# R Camp - Day 3 Homework

*Emily Elia*

*8/22/2019*

## Problem 1 - Simulating Experimental Data

### Experiment Set Up

For my experiment, I test the presence of an anti-corruption plan in a mayorial candidate's platform on candidate likability by potential voters. In my experiment, a population of 2000 voters are presented with two vignettes: one about an incumbent seeking reelection, and one about a challenger candidate. Everyone receives the same vignette for the incumbent, which includes basic information about the incumbent's history and plans for the future if reelected. Everything included in the vignette is nonpartisan, nonpolarizing information, such as "During Mayor Smith's term, he has created 150 jobs and oversaw a infrasturcture improvement plan." Then, everyone receives a vignette about the challenger candidate. This vignette also includes basic, non-partisan information about the candidate's qualifications and plans if elected. However, the oppostion vignette has a control version and a treatment version which is randomly assigned to experiment participants. The control version talkes about the qualifications and plans, while the treatment version has an additional line stating that the candidate intends to implement a comprehensive plan to catch and combat potential corruption amongst public officials in the municipality. After reading the vignette, subjects are then asked to indicate on a scale of 0-100 how likely they would be to elect the oppostion challenger candidate over the incumbent.

For the experiment, I include two co-variates: gender and age. I include these two covariates because literature on corruption perceptions often shows evidence that females are more adverse to corruption than males, and that older people are more adverse to corruption than younger people. I generate gender by simulating a binary variable, female, where 1=female and 0=male. I then generate age by drawing from a random poisson distribution with a lambda of 40 so that the age distribution falls mostly in the middle age range from 30 to 50 with fewer elderly subjects while also limiting the distribution to most closely resemble a voting age population where age does not dip below ~18-20 years of age.

I simulate my covariates by drawing from random distributions. I then assign fixed values to my covariate coefficients and create a fixed treatment effect. I build my potential outcome variables - one for those receiving the treatment and one for thos in the control group - with my coefficients and covariates.

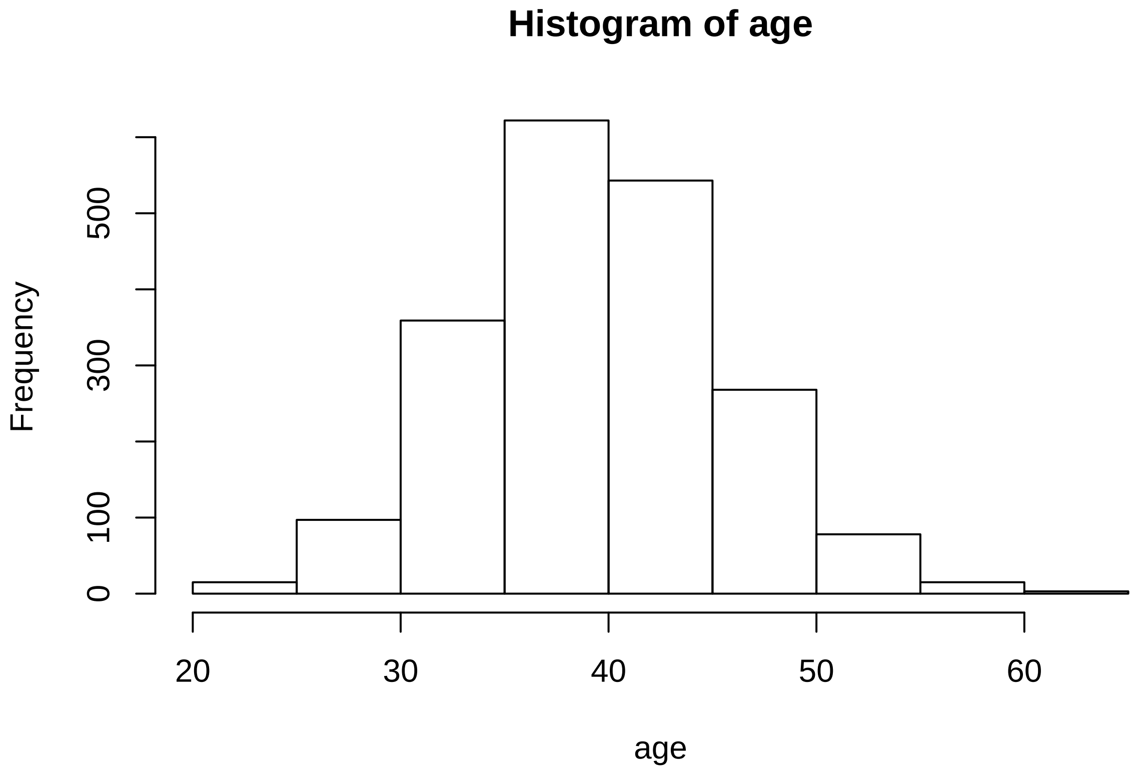
```
rm(list=ls())

# Global parameters
set.seed(212121)
N = 2000                                    # Population size
N_samp <- 50                                # Sample size

# Population parameters
female <- rbinom(n=N, 1, 0.4)              # Some binary covariate
age <- rpois(n=N, lambda=40)         # Some continuous covariate
mean(female)
```

```
## [1] 0.3775
```

```
mean(age)
```

```
## [1] 39.991
```

```
hist(age)
```



```
a = 0.5                                    # Intercept: Ground mean
b = 2                                      # (Fixed=constant) Effects of female
b2 = 1.50                                   # (Fixed=constant) Effects of age
tau = 5                                    # (Fixed=constant) Treatment effect
e = rnorm(n=N, mean=0, sd=0.25)               # N(0,1) error

Y_0 <- a + tau*0 + (b*female) + (b2*age) + e  # Potential outcome when not treated
Y_1 <- a + tau*1 + (b*female) + (b2*age) + e  # Potential outcome when treated

range(Y_0) #33.68 - 102.70
```

```
## [1] 30.74945 96.33202
```

```
range(Y_1) #38.68 - 107.70
```

```
## [1]  35.74945 101.33202
```

```
median(Y_0) #60.92
```

```
## [1] 60.73535
```

```
median(Y_1) #65.92
```

