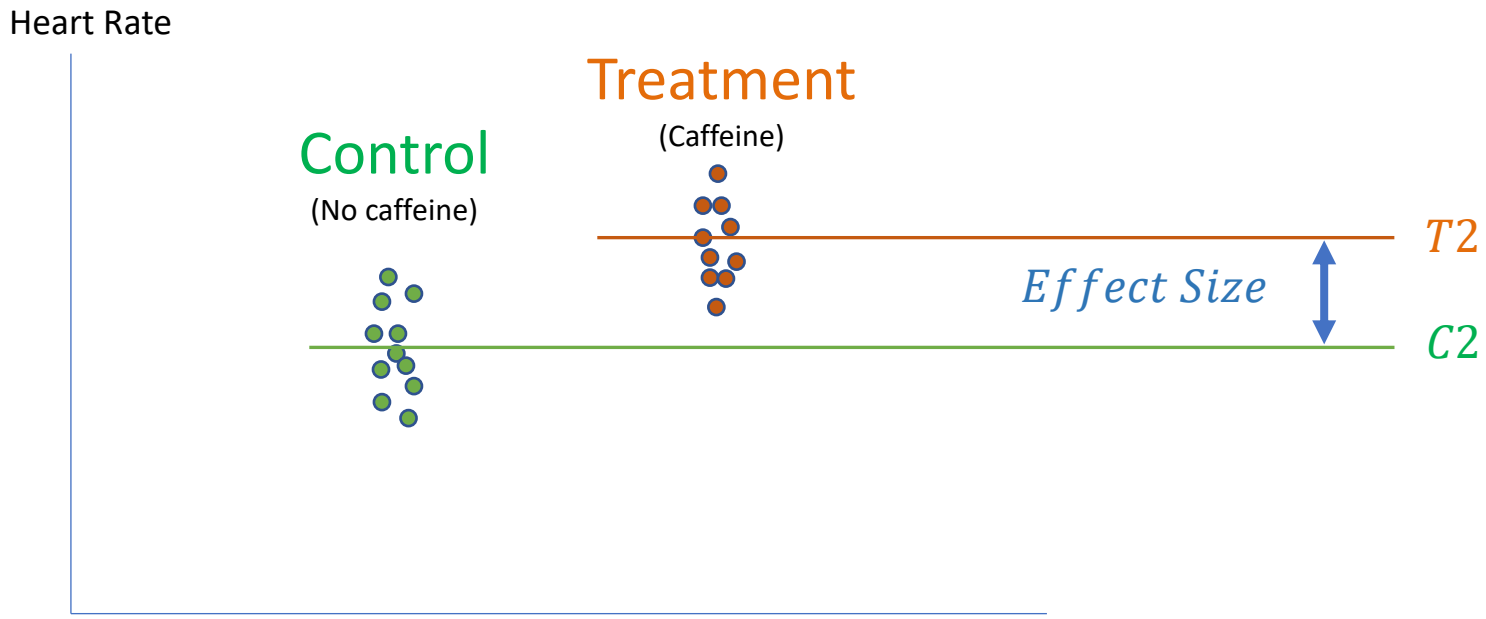
TESTS FOR GROUP EQUIVALENCE OR “BALANCE”

THE PROGRAM EVALUATION FRAMEWORK: "DISCRETE" TREATMENT GROUPS (YES/NO)

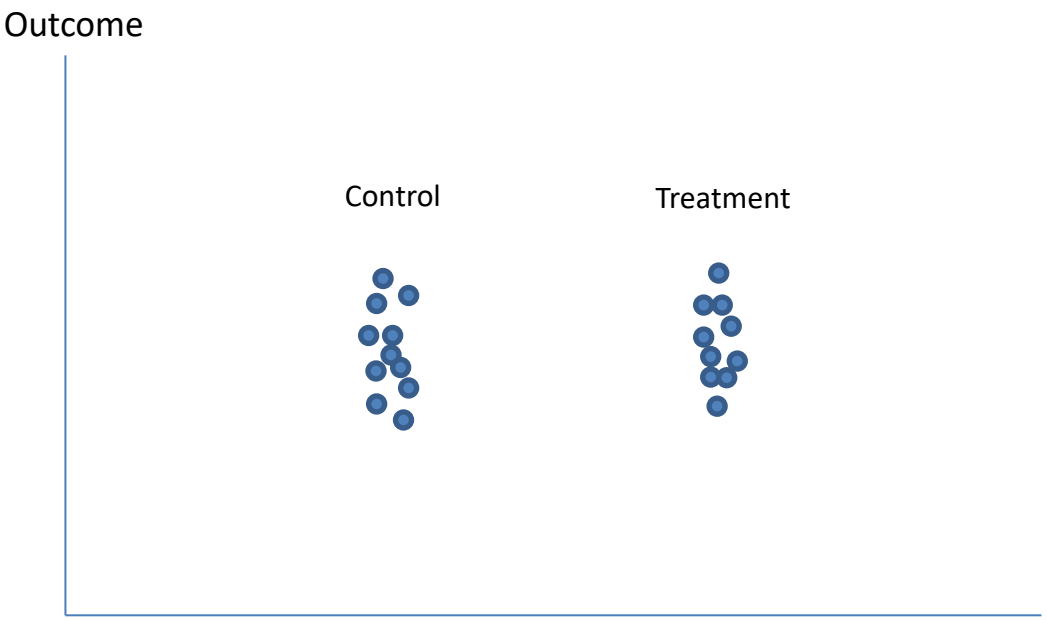
$$Mean(T2) - Mean(C2) = Program\ Effect$$

