# Homework 2: Evaluating Randomized Experiments

*Neeraj Sharma*

*05/02/2020*

## Question 1: Estimate the treatment effect and the associated standard error on the raw data given to you, assuming no issues with the data.

By definition, the Average Treatment Effect is defined to be $\frac{1}{N}\sum_i y_1(i) - y_0(i)$ where $y_1(i)$ and $y_0(i)$ are the values of the outcome variable (in this case hours exercised) in each treatment scenario. Unfortunately, it is often impractical to utilize this straightforward approach. This formula presumes that one can quantify the outcome of a given individual when treatment is both given and withheld. However, because each individual can only be slotted into one category, this approach cannot be perfectly achieved.

Thus, a random experiment with a control and treatment group can be conducted to smooth out differences among populations in order to isolate the treatment effect specifically. With large enough sample groups, the difference between the mean of the outcome of the treatment group and mean of the outcome of the control group yields the Average Treatment Effect.

Here are the relevant summary statistics of the data set without any modification.

Table 1: Summary of Hours Exercised, No Data Cleaning

| Treatment | Count | Mean | St Dev | St Err |
|---|---|---|---|---|
| 0 | 500 | 20.952 | 84.02863 | 3.757875 |
| 1 | 500 | 27.546 | 104.36699 | 4.667434 |

The mean hours exercised for individuals in the treatment group is 27.546 and the mean hours exercised for individuals in the non-treatment group is 20.952. Assuming randomness was properly implimented in this study, the difference between these two numbers will be the Average Treatment Effect of the treatment based on the analysis I provide above. Thus, the Average Treatment Effect is **6.594** with a standard error of **0.9096**.

## Question 2: It's always a good idea to check a data set for errors. Clean this data set as you think appropriate. As an answer to this question, note all the kinds of changes you made to the data, a few words explaining your reason for the change, and which observations you changed (noting the observation number included as a variable in the data set for identification purposes). If it is totally obvious which observations you changed and there are a large number in the category (e.g. if you decided to drop all White study participants), you can just note what you did for that change ("I dropped all white participants") and explain why.

I noticed several types of data errors and compiled a representative sample of troubled observations here.

Table 2: Representative sample of unclean observations

| subject_id | hours | treatment | community_center | female | age | bmi | education | race_ethnicity |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | Woodlawn | 0 | 44 | 28.6634560 | high school | BLACK |
| 9 | 14 | 0 | WOODLAWN | 0 | 19 | 29.3965780 | high school | black |
| 16 | 180 | 0 | hyde park | 1 | 32 | 26.9350070 | higher degree | BLACK |

| subject_id | hours | treatment | community_center | female | age | bmi | education | race_ethnicity |
|---|---|---|---|---|---|---|---|---|
| 26 | 4 | 0 | hyde park | 0 | -99 | 33.4841000 | higher degree | black |
| 31 | 13 | 1 | Woodlawn | female | 22 | 22.7878760 | higher degree | BLACK |
| 52 | 1 | 0 | hyde park | male | 51 | 35.4018250 | higher degree | white |
| 65 | 11 | 1 | Woodlawn | 0 | 20 | 31.7927250 | high school | NA |
| 100 | 7 | 0 | Hyd Park | 1 | 39 | 34.9566570 | higher degree | white |
| 103 | 8 | 1 | Woodlawn | 1 | 37 | 0.2565794 | high school | BLACK |

Given these errors, I perform the following modifying operations to clean the dataset. See appendicies for the specific code I use to accomplish these modifications.

| Variable | Description of Modification |
|---|---|
| subject_id | No change |
| hours | Hours over 60 are minutes; reformatted to be hours. |
| treatment | No change. |
| community_center | All variant spellings recoded to "Woodlawn" and "Hyde Park." |
| female | 0/1/male/female recoded uniformly as 0/1. |
| age | -99 means missing age data; recoded to be NA. |
| bmi | BMIs of less than 1 have undetermined error; recoded to be NA.* |
| education | No change. |
| race_ethnicity | Race/Ethnicity "BLACK" capitalization recoded to "Black." |
| changed? | `changed?` notes and counts changes that occured in cleaning an observation. |

*The source of error on BMI is very difficult to determine. There are multiple hypotheses I have for the questionable data. It could be an error in data entry where the decimal was misplaced in which case multiplying by 100 would resolve the issue. It could be that the data was entered in a different unit than normal (pounds vs kilograms or meters vs feet) and the application of the BMI formula resulted in very low numbers. It is impossible to certify which error occured, if any, so the safest option is to ignore these 26 rows. This is a small enough number that it will not significantly harm my interpretation.

