COMM 215 Case Analysis Project
Section: C

Professor: Wissam Nawfal M.Sc.

Date: December 11, 2015

Prepared by:

Anael Abitbol (27390440)
Jonathan Amar (27566107)
Ryan Bickerton (27443455)
Matthew Cowan (27756704)
Gabriel Faubert (7760817)
Mrittika Ghoose (27759576)
Noah Green (27444087)
Eli Levitas (27419600)
Robert Lussier (27498381)
Samantha Maldoff (27006802)
Jonathan Marciano (27429339)

Executive Summary

This study is being conducted thanks to a private experimental biology laboratory that has begun experimenting and testing a revolutionary new drug. After conducting many specific tests in order to produce results for the Lab's data, Component 1 takes the shortest amount of time to become effective. This report provides different statistical tests on the effectiveness of the different drug components in order to provide the private lab with the proper results.

Introduction

We are a team of data analysts and business consultants and we have been assigned a consulting contract of a private experimental biology laboratory called The Lab. The Lab has been conducting different experiments in order to understand the effectiveness of three components of a revolutionary drug and one final comprehensive experiment, which combines all of these three components together. The Lab has given us information to work with, which includes the success rate for each time that every drug has been used, as well as the time it took for each drug to start working. We have also been provided with the success rate for the final comprehensive experiment. Our team's job is to provide The Lab with a statistical report on the results of their experiment in order to help them decide whether or not to continue with experimentation and to better understand the possible implementation of one component of the drug compared to combining all three components.

Discussion

When analyzing the different aspects of descriptive statistics (see Appendix 1), we can conclude that Component 1 has the largest variance, meaning that the data is very spread out and it also takes the shortest amount of time to take effect. Since the variance in Component 1 is higher than any other, we understand that this component may not be as trustworthy as the others because the time it takes for the drug to be effective is hard to guess. From these descriptive statistics, we are not confident that combining all three components would be beneficial; as their maximum combined effectiveness does not reach 100%. However, further testing may prove otherwise. When observing the effectiveness of drug ranges (see Appendix 2), it can be concluded that the data does not follow a normal distribution, and all three graphs are left-skewed. Component 3's histogram's frequencies fall within the largest ranges, between 80% and 100%. Similarly, when examining the range of time of each drug component's effectiveness, we found all results to again, be left-skewed. The interquartile range for Component 2 is the highest, meaning that the data set is very spread out from the median effectiveness value. After creating a box and whisker plot to compare the range of effectiveness of each drug, we can conclude that Component 2 is the most effective component since the other component's minimum effectiveness is higher (see Appendix 4). The results from the second box and whisker plot (see Appendix 5) conclude that Component 2 took the longest amount of time to complete, at 2603 seconds, while Component 1 took the shortest amount of time, at 1940 seconds. In each scatter plot, we see that as time goes up, it doesn't necessarily mean that the percentage of effectiveness goes up as well. Therefore, when trying to understand the time it took each drug component to become affective at its level of effectiveness, we concluded that there is no clear relationship between time in seconds and percentage of effectiveness in either of them (see Appendix 6). The scatter plot for component 2 shows more outliers than any other component, and the scatter plot for component 1 shows the most evenly distributed set of data compared to the other components. We found a weak positive relationship when observing the average effectiveness of all three components for each attempt. We calculated the correlation coefficient using Excel and the result was 0.394, showing a weak positive relationship between the time it took each drug component to become affective at its level of effectiveness (see Appendix 7). When testing and questioning the significance and correlation between x and y, the coefficient of determination and the regression slope coefficient, we determined that they are all equivalent and produce the same conclusion. There is not sufficient evidence to strongly suggest that time and the results are correlated, or that the time may explain the variance in the result percentages. However, we would still advise that the component 3 is the one where the effectiveness decreases the fastest, making it the worst component in terms of duration of drug efficiency and component 2 has the highest increasing slope, making it the one that increases the most in an efficient way over time. We used the predictive equation (see Appendix 9) to describe the relationship between the average effectiveness of each drug component and the effectiveness of the final comprehensive experiment. By using the correlation coefficient of 0.394, we discovered that the coefficient of determination is 15.53% and this shows us that the total movement in y is explained by x. We can conclude that 16% of the variation in the experiment can be explained by the average effectiveness of the three components (see Appendix 9). We've concluded that we are 99% confident that the final experiment is effective within a specific range that is higher or lower than the population mean. Finally, since z observed is greater than z critical, we do not reject Ho. According to our multiple regression model (see Appendix 10), the adjusted R square value says that the success of the drug is

