13 March Assignmnet

May 16, 2023

[ ]:
```
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 method used to compare␣
↪the means of three or more groups. The assumptions required to
use ANOVA are:.

Normality: The data should be normally distributed.

Homogeneity of variance: The variance of the dependent variable should be equal␣
↪across all levels of the independent variable.

Independence: Observations should be independent of each other.

If these assumptions are not met, the validity of the results may be impacted.␣
↪For example, if the normality assumption is violated, then the
results may not be reliable. Similarly, if the homogeneity of variance␣
↪assumption is violated, then the results may not be accurate.
```

[ ]:

[ ]:

[ ]:
```
Q2. What are the three types of ANOVA, and in what situations would each be␣
↪used?
```

ANS -

[ ]:
```
       There are three types of ANOVA: **One-way ANOVA**, **Two-way ANOVA␣
↪without replication**, and **Two-way ANOVA with replication**.

-One-way ANOVA : is used when you want to test two groups to see if there's a␣
↪difference between them.
```

```
-Two-way ANOVA without replication : is used when you have one group and you’re␣
↪double-testing that same group.

-Two-way ANOVA with replication : is used when you have two groups, and the␣
↪members of those groups are doing more than one thing.
```

[ ]:

[ ]:

```
[ ]: 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 is a statistical technique␣
↪that partitions the total variance of a dataset into different
components that are associated with different sources of variation. The␣
↪observed variance in a particular variable is partitioned into
components attributable to different sources of variation. In its simplest␣
↪form, ANOVA provides a statistical test of whether two or more
population means are equal, and therefore generalizes the t-test beyond two␣
↪means1.

    The partitioning of variance is important because it allows us to determine␣
↪which sources of variation are significant and which are not.
This information can be used to identify factors that are contributing to the␣
↪variability in the data and to develop strategies for reducing
this variability
```

[ ]:

[ ]:

```
[ ]: 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 -

```
[39]: import numpy as np
import pandas as pd
from scipy.stats import f

# create a sample dataset
data = {'group': ['A', 'A', 'B', 'B', 'C', 'C'],
        'value': [1, 2, 3, 4, 5, 6]}
```

```
df = pd.DataFrame(data)

# calculate the mean of all observed values
ymean = df['value'].mean()

# calculate the predicted value for each observation
df['yhat'] = df.groupby('group')['value'].transform('mean')

# calculate SSE
df['sse'] = (df['value'] - df['yhat'])**2
sse = df['sse'].sum()

# calculate SSR
df['ssr'] = (df['yhat'] - ymean)**2 * len(df['value'].unique())
ssr = df['ssr'].sum()

# calculate SST
sst = sse + ssr

print(f"SSE: {sse}")
print(f"SSR: {ssr}")
print(f"SST: {sst}")
```

```
SSE: 1.5
SSR: 96.0
SST: 97.5
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
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 -

[ ]:

```
            In one-way ANOVA, the F-statistic is used to test the null␣
  ↪hypothesis that all group means are equal. The p-value is the probability
of observing a test statistic as extreme as the one calculated from your data,␣
  ↪assuming that the null hypothesis is true. If the p-value is
less than the significance level (usually 0.05), then you can reject the null␣
  ↪hypothesis and conclude that there is evidence of a difference
between at least two groups .
```

```
  An F-statistic of 5.23 and a p-value of 0.02 suggests that there is
evidence of a difference between at least two groups [2]. However, its
important to note that ANOVA only tells you whether there is a difference
between groups, not which groups are different from each other.
To determine which groups are different from each other, you would need to
perform post-hoc test.
```

[ ]:

[ ]:

[ ]:
```
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 -

[ ]:
```
      In repeated measures ANOVA, missing data can be a serious problem. One
of the biggest problems with traditional repeated measures
ANOVA is missing data on the response variable. The problem is that repeated
measures ANOVA treats each measurement as a separate variable.
Because it uses listwise deletion, if one measurement is missing, the entire
case gets dropped.

   There are different methods to handle missing data in repeated measures
ANOVA. One method is to use listwise deletion which means that if
any of the observations for a subject are missing, the entire subject will be
omitted from the analysis. Another method is to use imputation
methods such as mean imputation or regression imputation.