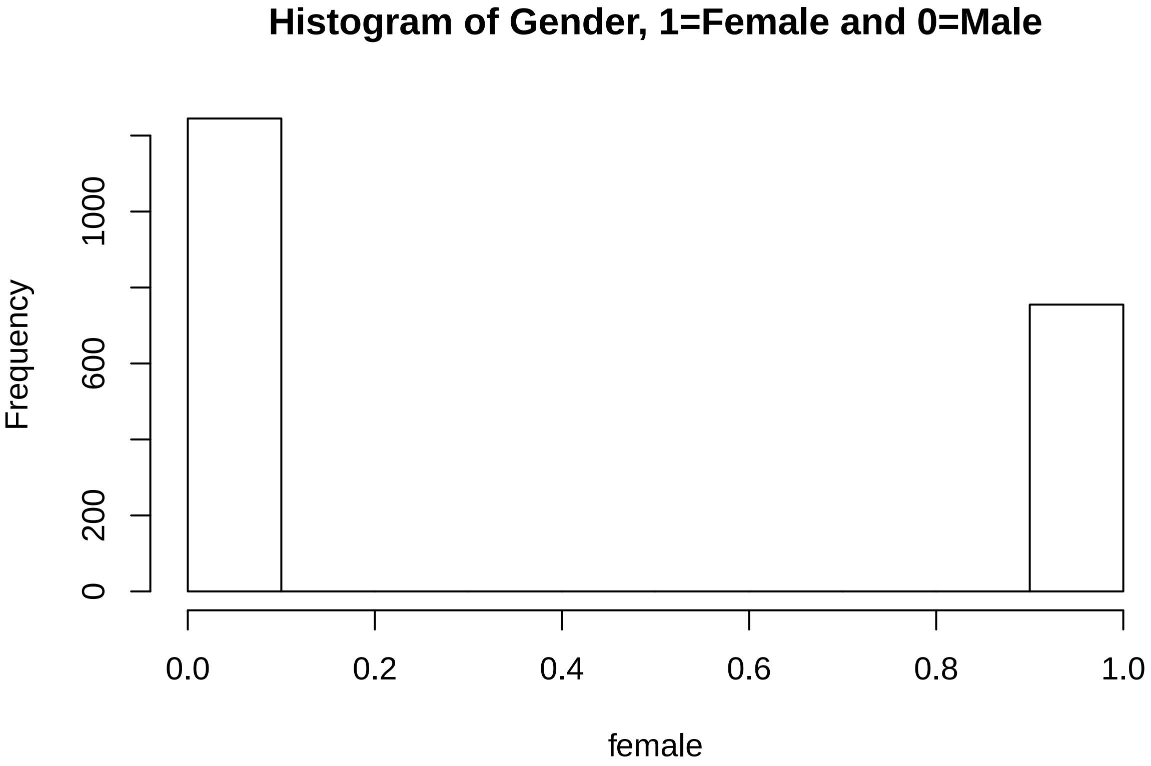
```
## [1] 65.73535
```

```
pop_dat <- data.frame(Y_0, Y_1, female, age) # Population level data
head(pop_dat)
```

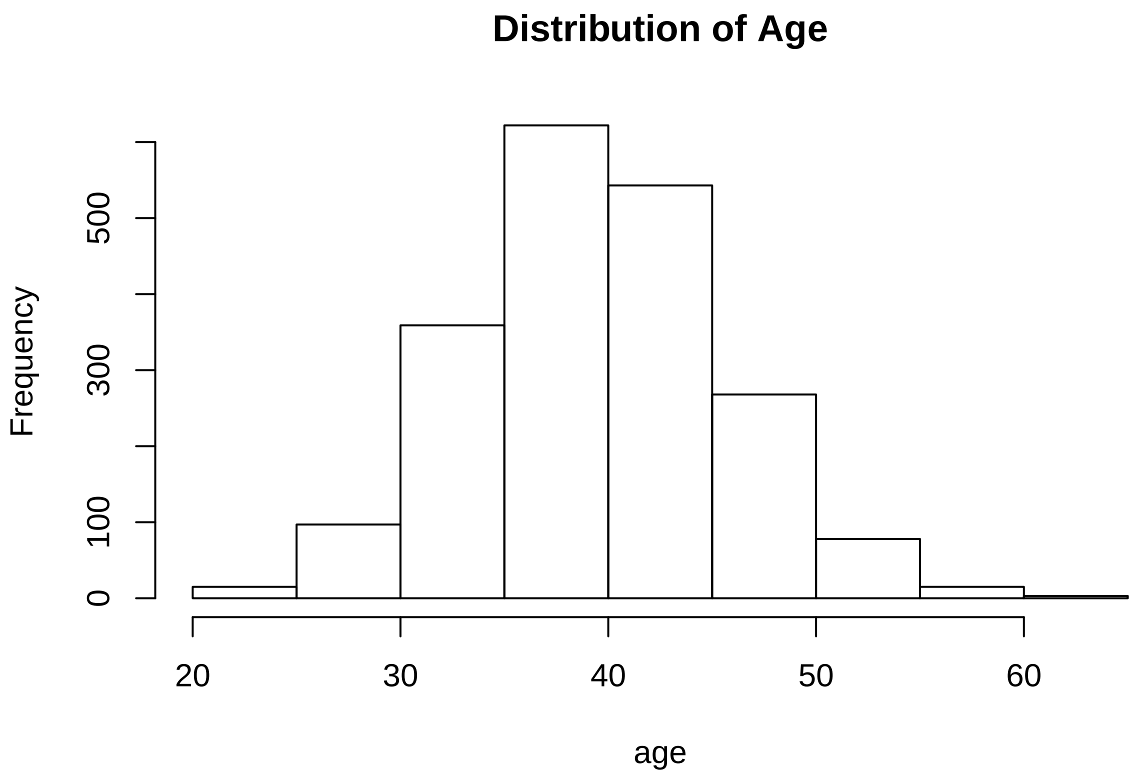
```
##        Y_0      Y_1 female age
## 1 77.32678 82.32678      1  50
## 2 54.33245 59.33245      0  36
## 3 61.89548 66.89548      0  41
## 4 63.64301 68.64301      0  42
## 5 54.30083 59.30083      0  36
## 6 63.17072 68.17072      0  42
```

### Visualizing Covariates

```
hist(female, main=("Histogram of Gender, 1=Female and 0=Male"))
```



```
hist(age, main=("Distribution of Age"))
```



## Initial Sample Size - 50

After building my variables and potential outcomes, I randomly sample 50 subjects from my N of 2000 voters. In Problem 2, I alter the sample size and examine results further.

```
sample_ind <- sample(1:nrow(pop_dat), size=N_samp)      # Sampling index
sample_dat <- pop_dat[sample_ind, ]                     # Only keep obs that match the index
head(sample_dat)
```

```
##            Y_0      Y_1 female age
## 131  54.48304 59.48304      0  36
## 1899 78.56214 83.56214      1  51
## 482  56.28647 61.28647      0  37
## 874  58.94504 63.94504      0  39
## 788  51.20042 56.20042      0  34
## 878  56.32351 61.32351      0  37
```

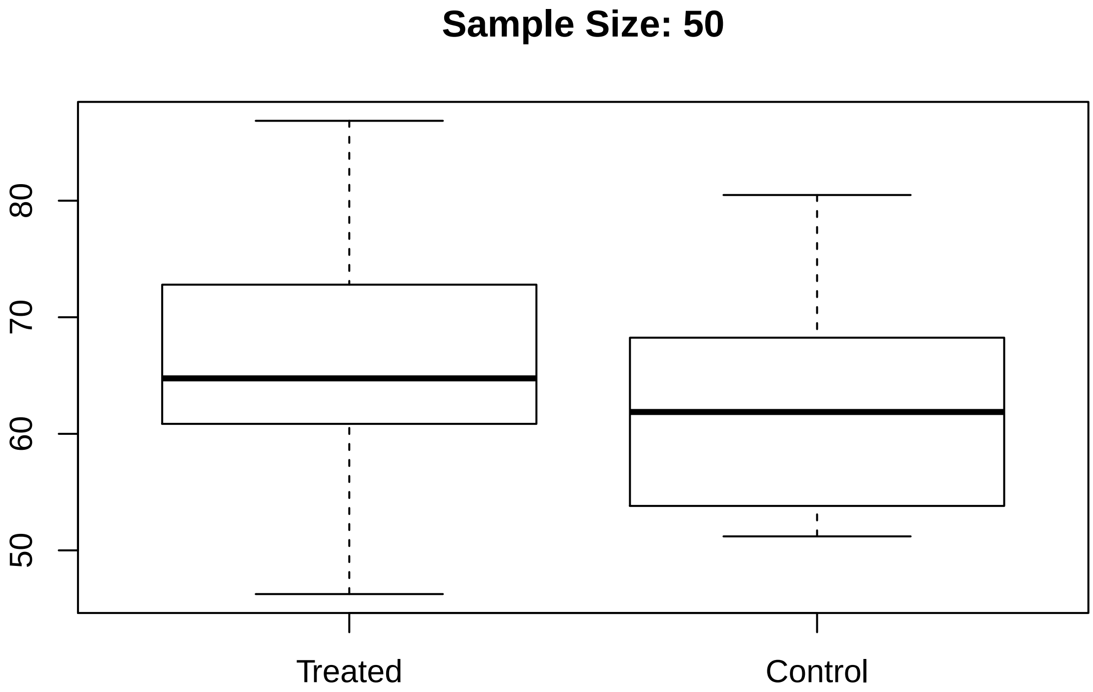
I then randomly assign my treatment to half of my sample population and calculate what would be the observed outcomes for treatment and control groups.

```
d <- ifelse(runif(N_samp)<=0.5, 1, 0)                       # Treatment assignment indicator (1=Treate
sample_dat$Y_obs <- d*sample_dat$Y_1 + (1-d)*sample_dat$Y_0  # Observed outcomes
# Y_obs = d*Y_1 + (1-d)*Y_0
sample_dat$Status <- ifelse(d==1, "Treated", "Control")      # Copying the treatment status into sample

head(sample_dat)
```

