

# What does it mean to test a hypothesis?

*Jake Bowers<sup>1,2</sup>*

EGAP Learning Days 4: Chile

<sup>1</sup> Political Science & Statistics & NCSA @ University of Illinois

<sup>2</sup> White House Social and Behavioral Sciences Team

*jwbowers@illinois.edu – <http://jakebowers.org>*

## Counterfactual Causal Inference

How can we use what we see to learn about what we want to know ?

City	Pair	Treatment	Turnout		Newspaper	$y_1$	$y_0$
			Baseline	Outcome			
Saginaw	1	0	17	16		?	16
Sioux City	1	1	21	22	Sioux City Journal	22	?
Battle Creek	2	0	13	14		?	14
Midland	2	1	12	7	Midland Daily News	7	?
Oxford	3	0	26	23		?	23
Lowell	3	1	25	27	Lowell Sun	27	?
Yakima	4	0	48	58		?	58
Richland	4	1	41	61	Tri-City Herald	61	?

Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment randomized within pair with the newspaper ads as 1 and lack of treatment as 0. The potential outcomes are  $y_1$  for treatment and  $y_0$  for control. Panagopoulos (2006) provides more detail on the design of the experiment.

## Counterfactual Causal Inference

How can we use what we see to learn about potential outcomes?

City	Pair	Treatment	Turnout		Newspaper	$y_1$	$y_0$
			Baseline	Outcome			
Saginaw	1	0	17	16		?	16
Sioux City	1	1	21	22	Sioux City Journal	22	?
Battle Creek	2	0	13	14		?	14
Midland	2	1	12	7	Midland Daily News	7	?
Oxford	3	0	26	23		?	23
Lowell	3	1	25	27	Lowell Sun	27	?
Yakima	4	0	48	58		?	58
Richland	4	1	41	61	Tri-City Herald	61	?

Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment randomized within pair with the newspaper ads as 1 and lack of treatment as 0. The potential outcomes are  $y_1$  for treatment and  $y_0$  for control. Panagopoulos (2006) provides more detail on the design of the experiment.

## Counterfactual Causal Inference

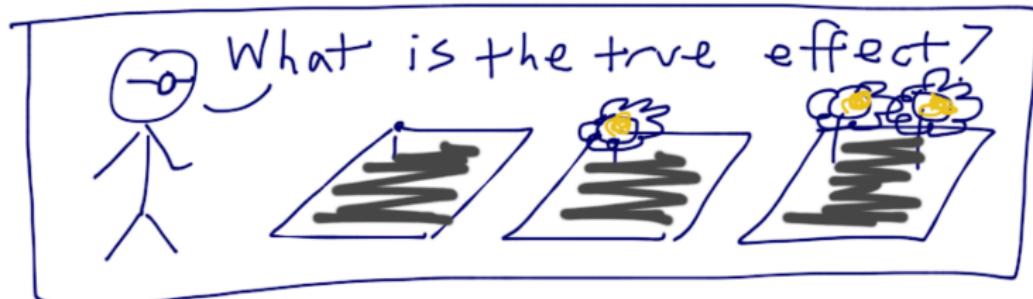
How can we use what we see to learn about a causal effect  $i = f(y_{i,1}, y_{i,0})$ ?

City	Pair	Treatment	Baseline	Outcome	Turnout		$y_1$	$y_0$
					Newspaper			
Saginaw	1	0	17	16			?	16
Sioux City	1	1	21	22	Sioux City Journal		22	?
Battle Creek	2	0	13	14			?	14
Midland	2	1	12	7	Midland Daily News		7	?
Oxford	3	0	26	23			?	23
Lowell	3	1	25	27	Lowell Sun		27	?
Yakima	4	0	48	58			?	58
Richland	4	1	41	61	Tri-City Herald		61	?

Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment randomized within pair with the newspaper ads as 1 and lack of treatment as 0. The potential outcomes are  $y_1$  for treatment and  $y_0$  for control. Panagopoulos (2006) provides more detail on the design of the experiment.

**Small group exercise:** Can you think of **more than one way** to learn about the counterfactual causal effect of treatment using what we observe?

# What is the true effect of the treatment assignment?



We don't know.



