption that must be met for the *t*-test, in that the variances of the two samples to be compared must be equal. A Welch test should be used if the variances are not equal.

With the Welch test, the variance of the difference between the two means, as well as the degrees of freedom, are calculated differently from in the *t*-test calculations. As a result, the 95% CI for the difference between the two means, as well as the *p*-value, will be different from those obtained using a *t*-test.

When the variances are unequal, then the (inappropriate) *t*-test and the (appropriate) Welch test will often give quite different results. The differences in the results become smaller, however, as the differences between the variances become smaller, if the sample sizes are equal and if the sample sizes become larger. The extent of this similarity (between the two test results), however, varies depending on the size of the differences in the sample variances.

Performance of the Welch test does not require that the population variances should be unequal. The Welch test may be performed whether or not the population variances are equal.

The *t*-test, however, requires that the two population variances may be assumed to be equal. How do we decide whether the variances are 'equal' or not? Many older statistical textbooks suggest that the *F*-test be conducted to assess the equality of variances. However, more modern text books, such as Pagano and Gauvreau,[1] discourage the use of the *F*-test to assess whether or not the two population variances may be considered to be equal. The *F*-test may lack sufficient power to correctly point the analyst away from an inappropriate *t*-test in many cases.[1,2] This is a particular risk when sample sizes are on the small side (say <20). We encourage our students to use visual inspection of the sample variances, and, if in doubt, to perform the Welch test rather than the *t*-test.

If the sample sizes of the two samples being compared are equal, then the Welch and *t*-tests give almost identical results. When these two sample sizes are both large and equal (say >30), the degrees of freedom (DF) are somewhat different (Welch v. *t*-test), but are so large as to not make any difference in practice.

When the sample sizes are equal but small (say <30), the DF are almost the same, so that once again, it makes no difference which of the two tests is used.

Some authors go so far as to suggest that the Welch test be used, rather than the *t*-test, in all situations.[3]

We recommend the following approach (also illustrated in Fig. 1):

- If sample variances are the same or vary by a very small degree only, use the *t*-test
- In all other cases, or if in doubt, use the Welch test.

Conditions for the paired *t*-test

For the paired *t*-test, the only assumption that must be met is that a single data set made up of the differences between each of the two paired readings should have a bell-shaped frequency distribution ('unimodal and not too skewed'). This assumption becomes less important for larger sample sizes owing to the robust nature of the *t*-test.

Hence one should first calculate the differences between the two readings for each study participant, and then examine the frequency distribution of this newly calculated set of differences. This may be done by inspection of a frequency histogram, or, especially for small samples (<20), by performing a Shapiro-Wilk test.

This is especially important for small (say <20 pairs) studies. If the frequency distribution is clearly skewed, or not unimodal, then one should rather consider performing a non-parametric test, or perhaps transforming the differences to a form that has a bell-shaped curve (if this is a meaningful thing to do; it may not be).

Non-parametric analogues of the parametric tests

The usual non-parametric equivalent of the *t*-test is the Mann-Whitney-Wilcoxon (MWW) rank sums test. The usual non-parametric equivalent of the paired *t*-test is the Wilcoxon signed-ranks test. Both these tests can be easily performed in Stata. The null hypotheses may differ from those for the parametric tests, however, so they may not be truly 'analogous'. In addition, be sure to check that the assumptions for the non-parametric tests have been met before performing these tests. Just because these tests are 'distribution free' does not mean that they are assumption free.

