Lecture 2: Experimental Methods
Randomization Inference

Raymond Duch

May 2, 2018

Director CESS Nuffield/Santiago

Lecture 2

- Hypothesis testing
- Confidence bounds
- Block random assignment

Observed Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30

Potential Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1	10	15	5
Village 2	15	15	0
Village 3	20	30	10
Village 4	20	15	-5
Village 5	10	20	10
Village 6	15	15	0
Village 7	15	30	15
Average	15	30	15

Table 3.1 Sampling distribution of estimated ATEs generated when two of the seven villages listed in Table 2.1 are assigned to treatment

	Estimated ATE	Frequency with which an estimate occurs
	-1	2
	0	2
	0.5	1
	1	2
	1.5	2
	2.5	1
	6.5	1
	7.5	3
	8.5	3
	9	1
	9.5	1
	10	1
	16	1
Average	5	
Total		21

Example

- Based on the numbers in Table 3.1, we calculate the standard error as follows:

Sum of squared deviations

$$
\begin{aligned}
&= (-1-5)^2 + (-1-5)^2 + (0-5)^2 + (0.5-5)^2 + (1-5)^2 + (1-5)^2 \\
&+ (1.5-5)^2 + (1.5-5)^2 + (2.5-5)^2 + (6.5-5)^2 + (7.5-5)^2 + (7.5-5)^2 \\
&+ (7.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (9-5)^2 + (9.5-5)^2 \\
&+ (10-5)^2 + (16-5)^2 = 445
\end{aligned}
$$

$$
\textit{Square root of the average squared deviation} = \sqrt{\frac{1}{21}(445)} = 4.60
$$

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left\{\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1)\right\}}$$

$$SE(\widehat{ATE}) = \sqrt{\frac{5}{1}6\{\frac{(2)(14.29)}{5} + \frac{(5)(42.86)}{2} + (2)(7.14)\}} = 4.60$$

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i(0))}{N-m} + \frac{\widehat{Var}(Y_1(1))}{m}}$$

Hypothesis Testing

- We can test certain conjectures that provide us a complete schedule of potential outcomes
- One such conjecture is the *sharp null hypothesis* that the treatment effect is zero for all observations
- Under this hypothesis, $Y_i(1) = Y_i(0)$, in which case we observe *both* potential outcomes for every observation
- Simulated randomizations provide an exact sampling distribution of the estimated average treatment effect under the sharp null hypothesis

Observed Outcome Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30

Example: Randomization

- From this table generate an estimate of the ATE of 6.5
- How likely are we to obtain estimate as large as or larger than 6.5 if the true effect were zero for all observations?
- The probability, or p-value, of interest in this case addresses a *one-tailed hypothesis*, namely that female village council heads increase budget allocations to water sanitation
- All applied analyses will be conducted in R

Example: Randomization

- Based on the observed outcome in the table, we may calculate the 21 possible estimates of the ATE that could have been generated if the null hypothesis were true: {-7.5, -7.5, -7.5, -4.0, -4.0, -4.0, -4.0, -4.0, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, 3.0, 3.0, 6.5, 6.5, 6.5, 10.0, 10.0}
- How likely are we to obtain an estimate as large as or larger than 6.5 if the true effect were zero for all observations?

Example: 1-tailed test

- The probability, or p-value, of interest in this case addresses a *one-tailed hypothesis*, namely that female village council heads increase budget allocations to water sanitation
- Five of the estimates are as large as 6.5. Therefore, when evaluating the one-tailed hypothesis that female village heads *increase* water sanitation budgets, we would conclude that the probability of obtaining an estimate as large as 6.5 if the null hypothesis were true is $5/21 = 24\%$

Example: 2-tailed test

- If we sought to evaluate the *two-tailed hypothesis* - whether female village council heads either increase or decrease the budget allocation for water sanitation
- We would calculate the p-value of obtaining a number that is greater than or equal to 6.5 or less than or equal to -6.5. A two-tailed hypothesis test would count all instances in which the estimates are at least as great as 6.5 *in absolute value*. Eight of the estimates qualify, so the two-tailed p-value is $8/21 = 38\%$

Lady Testing Tea

- Ronald Fisher, *The Design of Experiments*
- Randomized Tea Experiment:
  - 8 identical cups prepared
  - 4 cups randomly prepared with milk first
  - 4 cups randomly prepared with milk after
- Lady correctly classified all cups!

