An Introduction to Tests of Mean Difference

In these exercises shift focus onto statistical techniques that have been developed, primarily, for use in investigating mean group/categorical differences in responses on a continuous dependent variable. Specifically, we'll focus on a method that we can apply when we have **one continuous dependent variable** and **one categorical independent variable *which has ONLY two levels*** – known as a t-test.

Relevant chapters for these exercise in the Andy Field Textbook are:
Chapter 6: Beast of Bias

Chapter 10: Comparing two means

Comparing two means (using t-tests)

In the following exercises we're going to test very simple hypotheses about mean group differences. Namely, that differences in performance on a single continuous dependent variable (in terms of its' Mean and Variability) exist between only two levels (AKA groups/categories) within a single categorical independent variable. You may see this type of test referred to as **pairwise comparisons** (because it compares performance between a 'pair' of groups) or **t-tests** (because they use the t-statistic to estimate the probability of the data under the null hypothesis).

The type of t-test you use depends on whether your independent variable varies its levels within- or between-participants in your study. We'll look at the between-subject variety first:

Exercise 1: Independent Groups t-test

Independent Groups (also known as Between-subjects) t-tests are used in scenarios where the levels of our Independent variables are represented by two different mutually exclusive groups of participants. Such as in the example described below.

A researcher is interested in investigating whether there is a difference between short-term memory performances in the morning as opposed to the afternoon. They recruit a group who all perform a test of short-term memory (a digit-span task), the dependent variable. Using stratified random sampling, half of the participants come to the lab in the morning to do this test and half to come to the lab in the afternoon.

To test their hypothesis (i.e. that short-term memory performance will differ depending on time of day) they want to use an independent groups t-test to compare test scores between their AM and PM groups. To do this…

1. First open the **session3_data_Ex2.sav** file

Take a moment to note the structure of the data in this file:

- Each row represents a different participant
- The first column (*time*) identifies which level/group within the independent variable that a participant was randomised to (1 = AM & 2 = PM, coded with value labels).
- The second column (*span*) provides the short-term memory scores (the dependent variable)

Extra tip/pointer:

Note that the terminology we use to discuss/describe our study's design will sometimes vary from that which SPSS will use in its coding and output. In designs where our independent variable is categorical, SPSS will normally refer to such variables as **Grouping Variables** or **Factors** and the groups/categories within them as **Levels**. These levels will, as we've seen previously in our datasets, be assigned a numeric code (e.g. Level 1 = Male, Level 2 = female).

Whilst we can assign value labels to these codes in our dataset these will not always be 'pulled through' by SPSS into the output. SPSS sometimes will just give you the code number for a level within a variable in the output. So, at times you'll have to remember what these numbers represent when you see them in your. For example, that '1' denotes AM and '2' denotes PM.

To run the Independent groups t-test:

2. Select **Analyze > Compare Means > Independent-Samples T Test**
3. Select your dependent variable (i.e. *span*) and move it into the **Test Variable(s):** box and then select your independent variable (i.e. *time*) as the **Grouping Variable:**
4. Next, define for SPSS the order you want it to use when contrasting between the two levels within the independent variable, for values of the dependent variable. For example, do you want level 1 (i.e. AM) subtracted from level 2 (i.e. PM) or vice versa? To specify this, click on **Define Groups…** and in the **Use specified values** fields state the numerical codes within your independent variable that you want SPSS to use for '**Group 1'** and '**Group 2**'[1]. In this instance, put '*1*' in **Group 1** (i.e. the AM group's code number) and '*2*' in **Group 2** (i.e. the PM group's code number). This order means that: if the mean difference SPSS reports in the output is positive then the average response in the AM group is higher than the average response in the PM group. The opposite is true if the mean difference is negative. To summarise, thinking about the order you enter your levels in here **matters** for interpretation!
5. Then click **Continue** to close this pop-up window and **OK** to run the test.

