Data wrangling

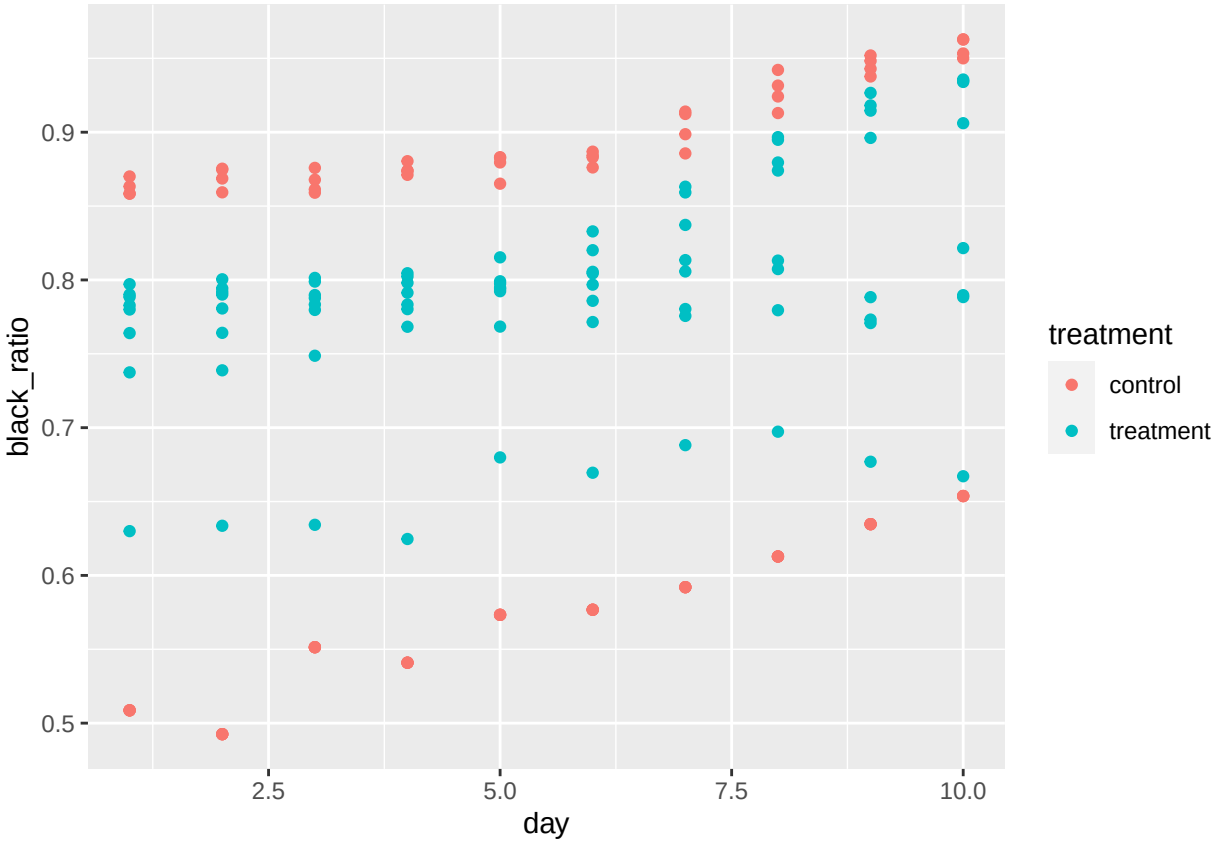
Banana

Data Cleaning

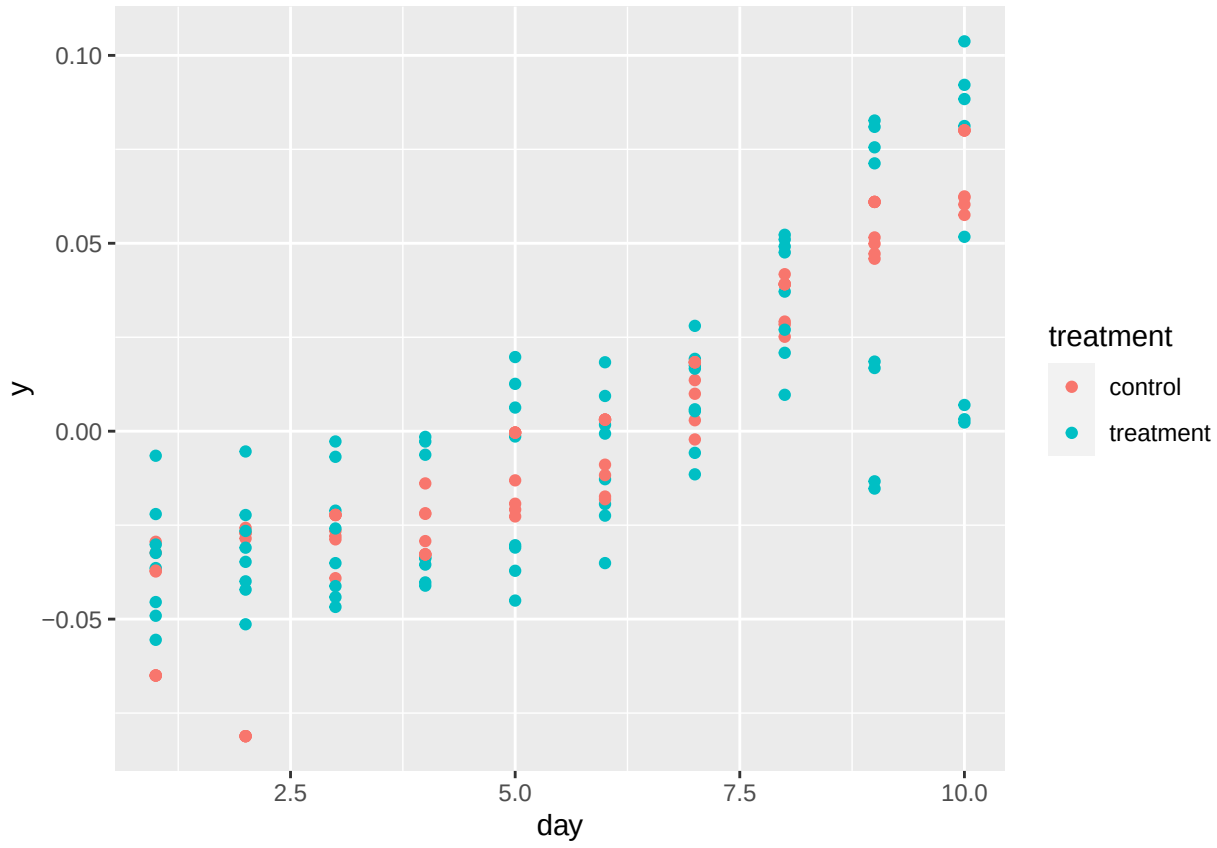
```
#data is given in a long format
d <- fread('banana.csv',header=FALSE, sep=",")
names(d) <- c('block', 'banana_number', 'treatment', 'day', 'humidity',
              'temperature', 'black_ratio', 'weight')
```

- The following figure indicates there exists both individual fixed effect (i.e., value that varies among differnt subjects, but fixed over time) and block effect (i.e., value that varies between blocks but fixed within blocks)
- The individual fixed effect could have caused by how the picture was taken by different individuals and block fixed effect could have been the result of different enviromental condition (i.e., temperature of the house or humidity or lighting).
- The individual and block fixed had been taken out in the following figure

```
#remove the fixed effect
d[,b_mu:=mean(black_ratio), by = .(block,banana_number)]
d[,y := black_ratio - b_mu]

d %>% ggplot(aes(x = day, y = black_ratio)) +
  geom_point(aes(color = treatment))
```




```
d %>% ggplot(aes(x = day, y = y)) +
  geom_point(aes(color = treatment))
```




```
#save the result of the detrended data

db <- d[,c('y','block', 'banana_number', 'treatment', 'day', 'humidity', 'temperature','black_ratio','w

write.csv(db,"db.csv")
```

