Anxiety Level Analysis

Haochong(Rogers) Yang

05/08/2022

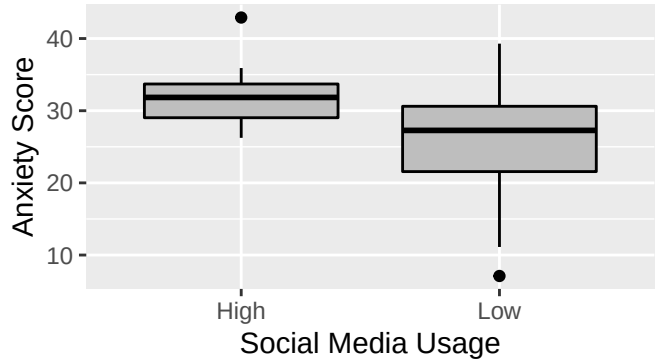
There have been many questions regarding whether or not usage of social media increases anxiety levels. A study was conducted to examine the relationship between social media usage and student anxiety. The following scores are measures of anxiety levels of students on a Monday. Higher scores indicate higher anxiety; and, if a student used social media more than 1 hour per day then their usage was categorized as "High".

```
social_media_usage <- c(rep("Low", 30), rep("High", 16));
anxiety_score <- c(24.64, 39.29, 16.32, 32.83, 28.02,
                   33.31, 20.60, 21.13, 26.69, 28.90,
                   26.43, 24.23, 7.10,  32.86, 21.06,
                   28.89, 28.71, 31.73, 30.02, 21.96,
                   25.49, 38.81, 27.85, 30.29, 30.72,
                   21.43, 22.24, 11.12, 30.86, 19.92,
                   33.57, 34.09, 27.63, 31.26,
                   35.91, 26.68, 29.49, 35.32,
                   26.24, 32.34, 31.34, 33.53,
                   27.62, 42.91, 30.20, 32.54)

anxiety_data <- tibble(social_media_usage, anxiety_score)
glimpse(anxiety_data)
```

```
## Rows: 46
## Columns: 2
## $ social_media_usage <chr> "Low", "Low", "Low", "Low", "Low", "Low", "Low", "L~
## $ anxiety_score      <dbl> 24.64, 39.29, 16.32, 32.83, 28.02, 33.31, 20.60, 21~
```

I constructed boxplots of `anxiety_score` for the two levels of social media usage.

```
anxiety_boxplot <- anxiety_data %>%
  ggplot(aes(x = social_media_usage, y = anxiety_score)) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = "Social Media Usage", y = "Anxiety Score")
anxiety_boxplot
```



Students who use social media more frequently tend to be more anxious than students that use less social media. The distribution for high social media usage is left-skewed, with a median around 33, and an inter-quartile range of 5. The distribution for low social media usage is roughly symmetric, with a median around 27, and an inter-quartile range of 10. The median of anxiety score distribution of low social media usage is lower than that of high usage, and the inter-quartile range is also larger than that of high usage.

Do these data support the claim that the median anxiety level is different for those who use social media in high frequency compared to those who use social media in lower frequency?

The null-hypothesis(H_0) is that the median anxiety level is the same for those who use social media in high frequency ($median_{high}$) compared to those who use social media in lower frequency($median_{low}$).

$$H_0 : median_{high} = median_{low}$$

The alternate hypothesis(H_1) is that the median anxiety level is different for those who use social media in high frequency($median_{high}$) than those who use social media in lower frequency($median_{low}$).

