Choosing the right statistical test

1. t-test

- **Purpose**: To **compare the means of two groups.**
- **Use When**:
  - You have **two independent groups** and want to compare their means.
  - You have a **continuous variable** (e.g., **height, weight, salary**) and a **binary categorical variable** (e.g., male/female, treatment/control).
  - The **data is normally distributed** and has **similar variances** in both groups.

Types of t-tests:

- **One-sample t-test**: Compares the mean of a sample to a known value or population mean.
  - Ex. Testing if the average height of a group is 170 cm.
- **Independent t-test**: Compares the means of two independent groups.
  - Ex. Compare the average test scores between males and females.
- **Paired t-test**: Compares means from the same group at different times (e.g., before/after).
  - Ex. Compare blood pressure before and after treatment for the same patients.

2. ANOVA (Analysis of Variance)

- **Purpose**: To compare the means of **three or more groups**.
- **Use When**:
  - You have more than two groups (three or more) and want to test if their means are different.
  - The **dependent variable** is **continuous**, and the independent variable is **categorical** with more than two levels.
  - The data is **normally distributed**, and variances are **equal** across the groups.

Types of ANOVA:

- **One-way ANOVA**: Compares the means of three or more independent groups (single independent variable).
  - Ex. Compare the average salary of employees based on three different departments (HR, Sales, Marketing).

- **Two-way ANOVA**: Compares the means of groups based on two independent variables (factors) and checks for interaction between them.
    - Compare the test scores based on teaching methods and gender.

3. Chi-Square Test

- **Purpose**: To examine if there is an association or independence between **two categorical variables.**
- **Use When**:
    - You have **categorical data** (nominal or ordinal) and want to test the association between two variables.
    - The data is in the form of **counts or frequencies** (e.g., number of people in each category).

Types of Chi-Square Tests:

- **Chi-Square Test of Independence**: Tests if two categorical variables are independent (no association).
    - Ex. *Examine if there is an association between gender (male/female) and smoking status (smoker/non-smoker).*
- **Chi-Square Goodness-of-Fit Test**: Tests if a sample data matches an expected distribution.
    - Ex. *Test if a die is fair by comparing the frequency of each number rolled to the expected uniform distribution.*

Quick Guide to Choose the Test:

1. **If your data is categorical**:
    - **Chi-Square** (Goodness-of-Fit or Test of Independence).
2. **If your data is continuous**:
    - **t-test** for comparing **two groups**.
    - **ANOVA** for comparing **three or more groups**.
    - **Paired t-test** for repeated measures or matched pairs.

Choose the Right Test

For Continuous Data:

Question	Test	Example
Compare means of **two groups**.	Independent t-test	Compare test scores of two schools.
Compare means of **paired groups**.	Paired t-test	Compare pre-test and post-test scores.
Compare means of **three or more groups**.	ANOVA	Compare test scores of three teaching methods.
Test for **correlation**.	Pearson's correlation	Is there a relationship between height and weight?

For Categorical Data:

Question	Test	Example
Test for **independence** between two categorical variables.	Chi-Square Test of Independence	Is there a relationship between gender and voting preference?
Test if observed frequencies match expected frequencies.	Chi-Square Goodness-of-Fit Test	Does the distribution of colors in a bag of candies match the expected distribution?
Compare proportions of **two groups**.	Z-test for proportions	Do two groups have different success rates?

Decision Tree for Choosing a Test

1. **What is your data type?**
   - **Continuous**
   - **Categorica**
2. **For Continuous Data**:
   - Are you comparing **two groups**?
     - Yes: Use **t-test**.
       - Independent groups: **Independent t-test**.
       - Paired groups: **Paired t-test**.
     - No: Are you comparing **three or more groups**?
       - Yes: Use **ANOVA**.
       - No: Are you testing for a **relationship**?
         - Yes: Use **Pearson's correlation**.
3. **For Categorical Data**:
   - Are you testing for **independence** between two variables?
     - Yes: Use **Chi-Square Test of Independence**.
     - No: Are you testing if observed frequencies match expected frequencies?

- Yes: Use **Chi-Square Goodness-of-Fit Test**.
- No: Are you comparing **proportions**?
  - Yes: Use **Z-test for proportions**.

Summary Table

Data Type	Question	Test
Continuous	Compare two groups.	Independent t-test.
Continuous	Compare paired groups.	Paired t-test.
Continuous	Compare three or more groups.	ANOVA.
Continuous	Test for a relationship.	Pearson's correlation.
Categorical	Test for independence.	Chi-Square Test of Independence.
Categorical	Test goodness-of-fit.	Chi-Square Goodness-of-Fit Test.
Categorical	Compare proportions.	Z-test for proportions.

