Problem Set 2: Causal Inference

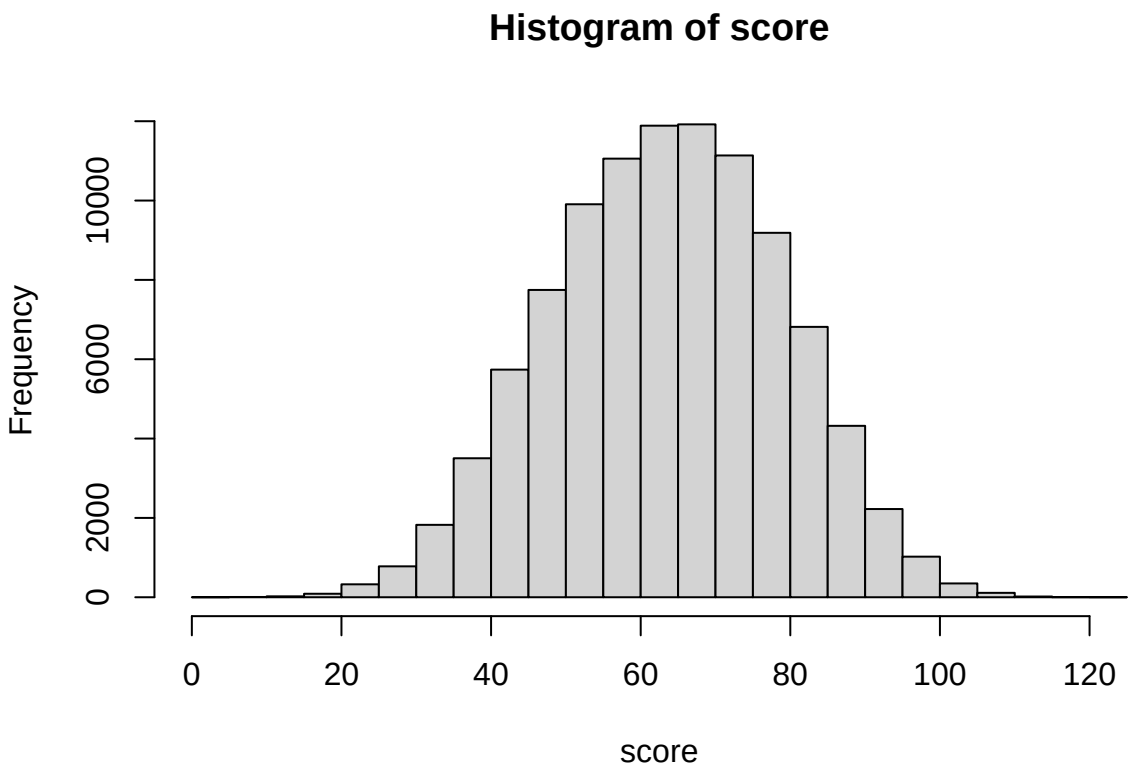
1 Setup 1

Consider the following DGP.

```
set.seed(12577) # for replication
n <- 100000 # observations
gender <- rbinom(n, 1, 0.5) # female 1
ability <- rnorm(n, 0, 1)
# income follows the log normal distributing
income <- rlnorm(n, meanlog = 1, sdlog = 0.5)
# multiplying the log normal dist with 20000
income <- income * 20000
lowincome <- ifelse(income < median(income), 1, 0)
highwork <- rbinom(n, 1, 0.3 + 0.2 * lowincome + 0.1 * gender)
# score on the first exam
score <- 65 - (highwork * 20) - (lowincome * 5) + gender * 10 + ability * 10 + rnorm(n, 5, 2.5)
```

Histogram of score.

```
hist(score)
```



1. Run a quick regression where you regress score on lowincome, gender, highwork, and ability. Are the estimates on these variables close enough to the population parameter used for construction?

2. Next, generate a random treatment assignment (each person has the same probability of being treated). You can do this using rbinom(n, 1, 0.5) in R. This splits the treatment and control groups randomly (e.g., independence assumption). Discuss how this assumption can help in causal inference in this context.

3. On average, treatment program boosts score by 15. You'd want to generate the new score (score on the second exam) using score2 <- score + (15 * treat) + rnorm(n, 5, 2.5).

4. Assess balance in terms of the covariates (lowincome, gender, highwork, ability, and score on the first exam) across treatment and control groups.

5. Estimate the average treatment effect (use difference across means).

6. Replicate estimation of ATE as given in question 5 using a bootstrap process consisiting of 99 replications.

Here's a short algorithm to do so.

a. Treat the initial data as the population (say, this is data1).
b. Sample with replacement from data1 with sample size same as that of data1. You can do this using sample(x, n, replace = T).
c. Using the sample that you get in b, estimate the ATE. Store your estimate.
d. Repeat steps b and c 99 times. Each time you want to store the ATE estimate from the newer sample.
e. You'll have 99 estimates of ATE now.

This process is called bootstrapping, which has several advantages. One is estimating the standard error.

Create a histogram of your bootstrapped estimates. Label the 5th and 95th percentiles of your bootstrapped estimates.

Next, mark the initial ATE estimate that you estimated in question 5 using a vertical dashed line. Comment on the statistical significance.

7. Estimate ATE using regression. Report the standard error and discuss statistical significance.
8. Swap the treatment and control group (role reversal). Re-calibrate the new score and re-estimate ATE. Comment on the difference between ATE and ATT (average treatment effect on the treated).

2 Setup 2

Now, say there were complaints that the treatment administration is not ethical. This can be driven by the case that you cannot deny students from the control group who are willing to come to the program. Due to this, you had to change the assignment of treatment to voluntary.

Treatment: Those who showed up.

Control: Those who did not.

This obviously creates selection issues, but we are going to assume that we know the selection mechanism. The objective here is to see that if we know what variables are determining the treatment, we can account (or condition) on them to get an unbiased estimate of treatment effect. In other words, we are invoking the conditional independence assumption.

Here's how I'd you to proceed.

1. Create a new treatment such that treatment is determined by highwork and gender.

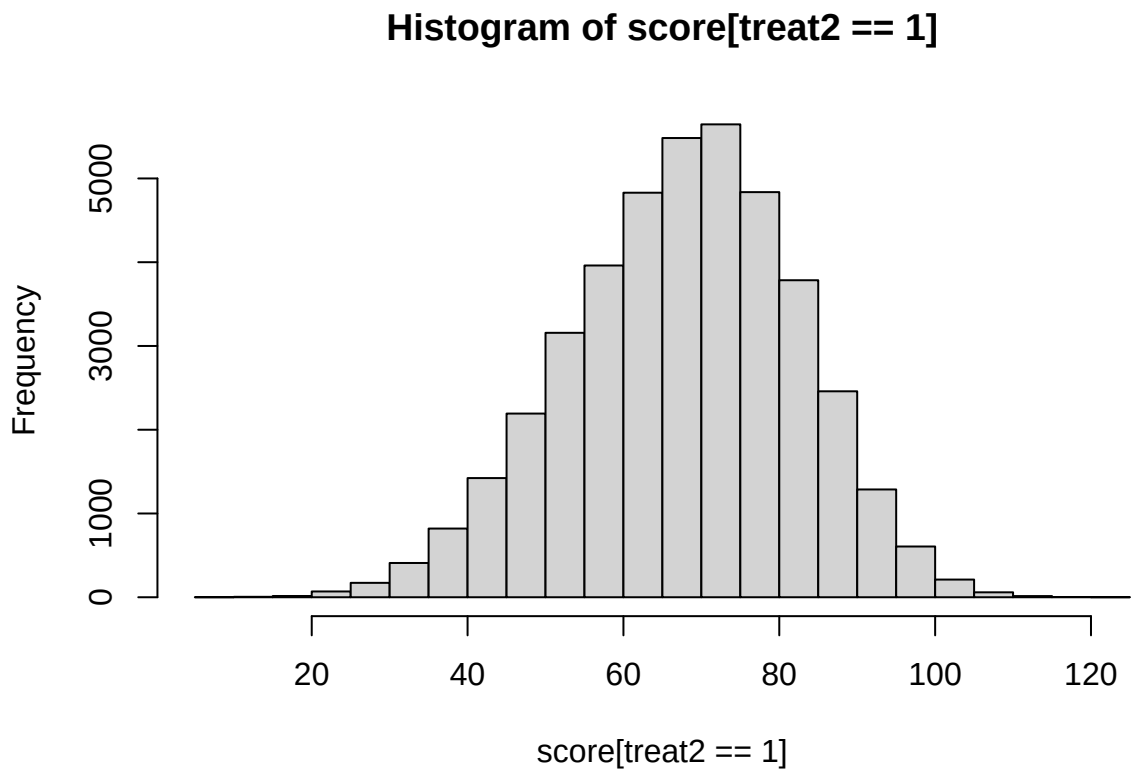
```
treat2 <- rbinom(n, 1, 0.5 - 0.3 * highwork + 0.1 * gender)
table(treat2)
```

```
## treat2
##     0     1
## 58554 41446
```

