Objectives

We will go through the various types of tests for asking the question 'is the mean of this sample significantly different from… (something)'. We will then look at the ***t* test,** a very popular 'parametric' test. This has various forms, depending on the kind of data you want to analyse. We will look at nonparametric tests in Practical 4.

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

3.1 Background

Reminders

We've already discussed the differences between **one- and two-tailed tests** (Handout 1, *One-tailed and two-tailed tests*).

We've already talked about making **multiple comparisons** between groups (Handout 1, *The danger of running multiple significance tests*).

Paired and unpaired tests (related and unrelated data)

When we come to look at the difference between two samples of data, the samples can be *related* or *unrelated*. Suppose we want to compare the speed with which people can rotate figures mentally in two conditions: on land and underwater. (1) We could take a group of landlubbers and a group of divers, and compare them. There would be no particular relationship between individual data points from the land sample and the underwater sample. We would use statistical methods that are described as *unrelated*, *unpaired*, or *between-subjects*. (2) Alternatively, we could measure the *same* group of people in two conditions, on land and underwater. In this situation, there is a relationship between one subject's score on land and the same subject's score underwater — they are likely to be more similar than they would be by chance alone, because they come from the same person. Our statistical methods must reflect this fact; the techniques we would use are described as *related*, *paired*, or *within-subjects*.

It is absolutely **not** acceptable to fail to take account of relationships between data like this. A classic example of this sort of error is something called **pseudoreplication.** Suppose you test Alice, Bob, and Celia on land, and Eric, Frankie, and Greg underwater. You obtain 6 observations, $n = 3$ for each group. Your groups are not related. So far, so good. But suppose you want more than 6 observations; you might measure each subject three times. This would give you observations A_1, A_2, A_3, B_1… on land, and E_1, E_2, E_3, D_1… underwater. The error is to analyse this as if you had 18 observations ($n = 9$ for each group). This is wrong, because A_1, A_2 and A_3 are all *related* — more so than A_1 and B_1, or A_1 and E_1. We will not cover the analytical techniques required for this sort of situation, where we have multiple variables (in this case, land / underwater as a between-subjects variable, observation 1 / observation 2 / observation 3 as a within-subjects variable) — that's covered in the Part II course. If you have data like this, the simplest thing is to obtain some sort of 'overall' score for each subject (e.g. take Alice's overall score to be the mean of A_1, A_2, and A_3) and analyse those.

If you have data from *only* one subject, then you can consider the data to be 'unrelated' for the purposes of analysis, but your conclusions *only apply to that subject*. For example, if you measured my ability to remember sequences of digits (my digit span) ten times when I'm on dry land and ten times when I'm underwater, you could treat the data as unrelated — they have no relationship to each other *beyond* the fact that they come from the same subject, and that's part of your analytical 'context'

anyway. You would have a sample (n = 10) of my dry-land digit span, and a sample (n = 10) of my underwater digit span. If the dry-land scores were significantly higher than the underwater scores, you could conclude that *my* digit span was better on dry land than underwater — but this would tell you absolutely nothing about people in general, because I might not be a representative person. You would only know this by testing more people. (If you're wondering, the rather foolish situation in which you would need to deal with further 'relatedness' when you're only testing one subject might be something like this: you test me in a car on dry land, in a car underwater, drunk on dry land, drunk underwater, tired on dry land, tired underwater… then to ask the 'dry land versus underwater' question, you would treat the 'car' pair of observations as related, the 'drunk' pair as related, and so on.)

Parametric and non-parametric tests

In the tests we'll cover here, we analyse differences involving one or two samples by making *assumptions* about the populations they come from. Remember the jargon (Handout 1): we estimate *parameters* of populations by using *statistics* of samples. The tests we'll cover in this handout make assumptions about the parameters of the populations — for example, assuming that the underlying population is normally distributed. They are therefore called **parametric** tests.

If the assumptions of a parametric test are *not* justified — if our data are a bit odd — then we have two alternatives. (1) We can **transform** the data to make them fit the assumptions better. We won't cover this approach in the IB course, but it's important for 'real-life' data analysis. (2) We can use a test that does not make these assumptions about the distribution of the population — a **nonparametric** or **distribution-free** test.

If a test's assumptions are met, it should give an accurate value of p. We say that a test is **robust** if it gives a *good* estimate of p even if we violate its assumptions. (We may also say that it's **liberal** if it underestimates p when certain assumptions are violated — that is, says things are 'significant' more often than it should — or **conservative** if it does the opposite.)

In general, parametric tests have *more power*. If the assumptions of a parametric test are met, it's therefore better to use the parametric test. Many parametric tests are also quite *robust*, so people don't get too worried if the assumptions are not quite met, but not grossly violated. Parametric tests can also be used for *complex analyses* that can be quite hard to do with non-parametric tests. Transformations are a way of 'rescuing' the parametric test by making the data fit the test's assumptions better; this is why transformations are widely used. Non-parametric tests are sometimes viewed as a bit of a last resort, because they have lower power. (On the other hand, if you find a significant effect with a low-power test, you have no problem, and some statisticians argue that non-parametric tests are a generally Good Thing, though it's probably fair to say that most researchers prefer parametric tests.) Occasionally, if the data are 'odd', nonparametric tests have *more* power.

