جامعة نيويورك أبوظبي

NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

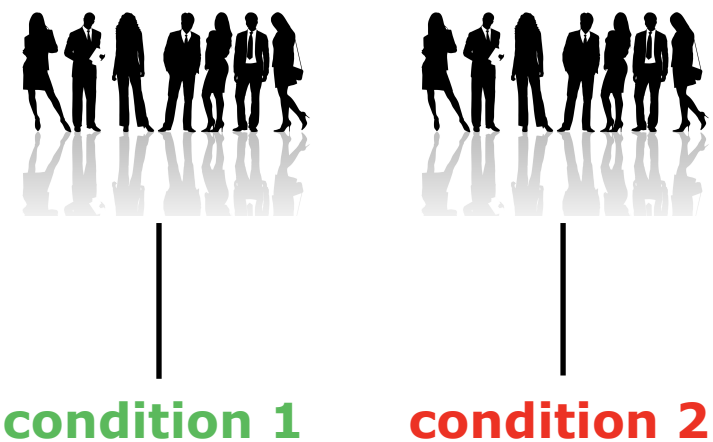
Class 17: Paired *t*-test, and t-test as linear regression

Prof. Jon Sprouse
Psychology

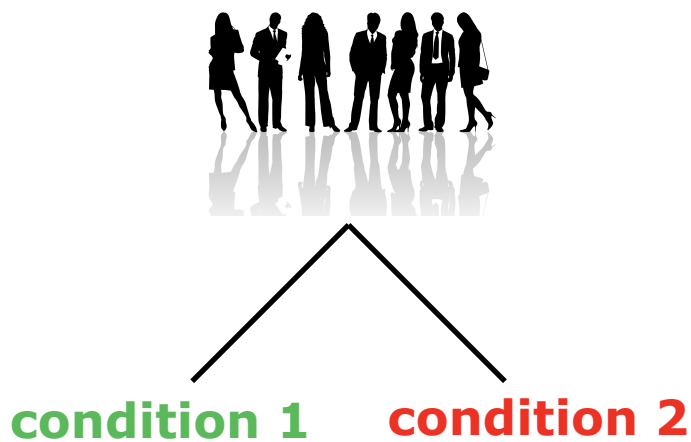
Independent Measures vs Repeated Measures

Remember, for experimental design we have two options:

Independent Measures



Repeated Measures



Independent Measures: If each participants sees only one condition, we call it independent measures. It is also called a **between-subjects** design.

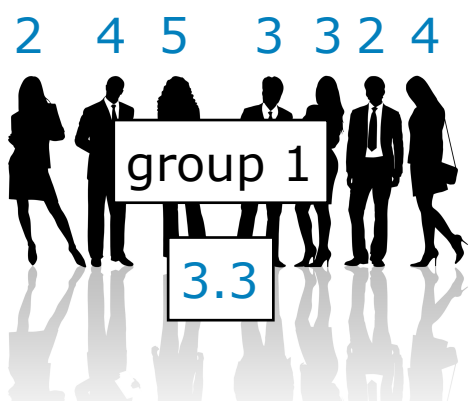
Repeated Measures: If each participants sees every condition, we call it repeated measures. It is also called a **within-subjects** design.

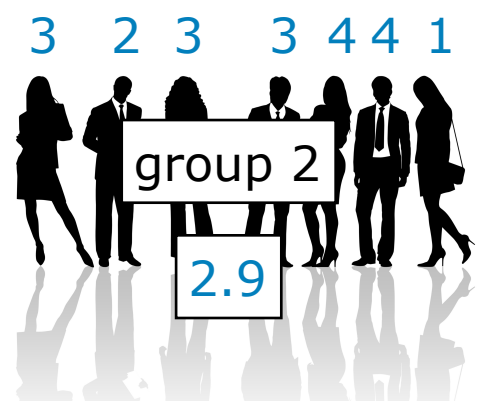
The disadvantage of independent measures

To see the advantage of repeated measures, we need to recognize the disadvantage with independent measures.

Let's imagine that you are testing two pills. The green pill is a placebo. It has nothing in it. The red pill has a medicine that should make ratings of wellbeing in people higher.

baseline wellbeing:





baseline $\bar{y}$:

