A Reasoned Basis for Inference: Fun with Fisher

Jake Bowers[1,2]

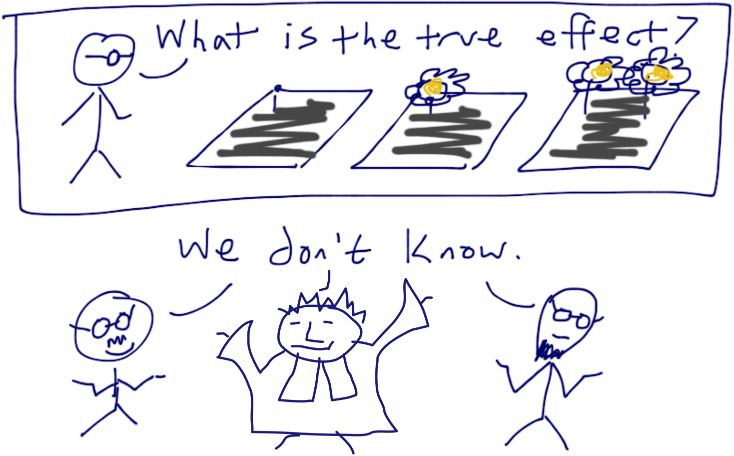
Aronow and Crawford Seminar, Yale University

[1] *Political Science & Statistics & NCSA @ University of Illinois*

[2] *White House Social and Behavioral Sciences Team*

jwbowers@illinois.edu — http://jakebowers.org

What is the true effect of the treatment assignment?



What is the true effect of the treatment assignment?

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	y_{i1}	y_{i0}
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i1}$	$\bar{y}_{i0}$

$$\widehat{ATE} = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

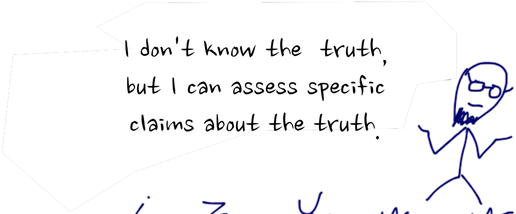
What is the true effect of the treatment assignment?

I dew nut knew thee truth, but, given pryors, I cane predikte itf probabeeleetee.

i	Z_i	Y_i	y_{i1}	y_{i0}
A	0	16		16
B	1	22	22	
C	0	7		7
D	1	14	14	

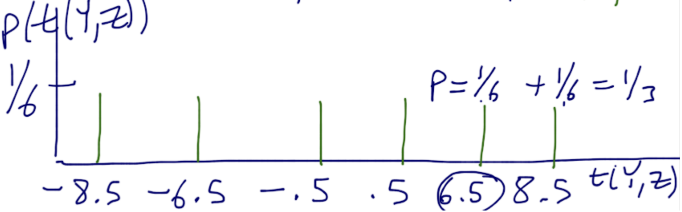
$$p\left(\boxed{\;}, f(y_{i1} - y_{i0})\right) =$$

What is the true effect of the treatment assignment?



$H_0: y_{i,1} = y_{i,0}$

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	22
C	0	7	?	7
D	1	14	14	14



Testing the Sharp Null of No Effects

```
Z <- c(0,1,0,1)
Y <- c(16,22,7,14)
Om <- matrix(0,ncol=choose(4,2),nrow=length(Z))
whotrted <- combn(1:4,2)
for(i in 1:choose(4,2)){ Om[cbind(whotrted[,i],i)]<-1 }
meandifftz <- function(y,z){ mean(y[z==1]) - mean(y[z==0]) }
thedist<-apply(Om,2, function(z){ meandifftz(Y,z) })
rbind(Om,thedist)

        [,1] [,2] [,3] [,4] [,5] [,6]
         1.0  1.0  1.0  0.0  0.0  0.0
         1.0  0.0  0.0  1.0  1.0  0.0
         0.0  1.0  0.0  1.0  0.0  1.0
         0.0  0.0  1.0  0.0  1.0  1.0
thedist  8.5 -6.5  0.5 -0.5  6.5 -8.5

table(thedist)

thedist
-8.5 -6.5 -0.5  0.5  6.5  8.5
   1    1    1    1    1    1

theobs <- meandifftz(Y,Z)
mean(thedist >= theobs)

[1] 0.3333333
```

The Newspapers Study

The Newspapers Study: $H_0 : y_{i1} = y_{i0}, p = 6/16$.

