Section 4 : Matching

Andrew Bertoli

25 September 2013

Roadmap

1. Matching
2. Questions
3. Testing Multiple Hypotheses
4. P-Hacking

Matching

Assumptions

Different matching designs require different assumptions. However, to estimate most sample parameters, we must assume

1. $Y_{0,1} \perp T|X$
2. $0 < P(T = 1|X = x) < 1$ for all $x \in X$

The first assumption alone implies Ignorability. Both assumptions together imply Strong Ignorability.

Matching

If you just care about the treated units, you can assume

1. $Y_0 \perp T|X$
2. $P(T = 1|X = x) < 1$ for all $x \in X$

Matching

Important Decisions

1. Should we match with or without replacement?
2. What distance metric should we use?
3. How close must two units be for us to count them as a reasonable match?
4. How should we deal with ties?

Matching

Matching With or Without Replacement

When we match with replacement, we match every treated unit to the nearest control unit (assuming we are interested in the ATT). In theory, we could match every treated unit to the same control unit.

When we match without replacement, we allow each control unit to be used at most once. This constraint can lead to some really poor matches.

The choice comes down to a bias-variance tradeoff. Matching with replacement should reduce bias, but it might increase the variance of our estimator if a few control units are matched to many treated units.

The best choice is almost always matching with replacement. If you do end up matching a few control units to many treated units, it probably indicates a support problem.

Matching

Distance Metrics

1. Univariate Matching

a) $d_{k,j} = (X_k - X_j)^2$ (Squared Difference)

b) $d_{k,j} = |X_k - X_j|$ (Absolute Difference)

2. Multivariate Matching

a) Exact match on each covariate

b) Propensity Score

c) Other Ways (to be discussed later)

Matching

Exact Matching

1. Covariates must be discrete
2. You can only match on a few factors (curse of dimensionality)
3. Works when you have a small number of controls

Matching

Propensity Score

Assuming that we do not know the real propensity score, we must estimate it from the covariates.

We will use a model like this one from last class

> pscore=glm(Treat Age + Gender + Parents.Eaters, family=binomial(link=logit),data=data)$fitted.values

Be sure to test that the covariates are balanced after matching. If they are not, then there is probably something wrong with the propensity score.

Matching

Dealing with Poor Matches

There will sometimes be treated units that do not have good matches from the control group.

When this happens, it sometimes makes sense to drop treated units these units.

We set a caliper and drop all matches where the distances between the units is greater than the caliper.

When we discard matches, this means we are changing our parameter of interest from the ATT to the treatment effect for treated units with adequate controls.

Matching

Ties

Ties may sometimes occur when we are dealing with discrete covariates.

There are two options

1. Flip a coin

2. Match both control units to the treated unit, but give each of these controls a weight of $1/2$

The second option is preferable because flipping a coin increases the variance of the estimator. If you flip a coin, your estimate of the standard error will be wrong.

Matching

Advantages of Matching

1. Separates design from analysis (in theory)

a) Researchers can spend a lot of time on matching but test the outcome only once.

b) However, researchers can still modify their designs after observing the results.

2. Achieving optimal balance is an algorithmic problem

3. Makes it easier to see when there is a clear lack of support in the data

Matching

Example

Recall the question of whether electronic voting machines in several counties in Florida caused Bush to receive more votes.

Let's compare regression analysis to matching.

Matching

```
> glm2 <- glm(b04pc ~ etouch + b00pc + b00pc_sq + d96pc1 + v_change +
+             income + hispanic +  b00pc_e + b00pcsq_e, data=dta)
> summary(glm2)

Call:
glm(formula = b04pc ~ etouch + b00pc + b00pc_sq + d96pc1 + v_change +
    income + hispanic + b00pc_e + b00pcsq_e, data = dta)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.05936  -0.01192   0.00058   0.01116   0.04883

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.178e-01  9.334e-02  -2.333   0.0232 *
etouch       3.992e-01  1.457e-01   2.740   0.0082 **
b00pc        2.043e+00  3.194e-01   6.398 3.19e-08 ***
b00pc_sq    -6.775e-01  2.783e-01  -2.434   0.0181 *
d96pc1      -1.531e-01  1.163e-01  -1.317   0.1932
v_change    -1.584e-07  1.313e-07  -1.206   0.2328
income      -7.725e-07  7.475e-07  -1.033   0.3058
hispanic    -5.940e-02  2.829e-02  -2.100   0.0402 *
b00pc_e     -1.244e+00  5.480e-01  -2.269   0.0271 *
b00pcsq_e    9.178e-01  5.146e-01   1.784   0.0798 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.0004512571)
```

Matching

```
> rr <- Match(Y=Y, Tr=Tr, X=cbind(dta$b00pc, dta$b00pc_sq, dta$d96pc1, dta$v_change,
+ dta$income, dta$hispanic, dta$b00pc_e, dta$b00pcsq_e), estimand="ATT")
> summary(rr)

Estimate...  -0.014285
AI SE......  0.020608
T-stat.....  -0.69317
p.val......  0.4882

Original number of observations..............  67
Original number of treated obs...............  15
Matched number of observations...............  15
Matched number of observations  (unweighted).  15
```

Matching

Problems of Matching

1. Potential for bias

2. Creates opportunities for dishonest research

Matching

Pop Quiz

Recall the study about fast food and heart disease.