placebo 0 0 0 0 0 0 0

medicine .2 .1 .3 .1 .2 .2 .4

measure: 2 4 5 3 3 2 4

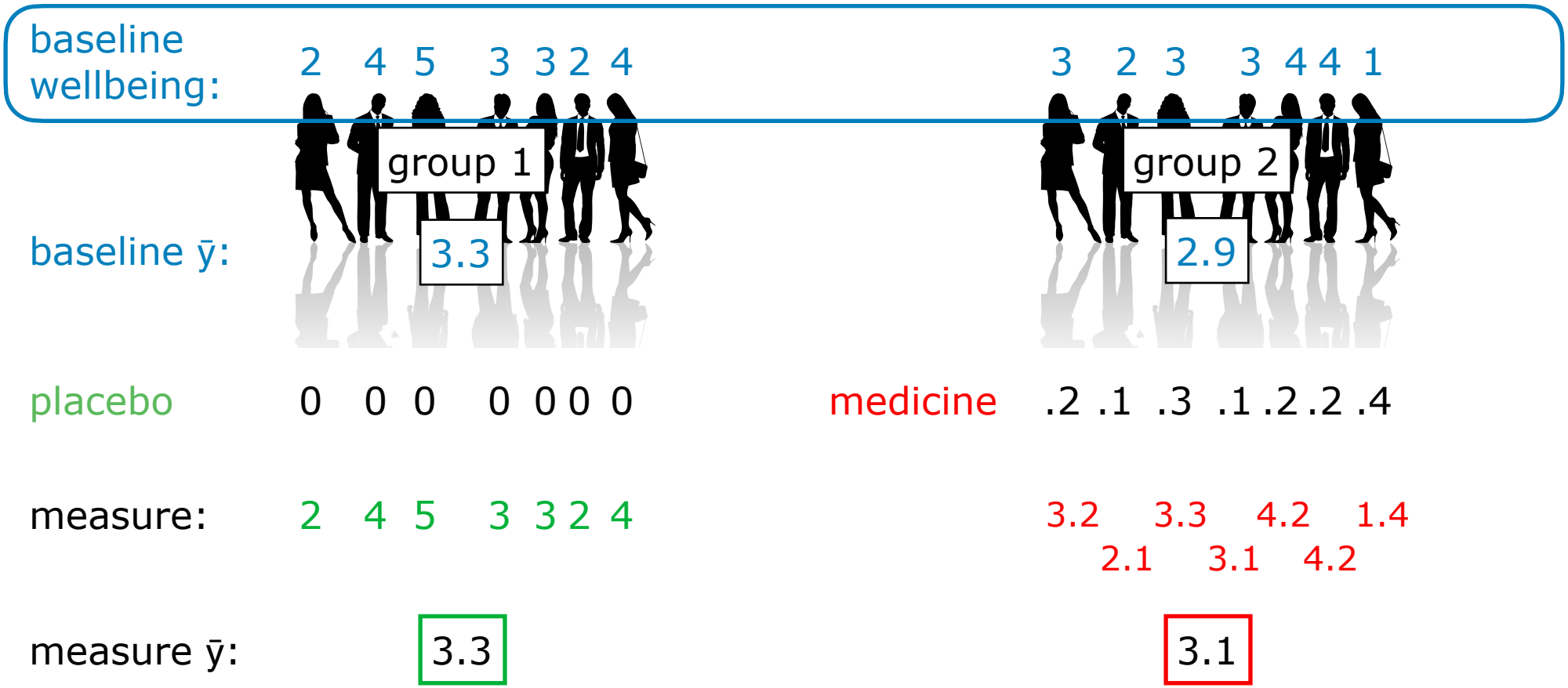
3.2 3.3 4.2 1.4
2.1 3.1 4.2

measure $\bar{y}$: 3.3

3.1

The disadvantage of independent measures

All people come with their baseline characteristics. And these vary between individuals.



The disadvantage of independent measures

All people come with their baseline characteristics. And these vary between individuals.

They also each have a different response to your treatment condition.

baseline wellbeing:

baseline ȳ:

placebo	0 0 0 0 0 0 0	medicine	.2 .1 .3 .1 .2 .2 .4

measure: 2 4 5 3 3 2 4

3.2 3.3 4.2 1.4
2.1 3.1 4.2

measure ȳ: 3.3

3.1

The disadvantage of independent measures

All people come with their baseline characteristics. And these vary between individuals.

They also each have a different response to your treatment condition.

When you run an independent measures experiment, you can't see these. **All you can see are your measurements!**