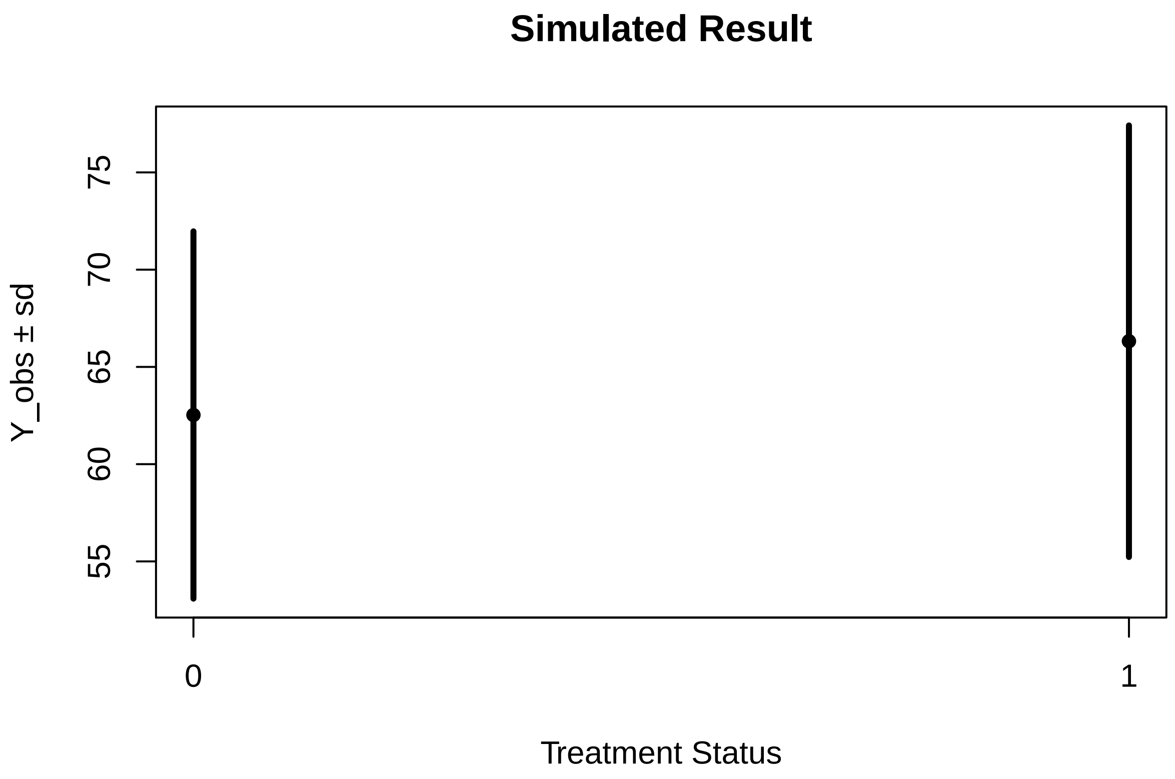
```
##            Y_0      Y_1 female age    Y_obs  Status
## 131  54.48304 59.48304      0  36 59.48304 Treated
## 1899 78.56214 83.56214      1  51 78.56214 Control
## 482  56.28647 61.28647      0  37 56.28647 Control
## 874  58.94504 63.94504      0  39 63.94504 Treated
## 788  51.20042 56.20042      0  34 51.20042 Control
## 878  56.32351 61.32351      0  37 61.32351 Treated
```

To examine the impact of an anti-corruption plan on subjects' likeliness to vote for the challenger candidate, I first use a box plot to show the impact of the treatment on treated subjects vs. control subjects. I then use a second plot to showcase the difference.

```
boxplot(sample_dat$Y_obs[sample_dat$Status=="Treated"],
        sample_dat$Y_obs[sample_dat$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 50"))
```



```
t <- c(0,1)
y_mean <- c(mean(sample_dat$Y_obs[sample_dat$Status=="Control"]),
            mean(sample_dat$Y_obs[sample_dat$Status=="Treated"]))
y_sdv <- c(sd(sample_dat$Y_obs[sample_dat$Status=="Control"]),
           sd(sample_dat$Y_obs[sample_dat$Status=="Treated"]))

plot(y_mean ~ t, pch=16, ylim=range(c(y_mean-y_sdv, y_mean+y_sdv)),
     xlab="Treatment Status", ylab="Y_obs ± sd",xaxt="n") # Without x-axis lable
axis(1, at = seq(00, 1, by = 1), las=1) # If las=2, numbers will be flipped by 90 degree
arrows(t, y_mean-y_sdv, t, y_mean+y_sdv, length=0, angle=90, lwd=3)
title("Simulated Result")
```



## Problem 2 - Altering Sample Sizes

For Problem 2, I repeat the experiment four times but I alter the sample sizes. I use a sample size of 100, of 150, of 200, and of 250 to compare to my first sample size of 50. I then demonstrate the difference in experimental outcomes by sample size through boxplots.

