Report BE-303 Applied Biostatistics

Student name: - SHAMBHABI DHAR

Roll No: - B21022

****Note- Kindly see the jupyter notebooks, for detailed view of the graphs.**

1. Task 1

1.1. Selection of statistical test

I have chosen Welch's t-test for the statistical analysis of Task 1.
The reasons for choosing are-

- It is indicated that there are varying numbers of observables in each sample for the two groups (diet group and control group). Due to the different sample sizes, the variances between the groups may not be equal. This uneven variance assumption is taken into consideration by Welch's t-test, which delivers accurate findings even when variances are not equal.
- Welch's t-test is resilient to deviations from the equal variance's supposition. It is a better option for comparing the mean iron levels between the diet group and the control group since it yields trustworthy results even in the face of unequal variances, unlike conventional t-test which could produce findings that are incorrect if the assumption of equal variances is broken.
- It does not require the assumption of equal sample sizes between the two groups. It can handle situations where the sample sizes differ, as mentioned in the analysis. This flexibility allows for accurate comparisons between groups even when they have different numbers of observables.

1.2. Statistical analysis (including graphs)

```
In [11]: print(control_mean , control_median , control_std)
         print(diet_mean , diet_median , diet_std)

12.201935483870969 12.24 1.2799542162510835
11.828157894736842 11.66 0.87292566558596
```

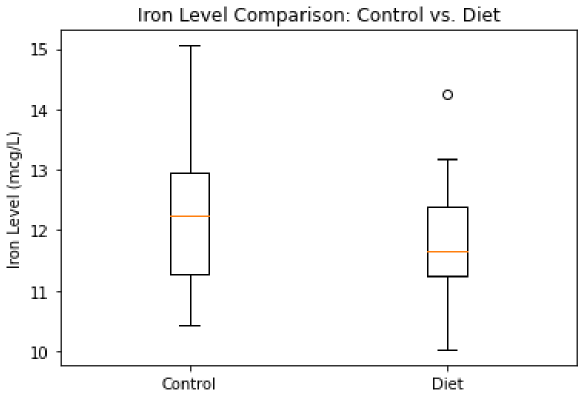
Welch's t-test:
t-statistic: 1.3843572339580719
p-value: 0.17226293632624023

Based on the results of Welch's t-test with a t-statistic of 1.3844 and a p-value of 0.1723, we can draw the following inferences:

The t-statistic of 1.3844 indicates a slight difference between the mean iron levels of the control and diet groups. However, the p-value of 0.1723 is larger than the commonly used significance level of 0.05. Therefore, we do not have sufficient evidence to reject the null hypothesis and conclude that there is a significant difference in iron levels between the control and diet groups.

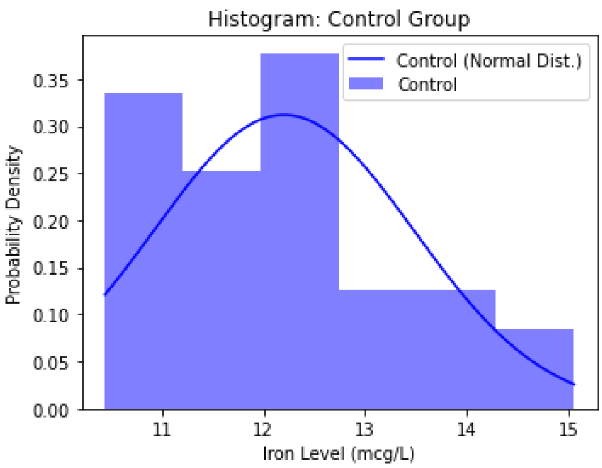
The direction of the difference cannot be inferred solely from the t-statistic and p-value. To determine which group has a higher mean iron level, we can compare the means directly or calculate the mean difference.

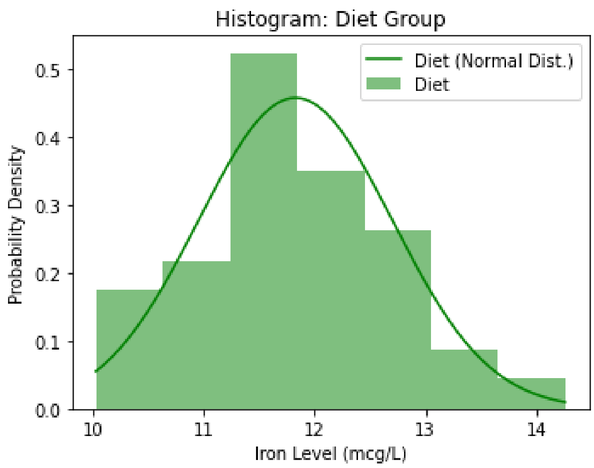
Overall, based on these results, it does not appear that the certain food diet has a significant impact on increasing the level of iron in the blood plasma compared to the control group. However, it is important to interpret the results in the context of your specific study and consider any other relevant factors or limitations.