## Question 3: With your cleaned data set, re-estimate the treatment effect and estimated standard error, assuming the randomization worked fine.

Table 4: Summary of Hours Exercised, Data Cleaned

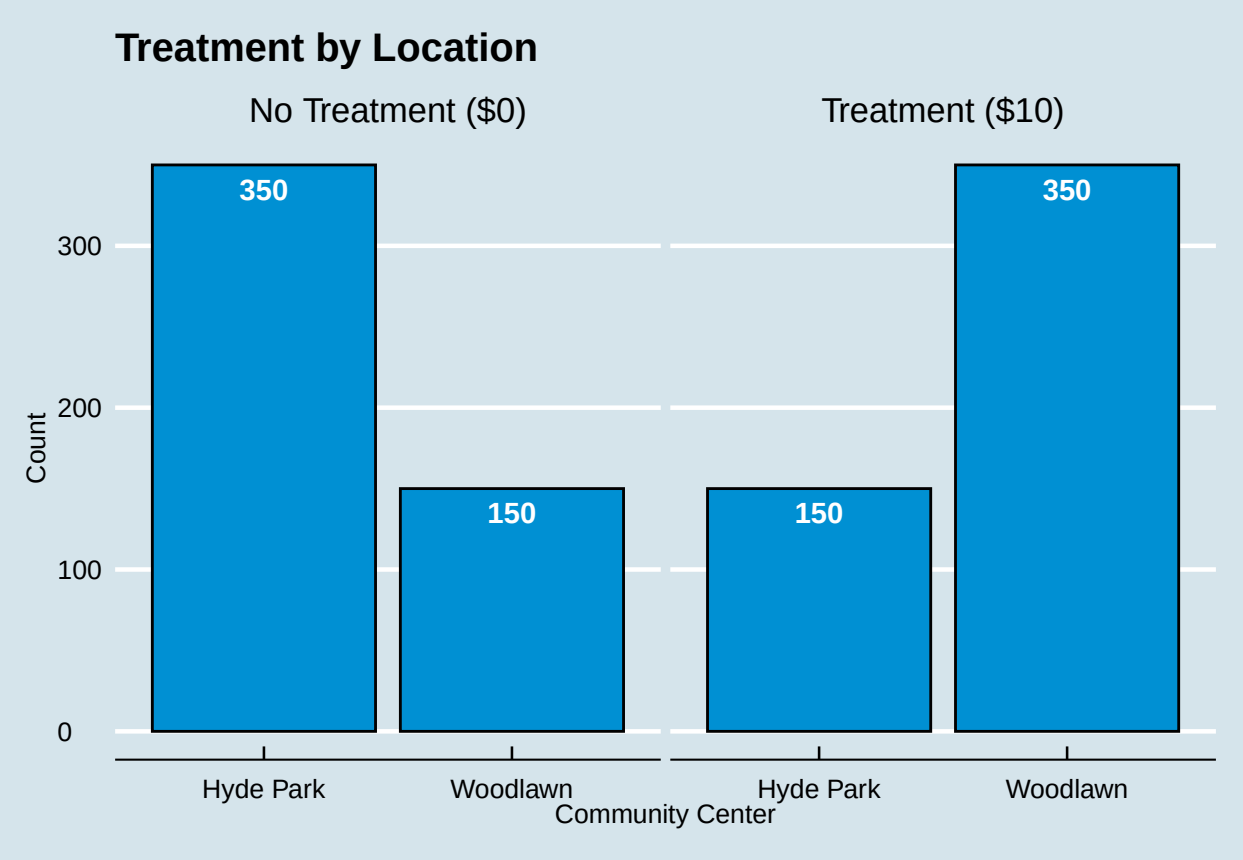
| Treatment | Count | Mean | St Dev | St Err |
|---|---|---|---|---|
| 0 | 500 | 5.612 | 3.387093 | 0.1514754 |
| 1 | 500 | 7.486 | 3.669432 | 0.1641020 |

The mean number of hours exercised in both the treatment and control groups has fallen dramatically upon cleaning the data. Numerous entries were coded in minutes instead of hours and those observations were dragging the values up significantly. Those errors have since been corrected. Thus, the Average Treatment Effect upon cleaning the data yet assuming proper randomization is is **1.874** with a standard error of **0.0126**.