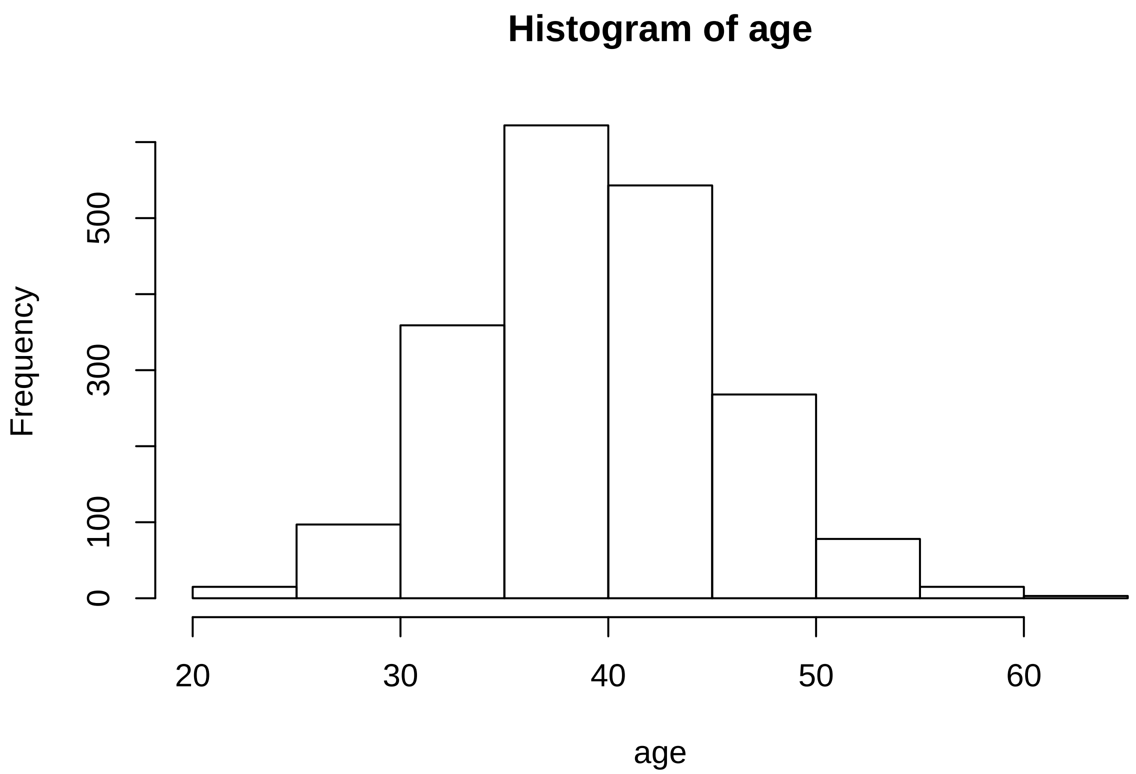
```
##### Simulation and Model #####

### Sample Size = 100 ###

# Global parameters
set.seed(212121)
N = 2000                                    # Population size
N_samp2 <- 100                                # Sample size

# Population parameters
female <- rbinom(n=N, 1, 0.4)              # Some binary covariate
age <- rpois(n=N, lambda=40)         # Some continuous covariate
mean(female)
```

```
## [1] 0.3775
```

```
mean(age)
```

```
## [1] 39.991
```

```
hist(age)
```



```
a = 0.5                                    # Intercept: Ground mean
b = 2                                      # (Fixed=constant) Effects of female
b2 = 1.50                                   # (Fixed=constant) Effects of age
tau = 5                                    # (Fixed=constant) Treatment effect
e = rnorm(n=N, mean=0, sd=0.25)               # N(0,1) error

Y_0 <- a + tau*0 + (b*female) + (b2*age) + e  # Potential outcome when not treated
Y_1 <- a + tau*1 + (b*female) + (b2*age) + e  # Potential outcome when treated

range(Y_0) #33.68 - 102.70
```

```
## [1] 30.74945 96.33202
```

```
range(Y_1) #38.68 - 107.70
```

```
## [1]  35.74945 101.33202
```

```
median(Y_0) #60.92
```

```
## [1] 60.73535
```

```
median(Y_1) #65.92
```

```
## [1] 65.73535
```

```
pop_dat <- data.frame(Y_0, Y_1, female, age) # Population level data
head(pop_dat)
```

```
##        Y_0      Y_1 female age
## 1 77.32678 82.32678      1  50
## 2 54.33245 59.33245      0  36
## 3 61.89548 66.89548      0  41
## 4 63.64301 68.64301      0  42
## 5 54.30083 59.30083      0  36
## 6 63.17072 68.17072      0  42
```

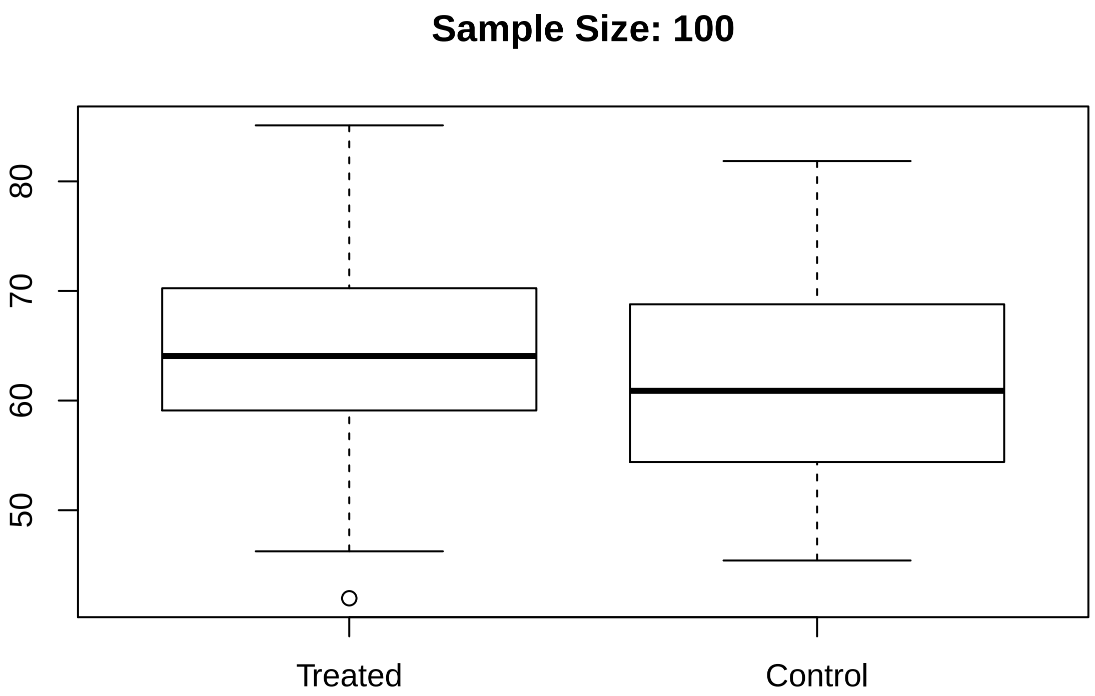
```
#*** Values are defined at the population level up to this point

# Now, we consider our sample from here
sample_ind2 <- sample(1:nrow(pop_dat), size=N_samp2)     # Sampling index
sample_dat2 <- pop_dat[sample_ind2, ]                    # Only keep obs that match the index
head(sample_dat2)
```

```
##            Y_0      Y_1 female age
## 131  54.48304 59.48304      0  36
## 1899 78.56214 83.56214      1  51
## 482  56.28647 61.28647      0  37
## 874  58.94504 63.94504      0  39
## 788  51.20042 56.20042      0  34
## 878  56.32351 61.32351      0  37
```

```
d <- ifelse(runif(N_samp2)<=0.5, 1, 0)                        # Treatment assignment indicator (1=Treat
sample_dat2$Y_obs <- d*sample_dat2$Y_1 + (1-d)*sample_dat2$Y_0  # Observed outcomes
# Y_obs = d*Y_1 + (1-d)*Y_0
sample_dat2$Status <- ifelse(d==1, "Treated", "Control")      # Copying the treatment status into sampl

head(sample_dat2)
```

```
##            Y_0      Y_1 female age    Y_obs  Status
## 131  54.48304 59.48304      0  36 59.48304 Treated
## 1899 78.56214 83.56214      1  51 83.56214 Treated
## 482  56.28647 61.28647      0  37 61.28647 Treated
## 874  58.94504 63.94504      0  39 63.94504 Treated
## 788  51.20042 56.20042      0  34 51.20042 Control
## 878  56.32351 61.32351      0  37 61.32351 Treated
```

```
### Visualzing ###

# box plot

boxplot(sample_dat2$Y_obs[sample_dat2$Status=="Treated"],
        sample_dat2$Y_obs[sample_dat2$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 100"))
```



