Horizontal Stratification of Higher Education and Gender Earnings Gap in the Early Labor Market

Inchan Hwang, inchan@yonsei.ac.kr

2 Jan 2025

Basics of Statistics

Statistics is a tool used to extract information from data and support decision-making. Statistics are widely used in the social sciences to analyze patterns, understand relationships, and draw conclusions from data. Fundamentally, traditional statistics aim to summarize data effectively. This session is introduction of two basic statistical methods: (1) how to compare several group's outcome variables (t-test and ANOVA) (2) how to examine the relationship between explanatory variable and outcome variable. (regression and logistic regression)

For instance, consider a class of fifteen students with their scores in Korean and Math. These scores can be grouped by class (Class 1, Class 2, and Class 3) and analyzed to better understand the overall performance of each group.

In this section, we will explore how data can be summarized using these basic statistical measures.

- **Population**: The entire set of data we are interested in.
- **Sample**: A subset of the population selected for statistical analysis.
- **Mean**: A measure representing the central value of the data.
- **Standard Deviation**: A measure indicating how much the data varies from the mean.

```r
# Generating example data
data <- data.frame(
  Student = paste("Student", 1:15), # Names of students
  Class = rep(c("Class 1", "Class 2", "Class 3"), each = 5), # Class assignments
  Korean = c(100, 100, 100, 100, 95, 82, 87, 75, 89, 91, 80, 50, 66, 77, 84), # Korean score
  Math = c(100, 100, 100, 90, 95, 84, 79, 90, 86, 83, 78, 82, 88, 81, 85)    # Math scores
)

# Viewing the data
data
```

```
##       Student   Class Korean Math
## 1   Student 1 Class 1    100  100
## 2   Student 2 Class 1    100  100
## 3   Student 3 Class 1    100  100
## 4   Student 4 Class 1    100   90
## 5   Student 5 Class 1     95   95
## 6   Student 6 Class 2     82   84
```

```
## 7   Student 7 Class 2     87   79
## 8   Student 8 Class 2     75   90
## 9   Student 9 Class 2     89   86
## 10 Student 10 Class 2     91   83
## 11 Student 11 Class 3     80   78
## 12 Student 12 Class 3     50   82
## 13 Student 13 Class 3     66   88
## 14 Student 14 Class 3     77   81
## 15 Student 15 Class 3     84   85
```

```r
# Calculating mean and standard deviation for each subject
mean(data$Korean)
```

```
## [1] 85.06667
```

```r
sd(data$Korean)
```

```
## [1] 14.24513
```

```r
mean(data$Math)
```

```
## [1] 88.06667
```

```r
sd(data$Math)
```

```
## [1] 7.601378
```

Summarizing data using **mean** and **standard deviation** allows us to capture the central tendency and variability within each class for both Korean and Math scores. This provides a clear picture of how students in Class 1, Class 2, and Class 3 perform overall.

t-Test

What is a t-Test?

t-Test is a statistical method used to compare the means of two groups. It helps to determine whether there is a statistically significant difference between the groups. The formula for the independent t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where: - $\bar{X}_1$ and $\bar{X}_2$: Means of the two groups - s_1^2 and s_2^2: Variances of the two groups - n_1 and n_2: Sample sizes of the two groups

Applying t-Test to the Data

Now, we will apply the t-test to compare the Math scores between Class 1 and Class 2:

```r
# Independent t-test comparing Math scores between Class 1 and Class 2
independent_ttest <- t.test(Math ~ Class, data = subset(data, Class %in% c("Class 1", "Class
independent_ttest # Display the test result
```

```
##
##  Two Sample t-test
##
## data:  Math by Class
## t = 4.6763, df = 8, p-value = 0.001589
## alternative hypothesis: true difference in means between group Class 1 and group Class 2
## 95 percent confidence interval:
##   6.386613 18.813387
## sample estimates:
## mean in group Class 1 mean in group Class 2
##                  97.0                  84.4
```

The results of the t-test show whether there is a significant difference in the Math scores between the two classes.

Analysis of Variance (ANOVA)

What is ANOVA?

ANOVA (Analysis of Variance) is used to compare the means of three or more groups. It tests whether there are significant differences among group means. The formula for ANOVA can be expanded as:

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{\frac{SS_{between}}{df_{between}}}{\frac{SS_{within}}{df_{within}}} = \frac{MS_{between}}{MS_{within}}$$

Where: - $SS_{between}$: Sum of squares between groups (measures variability due to group differences) - $df_{between}$: Degrees of freedom for the between-group variability $(k-1)$, where k is the number of groups - SS_{within}: Sum of squares within groups (measures variability within each group) - df_{within}: Degrees of freedom for the within-group variability $(N-k)$, where N is the total number of observations

One-Way ANOVA

Compares the means of groups based on one factor.

```r
# Performing ANOVA on the example data
anova_result <- aov(Math ~ Class, data = data) # Testing mean differences in Math scores bet
summary(anova_result) # Display the ANOVA table
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Class        2  604.9   302.5   17.79 0.000257 ***
## Residuals   12  204.0    17.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression

What is Regression?