Jerry_Huang_HW1

Jerry Huang

2024-06-05

```
# install.packages("readr")
# install.packages("dplyr")
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Problem 1 - Hit or Miss?

Question 1:

```
# getwd()
# setwd()
leaders <- read_csv("leaders.csv", show_col_types = FALSE)
attempts <- nrow(leaders)
attempts
```

```
## [1] 250
```

```
countries_attempts <- unique(leaders$country)
length(countries_attempts)
```

```
## [1] 88
```

```
num_countries <- length(countries_attempts)
attempts_per_country_per_year <- leaders %>%
  group_by(country, year) %>%
  summarise(attempts = n(), .groups = 'drop')
average_attempts_per_year <- mean(attempts_per_country_per_year$attempts)
average_attempts_per_year
```

```
## [1] 1.01626
```

There are 250 attempts recorded in data. There are 88 countries experience at least one leader assassination attempt. The average number of such attempts per year among these countries is 1.01626.

Question 2:

```
leaders$success <- as.integer(grepl("dies", leaders$result))
num_success <- sum(leaders$success)
success_rate <- mean(leaders$success) # Overall success rate
success_rate
```

```
## [1] 0.216
```

```
test_result <- binom.test(num_success, attempts, p = 0.5, alternative = "two.sided")
test_result
```

```
##
##  Exact binomial test
##
## data:  num_success and attempts
## number of successes = 54, number of trials = 250, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1666307 0.2722389
## sample estimates:
## probability of success
##                  0.216
```

The overall success rate is 0.216. It does not speak to the validity of the assumption that the success of assassination attempts is randomly determined, because the p-value is so small which reject the null hypothesis that the true probability of success is 0.5.

Question 3:

```
average_score_before_success <- mean(leaders$politybefore[leaders$success == "1"])
average_score_before_success
```

```
## [1] -0.7037037
```

```
average_score_before_failure <- mean(leaders$politybefore[leaders$success == "0"])
average_score_before_failure
```

```
## [1] -1.743197
```

```
diff_in_score <- average_score_before_success - average_score_before_failure
diff_in_score
```

```
## [1] 1.039494
```

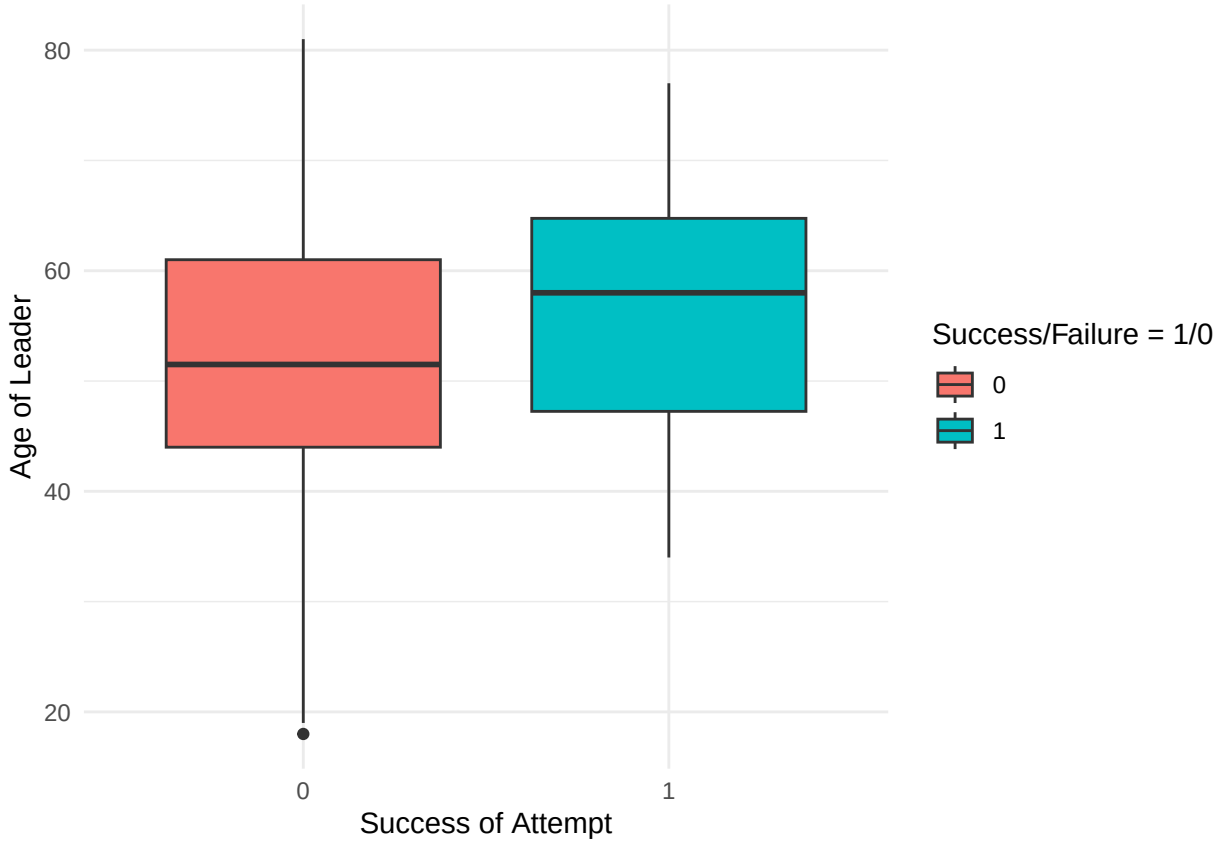
```
age_success <- leaders$age[leaders$success == "1"]
age_failure <- leaders$age[leaders$success == "0"]