only 15.5% responsible due to three components that make up the new drug. From our results, we can conclude that the component 2 of the drug is the most vital part of the medication because its relationship explains more of the drugs success than any other component.

Conclusion

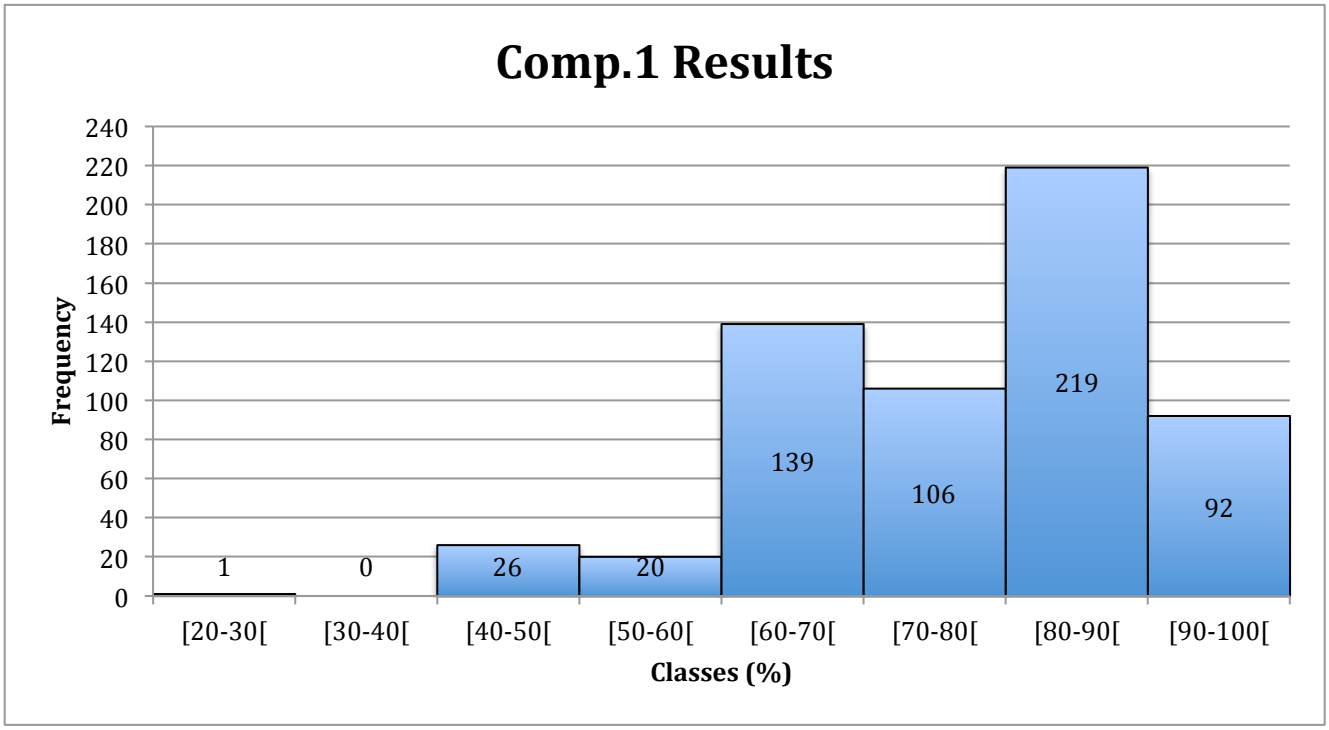
After conducting various statistical tests using the data given to us by the Lab's experiment, multiple conclusions have been drawn. The tests we have conducted include the mean effectiveness of each component, histograms, scatter plots, and correlating the relationship between drug effectiveness and time of effectiveness, we can conclude that the average time it takes for each drug component to become effective moves at a steady pace. We are 99% confident that the final experiment is effective on an average of 70% or more. The time it takes all three components to reach their effectiveness are quite different, which could explain the lower average effectiveness of the final drug. Finally, we suggest that combining all of the three drug components is the smart choice. We believe this is most optimal solution because all of the drug components take a different amount of time to become effective, and combining all of them would create a drug that is effective and long lasting.

