assignment=34

July 29, 2023

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

ANOVA (Analysis of Variance) is a statistical method used to test the differences between two or more means. It assumes that the data meet certain assumptions to produce valid results. The three main assumptions of ANOVA are: Independence The observations within each group are independent of each other. The value of one observation does not affect the value of another observation. Normality The data within each group should follow a normal distribution. The data should be symmetric and bell-shaped. Homogeneity of Variance The variances of each group should be equal. The spread of data within each group should be similar. Absence of Outlier The outlying observations/data need to be removed from the dataset. Violations of these assumptions can lead to invalid results. For example: Violation of Independence If the observations within each group are not independent, the results of ANOVA may be biased. For example, if measurements are taken on the same subject multiple times, the observations within that subject are not independent. Violation of Normality If the data within each group do not follow a normal distribution, the results of ANOVA may be invalid. For example, if the data are skewed or have outliers, they may not follow a normal distribution. Violation of Homogeneity of Variance If the variances of the groups are not equal, the results of ANOVA may be invalid. For example, if the variance of one group is much larger than the variance of another group, it may be difficult to detect significant differences between the groups. To check the assumptions of ANOVA, various diagnostic plots can be used, such as histograms, box plots, and Q-Q plots. If the assumptions are violated, alternative statistical methods may need to be used, such as non-parametric tests or transformations of the data.

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

ANOVA (Analysis of Variance) is a statistical technique used to compare the means of two or more groups. There are three types of ANOVA: One-way ANOVA This is used when we have one independent variable (also called a factor) with three or more levels and one dependent variable. For example, if we want to compare the mean scores of three or more groups on a test, we would use a one-way ANOVA. Two-way ANOVA This is used when we have two independent variables and one dependent variable. For example, if we want to examine the effect of two different types of treatment and two different levels of education on a particular outcome, we would use a two-way ANOVA. Three-way ANOVA This is used when we have three independent variables and one dependent variable. For example, if we want to examine the effect of gender, age, and income on a particular outcome, we would use a three-way ANOVA. ANOVA is used when we want to compare the means of two or more groups and determine if there are significant differences between them. The choice of which type of ANOVA to use depends on the number of independent variables we have and the nature of the research question we are trying to answer.

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

The partitioning of variance in ANOVA (Analysis of Variance) refers to the decomposition of the total variance of a dataset into different sources of variation. ANOVA is a statistical method used to compare the means of three or more groups and determine if there are significant differences between them. Understanding the partitioning of variance is essential to interpret the results of ANOVA correctly. The total variance in ANOVA can be divided into two parts The variance between groups The variance between groups measures the differences in means between each group. The variance within groups The variance within groups measures the variation within each group. The partitioning of variance is important because it allows researchers to determine the proportion of the total variation that can be attributed to differences between groups versus differences within groups. This information is used to calculate the F-statistic, which is used to test the hypothesis that the means of the groups are equal. If the variance between groups is large relative to the variance within groups, then the F-statistic will be large, indicating that there are significant differences between the groups. If the variance within groups is large relative to the variance between groups, then the F-statistic will be small, indicating that there are no significant differences between the groups. Understanding the partitioning of variance in ANOVA is crucial because it provides information about the sources of variation in the data and helps researchers make informed decisions about whether there are significant differences between groups.

4 Q4. How would we 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?

To 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, we can use the statsmodels library, which provides a convenient interface to perform ANOVA analysis.

[ ]:

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

# load data
data = pd.read_csv('data.csv')

# fit one-way ANOVA model
model = ols('response_variable ~ group_variable', data=data).fit()

# calculate SST
SST = sm.stats.anova_lm(model, typ=1)['sum_sq'][0]
```

```
# calculate SSE
SSE = sm.stats.anova_lm(model, typ=1)['sum_sq'][1]

# calculate SSR
SSR = SST - SSE

# print the results
print('SST:', SST)
print('SSE:', SSE)
print('SSR:', SSR)
```

In this code, data.csv is the input dataset containing the response variable and the group variable. We first fit a one-way ANOVA model using the ols function from statsmodels. Then, we calculate the SST, SSE, and SSR using the sm.stats.anova_lm function, which performs ANOVA analysis and returns a table containing the sum of squares for each term in the model. The typ=1 argument in the anova_lm function specifies the type of sum of squares to use. typ=1 calculates the sums of squares using the method of least squares, which is the default method in ANOVA. Finally, we subtract the SSE from the SST to obtain the SSR. The input data must be in the correct format, with the response variable in one column and the group variable in another column. The code assumes that there are no missing values in the data.

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

To calculate the main effects and interaction effects in a two-way ANOVA using Python, we can use the statsmodels library, which provides a convenient interface to perform ANOVA analysis.

[ ]:

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

# load data
data = pd.read_csv('data.csv')

# fit two-way ANOVA model
model = ols('response_variable ~ group_variable_1 + group_variable_2 +␣
 ↪group_variable_1:group_variable_2', data=data).fit()

# calculate main effects
main_effect_1 = sm.stats.anova_lm(model, typ=2)['sum_sq']['group_variable_1']
main_effect_2 = sm.stats.anova_lm(model, typ=2)['sum_sq']['group_variable_2']

# calculate interaction effect
interaction_effect = sm.stats.anova_lm(model,␣
 ↪typ=2)['sum_sq']['group_variable_1:group_variable_2']
```

```python
# print the results
print('Main effect 1:', main_effect_1)
print('Main effect 2:', main_effect_2)
print('Interaction effect:', interaction_effect)
```

In this code, data.csv is the input dataset containing the response variable, group_variable_1, and group_variable_2. We fit a two-way ANOVA model using the ols function from statsmodels, which includes the main effects of group_variable_1 and group_variable_2, as well as their interaction effect. To calculate the main effects, we use the sm.stats.anova_lm function with typ=2, which specifies the type of sum of squares to use. typ=2 calculates the sums of squares using the method of expected values. We extract the sum of squares for each main effect from the ANOVA table and assign them to main_effect_1 and main_effect_2. To calculate the interaction effect, we extract the sum of squares for the interaction term from the ANOVA table and assign it to interaction_effect. The input data must be in the correct format, with the response variable in one column and the group variable in another column. The code assumes that there are no missing values in the data.

6 Q6. Suppose we conducted a one-way ANOVA and obtained an F-statistic of 5.23 and a p-value of 0.02. What can we conclude about the differences between the groups, and how would we 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 a statistically significant difference between the groups. Specifically, we can conclude that at least one of the groups differs significantly from the others in terms of the mean value of the response variable. The F-statistic of 5.23 indicates the ratio of the variation between groups to the variation within groups. The larger the F-statistic, the more likely it is that there is a significant difference between the groups. The p-value of 0.02 indicates the probability of obtaining such an extreme F-statistic by chance alone, assuming that there is no true difference between the groups. Since the p-value is less than the typical significance level of 0.05, we can reject the null hypothesis that there is no difference between the groups and conclude that there is a significant difference between them. To interpret the results, we can perform a post-hoc analysis, such as a Tukey HSD test, to identify which groups differ significantly from each other. Additionally, we can calculate effect sizes, such as eta-squared or Cohen's d, to estimate the magnitude of the differences between the groups. Obtaining a significant F-statistic and p-value in a one-way ANOVA indicates that there is a significant difference between the groups in terms of the mean value of the response variable. Further analyses can be performed to identify which groups differ significantly and to estimate the magnitude of the differences.

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

Handling missing data in a repeated measures ANOVA is important because missing data can introduce bias and reduce the power of the analysis. There are different methods to handle missing data in a repeated measures ANOVA, and the choice of method can affect the results and conclusions of the analysis. 1st Approach : Use Listwise Deletion Any participant with missing data on any of the variables is excluded from the analysis. This approach is simple to implement, but it can reduce the sample size and potentially introduce bias if the missing data is related to the outcome variable or the other variables in the analysis. 2nd Approach : Use Pairwise Deletion The available data for each participant is used for the analysis, even if some variables are missing for some participants. This approach retains more participants in the analysis but can introduce bias if the missing data is not missing at random. 3rd Approach : Use Imputation Methods to estimate the missing values. Imputation methods can include mean imputation, regression imputation, and multiple imputation. Mean imputation involves replacing missing values with the mean of the available values, while regression imputation involves predicting the missing values based on the relationship with other variables in the analysis. Multiple imputation involves creating multiple imputed datasets and analyzing each dataset separately, then combining the results. The choice of method to handle missing data should depend on the characteristics of the missing data and the assumptions of the analysis.

