STAT1005 - Assignment 2

**Instructions:**

* Please complete all questions put all your code and outputs into a single separate Python Notebook.
* Please rename your python notebook file as “{*UID*}-{*Name*}-assignment2.ipynb”, e.g., “3030016546-Chan\_Tai\_Man-assignment2.ipynb”.
* Please upload your ipynb file to the Moodle.
* Write down your explanation using Markdown cells.

**Submission:**

1. In Jupyter Notebook, save your file by clicking **File 🡪 Downloaded as 🡪 Notebook (.ipynb).**

2. Submit your **ipynb file** to Moodle.

**Enquiry:**

If you encounter any problems with this assignment, please feel free to post your questions on Moodle’s **Assignment Discussion Forum or email to TA,** MrCheung Wai Ki,Keith ([u3580369@connect.hku.hk](mailto:u3580369@connect.hku.hk)).

**Note:**

* You can initialise the ipynb file by your own, or directly download it from GitHub: <https://bit.ly/3BvlOcs>, or use it on Colab <https://bit.ly/3BqAOZ7>.

**Questions:**

1. Student Bob takes bus to school every day, but the wait time varies. In order to estimate the average wait time, he collected a sample of 30 observations (in minutes) [2 marks]:

wait\_time = [ 1, 4, 1, 8, 11, 4, 12, 4, 3, 11, 2, 6, 11, 6,

0, 9, 4, 3, 2, 9, 7, 8, 8, 6, 9, 1, 5, 4, 4, 4]

1. Write script to calculate the *sample mean* and *sample standard deviation* and print out the results.
2. Calculate the *unbiased* estimate of *population standard deviation* and print it out.
3. Calculate *standard error of the mean* and print out.
4. Approximate the distribution of *sample mean* via the Central Limit Theorem and plot out its probability density function.
5. Calculate the 90% confidence interval via the distribution of sample mean obtained in (d).
6. Based on the same sample in Q1, we now want to estimate the confidence interval with bootstrapping method. Please use fixed random seed by np.random.seed(0). [2 marks]
7. Write script to generate ONE bootstrapping sample and print the sample out. Also print the mean and standard deviation of this bootstrapped sample.
8. Write script to generate 1,000 bootstrapping samples. For each bootstrapped sample, calculate its mean. Print out the mean and standard deviation of the 1,000 bootstrapped sample means.
9. Plot the histogram of the bootstrapped sample means.
10. Calculate the lower and upper bound of the 95% confidence interval and print them out.
11. As a teacher of the STAT1005, I want to find out what factors contribute to the course performance. In order to check if boys and girls performs differently, we randomly collected the final marks for 20 boys and 20 girls, as below [2 marks]:

girls\_scores = [77, 95, 89, 94, 99, 91, 90, 77, 85, 90,

93, 96, 87, 77, 85, 98, 83, 89, 98, 72]

boys\_scores = [91, 65, 87, 90, 84, 77, 76, 91, 75, 71,

89, 97, 59, 84, 92, 87, 90, 79, 96, 89]

1. Calculate difference of *group mean*, and unbiased estimate of *standard deviation* from each group.
2. Calculate the pooled standard deviation and print it out. Hint: check slide P20.
3. Calculate the *t* statistic, at the standardized scale.
4. Use scipy.stats.ttest\_ind() to perform the *t* test, and print the t statistic and p value based on *two-sided* alternative. Is the returned *t* statistic the same as the calculated in (c)?
5. Multiple testing [1 mark].
6. If the null hypothesis is true (e.g., that boys have no difference with girls in Q3), what is the probability to observe p<0.05? Print out the result.
7. If the null hypothesis is true and we repeating the above tests 20 tests (e.g., by repeating sampling 20 students in each group from our class), what is probability to see at least one test with p<0.05? Hint: think the opposite that all p >0.95.
8. In a hypothetic situation, the final marks for STAT1005 have a standard deviation of 10 marks for both boys and girls, and girls on average have 5 marks higher than boys. [1 mark]
9. What is the standardized effect size?
10. What is power to detect this effect size at significance level of 0.05 by randomly sampling 20 students in each group. Print out the results and interpret the meaning of the results. Use the *two-sided* alternative. Hint: use the statsmodels.stats.power.TTestIndPower() class & its power() function.
11. Now, we consider using regression-based testing to analyze the final scores between boys and girls in Q3. Besides the gender, we also obtained the study time as below (Note, this is hypothetical data) [2 mark]

df = pd.DataFrame({

'scores': [77, 95, 89, 94, 99, 91, 90, 77, 85, 90, 93, 96, 87, 77, 85,

98, 83, 89, 98, 72, 91, 65, 87, 90, 84, 77, 76, 91, 75, 71,

89, 97, 59, 84, 92, 87, 90, 79, 96, 89],

'is\_girl': [1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,

1., 1., 1., 1., 1., 1., 0., 0., 0., 0., 0., 0., 0., 0.,

0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],

'study\_time': [3.6, 4.3, 3.7, 4.0, 5.4, 3.0, 5.0, 2.4, 3.8, 4.0, 5.1,

3.7, 3.6, 2.6, 4.2, 4.4, 3.3, 3.6, 4.9, 2.6, 3.6, 2.2,

4.3, 4.4, 4., 2.5, 2.6, 3.7, 2.5, 2.5, 3.7, 4.6, 0.7,

3.1, 4.0, 3.8, 3.5, 3.1, 5.5, 4.4],

'intercept': [1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,

1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,

1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]

})

1. Like the analysis in Q3, we want to test if girls have different performance compared to boys with regression-based test, by only using [“is\_girl”, “intercept”] as covariate. Write script for this and print out the *t* statistic and p value. Hint: use statsmodels.api’s OLS.
2. Now write script to add the study time (hours per week) as additional covariate, namely using [“is\_girl”, “study\_time”, “intercept”] as the covariates. Print out the p value of “is\_girl” and interpret the difference between (a).