

algorithms-corrected

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 0.24
1    Liver 0.284 0.009 0.1040 0.23
2 Intestine 0.010 0.013 0.0590 0.42
3 Intestine 0.008 0.016 0.0588 0.49
4    Gills 0.015 0.015 0.0937 0.74
5    Gills 0.015 0.018 0.0998 0.73
6    Flesh 0.013 0.017 0.0500 0.70
7    Flesh 0.005 0.019 0.0455 0.69
```

```
[ ]: # 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)	3.0	0.127069	0.042356	248.42522	0.000053
Residual	4.0	0.000682	0.000171	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 (0.000053). This suggests that there is a significant difference in the dependent variable among the different types of the “Samples” variable.

The F-statistic (248.43) 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 0.9642 -0.0472 0.0592  False
    Flesh Intestine      0.0    1.0 -0.0532 0.0532  False
    Flesh      Liver      0.293 0.0001  0.2398 0.3462   True
    Gills Intestine     -0.006 0.9642 -0.0592 0.0472  False
    Gills      Liver      0.287 0.0001  0.2338 0.3402   True
Intestine      Liver      0.293 0.0001  0.2398 0.3462   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, and Liver.

According to the analysis, the means of Flesh and Gills, Gills and Intestine, and Flesh and Intestine 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, and Intestine and Liver 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)	3.0	0.000063	0.000021	5.451613	0.067475
Residual	4.0	0.000016	0.000004	NaN	NaN

Based on the output, the p-value is 0.067475, which is greater than the typical significance level of 0.05. This suggests that there is no significant difference in the effect of Cadmium between the samples. Therefore, the null hypothesis is not rejected.

```
[ ]: # 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)	3.0	0.017564	0.005855	2.889024	0.165946
Residual	4.0	0.008106	0.002026	NaN	NaN

For Zinc and samples, the p-value for the ANOVA test is 0.165946 which is greater than 0.05. Therefore, there is sufficient evidence for not rejecting 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)	3.0	0.3238	0.107933	166.051282	0.000119
Residual	4.0	0.0026	0.000650	NaN	NaN

Based on the ANOVA results, it can be concluded that there is a significant difference between the lead in the different samples ($F = 166.051$, $p < 0.000119$). To determine which groups are significantly different from each other, a post-hoc Tukey HSD test can be performed.

```
[ ]: # 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	0.4825	-0.0638	0.1438	False
Flesh	Intestine	-0.24	0.0025	-0.3438	-0.1362	True
Flesh	Liver	-0.46	0.0002	-0.5638	-0.3562	True
Gills	Intestine	-0.28	0.0014	-0.3838	-0.1762	True
Gills	Liver	-0.5	0.0001	-0.6038	-0.3962	True
Intestine	Liver	-0.22	0.0034	-0.3238	-0.1162	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.