We'll cover some non-parametric tests in Handout 4.

3.2 The one-sample *t* test

Overview

Suppose we have **one group** of n men and want to know if they are unusually tall. We can measure their height, and ask the question 'does the mean of this sample differ significantly from μ metres?', where μ is the average height of our reference population (all the men in the UK, perhaps). To do this, we define the null hypothesis that the sample comes from a population with mean height μ metres. We calculate the sample mean $\bar{x}$ and the sample standard deviation s_X. From this, we can calculate the **standard error of the mean,** $s_{\bar{x}} = s_X / \sqrt{n}$. Then, we can calculate a ***t* statistic:**

$$t_{n-1} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

This is called a *t* statistic with ***n* – 1 degrees of freedom (*df*).** We look up our *t* statistic in our tables to find a **critical value** of *t* for this many *df* and our desired level of α. (If we want a two-tailed test with a level of α, we have to allocate $\alpha/2$ to each tail.) If our value of *t* is bigger than this critical value, we reject the null hypothesis.

Significant values of *t* can be big positive numbers or big negative numbers. Non-significant values of *t* are close to zero.

The *t* test is always obtained by taking a **number,** subtracting from it a **test value,** and dividing the result by the **standard error of the number.** We'll see several different forms of the *t* test for different types of data (one sample, two samples, etc.), but they all have the same general format.

Some people use the subscript on the *t* to refer to the number of degrees of freedom (e.g. '$t_6 = 2.5$, two-tailed $p < 0.05$'); others use it to denote critical values ('for $df = 6$ and two-tailed $\alpha = 0.05$, $t_{\alpha/2} = t_{0.025} = 2.447$; our $t = 2.5$, so $p < 0.05$'). I prefer the first of these, as you can probably tell.

How did we arrive at this? You don't need to know, but if you're interested, see section 3.11 (*Deriving the one-sample t test*).

What is the standard error of the mean (SEM)?

Suppose we have a population with mean μ and variance σ^2, and we repeatedly take very many samples from it, with each sample containing *n* observations. We can say some things about the samples that we take. For each sample, we can calculate a sample mean $\bar{x}$. So we can collect lots of different sample means — many values of $\bar{x}$. Now we can ask what might at first appear to be an odd question: what will be the *distribution* of these sample means? The mean of all the sample means (the mean of all the values of $\bar{x}$), written $\mu_{\bar{x}}$, will be the same as the population mean, μ. The standard deviation of all these sample means (the standard deviation of all the values of $\bar{x}$), written $\sigma_{\bar{x}}$, is usually called the **standard error of the mean (SEM).** It's a measure of how much the value of the sample mean $\bar{x}$ may vary from sample to sample taken from the same population. It can be used to compare the observed mean to a hypothesized value — as we saw above, it's the basis of the *t* test. If we know the population standard deviation σ and the sample size *n*, we can calculate the SEM like this:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If we don't know σ, we can estimate the population SEM using the sample standard deviation *s*:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

3.3 The two-sample, paired *t* test

It's very easy to extend the one-sample *t* test to **two related groups.** Suppose you measure the heights of *n* girls when they're 10, and then again when they're 11, so you have two measurements for each girl. These two measurements are clearly related (more so than two measurements for two different girls). We want to know if our girls are growing normally. For each girl, we can therefore calculate the **difference** or **difference score** between the two related measurements — we just subtract one from the other. We will obtain *n* difference scores (the amount that each girl has grown). Suppose we know that the average girl grows 5 cm between the ages of 10 and 11 ($\mu = 5$). We can just run a *t* test on the difference scores, exactly as before:

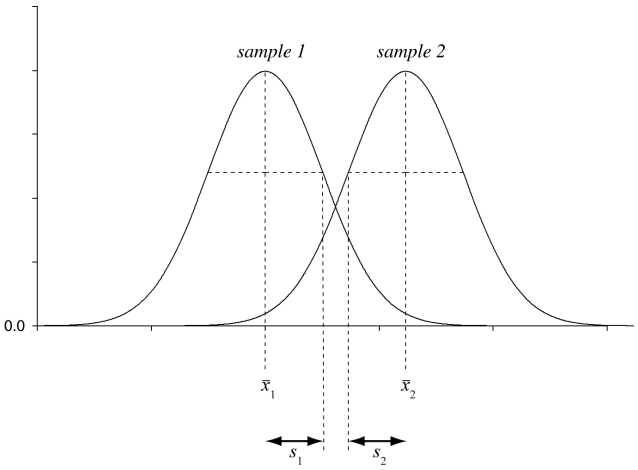
$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

If our value of t exceeds the relevant critical value for $n - 1$ *df* and an appropriate α, we reject the null hypothesis that our girls come from a normal-growing population.

The paired t test is used for **related (or matched) samples.** Two samples are related whenever you can use one sample to make better-than-chance predictions of the other. In this example, knowing one girl's height aged 10 allows you to make a better-than-chance prediction of the same girl's height aged 11, but doesn't allow you to predict another 11-year-old girl's height. In this example, the two samples come from the same subject, but sometimes related samples don't come from the same subject. For example, if you ask different couples to rate their satisfaction about their relationships, it is likely that if the man is very dissatisfied with the relationship, the woman is too, so their scores would be related (but would not be related to scores from a different couple).

