Lecture Notes: Causal Inference Fall 2020

Claire Duquennois

7/28/2020

Contents

1 The Experimental Ideal: Randomized Control Trials **1**
1.1 Random assignment and the selection problem 1
1.2 Key assumption and testing it's validity 4
1.3 Controls in RCT specifications 8
1.4 Heterogeneity 10
1.5 Alphabet Soup! 11
1.6 RCT Challenges 12
1.7 RCT critiques 14
1.8 Fishing for Stars 15
1.9 Research Transparency 20

1 The Experimental Ideal: Randomized Control Trials

It should be clear by now that achieving coefficient estimates that can be interpreted as causal effects is HARD! Even with rich datasets and powerful statistical tools such as fixed effect and instrumental variables, omitted unobserved variables and hidden correlations can still generate concerning bias in our estimates.

So where do we go from here? In this section we will cover what is sometimes referred to as the "Gold standard" of experimental designs for causal inference: Randomized control trials (RCT's).

1.1 Random assignment and the selection problem

The idea behind RCT designs is to use random assignment into treatment to solve the problem of selection bias based on unobservables.

Revisiting what we discussed in the section on conditional independence, suppose the treatment effect is the same for everyone so that $Y_i(1) - Y_i(0) = \tau$, a constant. If this is the case, for observation i we have that

$$\begin{aligned} Y_i &= Y_i(0) + \tau D_i \\ Y_i &= E[Y_i(0)] + \tau D_i + Y_i(0) - E[Y_i(0)] \\ Y_i &= \alpha + \tau D_i + \eta_i \end{aligned} \tag{1}$$

where $\alpha = E[Y_i(0)]$, $\tau = Y_i(1) - Y_i(0)$, and η_i is the random part of $Y_i(0)$ since $\eta_i = Y_i(0) - E[Y_i(0)]$. We can then see that the expected outcomes for someone with treatment ($D_i = 1$), and without treatment ($D_i = 0$) are given by

$$\begin{aligned} E[Y_i(1)] &= \alpha + \tau + E[\eta_i | D_i = 1] \\ E[Y_i(0)] &= \alpha + E[\eta_i | D_i = 0] \end{aligned} \tag{2}$$

so that we can break down the difference between these outcomes as

$$E[Y_i(1)] - E[Y_i(0)] = \underbrace{\tau}_{\text{treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{selection bias}}. \tag{3}$$

It is thus clear that selection bias will bias our estimate of τ if those who would select into treatment have a different expected outcome compared to those who would not select into treatment, such that $E[Y_i(0)|D_i = 1] \neq E[Y_i(0)|D_i = 0]$. For instance, returning to our blood pressure medication example, if people with high blood pressure are more likely to take blood pressure medication, we have that $E[Y_i(0)|D_i = 1] > E[Y_i(0)|D_i = 0]$ which would lead a naive estimate of the treatment effect to underestimate the effect of the drug.

This is because treatment is not random, that is the outcome is not independent of the treatment status: $\{Y_i(1), Y_i(0)\} \not\perp D_i$. Since subjects select into treatment, there is no reason to believe that those who select into getting treated would have the same expected outcome as those who did not make that selection, if they were to be treated, that is to say, it is possible (and even likely) that

$$\underbrace{E[Y_i(0)|D_i = 0]}_{\text{observed}} \neq \underbrace{E[Y_i(0)|D_i = 1]}_{\text{unobserved}} \neq E[Y_i(0)]$$

and

$$\underbrace{E[Y_i(1)|D_i = 0]}_{\text{unobserved}} \neq \underbrace{E[Y_i(1)|D_i = 1]}_{\text{observed}} \neq E[Y_i(1)]$$

While the conditional independence assumption allows us to control for selection bias by conditioning on observed characteristics, we now know that there are often important unobserved characteristics that we cannot control for that will also bias our estimates.

Random assignment solves all of these selection bias problems.

Random assignment of D_i solves the selection problem because it makes D_i independent of potential outcomes. Formally, with random assignment, treatment status,D_i, is orthogonal to potential outcomes,

$$\{Y_i(1), Y_i(0)\} \perp D_i.$$

What this means is that with random assignment, though we still do not observe the potential outcomes of those who received treatment had they not been treatment and those left untreated had they been treated, we know that in expectation,

$$\underbrace{E[Y_i(0)|D_i = 0]}_{\text{observed}} = \underbrace{E[Y_i(0)|D_i = 1]}_{\text{unobserved}} = E[Y_i(0)]$$

and

$$\underbrace{E[Y_i(1)|D_i = 0]}_{\text{unobserved}} = \underbrace{E[Y_i(1)|D_i = 1]}_{\text{observed}} = E[Y_i(1)]$$

When this is the case, the average causal effect, $\bar{\tau}$, is thus

$$\bar{\tau} = E[Y_i(1)] - E[Y_i(0)] = \underbrace{E[Y_i(1)|D_i = 1]}_{\text{observed}} - \underbrace{E[Y_i(0)|D_i = 0]}_{\text{observed}} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0].$$

We can easily estimate $\bar{\tau}$ by taking the difference between the average value of Y_i in the treatment group and the average value of Y_i in the control group. Because it allows estimation of the **Average Treatment Effect (ATE)**, the randomized control trial is considered the "gold standard" of evidence in medicine, and in many areas of social science as well.

In terms of empirics, the basics of running an RCT regression are about as straightforward as it gets. As modeled above, you can estimate

$$Y_i = \alpha + \tau D_i + \eta_i$$

where $\alpha = E[Y_i(0)]$, $\tau = Y_i(1) - Y_i(0)$, and η_i is the random error term.[1] The treatment effect will be given by

$$E[Y_i(1)] - E[Y_i(0)] = \underbrace{\tau}_{\text{treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{selection bias}}. \tag{4}$$

As long as treatment was properly randomized, $E[\eta_i|D_i = 1] = E[\eta_i|D_i = 0]$ and there will be no selection bias giving us an unbiased estimate of τ.

1.1.1 Simulation

Suppose I am a principal of a large school and I am interested in determining how access to small reading groups with a paraprofessional helps improve 4th grade test scores. I decide to take all of the 4th graders in my school and randomly assign 30 percent of them to treatment (participating in the reading groups) with the remainder assigned to the control group which continued with class as normal.

