CDCS2024NHT-teaching session2

Fang Yang

2024-05-16

1 Preparation

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

We will start with the same dateset from last week.

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

data <- data %>%
  drop_na(Time) %>%
  mutate(Group=factor(Group)) %>%
  mutate(Group=fct_relevel(Group,c("Unistudent","FTemployee"))) %>%
 arrange(Group)

levels(data$Group)


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

2 two-sample t-test

```
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 = 9.3563, df = 153, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Unistudent and group FTemployee is gre
## 95 percent confidence interval:
##  12.45888      Inf
## sample estimates:
## mean in group Unistudent mean in group FTemployee
##                 49.69696                 34.56092
```

Let μ_s denote the population mean time spent on Instagram by university students daily and μ_e denote the population mean time spent on Instagram everyday by full-time employees. To test whether μ_s is greater than μ_e, we performed a one-sided two-sample t-test of $H_1 : \mu_s > \mu_e$ against $H_0 : \mu_s = \mu_e$. This allows us to discern whether the observed numeric difference between the two sample means was due to an effect or if it was due to random sampling variation. We used A significance level of $\alpha = .05$.

3 Effect size

```
# use cohen_d() function from the 'effectsize' package to calculate effect size

D <- cohens_d(data$Time ~ data$Group,
         mu = 0,
         alternative = "two.sided",
         var.equal = TRUE)

D
```

```
## Cohen’s d |       95% CI
## ------------------------
## 1.50      | [1.14, 1.86]
##
## - Estimated using pooled SD.
```

The sample data provided very strong evidence against the null hypothesis and in favour of the alternative hypothesis that on average university students indeed spend more time than full-time employees everyday on Instagram.

The effect size was found to be large (Cohen’s $D =$ `D`). Therefore, we conclude that not only that the difference between the average time spent on Instagram daily by the two groups was statistically significant, it is also of practical importance.

4 Assumptions Check

4.1 Equality of Variance

```
var.test(data$Time ~ data$Group)
```

```
##
##  F test to compare two variances
##
## data:  data$Time by data$Group
## F = 0.15684, num df = 78, denom df = 75, p-value = 2.367e-14
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.09973962 0.24611638
## sample estimates:
## ratio of variances
##          0.1568351
```

The results suggest that the equality of variances could not be rejected (F(78, 72) = 1.06, p = .79), thus it is appropriate to use two-sample t-test to analyse our data.

4.2 Independence

Independence can be assumed in the design, as the two groups of participants were randomly selected .