## Question 4: Evaluate whether the randomization appears legitimate. If not, what is your evidence? (Hint: something went wrong.)

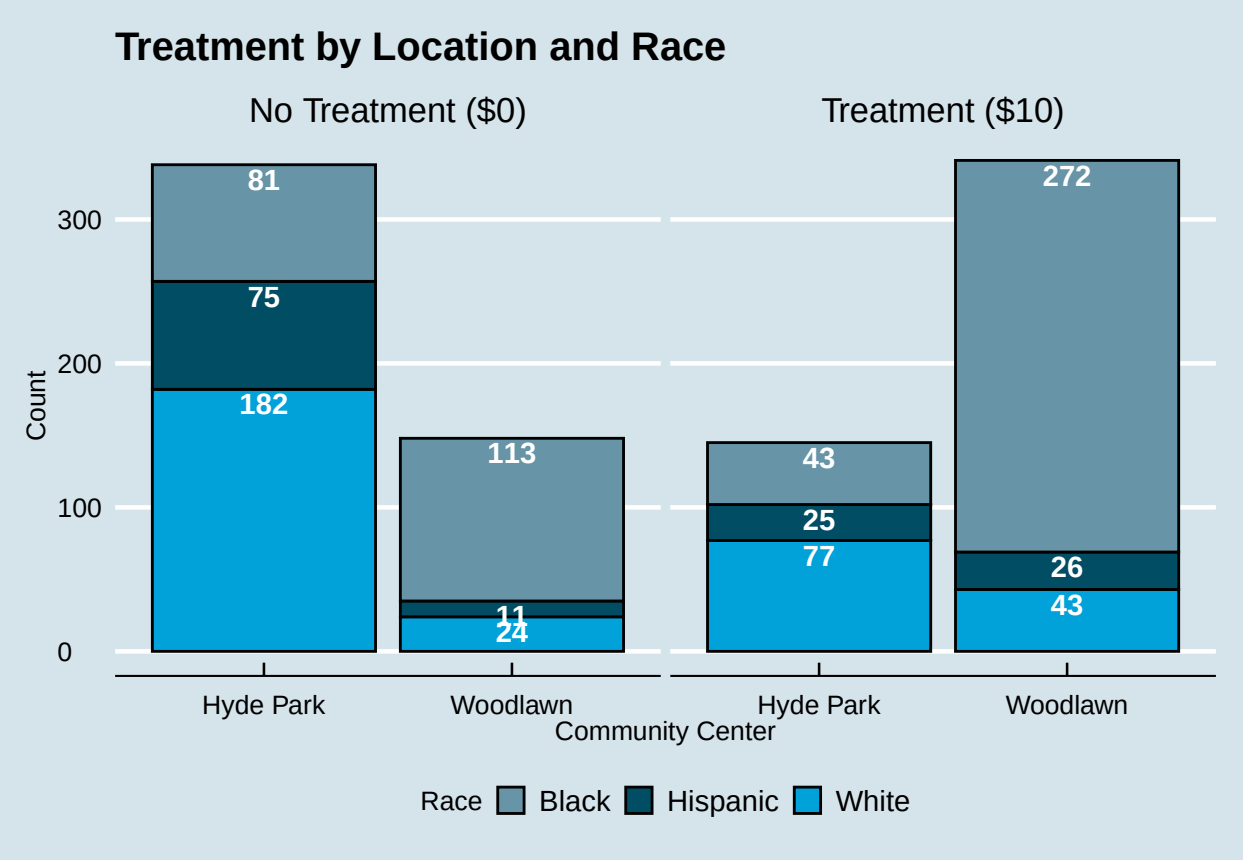
An important piece of insight I gained from Eric's office hours were that given a sufficiently large sample size, perfect random assignment means by definition that one could randomly subdivide the data into equally

sized groups of people and the distribution of the covariates should be the same. Thus, if I were to create histograms of some of the specified covariates in this dataset like BMI, age, race, or community, I should observe similar distributions and summary statistics to justify the representativeness and randomness of the allocation of the sample groups. Eric noted that on this problem set specifically is straightforward enough with regards to the identification of systematic differences that imply a non-random allocation that full t-test would not be necessary.