- The percent of `black-ratio` should be non-decreasing function of days, but we have noticed that `black_ratio` can decrease when the tip of the banana withers which cause size of the banana cropped during the image processing part.
- To account for such error introduced while measuring the `black_ratio`, one of off-line abrupt change detection algorithm was employeed to detect changes in `blac_ratio` while accounting for such small noise.

```
#included the detrended data
db <- d[,c('y','block', 'banana_number', 'treatment', 'day', 'humidity', 'temperature','black_ratio','w

## change the format to process that data
db[,b_name := paste0(block,banana_number)]
db <- db[order(rank(b_name), y)]
dim(db)
```

```
## [1] 160  10
```

```r
head(db)
```

```
##                y block banana_number treatment day humidity temperature
## 1: -0.0811424854     B             1   control   2       45          77
## 2: -0.0650039843     B             1   control   1       43          73
## 3: -0.0327769297     B             1   control   4       46          73
## 4: -0.0222880071     B             1   control   3       50          79
## 5: -0.0003719828     B             1   control   5       48          73
## 6:  0.0030826880     B             1   control   6       50          77
##    black_ratio weight b_name
## 1:   0.4925462    777     B1
## 2:   0.5086847    152     B1
## 3:   0.5409117    777     B1
## 4:   0.5514006    777     B1
## 5:   0.5733167    777     B1
## 6:   0.5767713    777     B1
```