As before, I generate a set of simulated data with a data generating process that we fully understand

```
library(MASS)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
set.seed(1999)

scores5 <- as.data.frame(rep(c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10), times = 30))
names(scores5) <- c("class")
scores5 <- fastDummies::dummy_cols(scores5, select_columns = "class")

scores5$error <- rnorm(300, mean = 0, sd = 10)

# treatment indicator
scores5$treat <- rbinom(300, 1, 0.3)



# mean reading score
alpha = 75

# treatment effect
tau = 10
```

[1] Note: As you can see, the treatment effect is really just the difference in means between the treated and control groups. You could calculate it without running a regression. Using the regression format however is convenient as it gives you standard errors, allows you to add controls, and look at heterogeneity in effects.

```r
# the data generating process: notice the class does affect a students score
scores5$read4 <- alpha + tau * scores5$treat + scores5$error + 4 * scores5$class_1 +
    (-6) * scores5$class_2 + 8 * scores5$class_3 + (-4) * scores5$class_4 + 7 * scores5$class_5 +
    (-2) * scores5$class_6 + 5 * scores5$class_7 + (-10) * scores5$class_8 + 8 *
    scores5$class_9 + 4 * scores5$class_10

rct1 <- felm(read4 ~ treat, scores5)

stargazer(rct1, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

Table 1:

	Dependent variable:
	read4
treat	11.229^{***}
	(1.442)
Constant	77.150^{***}
	(0.750)
Observations	300
R^2	0.169
Adjusted R^2	0.166
Residual Std. Error	11.092 (df = 298)
Note:	*p<0.1; **p<0.05; ***p<0.01

As you can see, even though the class the student is in does affect their score, because the treatment was randomized, this simple estimation strategy allows us to recover an unbiased estimate of the true treatment effect ($\tau = 10$).

1.2 Key assumption and testing it's validity

The key assumption for $\hat{\tau}$ to be an unbiased estimator of τ is that

$$E[\eta_i | D_i = 1] = E[\eta_i | D_i = 0] = 0.$$

Though we cannot test this assumption directly, we can check to see if observable characteristics among treatment and control groups are the same on average. Note that when doing this type of **balance test**, you want to make sure you either check the value of these characteristics at baseline, prior to treatment, or you check the value of characteristics that would be unaffected by treatment. A balance test can be presented as a table of the following regressions

$$X_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

where X_i is a vector of characteristics being tested. More commonly, balance tables are often presented as simple t-test tables testing the difference in means between the treatment and control groups.[2]

It is worth noting that balance tests are often run on many variables. Because of this, it is not surprising if one or two come up with a statistically significant difference. Probability would predict that if you do a

[2]These two approached are statistically equivalent, this is simply a difference in presentation choice.

balance test on 20 variables, it would be unsurprising for two to come out significant at the 10 percent level and/or one at the 5 percent level just by random chance.[3]

1.2.1 Simulation

Suppose the principal is concerned that there were some problems with the randomization. She has access to some additional data (which I will simulate below). She adds it to her data set and does a balance test.

```
# smulating covariates

# third grade test scores. Notice I am generateing simulated academic scores
# that have a correlation to their 'untreated' performance in 4th grade reading
scores5$read3 <- alpha + scores5$error + rnorm(300, 3, 2)
scores5$math3 <- alpha + scores5$error + rnorm(300, 15, 2)
scores5$hist3 <- alpha + scores5$error + rnorm(300, 5, 2)
scores5$pe3 <- rnorm(300, 90, 2)

# other 4th grade test scores: notice I am generating scores that correlated
# with their subject performance in 3rd grade. Also, the treatment is affecting
# other 4th grade academic scores
scores5$hist4 <- 4 * scores5$treat + scores5$hist3 + rnorm(300, -2, 2)
scores5$pe4 <- scores5$pe3 + rnorm(300, 0, 5)
scores5$math4 <- 2 * scores5$treat + scores5$math3 + rnorm(300, -5, 3)

# student characteristics
scores5$female <- rbinom(300, 1, 0.5)
scores5$age <- runif(300, 9, 10)
scores5$height <- rnorm(300, 1.3, 0.2)

scoresmini <- scores5[, c("treat", "read4", "read3", "math3", "hist3", "pe3", "hist4",
    "pe4", "math4", "female", "age", "height")]

cor(scoresmini)
```

```
##              treat        read4       read3        math3        hist3
## treat   1.000000000  0.411103370  0.06210422  0.043810660  0.048985855
## read4   0.411103370  1.000000000  0.76370646  0.756595945  0.767164090
## read3   0.062104220  0.763706465  1.00000000  0.954947392  0.957691240
## math3   0.043810660  0.756595945  0.95494739  1.000000000  0.951163232
## hist3   0.048985855  0.767164090  0.95769124  0.951163232  1.000000000
## pe3    -0.134794987 -0.078928737  0.01977274  0.025802541  0.002345894
## hist4   0.210104738  0.799439890  0.92717436  0.910566252  0.965319194
## pe4    -0.043975444  0.001522679  0.06059868  0.067988360  0.044767865
## math4   0.140564617  0.764954614  0.92332515  0.951489208  0.917393931
## female  0.003009974 -0.019753464  0.01298834  0.002850811 -0.014423484
## age     0.069344630 -0.046119854 -0.09536580 -0.096909858 -0.103607528
## height -0.049395136  0.055676186  0.04243414  0.032683477  0.040917764
##                 pe3       hist4          pe4       math4       female
## treat  -0.134794987  0.21010474 -0.043975444  0.14056462  0.003009974
## read4  -0.078928737  0.79943989  0.001522679  0.76495461 -0.019753464
## read3   0.019772739  0.92717436  0.060598681  0.92332515  0.012988341
## math3   0.025802541  0.91056625  0.067988360  0.95148921  0.002850811
```

[3]There are corrections that can be implemented if the unbalanced variables are are of particular concern (look up Bonferroni correction).

