Coding Problem

Yuhuan Huang

2025-04-04

Q1

```
#Q1
set.seed(101)
num = 10000
e <- rnorm(num,mean=0,sd=2)
y <- 100 + exp(e)
#print(y)
print(mean(y))
```

```
## [1] 107.0885
```

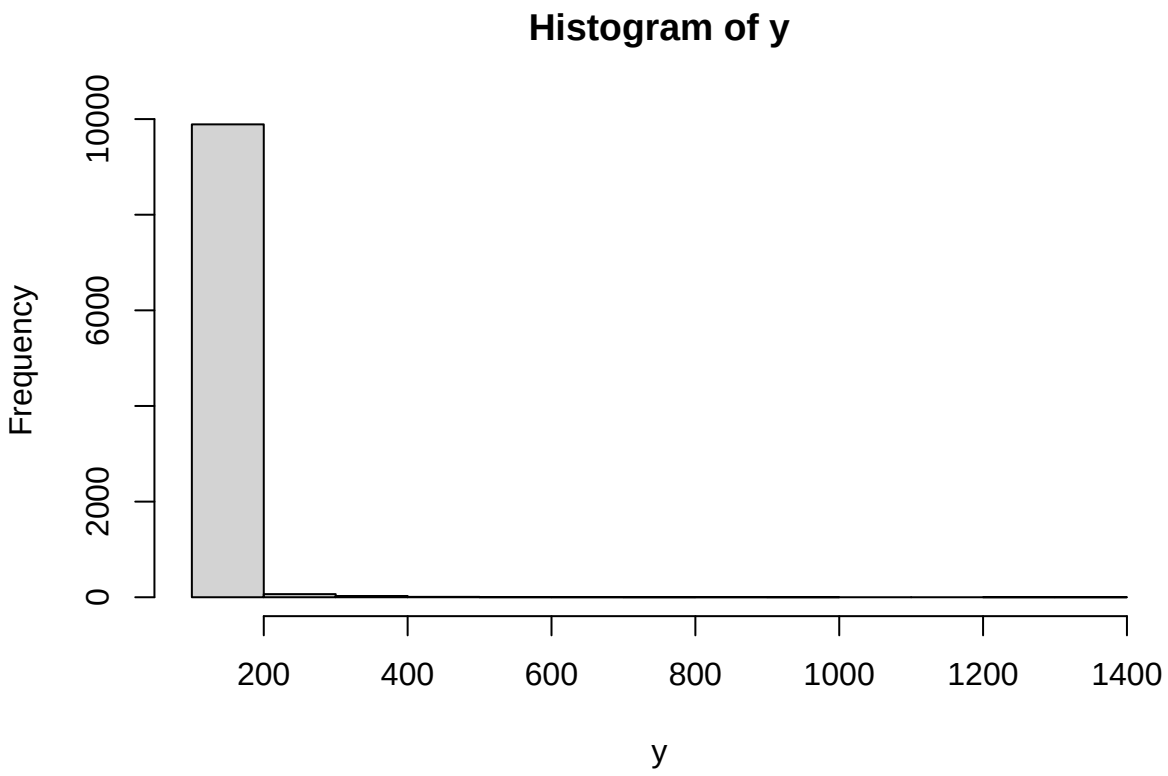
```
print(var(y))
```

```
## [1] 1080.825
```