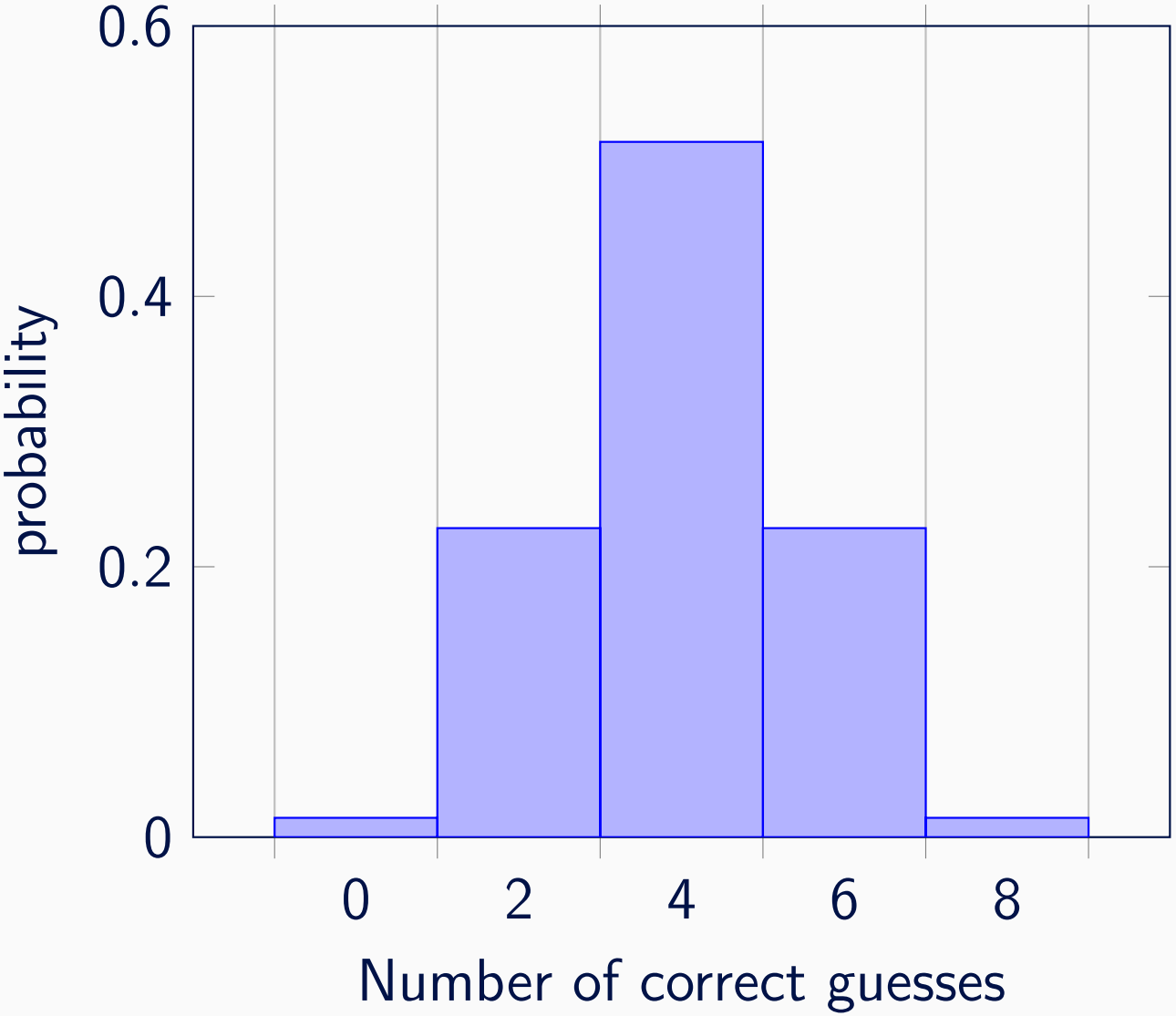
Lady Tasting Tea

cups	lady's guess	actual order	scenarios			...
1	M	M	T	T	T	
2	T	T	T	T	M	
3	T	T	T	T	M	
4	M	M	T	M	M	
5	M	M	M	M	T	
6	T	T	M	M	T	
7	T	T	M	T	M	
8	M	M	M	M	T	
number of correct guesses		8	4	6	2	...

Lady Tasting Tea

```
> ## truth: enumerate the number of assignment combinations

> true <- c(choose(4, 0) * choose(4, 4), choose(4, 1) *
    choose(4, 3), choose(4, 2) * choose(4, 2), choose(4, 3)
    * choose(4, 1), choose(4, 4) * choose(4, 0))

> ## compute probability: divide it by the total number of
    events

> true <- true/sum(true)

> ## number of correctly classified cups as labels

> names(true) <- c(0,2,4,6,8)

> true
        0         2         4         6         8
0.01428571 0.22857143 0.51428571 0.22857143 0.01428571
```

Lady Tasting Tea



Lady Tasting Tea: Simulate

```
> ## Simulations

> sims <- 1000

> guess <- c("M", "T", "T", "M", "M", "T", "T", "M") # lady's guess

> correct <- rep(NA, sims) # place holder for number of correct guesses

> for (i in 1:sims) {
+   cups <- sample(c(rep("T", 4), rep("M", 4)), replace = FALSE)
+   correct[i] <- sum(guess == cups) # number of correct guesses