## 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	?

$$\widehat{ATE} = \bar{Y}_1 | z_i=1 - \bar{Y}_0 | z_i=0$$
$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

# What is the true effect of the treatment assignment?

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



$i$	$\epsilon_i$	$y_i$	$y_{i1}$	$y_{i0}$
A	0	16	16	16
B	1	22	22	22
C	0	7	7	7
D	1	14	14	14

$$P(\text{[speech bubble]} \rightarrow f(y_1 - y_0)) = \text{[wavy line]}$$

# What is the true effect of the treatment assignment?

I don't know the truth,  
but I can assess specific  
claims about the truth.


$$H_0: y_{i1} = y_{i0}$$

i	$z_i$	$y_i$	$y_{i1}$	$y_{i0}$
A	0	16	?	16
B	1	22	22	22
C	0	7	?	7
D	1	14	14	14

$$P(t(y, z))$$

$$\frac{1}{6}$$

$$-8.5$$

$$-6.5$$

$$-.5$$

$$.5$$

$$P = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$6.5$$

$$8.5$$

$$t(y, z)$$



# Testing the Sharp Null of No Effects

```
Z <- c(0,1,0,1) ## Define treatment vector
Y <- c(16,22,7,14) ## Define outcome vector
## There are choose(4,2) ways to assign treatment
## Make a matrix containing all of the ways to assign treatment
Om <- matrix(0, ncol=choose(4,2), nrow=length(Z)) ## First fill with zeros
whotrted <- combn(1:4,2) ## Generate indicators of who is treated
for(i in 1:choose(4,2)){ Om[cbind(whotrted[,i],i)]<-1 }

## Create a function to summarize the hypothesized
## treatment to outcome relationship
meandifftz <- function(y,z){ mean(y[z==1]) - mean(y[z==0]) }
rankdifftz <- function(y,z){ q<-rank(y); mean(q[z==1]) - mean(q[z==0]) }
## Apply the function to all possible experiments
mdist<-apply(Om,2, function(z){ meandifftz(Y,z) })
rdist<-apply(Om,2, function(z){ rankdifftz(Y,z) })
rbind(Om,mdist,rdist)

 [,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
mdist  8.5 -6.5  0.5 -0.5  6.5 -8.5
rdist  2.0 -1.0  0.0  0.0  1.0 -2.0
```

## continued

```
table(mdist)
```

mdist

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

```
table(rdist)
```

rdist

-2	-1	0	1	2
1	1	2	1	1

```
mobs <- meandifftz(Y,Z) ## observed value of test stat  
robs <- rankdifftz(Y,Z) ## observed value of test stat  
mean(mdist >= mobs) ## p-value
```

```
[1] 0.3333333
```

```
mean(rdist >= robs) ## p-value
```

```
[1] 0.3333333
```

## Faster Method

```
library(RItools, lib.loc="libraries") ## see fishersinference.Rnw to get the dev versi  
## Specify design: only 1 block  
randomassignment<-simpleRandomSampler(z=Z, b=rep(1,4))  
  
meandiffTZ <- function(ys, z) {  
  mean(ys[!(z)]) - mean(ys[z])  
}  
  
meandifftest<-RItest(y=Y, z=Z, test.stat=meandiffTZ, sampler = randomassignment)  
meandifftest  
  
Call: RItest(y = Y, z = Z, test.stat = meandiffTZ, sampler = randomassignment)
```

	Value	Pr(>x)
Observed Test Statistic	6.5	0.3333

## The Newspapers Study

City	Pair	Treatment	Baseline	Outcome	Turnout		$y_1$	$y_0$
					Newspaper			
Saginaw	1	0	17	16			?	16
Sioux City	1	1	21	22	Sioux City Journal		22	?
Battle Creek	2	0	13	14			?	14
Midland	2	1	12	7	Midland Daily News		7	?
Oxford	3	0	26	23			?	23
Lowell	3	1	25	27	Lowell Sun		27	?
Yakima	4	0	48	58			?	58
Richland	4	1	41	61	Tri-City Herald		61	?

Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment randomized within pair with the newspaper ads as 1 and lack of treatment as 0. The potential outcomes are  $y_1$  for treatment and  $y_0$  for control. Panagopoulos (2006) provides more detail on the design of the experiment.