measure: 2 4 5 3 3 2 4

measure ȳ: 3.3

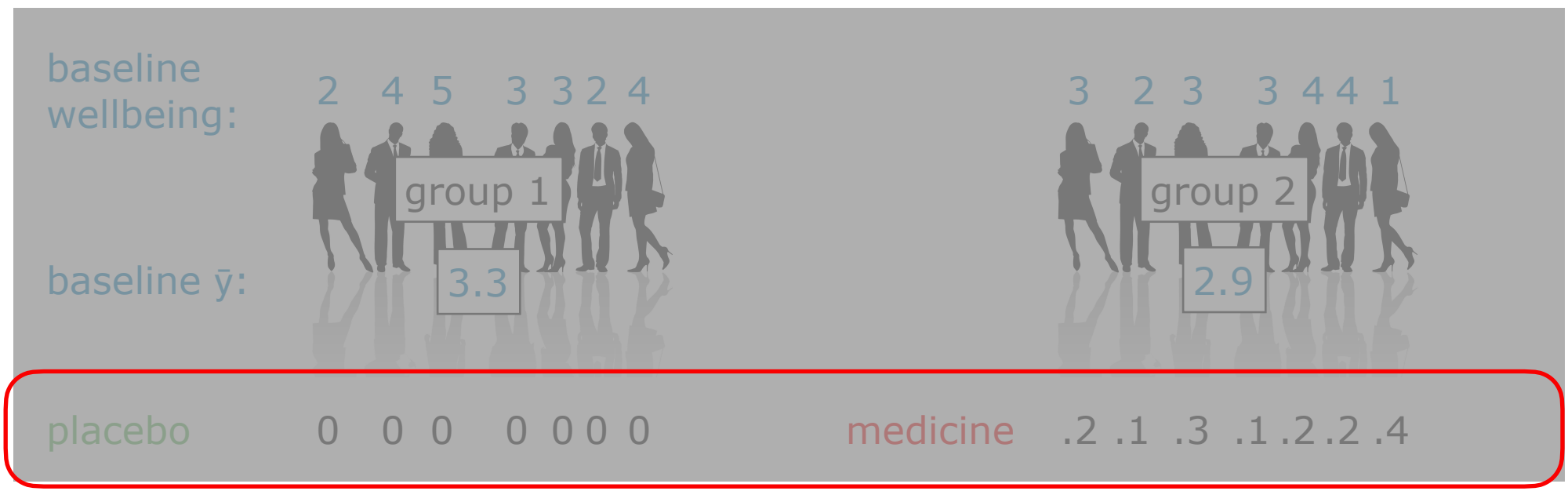
3.2 3.3 4.2 1.4
2.1 3.1 4.2

3.1

The disadvantage of independent measures

Looking at your measures, could you conclude that your medicine increased ratings?

No. The mean rating for your medicine group is actually lower than the placebo group. But we know that the medicine increased ratings for this group because we constructed the example. What happened?



measure: 2 4 5 3 3 2 4

3.2 3.3 4.2 1.4
2.1 3.1 4.2

measure ȳ: 3.3

3.1

The disadvantage of independent measures

The problem is that the effect of our medicine is smaller than the variation that humans naturally have between groups.

Group 2's baseline mean is 0.4 lower than group 1 to begin with, before either get a treatment. The treatment has an effect of 0.2. It is not enough to overcome the baseline difference.

Repeated measures solves this!

In repeated measures, you use the same group twice. That means you don't have to worry about variation in the baseline. Because the same baseline is there for both conditions.

Notice that these are now both "group 2". The same set of people!

baseline wellbeing: 3 2 3 3 4 4 1

baseline $\bar{y}$: 2.9

placebo 0 0 0 0 0 0 0

measure: 3 2 3 3 4 4 1

measure $\bar{y}$: 2.9

3 2 3 3 4 4 1

2.9

medicine .2 .1 .3 .1 .2 .2 .4

3.2 3.3 4.2 1.4
2.1 3.1 4.2

3.1

The advantage of repeated measures designs

Looking at your measures alone, just like a real experiment, could you conclude that your medicine increased ratings?

Yes, probably. There is an increase. We just need to formalize our method for determining if that increase is statistically significant.

measure: 3 2 3 3 4 4 1

measure ȳ: 2.9

3.2 3.3 4.2 1.4
2.1 3.1 4.2

3.1

Advantages and Disadvantages

Independent Measures

Repeated Measures

Independent Measures: The advantage of independent measures designs is that you don't have to worry about one condition affecting the other. (Like giving the same people two different medicines!) The disadvantage is that we have less sensitivity to detect small effect sizes.

Repeated Measures: The advantage of repeated measures is that we have more sensitivity to detect small effect sizes. But the disadvantage is that the two conditions could potentially affect each other. Before using this design, you have to be sure that there is no chance that they could interact.

Formalizing our measure of the effect for repeated measures

How do we measure the effect?

Independent Measures

condition 2 **condition 1**

condition 2	condition 1
3.2	2
2.1	4
3.3	5
3.1	4
4.2	4
4.2	2
1.4	4
$\bar{y}$: 3.1	$\bar{y}$: 3.3

effect: $\bar{y} - \bar{y}$

Repeated Measures

condition 2 **condition 1**

condition 2		condition 1		**D**
3.2	-	3	=	0.2
2.2	-	2	=	0.2
3.3	-	3	=	0.3
3.1	-	3	=	0.1
4.2	-	4	=	0.2
4.2	-	4	=	0.2
1.4	-	1	=	0.4
			$\bar{D}$ =	0.2

effect: $\bar{D}$

How do we measure the effect?

Independent Measures

condition 2 **condition 1**

Lisa	3.2	Kenny	2
Brian	2.1	April	4
Susan	3.3	Bret	5
Rich	3.1	Karen	4
Mary	4.2	Kevin	4
John	4.2	Madie	2
Rob	1.4	Heather	4

$\bar{y}$: 3.1 $\bar{y}$: 3.3

effect: $\bar{y} - \bar{y}$

In an independent measures design, we have to look at group means. We have no other choice.

We **cannot** subtract the measure for one participant from the measure for another because **they are different people**.

So we calculate the means for each group, and compare them. It is **the only way we can do it!**

How do we measure the effect?

Repeated Measures

In a repeated measures design, we can look within each participant for the effect. Because we are measuring them in both conditions.

So we can define our effect as the difference between each measurement. We call these **difference scores**, and we can give them the symbol **D**.

Once we calculate a difference score for each participant, we can calculate a mean of the difference scores for our effect. We give it the symbol $\overline{\mathbf{D}}$.

	condition 2		condition 1		D
Lisa	3.2	-	3	=	0.2
Brian	2.2	-	2	=	0.2
Susan	3.3	-	3	=	0.3
Rich	3.1	-	3	=	0.1
Mary	4.2	-	4	=	0.2
John	4.2	-	4	=	0.2
Rob	1.4	-	1	=	0.4
					$\overline{\mathbf{D}}$ = 0.2

effect: $\overline{\mathbf{D}}$

[mean, subtract] vs [subtract, mean]

Independent Measures

condition 2 **condition 1**

effect: $\bar{y} - \bar{y}$

mean $\begin{pmatrix} 3.2 \\ 2.1 \\ 3.3 \\ 3.1 \\ 4.2 \\ 4.2 \\ 1.4 \end{pmatrix}$ $-$ **mean** $\begin{pmatrix} 2 \\ 4 \\ 5 \\ 4 \\ 4 \\ 2 \\ 4 \end{pmatrix}$

$\bar{y}$: 3.1 $\bar{y}$: 3.3

Repeated Measures

condition 2 **condition 1**

effect: $\bar{D}$

3.2	-	3	=	0.2
2.2	-	2	=	0.2
3.3	-	3	=	0.3
3.1	-	3	=	0.1
4.2	-	4	=	0.2
4.2	-	4	=	0.2
1.4	-	1	=	0.4

mean **D**

$\bar{D} = 0.2$

Creating a *t*-test for D scores

We only have one sample… so…

In a repeated measures design, we only have one sample. So we can use the one sample *t*-test as a starting point for our test!

one sample *t*-test

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

But the difference is that we are using the mean of difference scores ($\bar{D}$) as our effect, not sample means. So we need to adjust the formula:

paired *t*-test

$$t = \frac{\bar{D} - \Delta_0}{s_{\bar{D}}}$$

Notice that the standard error term is the standard error of means of difference scores!

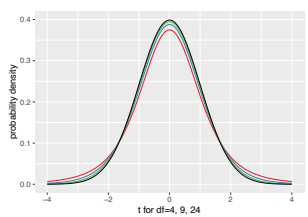
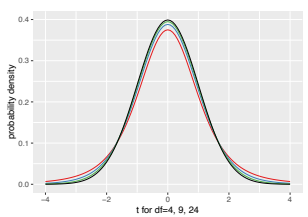
$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

	condition 2		condition 1		D
Lisa	3.2	-	3	=	0.2
Brian	2.2	-	2	=	0.2
Susan	3.3	-	3	=	0.3
Rich	3.1	-	3	=	0.1
Mary	4.2	-	4	=	0.2
John	4.2	-	4	=	0.2
Rob	1.4	-	1	=	0.4

$\bar{D} = 0.2$

$s_D = 0.1$

one sample vs paired *t*-tests

	one sample *t*-test	**paired *t*-test**
Scientific question	Does our sample differ from a population with a known mean (but unknown SD)?	Does our sample of **difference scores** differ from a population of difference scores with a known mean?
Null Hypothesis	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$)	The mean of the population of difference scores that the sample comes from is equal to the mean of the known population
Equation	$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$	$t = \frac{\bar{D} - \Delta_0}{s_{\bar{D}}}$
Descriptive information	The *t* statistic tells us how much our sample mean differs from the population mean in terms of sample SE (as an estimate)	The *t* statistic tells us how much our mean of differences differs from the population in terms of the standard error of differences
Null distribution	*t*-distribution based on df 	*t*-distribution based on df 