The true model is

$\pi_i = \frac{1}{300} \cdot Age_i + 0.1 \cdot Gender_i + 0.2 \cdot Parents.Eaters_i + 0.3 \cdot Stress_i$

P(H. Disease)$= 0.3 \cdot Treat + \frac{1}{400} \cdot Age + 0.05 \cdot Gender + 0.1 \cdot Stress$

Since we don't observe stress, we estimate the propensity score by using the model

```
> pscore=glm(Treat  Age + Gender + Parents.Eaters, family=
binomial(link=logit),data=data)$fitted.values
```

Question: What is the direction of the bias in our estimator?

Answer: Our estimator is biased upwards.

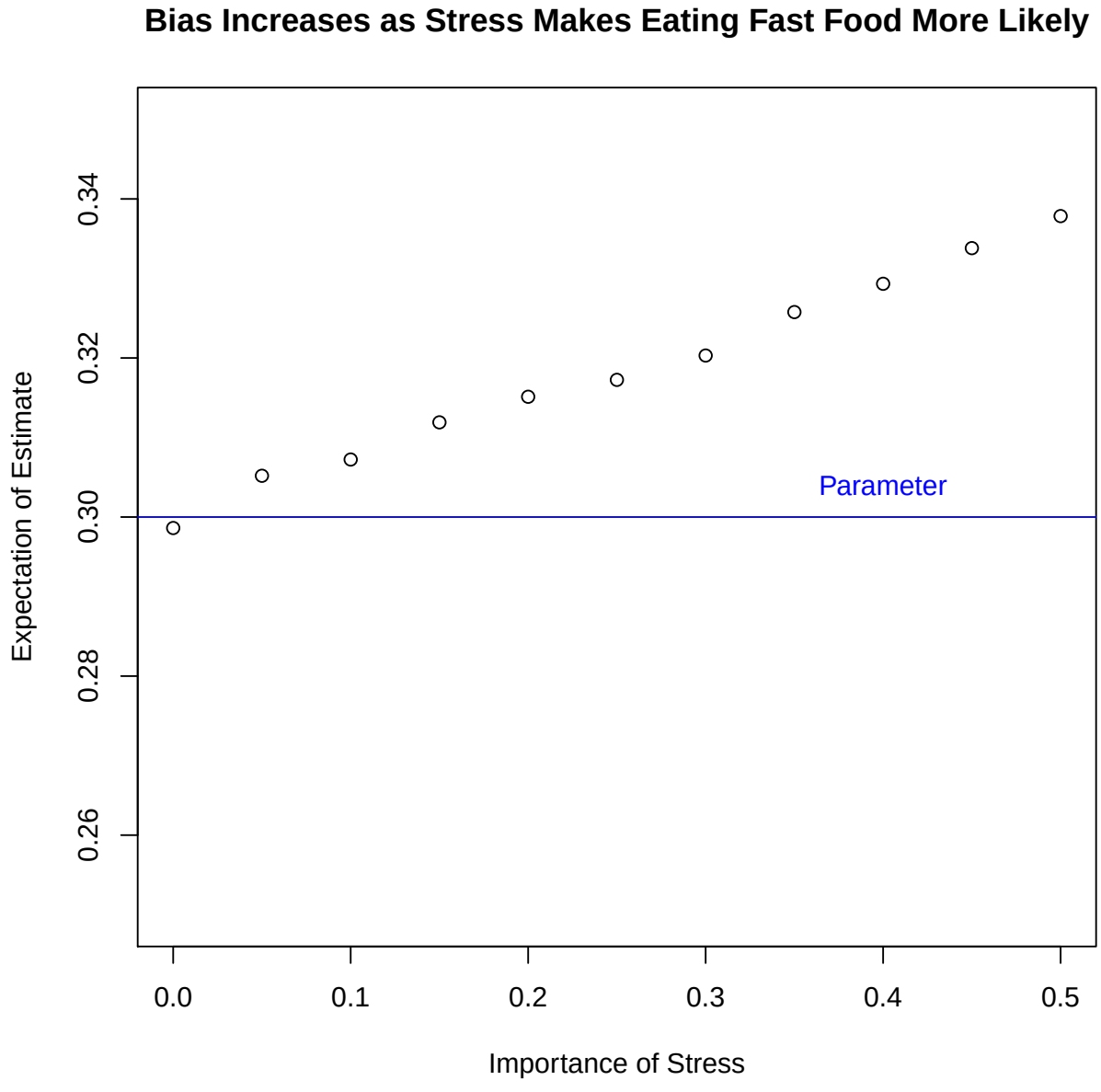
Matching

Pop Quiz

Question: What happens to the bias of our estimator as the impact of stress on the probability of eating fast food every day increases.

Answer: The bias of our estimator increases.

