Problem_Set_5

Santiago Vidal Calvo

2025-12-07

```r
# Part 1: Simulation

# Create a simulated data set with a dependent variable that is a linear

# > function of a treatment variable and a confounding variable. Fit a

# > linear model for the true data generating process and print the

# > summary table.

set.seed(123)

n <- 5000

# Confounder C

C <- rnorm(n, mean = 0, sd = 1)

# Treatment T depends on C

Treated <- 0.8 * C + rnorm(n, 0, 1)

# Outcome Y depends on T and C

# True model: Y = 1 + 2*T + 1*C + error

Y <- 1 + 2 * Treated + 1 * C + rnorm(n, 0, 1)

sim_data <- data.frame(
Y = Y,
T = Treated,
C = C
)

head(sim_data)
```

```
##           Y          T           C
## 1 0.9251408 -0.9425544 -0.56047565
## 2 2.4899135  0.9434515 -0.23017749
## 3 3.6857039  0.1000171  1.55870831
## 4 3.5772073  1.5374253  0.07050839
## 5 3.3936206  1.0196214  0.12928774
## 6 7.2614169  1.7071830  1.71506499
```

```r
# This data.frame contains the outcome Y, the treatment T, and the

# > confounder C, all generated from the specified linear DGP.
```

```r
# Fit the true model Y ~ T + C and print the summary table

mod_true <- lm(Y ~ T + C, data = sim_data)
summary(mod_true)
```

```
##
## Call:
## lm(formula = Y ~ T + C, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3523 -0.6578  0.0034  0.6994  3.0994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00819    0.01416   71.21   <2e-16 ***
## T            2.00289    0.01412  141.85   <2e-16 ***
## C            1.00252    0.01812   55.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 4997 degrees of freedom
## Multiple R-squared:  0.9143, Adjusted R-squared:  0.9142
## F-statistic: 2.664e+04 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```r
# In this model, the true coefficient on T is 2 by construction and the

# > true coefficient on C is 1. The estimated coefficients should be

# > close to these values, and the summary output gives us the standard

# > errors and t statistics for each estimate.
```

```r
# Part 1(a)

# Using the true model, demonstrate that the coefficient for your

# > treatment variable follows the central limit theorem. That is,

# > demonstrate that the coefficient's sampling distribution is

# > approximately normal.

set.seed(456)

B <- 1000          # number of simulated datasets
beta_T <- numeric(B)

for (b in seq_len(B)) {
C_b <- rnorm(n, 0, 1)
T_b <- 0.8 * C_b + rnorm(n, 0, 1)
```

```r
Y_b <- 1 + 2 * T_b + 1 * C_b + rnorm(n, 0, 1)
dat_b <- data.frame(Y = Y_b, T = T_b, C = C_b)
fit_b <- lm(Y ~ T + C, data = dat_b)
beta_T[b] <- coef(fit_b)["T"]
}

# Look at the mean and standard deviation of the sampled coefficients

mean(beta_T)
```

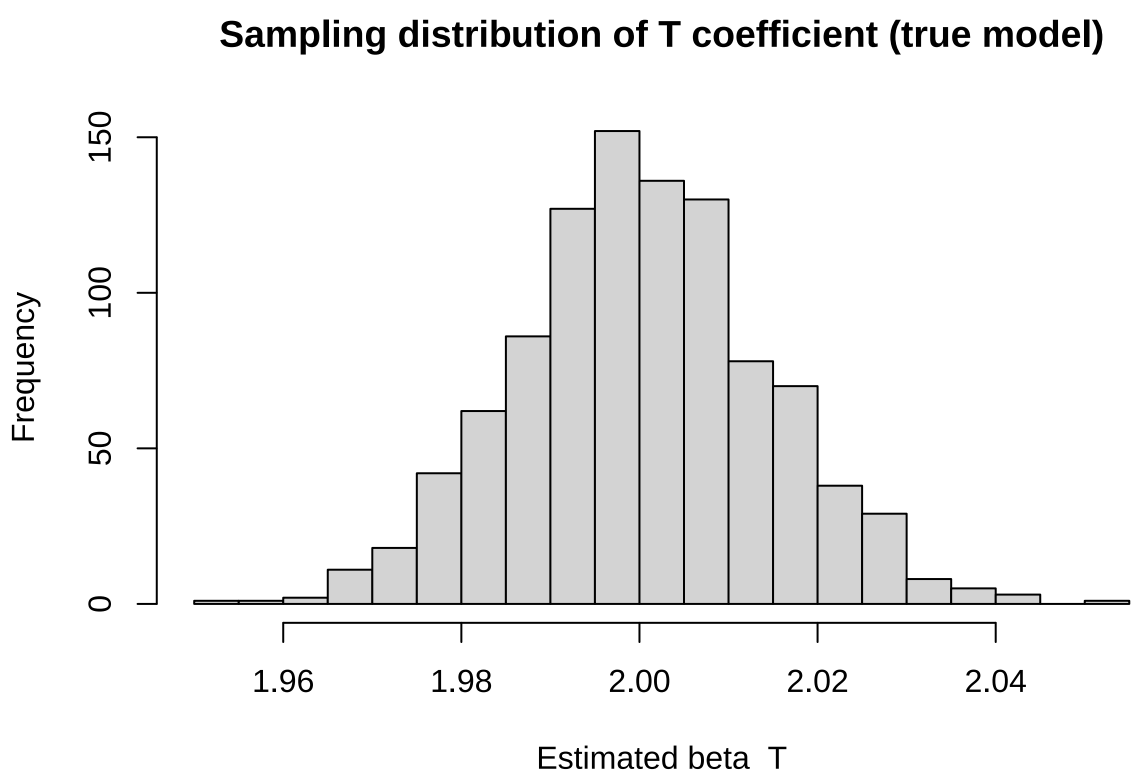
```
## [1] 2.000429
```

```r
sd(beta_T)
```

```
## [1] 0.01417777
```

```r
# Plot a histogram of the sampling distribution of the T coefficient

hist(
beta_T,
breaks = 30,
main = "Sampling distribution of T coefficient (true model)",
xlab = "Estimated beta_T"
)
```




```r
# By the central limit theorem, the sampling distribution of the T

# > coefficient should be approximately normal when we repeatedly

# > sample large datasets from the same DGP.
```

```r
# The histogram looks bell-shaped and centered near the true value of

# > 2, and the mean(beta_T) is very close to 2, which supports the CLT

# > intuition for this regression coefficient.
```

```r
# Part 1(b)

# Compute the bootstrapped standard error for the coefficient of the

# > treatment variable.

set.seed(789)

B_boot <- 1000
beta_boot <- numeric(B_boot)

# We bootstrap the original simulated dataset sim_data

for (b in seq_len(B_boot)) {
idx <- sample(seq_len(n), size = n, replace = TRUE)
boot_dat <- sim_data[idx, ]
fit_boot <- lm(Y ~ T + C, data = boot_dat)
beta_boot[b] <- coef(fit_boot)["T"]
}

boot_se <- sd(beta_boot)
boot_se
```

```
## [1] 0.01402459
```

```r
# Compare to the model-based standard error from the original model

se_model <- summary(mod_true)$coef["T", "Std. Error"]
se_model
```

```
## [1] 0.01412003
```

```r
# The bootstrap standard error boot_se is very close to the analytic

# > standard error se_model from the regression output.

# This shows that the model-based SE is doing a good job approximating

# > the true sampling variability of the T coefficient under this DGP.
```

```r
# Part 2: Data Analysis

# For this part of the assignment, use any data set you like.

# > Here I follow the instruction to use the thermometers data from

# > class (thermometers.csv).

thermo <- read.csv("/Users/santividal5/Desktop/R/thermometers.csv")
```

```r
thermo$party_id <- factor(thermo$party_id)
thermo$sex <- factor(thermo$sex)
thermo$race <- factor(thermo$race)
thermo$educ <- factor(thermo$educ)

head(thermo)
```

```
##   birth_year    sex  race    party_id                 educ ft_black ft_white
## 1       1931 Female White    Democrat               4-year       51       50
## 2       1952 Female White  Republican               2-year       98       90
## 3       1931   Male White Independent High school graduate       87       90
## 4       1952   Male White  Republican               4-year       90       85
## 5       1939 Female White    Democrat               2-year      100       50
## 6       1959 Female Black    Democrat            Post-grad       98       70
##   ft_hisp ft_asian ft_muslim ft_jew ft_christ ft_fem ft_immig ft_gays ft_unions
## 1      79       50        50     50        50     99       95      50        80
## 2      95      100        61    100        98     65       96      82        62
## 3      91       88        49     25        50     74       77      77       100
## 4      90       96        80     91        94     25       91      71        20
## 5     100      100       100    100        28    100      100     100       100
## 6      99      100       100    100       100     73      100      54        80
##   ft_police ft_altright ft_evang ft_dem ft_rep
## 1        76           1       50     88     21
## 2        95          50       96     86     96
## 3        78           0        2     91     20
## 4        94          50       70     22     83
## 5        28          NA       NA     99     NA
## 6        24           4       53     53      4
```

```r
# This confirms that the thermometer data loaded correctly and that the

# > main variables (party_id and thermometer scores) are present.
```

```r
# Part 2(a)

# 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.

#

# I test whether Democrats and Republicans differ in their mean feeling

# > thermometer toward immigrants (ft_immig).

# I use a two-sample t-test with unequal variances and a 5% significance

# > level, which is standard in this setting.

# Keep only Democrats and Republicans
```

```r
thermo_DR <- subset(thermo, party_id %in% c("Democrat", "Republican"))

tapply(
thermo_DR$ft_immig,
thermo_DR$party_id,
mean,
na.rm = TRUE
)
```

```
##    Democrat Independent    Not sure       Other  Republican
##    71.65829          NA          NA          NA    50.20192
```

```r
t.test(
ft_immig ~ party_id,
data = thermo_DR
)
```

```
##
##  Welch Two Sample t-test
##
## data:  ft_immig by party_id
## t = 22.673, df = 2685.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Democrat and group Republican is not e
## 95 percent confidence interval:
##  19.60077 23.31196
## sample estimates:
##   mean in group Democrat mean in group Republican
##                 71.65829                 50.20192
```

```r
# Interpretation:

# The t-test output shows the estimated difference in means, a t

# > statistic, and a p-value.

# The mean ft_immig for Democrats is substantially higher than for

# > Republicans, and the p-value is effectively zero at the 5% level.

# Statistically, we reject the null hypothesis that Democrats and

# > Republicans have the same average warmth toward immigrants.

# Substantively, this suggests that in this survey Democrats feel

# > noticeably warmer toward immigrants than Republicans do.
```

```r
# Part 2(b)

# Using the same data, fit a linear model. Interpret the coefficient,

# > standard error, t-value, and p-value.

#
```

```r
# I fit a simple linear model where the dependent variable is ft_immig

# > and the predictor is party_id (with Democrats as the baseline).

mod_party <- lm(ft_immig ~ party_id, data = thermo_DR)
summary(mod_party)
```

```
##
## Call:
## lm(formula = ft_immig ~ party_id, data = thermo_DR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.658 -19.658   2.798  19.342  49.798
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         71.6583     0.6233  114.97   <2e-16 ***
## party_idRepublican -21.4564     0.9320  -23.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.48 on 3021 degrees of freedom
##   (123 observations deleted due to missingness)
## Multiple R-squared:  0.1493, Adjusted R-squared:  0.149
## F-statistic:   530 on 1 and 3021 DF,  p-value: < 2.2e-16
```

```r
# Interpretation of key pieces:

# - The intercept is the estimated mean ft_immig for Democrats (the

# > baseline category).

# - The coefficient on party_idRepublican is the estimated difference

# > in means between Republicans and Democrats.

# It is negative and large in magnitude, which means Republicans give

# > lower immigrant thermometer scores on average.

# - The standard error for this coefficient measures how much that

# > estimated difference would vary across repeated samples.

# - The t-value is the estimated coefficient divided by its standard

# > error; a large absolute t-value indicates strong evidence that the

# > true difference is not zero.

# - The p-value associated with the t-value is extremely small, so we

# > reject the null hypothesis that Democrats and Republicans have
```

```
# > equal mean ft_immig scores.
# Substantively, this matches the two-sample t-test: party ID is strongly
# > associated with how warmly respondents feel toward immigrants, with
# > Democrats rating immigrants much more positively than Republicans.
```