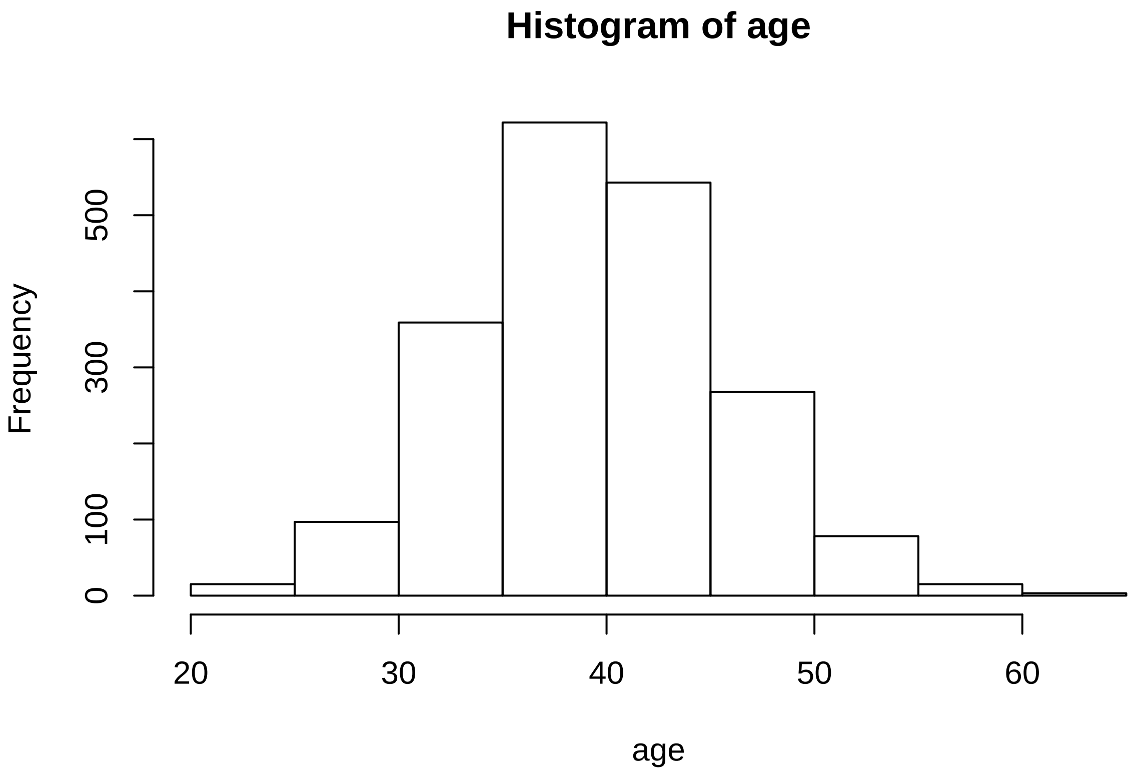
```
### Sample Size = 150 ###

# Global parameters
set.seed(212121)
N = 2000                                   # Population size
N_samp3 <- 150                               # Sample size

# Population parameters
female <- rbinom(n=N, 1, 0.4)             # Some binary covariate
age <- rpois(n=N, lambda=40)        # Some continuous covariate
mean(female)
```

```
## [1] 0.3775
```

```
mean(age)
```

```
## [1] 39.991
```

```
hist(age)
```



```
a = 0.5                                    # Intercept: Ground mean
b = 2                                      # (Fixed=constant) Effects of female
b2 = 1.50                                   # (Fixed=constant) Effects of age
tau = 5                                    # (Fixed=constant) Treatment effect
e = rnorm(n=N, mean=0, sd=0.25)               # N(0,1) error

Y_0 <- a + tau*0 + (b*female) + (b2*age) + e  # Potential outcome when not treated
Y_1 <- a + tau*1 + (b*female) + (b2*age) + e  # Potential outcome when treated

range(Y_0) #33.68 - 102.70
```

```
## [1] 30.74945 96.33202
```

```
range(Y_1) #38.68 - 107.70
```

```
## [1]  35.74945 101.33202
```

```
median(Y_0) #60.92
```

```
## [1] 60.73535
```

```
median(Y_1) #65.92
```

```
## [1] 65.73535
```

```
pop_dat <- data.frame(Y_0, Y_1, female, age) # Population level data
head(pop_dat)
```

```
##        Y_0      Y_1 female age
## 1 77.32678 82.32678      1  50
## 2 54.33245 59.33245      0  36
## 3 61.89548 66.89548      0  41
## 4 63.64301 68.64301      0  42
## 5 54.30083 59.30083      0  36
## 6 63.17072 68.17072      0  42
```

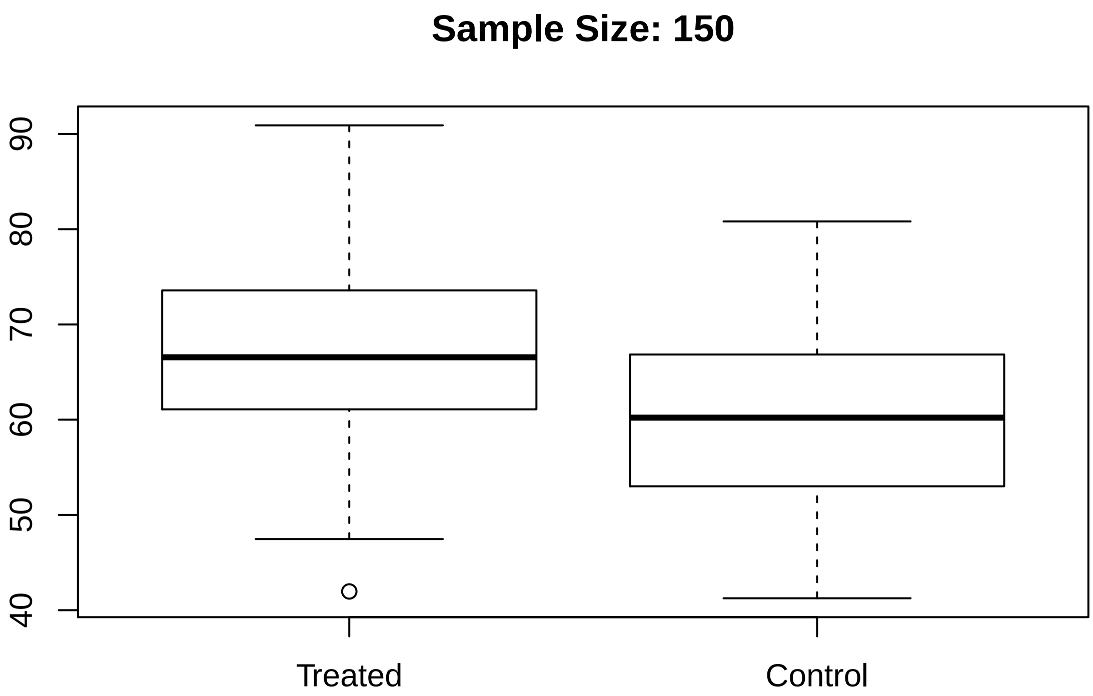
```
#*** Values are defined at the population level up to this point

# Now, we consider our sample from here
sample_ind3 <- sample(1:nrow(pop_dat), size=N_samp3)      # Sampling index
sample_dat3 <- pop_dat[sample_ind3, ]                     # Only keep obs that match the index
head(sample_dat3)
```

```
##           Y_0      Y_1 female age
## 131  54.48304 59.48304      0  36
## 1899 78.56214 83.56214      1  51
## 482  56.28647 61.28647      0  37
## 874  58.94504 63.94504      0  39
## 788  51.20042 56.20042      0  34
## 878  56.32351 61.32351      0  37
```

```
d <- ifelse(runif(N_samp3)<=0.5, 1, 0)                       # Treatment assignment indicator (1=Treat
sample_dat3$Y_obs <- d*sample_dat3$Y_1 + (1-d)*sample_dat3$Y_0  # Observed outcomes
# Y_obs = d*Y_1 + (1-d)*Y_0
sample_dat3$Status <- ifelse(d==1, "Treated", "Control")      # Copying the treatment status into sampl

head(sample_dat3)
```

```
##           Y_0      Y_1 female age    Y_obs  Status
## 131  54.48304 59.48304      0  36 59.48304 Treated
## 1899 78.56214 83.56214      1  51 83.56214 Treated
## 482  56.28647 61.28647      0  37 61.28647 Treated
## 874  58.94504 63.94504      0  39 63.94504 Treated
## 788  51.20042 56.20042      0  34 56.20042 Treated
## 878  56.32351 61.32351      0  37 56.32351 Control
```

```
### Visualzing ###

# box plot

boxplot(sample_dat3$Y_obs[sample_dat3$Status=="Treated"],
        sample_dat3$Y_obs[sample_dat3$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 150"))
```