Matching


Bias Increases as Stress Makes Eating Fast Food More Likely
Expectation of Estimate
0.26
0.28
0.30
0.32
0.34
Parameter
0.0
0.1
0.2
0.3
0.4
0.5
Importance of Stress

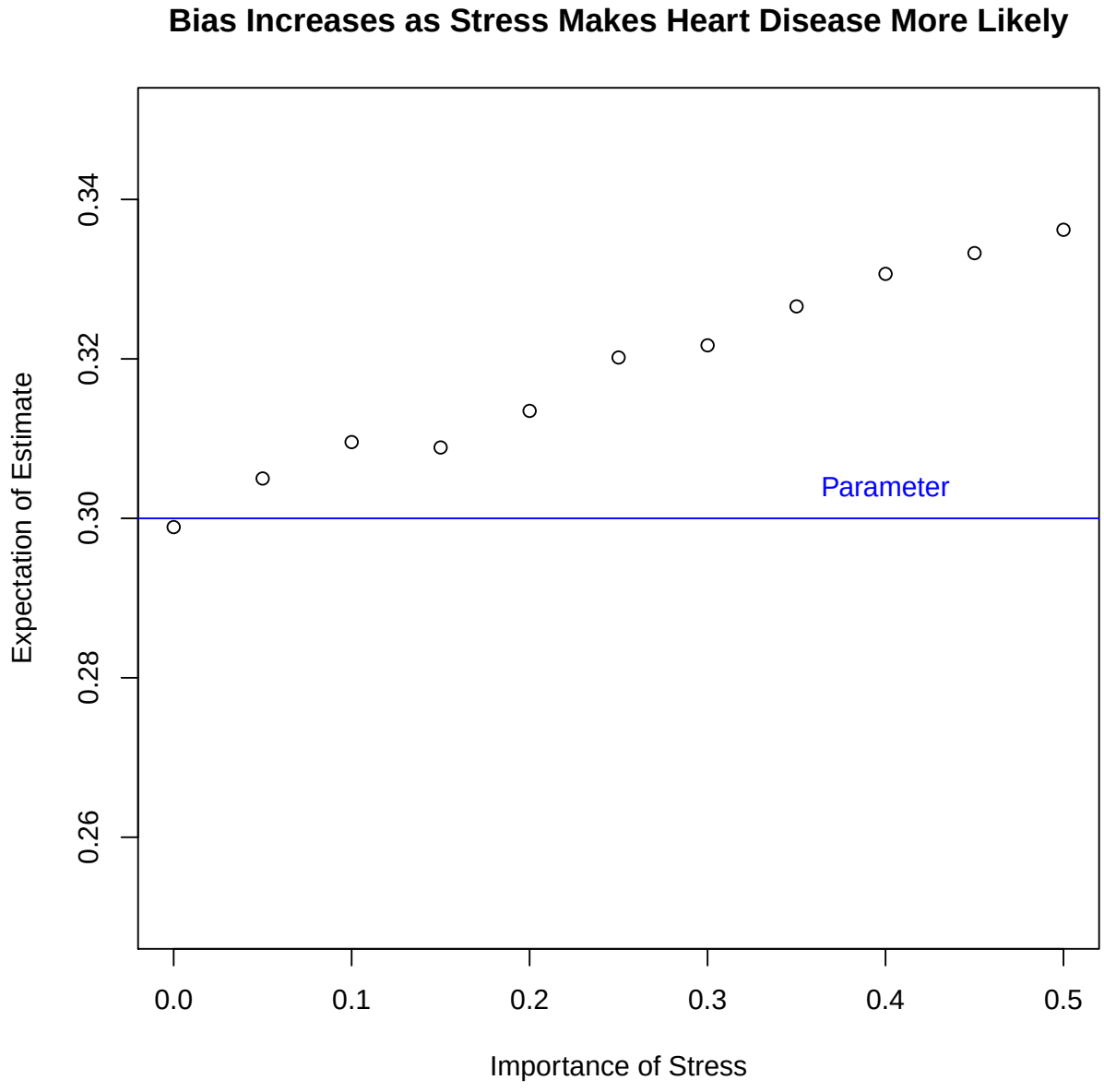
Matching

Pop Quiz

Question: What happens to the bias of our estimator as the impact of stress on the probability of getting heart disease increases.

Answer: The bias of our estimator increases.

Matching


Bias Increases as Stress Makes Heart Disease More Likely
Expectation of Estimate
0.26
0.28
0.30
0.32
0.34
Parameter
0.0
0.1
0.2
0.3
0.4
0.5
Importance of Stress

Matching

Results depend on what you choose to control for.

P-Values After Matching Under No Treatment Effect

Age	Gender	Parents.Eaters	p-value
			0.69
X			0.062
	X		1
		X	0.00
X	X		0.13
X		X	0.062
	X	X	1
X	X	X	0.13

Questions

1. Lecture
2. Readings
3. Homeworks

Testing Multiple Hypotheses

Say we are running a number of hypothesis tests.

For any given test, there is a 5% probability that the results will be significant without a treatment effect.