Here's an example: suppose the initial heights of the girls in cm are {125, 148, 132, 135, 139, 129} and after a year they are {129, 153, 135, 140, 148, 136}. The difference scores (age 11 minus age 10) are {4, 5, 3, 5, 9, 7}. The mean of this sample of difference scores is $\bar{x} = 5.5$; the sample SD is $s_X = 2.17$; $n = 5$. We want to know if our group differs from a population with mean $\mu = 5$. We can calculate that $t = (5.5 - 5) / (2.17 / \sqrt{5}) = 0.51$. This t statistic has $n - 1 = 4$ *df*. For a two-tailed $\alpha = 0.05$, the critical value of t is 2.776. Our t is less than this, so we do not reject the null hypothesis; the girls are growing normally.

3.4 The two-sample, unpaired *t* test, for equal sample variances



The essence of a two-sample t test. We have two samples with means $\bar{x}_1$ and $\bar{x}_2$. If the distance (difference) between means ($\bar{x}_2 - \bar{x}_1$) is big enough, we say that the two samples are significantly different (which is to say, the two samples come from underlying populations whose means are different). We measure the distance between the means — somehow — in terms of the standard deviations of the samples, s_1 and s_2.

Overview

If we have **two independent (unrelated) groups,** X_1 and X_2, with **equal variances** ($s_1^2 = s_2^2$), we can ask if they are significantly different from each other. The null hypothesis is that the two underlying populations have the same mean ($\mu_1 = \mu_2$). We can calculate a t statistic, which has the same general form as before: it's the **difference** between means divided by the **standard error** of that difference, and this time it has $(n_1 + n_2 - 2)$ degrees of freedom.

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where s_p^2 (called the *pooled variance*) is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If $n_1 = n_2$, then the formula is a bit simpler:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This test assumes that the two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$), whether or not $n_1 = n_2$. If this assumption is violated, we must use the *unequal variances* version of this test (see below).

Example

Suppose we collect young horses and assign them to one of two groups at random. We feed one group (n = 10) FastDope, a drug that we suspect of having performance-enhancing properties. The other group (n = 10) are given a placebo. They are then timed running along a 1 km racetrack and their speed is calculated in $m \cdot s^{-1}$. The null hypothesis is that the speeds of the drugged and undrugged groups do not differ. We find that the speeds of the drugged group (group 1) are {12.2, 13.3, 12.6, 12.0, 11.6, 13.7, 13.6, 14.9, 13.0, 13.2} and the speeds of the placebo group (group 2) are {12.1, 10.1, 12.3, 9.1, 9.7, 10.1, 8.6, 9.2, 13.4, 13.9}. Since $n_1 = n_2$, we can use the simpler of the two formulae for t, and can therefore calculate

$$t_{n_1+n_2-2} = t_{19} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{13.01 - 10.85}{\sqrt{\frac{(0.96)^2}{10} + \frac{(1.91)^2}{10}}} = 3.20$$

For 19 *df*, the critical value of t for a two-tailed $\alpha = 0.05$ is 2.093. Since our t statistic exceeds this critical value, we reject the null hypothesis; the drugged group ran faster.

How did we derive this t test? If you're interested, see section 3.12…

3.5 The two-sample, unpaired *t* test, for unequal sample variances

If the two sample variances are not equal **(heterogeneous variances),** we have a bit of a problem. First, the number we calculate will not have a t distribution, so if we look it up using t tables we'll get the wrong answer. Second, it makes no sense to use s_p^2 in our formula (to 'pool' the variances of the two groups) since that procedure also assumes equal variances (as explained in section 3.12 if you're really interested). But we can still run a t test, although we'll lose a bit of power. We use our simpler formula and call the result t':

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We then test it just as if it were a t score, but with a **different number of degrees of freedom.** If we're doing it by hand,

degrees of freedom = (n_1 – 1) or (n_2 – 1), whichever is smaller.

If you have a computer, you can get a slightly better answer, which will lie somewhere between the hand-calculated version above and the original, uncorrected formula (using $df = n_1 + n_2 - 2$). It's called the Welch–Satterthwaite approximation (see Howell, 1997, p. 197), and it'll give us slightly more power. But you'll be doing it by hand in the exam and the W–S technique is too laborious to do by hand.

3.6 So are the variances equal or not?

If you want to know whether to use the *equal variances* or *unequal variances* version of the two-sample unpaired t test, you obviously need to know whether your population variances are equal or not, and the only way you can usually find that out is to test whether your sample variances are equal or not. Actually, what we do is to ask if our sample variances are *significantly different* from each other; if they are, we use the 'unequal variances' t test; if they're not, we use the 'equal variances' t test.

There are several methods available for testing differences between variances. Firstly, **look at the data;** it may be obvious. A good formal statistical test is Levene's test, provided by all good statistical packages, but it's a bit too much work to calculate by hand. Even the pen-and-paper version suggested by Howell (1997, p. 198) would take a lot of time in the exam. So we'll use the ***F* test.** This may not be the 'best' test (it has problems if the data are not quite normally distributed, though if they are, it's the most powerful at detecting differences in variances) but it's quick and good enough for our purposes — to decide whether the variances are too different for the 'equal variances' version of the *t* test.

