Bihan Q

Covariate, Pairwise, Anova, Tied Obs

Q1(a)

Consider the ChickWeight data in R. The body weights of the chicks were measured at birth (i.e., time=0) and every second day thereafter until day 20. They were also measured on day 21. There were four groups of chicks on different protein diets.

1.Determine whether there is a significant difference in the mean weights of the four groups on Day 18:

a)Without adjusting for Birth Weight

In [46]:
```
# packages
import pandas as pd
```

In [47]:
```
# get data
import statsmodels.api as sm
ChickWeight = sm.datasets.get_rdataset('ChickWeight','datasets')
chick = pd.DataFrame(ChickWeight.data)
print(chick.shape)
chick.head(40)
```

```
(578, 4)
```

Out[47]:

	weight	Time	Chick	Diet
0	42	0	1	1
1	51	2	1	1
2	59	4	1	1
3	64	6	1	1
4	76	8	1	1
5	93	10	1	1
6	106	12	1	1
7	125	14	1	1
8	149	16	1	1
9	171	18	1	1
10	199	20	1	1
11	205	21	1	1
12	40	0	2	1
13	49	2	2	1
14	58	4	2	1
15	72	6	2	1
16	84	8	2	1
17	103	10	2	1
18	122	12	2	1
19	138	14	2	1
20	162	16	2	1
21	187	18	2	1
22	209	20	2	1
23	215	21	2	1
24	43	0	3	1
25	39	2	3	1
26	55	4	3	1
27	67	6	3	1
28	84	8	3	1
29	99	10	3	1
30	115	12	3	1
31	138	14	3	1
32	163	16	3	1
33	187	18	3	1

	weight	Time	Chick	Diet
34	198	20	3	1
35	202	21	3	1
36	42	0	4	1
37	49	2	4	1
38	56	4	4	1
39	67	6	4	1

In [48]:
```
chick['Diet'].value_counts()
```

Out[48]:
```
1    220
2    120
3    120
4    118
Name: Diet, dtype: int64
```

In [49]:
```
chick.describe()
```

Out[49]:

	weight	Time	Chick	Diet
count	578.000000	578.000000	578.000000	578.000000
mean	121.818339	10.717993	25.750865	2.235294
std	71.071960	6.758400	14.568795	1.162678
min	35.000000	0.000000	1.000000	1.000000
25%	63.000000	4.000000	13.000000	1.000000
50%	103.000000	10.000000	26.000000	2.000000
75%	163.750000	16.000000	38.000000	3.000000
max	373.000000	21.000000	50.000000	4.000000

In [50]:
```
chick_18 = chick[chick['Time'] == 18]
print(chick_18.shape)
chick_18.head(3)
```

```
(47, 4)
```

Out[50]:

