13 Mar Assignment

March 17, 2023

1 Assignment 37

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

Ans.ANOVA (Analysis of Variance) is a statistical technique used to test for significant differences between three or more groups. To use ANOVA, certain assumptions must be met, and any violation of these assumptions can impact the validity of the results. The four assumptions of ANOVA are:

1. Normality: The data within each group should follow a normal distribution.
2. Homogeneity of Variance: The variances of each group should be equal.
3. Independence: The observations within each group should be independent of each other.
4. Random Sampling: The data should be obtained through random sampling.

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

1. Violation of Normality: If the data within each group does not follow a normal distribution, the ANOVA results may not be accurate. For example, if the data is skewed or has outliers, it may violate this assumption. In such cases, it may be necessary to transform the data or use a non-parametric alternative to ANOVA.
2. Violation of Homogeneity of Variance: If the variances of each group are not equal, the ANOVA results may be affected. For example, if one group has a much larger variance than the others, it may affect the overall variance of the data, leading to inaccurate results. In such cases, it may be necessary to use a modified version of ANOVA or a non-parametric alternative.
3. Violation of Independence: If the observations within each group are not independent, it may impact the validity of the ANOVA results. For example, if the same subject is measured in multiple groups, the data within each group may not be independent. In such cases, it may be necessary to use a different statistical test that accounts for the lack of independence.
4. Violation of Random Sampling: If the data is not obtained through random sampling, the ANOVA results may not be generalizable to the population of interest. For example, if the data is obtained through convenience sampling, it may not be representative of the population. In such cases, it may be necessary to use a different statistical test or draw a random sample from the population.

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

Ans. The three types of ANOVA are:

1. One-Way ANOVA: One-way ANOVA is used when there is only one independent variable or factor with three or more levels. It is used to test whether there are any statistically significant differences between the means of two or more independent groups. For example, if we want to compare the mean weight loss of three different diets, we can use one-way ANOVA.
2. Two-Way ANOVA: Two-way ANOVA is used when there are two independent variables or factors. It is used to test for the effects of each variable on the dependent variable, as well as their interaction effects. For example, if we want to test the effects of both gender and age on the income of employees, we can use two-way ANOVA.
3. Three-Way ANOVA: Three-way ANOVA is used when there are three independent variables or factors. It is used to test for the main effects of each variable, as well as their interaction effects. For example, if we want to test the effects of different teaching methods, class sizes, and teacher qualifications on student test scores, we can use three-way ANOVA. #### In summary, the type of ANOVA used depends on the number of independent variables or factors being studied. One-way ANOVA is used when there is only one factor, two-way ANOVA is used when there are two factors, and three-way ANOVA is used when there are three factors.

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

Ans.The partitioning of variance in ANOVA refers to the process of decomposing the total variance in a data set into different sources of variation, including the variation within groups and the variation between groups. The variance between groups can be further decomposed into the variance explained by the independent variable and the residual variance.

It is important to understand the partitioning of variance in ANOVA because it provides insights into the sources of variation in a data set and helps to determine whether the independent variable has a significant effect on the dependent variable. By examining the proportion of variance explained by the independent variable and the residual variance, we can assess the strength of the relationship between the independent variable and the dependent variable.

Additionally, understanding the partitioning of variance can help in interpreting ANOVA results and communicating the findings to others. It enables researchers to explain how much of the variation in the data is due to the independent variable and how much is due to other factors or random error.

Overall, the partitioning of variance in ANOVA is an important concept as it helps researchers to better understand the relationship between variables and to draw meaningful conclusions from their data.

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?

Ans.In Python, you can calculate the total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR) in a one-way ANOVA using the statsmodels package. Here's how you can do it:

1.First, import the necessary packages:

- import pandas as pd
- import statsmodels.api as sm
- from statsmodels.formula.api import ols

2.Next, load your data into a pandas dataframe:

3.Assuming you have one independent variable and one dependent variable, you can run a one-way ANOVA using the ols function: model = ols('dependent_variable ~ independent_variable', data=data).fit()

4.Then, you can extract the total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR) using the anova_lm function:

- anova_table = sm.stats.anova_lm(model, typ=2)
- SST = anova_table['sum_sq'][0]
- SSE = anova_table['sum_sq'][1]
- SSR = anova_table['sum_sq'][2]

Here, typ=2 specifies that we want to calculate the sum of squares using the Type 2 method, which partitions the sum of squares based on the order of the terms in the model formula.

