project_covid_data

Danny Han

2025-11-08

Install Sampling and Survey Package

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages(c("tidyverse", "psych", "effectsize", "knitr"))
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
## package 'psych' successfully unpacked and MD5 sums checked
## package 'effectsize' successfully unpacked and MD5 sums checked
## package 'knitr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\toadcode\AppData\Local\Temp\RtmpQxuVmS\downloaded_packages
```

Load the Required Libraries

```
library(tidyverse)
library(psych)
library(effectsize)
library(knitr)
```

Import the Data File into R

```
# Read/Import the data file in R
student_mental_health <- read.csv("student_mental_health.csv")[1:1193, ]
attach(student_mental_health)
```

Introduction

Research Question: "Did students' perceived stress levels (PSS) significantly change from before to after the COVID-19 outbreak?"

This analysis explores whether university students experienced higher perceived stress levels after the onset of COVID-19. The Perceived Stress Scale (PSS) is used to measure stress before and after the pandemic. By comparing each participant's pre- and post-COVID stress scores, we can assess if a statistically significant increase occurred.

Select & Prepare PSS Variables

We extract the 10 pre-COVID and 10 post-COVID PSS items to focus only on the variables relevant to this analysis.

```
# Select only the relevant columns
pss_vars_pre  <- paste0('Pre_PSS_', 1:10)
pss_vars_post <- paste0('Post_PSS_', 1:10)

pss_data <- student_mental_health %>%
  select(all_of(c(pss_vars_pre, pss_vars_post)))
```

Reverse-Code Items 4, 5, 7, 8

In the PSS, some items are phrased positively (e.g., "I felt confident about handling personal problems") and must be reverse-coded. We subtract these from the maximum value (4) to ensure that higher scores consistently represent higher stress.

```
reverse_items <- c(4, 5, 7, 8)

for (i in reverse_items) {
    pss_data[[paste0('Pre_PSS_', i)]] <- 4 - pss_data[[paste0('Pre_PSS_', i)]]
    pss_data[[paste0('Post_PSS_', i)]] <- 4 - pss_data[[paste0('Post_PSS_', i)]]
}
```

Compute Total Scores

Next, we sum across the 10 PSS items for each participant to get their overall stress score before and after COVID. Higher totals indicate greater perceived stress.

```
pss_data <- pss_data %>%
mutate(
    pre_total = rowSums(select(., all_of(pss_vars_pre)), na.rm = TRUE),
    post_total = rowSums(select(., all_of(pss_vars_post)), na.rm = TRUE)
)
```

Descriptive Statistics

We compute summary statistics (mean, standard deviation, range) to understand the central tendency and variability of stress scores.

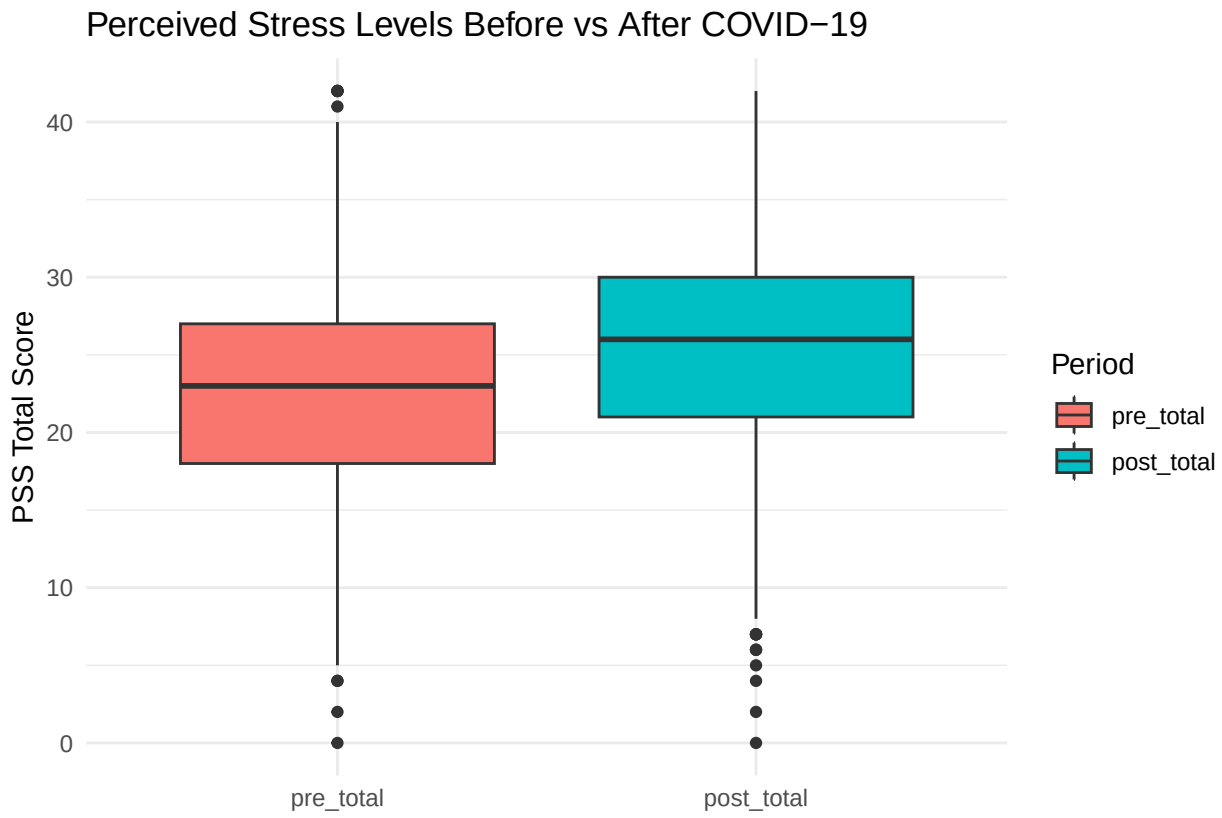
```
describe(pss_data[, c('pre_total', 'post_total')])
```