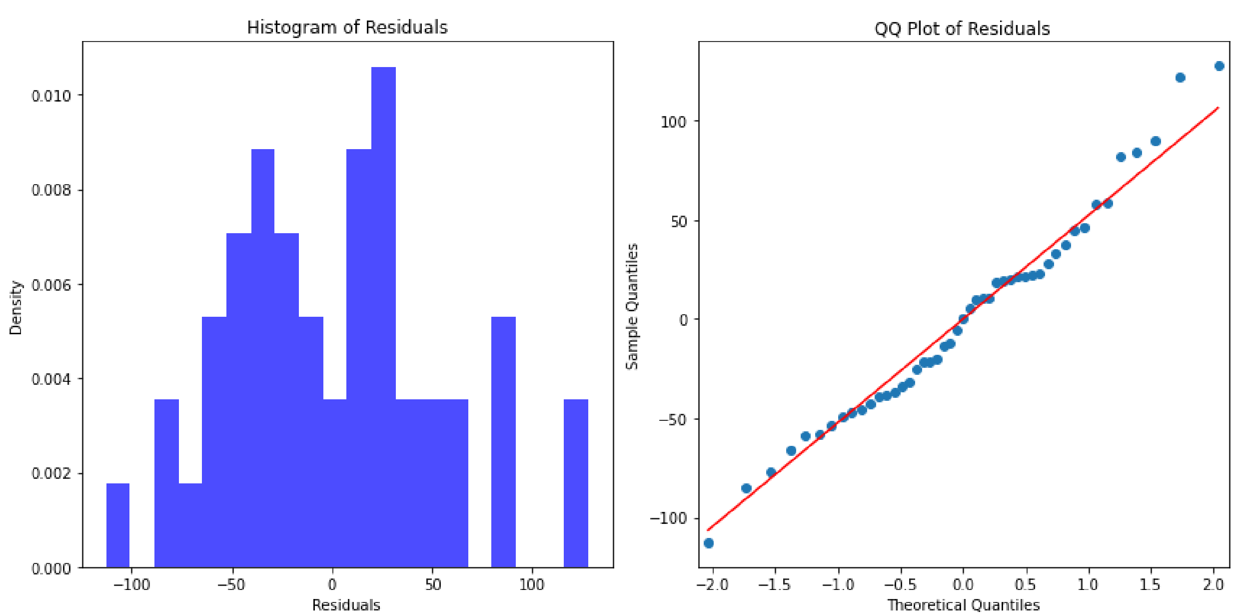
	weight	Time	Chick	Diet
9	171	18	1	1
21	187	18	2	1
33	187	18	3	1

In [51]:
```
# with fixed time, weight is Y, diets are groups and independent
from statsmodels.formula.api import ols
model = ols('weight ~ Diet', data=chick_18).fit()
anova_table = sm.stats.anova_lm(model)

# Print the ANOVA table
```

```
print("One-Way ANOVA Table:")
print(anova_table)
```

```
One-Way ANOVA Table:
            df         sum_sq       mean_sq         F    PR(>F)
Diet       1.0   23263.992032  23263.992032  8.161704  0.006455
Residual  45.0  128267.284564   2850.384101       NaN       NaN
```

In [52]:
```
#check non-normality
import matplotlib.pyplot as plt

# hist
model = ols('weight ~ Diet', data=chick_18).fit()
residuals = model.resid
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.hist(residuals, bins=20, density=True, color='b', alpha=0.7)
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Density')

# qq
plt.subplot(1, 2, 2)
sm.qqplot(residuals, line='s', ax=plt.gca())
plt.title('QQ Plot of Residuals')

plt.tight_layout()
plt.show()

# Bartlett's test
import pingouin as pg
bartlett_test_result = pg.homoscedasticity(chick_18, dv='weight', group='Diet')
print("Bartlett's Test for Homogeneity of Variances:")
print(bartlett_test_result)
```




```
Bartlett's Test for Homogeneity of Variances:
               W      pval  equal_var
levene  1.058972  0.376444       True
```

In [53]:
```
# shapiro
from scipy.stats import shapiro
```

```
model = ols('weight ~ Diet', data=chick_18).fit()
residuals_model_with_covariate = model.resid
shapiro_test_statistic, shapiro_test_pvalue = shapiro(residuals_model_with_covariate)

print("Shapiro-Wilk Test Statistic:", shapiro_test_statistic)
print("p-value:", shapiro_test_pvalue)
```

```
Shapiro-Wilk Test Statistic: 0.980424702167511
p-value: 0.6098649501800537
```

In [54]:

```
# Bartlett's test
bartlett_test_result = pg.homoscedasticity(chick_18, dv='weight', group='Diet')
print("Bartlett's Test for Homogeneity of Variances:")
print(bartlett_test_result)
```

```
Bartlett's Test for Homogeneity of Variances:
               W      pval  equal_var
levene  1.058972  0.376444       True
```

Answer:

1. Under satisfied normality and equal variance assumption, 26.098205 as F-stats and 4.418863e-07 as p-value indicate **a significant difference in the mean weights of the four groups on Day 18.**

Q1(b)

b)Adjusting for Birth Weight. Give the LS Means (i.e., adjusted for Birth Weight).

