Statistical Analysis

Andrew Lam

4/21/2017

```
# Load packages
library(data.table)
```

Data Cleaning

Made the following changes in "assignments-refined.csv":

1. For pilot group, converted NAs to 1 in pilot column
2. For pilot group, converted NAs to 0 or 1 in treat column
3. In pilot group, only kept one week of treatment data (removed data from 3/29 to 4/4)

```
# Load data
assignments <- read.csv("./assignments-refined.csv")
d <- read.csv("./everyone-refined.csv")
d <- data.table(d)
# d$timestamp <-as.Date(d$timestamp,format='%Y/%m/%d')
```

```
# d[username == "jljones.dt@gmail.com" | username == "victorwwang@gmail.com" | username == "jill.wishar

# merge(d, assignments, by.x = "username", by.y = "email")

# assignments[match(trimws(d$username), trimws(assignments$email)), "treat"]
```

Average Outcomes

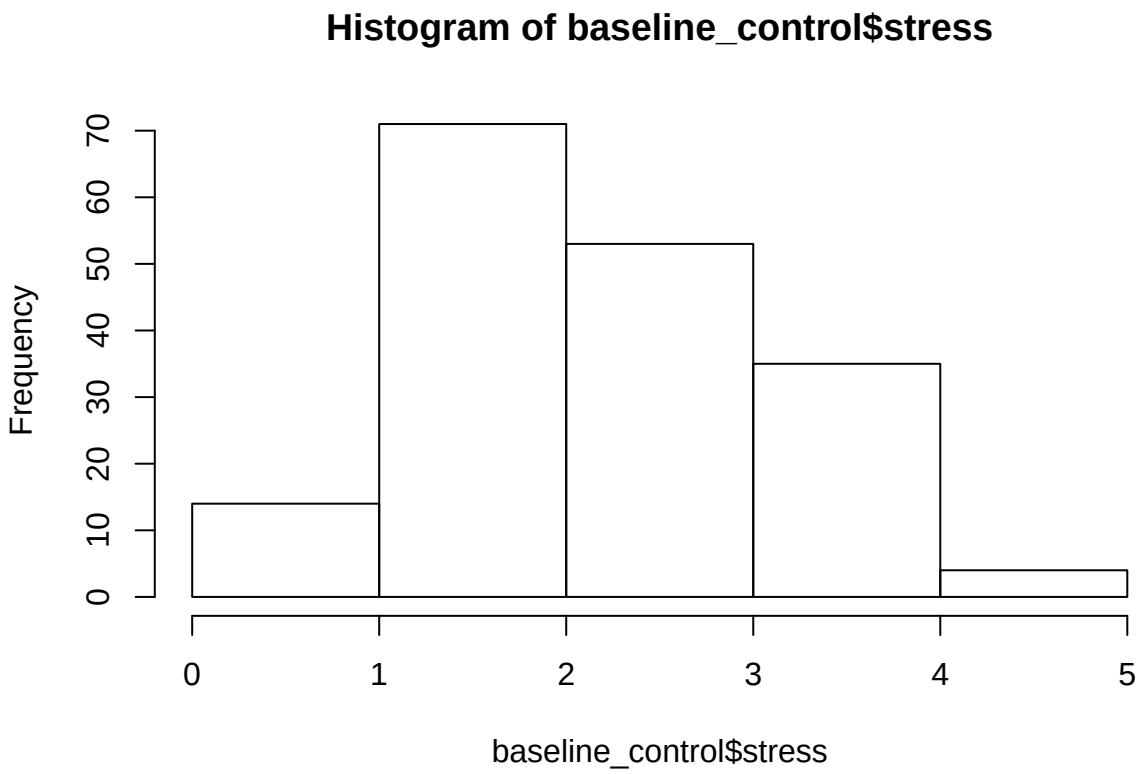
```
# Baseline - Control
baseline_control <- d[assignment == 'b' & treat == 0, ]

summary(baseline_control$stress)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.684   3.000   5.000
```

```
sd(baseline_control$stress)
```

```
## [1] 0.9543039
```

```
hist(baseline_control$stress, breaks=seq(0,5))
```