leaders %>%
  group_by(success) %>%
  summarize(
    count = n(),
    mean_age = mean(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
```

```
    min_age = min(age, na.rm = TRUE),
    max_age = max(age, na.rm = TRUE)
  )
```

```
## # A tibble: 2 x 6
##   success count mean_age sd_age min_age max_age
##     <int> <int>    <dbl>  <dbl>   <dbl>   <dbl>
## 1       0   196     52.7   12.3      18      81
## 2       1    54     56.5   10.4      34      77
```

```
ggplot(leaders, aes(x = factor(success), y = age, fill = factor(success))) +
  geom_boxplot() +
  labs(x = "Success of Attempt", y = "Age of Leader", fill = "Success/Failure = 1/0") +
  theme_minimal()
```




```
t_test_result <- t.test(age ~ success, data = leaders)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by success
## t = -2.2436, df = 97.893, p-value = 0.02711
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -7.064505 -0.432850
## sample estimates:
## mean in group 0 mean in group 1
```

```
##        52.71429        56.46296
```

There is a difference in "polity score before" between successful and failed attempts. Meanwhile, there is a difference in the age of targeted leaders between successful and failed attempts. As we compare the two groups (success or failure), the mean_age for success attempted assassination is greater than that of failure attempted assassination. Moreover, the success group has a smaller sd_age, meaning that it is more clustered between the age 34 to 77. If the success or failure of assassination attempts is assumed to be essentially random, then there should be no different in the distribution of age between two groups.

Thus, I then used Welch Two Sample t-test to investigate whether there is a statistically significant difference in the average age between leaders who were targeted in successful versus failed attempts. The result turns out to be 0.02711, which is smaller than $\alpha = 0.05$. Therefore, the age of the leaders play an important role in whether the assassination is successful or not.

Question 4

```
leaders <- leaders %>%
  mutate(warbefore = as.integer(interwarbefore == 1 | civilwarbefore == 1))

average_score_before_inWar <- mean(leaders$warbefore[leaders$success == "1"])
average_score_before_inWar
```

```
## [1] 0.3518519
```

```
average_score_before_noWar <- mean(leaders$warbefore[leaders$success == "0"])
average_score_before_noWar
```

```
## [1] 0.372449
```

```
diff_in_score_war <- average_score_before_inWar - average_score_before_noWar
diff_in_score_war
```

```
## [1] -0.02059713
```

```
t_test_war_success <- t.test(leaders$warbefore[leaders$success == 1], leaders$warbefore[leaders$success
print(t_test_war_success)
```

```
##
##  Welch Two Sample t-test
##
## data:  leaders$warbefore[leaders$success == 1] and leaders$warbefore[leaders$success == 0]
## t = -0.27769, df = 84.851, p-value = 0.7819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1680747  0.1268805
## sample estimates:
## mean of x mean of y
## 0.3518519 0.3724490
```

```
t_test_warbefore_polity <- t.test(leaders$politybefore[leaders$warbefore == 1], leaders$politybefore[lea
print(t_test_warbefore_polity)
```