In [55]:

```
# get inital weight
birth_weight = chick[chick['Time'] == 0][['Chick', 'weight']]
birth_weight.rename(columns={'weight': 'weight_inital'}, inplace=True)
chick_ajusted = pd.merge(chick, birth_weight, on='Chick', how='left')
chick_ajusted_18 = chick_ajusted[chick_ajusted['Time'] == 18]
print(chick_ajusted_18.shape[0]==chick_18.shape[0])
chick_ajusted_18.head(3)
chick_ajusted_18['weight_inital'].value_counts()
```

```
True
```

Out[55]:

```
41    18
42    14
39     6
40     5
43     4
Name: weight_inital, dtype: int64
```

In [56]:

```
chick_ajusted_18['Diet'].value_counts()
```

Out[56]:

```
1    17
2    10
3    10
4    10
Name: Diet, dtype: int64
```

In [57]:

```python
# two-anova (python)
model2 = ols('weight ~ Diet+weight_inital', data=chick_ajusted_18).fit()
anova_table2 = sm.stats.anova_lm(model2)
print(anova_table2)
```

```
                  df         sum_sq       mean_sq         F    PR(>F)
Diet             1.0   23263.992032  23263.992032  8.573318  0.005384
weight_inital    1.0    8871.784187   8871.784187  3.269457  0.077419
Residual        44.0  119395.500377   2713.534099       NaN       NaN
```

In [58]:

```python
# anova (r)
from IPython.display import Image
image_file_path = f"C:/Users/11139/Desktop/STAT5391/anova2.png"
Image(filename=image_file_path)
```

Out[58]:

```r
# anova
library(car)
m <- aov(data=chick_adjusted_18, weight ~ Diet+weight_initial)
summary(m)
```
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)   
Diet            3  36690   12230   4.729 0.00623 **
weight_initial  1   6229    6229   2.409 0.12818   
Residuals      42 108612    2586                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [59]:

```python
# pairwise test
import pingouin as pg
from statsmodels.stats.anova import AnovaRM
# ls_means = pg.pairwise_ttests(data=chick_ajusted_18, dv='weight', between='Diet', co
from IPython.display import Image
image_file_path = f"C:/Users/11139/Desktop/STAT5391/pairwise_test.png"
Image(filename=image_file_path)
```
```

Out[59]:

```
library(stats)
library(stats)
pairwise_tests <- pairwise.t.test(
  x = chick_adjusted_18$weight,
  g = chick_adjusted_18$Diet,
  p.adjust.method = "fdr",
  covariates = chick_adjusted_18$weight_initial
)

# Print the results
print(pairwise_tests)

```
```
```

```
        Pairwise comparisons using t tests with pooled SD

data:  chick_adjusted_18$weight and chick_adjusted_18$Diet

  1      2      3
2 0.2379 -      -
3 0.0049 0.1119 -
4 0.1119 0.5142 0.2379

P value adjustment method: fdr
```

In [60]:

```python
# LSM
image_file_path = f"C:/Users/11139/Desktop/STAT5391/LSM.png"
Image(filename=image_file_path)
```

Out[60]:

```
model <- lm(weight ~Diet+weight_initial, chick_adjusted_18)
anova_result <- anova(model)
lsmeans_result <- lsmeans(model, "Diet", cov.reduce=TRUE)
lsmeans_result

320:1   Chunk 8
```

Console   Background Jobs

R 4.2.1 · ~/

```
 4         202 16.1 42       169       234

Confidence level used: 0.95
> lsmeans_result <- lsmeans(model, "Diet", cov.reduce=TRUE)
> lsmeans_result
 Diet lsmean   SE df lower.CL upper.CL
 1       164 12.8 42      138      190
 2       183 16.3 42      150      216
 3       230 16.2 42      197      262
 4       202 16.1 42      169      234