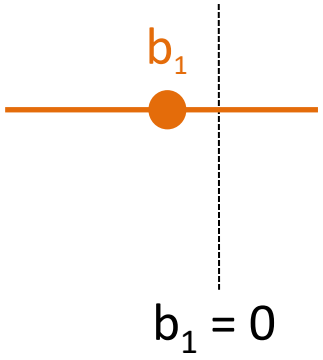


$b_1 = MEAN_{treat} - MEAN_{control}$

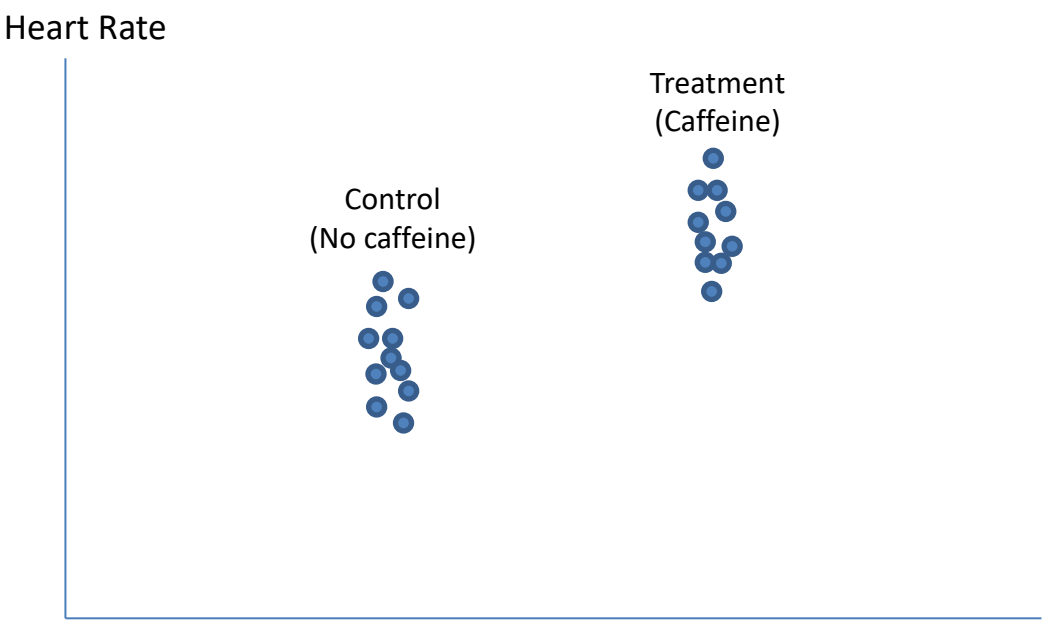
$b_1 = T2 - C2$

No Program Impact

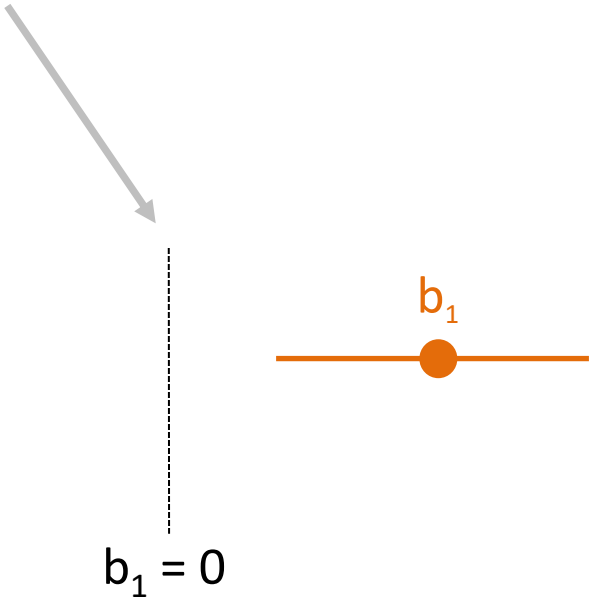




b_1 = $MEAN_{treat}$ - $MEAN_{control}$

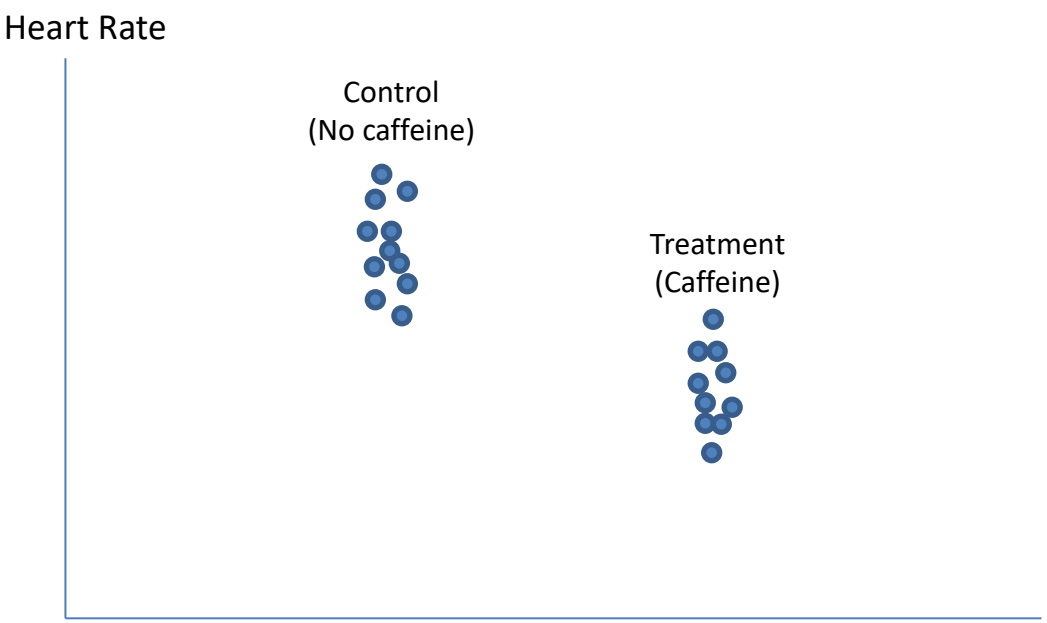


STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

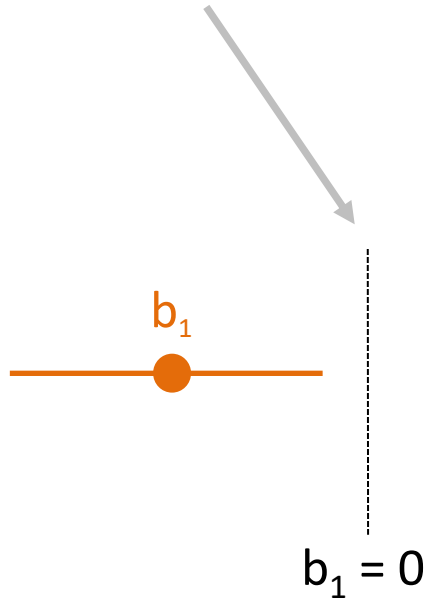


SIGNIFICANT
(POSITIVE PROGRAM IMPACT)

$b_1 = MEAN_{treat} - MEAN_{control}$



STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)



SIGNIFICANT
(NEGATIVE PROGRAM IMPACT)

Recall from Unit on Dummy Variable Models

(basic set-up for a comparison of group means in regression)

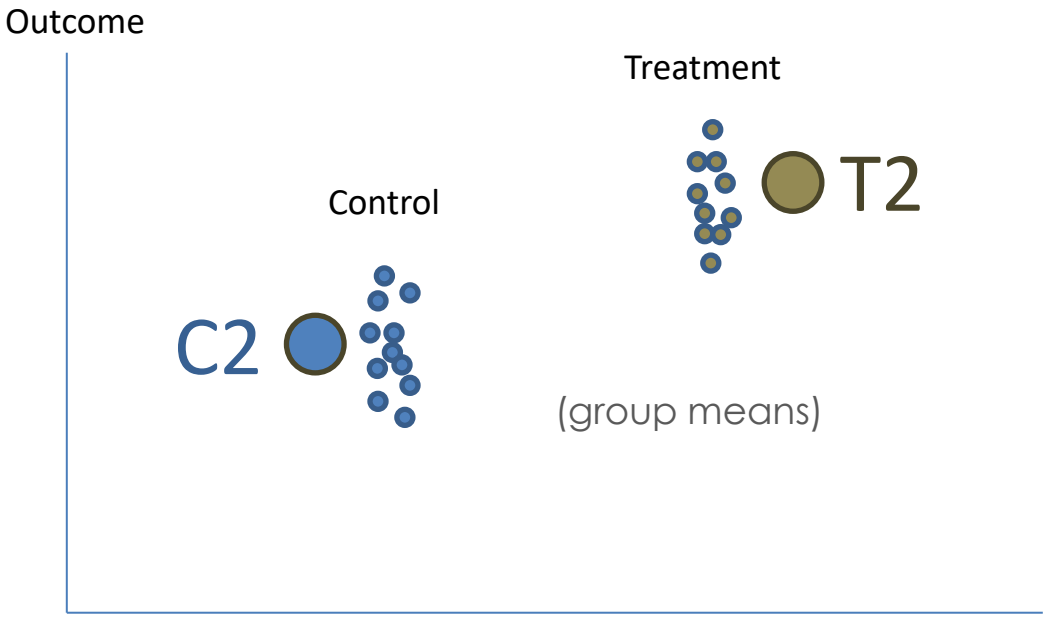
b_0 will measure Y-bar of the omitted group C2