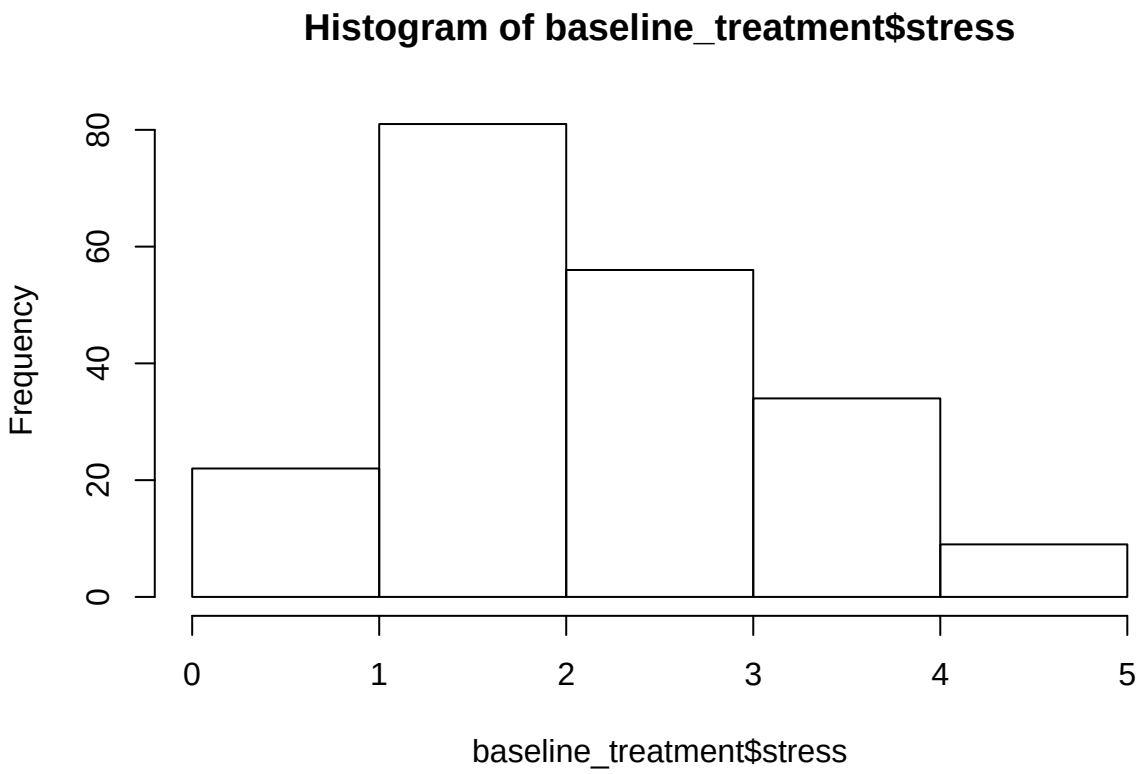

```
# Baseline - Treatment
baseline_treatment <- d[assignment == 'b' & treat == 1, ]

summary(baseline_treatment$stress)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.639   3.000   5.000
```

```
sd(baseline_treatment$stress)
```

```
## [1] 1.028496
```

```
hist(baseline_treatment$stress, breaks=seq(0,5))
```