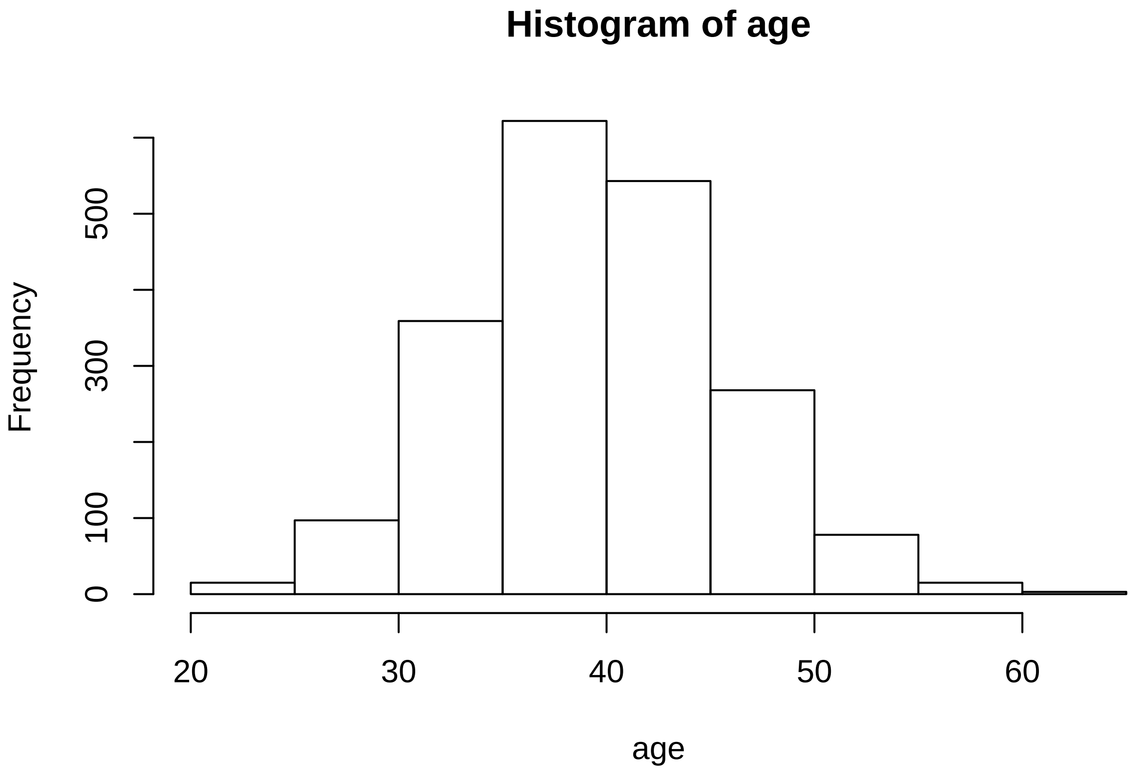
```
### Sample Size = 200 ###


# Global parameters
set.seed(212121)
N = 2000                                        # Population size
N_samp4 <- 200                                    # Sample size

# Population parameters
female <- rbinom(n=N, 1, 0.4)              # Some binary covariate
age <- rpois(n=N, lambda=40)        # Some continuous covariate
mean(female)
```

```
## [1] 0.3775
```

```
mean(age)
```

```
## [1] 39.991
```

```
hist(age)
```



```
a = 0.5                                    # Intercept: Ground mean
b = 2                                      # (Fixed=constant) Effects of female
b2 = 1.50                                   # (Fixed=constant) Effects of age
tau = 5                                    # (Fixed=constant) Treatment effect
e = rnorm(n=N, mean=0, sd=0.25)               # N(0,1) error

Y_0 <- a + tau*0 + (b*female) + (b2*age) + e  # Potential outcome when not treated
Y_1 <- a + tau*1 + (b*female) + (b2*age) + e  # Potential outcome when treated

range(Y_0) #33.68 - 102.70
```

```
## [1] 30.74945 96.33202
```

```
range(Y_1) #38.68 - 107.70
```

```
## [1]  35.74945 101.33202
```

```
median(Y_0) #60.92
```

```
## [1] 60.73535
```

```
median(Y_1) #65.92
```

```
## [1] 65.73535
```

```
pop_dat <- data.frame(Y_0, Y_1, female, age) # Population level data
head(pop_dat)
```

```
##        Y_0      Y_1 female age
## 1 77.32678 82.32678      1  50
## 2 54.33245 59.33245      0  36
## 3 61.89548 66.89548      0  41
## 4 63.64301 68.64301      0  42
## 5 54.30083 59.30083      0  36
## 6 63.17072 68.17072      0  42
```