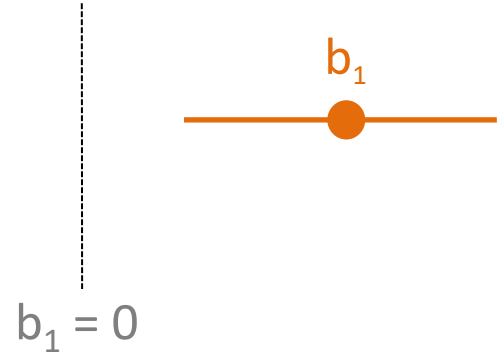
$b_0 + b_1$ represents the Y-bar of T2

$$Y = b_0 + b_1(\text{TreatDummy}) + e$$

$$b_1 = \text{MEAN}_{\text{treat}} - \text{MEAN}_{\text{control}}$$

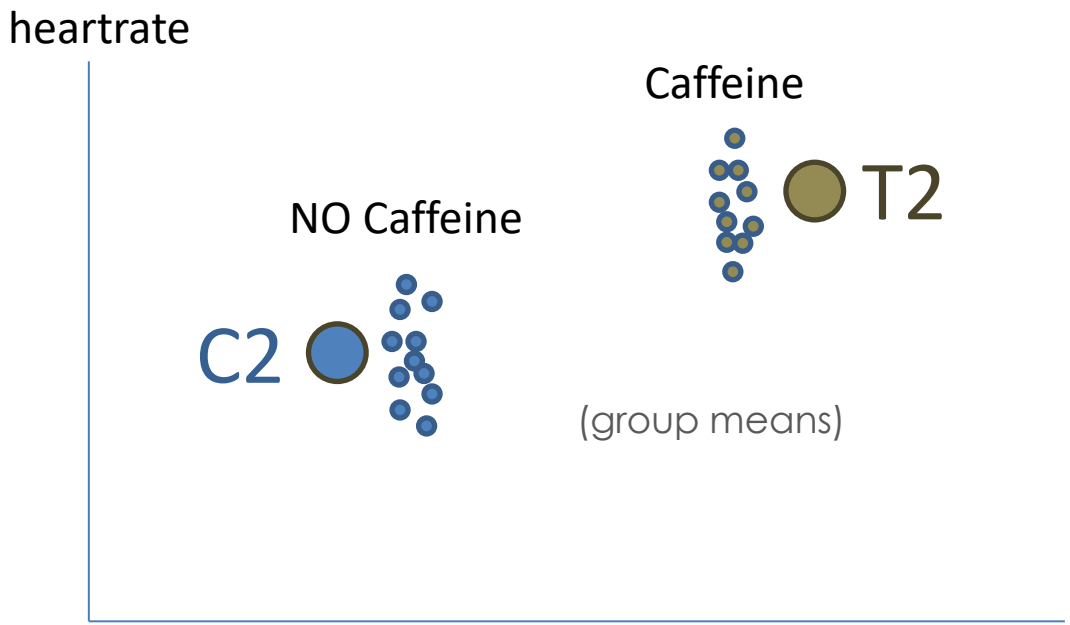
$$b_1 = \text{T2} - \text{C2}$$

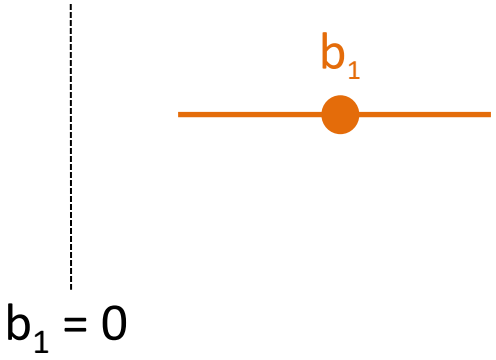




The default hypothesis test in the regression then uses b_1 to **test for a meaningful difference between T2 and C2**

$$b_1 = MEAN_{treat} - MEAN_{control}$$





WHEN ARE DIFFERENCES CAUSAL?

Does caffeine increase heart rate?

`caffeine` → `heart rate`

Or do people with high stress jobs and sleep deprivation tend to drink a lot of coffee?