4.3 Normality

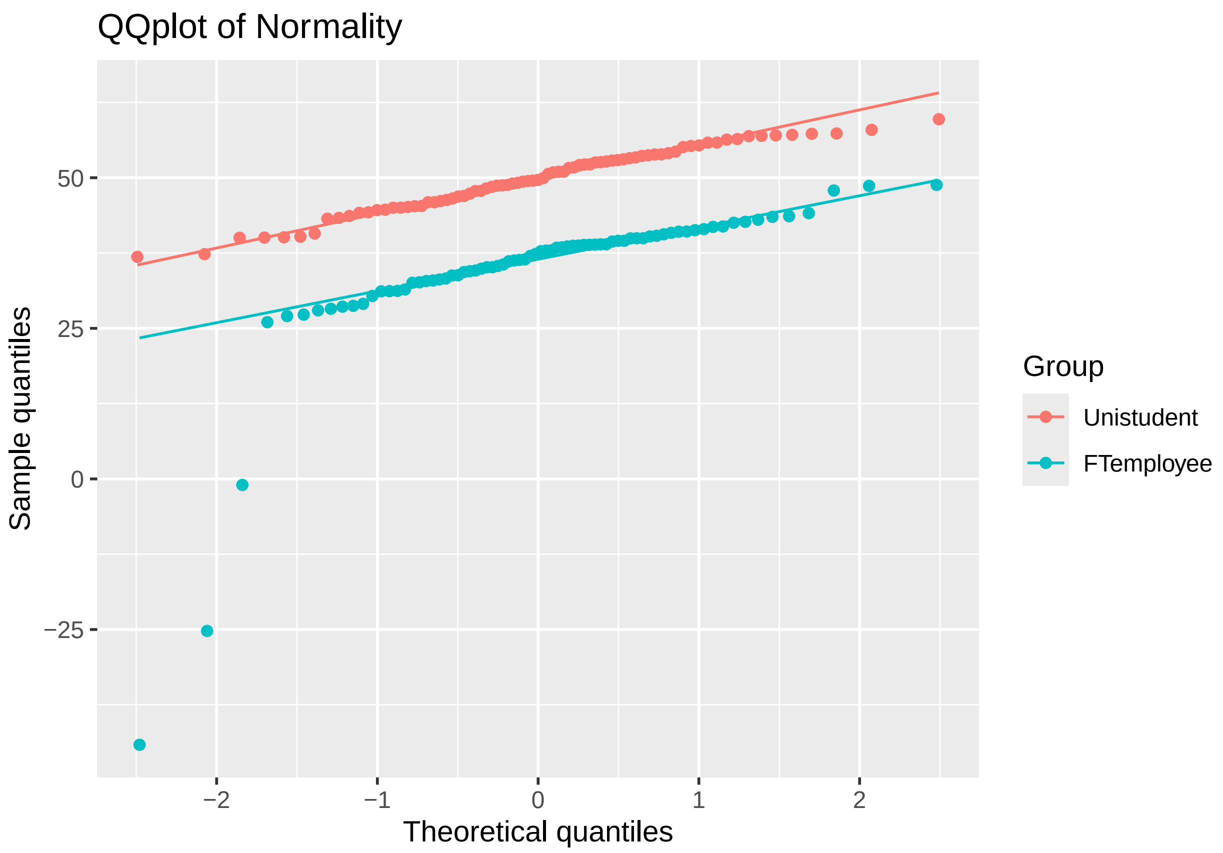
Firstly we check skewness statistics. The absolute value of the skewness should be smaller than 1.

```
skewness <- data %>%
  group_by(Group) %>%
  summarise(Skew = skew(Time))
skewness    # Skewness is OK as it's < 1 in their absolute values
```

```
## # A tibble: 2 x 2
##   Group        Skew
##   <fct>       <dbl>
## 1 Unistudent -0.345
## 2 FTemployee -4.07
```

Next we visualise the normality of the data using qq plot.

```
plt_qq <- ggplot(data, aes(colour = Group, sample = Time)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles",
       title = "QQplot of Normality")
plt_qq
```



4.3.1 Think Point

The plots look okay but not the perfect. How can we be certain whether the data are normally distributed?

Hint: we can test this use the shapiro.test()

4.4 Shapiro.test

First we subset the data by group.

```
Unistudent <- data %>%
filter(Group == "Unistudent")

FTemployee <- data %>%
filter(Group == "FTemployee")
```

Next we run a Shapiro-Wilk test for each group. Note that the H_0 of the test is that the data is normally distributed. Thus, if you find a large p-value, that means the results fail to reject the H_0, in other words, supporting the assumption that the data is normally distributed.

```
shapiroUnistudent <- shapiro.test(Unistudent$Time)
shapiroUnistudent #Shapiro-Wilk: W = 0.98, p = .12 (> alpha), fail to reject H0
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  Unistudent$Time
## W = 0.97505, p-value = 0.1239
```

```
shapiroFTemployee  <- shapiro.test(FTemployee$Time)
shapiroFTemployee #Shapiro-Wilk: W = 0.98, p = .32 (> alpha),fail to reject H0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  FTemployee$Time
## W = 0.54877, p-value = 6.374e-14
```

we can report the following.

At the 5% significance level, we performed a Shapiro-Wilk test against the null hypothesis of normality of the data from the two groups, University students and Full-time employees. For each group, the sample data did not provide sufficient evidence to reject the null hypothesis of normality in the population. Therefore, we conclude that the results showed that the data were approximately normally distributed for both University Student group (W = 0.96, p = .53) and Full-time Employee group (W = 0.98, p = .32).

