What is a hypothesis test? Why test hypotheses?

Jake Bowers and EGAP Learning Days Instructors

9 April 2019 — Bogotá

University of Illinois @ Urbana-Champaign among other affiliations

Overview

Key Points for this lecture

Statistical inference (e.g. Hypothesis tests and confidence intervals) is **inference** — reasoning about the unobserved.

p-values require probability distributions.

Randomization (or Design) + a Hypothesis + a Test Statistic Function can provide probability distributions representing the hypothesis (and thus *p*-values).

Using randomization to reason about causal Inference

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

			Turnout				
City	Pair	Treat	Baseline	Outcome	Newspaper	y_1	y_0
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	?

Table 1: 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.

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

			Turnout				
City	Pair	Treat	Baseline	Outcome	Newspaper	y_1	y_0
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	?

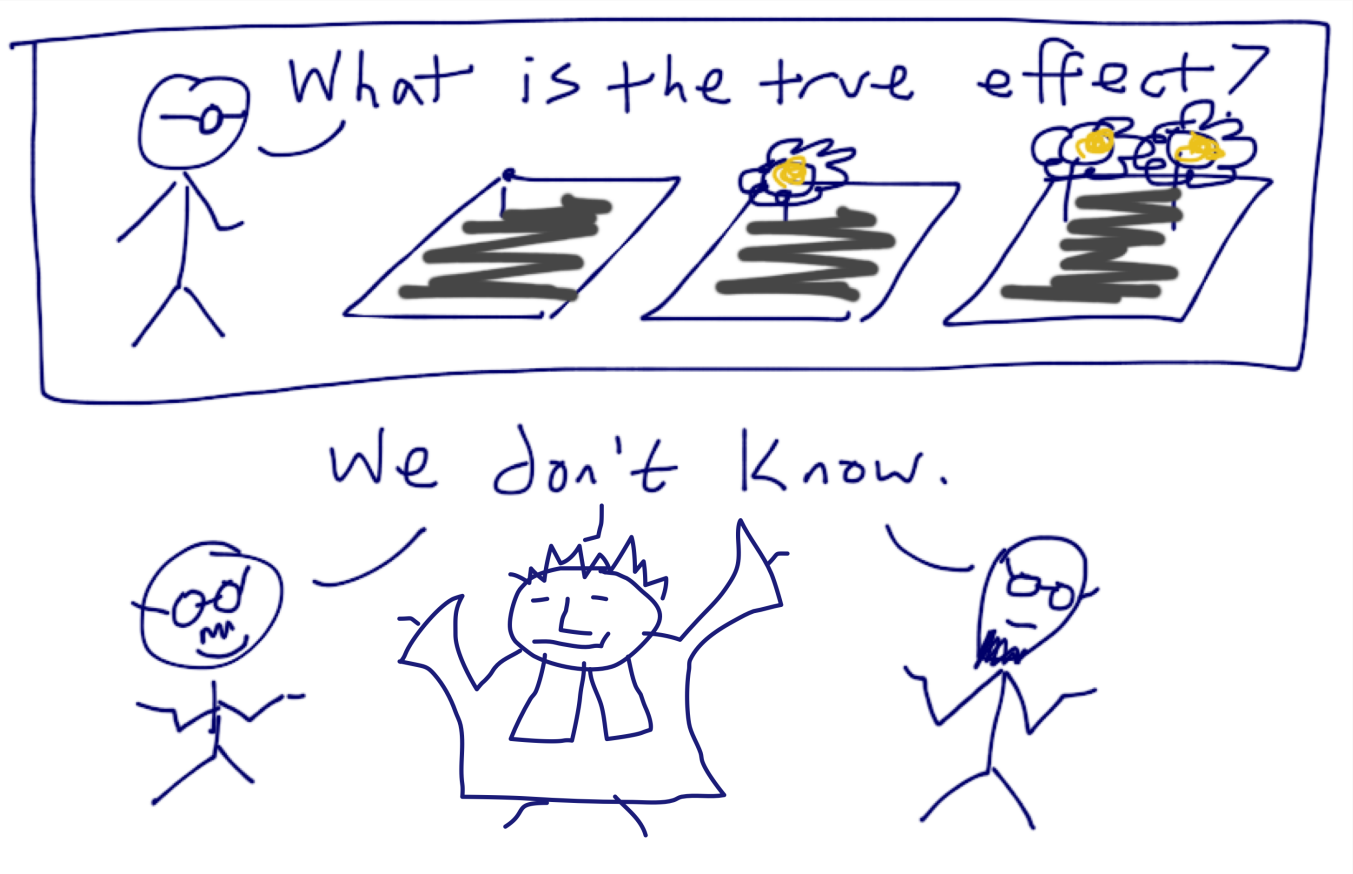
Table 1: Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment randomized within pair with the newspaper ads as 1 and lack of treatment as 0. The potential outcomes are y_1 for treatment and y_0 for