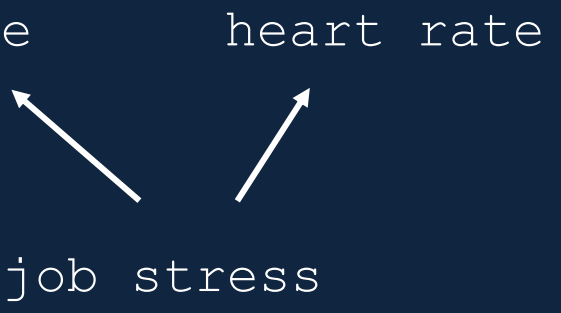
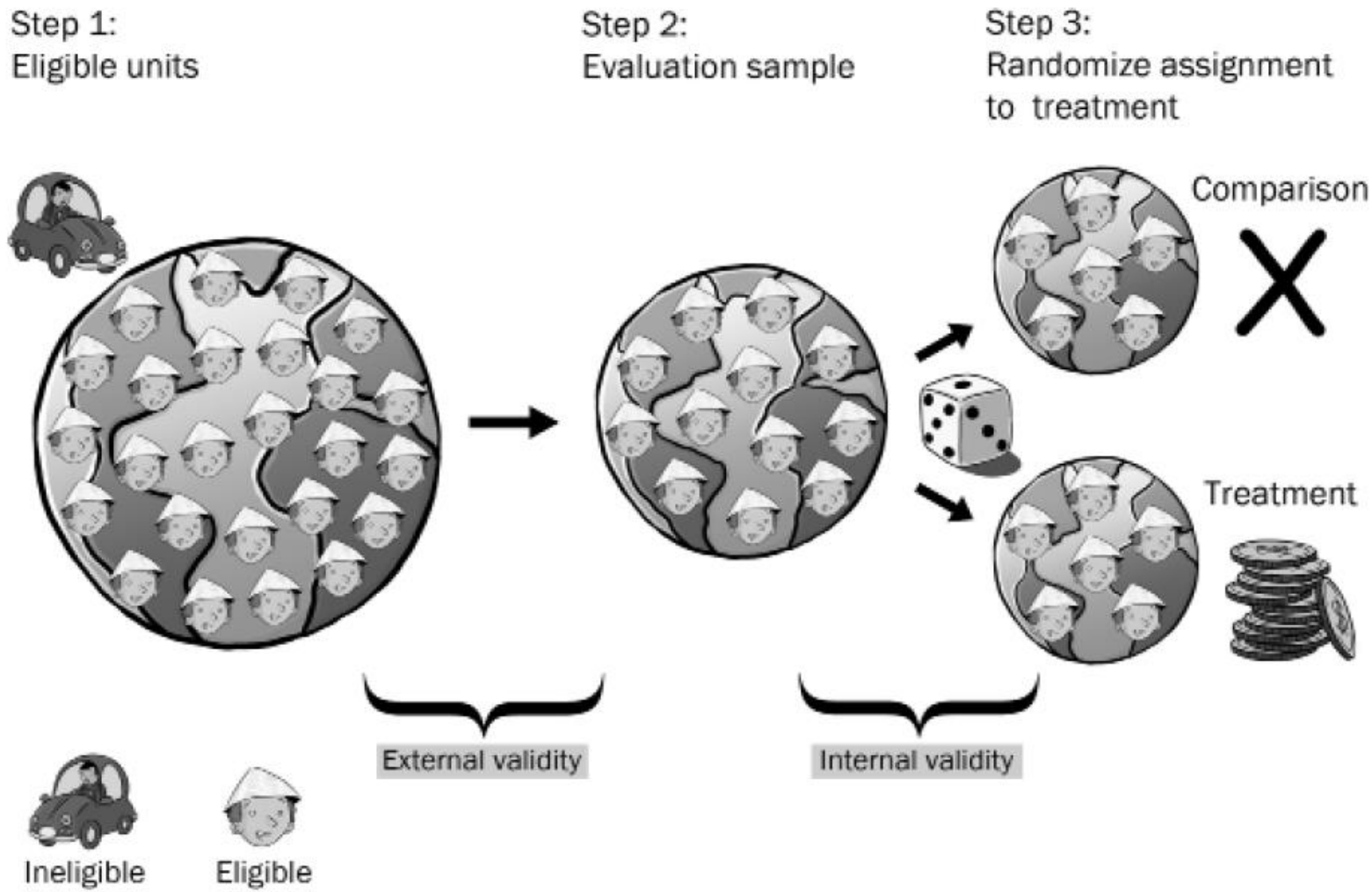
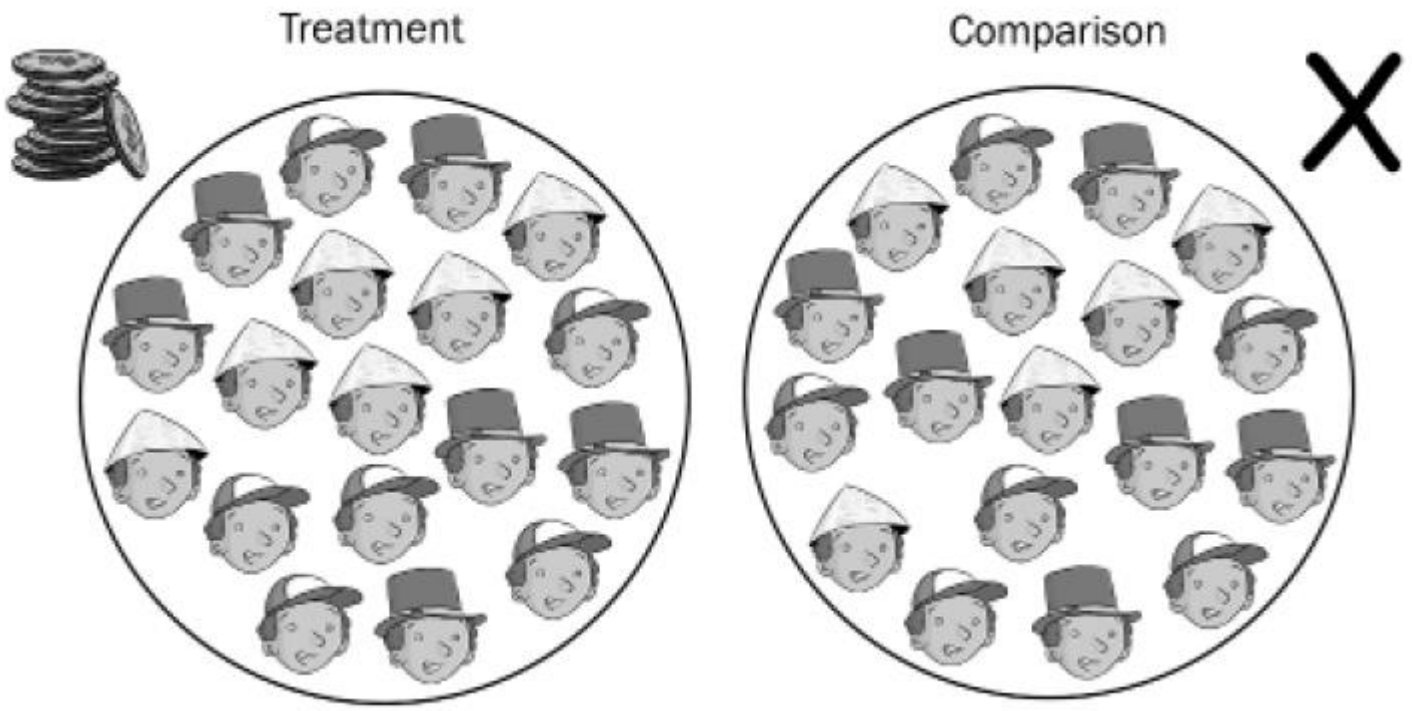