Overview Statistical Inference for Causal Quantities

Testing Fishers Sharp Null Hypothesis of No Effects

A Real Field Experiment with 8 Cities: The Newspapers Experiment

Sharp Hypotheses of Some Effects. Sharp Hypotheses as Causal Models

The Constant Additive Effect Model

The No Effects and No Interference Model

Aspects of the approach: The tests do not require asymptopia.

Information arises from both theory and instruments

Summary and Discussion

Appendix

What range of effects might be surprising?

Hypothesize a model of potential outcomes For $H_0 : y_{i1} = y_{i0} + \tau$, what τ might be surprising?

Map the model to observation via design What would $\tau = 6$ imply for what we observe? If $\tau = 6$ and $Y_i = Z_i y_{i1} + (1 - Z_i) y_{i0}$ then $Y_i - Z_i 6 = y_{i0}$.

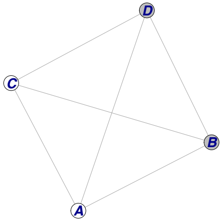
Generate the randomization distribution of this hypothesis As before

Summarize information against the hypothesis For hypotheses of $\tau \geq 6$ we have $p \leq .125$ using a mean difference test statistic.

What effects might be least surprising?

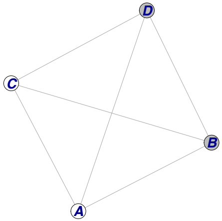
The idea of a "best guess" maps onto the Hodges-Lehmann point estimate: the hypotheses for which $E(t(Z,Y) = 0$: ex. the difference of means is zero. Here $\tau = 1.5$ for the mean difference and $\tau = 3.25$ for the rank-based test (which equalizes medians).

Statistical inference for counterfactual quantities with interference?



i	Z_i	Y_i	$y_{i,1100}$	$y_{i,0101}$	$y_{i,1001}$	$y_{i,0110}$	$y_{i,1010}$	$y_{i,0011}$
A	0	16	?	16	?	?	?	?
B	1	22	?	22	?	?	?	?
C	0	7	?	7	?	?	?	?
D	1	14	?	14	?	?	?	?

Statistical inference for counterfactual quantities with interference?



i	Z_i	Y_i	$y_{i,1100}$	$y_{i,0101}$	$y_{i,1001}$	$y_{i,0110}$	$y_{i,1010}$	$y_{i,0011}$	$y_{i,0000} \equiv y_{i,\mathbf{0}}$
A	0	16	?	16	?	?	?	?	16
B	1	22	?	22	?	?	?	?	22
C	0	7	?	7	?	?	?	?	7
D	1	14	?	14	?	?	?	?	14

The sharp null of no effects is a model of no interference:

$H_0 : y_{i,1100} = y_{i,0101} = y_{i,1001} = y_{i,0110} = y_{i,1010} = y_{i,0011} = y_{i,0000}$,
$y_{i,\mathbf{0}} = \mathcal{H}(y_{i,\mathbf{z}}, \mathbf{0}) = y_{i,\mathbf{z}}$, $p = 0.33$.

Introducing the **uniformity trial** $\equiv \mathbf{y}_{i,0000}$ (Rosenbaum, 2007).

Theoretical models of potential outcomes can produce sharp hypotheses

Theoretical models of potential outcomes can produce sharp hypotheses

Theoretical models of potential outcomes can produce sharp hypotheses

A General Fisherian Inference Algorithm

1. Write a model ($\mathcal{H}(\mathbf{y_0}, \mathbf{z}, \theta)$) converting uniformity trial into observed data (i.e. a causal model).
2. Solve for $\mathbf{y_0}$: $\mathcal{H}(\mathbf{y_z}, \mathbf{0}, \theta_0) = \mathbf{y_0}$
3. Select a test statistic that is effect increasing in all relevant dimensions.
4. Compute p-values for substantively meaningful range of θ. Or calculate boundaries of regions.

Fisherian Tests work when asymptopia is out of reach

Fisherian Tests work when asymptopia is out of reach

Robust test statistics can increase power.

For the simple mean difference test statistic, we have $p = .375$, for a rank sum test (the sum of the ranks of the treated units), we have a $p = .4375$, and for an M-estimator based test (like mean-differences but with weights roughly inversely proportional to the Cook's D influence measure) we have $p = .3125$.

Covariance adjusted tests can increase power.

