statistics-advance-6

August 13, 2023

[ ]:

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

Analysis of Variance (ANOVA) is a statistical technique used to compare means of multiple groups simultaneously.

It's an extension of the t-test and is commonly used when you have more than two groups to compare.

To use ANOVA effectively and interpret the results accurately, certain assumptions need to be met. These assumptions include:

Independence: The observations within each group should be independent of each other. This means that the data points in one group should not be influenced by or related to the data points in another group.

Normality: The distribution of the dependent variable within each group should be approximately normal (follow a bell-shaped curve).

Homogeneity of Variance (Homoscedasticity): The variance of the dependent variable should be roughly equal across all groups. This means that the spread of data points around the group means should be similar for each group.

Examples of violations of these assumptions that could impact the validity of ANOVA results:

Independence Violation: If observations within groups are not independent, it can lead to pseudoreplication. For instance, if you have multiple measurements from the same subject or if data points are time-dependent, the assumption of independence is violated.

Normality Violation: If the data in each group do not follow a normal distribution, ANOVA results might be less reliable. For example, if the data is heavily skewed or has extreme outliers, the normality assumption could be violated

Homoscedasticity Violation: If the variances across groups are not similar, the F-test used in ANOVA might not be accurate. Unequal variances can affect the power of the test, potentially leading to false positives or negatives. Levene's or Bartlett's test can be used to assess homoscedasticity.

[ ]:

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

One-Way ANOVA:

Situation: Used when you have one categorical independent variable (factor) with three or more levels (groups), and you want to compare the means of these groups to determine if they are significantly different from each other.

Example: You want to compare the average scores of students from three different schools to see if there's a significant difference in performance.

Two-Way ANOVA:

Situation: Used when you have two categorical independent variables (factors), often called "factors" and "levels," and you want to examine their effects on a continuous dependent variable. This type includes both main effects and interaction effects.

Example: You want to investigate the effects of both gender and treatment type on the effectiveness of a new drug. Here, gender and treatment type are the two independent variables, and the dependent variable is the drug's effectiveness.

Repeated Measures ANOVA (or within-subjects ANOVA):

Situation: Used when the same subjects are measured multiple times under different conditions or at different time points. It's appropriate when you want to analyze the effects of a repeated factor on a dependent variable.

Example: You're studying the impact of three different training programs on participants' performance. Instead of having separate groups, you measure each participant's performance before and after each training program.

[ ]:

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

Partitioning of variance in ANOVA refers to the process of breaking down the total variability observed in the data into different sources of variability or components. These components help us understand how much of the total variability can be attributed to different factors or sources within an experimental design. ANOVA partitions the variance into three main components:

Between-Groups Variance (SSB): This component of variance measures the variability between different groups or treatment conditions. It quantifies the differences among group means and provides insight into whether the groups are significantly different from each other.

Within-Groups Variance (SSW): Also known as the residual or error variance, this component of variance measures the variability within each group. It represents the variability that cannot be explained by the treatment conditions and is often considered as random variation or noise.

Total Variance (SST): This is the overall variability in the data, encompassing both the variability due to group differences and the within-group variability. It's the sum of the between-groups variance and the within-groups variance.