}

> ## comparison
> prop.table(table(correct)) - true
correct
           0            2            4            6            8
 0.001714286  0.004428571 -0.015285714  0.007428571  0.001714286
```

Fisher Exact Test

```
> ## rows: actual assignments
> ## columns: reported guesses

> ## all correct
> x <- matrix(c(4, 0, 0, 4), byrow = TRUE, ncol = 2, nrow = 2)

> ## six correct
> y <- matrix(c(3, 1, 1, 3), byrow = TRUE, ncol = 2, nrow = 2)

> rownames(x) <- colnames(x) <- rownames(y) <- colnames(y) <- c("M",    T")

>␣x
 ␣␣M␣T
M␣4␣0
T␣0␣4
>␣y
 ␣␣M␣T
M␣3␣1
T␣1␣3
```

Fisher Exact Test

```
> fisher.test(x, alternative = "greater") # one-
    sided
```

Fisher's Exact Test for Count Data

data: x
p-value = 0.01429
alternative hypothesis: true odds ratio is greater than 1 95 percent
confidence interval:
2.003768 Inf
sample estimates:
odds ratio
Inf

Fisher Exact Test

```
> fisher.test(x, alternative = "greater") # one-
    sided
```

Fisher's Exact Test for Count Data

data: y
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1 95 percent
confidence interval:
0.3135693 Inf
sample estimates:
odds ratio
6.408309