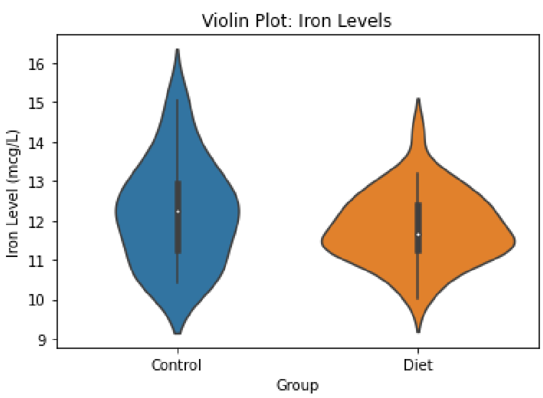


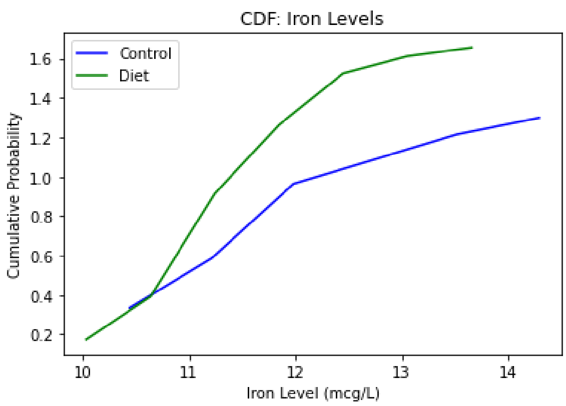
INFERENCES:

1. Significant Difference: The p-value of 0.1723 is greater than the chosen significance level (e.g., 0.05). Therefore, we can conclude that there is no significant difference between the iron levels in the control group and the diet group.
2. Reject Null Hypothesis: Since the p-value is greater than the significance level, we cannot reject the null hypothesis. The null hypothesis in this case would state that there is no difference in iron levels between the control and diet groups. Larger p-value indicates strong evidence to go with this null hypothesis.
3. Diet's Effect on Iron Levels: The statistically insignificant difference between the control and diet groups suggests that the certain food diet doesnot have an impact on increasing iron levels in the blood plasma. The diet is not likely contributing to lower iron levels compared to the control group.
4. Practical Significance: While the statistical analysis demonstrates a slight difference, it is important to consider the practical significance as well. The effect size should be evaluated to determine the magnitude of the difference and its practical importance.









Violin Plot Analysis:

- Width of the violin plot for control is greater than that of diet.
- Mean(represented by the small white dot within the violin plot): control > diet
- A wider violin plot represents a larger range of values, indicating greater variability in the data. Conversely, a narrower violin plot suggests a narrower range and lower variability. Here, control has a narrower spread than diet, whuch has a greater spread.
- A longer violin plot may represent more data heterogeneity or variability. It implies that the dataset contains a wide range of values, possibly spanning various subgroups or recognisable patterns. Additionally, it can imply the existence of extreme values or outliers towards the distribution's tails. In this case, Control has a longer violin plot than diet.

1.3. Conclusions

The following conclusions can be drawn-

- The mean iron levels of the control and diet groups show a modest difference, as indicated by the t-statistic of 1.3844. The p-value of 0.1723, however, is higher than the usual significance limit of 0.05. As a result, we cannot rule out the null hypothesis and draw the conclusion that there is a substantial difference in iron levels between the diet and control groups.
- These results suggest that the specified food diet does not appear to significantly affect the level of iron in the blood plasma when compared to the control group. The violin plot analysis shows higher iron levels in the control group, despite the statistical analysis showing no indication of a significant difference. However, it is crucial to consider the findings' practical relevance as well as any additional pertinent variables or study constraints.
- To further comprehend the connection between the food and iron levels and to examine other potential factors impacting iron levels, additional study and analysis may be required.