```
## hist3   0.002345894  0.96531919  0.044767865  0.91739393 -0.014423484
## pe3     1.000000000 -0.02968731  0.472814327  0.02354280  0.026070397
## hist4  -0.029687311  1.00000000  0.017628038  0.89351172 -0.017902077
## pe4     0.472814327  0.01762804  1.000000000  0.05519176 -0.013560064
## math4   0.023542799  0.89351172  0.055191759  1.00000000  0.034182940
## female  0.026070397 -0.01790208 -0.013560064  0.03418294  1.000000000
## age    -0.045765250 -0.08045850 -0.036665854 -0.11731677 -0.070468503
## height  0.046935386  0.03323997 -0.019725022  0.01223729 -0.079529570
##                age      height
## treat   0.06934463 -0.04939514
## read4  -0.04611985  0.05567619
## read3  -0.09536580  0.04243414
## math3  -0.09690986  0.03268348
## hist3  -0.10360753  0.04091776
## pe3    -0.04576525  0.04693539
## hist4  -0.08045850  0.03323997
## pe4    -0.03666585 -0.01972502
## math4  -0.11731677  0.01223729
## female -0.07046850 -0.07952957
## age     1.00000000  0.01430370
## height  0.01430370  1.00000000
```

```
# as you can see, we have simulated some complex interrelationships between
# theses variables.

# Balance test: I generate a loop to run all the covariate regressions.

namevec <- names(scores5)
namevec <- namevec[!namevec %in% c("class", "error", "treat", "read4")]
```

```
allModelsList <- lapply(paste(namevec, "~treat"), as.formula)

allModelsResults <- lapply(allModelsList, function(x) lm(x, scores5))


stargazer(allModelsResults[[1]], allModelsResults[[2]], allModelsResults[[3]], allModelsResults[[4]],
    allModelsResults[[5]], type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

```
stargazer(allModelsResults[[6]], allModelsResults[[7]], allModelsResults[[8]], allModelsResults[[9]],
    allModelsResults[[10]], type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

```
stargazer(allModelsResults[[11]], allModelsResults[[12]], allModelsResults[[13]],
    allModelsResults[[14]], allModelsResults[[15]], type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

```
stargazer(allModelsResults[[16]], allModelsResults[[17]], allModelsResults[[18]],
    allModelsResults[[19]], allModelsResults[[20]], type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu

Table 2:

	Dependent variable:				
	class_1	class_2	class_3	class_4	class_5
	(1)	(2)	(3)	(4)	(5)
treat	−0.019	0.049	−0.052	−0.052	−0.036
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
Constant	0.105***	0.087***	0.114***	0.114***	0.110***
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Observations	300	300	300	300	300
R^2	0.001	0.005	0.006	0.006	0.003
Adjusted R^2	−0.003	0.002	0.003	0.003	−0.001
Residual Std. Error (df = 298)	0.301	0.300	0.300	0.300	0.301
F Statistic (df = 1; 298)	0.226	1.578	1.805	1.805	0.825

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3:

	Dependent variable:				
	class_6	class_7	class_8	class_9	class_10
	(1)	(2)	(3)	(4)	(5)
treat	0.066*	0.015	−0.052	0.015	0.066*
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
Constant	0.082***	0.096***	0.114***	0.096***	0.082***
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Observations	300	300	300	300	300
R^2	0.010	0.001	0.006	0.001	0.010
Adjusted R^2	0.006	−0.003	0.003	−0.003	0.006
Residual Std. Error (df = 298)	0.300	0.301	0.300	0.301	0.300
F Statistic (df = 1; 298)	2.866*	0.151	1.805	0.151	2.866*

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4:

	Dependent variable:				
	read3	math3	hist3	pe3	hist4
	(1)	(2)	(3)	(4)	(5)
treat	1.329	0.934	1.038	−0.613**	4.672***
	(1.237)	(1.233)	(1.226)	(0.261)	(1.260)
Constant	78.760***	90.965***	80.911***	90.044***	79.153***
	(0.643)	(0.641)	(0.637)	(0.136)	(0.654)
Observations	300	300	300	300	300
R^2	0.004	0.002	0.002	0.018	0.044
Adjusted R^2	0.001	−0.001	−0.001	0.015	0.041
Residual Std. Error (df = 298)	9.511	9.483	9.425	2.006	9.685
F Statistic (df = 1; 298)	1.154	0.573	0.717	5.515**	13.762***
Note:					*p<0.1; **p<0.05; ***p<0.01

% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

Table 5:

	Dependent variable:				
	pe4	math4	female	age	height
	(1)	(2)	(3)	(4)	(5)
treat	−0.548	3.122**	0.003	0.043	−0.022
	(0.722)	(1.274)	(0.065)	(0.036)	(0.026)
Constant	89.953***	86.029***	0.466***	9.490***	1.295***
	(0.375)	(0.662)	(0.034)	(0.019)	(0.013)
Observations	300	300	300	300	300
R^2	0.002	0.020	0.00001	0.005	0.002
Adjusted R^2	−0.001	0.016	−0.003	0.001	−0.001
Residual Std. Error (df = 298)	5.549	9.794	0.501	0.279	0.198
F Statistic (df = 1; 298)	0.577	6.007**	0.003	1.440	0.729
Note:					*p<0.1; **p<0.05; ***p<0.01

Notice that some of the variables you included in the balance test are problematic. Math and history scores in 4th grade could have potentially been affected by the treatment as well (and indeed they come out as statistically significant because our simulation modeled that they would be affected). These types of variables should not be included in a balance table since they are themselves outcomes of the treatment. The other variables seem quite balanced. Note that a few come out as statistically significant: class_6 and class_10 at 10%, and pe_3 at 5%. This is the result of random chance as discussed above (we know this for certain since we modeled the data). If you had not modeled the data, the fact that pe_3 was determined prior to treatment, and is not generally a variable we would expect to correlated with reading scores, should reassure you that it is the result of random chance. Class_6 and Class_10 would be more concerning since it might signal that some teachers were better able to get their students into the small groups but the coefficients are not large, nor are they highly significant which should reassure you that they are the result of random chance.

One way to see this, because we are working with simulated data is to change the seed in the simulation. Change the seed in the simulation (try 5000 for example) and you will find that some other variables will likely be significant due to random chance.

1.3 Controls in RCT specifications

Because the treatment was randomized, estimating

$$Y_i = \alpha + \tau D_i + \epsilon$$

gives us an unbiased estimate of τ without having to worry about controlling for omitted variables. That said, it is common to see specifications in RCT projects that include a vector of control variables. One reason to add controls is to verify that our estimated coefficient does not change significantly when controls are added. Indeed, if it did, this would suggest that treatment was correlated with one of those controls which would be concerning in the context of an RCT. Secondly, adding controls can make our estimated more precise as they will shrink the standard errors.

