# Math 474:

# Probability and Statistics

# Fall 2023

# Extra Credit Project (20 pts)

# Due on 11/30/2023

**Submitted To:** Prof. Tomasz R. Bielecki

**Submitted By:** Manpreet Kaur

**Student ID:** A20551672

**Section:** MATH-474-02

# Summary Report

This is a project based on a real data set "Percentage of the Population 3 Years Old and Over Enrolled in School, by Age, Sex, Race, and Hispanic Origin: October 1947 to 2012" that can be found in the file TableA-2.xls in the Blackboard.

Some Important Points:

- Choose any age group, i.e. a column from D-O, and fix it throughout the project.
- For all evaluations use only the data from 1985-2012.
- Assume that all random variables are normally distributed.
- Everywhere use level $\alpha = 0.05$ for the confidence intervals or hypothesis testing.

**Using the given dataset,**

**Part 1:**

**Find 95% confidence interval of the mean for the following races, both sexes: Hispanic, White and Asian. Interpret the results.**

**Part 2:**

**Consider: the null hypothesis - the population means (true means) of female and male enrolled in school are equal; the alternative is that the means are not equal. Check this hypothesis for 'all races' and for 'Asians.' Interpret the results.**

From here onwards, I have explained how I have solved these above-mentioned two parts using Microsoft Excel and the source excel file is also attached with this doc file.

# Part 1:

To find the 95% confidence interval of the mean for the both sexes for races Hispanic, White and Asian, I have followed the steps as mentioned below:

1. As instructed to choose any age group, I chose **'5 and 6 years'** to perform computations for the project.
2. From the main dataset present in the TableA-2.xls file, segregated the data for the races - Hispanic, White and Asian for the same age group which further consisted of data recorded for both sexes only.
3. Now, calculated the **mean** of the both sexes data for the three races by using the excel function **AVERAGE()** which resulted in:

   Mean for Race - White: 95.28

   Mean for Race - Asian: 95.72

   Mean for Race - Hispanic: 93.98
4. After calculating the mean, **Std Error** by dividing sample standard deviation with the square root of n was calculated for which excel functions **STDEV.S()** to calculate the sample standard deviation and **SQRT()** to calculate square root of n were used.

   The output for Std Error was:

   Std Error for Race - White: 0.151

   Std Error for Race - Asian: 0.560

   Std Error for Race - Hispanic: 0.314
5. Next, was the **T-Value** which was calculated using the excel function **TINV()** by inputting the values of alpha and degree of freedom.

   T value calculated was 2.048 for White and Hispanic races and 2.093 for Asian.
6. As we know, formula to calculate the Confidence interval is:

   **Mean ± (Std Error * T-Value)**

   So, the 95% confidence interval for the three races was calculated using this formula and the results obtained were:

   For Race - White: **[94.97, 95.58]**

   For Race - Asian: **[94.55, 96.89]**

   For Race - Hispanic: **[93.33, 94.62]**

# Part 2:

As given,

The null hypothesis: the population means (true means) of female and male enrolled in school are equal; $H_0$: $\mu 1 = \mu 2$

The alternate hypothesis: the population means (true means) of female and male enrolled in school are not equal; $H_1$: $\mu 1 \neq \mu 2$

These can also be written as:

The null hypothesis: the difference of population means (true means) of female and male enrolled in school is equal to zero; $H_0$: $\mu D = \mu 1 - \mu 2 = 0$

The alternate hypothesis: the difference of population means (true means) of female and male enrolled in school is not equal to zero; $H_1$: $\mu D = \mu 1 - \mu 2 \neq 0$

To check this hypothesis for 'all races' and for 'Asians', I have done the following steps:

1. As instructed to choose any age group, I chose **'5 and 6 years'** to perform computations for the project.
2. From the main dataset present in the TableA-2.xls file, segregated the data for the races - All races and Asians for the same age group which further consisted of data recorded for Males and Females both.
3. After that, for both the races, calculated the difference between the values for Males and Females and named it as the **Difference** column.
4. Now, calculated the **mean** of the difference column for the two races races by using the excel function **AVERAGE()** which resulted in:
   Mean difference for Race - All races: -0.228
   Mean difference for Race - Asian: -0.085
5. After calculating the mean, **Std Error** by dividing sample standard deviation with the square root of n was calculated for which excel functions **STDEV.S()** to calculate the sample standard deviation and **SQRT()** to calculate square root of n were used.
   The output for Std Error was:
   Std Error for Race - All races: 0.145
   Std Error for Race - Asian: 0.778

6. To test the hypothesis using the P-value method, **T-statistic** plays a very important role.
   The formula to calculate T-statistic is:
   **t = (mean difference - 0)/std error** (here $\mu D = 0$)
   So I calculated T-statistic and then used it to get the P-value.
   T-statistic values for All races and the Asian race were calculated as -1.57 and -0.109 respectively.
7. Then the excel function **T.DIST()** to calculate **P-value** was used to which absolute value of T-statistic, degree of freedom and 2(for two tailed test) were given as input.
   P-Value results computed were:
   P-value for All races: 0.127
   P-value for Asian race: 0.914
8. Now, as the P-value for both the cases I got was greater than 0.05 (significance level), hence I failed to reject the null hypothesis i.e I accepted the null hypothesis.
9. This implied that the population means of male and females of both **All races** and the **Asian** race enrolled in school are equal.