Confidence level used: 0.95
>
```

Answer:

Two-way Anova:

1. Compared to before adjusted, after we add inital_birth_weight as second variable, diet with p-value 0.00623 indicates that **there is a statistically difference in the mean weights among diet groups on Day18.**
2. From p-value = 0.12818, **different inital weight groups does not have a significant difference on mean weight on Day18.**
3. Python and R gives little different numbers, but similar in result significance.

Pairwise:

1. Only p-value for the comparison between **Group 1 and Group 3** is 0.0049. **There is statistically significant difference in the mean weights between Group 1 and Group 3 after adjusting for the covariate 'weight_inital' at Day18.** For other groups, there are no signficiant difference.

LSM:

1. Compared to original weights for four groups 1-4 (158.94,187.70,233.10,202.90), we get new adjusted weight for group 1-4 is 164,183,230,and 202. **Four groups of diet are different.**

1(c)

c)Check the validity of your assumptions, including parallelism. Suggest measures that you would take if the assumptions are not satisfied.

In [61]:

```
# Shapiro-Wilk
# from scipy.stats import shapiro
# shapiro_test_statistic, shapiro_test_pvalue = shapiro(residuals_model_with_covariate
image_file_path = f"C:/Users/11139/Desktop/STAT5391/shapiro.png"
Image(filename=image_file_path)
```

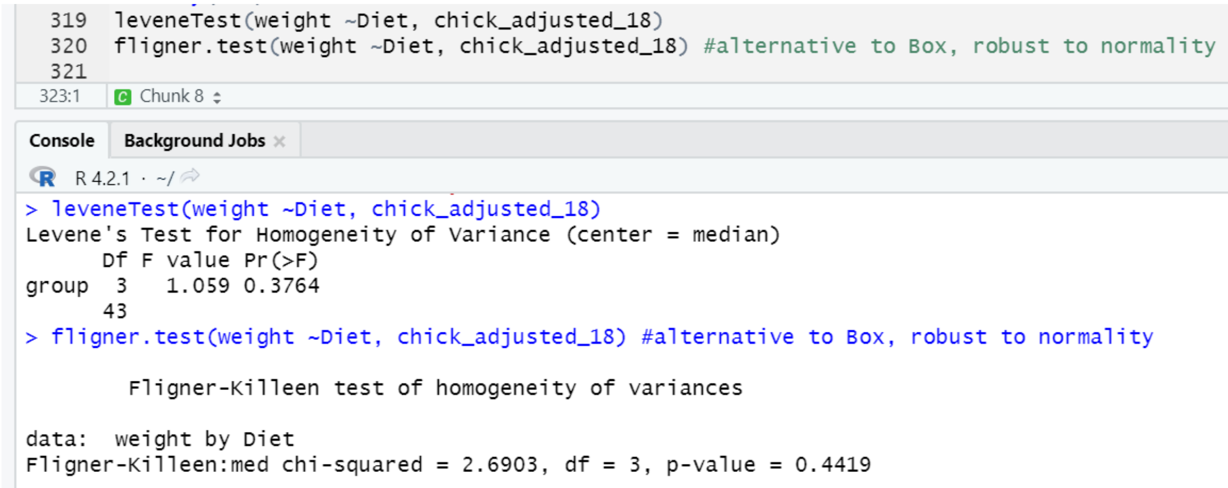
Out[61]:

```
# anova
library(car)
m <- aov(data=chick_adjusted_18, weight ~ Diet+weight_initial)
summary(m)

shapiro.test(resid(aov(data=chick_adjusted_18, weight ~ Diet+weight_initial)))
```
```

	Shapiro-Wilk normality test

data:  resid(aov(data = chick_adjusted_18, weight ~ Diet + weight_initial))
W = 0.98613, p-value = 0.8441
```

In [62]:

```
# Levene's and Box's test
image_file_path = f"C:/Users/11139/Desktop/STAT5391/Box.png"
Image(filename=image_file_path)
```

Out[62]:



```
leveneTest(weight ~Diet, chick_adjusted_18)
fligner.test(weight ~Diet, chick_adjusted_18) #alternative to Box, robust to normality