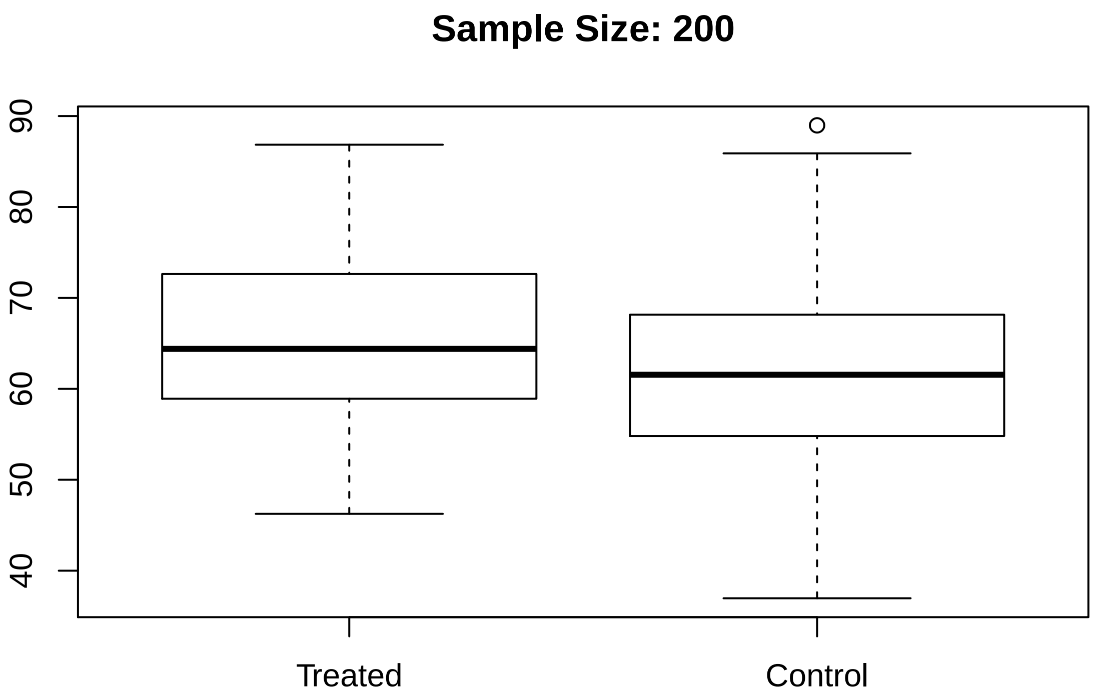
```
#*** Values are defined at the population level up to this point

# Now, we consider our sample from here
sample_ind4 <- sample(1:nrow(pop_dat), size=N_samp4)    # Sampling index
sample_dat4 <- pop_dat[sample_ind4, ]                   # Only keep obs that match the index
head(sample_dat4)
```

```
##           Y_0      Y_1 female age
## 131  54.48304 59.48304      0  36
## 1899 78.56214 83.56214      1  51
## 482  56.28647 61.28647      0  37
## 874  58.94504 63.94504      0  39
## 788  51.20042 56.20042      0  34
## 878  56.32351 61.32351      0  37
```

```
d <- ifelse(runif(N_samp4)<=0.5, 1, 0)                       # Treatment assignment indicator (1=Treat
sample_dat4$Y_obs <- d*sample_dat4$Y_1 + (1-d)*sample_dat4$Y_0  # Observed outcomes
# Y_obs = d*Y_1 + (1-d)*Y_0
sample_dat4$Status <- ifelse(d==1, "Treated", "Control")     # Copying the treatment status into sampl

head(sample_dat4)
```

```
##           Y_0      Y_1 female age    Y_obs  Status
## 131  54.48304 59.48304      0  36 54.48304 Control
## 1899 78.56214 83.56214      1  51 83.56214 Treated
## 482  56.28647 61.28647      0  37 56.28647 Control
## 874  58.94504 63.94504      0  39 63.94504 Treated
## 788  51.20042 56.20042      0  34 56.20042 Treated
## 878  56.32351 61.32351      0  37 56.32351 Control
```

```
### Visualzing ###

# box plot

boxplot(sample_dat4$Y_obs[sample_dat4$Status=="Treated"],
        sample_dat4$Y_obs[sample_dat4$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 200"))
```



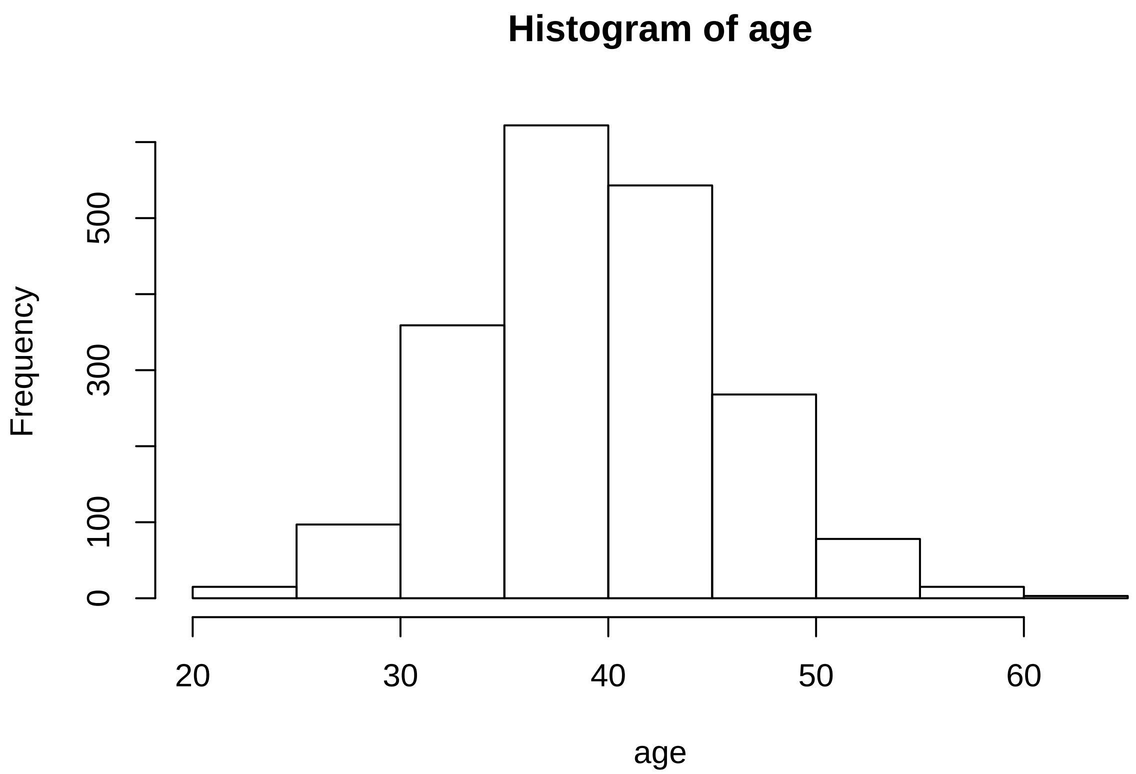
```
### Sample Size = 250 ###


# Global parameters
set.seed(212121)
N = 2000                                    # Population size
N_samp5 <- 250                                # Sample size

# Population parameters
female <- rbinom(n=N, 1, 0.4)               # Some binary covariate
age <- rpois(n=N, lambda=40)          # Some continuous covariate
mean(female)
```

```
## [1] 0.3775
```

```
mean(age)
```

```
## [1] 39.991
```

```
hist(age)
```



```
a = 0.5                                           # Intercept: Ground mean
b = 2                                             # (Fixed=constant) Effects of female
b2 = 1.50                                          # (Fixed=constant) Effects of age
tau = 5                                           # (Fixed=constant) Treatment effect
e = rnorm(n=N, mean=0, sd=0.25)                      # N(0,1) error

Y_0 <- a + tau*0 + (b*female) + (b2*age) + e  # Potential outcome when not treated
Y_1 <- a + tau*1 + (b*female) + (b2*age) + e  # Potential outcome when treated

range(Y_0) #33.68 - 102.70
```

```
## [1] 30.74945 96.33202
```

```
range(Y_1) #38.68 - 107.70
```

```
## [1]  35.74945 101.33202
```

```
median(Y_0) #60.92
```

```
## [1] 60.73535
```

```
median(Y_1) #65.92
```

```
## [1] 65.73535
```

```
pop_dat <- data.frame(Y_0, Y_1, female, age) # Population level data
head(pop_dat)
```

```
##        Y_0      Y_1 female age
## 1 77.32678 82.32678      1  50
## 2 54.33245 59.33245      0  36
## 3 61.89548 66.89548      0  41
## 4 63.64301 68.64301      0  42
## 5 54.30083 59.30083      0  36
## 6 63.17072 68.17072      0  42
```