The F test

The *F* statistic is a ratio of two variances. If the two variances are equal, $F = 1$. If they're not, $F \neq 1$. How much more / less than 1 does it need to be before we declare the difference 'significant'? We find that from tables of critical values of *F*. The *F* distribution is based on *two* numbers for the degrees of freedom: one for the numerator, and one for the denominator. We might write this as $F_{a,b}$ where *a* is the number of *df* for the numerator and *b* is the number of *df* for the denominator:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2}$$

In practice, tables of *F* don't give critical values for $F < 1$; they only give critical values for $F > 1$ (if you had $F < 1$, you could always take the reciprocal, 1 / F, and test that). So to make sure that our $F > 1$, we always put the **biggest variance on the top** (numerator) of the ratio, and the smallest variance on the bottom (denominator). So if the variances are different, the *F* statistic will be bigger than 1. In other words:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} \text{ if } s_1^2 > s_2^2$$

$$F_{n_2-1,n_1-1} = \frac{s_2^2}{s_1^2} \text{ if } s_2^2 > s_1^2$$

So you can run an *F* test on your data before choosing a *t* test; if it's significant (especially if $n_1 \neq n_2$), use the unequal variances *t* test; if it isn't, use the equal variances *t* test.

One more thing, though — if you want to test whether the variances are *different* with $\alpha = 0.05$ (two-tailed), you must run the *F* test itself with $\alpha = 0.025$. If you run the test with tabled values for $\alpha = 0.05$ (one-tailed), your actual two-tailed α will be 0.1. Why? Well, asking whether the variances are different without specifying the direction of the difference is a two-tailed test. The critical values of *F*, however, are for a one-tailed test (because we only test significance when $F > 1$, rather than $F < 1$). You've forced it to become a two-tailed test by calculating *F* in such a way that that $F > 1$; you must therefore double the stated one-tailed α to get the two-tailed α.

Relationship between the F test and the t test

The *t* test is actually a special case of the *F* test:

$$F_{1,k} = t_k^2 \text{ and } t_k = \sqrt{F_{1,k}}$$

where *k* is the number of degrees of freedom. In other words, a *t* test on *k df* is directly equivalent to an *F* test on 1 and *k df*. The difference that the *t* distribution is symmetrical about zero, since it deals with the differences between things, so values of *t* can be positive or negative. The *F* test deals with squared values, which are always positive, so *F* ratios are always positive (see Keppel, 1991, p. 121).

3.7 Assumptions of the *t* test

For any *t* test:

- You're testing hypotheses about the mean, which only makes sense if the mean is meaningful (it may not be if the measurement scale you used wasn't an inter-

val or ratio scale — see Handout 1).

- The maths behind the *t* test assumes that the underlying populations of the scores (or difference scores, for the paired *t* test) are **normally distributed.** If this assumption is violated, you can't use *any* form of *t* test. (Rule of thumb: if $n > 15$ and the data don't look too weird, it's probably OK to use a *t* test; if $n > 30$, it should be fine.)

For two independent samples, to use the equal-variance *t* test, we assume

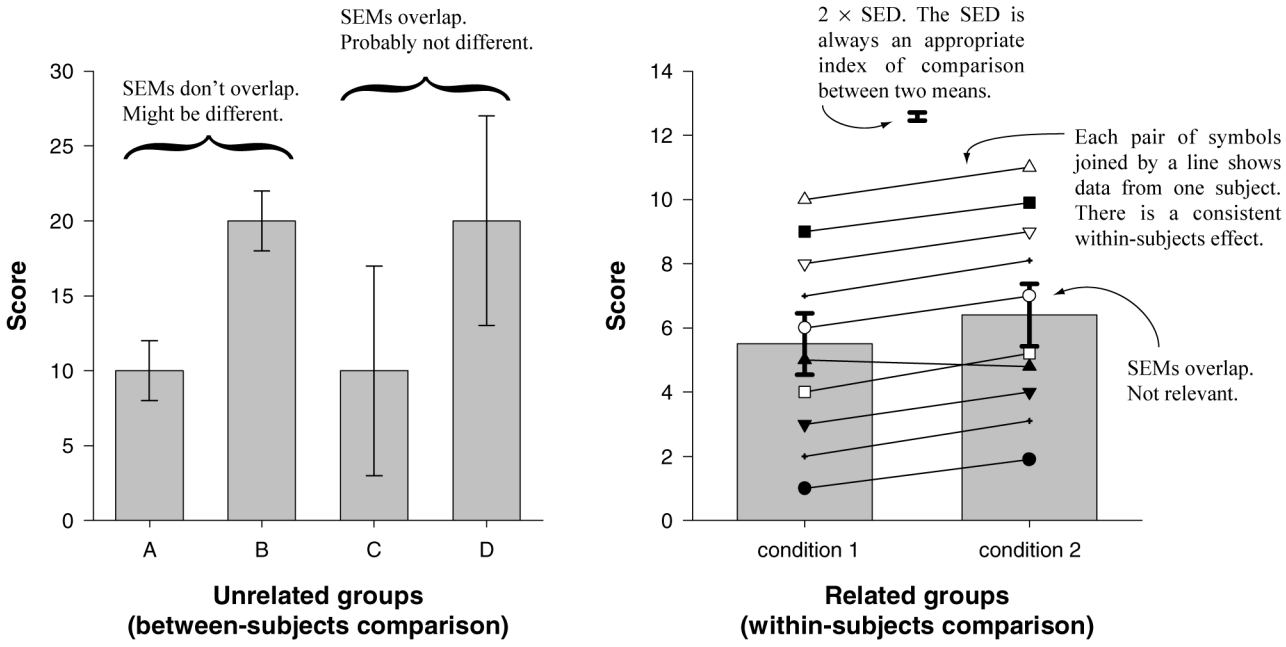
- The two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$), whether or not $n_1 = n_2$.