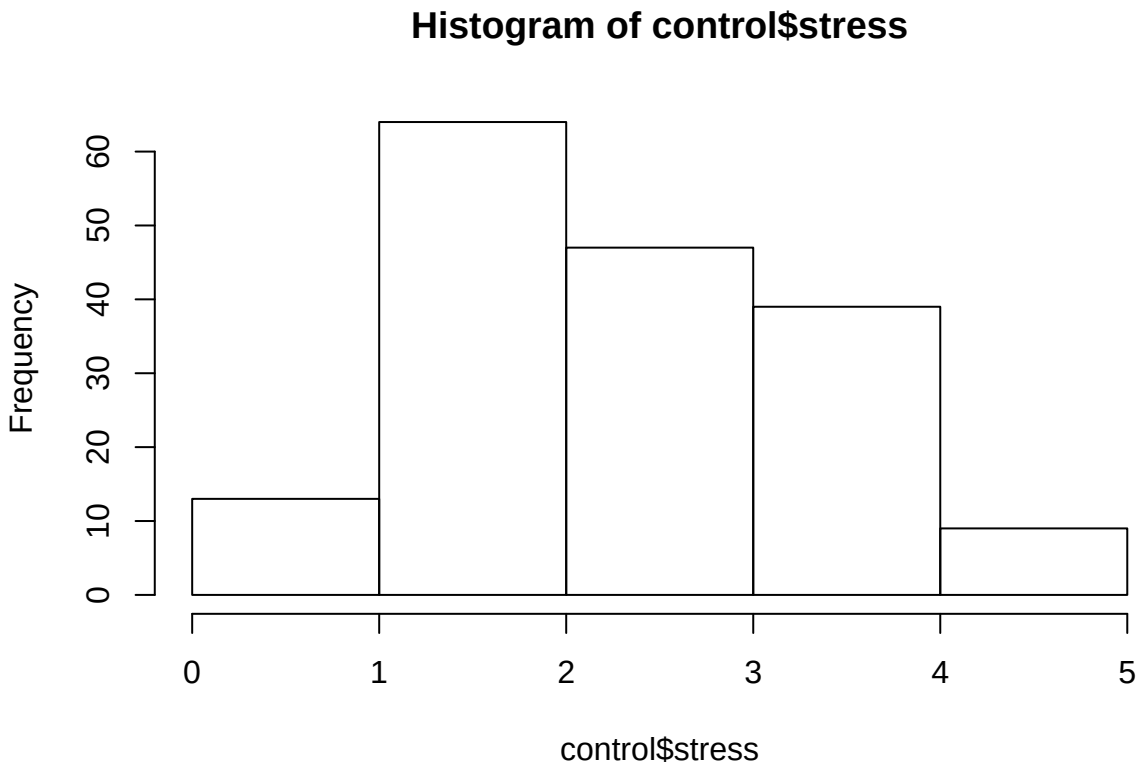

```
# Control
control <- d[assignment == 'c', 'stress']

summary(control$stress)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.808   4.000   5.000
```

```
sd(control$stress)
```

```
## [1] 1.039198
```

```
hist(control$stress, breaks=seq(0,5))
```



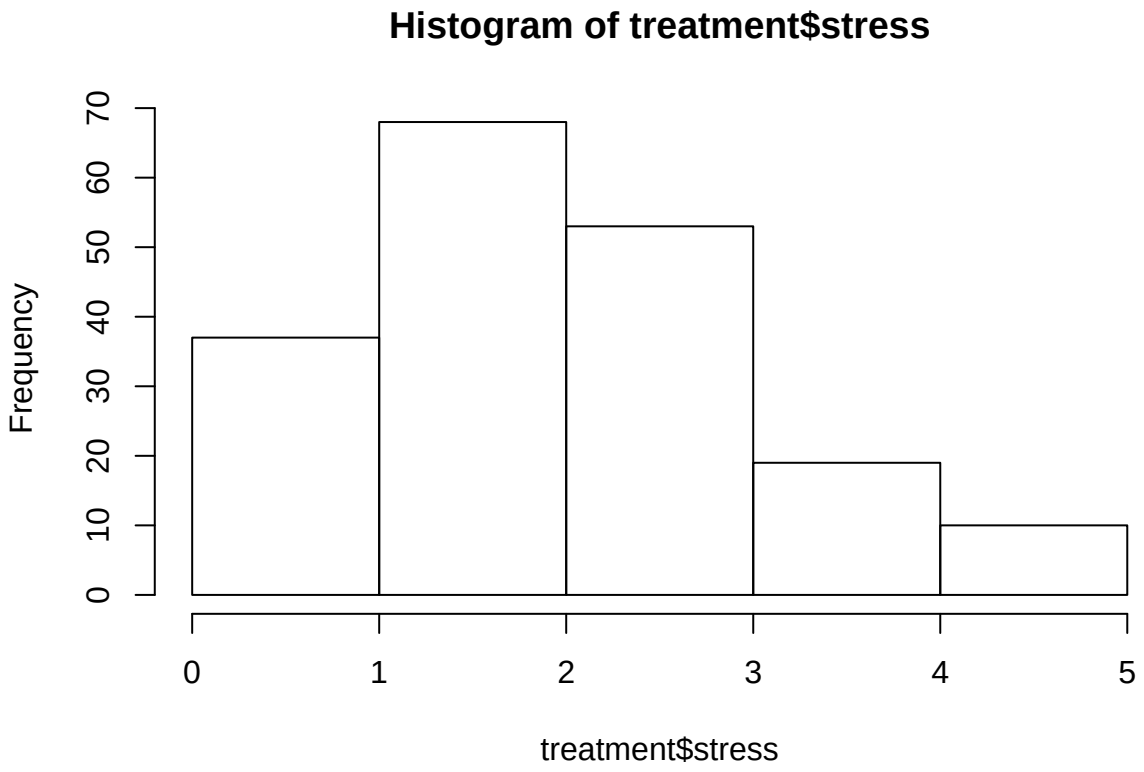

```
# Treatment
treatment <- d[assignment == 't', 'stress']

summary(treatment$stress)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.449   3.000   5.000
```

```
sd(treatment$stress)
```

```
## [1] 1.083273
```

```
hist(treatment$stress, breaks=seq(0,5))
```



ATE

```
# treatment - baseline_treatment
mean(treatment$stress) - mean(baseline_treatment$stress)
```

```
## [1] -0.189416
```

```
# control - baseline_control
mean(control$stress) - mean(baseline_control$stress)
```

```
## [1] 0.1245237
```

```
# difference in difference
ate <- (mean(treatment$stress) - mean(baseline_treatment$stress)) - (mean(control$stress) - mean(baseli
```

CACE

```
# CACE = ATE / alpha
# alpha = share of treatment subjects actually treated (application rate, compliance rate)
# compliers = defined as those who never answered no for the "dnd" variable; assumes that the subject k
# alternatively, compliers could be defined as those who answered all 7 surveys during the treatment pe
```

```
# Number of people in the treatment group that did not comply
alpha_num <- nrow(unique(d[dnd == "No" & treat == 1, "username"]))
alpha_num
```

```
## [1] 15
```

```r
# Total number of people in treatment
alpha_denom <- nrow(unique(d[dnd == "Yes" | dnd == "No", "username"]))
```

```r
# Calculation for alpha
alpha <- (alpha_denom-alpha_num) / alpha_denom
```

```r
# CACE
ate / alpha
```

```
## [1] -0.6082582
```

Hypothesis Tests

```r
# Randomization inference
po.control <- c(seq(from = 1, to = 5))
po.treatment <- po.control
po.control
```

```
## [1] 1 2 3 4 5
```

```r
randomize <- function() {
sample(c(rep(0,30),rep(1,30)))
}

randomize()
```

```
##  [1] 0 1 1 0 1 1 0 1 1 1 1 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 1 1 1 1
## [36] 0 1 1 1 1 0 0 0 0 1 1 1 0 0 1 0 0 1 1 0 1 0 0 1 0
```

```r
treatment <- randomize()
treatment
```

```
##  [1] 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1 0 1 1 1 1 0 0 1 0 0 1 0 1 1 1 0 1 1 0 1
## [36] 0 0 0 1 0 1 0 1 0 1 1 0 1 1 0 1 1 0 1 0 0 1 1 0 1
```

```r
table(treatment)
```

```
## treatment
##  0  1
## 30 30
```

```r
outcomes <- po.treatment * treatment + po.control*(1-treatment)
outcomes
```

```
##  [1] 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
## [36] 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
```