5 Exercise: Paired t-test

5.1 preparation

```
vocabdata <- read_csv("vocabdata.csv")
```

```
## Rows: 240 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): participant, exam_time
## dbl (1): vocab_test_score
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
dim(vocabdata)
```

```
## [1] 240   3
```

```
glimpse(vocabdata)
```

```
## Rows: 240
## Columns: 3
## $ participant      <chr> "ID01", "ID02", "ID03", "ID04", "ID05", "ID06", "ID07~
## $ exam_time        <chr> "after", "after", "after", "after", "after", "after",~
## $ vocab_test_score <dbl> 58, 86, 93, 44, 45, 47, 51, 52, 50, 64, 61, 71, 67, 5~
```

```r
summary(vocabdata)
```

```
##  participant         exam_time         vocab_test_score
##  Length:240         Length:240         Min.   :39.00   
##  Class :character   Class :character   1st Qu.:58.00   
##  Mode  :character   Mode  :character   Median :73.00   
##                                        Mean   :69.31   
##                                        3rd Qu.:80.00   
##                                        Max.   :93.00   
```

```r
str(vocabdata)
```

```
## spc_tbl_ [240 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ participant     : chr [1:240] "ID01" "ID02" "ID03" "ID04" ...
##  $ exam_time       : chr [1:240] "after" "after" "after" "after" ...
##  $ vocab_test_score: num [1:240] 58 86 93 44 45 47 51 52 50 64 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   participant = col_character(),
##   ..   exam_time = col_character(),
##   ..   vocab_test_score = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
vocabdata$exam_time <-as.factor(vocabdata$exam_time)
levels(vocabdata$exam_time)
```

```
## [1] "after"  "before"
```

```r
table(is.na(vocabdata))
```

```
## 
## FALSE 
##   720
```

```r
vocabdata <- vocabdata %>% drop_na(vocab_test_score)

# check the data
tbl_stats_vocabdata <- vocabdata %>%
  group_by(exam_time) %>%
  summarise(n = n(),
            M = mean(vocab_test_score),
            SD = sd(vocab_test_score),
            Min = min(vocab_test_score),
            Max = max(vocab_test_score))
tbl_stats_vocabdata
```

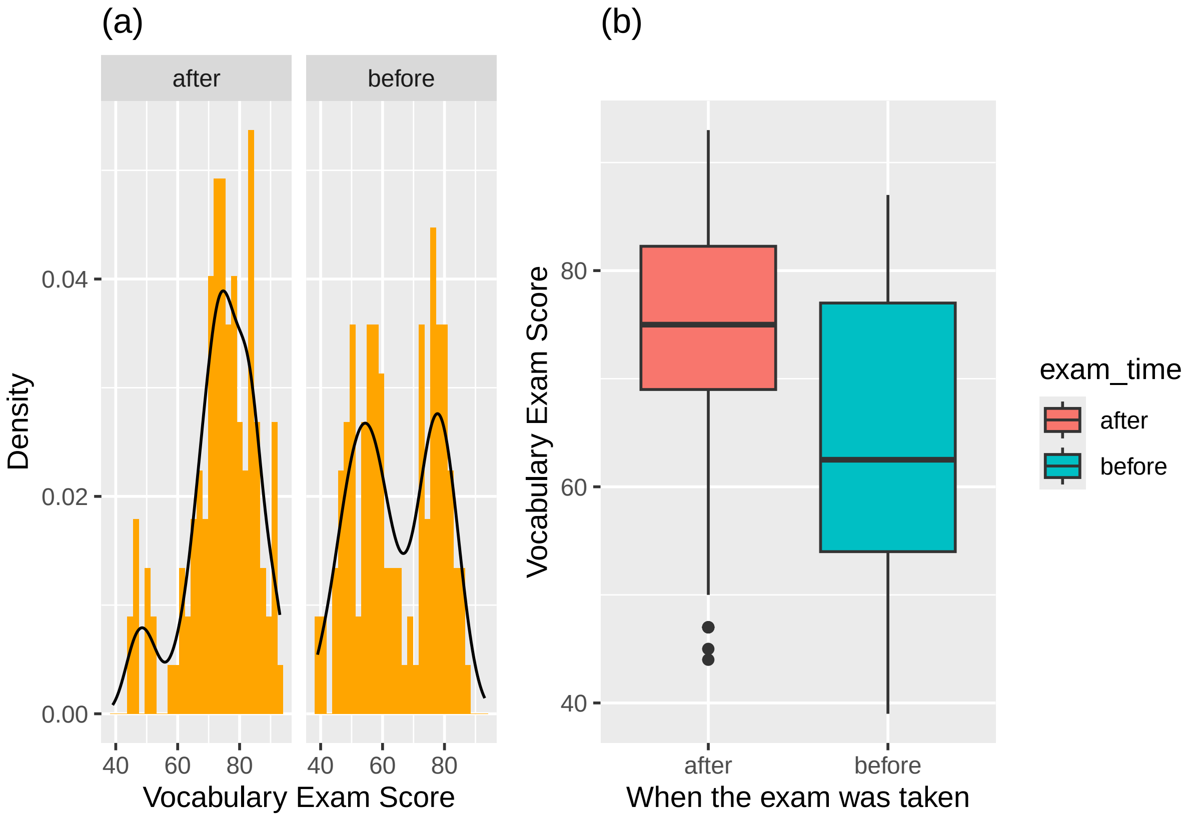
```
## # A tibble: 2 x 6
##   exam_time     n     M    SD   Min   Max
##   <fct>     <int> <dbl> <dbl> <dbl> <dbl>
## 1 after       120  74.1  11.3    44    93
## 2 before      120  64.5  13.2    39    87
```