This means that if there is no real treatment effect and we run 13 independent hypothesis tests, there is more than a 50% chance that at least one will be significant at the 5% level.

Bottom Line: We need some way to make our tests more conservative to make sure our false positive rate does not rise above 5%.

Testing Multiple Hypotheses

Approach 1: The Bonferroni Method

If we have m hypothesis tests, we only reject the null when $p_i < 0.05/m$.

This is the same as multiplying all our p-values by the number of hypothesis tests we run.

Example: If we run 10 hypothesis tests, we only reject the null when $p < 0.005$

Problem: This method is too conservative. If we run 100 hypothesis tests, the standard of evidence needed to reject a null will be extremely high.

Testing Multiple Hypotheses

Approach 2: The Benjamini-Hochberg (BH) Method

Order the m p-values from smallest to highest, $p_{(1)}, ..., p_{(m)}$

Define

$\ell_i = \frac{0.05(i)}{C_m m}$

where $C_m = 1$ if the tests are independent and $C_m = \sum_{i=1}^{m} 1/i$ otherwise.

For each $p_{(i)}$, check if $p_{(i)} < \ell_i$. We take the largest p-value where this inequality holds, and we reject it and all p-values smaller than it.

Testing Multiple Hypotheses

Example: Say we do 5 hypothesis tests and get the following p-values:

(1) 0.015, (2) 0.008, (3) 0.56, (4) 0.039, and (5) 0.014

The Bonferroni Method

Reject only the hypothesis tests where $p < 0.05/5 = 0.01$.

Thus, we only reject (2) p=0.008.

Testing Multiple Hypotheses

Example: Say we do 5 hypothesis tests and get the following p-values:

(1) 0.015, (2) 0.008, (3) 0.56, (4) 0.039, and (5) 0.014

The Benjamini-Hochberg (BH) Method (Independent Tests)

First, reorder the p-values

(1*) 0.008, (2*) 0.014, (3*) 0.015, (4*) 0.039, (5*) 0.56

So $\ell_1 = \frac{0.05(1)}{1*5} = 0.01$, $\ell_2 = \frac{0.05(2)}{1*5} = 0.02$, $\ell_3 = 0.03$, $\ell_4 = 0.04$, $\ell_5 = 0.05$

So we would reject (1*), (2*), (3*), and (4*).

Testing Multiple Hypotheses

Example: Say we do 5 hypothesis tests and get the following p-values:

(1) 0.015, (2) 0.008, (3) 0.56, (4) 0.039, and (5) 0.014

The Benjamini-Hochberg (BH) Method (Dependent Tests)

First, reorder the p-values

(1*) 0.008, (2*) 0.014, (3*) 0.015, (4*) 0.039, (5*) 0.56

So $\ell_1 = \frac{0.05(1)}{\sum_{i=1}^{5} \frac{1}{i} * 5} = 0.004$, $\ell_2 = \frac{0.05(2)}{\sum_{i=1}^{5} \frac{1}{i} * 5} \approx 0.008$, $\ell_3 \approx 0.013$,
$\ell_4 \approx 0.018$, $\ell_5 \approx 0.022$

So we would reject nothing.

Testing Multiple Hypotheses

Problems

1. Usually impossible to know if researchers are reporting all their hypothesis tests

2. Often many researchers working on one data set, and there is no correction for the aggregate number of hypothesis tests

P-Hacking

Basic Idea: Researchers run tests until their p-values are below 0.05 or 0.01, and then report only the results for significant tests with no correction for multiple testing.

Very common for researchers who use regression or matching, since they can choose their control variables selectively.

Can also be a problem in natural experiments when researchers have the freedom to make choices about their designs, like where to set a regression discontinuity window.