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

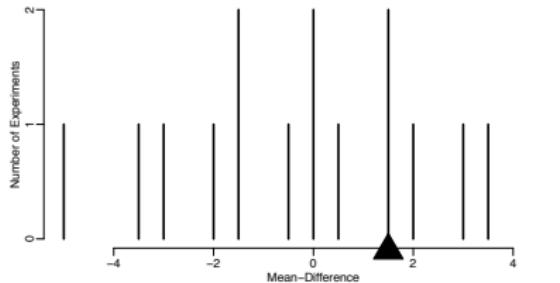
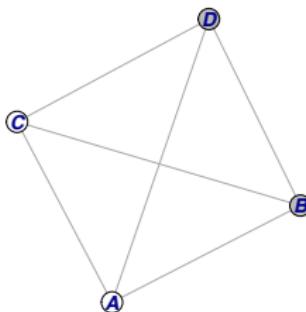


Figure 2: The randomization distribution of the mean-difference test statistic under the null hypothesis of no effects is shown by the tall black lines. The black triangle shows the value of the mean-difference observed in the actually fielded experiment.

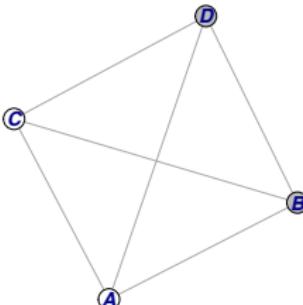


## Statistical inference for counterfactual quantities with interference?



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

## Statistical inference for counterfactual quantities with interference?



$i$	$Z_i$	$Y_i$	$y_{i,1100}$	$y_{i,0101}$	$y_{i,1001}$	$y_{i,0110}$	$y_{i,1010}$	$y_{i,0011}$	$y_{i,0000} \equiv y_{i,0}$
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:

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

Introducing the uniformity trial  $\equiv y_{i,0000}$  (Rosenbaum, 2007).



## Fisherian Tests work when asymptopia is out of reach

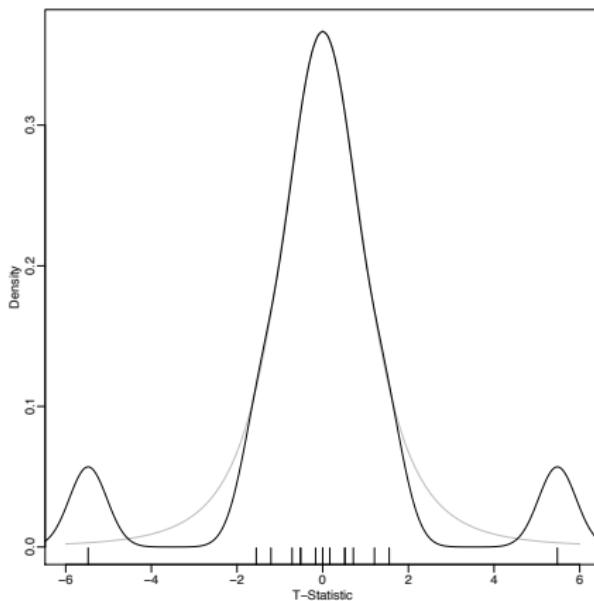


Figure 5: Under the null hypothesis of no effects,  $t$ -statistics calculated after repeating the experimental assignment process should be distributed around zero. The exact randomization distribution is tri-modal for the Newspapers study and does not match the  $t$ -distribution that we would expect under a larger sample.

## Fisherian Tests work when asymptopia is out of reach

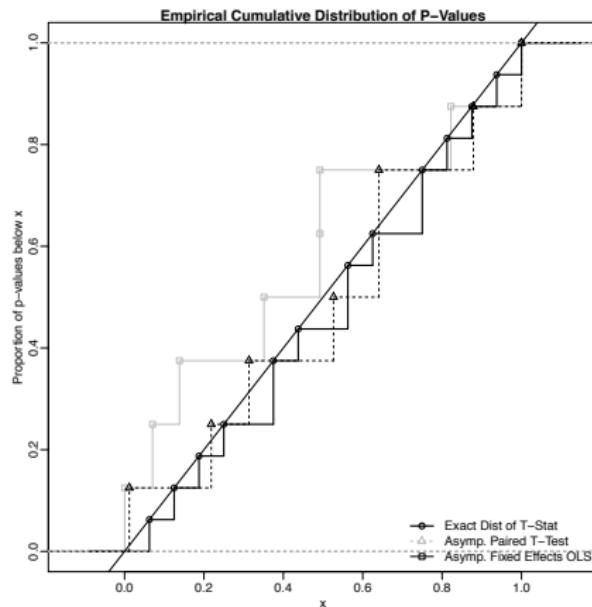


Figure 6: Cumulative distributions of p-values for testing the null hypothesis of no effects. A valid test would be at or below the diagonal line. Tests which would encourage rejection too quickly have points above the line.