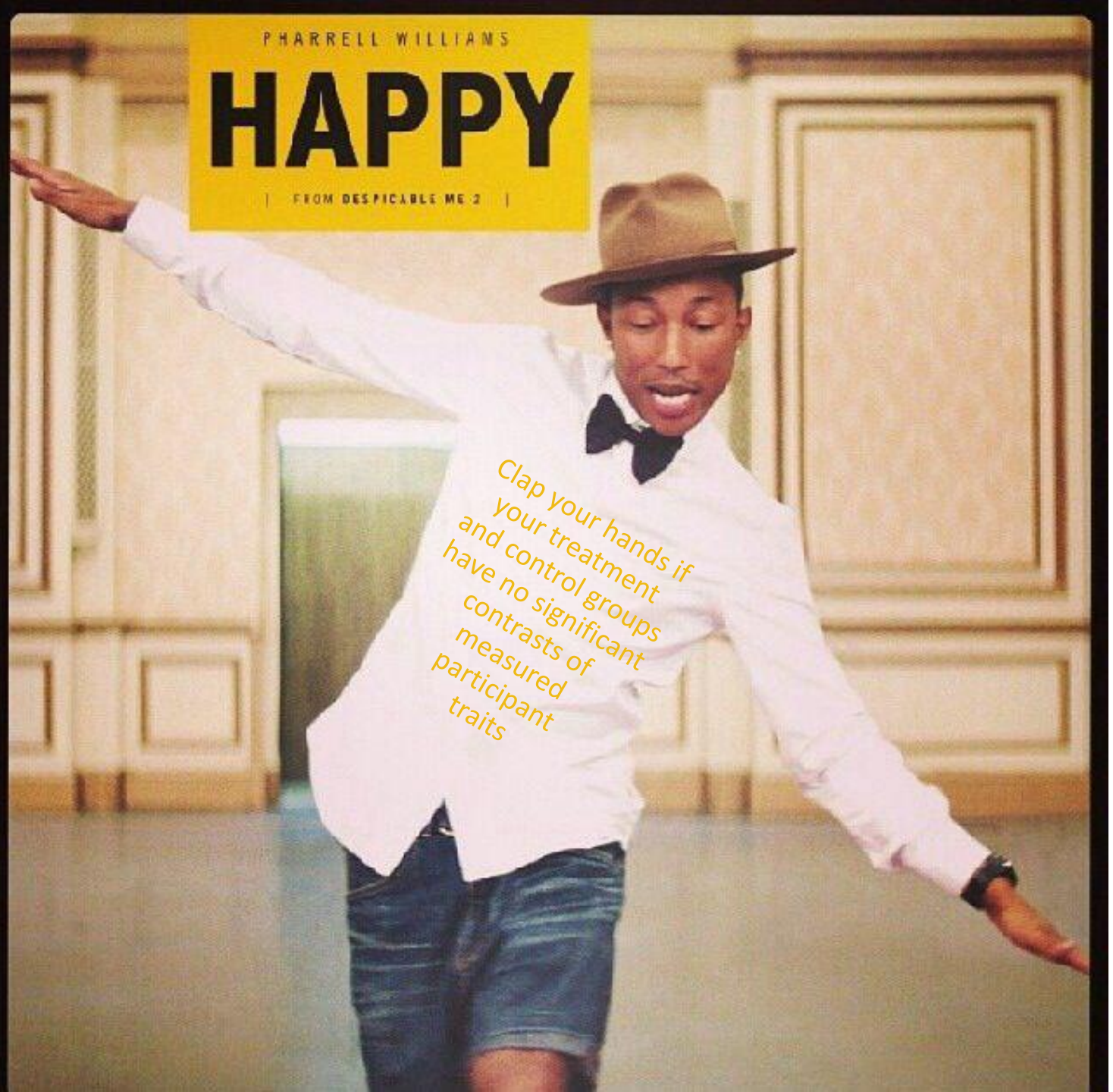
Figure 4.3 Steps in Randomized Assignment to Treatment



Our counterfactual framework is
valid / robust when the
groups only DIFFER BY THE TREATMENT
but are OTHERWISE "IDENTICAL"

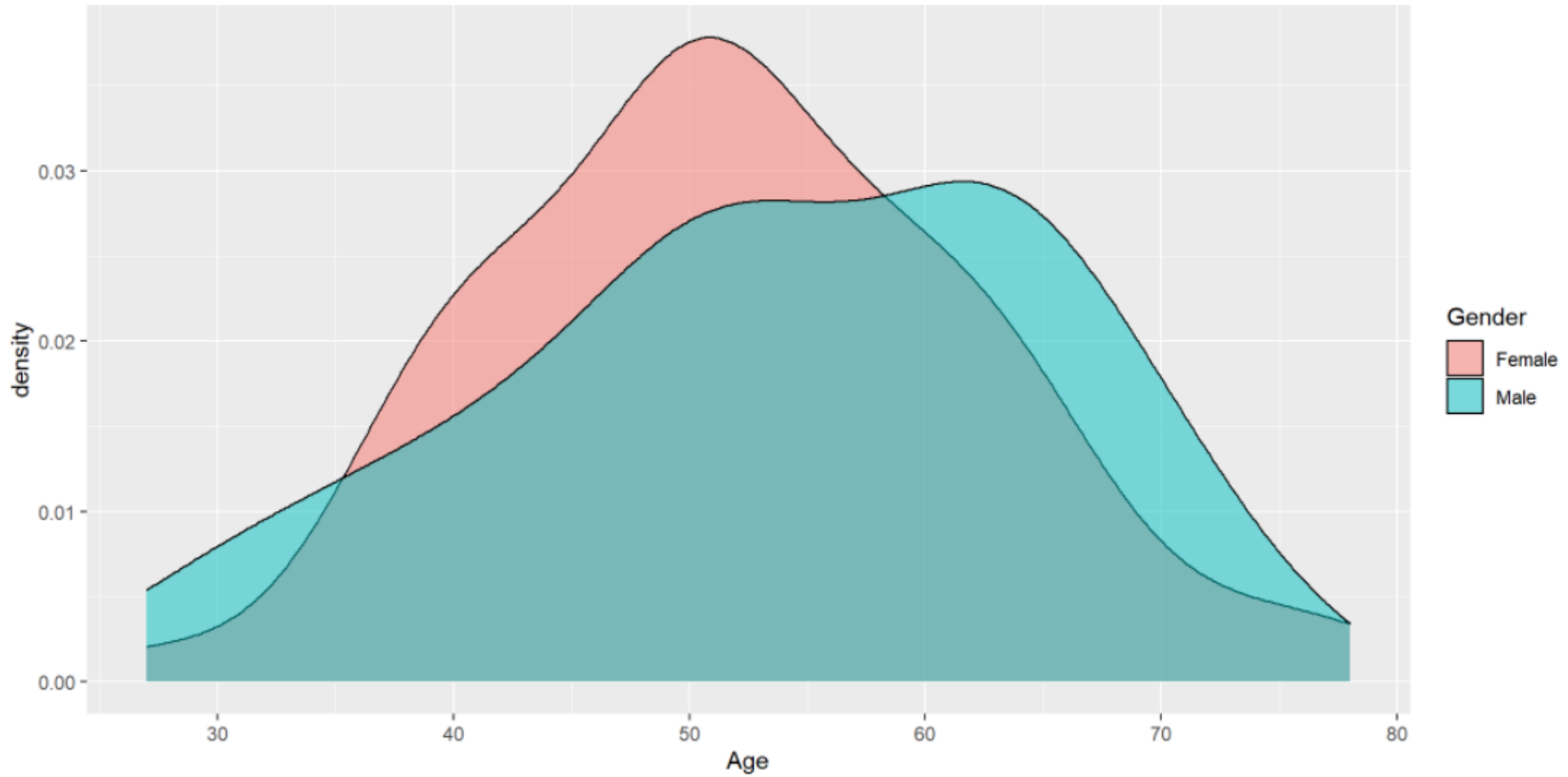


When true, we can interprett the differences in group outcomes
after the treatment period to be caused by the treatment



"HAPPY" RANDOMIZATION

Gender	min.age	median.age	mean.age	max.age
Female	27	51	52.05	78
Male	29	53	53.63	72