Note that the anova_lm function returns a table with various statistics, including the sum of squares for each source of variation. The first row of this table gives the total sum of squares (SST), which is the sum of squares of the deviation of each observation from the grand mean. The second row gives the explained sum of squares (SSE), which is the sum of squares of the deviation of each group mean from the grand mean. The third row gives the residual sum of squares (SSR), which is the sum of squares of the deviation of each observation from its group mean.

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

Ans.. To calculate the main effects and interaction effects in a two-way ANOVA using Python, you can use the statsmodels package. Here's how you can do it:

1.First, import the necessary packages:

- import pandas as pd
- import statsmodels.api as sm
- from statsmodels.formula.api import ols

2.Next, load your data into a pandas dataframe: data = pd.read_csv('my_data.csv')

3.Assuming you have two independent variables and one dependent variable, you can run a two-way ANOVA using the ols function: model = ols('dependent_variable ~ independent_variable_1 + independent_variable_2 + independent_variable_1 * independent_variable_2', data=data).fit()

4.Here, the * symbol specifies an interaction effect between independent_variable_1 and independent_variable_2.

5.Then, you can extract the main effects and interaction effects using the anova_lm function:

- anova_table = sm.stats.anova_lm(model, typ=2)
- main_effect_1 = anova_table['sum_sq'][0]
- main_effect_2 = anova_table['sum_sq'][1]
- interaction_effect = anova_table['sum_sq'][2]

Here, typ=2 specifies that we want to calculate the sum of squares using the Type 2 method, which partitions the sum of squares based on the order of the terms in the model formula.

Note that the anova_lm function returns a table with various statistics, including the sum of squares for each source of variation. The first row of this table gives the sum of squares for the main effect of independent_variable_1, the second row gives the sum of squares for the main effect of independent_variable_2, and the third row gives the sum of squares for the interaction effect between independent_variable_1 and independent_variable_2.

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?

Ans. If you conducted a one-way ANOVA and obtained an F-statistic of 5.23 and a p-value of 0.02, you can conclude that there is evidence of a significant difference between the groups.

The F-statistic is a measure of the ratio of the variance between the groups to the variance within the groups. A higher F-statistic indicates that the variance between the groups is relatively larger compared to the variance within the groups, suggesting that there may be significant differences between the groups.

The p-value indicates the probability of observing the obtained F-statistic or a more extreme F-statistic under the null hypothesis that there are no differences between the groups. A p-value of 0.02 suggests that the probability of observing the obtained F-statistic or a more extreme F-statistic under the null hypothesis is only 2%, which is lower than the commonly used threshold of 5% for rejecting the null hypothesis.

Therefore, based on these results, you can conclude that there are significant differences between the groups, and you can reject the null hypothesis that there are no differences between the groups. However, you would need to conduct further post-hoc tests to determine which specific groups differ significantly from each other.

It's important to note that ANOVA is only able to tell us if there is a significant difference between at least two groups, but it does not tell us which groups are significantly different. This is why further post-hoc tests are necessary.

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?

Ans. Handling missing data in repeated measures ANOVA is an important consideration because missing data can affect the validity and reliability of the results. There are different ways to handle missing data in repeated measures ANOVA, each with its own advantages and potential consequences.

One common approach to handle missing data is to use listwise deletion, which involves removing any participants with missing data from the analysis. This method can be simple to implement, but it can result in reduced statistical power if a large number of participants are excluded, and it can introduce bias if the missing data are not missing completely at random (MCAR). This means that the probability of missing data does not depend on any unobserved or observed variables.