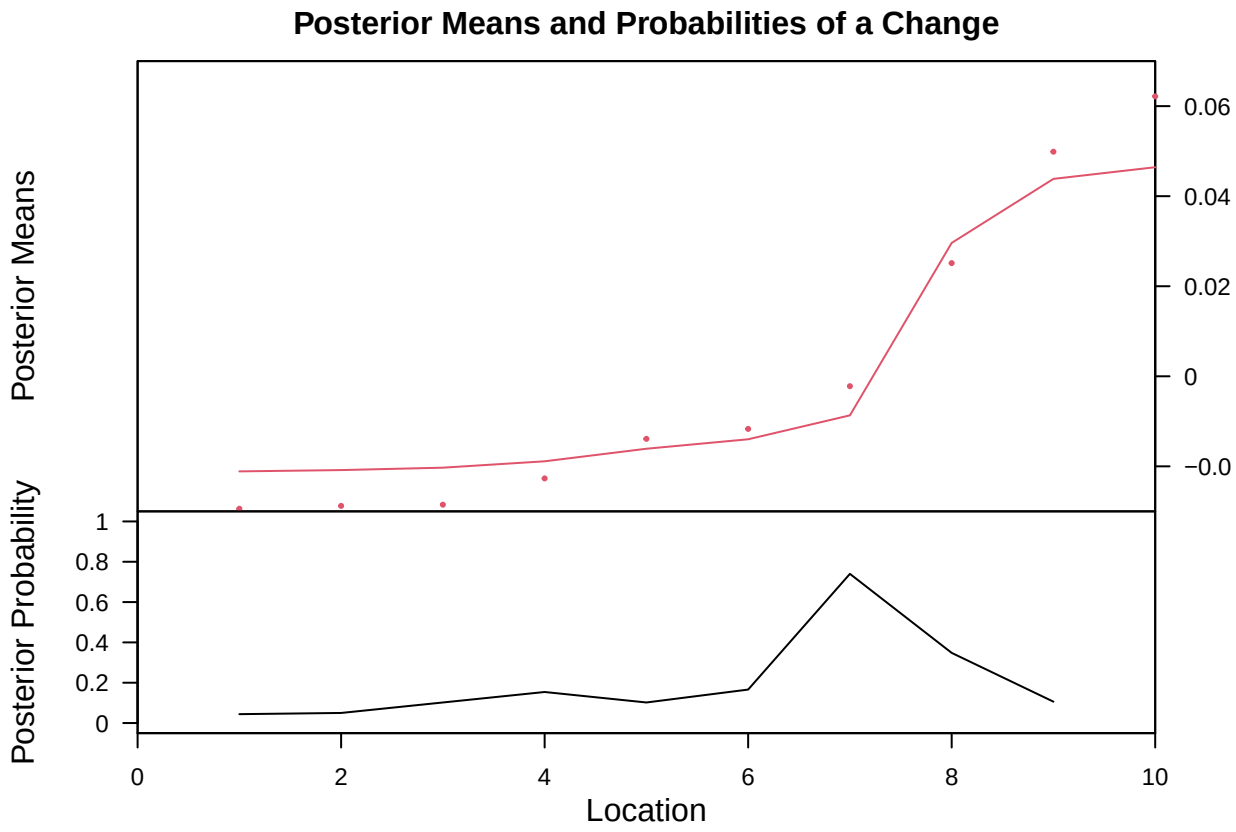
```r
b_result <- data.frame(block = numeric(0),banana_number= numeric(0),
                       treatment = numeric(0),
                       d_turn = numeric(0) )

write.csv(db,"db.csv")
```

Estimate the treatment effect

- The response is the day at which rapid deteroriation of banana condition was observed and the following figure shows there were detected

TODO: Need to figure out how to suppress the output

Code for mannual confirmation




```
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability        X1
## 1        0.044 -0.021123
## 2        0.050 -0.020829
## 3        0.102 -0.020301
## 4        0.154 -0.018880
## 5        0.102 -0.016093
## 6        0.166 -0.013990
## 7        0.740 -0.008681
## 8        0.348  0.029618
## 9        0.106  0.043844
## 10          NA  0.046435

## [1] 7
```