[1]:

```
# 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?
```

```python
# In a one-way ANOVA, you can calculate the Total Sum of Squares (SST), Explained Sum of Squares (SSE), and Residual Sum of Squares (SSR) to understand the variance components and assess the goodness of fit of the model.


import numpy as np
import scipy.stats as stats

# Example data for three groups (replace with your own data)
group1 = np.array([15, 18, 20, 22, 23])
group2 = np.array([25, 26, 28, 30, 31])
group3 = np.array([35, 38, 40, 42, 43])

# Combine data from all groups
all_data = np.concatenate((group1, group2, group3))

# Calculate the overall mean
overall_mean = np.mean(all_data)

# Calculate the Total Sum of Squares (SST)
sst = np.sum((all_data - overall_mean) ** 2)

# Calculate the group means
group1_mean = np.mean(group1)
group2_mean = np.mean(group2)
group3_mean = np.mean(group3)

# Calculate the Explained Sum of Squares (SSE)
sse = np.sum((group1_mean - overall_mean) ** 2) * len(group1) + \
      np.sum((group2_mean - overall_mean) ** 2) * len(group2) + \
      np.sum((group3_mean - overall_mean) ** 2) * len(group3)

# Calculate the Residual Sum of Squares (SSR)
ssr = sst - sse

# Print the results
print("Total Sum of Squares (SST):", sst)
print("Explained Sum of Squares (SSE):", sse)
print("Residual Sum of Squares (SSR):", ssr)
```

```
Total Sum of Squares (SST): 1116.9333333333332
Explained Sum of Squares (SSE): 1008.5333333333333
Residual Sum of Squares (SSR): 108.39999999999986
```

[2]:

```python
# Q5. In a two-way ANOVA, how would you calculate the main effects and interaction effects using Python?
# In a two-way ANOVA, you can calculate the main effects and interaction effects to understand the influences of the two independent variables and their combined effects on the dependent variable.

# Let's assume you have a dataset where you have two independent variables: Factor A (with levels A1 and A2) and Factor B (with levels B1 and B2), and a dependent variable (DV).








import numpy as np
import scipy.stats as stats

# Example data (replace with your own data)
factor_a = np.array([1, 2, 3, 4, 5, 6, 7, 8])
factor_b = np.array([10, 15, 20, 25, 30, 35, 40, 45])
dv = np.array([[12, 15, 18, 20, 25, 30, 32, 35],
               [18, 22, 24, 28, 33, 38, 40, 43],
               [25, 30, 32, 35, 40, 45, 48, 52],
               [30, 35, 38, 40, 45, 50, 52, 55]])

# Calculate the overall mean
overall_mean = np.mean(dv)

# Calculate the main effects
main_effect_a = np.mean(np.mean(dv, axis=1) - overall_mean)
main_effect_b = np.mean(np.mean(dv, axis=0) - overall_mean)

# Calculate the interaction effect
interaction_effect = np.sum(dv) - (main_effect_a + main_effect_b + overall_mean)

# Print the results
print("Main Effect of Factor A:", main_effect_a)
print("Main Effect of Factor B:", main_effect_b)
print("Interaction Effect:", interaction_effect)
```

```
Main Effect of Factor A: 0.0
Main Effect of Factor B: 0.0
Interaction Effect: 1051.09375
```

[ ]:

```python
# 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?
```

In a one-way ANOVA, the F-statistic is used to test the null hypothesis that the means of the groups are equal against the alternative hypothesis that at least one group mean is different from the others.

The p-value associated with the F-statistic helps you determine whether the observed differences between the group means are statistically significant.

F-Statistic: 5.23 p-value: 0.02

Interpretation:

Since the p-value (0.02) is less than the conventional significance level of 0.05, you would reject the null hypothesis. This suggests that there is enough evidence to conclude that there are statistically significant differences between the group means.

In other words:

There is evidence to suggest that at least one group mean is significantly different from the others.

The variability between the group means is larger than what you would expect due to random chance alone.

However, the ANOVA itself doesn't tell you which specific group(s) have different means.

[ ]:

```
# Q7. In a repeated measures ANOVA, how would you handle missing data?

# what are the potential consequences of using different methods to handle missing data?
```

Handling missing data in a repeated measures ANOVA is important to ensure the validity and reliability of your analysis. Missing data can occur due to various reasons, such as participants dropping out, equipment failure, or skipped responses. There are different methods to handle missing data, each with its potential consequences.

Mean Imputation:

Method: Replace missing data with the mean value of the observed data for that variable.

Consequences: It can distort the relationships between variables, underestimate variability, and underestimate standard errors. It assumes that missing data occur completely at random, which might not be the case.

[ ]:

```
# 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
```

Post-hoc tests are conducted after finding a significant result in an Analysis of Variance (ANOVA) to determine which specific group means are significantly different from each other.

