prob_set 5 Baixi

2025-11-26

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(broom)
```

Part 1

1.

```
#set seed for reproductiability
set.seed(1212)

#set numbers of observations and iterations
n <- 1000
n_iter <- 1000


C <- rnorm(n) #confounder
X <- rnorm(n)+ (0.5)*C + rnorm(n)
Y <- (-0.6)*X + 0.2*C + rnorm(n)

#fit the true model
model_t <- lm(Y ~ X + C)

summary(model_t)
```

```
##
## Call:
## lm(formula = Y ~ X + C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0727 -0.6891  0.0360  0.6728  3.2600
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06355    0.03239  -1.962     0.05 *
## X           -0.59254    0.02330 -25.431  < 2e-16 ***
## C            0.21759    0.03472   6.267 5.47e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 997 degrees of freedom
## Multiple R-squared:  0.3958, Adjusted R-squared:  0.3946
## F-statistic: 326.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

```r
#simulate the coefficient and save in a vector
model_coef <- replicate(n_iter,{

  C <- rnorm(n)
  X <- rnorm(n)+ (0.5)*C + rnorm(n)
  Y <- (-0.6)*X + 0.2*C + rnorm(n)

  model_t <- lm(Y ~ X + C)

  return(coef(model_t)['X'])
})

#get the mean and standard error
mean(model_coef)
```

```
## [1] -0.5991369
```

```r
sd(model_coef)
```

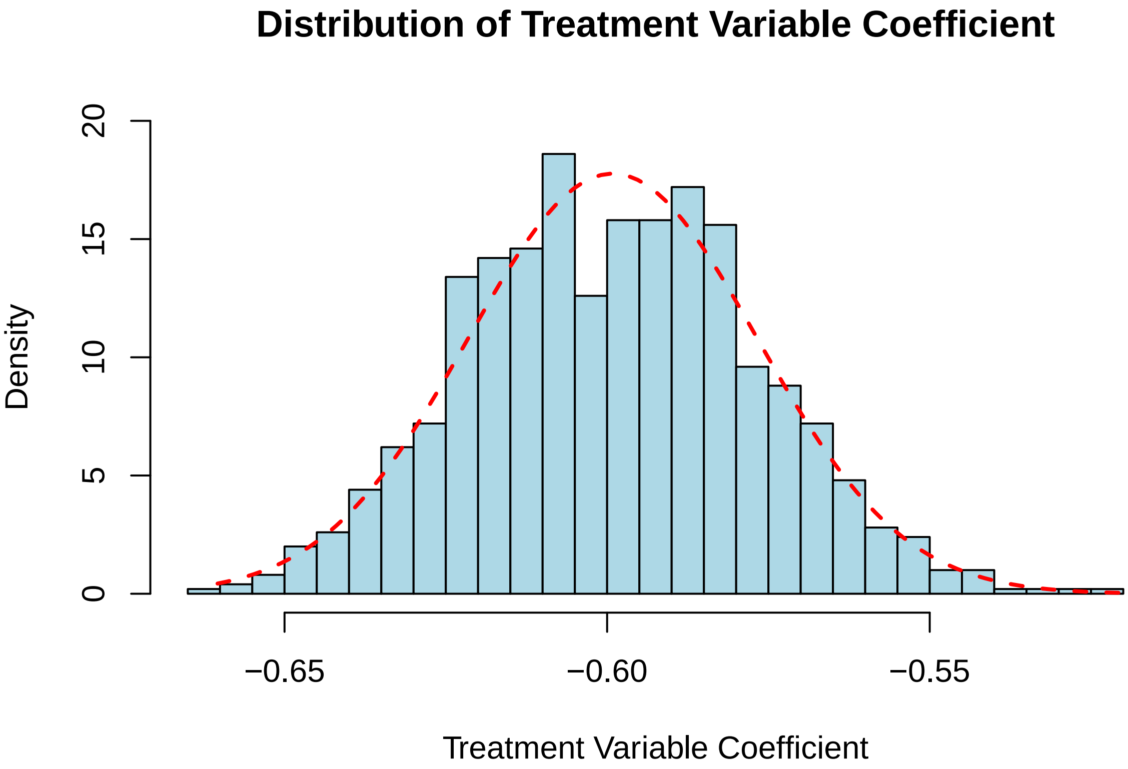
```
## [1] 0.02245439
```

```r
#draw a histogram
hist_c <- hist(model_coef,
     col = "lightblue",
     main = "Distribution of Treatment Variable Coefficient",
     xlab = "Treatment Variable Coefficient",
     freq = FALSE,
     ylim = c(0, 20),
     breaks = 30)

#add a line of normal distribution
x_grid <- seq(min(model_coef), max(model_coef), length = 100)


y_vals <- dnorm(x_grid, mean = mean(model_coef), sd = sd(model_coef))

lines(x_grid, y_vals, col = "red", lwd = 2, lty = 2)
```



The simulation results shows that, under the sample of 1000 observations, the mean of the simulated coefficient is -0.599, which converge to the true parameter of -0.6. In the histogram, the figure of sampling distribution approximately follows the normal density curve. Thus we can conclude the coefficient for treatment variable X follows the central limit theorem

2.

From the last section we already got the sample standard error of 0.022

```
#set seed for reproductiability
set.seed(1212)

#get the one-time data for bootstrap
n <- 1000
C1 <- rnorm(n)
X1 <- rnorm(n) + 0.5 * C1 + rnorm(n)
Y1 <- -0.6 * X1 + 0.2 * C1 + rnorm(n)

#save in a dataframe
data1 <- data.frame(Y1, X1, C1)

#start bootstrap
n_iter = 1000

boot_coef <- replicate( n_iter, {
  boot_index <- sample(x = 1:n, size = n, replace = TRUE)

  data_boot <- data1[boot_index, ]

  model_boot <- lm(Y1 ~ X1 + C1, data = data_boot)
```

```r
  return(coef(model_boot)['X1'])
})

se_boot <- sd(boot_coef)
print(se_boot)
```

```
## [1] 0.02362098
```

3.

```r
#set seed for reproductiability
set.seed(1212)

#fit the confounder-omitted model
model_o <- lm(Y ~ X)

summary(model_o)
```

```
## 
## Call:
## lm(formula = Y ~ X)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3.07582 -0.70930  0.03389  0.69430  2.98172 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.06727    0.03300  -2.039   0.0417 *  
## X           -0.54287    0.02233 -24.315   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.043 on 998 degrees of freedom
## Multiple R-squared:  0.372,  Adjusted R-squared:  0.3714 
## F-statistic: 591.2 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
#simulate the coefficient and save in a vector
model_o_coef <- replicate(n_iter,{

  C <- rnorm(n)
  X <- rnorm(n)+ (0.5)*C + rnorm(n)
  Y <- (-0.6)*X + 0.2*C + rnorm(n)

  model_o <- lm(Y ~ X )

  return(coef(model_o)['X'])
})

#get the mean and standard error
mean(model_o_coef)
```

```
## [1] -0.554546
```

```
sd(model_o_coef)
```

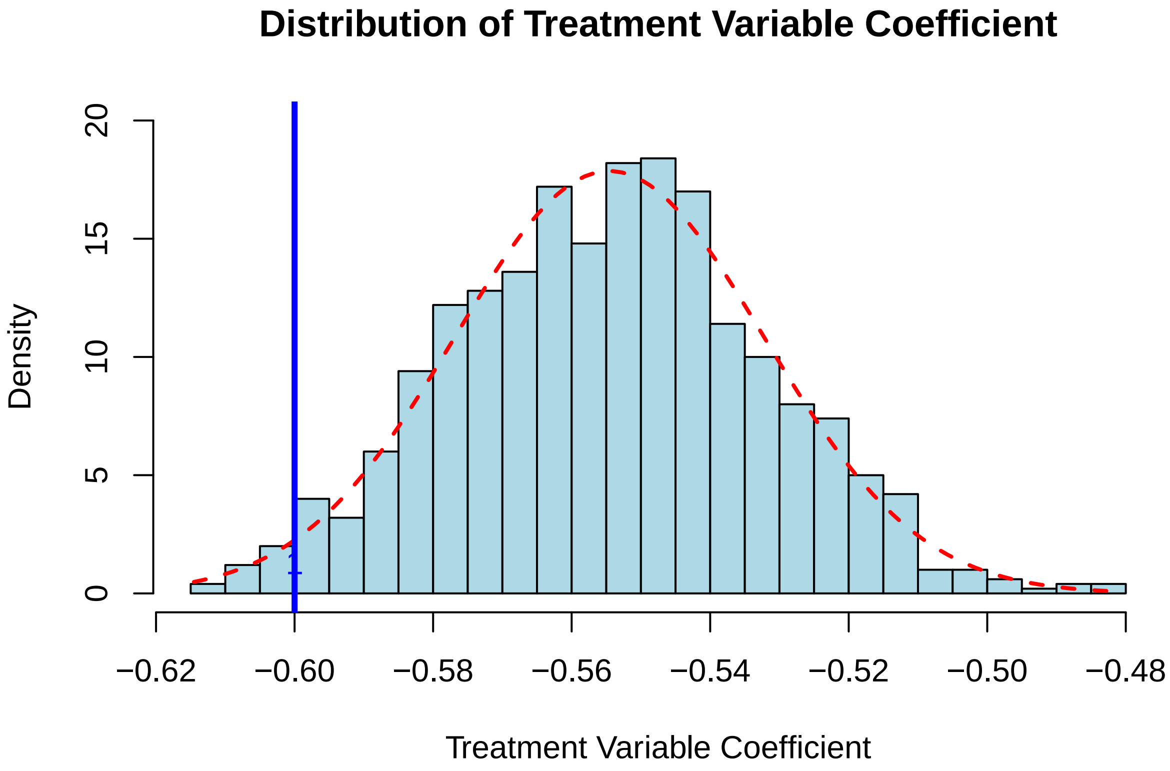
```
## [1] 0.02233352
```

```
#draw a histogram
hist_c1 <- hist(model_o_coef,
     col = "lightblue",
     main = "Distribution of Treatment Variable Coefficient",
     xlab = "Treatment Variable Coefficient",
     freq = FALSE,
     ylim = c(0, 20),
     breaks = 30)

#add a line of normal distribution
x_grid1 <- seq(min(model_o_coef), max(model_o_coef), length = 100)

y_vals1 <- dnorm(x_grid1, mean = mean(model_o_coef), sd = sd(model_o_coef))

lines(x_grid1, y_vals1, col = "red", lwd = 2, lty = 2)
abline(v = -0.6, col = "blue", lwd = 3)
text(-0.6, 0, col = "blue", pos = 3)
```



The mean of the sampling distribution is about -0.55, which is deviated from the true value -0.6. It implies that the significance of the statistical test based on sampling distribution does not ensures your causal identification of the variables. Omitted variable will lead to biased estimate, but the central limited theorem still works to generate a normal sampling distribution, although the center is deviated. We may end up having high statistical confidence in a biased estimate.

Part 2

1.

I plan to use the thermometers dataset to test whether there's a difference in attitude towards Asian between male and female.

- H_0: $\mu_{female} - \mu_{male} = 0$ (There is no difference in the mean of attitude on Asian between male and female)
- H_a: $\mu_{female} - \mu_{male} \neq 0$ (There is significant difference.)

I plan to use the t-test because we are using a dataset of observational data, so we don't know the population parameter of standard deviation. I used Welch's method (instead of Student's) because it does not assume that the variance of the male and female groups is identical. This makes the test more robust to heteroscedasticity.

```
thermo <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/thermomet

#t test
t.test(thermo$ft_asian ~ thermo$sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  thermo$ft_asian by thermo$sex
## t = -0.21121, df = 4841.1, p-value = 0.8327
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -1.342916  1.081701
## sample estimates:
## mean in group Female   mean in group Male
##             74.12446             74.25506
```

The result shows that we cannot reject the null hypothesis. The mean of attitude thermometer of female is 74.12446, while the figure of male is 74.25506. Although there's a difference of about 0.13 in sample estimates, but the p-value is 0.8327, which is well above the significance threshold of 0.05. Besides, the confidence interval [-1.34,1.08] also contains zero. Thus, we fail to reject the hypothesis that there is no difference in the mean of attitude on Asian between male and female. The difference 0.13 of sample estimates is trivial and it is just the variation in random sampling.

```
#fit a linear model
model1 <- lm(ft_asian ~ sex, data = thermo)
summary(model1)
```

```
##
## Call:
## lm(formula = ft_asian ~ sex, data = thermo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.255 -20.124   4.745  17.745  25.876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.1245     0.4293 172.666   <2e-16 ***
## sexMale       0.1306     0.6202   0.211    0.833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 21.56 on 4842 degrees of freedom
##   (145 observations deleted due to missingness)
## Multiple R-squared:  9.159e-06,  Adjusted R-squared:  -0.0001974
## F-statistic: 0.04435 on 1 and 4842 DF,  p-value: 0.8332
```

The coefficient equals to the difference-in-mean in t-test, it means that being a male will predict 0.13 points higher of feeling on Asian compared to being a female, however this effect is insignificant. The standard error is 0.6202 which is well above the coefficient 0.13. It shows that the estimate is heavily influenced by random noise. The t-value is 0.211, and the p-value is 0.8332, both are far from the threshold of being significant. Overall, the linear model outcome also shows that gender explains virtually nothing about attitudes toward Asians.