Final Project

Liying Lu

2020-04-08

Question: Is there an earnings gap between male and female college graduates 6 years after graduation, and does the gap widen or shrink 4 years later?

Preprocessing the dataset

- A reduced dataset `college_reduced` is created to store only the columns which we will use in this final project.

```
college_reduced <- college %>%
  select(
    MN_EARN_WNE_MALE0_P10,MN_EARN_WNE_MALE1_P10,MN_EARN_WNE_MALE0_P6,
    MN_EARN_WNE_MALE1_P6,COSTT4_A,ENRL_ORIG_YR8_RT,REGION,SATVR25,
    SATVR75,SATMT25,SATMT75,SATWR25,SATWR75,SATVRMID,SATMTMID,SATWRMID,
    ACTCM25,ACTCM75,ACTEN25,ACTEN75,ACTMT25,ACTMT75,ACTWR25,ACTWR75,
    ACTCMMID,ACTENMID,ACTMTMID,ACTWRMID,INSTNM,TUITIONFEE_IN,TUITIONFEE_OUT
)
```

- The columns are renamed to have more readable names.

```
college_renamed <- college_reduced %>%
  rename(
        mean_earning_female_10yrs = MN_EARN_WNE_MALE0_P10,
        mean_earning_male_10yrs = MN_EARN_WNE_MALE1_P10,
        mean_earning_female_6yrs = MN_EARN_WNE_MALE0_P6,
        mean_earning_male_6yrs = MN_EARN_WNE_MALE1_P6,
        Institution = INSTNM,
        In_state_tuition = TUITIONFEE_IN,
        Out_state_tuition = TUITIONFEE_OUT,
        tuition_yearly = COSTT4_A,
        Enrolled_8_years = ENRL_ORIG_YR8_RT
  )
```

Prepare the data set for analysis:

- The columns needed for this questions are extracted into a new dataset.

```
mean_earning <- college_renamed %>%
  select(mean_earning_female_10yrs, mean_earning_male_10yrs,
         mean_earning_female_6yrs, mean_earning_male_6yrs)
```

- An additional column `gender` is created to label the observations with either "female" or "male".

```
# create the female columns
female <- mean_earning %>%
  select(mean_earning_female_10yrs, mean_earning_female_6yrs) %>%
  mutate(gender = "female")
# create the male columns
male <- mean_earning %>%
  select(mean_earning_male_10yrs, mean_earning_male_6yrs) %>%
  mutate(gender = "male")
```

- The `mean_earning_female_10yrs` and `mean_earning_male_10yrs` columns in `female` and `male` are renamed to `mean_earning_10yrs`. The same will be done to the data for mean-earning after 6 years, renaming them to `mean_earning_6yrs`.

```
# rename the female dataset
female_renamed <- female %>%
  rename(mean_earning_10yrs = mean_earning_female_10yrs,
         mean_earning_6yrs = mean_earning_female_6yrs)
# rename the male dataset
male_renamed <- male %>%
  rename(mean_earning_10yrs = mean_earning_male_10yrs,
         mean_earning_6yrs = mean_earning_male_6yrs)
```

- The columns `female_renamed` and `male_renamed` are then row combinded to `mean_earning_clean` for the data analysis in the next section.

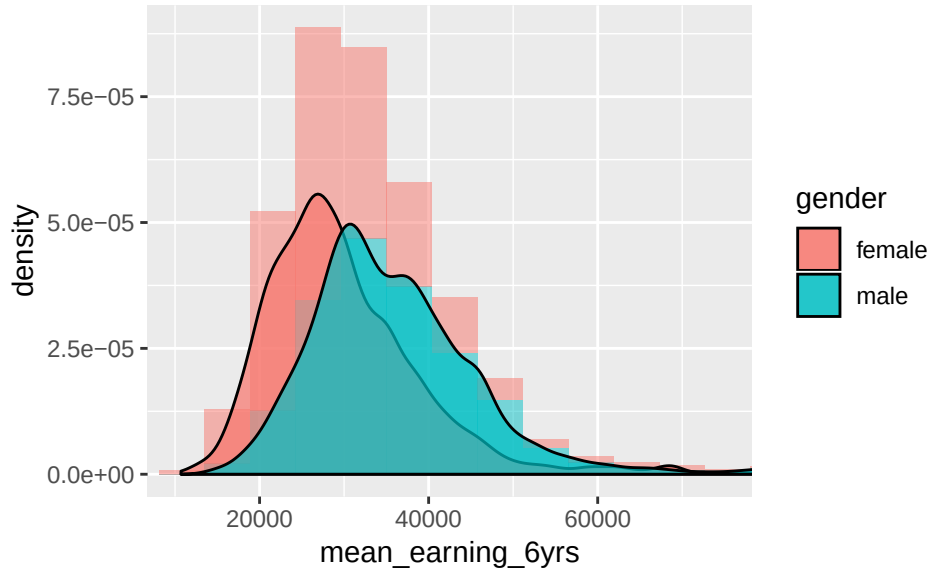
```
mean_earning_clean <- rbind(female_renamed, male_renamed)
```