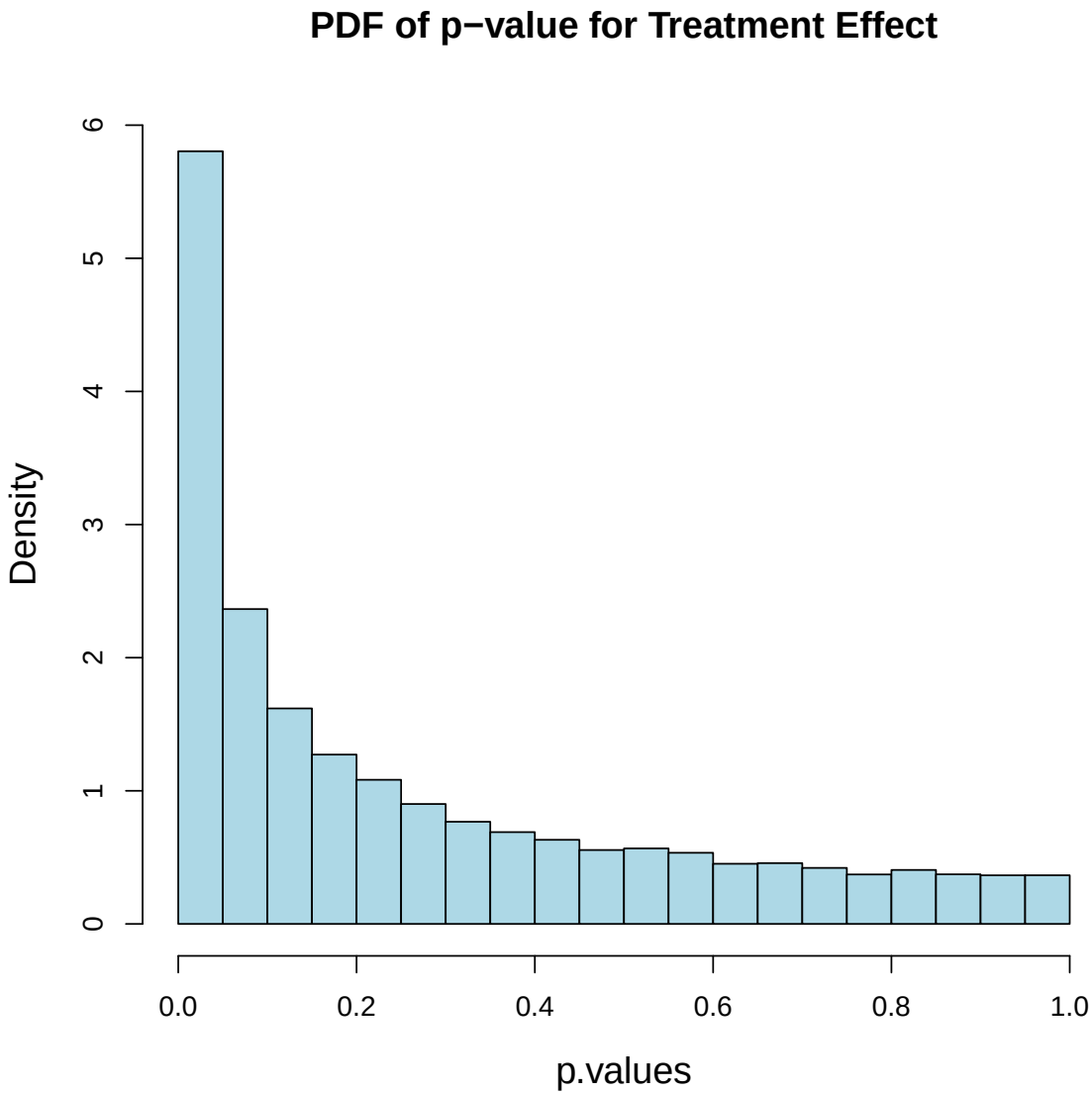
P-Hacking

Question: What is the distribution of a p-value when there is a treatment effect?

P-Hacking

```
p.values=rep(0,50000)

for(i in 1:50000){

t=rnorm(100,1,5)

c=rnorm(100,0,5)

p.values[i]=t.test(t,c)$p.value

}

hist(p.values, freq=FALSE, ylim=c(0,6), main="PDF of p-value
for Treatment Effect", cex.lab=1.3, cex.main=1.3,
col="lightblue")
```

P-Hacking


PDF of p−value for Treatment Effect
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
p.values

