**Homework 5**

**Your student ID #**

|  |
| --- |
| **Homework Guidance**  To ensure the integrity and educational value of your work, please adhere to the following guidelines as you complete Homework 1:  **1) Originality and Collaboration**: You are encouraged to work independently and ensure that the submissions are your own. While collaboration with your peers for understanding concepts and discussing problems is allowed, directly copying work from your colleagues is strictly prohibited. Your submission should reflect your understanding and your ability to apply what you have learned.  **2) Use of GPT-like Platforms**: You are permitted to use GPT-like platforms for assistance with your homework. However, this tool should only be used when you fully understand the answers it provides. The purpose of using such platforms is to enhance your learning, not to bypass the learning process. ***Keep in mind that midterm and final exams will be conducted without internet access***. If there is a significant discrepancy between the code you submit for homework and your ability to write similar code during an exam, ***it will be considered cheating***. Such instances will result in a score of zero for the involved exam component.  **3) Submission Quality**: Your focus should be on submitting code that you comprehend thoroughly. Fancy or complex code that goes beyond your level of understanding is not the goal. We value honesty and genuine effort. Make sure that you can explain and justify every line of code you submit. This approach will not only help you in your homework but also prepare you for the no-internet exams. |

Your coding HWs (check week05 pdf files) from 1 should be answered here:

1.

2. Evaluate the normality of the flow data sets from both before and after the regulation using the PPCC (Probability Plot Correlation Coefficient) test. Should any of the datasets deviate from a normal distribution, apply an appropriate transformation to the data and conduct the test again to identify a transformation that results in a distribution resembling normality.

Define and use the following ppcc\_test function for questions 2, 3 and 4:

def calculate\_ppcc(data):

"""Calculate the Probability Plot Correlation Coefficient (PPCC)."""

sorted\_data = np.sort(data)

theoretical\_quantiles = stats.norm.ppf((np.arange(1, len(data) + 1) - 0.5) / len(data))

ppcc, \_ = stats.pearsonr(sorted\_data, theoretical\_quantiles)

return ppcc

def ppcc\_test(data, alpha=0.05, num\_simulations=10000):

"""Calculate the PPCC, simulate p-value, evaluate against alpha, and plot both simulated PPCC histogram and Q-Q plot."""

original\_ppcc = calculate\_ppcc(data)

simulated\_ppccs = []

for \_ in range(num\_simulations):

simulated\_data = np.random.normal(np.mean(data), np.std(data), len(data))

simulated\_ppcc = calculate\_ppcc(simulated\_data)

simulated\_ppccs.append(simulated\_ppcc)

p\_value = np.sum(np.array(simulated\_ppccs) <= original\_ppcc) / num\_simulations

fig, axs = plt.subplots(1, 2, figsize=(12, 5))

axs[0].hist(simulated\_ppccs, bins=30, alpha=0.7, label='Simulated PPCCs')

axs[0].axvline(original\_ppcc, color='r', linestyle='dashed', linewidth=2, label=f'Original PPCC: {original\_ppcc}')

axs[0].set\_title('Distribution of Simulated PPCCs')

axs[0].set\_xlabel('PPCC')

axs[0].set\_ylabel('Frequency')

axs[0].legend()

stats.probplot(data, dist="norm", plot=axs[1])

axs[1].set\_title('Q-Q Plot for Normality')

axs[1].legend(['Data', 'Theoretical Quantiles'])

plt.tight\_layout()

plt.show()

print(f"Simulated p-value: {p\_value}, Alpha: {alpha}")

if p\_value < alpha:

print(f"H\_0 (data is normally distributed) rejected, which means the data is considered to be not normal distribution with the significance level of {alpha}, and the p-value is {p\_value}.")

else:

print(f"H\_0 (data is normally distributed) not rejected, indicating the data can be considered normal distribution with the significance level of {alpha}, and the p-value is {p\_value}.")

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **Value** | **Period** | 1969 | 1350 | After |
| 1950 | 4910 | Before | 1970 | 2350 | After |
| 1951 | 3660 | Before | 1971 | 3140 | After |
| 1952 | 3910 | Before | 1972 | 3060 | After |
| 1953 | 1750 | Before | 1973 | 3630 | After |
| 1954 | 1050 | Before | 1974 | 3890 | After |
| 1955 | 2670 | Before | 1975 | 3780 | After |
| 1956 | 2880 | Before | 1976 | 3180 | After |
| 1957 | 2600 | Before | 1977 | 2260 | After |
| 1958 | 3520 | Before | 1978 | 3430 | After |
| 1959 | 1730 | Before | 1979 | 5290 | After |
| 1960 | 2340 | Before | 1980 | 2870 | After |
| 1961 | 2600 | Before |
| 1962 | 3410 | Before |
| 1963 | 1870 | Before |
| 1964 | 1730 | Before |
| 1965 | 2730 | Before |
| 1966 | 1550 | Before |
| 1967 | 4060 | Before |
| 1968 | 2870 | Before |
|  |  |  |

3. Test the arsenic data and transformed data (week04) for normality.