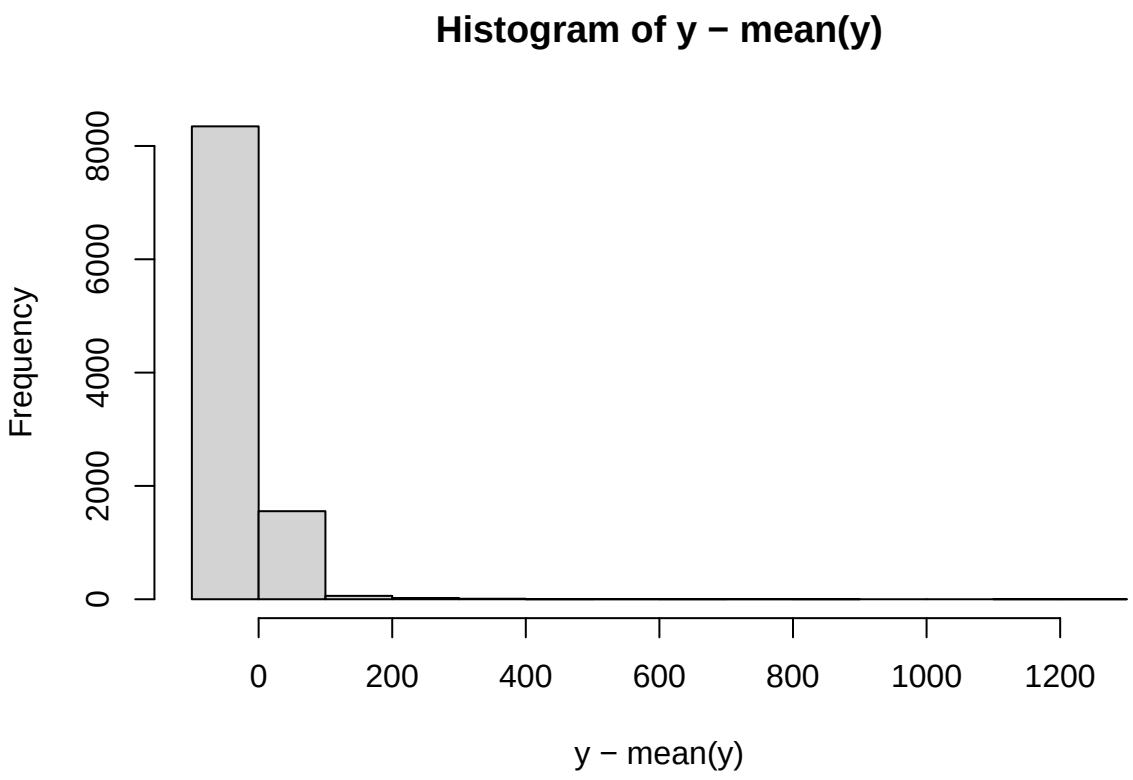
```
temp2 <- var(y)
```

Q2

```
#Q2
hist(y)
```


Histogram of y
Frequency
0
2000
6000
10000
200
400
600
800
1000
1200
1400
y

```
hist(y-mean(y))
```



Q3

```
#Q3

#construct data frame
df_data <- data.frame(y, e)
df_data0 <- data.frame(y, e)

#random 50-50 students
ran_indices <- sample(1:num,100)
print(ran_indices)
```

```
##   [1] 3681 4901 3738  806  305  599 2791 3129 7606 8474 9570 6962 7656 2232 5798
##  [16] 5332  411 5276 5363 2965  783 7597 2098 5757 3273  686 2001  705 3834 5264
##  [31] 3875 8260 8009  150 8021 2528 9034 4215  974 2494 5800 4990 1557 6484 5883
##  [46] 5117 7464 1152 6544 9071 6198 5752 4694 2988 2277 7695 9286 3771 9909 9680
##  [61] 8894 3809 5247 2208 3983 1183 9215 4999 4905  535  343 2501 6126 2580 2196
##  [76]  256 2610 2245 2218 3296 7865 9482 5758 9790 1480  484 2423 1458  364 8571
##  [91] 9461 1830 6065 1707 3822 7475 9842 3911 7858 6848
```

```
treated_indices <- sample(ran_indices, 50)
untreated_indices <- setdiff(ran_indices,treated_indices)
print(treated_indices)
```

```
##  [1] 2501 3738 7656 5264 1557 9286  411 3681 2277 5758  686 8571 4999 1458 2610
```

```
## [16] 6484  974 8260  364 4990 5798 1152 7597 9842 9680 2423 9461 6198 1480 3834
## [31] 1707 6962 5247 7464 7475 3771 3296 3809 3273 8894 2494 8021 9790 9071 9034
## [46] 6848  599 2988  150 6065
```

```
print(untreated_indices)
```

```
##  [1] 4901  806  305 2791 3129 7606 8474 9570 2232 5332 5276 5363 2965  783 2098
## [16] 5757 2001  705 3875 8009 2528 4215 5800 5883 5117 6544 5752 4694 7695 9909
## [31] 2208 3983 1183 9215 4905  535  343 6126 2580 2196  256 2245 2218 7865 9482
## [46]  484 1830 3822 3911 7858
```

```
#add treatment D
df_data$D <- 0
df_data$D[treated_indices] = 1

df_subdata <- df_data[ran_indices,]
df_subdata$new_y <- df_subdata$y + 5*df_subdata$D

#regression
model <- lm(new_y ~ D, data = df_subdata)
summary(model)
```

```
##
## Call:
## lm(formula = new_y ~ D, data = df_subdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.07 -22.63  -8.85  -7.35 353.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  108.916      7.712   14.12   <2e-16 ***
## D             20.177     10.906    1.85   0.0673 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.53 on 98 degrees of freedom
## Multiple R-squared:  0.03375,    Adjusted R-squared:  0.02389
## F-statistic: 3.423 on 1 and 98 DF,  p-value: 0.06731
```

```
print(coef(model)["D"])
```

```
##        D
## 20.17685
```