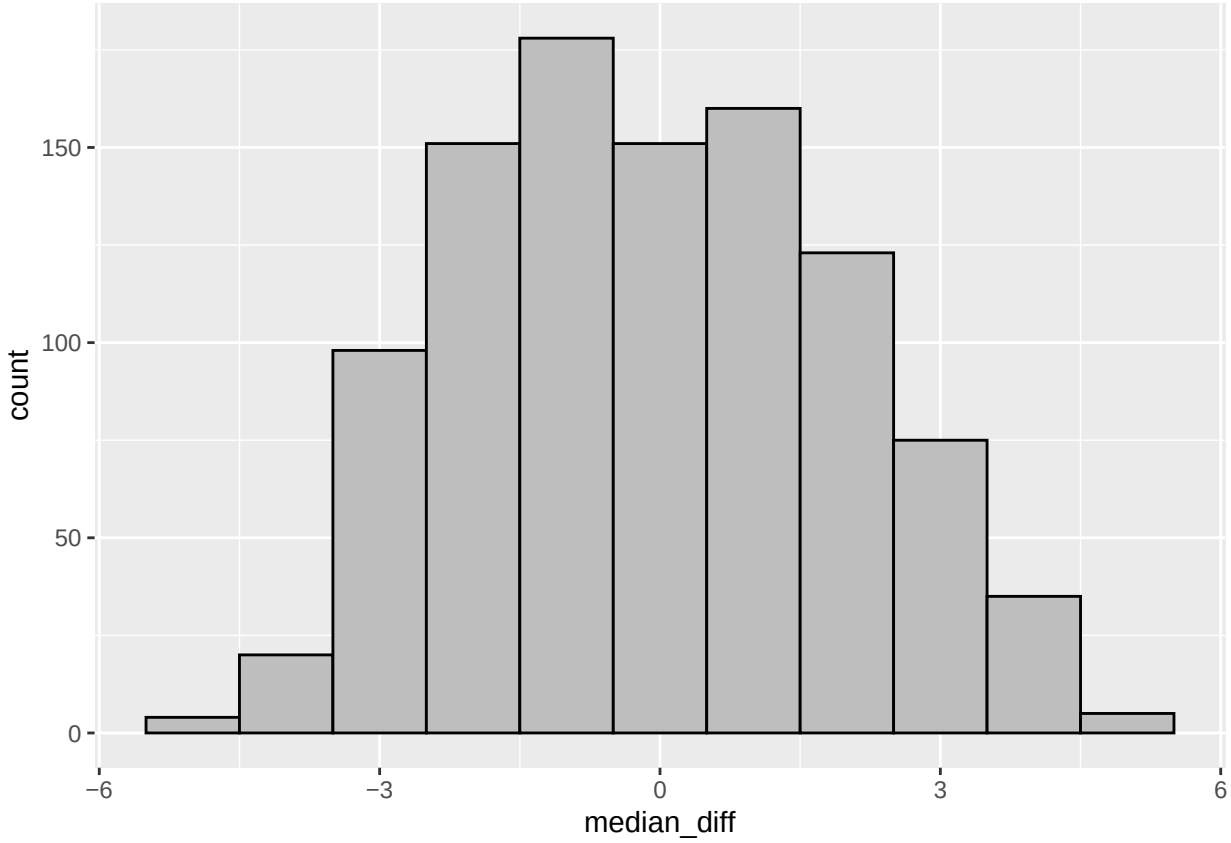
$$H_1 : median_{high} \neq median_{low}$$

```
# Note: including the .groups="drop" option in summarise() will suppress a friendly
# warning R prints otherwise "`summarise()` ungrouping output (override with
#`.groups` argument)".
  # Including the .groups="drop" option is optional, but you should include it if you
#don't want to see that warning.
test_stat <- anxiety_data %>% group_by(social_media_usage) %>%
  summarise(medians = median(anxiety_score), .groups="drop") %>%
  summarise(value = diff(medians))
test_stat <- as.numeric(test_stat)
test_stat
```

```
## [1] -4.57
```

```
set.seed(523)
repetitions <- 1000;
simulated_values <- rep(NA, repetitions)

for(i in 1:repetitions){
  simdata <- anxiety_data %>% mutate(social_media_usage = sample(social_media_usage))

  sim_value <- simdata %>% group_by(social_media_usage) %>%
```

```
    summarise(medians = median(anxiety_score), .groups="drop") %>%
    summarise(value = diff(medians))

  simulated_values[i] <- as.numeric(sim_value)
}

sim <- tibble(median_diff = simulated_values)

sim %>% ggplot(aes(x=median_diff)) + geom_histogram(binwidth=1, color="black", fill="gray")
```




```
# Calculate p-value
num_more_extreme <- sim %>% filter(abs(median_diff) >= abs(test_stat)) %>% summarise(n())

p_value <- as.numeric(num_more_extreme / repetitions)
p_value
```

```
## [1] 0.009
```

The for loop creates a value that stores the difference in median anxiety scores. The for loop first shuffles the value of media usage and pairs it with a random value from anxiety score. Then it groups all the observations by variable "social_media_usage" and finds the median of "anxiety_score" in two groups "low" and "high" social media usage. And then it calculates the different between the two medians. Lastly, it stores the value of difference between two medians into the object we created out of the for loop. The for loops repeated 1000 times and created 1000 values of the difference in medians.

Conclusion: Since the p-value is 0.009, which is larger than 0.001 and smaller than 0.010, so there's strong evidence against the null hypothesis(H_0). This can be interpreted as, if we assume that there is no difference in median anxiety score between high usage group and low usage group, the chance for us to observe a value that is as extreme as -4.57 is very low. Using our hypothesis the test statistic is very unlikely to happen, which rejected our hypothesis. This means that the median anxiety level is different for those who use social media in high frequency($median_{high}$) compared to those who use social media in lower frequency($median_{low}$).