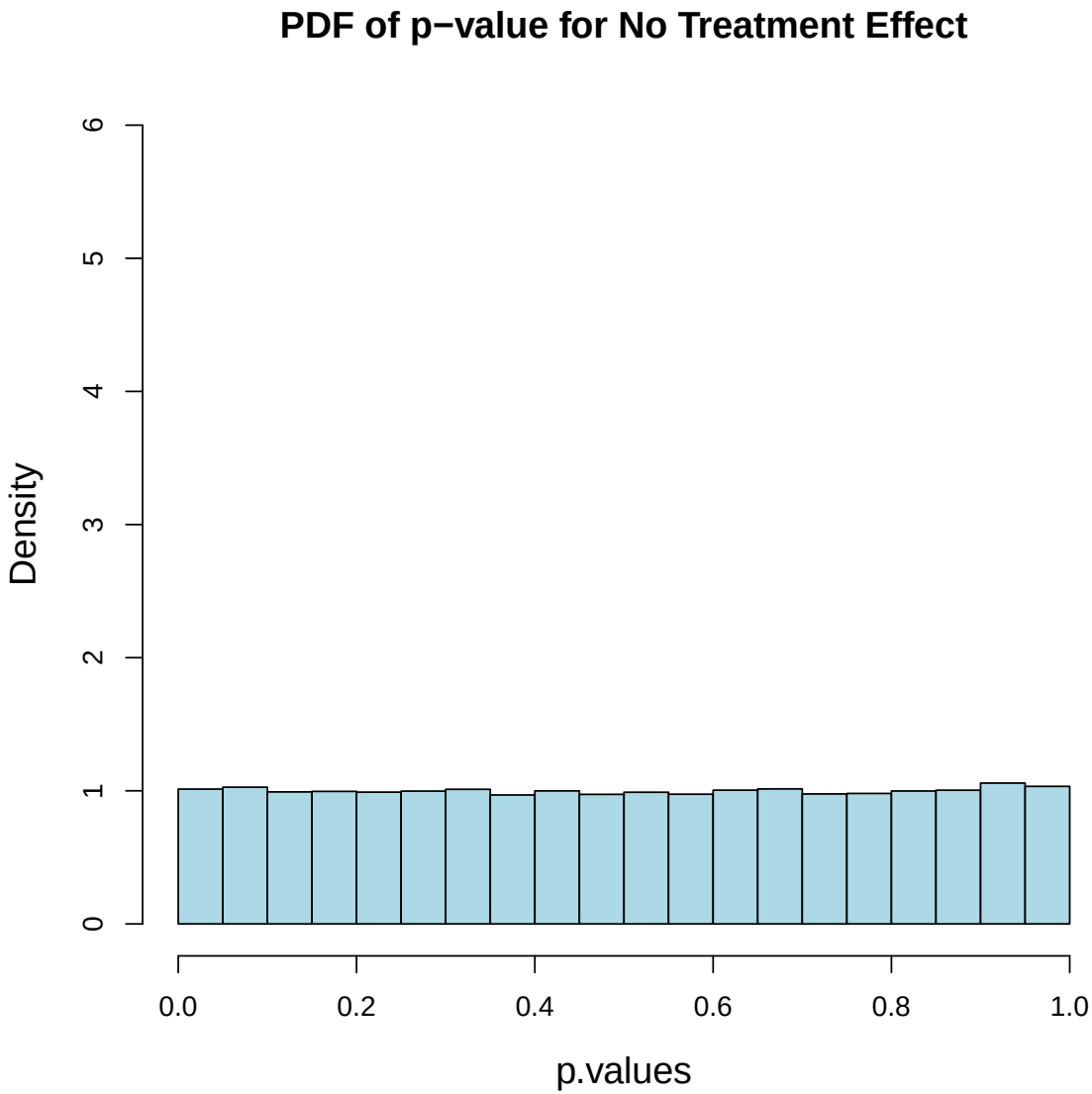
P-Hacking

Question: What is the distribution of a p-value when there is no treatment effect?

P-Hacking

```
p.values=rep(0,50000)

for(i in 1:50000){

t=rnorm(100,0,5) # No treatment effect

c=rnorm(100,0,5)

p.values[i]=t.test(t,c)$p.value

}

hist(p.values, freq=FALSE, ylim=c(0,6), main="PDF of p-value
for No Treatment Effect", cex.lab=1.3, cex.main=1.3,
col="lightblue")
```