Visualization:

The first step is to visualize the both the mean-earning after 6 years and 10 years in histograms and probability mass functions (PMF). The scale of the x-axis is adjusted so that the graph focuses on the majority of the data.

- A histogram and the PMF of the mean-earning after 6 years are created.

```
ggplot(mean_earning_clean,
       aes(x = mean_earning_6yrs,
           y = ..density..,
           fill = gender)) +
  geom_histogram(alpha = 0.5) +
  geom_density(alpha = 0.7) +
  # adjust the x-axis
  coord_cartesian(xlim =combine(10000, 75000))
```



- The mean-earning after 6 years for female is right-skewed, centers at about $28,000, and ranges from about $10,000 to $80,000.
- The mean-earning after 6 years for male is right-skewed, centers at about $30,000, and ranges from about $15,000 to $80,000.
- A histogram and the PMF of the mean-earning after 10 years are created.

```
ggplot(mean_earning_clean,
       aes(x = mean_earning_10yrs,
           y = ..density..,
           fill = gender)) +
  geom_histogram(alpha = 0.5) +
  geom_density(alpha = 0.7) +
  # adjust the x-axis
  coord_cartesian(xlim =combine(10000, 100000))
```

- The mean-earning after 6 years for female is right-skewed, centers at about $30,000, and ranges from about $12,500 to $100,000.
- The mean-earning after 6 years for male is right-skewed as well, centers at about $40,000, and ranges from about $18,000 to $100,000.

Based on the visualization, there is an income gap between female and male college graduates both 6 years and 10 years after graduation. The median income for male graduates is higher than that of the female graduates, and the gap increases after 10 years.

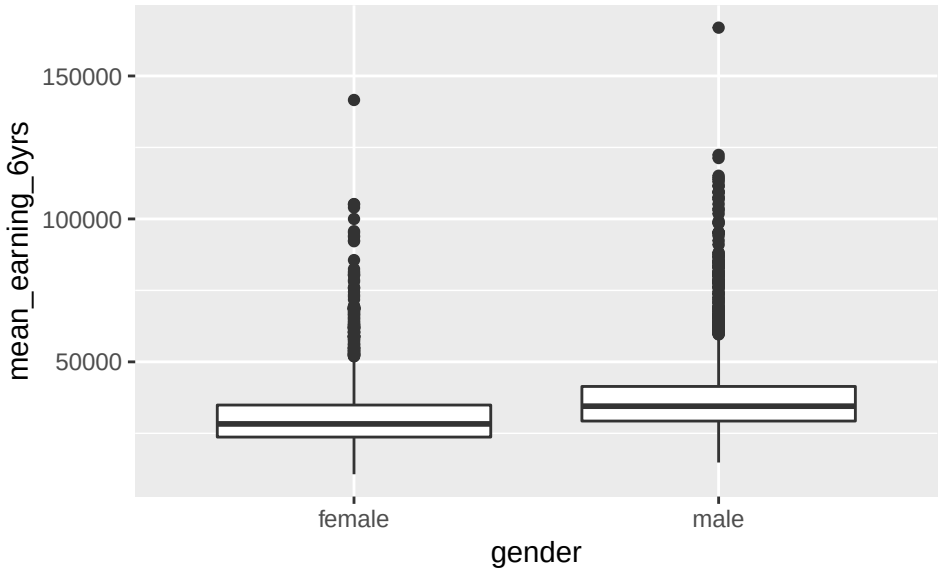
Outliers

There may be outliers in the data which could affect the reliability of statistical analysis. Thus, boxplots are used here to identify the outliers.

