Statistics

Practice Questions

PW SKILLS

1. Generate a list of 100 integers containing values between 90 to 130 and store it in the variable `int_list`.

After generating the list, find the following:

(i) Write a Python function to calculate the mean of a given list of numbers.
Create a function to find the median of a list of numbers.

(ii) Develop a program to compute the mode of a list of integers.

(iii) Implement a function to calculate the weighted mean of a list of values and their corresponding weights.

(iv) Write a Python function to find the geometric mean of a list of positive numbers.

(v) Create a program to calculate the harmonic mean of a list of values.

(vi) Build a function to determine the midrange of a list of numbers (average of the minimum and maximum).

(vii) Implement a Python program to find the trimmed mean of a list, excluding a certain percentage of outliers.

2. Generate a list of 500 integers containing values between 200 to 300 and store it in the variable `int_list2`.

After generating the list, find the following:

(i) Compare the given list of visualization for the given data:

1. Frequency & Gaussian distribution
2. Frequency smoothened KDE plot
3. Gaussian distribution & smoothened KDE plot

(ii) Write a Python function to calculate the range of a given list of numbers.

(iii) Create a program to find the variance and standard deviation of a list of numbers.

(iv) Implement a function to compute the interquartile range (IQR) of a list of values.

(v) Build a program to calculate the coefficient of variation for a dataset.

(vi) Write a Python function to find the mean absolute deviation (MAD) of a list of numbers.

(vii) Create a program to calculate the quartile deviation of a list of values.

(viii) Implement a function to find the range-based coefficient of dispersion for a dataset.

3. **Write a Python class representing a discrete random variable with methods to calculate its expected value and variance.**

4. **Implement a program to simulate the rolling of a fair six-sided die and calculate the expected value and variance of the outcomes.**

5. **Create a Python function to generate random samples from a given probability distribution (e.g., binomial, Poisson) and calculate their mean and variance.**

6. **Write a Python script to generate random numbers from a Gaussian (normal) distribution and compute the mean, variance, and standard deviation of the samples.**

7. **Use seaborn library to load `tips` dataset. Find the following from the dataset for the columns `total_bill` and `tip`:**

(i) Write a Python function that calculates their skewness.

(ii) Create a program that determines whether the columns exhibit positive skewness, negative skewness, or is approximately symmetric.

(iii) Write a function that calculates the covariance between two columns.

(iv) Implement a Python program that calculates the Pearson correlation coefficient between two columns.

(v) Write a script to visualize the correlation between two specific columns in a Pandas DataFrame using scatter plots.

8. **Write a Python function to calculate the probability density function (PDF) of a continuous random variable for a given normal distribution.**

9. **Create a program to calculate the cumulative distribution function (CDF) of exponential distribution.**

10. **Write a Python function to calculate the probability mass function (PMF) of Poisson distribution.**

11. **A company wants to test if a new website layout leads to a higher conversion rate (percentage of visitors who make a purchase). They collect data from the old and new layouts to compare.**

To generate the data use the following command:

```python
import numpy as np
# 50 purchases out of 1000 visitors
old_layout = np.array([1] * 50 + [0] * 950)
# 70 purchases out of 1000 visitors
new_layout = np.array([1] * 70 + [0] * 930)
```

Apply z-test to find which layout is successful.

12. **A tutoring service claims that its program improves students' exam scores. A sample of students who participated in the program was taken, and their scores before and after the program were recorded.**

Use the below code to generate samples of respective arrays of marks:

```python
before_program = np.array([75, 80, 85, 70, 90, 78, 92, 88, 82, 87])
after_program = np.array([80, 85, 90, 80, 92, 80, 95, 90, 85, 88])
```

Use z-test to find if the claims made by tutor are true or false.

13. **A pharmaceutical company wants to determine if a new drug is effective in reducing blood pressure. They conduct a study and record blood pressure measurements before and after administering the drug.**

Use the below code to generate samples of respective arrays of blood pressure:

```python
before_drug = np.array([145, 150, 140, 135, 155, 160, 152, 148, 130, 138])
after_drug = np.array([130, 140, 132, 128, 145, 148, 138, 136, 125, 130])
```

Implement z-test to find if the drug really works or not.

14. **A customer service department claims that their average response time is less than 5 minutes. A sample of recent customer interactions was taken, and the response times were recorded.**

Implement the below code to generate the array of response time:
```python
response_times = np.array([4.3, 3.8, 5.1, 4.9, 4.7, 4.2, 5.2, 4.5, 4.6, 4.4])
```
Implement z-test to find the claims made by customer service department are tru or false.