The *t* test is fairly robust to violations of this assumption if $n_1 = n_2$, but not if $n_1 \neq n_2$.

3.8 Graphical representation of between- and within-subject changes

'Error bars' (or 'mean ± variation') — the SEM is commonly used

The SEM is frequently used when people publish data. They may quote a measurement of '25.4 ± 1.2 g', or display a datum on a graph with a value of 25.4 units and error bars that are each 1.2 units long. These 'variation' indices could be one of several things — mean ± SD, mean ± 95% CI, mean ± SEM… The paper should state somewhere which one is being used, but usually it's the SEM. Why? First, it's smaller than the SD, so it conveys an impression of improved precision (remember that **accuracy** is how close a measurement is to a 'true' value and **precision** is how well it is defined; thus, 2.5000000003×10^8 m·s^{-1} is a more precise but far less accu-



Unrelated groups (between-subjects comparison)

Height of bar = mean
Error bar = ±1 SEM (1 SEM above, 1 SEM below the mean).

If the *n*s and SEMs of two groups are the same, then *t* = (difference between means) divided by (√2 × SEM). And if the SEMs of the two groups are the same and the SEMs overlap, then the means differ by <2 × SEM, so $t < 2 / \sqrt{2} = 1.4$. And $t < 1.4$ is never significant even at the 0.1 level.

So for independent groups, if the SEM error bars overlap, there's probably not a significant difference.

Related groups (within-subjects comparison)

For within-subject comparisons, the SEM of each condition is not helpful. The vertical bars show group means; their error bars show ±1 SEM. You would think that the groups don't differ. But in fact, the same subjects were tested in condition 1 and condition 2. The subjects all scored very differently, but there is a consistent improvement from condition 1 to condition 2. If we ran a paired-sample *t* test on the difference scores, we would find a highly significant difference between the two conditions. The appropriate index of variation to compare the two conditions is the standard error of the difference between means (SED), shown at the top.

Another way of plotting these data would just be to plot the difference scores, with their SEM; readers could then visually compare that mean to zero. However, that would not show the baseline scores.

The SED is always an appropriate index of comparison; a *t* test is calculated as (difference between means) divided by (appropriate SED). But different comparisons require different SEDs. If your error bars don't convey the right impression, consider using SEDs (as in the top-right example; you could say "the error bar is 2 × the standard error of the difference for the comparison between …").

Graphical presentation and interpretation of between- and within-subject changes.

rate measurement of the speed of light than 3.0×10^8 m·s^{-1}). In fact, using the SEM is perfectly fair and correct: the precision of an estimator is generally measured by the standard error of its sampling distribution (Winer, 1971, p. 7). Secondly — more importantly — if the SEM error bars of two groups overlap, it's very unlikely that the two groups are significantly different. (This is explained somewhat in the figure.) The opposite isn't necessarily true, though — **just because two sets of error bars don't overlap doesn't mean they are significantly different** (they have to 'not overlap' by a certain amount, and that depends on the sample size, and so on).

Within-subjects comparisons and the SED

For **within-subjects** comparisons, SEMs calculated for each condition are highly misleading (see figure). For this comparison — indeed, for any comparison — the SED is an appropriate index of comparison, because that's what the *t* test is based on (*t* = difference between means / SED). So **if the difference between two means is greater than twice the SED, *t* > 2.** And for a healthy *n*, *t* > 2 is significant at the two-tailed $\alpha = 0.05$ level (have a quick glance at your tables of critical values of *t*).

The SED is therefore a very good index of variation that can be used to make visual comparisons directly, particularly if you draw error bars that are 2SED long — if the means to be compared are further apart than the length of this bar, there's a good chance the difference is significant. However, it's a bit more work to calculate the SED, which is why you don't see it very often.

If you want to work out an SED, just choose the appropriate *t* test and calculate the denominator of the *t* test. For between-group comparisons where the group SEMs are SEM_1 and SEM_2, you'll see that $SED = \sqrt{(SEM_1^2 + SEM_2^2)}$.

To summarize, for within-subject changes:

1. The mean within-subject change equals the difference of the group means.
2. The variance of the within-subject change may differ greatly from the variance of any one condition (group).
3. Present within-subject changes when the baseline varies a lot, or you want to show variance of the within-subject measure.
4. Present group means when the baseline matters.