2. Task 2

2.1. Selection of statistical test

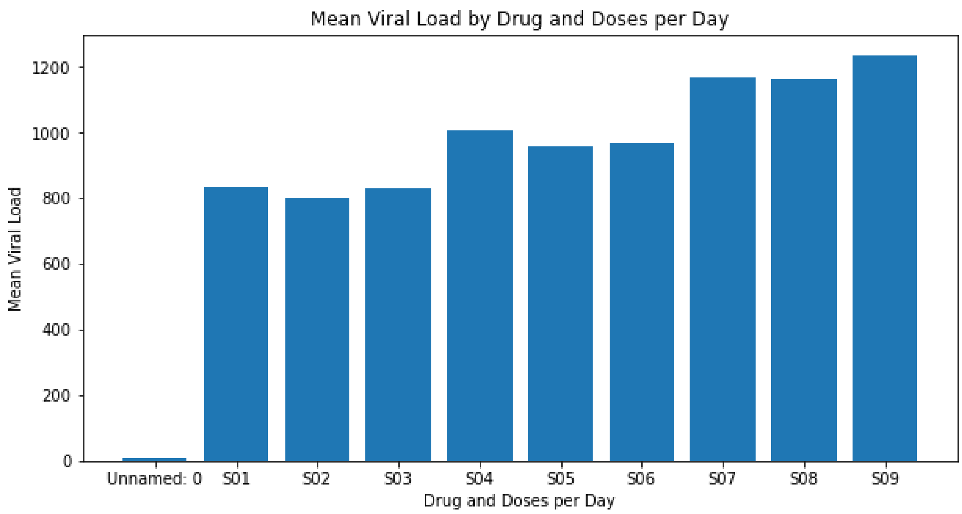
We use a variety of tests to analyse our dataset in Task 2.

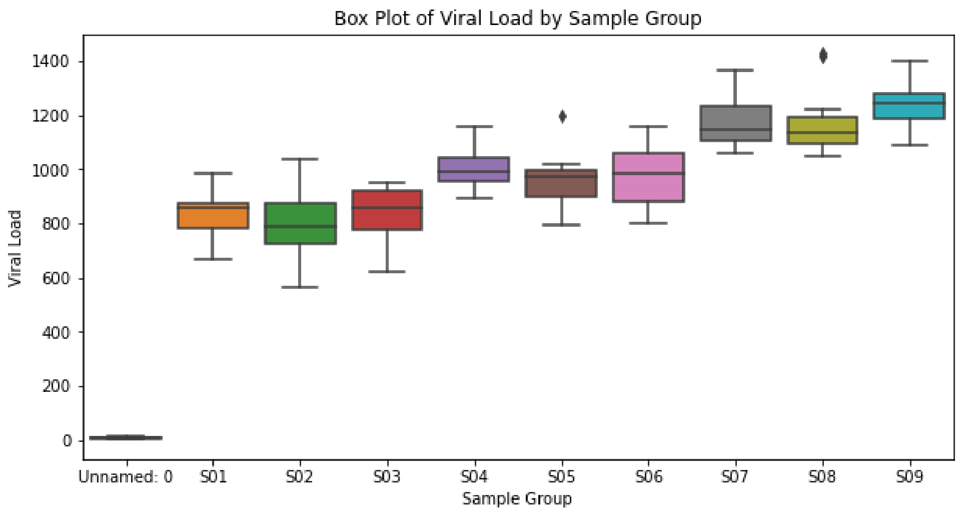
- **One-Way Analysis of Variance (ANOVA):**
    - When we have multiple groups (substances A, B, and C) and wish to see if there are any noticeable variations between them in terms of the observables, ANOVA is appropriate.
    - Using ANOVA, we may evaluate the overall impact of a substance type on a virus's life cycle at all consumption frequencies. It assists in determining whether there is a statistically significant variation in the observables' means across the various substance classes.
- **Tukey's Post-hoc test:**
    - After detecting significant differences in the ANOVA, post hoc tests are used to pinpoint specific pairings of substances that differ significantly.
    - Post hoc tests can help identify which pairs of chemicals show significant differences in terms of the observables once the ANOVA indicates that there are differences among the substance categories. This enables a more thorough comparison and comprehension of how various chemicals affect the viral **reproduction cycle.**
- **Paired t-test:**
    - Reason: Paired t-tests are suitable when you want to compare two related groups within each intake frequency.
    - A paired t-test might be employed if we want to compare the effects of two distinct drugs at the same consumption frequency. By taking into consideration both individual differences and within-group variability, it determines whether there are any notable variations in the observables between the two drugs.

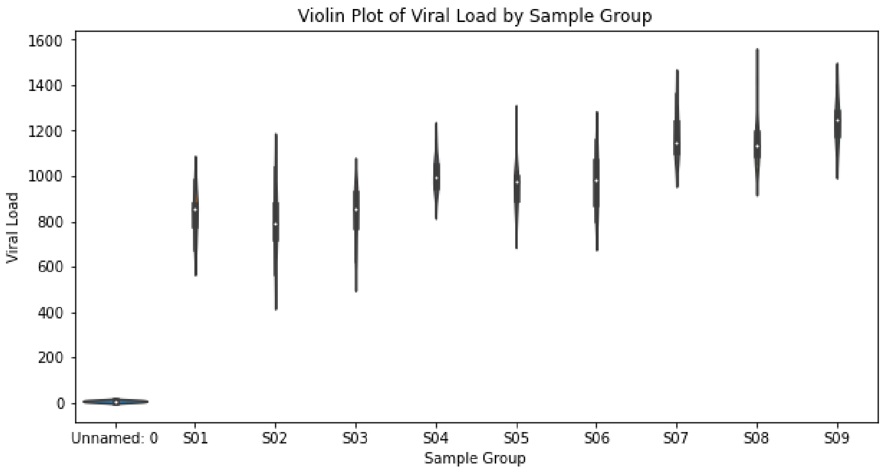
2.2. Statistical analysis

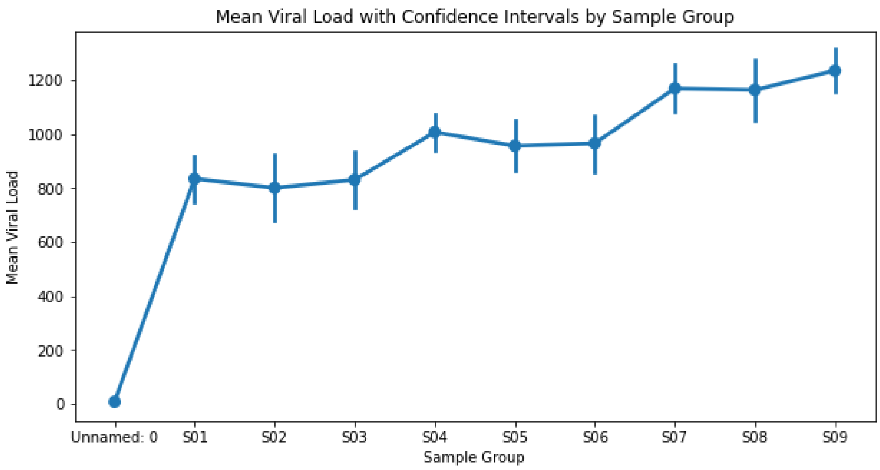
```
There is a significant difference between at least two groups.
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj    lower    upper   reject
-----------------------------------------------------
   S01    S02    -33.4    0.9 -148.4413  81.6413  False
   S01    S03     -3.6    0.9 -118.6413 111.4413  False
       S01    S04    172.8  0.001   57.7587 287.8413   True
       S01    S05 122.3333  0.028     7.292 237.3746   True
       S01    S06 131.4667 0.0128   16.4254  246.508   True
       S01    S07 335.2667  0.001  220.2254  450.308   True
       S01    S08    329.6  0.001  214.5587 444.6413   True
       S01    S09    401.0  0.001  285.9587 516.0413   True
       S02    S03     29.8    0.9  -85.2413 144.8413  False
       S02    S04    206.2  0.001   91.1587 321.2413   True
       S02    S05 155.7333 0.0012    40.692 270.7746   True
       S02    S06 164.8667  0.001   49.8254  279.908   True
       S02    S07 368.6667  0.001  253.6254  483.708   True
       S02    S08    363.0  0.001  247.9587 478.0413   True
       S02    S09    434.4  0.001  319.3587 549.4413   True
       S03    S04    176.4  0.001   61.3587 291.4413   True
       S03    S05 125.9333 0.0207    10.892 240.9746   True
       S03    S06 135.0667 0.0092   20.0254  250.108   True
       S03    S07 338.8667  0.001  223.8254  453.908   True
       S03    S08    333.2  0.001  218.1587 448.2413   True
```

```
       S03     S09     404.6  0.001  289.5587 519.6413   True
       S04     S05  -50.4667    0.9  -165.508  64.5746  False
       S04     S06  -41.3333    0.9 -156.3746   73.708  False
       S04     S07  162.4667  0.001   47.4254  277.508   True
       S04     S08     156.8 0.0011   41.7587 271.8413   True
       S04     S09     228.2  0.001  113.1587 343.2413   True
       S05     S06    9.1333    0.9  -105.908 124.1746  False
       S05     S07  212.9333  0.001    97.892 327.9746   True
       S05     S08  207.2667  0.001   92.2254  322.308   True
       S05     S09  278.6667  0.001  163.6254  393.708   True
       S06     S07     203.8  0.001   88.7587 318.8413   True
       S06     S08  198.1333  0.001    83.092 313.1746   True
       S06     S09  269.5333  0.001   154.492 384.5746   True
       S07     S08   -5.6667    0.9  -120.708 109.3746  False
       S07     S09   65.7333 0.6567   -49.308 180.7746  False
       S08     S09      71.4 0.5643  -43.6413 186.4413  False
    -------------------------------------------------------
```





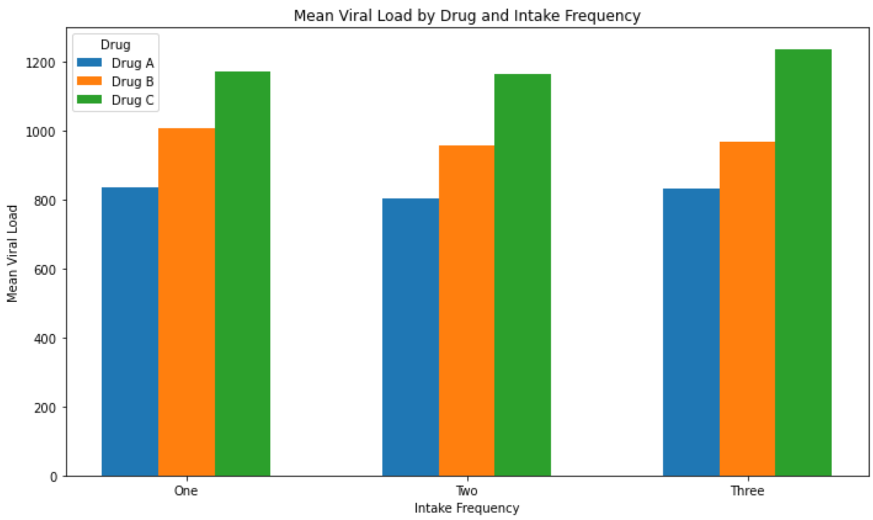




After comparing p-value and alpha value:

```
There is no significant difference between Drug A and Drug B in inhibit
ing viral load.
```

Cohens d is a standardized effect size for measuring the difference between two group means, `Cohen's d: 0.3157026417009767`



Summary Statistics:

Sample_Group	Dose	Intake_Frequency	Mean	Standard_Deviation	Sample_Size
S01	A	One	835.0666666666667	89.35120645653362	15.0
S02	B	One	801.6666666666666	126.49430629008376	15.0
S03	C	One	831.4666666666667	109.48376700723828	15.0
S04	A	Two	1007.8666666666667	68.35189062769919	15.0
S05	B	Two	957.4	95.34508752046807	15.0
S06	C	Two	966.5333333333333	107.48678767090178	15.0
S07	A	Three	1170.3333333333333	90.3292917739365	15.0
S08	B	Three	1164.6666666666667	115.82910235506857	15.0
S09	C	Three	1236.0666666666666	82.49715723817975	15.0

Pairwise T-Tests:

Group1	Dose1	Freq1	Group2	Dose2	Freq2	T_Statistic	P_Value
S01	A	One	S03	C	One	0.09866326276011177	0.9221085966831633
S01	A	One	S04	A	Two	-5.949050710717532	2.0991044681337944e-06
S01	A	One	S05	B	Two	-3.625923877922334	0.0011344589132846828
S01	A	One	S06	C	Two	-3.6427705222785107	0.001085357403749502
S01	A	One	S07	A	Three	-10.219828937044038	5.942559261839753e-11
S01	A	One	S08	B	Three	-8.726211855387541	1.780215893927293e-09
S01	A	One	S09	C	Three	-12.770692413106488	3.3829732259150916e-13
S02	B	One	S03	C	One	-0.6898899261310174	0.4959421164286606
S02	B	One	S04	A	Two	-5.554369135737531	6.110293048959636e-06
S02	B	One	S05	B	Two	-3.8077106417771502	0.0007019788123136649
S02	B	One	S06	C	Two	-3.8466655988061014	0.0006329293925833884
S02	B	One	S07	A	Three	-9.186060995965414	6.058783196021892e-10
S02	B	One	S08	B	Three	-8.196938152905897	6.371255768134509e-09
S02	B	One	S09	C	Three	-11.140514903792283	8.421885231843345e-12
S03	C	One	S04	A	Two	-5.2932690243231235	1.2444881215746578e-05
S03	C	One	S05	B	Two	-3.3595259949942444	0.002266846443275985
S03	C	One	S06	C	Two	-3.409491529747493	0.0019931134206586454
S03	C	One	S07	A	Three	-9.246539816877233	5.268953279543267e-10
S03	C	One	S08	B	Three	-8.096689165152357	8.145306818683346e-09
S03	C	One	S09	C	Three	-11.430875736565584	4.645926964395509e-12
S04	A	Two	S05	B	Two	1.6660919050282847	0.1068454269005213
S04	A	Two	S06	C	Two	1.2567488062031342	0.21922811783767399
S04	A	Two	S07	A	Three	-5.554865176143489	6.102058412352527e-06
S04	A	Two	S08	B	Three	-4.5153567474852006	0.0001042045407763848
S04	A	Two	S09	C	Three	-8.2495956791776	5.60294441912632e-09
S05	B	Two	S06	C	Two	-0.2461937427101593	0.8073281466237519
S05	B	Two	S07	A	Three	-6.279049306032996	8.666114443874442e-07
S05	B	Two	S08	B	Three	-5.3507621196298985	1.0638203014926822e-05
S05	B	Two	S09	C	Three	-8.560129806266595	2.6455211669627593e-09
S06	C	Two	S07	A	Three	-5.621804725790053	5.0873391593690716e-06
S06	C	Two	S08	B	Three	-4.856190628065189	4.109370448735146e-05
S06	C	Two	S09	C	Three	-7.704267401605672	2.157421416268622e-08
S07	A	Three	S08	B	Three	0.14941363298918126	0.8822982688978322
S07	A	Three	S09	C	Three	-2.0810886255974035	0.04668902779567866
S08	B	Three	S09	C	Three	-1.944598962575364	0.061930764107910014