Using the difference pre-vs-post as the outcome (comparing treated pre-vs-post with paired control pre-vs-post), and using the robust test statistic for $H_0 : y_{i1} = y_{i0}$ $p = 4/16 = .25$. Using $e_i = (Y_i - Y_{i,t-1}) - (\hat{\beta}_0 + \hat{\beta}_1 \text{pop} + \hat{\beta}_2 \text{num candidates})$ and the robust test statistic to test $H_0 : y_{i1} = y_{i0}$ we have $p = 2/16 = .125$.

Key features of Fisher's approach

Flexible Any scientific model than can generate implications for all units' potential outcomes can, in principle, produce testable parameters.

Design based Requires knowledge of probability of Z not Y or $Y|X$ or $\beta|\gamma$.

Finite Sample Oriented Does not require asymptopia. Can use asymptopia when there for a visit.

Can be slow In between 8 cities and asymptopia is a land of many permutations.

Probably conservative Uses relatively little of the total information we have available about the science.

If you want to know more read Paul Rosenbaum's work The version of Fisher's approach I discuss here is built on work by Paul Rosenbaum. Read his work if you want to learn more.

New Questions

- How to choose test statistics for multidimensional sharp-hypothesis testing? Are there multi-dimensional “effect increasing” characteristics that we can assess for a given model?
- Are there general classes of scientific/counterfactual models?
- How should we interpret and display results?

On models

- Models are mathematical functions, multiple functions can have similar adjustments to the data.
- Assessing more than one model may enhance insight (Rosenbaum, 2010).
- When more than one model is plausible, what should you do?
- Our method can help eliminate the implausible, not accept the plausible.

New Questions

- Where do models of counterfactuals come from? Do we have advice about going from words to math?
- Math has its own logic. Some expressions for models may not be sensitive to changes in parameters. How can we assess what a given model is telling? How can we go from math to words before testing hypotheses?
- The KS-statistic is low powered for tail-differences. Recall that we are testing $t(\mathcal{H}(), \mathbf{z})$ not just $\mathcal{H}()$. Some results might tell us that our test is low powered against certain alternatives more than that we have identified a region of plausibility. How to find a better test statistic?
- How can this work learn from other modes of statistical inference and other representations of causal inference? What are the connections to ATE and other estimation frameworks (Spatial Econometrics, Network Analysis (ERGMs), etc...)?

New Questions

- Where do models of counterfactuals come from? Do we have advice about going from words to math?
- Math has its own logic. Some expressions for models may not be sensitive to changes in parameters. How can we assess what a given model is telling? How can we go from math to words before testing hypotheses?
- The KS-statistic is low powered for tail-differences. Recall that we are testing $t(\mathcal{H}(), \mathbf{z})$ not just $\mathcal{H}()$. Some results might tell us that our test is low powered against certain alternatives more than that we have identified a region of plausibility. How to find a better test statistic?
- How can this work learn from other modes of statistical inference and other representations of causal inference? What are the connections to ATE and other estimation frameworks (Spatial Econometrics, Network Analysis (ERGMs), etc...)?

New Questions

- Where do models of counterfactuals come from? Do we have advice about going from words to math?
- Math has its own logic. Some expressions for models may not be sensitive to changes in parameters. How can we assess what a given model is telling? How can we go from math to words before testing hypotheses?
- The KS-statistic is low powered for tail-differences. Recall that we are testing $t(\mathcal{H}(), \mathbf{z})$ not just $\mathcal{H}()$. Some results might tell us that our test is low powered against certain alternatives more than that we have identified a region of plausibility. How to find a better test statistic?
- How can this work learn from other modes of statistical inference and other representations of causal inference? What are the connections to ATE and other estimation frameworks (Spatial Econometrics, Network Analysis (ERGMs), etc...)?

New Questions

- Where do models of counterfactuals come from? Do we have advice about going from words to math?
- Math has its own logic. Some expressions for models may not be sensitive to changes in parameters. How can we assess what a given model is telling? How can we go from math to words before testing hypotheses?
- The KS-statistic is low powered for tail-differences. Recall that we are testing $t(\mathcal{H}(), \mathbf{z})$ not just $\mathcal{H}()$. Some results might tell us that our test is low powered against certain alternatives more than that we have identified a region of plausibility. How to find a better test statistic?
- How can this work learn from other modes of statistical inference and other representations of causal inference? What are the connections to ATE and other estimation frameworks (Spatial Econometrics, Network Analysis (ERGMs), etc...)?