Fisher Exact Test

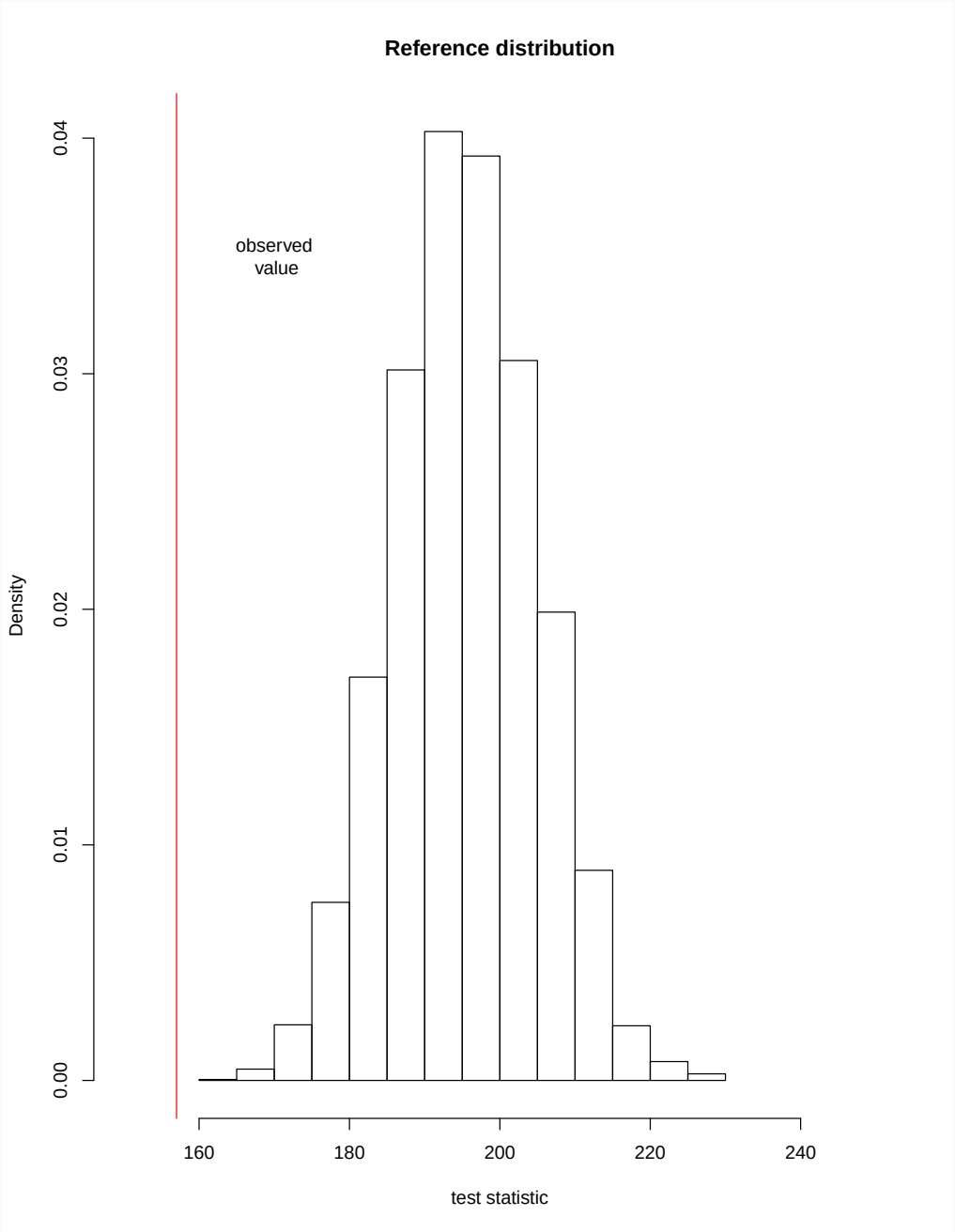
```
> fisher.test(x) # two-sided
```

Fisher's Exact Test for Count Data

data: x
p-value = 0.02857
alternative hypothesis: true odds ratio is greater than 1 95 percent
confidence interval:
1.339059 Inf
sample estimates:
odds ratio
Inf