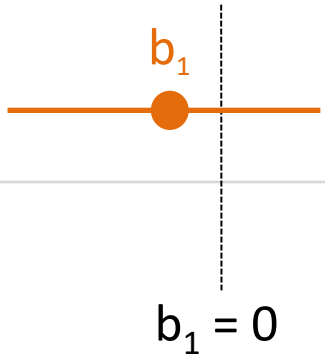
$$age = b_0 + b_1(gender) + e$$

$$b_1 = age_{MEN} - age_{WOMEN}$$

```
t.test( Age ~ Gender, data=d )
```

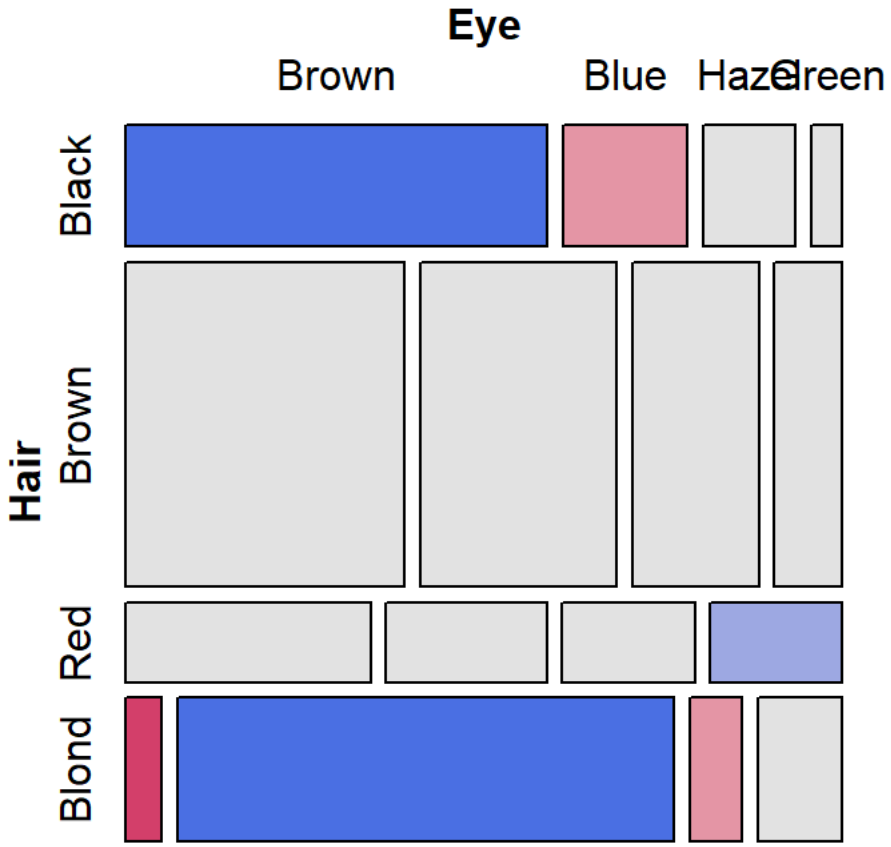
```
##
##  Welch Two Sample t-test
##
## data:  Age by Gender
## t = -0.69933, df = 78.271, p-value = 0.4864
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.090387  2.923789
## sample estimates:
## mean in group Female   mean in group Male
##             52.05085             53.63415
```

p-value > 0.05 so the study groups are NOT different



```
m <- table( hair.color, eye.color )
```

	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16



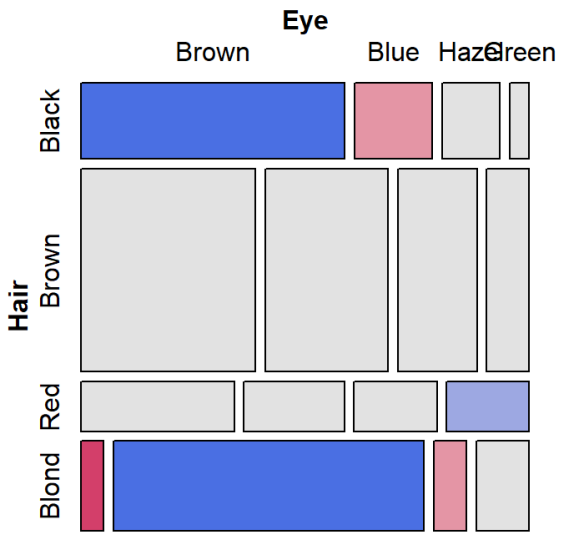
The chi-square test

The chi-square statistic provides a test for independence of two factors:

```
chisq.test( m )
```

```
##
##  Pearson's Chi-squared test
##
## data:  m
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```

p-value < 0.05 so the factors ARE dependent upon each other, i.e. hair color and eye color are "correlated"




