algorithms

March 7, 2023

[ ]:
```
#Load statistical analysis
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import MultiComparison, pairwise_tukeyhsd
```

[ ]:
```
#Load dataset
anova_data = pd.read_csv('raw_data.csv')
anova_data
```

[ ]:

	Samples	Cu	Cd	Zn	Pb
0	Liver	0.320	0.012	0.2311	0.240
1	Liver	0.284	0.009	0.1040	0.230
2	Intestine	0.010	0.013	0.0590	0.420
3	Intestine	0.008	0.016	0.0588	0.490
4	Gills	0.015	0.015	0.0937	0.740
5	Gills	0.015	0.018	0.0998	0.730
6	Flesh	0.013	0.017	0.0500	0.700
7	Flesh	0.005	0.019	0.0455	0.690
8	Water	0.017	0.027	0.0142	0.010
9	Water	0.011	0.025	0.0120	0.011

[ ]:
```
# Fit the one-way ANOVA model using the ols method for Cu
model = ols("Cu ~ C(Samples)", data=anova_data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
print(aov_table)
```

```
               df    sum_sq   mean_sq       F    PR(>F)
C(Samples)   4.0  0.134854  0.033713  240.81  0.000007
Residual     5.0  0.000700  0.000140     NaN       NaN
```

According to the result, the ANOVA has found a significant effect of "Copper" on the "Samples", as indicated by the very small p-value (0.000007). This suggests that there is a significant difference in the dependent variable among the different types of the "Samples" variable.

The F-statistic (240.81) also indicates a large effect size, suggesting that the variability in the dependent variable among the different types of the "Samples" variable is much greater than the variability within each type of the "Samples" variable.

[ ]:
```
# Perform a multiple comparison test using the Duncan method for Cu
mc = MultiComparison(anova_data['Cu'], anova_data['Samples'])
mc_results = mc.tukeyhsd()
print(mc_results)
```

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower   upper  reject
-----------------------------------------------------
    Flesh     Gills    0.006 0.9829 -0.0415  0.0535  False
    Flesh Intestine      0.0    1.0 -0.0475  0.0475  False
    Flesh     Liver    0.293    0.0  0.2455  0.3405   True
    Flesh     Water    0.005 0.9913 -0.0425  0.0525  False
    Gills Intestine   -0.006 0.9829 -0.0535  0.0415  False
    Gills     Liver    0.287    0.0  0.2395  0.3345   True
    Gills     Water   -0.001    1.0 -0.0485  0.0465  False
Intestine     Liver    0.293    0.0  0.2455  0.3405   True
Intestine     Water    0.005 0.9913 -0.0425  0.0525  False
    Liver     Water   -0.288    0.0 -0.3355 -0.2405   True
-----------------------------------------------------
```

The Tukey HSD test is used to compare the means of all possible pairs of groups. In this case, the groups being compared are Flesh, Gills, Intestine, Liver, and Water.

According to the analysis, the means of Flesh and Water, Flesh and Gills, and Gills and Water do not differ significantly from each other, as indicated by the high p-values (greater than 0.05) and the "False" values in the "reject" column.

However, the means of Flesh and Liver, Gills and Liver, Intestine and Liver, and Liver and Water do differ significantly from each other, as indicated by the low p-values (less than 0.05) and the "True" values in the "reject" column. Meaning the null hypothesis in each of the pair is rejected.

[ ]:
```
# Fit the one-way ANOVA model using the ols method for Cd
model = ols("Cd ~ C(Samples)", data=anova_data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
print(aov_table)
```

```
              df    sum_sq   mean_sq          F    PR(>F)
C(Samples)   4.0  0.000261  0.000065  18.671429  0.003297
Residual     5.0  0.000017  0.000003        NaN       NaN
```

Based on the output, the p-value is 0.003297, which is less than the typical significance level of 0.05. This suggests that there is a significant difference in the effect of Cadmium between at least two of the samples.

Therefore, we can reject the null hypothesis, and conclude that there is a significant difference in the effect of Cadmium among at least some of the samples.

[ ]:
```
# Perform a multiple comparison test using the Duncan method for Cd
mc = MultiComparison(anova_data['Cd'], anova_data['Samples'])
mc_results = mc.tukeyhsd()
```