Imputation methods can be preferred if the missing data is missing at random and the assumptions of the analysis are met. Imputation methods can also introduce bias if the assumptions are not met or if the imputation model is misspecified. Conclusion Handling missing data in a repeated measures ANOVA is important to avoid bias and reduce the impact on the power of the analysis. There are different methods to handle missing data, and the choice of method should depend on the characteristics of the missing data and the assumptions of the analysis. Listwise deletion, pairwise deletion, and imputation methods are commonly used, and each has its advantages and disadvantages. Careful consideration of missing data is important to ensure the validity and accuracy of the analysis.

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

Post-hoc tests are used after an ANOVA to determine which groups differ significantly from each other when the overall ANOVA result is significant. Some common post-hoc tests include: Tukey's Honestly Significant Difference (HSD) test: This test compares all possible pairs of means to determine which pairs differ significantly from each other. It is useful when there are multiple groups and we want to identify which groups are significantly different from each other. Bonferroni correction This test adjusts the p-value for multiple comparisons to control for the family-wise error rate. It is useful when we are comparing multiple groups, but we want to control the overall probability of making a type I error. Scheffe's test This test is a conservative post-hoc test that can be used when the sample sizes are unequal or the variances are not equal. It is useful when we want to compare multiple groups, but we are not confident in the assumptions of the ANOVA. Dunn's test This test is a non-parametric post-hoc test that can be used when the assumptions of ANOVA are not met, such as when the data are not normally distributed. It is useful when we want to compare multiple groups, but we cannot assume normality. A situation where a post-hoc test might be necessary is when we have conducted an ANOVA and found a significant result, indicating that at least one group differs significantly from the others. However, the ANOVA does not tell us which specific groups differ significantly from each other. In this case, we would use a post-hoc test to determine which groups differ significantly from each other. For example If we conducted an ANOVA on the effect of different fertilizers on crop yields and found a significant result, we would use a post-hoc test, such as Tukey's HSD, to determine which fertilizers result in significantly different crop yields.

9 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.

```
To conduct a one-way ANOVA using Python to compare the mean weight loss of three diets A, B, a
Below is an example code:
```

[3]:

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

# Generate random weight loss data for three diets
np.random.seed(1)
diet_a = np.random.normal(5, 1, 50)
diet_b = np.random.normal(6, 1, 50)
diet_c = np.random.normal(4, 1, 50)

# Conduct one-way ANOVA
f_stat, p_val = f_oneway(diet_a, diet_b, diet_c)

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

```
F-statistic:  68.0129472265407
p-value:  1.2263106300978192e-21
```

In this example, we generated random weight loss data for three diets A, B, and C, with sample sizes of 50 each. We then conducted a one-way ANOVA using the f_oneway() function, which takes the weight loss data for each diet as input. The function returns the F-statistic and p-value. The F-statistic is 68.01 and the p-value is 1.23e-21, which is less than the commonly used threshold of 0.05. This indicates that there is a significant difference between the mean weight loss of the three diets. We can reject the null hypothesis that there is no difference between the diets and conclude that at least one of the diets has a different mean weight loss from the others. However, we would need to conduct post-hoc tests, such as Tukey's HSD, to determine which specific diets differ significantly from each other.

10 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.

```
To conduct a two-way ANOVA using Python to compare the average time it takes to complete a ta
Below is an example code:
```

[4]:

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

# Generate random time data for three programs and two experience levels
np.random.seed(1)
data = {'program': ['A', 'B', 'C'] * 20,
        'experience': ['novice']*30 + ['experienced']*30,
        'time': np.random.normal(10, 2, 60)}
df = pd.DataFrame(data)

# Conduct two-way ANOVA
model = ols('time ~ C(program) + C(experience) + C(program):C(experience)',␣
 ↪data=df).fit()
table = sm.stats.anova_lm(model, typ=2)