Q4

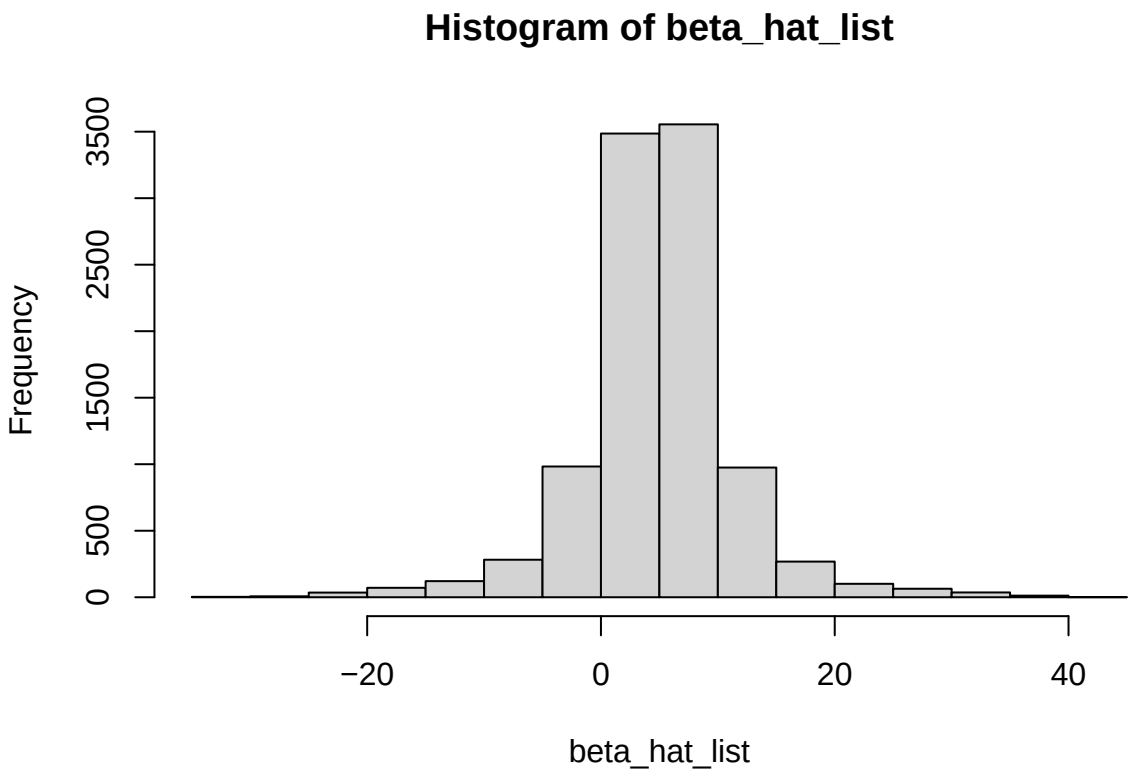
```
beta_hat_list <- numeric(10000)

for (i in 1:10000){
```

```
  temp_df_data <- df_data0

  #random 50-50 students
  temp_ran_indices <- sample(1:num,100)
  #print(temp_ran_indices)
  temp_treated_indices <- sample(temp_ran_indices, 50)
  temp_untreated_indices <- setdiff(temp_ran_indices,temp_treated_indices)
  #print(temp_treated_indices)
  #print(temp_untreated_indices)

  #add treatment D
  temp_df_data$D <- 0
  temp_df_data$D[temp_treated_indices] = 1

  temp_df_subdata <- temp_df_data[temp_ran_indices,]
  temp_df_subdata$new_y <- temp_df_subdata$y + 5*temp_df_subdata$D

  #regression
  temp_model <- lm(new_y ~ D, data = temp_df_subdata)
  #print(coef(temp_model)["D"])
  beta_hat_list[i] = coef(temp_model)["D"]
}

#print(beta_hat_list)
print(mean(beta_hat_list))
```