```
rct1 <- felm(read4 ~ treat, scores5)
rct2 <- felm(read4 ~ treat + read3 + female + pe3 + math3 + hist3, scores5)
rct3 <- felm(read4 ~ treat + read3 + female + pe3 + math3 + hist3 | class, scores5)

stargazer(rct1, rct2, rct3, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:04:59 PM

Why does adding control variables add precision? Think about the formula for the variance/standard error of our estimator:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x(1-R_j^2)}$$

$$se(\beta_1) = \frac{\hat{\sigma}}{\sqrt{SST_x(1-R_j^2)}}$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_i^n \hat{u}_i^2$$

If we include more x's in our regression, we can reduce $\hat{u}_1^2$, i.e. the unexplained variation in Y goes down , which means $se(\hat{\beta}_j)$ decreases which means our $\hat{\beta}$ can be estimated more precisely.

1.4 Heterogeneity

You may believe that the treatment you are investigating may affect certain individuals and subgroups of the population more than others. We can measure heterogeneity of the program effects for individuals with specific characteristics by interacting these characteristics with the treatment variable.

In the regression framework, we simply add the relevant interaction terms,

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 D_i \times x_i + \epsilon,$$

where x_i could represent an indicator variable for being female, then β_3 gives us the differential effect of the treatment for females relative to non-females.

In our simulation, our DGP did not model any heterogenous effects. Below, I start by searching for heterogeneity by gender using our existing simulation data. I then generate two new sets of 4th grade reading scores with a DGP that model heterogenous effects.

Table 6:

	Dependent variable:		
	read4		
	(1)	(2)	(3)
treat	11.229***	10.028***	10.152***
	(1.442)	(0.824)	(0.153)
read3		0.227	0.319***
		(0.153)	(0.028)
female		−0.426	0.120
		(0.727)	(0.135)
pe3		−0.235	0.074**
		(0.181)	(0.034)
math3		0.287**	0.339***
		(0.143)	(0.026)
hist3		0.472***	0.339***
		(0.149)	(0.027)
Constant	77.150***	16.382	
	(0.750)	(16.606)	
Observations	300	300	300
R^2	0.169	0.740	0.992
Adjusted R^2	0.166	0.735	0.991
Residual Std. Error	11.092 (df = 298)	6.255 (df = 293)	1.133 (df = 284)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
rct1 <- felm(read4 ~ treat, scores5)
rcthet1 <- felm(read4 ~ treat + female + female * treat, scores5)

nf <- 20

# the data generating process: notice the class does affect a students score
scores5$read4het1 <- (alpha + nf * scores5$treat + scores5$error + 4 * scores5$class_1 +
    (-6) * scores5$class_2 + 8 * scores5$class_3 + (-4) * scores5$class_4 + 7 * scores5$class_5 +
    (-2) * scores5$class_6 + 5 * scores5$class_7 + (-10) * scores5$class_8 + 8 *
    scores5$class_9 + 4 * scores5$class_10 + (-20) * scores5$female * scores5$treat)

rct2 <- felm(read4het1 ~ treat, scores5)
rcthet2 <- felm(read4het1 ~ treat + female + female * treat, scores5)

nf2 <- 30

scores5$read4het2 <- (alpha + nf2 * scores5$treat + scores5$error + 4 * scores5$class_1 +
    (-6) * scores5$class_2 + 8 * scores5$class_3 + (-4) * scores5$class_4 + 7 * scores5$class_5 +
    (-2) * scores5$class_6 + 5 * scores5$class_7 + (-10) * scores5$class_8 + 8 *
    scores5$class_9 + 4 * scores5$class_10 + (-40) * scores5$female * scores5$treat)

rct3 <- felm(read4het2 ~ treat, scores5)
rcthet3 <- felm(read4het2 ~ treat + female + female * treat, scores5)

stargazer(rct1, rcthet1, rct2, rcthet2, rct3, rcthet3, type = "latex", omit.stat = "ser")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:05:00 PM

Table 7:

	Dependent variable:					
	read4		read4het1		read4het2	
	(1)	(2)	(3)	(4)	(5)	(6)
treat	11.229***	12.829***	11.847***	22.829***	12.464***	32.829***
	(1.442)	(1.980)	(1.636)	(1.980)	(2.046)	(1.980)
female		0.412		0.412		0.412
		(1.504)		(1.504)		(1.504)
treat:female		−3.413		−23.413***		−43.413***
		(2.893)		(2.893)		(2.893)
Constant	77.150***	76.959***	77.150***	76.959***	77.150***	76.959***
	(0.750)	(1.026)	(0.850)	(1.026)	(1.063)	(1.026)
Observations	300	300	300	300	300	300
R^2	0.169	0.173	0.150	0.342	0.111	0.560
Adjusted R^2	0.166	0.165	0.147	0.336	0.108	0.556
Note:						*p<0.1; **p<0.05; ***p<0.01

You can see that the basic, non-interacted specifications in columns 1,3 and 5 return similar estimates of

the treatment effect. Nevertheless, these three DGP are quite different. Columns 2,4,and 6 show that these similar average effects hide some important heterogenity in effects in the 2nd and 3rd DGP. In the second DGP, we see that the treatment effect is entirely driven by a large effect on non-female students ($\hat{\beta}_1 = 22.8$), but the treatment has no effect on female students ($\hat{\beta}_1 + \hat{\beta}_3 = -0.6$). The third DGP is even more extreme as the positive aggregate treatment effect hides not only a difference, but actually a negative treatment effect on female students ($\hat{\beta}_1 + \hat{\beta}_3 = -10.6$) while non-female students have a large positive effect ($\hat{\beta}_1 = 32.8$).

1.5 Alphabet Soup!

1.5.1 Alphabet Soup!:IV's and RCT's from ITT to LATE

Instrumental variable estimations commonly pop up in RCT projects. The use of an instrumental variable estimator allows us to estimate what is referred to as the **LATE(Local Average Treatment Effect)** from **ITT(Intent To Treat)** estimates.