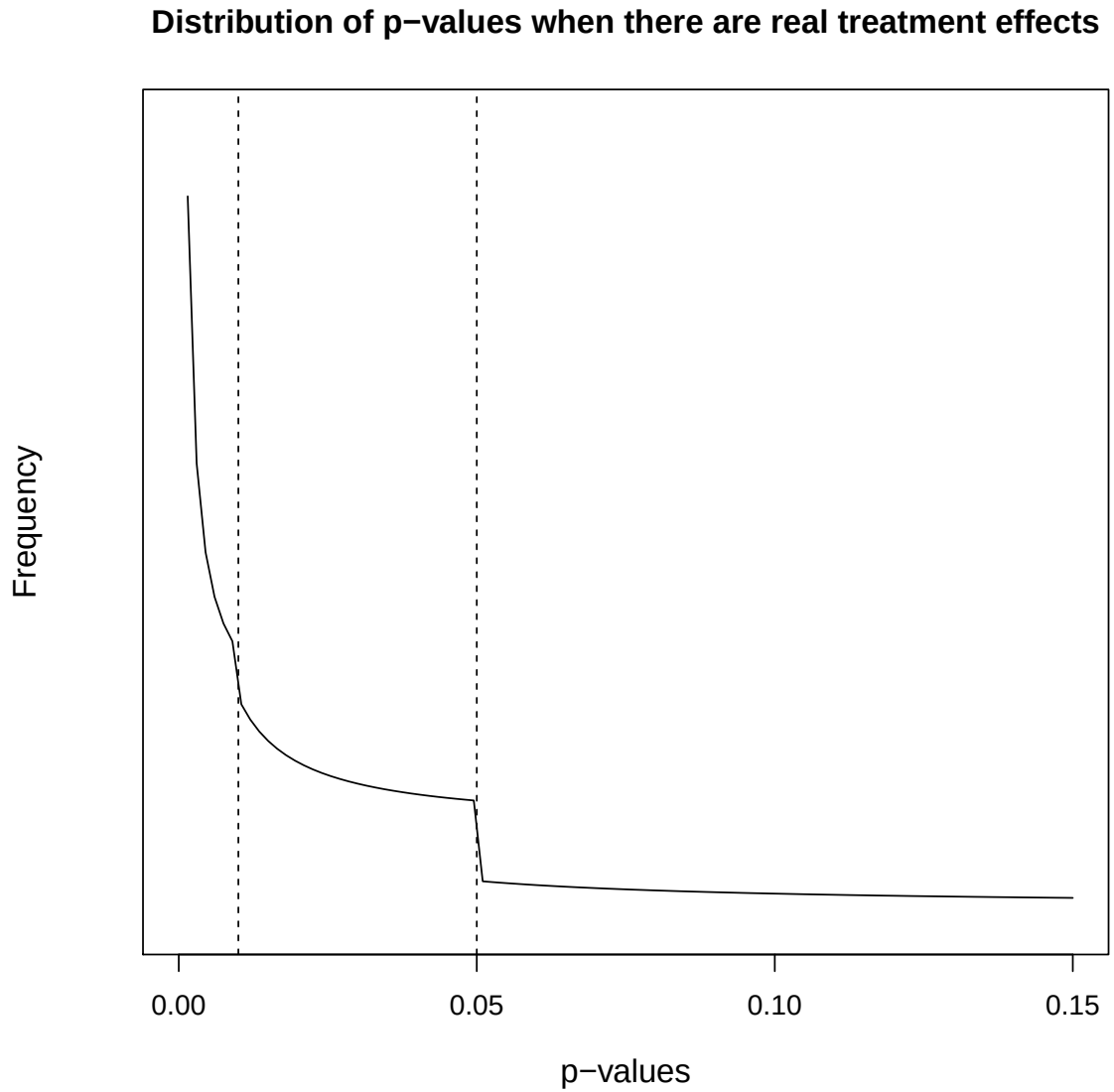
P-Hacking


PDF of p−value for No Treatment Effect
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
p.values

P-Hacking

Question: What is the distribution of p-values in journal articles when researchers are finding real treatment effects?

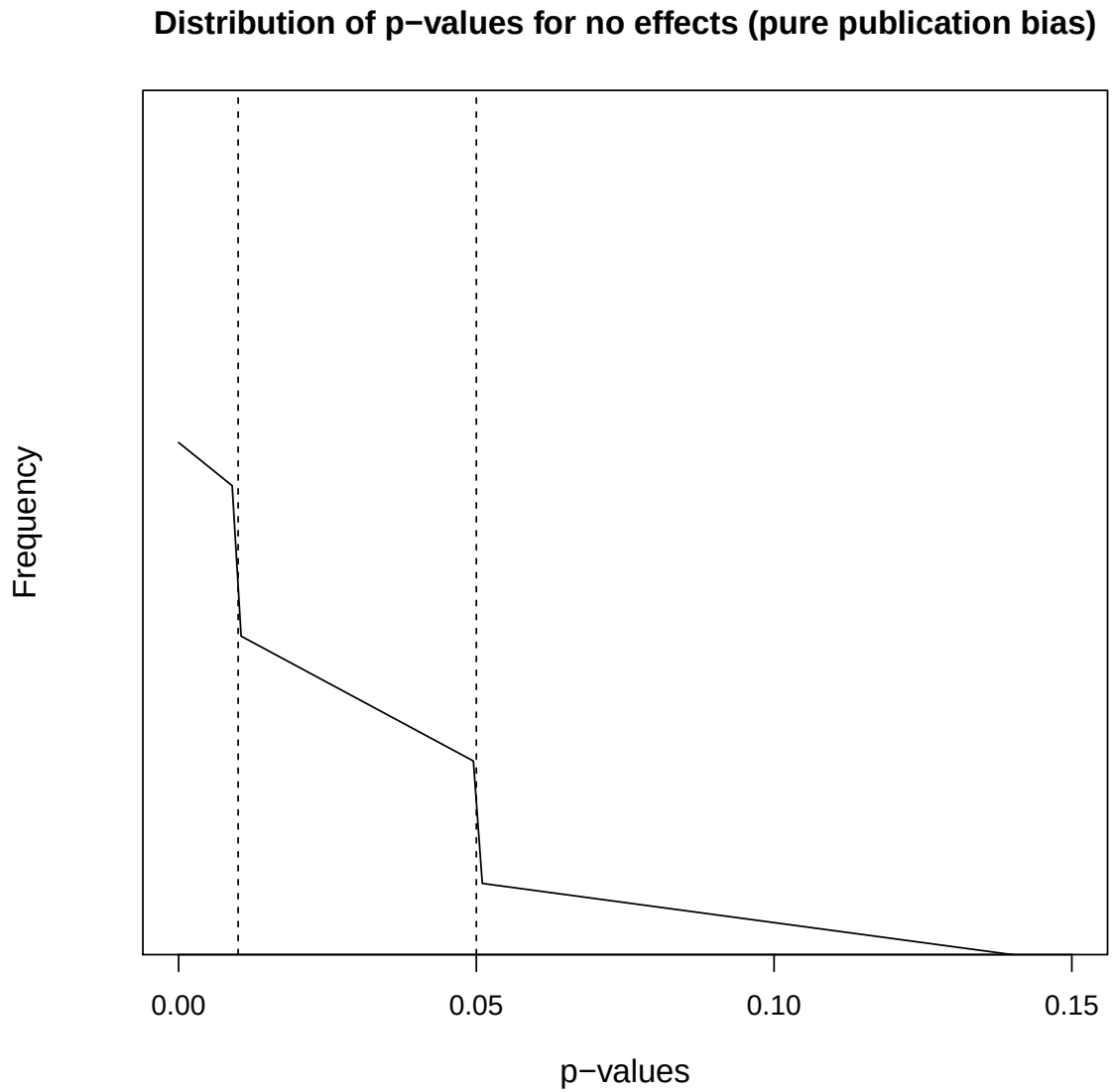
P-Hacking


Distribution of p−values when there are real treatment effects
Frequency
0.00
0.05
0.10
0.15
p−values