3.9 Confidence intervals

One sample — confidence intervals on the population mean, μ

We can use the *t* formula to establish confidence intervals for particular measurements. Suppose when we measured the heights of a group of $n = 10$ UK men and found $\bar{x} = 1.82$ m, $s = 0.08$ m. We could calculate the 95% confidence interval like this. Since

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

we can work out 95% critical values for *t* (i.e. $\alpha = 0.025$ each tail) with $n - 1 = 9$ *df*. From our tables, these critical values are ±2.262. We can plug these into the formula above to find an expression for μ as a 95% confidence interval:

$$\pm 2.262 = \frac{1.82 - \mu}{\frac{0.08}{\sqrt{10}}}$$

$$\mu = 1.82 \pm 0.06$$

What would this mean? That there is a **95% chance that the true mean** height of UK men is in the range 1.76 to 1.88 m. We could also write this as a general formula:

$$\mu = \bar{x} \pm t_{critical\,(n-1)df} \frac{s_X}{\sqrt{n}}$$

Two samples — confidence intervals on a difference between means, $\mu_1 - \mu_2$

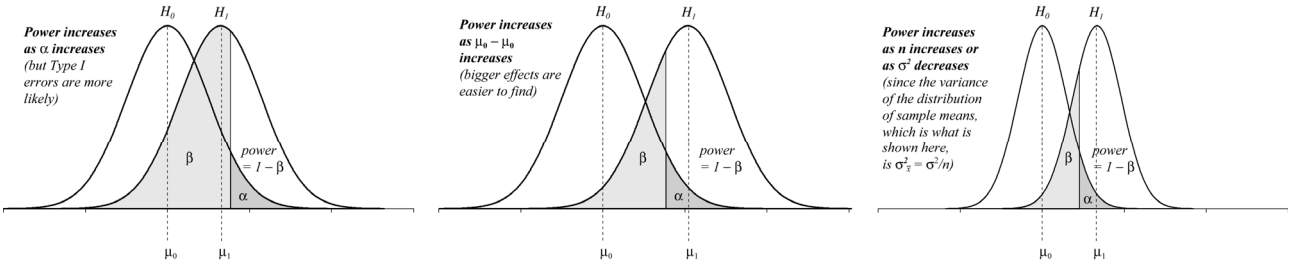
Similarly, if we have two samples whose mean difference is $\bar{x}_1 - \bar{x}_2$, we can use the formula for a two-sample *t* test to find the interval within which there is a 95% chance of finding the underlying population difference, $\mu_1 - \mu_2$.

3.10 Power and things that affect it

We won't talk about power in any great detail; certainly, you're not expected to calculate power. But it is helpful to understand what power is. Remember (from Handout 1) that α is the probability of rejecting the null hypothesis H_0 when it is in fact true (a Type I error); β is the probability of not rejecting H_0 when it is in fact false (a Type II error); power is $(1 - \beta)$, or the probability of rejecting H_0 when it is in fact false. If your power is 0.8, it means that you will detect 'genuine' effects with p = 0.8.

The consequences of Type II errors can be just as serious as those of Type I errors. If you run an expensive experiment with a very low power, you have a very small chance of finding the effect that you're looking for even if it does exist; if you then *don't* find it, you've probably wasted your time and money. (If you ever plan to run a seriously expensive experiment, make sure you understand how to do power calculations to work out how big your sample size should be, or ask a statistician to do it for you!)

Several things affect power: the size of the effect you're looking for (the difference between μ_0 and μ_1 — bigger effects give higher power), the sample size (n — the more observations you have, the higher the power), the variance of the sample (σ^2 — smaller variances give higher power), and of course your chosen level of α (higher α means lower β and therefore higher power, although higher α increases the chance of a Type I error). Have a look at the piccie (below).



Factors affecting power. If H_0 is true, and we take a set of samples each with mean $\bar{x}$, the mean of all the values of $\bar{x}$ will be μ_0. If H_1 is true, the mean of $\bar{x}$ will be μ_1. The distribution of all the values of $\bar{x}$ — the so-called 'sampling distribution of the mean' will be the curve labelled H_0 (if H_0 is true) or H_1 (if H_1 is true instead). The area under each curve is 1. Our job is to try to distinguish whether H_0 or H_1 is true on the basis of a single sample mean $\bar{x}$. We do this by setting α, the proportion of times that we reject H_0 when it is true. Setting α creates a criterion and thereby determines β, the chance of rejecting H_1 when it is true. In turn, this determines power, since this is $1 - \beta$ (the rest of the area under the H_1 curve). However, things other than α also affect power (middle and right-hand figures).

One thing that you should remember from this is that **significance levels do not indicate effect size.** Extremely large samples have power to detect very small effects with very small p values. Suppose a carefully-controlled study of a million people finds that running two miles a day decreases the risk of puffy ankles by 1% ($p < 0.001$). This is a study with high power finding a small effect that probably isn't important. On the other hand, **absence of evidence is not evidence of absence** — underpowered studies may fail to find large effects. A study of twenty 50-year-old men with heart disease might find no evidence that aspirin decreases the risk of a heart attack over the next five years ($p > 0.1$). This is a study with very low power failing to detect quite a substantial and important effect (aspirin does indeed reduce this risk).