After checking the p-value:

ANOVA is applicable. There are significant differences among the groups

ANOVA Results:

ANOVA Results:

	sum_sq	df	F	PR(>F)
C(Sample_Group)	3.145534e+06	8.0	39.464422	8.136908e-31
Residual	1.255363e+06	126.0	NaN	NaN

The ANOVA results indicate that there is a significant difference among the sample groups. The p-value for the C(Sample_Group) effect is very small (8.136908e-31), which is less than the typical significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there are significant differences in the mean viral load among the different sample groups. The ANOVA table also provides information about the sum of squares, degrees of freedom, and F-statistic for the C(Sample_Group) effect. The sum_sq column represents the sum of squares, the df column represents the degrees of freedom, the F column represents the F-statistic, and the PR(>F) column represents the p-value. The Residual row in the ANOVA table represents the sum of squares and degrees of freedom for the residual (unexplained) variation in the data. Overall, the ANOVA analysis indicates that there are significant differences in the mean viral load across the different sample groups.

Tukey's HSD post Hoc Test:

```
Tukey's HSD Post Hoc Test:
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
```

group1	group2	meandiff	p-adj	lower	upper	reject
S01	S02	-33.4	0.9	-148.4413	81.6413	False
S01	S03	-3.6	0.9	-118.6413	111.4413	False
S01	S04	172.8	0.001	57.7587	287.8413	True
S01	S05	122.3333	0.028	7.292	237.3746	True
S01	S06	131.4667	0.0128	16.4254	246.508	True
S01	S07	335.2667	0.001	220.2254	450.308	True
S01	S08	329.6	0.001	214.5587	444.6413	True
S01	S09	401.0	0.001	285.9587	516.0413	True
S02	S03	29.8	0.9	-85.2413	144.8413	False
S02	S04	206.2	0.001	91.1587	321.2413	True
S02	S05	155.7333	0.0012	40.692	270.7746	True
S02	S06	164.8667	0.001	49.8254	279.908	True
S02	S07	368.6667	0.001	253.6254	483.708	True
S02	S08	363.0	0.001	247.9587	478.0413	True
S02	S09	434.4	0.001	319.3587	549.4413	True
S03	S04	176.4	0.001	61.3587	291.4413	True
S03	S05	125.9333	0.0207	10.892	240.9746	True
S03	S06	135.0667	0.0092	20.0254	250.108	True
S03	S07	338.8667	0.001	223.8254	453.908	True
S03	S08	333.2	0.001	218.1587	448.2413	True
S03	S09	404.6	0.001	289.5587	519.6413	True
S04	S05	-50.4667	0.9	-165.508	64.5746	False
S04	S06	-41.3333	0.9	-156.3746	73.708	False
S04	S07	162.4667	0.001	47.4254	277.508	True
S04	S08	156.8	0.0011	41.7587	271.8413	True
S04	S09	228.2	0.001	113.1587	343.2413	True
S05	S06	9.1333	0.9	-105.908	124.1746	False
S05	S07	212.9333	0.001	97.892	327.9746	True
S05	S08	207.2667	0.001	92.2254	322.308	True
S05	S09	278.6667	0.001	163.6254	393.708	True
S06	S07	203.8	0.001	88.7587	318.8413	True
S06	S08	198.1333	0.001	83.092	313.1746	True
S06	S09	269.5333	0.001	154.492	384.5746	True
S07	S08	-5.6667	0.9	-120.708	109.3746	False
S07	S09	65.7333	0.6567	-49.308	180.7746	False
S08	S09	71.4	0.5643	-43.6413	186.4413	False

The Tukey's HSD post hoc test results indicate the significant differences between pairs of sample groups. The reject column specifies whether the null hypothesis of equal means is rejected for each pair. Based on the post hoc test results:

Sample Group S01 (Drug A, Intake frequency 1) shows significant differences in mean viral load compared to Sample Groups S04, S05, S06, S07, S08, and S09.

Sample Group S02 (Drug B, Intake frequency 1) shows significant differences in mean viral load compared to Sample Groups S03, S04, S05, S06, S07, S08, and S09.

Sample Group S03 (Drug C, Intake frequency 1) shows significant differences in mean viral load compared to Sample Groups S04, S05, S06, S07, S08, and S09.