## What does it mean to test a hypothesis?

What is a  $p$ -value?

What do we learn from a hypothesis test?

What do we need to assume to make a hypothesis test (in this simplest case) interpretable?



## What range of effects might be surprising?

Hypothesize a model of potential outcomes For  $H_0 : y_{i1} = y_{i0} + \tau$ , what  $\tau$  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.

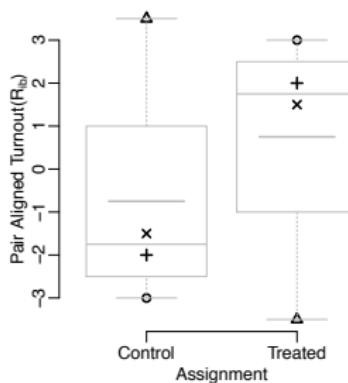


Figure 3: The distribution of outcomes in the control and treatment groups. The values of turnout are “pair aligned” or “pair mean centered” so that we can focus attention on the paired differences rather than on the differences in turnout levels between pairs.

## 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).

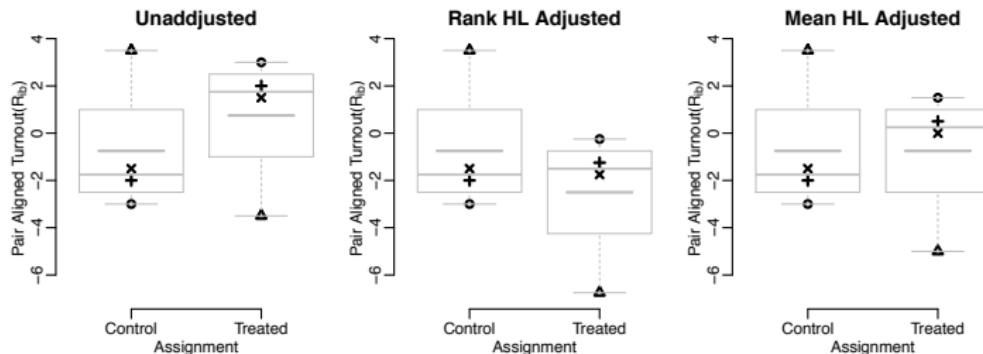
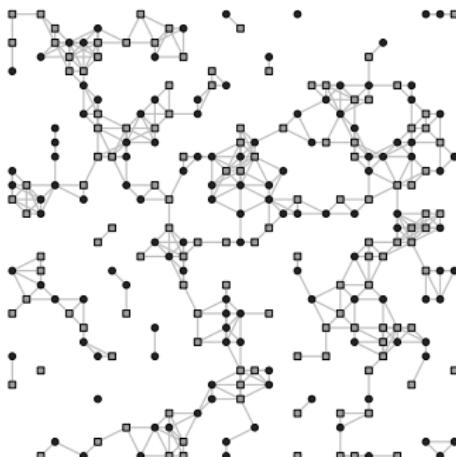


Figure 4: The unadjusted comparison of controls versus treated (left most plot) as compared to the results of adjusting the treatment group based on the Hodges-Lehmann (HL) point estimates derived from two different test statistics (Wilcoxon paired signed ranks and the mean difference). Notice that the control group remains the same and it is the hypotheses which imply changes in the distribution of the treatment group compared to the unadjusted, observed, outcomes.

## Theoretical models of potential outcomes can produce sharp hypotheses



**Figure:** A simulated data set with 256 units and 512 connections. The  $256/2 = 128$  treated units ( $Z_i = 1$ ) are shown as filled circles and an equal number of control units ( $Z_i = 0$ ) are shown as gray squares.