```
table( study.group, f1 )
```

If we are testing for happy randomization or study group balance we want these to be independent – the proportions of each category in the factor f1 are approximately the same in the treatment and control groups

Bonferroni Correction:

When we want to be 95% confident that two groups are the same, and we can measure those groups using a set of contrasts, then our decision rule is no longer to reject the null (that the groups are the same) if the p-value < 0.05. A “contrast” is a comparison of means of any measured characteristic between two groups.

If we have a 5% chance of observing a p-value of less than 0.05 for each contrast, then the probability of observing at least one contrast with a p-value that small is greater than 5%! It is actually n*0.05 (minus prob of observing multiple < 0.05 at same time) where n is the number of contrasts.

So if we want to be 95% confident that the groups are different (not just the contrasts), we have to adjust our decision rule to α/n.

For example, if we have 10 contrasts, then our decision rule is now 0.05/10, or 0.005. The p-value of at least one contrast must be below 0.005 for us to conclude that the groups are different.

Table 4.1 Case 3—Balance between Treatment and Comparison Villages at Baseline

Household characteristics	Treatment villages (N = 2964)	Comparison villages (N = 2664)	Difference	*t*-stat
Health expenditures ($ yearly per capita)	14.48	14.57	−0.09	−0.39
Head of household's age (years)	41.6	42.3	−0.7	−1.2
Spouse's age (years)	36.8	36.8	0.0	0.38
Head of household's education (years)	2.9	2.8	0.1	2.16*
Spouse's education (years)	2.7	2.6	0.1	0.006
Head of household is female = 1	0.07	0.07	−0.0	−0.66
Indigenous = 1	0.42	0.42	0.0	0.21
Number of household members	5.7	5.7	0.0	1.21
Has bathroom = 1	0.57	0.56	0.01	1.04
Hectares of land	1.67	1.71	−0.04	−1.35
Distance to hospital (km)	109	106	3	1.02

"contrasts"

Source: Authors' calculation.

* Significant at the 5 percent level.

The most important table in every study: comparisons of treatment and control group characteristics

For the counterfactual to be **valid**, the groups can ONLY differ by the treatment, not by any measured traits.

Is this problematic?

What is the appropriate test for "identical" or equivalent groups?

We should observe no differences in measured traits.
