Q1. Explain the assumptions required to use ANOVA and provide examples of violations that could impact the validity of the results.

Assumptions for using ANOVA:

Independence: Observations in each group are independent of each other.

Normality: Each group of observations follows a normal distribution.

Homogeneity of variance: The variances of each group are approximately equal.

Violations that could impact the validity of ANOVA results:

Violation of independence: When the observations in each group are not independent of each other, the results of ANOVA may be invalid. For example, if multiple measurements are taken from the same subject, the independence assumption is violated. Violation of normality: If the data in any group does not follow a normal distribution, the results of ANOVA may be inaccurate. For example, if the data is skewed or has outliers, the normality assumption may be violated. Violation of homogeneity of variance: If the variances of the groups are not approximately equal, the results of ANOVA may be biased. For example, if the variance of one group is much larger than the others, the homogeneity of variance assumption may be violated.

Q2. What are the three types of ANOVA, and in what situations would each be used?

There are three types of ANOVA: one-way ANOVA, two-way ANOVA, and repeated measures ANOVA. Each type of ANOVA is used in different situations depending on the research question and the design of the study. One-way ANOVA: One-way ANOVA is used when there is one independent variable (IV) with two or more levels, and the dependent variable (DV) is continuous. For example, if we want to compare the mean test scores of students from three different schools, we would use one-way ANOVA. The independent variable would be the school, and the dependent variable would be the test scores.

Two-way ANOVA (or factorial ANOVA): It is used when there are two independent variables with two or more levels, and the dependent variable is continuous. For example, if we want to compare the mean test scores of students from three different schools who also have different genders, we would use two-way ANOVA. The independent variables would be school and gender, and the dependent variable would be test scores.

Repeated measures ANOVA: Repeated measures ANOVA is used when the same individuals are measured on the same dependent variable at different points in time or under different conditions. For example, if we want to compare the effectiveness of three different medications on the same group of patients, we would use repeated measures ANOVA. The dependent variable would be the patient's health status, and the independent variable would be the type of medication they received

Q3. What is the partitioning of variance in ANOVA, and why is it important to understand this concept?

The partitioning of variance in ANOVA is a way to break down the total variance in a dependent variable into its different components, which helps us understand the sources of variability in our data. There are three components to this partitioning: the sum of squares total (SST), the sum of squares due to treatment (SSTr), and the sum of squares error (SSE). The SST represents the total variability in the data, the SSTr represents the variability that can be explained by the independent variable(s), and the SSE represents the variability that is due to error or random fluctuations.

Understanding the partitioning of variance is important because it allows us to test the significance of the effects of the independent variable(s) on the dependent variable. By comparing the variance due to the independent variable(s) with the variance due to error, we can calculate an F-statistic and determine whether the effects of the independent variable(s) are statistically significant.

In addition, partitioning of variance helps us explore interaction effects between multiple independent variables on the dependent variable, which can provide valuable insights into the complex relationships between different factors. Overall, understanding the partitioning of variance in ANOVA is crucial for

Q4. How would you calculate the total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR) in a one-way ANOVA using Python?

In [ ]:

```
import pandas as pd
from scipy.stats import f_oneway

data = pd.read_csv('/content/data1.csv')

group_means = data.groupby('group')['value'].mean()
grand_mean = data['value'].mean()

ss_total = ((data['value'] - grand_mean) ** 2).sum()

ss_explained = ((group_means - grand_mean) ** 2 * data['group'].value_counts()).sum()

ss_residual = ss_total - ss_explained

f_stat, p_value = f_oneway(data[data['group'] == 'A']['value'],
                           data[data['group'] == 'B']['value'],
                           data[data['group'] == 'C']['value'])

print('SST:', ss_total)
print('SSE:', ss_explained)
print('SSR:', ss_residual)
print('F-statistic:', f_stat)
print('p-value:', p_value)
```

Q5. In a two-way ANOVA, how would you calculate the main effects and interaction effects using Pytho

In [ ]:

```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

data = pd.read_csv('/content/data2.csv')

model = ols('value ~ group1 + group2 + group1:group2', data).fit()

anova_results = anova_lm(model, typ=2)

print('Main effect of group1:', anova_results.loc['group1', 'sum_sq'])
print('Main effect of group2:', anova_results.loc['group2', 'sum_sq'])
print('Interaction effect:', anova_results.loc['group1:group2', 'sum_sq'])
```

Q6. Suppose you conducted a one-way ANOVA and obtained an F-statistic of 5.23 and a p-value of 0.02. What can you conclude about the differences between the groups, and how would you interpret these results?