Resume Experiment

```
> setwd("~/Dropbox/Experimental Methodology/DPIR 2017/qss-master/CAUSALITY")
> resume <- read.csv("resume.csv")
> sims <- 5000
> z <- rep(NA, sims)
> for (i in 1:sims) {
+     ## shuffles treatment assignment
+     treat <- sample(resume$race, replace = FALSE)
+     ## test statistic
+     z[i] <- sum(resume$call[treat == "black"])
+ }
> ## observed z
> z.obs <- sum(resume$call[resume$race == "black"])
> z.obs
[1] 157
> ## one-sided p-value; proportion of simulated draws less than observed value
> pvalue <- mean(z[i] <= z.obs)
> pvalue
[1] 0
> ## histogram of reference distribution
> hist(z, freq = FALSE, xlim = c(150, 250),
+      xlab = "test statistic", main = "Reference distribution")
> abline(v = z.obs, col = "red")
> text(170, 0.035, "observed\n value")
```


Reference distribution
observed
value
Density
0.00
0.01
0.02
0.03
0.04
160
180
200
220
240
test statistic

Commments on Randomization

- One obtains arbitrarily precise p-values without relying on distributional assumptions
- The same method can be used for a wide variety of applications and test statistics (e.g., the difference-in-means estimator, regression, difference-in-variance, etc.)
- It forces the researcher to take a moment to think carefully about what the null hypothesis is and how it should be tested

Confidence Intervals

- We cannot estimate the dispersion of the estimates without making simplifying assumptions
- A simple approach is to assume that the treatment effect for every subject is equal to the estimated ATE
- For subjects in the control condition, missing $Y_i(1)$ values are imputed by adding the estimated ATE to the observed values of $Y_i(0)$
- For subjects in the treatment condition, missing $Y_i(0)$ values are imputed by subtracting the estimated ATE from the observed values of $Y_i(1)$
- This approach yields a complete schedule of potential outcomes, which we may then use to simulate all possible random allocations

Effect of winning visa lottery on attitudes toward people from other countries

- We cannot estimate the dispersion of the estimates without making simplifying assumptions
- Winners and losers were asked to rate the Saudi, Indonesian, Turkish, African, European, and Chine people on a five-point scale ranging from very negative (-2) to very positive (+2)
- Adding the responses to all six items creates an index ranging from -12 to +12
- Average in the treatment group is 2.34
- 1.87 in the control group

Pakistani Muslims Lottery

Ratings of people from other countries	Control(%)	Treatment (%)
-12	0	0.2
-9	0.22	0
-8	0	0.2
-6	0.45	0.2
-5	0	0.2
-4	0.45	0.59
-3	0	0.2
-2	1.12	0.98
-1	1.56	2.75
0	27.23	18.63
1	18.3	13.14
2	24.33	25.29
3	8.48	10.98
4	5.8	9.61
5	3.35	3,92
6	3.79	7.25
7	2.23	2.55
8	0.89	1.37
9	0.22	0.78
10	0.45	0
11	0.67	0.2
12	0.45	0.98
TOTAL	100	100
N	(448)	(510)