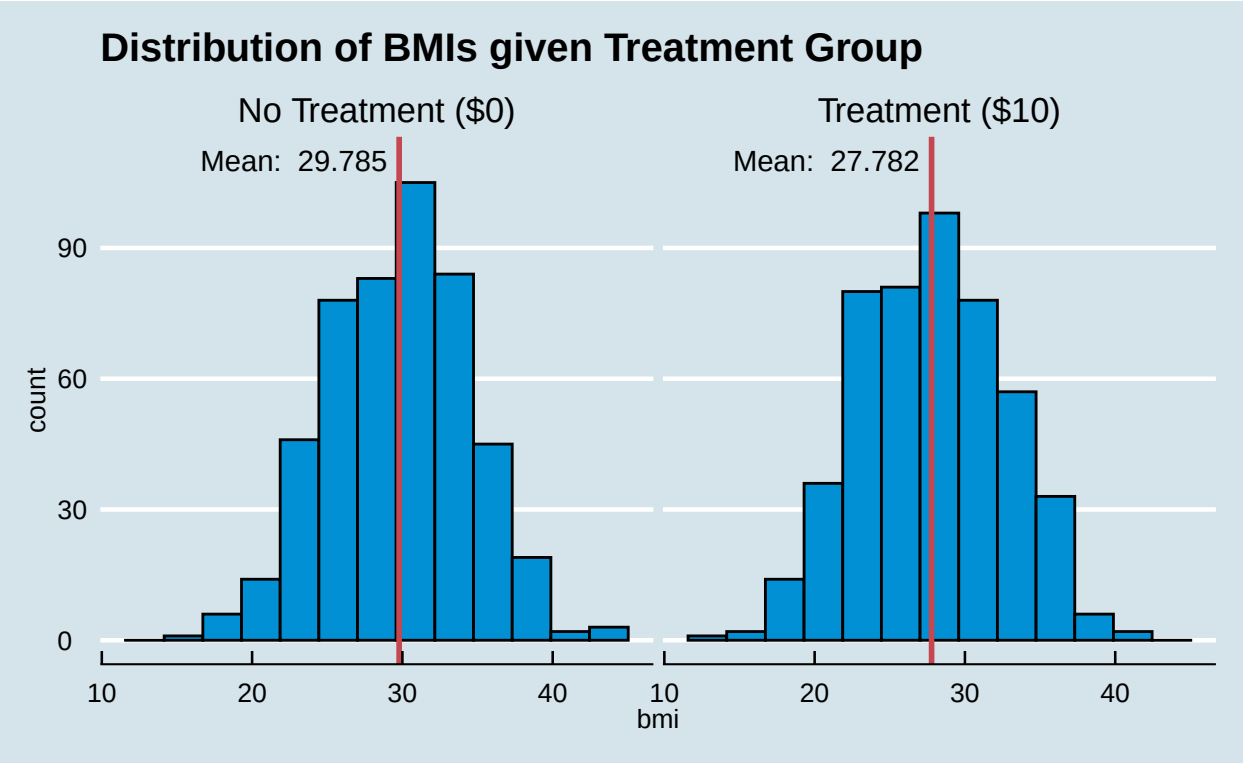



This plot clearly shows a bias in the group allocations based on survey site. Were this data to be randomly created, we would expect to see an even distribution of individuals in the treatment and non-treatment groups across Woodlawn and Hyde Park. Instead, we find a disproportionate amount of the control group was assigned at Hyde Park and conversely, a disproportionate number of people in the treatment group were assigned in Woodlawn. A random experiment would not exhibit this property.

Further analysis on other covariates supports the conclusion that race is not controlled between treatment groups.




The third and final variable I perform significant analysis on to understand the pattern of randomness is BMI.



There is an apparent difference between the mean BMI in the treatment and non-treatment group, but it is unclear if it is significant enough to warrant inclusion as a source of failed randomization. A simple T-Test can shed more light.

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | alternative |
|---|---|---|---|---|---|---|---|---|
| 2.003217 | 29.78517 | 27.78196 | 6.527694 | 1e-10 | 971.5728 | 1.400993 | 2.60544 | two.sided |

The T-Test reveals that there is a statistically significant difference between the means of the two sample groups given the p-value of 1e-10.