```
##            vars    n  mean   sd median trimmed  mad min max range  skew
## pre_total     1 1193 22.51 6.79     23   22.50 5.93   0  42    42  0.01
## post_total    2 1193 25.38 6.98     26   25.44 7.41   0  42    42 -0.14
##            kurtosis  se
## pre_total     -0.03 0.2
## post_total     0.02 0.2
```

To visually assess how stress levels shifted, we create a boxplot comparing pre- and post-COVID distributions. This helps spot trends such as increased median or variability in scores.

```
# Boxplot comparison (pre shown on the left)
pss_data %>%
```

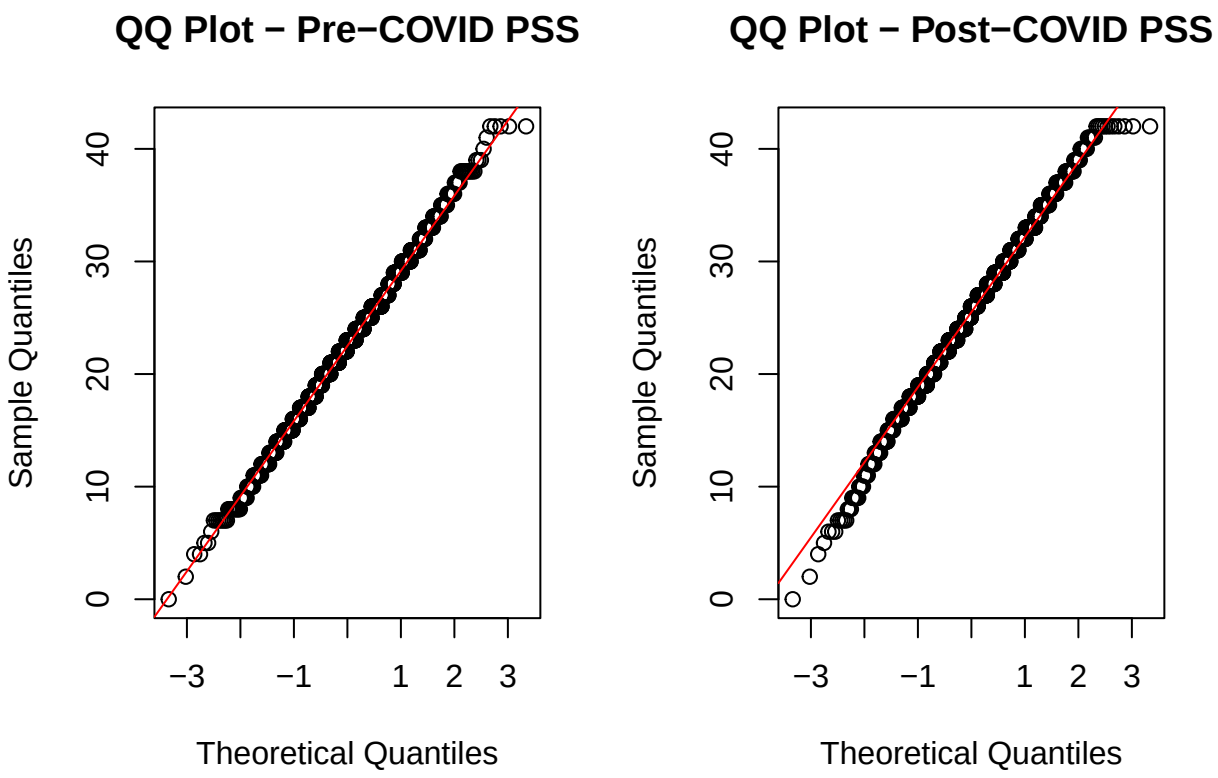
```r
  pivot_longer(cols = c(pre_total, post_total),
               names_to = 'Period', values_to = 'Score') %>%
  mutate(Period = factor(Period, levels = c('pre_total', 'post_total'))) %>%  # enforce order
  ggplot(aes(x = Period, y = Score, fill = Period)) +
  geom_boxplot() +
  labs(
    title = 'Perceived Stress Levels Before vs After COVID-19',
    y = 'PSS Total Score',
    x = ''
  ) +
  theme_minimal()
```