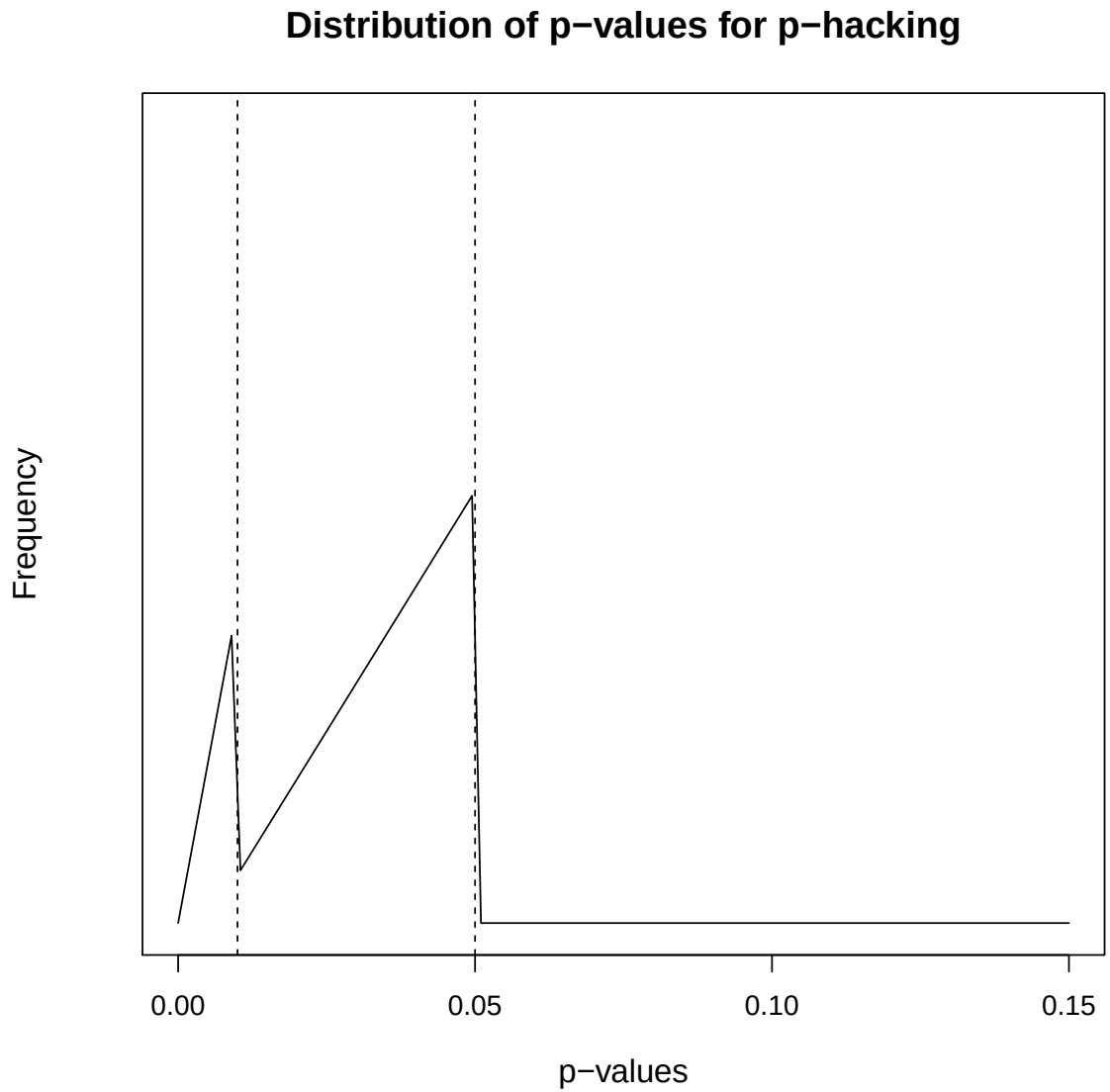
P-Hacking

Question: What is the distribution of p-values in journal articles when there are no treatment effects, but researchers are not tweaking their models to get significant results?

P-Hacking


Distribution of p−values for no effects (pure publication bias)
Frequency
0.00
0.05
0.10
0.15
p−values

P-Hacking


Distribution of p−values for p−hacking
Frequency
0.00
0.05
0.10
0.15
p−values

P-Hacking

Recommendations

1. The less freedom you have in the analysis phase of your study, the better

2. Anytime you face a decision that will affect your results, justify your choice and report the results for the other reasonable choices you could have made

3. Pre-analysis plans can help limit your choices in the analysis phase, strengthening the credibility of your study