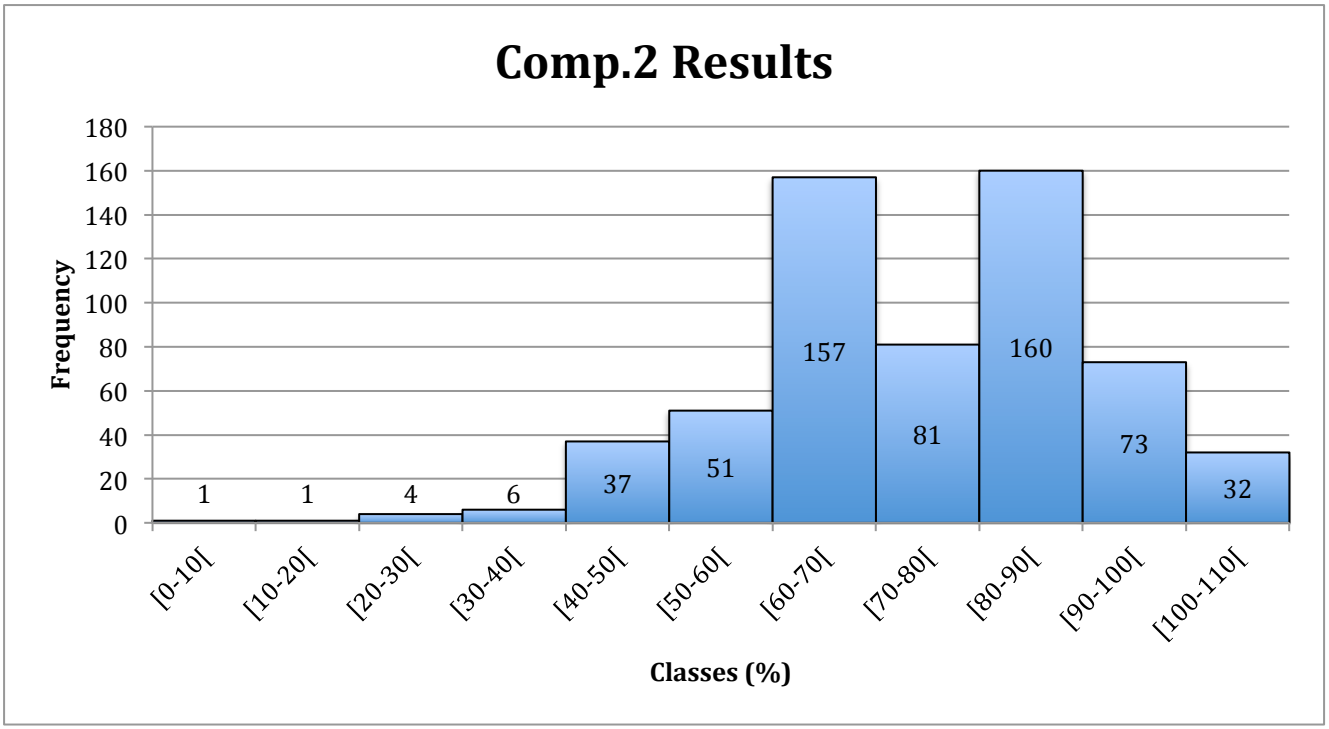
Appendix 1: Descriptive Statistics

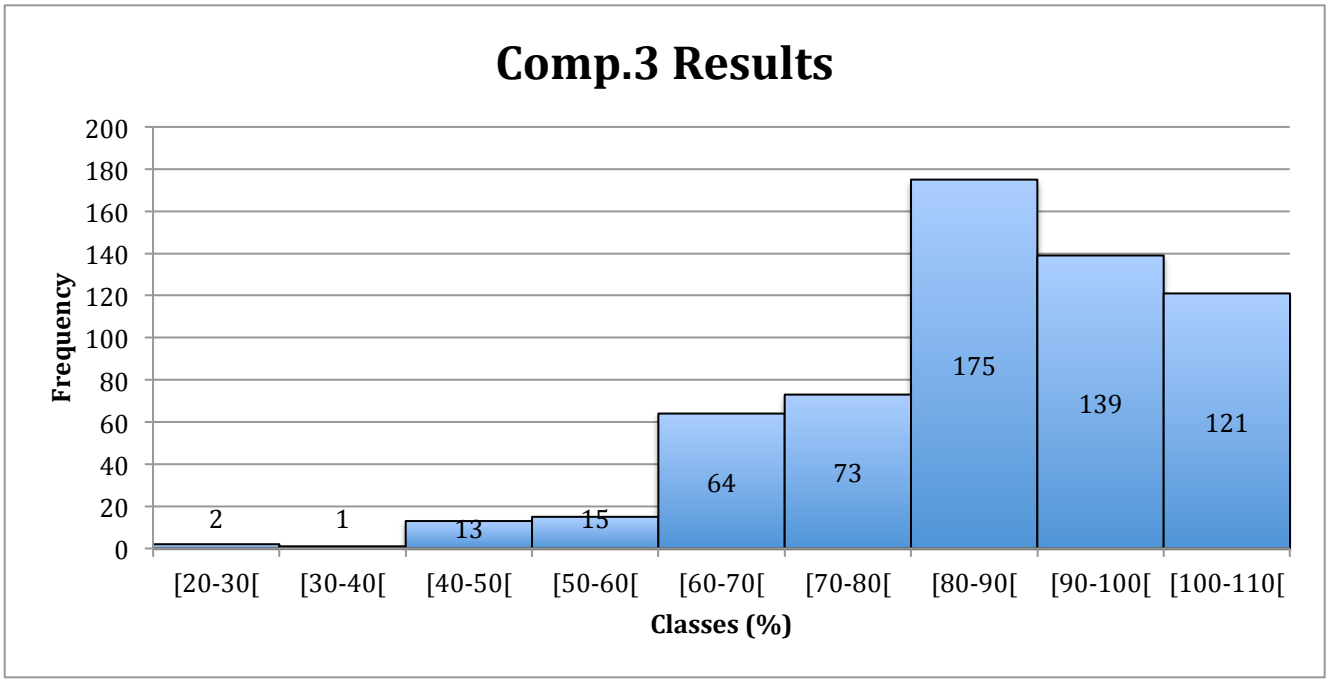
Effectiveness of each component test, including final				
	COMP 1.	COMP 2.	COMP 3.	Final Test
Mean	76.08%	72.96%	83.90%	71.77%
Mode	80.00%	66.67%	93.33%	88.33%
Median	80.00%	73.33%	86.67%	73.75%
St. Dev	13.37%	16.42%	14.40%	14.76%
Variance	0.017881161	0.026947825	0.020748258	0.021787925
Upper	116.67%	126.68%	123.33%	114.26%
Lower	36.67%	20.00%	43.33%	31.78%
1rst Quartile	66.67%	60.00%	73.33%	62.71%
2nd Quartile	80.00%	73.33%	86.67%	73.75%
3rd Quartile	86.67%	86.67%	93.33%	83.33%
Min	20.00%	0.00%	26.67%	25.00%
Max	100%	100%	100%	97.92%
Range	80.0%	100.00%	73.33%	72.92%
IQR	20.00%	26.67%	20.00%	20.62%