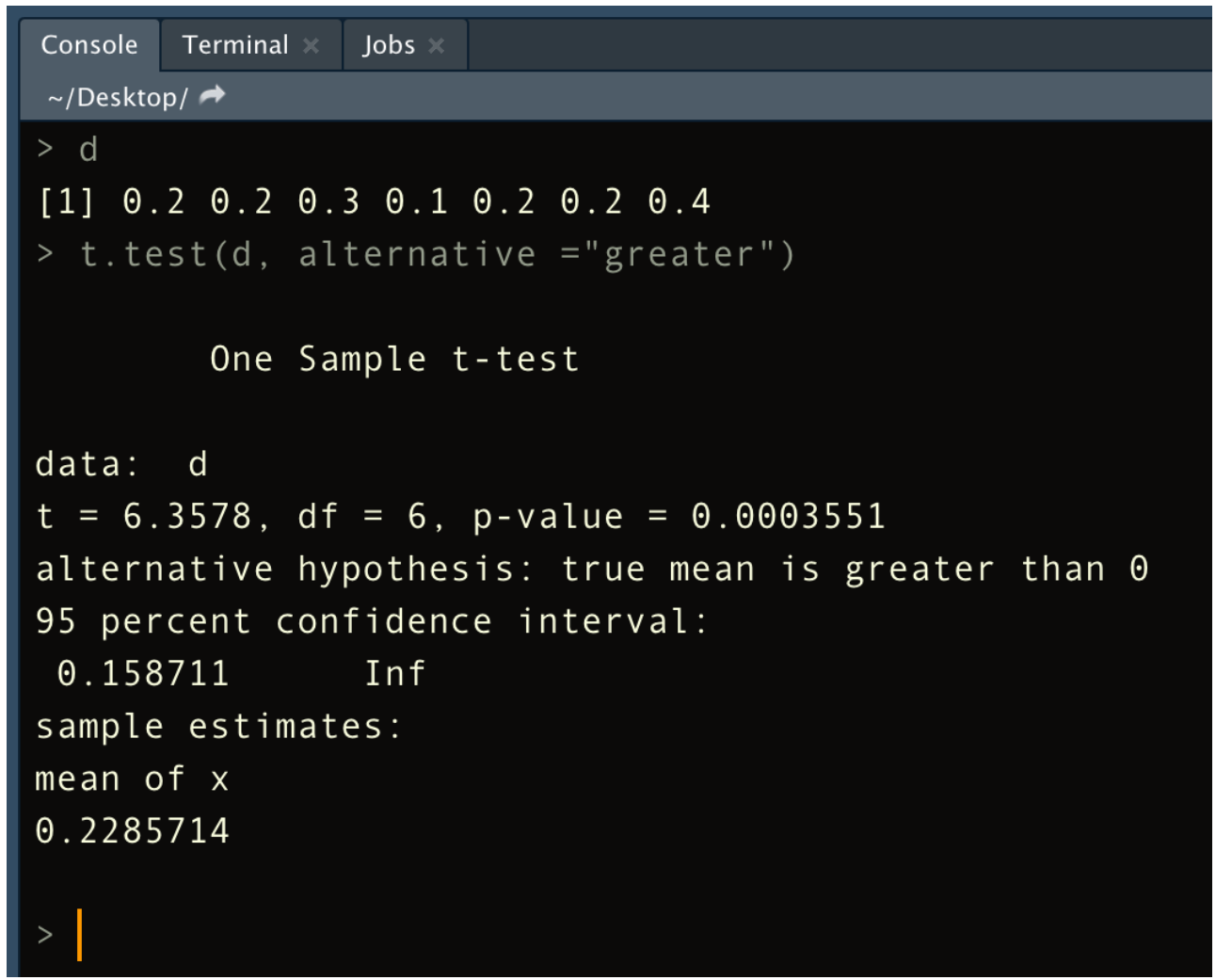
Running it in R

This is just a one-sample *t*-test performed on difference scores. So all we need is a vector of difference scores, and our old friend the function t.test().

vector of Ds

t.test() knows it is one-sample because there is only one vector in its input. We tell it that the hypothesis is that the differences are greater (than zero).

It gives us the *t*, the df, and the *p*-value. (Ignore the CI - CIs are two-sided by definition!)



The paired *t*-test on raw scores

Though conceptually the paired *t*-test is a one-sample *t*-test on difference scores, there is a way to compute the *t*-test directly from raw scores. It is similar to the independent samples formula because it involves two samples:

independent samples

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

paired

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

The difference between the two is this purple term that is subtracted from the standard error. This term makes the standard error smaller. This in turn makes the *t* larger. This is what we want. It makes the test more sensitive, just like the way difference scores are more sensitive!

You can see that this term is based on the correlation between the two sets of measures. If the correlation is high, a lot will be subtracted from the standard error. If there is no correlation, this term will be 0, and it will be identical to the independent samples *t*-test. This is good - if there is no correlation, the two samples are independent! The full denominator is just another arithmetic shortcut for estimating $s_{\bar{D}}$.