```
##
##  Welch Two Sample t-test
##
## data:  leaders$politybefore[leaders$warbefore == 1] and leaders$politybefore[leaders$warbefore == 0]
## t = 0.16029, df = 179.07, p-value = 0.8728
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -1.566129  1.843050
## sample estimates:
## mean of x mean of y
## -1.431159 -1.569620
```

```
t_test_war_age <- t.test(leaders$age[leaders$warbefore == 1], leaders$age[leaders$warbefore == 0], alte
print(t_test_war_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  leaders$age[leaders$warbefore == 1] and leaders$age[leaders$warbefore == 0]
## t = 0.48313, df = 212.54, p-value = 0.6295
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.266807  3.738738
## sample estimates:
## mean of x mean of y
##  53.98913  53.25316
```

1) The analysis on the average successful assassination attempt with whether a country is in either civil or international war during the three years prior to an assassination attempt. The t-test has shown that there is no significant difference on average whether a country is in war between if the assassination is successful or not.
2) The analysis on whether the average polity score over three years prior to an assassination attempt differs on average between whether a country is in either civil or international war during the three years prior to an assassination attempt.The t-test has shown that there is no significant difference in mean of polity score before assassination between if the assassination is successful or not. The polity score before is similar on whether or not the country is in war.
3) The analysis on the leaders age with whether a country is in either civil or international war during the three years prior to an assassination attempt. The t-test has shown that there is no significant difference in mean of age of leaders between if the assassination is successful or not. Therefore, the analysis suggest that neither the success of assassination attempts, the polity score before, nor the age of leaders show significant differences based on the war status of their countries.

Question 5:

```
leaders <- leaders %>%
  mutate(warafter = as.integer(interwarafter == 1 | civilwarafter == 1))
leaders$warhappen <- as.integer(leaders$warbefore == 0 & leaders$warafter == 1)

# Comparing the polity scores before and after the assassination for successfully assassinated leaders.
t_test_politychange_success <- t.test(leaders$polityafter[leaders$success == 1], leaders$politybefore[l
print(t_test_politychange_success)
```

```
##
##  Welch Two Sample t-test
##
## data:  leaders$polityafter[leaders$success == 1] and leaders$politybefore[leaders$success == 1]
## t = -0.047464, df = 105.84, p-value = 0.9622
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.508200  2.390916
## sample estimates:
```

```
##  mean of x  mean of y
## -0.7623457 -0.7037037
```

```
# Does successful leader assassination lead countries to war? Does successful leaders assassination cau
t_test_warchange_success <- t.test(leaders$warhappen[leaders$success == 1], leaders$warhappen[leaders$su
print(t_test_warchange_success)
```

```
##
##  Welch Two Sample t-test
##
## data:  leaders$warhappen[leaders$success == 1] and leaders$warhappen[leaders$success == 0]
## t = 0.13127, df = 82.175, p-value = 0.8959
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08291465  0.09463045
## sample estimates:
##  mean of x  mean of y
## 0.09259259 0.08673469
```

There is no significant evidence saying that a successful leader assassination will cause democratization. The t-test comparing the difference in means between the polity score after assassination and before assassination given successful assassination has a p-value of 0.9622, which is quite larger than the significant level $\alpha = 0.05$.

Meanwhile, there is no significant result saying that successful leader assassination will lead countries to war. I assume that the war happen after assassination by selecting groups with warbefore == 0 and warafter == 1. This helps me to focus on how success and failure lead to the warfare. The t-test comparing the difference in means between successful and failed assassination given the war happen (there is a war happening after assassination) has a p-value of 0.8959. Therefore, we fail the reject the null hypothesis that there is no difference in means between two groups.

Problem 2 - ITE and ATE

Question 1:

$$ITE_i = Y_i(1) - Y_i(0)$$

```
patients <- data.frame(
  Name = c("Rheia", "Kronos", "Demeter", "Hades", "Hestia", "Poseidon", "Hera", "Zeus", "Artemis",
           "Apollo", "Leto", "Ares", "Athena", "Hephaestus", "Aphrodite", "Cyclope", "Persephone",
           "Hermes", "Hebe", "Dionysus"),
  D_i = c(0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1),
  Y_i_0 = c(0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1),
  Y_i_1 =  c(1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0),
  Y_i = c(0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0)
)
result<-patients$Name[patients$Y_i == 1]
result
```