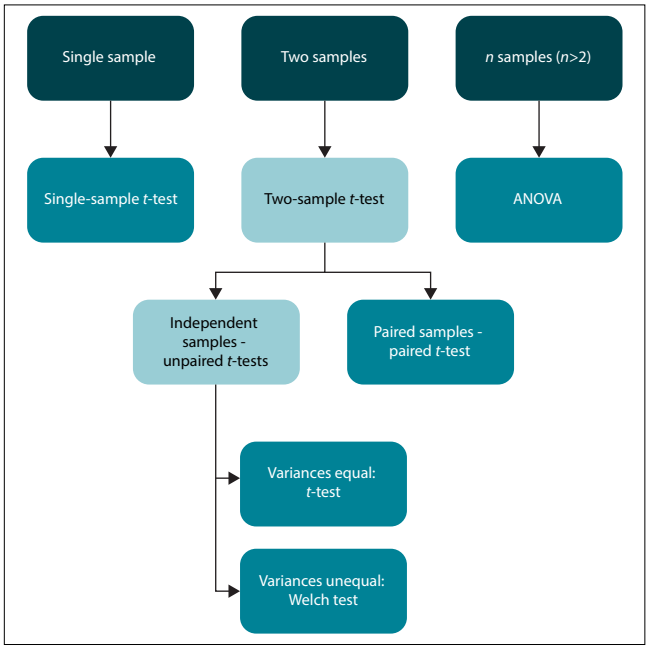

Fig. 1. A flow chart to assist with the selection of an appropriate parametric hypothesis test of population means.

Single sample *t*-tests

As illustrated in Fig. 1, one can also perform single-sample hypothesis tests where one compares the population mean to a fixed known value such as a gold standard, benchmark or target. For example, the mean normal birth weight of a sample of babies may be compared to an expected standard as defined by the World Health Organization. In this case a single sample *t*-test could be used. The null hypothesis for such a test is of the form:

H_0: μ = standard (where 'standard' = the gold standard, benchmark or target).

Once again, if the sample is small (say <30) the data should be unimodal and not too skewed. If this is not the case (for small samples), then a non-parametric test such as a single-sample sign test should rather be considered. As the sign test is less powerful (and actually assumes that the standard measure is a median rather than a mean), you should try to rather ensure that you have a large enough sample size (>30, say) so that you may use the *t*-test.

A brief note on the *z*-test (also sometimes referred to as a 'normal test')

The *z*-test is rarely used nowadays. It is performed in the same way that one performs the *t*-test, except that it makes use of the population variance rather than the sample variance in calculating the CI and the *p*-value. This would seem desirable.

However, it is very rare that one knows the population variance (and does not know the population mean). The *t*-test procedure overcomes this problem (that we do not know the population variance) by substituting the sample variances for the population variances. As the sample variances are likely to be inaccurate estimates of the population variance (since they are subject to sampling error), this may sometimes result in unduly low estimates of the sample variances, resulting in type I errors.

With the *t*-test or Welch test this risk is mitigated by calculating wider CIs from the sample variances, and higher *p*-values, than would have been the case for a *z*-test. Of course, if the sample sizes are large then the *z*-test and *t*-test results will be similar even if one performs the *z*-test by using the sample variances substituted for the unknown population variances. For this reason, in some older textbooks, it was sometimes stated that one might perform a *z*-test if the samples were both, say, >30.

This option was considered desirable since the *z*-test does not require the normality assumption of the *t*-test if sample sizes are >30. However, simulation studies have shown that, for sample sizes

>30, the *t*-test is sufficiently robust to give valid answers even if the samples are not drawn from normally distributed populations.

It is, therefore, acceptable to perform the *t*-test when the population variances are not known, as is usually the case for large sample sizes, even though the normality assumption is violated. It is perhaps for these reasons (the relative obsolescence of the *z* test) that Stata does not offer a *z*-test option as part of its standard package.

Performing the analyses using Stata statistical software

One-tailed or two-tailed tests?

Having decided which hypothesis test to use, the next consideration is to decide whether one wishes to perform a one-tailed or a two-tailed test. This decision does not affect the Stata command that you will use. Stata will present all the results for single upper tail, single lower tail and two-tailed tests.

One perfoms two-tailed tests if one is testing the null hypothesis that there is no difference between the population means. Since one does not wish to prejudge the outcome of the analysis, one is usually expected to perform a two-tailed test.

One encounters single-tailed tests, for example, in quality-control studies. Let us assume that a provincial health department set a performance standard of at least 90% immunisation coverage for the measles vaccine in children aged 1 year for 2016.

The researcher would only be concerned whether the coverage were to be less than the performance standard set. There may be no interest in whether the standard had been exceeded (that would be a good outcome). In such a case, a single-tailed test of the null hypothesis that the coverage is <90% would often be performed.

Data layout

Data may be entered into Stata in either the wide or the long format. For paired tests, the data need to be in the wide format. For tests with independent samples, the data may be in either the long or the wide format. Below we give an example of data that are in the long (Table 1 A) v. wide (Table 1 B) format for independent samples (females and males are the two independent groups being compared).

The code used in Table 1 A for 'female' is 1 = female, while 0 = male. Note that in this case we do not have a 'participant_id' for the wide format, as the female with bodymass 65.3 kg cannot be the same person as the male with bodymass 60.9 kg.

It is usual in Stata to have data entered in the long format, as we would normally like to have participant_id for each entry, with rows of data that pertain to a particular participant. If there are variables that need to be entered in the wide format (such as repeated measures) this may easily be accommodated, so that the resulting dataset may contain some variables in the long format and some in the wide format.

Table 3 gives an example of both independent and paired data in the same table. Since there are also independent data present in the table, with these paired data, there is a participant_id entry specific to each participant row.

Pre- and post-treatment systolic blood pressures are measured in the same individuals and are thus paired data, which are presented alongside each other (in the wide format) in this case.

Selecting a value for alpha

It is usual in biostatistics to use a *p*-value of 0.05 as a cut point for deciding if a result is statistically significant. The *p*-value that we obtain from the hypothesis test that we perform is the probability of obtaining the observed results, or more extreme results, by chance or sampling error if there really is no difference between the population means. Using p=0.05 as our cut point means we would reject the null hypothresis if $p \leq 0.05$. If this is the case, we might state that the test results are 'statistically significant'. If $p > 0.05$, we would fail to reject the null hypothesis, concluding that the results are not statistically significant.

This special value of *p*, namely 0.05, that is used to decide whether or not our results are statistically significant is called α (alpha). Note that we should never 'accept' a null hypothesis and/or conclude that two parameters are equal.

Table 1. The same data entered in the long format (A), and re-entered in the wide format (B)

A. The long format layout			B. The wide format layout	
Participant_id	Bodymass	Female	Female_bodymass	Male_bodymass
1	65.3	1	65.3	60.9
2	60.9	0	54.4	67.2
3	54.4	1	59.1	
4	59.1	1		
5	67.2	0		
etc.	etc.	etc.		

Table 2. A hypothetical example of a Likert-type questionnaire item

I am satisfied with the clinic opening times						
Strongly disagree	1	2	3	4	5	Strongly agree

Table 3. Data with a mix of long and wide formats, with paired data in the wide format

Participant_id	Bodymass	Female	pretreatmentsystolicbp	posttretamentsystolicbp
1	65.3	1	90	88
2	60.9	0	95	85
3	54.4	1	88	80
4	59.1	1	98	100
5	67.2	0	90	90
etc.	etc.	etc.	etc.	etc.

The reason why we failed to reject the null hypothesis might indeed be because the null hypothesis is true. However, it may also be due to the fact that we have samples that are too small, or that our test is underpowered, or that our measurement methods have been imprecise, and so on. We are not entitled to assume that the reason for statistical non-significance is that the null hypothesis is true.

Now, with a cut point of α=0.05, there is a probability that 5% of all null hypotheses that are true will test positive (type I errors) just because of sampling error (due to the variable composition of the samples that we are using and for no other reason). In other words, it is likely that we are wrong 5% of the time when we reject a null hypothesis and claim that our results are 'statistically significant'.

If we set α=0.05, and then perform 100 pairwise *t*-tests of 100 known false hypotheses, using independent samples, the type I error would be 0.05 per test performed. The probability of committing at least one type I error when performing so many tests will exceed our planned level of 0.05.

When a researcher decides to collect data for a large number of variables, and then aimlessly perform pairwise hypothesis tests on all of them, in the hopes of finding 'something significant', the likelihood of finding statistical significance as a result of type I errors is increased. Most textbooks are agreed that this is especially problematic when the decision to perform these hypothesis tests is made after data have been collected, although many do not agree that this is problematic if the tests are specified before data are collected.

However, just because a person lists every possible hypothesis test possible in the protocol, before collecting the data, ('just in case'?) this does not diminish the risk of type I errors. Hence it is our view that in all cases where multiple pairwise hypothesis testing is carried out (either routinely or without good a priori arguments for their salience), a lower value of α should be required in order to establish statistical significance.

The Bonferroni adjusted value of α is a widely used adjustment (there are others) and is easily calculated as 0.05/T, where T is the proposed number of comparisons to be made. For five pairwise comparisons we would therefore use α=0.01 as the value to determine statistical significance (rather than 0.05).

A brief note on the analysis of Likert-style questionnaire data sets

Likert-style questionnaires are of the following type where respondents are asked to check a single cell that best indicates their answer, as illustrated by Table 2.

One may then have two sets of responses, for example, one from a group of respondents working in the formal sector, and one from those who also work, but in the informal sector, on this issue. One wonders if there is a statistically significant difference between the responses of these two groups. The values 1 - 5 are somewhat arbitrary, and are definitely neither continuous nor quantitative, although they are ordinal. Theoretically, one should not perform a *t*-test on these results since the data are qualitative.

The 'amount' of satisfaction represented by a move from 2 to 3 may not be the same as that between, say, 3 and 4 (if a 'satisfaction' amount could be quantitatively measured, which it cannot). Given that the data are ordinal at best, it would also be mathematically incorrect to calculate means or to perform addition, subtraction, multiplication or division on the data.

In addition, the responses to these Likert-type items are frequently skewed and bunched at one or other end. They may also be bimodal. Typically, they are not normally distributed.

Hence it would appear that a *t*-test or Welch test would be inappropriate on a number of counts. However, if one were to perform a non-parametric test, one would expect to lose power and run an increased risk of a type II error.

Alternatively, one may count the number of responses in each cell for each of the two comparison groups and then perform a χ^2 test. Unfortunately, if this approach is taken, one loses the information available from the ordinality of the responses; the χ^2 test will treat the cells as if they were purely nominal counts, with no ordinal information being taken into account. Once again, power will be lost.

De Winter and Dodou[4] have evaluated the use of the *t*-test and also the non-parametric Mann-Whitney-Wilcoxon rank sum test in the situation where one has data from a five-option Likert item such as the example shown in Table 2. They performed this evaluation through empirical study (simulations) rather than theoretical argument. They conclude that, as long as one has samples of at least 10 respondents in each group, either the *t*-test or the Mann-Whitney-Wilcoxon test may be used, in spite of the theoretical reservations that one might have regarding the use of the *t*-test in this situation. They showed that the two tests had similar power even if the sample sizes of the two comparison groups were markedly different. When the data frequency distributions were skewed or peaked (as is commonly the case), the Mann-Whitney-Wilcoxon test had greater power than the *t*-test.

We would, therefore, recommend that the Mann-Whitney-Wilcoxon test be used to analyse data from these Likert-style questionnaires with five ordinal selection categories. Not only is the use of this test theoretically easier to justify, but in most practical cases it would be expected to have greater power.

Stata commands (shown here between < and >; when typing the command omit < and >)

