# Statistics Advance-6

January 27, 2024

## 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 test used to determine whether there are any significant differences between means of two or more groups. The basic assumptions of ANOVA are as follows:

Normality: The data should be normally distributed within each group.

Homogeneity of variance: The variance of the data in each group should be approximately equal.

Independence: The observations in each group should be independent of each other.

If any of these assumptions are violated, the results of the ANOVA may not be valid. Here are some examples of violations that could impact the validity of the results:

Non-normality: If the data in any of the groups is not normally distributed, this can affect the validity of the ANOVA results. For example, if the data in one group is skewed, the mean of that group may not be a good representation of the data.

Heterogeneity of variance: If the variance of the data in each group is significantly different, this can affect the validity of the ANOVA results. For example, if one group has a much larger variance than the other groups, this can make it more difficult to detect differences between means.

Dependence: If the observations in any of the groups are not independent, this can affect the validity of the ANOVA results. For example, if the observations in one group are paired (e.g., before and after measurements), this violates the independence assumption and can lead to spurious results.

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

There are three types of ANOVA:

1. One-way ANOVA: This type of ANOVA is used when there is only one independent variable, which has three or more levels (or groups). One-way ANOVA is used to determine if there are any significant differences between the means of the different levels/groups. For example, one-way ANOVA can be used to compare the average weight loss among three different diet groups.

2. Two-way ANOVA: This type of ANOVA is used when there are two independent variables, both of which have two or more levels. Two-way ANOVA is used to determine if there are any significant main effects (i.e., the effect of each independent variable on the dependent variable) and/or interaction effects (i.e., the combined effect of the two independent variables on the dependent variable). For example, two-way ANOVA can be used to compare the average sales of two different products (product A and product B) in two different regions (North and South).

3. MANOVA (Multivariate Analysis of Variance): This type of ANOVA is used when there are two or more dependent variables (i.e., outcome variables) and one or more independent variables. MANOVA is used to determine if there are any significant differences between the means of the dependent variables across the different levels of the independent variable. For example, MANOVA can be used to compare the average scores on multiple personality traits (e.g., extroversion, agreeableness, neuroticism) between different age groups (young, middle-aged, and old).

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

Partitioning of variance in ANOVA (Analysis of Variance) refers to the process of decomposing the total variation in the data into different sources of variation. In other words, ANOVA breaks down the total variation in the dependent variable into the variation explained by the independent variable(s) and the variation that is not explained by the independent variable(s). This is important because it helps us understand how much of the variation in the dependent variable can be attributed to the independent variable(s).

The partitioning of variance in ANOVA is typically represented by the following equation:

Total variation = Variation explained by independent variable(s) + Variation not explained by independent variable(s)

The variation explained by the independent variable(s) is referred to as the “between-group” variation, while the variation not explained by the independent variable(s) is referred to as the “within-group” variation.

Understanding the partitioning of variance is important in ANOVA because it helps us determine if there is a significant difference between the means of the different groups being compared. If the variation explained by the independent variable(s) (i.e., the between-group variation) is much larger than the variation not explained by the independent variable(s) (i.e., the within-group variation), then it suggests that there is a significant difference between the means of the groups being compared.

## 4 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?

[1]:

```python
import scipy.stats as stats

# create three sample groups
group1 = [10, 12, 14, 16, 18]
group2 = [8, 11, 14, 17, 20]
group3 = [9, 12, 15, 18, 21]

# concatenate the groups
data = group1 + group2 + group3

# calculate the mean of the data
mean = sum(data) / len(data)

# calculate the total sum of squares (SST)
SST = sum([(x - mean)**2 for x in data])

# calculate the sum of squares between (SSB)
SSB = len(group1) * (sum([(x - mean)**2 for x in group1]) / len(group1))
SSB += len(group2) * (sum([(x - mean)**2 for x in group2]) / len(group2))
SSB += len(group3) * (sum([(x - mean)**2 for x in group3]) / len(group3))

# calculate the explained sum of squares (SSE)
SSE = SSB

# calculate the residual sum of squares (SSR)
SSR = SST - SSE

print("SST =", SST)
print("SSE =", SSE)
print("SSR =", SSR)
```

```
SST = 223.33333333333337
SSE = 223.33333333333337
SSR = 0.0
```

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

[ ]:

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

```
# Load data into a pandas dataframe
data = pd.read_csv('data.csv')

# Fit the two-way ANOVA model
model = ols('dependent_variable ~ independent_variable_1 +␣
 ↪independent_variable_2 + independent_variable_1 * independent_variable_2',␣
 ↪data=data).fit()

# Calculate the main effects
main_effects = sm.stats.anova_lm(model, typ=1)
print(main_effects)

# Calculate the interaction effect
interaction_effect = sm.stats.anova_lm(model, typ=2)
print(interaction_effect)
```

# 6 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 at least one significant difference between the means of the groups being compared. The F-statistic tells us the ratio of the variance between the groups to the variance within the groups. A larger F-statistic indicates a greater difference between the group means compared to the variation within the groups. The p-value tells us the probability of obtaining a result as extreme as the observed result if there were no true differences between the groups.