Now look at the output created and follow the next steps to interpret it…

A. Check the **Levene's test for equality of variances** in the output. This tests whether we've satisfied the assumption of homogeneity of variance (necessary for an Independent Groups t-test). The **Sig.** value for the Levene's test must be p > .05. It is in this example (p = .388), so we can assume equal variances.[2]

B. The **Group Statistics** output shows a Mean (SD) of 6.26 (0.98) for span length in the morning and 5.84 (1.57) for the afternoon. The Mean Difference (and 95% Confidence Intervals around it) in the **Independent Samples Test table** is indicated as 0.42 (-0.30 to 1.14).

   Interpreting just this mean difference and its confidence interval, is memory likely better in the morning? That group mean is larger in the morning than the afternoon but remember these are estimates, with uncertainty around them. Thus, we don't know for sure whether any difference is statistically significant until we look at the t-test result…

[1] The difference is always calculated as Group 1 minus Group 2.

[2] If it wasn't we'd look at the **Equal variances not assumed** row when interpreting the t-test results instead. This applies a more conservative correction to this test of statistical significance.

Extra tip/pointer:

The fact the confidence interval here 'crosses zero' (i.e. it ranges from a mean difference estimate of -0.30 (in favour of PM > AM) to 1.14 (in favour of AM > PM) should ring alarm bells here for any potential prediction of statistical significance. This is because the confidence interval's range includes zero, the value we have rule out as a plausible mean difference in the wider population, if we want to reject the null hypothesis (i.e. $H_0 = 0$)

C. The t-test result, t(51) = 1.18, *p* = .243, indicates the difference between AM and PM performance is not statistically significant (i.e. *p* > .05). So, despite the means showing a difference (as mentioned in B) we *cannot* reject the null hypothesis that in the wider population there is a greater than 5% chance that the true difference in the population means might be zero. Or to put it another way, it is not improbable that there is no meaningful difference in span length between AM and PM in the wider population.

However, we didn't check our data for outliers, did we? Use **Explore** in the **Descriptives** options (as we did in previous see) to see if there are any outliers. *Hint:* You should find one extreme value for memory performance in the afternoon.

- What assumptions does the inclusion of this outlier violate?

The assumption that data in the dependent variable has a normal distribution has been violated. This is indicated in the significant tests of normality, histograms and boxplots

- Exclude the outlying value in the afternoon (i.e. Using **Data > Select Cases**, like we did in previous weeks) and re-run the Independent Groups t-test

- Have your results changed, if so how?

The mean has decreased for PM (SD decreased dramatically also). Consequently, the mean difference between AM and PM span length is now statistically significant (i.e. p < .05):

Mean Difference (95% CI) = 0.66 (0.13, 1.21), t (50) = 2.48, *p* = .017

- Consequently, would this change your interpretation?

After excluding the outlier there is stronger evidence in favour of digit span performance in the morning being better than in the afternoon. However, note that the actual size of this difference is still very small (i.e. the parameter estimate for the mean difference is somewhere between 0.13 and 1.21 more items on average being recalled in the AM. Therefore, whilst this result is statistically significant the effect size here questions how practically meaningful a finding this is in terms of wider applications in the real world?

- Meanwhile, has the external validity of our study been improved or undermined by excluding this outlier?

Because we've just excluded an extreme **but plausible** outlier we may have undermined our external validity. If we had more information on this outlier we might want to consider if there is any explanation for this highly anomalous result. However, given we don't, and it is conceivable someone may have a span length of 12, we have weak grounds to exclude it automatically. At best I'd report the results with and without the outlier included. That, unfortunately, is going to create a lot of ambiguity (BUT transparency) in our conclusions!

In fact, there is a better way of handling this outlier though, which doesn't require excluding. This involves using a more "robust" methods to estimate the mean and SD of the group differences:

Using Bootstrapping in an Independent Samples t-test

First, 'switch-off' the exclusion criteria you just applied (i.e. go to **Data > Select Cases** and select **All Cases**), so the outlier is once again included in your sample.

Instead of excluding this case, we'll use a technique call **Bootstrapping**, which can be applied to most statistical tests in SPSS to make them more robust to violations of assumptions (such as violations of normality).[3] This technique essentially re-runs your analysis multiple times (usually at least a 1000) on random resamples of your sample (with replacement) to work out the 'average' relationship observed across the majority of these resamples. Using bootstrapping will minimize the influence of more extreme values and also allows us to make better estimates of the actual sampling distribution in the wider population. To apply bootstrapping in your t-test:

1. Select **Analyze > Compare Means > Independent-Samples T Test**
2. Your original analysis should still be 'loaded' in SPSS's memory so no need to re-input it.
3. Click on the **Bootstrap…** option on the right of the window.
4. In the window that pops up:
   - tick the option to **Perform Bootstrapping**
   - Set the **Number of Samples** at 2000
   - switch to the **Bias Corrected accelerated (BCa)** method of bootstrapping which is apparently more accurate (see reference to this in the **Andy Field textbook, in section 6.12.3**)
5. Press **Continue** and then press **OK** to run the t-test

What additional information do you find has been added to the **Group Statistics** table in your output, and what do you think this is telling you?

You've now got a confidence interval around both the mean and SD estimates. This is a range of values that the bootstrapped estimates for these parameters to fell between in 95% of the 1000 resamples you just re-ran your test on.

You now also have an output table that reports the **mean** difference between AM and PM for memory performance, calculated from the bootstrapped samples, with a *p*-value and BCa confidence interval reported for the bootstrapped mean difference too. Given your three sets of results, which would you favour writing up?

[3] Applying bootstrapping alone doesn't constitute a 'Robust' equivalent of the standard t-test. See **Andy Field, Chapter 6, section 6.12.3** for a fuller description of "Robust Methods" and their use in statistical analysis to overcome violated assumptions.

Note: there's not necessarily a "right" answer here. I'm more interested in how you justify your choice.

The mean difference after boostrapping, which will vary slightly every time you run it because it is based on **averaging** across X number of *random* resamples is still not statistically significant. In my example I got a Mean difference of 0.42 (BCa 95% CI of -0.30 to 1.14).

You might argue that excluding the outlier is therefore still necessary because it is obscuring an otherwise significant results. ***However,*** bear in mind that limits your generalisability and leads to a significant result with only a very small effect-size. Hence, I would probably argue that you've tried to take account for the violated assumption of non-normality by using bootstrapping and, because the result using this method is not statistically significant, you have insufficient evidence to reject the null-hypothesis. After all, we don't want to be accused of p-hacking!

The case for excluding the outlier would be a lot stronger if you knew there was some good reason for the outlier arising, which was suggesting of the participant violating the experimental protocol (e.g. the participant had woken up later and hadn't already been up all morning).

Extra tip/pointer:

Depending on the version of SPSS you are using, once you switch bootstrapping 'on' in SPSS it may presume it should apply this to all subsequent analysis where bootstrapping is possible to apply – **even if you're not doing the same test.** The resampling process used in bootstrapping is much more computationally intensive than running a standard version of any given type of analysis. Therefore, it can really slow SPSS down on more advanced analyses than your basic t-test, particularly on old/slow computers. Therefore I would leave this option switched 'off' until there's a call for it and check it is switched 'off' again after use.

Calculating Standardised Effect sizes

To calculate a standardised effect size for an independent groups pairwise comparison we can use the formula for Cohen's *d*. A worked example of how to calculate it for this analysis, with the outlier retained, is illustrated below:

$$d = \frac{M_{AM} - M_{PM}}{\sqrt{\frac{(N_{AM}-1)SD_{AM}^2 + \ (N_{PM}-1)SD_{PM}^2}{(N_{AM}+N_{PM}-2)}}}$$

d = 6.2668 – 5.8433 / √(((27 – 1)0.97781^2 + (26 – 1)1.57383^2))/(27 + 26 – 2))
d = 0.4235/√(((26)0.97781^2 + (25)1.57383^2))/51)
d = 0.4235/√((24.85 + 61.92))/51)
d = 0.4235/√1.70
d = 0.32

Notes: AM refers to values on the 'AM' row of the Descriptives table, PM refers instead to the 'PM' row. M refers to the Mean values and SD to the Std. Deviation on these rows and N refers to the number of participants.