Since ANOVA only tells us that there are differences among groups but doesn't identify which groups are different, post-hoc tests help provide more detailed insights.

Tukey's Honestly Significant Difference (HSD) Test:

Use: When you have a moderate to large number of groups and want to compare all possible pairwise differences.

Situation: You conduct an ANOVA comparing the effectiveness of three different teaching methods on student performance. The ANOVA indicates a significant difference among the groups, but you want to know which specific pairs of methods are significantly different.

Duncan's Multiple Range Test:

Use: When you want to test the differences between group means while controlling for the number of groups compared.

Situation: You are studying the yield of five different fertilizer treatments in agriculture. After conducting an ANOVA, you use Duncan's test to compare the groups and determine which treatments have significantly different yields.

Scheffe's Test:

Use: When you have unequal group sizes and want to control the experimentwise error rate.

Situation: You perform an ANOVA to compare the performance of several different machine learning algorithms on different datasets. Due to variations in dataset sizes, you choose Scheffe's test to account for the unequal group sizes when conducting pairwise comparisons.

Games-Howell Test:

Use: When the assumption of equal variances is violated and you have unequal sample sizes.

Situation: You're investigating the effects of three different medications on reducing blood pressure. The ANOVA shows a significant difference, but the Levene's test indicates unequal variances. In this case, you'd use the Games-Howell test for pairwise comparisons.

Bonferroni Correction:

Use: When conducting multiple pairwise comparisons while controlling for the familywise error rate (reducing the chance of making a Type I error).

Situation: You are comparing the means of five different treatment groups after an ANOVA. To reduce the risk of making false positive conclusions, you apply the Bonferroni correction to adjust the significance level for each comparison.

[ ]:

```
# 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.
```

[3]:

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

# Example data for weight loss in each diet group (replace with your own data)
diet_a = np.array([2, 3, 4, 5, 1, 2, 3, 2, 4, 3, 2, 4, 5, 1, 2, 3, 2, 4, 3, 2,␣
 ↪4, 5, 1, 2, 3, 2, 4, 3, 2, 4, 5, 1, 2, 3, 2, 4, 3, 2, 4, 5, 1, 2, 3, 2, 4,␣
 ↪3, 2, 4])
diet_b = np.array([3, 4, 5, 3, 2, 3, 2, 1, 4, 3, 4, 5, 3, 2, 3, 2, 1, 4, 3, 4,␣
 ↪5, 3, 2, 3, 2, 1, 4, 3, 4, 5, 3, 2, 3, 2, 1, 4, 3, 4, 5, 3, 2, 3, 2, 1, 4,␣
 ↪3, 4])
diet_c = np.array([1, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2,␣
 ↪3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 4, 5, 1, 2, 3, 2,␣
 ↪1, 2, 3])

# Combine data from all diet groups
all_data = np.concatenate((diet_a, diet_b, diet_c))

# Create labels for each diet group
group_labels = ['A'] * len(diet_a) + ['B'] * len(diet_b) + ['C'] * len(diet_c)

# Perform one-way ANOVA
f_statistic, p_value = stats.f_oneway(diet_a, diet_b, diet_c)

# Print the results
print("F-Statistic:", f_statistic)
print("p-value:", p_value)

# Interpret the results
if p_value < 0.05:
    print("There are significant differences between the mean weight loss of␣
 ↪the three diets.")
else:
    print("There are no significant differences between the mean weight loss of␣
 ↪the three diets.")
```

```
F-Statistic: 2.2466217926584706
p-value: 0.10955562521825858
There are no significant differences between the mean weight loss of the three
diets.
```

Interpretation of the results:

If the p-value is less than 0.05 (or your chosen significance level), you would conclude that there are significant differences between the mean weight loss of the three diets.

If the p-value is greater than or equal to 0.05, you would conclude that there are no significant differences between the mean weight loss of the three diets.

[ ]:

```
# 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.
```

[5]:

```
import numpy as np
import pandas as pd

# Create data
software_programs = np.repeat(['Program A', 'Program B', 'Program C'], 30)
experience_level = np.tile(['Novice', 'Experienced'], 45)
completion_time = np.random.randint(30, 120, size=90)  # Random completion times

# Create a DataFrame
data = pd.DataFrame({'Software': software_programs, 'Experience':␣
 ↪experience_level, 'Time': completion_time})

# Use the statsmodels library to perform the two-way ANOVA␣
 ↪<================================================

import statsmodels.api as sm
from statsmodels.formula.api import ols

# Fit the ANOVA model
model = ols('Time ~ C(Software) + C(Experience) + C(Software):C(Experience)',␣
 ↪data=data).fit()

# Get ANOVA table
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

```
                                 sum_sq    df         F    PR(>F)
C(Software)                 3018.466667   2.0  3.201318  0.045722
C(Experience)                227.211111   1.0  0.481950  0.489456
C(Software):C(Experience)   5515.755556   2.0  5.849886  0.004183
```

```
Residual                     39601.066667  84.0       NaN       NaN
```

Software Effect (Main Effect of Software):

p-value (PR(>F)): 0.045722 (less than 0.05)

Interpretation: The p-value is smaller than 0.05, indicating that there is a significant main effect of software programs on completion time. In other words, at least one software program significantly affects completion time.

Experience Effect (Main Effect of Experience):

p-value (PR(>F)): 0.489456 (greater than 0.05)

Interpretation: The p-value is larger than 0.05, suggesting that there is no significant main effect of experience level on completion time. Experience level does not have a significant influence on completion time in this analysis.

Interaction Effect (Software:Experience):

p-value (PR(>F)): 0.004183 (less than 0.05)

Interpretation: The p-value is smaller than 0.05, indicating a significant interaction effect between software programs and experience levels on completion time. This means that the effect of software programs on completion time is different depending on the experience level of the employees, and vice versa.

Residual (Error):

This row shows the variation in completion time that cannot be explained by the factors included in the model.

In summary:

There is a significant main effect of software programs on completion time. There is no significant main effect of experience level on completion time. There is a significant interaction effect between software programs and experience levels on completion time.

[ ]:
```
# 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.
```

[6]:
```
# Conducting a Two-Sample T-Test ========================================>



import numpy as np
import scipy.stats as stats

# Example data (replace with your own data)
control_group = np.array([85, 78, 92, 88, 75, 95, 82, 90, 87, 81, 79, 83, 86,␣
 ↪89, 93, 76, 80, 84, 91, 94])
experimental_group = np.array([88, 82, 96, 91, 76, 98, 85, 93, 90, 84, 81, 85,␣
 ↪88, 92, 95, 78, 82, 86, 94, 97])

# Conduct two-sample t-test
t_statistic, p_value = stats.ttest_ind(control_group, experimental_group)

# Print t-statistic and p-value
print("t-statistic:", t_statistic)
print("p-value:", p_value)

# Interpret the results
if p_value < 0.05:
    print("There is a significant difference in test scores between the two␣
 ↪groups.")
else:
    print("There is no significant difference in test scores between the two␣
 ↪groups.")
```

```
t-statistic: -1.338384736778457
p-value: 0.18872445240562336
There is no significant difference in test scores between the two groups.
```

a two-sample t-test and found that the p-value is 0.1887, which is greater than 0.05. This indicates that there is no significant difference in test scores between the control and experimental groups. Since your t-test did not yield a significant result, there is no need to perform a post-hoc test.