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

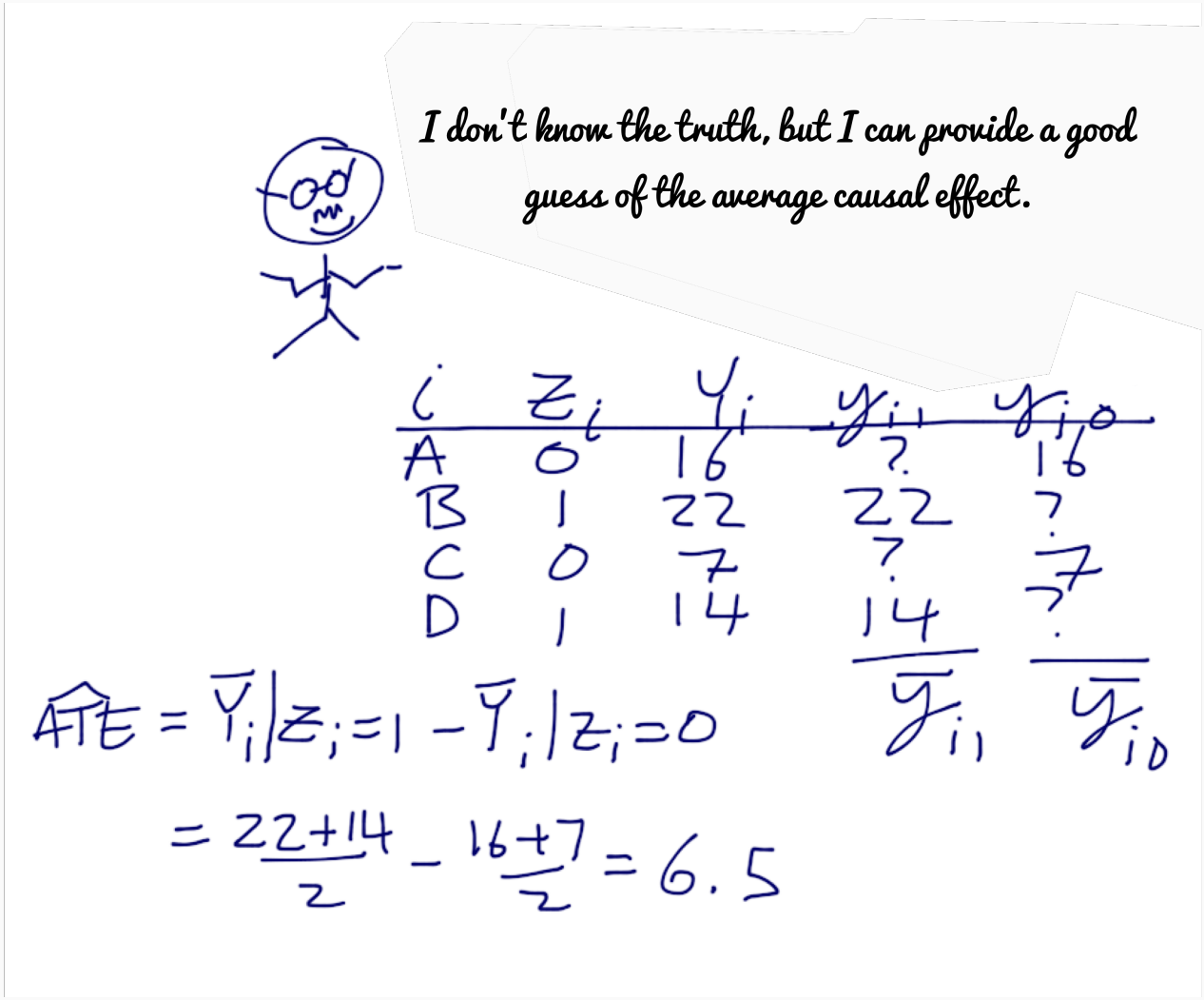
			Turnout				
City	Pair	Treat	Baseline	Outcome	Newspaper	y_1	y_0
Saginaw	1	0	17	16		?	16
Sioux City	1	1	21	22	Sioux City Journal	22	?
Battle Creek	2	0	13	14		?	14
Midland	2	1	12	7	Midland Daily News	7	?
Oxford	3	0	26	23		?	23
Lowell	3	1	25	27	Lowell Sun	27	?
Yakima	4	0	48	58		?	58
Richland	4	1	41	61	Tri-City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

Table 1: 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

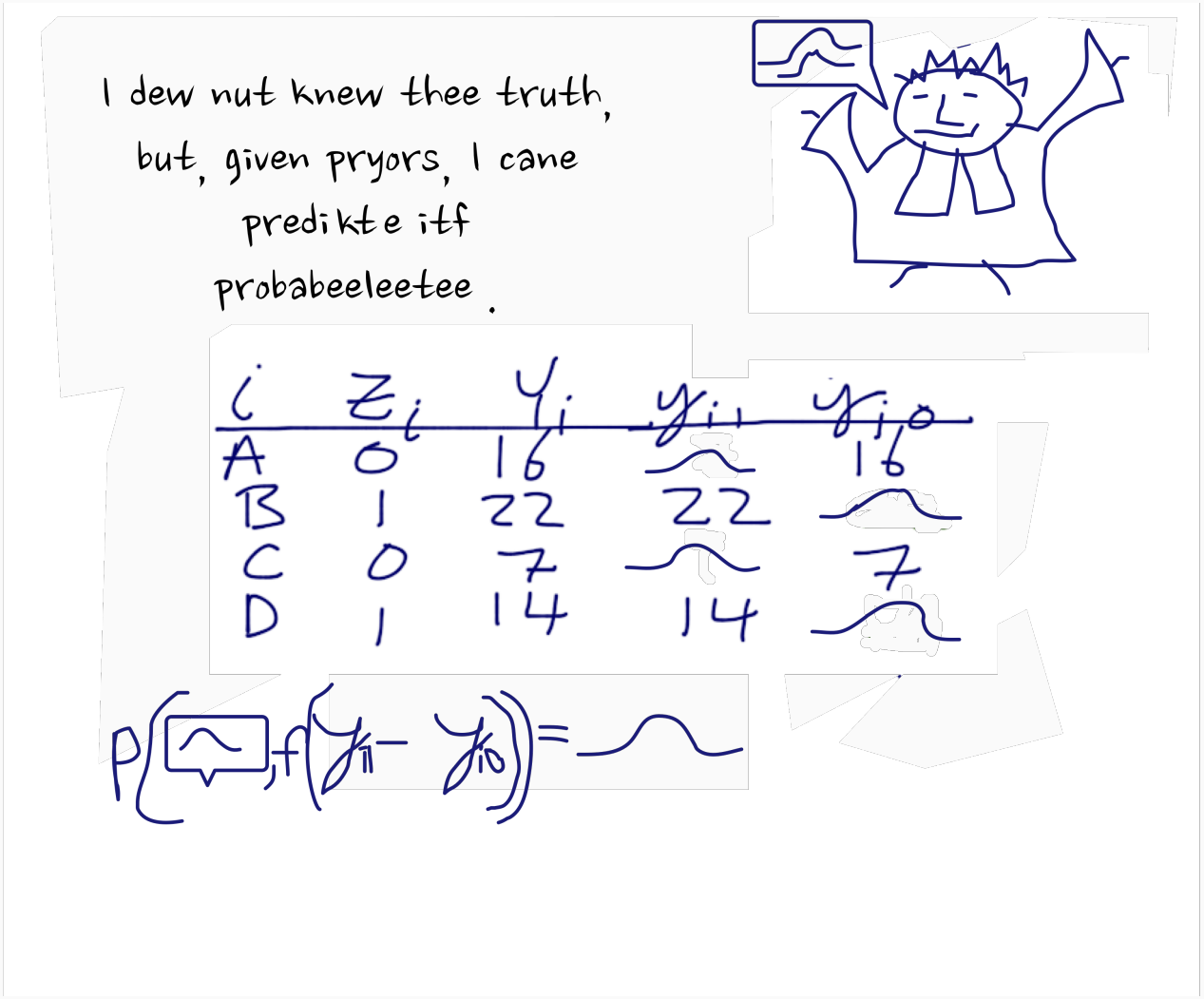
What is the true effect of the treatment assignment?