## Theoretical models of potential outcomes can produce sharp hypotheses

$$\mathcal{H}(\mathbf{y}_0, \mathbf{z}, \beta, \tau) = \left[ \beta + (1 - z_i)(1 - \beta) \exp(-\tau^2 \mathbf{z}^T \mathbf{S}) \right] \mathbf{y}_0 \quad (1)$$

$$\mathcal{H}(\mathbf{y}_z, \mathbf{0}, \beta, \tau) = \left[ \beta + (1 - z_i)(1 - \beta) \exp(-\tau^2 \mathbf{z}^T \mathbf{S}) \right]^{-1} \mathbf{y}_z \equiv \mathbf{y}_0 \quad (2)$$

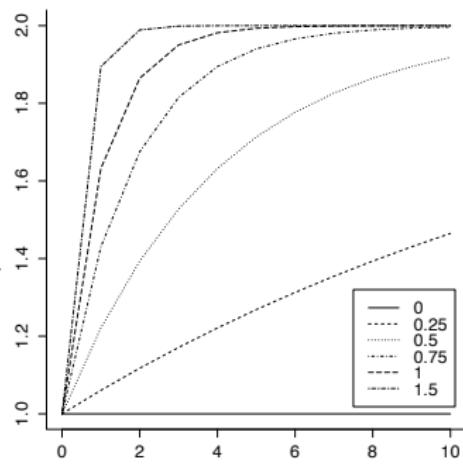
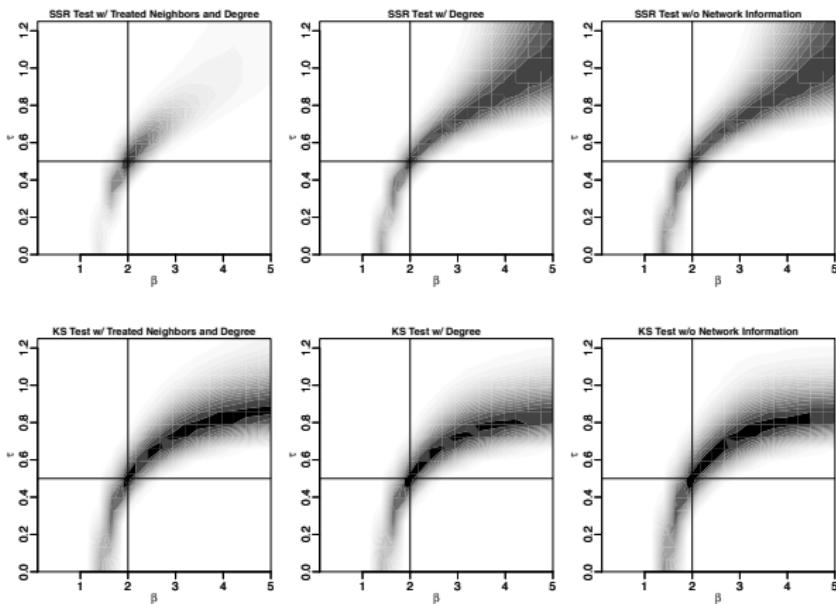


Figure: Growth curve of spillover effects for the expression  $\beta + (1 - \beta) \exp(-\tau^2 \mathbf{z}^T \mathbf{S})$  as the number of treated neighbors,  $\mathbf{z}^T \mathbf{S}$ , increases for  $\beta = 2$  and a selection of  $\tau$  values.

# Theoretical models of potential outcomes can produce sharp hypotheses



**Figure:** Proportion of  $p$ -values less than .05 for randomization tests of joint hypotheses about  $\tau$  and  $\beta$ . Darker values mean less rejection. Truth is at  $\tau = .5, \beta = 2$ . All tests reject the truth no more than 5% of the time at  $\alpha = .05$ . All simulations, not Normal approximations.

## 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  $\theta$ . Or calculate boundaries of regions.



## 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  $p = .3125$ .

Cook's $D$	MM-Weights	City	Pair	Treatment	Turnout
0.13	1.00	Saginaw	1	0	16
0.13	1.00	Sioux City	1	1	22
0.48	0.94	Battle Creek	2	0	14
0.48	0.94	Midland	2	1	7
0.04	1.00	Oxford	3	0	23
0.04	1.00	Lowell	3	1	27
0.01	1.00	Yakima	4	0	58
0.01	1.00	Richland	4	1	61

Table 2: Not all cases have the same influence on the mean difference. Cook's  $D$  summarizes this influence — Battle Creek and Midland have disproportionate influence. A robust fit (details in the text) downweights Battle Creek and Midland.

## 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.

Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment." Paper presented at the MPSA 2006.