The format for reporting this result would be:

t(51) = 1.18, p = .243, *d* = 0.32

Can you replicate this calculation and formatting for this test result when the outlier is excluded?

$d = 6.2668 - 5.5970 / \surd(((27 - 1)0.97781^2 + (25 - 1)0.96830^2))/(27 + 25 - 2))$
$d = 0.6698/\surd(((26)0.97781^2 + (24)0.96830^2))/50)$
$d = 0.6698/\surd((24.85 + 22.50))/50)$
$d = 0.6698/\surd 0.947$
$d = 0.72$

t(50) = 2.48, p = .017, *d* = 0.72

Exercise 2: Paired Samples t-test

Paired samples (also known as within-subjects) t-tests are used in scenarios where the levels (AKA groups/categories) within your independent variable are represented by two different measurements that are related in some way. Usually they are measurements of the dependent variable in the same participant under different conditions. However, the use of this test is in fact appropriate whenever you can argue strongly for a degree of relatedness (and thus a lack of independence) between levels in your IV. For example, repeated measures of the same group's work over a series of weeks or measures of different types of highly related outcomes in the same individual (e.g. standardised scores on a test of *sustained attention* and a test of *divided attention* are related in the sense that they are both types of *attention*). Here's a specific example:

A researcher is interested in investigating whether there are systematic gender biases in most standardised IQ tests - because most IQ tests were developed a long time ago primarily by men, to assess abilities they were biased to value most highly!

To test the theory of an in-built male bias they recruit a sample of dizygotic twins (i.e. each set of twins contains 1 male and 1 female sibling). They measure both siblings on a standardised IQ test battery. They hypothesise that because only gender is a key difference in these otherwise very similar individuals (i.e. they share 50% of their genome and 100% of their home environment) evidence of differences in IQ test performance would support their claims for gender biases in standardised IQ tests.

To test their hypothesis they want to use a paired-samples t-test to compare IQ scores between the siblings within a pair. To do this…

To analyse this data, open the **session3_data_Ex3.sav** file

Take a moment to note the structure of the data in this file:

- Each row represents a different 'participant'. However, in this example the term 'participant' is rather misleading. More correctly, each row represents one *Independent unit of observation* within our sample. In our example that 'unit' is a set of twins (i.e. two people not one). Other examples of research where a unit of observation might not simply equate to one participant include:
    - Comparing between school's performance on academic league tables
    - Comparing between hospital's performance on patient waiting-times
    - Comparing between families from different ethnicities on their level of social cohesion.

Now, let's take a moment to understand the structure of the data in this dataset:

- The first column (*girls*) provides the IQ test battery scores for the female sibling
- The second column (*boys*) provides IQ test battery scores for the male sibling
- Note that a within-subject dataset has a **very different** structure from the one we had when we were handling between-subject data. Now, instead of 1 column representing the independent variable and 1 column representing the dependent variable, we have two columns that both indicate performance on the Dependent Variable (IQ), under specific levels of the Independent Variable. In other words, there is no single column categorising *Gender* here (even though this is our independent variable in this study). Instead we have columns for each level within Gender (i.e. 1 column for *girl's* performance and 1 column for *boy's* performance). This formatting is necessary because in this design we are treating *Gender* as a repeated measure (i.e. we have multiple observation, one per gender category) 'within' each unit of observation (i.e. pair of twins).
- This creates a major differences in how SPSS handles analysing data from these two types of study design:
    - For between-subjects designs (i.e. **previous** exercise) SPSS needs you to tell it:
        1. Which column represents the dependent variable
        2. Which column represents the independent variable(s)?
    - For within-subjects designs (i.e. **this** exercise) SPSS needs you to tell it:
        1. The name of the Independent Variable(s) and how many levels it has?
        2. **Then**, which columns corresponds with measuring the dependent variables under a given level of your Independent Variable[4]

To run the paired samples t-test:

1. Select **Analyze > Compare Means > Paired-Samples T Test**
2. Enter the column representing the first level of your IV (i.e. *Girls IQ*) into the **Variable 1** box.
3. Enter the column representing the second level of your IV (i.e. *Boys IQ*) into the **Variable 2** box.
4. Then click **Continue** to close this pop-up window and **OK** to run the test.