[ ]:

```
# 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
```

[23]:

```
import pandas as pd
from statsmodels.stats.anova import AnovaRM
```

```
# Your data
data = {
    'Day': list(range(1, 31)),
    'Store_A': [120, 130, 125, 140, 135, 130, 125, 140, 150, 130, 145, 135,␣
 ↪125, 130, 140, 135, 130, 125, 140, 150, 130, 145, 135, 125, 130, 140, 135,␣
 ↪130, 125, 140],
    'Store_B': [110, 125, 120, 135, 130, 125, 120, 135, 145, 125, 140, 130,␣
 ↪120, 125, 135, 130, 125, 120, 135, 145, 125, 140, 130, 120, 125, 135, 130,␣
 ↪125, 120, 135],
    'Store_C': [105, 115, 110, 125, 120, 115, 110, 125, 135, 115, 130, 120,␣
 ↪110, 115, 125, 120, 115, 110, 125, 135, 115, 130, 120, 110, 115, 125, 120,␣
 ↪115, 110, 125]
}

# Create a DataFrame
df = pd.DataFrame(data)

# Reshape the data for repeated measures ANOVA
long_df = pd.melt(df, id_vars=['Day'], value_vars=['Store_A', 'Store_B',␣
 ↪'Store_C'], var_name='Store', value_name='Sales')

# Conduct repeated measures ANOVA
anovarm = AnovaRM(long_df, 'Sales', 'Day', within=['Store'])
results = anovarm.fit()

print(results)
```

```
                Anova
=====================================
      F Value  Num DF  Den DF Pr > F
-------------------------------------
Store 6271.0000 2.0000 58.0000 0.0000
=====================================
```

[24]:

```
import pandas as pd
from statsmodels.stats.anova import AnovaRM
import numpy as np
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Your data
data = {
    'Day': list(range(1, 31)),
    'Store_A': [120, 130, 125, 140, 135, 130, 125, 140, 150, 130, 145, 135,␣
 ↪125, 130, 140, 135, 130, 125, 140, 150, 130, 145, 135, 125, 130, 140, 135,␣
 ↪130, 125, 140],
    'Store_B': [110, 125, 120, 135, 130, 125, 120, 135, 145, 125, 140, 130,␣
 ↪120, 125, 135, 130, 125, 120, 135, 145, 125, 140, 130, 120, 125, 135, 130,␣
 ↪125, 120, 135],
    'Store_C': [105, 115, 110, 125, 120, 115, 110, 125, 135, 115, 130, 120,␣
 ↪110, 115, 125, 120, 115, 110, 125, 135, 115, 130, 120, 110, 115, 125, 120,␣
 ↪115, 110, 125]
}

# Create a DataFrame
df = pd.DataFrame(data)

# Reshape the data for repeated measures ANOVA
long_df = pd.melt(df, id_vars=['Day'], value_vars=['Store_A', 'Store_B',␣
 ↪'Store_C'], var_name='Store', value_name='Sales')

# Conduct repeated measures ANOVA
anovarm = AnovaRM(long_df, 'Sales', 'Day', within=['Store'])
results = anovarm.fit()

print("Repeated Measures ANOVA Results:")
print(results)
print()

# Check if the results are significant
if results.anova_table['Pr > F'][0] < 0.05:
    print("There are significant differences in average daily sales between the␣
 ↪three stores.")

    # Perform post hoc test (Tukey's HSD)
    posthoc = pairwise_tukeyhsd(long_df['Sales'], long_df['Store'])
    print("Post Hoc Test Results:")
    print(posthoc)
else:
    print("There are no significant differences in average daily sales between␣
 ↪the three stores.")
```

```
Repeated Measures ANOVA Results:
                Anova
=====================================
       F Value  Num DF  Den DF Pr > F
-------------------------------------
Store 6271.0000 2.0000 58.0000 0.0000
=====================================



There are significant differences in average daily sales between the three
stores.
```

```
Post Hoc Test Results:
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
========================================================
 group1  group2 meandiff p-adj   lower    upper   reject
--------------------------------------------------------
Store_A Store_B  -5.1667 0.0364 -10.0685  -0.2648   True
Store_A Store_C    -15.0    0.0 -19.9019 -10.0981   True
Store_B Store_C  -9.8333    0.0 -14.7352  -4.9315   True
--------------------------------------------------------
```

[ ]: