Section 2: Randomization Inference & Propensity Scores

Stephen Pettigrew

Harvard University

October 6, 2015

Fisherian (Randomization) Inference

Why use it?

- Super helpful with small sample sizes
- Doesn't rely on distributional assumptions like asymptotic normality

Fisherian (Randomization) Inference

Why use it?

- Super helpful with small sample sizes
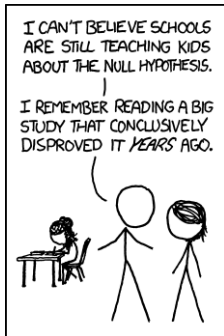- Doesn't rely on distributional assumptions like asymptotic normality

What are the drawbacks?

- Requires a much stronger null hypothesis that might be an unrealistic benchmark to test against

The Sharp Null Hypothesis
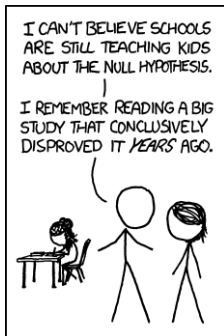
Formally, the sharp null hypothesis implies that:

$$\tau_i = \tau_0 \ \ \forall i$$

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.

I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

The Sharp Null Hypothesis

Formally, the sharp null hypothesis implies that:

$$\tau_i = \tau_0 \ \ \forall i$$

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.

I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.
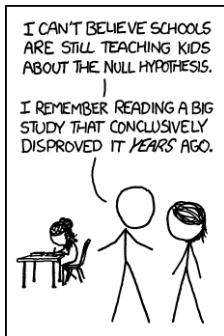
Recall that with hypothesis testing in the frequentist framework, your null hypothesis can be that β (or the ATE or whatever) is equal to any number, α (not just zero)

It's the same in the Fisherian framework–the sharp null can be that the treatment effect equals α

The Sharp Null Hypothesis

Formally, the sharp null hypothesis implies that:

$$\tau_i = \tau_0 \;\; \forall i$$

Recall that with hypothesis testing in the frequentist framework, your null hypothesis can be that β (or the ATE or whatever) is equal to any number, α (not just zero)

It's the same in the Fisherian framework–the sharp null can be that the treatment effect equals α

Typically though we're interested in the sharp null:

$$\tau_i = \tau_0 = 0 \;\; \forall i$$

Stephen Pettigrew (Harvard University)   Section 2: Randomization Inference & Propen   October 6, 2015   3 / 11

Testing the Sharp Null

Basic approach:

1. Choose a test statistic–i.e. a scaler quantity that you can calculate from the treatment assignment vector and other observed data
2. Calculate that test statistic given the observed data and treatment assignment vector

Testing the Sharp Null

Basic approach:

1. Choose a test statistic–i.e. a scaler quantity that you can calculate from the treatment assignment vector and other observed data

2. Calculate that test statistic given the observed data and treatment assignment vector

3. Use your knowledge of the treatment assignment process to write out all possible treatment assignment vectors

   If there's a ton of those, then randomly sample a bunch

Testing the Sharp Null

Basic approach:

1. Choose a test statistic–i.e. a scaler quantity that you can calculate from the treatment assignment vector and other observed data

2. Calculate that test statistic given the observed data and treatment assignment vector

3. Use your knowledge of the treatment assignment process to write out all possible treatment assignment vectors

   If there's a ton of those, then randomly sample a bunch

4. For each of these vectors, replace the observed treatment vector with the hypothetical treatment vector. Recalculate and store the test stat with this 'new' data

Testing the Sharp Null

Basic approach:

1. Choose a test statistic–i.e. a scaler quantity that you can calculate from the treatment assignment vector and other observed data

2. Calculate that test statistic given the observed data and treatment assignment vector

3. Use your knowledge of the treatment assignment process to write out all possible treatment assignment vectors

   If there's a ton of those, then randomly sample a bunch

4. For each of these vectors, replace the observed treatment vector with the hypothetical treatment vector. Recalculate and store the test stat with this 'new' data

5. Get Fisher's exact p-value by calculating the percentage of hypothetical test statistics that were at least as large as the observed test statistic

How to do it in R

Step 0: get some data

```
set.seed(25)
data <- data.frame(y.obs=rnorm(15,mean = 5,sd = 2))

# we'll assume a completely randomized experiment
# where 8 units are assigned treatment
data$treat <- 0
data$treat[sample(1:nrow(data), 8)] <- 1
```

How to do it in R

Steps 1 & 2: pick a test stat and calculate its value
the observed data

How to do it in R

Steps 1 & 2: pick a test stat and calculate its value the observed data

```
# absolute value of diff of means
# as the test statistic
test.stat <- abs(mean(data$y.obs[data$treat == 1]) -
                 mean(data$y.obs[data$treat == 0]))
[1] 0.9348497
```

How to do it in R

Step 3: write out all the possible **D** vectors

How to do it in R

Step 3: write out all the possible **D** vectors

```
#install.packages("ri")
require(ri)
possible.d.vectors <- genperms(data$treat)
```

How to do it in R

Step 3: write out all the possible **D** vectors

```
#install.packages("ri")
require(ri)
possible.d.vectors <- genperms(data$treat)
```

How many possible **D** vectors are there?

How to do it in R

Step 3: write out all the possible **D** vectors

```
#install.packages("ri")
require(ri)
possible.d.vectors <- genperms(data$treat)
```

How many possible **D** vectors are there? $\binom{15}{8} = 6435$

How to do it in R

Step 3: write out all the possible **D** vectors

```
#install.packages("ri")
require(ri)
possible.d.vectors <- genperms(data$treat)
```

How many possible **D** vectors are there? $\binom{15}{8} = 6435$

Note: if there's more than 10,000 possible **D** vectors, `genperms()` will randomly sample 10,000 of them
The `maxiter` option in `genperms()` can change this

How to do it in R

Step 4: calculate the test stat for each possible **D**

How to do it in R

Step 4: calculate the test stat for each possible **D**

```
all.test.stats <- c()
for(ii in 1:ncol(possible.d.vectors)){
  current.d.vector <- possible.d.vectors[,ii]
  all.test.stats[ii] <- abs(mean(data$y.obs[current.d.vector == 1]) -
                            mean(data$y.obs[current.d.vector == 0]))
}
```

How to do it in R

Step 4: calculate the test stat for each possible **D**

```
all.test.stats <- c()
for(ii in 1:ncol(possible.d.vectors)){
  current.d.vector <- possible.d.vectors[,ii]
  all.test.stats[ii] <- abs(mean(data$y.obs[current.d.vector == 1]) -
                            mean(data$y.obs[current.d.vector == 0]))
}
```

How to do it in R

Step 5: calculate the exact p-value

How to do it in R

Step 5: calculate the exact p-value

```
mean(test.stat <= all.test.stats)
[1] 0.3633256
```

How to do it in R

Step 5: calculate the exact p-value

```
mean(test.stat <= all.test.stats)
[1] 0.3633256
```

Is this one sided or two?

How to do it in R

Step 5: calculate the exact p-value

```
mean(test.stat <= all.test.stats)
[1] 0.3633256
```

Is this one sided or two? two, because we took the absolute value

How to do it in R

Step 5: calculate the exact p-value

```
mean(test.stat <= all.test.stats)
[1] 0.3633256
```

Is this one sided or two? two, because we took the absolute value

What does this p-value mean?

How to do it in R

Step 5: calculate the exact p-value

```
mean(test.stat <= all.test.stats)
[1] 0.3633256
```

Is this one sided or two? two, because we took the absolute value

What does this p-value mean?

The probability of getting a test statistic as big or bigger than the one we observed, given that the sharp null hypothesis is true

In other words, the sharp null says that the y_{obs} will not change if we change the treatment vector. So if the randomization had gone differently, there is a 36% chance that we'd have seen a test statistic at least this big

In this case, out of the 6435 possible randomizations, 2338 of them would have yielded a test stat as big or bigger than 0.9348, even if the treatment effect is zero for all the observations

How to do it in R

(Bonus) Step 6: display the result graphically
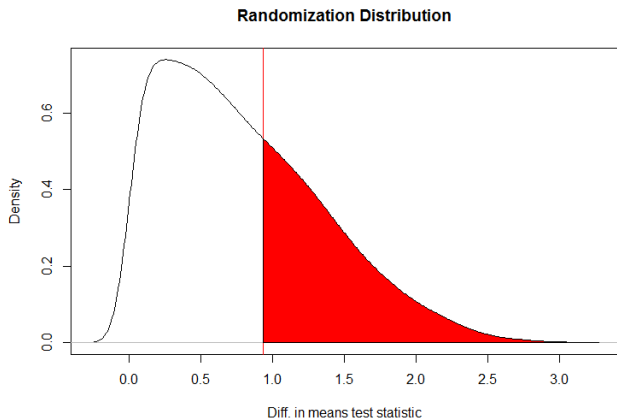
Randomization Distribution

Diff. in means test statistic

Fisher vs. frequentist p-values

For the same test statistic, are you more likely to reject the null based on a Fisher exact p-value or a frequentist one?

Fisher vs. frequentist p-values

For the same test statistic, are you more likely to reject the null based on a Fisher exact p-value or a frequentist one?

The sharp null under the Fisher approach is much more restrictive, so it's easier to find evidence to reject it

Fisher vs. frequentist p-values

For the same test statistic, are you more likely to reject the null based on a Fisher exact p-value or a frequentist one?

The sharp null under the Fisher approach is much more restrictive, so it's easier to find evidence to reject it

```
# Fisher exact p-value:
mean(test.stat <= all.test.stats)
[1] 0.3633256

# Frequentist diff of means estimator:
t.test(data$y.obs[data$treat == 1],
       data$y.obs[data$treat == 0])$p.value
[1] 0.3780406
```