- In the **Paired Samples Statistics** table we see the means are both above 100 (boys 100.7 girls 101.6), with girls approximately 1 point up on their brothers.

- The result that gives you a test of your hypothesis are reported in the **Paired Samples Test** table. Here we can see the difference in means between genders for twins is not statistically significant (t (26) = 1.310, *p* = .202). Therefore, we are not able to reject the null hypothesis that the mean difference in IQ within mixed sex twins due to gender is zero in the wider population.

…Or can we? What else should you check?

(*Hint*: have you **Explored** the data and are you aware that anyone with an IQ of < 70 would normally be clinically diagnosed as having an Intellectual Disability?).

[4] That means that in this within-subjects design we never actually specifically tell SPSS what the dependent variable is! So be sure you remember (i.e. have a metadata file, as discussed in Practical 1)

Run these checks and any further re-analysis this prompts you to do. What conclusions would you now draw regarding your hypothesis that there are gender differences (biases) in IQ tests?

If you explore the data you notice two girls with an IQ < 70, with their siblings close to this cut-off too. This cut-off point is usually taken as an indicator of severe learning disability, in clinical practice. Therefore, you'd have strong grounds to argue that these sibling pairs represented a very different population from the majority of the sample. This is an example of a case where there is a known theoretical argument for removing these cases (i.e. these twinset's responses are atypical due to an underlying medical diagnosis that confounds interpretation).

Their exclusion will limit the generalisability of your study slightly (i.e. your results won't be generalisable to individuals with an intellectual disability) but, providing you acknowledge this limitation in your discussion, it would seem you have a plausible reason here to exclude these cases from your analysis.

If you do so, there is now a statistically significant difference apparent in the IQ scores by gender. The mean difference and 95% confidence interval are 1.70 [0.46 to 2.94], indicating the t-test results is now statistically significant: t(24) = 2.834, p = .009.

Interestingly, whilst the difference between these groups is statistically significant, the results still **do not support** the original hypothesis because it shows Girls having a significantly higher IQ than Boys. In other words, a gender advantage in IQ test performance that is actually running in the **opposite** direction to what was originally predicted!

This is why you must always interpret the size **and direction** of a significant result. Just 'hunting' for significant result (i.e. p < .05) can sometimes blind people to the fact they also need to check if their alternative hypothesis remains a good explanation for their data, having managed to reject the null hypothesis.

Calculating Effect sizes

Again, Cohen's *d* can be reported as a standardised estimate of the effect size for paired samples t-test. The formula is slightly different this time and I've illustrated how to calculate it for the statistically non-significant result. Can you recalculate for the significant result (i.e. the instance with outliers excluded)?

$$d = \frac{M_1 - M_2}{SD_{pooled}} \qquad SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

d = 101.698 – 100.7191 / √((11.23905^2 + 8.07353^2)/2)
d = 0.9789 / √(191.498/2)
d = 0.9789 / √(95.749)
d = 0.9789 / 9.785
d = 0.10

Notes: The '1' in M_1 refers to values in the row for Level 1 in the Descriptives table (i.e. Girls IQ), and the '2' in M_2 refers instead to values in the row for level 2 (i.e. Boys IQ). M stand for Mean and SD to the Std. Deviation. The SD_{Pooled} is not something **this** output easily gives us, but you'll see above we can use the SDs for each condition instead to calculate this value (i.e. the expanded equation on the second line of the example).

$d = 104.3290 - 102.6284 / \surd((6.28319^2 + 4.39385^2)/2)$

$d = 1.7006 / \surd(58.784/2)$

$d = 1.7006 / \surd 29.392$

$d = 1.7006 / 5.4214$

$d = 0.31$