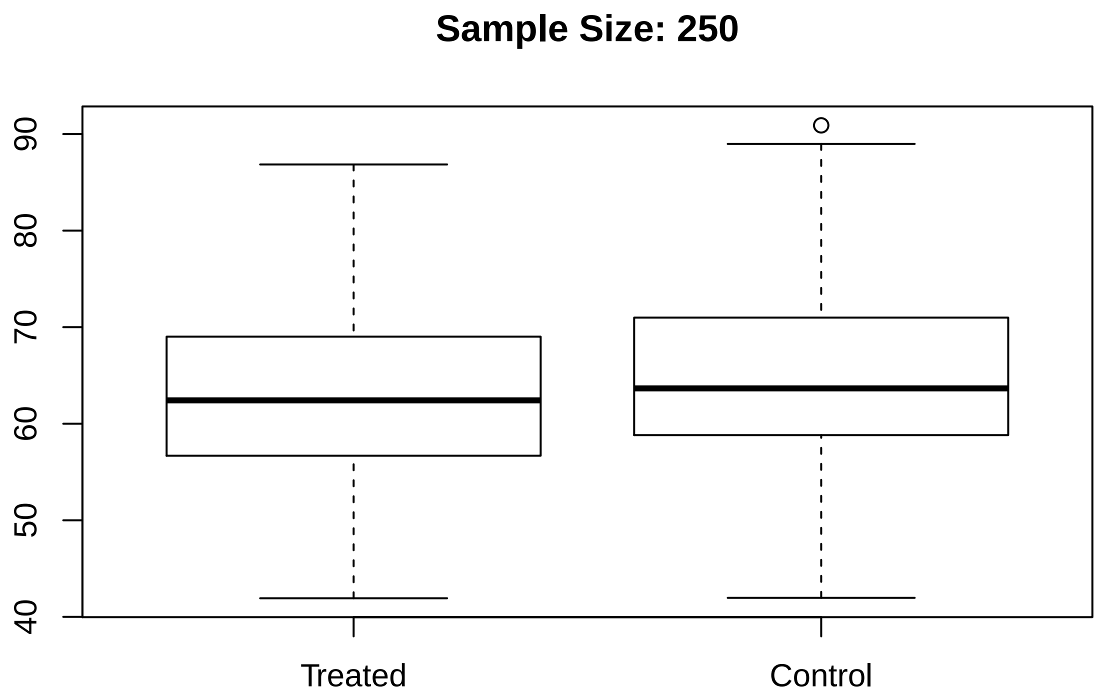
```
#*** Values are defined at the population level up to this point

# Now, we consider our sample from here
sample_ind5 <- sample(1:nrow(pop_dat), size=N_samp5)      # Sampling index
sample_dat5 <- pop_dat[sample_ind5, ]                     # Only keep obs that match the index
head(sample_dat5)
```

```
##           Y_0      Y_1 female age
## 131  54.48304 59.48304      0  36
## 1899 78.56214 83.56214      1  51
## 482  56.28647 61.28647      0  37
## 874  58.94504 63.94504      0  39
## 788  51.20042 56.20042      0  34
## 878  56.32351 61.32351      0  37
```

```
d <- ifelse(runif(N_samp5)<=0.5, 1, 0)                        # Treatment assignment indicator (1=Treat
sample_dat5$Y_obs <- d*sample_dat5$Y_1 + (1-d)*sample_dat5$Y_0  # Observed outcomes
# Y_obs = d*Y_1 + (1-d)*Y_0
sample_dat5$Status <- ifelse(d==1, "Treated", "Control")      # Copying the treatment status into sampl

head(sample_dat5)
```

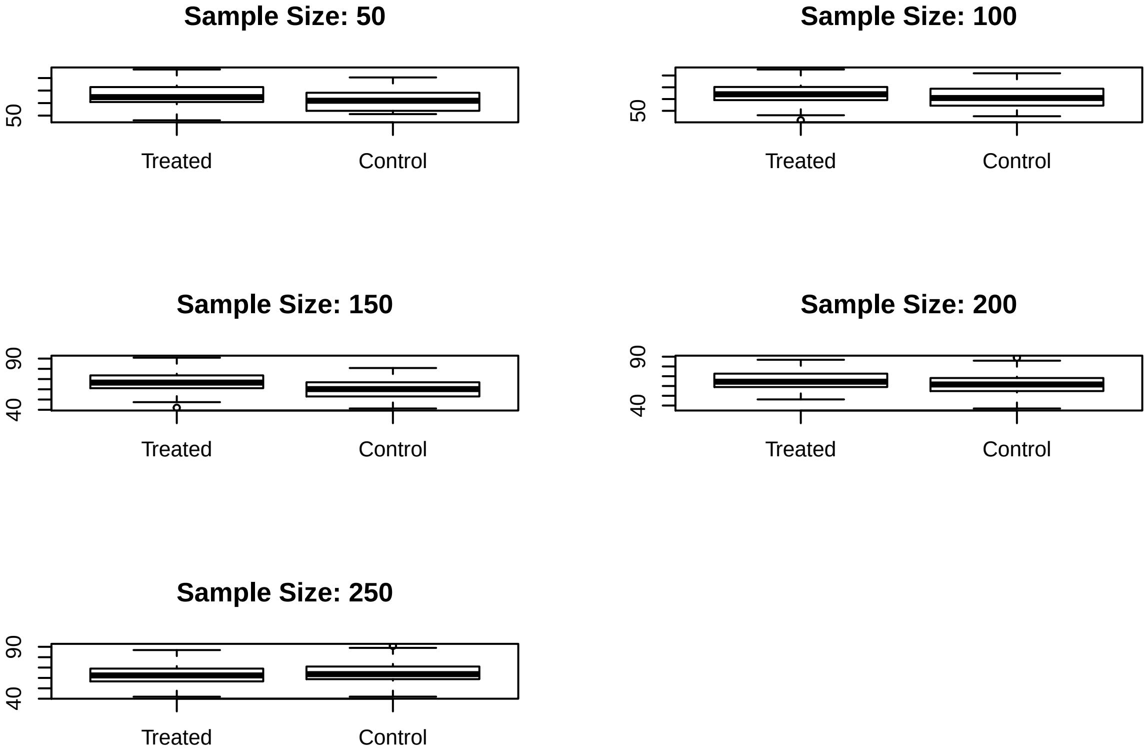
```
##           Y_0      Y_1 female age    Y_obs  Status
## 131  54.48304 59.48304      0  36 59.48304 Treated
## 1899 78.56214 83.56214      1  51 83.56214 Treated
## 482  56.28647 61.28647      0  37 61.28647 Treated
## 874  58.94504 63.94504      0  39 58.94504 Control
## 788  51.20042 56.20042      0  34 56.20042 Treated
## 878  56.32351 61.32351      0  37 61.32351 Treated
```

```
### Visualzing ###

# box plot

boxplot(sample_dat5$Y_obs[sample_dat4$Status=="Treated"],
        sample_dat5$Y_obs[sample_dat4$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 250"))
```



Combining Plots

```
##### Combining all 5 plots to show difference in sample sizes #####

par(mfrow=c(3,2))

# SS = 50 #
boxplot(sample_dat$Y_obs[sample_dat$Status=="Treated"],
        sample_dat$Y_obs[sample_dat$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 50"))

# SS = 100 #
boxplot(sample_dat2$Y_obs[sample_dat2$Status=="Treated"],
        sample_dat2$Y_obs[sample_dat2$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 100"))

# SS = 150 #
boxplot(sample_dat3$Y_obs[sample_dat3$Status=="Treated"],
        sample_dat3$Y_obs[sample_dat3$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 150"))

# SS = 200 #
boxplot(sample_dat4$Y_obs[sample_dat4$Status=="Treated"],
        sample_dat4$Y_obs[sample_dat4$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 200"))

# ss = 250 #
boxplot(sample_dat5$Y_obs[sample_dat4$Status=="Treated"],
        sample_dat5$Y_obs[sample_dat4$Status=="Control"],
        names=c("Treated", "Control"), main=("Sample Size: 250"))
```


Sample Size: 50
50
Treated
Control
Sample Size: 100
50
Treated
Control
Sample Size: 150
90
40
Treated
Control
Sample Size: 200
90
40
Treated
Control
Sample Size: 250
90
40
Treated
Control