If we conducted a one-way ANOVA and obtained an F-statistic of 5.23 and a p-value of 0.02, we can conclude that there is significant evidence to suggest that at least one of the groups has a different mean from the others. The p-value of 0.02 means that if there were truly no difference between the groups, we would only observe an F-statistic as large as 5.23 by chance in 2% of samples. This is below the conventional threshold of 0.05, which means we reject the null hypothesis of equal means.

In summary, a one-way ANOVA with an F-statistic of 5.23 and a p-value of 0.02 suggests that there is significant evidence to suggest that at least one of the groups has a different mean from the others. Further analysis can help to determine which specific group(s) differ and the practical significance of these differences.

Q7. In a repeated measures ANOVA, how would you handle missing data, and what are the potential consequences of using different methods to handle missing data?

In a repeated measures ANOVA, missing data can occur when some participants have missing values for one or more of the repeated measures. There are several ways to handle missing data in a repeated measures ANOVA, each with their own potential consequences: Complete case analysis: This method involves only analyzing participants who have complete data for all of the repeated measures. This can lead to loss of statistical power and potentially biased results if the missing data is not missing completely at random.

Mean imputation: This method involves replacing missing values with the mean value for that measure across all participants. This can introduce bias if the missing data is related to other variables in the dataset.

Last observation carried forward (LOCF): This method involves replacing missing values with the last observed value for that measure for that participant. This can introduce bias if the missing data is related to other variables in the dataset, and may not accurately reflect the participant's true score.

Multiple imputation: This method involves creating multiple imputed datasets where missing values are replaced with plausible values based on the observed data and the underlying data distribution. This method can provide more accurate estimates of the parameters of interest and may increase statistical power.

It is important to carefully consider the potential consequences of each method when handling missing data in a repeated measures ANOVA. Choosing an appropriate method can help to minimize bias and increase the accuracy and reliability of the results.

Q8. What are some common post-hoc tests used after ANOVA, and when would you use each one? Provide an example of a situation where a post-hoc test might be necessary.

After conducting an ANOVA, post-hoc tests are used to determine which specific groups differ significantly from each other. There are several commonly used post-hoc tests, including: Tukey's HSD (honest significant difference): This test is used to compare all possible pairs of group means and control the familywise error rate. It is appropriate when there are equal sample sizes and variances across groups.

Bonferroni correction: This test adjusts the significance level for each individual comparison to control the overall type I error rate. It is appropriate when there are few comparisons to be made.

Scheffé's method: This test is more conservative than Tukey's HSD and is appropriate when there are unequal sample sizes and variances across groups.

Games-Howell test: This test is a non-parametric alternative to Tukey's HSD and is appropriate when assumptions of normality and homogeneity of variance are violated.

Q9. A researcher wants to compare the mean weight loss of three diets: A, B, and C. They collect data from 50 participants who were randomly assigned to one of the diets. Conduct a one-way ANOVA using Python to determine if there are any significant differences between the mean weight loss of the three diets. Report the F-statistic and p-value, and interpret the results.

In [2]:

```
import numpy as np
from scipy.stats import f_oneway

# weight loss data for each diet
diet_A = np.array([2.1, 1.5, 3.2, 2.8, 1.7, 1.9, 2.5, 2.0, 2.9, 2.3,
                   1.8, 2.4, 1.6, 2.2, 2.6, 3.0, 1.4, 2.7, 2.1, 1.9,
                   2.8, 1.6, 2.3, 1.7, 2.5])
diet_B = np.array([2.7, 3.1, 2.5, 2.2, 2.8, 1.9, 2.1, 2.6, 2.0, 2.3,
                   2.8, 1.5, 2.2, 1.8, 2.9, 2.4, 1.7, 2.3, 2.6, 2.1,
                   1.4, 2.0, 1.6, 2.7, 2.3])
diet_C = np.array([1.8, 1.9, 1.6, 2.2, 2.0, 2.5, 2.1, 2.3, 2.8, 2.6,
                   2.4, 2.0, 2.7, 2.2, 1.5, 1.7, 2.9, 2.1, 1.8, 2.3,
                   2.5, 2.6, 2.1, 1.4, 2.2])

f_stat, p_val = f_oneway(diet_A, diet_B, diet_C)
print("F-statistic: {:.2f}".format(f_stat))
print("p-value: {:.4f}".format(p_val))
```

```
F-statistic: 0.25
p-value: 0.7771
```

Q10. A company wants to know if there are any significant differences in the average time it takes to complete a task using three different software programs: Program A, Program B, and Program C. They randomly assign 30 employees to one of the programs and record the time it takes each employee to complete the task. Conduct a two-way ANOVA using Python to determine if there are any main effects or interaction effects between the software programs and employee experience level (novice vs. experienced). Report the F-statistics and p-values, and interpret the results.