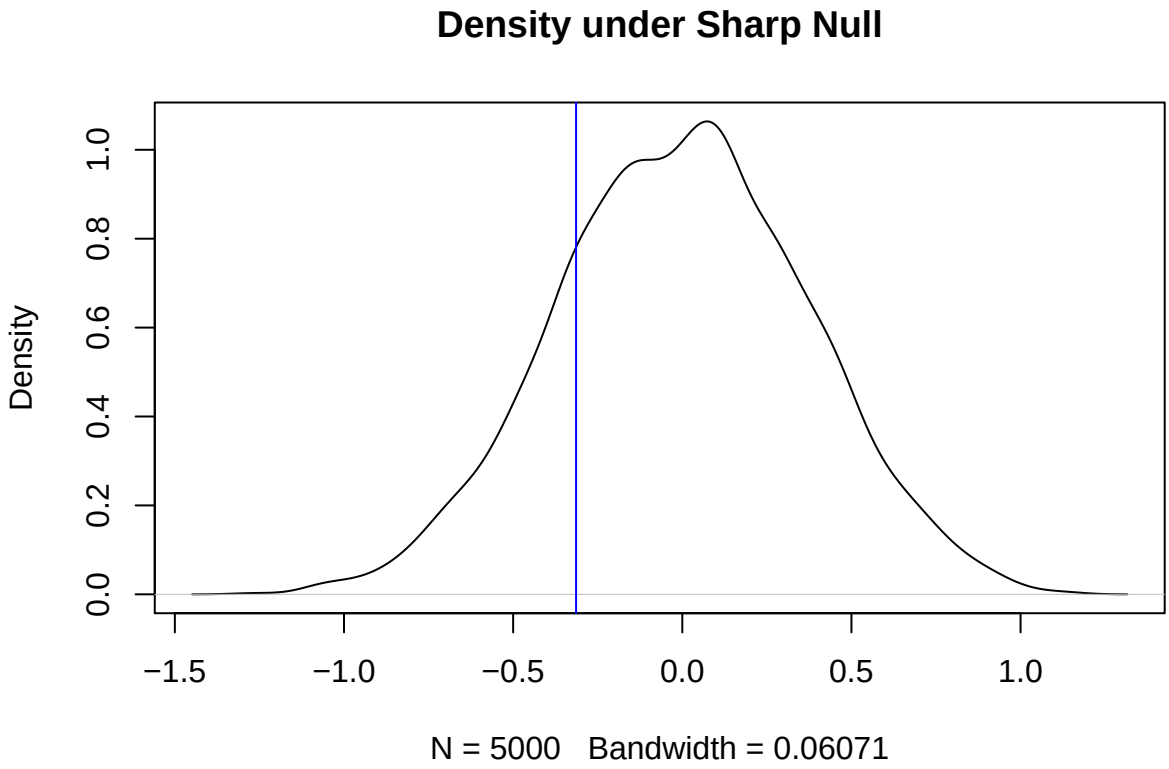
```r
est.ate <- function(outcome, treat) {
mean(outcome[treat==1]) - mean(outcome[treat==0])
}
```

```r
est.ate(outcomes, treatment)
```

```
## [1] 0.4
```

```r
distribution.under.sharp.null <- replicate(5000, est.ate(outcomes,
randomize() ) )
```

```r
plot(density(distribution.under.sharp.null),
main = "Density under Sharp Null")
abline(v = ate, col = "blue")
```




```r
mean(ate >= distribution.under.sharp.null) #p-value
```

```
## [1] 0.2126
```