Get the data

```
db[,b_name := paste0(block,banana_number)]
db <- db[order(rank(b_name), y)]
dim(db)
```

```
## [1] 160  10
```

```
head(db)
```

```
##                y block banana_number treatment day humidity temperature
## 1: -0.0811424854     B             1   control   2       45          77
## 2: -0.0650039843     B             1   control   1       43          73
## 3: -0.0327769297     B             1   control   4       46          73
## 4: -0.0222880071     B             1   control   3       50          79
## 5: -0.0003719828     B             1   control   5       48          73
## 6:  0.0030826880     B             1   control   6       50          77
##    black_ratio weight b_name
## 1:   0.4925462    777     B1
## 2:   0.5086847    152     B1
## 3:   0.5409117    777     B1
## 4:   0.5514006    777     B1
## 5:   0.5733167    777     B1
## 6:   0.5767713    777     B1
```

```
b_result <- data.frame(block = numeric(0),banana_number= numeric(0),
                       treatment = numeric(0),
                       d_turn = numeric(0) )
#has 160 rows
for (val in 1:16)
{
 start <- 1 + (val-1)*10
 end <- (10*val)
 #get the subject information
 val_block   <- db[start,block]
 val_subject <- db[start,banana_number]
 val_treat   <- db[start,treatment]

 # # print(paste(start,":",end))
 d_temp <- db[start:end,]
 d_turn <- getChangeDate(d_temp)
 b_result[val,] <- c(val_block,val_subject,val_treat,d_turn)
}
```

```
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
```

```
## 1        0.092 NaN
## 2        0.294 NaN
## 3        0.256 NaN
## 4        0.258 NaN
## 5        0.170 NaN
## 6        0.242 NaN
## 7        0.292 NaN
## 8        0.224 NaN
## 9        0.100 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.178 NaN
## 2        0.228 NaN
## 3        0.138 NaN
## 4        0.116 NaN
## 5        0.184 NaN
## 6        0.282 NaN
## 7        0.522 NaN
## 8        0.162 NaN
## 9        0.074 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.052 NaN
## 2        0.046 NaN
## 3        0.048 NaN
## 4        0.944 NaN
## 5        0.072 NaN
## 6        0.126 NaN
## 7        0.154 NaN
## 8        0.288 NaN
## 9        0.138 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
```

```
## 1        0.118 NaN
## 2        0.334 NaN
## 3        0.152 NaN
## 4        0.134 NaN
## 5        0.128 NaN
## 6        0.186 NaN
## 7        0.224 NaN
## 8        0.518 NaN
## 9        0.144 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.092 NaN
## 2        0.294 NaN
## 3        0.256 NaN
## 4        0.258 NaN
## 5        0.170 NaN
## 6        0.242 NaN
## 7        0.292 NaN
## 8        0.224 NaN
## 9        0.100 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.092 NaN
## 2        0.294 NaN
## 3        0.256 NaN
## 4        0.258 NaN
## 5        0.170 NaN
## 6        0.242 NaN
## 7        0.292 NaN
## 8        0.224 NaN
## 9        0.100 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
```

```
## 1        0.056 NaN
## 2        0.140 NaN
## 3        0.646 NaN
## 4        0.130 NaN
## 5        0.102 NaN
## 6        0.102 NaN
## 7        0.144 NaN
## 8        0.204 NaN
## 9        0.664 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.092 NaN
## 2        0.294 NaN
## 3        0.256 NaN
## 4        0.258 NaN
## 5        0.170 NaN
## 6        0.242 NaN
## 7        0.292 NaN
## 8        0.224 NaN
## 9        0.100 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.066 NaN
## 2        0.058 NaN
## 3        0.112 NaN
## 4        0.168 NaN
## 5        0.152 NaN
## 6        0.234 NaN
## 7        0.682 NaN
## 8        0.210 NaN
## 9        0.098 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
```

```
## 1        0.050 NaN
## 2        0.048 NaN
## 3        0.080 NaN
## 4        0.140 NaN
## 5        0.252 NaN
## 6        0.456 NaN
## 7        0.508 NaN
## 8        0.202 NaN
## 9        0.082 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.046 NaN
## 2        0.044 NaN
## 3        0.046 NaN
## 4        0.058 NaN
## 5        0.070 NaN
## 6        0.398 NaN
## 7        0.678 NaN
## 8        0.322 NaN
## 9        0.100 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.042 NaN
## 2        0.042 NaN
## 3        0.040 NaN
## 4        0.072 NaN
## 5        0.148 NaN
## 6        0.568 NaN
## 7        0.568 NaN
## 8        0.218 NaN
## 9        0.090 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
```

```
## 1        0.050 NaN
## 2        0.058 NaN
## 3        0.074 NaN
## 4        0.090 NaN
## 5        0.094 NaN
## 6        0.654 NaN
## 7        0.512 NaN
## 8        0.110 NaN
## 9        0.078 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.054 NaN
## 2        0.050 NaN
## 3        0.060 NaN
## 4        0.082 NaN
## 5        0.092 NaN
## 6        0.572 NaN
## 7        0.464 NaN
## 8        0.288 NaN
## 9        0.116 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability  X1
## 1        0.024 NaN
## 2        0.028 NaN
## 3        0.034 NaN
## 4        0.070 NaN
## 5        0.146 NaN
## 6        0.056 NaN
## 7        0.986 NaN
## 8        0.138 NaN
## 9        0.076 NaN
## 10          NA NaN
##
##
## Bayesian Change Point (bcp) summary:
##
##
## Probability of a change in mean and posterior means:
##
##    Probability        X1
```

```
## 1        0.044 -0.021123
## 2        0.050 -0.020829
## 3        0.102 -0.020301
## 4        0.154 -0.018880
## 5        0.102 -0.016093
## 6        0.166 -0.013990
## 7        0.740 -0.008681
## 8        0.348  0.029618
## 9        0.106  0.043844
## 10          NA  0.046435
```

```
b_result
```

```
##    block banana_number treatment d_turn
## 1      B             1   control      2
## 2      B             2 treatment      7
## 3      B             3 treatment      4
## 4      B             4 treatment      8
## 5      B             5   control      2
## 6      B             6   control      2
## 7      B             7 treatment      3
## 8      B             8   control      2
## 9      N             1   control      7
## 10     N             2 treatment      7
## 11     N             3 treatment      7
## 12     N             4 treatment      6
## 13     N             5   control      6
## 14     N             6   control      6
## 15     N             7 treatment      7
## 16     N             8   control      7
```