In general, plots of most other covariates look fairly decently normal. For brevity, I only provide summary statistics to articulate this point:

Table 6: Summary of random covariates

| treatment | Mean Female | Mean Age | less than high school | high school | higher degree |
|---|---|---|---|---|---|
| No Treatment ($0) | 0.636 | 32.86141 | 54 | 208 | 238 |
| Treatment ($10) | 0.582 | 31.68421 | 45 | 253 | 202 |

## Question 5: Offer your best hypothesis/hypotheses as to what went wrong with the randomization? What evidence do you have to support your hypothesis(es)? For each of these hypotheses, describe your best strategy for estimating a plausible treatment effect, in spite of the bad randomization. (But don't actually estimate that treatment effect.)

My hypothesis is that Justin was trigger happy with the treatment assignment because he was stationed at the Woodlawn community center which is overrepresented in the treatment group. Furthermore, I predict that the differences in race and BMI are also side effects of this incorrect sampling.

## Question 6: Given your answer to question five come up with your best estimate of the true treatment effect in the experiment as well as its standard error.

```
## Analysis of Variance Table
##
## Model 1: hours ~ treatment
## Model 2: hours ~ treatment + bmi
## Model 3: hours ~ treatment + bmi + hispanic + white
## Model 4: hours ~ treatment + bmi + hispanic + white + community_center
##   Res.Df     RSS Df Sum of Sq       F    Pr(>F)
## 1    972 12167.9
## 2    971 10750.1  1   1417.83 147.304 < 2.2e-16 ***
## 3    969 10260.1  2    489.98  25.453 1.686e-11 ***
## 4    968  9317.2  1    942.93  97.965 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = hours ~ treatment, data = fittingd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -6.502 -2.607 -0.502  2.393 14.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.6070     0.1605  34.936   <2e-16 ***
## treatment     1.8951     0.2267   8.358   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.538 on 972 degrees of freedom
## Multiple R-squared:  0.06705,    Adjusted R-squared:  0.06609
## F-statistic: 69.85 on 1 and 972 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = hours ~ treatment + bmi, data = fittingd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2284 -2.4662 -0.1754  2.0693 13.8056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.11873    0.68072   19.27  < 2e-16 ***
## treatment    1.38985    0.21785    6.38 2.74e-10 ***
## bmi         -0.25220    0.02229  -11.32  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.327 on 971 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1741
## F-statistic: 103.5 on 2 and 971 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = hours ~ treatment + bmi + hispanic + white, data = fittingd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5520 -2.3506 -0.1716  2.0682 12.5675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.01374    0.66591  19.543  < 2e-16 ***
## treatment    1.64346    0.21749   7.557 9.57e-14 ***
## bmi         -0.27308    0.02215 -12.327  < 2e-16 ***
## hispanic     0.36748    0.32124   1.144    0.253
## white        1.61228    