Dependent T-Tests:

Dealing with multiple nested observations.

The Problem: Keith's Medical Reasoning Test

Let's pretend that I have created a test of diagnostic decision making for the AMA. On average, people score about 500 points on this test. If someone scores over 600 points, they show very exceptional diagnostic and medical reasoning.

To the right are the data from 10 subjects who took the medical reasoning test.

Group 1 is the control group, who were not fatigued prior to taking the test.

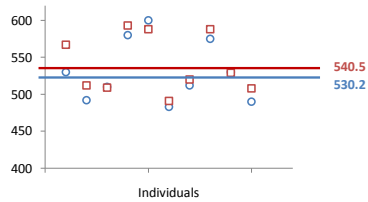
Group 2 is the experimental group, the same individuals worked for 12 operational hours prior to re-taking the test.

Subject #	Group 1	Group 2
1	567	530
2	512	492
3	509	510
4	593	580
5	588	600
6	491	483
7	520	512
8	588	575
9	529	530
10	508	490

Experimental design.

- What is the null hypothesis?
- What is the alternative hypothesis?
- What is the alpha level?
  - Is the test 1- or 2-tailed?
- What is our IV?
- What is our DV?

- How do we test this?



The mean of group 1 (540.5) and the mean of group 2 (530.2) are very close to each other. Given the large overlap (spread) it is unlikely that this effect will be significant.

The Independent T-Test

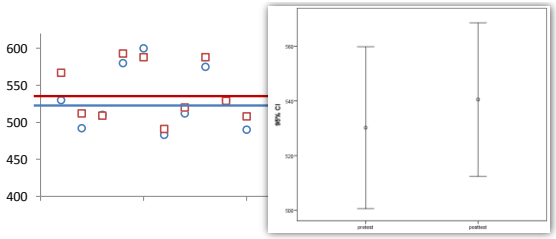
$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_{\bar{x}_1 - \bar{x}_2}}$$

$x_1 - x_2$= the difference between our sample means

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

$S_{x1\text{-}x2}$ = the common standard deviation, assuming samples are the same size.

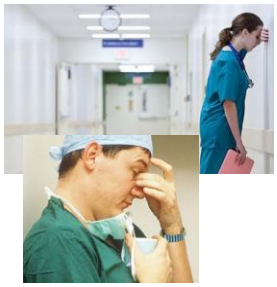
Thus, what the t-statistic is asking is: **Is the difference between our sample means an unusually large difference given the variance?**



d.f.	t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
18	-.571	p = .575	-10.30000	18.02347

There is no significant difference between the mean of Group 1 and the mean of Group 2, $t(18) = -.571$, $p = .58$.

So, what do we conclude?



- Based on these data, and the statistical methods we used, we would concluded that fatigue has no effect of medical decision making.

But!

- ...there is a problem with this study. What is wrong with the study the way we tested it?
- We used an **independent t-test** when we have **dependent samples** and this a **problem.**

This should be a dependent t-test!

- I think that fatigue really had an effect!
- Almost everyone got worse after fatigue…
- ***Remember that Group 1 and Group 2 were the same people at different times!***

Subject #	Group1 (Pre)	Group2 (Post)	Gender	Hand	Age
1	567	530	M	R	24
2	512	492	F	R	26
3	509	510	M	R	25
4	593	580	M	L	27
5	588	600	F	L	24
6	491	483	F	R	25
7	520	512	M	R	26
8	588	575	F	R	26
9	529	530	F	L	25
10	508	490	M	R	24

Why is dependence a problem?

- **Individual differences**: Data are noisy because people are different.
  - Some people score high, some people score low, a lot people are in between.
  - Ability
  - Experience
  - Acute effects (hunger, caffeine, etc.)
- In a hypothesis test, we want to know what variation in the DV is explained by the IV, and what variance is due to other factors.

Why is dependence a problem?

- **Scores that are nested within subjects create a problem, because these scores are statistically dependent on each other!**
  - Dependence means that the occurrence of one event makes another event more or less probable.
- Thus, to understand what variance in the DV is explained by the IV, we need to control for dependence.

Dependence and explained variance.