Recall the blood pressure medication example. I start with a pool of subjects that I randomize into treatment and control groups. I go to the subjects in the treatment group and I tell them to take the medication I give them. Have I actually treated them? This depends on how I define the treatment. If the treatment is defined as **taking the medication** then no, I have not treated them. If the treatment is defined as **instructing them to take the medication** then yes. I have treated them.

If I am interested in estimating the effects of actually taking the medication, $\tau = E[Y_i(1)] - E[Y_i(0)]$ is not what I am looking for. This estimate tells me the difference in the outcomes between those who were **told** to take the medication and those who were not. Thus τ is an estimate of the **intent-to-treat (ITT)** because it compares the outcomes of those I **intended** to give the medication to, to those I did not intend to give the medication to.

Whenever you have non-complieres, ie subjects that don't do as they were told, this **ITT** estimate will be different from the **LATE** estimate. The LATE estimate is, as the name suggests, the effect of the treatment on those who are induced into getting treated by our intervention. In our medicine example, the LATE of the blood control medication is the effect of actually taking the medication (not just the effect of being in the treatment group). We can recover the LATE estimates using an IV estimation where being treated is instrumented by being in the treatment group (see section XXXXXXX) for the complete details.

So what do we care about? The ITT or the LATE? Well, it really depends. A great example to think about the difference between ITT and LATE that you may have considered already is in the choice of contraceptives. When condoms are used perfectly, not only do they protect against STD's but they are also very effective as a contraceptive. But of course there is a big difference between perfect use and typical use. Though many may intend to use condoms as their main contraceptive methods, they are often not used perfectly and are thus much less effective. The ITT is substantially lower than the LATE. If you are a health care provider at the university's health center, which estimate do you consider when advising your patients on contraceptive options? Often times practitioners care more about the ITT then the LATE since for many treatments there will be non-compilers and if you are interested in estimating population level effects of a policy, these non-compilers are part of what you will need to contend with.

Thus the CDC website on contraception effectiveness reports typical (ITT) while the manufacturers of contraceptives will likely list effectiveness with perfect use (LATE) on their boxes.

You may also encounter what are refered to as TOT (Treatment on the Treated) estimates. TOT is an estimate on people who take the treatment, including always-takers (ie people in the control group who get treated). If no one in the control group is treated then the LATE=TOT.

1.6 RCT Challenges

1.6.1 Spillovers! SUTVA

Beyond the assumption of random assignment of D, there is an implicit assumption embedded in the previous section that is known rather awkwardly as the *stable unit treatment value assumption* or *SUTVA*. SUTVA

Table 3–2 Percentage of women experiencing an unintended pregnancy during the first year of typical use and the first year of perfect use of contraception, and the percentage continuing use at the end of the first year. United States.

Method (1)	% of Women Experiencing an Unintended Pregnancy within the First Year of Use		% of Women Continuing Use at One Year[3]
	Typical Use[1] (2)	Perfect Use[2] (3)	(4)
No method[4]	85	85	
Spermicides[5]	28	18	42
Fertility awareness-based methods	24		47
Standard Days method[6]		5	
TwoDay method[6]		4	
Ovulation method[6]		3	
Symptothermal method[6]		0.4	
Withdrawal	22	4	46
Sponge			36
Parous women	24	20	
Nulliparous women	12	9	
Condom[7]			
Female (fc)	21	5	41
Male	18	2	43
Diaphragm[8]	12	6	57
Combined pill and progestin-only pill	9	0.3	67
Evra patch	9	0.3	67
NuvaRing	9	0.3	67
Depo-Provera	6	0.2	56
Intrauterine contraceptives			
ParaGard (copper T)	0.8	0.6	78
Mirena (LNg)	0.2	0.2	80
Implanon	0.05	0.05	84
Female sterilization	0.5	0.5	100
Male sterilization	0.15	0.10	100

Emergency Contraception: Emergency contraceptive pills or insertion of a copper intrauterine contraceptive after unprotected intercourse substantially reduces the risk of pregnancy.[9] (See Chapter 6.).

Lactational Amenorrhea Method: LAM is a highly effective, *temporary* method of contraception.[10] (See Chapter 18.)

Source: Trussell J. Contraceptive Efficacy. In Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, Policar M. *Contraceptive Techology: Twentieth Revised Edition*. New York NY: Ardent Media, 2011.

Figure 1: ...

basically states that unit i's potential outcomes are unaffected by whether unit j is treated or untreated, ie there are no spillovers. Thus, SUTVA is often referred to as the "no interference" assumption. Basically what assumption requires is that for our treatment estimates to be correct, the affect of treatment on Jill does not depend on the whether or not Jill's neighbor received the treatment. In other words, SUTVA tends to be violated whenever the treatment, D_i involves spillovers between groups, aka some type of externality.

There are many instances where SUTVA does not hold. A classic example of SUTVA not holding is the case of vaccines. If D_i represents inoculation of unit i with the measles vaccine, and Y_i represents whether unit i gets the measles, clearly $Y_i(D_i)$ depends on all of the D's, not just D_i. For example, if all the units are vaccinated, except i then $Y_i(1) - Y_i(0) = 0$, ie the vaccine has no treatment effect on unit i, since there is no one for i to catch the measles from.

1.6.2 Attrition Bias

Another factor that can bias RCT estimates is what is referred to attrition bias. What happens to your estimates if some of the people in your experiment vanish before you can collect their end line outcome data?

If attrition is completely random, then it is not really a problem. Your sample size will be smaller but it will not bias your results. But what will happen if attrition is a function of treatment status?

RCT estimates are likely biased when attrition correlates with treatment status. To see why, lets return to our simulation of the reading group RCT. We know that the true treatment effect of participating in the small reading group is $\tau = 10$. However suppose that instead of observing the reading scores of all students, there quite a few scores that are missing because some students did not show up on test day. Furthermore, the students who do not show up are the students who are nervous about the test because they know they will do poorly on the reading exam.