3.11 Supplementary material: deriving the one-sample *t* test

The sampling distribution of the mean and the central limit theorem

Suppose we have a population with mean μ and variance σ^2. If we repeatedly take samples of n observations, we can say some things about the samples that we take. For each sample, we can calculate a sample mean $\bar{x}$. So we can collect lots of different sample means — many values of $\bar{x}$. Now we can ask what might at first appear to be an odd question: what will be the distribution of the sample means (also known as the **sampling distribution of the mean**)? What will be the mean of all the sample means (the mean of all the values of $\bar{x}$, written $\mu_{\bar{x}}$)? What will be the standard deviation of all these sample means (the standard deviation of all the values of $\bar{x}$, written $\sigma_{\bar{x}}$)? What we need to know is contained in a fact called the **central limit theorem.** There are various ways of stating this. The simplest is that if W_1, W_2, … W_n are independent, identically distributed random variables and $Y = W_1 + W_2 + … + W_n$, then the probability density function of Y approaches the normal distribution as $n \rightarrow \infty$. (This explains why the normal distribution so closely approximates so many biological, sociological, economic, and other variables that are themselves the sum of the effects of many other variables.) A more thorough version of the central limit theorem applicable to our present needs is this:

> Given a population with mean μ and variance σ^2, from which we take samples of size n, the distribution of sample means will have a mean $\mu_{\bar{x}} = \mu$, a variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$, and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. As the sample size n increases, the distribution of the sample means will approach the normal distribution.

This is very important. It doesn't matter whether or not the population is normally distributed; if you sample from it, the distribution *of* the sample means always approaches the normal distribution. (If the population is normally distributed and unimodal, the sample means will be normally distributed even if n is small; if the population is very skewed, n may have to be quite large — e.g. >30 — before the distribution of the means starts to become normally distributed.)

If we know the population SD, σ, we can test hypotheses very simply with a Z test

It's unusual for us to know the population standard deviation, σ. But sometimes we do. For example, we know that IQ in the general population has a mean of 100 and a standard deviation of 15. In this case, we saw in the Background Knowledge handout that we could calculate the probability that a single individual with an IQ of 89 came from the general population. We could calculate a Z score:

$$z = \frac{x - \mu}{\sigma}$$

which in this case would be $z = (83 - 100) / 15 = -1.13$; we could look this up in our tables and find that the probability that a single IQ score of 83 or less could come from the general population is 0.129. We would not reject the null hypothesis that this subject was drawn from the general population.

But suppose that we have five subjects, and their IQs are 89, 94, 73, 82, and 77. Are these *five* subjects drawn from a healthy population (mean 100, SD 15)? The null hypothesis is that they are (null hypothesis: population mean $\mu = 100$). So what we do is this. We calculate our sample mean $\bar{x} = 83$ and sample size $n = 5$. We know from the central limit theorem that if we repeatedly took samples of size 5 from a population with $\mu = 100$ and $\sigma = 15$, that these sample means ($\bar{x}$) themselves would have a mean of $\mu_{\bar{x}} = 100$ and a standard deviation $\sigma_{\bar{x}} = 15 / \sqrt{5} = 6.71$. We also know from the central limit theorem that the distribution of the sample means ($\bar{x}$) approaches a normal distribution. So we could obtain a Z score again:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{83 - 100}{6.71} = -2.53$$

Using our tables of *Z* scores, we'd find that the probability of obtaining a *Z* score of –2.53 or more extreme is 0.0057. If we set our α to be 0.05 with a two-tailed test (α = 0.025 each tail), we'd reject the null hypothesis, and conclude that our group of five subjects were not drawn from the general healthy population; the group mean of 83 was *significantly different* from 100. (It should be fairly obvious that our likelihood of finding a significant difference depends on the sample size *n*; larger samples have more **power** to detect a significant difference.)

More often, we do not know the population SD, σ, and can't use a Z test…

It's much more common that we don't know the population SD, σ, or the population variance, σ^2, so we have to estimate it from the sample SD, *s*, or the sample variance, s^2. Unfortunately, this complicates matters a bit. Although in the long run, the average value of the sample variance s^2 is equal to σ^2 (it's an *unbiased estimator*; see the Background Knowledge handout if you're interested), the distribution of s^2 is *positively skewed*. That means that although the average value of s^2 equals σ^2, more than half the values of s^2 are less than σ^2 (and less than half are more than σ^2 — though the values that are more than σ^2 are *much* more than σ^2, to balance things out). So any *individual* value of s^2 is likely to underestimate σ^2.

What we have to do to compensate is to change test from a *Z* test to something called a *t* test. Instead of calculating a *Z* score based on σ:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

we calculate a *t* score based on *s*:

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

Since s^2 is more likely than not to be smaller than σ^2, *t* is more likely than not to be bigger than *z*. The *t* score is not normally distributed; it has its own distribution. This distribution was worked out by William Gossett in 1908. Gossett worked for Guinness and they wouldn't let him publish under his own name, so he published under the pseudonym of Student. The distribution is therefore called **Student's *t* distribution.** There are in fact infinitely many *t* distributions, one for each *degree of freedom* (*df*; see below). For a one-sample *t* test, the number of degrees of freedom is $n - 1$, where *n* is the number of observations in the sample. As $n \to \infty$, $df \to \infty$, the distribution of s^2 becomes less and less skewed, and the *t* distribution becomes more and more like the normal distribution, *Z*. Anyway, we don't routinely need to calculate the distribution of *t* because we have it in the form of pre-calculated tables. If our calculated value of *t* exceeds the relevant critical value for the appropriate number of degrees of freedom and α, we reject the null hypothesis.

