Problem Set 5

Luke Haner

2025-12-01

#**Part 1 Simulation**

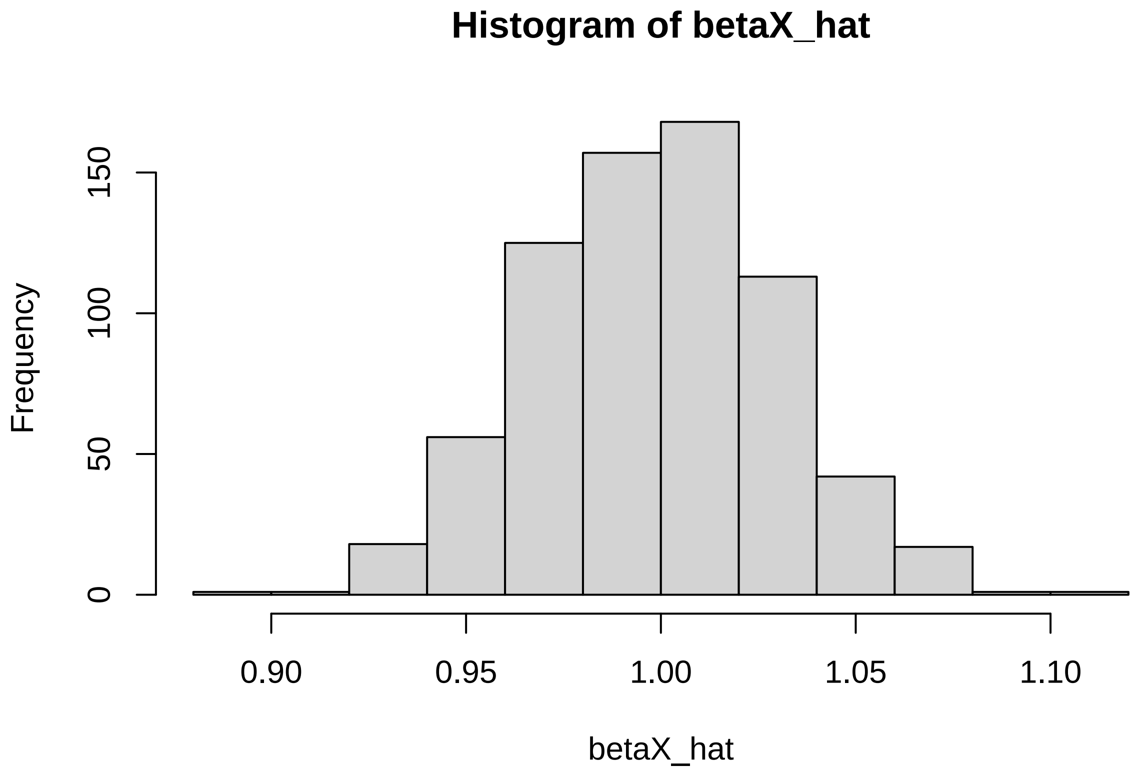
```r
#Create a simulated data set with a dependent variable that is a linear function of a treatment variabl

#setting for reproducibility
set.seed(123)
n <- 1000 # sample size

#Confounding Variable
C <- rnorm(n)

#Treatment
X <- 0.5*C + rnorm(n)

#Dependent Variable
Y <- 1 * X + 0.7*C + rnorm(n)

#Create data frame
sim_data <- data.frame(C, X, Y)


#Fit a linear model for the true data generating process and print the summary table
true_model <- lm(Y ~ X + C, data = sim_data)

summary(true_model)
```

```
##
## Call:
## lm(formula = Y ~ X + C, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8360 -0.6277 -0.0370  0.6538  3.3787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02093    0.03098  -0.676    0.499
## X            1.02751    0.03079  33.377   <2e-16 ***
## C            0.66476    0.03609  18.417   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 0.9788 on 997 degrees of freedom
## Multiple R-squared:  0.735,  Adjusted R-squared:  0.7345
## F-statistic:  1383 on 2 and 997 DF,  p-value: < 2.2e-16
```

#1a. Using the true model, demonstrate that the coefficient for your treatment variable follows the central limit theorem. That is, demonstrate that the coefficients sampling distribution is approximately normal.

```r
#central limit theorem: as the sample size increase, the standardized sample mean of X can be approxima

set.seed(123)
simulations <- 700 # number of simulations

#place to store estimated X coefficients
betaX_hat <- numeric(simulations)

#putting data in for loop to run simulations
for (i in 1:simulations) {
  C <- rnorm(n)
  X <- 0.5 * C + rnorm(n)
  Y <- 1 * X + 0.7*C + rnorm(n)

  betaX_hat[i] <- coef(lm(Y ~ X + C))["X"] #fitting true model and storing treatment coefficient X
}

#Plotting distribution

hist(betaX_hat)
```