In the code below, I replace the 4th grade reading score of low performing students with `NA` and re-estimate my treatment effects

```
scores5$read4miss <- NA
scores5$read4miss[scores5$read4 > 75] <- scores5$read4[scores5$read4 > 75]

scores5$obsnew <- 0
scores5$obsnew[scores5$treat == 1 & scores5$read4 > 75 & scores5$read4 < 85] <- 1

rctmiss <- felm(read4miss ~ treat, scores5)
rctmiss2 <- felm(read4miss ~ treat, scores5[scores5$obsnew == 0, ])


stargazer(rctmiss, rctmiss2, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 30, 2022 - 1:05:00 PM

Clearly this attrition pattern has biased our estimate of the treatment effect downward. Why is this? Students who scored between 65 and 75 in the control group are unobserved but their treatment group counterparts are observed since they are scoring between 75 and 85 as a result of the treatment. Thus while the treatment is shifting the distribution upwards by 10 points, it is also making the "observed" left tale longer which biases our estimates downwards from the true 10 point effect.

So what can be done to address attrition bias? The first solution is to minimize attrition throughout the data collection process. If you are still faced with attrition, you can also check to see if there the attrition is similar across treatment and control groups and make sure that it does not correlate with observables. Finally, even with problematic attrition, it is often possible to bound the extent of the bias by making hypothetical assumptions about who is dropping out of control and treatment. For instance, if the principal knows that students who will score below 75 don't take the test, she can recover the true treatment effect by excluding

Table 8:

	Dependent variable:	
	read4miss	
	(1)	(2)
treat	6.549***	10.240***
	(1.113)	(1.102)
Constant	84.963***	84.963***
	(0.667)	(0.597)
Observations	195	177
R^2	0.152	0.330
Adjusted R^2	0.148	0.327
Residual Std. Error	7.455 (df = 193)	6.677 (df = 175)
Note:	*p<0.1; **p<0.05; ***p<0.01	

the students who became observed because of the treatment from the estimation, as we see in the second column (of course it is unlikely that in the real world, a researcher would know the exact model of attrition).

1.7 RCT critiques

Randomized Control trial present some huge advantages when it comes to causal inference. With randomization, omitted variable are no longer a problem and we can be much more confident that our estimates are the causal effect of the treatment we are studying. Because of these advantages RCT have become a widely used tool in economics and the social sciences today. The 2019 Nobel Laureat in Economics was give to Abhijit Banerjee, Esther Duflo and Michael Kremer for their role in bringing this experimental approach to economic research (particularly in development economics). It is also worth noting that RCT are relatively easy to explain to policy makers, and even the general public. This presents some important advantages when it comes to communicating research results to the wider world.

Nevertheless, despite their many advantages, RCT's do present certain limitations. RCT's can be quite costly to conduct and the logistics of running an RCT are quite demanding. Furthermore, there are many very important social and economic questions where running an RCT would simply not be ethical. Understanding the effects of juvenile incarceration on human capital and future crime is clearly a question of first order importance. Randomizing which juveniles get incarcerated for the purpose of research is also clearly unethical, leaving researcher to tackle these questions with the other empirical approaches we discuss in this course.

In addition to these practical and ethical limitations, there are also some more conceptual concerns with RCT's that you should be aware of.

1.7.1 External validity