  The potential consequences of using different methods to handle missing data
are that it can lead to biased estimates of parameters and
standard errors1. It can also lead to a loss of power and efficiency 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 -

[ ]:
```
  There are several post-hoc tests that can be used after ANOVA. Some of the
most common ones are:
```

```
-Bonferroni Procedure
-Duncan's new multiple range test (MRT)
-Fisher's Least Significant Difference (LSD)
-Holm-Bonferroni Procedure
-Newman-Keuls
-Rodger's Method
-Scheffe's Method
-Tukey's Procedure
-Each of these tests has its own strengths and weaknesses. For example, Tukey's
↪HSD ("honestly significant difference") is the most common
post hoc test for ANOVA. It is useful when you want to make every possible
↪pairwise comparison.

   A post-hoc test might be necessary when you have found a significant
↪difference between groups in an ANOVA analysis. The post-hoc test will
help you determine which groups are significantly different from each other.
```

[ ]:

[ ]:

[ ]:
```
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 -

[20]:
```python
import scipy.stats as stats

# Create data arrays for each group
group_a = [1, 2, 3, 4, 5]
group_b = [2, 3, 4, 5, 6]
group_c = [3, 4, 5, 6, 7]

# Perform one-way ANOVA test
f_statistic, p_value = stats.f_oneway(group_a, group_b, group_c)

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

```
F-statistic: 2.0
```

p-value: 0.177978515625

[ ]:
```
    If the p-value is less than your chosen significance level (e.g., 0.05),␣
  ↪then you can reject the null hypothesis and conclude that there is
a statistically significant difference between at least two of the population␣
  ↪means. Otherwise, you fail to reject the null hypothesis and
conclude that there is insufficient evidence to suggest that there is a␣
  ↪difference between any of the population means.
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
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 -

[21]:
```
import numpy as np
import pandas as pd
```

[22]:
```
df = pd.DataFrame({
    'Software': ['A', 'B', 'C'] * 10,
    'Experience': ['Novice'] * 15 + ['Experienced'] * 15,
    'Time': [10, 12, 13, 11, 14, 15, 12, 13, 14, 15] * 3
})
```

[23]:
```
from statsmodels.formula.api import ols

model = ols('Time ~ C(Software) + C(Experience) + C(Software):C(Experience)',␣
  ↪data=df).fit()
```

[24]:
```
from statsmodels.stats.anova import anova_lm

anova_results = anova_lm(model)
```

[ ]:
```
    The F-statistics and p-values can be obtained from anova_results. The main␣
  ↪effects and interaction effects between software programs and
```

```
employee experience level can be interpreted based on these values.
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
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 -

[25]:

```
import scipy.stats as stats

control_group = [80, 85, 90, 95, 100]
experimental_group = [90, 95, 100, 105, 110]

t_statistic, p_value = stats.ttest_ind(control_group, experimental_group)

print("t-statistic: ", t_statistic)
print("p-value: ", p_value)
```

```
t-statistic:  -2.0
p-value:  0.08051623795726257
```

[ ]:

```
      If the p-value is less than the significance level (usually 0.05), then␣
↪we can reject the null hypothesis and conclude that there is a
significant difference between the two groups.

If the results are significant, you can follow up with a post-hoc test to␣
↪determine which group(s) differ significantly from each other.
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
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 -

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

      # create a dataframe with the data
      data = {'store': ['A', 'B', 'C'] * 30,
              'sales': [10, 12, 8, 11, 13, 9, 9, 11, 7] * 10}
      df = pd.DataFrame(data)

      # fit the model
      model = ols('sales ~ C(store)', data=df).fit()

      # perform the ANOVA
      anova_table = sm.stats.anova_lm(model, typ=2)

      print(anova_table)
```

```
           sum_sq    df      F        PR(>F)
C(store)    240.0   2.0  174.0  3.933732e-31
Residual     60.0  87.0    NaN           NaN
```

```
[34]: from statsmodels.stats.multicomp import pairwise_tukeyhsd

      # perform Tukey's HSD test
      tukey_results = pairwise_tukeyhsd(df['sales'], df['store'])

      print(tukey_results)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj  lower   upper  reject
---------------------------------------------------
     A      B      2.0   0.0  1.4887  2.5113   True
     A      C     -2.0   0.0 -2.5113 -1.4887   True
     B      C     -4.0   0.0 -4.5113 -3.4887   True
---------------------------------------------------
```

```
[ ]:
```