Assume a 95% confidence interval.

Test for Group Equivalence

seven contrasts reported

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

Characteristic	All students in the study: Choice students	All students in the study: Control students	All students in the study: p value[a]	All students with scores three or four years after application: Choice students	All students with scores three or four years after application: Control students	All students with scores three or four years after application: p value[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Bonferroni Correction:

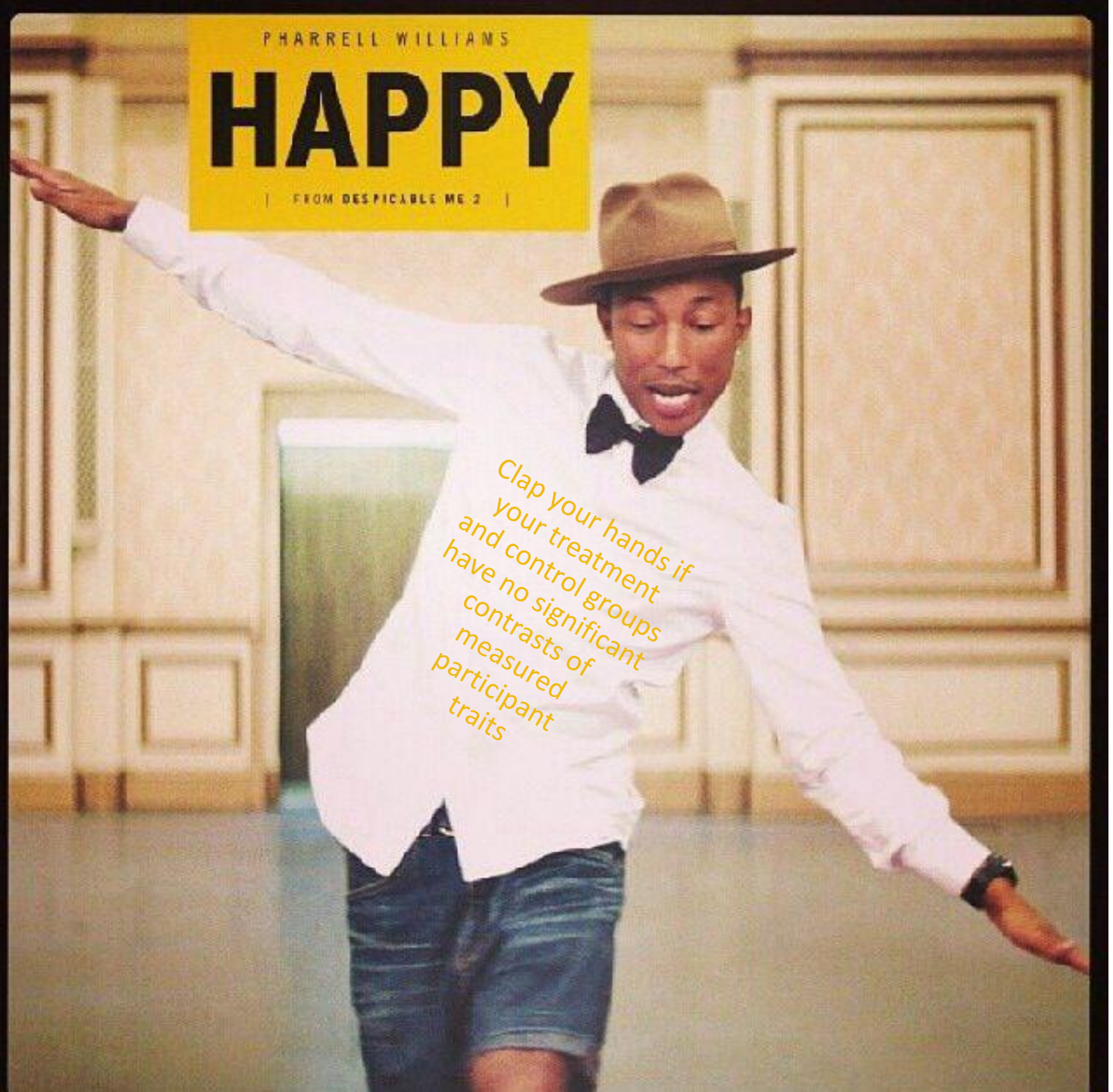
New alpha = (0.05 / 7) = 0.0071

Smallest p-value in table

0.04 > 0.0071

Do not reject :: Groups are equivalent

RANDOMIZATION VS STUDY GROUP EQUIVALENCE

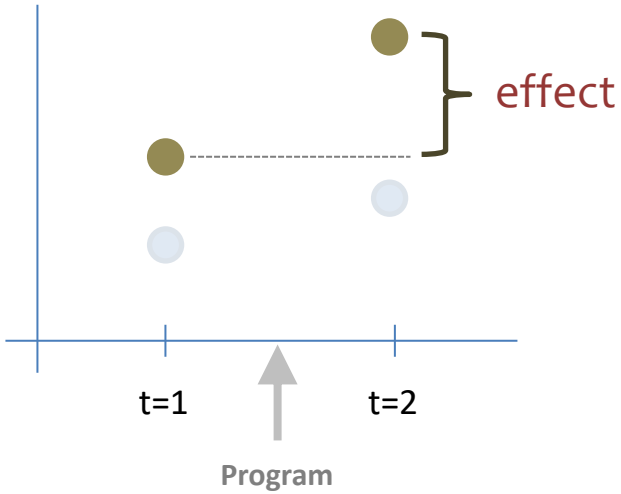


"HAPPY" RANDOMIZATION

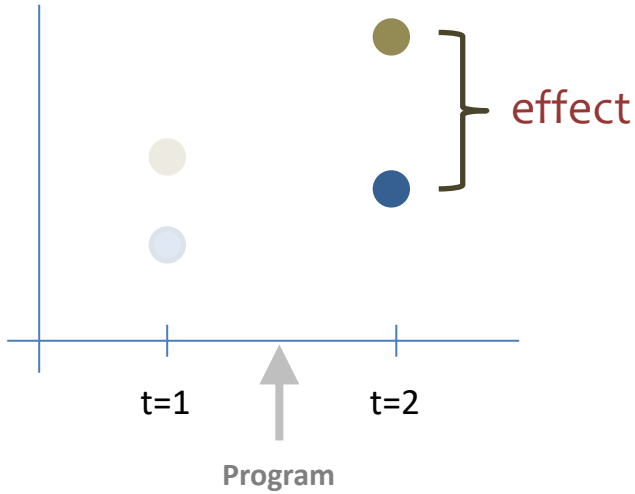
3 VALID COUNTERFACTUALS:

Each variety of counterfactual has a different formula for the program effect estimate.

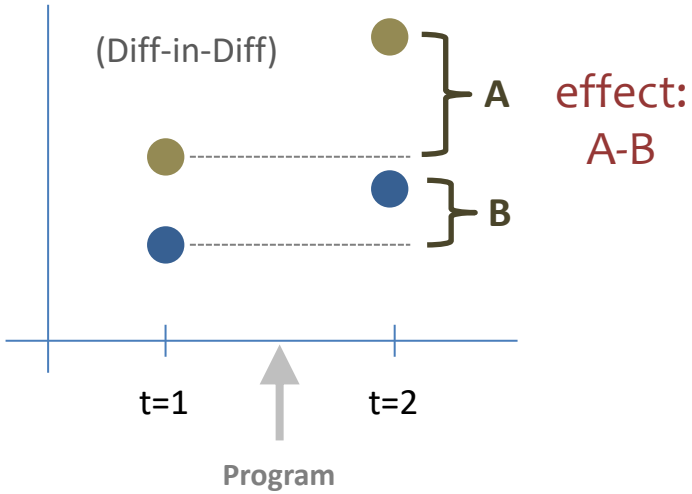
PRE-POST REFLEXIVE
Estimator
effect = T2-T1



POST-ONLY
Estimator
effect = T2-C2



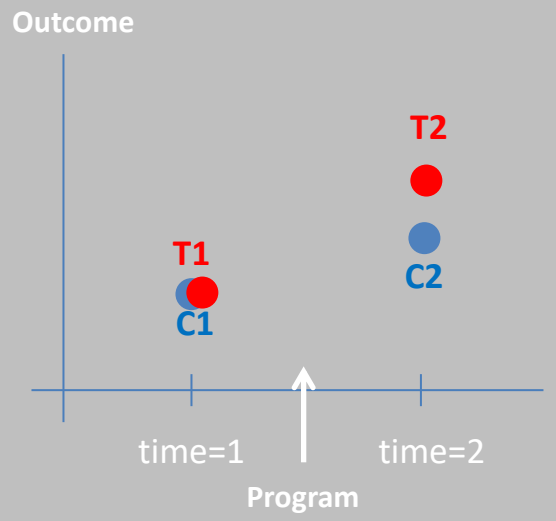
PRE-POST W COMPARISON
Estimator
effect = (T2-T1) – (C2-C1)

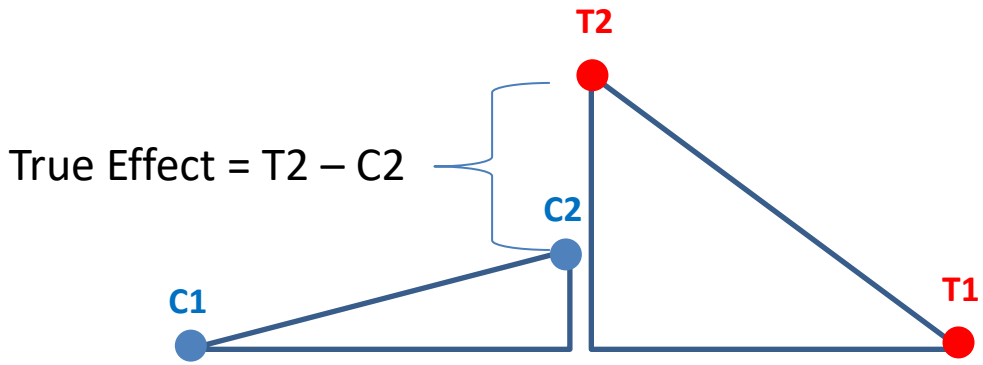


Treatment Groups, **T1**=before, **T2**=post-program measure

Control Groups, **C1**=before, **C2**=post-program measure

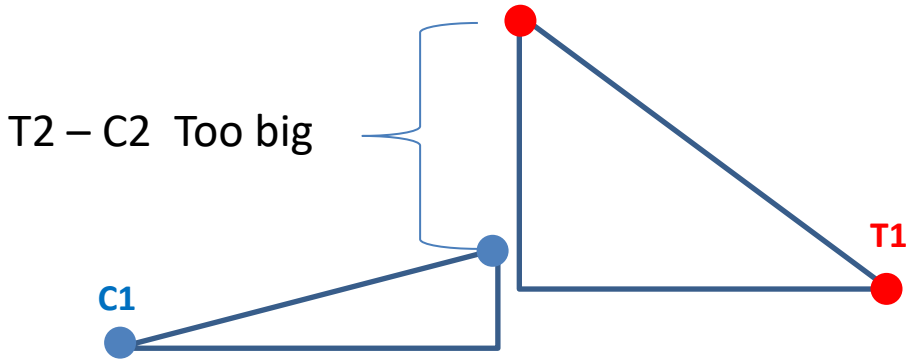
Post-Test Only Measure





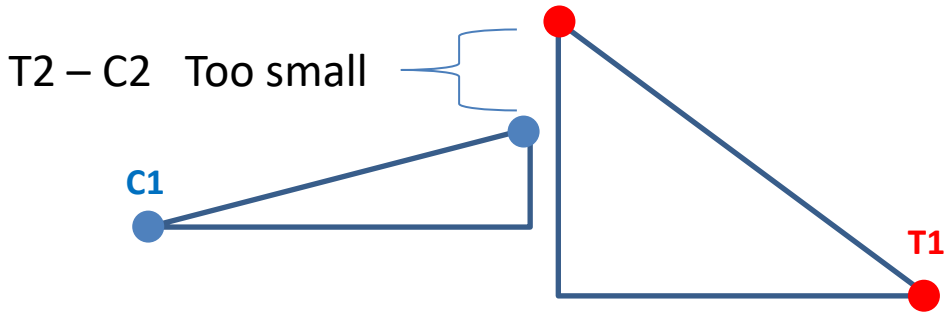
If C1 = T1

Measured effect accurately represents program impact



C1 < T1 biased

Measured effect overstates program impact



C1 > T1 biased

Measured effect understates program impact

Experiments

☑

Quasi-Experiments

☑

Observational Studies

☒

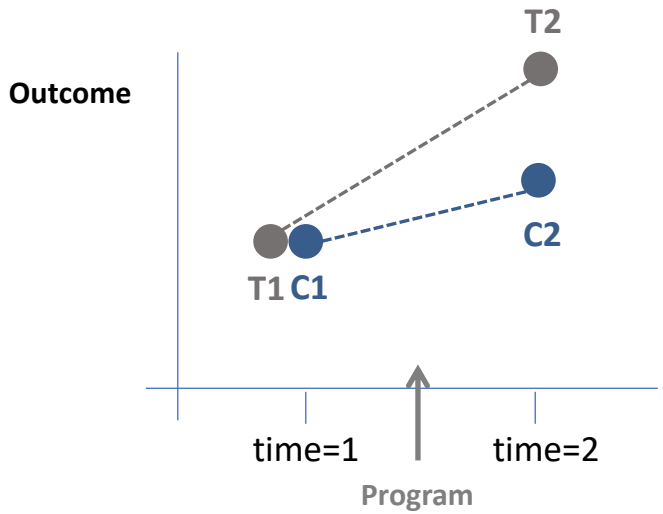
Validity of the post-test only estimator:

Post-Only

$(T2 - T1) - (C2 - C1) = T2 - C2$

IFF

$C1 - T1 = 0$
(equivalent at time 1)



If we have confidence that the two groups are identical prior to the treatment, then mathematically T2-C2 will still account for gains independent of the treatment. This condition is necessary for the post-test only estimator to be unbiased.

In experimental design, this is usually accomplished through randomization or lottery.

Observational students typically use matching models to create equivalent groups.