With a p-value of 0.02, we can interpret this result as follows: if there were no true differences between the groups, we would only expect to obtain a result as extreme as an F-statistic of 5.23 or higher by chance 2% of the time. Therefore, we reject the null hypothesis (that there are no true differences between the groups) and conclude that there is sufficient evidence to suggest that at least one of the group means is different from the others.

# 7 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 be handled in several ways:

1. Listwise deletion: This method involves excluding any cases with missing data from the analysis. This can be done using the dropna() function in pandas. While this approach is simple, it may result in a loss of statistical power if a large amount of data is missing.
2. Mean imputation: This method involves replacing missing values with the mean of the non-

missing values. This can be done using the fillna() function in pandas. While this approach is simple and easy to implement, it may underestimate the variability of the data and result in biased estimates.

3. Last observation carried forward (LOCF): This method involves imputing missing values with the last observed value. This can be done using the fillna(method='ffill') function in pandas. While this approach is useful for data with a temporal order, it may not be appropriate for all situations and may result in biased estimates.
4. Multiple imputation: This method involves imputing missing values multiple times using a statistical model, and then combining the results to obtain estimates and standard errors. This can be done using the fancyimpute library in Python. While this approach is more sophisticated and can produce more accurate estimates than mean imputation or LOCF, it is computationally intensive and requires careful consideration of the underlying assumptions.

The potential consequences of using different methods to handle missing data in a repeated measures ANOVA include bias in the estimated means, standard errors, and effect sizes, as well as a loss of statistical power. It's important to carefully consider the underlying assumptions and potential limitations of each method and choose the approach that is most appropriate for the specific dataset and research question. Additionally, it may be beneficial to conduct sensitivity analyses to assess the robustness of the results to different methods of handling missing data.

## 8 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 used after ANOVA to make pairwise comparisons between groups when the overall ANOVA result is statistically significant. The purpose of post-hoc tests is to determine which specific groups differ from each other and to control for the familywise error rate, which is the probability of making at least one Type I error (false positive) across all the pairwise comparisons. Here are some common post-hoc tests used after ANOVA, along with an example of a situation where each one might be necessary:

1. Tukey's HSD (honestly significant difference) test: This test is a conservative post-hoc test that is commonly used when the sample sizes are equal across groups. It controls for the familywise error rate by adjusting the significance level for each pairwise comparison. For example, if we have four groups (A, B, C, D), and the overall ANOVA result is significant, we might use Tukey's HSD test to determine which specific groups differ from each other. If the test shows that group A is significantly different from group B and group C, but not group D, we can conclude that group A is significantly different from groups B and C, but not group D.
2. Bonferroni correction: This test is a simple and commonly used method to adjust the significance level for each pairwise comparison. It divides the significance level (usually 0.05) by the number of comparisons being made. For example, if we have four groups (A, B, C, D), and the overall ANOVA result is significant, we might use the Bonferroni correction to determine which specific groups differ from each other. If the test shows that group A is significantly different from group B, group C, and group D, we can conclude that group A is significantly different from all the other groups.

3. Dunnett's test: This test is used when we have one control group and several treatment groups. It compares each treatment group to the control group, while controlling for the overall familywise error rate. For example, if we have one control group and three treatment groups (A, B, C), and the overall ANOVA result is significant, we might use Dunnett's test to determine which specific treatment groups differ from the control group. If the test shows that group A is significantly different from the control group, but groups B and C are not significantly different from the control group, we can conclude that group A is significantly different from the control group, but groups B and C are not.

4. Scheffe's test: This test is a conservative post-hoc test that is used when the sample sizes are unequal across groups. It controls for the familywise error rate by adjusting the significance level for each pairwise comparison. For example, if we have four groups (A, B, C, D), and the overall ANOVA result is significant, we might use Scheffe's test to determine which specific groups differ from each other. If the test shows that group A is significantly different from group B and group C, but not group D, we can conclude that group A is significantly different from groups B and C, but not group D.