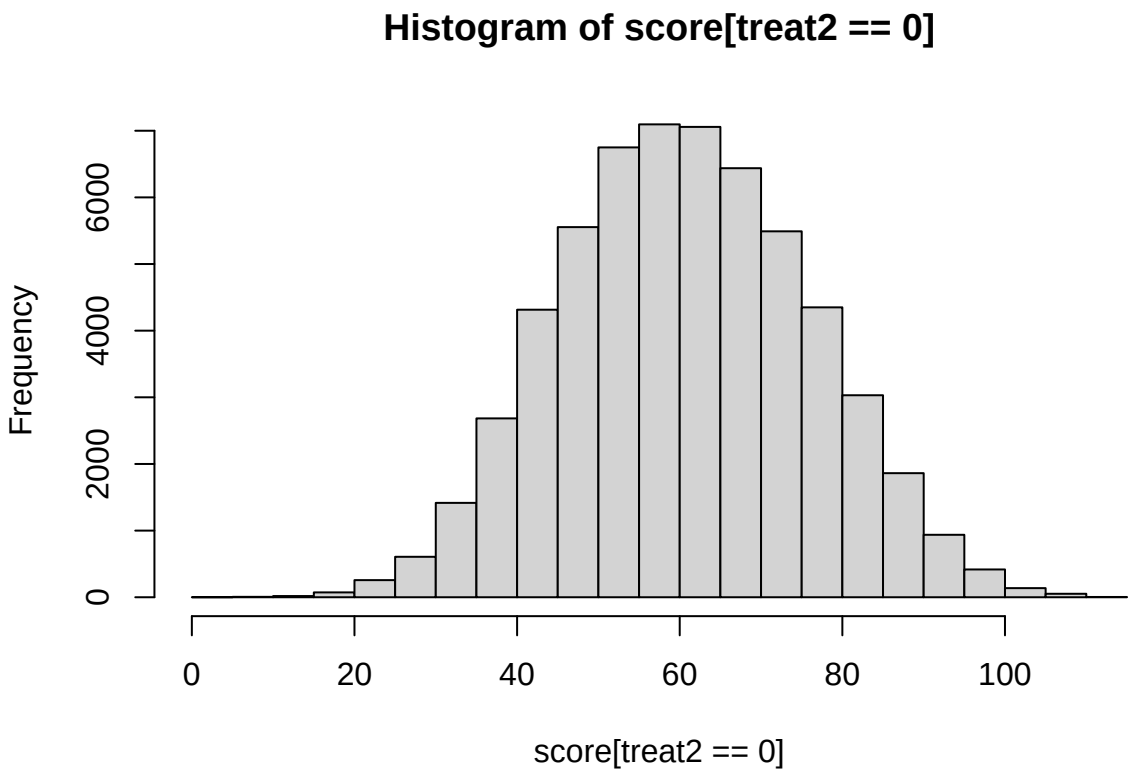
2. Re-generate the score variable in this case. Note that treatment still boosts score by 15 points on average.

```
library(patchwork)
score_later2 <- score + (15 * treat2) + rnorm(n, 5, 2.5)

f0  <- hist(score[treat2 == 1])
```




```
f1  <- hist(score[treat2 == 0])
```




```
mean(score[treat2 == 1])
```

```
## [1] 67.58388
```

```
mean(score[treat2 == 0])
```

```
## [1] 60.64669
```

3. Run the balance exercise like the one that you did in part 1. Comment on whether baseline covariates are balanced.

4. Estimate the ATE using difference in means across the treatment and control group. Comment on whether the estimate is unbiased.

5. We'd want to account for the selection mechanism that arguably is inducing bias to our estimation. We'd want to estimate ATEs in each strata and construct the weighted average of strata-specific ATEs to get an estimate of overall ATE.

a. Identify the stratas of concern.
b. Estimate strata-specific ATE.
c. Use estimates in b to construct a weighted average of strata-specific ATEs. Weights are given by the fraction of units in a given strata.
d. Compare and contrast the ATE from this approach versus the one that you estimated in 4.