Analysis

We have used block random assignement design.
We have three blocks, `B`, `N`, and `J` and the probability of assignement to treatment, $p(d = 1) = 0.5$ for all three which allows us to later pool the data across the block and estimate ATE (page 76 FE)

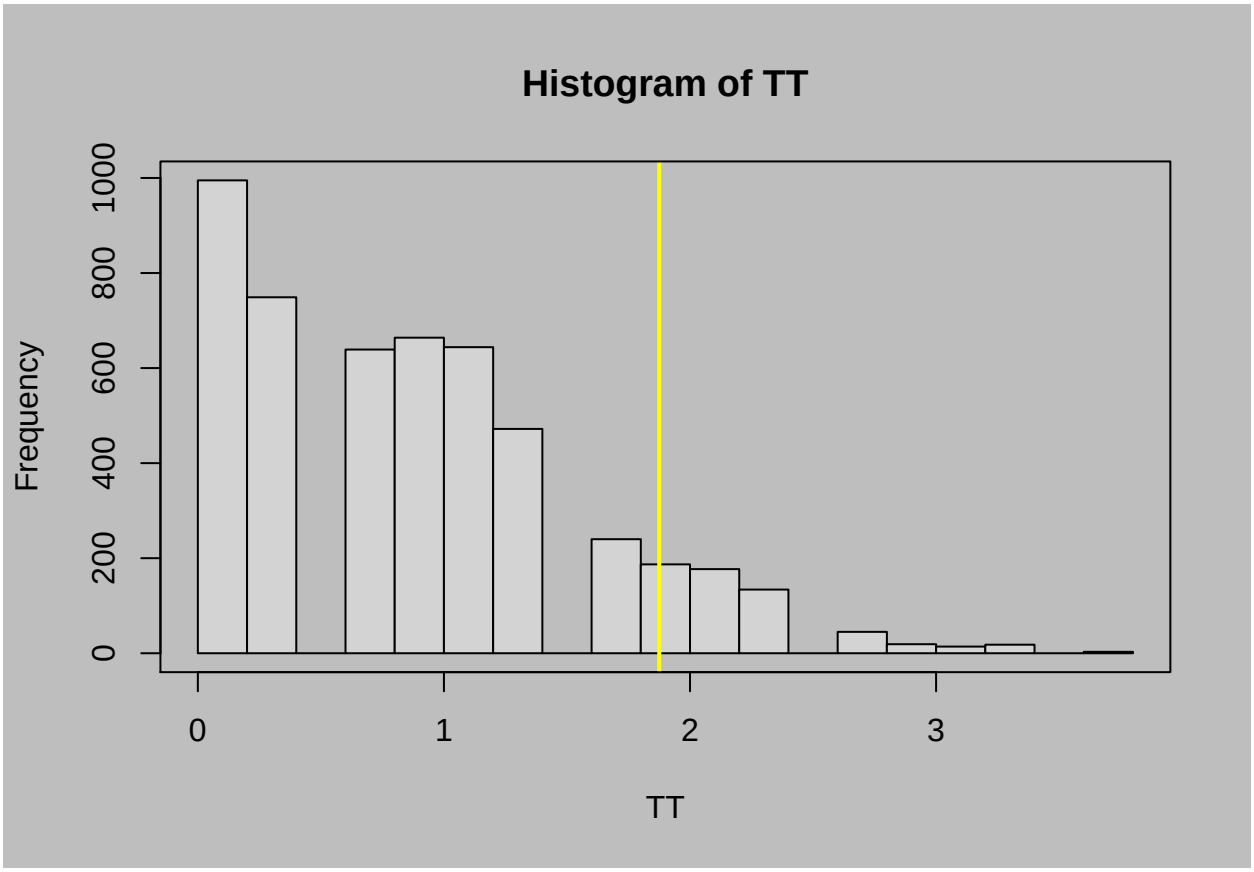
We have selected block radomization to reduce sampling variablity (see page 72 FE). The subjects within the group are from the same brunch LINK such that subjects in each block have similiar potential outcomes. This also ensured that certain subgroups are available for seperate analysis (see page 71 FE)