```r
# QQ plots for normality check
par(mfrow = c(1, 2))  # two plots side by side

qqnorm(pss_data$pre_total, main = 'QQ Plot - Pre-COVID PSS')
qqline(pss_data$pre_total, col = 'red')

qqnorm(pss_data$post_total, main = 'QQ Plot - Post-COVID PSS')
qqline(pss_data$post_total, col = 'red')
```




```
par(mfrow = c(1, 1))  # reset layout
```

Paired t-Test: Assumptions & Conditions

We look to use the paired t-test to determine whether the average stress score significantly changed from pre- to post-COVID for the same participants. This is appropriate because the data come from repeated measurements of the same individuals.

Before we proceed, we check that certain assumptions are met that validates the use of the paired t-test.

1. The data are quantitative. They are also paired, since one individual answers both sets of PSS questions with respect to the COVID period.
2. Once again, since participants answered the questions individually, we have the independence assumption for the differences. Also, we can safely grant the assumption that the sample size of $n = 1193$ is less than 10% of the entire population of all students who were enrolled in post-secondary studies before and after the COVID outbreak.
3. The above boxplots shows rough symetry around the median for both samples. The QQ-plot for each sample also justfies the approximation of normality.

Paired t-Test

We now proceed with the paired t-test.

```
t_test_result <- t.test(pss_data$pre_total, pss_data$post_total, paired = TRUE)
t_test_result
```

```
##
##  Paired t-test
##
## data:  pss_data$pre_total and pss_data$post_total
## t = -15.13, df = 1192, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.241302 -2.497173
## sample estimates:
## mean difference
##       -2.869237
```

Effect Size

Statistical significance alone doesn't show how large the change was. Cohen's d provides a standardized measure of the effect size, indicating the magnitude of change in stress.

```
cohens_d(pss_data$post_total, pss_data$pre_total, paired = TRUE)
```

```
## Cohen's d |       95% CI
## ------------------------
## 0.44      | [0.38, 0.50]
```

Results

```
# Compute descriptive statistics
desc_stats <- describe(pss_data[, c('pre_total', 'post_total')])
# Round all numeric columns to 4 decimals
desc_summary <- desc_stats %>%
  select(vars, n, mean, sd, median, min, max, range, skew, kurtosis) %>%
  mutate(across(where(is.numeric), ~ round(., 4)))

# Display formatted table
kable(desc_summary,
      caption = "Descriptive Statistics for Pre- and Post-COVID Perceived Stress Scores",
      align = 'c')
```

Table 1: Descriptive Statistics for Pre- and Post-COVID Perceived Stress Scores

	vars	n	mean	sd	median	min	max	range	skew	kurtosis
pre_total	1	1193	22.5063	6.7924	23	0	42	42	0.0148	-0.0308
post_total	2	1193	25.3755	6.9850	26	0	42	42	-0.1448	0.0243

Statistical Significance:

The paired t-test produced: t = -15.13, df = 1192, p < 2.2e-16

The negative t-value indicates that the post-COVID mean stress score (25.38) is higher than the pre-COVID mean stress score (22.51).

The p-value (< 0.001) is far below the 0.05 threshold, so we reject the null hypothesis that the mean stress level remained unchanged.

This provides strong statistical evidence that students' perceived stress increased after COVID-19 began

Practical Significance:

Cohen's d = 0.44 (95% CI [0.38, 0.50]).

This represents a medium effect size, meaning that while the difference is not massive, it is practically meaningful and consistent across participants.

On average, the increase in perceived stress was about 0.44 standard deviations higher post-COVID, which is a noticeable shift for a psychological measure.

Conclusion

Students reported significantly higher stress levels after COVID-19. Given the medium effect size and consistent upward shift across measures, this suggests the pandemic had a real and measurable psychological impact on university students.

In conclusion: The data strongly support that the COVID-19 outbreak was associated with an increase in perceived stress among university students, both statistically and practically significant.