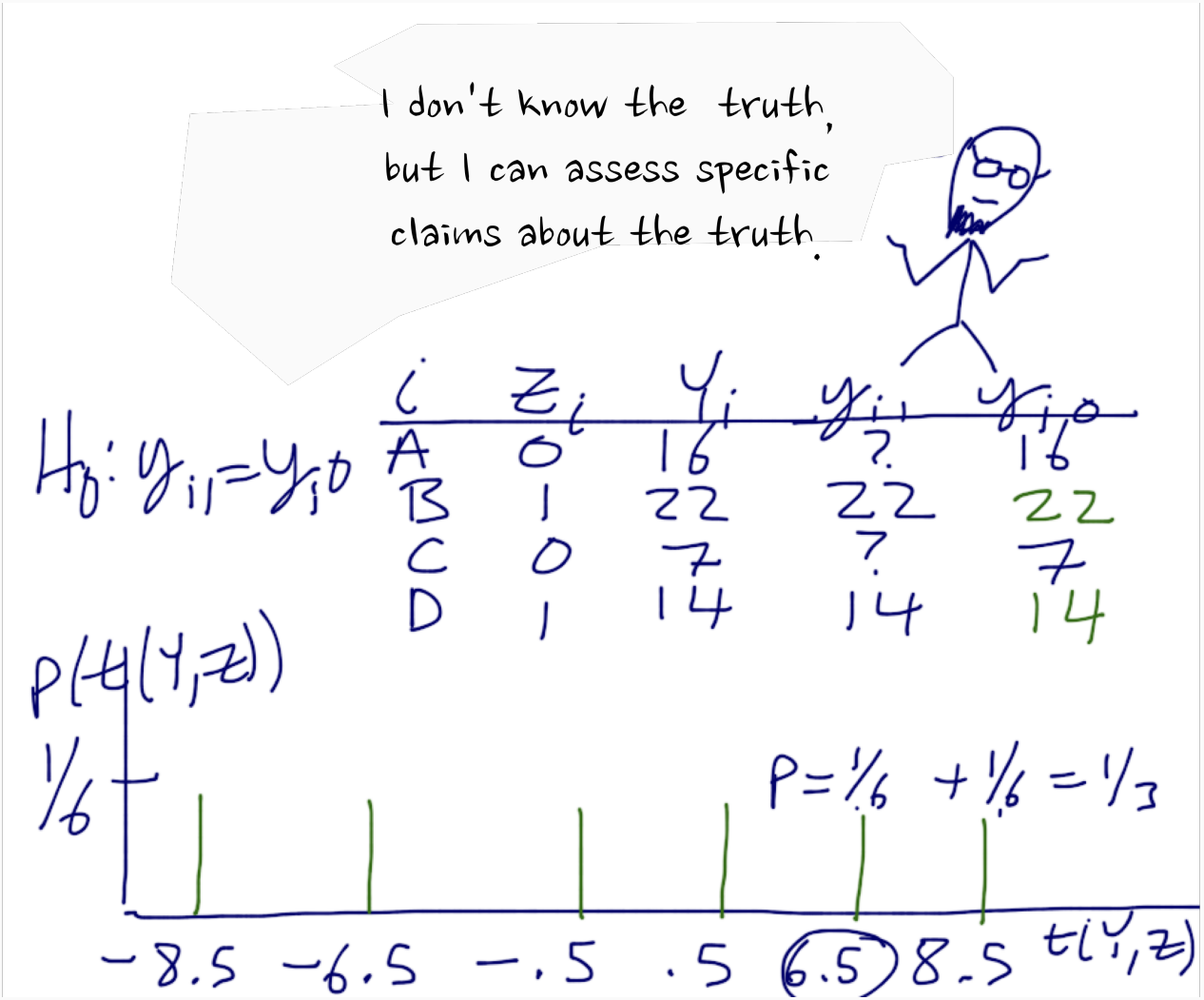
What is the true effect of the treatment assignment?



What is the true effect of the treatment assignment?



What is the true effect of the treatment assignment?



Ingredients of a hypothesis test

- A **hypothesis** is a statement about a relationship among potential outcomes (Strong or Weak)
- A **test statistic** summarizes the relationship between treatment and observed outcomes.
- The **design** allows us to link the hypothesis and the test statistic: calculate a test statistic that describes a relationship between potential outcomes.
- The **design** also generates a distribution of possible test statistics implied by the hypothesis
- A *p*-value describes the relationship between our observed test statistic and the possible hypothesized test statistics

A hypothesis is a statement about or model of a relationship between potential outcomes

```
kable(dat)
```

Y	Z	y0	tau	y1	Ybin
10	1	0	10	10	0
30	1	0	30	30	0
200	1	0	200	200	1
91	1	1	90	91	0
1	0	1	10	11	0
3	0	3	20	23	0
4	0	4	30	34	0
5	0	5	40	45	0
280	1	190	90	280	1
200	0	200	20	220	1

For example, the sharp, or strong, null hypothesis of no effects is $H_0 : y_{i,1} = y_{i,0}$

Test statistics summarize treatment to outcome relationships

```r
## The mean difference test statistic
meanTZ <- function(ys,z){
    mean(ys[z==1]) - mean(ys[z==0])
}

## The difference of mean ranks test statistic
meanrankTZ <- function(ys,z){
    ranky <- rank(ys)
    mean(ranky[z==1]) - mean(ranky[z==0])
}

observedMeanTZ <- meanTZ(ys=Y,z=Z)
observedMeanRankTZ <- meanrankTZ(ys=Y,z=Z)
observedMeanTZ
```

```
[1] 79.6
```

```r
observedMeanRankTZ
```

```
[1] 3.6
```

Linking test statistic and hypothesis.