```
##  [1] "Kronos"     "Zeus"       "Artemis"    "Apollo"     "Ares"
##  [6] "Athena"     "Hephaestus" "Aphrodite"  "Cyclope"    "Persephone"
```

Question 2:

```
patients$Individual_Causal_Effect <- patients$Y_i_1 - patients$Y_i_0
ATE = mean(patients$Individual_Causal_Effect)
print(paste("ATE =", ATE))
```

```
## [1] "ATE = 0"
```

We cannot conclude that "null hypothesis of no average causal effect" true. Even though ATE is found to be 0, we can only say that we fail to reject the null hypothesis of no average causal effect.

Question 3:

```
Average_ITE = mean(patients$Individual_Causal_Effect)
Average_ITE == ATE
```

```
## [1] TRUE
```

Yes, ATE is the same as the average of ITEs.

Question 4:

When there is no causal effect for any unit in the population, we say that we fail to reject the sharp null hypothesis. Yes, sharp causal null hypothesis does imply the null hypothesis of no average causal effect. It is because we have assume all observed outcomes equal to the potential outcomes with or without the treatment in the experiment. Thus, there will be no individual treatment effect. The absence of ATE does not imply the absence of ITE. It is possible that some individuals have positive effects and some have negative, which may be averaged to 0 in calculating ATE.

Question 5:

```
E_Y1_given_D1 <- mean(patients$Y_i_1[patients$D_i == "1"])
E_Y0_given_D0 <- mean(patients$Y_i_0[patients$D_i == "0"])

mean_assoc_difference<- E_Y1_given_D1 - E_Y0_given_D0
mean_assoc_difference
```

```
## [1] 0.1098901
```

The result of mean association difference is 0.11, so there is an association between treatment and outcome.

Problem 2 - Scenario 2

Question 1

```
Proportion_A_Treated <- 2/10
Proportion_A_Control <- 3/10
ATE_A <- Proportion_A_Treated - Proportion_A_Control

Proportion_B_Treated <- 1/10
Proportion_B_Control <- 3/10
ATE_B <- Proportion_B_Treated - Proportion_B_Control

ATE_A
```

```
## [1] -0.1
```

```
ATE_B
```

```
## [1] -0.2
```

The ATE for treatment A is -0.1. The ATE for treatment B is -0.2. The result is negative because the algorithm is Proportion_treated - Proportion_Control.

Question 2

10 people needed to be treated to save one life in treatment A. 5 people needed to be treated to save one life in treatment B.

Question 3

```
total_treated_A <- 10000000/10000
total_treated_B <- 10000000/4000

total_saved_A <- total_treated_A*abs(ATE_A)
total_saved_B <- total_treated_B*abs(ATE_B)