Degrees of freedom (df)

When we begin, we have *n* observations, and all of them are free to vary. When we obtained the sample variance, s^2, we calculated the deviations of each observation from the sample mean ($x - \bar{x}$), rather than from the population mean ($x - \mu$). Because the sum of the deviations about the mean, $\sum(x - \bar{x})$, is always zero, only $n - 1$ of the deviations are free to vary. We've 'used up' one of our degrees of freedom by calculating $\bar{x}$ using data from our sample. So s^2 is based on $n - 1$ degrees of freedom, and so is our *t* statistic.

3.12 Supplementary material: deriving the two-sample *t* test

The distribution of differences between means; deriving the two-sample t test

When we want to compare two groups, what we do is take two samples from two different populations, X_1 and X_2, and ask if the two populations have the same mean ($\mu_1 = \mu_2$) or not ($\mu_1 \neq \mu_2$). Suppose the populations have means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. If we draw pairs of samples, of size n_1 from population X_1 and of size n_2 from population X_2, we can calculate the *difference* between each pair of sample means $\bar{x}_1$ and $\bar{x}_2$, or $\bar{x}_1 - \bar{x}_2$. If we draw many pairs of samples, we can calculate the **distribution of the differences between sample means,** also called the sampling distribution of differences between means. The mean difference between sample means will be given by

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

From the central limit theorem, we know that the variance of sample means from X_1 will be $\sigma_{\bar{x}_1}^2 = \frac{\sigma_1^2}{n_1}$, and similarly $\sigma_{\bar{x}_2}^2 = \frac{\sigma_2^2}{n_2}$. The **variance sum law** states that the variance of a sum or difference of two variables is:

$$\sigma_{X_1 + X_2}^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$$

$$\sigma_{X_1 - X_2}^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

where ρ is the correlation between them; therefore, for two *independent* variables ($\rho = 0$), the variance of the sum or difference of the variables is the sum of their variances ($\sigma_{X_1 \pm X_2}^2 = \sigma_1^2 + \sigma_2^2$). Therefore, the variance of the difference between our two means will be

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

so the corresponding standard deviation is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$$

This is called the **standard error of the difference between means (SED).** We now know the mean and SD of the distribution of the differences between sample means; all that's left is to determine the shape of this distribution. Another theorem tells us that the sum or difference of two independent normally-distributed variables is itself normally distributed; we're basically done. If we knew the population SDs σ_1 and σ_2 — which is very unusual! — we could perform a *Z* test:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

… but that's very unlikely. So just as we used a *t* test in place of a *Z* test earlier, when we had to estimate σ based on the sample SD, *s*, we'll do the same now:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Generally, the null hypothesis is that the means are the same ($\mu_1 - \mu_2 = 0$), so we can simplify this a bit:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and there we have the formula that I stated at the top for when $n_1 = n_2$. What about the **degrees of freedom?** Well, we started with $n_1 + n_2$ degrees of freedom. We've calculated two sample variances, so we've lost 2 *df*; we're left with ($n_1 + n_2 - 2$) *df*.

Pooling variances when $n_1 \neq n_2$

Actually, the use of the *t* test for two independent samples requires the assumption that the two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$). We could denote this variance simply σ^2. This is often a reasonable assumption, particularly if we start with two groups of equivalent subjects ($\Rightarrow$ equal variances) and then do something to one or both groups that affects the mean of those groups; the variances will often be relatively unaffected. Anyway, when we use the *t* test, we are using the sample variances s_1^2 and s_2^2 to estimate σ^2. If our sample sizes are not equal ($n_1 \neq n_2$), then the larger sample will probably give us a *better* estimate of σ^2 (both s_1^2 and s_2^2 are meant to be estimating the same thing, since we're assuming $\sigma_1^2 = \sigma_2^2$, and the larger sample contains more information). Accordingly, we would be better off with a **weighted average,** in which the sample variances are weighted by their degrees of freedom ($n - 1$), the number of observations on which they are based. This weighted average is usually called the **pooled variance estimate,** s_p^2:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If we use that in our *t* test, we get the general formula for the two-sample unpaired *t* test that we began with:

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

It's just the same as the first formula in that it involves dividing the difference between means by the standard error of the difference between means (SED). The only difference is how we calculate the SED. If the sample sizes are equal ($n_1 = n_2$), then the two formulae are equivalent.

Another way of thinking about the pooled variance is in terms of *sums of squares*; we mentioned this in some of the wavy-line bits of Handout 2 (on what r^2 means in correlation). A variance is a 'sum of squares' (the sum of squared deviations from the mean) divided by the degrees of freedom. So when we multiple each sample variance by its own *df* we get the sample sums-of-squares. We also said that you could only add sample variances meaningfully when they were based on the same *df*, but you can add sums of squares any way you like — so to calculate the pooled variances, we convert the sample variances to the sample sums-of-squares, add them together, and divide by the overall number of *df* to get the overall (pooled) variance.

Bibliography

Howell, D. C. (1997). *Statistical Methods for Psychology*. Fourth edition, Wadsworth, Belmont, California.
Keppel, G. (1991). *Design and analysis: a researcher's handbook*. Third edition, Prentice-Hall, London.
Winer, B. J. (1971). *Statistical principles in experimental design*. Second edition, McGraw-Hill, New York.