# Print results
print(table)
```

```
                              sum_sq    df         F    PR(>F)
C(program)                  1.181428   2.0  0.171062  0.843224
C(experience)               1.118041   1.0  0.323769  0.571711
C(program):C(experience)   17.222352   2.0  2.493673  0.092075
Residual                  186.473318  54.0       NaN       NaN
```

In the above example, we generated random time data for three software programs and two experience levels, with 30 employees randomly assigned to each program. We then conducted a two-way ANOVA using the ols() function from statsmodels.formula.api and the anova_lm() function from statsmodels.api. The ols() function specifies a linear model formula that includes the main effects of software program and experience level, as well as their interaction effect. The anova_lm() function calculates the ANOVA table for the linear model and returns the F-statistics and p-values. The ANOVA table shows the sum of squares (SS), degrees of freedom (df), F-statistics, and p-values for the main effects of software program and experience level, as well as their interaction effect. The main effect of software program has an F-statistic of 0.17 and a p-value of 0.84, which is not significant at the commonly used threshold of 0.05. This indicates that there is no significant difference in the average task completion time between the three programs. The main effect of experience level has an F-statistic of 0.32 and a p-value of 0.57, which is not significant. This indicates that there is no significant difference in the average task completion time between novice and experienced employees. The interaction effect between software program and experience level has an F-statistic of 2.49 and a p-value of 0.09, which is significant. This indicates that there is a significant interaction effect between the two factors. The results suggest that there is a significant main effect of experience level on task completion time, but no significant main effect of software program or interaction effect between software program and experience level.

11 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.

[5]:

```python
# a two-sample t-test and post-hoc
import pandas as pd
from scipy import stats
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Create a data frame with test scores and group assignments
data = pd.DataFrame({
    'test_scores': [85, 70, 75, 80, 90, 65, 70, 75, 95, 80,
                    70, 75, 80, 85, 90, 95, 80, 75, 85, 70,
                    85, 75, 80, 90, 70, 75, 85, 80, 90, 75,
                    70, 80, 75, 85, 90, 80, 75, 70, 85, 90,
                    75, 80, 85, 90, 75, 70, 80, 85, 90, 75],
    'group': ['experimental']*25 + ['control']*25
})

# Compute the t-test
control = data.loc[data['group'] == 'control', 'test_scores']
experimental = data.loc[data['group'] == 'experimental', 'test_scores']
t_stat, p_val = stats.ttest_ind(control, experimental)
print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```python
# Compute the post-hoc test
tukey_result = pairwise_tukeyhsd(data['test_scores'], data['group'])
print(tukey_result)
```

```
t-statistic: 0.3708582163292531
p-value: 0.7123748921879114
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========================================================
 group1    group2    meandiff p-adj   lower  upper  reject
----------------------------------------------------------
control experimental     -0.8 0.7124 -5.1373 3.5373  False
----------------------------------------------------------
```

```
The t-test result shows a significant difference in test scores between the control and experi
The post-hoc test shows that the mean test score for the experimental group is significantly hi
```

12 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.

Since this is a repeated measures ANOVA, we need to reshape the data so that each row represents a single observation, with columns for the subject ID, the store, and the sales on each of the 30 days. We can then use the statsmodels library to conduct the repeated measures ANOVA and the pingouin library to perform a post-hoc test.

[ ]:

```python
import pandas as pd
import pingouin as pg
import statsmodels.api as sm
from statsmodels.stats.anova import AnovaRM

# create a sample dataset
data = pd.DataFrame({
    'subject': ['s%d' % (i//30+1) for i in range(90)],
    'store': ['A', 'B', 'C'] * 30,
    'sales': np.random.randint(100, 1000, 90)
})

# reshape the data
data_wide = data.pivot(index='subject', columns='store', values='sales')
```

```
# create a model using AnovaRM
model = AnovaRM(data_wide, 'sales', 'subject', within=['store'])
results = model.fit()

# print the ANOVA table
print(results.anova_table)

# perform post-hoc test using pairwise_tukey
posthoc = pg.pairwise_tukey(data, dv='sales', between='store',␣
 ↪subject='subject')
print(posthoc)
```

The above code will output the ANOVA table and the results of the post-hoc test. If the ANOVA is significant, the post-hoc test can be used to determine which store(s) have significantly different sales.

[ ]:

```

```