Boxplots:

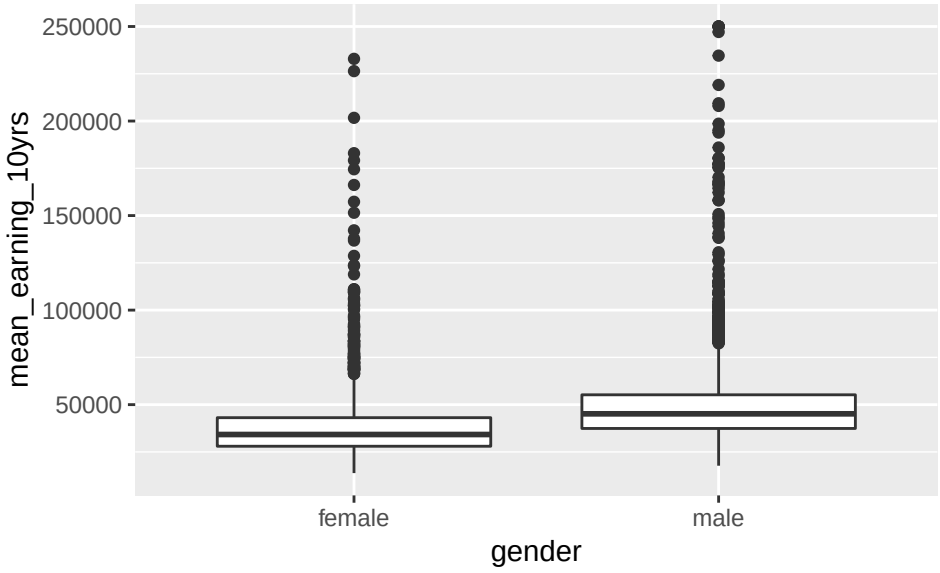
* Boxplot for the mean-earning after 6 years of graduation for college graduates is created.

```
ggplot(mean_earning_clean) +
  geom_boxplot(aes(x = gender, y = mean_earning_6yrs))
```



- Boxplot for the mean-earning after 10 years of graduation for college graduates is created.

```
ggplot(mean_earning_clean) +
  geom_boxplot(aes(x = gender, y = mean_earning_10yrs))
```



- Based on the above graphics, there are mumerous extremely high values of income in both the mean-earning after 6 years and after 10 years which affect the fair representation of the summary statistics mean and standard deviation. However, they are unlikely to be data entry errors so they remain in the data sets in the following analysis.
- The median will be used to compare the mean-earning because extreme values do not have significant effect on the median.

Summary statistics:

The summary statistics are calculated to further compare the income between the female and male graduates.

- The summary statistics of the mean-earning after 10 years are calculated.

```
mean_earning_6yrs_summary_stats <- mean_earning_clean %>%
  # filter the NA values in the data
  filter(!is.na(mean_earning_6yrs)) %>%
  # compute the summary statistics of mean_earning_6yrs
  group_by(gender) %>%
  summarize(
    mean = mean(mean_earning_6yrs),
    median = median(mean_earning_6yrs),
    sd = sd(mean_earning_6yrs),
    iqr = IQR(mean_earning_6yrs),
    min = min(mean_earning_6yrs),
    max = max(mean_earning_6yrs)
  )
mean_earning_6yrs_summary_stats
```

gender	mean	median	sd	iqr	min	max
female	30201.77	28300	10158.18	11200	10700	141600
male	36577.77	34500	11994.03	12100	14800	166900

- The summary statistics of the mean-earning after 10 years are calculated.

```
mean_earning_10yrs_summary_stats <- mean_earning_clean %>%
  filter(!is.na(mean_earning_10yrs)) %>%
  group_by(gender) %>%
  summarize(
    mean = mean(mean_earning_10yrs),
    median = median(mean_earning_10yrs),
    sd = sd(mean_earning_10yrs),
    iqr = IQR(mean_earning_10yrs),
    min = min(mean_earning_10yrs),
    max = max(mean_earning_10yrs)
  )
mean_earning_10yrs_summary_stats
```

gender	mean	median	sd	iqr	min	max
female	37290.80	34200	15020.84	15100	13800	232900
male	48755.33	45150	19967.72	17800	17700	250000

- In 6 years of graduation, the median income of male graduates is $6200 higher than that of female graduates.
- In 10 years of graduation, the median income of male graduates is $10950 higher than that of

female graduates.
- The summary statistics suggest that an income gap does exists between female and male graduates. There is an increase in the median income gap of $4750, or 76.6%, between male and female graduates from 6 to 10 years of graduation.

Hypothesis testing:

The previous visualizations and summary statistics all suggest that there is indeed an income gap between the male and female graduates after 6 years and 10 years of graduation. With that in mind, a hypothesis testing will be carried out to further study the question.

Hypothesis testing for the mean-earning of college graduates after 6 years of graduation:

Null hypothesis: The median for the mean-earnings after 6 years of graduation does not differ for the male and female graduates.
Alternative hypothesis: The median for the mean-earnings after 6 years of graduation does differ for the male and female graduates.

- A two-sided hypothesis test will be conducted with a significance level of 0.05. First, the observed difference in median are calculated.

```
# pull the median column
mean_earning_6yrs_medians <- mean_earning_6yrs_summary_stats %>%
  pull(median)
# calculate the observed diff_in_median:
# (median of male mean-earning) - (median of female mean-earning)
diff_median_6yrs_right <- mean_earning_6yrs_medians[2] - mean_earning_6yrs_medians[1]
diff_median_6yrs_left <- -(mean_earning_6yrs_medians[2] - mean_earning_6yrs_medians[1])
```

- A null distribution of the mean-earning after 6 years is simulated.

```
mean_earning_6yrs_medians_null <- mean_earning_clean %>%
  specify(formula = mean_earning_6yrs ~ gender) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in medians", order = combine("male", "female"))
```

- The p-value is calculated with a significance level of 0.05.

```
# p-value on the right side
pvalue_6yrs_right <- mean_earning_6yrs_medians_null %>%
  get_p_value(obs_stat = diff_median_6yrs_right, direction = "right")
# p-value on the left side
pvalue_6yrs_left <- mean_earning_6yrs_medians_null %>%
  get_p_value(obs_stat = diff_median_6yrs_left, direction = "left")
# two-sided p-value
pvalue_6yrs <- pvalue_6yrs_right + pvalue_6yrs_left
pvalue_6yrs
```

p_value

p_value
0

- Since the p-value of 0 is less than the significance level of 0.05, the null hypothesis is rejected in favor of the alternative hypothesis.
- Therefore, there is a difference between the median mean-earning for the male and female college graduates after 6 years of graduation.

Hypothesis testing for the mean-earning for college graduates after 10 years of graduation:

Null hypothesis: There is no difference between the median mean-earning between male and female college graduates after 10 years of graduation.
Alternative hypothesis: There is a difference between the median mean-earning between male and female college graduates after 10 years of graduation.

- The observed difference in medians for the mean-earning after 10 years are calculated for the hypothesis testing.

```
# pull the medians for mean-earning after 10 years.
mean_earning_10yrs_medians <- mean_earning_10yrs_summary_stats %>%
  pull(median)
# calculate the observed diff_in_median:
# (median of male mean-earning) - (median of female mean-earning)
diff_median_10yrs_right <- mean_earning_10yrs_medians[2] - mean_earning_10yrs_medians[1]
diff_median_10yrs_left <- -(mean_earning_10yrs_medians[2] - mean_earning_10yrs_medians[1])
```

- A null distribution of the mean-earning after 6 years is simulated.

```
mean_earning_10yrs_medians_null <- mean_earning_clean %>%
  specify(formula = mean_earning_10yrs ~ gender) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in medians", order = combine("male", "female"))
```

- The p-value is calculated with a significance level of 0.05.

```
# p-value on the right side
pvalue_10yrs_right <- mean_earning_10yrs_medians_null %>%
  get_p_value(obs_stat = diff_median_10yrs_right, direction = "right")
# p-value on the left side
pvalue_10yrs_left <- mean_earning_10yrs_medians_null %>%
  get_p_value(obs_stat = diff_median_10yrs_left, direction = "left")
# two-sided p-value
pvalue_10yrs <- pvalue_10yrs_right + pvalue_10yrs_left
pvalue_10yrs
```

p_value
0

- Since the p-value of 0 is less than the significance level of 0.05, the null hypothesis is rejected in favor of the alternative hypothesis.
- Therefore, there is a difference between the median mean-earning of the male and female college graduates after 10 years of graduation.

Both hypothesis tests suggest that there is a difference between the median mean-earning of the male and female college graduates after 10 years of graduation. This means there exists an income gap where male graduates have a higher median mean-earning than female graduates after 6 and 10 years of graduation.