5.2 Visualisation

```
plt_hist_vocabdata <- ggplot(vocabdata, aes(x = vocab_test_score,
                                            after_stat(density))) +
  geom_histogram(fill = "orange") +
  geom_density() +
  facet_wrap(~exam_time) +
  labs(x = "Vocabulary Exam Score",
       y = "Density",
       title = "(a)")
plt_hist_vocabdata
```

```
plt_box_vocabdata <- vocabdata %>%
  ggplot(aes(x=exam_time, y = vocab_test_score, fill = exam_time)) +
  geom_boxplot() +
  labs(x = "When the exam was taken",
    y = "Vocabulary Exam Score",
       title = "(b)") +
  theme()
plt_box_vocabdata
```

```
plt_hist_vocabdata | plt_box_vocabdata
```



5.3 Paired t-test

```
pairedt_test <- t.test(vocabdata$vocab_test_score ~ vocabdata$exam_time,
                 paired = TRUE,
                 mu = 0,
                 alternative = "greater",
                 )

pairedt_test
```

```
##
##  Paired t-test
##
## data:  vocabdata$vocab_test_score by vocabdata$exam_time
## t = 9.7623, df = 119, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  8.011314      Inf
## sample estimates:
## mean difference
##            9.65
```

5.3.1 Confidence Intervals

```
pairedt_test_CI <- t.test(vocabdata$vocab_test_score ~ vocabdata$exam_time,
                    paired = TRUE,
                    mu = 0,
                    alternative = "two.sided"
                    )
pairedt_test_CI
```

```
##
##  Paired t-test
##
## data:  vocabdata$vocab_test_score by vocabdata$exam_time
## t = 9.7623, df = 119, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   7.692682 11.607318
## sample estimates:
## mean difference
##            9.65
```

5.3.2 Effect Size

```
D_vocabdata <- cohens_d(vocabdata$vocab_test_score ~ vocabdata$exam_time,
         mu = 0,
         alternative = "two.sided",
         var.equal = TRUE)

D_vocabdata
```

```
## Cohen’s d |       95% CI
## ------------------------
## 0.79      | [0.52, 1.05]
##
## - Estimated using pooled SD.
```

5.3.3 Assumptions Check

```


skewness_pairedt <- vocabdata %>%
  group_by(exam_time) %>%
  summarise(Skew = skew(vocab_test_score))
skewness_pairedt
```