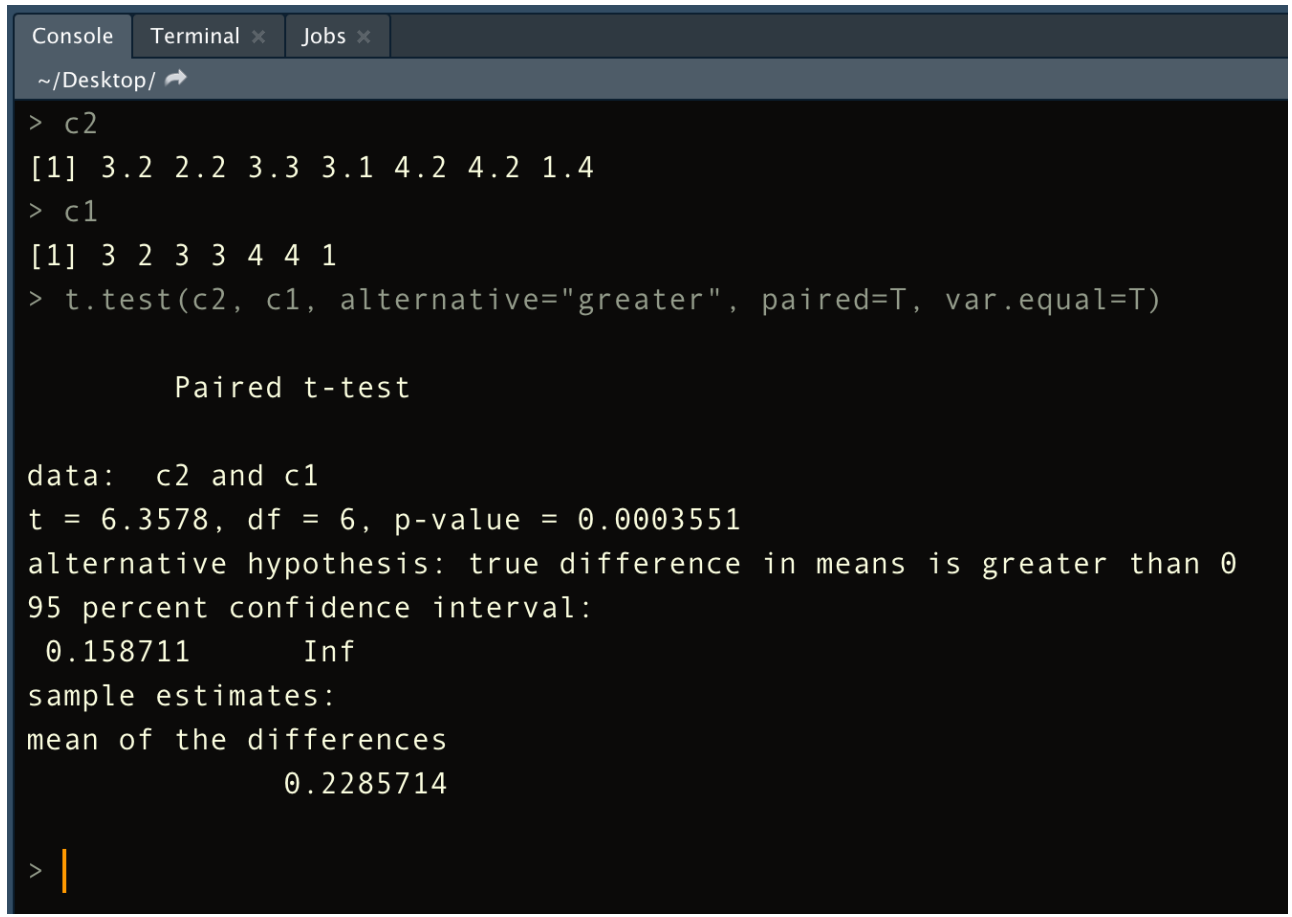
Running the raw score equation in R

The raw score equation is a two-sample test, but with the two samples paired. We can use the t.test() function for this too!

Two vectors, one for each condition

We have to enter both conditions. We also have to tell t.test() that we want a paired *t*-test by using paired=T. We also generally want var.equal=T.

Notice that the *t*, df, and *p*-value are identical to the one sample version!



OK, now for a change in topic:

t-tests are linear regression (what?!)

Let's create a data set to use as an example

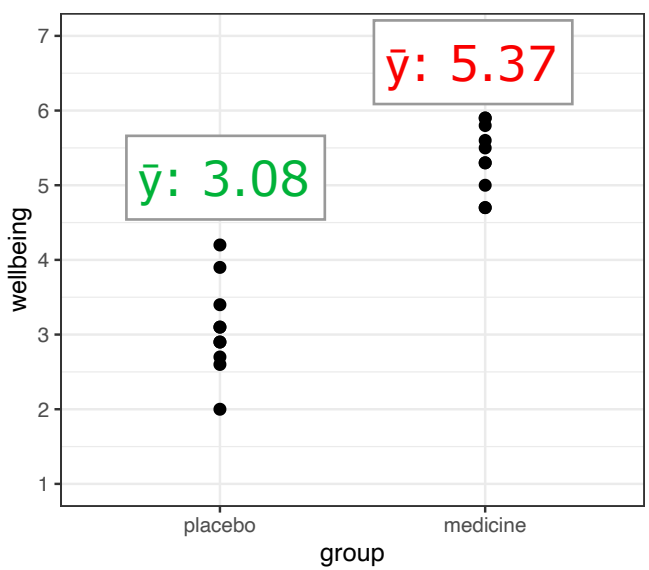
Let's use an independent measures design as our example. So we have two groups. I will use placebo and medicine as the two groups. But I will make the difference between the two groups a little bigger so we can see it. Let's say the response variable is a 7-point rating scale of "wellbeing".

```
placebo=round(rnorm(10, mean=3, sd=.75), 1)
```

```
medicine=round(rnorm(10, mean=5, sd=.75), 1)
```

```
data = tibble(group = rep(c("placebo", "medicine"), each=10), wellbeing = c(placebo, medicine))
```

This R code creates two vectors of 10 numbers, then combines them into a data set.



Here is a plot of our data points. I used this code:

```
ggplot(data, aes(x=group, y=wellbeing)) + geom_point(size =2) + theme_bw() + scale_y_continuous(limits=c(1,7), breaks=c(1:7), labels=c(1:7))
```

Linear regression on non-numeric variables requires **dummy coding**

Remember our equation for a line: $\hat{y} = b_1x + b_0$

This equation requires us to multiply x by b_1. This is easy when x is a numeric variable because its values are already numbers. But how can we multiply x by a non-numeric variable like "placebo" or "medicine"?

$\hat{y} = b_1x + b_0$

placebo = 0

medicine = 1

The answer is that we re-code the levels of the variable as numbers! This is called **dummy coding**.

This gives rise to two predicted values ($\hat{y}$), one for when the x has the value "placebo" and one for when the x has the value "medicine".

placebo: $\hat{y} = b_10 + b_0$

medicine: $\hat{y} = b_11 + b_0$

You should always make your boring condition 0 and your interesting condition 1. If they are both interesting, you just pick whichever you prefer.

Now let's create a linear model

I'll use lm() to calculate the coefficients for me.

$\hat{y} = b_1x + b_0$

The output of lm() tells us that:

$b_0 = 3.08$

$b_1 = 2.29$

$\hat{y} = 2.29x + 3.08$

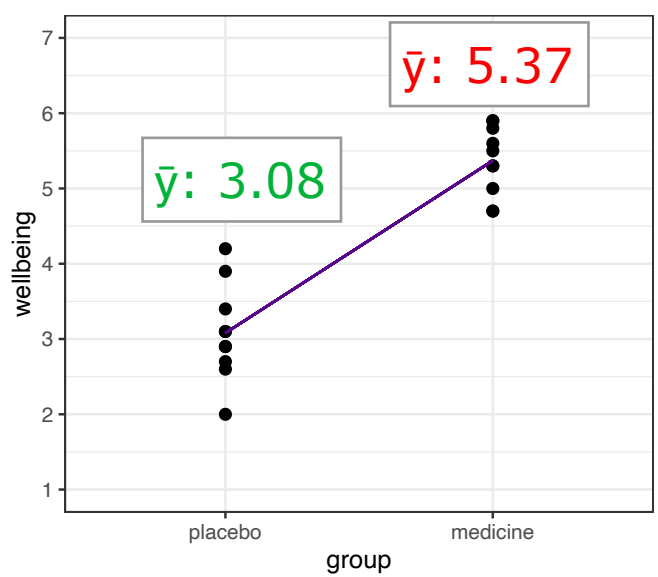
```
Console   Terminal   Jobs
~/Desktop/
> lm(wellbeing~group, data=data)

Call:
lm(formula = wellbeing ~ group, data = data)

Coefficients:
  (Intercept)  groupmedicine
         3.08           2.29

>
```

Do you recognize these values? I'll put the plot here again so you can see them!

$b_0 = 3.08$ $\quad \bar{y}: 3.08$

$b_1 = 2.29$ $\quad 5.37 - 3.08 = 2.29$

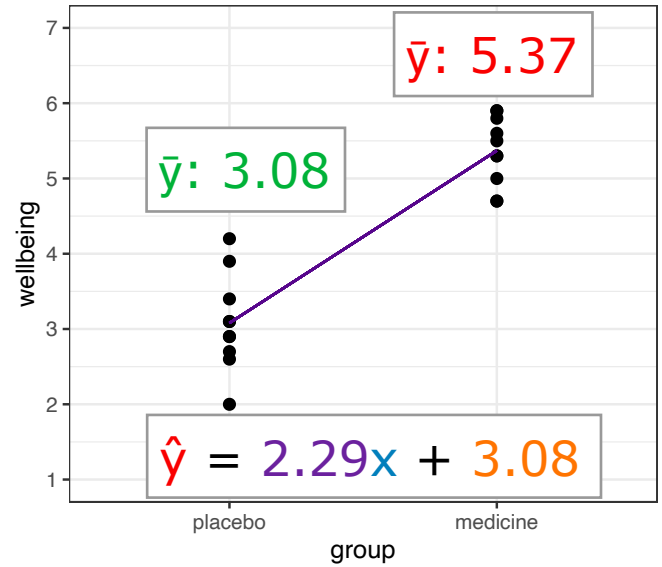
$\bar{y} \quad - \quad \bar{y}$



Why does the model work out this way?

This is our linear model:

Let's work through the two possible values of x: 0 and 1



placebo: $\hat{y} = (2.29)0 + 3.08$

$\hat{y} = 3.08$

The equation for the 0 condition (placebo), simply becomes the y-intercept b_0. So the y-intercept is the mean of the 0 condition!

medicine: $\hat{y} = (2.29)1 + 3.08$

$\hat{y} = 5.37$

The equation for the 1 condition (medicine) starts at the placebo condition and adds the slope. So the slope is the difference between means!

What about significance tests?

I am going to run a t-test on the slope of the linear model and run a t-test on the two conditions separately.

t-test on the linear model

```
Console  Terminal  Jobs
~/Desktop/
> summary(model)

Call:
lm(formula = wellbeing ~ group, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0800 -0.3725 -0.0250  0.3475  1.1200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.0800     0.1748  17.625 8.44e-13 ***
groupmedicine   2.2900     0.2471   9.266 2.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5526 on 18 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.8171
F-statistic: 85.86 on 1 and 18 DF,  p-value: 2.848e-08
```

t-test on the conditions

```
Console  Terminal  Jobs
~/Desktop/
> t.test(medicine, placebo, data = data, var.equal=T)

        Two Sample t-test

data:  medicine and placebo
t = 9.266, df = 18, p-value = 2.848e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.77078 2.80922
sample estimates:
mean of x mean of y
     5.37      3.08
```

They yield identical results. The *t* for the slope is 9.266 and the *t* for the conditions is 9.266!

Why do they yield identical results?