Sample Group S04 (Drug A, Intake frequency 2) shows significant differences in mean viral load compared to Sample Groups S07, S08, and S09.

Sample Group S05 (Drug B, Intake frequency 2) shows significant differences in mean viral load compared to Sample Groups S06, S07 and S08.

Sample Group S06 (Drug C, Intake frequency 2) shows significant differences in mean viral load compared to Sample Groups S07, S08, and S09.

Sample Group S07 (Drug A, Intake frequency 3) shows significant differences in mean viral load compared to Sample Group S09.

Sample Group S08 (Drug B, Intake frequency 3) shows significant differences in mean viral load compared to Sample Group S09.

These results help identify specific pairs of sample groups that have significantly different mean viral loads. The meandiff column provides the difference in means, and the p-adj column provides the adjusted p-value.

Please note that the interpretation of the post hoc test results should consider the specific research question, significance level, and context of the study.

44.1. Conclusions

Based on the data analysis performed using ANOVA and Tukey's HSD post hoc test, the following conclusions can be drawn:

- According to the data, the groups differ significantly in their ability to stop the virus from reproducing. The ANOVA's p-value of 8.136908e-31 indicates significant evidence against the null hypothesis of equal means. It may be inferred from this that the various pharmaceuticals (drugs A, B, and C) and ingestion frequency (1, 2, or 3 dosages per day) have a substantial impact on the viral reproduction cycle.
- Further findings from the Tukey's HSD post hoc test show that particular pairs of sample groups had significantly different mean virus loads. These variations are seen in several groups, showing varied degrees of success in thwarting the virus. The post hoc test results, however, show that there is no statistically significant difference between Drug A and Drug B in reducing viral load.
- The Cohen's d value for the effect size is 0.3157, which denotes a negligibly little effect. As a result, it appears that the variations in mean viral load between the sample groups are only mildly different.

In conclusion, the analysis demonstrates that different medical substances and intake frequencies have a significant impact on inhibiting the reproduction cycle of the virus. The findings suggest that certain substances and intake frequencies are more effective than others in reducing viral load. These results have implications for designing treatment strategies and optimizing drug regimens to effectively inhibit the virus's reproduction cycle.

45. Task 3

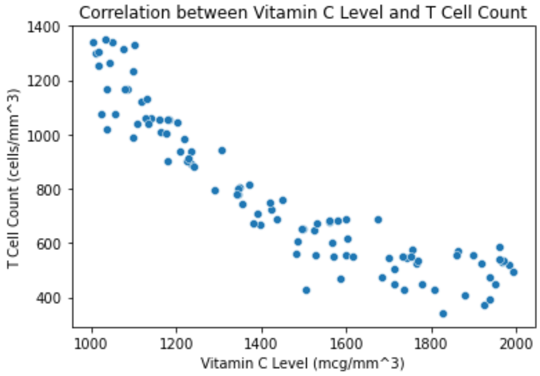
45.1. <u>Selection of statistical test</u>

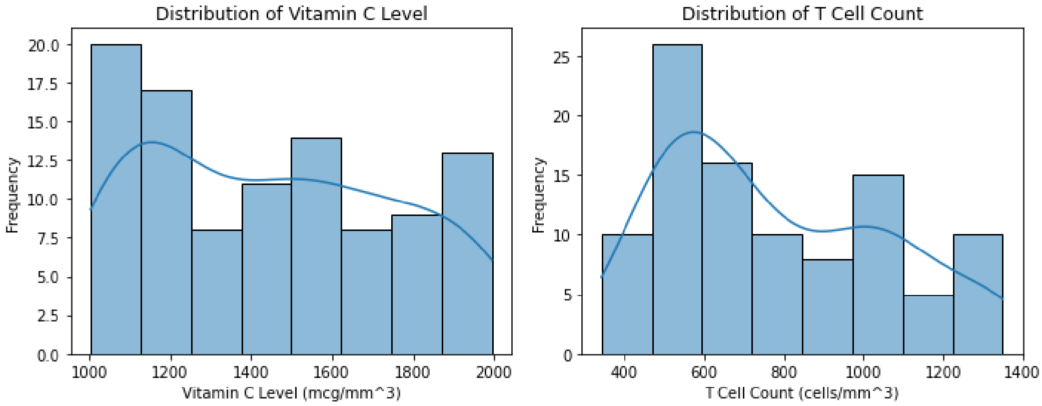
Regression Analysis is my chosen statistical set for Task 3. The reasons for it are:

- Regression analysis is a statistical method that shows the relationship between two or more variables. Usually expressed in a graph, the method tests the relationship between a dependent variable against independent variables.
- We need to find a correlation between the amount Vitamin C level and no. of T cells in the blood plasma.
- Here, independent variable, x = Vitamin C level in blood plasma and
- Dependent, y = amount of T cells in the blood plasma

45.2. <u>Statistical analysis</u>

Correlation coefficient: -0.9151056797968393
A correlation coefficient of -0.9151 indicates a strong negative correlation between the two variabl
es being analysed. It suggests that there is a strong inverse relationship between the variables, i.
e., the vitamin C levels in blood plasma and the amount of T cells.