Time taken for each drug to become affective			
	COMP 1.	COMP 2.	COMP 3.
Mean	1911.509121	2382.461028	2269.613599
Mode	1940	2603	2489
Median	2700	2700	2700
St. Dev	594.0729967	445.4340446	527.2499596
Variance	352922.7254	198411.4881	277992.5199
Upper	3987.25	3459.5	3816.75
Lower	-62.75	1431.5	1204
1rst	1456	2192	1950.5
2nd	1940	2603	2489
3rd	2468.5	2699	2697
Min	196	88	301
Max	2717	3144	3221
Range	2521	3056	2920
IQR	1012.5	507	746.5

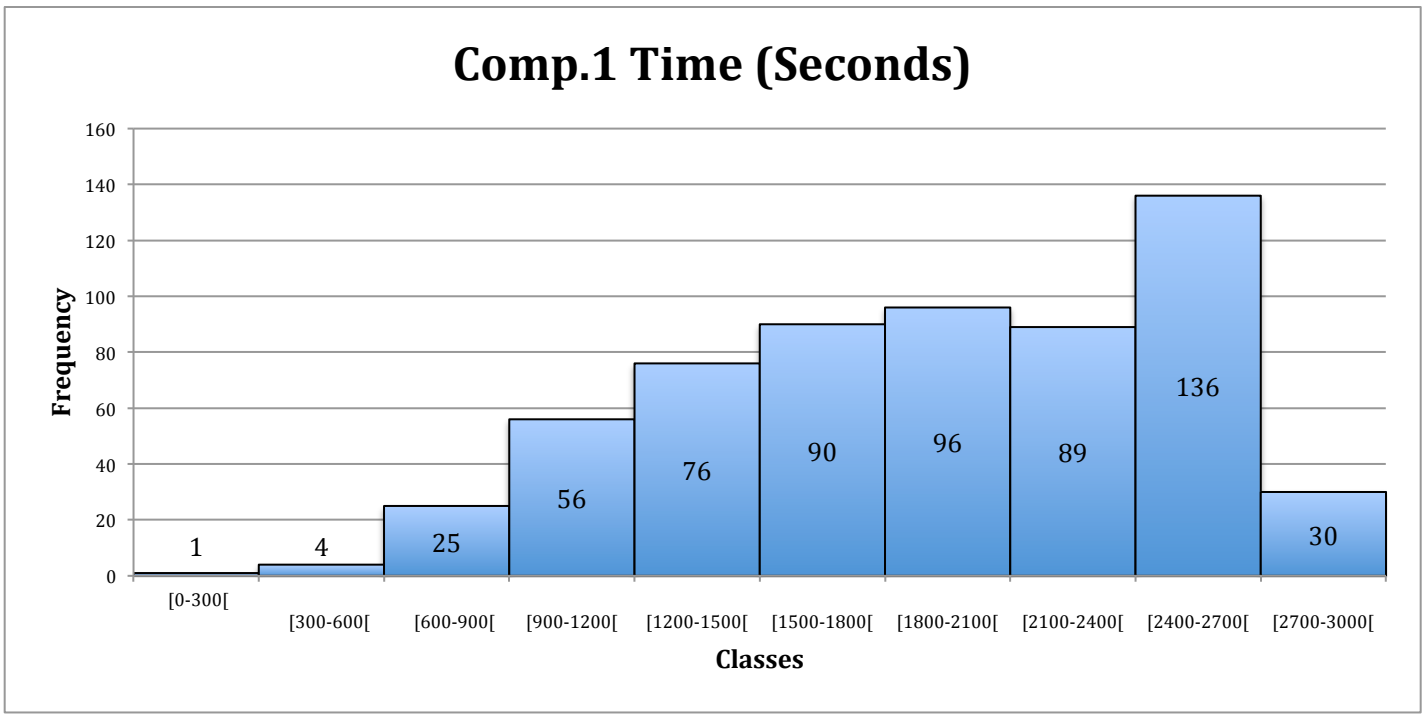
Appendix 2: The Range of Effectiveness of Each Drug Component and the Final Experiment