In [3]:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

np.random.seed(1234)

program = np.random.choice(['A', 'B', 'C'], size=90)
experience = np.random.choice(['Novice', 'Experienced'], size=90)
time = np.random.normal(loc=10, scale=2, size=90)

data = pd.DataFrame({'Program': program, 'Experience': experience, 'Time': time})
model = ols('Time ~ C(Program) + C(Experience) + C(Program):C(Experience)', data=data).f
anova_table = sm.stats.anova_lm(model, typ=2)

df = pd.DataFrame(anova_table)
print(df)
```

```
                            sum_sq    df         F    PR(>F)
C(Program)                4.970375   2.0  0.931445  0.398015
C(Experience)            18.443605   1.0  6.912639  0.010176
C(Program):C(Experience)  1.409813   2.0  0.264198  0.768457
Residual                224.120297  84.0       NaN       NaN
```

Q11. An educational researcher is interested in whether a new teaching method improves student test scores. They randomly assign 100 students to either the control group (traditional teaching method) or the experimental group (new teaching method) and administer a test at the end of the semester. Conduct a two-sample t-test using Python to determine if there are any significant differences in test scores between the two groups. If the results are significant, follow up with a post-hoc test to determine which group(s) differ significantly from each other.

In [5]:

```
import numpy as np
from scipy.stats import ttest_ind

np.random.seed(123)
control_scores = np.random.normal(loc=70, scale=10, size=100)
experimental_scores = np.random.normal(loc=75, scale=10, size=100)

t_stat, p_val = ttest_ind(control_scores, experimental_scores)

print("t-statistic: ", t_stat)
print("p-value: ", p_val)

import statsmodels.api as sm
from statsmodels.stats.multicomp import pairwise_tukeyhsd

data = np.concatenate([control_scores, experimental_scores])
group = np.concatenate([np.zeros_like(control_scores), np.ones_like(experimental_scores)
df = pd.DataFrame({"group": group, "score": data})

tukey = pairwise_tukeyhsd(endog=df['score'], groups=df['group'], alpha=0.05)

print(tukey)
```

```
t-statistic:  -3.0316172004188147
p-value:  0.0027577299763983324
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower  upper  reject
-------------------------------------------------
   0.0    1.0   4.5336 0.0028 1.5846 7.4826   True
-------------------------------------------------
```

Q12. A researcher wants to know if there are any significant differences in the average daily sales of three retail stores: Store A, Store B, and Store C. They randomly select 30 days and record the sales for each store on those days. Conduct a repeated measures ANOVA using Python to determine if there are any significant differences in sales between the three stores. If the results are significant, follow up with a post-hoc test to determine which store(s) differ significantly from each other.

In [6]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd

np.random.seed(123)
store_a = np.random.normal(50, 10, 30)
store_b = np.random.normal(60, 10, 30)
store_c = np.random.normal(70, 10, 30)
data = pd.DataFrame({'store_a': store_a, 'store_b': store_b, 'store_c': store_c})

data_melt = pd.melt(data.reset_index(), id_vars=['index'], value_vars=['store_a', 'store
data_melt.columns = ['index', 'store', 'sales']

model = ols('sales ~ C(store) + C(index)', data=data_melt).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print(anova_table)

posthoc = pairwise_tukeyhsd(data_melt['sales'], data_melt['store'])
print(posthoc)
```

```
               sum_sq    df          F        PR(>F)
C(store)  5331.865905   2.0  18.664017  5.520945e-07
C(index)  3482.528266  29.0   0.840722  6.896639e-01
Residual  8284.610559  58.0        NaN           NaN
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
 group1  group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
store_a store_b  10.9677 0.0013  3.8075 18.1279   True
store_a store_c  18.7645    0.0 11.6043 25.9247   True
store_b store_c   7.7968 0.0295  0.6366  14.957   True
----------------------------------------------------
```

In [ ]:

```

```