Estimate our 95% interval

- We add 0.47 to the observed outcomes in the control group in order to approximate their values
- We subtract 0.47 for the treatment group's observed outcomes in order to approximate their values
- Simulating 100,000 random allocations using this schedule of potential outcomes and sorting the estimated ATEs in ascending order
- We find that the 2,500th estimate is 0.16 and the 97,501st estimate is 0.79, so the 95% interval is [0.16, 0.79]

```r
##################################################
# R Starter Code - Randomization Inference
# POLS 4368 Section - Feb 12 2013
##################################################

## Load the RI package

# install.packages("ri", dependencies=TRUE)
library(ri)
set.seed(1234567)

########################################
## Generate data, or read-in data
########################################
N <- 50
m <- 25

d <- ifelse(1:N %in% sample(1:N, m), 1, 0)
Y0 <- runif(N,0,1)
Y1 <- Y0 + rnorm(N,2,2)
Y <- Y1*d + Y0*(1-d)

cbind(Y0,Y1,d,Y)          # look at your data

## Conduct analysis of actual experiment
## Estimate the ATE

# nonparametric
mean(Y[d==1]) - mean(Y[d==0])

# or fitting data to ols
lm(Y~d)
```

```
# Define inputs (Z, Y, any blocking variable, or pre-treatment variables)
# Z must be a binary variable 0=control, 1=treatment
Z <- d
probs <- genprobexact(Z)
ate <- estate(Y,Z,prob=probs)

# Set the number of simulated random assignments
perms <- genperms(Z,maxiter=10000)
# Create potential outcomes UNDER THE SHARP NULL OF NO EFFECT FOR ANY UNIT
Ys <- genouts(Y,Z,ate=0)

# Generate the sampling distribution based on schedule of potential outcome
# implied by the sharp null hypothesis
distout <- gendist(Ys,perms,prob=probs)

ate                                  # estimated ATE
sum(distout >= ate)                  # one-tailed comparison used to calculate p-
    value (greater than)
sum(abs(distout) >= abs(ate))        # two-tailed comparison used to calculate p-
    value

dispdist(distout,ate)                # display p-values, 95% confidence interval,
     standard error under the null, and graph the sampling distribution under
    the null
#-------------------------------------------------------------------------
# estimation of confidence intervals assuming ATE=estimated ATE
#-------------------------------------------------------------------------
Ys <- genouts(Y,Z,ate=ate)           # create potential outcomes UNDER THE
    ASSUMPTION THAT ATE=ESTIMATED ATE
distout <- gendist(Ys,perms,prob=probs)  # generate the sampling distribution
    based on the schedule of potential outcomes implied by the null hypothesis
dispdist(distout,ate)                # display p-values, 95% confidence interval,
     standard error under the null, and graph the sampling distribution under
    the null
```

Example from Matthew Blackwell

- Suppose we are targeting 6 people for donations to Harvard.
- As an encouragement, we send 3 of them a mailer with inspirational stories of learning from our graduate students.
- Afterwards, we observe them giving between \$0 and \$5.
- Simple example to show the steps of RI in a concrete case.

Randomization Distribution

	Mailer	Contrib		
Unit	**D**	**Y**	**Y(0)**	**Y(1)**
Donald	1	3	(3)	3
Carly	1	5	(5)	5
Ben	1	0	(0)	0
Ted	0	4	4	(4)
Marco	0	0	0	(0)
Scott	0	1	1	(1)

T(diff)=|8/3 - 5/3| = 1

Randomization Distribution

	Mailer	Contrib		
Unit	D	**Y**	**Y(0)**	**Y(1)**
Donald	1	3	(3)	3
Carly	1	5	(5)	5
Ben	0	0	(0)	0
Ted	1	4	4	(4)
Marco	0	0	0	(0)
Scott	0	1	1	(1)

T(diff)=|12/3 - 1/3| = 3.67

T(diff)=|8/3 - 5/3| = 1

T(diff)=|9/3 - 4/3| = 1.67

Randomization Distribution

D1	D2	D3	D4	D5	D6	| Diff in means |
1	1	1	0	0	0	1.00
1	1	0	1	0	0	3.67
1	1	0	0	1	0	1.00
1	1	0	0	0	1	1.67
1	0	1	1	0	0	0.33
1	0	1	0	1	0	2.33
1	0	1	0	0	1	1.67
1	0	0	1	1	0	0.33
1	0	0	1	0	1	1.00
1	0	0	0	1	1	1.67
0	1	1	1	0	0	1.67
0	1	1	0	1	0	1.00
0	1	1	0	0	1	0.33
0	1	0	1	1	0	1.68

In R: ri

```
library(ri)
y <- c(3, 5, 0, 4, 0, 1)
D <- c(1, 1, 1, 0, 0, 0)
T_stat <- abs(mean(y[D == 1]) - mean(y[D == 0]))
Dbold <- genperms(D)
Dbold[, 1:6]
```