## 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.

[4]:

```python
import scipy.stats as stats

# Sample data
diet_A = [4.2, 5.1, 3.7, 6.2, 4.9, 2.8, 5.4, 4.7, 3.9, 4.1,
          3.3, 5.5, 4.8, 5.3, 3.9, 6.1, 4.3, 3.5, 5.2, 5.0,
          5.6, 3.8, 5.8, 4.0, 5.7]
diet_B = [2.9, 1.9, 2.5, 2.3, 3.2, 2.7, 1.8, 3.1, 2.6, 3.0,
          2.4, 2.0, 2.8, 2.2, 2.1, 2.6, 2.7, 1.6, 2.4, 2.9,
          3.0, 1.8, 2.1, 2.3, 2.6]
diet_C = [1.3, 0.9, 1.1, 1.6, 1.4, 1.8, 1.2, 1.5, 2.0, 1.4,
          1.7, 1.3, 1.9, 1.0, 1.6, 1.2, 1.8, 1.4, 1.1, 1.7,
          1.5, 1.6, 1.9, 1.2, 1.4]

# One-way ANOVA
f_statistic, p_value = stats.f_oneway(diet_A, diet_B, diet_C)

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

```
F-statistic: 176.89689491604346
```

```
p-value: 1.632278437376119e-28
```

## 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 in Python, we first need to import the necessary packages and read in the data:

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

data = pd.read_csv("task_completion_times.csv")
```

Next, we can use the ols function from the statsmodels.formula.api module to create a model formula for the two-way ANOVA:

[ ]:
```
model = ols('Time ~ C(Program) + C(Experience) + C(Program):C(Experience)',␣
  ↪data=data).fit()
```

Here, we're using the C function to indicate that Program and Experience are categorical variables, and : to indicate the interaction term between Program and Experience.

We can then use the anova_lm function from the statsmodels.api module to obtain the ANOVA table and calculate the F-statistics and p-values:

[ ]:
```
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

## 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.

[ ]:

```
# To conduct a two-sample t-test in Python, we first need to import the necessary packages and read in the data:
import pandas as pd
from scipy.stats import ttest_ind

data = pd.read_csv("test_scores.csv")

#The data should be in a CSV file with two columns: "Group" (control or experimental) and "Score" (the test score).
#We can then use the ttest_ind function from the scipy.stats module to conduct a two-sample t-test:


control_scores = data[data["Group"] == "control"]["Score"]
experimental_scores = data[data["Group"] == "experimental"]["Score"]
t, p = ttest_ind(control_scores, experimental_scores)
print("t = {:.2f}, p = {:.4f}".format(t, p))
```

## 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 n 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- h c test to determine which store(s) differ significantly from each other.

[ ]:

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

# Create a dataframe with sales data
sales_data = {'Store': ['A']*30 + ['B']*30 + ['C']*30,
              'Day': list(range(1, 31))*3,
              'Sales': [100, 120, 130, 110, 130, 140, 90, 110, 120, 80,
                        100, 130, 140, 120, 140, 150, 110, 130, 140, 100,
                        90, 110, 120, 100, 120, 130, 80, 100, 110, 70,
                        80, 100, 110, 90, 110, 120, 70, 90, 100, 60,
                        70, 90, 100, 80, 100, 110, 60, 80, 90, 50,
                        60, 80, 90, 70, 90, 100, 50, 70, 80, 40,
                        50, 70, 80, 60, 80, 90, 40, 60, 70, 30,
                        40, 60, 70, 50, 70, 80, 30, 50, 60, 20,
                        30, 50, 60, 40, 60, 70, 20, 40, 50, 10,
                        20, 40, 50, 30, 50, 60, 10, 30, 40, 0]}

sales_df = pd.DataFrame(sales_data)

# Conduct one-way repeated measures ANOVA
rm_anova = sm.stats.anova.AnovaRM(data=sales_df, depvar='Sales', subject='Day',
                                  within=['Store'], aggregate_func='mean')
result = rm_anova.fit()

# Print ANOVA table
print(result.summary())

# Conduct post-hoc tests using Tukey's HSD test
from statsmodels.stats.multicomp import MultiComparison

comp = MultiComparison(sales_df['Sales'], sales_df['Store'])
tukey_result = comp.tukeyhsd()

# Print post-hoc results
print(tukey_result.summary())
```

[ ]:

```
```