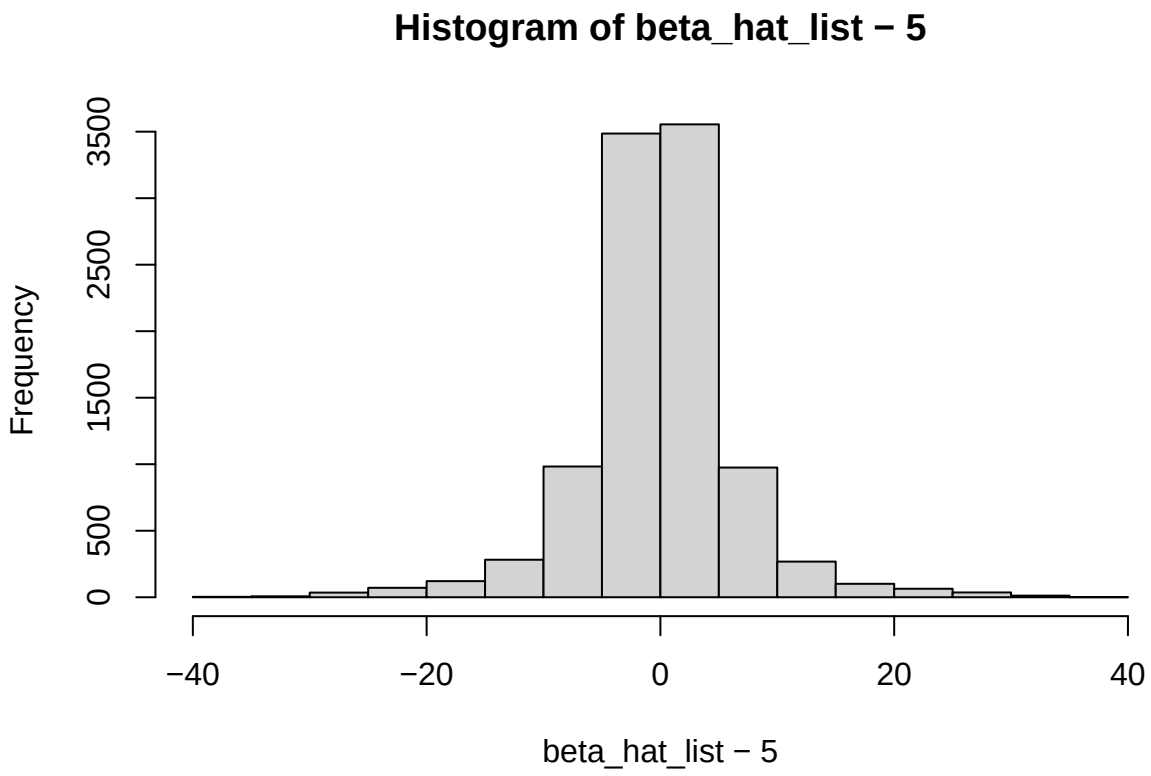
```
## [1] 4.932684
```

Q5

```
hist(beta_hat_list)
```




```
hist(beta_hat_list-5)
```



Q6

```
negative_num <- sum(beta_hat_list<0)
p_negative <- negative_num/length(beta_hat_list)
print(p_negative)
```

```
## [1] 0.1502
```