323:1   Chunk 8
```

Console   Background Jobs

R 4.2.1 · ~/

```
> leveneTest(weight ~Diet, chick_adjusted_18)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3   1.059 0.3764
      43
> fligner.test(weight ~Diet, chick_adjusted_18) #alternative to Box, robust to normality

        Fligner-Killeen test of homogeneity of variances

data:  weight by Diet
Fligner-Killeen:med chi-squared = 2.6903, df = 3, p-value = 0.4419
```

In [63]:

```python
# Two-way ANOVA with interaction
two_way_anova_result = sm.stats.anova_lm(ols('weight ~ Diet * weight_inital' , data=ch
print("Two-Way ANOVA with Interaction:")
print(two_way_anova_result)
```

```
Two-Way ANOVA with Interaction:
                       df         sum_sq       mean_sq         F    PR(>F)
Diet                  1.0   23263.992032  23263.992032  8.393622  0.005901
weight_inital         1.0    8871.784187   8871.784187  3.200930  0.080642
Diet:weight_inital    1.0     215.522671    215.522671  0.077760  0.781693
Residual             43.0  119179.977706   2771.627389       NaN       NaN
```

Answer:

Shapiro-Wilk:

- Shapiro-Wilk p-value (0.8441) shows normality assumption is **satisfied**.

Bartlett's test:

- Assumption of constant variance is **satisfied**.

Interaction(parallelism):

- The p-value for the interaction term (0.781693), suggesting **no interaction** effect between diet and weight_inital.

2(a)

2.For 1a), perform pairwise comparisons among the 4 groups using each of the following, and comment on the results a)Bonferroni method

In [64]:

```python
from IPython.display import Image
image_file_path = f"C:/Users/11139/Desktop/STAT5391/Bonferroni.png"
Image(filename=image_file_path)
```

Out[64]:

```
# bonferroni
model <- aov(weight ~ Diet, data = chick_18)
pairwise_results <- pairwise.t.test(chick_18$weight, chick_18$Diet, p.adjust.method =
"bonferroni")
print(pairwise_results)
```

342:1 Chunk 9

Console Background Jobs

R 4.2.1 · ~/

```
> print(pairwise_results)

        Pairwise comparisons using t tests with pooled SD

data:  chick_18$weight and chick_18$Diet

  1      2      3
2 1.0000 -      -
3 0.0049 0.3358 -
4 0.2313 1.0000 1.0000

P value adjustment method: bonferroni
>
```

Answer:

- There is significant difference in the mean weights of the diet1 and diet3 on Day 18.
- All other groups are not significantly different from others on Day18..

2(b)

2.For 1a), perform pairwise comparisons among the 4 groups using each of the following, and comment on the results b)Tukey method

In [65]:

```python
from statsmodels.stats.multicomp import pairwise_tukeyhsd

post_hoc = pairwise_tukeyhsd(chick_18['weight'], chick_18['Diet'], alpha=0.05)
#table
print("ANOVA Table:")
print(anova_table)
print("\nPairwise Comparisons (Bonferroni Correction):")
print(post_hoc)
```

```
ANOVA Table:
              df         sum_sq       mean_sq         F    PR(>F)
Diet         1.0   23263.992032  23263.992032  8.161704  0.006455
Residual    45.0  128267.284564   2850.384101       NaN       NaN