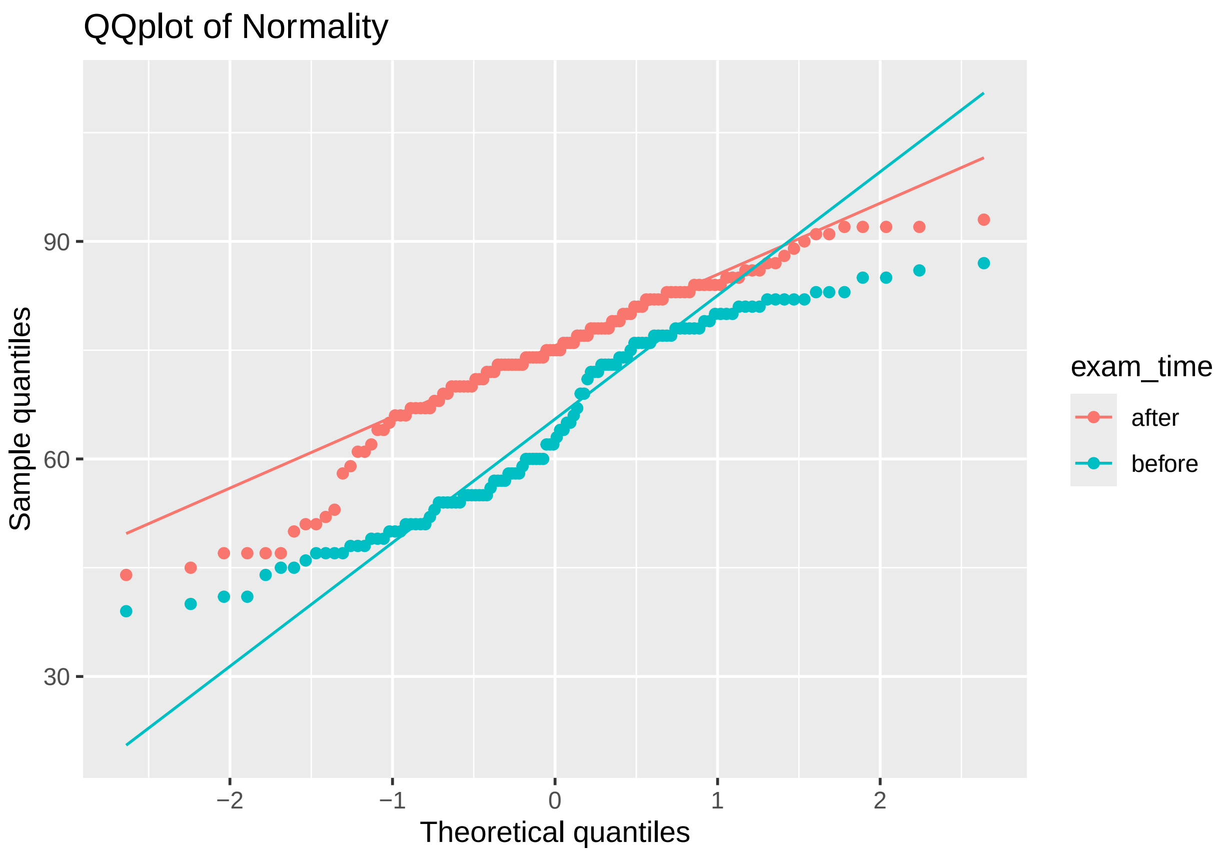
5.3.3.1 Skewness

```
## # A tibble: 2 x 2
##   exam_time    Skew
##   <fct>       <dbl>
## 1 after     -0.791
## 2 before    -0.0365
```

5.3.3.2 Normality

Fist visualise the normality via qqplot.

```
plt_qq_vocabdata <- ggplot(vocabdata, aes(colour = exam_time,
                                          sample = vocab_test_score)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles",
       title = "QQplot of Normality")
plt_qq_vocabdata
```



```
beforedata <- vocabdata %>%
filter(exam_time == "before")

afterdata <- vocabdata %>%
filter(exam_time == "after")
```

```
shapirobeforedata <- shapiro.test(beforedata$vocab_test_score)
shapirobeforedata
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beforedata$vocab_test_score
## W = 0.93567, p-value = 2.21e-05
```

```
shapiroafterdata <- shapiro.test(afterdata$vocab_test_score)
shapiroafterdata
```

```
##
##  Shapiro-Wilk normality test
##
## data:  afterdata$vocab_test_score
## W = 0.94147, p-value = 5.388e-05
```

5.3.3.3 Think Point The data do not meet the t-test assumption of normally distributed data. Can you still use paired t-test?