Another approach is to use pairwise deletion, which involves analyzing only the available data for each pair of variables. This method can result in less loss of data and less reduction in statistical power compared to listwise deletion, but it can also lead to biased results if the missing data are not MCAR. Additionally, this method can result in different results depending on which pairs of variables are included in the analysis.

Another method is to use imputation, which involves estimating the missing data based on the observed data. Imputation methods can be based on simple techniques such as mean or median imputation, or more complex methods such as multiple imputation or maximum likelihood estimation. Imputation can help preserve statistical power and reduce bias, but it can also introduce additional uncertainty into the analysis, particularly if the imputation method is not appropriate for the missing data pattern or if the imputation model is misspecified.

Ultimately, the best method for handling missing data in repeated measures ANOVA depends on the nature of the missing data and the goals of the analysis. It is important to carefully consider the assumptions and potential consequences of each method and to document the method used and any assumptions made in the analysis.

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.

Ans.Post-hoc tests are used in ANOVA to determine which specific groups differ significantly from each other after a significant F-test has been obtained. There are several common post-hoc tests, each with its own assumptions and strengths.

1. Tukey's HSD (honestly significant difference): Tukey's HSD is a commonly used post-hoc test that can be used to compare all possible pairwise differences between group means. It has the advantage of controlling for Type I error rate, and it is appropriate when the sample sizes are equal and the variances are homogenous.
2. Bonferroni correction: The Bonferroni correction is a conservative method that adjusts the alpha level of each comparison by dividing the overall alpha level by the number of comparisons. This approach controls for the overall type I error rate, but it can be overly conservative and reduce statistical power.
3. Dunnett's test: Dunnett's test is used when there is one control group and several experimental groups, and the goal is to compare each experimental group to the control group. This method controls for the family-wise error rate, but it assumes that the variances in the experimental groups are equal.
4. Scheffe's test: Scheffe's test is a conservative method that can be used to compare all possible pairwise differences between group means. It has the advantage of controlling for the family-wise error rate, but it can have low statistical power. #### A situation where a post-hoc test might be necessary is when you have conducted an ANOVA and obtained a significant F-test, indicating that there are significant differences between at least two groups. However, the ANOVA itself does not tell you which specific groups differ significantly from each other. In this case, a post-hoc test can be used to determine which specific groups differ significantly from each other. For example, if you conducted an ANOVA on the effect of different types of fertilizer on plant growth and obtained a significant F-test, you would need to conduct a post-hoc test to determine which specific types of fertilizer led to significantly different plant growth.

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.

Ans.

[3]:
```
import numpy as np
from scipy.stats import f_oneway

# create sample data
diet_A = np.array([10, 12, 11, 9, 8, 10, 11, 12, 11, 13, 9, 8, 12, 10, 11, 12,␣
 ↪9, 11, 10, 11,
                   12, 9, 10, 11, 12, 10, 11, 9, 12, 10, 11, 9, 10, 12, 11, 8,␣
 ↪10, 9, 11, 12,
                   9, 10, 12, 11, 10, 11, 9, 12, 11, 10])
diet_B = np.array([8, 9, 7, 10, 12, 8, 9, 10, 11, 7, 10, 9, 11, 8, 10, 9, 12,␣
 ↪7, 11, 9,
```

```
                    10, 11, 7, 8, 9, 10, 11, 12, 8, 9, 7, 11, 10, 9, 8, 10, 11,␣
 ↪7, 9, 8,
                    10, 11, 9, 7, 8, 10, 11, 9, 7])
diet_C = np.array([6, 5, 7, 4, 3, 6, 5, 4, 3, 7, 4, 6, 5, 7, 6, 4, 5, 3, 6, 4,
                    5, 3, 7, 6, 5, 4, 3, 6, 7, 5, 4, 6, 5, 3, 4, 6, 7, 5, 4, 6,
                    5, 7, 4, 3, 6, 5, 4, 3])

# conduct ANOVA
f_stat, p_value = f_oneway(diet_A, diet_B, diet_C)

# print results
print("F-statistic:", f_stat)
print("p-value:", p_value)
```

```
F-statistic: 220.11794057678543
p-value: 1.6134190017836547e-44
```

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.