Q7

```
# Q7
alpha = 0.05
t_alpha = qnorm(1-alpha/2)
k = 0.8
t_1_minus_k = qnorm(k)
p = 0.5
sigma_square = temp2
MDE = 5

N_star = ( (t_1_minus_k+t_alpha)/MDE )^2 * sigma_square/(p*(1-p))
print(N_star)
```

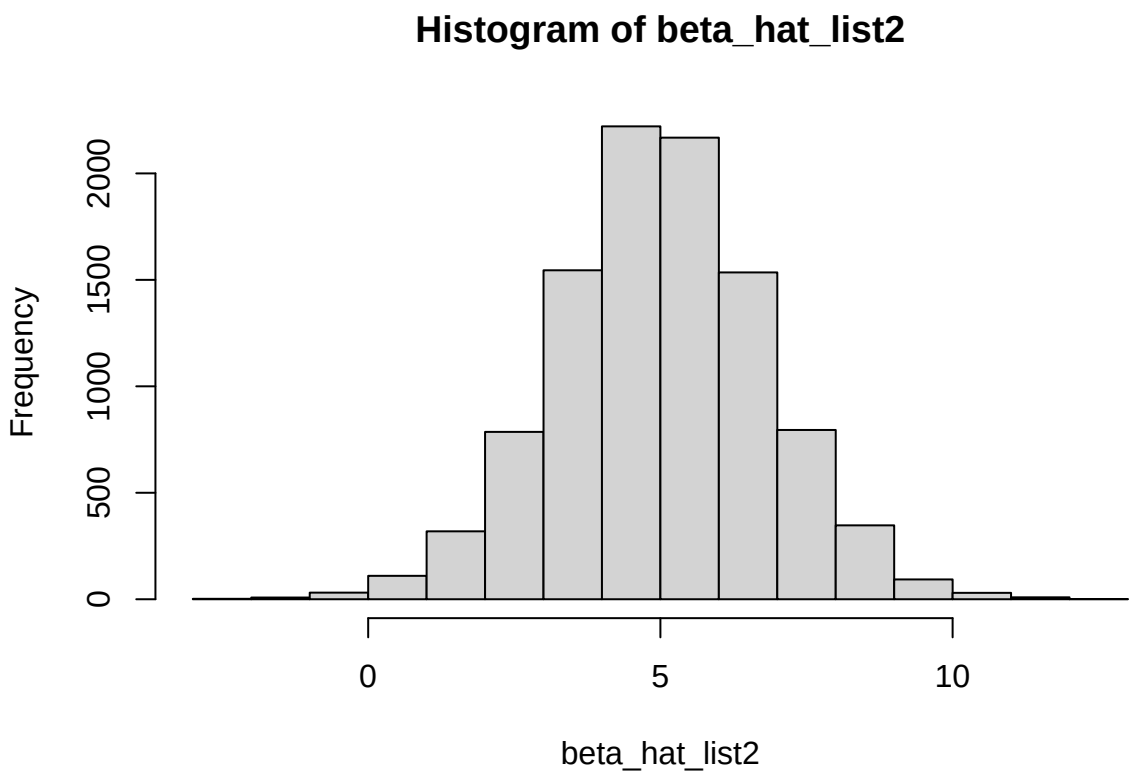
```
## [1] 1357.323
```

Q8

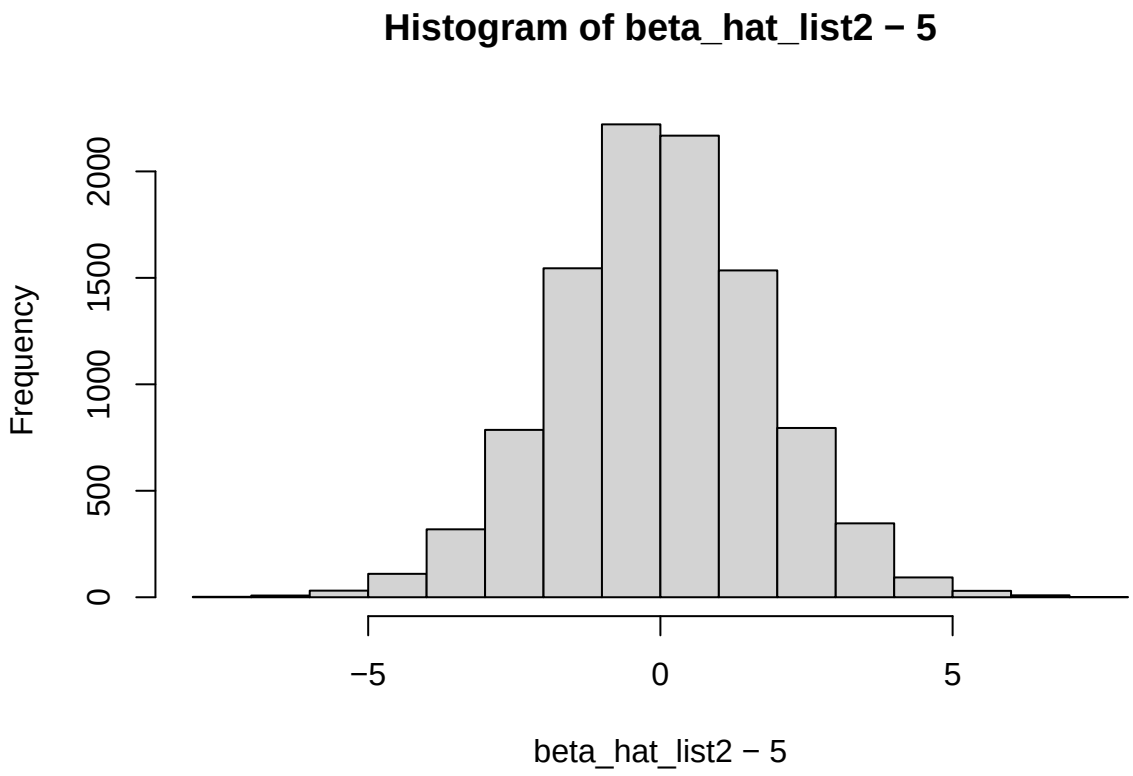
```r
beta_hat_list2 <- numeric(10000)
p_value_list <- numeric(10000)

for (i in 1:10000){
  temp_df_data <- df_data0

  #random 50-50 students
  temp_ran_indices <- sample(1:num,N_star)
  #print(temp_ran_indices)
  temp_treated_indices <- sample(temp_ran_indices, N_star/2)
  temp_untreated_indices <- setdiff(temp_ran_indices,temp_treated_indices)
  #print(temp_treated_indices)
  #print(temp_untreated_indices)

  #add treatment D
  temp_df_data$D <- 0
  temp_df_data$D[temp_treated_indices] = 1

  temp_df_subdata <- temp_df_data[temp_ran_indices,]
  temp_df_subdata$new_y <- temp_df_subdata$y + 5*temp_df_subdata$D

  #regression
  temp_model <- lm(new_y ~ D, data = temp_df_subdata)
  #print(coef(temp_model)["D"])
  beta_hat_list2[i] = coef(temp_model)["D"]
  p_value_list[i] <- summary(temp_model)$coefficients["D", "Pr(>|t|)"]
}

#print(beta_hat_list2)
print(mean(beta_hat_list2))
```

```
## [1] 5.000886
```

```r
hist(beta_hat_list2)
```




```
hist(beta_hat_list2-5)
```




```
#print(p_value_list)
# rejects H0: beta=0 when p_value < 0.05
rej <- sum(p_value_list<0.05)
p_rej <- rej/10000
print(p_rej)
```

```
## [1] 0.8032
```