Sharp null hypothesis

We began by testing sharp null hypothesis of no treatment effect

the treatment effect was measured by the day beyond which the black ratio rapidly increased.

```r
#get control and treatment data
b_control <- as.integer(b_result[b_result[,"treatment"] == "control","d_turn"])
b_treatment <- as.integer(b_result[b_result[,"treatment"] == "treatment","d_turn"])

#put them into a vector
Z <- c(b_control,b_treatment)

#get the length of each vectors
n <- length(b_control)
m <- length(b_treatment)
N <- length(Z)

# Number of permutations
K = 5000

# Test statistic
getMean <- function(A,B) abs(mean(A) - mean(B))

# Test statistic for the observed sample
b_mean <- getMean(b_control,b_treatment)

# Vector of test statistics for each permutation
TT <- vector()

# Permutation test
for(i in 1:K){
 #set.seed(i)
 Z.pi <- sample(Z, N, replace = FALSE)
 TT[i] <- getMean(Z.pi[1:n], Z.pi[(n+1):(n+m)])
}

# Visualising the permuted test statistics
par(bg = 'grey')
hist(TT)
abline(v = b_mean , lwd = 2, col = "yellow")
box()
```




```
# approximate p-value
mean(TT>b_mean )
```

```
## [1] 0.082
```

Regression Estimator for Block Randomization

We have 3 blocks, B, N, and J with 8 samples in each block.

$$Y_i = c + D + B + N + J$$

TODO improve description where c and constant and D, B, N, and J are dummy varible with 1 and 0

```
mod_b <- lm(d_turn ~ treatment + as.factor(block), data = b_result)
coeftest(mod_b, vcov = vcovHC(mod_b, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         2.81250    0.49243  5.7115 7.155e-05 ***
## treatmenttreatment  1.87500    0.75080  2.4973  0.026723 *
```

```
## as.factor(block)N   2.87500    0.75080  3.8292  0.002088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b_result
```

```
##    block banana_number treatment d_turn
## 1      B             1   control      2
## 2      B             2 treatment      7
## 3      B             3 treatment      4
## 4      B             4 treatment      8
## 5      B             5   control      2
## 6      B             6   control      2
## 7      B             7 treatment      3
## 8      B             8   control      2
## 9      N             1   control      7
## 10     N             2 treatment      7
## 11     N             3 treatment      7
## 12     N             4 treatment      6
## 13     N             5   control      6
## 14     N             6   control      6
## 15     N             7 treatment      7
## 16     N             8   control      7
```

```
b_result %>% ggplot(aes(x = d_turn, fill=treatment)) + geom_bar() +  facet_wrap(~block,ncol = 1)
```