1. For H_0: data were drawn at random from a population with normal distribution.
Preferred if a sample size <20. Reject H_0 if $p \leq 0.05$.
(The Shapiro-Wilk test) (e.g. weight of females v. weight of males):
<swilk weight if sex==0>
<swilk weight if sex==1>
2. For H_0: data were drawn at random from a population with normal distribution.
Preferred if a sample size >20. Perform visual inspection.
<histogram weight if sex==0>
<histogram weight if sex==1>
3. For obtaining the variances to decide by inspection if they are equal or not:
If using Stata 14 or earlier:
<sum weight if sex==0> Then, in order to obtain the variance (= standard deviation (SD)^2):
<di sd^2> (obtain SD from the output of previous command and insert).
<sum weight if sex==1> Then, in order to to obtain the variance (=SD^2):
<di sd^2> (obtain SD from the output of previous command and insert).
4. For obtaining the variances to decide by inspection if they are equal or not:
If using Stata 15 or later (more convenient way to obtain the variances):
<ci variances weight if sex==0> This gives the variance directly.
<ci variances weight if sex==1> This gives the variance directly.
5. For a *t*-test: independent samples, data in wide format (less usual):
<ttest $\text{weight}_{\text{male}}$ = $\text{weight}_{\text{female}}$, unpaired> Must put 'unpaired'
6. For a *t*-test: independent samples, data in long format (more usual):
<ttest weight, by(sex)> No need to put "unpaired"
7. For a Welch test: independent samples, data in wide format (less usual):
Stata calls the Welch test a *t*-test with unequal variances.
<ttest $\text{weight}_{\text{male}}$ = $\text{weight}_{\text{female}}$, unpaired unequal> Must include 'unpaired'.
8. For a Welch test: independent samples, data in long format (more usual):
<ttest weight, by(sex) unequal> No need to include 'unpaired'.
9. For a paired *t*-test:
Consider a paired *t*-test for prediet and postdiet weights ('prewt' and 'postwt'):
<gen diff = postwt-prewt> This generates a new variable called 'diff' that contains all the individual differences between pre- and postdiet weights).
<swilk diff> This is the preferred way to assess if the differences are from a normally distributed population when sample size <20 pairs of data.
<histogram diff> For samples >20 pairs a frequency histogram is produced and may be inspected.
<ttest postwt = prewt> This will then result in the paired *t*-test being performed. No need to type 'paired' in as in this wide format the paired test is the default in Stata.
10. For a single sample *t*-test:
<ttest variable = GS> 'Variable' is the name of the single-sample variable. 'GS' is the gold standard/benchmark/target. Remember that these tests may often be single-tailed tests, especially in the context of quality control.
11. For a Mann-Whitney-Wilcoxon test:
<ranksum weight, by(sex)> No need to type in either 'unequal' or 'unpaired'. Data must be in the long format.
12. For a Wilcoxon signed-rank test:
<signrank $\text{weight}_{\text{male}}$ = $\text{weight}_{\text{female}}$ > Data must be in the wide format.

Stata version 15 outputs

Example 1: Output from a *t*-test (independent samples, population variances assumed equal)

The first example, shown in Fig. 2, shows the output from a *t*-test (independent samples, population variances assumed equal). Note that the two-tailed *p*-value is given by 'Pr(|T|) = 0.2400' (not statistically significant since $p>0.05$). The point difference between the two sample means is 1.221124, and the 95% CI for the difference between the means is given by –0.8379923 - 3.280239, which includes the null value of zero. This is expected since $p>0.05$ and the result is not statistically significant.

(The *p*-values given for Ha: diff <0 and Ha: diff >0 are for single-tailed tests and need not concern us here).

Example 2: Output from a Welch test

The difference between the variances of the two samples (see Fig. 3) is very bizarre in this contrived example (6.43^2 and 0.74^2 or 41.34 v. 0.55). The resulting *p*-value (0.1381) for the two-tailed test is much lower than the *p*-value that would have been obtained had a *t*-test been performed (0.2306).

In spite of the fact that the degrees of freedom for the Welch test are lower than those for the *t*-test (12 v. 19), this does not mean that the Welch test is necessarily less powerful. The way in which the variance of the difference between the two means is calculated for the Welch test means that sometimes this variance may be smaller for the Welch test than it would have been for the *t*-test. As a result, one

cannot generalise about a difference in power (*t*-tests v. Welch tests), but should rather use the more appropriate of the two tests.

Note that the two-tailed *p*-value is given by 'Pr(|T|) = 0.1381' (not statistically significant). The point difference between the two sample means is 2.855769, and the 95% CI for the difference between the means is given by –1.0546 - 6.766139, which includes the null value of zero. This inclusion of the null value is expected since $p>0.05$.

In addition, note the variances SD^2 are very different and the sample sizes are both different and small (13 males and 8 females). The (inappropriate in this case) *t*-test would have yielded a *p*-value of 0.2306.

(The *p*-values given for Ha: diff <0 and Ha: diff >0 are for single tail tests and need not concern us here).

Example 3: Output from a paired *t*-test

In the example presented in Fig. 4, the difference between two weights measured in gold miners, 1 year apart, was generated. There were 510 participants (hence two samples of 510 readings, and a single sample of 510 differences). The Shapiro-Wilk test yielded a *p*-value of <0.001, indicating that the data were not drawn at random from a normal distribution. However, the frequency histogram yielded a unimodal graph that was not skew, nor particularly kurtotic (peaked).