One of the major concerns that comes up with regards to RCT generated estimates are concerns about the external validity of the results: are the estimates generalizable to other contexts? For practical reasons, RCT's (especially in development economics) are often conducted in a limited geographical area with a relatively small sample size.^ [Small as compared to the kind of sample you might have when working with large administrative data.] It is thus not always clear if you would expect to get the same results in a different context or if the program were implemented on a larger scale. Because RCT's are often implemented with a lot more care and resources then a large scale policy might be, it is also not clear if we would expect the same results. This concern is further complicated by experimenter demand effects.

1.7.2 Experimenter demand effects

Experimenter demand effects refers to the concern that subjects may be behaving differently because of the experimental context then they would otherwise. Hawthorne effects, the idea that individuals might modify their behavior simply because they are being observed are a concern, particularly if they would affect the treatment and control differently. It is also worth thinking about whether subjects might change their behavior in order to conform to what they believe the researcher expects of them. This is a particularly important question when experiments are incentivized and a subject could perceive that they would receive more rewards for certain types of behaviors.

For these reasons, it is generally advisable to make sure that the salience of the evaluation is minimized as much as possible and that they are the same for both the treated and control groups.

1.7.3 General equilibrium effects

Another important concern with small scale RCT's is that many of the policies we are interested in in economics affect variables that are not determined by a single individual's choices, such as prices. Most RCT's are small enough in scale that they are unlikely to affect market level variables such as prices. It is conceivable though that if the intervention were implemented at scale, some interventions could affect market level variables which in turn could potentially counteract some of the benefits that were estimated in the small scale experiment. In a small experiment, I may be able to encourage seasonal migration to urban areas with a subsidy and find that migrants are able to find jobs there and earn some returns to migrating. If this was implemented at scale, the outcome might be very different. A national subsidy could increase labor supply in the urban area substantially possibly leading to unemployed migrants and wage adjustments. The effect of the nationally implemented policy could be very different from those estimated with the smaller scale RCT.

1.8 Fishing for Stars

This next point is not exclusively an RCT problem, nor even an Economics problem. Though because of the sunk costs involved in doing an RCT, incentives have made some of these problems particularly pronounced in projects with RCT designs, while others are more pronounced in other research designs.

The points made here will be most relevant for research done in academic settings because they are driven by the incentive to publish. Even if you do not end up working in this setting, what you want to think about is how the incentive structure you are working in (business, policy...) could generate incentives that affect how you, your colleagues, your bosses, your subordinates, conduct analyses.

1.8.1 Publication Bias