total_saved_A
```

```
## [1] 100
```

```
total_saved_B
```

```
## [1] 500
```

Within the same budget, the total number people saved using treatment B is greater, so treatment B is suggested to use.

Problem 3 - ATT and ATE

Question 1

ATT measures the average treatment effect for those who have been treated (Di=1). In other words, It is a conditional expectation that reflects the average treatment effect, but only for those who have received the treatment. ATT is different from ATE because it is restricted and only reflects on the units being treated truly, whereas ATE will measure the difference between two potential outcomes for all units.

Question 2

$$\tau^t = E[Y_i(1) - Y_i(0)|D_i = 1]$$

Initially, we apply the linearity of expectation to decompose the formula into

$$\tau^t = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

Since consistency holds for all treatment levels, we then know that the potential outcomes will equal to the observed outcome. We will have:

$$E[Y_i(1)|D_i = 1] = E[Y_i|D_i = 1]$$

Secondly, we know that the weak ignorability holds only for the control outcome, meaning that the control outcome is independent from whether being assigned with treatment or not. We will have:

$$E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$$

Also, because of consistency holds for all treatment levels, we will then have:

$$E[Y_i(0)|D_i = 0] = E[Y_i|D_i = 0]$$

Eventually, we combine everything together. We will have:

$$\tau^t = E[Y_i(1) - Y_i(0)|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Question 3

$$ATE - ATT = E[Y_i(1) - Y_i(0)] - E[Y_i(1) - Y_i(0)|D_i = 1]$$

Initially, we use the linearity of expectation to decompose both ATE and ATT. We will have:

$$ATE - ATT = E[Y_i(1)] - E[Y_i(0)] - (E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1])$$

Since we know that weak ignorability holds for the control outcome only, we will simplify ATE and ATT into:

$$ATE - ATT = E[Y_i(1)] - E[Y_i(0)|D_i = 0] - (E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0])$$

Also, as we know that consistency holds for all treatments, we will have:

$$ATE - ATT = E[Y_i(1)] - E[Y_i|D_i = 0] - (E[Y_i|D_i = 1] - E[Y_i|D_i = 0])$$

Eventually, we will have:

$$ATE - ATT = E[Y_i(1)] - E[Y_i(1)|D_i = 1]$$

It will be necessary to assume $Y_i(1)$ and D_i are independent for the ATT to be equal to the ATE. It tells us that the observed outcome $E[Y_i(1)|D_i = 1]$ is the same as the counterfactual outcome $E[Y_i(1)|D_i = 0]$ when the treatment D_i is independent of the potential outcome Y_i. If that is true, we will have:

$$E[Y_i(1)] = E[Y_i(1)|D_i = 1]$$

Therefore, if assuming ignorability for both treatment and control outcome, we will have:

$$ATE - ATT = E[Y_i(1)] - E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 1] = 0$$

The assumption about $Y_i(1)$ and D_i are independent is enough because we have already assumed consistency is true for all treatment levels and the weak ignorability for control outcome.

Problem 4 - Bernoulli Trial and ATE - Part a

Question 1:

D_i follows a Bernoulli distribution.

Question 2:

We define the Number of treated units here as:

$$N_t = \sum_{i=1}^{n} D_i$$

The expected value can be written as:

$$E[N_t] = E\left[\sum_{i=1}^{n} D_i\right]$$

It can then be simplified using the linearity of expectation:

$$E[N_t] = \sum_{i=1}^{n} E[D_i] = \sum_{i=1}^{n} p = np$$

Thus, the expected value of the number of treated units is:

$$E[N_t] = np$$

Question 3:

The variance of the number of treated units can be written as:

$$Var[N_t] = Var\left[\sum_{i=1}^{n} D_i\right]$$

Then, it can be decomposed to:

$$Var[N_t] = \sum_{i=1}^{n} Var\left[D_i\right] = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Question 4:

We want that $E[N_t] = n_t$ Proof: We know that the expected value for a Bernoulli random variable is p $E[D_i] = 1 * p + 0 * (1-p)$

$$E[N_t] = E\left[\sum_{i=1}^{n} D_i\right] = \sum_{i=1}^{n} E[D_i]$$

So the sum of the expected Bernoulli random variable, which is the total number of treated units is:

$$E[N_t] = \sum_{i=1}^{n} E[D_i] = np$$

Therefore, we need

$$p = \frac{n_t}{n}$$

Problem 4 - Part b

Trying to prove that $E[\hat{\tau}_{IPW}] = \tau = E[Y_i(1)] - E[Y_i(0)]$ Proof:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{p}Y_i - \frac{1-D_i}{1-p}Y_i\right)$$

$$E[\hat{\tau}_{IPW}] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{p}\right)\right] - E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-D_i)Y_i}{1-p}\right)\right]$$

$$E[\hat{\tau}_{IPW}] = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\frac{D_i Y_i}{p}\right)\right] - \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\frac{(1-D_i)Y_i}{1-p}\right)\right]$$

For the first term, by consistency we know that $Y_i = Y_i(1)$ when $D_i = 1$ and by ignorability we know that $E[Y_i(1)] = E[Yi(1)|D_i = 1]$, we will have:

$$E\left[\left(\frac{D_i Y_i}{p}\right)\right] = \frac{1}{p}E\left[Y_i(1)D_i\right] = \frac{1}{p}E\left[Y_i(1)\right]E[D_i] = \frac{1}{p}E\left[Y_i(1)\right]p = E[Y_i(1)]$$

Similarly, by consistency we know that $Y_i = Y_i(0)$ when $D_i = 0$ and by ignorability we know that $E[Y_i(0)] = E[Yi(0)|D_i = 0]$, we will have:

$$E\left[\left(\frac{(1-D_i)Y_i}{1-p}\right)\right] = \frac{1}{1-p}E\left[Y_i(0)(1-D_i)\right] = \frac{1}{1-p}E\left[Y_i(0)\right]E[1-D_i] = \frac{1}{1-p}E\left[Y_i(0)\right](1-p) = E[Y_i(0)]$$

Thus, we will have:

$$E[\hat{\tau}_{IPW}] = E[Y_i(1)] - E[Y_i(0)] = \tau$$