This shows the importance of using the histogram, rather than the very sensitive Shapiro-Wilk test, to decide whether or not to proceed with the parametric *t*-test. In any event, with such a large sample size, the argument is academic; The *t*-test would have been an appropriate test to use in any case.

Note that the two-tailed *p*-value is given by 'Pr(|T|) = 0.0441'. This result is, of course, statistically significant ($p<0.05$). The difference between the two sample means is 0.65, and the 95% CI for the difference between the means is 0.018888 - 1.281112 (which excludes the null value of zero). This is expected for a statistically significant result.

(The *p*-values given for Ha: mean (diff) <0 and Ha: mean (diff) >0 are for single-tailed tests).

Presenting and interpreting the results

As a default, one might consider presenting one's results correct to two decimal places, with *p*-values correct to three decimal places. Stata *p*-values of, say, $p=0.0000$ should rather be presented as $p<0.001$, since, theoretically, *p* cannot be zero. The minimum information that should be presented includes the name of the test performed and the point estimate for the difference between the means, along with the *p*-value and the 95% CI for the difference between the two means. In the case of multiple tests having been performed, if a Bonferroni adjusted *p*-value is presented then this should be stated. Alternatively, if the *p*-value is unadjusted then the Bonferroni-adjusted α value should be stated alongside the results.

```
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      31    25.06871    .7385578    4.112116    23.56037    26.57705
       1 |      29    23.84759    .712573     3.837323    22.38795    25.30723
---------+--------------------------------------------------------------------
combined |      60     24.4785    .5158309    3.995609    23.44632    25.51068
---------+--------------------------------------------------------------------
    diff |            1.221124    1.028675               -.8379923    3.280239
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.1871
Ho: diff = 0                                     degrees of freedom =       58

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8800         Pr(|T| > |t|) = 0.2400          Pr(T > t) = 0.1200
```

Fig. 2. Results of a t-test with independent samples (Stata output). Mean Body Mass Index for males (Group 0) v. females (Group 1).

```
Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      13    24.23077    1.783665    6.431094    20.3445    28.11704
       1 |       8      21.375    .2630521    .7440238    20.75298   21.99702
---------+--------------------------------------------------------------------
combined |      21    23.14286    1.134493    5.198901    20.77635   25.50937
---------+--------------------------------------------------------------------
    diff |            2.855769    1.802958               -1.0546    6.766139
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.5839
Ho: diff = 0                     Satterthwaite's degrees of freedom =  12.5175

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9309         Pr(|T| > |t|) = 0.1381          Pr(T > t) = 0.0691
```

Fig. 3. Results of a Welch test (Stata output). Mean age for males (Group 0) v. females (Group 1) if BMI >25.

```
Paired t test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hb2 |      20          15    .3769685    1.685854      14.211      15.789
     hb1 |      20       14.35    .3101358    1.386969    13.70088    14.99912
---------+--------------------------------------------------------------------
    diff |      20         .65    .3015312    1.348488    .018888     1.281112
------------------------------------------------------------------------------
    mean(diff) = mean(hb2 - hb1)                                  t =   2.1557
Ho: mean(diff) = 0                               degrees of freedom =       19

Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
Pr(T < t) = 0.9779         Pr(|T| > |t|) = 0.0441          Pr(T > t) = 0.0221
```

Fig. 4. Results of a paired t-test (Stata output). Haemoglobin levels in 20 athletes before and after taking a naturopathic product for 4 weeks (fictitious data).

1. Pagano M, Gauvreau K. Principles of Biostatistics, 2nd ed. Pacific Grove: Duxbury, 2000.
2. Moser BK, Stevens GR. Homogeneity of variance in the two-sample means test. Am Stat 1992;46(1):19-21.
3. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's *t*-test instead of Student's *t*-tests. Rev Int Psychol Soc 2017;30(1):92-101.
4. De Winter JCF, Dodou D. Five-Point Likert Items: *t* test versus Mann-Whitney-Wilcoxon. Practic Assess Res Eval 2012;15(11):1-16. http://pareonline.net/getvn.asp?v=15&n=11 (accessed 20 August 2017).

Accepted 31 August 2017.