Effect size:

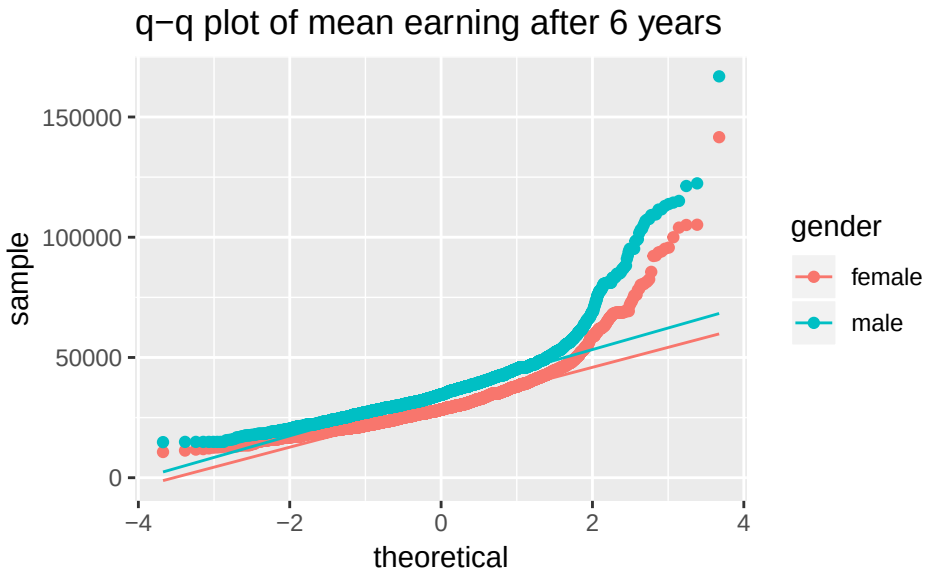
Besides conducting an hypothesis testing on the difference in medians for the mean-earning for the two groups, the effect size between the mean of the mean-earning of the two groups is also measured. Measuring the effect size is an important step to dertermine whether the income gap shrinks or widen from 6 to 10 years of graduation.

- Load the `effsize` library needs to be loaded to carry out this test.

```
library(effsize)
```

- The Q-Q plots show that the data do not have a normal distribution which is the condition of using the Cohen's d test.

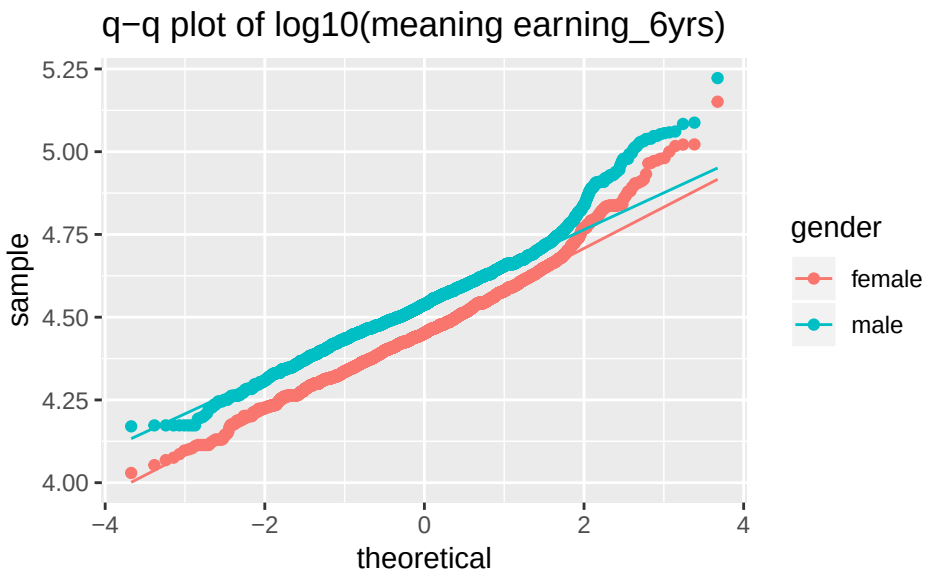
```
# q-q plot for mean_earning_6yrs
ggplot(mean_earning_clean) +
  geom_qq(aes(sample = mean_earning_6yrs, color = gender)) +
  geom_qq_line(aes(sample = mean_earning_6yrs, color = gender)) +
  labs(title = "q-q plot of mean earning after 6 years")
```