Ans.

[4]:
```
# Creating CSV File

import csv
import random

# Generate random task completion times for each program and experience level
program_list = ["Program A", "Program B", "Program C"]
exp_level_list = ["Novice", "Experienced"]
times_list = []

for i in range(30):
    for program in program_list:
        for exp_level in exp_level_list:
            time = random.randint(10, 60) # Generate random time between 10 and␣
 ↪60 seconds
            times_list.append([program, exp_level, time])

# Write data to CSV file
with open("task_completion_times.csv", "w", newline="") as csvfile:
    writer = csv.writer(csvfile)
```

```
    writer.writerow(["Program", "Experience Level", "Completion Time␣
 ↪(seconds)"])
    for row in times_list:
        writer.writerow(row)
```

[5]:
```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

# Read in CSV file
df = pd.read_csv("task_completion_times.csv")

# Conduct two-way ANOVA
model = ols('Q("Completion Time (seconds)") ~ C(Program) + C(Q("Experience␣
 ↪Level")) + C(Program):C(Q("Experience Level"))', data=df).fit()
anova_table = anova_lm(model, typ=2)

# Print ANOVA table
print(anova_table)
```

```
                                    sum_sq     df          F    PR(>F)
C(Program)                      116.877778    2.0   0.264304  0.768048
C(Q("Experience Level"))          4.050000    1.0   0.018317  0.892499
C(Program):C(Q("Experience Level"))  52.500000    2.0   0.118722  0.888127
Residual                        38472.233333  174.0        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.

Ans.

[6]:
```
# Generating CSV File

import csv
import random

# Set random seed for reproducibility
random.seed(123)

# Generate test scores for control group (traditional teaching method)
control_scores = [random.randint(60, 100) for _ in range(50)]

# Generate test scores for experimental group (new teaching method)
```

```
experimental_scores = [random.randint(65, 105) for _ in range(50)]

# Combine the scores and group information into a list of tuples
data = [("control", score) for score in control_scores] + [("experimental",␣
 ↪score) for score in experimental_scores]

# Shuffle the data and save it to a CSV file
random.shuffle(data)
with open("test_scores.csv", "w", newline="") as f:
    writer = csv.writer(f)
    writer.writerow(["group", "score"])
    writer.writerows(data)
```

[7]:
```
import pandas as pd
from scipy.stats import ttest_ind, f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Load data from CSV file
data = pd.read_csv("test_scores.csv")

# Split the data into two groups based on the "group" column
control_group = data[data["group"] == "control"]["score"]
experimental_group = data[data["group"] == "experimental"]["score"]

# Conduct a two-sample t-test
t_stat, p_val = ttest_ind(control_group, experimental_group)
print("t-statistic:", t_stat)
print("p-value:", p_val)

# Conduct a post-hoc test
f_stat, p_val = f_oneway(control_group, experimental_group)
print("F-statistic:", f_stat)
print("p-value:", p_val)
tukey_result = pairwise_tukeyhsd(data["score"], data["group"])
print(tukey_result)
```

```
t-statistic: -4.143232845439682
p-value: 7.271705857841821e-05
F-statistic: 17.166378411530207
p-value: 7.271705857841854e-05
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
 group1    group2    meandiff p-adj  lower   upper  reject
-------------------------------------------------------------
control experimental    10.14 0.0001 5.2833 14.9967   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.

Ans.

[8]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols

# create a dataframe with the sales data
data = pd.DataFrame({
    'store': ['A'] * 30 + ['B'] * 30 + ['C'] * 30,
    'sales': np.random.randint(100, 1000, size=90)
})

# fit the ANOVA model
model = ols('sales ~ C(store)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

# print the ANOVA table
print(anova_table)
```

```
                sum_sq    df         F    PR(>F)
C(store)  4.595756e+03   2.0  0.033912  0.966669
Residual  5.895122e+06  87.0       NaN       NaN
```