What we observe for each person, i, (Y_i) is either what we would have observed in treatment $(y_{i,1})$ **or** what we would have observed in control $(y_{i,0})$.

$$Y_i = Z_i y_{i,1} + \left(1 - Z_i\right) * y_{i,0}$$

So, if $y_{i,1} = y_{i,0}$ then: $Y_i = y_{i,0}$: What we actually observe is what we would have observed in the control condition.

Generating the distribution of hypothetical test statistics

We need to know how to repeat our experiment:

```
repeatExperiment <- function(N){
    complete_ra(N)
}
```

Then we repeat it, calculating the implied test statistic each time:

```
set.seed(123456)
possibleMeanDiffsH0 <- replicate(10000,meanTZ(ys=Y,z=repeatExperiment(N=10)))
set.seed(123456)
possibleMeanRankDiffsH0 <- replicate(10000,meanrankTZ(ys=Y,z=repeatExperiment(N=10)))
```

Plot the randomization distributions under the null

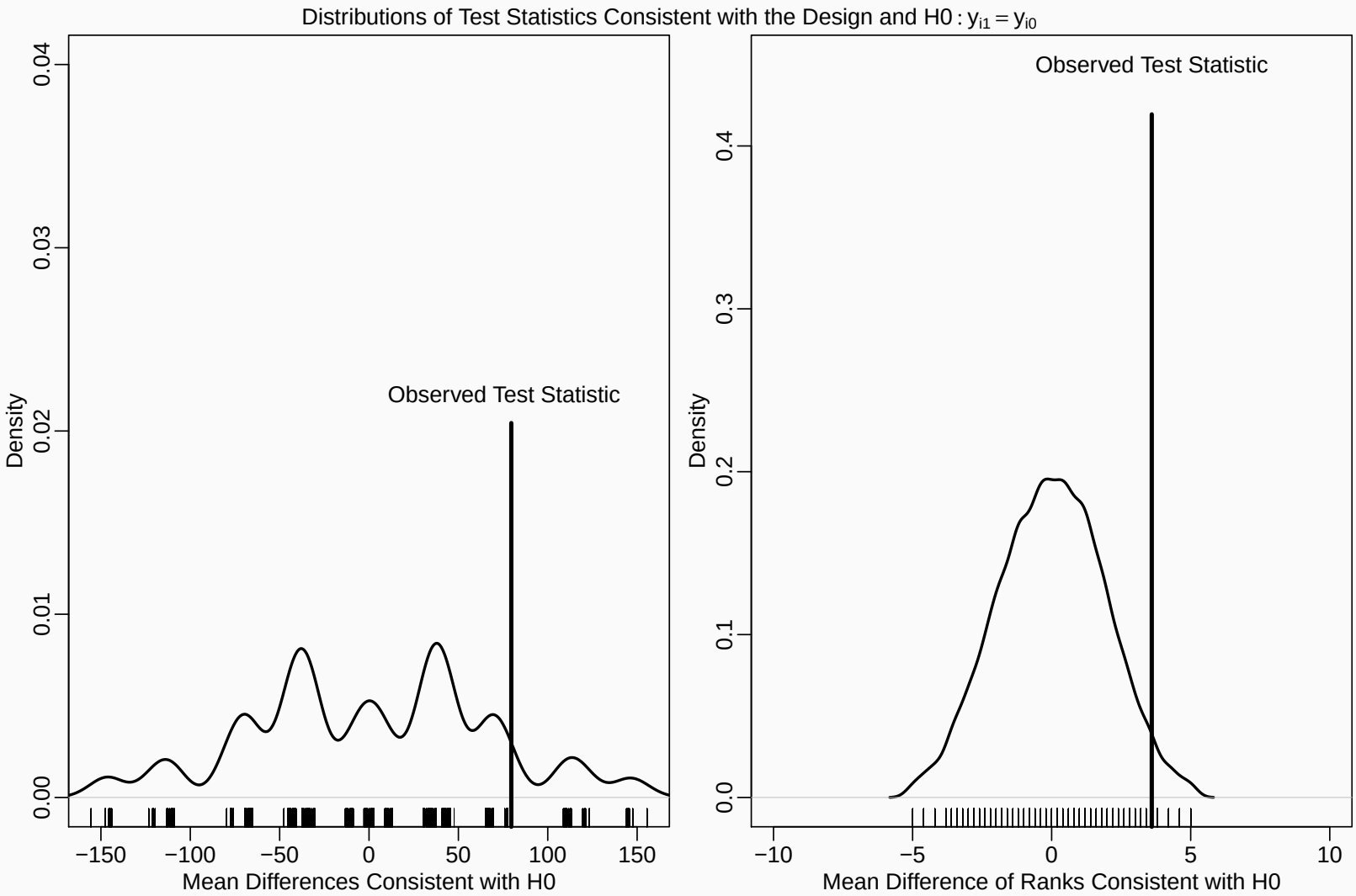


Figure 1: An example of using the design of the experiment to test a hypothesis.

P-values summarize the plots

```
pMeanTZ <- mean( possibleMeanDiffsH0 >= observedMeanTZ )
pMeanRankTZ <- mean( possibleMeanRankDiffsH0 >= observedMeanRankTZ )
pMeanTZ
```

```
[1] 0.0911
```

```
pMeanRankTZ
```

```
[1] 0.0325
```

How to do this in R.

```
## using the coin package
library(coin)
set.seed(12345)
pMean2 <- pvalue(oneway_test(Y~factor(Z),data=dat,distribution=approximate(B=1000)))
```

```
Warning in approximate(B = 1000): 'B' is deprecated; use 'nresample' instead
```

```
dat$rankY <- rank(dat$Y)
pMeanRank2 <- pvalue(oneway_test(rankY~factor(Z),data=dat,distribution=approximate(B=1
```

```
Warning in approximate(B = 1000): 'B' is deprecated; use 'nresample' instead
```

```
pMean2
```

```
[1] 0.181
99 percent confidence interval:
 0.1507 0.2144
```

```
pMeanRank2
```

```
[1] 0.065
99 percent confidence interval:
```

How to do this in R.

```
## using the ri2 package
library(ri2)
thedesign <- declare_ra(N=N)
pMean4 <- conduct_ri( Y ~ Z, declaration = thedesign,
             sharp_hypothesis = 0, data = dat, sims = 1000)
summary(pMean4)
```

```
  term estimate two_tailed_p_value
1    Z     79.6              0.1825
```

```
pMeanRank4 <- conduct_ri( rankY ~ Z, declaration = thedesign,
             sharp_hypothesis = 0, data = dat, sims = 1000)
summary(pMeanRank4)
```

```
  term estimate two_tailed_p_value
1    Z      3.6            0.06349
```

Next topics:

- Testing weak null hypotheses $H_0 : \bar{y}_1 = \bar{y}_0$
- Rejecting null hypotheses (and making false positive and/or false negative errors)
- Power of hypothesis tests
- Maintaining correct false positive error rates when testing more than one hypothesis.

Testing the weak null of no average effects

The weak null hypothesis is a claim about aggregates, and is nearly always stated in terms of averages: $H_0 : \bar{y}_1 = \bar{y}_0$ The test statistic for this hypothesis nearly always is the difference of means (i.e. `meanTZ()` above.

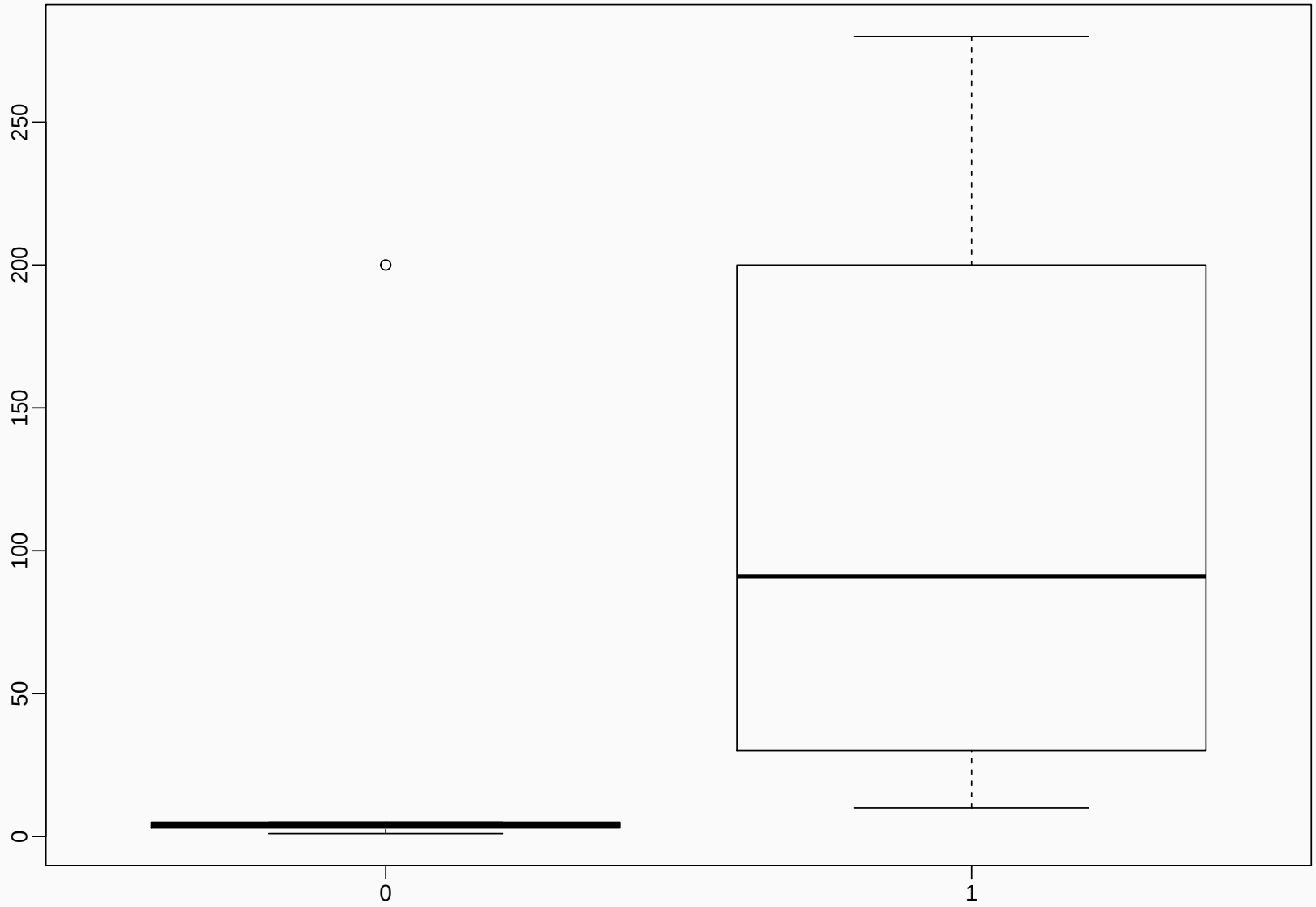
```
lm1 <- lm(Y~Z,data=dat)
lm1P <- summary(lm1)$coef["Z","Pr(>|t|)"]
ttestP1 <- t.test(Y~Z,data=dat)$p.value
library(estimatr)
ttestP2 <- difference_in_means(Y~Z,data=dat)
c(lm1P, ttestP1, ttestP2$p.value)
```

```
                    Z
0.2542 0.2566 0.2566
```

Why is the OLS p-value different? What assumptions is it making?

Testing the weak null of no average effects

```r
boxplot(Y~Z,data=dat)
```



Testing the weak null of no average effects

```r
## By hand:
varEstATE <- function(Y,Z){
    var(Y[Z==1])/sum(Z) + var(Y[Z==0])/sum(1-Z)
}
seEstATE <- sqrt(varEstATE(dat$Y,dat$Z))
obsTStat <- observedMeanTZ/seEstATE
c(observedTestStat=observedMeanTZ,stderror=seEstATE,tstat=obsTStat,
  pval=2*min(pt(obsTStat,df=8,lower.tail = TRUE),
         pt(obsTStat,df=8,lower.tail = FALSE))
  )
```

```
observedTestStat        stderror           tstat            pval
         79.6000         64.8051          1.2283          0.2542
```

Rejecting hypotheses and making errors

How should we interpret 0.0911? What about 0.0325?

What does it mean to "reject" $H_0 : y_{i,1} = y_{i,2}$ at $\alpha = .05$?

"In typical use, the level of the test [α] is a promise about the test's performance and the size is a fact about its performance..." (Rosenbaum 2010, Glossary)

Decision imply errors

If errors are necessary, how can we diagnose them? How to learn whether our hypothesis testing procedure might generate too many false positive errors?

Diagnose by simulation:

Diagnosing false positive rates by simulation

Across repetitions of the design:

- Create a true null hypothesis.
- Test the true null.
- The *p*-value should be large.

The proportion of small *p*-values should be no larger than α.

Diagnosing false positive rates by simulation

Example with a binary outcome.

```
collectPValues <- function(y,z,thedistribution=exact()){
    ## Make Y and Z have no relationship by re-randomizing Z
    newz <- repeatExperiment(length(y))
        thelm <- lm(y~newz,data=dat)
    ttestP2 <- difference_in_means(y~newz,data=dat)
    owP <- pvalue(oneway_test(y~factor(newz),distribution=thedistribution))
    ranky <- rank(y)
    owRankP <- pvalue(oneway_test(ranky~factor(newz),distribution=thedistribution))
    return(c(lmp=summary(thelm)$coef["newz","Pr(>|t|)"],
         neyp=ttestP2$p.value[[1]],
         rtp=owP,
         rtpRank=owRankP))
}
```

```
set.seed(12345)
pDist <- replicate(5000,collectPValues(y=dat$Ybin,z=dat$Z))
```

Diagnosing false positive rates by simulation

```
apply(pDist,1,table)
```

```
$lmp

0.0399685237139576 0.0399685237139577 0.0399685237139578  0.544737300804491
               372                288                173                4167

$neyp

0.0704839969102199  0.545424309672161
               833               4167

$rtp

0.166666666666667                 1
               833              4167

$rtpRank

0.166666666666667                 1
               833              4167
```

Diagnosing false positive rates by simulation

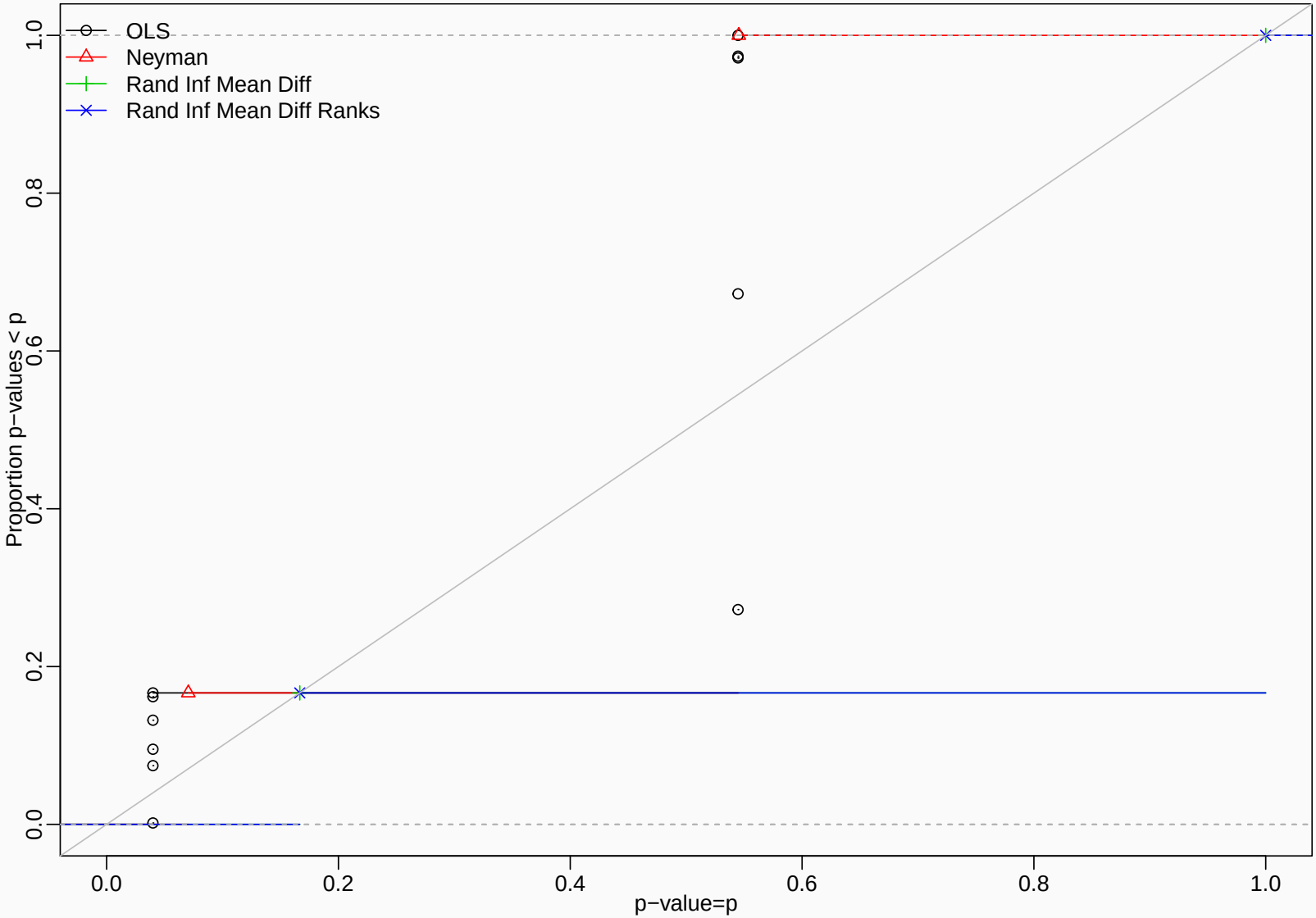
```
apply(pDist,1,function(x){ mean(x<.1)})
```

```
    lmp    neyp     rtp rtpRank
 0.1666  0.1666  0.0000  0.0000
```

```
apply(pDist,1,function(x){ mean(x<.25)})
```

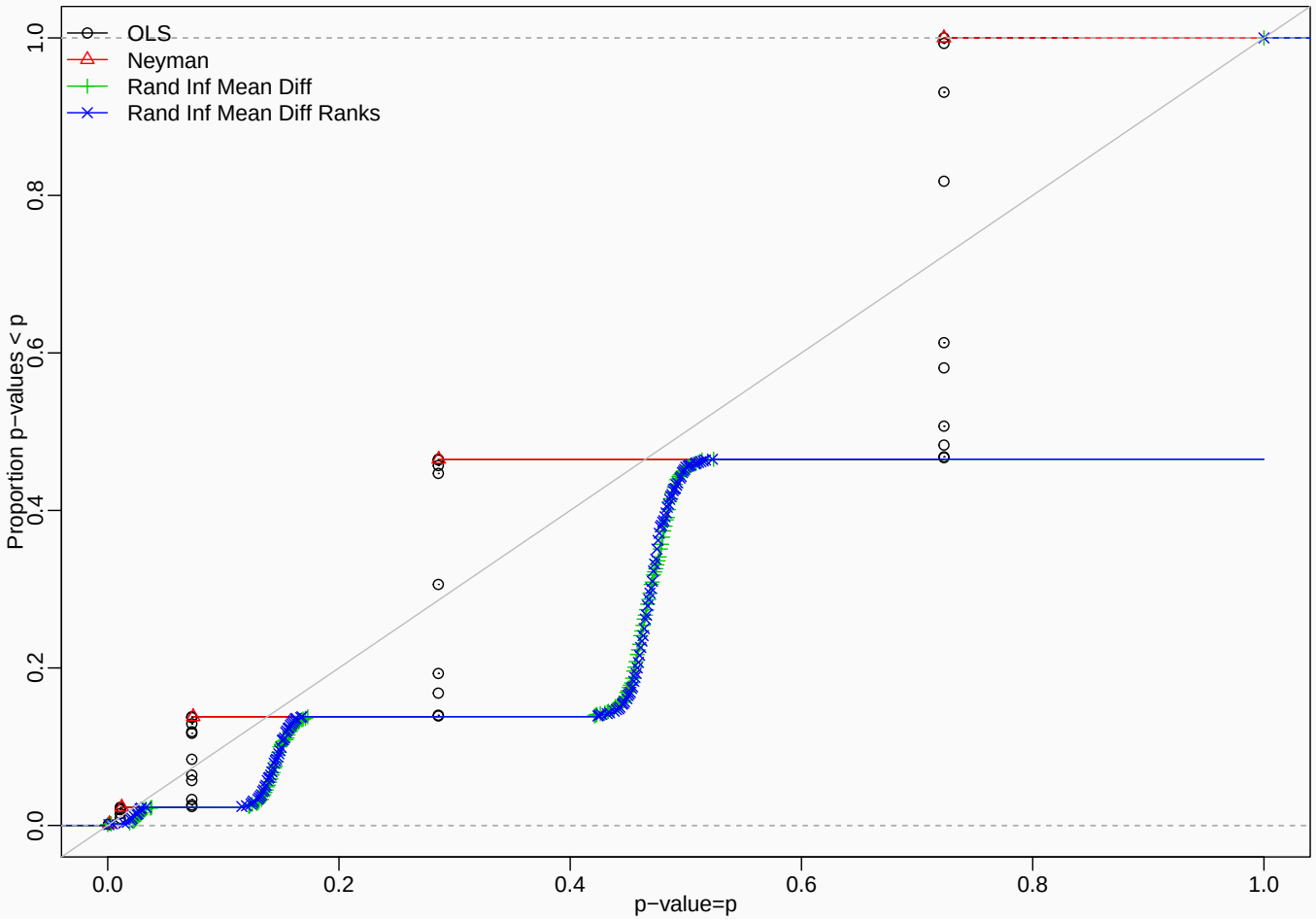
```
    lmp    neyp     rtp rtpRank
 0.1666  0.1666  0.1666  0.1666
```

Diagnosing false positive rates by simulation



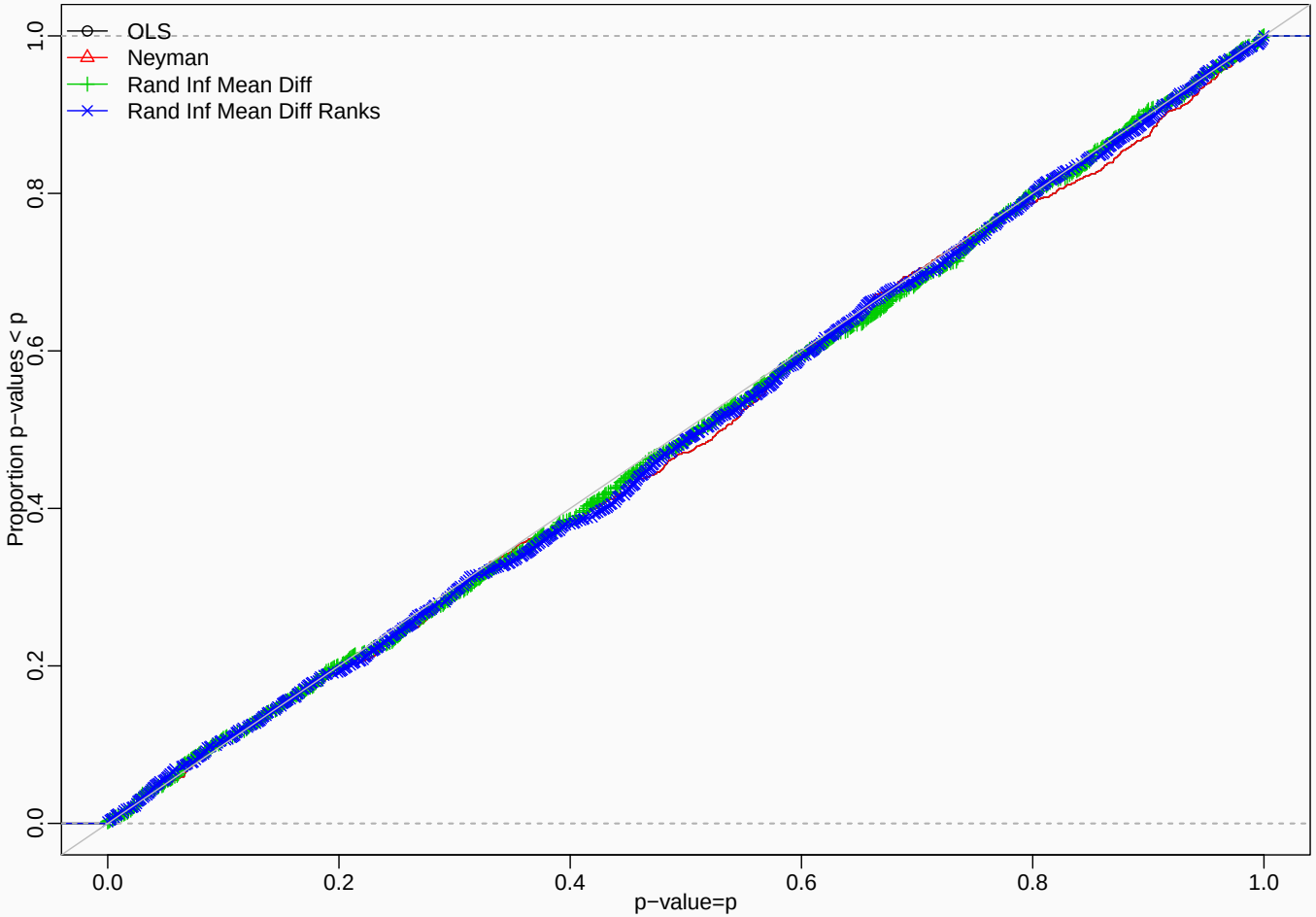
False positive rate with $N = 60$ and binary outcome

```r
set.seed(12345)
##pDistBig <- replicate(1000,collectPValues(y=bigdat$Ybin,z=bigdat$Z,thedistribution=a
library(parallel)
pDistBigLst <- mclapply(1:1000,function(i){ collectPValues(y=bigdat$Ybin,z=bigdat$Z,th
pDistBig <- simplify2array(pDistBigLst)
```



False positive rate with $N = 60$ and continuous outcome

```
set.seed(123456)
pDistBigLst2 <- mclapply(1:1000,function(i) {collectPValues(y=bigdat$Y,z=bigdat$Z,theo
pDistBig2 <- simplify2array(pDistBigLst2)
```



Topics for later

- Power of tests

Summary:

A good test (1) casts doubt on the truth rarely and (2) easily distinguishes signal from noise (casts doubt on falsehoods often).

We can learn whether our testing procedure controls false positive rates given our design.

When false positive rates are not controlled, what might be going wrong? (often has to do with asymptotics)

What else to know about hypothesis tests.

Here we list a few other important but advanced topics connected to hypothesis testing:

- Even if a given testing procedure controls the false positive rate for a single test, it may not control the rate for a group of multiple tests. See 10 Things you need to know about multiple comparisons for a guide to the approaches to controlling such rejection-rates in multiple tests.
- A $100\alpha\%$ confidence interval can be defined as the range of hypotheses where all of the p-values are greater than or equal to α. This is called inverting the hypothesis test. (Rosenbaum (2010)). That is, a confidence interval is a collection of hypothesis tests.
- A point estimate based on hypothesis testing is called a Hodges-Lehmann point estimate. (Rosenbaum (1993),Hodges and Lehmann (1963))

What else to know about hypothesis tests.

- A set of hypothesis tests can be combined into one single hypothesis test (Hansen and Bowers (2008),Caughey et al. (2017))
- In equivalence testing, one can hypothesize that two test-statistics are equivalent (i.e. the treatment group is the same as the control group) rather than only about one test-statistic (the difference between the two groups is zero) {Hartman and Hidalgo (2018)}
- Since a hypothesis test is a model of potential outcomes, one can use hypothesis testing to learn about complex models, such as models of spillover and propagation of treatment effects across networks (Bowers et al. (2013), Bowers et al. (2016), Bowers et al. (2018))

References i

Jake Bowers, Mark M Fredrickson, and Costas Panagopoulos. Reasoning about Interference Between Units: A General Framework. *Political Analysis*, 21(1):97–124, 2013.

Jake Bowers, Mark Fredrickson, and Peter M Aronow. Research note: A more powerful test statistic for reasoning about interference between units. *Political Analysis*, 24(3):395–403, 2016.

Jake Bowers, Bruce A Desmarais, Mark Frederickson, Nahomi Ichino, Hsuan-Wei Lee, and Simi Wang. Models, methods and network topology: Experimental design for the study of interference. *Social Networks*, 54: 196–208, 2018.

References ii

Devin Caughey, Allan Dafoe, and Jason Seawright. Nonparametric combination (npc): A framework for testing elaborate theories. *The Journal of Politics*, 79(2):688–701, 2017.

Ben B. Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

Erin Hartman and F Daniel Hidalgo. An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013, 2018.

J.L. Hodges and E.L. Lehmann. Estimates of location based on rank tests. *Ann. Math. Statist*, 34:598–611, 1963.

Costas Panagopoulos. The impact of newspaper advertising on voter turnout: Evidence from a field experiment. Paper presented at the MPSA 2006., 2006.

References iii

P R Rosenbaum. Design of observational studies. *Springer series in statistics*, 2010.

Paul R. Rosenbaum. Hodges-lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, 88 (424):1250–1253, 1993. ISSN 01621459.