Publication Bias is widespread in many scientific fields. It refers to a pattern whereby papers with certain types of findings are more likely to get published in academic journals, leaving a lot of other types of papers unpublished. In particular, papers that report null-results (ie, no statistically significant effect was detected) are much more difficult to publish, so much so that they go unwritten altogether leading to what some researchers refer to as the "file drawer" problem: Well done research that finds null-results is simply unobserved, languishing unwritten in the drawers (hard-drives) of universities, potentially giving an incomplete picture to the answer to important questions and leading to wasted effort as researchers re-do analysis that has already done but just was never publicized.

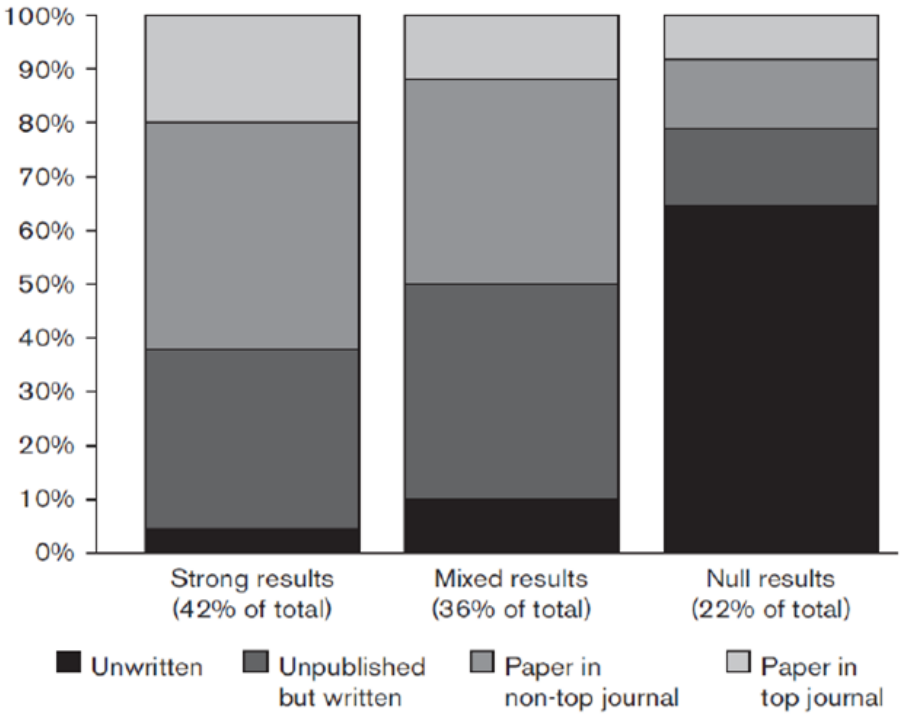
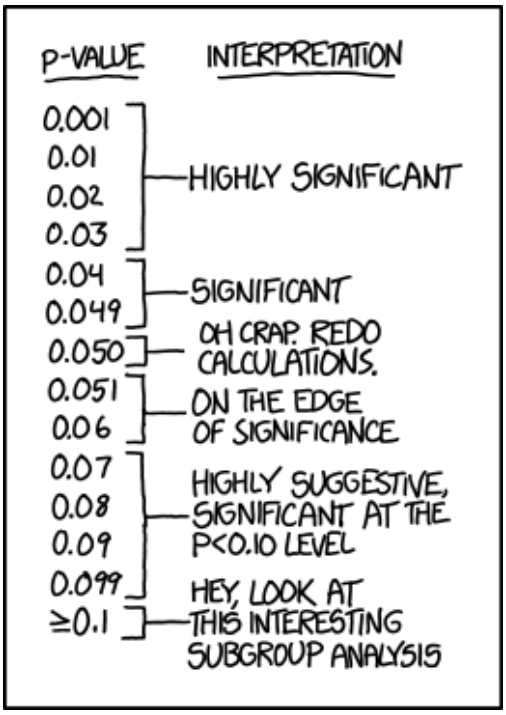


FIGURE 3.1. Publication rates and rates of writing-up of results from experiments with strong, mixed, and null results. These 221 experiments represent nearly the complete universe of studies conducted by the Time-sharing Experiments for the Social Sciences. The figure is from Mervis (2014), based on data from Franco, Malhotra, and Simonovits (2014). Reprinted with permission from AAAS.

1.8.2 P-Hacking



Since there are often different (valid) ways to specify a regression, researchers will often favor the (valid) specification that yields the most "publishable" results.

A pattern whereby there is an excess mass of papers reporting p-values that are right under the "significance" (z=1.96) threshold has been documented in many fields.

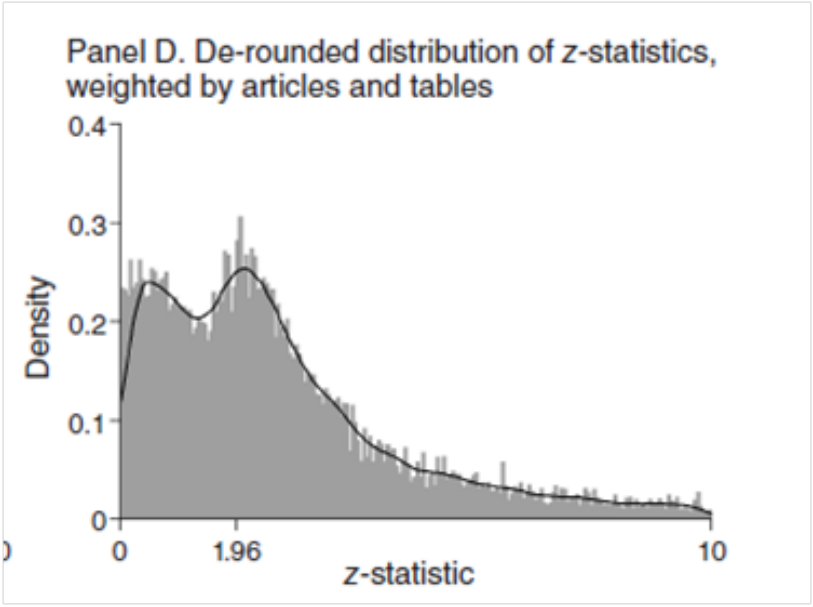


Figure 2: From Brodeur et al.2016

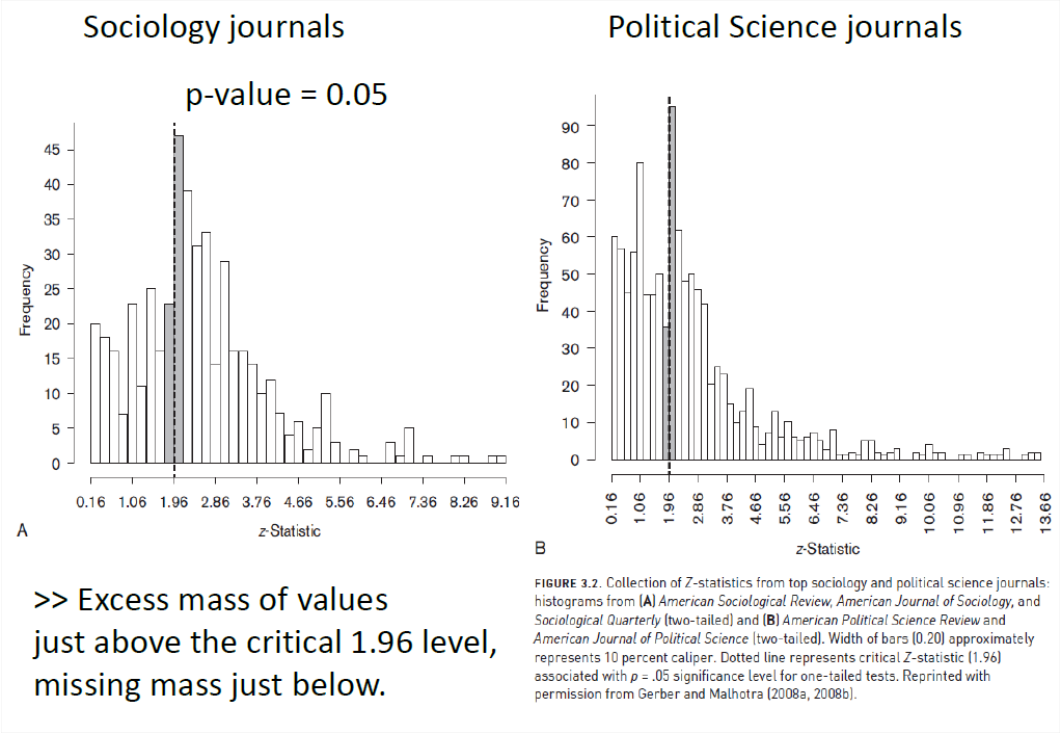
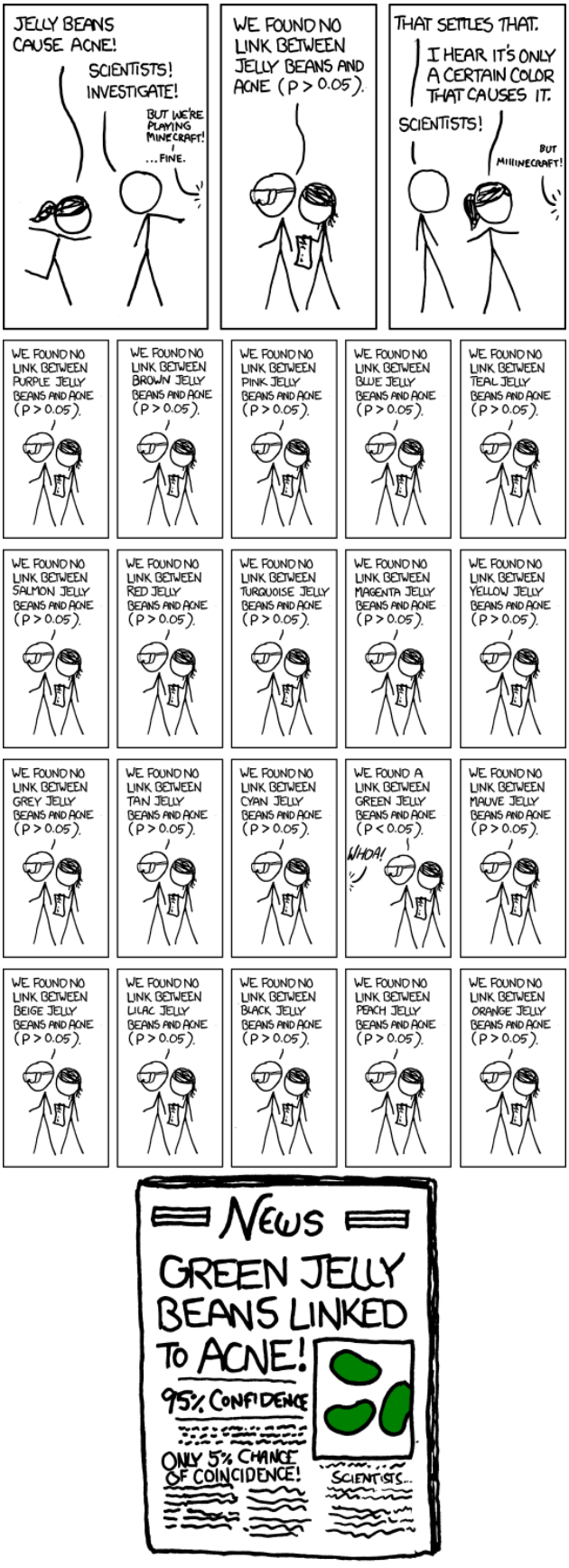


FIGURE 3.2. Collection of *Z*-statistics from top sociology and political science journals: histograms from **(A)** *American Sociological Review, American Journal of Sociology,* and *Sociological Quarterly* (two-tailed) and **(B)** *American Political Science Review* and *American Journal of Political Science* (two-tailed). Width of bars (0.20) approximately represents 10 percent caliper. Dotted line represents critical *Z*-statistic (1.96) associated with $p = .05$ significance level for one-tailed tests. Reprinted with permission from Gerber and Malhotra (2008a, 2008b).

1.8.3 Cherry-Picking

When researchers have incurred large costs on a project, there are very strong incentives to find significant effects. One way to increase the chance of finding significant effects, is to collect data on a huge number of outcome variables. Without insight into the actual research process, it then becomes difficult to tell if the researcher found what their hypothesis predicted or if they simply ran many many regressions.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

1.9 Research Transparency

Fortunately there is a push to address these issues in economics, particularly in the RCT field where these incentives are very strong and the researcher has so much control over the data collection process.

- Increasingly journals require that the data and code used to generate the research results be made available to other researchers. This allows people the opportunity to verify the methodology used by the researcher.
- When a project runs regressions on an inordinate number of outcome variables, it is increasingly expected that the researcher will report multiple inference adjusted standard errors (in addition to the usual standard errors). These adjusted standard errors generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.
- Pre-analysis plans are increasingly becoming the norm. A pre-analysis plan is a document detailing the the questions, methods, and estimations a researcher plans to implement in a research project before they actually receive any data. The plan is then dated and filed in a public repository so that when the data is later analysed, evaluators can assess how far the researcher had to deviate from their initial project plan.
- Some journals are starting to accept articles based on the pre-analysis plan, before seeing any results.