Key Takeaways

- **Continuous Data**: Use t-tests, ANOVA, or correlation.
- **Categorical Data**: Use Chi-Square tests or Z-tests for proportion

---

Python Code for Common Statistical Tests

Test	Use Case	Function	Python Code
Independent t-test	Compare means of two **independent groups**.	`scipy.stats.ttest_ind`	`python from scipy.stats import ttest_ind t_stat, p = ttest_ind( group1, group2 )`
Paired t-test	Compare means of **paired/matched groups**.	`scipy.stats.ttest_rel`	`python from scipy.stats import ttest_rel t_stat, p = ttest_rel ( pre, post )`
One-Way ANOVA	Compare means of **three or more groups**.	`scipy.stats.f_oneway`	`python from scipy.stats import f_oneway f_stat, p = f_oneway ( group1, group2, group3 )`

Test	Use Case	Function	Python Code
Chi-Square (Independence)	Test association between **two categorical variables**.	`scipy.stats.chi2_contingency`	`python from scipy.stats import chi2_contingency` `chi2, p, dof, expected = chi2_contingency( observed_table )`
Chi-Square (Goodness-of-Fit)	Test if observed frequencies match expected frequencies.	`scipy.stats.chisquare`	`python from scipy.stats import chisquare` `chi2, p = chisquare ( observed, expected )`
Pearson's Correlation	Test linear relationship between **two continuous variables**.	`scipy.stats.pearsonr`	`python from scipy.stats import pearsonr` `corr, p = pearsonr ( x, y)`
Z-Test for Proportions	Compare proportions of **two independent groups**.	`statsmodels.stats.proportion.proportions_ztest`	`python from statsmodels.stats.proportion import proportions_ztest` `stat, p = proportions_ztest([success1, success2], [n1, n2])`

- `chi2, p, dof, expected= chi2_contingency( observed_table )`
  - We get 4 values in `chi2_contingency()`

1. Independent t-test

```
from scipy.stats import ttest_ind

# Sample data for two independent groups
group1 = [25, 30, 35, 40, 45]  # Group A
group2 = [30, 35, 40, 45, 50]  # Group B

# Perform t-test
t_stat, p_value = ttest_ind(group1, group2)
print(f"T-statistic: {t_stat:.4f}, P-value: {p_value:.4f}")
```

2. Paired t-test

```
from scipy.stats import ttest_rel

# Paired data (before and after)
pre = [25, 30, 35, 40, 45]  # Before treatment
```

```
post = [30, 35, 40, 45, 50]  # After treatment

# Perform paired t-test
t_stat, p_value = ttest_rel(pre, post)
print(f"T-statistic: {t_stat:.4f}, P-value: {p_value:.4f}")
```

3. One-Way ANOVA

```
from scipy.stats import f_oneway

# Sample data for three groups
group1 = [25, 30, 35, 40, 45]  # Group A
group2 = [30, 35, 40, 45, 50]  # Group B
group3 = [35, 40, 45, 50, 55]  # Group C

# Perform ANOVA
f_stat, p_value = f_oneway(group1, group2, group3)
print(f"F-statistic: {f_stat:.4f}, P-value: {p_value:.4f}")
```

4. Chi-Square Test of Independence

```
from scipy.stats import chi2_contingency

# Contingency table (observed frequencies)
observed = [[30, 20],  # Row 1
            [25, 35]]  # Row 2

# Perform Chi-Square test
chi2_stat, p_value, dof, expected = chi2_contingency(observed)
print(f"Chi2-statistic: {chi2_stat:.4f}, P-value: {p_value:.4f}")
```

5. Chi-Square Goodness-of-Fit Test

```
from scipy.stats import chisquare

# Observed and expected frequencies
observed = [30, 20, 25, 35]
expected = [25, 25, 25, 25]

# Perform Chi-Square test
chi2_stat, p_value = chisquare(observed, expected)
print(f"Chi2-statistic: {chi2_stat:.4f}, P-value: {p_value:.4f}")
```

6. Pearson’s Correlation

```
from scipy.stats import pearsonr

# Sample data
x = [25, 30, 35, 40, 45]
y = [50, 55, 60, 65, 70]

# Compute correlation
corr, p_value = pearsonr(x, y)
print(f"Correlation: {corr:.4f}, P-value: {p_value:.4f}")
```

7. Z-Test for Proportions

```
from statsmodels.stats.proportion import proportions_ztest

# Success counts and sample sizes
success = [30, 40]  # Successes in Group 1 and Group 2
nobs = [100, 100]   # Total observations in each group

# Perform Z-test
stat, p_value = proportions_ztest(success, nobs)
print(f"Z-statistic: {stat:.4f}, P-value: {p_value:.4f}")
```

Key Takeaways

- Use **t-tests** for comparing means of 2 groups.
- Use **ANOVA** for comparing means of 3+ groups.
- Use **Chi-Square** for categorical data (independence or goodness-of-fit).
- Use **Pearson’s correlation** for linear relationships between continuous variables.
- Use **Z-test for proportions** to compare proportions between groups.