CDCS2024NHT-teaching

Fang Yang

2024-05-09

1 Preparation

```
# Load packages
library(tidyverse)
library(patchwork)
library(kableExtra)
```

```
# Read data
data <- read_csv("Instadata.csv")
```

2 Data Wrangling

```
# Inspect data
dim(data)
```

```
## [1] 160   3
```

```
glimpse(data)
```

```
## Rows: 160
## Columns: 3
## $ ...1  <dbl> 1, 102, 3, 28, 5, 152, 7, 32, 9, 96, 11, 98, 13, 2, 15, 42, 17, ~
## $ Group <chr> "Unistudent", "FTemployee", "Unistudent", "FTemployee", "Unistud~
## $ Time  <dbl> 52.22, -44.14, 45.03, -25.24, 53.88, -1.00, 44.16, 26.02, 57.29,~
```

```
summary(data)
```

```
##       ...1           Group                Time        
##  Min.   :  1.00   Length:160         Min.   :-44.14  
##  1st Qu.: 40.75   Class :character   1st Qu.: 37.33  
##  Median : 80.50   Mode  :character   Median : 43.19  
##  Mean   : 80.50                      Mean   : 42.28  
##  3rd Qu.:120.25                      3rd Qu.: 49.78  
##  Max.   :160.00                      Max.   : 59.71  
##                                      NA's   :5
```

```r
data$Group <- as.factor(data$Group)
levels(data$Group)
```

```
## [1] "FTemployee" "Unistudent"
```

3 Descriptive Statistics

3.1 Contingency Table

```r
tbl_stats <- data %>%
  group_by(Group) %>%
  summarise(n = n(),
            M = mean(Time),
            SD = sd(Time),
            Min = min(Time),
            Max = max(Time))
tbl_stats
```

```
## # A tibble: 2 x 6
##   Group          n     M    SD   Min   Max
##   <fct>      <int> <dbl> <dbl> <dbl> <dbl>
## 1 FTemployee    80    NA    NA    NA    NA
## 2 Unistudent    80    NA    NA    NA    NA
```

3.1.1 Think point

Any problems?

How to fix them?

3.1.2 Missing values

```r
data <- data %>% drop_na(Time)

# check the data
tbl_stats <- data %>%
  group_by(Group) %>%
  summarise(n = n(),
            M = mean(Time),
            SD = sd(Time),
            Min = min(Time),
            Max = max(Time))
tbl_stats
```

```
## # A tibble: 2 x 6
##   Group          n     M    SD   Min   Max
##   <fct>      <int> <dbl> <dbl> <dbl> <dbl>
## 1 FTemployee    76  34.6 13.3  -44.1  48.8
## 2 Unistudent    79  49.7  5.28  36.9  59.7
```

3.1.3 Impossible values

```
isTRUE(data$Time > 0)
```

```
## [1] FALSE
```

```
data <- data %>%
  filter(Time > 0)

# check the data
tbl_stats <- data %>%
  group_by(Group) %>%
  summarise(n = n(),
            M = mean(Time),
            SD = sd(Time),
            Min = min(Time),
            Max = max(Time))
tbl_stats
```

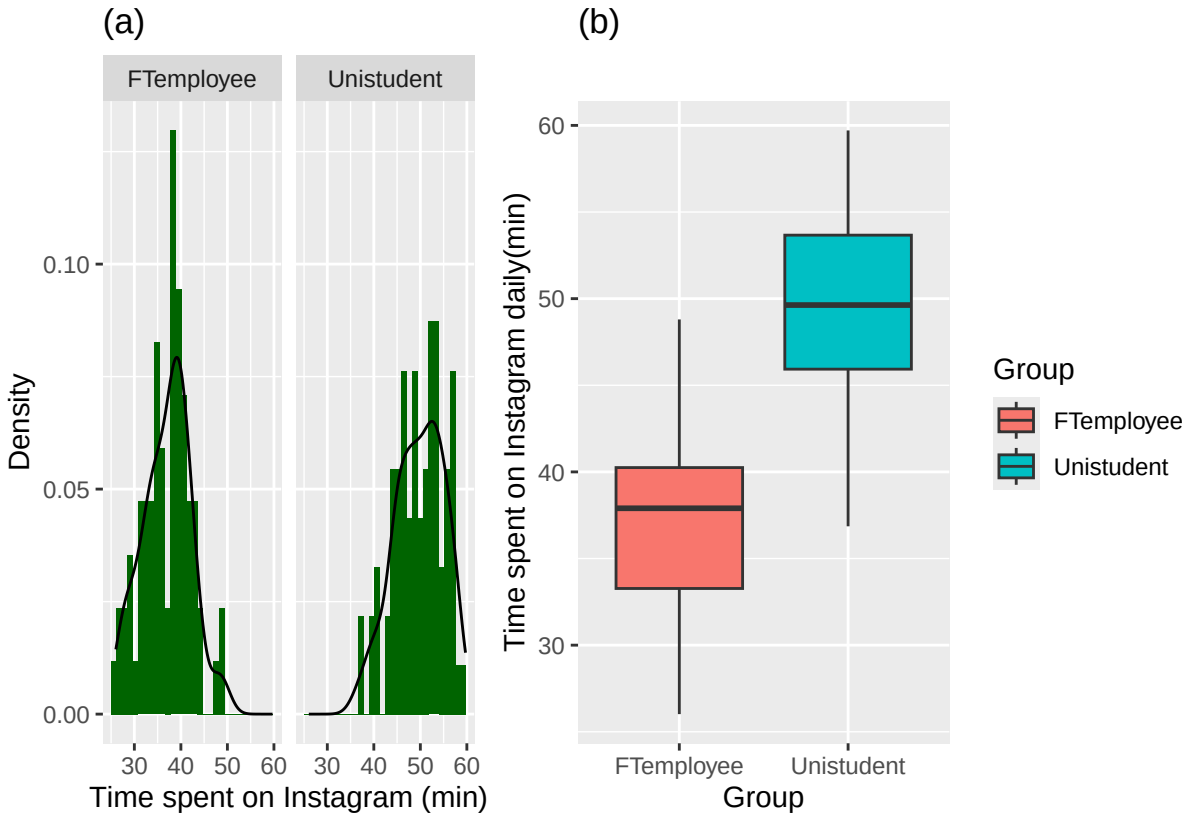
```
## # A tibble: 2 x 6
##   Group          n     M    SD   Min   Max
##   <fct>      <int> <dbl> <dbl> <dbl> <dbl>
## 1 FTemployee    73  36.9  5.12  26.0  48.8
## 2 Unistudent    79  49.7  5.28  36.9  59.7
```