Pairwise Comparisons (Bonferroni Correction):
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
======================================================
group1 group2 meandiff p-adj   lower    upper   reject
------------------------------------------------------
     1      2  28.7588 0.5085 -26.2809  83.7986  False
     1      3  74.1588 0.0044  19.1191 129.1986   True
     1      4  43.9588 0.1586 -11.0809  98.9986  False
     2      3     45.4 0.2173 -16.3638 107.1638  False
     2      4     15.2 0.9123 -46.5638  76.9638  False
     3      4    -30.2 0.5638 -91.9638  31.5638  False
------------------------------------------------------
```

Answer:

- Same as Bonferroni correction, by Tukey, there is significant difference in the mean weights of the diet1 and diet3 on Day 18.

Q3

Repeat 1a) using the Kurskal-Wallis test

In [66]:

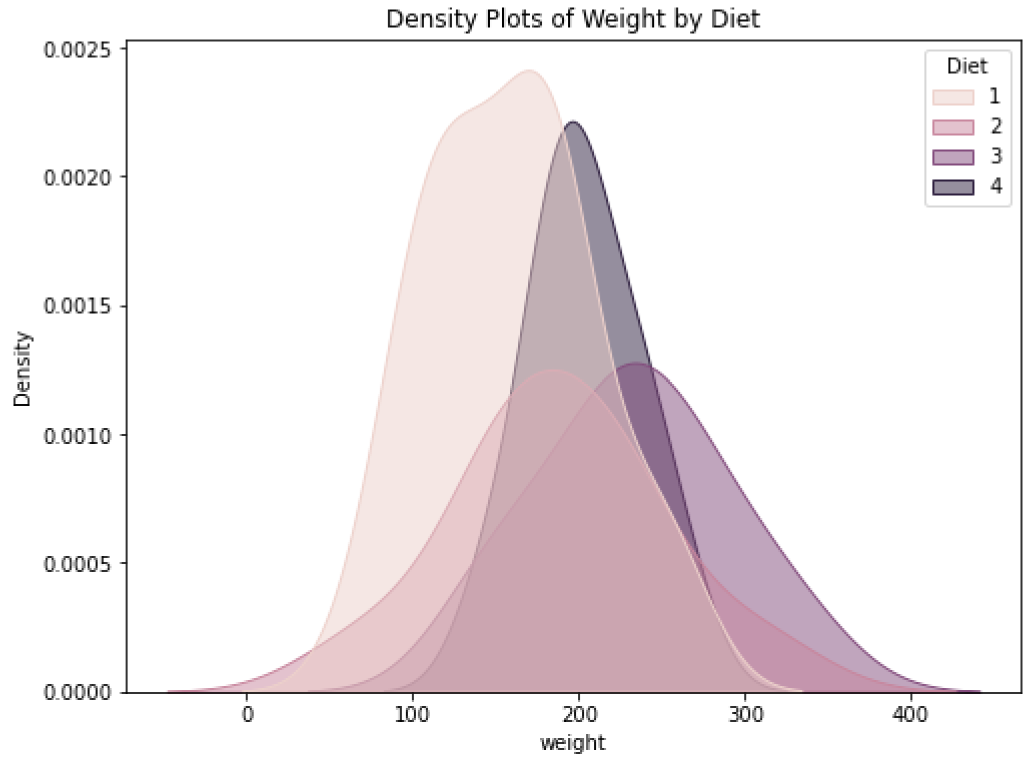
```
# exam of tie data
chick_18['weight'].value_counts()
```

Out[66]:

```
187    2
163    2
234    2
146    2
184    2
185    2
171    1
227    1
204    1
263    1
294    1
332    1
232    1
157    1
230    1
214    1
262    1
198    1
174    1
210    1
261    1
151    1
231    1
207    1
205    1
154    1
199    1
160    1
250    1
134    1
100    1
112    1
81     1
248    1
123    1
120    1
107    1
307    1
148    1
72     1
203    1
Name: weight, dtype: int64
```

In [67]:

```
# density plot for each group
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8, 6))
sns.kdeplot(data=chick_18, x='weight', hue='Diet', fill=True, alpha=0.5)
plt.title('Density Plots of Weight by Diet')
plt.show()
```



In [68]:

```
import pandas as pd
from scipy.stats import kruskal
result = kruskal(*[group["weight"] for name, group in chick_18.groupby("Diet")])
print(result)
```

```
KruskalResult(statistic=10.623444357500071, pvalue=0.01394646461394125)
```

Answer:

Discussion for assumption of Kruskal:

- Total sample size larger than 5.
- We test with numerical data weight.
- Deal with tied values, correction automatically applied.
- The distributions are similar across groups from density plot.

Result:

- The result of Kruskal test suggests that there is a significant difference in the mean weights of the four groups on Day 18.