0.23989   6.721 3.08e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.254 on 969 degrees of freedom
## Multiple R-squared:  0.2133, Adjusted R-squared:  0.2101
## F-statistic: 65.69 on 4 and 969 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = hours ~ treatment + bmi + hispanic + white + community_center,
##     data = fittingd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9733 -2.1933 -0.1751  1.8392 11.3179
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.56240    0.77805  11.005  < 2e-16 ***
## treatment         1.02296    0.21663   4.722 2.68e-06 ***
## bmi              -0.16998    0.02355  -7.218 1.07e-12 ***
## hispanic          1.21083    0.31791   3.809 0.000148 ***
## white             2.66216    0.25212  10.559  < 2e-16 ***
## community_center  2.69027    0.27181   9.898  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102 on 968 degrees of freedom
## Multiple R-squared:  0.2856, Adjusted R-squared:  0.2819
## F-statistic: 77.41 on 5 and 968 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 974 x 10
##    hours treatment   bmi .fitted .se.fit .resid    .hat .sigma .cooksd
##    <dbl>     <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>
##  1     2         0  28.7    5.89   0.153 -3.89  0.00211   3.33 9.67e-4
##  2     2         0  25.2    6.75   0.182 -4.75  0.00299   3.33 2.05e-3
##  3    12         1  31.5    6.57   0.172  5.43  0.00266   3.32 2.38e-3
##  4    13         1  30.1    6.92   0.159  6.08  0.00229   3.32 2.56e-3
##  5     2         1  31.0    6.70   0.167 -4.70  0.00251   3.33 1.67e-3
##  6     2         1  20.9    9.23   0.214 -7.23  0.00415   3.32 6.58e-3
##  7     7         1  27.8    7.50   0.151 -0.496 0.00205   3.33 1.52e-5
##  8     4         1  29.3    7.12   0.154 -3.12  0.00215   3.33 6.34e-4
##  9    14         0  29.4    5.70   0.151  8.30  0.00206   3.32 4.29e-3
## 10     2         0  27.9    6.08   0.157 -4.08  0.00222   3.33 1.12e-3
## # ... with 964 more rows, and 1 more variable: .std.resid <dbl>
```

```
## # A tibble: 974 x 13
##    hours treatment   bmi hispanic white community_center .fitted .se.fit
##    <dbl>     <dbl> <dbl>    <dbl> <dbl>            <dbl>   <dbl>   <dbl>
##  1     2         0  28.7        0     0                1    6.38   0.221
##  2     2         0  25.2        1     0                0    5.48   0.318
##  3    12         1  31.5        0     0                1    6.92   0.202
##  4    13         1  30.1        1     0                0    5.68   0.318
##  5     2         1  31.0        0     0                1    7.01   0.195
##  6     2         1  20.9        0     1                0    8.69   0.332
##  7     7         1  27.8        0     0                1    7.55   0.166
##  8     4         1  29.3        0     0                0    4.61   0.264
##  9    14         0  29.4        0     0                1    6.26   0.225
## 10     2         0  27.9        0     0                0    3.82   0.245
## # ... with 964 more rows, and 5 more variables: .resid <dbl>, .hat <dbl>,
## #   .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

## Appendicies

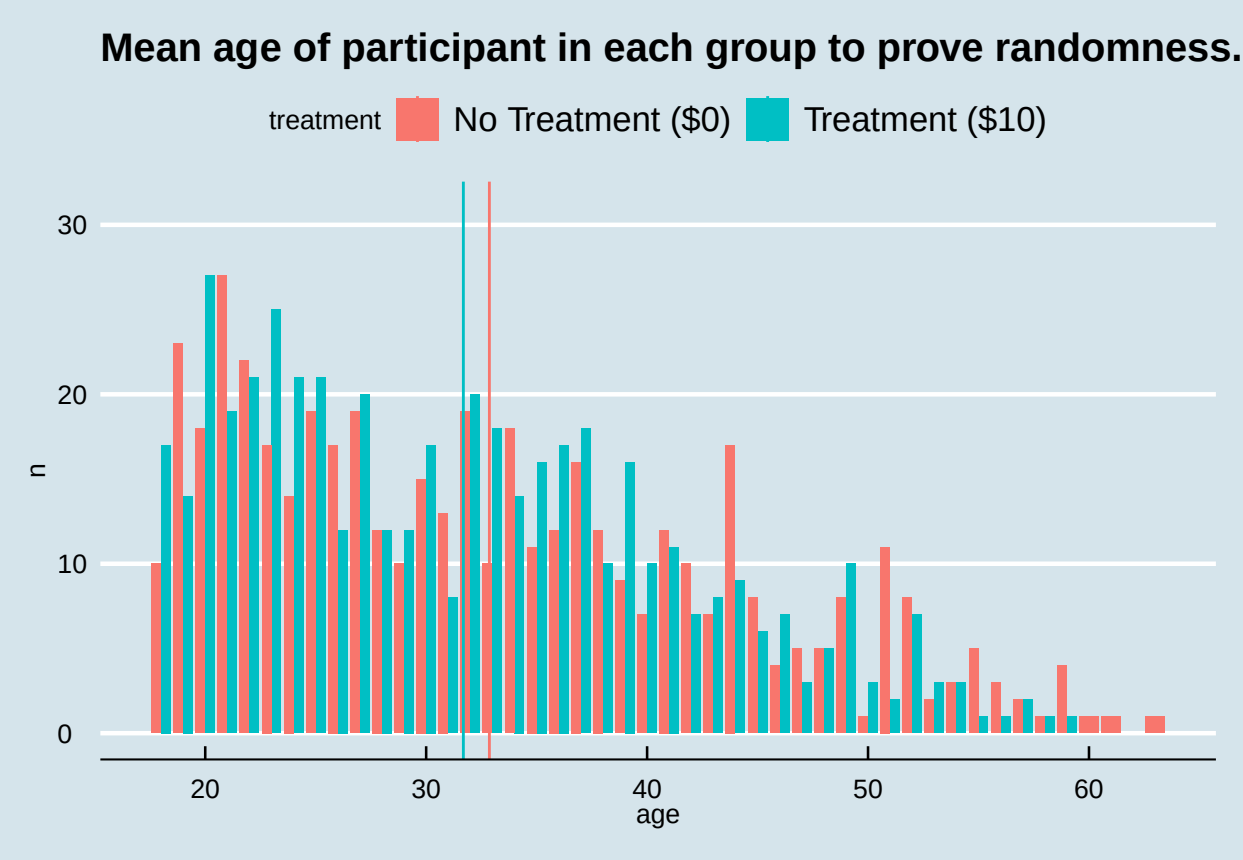
### Code to clean data in question 2