#1.b Compute the bootstrapped standard error for the coefficient of the treatment variable

```r
#Bootstrapped standard errors
#1.Collect data and define statistic
#2.Resample your data with replacement
#3.Calculate the statistic for each sample
#4.calculate the standard deviation of your collected statistics


set.seed(123)
B <- 1000 #number of bootstrap samples
boot_betaX <- numeric(B) #making place to store X coefficient from each bootstrap sample

for (i in 1:B) {
  boot_sample <- sim_data[sample(1:nrow(sim_data), size = nrow(sim_data), replace = TRUE),] #resampling

  boot_betaX[i] <- coef(lm(Y ~ X + C, data = boot_sample))["X"] #fitting model and taking out X coeffic
}

boot_se <- sd(boot_betaX) #taking standard deviation of collected statistics

boot_se #Approximately matches standard error from true model, 0.0278351 compared to 0.02971
```

```
## [1] 0.03042093
```

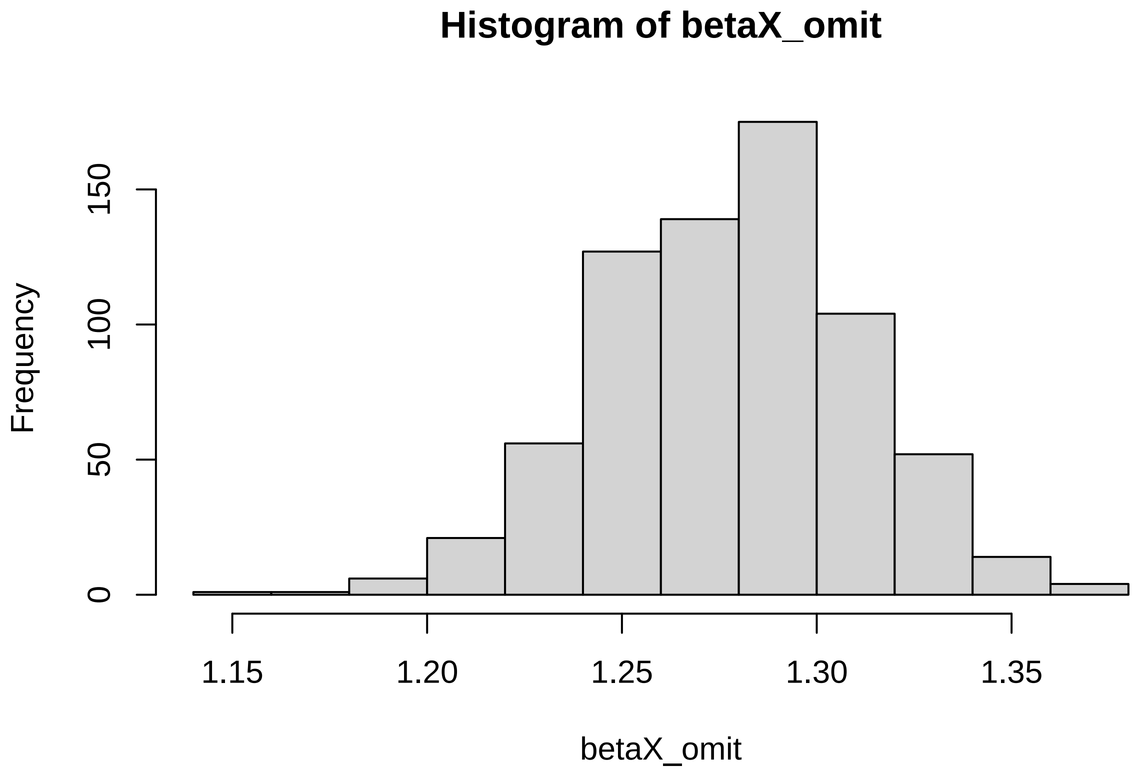
#1c. Fit a model that omits the confounding variable. Repeat part a for this new model

and plot the sampling distribution of the treatment variables coefficient. How do your results differ? What does this imply about statistical tests based on a coefficients sampling distribution?

```r
#fit model that omits the confounding variable
omit_model <- lm(Y ~ X, data = sim_data)

summary(omit_model)
```

```
##
## Call:
## lm(formula = Y ~ X, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2816 -0.7243  0.0056  0.8086  3.4599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02458    0.03585  -0.686    0.493
## X            1.31188    0.03082  42.570   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.133 on 998 degrees of freedom
## Multiple R-squared:  0.6449, Adjusted R-squared:  0.6445
## F-statistic:  1812 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
set.seed(123)
simulations <- 700 # number of simulations

#place to store estimated X coefficients
betaX_omit <- numeric(simulations)

#putting data in for loop to run simulations
for (i in 1:simulations) {
  C <- rnorm(n)
  X <- 0.5 * C + rnorm(n)
  Y <- 1 * X + 0.7*C + rnorm(n)

  betaX_omit[i] <- coef(lm(Y ~ X))["X"] #fitting true model and storing treatment coefficient X
}

#Plotting distribution

hist(betaX_omit)
```



#**Interpretation** In the model omitting the confounding variable the sampling distribution of the treatment coefficient is biased upwards. The coefficient is 1.2554 rather than around the true value of one from them model that controls for confoundrs,. This is also visually noticeable from the histograms. This implies that statistical tests based on a coefficients sampling distribution can be heavily influenced by bias, even returning statistical significant results even when the estimated coefficient is not accurate due to homogeneity created by omitting a relevant variable. As such, it is important to think causally about which variables may be at play and to control for them in order to get the most precise model possible.

#**Part 2 Data Analysis**

For this part of the assignment, use any data set you like. It can be from the course materials, simulated, or drawn from another source

#**2a.Conduct a hypothesis test for a difference in means. You decide what the hypotheses are, whether you use a t-test or a z-test, and what the level of significance is. Explain your decisions, and interpret your results both substantively and statistically**

```
star <- read.csv("C:/Users/luke.haner/Downloads/STAR.csv")

#Hypothesis: Students in smaller classes have a higher graduation rate than students in regular classes
#Null Hypothesis: The is no difference between graduation rates between students in small and regular c

#make classtype numeric for analysis
star$classtype <- as.factor(star$classtype)

#Conduct hypothesis Test
t.test(graduated ~ classtype,
       data = star,
       alternative = "greater",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  graduated by classtype
## t = -0.37193, df = 1247.2, p-value = 0.645
## alternative hypothesis: true difference in means between group regular and group small is greater tha
## 95 percent confidence interval:
##  -0.03814884         Inf
## sample estimates:
## mean in group regular   mean in group small
##             0.8664731             0.8735043
```

#**2a. Interpretation/Discussion

I decided to revisit the star datatset covered in class, and looked specifically at graduation rates between students who were in small classes and those in regular classes. I utilized a t-test as it compares the means in two groups which would be the graduation rate from regular and small class types. Additionally, I choose to utilize a t-test as it is more robust to non-normality. Substantively, my results show that small class type has a very small increased graduation rate (0.8735 compared to 0.8665) than regular class type. Statistically however, this result is not significant meaning it is not distinguishable form 0, and I cannot reject the null hypothesis that class type has no effect on graduation rate. Overall, the results show both statistically and substantively that small class size has little to no effect on an increased graduation rate compared to regular class types.

#2b.Using the same data, fit a linear model. Interpret the coefficient, standard error, t-value, and p-value.

```
# Linear model predicting graduation rate based on class type class type
model <- lm(graduated ~ classtype, data = star)

#regression results
summary(model)
```

```
##
## Call:
## lm(formula = graduated ~ classtype, data = star)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8735  0.1265  0.1265  0.1335  0.1335
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.866473   0.012834  67.514   <2e-16 ***
## classtypesmall 0.007031   0.018940   0.371    0.711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3369 on 1272 degrees of freedom
## Multiple R-squared:  0.0001083,  Adjusted R-squared:  -0.0006777
## F-statistic: 0.1378 on 1 and 1272 DF,  p-value: 0.7105
```

#2b. Interpretation

The intercept coefficient (0.8665) or %86.65, represents the predicted graduation rate for students in regular class types. It's standard error of 0.012834 is small which means the estimate itself is highly accurate. I observe a t-value of 67.514, which is a measure of how far the estimate is in terms of standard error from zero, which in turns gives a very very small p-value, indicating high statistical significance.

The class type small coefficient (0.007031), or %0.7, represents that being in a small class increases predicted graduation rates by about 0.7 percentage points. It's standard error of 0.0189 raises that there may be some uncertainty around the estimate because it is large compared to the magnitude of the estimate itself.. The t-value of 0.371 is relatively small, since the estimate is close to zero, it produces a large p-value of 0.715 which indicates statistical insignificance. As such, the coefficient is not statistically difference than 0, and I cannot reject the null hypothesis that class type has no effect on graduation rates.

Overall, the results show that being in a small class type leads to a very small and statistically insignificant increase in graduation rates compared to regular class types. Both substantively and statistically, it is shown that class type does not have meaningful influence on graduation rates/outcomes.