- The basic formula for hypothesis testing statistics:

$$\frac{\text{(differences between groups caused by IV)}}{\textbf{(variability within the groups)}}$$

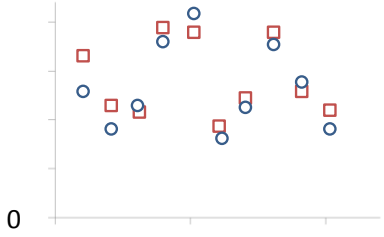
The bottom part of the equation (variability within groups) is composed of many parts:

1. *due to measurement error*
2. *due to individual differences*
3. *due to other factors…*

There are many things we can do statistically and experimentally to control for these factors.

$$\frac{\text{(differences between groups caused by IV)}}{\text{(error + individual differences)}}$$

Removing individual differences:



Dependent t-test

- First we can calculate the actual statistical dependence in our data using a *correlation* statistic:

Subject	Pretest	Posttest
1	567	530
2	512	492
3	509	510
4	593	580
5	588	600
6	491	483
7	520	512
8	588	575
9	529	530
10	508	490

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	pretest & posttest	10	.945	.000

Dependent t-test

- Instead of comparing the difference between means, we subtract paired scores from each other to create difference scores:

Subject	Pretest	Posttest	Difference
1	567	530	-37
2	512	492	-20
3	509	510	1
4	593	580	-13
5	588	600	12
6	491	483	-8
7	520	512	-8
8	588	575	-13
9	529	530	1
10	508	490	-18

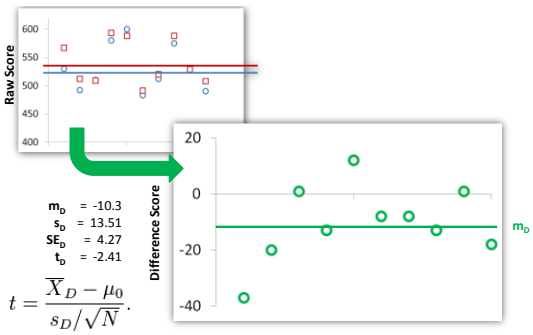
We now have 10 independent observations!

Knowing the difference score for a single subject does not allow us to predict the score for a different subject.

Experimental design.

- What is the null hypothesis?
- What is the alternative hypothesis?
- What is the alpha level?
  - Is the test 1- or 2-tailed?
- What is our IV?
- What is our DV?

- How do we test this?



By computing the appropriate t-statistic, a dependent t-test, we see that there is a significant effect of fatigue, *t*(9) = -2.41, *p* = .03.

4 Sentence summary:

- We were interested in the effects of fatigue on a test of diagnostic and medical decision making.
- A random sample of 10 medical residents were given a pre-test, worked a 12 hour shift and were then given a post-test.
- There was a significant decrease in subjects decision making ability from the pre-test (540.5) to the post-test (530.2), *t*(9) = -2.41, *p* = .03.
- Thus, we conclude that working a full 12-hour shift has a significant negative effect on medical residents diagnostic and medical decision making.

As an independent T-Test (Incorrect).

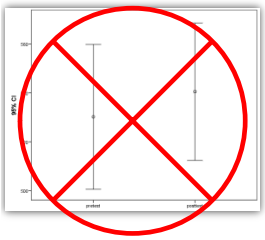
d.f.	t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
18	-.571	p = .575	-10.30000	18.02347

As a dependent T-Test (Correct).

d.f.	t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
9	-2.41	p = .039	-10.30000	4.27

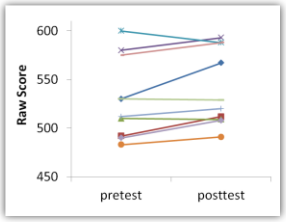
Because scores were nested within subjects, variability due to individual differences were "washing out" the effects of our IV.

By controlling for this nesting (creating independent difference scores) we get a clearer understanding of variation in the DV that is attributable to the IV.



(differences between groups caused by IV)
(**error** + **individual differences**)

An independent T-Test (Incorrect).



(differences within individuals caused by IV)
(**error**)

A dependent T-Test (Correct).

To Review…

I REPRESENT THE IMPORTANCE OF PROPER NESTING!!!

Observations that come from similar sources can be "nested".

Proper nesting allows you to control for a lack of independence in your observations.

Scores nested within students.

Students nested within classrooms.

Classrooms nested within schools.

Etc… .

A dependent t-test allows you to compare each person to herself, reducing individual differences.

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{N}}.$$