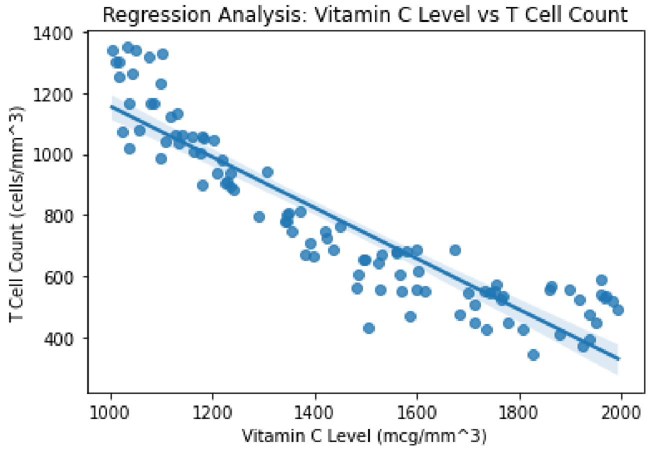
95% Confidence Interval: (-0.9949379947845668, -0.8352733648091119)
The range of numbers between which we are 95% convinced that the genuine population correlat
ion coefficient lies is represented by the 95% confidence interval of (-0.9949, -0.8353).

The confidence interval in this instance points to a likely negative true population correlation coeffi
cient between the two variables. A high negative correlation is indicated by the interval's lower bo
und (-0.9949), while a moderately negative correlation is indicated by the interval's upper bound (
-0.8353).

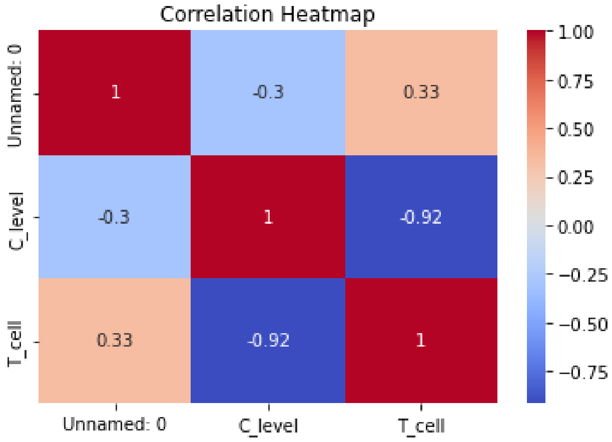
Based on the sample data, the confidence interval offers a range of likely values for the populatio
n correlation coefficient. The more precise our estimate of the genuine correlation coefficient, the

narrower the interval. The absence of zero from the interval lends even more credibility to the noti on that there is a negative correlation between the variables.

Regression Coefficients:
Slope: -0.8309276662860462
Intercept: 1987.044866437332
According to the regression coefficients, the line has an intercept of 1987.0449 and a slope of -0.8 309. The slope shows how much the dependent variable changes when the independent variable changes by one unit. In this instance, the slope indicates that the dependent variable falls by 0.83 09 units for every unit increase in the independent variable.
R-squared: 0.8374184051964356
The R-squared value of 0.8374 indicates that 83.74% of the variation in the dependent variable ca n be explained by the independent variable. This suggests that the model is a good fit for the data , as higher the R-squared value, better fit is the model.
P-value: 1.9293289425637757e-40
The two variables appear to be statistically related, as indicated by the p-value of 1.9293e-40. Sin ce a p-value of 0.05 is typically considered to be the cutoff for statistical significance, the incredibl y low p-value in this instance strongly supports the existence of a relationship.
Standard Error: 0.03698406300276937
The residuals around the regression line are variable, as indicated by the standard error of 0.0369 . The confidence intervals and hypothesis tests for the regression coefficients are computed using it.



Correlation Heatmap:



45.3. Conclusions

The results show a strong inverse relationship between vitamin C levels and T cell counts. A lower T cell count is linked to higher vitamin C levels. This association is supported by the regression analysis, which demonstrates that vitamin C levels have a considerable impact on the number of T cells. These findings emphasise the possible impact of vitamin C on T cell numbers and add to our understanding of the function of the immune system. It is crucial to take these findings into account in the context of the study and to be aware of any limitations or potential needs for additional research.

Also, an extremely low p-value of 1.9293e-40, further affirms that there is a significant relationship between the vitamin C levels and T cells and is unlikely to be due to random chance.