```
print(mc_results)
```

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower  upper  reject
-----------------------------------------------------
    Flesh     Gills  -0.0015 0.9195  -0.009  0.006  False
    Flesh Intestine  -0.0035 0.4312  -0.011  0.004  False
    Flesh     Liver  -0.0075 0.0501  -0.015    0.0  False
    Flesh     Water    0.008 0.0392  0.0005 0.0155   True
    Gills Intestine   -0.002 0.8154 -0.0095 0.0055  False
    Gills     Liver   -0.006 0.1096 -0.0135 0.0015  False
    Gills     Water   0.0095 0.0197   0.002  0.017   True
Intestine     Liver   -0.004 0.3308 -0.0115 0.0035  False
Intestine     Water   0.0115 0.0087   0.004  0.019   True
    Liver     Water   0.0155 0.0023   0.008  0.023   True
-----------------------------------------------------
```

For the comparison of Cadmium and the samples, the ANOVA results show that there is a significant difference among the groups (p=0.0033). However, the Tukey HSD post-hoc test did not reveal any significant differences between the groups after correction for multiple comparisons, except for the comparison between Flesh and Water, where the mean difference was 0.008 and the p-value was 0.0392. This indicates that the concentration of Cadmium in the flesh and water samples may be significantly different.

[ ]:
```
# Fit the one-way ANOVA model using the ols method for Zn
model = ols("Zn ~ C(Samples)", data=anova_data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
print(aov_table)
```

```
                df    sum_sq   mean_sq         F    PR(>F)
C(Samples)   4.0  0.027711  0.006928  4.271993  0.071533
Residual     5.0  0.008108  0.001622       NaN       NaN
```

For Zinc and samples, the p-value for the ANOVA test is 0.0715 which is greater than 0.05. Therefore, there is sufficient evidence to reject the null hypothesis that there is no significant difference in mean Zinc levels among the different samples.

[ ]:
```
# Fit the one-way ANOVA model using the ols method for Pb
model = ols("Pb ~ C(Samples)", data=anova_data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
print(aov_table)
```

```
                df    sum_sq   mean_sq           F    PR(>F)
C(Samples)   4.0  0.755608  0.188902  363.203422  0.000002
Residual     5.0  0.002601  0.000520         NaN       NaN
```

Based on the ANOVA results, it can be concluded that there is a significant difference between the lead in the different samples ($F = 363.203$, $p < 0.0001$). To determine which groups are significantly different from each other, a post-hoc Tukey HSD test can be perform.

[ ]:
```
# Perform a multiple comparison test using the Duncan method
mc = MultiComparison(anova_data['Pb'], anova_data['Samples'])
mc_results = mc.tukeyhsd()
print(mc_results)
```

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower   upper  reject
-----------------------------------------------------
    Flesh     Gills      0.04 0.4817 -0.0515  0.1315  False
    Flesh Intestine     -0.24 0.0007 -0.3315 -0.1485   True
    Flesh     Liver     -0.46    0.0 -0.5515 -0.3685   True
    Flesh     Water   -0.6845    0.0  -0.776  -0.593   True
    Gills Intestine     -0.28 0.0003 -0.3715 -0.1885   True
    Gills     Liver      -0.5    0.0 -0.5915 -0.4085   True
    Gills     Water   -0.7245    0.0  -0.816  -0.633   True
Intestine     Liver     -0.22 0.0011 -0.3115 -0.1285   True
Intestine     Water   -0.4445    0.0  -0.536  -0.353   True
    Liver     Water   -0.2245  0.001  -0.316  -0.133   True
-----------------------------------------------------
```

For Lead, the ANOVA results show a p-value of less than 0.05, indicating we have evidence to reject the null hypothesis that the means of the groups are equal. The Tukey HSD test shows that all pairwise comparisons between the groups are statistically significant, except for Flesh vs. Gills, which is not statistically significant.