3.2 Visualisation

```
plt_hist <- ggplot(data, aes(x = Time, after_stat(density))) +
  geom_histogram(fill = "darkgreen") +
  geom_density() +
  facet_wrap(~Group) +
  labs(x = "Time spent on Instagram (min)",
       y = "Density",
       title = "(a)")
plt_hist
```

```
plt_box <- data %>%
  ggplot(aes(x=Group, y = Time, fill=Group)) +
  geom_boxplot() +
  labs(x = "Group",
    y = "Time spent on Instagram daily(min)",
       title = "(b)") +
  theme()
plt_box
```

```
plt_hist | plt_box
```



4 Null Hypothesis Testing

Our Alternative Hypothesis of the research is that university students on average spend more time on Instagram daily then full-time exployees.

The Null Hypothesis is there is no different between the two groups.

$H0$: mu = 0 $H1$: mu > 0

Because our hypothesis is one-tailed, i.e., has a direction. we need to correctly specify the "alternative =" parameter. We will firstly check the reference group, and then specify the "alternative =" accordingly.

```
levels(data$Group)
```

```
## [1] "FTemployee" "Unistudent"
```

Our hypothesised direction is Uni student > FT employee. Given that our reference level is FT employee, we need to specify alternative as "less". This tell r that the alternative hypothesis assumes the group at the reference level (FT employee) is smaller than the group at the critical level (Uni student).

To make things easier, we relevel the factor so that University students are the reference group. We can then specify "alternative =" to "greater".

```
data$Group <- relevel(data$Group, ref= "Unistudent")
```

4.1 Two-sample t-test

Next we perform a two-sample t-test. We use the default significance level .05.

```
t_test <- t.test(data$Time ~ data$Group, mu = 0, alternative = "greater",
                 var.equal = TRUE)

t_test
```

```
##
##  Two Sample t-test
##
## data:  data$Time by data$Group
## t = 15.096, df = 150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Unistudent and group FTemployee is gr
## 95 percent confidence interval:
##  11.35358      Inf
## sample estimates:
## mean in group Unistudent mean in group FTemployee
##                 49.69696                 36.94534
```

4.1.1 Interpretation

At the 5% significance level, a two-sample t-test was conducted to investigate the amount of time spent on Instagram by university students and full-time employees. Results show that university students on average spend significantly more time (mean = 49.70 minutes) than than full-time employees (mean = 36.95 minutes) on Instagram everyday (t(df=150)= 15.10, p < .001).

4.2 Confidence Intervals

We have got the results of our t-test. But how confident are we about our results? To answer this question, we need to calculate the confidence intervals.

```
t_test_CI <- t.test(data$Time ~ data$Group, mu = 0, alternative = "two.sided",
                 var.equal = TRUE)
t_test_CI
```

```
##
##  Two Sample t-test
##
## data:  data$Time by data$Group
## t = 15.096, df = 150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Unistudent and group FTemployee is no
## 95 percent confidence interval:
##  11.08258 14.42066
## sample estimates:
## mean in group Unistudent mean in group FTemployee
##                 49.69696                 36.94534
```

4.2.1 Interpretation

We are 95% confident that university students on average spend between 11.08 and 14.42 more minutes on Instagram than full-time employees.

4.2.2 Think Point

What if we want to use a more rigid significance level (e.g., alpha = .01)?

Tip: you just need to specify conf.level = .99.

5 Exercise: YOUR TURN

Perform a two-sample t-test at significance level of .01 to answer the research question.

```
# Perform a t-test at alpha = .01

t_test_2 <- t.test(data$Time ~ data$Group, mu = 0, alternative = "less",
                   conf.level=.99,
                  var.equal = TRUE)

t_test_2
```

```
##
##  Two Sample t-test
##
## data:  data$Time by data$Group
## t = 15.096, df = 150, p-value = 1
## alternative hypothesis: true difference in means between group Unistudent and group FTemployee is les
## 99 percent confidence interval:
##      -Inf 14.73789
## sample estimates:
## mean in group Unistudent mean in group FTemployee
##                 49.69696                 36.94534
```

```
# get confidence intervals

t_test_2_CI <- t.test(data$Time ~ data$Group, mu = 0, alternative = "two.sided",
                   conf.level=.99,
                  var.equal = TRUE)

t_test_2_CI
```

```
##
##  Two Sample t-test
##
## data:  data$Time by data$Group
## t = 15.096, df = 150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Unistudent and group FTemployee is not
## 99 percent confidence interval:
##  10.54780 14.95543
## sample estimates:
## mean in group Unistudent mean in group FTemployee
##                 49.69696                 36.94534
```

How would you interpret the results?