```
# Clean the data to impose uniformity upon the variable encoding.
clean_data <- raw %>%
  mutate(`changed?` = seq(1, 1000) * 0) %>%
  # Fixing messed up hours readings. They start at 60 and upwards and that's 1 hr so I
  # fix based on that.
  mutate(`changed?` = if_else(hours >= 60, `changed?` + 1, `changed?`),
         hours = if_else(hours >= 60, hours / 60, hours)) %>%
  mutate(`changed?` = if_else(community_center %in% c("WOODLAWN",
                                                      "hyde park",
                                                      "woodlawn",
                                                      "Hyd Park",
                                                      "HYDE PARK",
                                                      "Hyde_Park",
                                                      "HYDE_PARK",
                                                      "hyde_park"),
                              `changed?` + 1, `changed?`),
         community_center = if_else(community_center %in% c("WOODLAWN", "woodlawn"),
                                    "Woodlawn", community_center),
         community_center = if_else(community_center %in% c("hyde park",
                                                            "Hyd Park",
                                                            "HYDE PARK",
                                                            "HYDE_PARK",
                                                            "hyde_park",
                                                            "Hyde_Park"),
                                    "Hyde Park", community_center)) %>%
  # Recoding female variable to be factor categorical variable from 0/1/male/female.
  mutate(`changed?` = if_else(female %in% c("female", "male"),
                              `changed?` + 1, `changed?`),
         female = as.double(if_else(female == "female",
                                    "1", if_else(female == "male", "0", female)))) %>%
  # -99 is missing age data so I reincode it at missing age data.
  # https://cran.r-project.org/web/packages/naniar/vignettes/replace-with-na.html
  mutate(`changed?` = if_else(age == -99, `changed?` + 1, `changed?`),
         age = na_if(age, -99)) %>%
  # Currently I have removed improper values, but I could also justify multiplying by 10.
  mutate(`changed?` = if_else(bmi < 1, `changed?` + 1, `changed?`),
         bmi = ifelse(bmi < 1, NA, bmi)) %>%
  # Fix all BLACK observations to normal capitalization structure.
  mutate(`changed?` = if_else(is.na(race_ethnicity),
                              `changed?`, if_else(race_ethnicity == "BLACK",
                                                  `changed?` + 1, `changed?`)),
    race_ethnicity = if_else(is.na(race_ethnicity),
                             race_ethnicity, str_to_sentence(race_ethnicity))) # %>%
# mutate(race_ethnicity = factor(race_ethnicity, c("Black", "Hispanic", "White"))) %>%
# mutate(treatment = factor(treatment, labels = c("No Treatment ($0)", "Treatment ($10)"))) %>%
# mutate(community_center = factor(community_center, labels = c("Woodlawn", "Hyde Park"))) %>%
# mutate(education = factor(education, levels = c("less than high school",
#                                                 "high school",
#                                                 "higher degree")))
```

```
## # A tibble: 2 x 2
```

```
##   treatment         `mean(age, na.rm = TRUE)`
##   <fct>                                 <dbl>
## 1 No Treatment ($0)                      32.9
## 2 Treatment ($10)                        31.7
```

```
## Warning: Removed 2 rows containing missing values (geom_col).
```




```
##
##  Welch Two Sample t-test
##
## data:  bmi by treatment
## t = 6.5277, df = 971.57, p-value = 1.074e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.400993 2.605440
## sample estimates:
## mean in group 0 mean in group 1
##        29.78517        27.78196
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment by community_center
## t = -13.377, df = 971.87, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4521386 -0.3364494
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        0.305499        0.699793
```