15. **A company is testing two different website layouts to see which one leads to higher click-through rates. Write a Python function to perform an A/B test analysis, including calculating the t-statistic, degrees of freedom, and p-value.**

Use the following data:
```python
layout_a_clicks = [28, 32, 33, 29, 31, 34, 30, 35, 36, 37]
layout_b_clicks = [40, 41, 38, 42, 39, 44, 43, 41, 45, 47]
```

16. **A pharmaceutical company wants to determine if a new drug is more effective than an existing drug in reducing cholesterol levels. Create a program to analyze the clinical trial data and calculate the t-statistic and p-value for the treatment effect.**

Use the following data of cholestrol level:
```python
existing_drug_levels = [180, 182, 175, 185, 178, 176, 172, 184, 179, 183]
new_drug_levels = [170, 172, 165, 168, 175, 173, 170, 178, 172, 176]
```

17. **A school district introduces an educational intervention program to improve math scores. Write a Python function to analyze pre- and post-intervention test scores, calculating the t-statistic and p-value to determine if the intervention had a significant impact.**

Use the following data of test score:

```python
pre_intervention_scores = [80, 85, 90, 75, 88, 82, 92, 78, 85, 87]
post_intervention_scores = [90, 92, 88, 92, 95, 91, 96, 93, 89, 93]
```

18. **An HR department wants to investigate if there's a gender-based salary gap within the company. Develop a program to analyze salary data, calculate the t-statistic, and determine if there's a statistically significant difference between the average salaries of male and female employees.**

Use the below code to generate synthetic data:

```python
# Generate synthetic salary data for male and female employees
np.random.seed(0)  # For reproducibility
male_salaries = np.random.normal(loc=50000, scale=10000, size=20)
female_salaries = np.random.normal(loc=55000, scale=9000, size=20)
```

19. **A manufacturer produces two different versions of a product and wants to compare their quality scores. Create a Python function to analyze quality assessment data, calculate the t-statistic, and decide whether there's a significant difference in quality between the two versions.**

Use the following data:

```python
version1_scores = [85, 88, 82, 89, 87, 84, 90, 88, 85, 86, 91, 83, 87, 84, 89, 86, 84, 88, 85, 86, 89, 90, 87, 88, 85]
version2_scores = [80, 78, 83, 81, 79, 82, 76, 80, 78, 81, 77, 82, 80, 79, 82, 79, 80, 81, 79, 82, 79, 78, 80, 81, 82]
```

20. **A restaurant chain collects customer satisfaction scores for two different branches. Write a program to analyze the scores, calculate the t-statistic, and determine if there's a statistically significant difference in customer satisfaction between the branches.**

Use the below data of scores:

```python
branch_a_scores = [4, 5, 3, 4, 5, 4, 5, 3, 4, 4, 5, 4, 4, 3, 4, 5, 5, 4, 3, 4, 5, 4, 3, 5, 4, 4, 5, 3, 4, 5, 4]
branch_b_scores = [3, 4, 2, 3, 4, 3, 4, 2, 3, 3, 4, 3, 3, 2, 3, 4, 4, 3, 2, 3, 4, 3, 2, 4, 3, 3, 4, 2, 3, 4, 3]
```

21. **A political analyst wants to determine if there is a significant association between age groups and voter preferences (Candidate A or Candidate B). They collect data from a sample of 500 voters and classify them into different age groups and candidate preferences. Perform a Chi-Square test to determine if there is a significant association between age groups and voter preferences.**

Use the below code to generate data:

```python
np.random.seed(0)
age_groups = np.random.choice(['18-30', '31-50', '51+', '51+'], size=30)
voter_preferences = np.random.choice(['Candidate A', 'Candidate B'], size=30)
```

22. A company conducted a customer satisfaction survey to determine if there is a significant relationship between product satisfaction levels (Satisfied, Neutral, Dissatisfied) and the region where customers are located (East, West, North, South). The survey data is summarized in a contingency table. Conduct a Chi-Square test to determine if there is a significant relationship between product satisfaction levels and customer regions.

Sample data:

```python
#Sample data: Product satisfaction levels (rows) vs. Customer regions (columns)
data = np.array([[50, 30, 40, 20], [30, 40, 30, 50], [20, 30, 40, 30]])
```

23. A company implemented an employee training program to improve job performance (Effective, Neutral, Ineffective). After the training, they collected data from a sample of employees and classified them based on their job performance before and after the training. Perform a Chi-Square test to determine if there is a significant difference between job performance levels before and after the training.

Sample data:

```python
# Sample data: Job performance levels before (rows) and after (columns) training
data = np.array([[50, 30, 20], [30, 40, 30], [20, 30, 40]])
```

24. A company produces three different versions of a product: Standard, Premium, and Deluxe. The company wants to determine if there is a significant difference in customer satisfaction scores among the three product versions. They conducted a survey and collected customer satisfaction scores for each version from a random sample of customers. Perform an ANOVA test to determine if there is a significant difference in customer satisfaction scores.

Use the following data:

```python
# Sample data: Customer satisfaction scores for each product version
standard_scores = [80, 85, 90, 78, 88, 82, 92, 78, 85, 87]
premium_scores = [90, 92, 88, 92, 95, 91, 96, 93, 89, 93]
deluxe_scores = [95, 98, 92, 97, 96, 94, 98, 97, 92, 99]
```