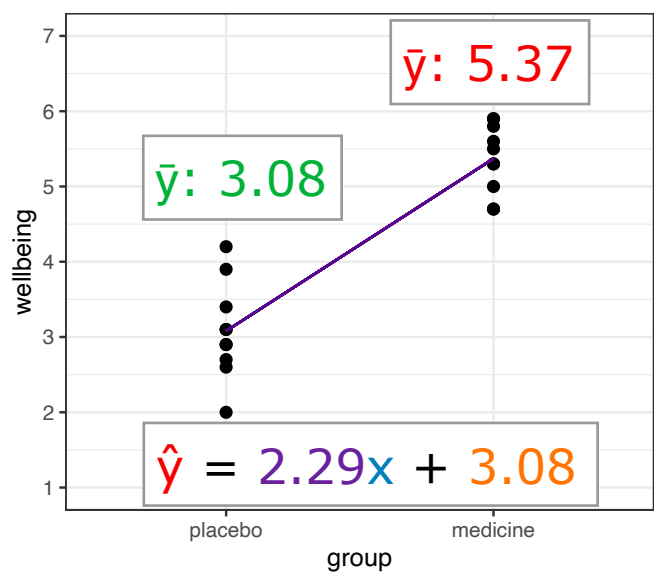
A *t*-test on a linear model asks whether the sample slope comes from a population with a known slope (typically 0).

$$t = \frac{b_1 - 0}{\frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}}}$$

A *t*-test on two conditions asks whether the two means come from two populations with known means (typically equal so their difference is 0).

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

In this case, the two questions are the same. The slope of our model is equal to the difference between means. So asking whether the slope is different from 0 is the same as asking whether the difference between means is 0.

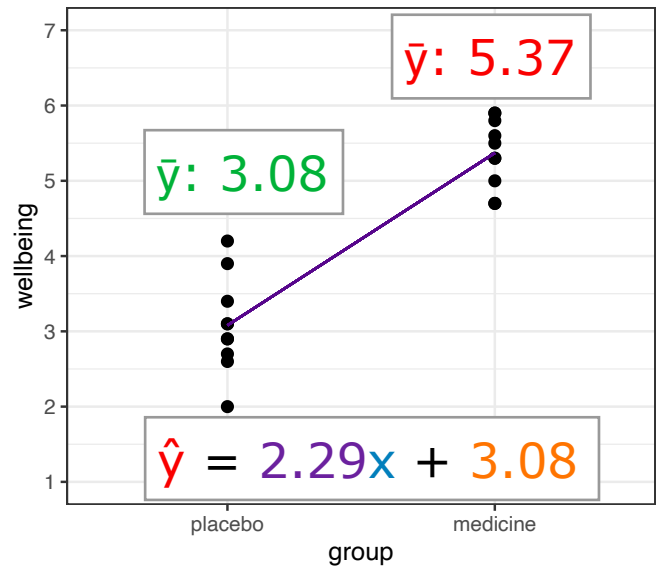


What does this mean?

Philosophically, it means that every time you run a *t*-test, you are actually creating a linear model of your data (with dummy coding).

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

=



This is a nice result for two reasons:

1. Instead of thinking about the statistics that we are doing in class as a collection of different tests, we can see it as one unified framework - linear modeling. Every test we do is a linear model!

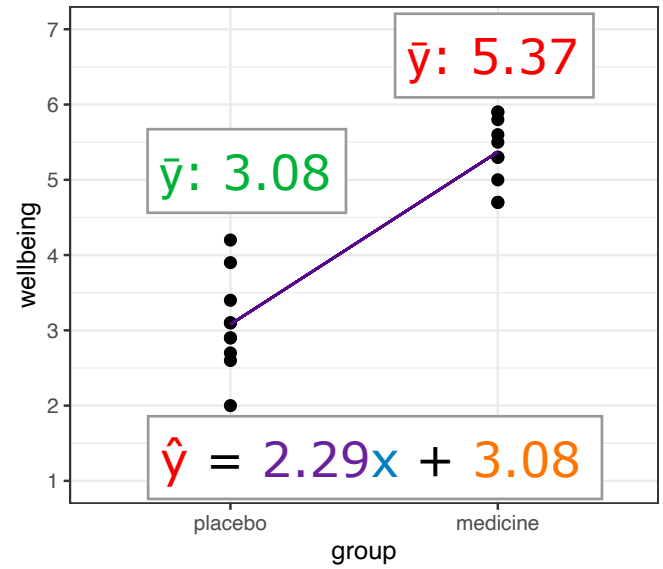
2. This shows us that our workflow for science is always the same: construct a theory, create a model, test the hypothesis. Even when it seemed like we were skipping modeling and going straight to t-tests, we were still modeling!

What does this mean?

Practically, it means you have two ways of doing the same thing. You can either choose to run a t-test (which implies a linear model), or you can choose to create a linear model explicitly, then calculate a t-test on its slope. There is no difference in the result.

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

=



t.test()

```
Console  Terminal  Jobs
~/Desktop/
> t.test(medicine, placebo, data = data, var.equal=T)

        Two Sample t-test

data:  medicine and placebo
t = 9.266, df = 18, p-value = 2.848e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.77078 2.80922
sample estimates:
mean of x mean of y
     5.37      3.08
```

lm()

```
Console  Terminal  Jobs
~/Desktop/
> summary(model)

Call:
lm(formula = wellbeing ~ group, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0800 -0.3725 -0.0250  0.3475  1.1200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.0800     0.1748  17.625 8.44e-13 ***
```