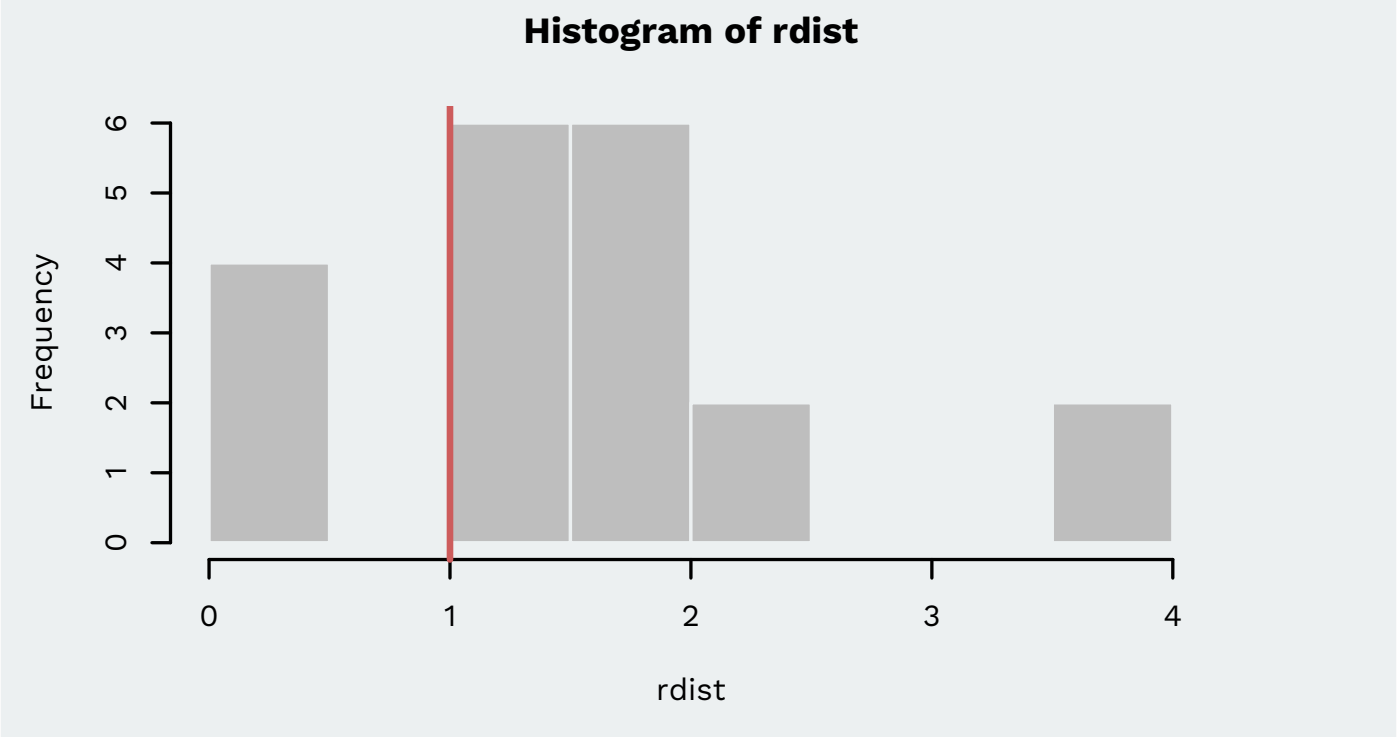
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
1	1	1	1	1	1	1
2	1	1	1	1	0	0
3	1	0	0	0	1	1
4	0	1	0	0	1	0
5	0	0	1	0	0	1
6	0	0	0	1	0	0

In R: ri

```
rdist <- rep(NA, times = ncol(Dbold))
for (i in 1:ncol(Dbold)) {
    D_tilde <- Dbold[, i]
    rdist[i] <- abs(mean(y[D_tilde == 1]) - mean(y[D_tilde ==
        0]))
}
rdist
```

```
[1] 1.0000000 3.6666667 1.0000000 1.6666667 0.3333333
2.3333333 1.6666667 [8] 0.3333333 1.0000000 1.6666667
1.6666667 1.0000000 0.3333333 1.6666667 [15] 2.3333333
0.3333333 1.6666667 1.0000000 3.6666667 1.0000000
```

P-value




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

[1] 0.8