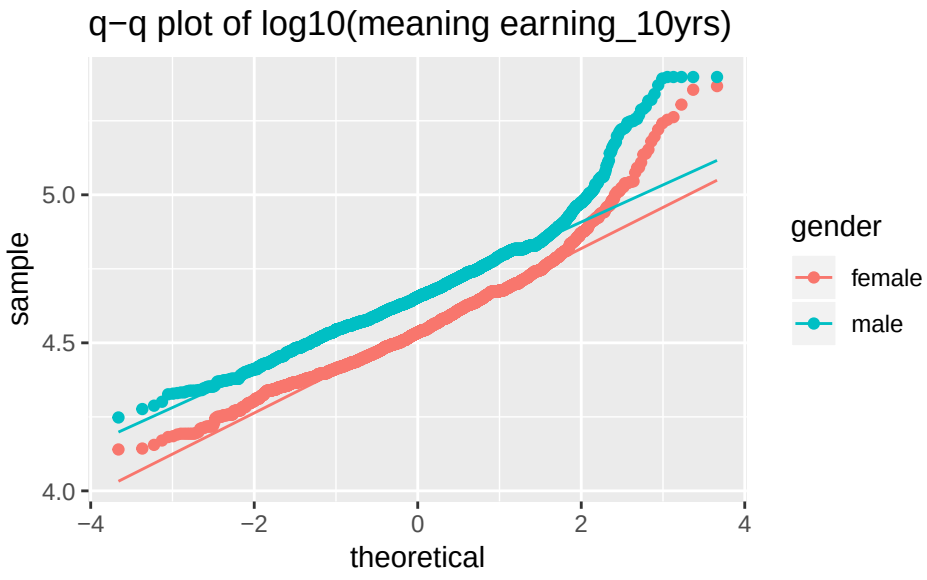
```r
# q-q plot for mean_earning_10yrs
ggplot(mean_earning_clean) +
  geom_qq(aes(sample = mean_earning_10yrs, color = gender)) +
  geom_qq_line(aes(sample = mean_earning_10yrs, color = gender)) +
  labs(title = "q-q plot of mean earning after 10 years ")
```

- The mean-earning columns are transformed to a logarithmic scale of base 10 using `log10()` to satisfy the condition of a normal distribution for a valid Cohen's d test.
- Create the Q-Q plots for the transformed mean-earning columns to ensure their normal distribution.
- The transformed data is not completely normal distributed but has a significant improvement from the original data. Most of the outliers in the transformed data lay in at the higher end, so the Cohen's d calculation is reasonably meaningful.

```r
ggplot(mean_earning_clean) +
  geom_qq(aes(sample = log10(mean_earning_6yrs), color = gender)) +
  geom_qq_line(aes(sample = log10(mean_earning_6yrs), color = gender)) +
  labs(title = "q-q plot of log10(meaning earning_6yrs)")
```




```
ggplot(mean_earning_clean) +
  geom_qq(aes(sample = log10(mean_earning_10yrs), color = gender)) +
  geom_qq_line(aes(sample = log10(mean_earning_10yrs), color = gender)) +
  labs(title = "q-q plot of log10(meaning earning_10yrs)")
```



- The cohen'd statistic is calculated for the transformed data.

```
# calculation for mean-earning after 6 years
cohen.d(formula = log10(mean_earning_6yrs) ~ gender,
        data = mean_earning_clean,
        method = "pooled")
```

```
##
## Cohen's d
##
```

```
## d estimate: -0.6712563 (medium)
## 95 percent confidence interval:
##      lower      upper
## -0.7152834 -0.6272291
```

```
# calculation for mean-earning after 10 years
cohen.d(formula = log10(mean_earning_10yrs) ~ gender,
        data = mean_earning_clean,
        method = "pooled")
```

```
##
## Cohen's d
##
## d estimate: -0.8414413 (large)
## 95 percent confidence interval:
##      lower      upper
## -0.8872989 -0.7955837
```

- The result of the Cohen'd calculation shows that the large effect size of the difference between the mean of the mean-earning after 10 years of graduation is greater than the medium effect size of that after 6 years of graduation for the male and female graduates.
- Therefore, the income gap has widen after 10 years of graduation as compared to 6 years of graduation.

Conclusion:

Based on the hypothesis tests for the difference between the medians of the mean-earning for female and male graduates in 6 and 10 years of graduation, there exists an income gap where the income of male is greater than that of female in both 6 and 10 years of graduation. The effect size of the mean suggests that the income gap has widened after 10 years of graduation as compared to 6 years of graduation.