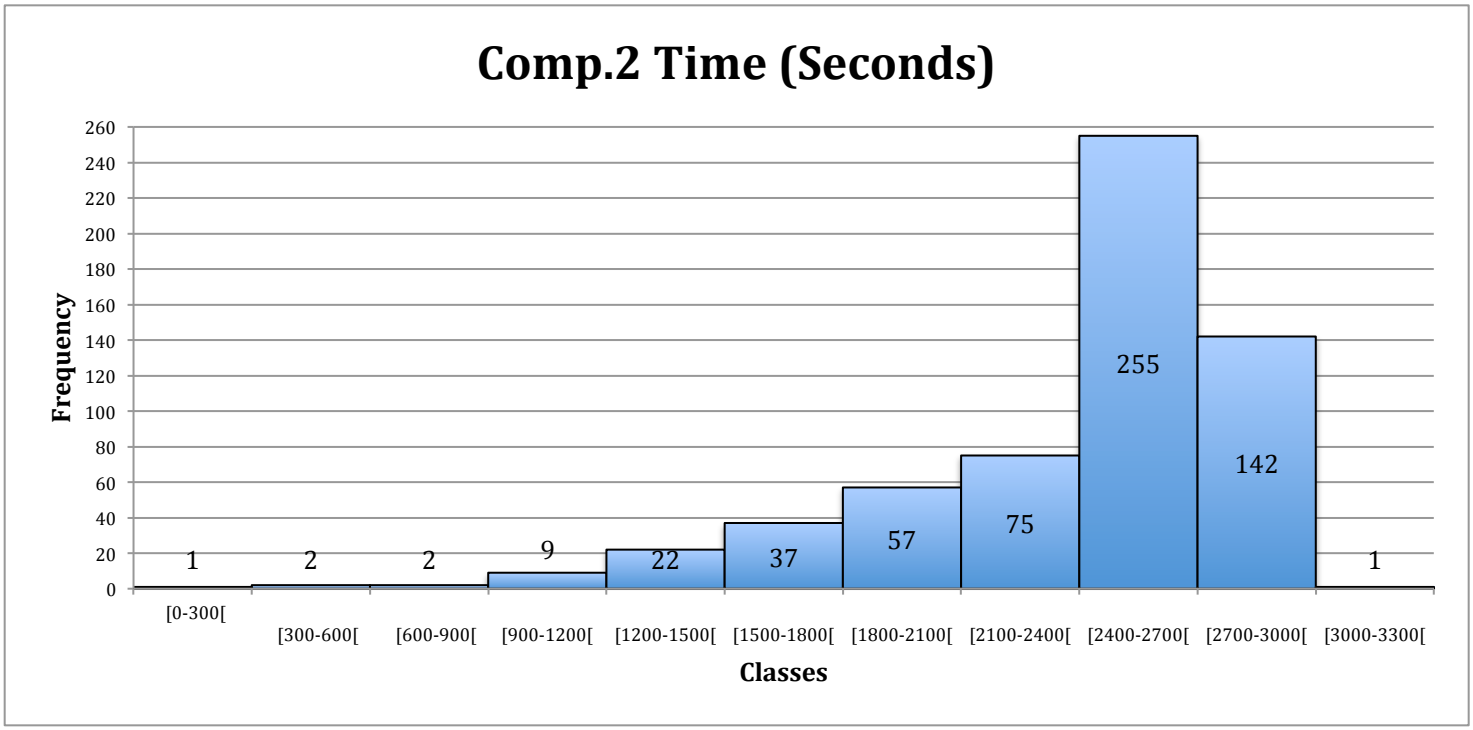


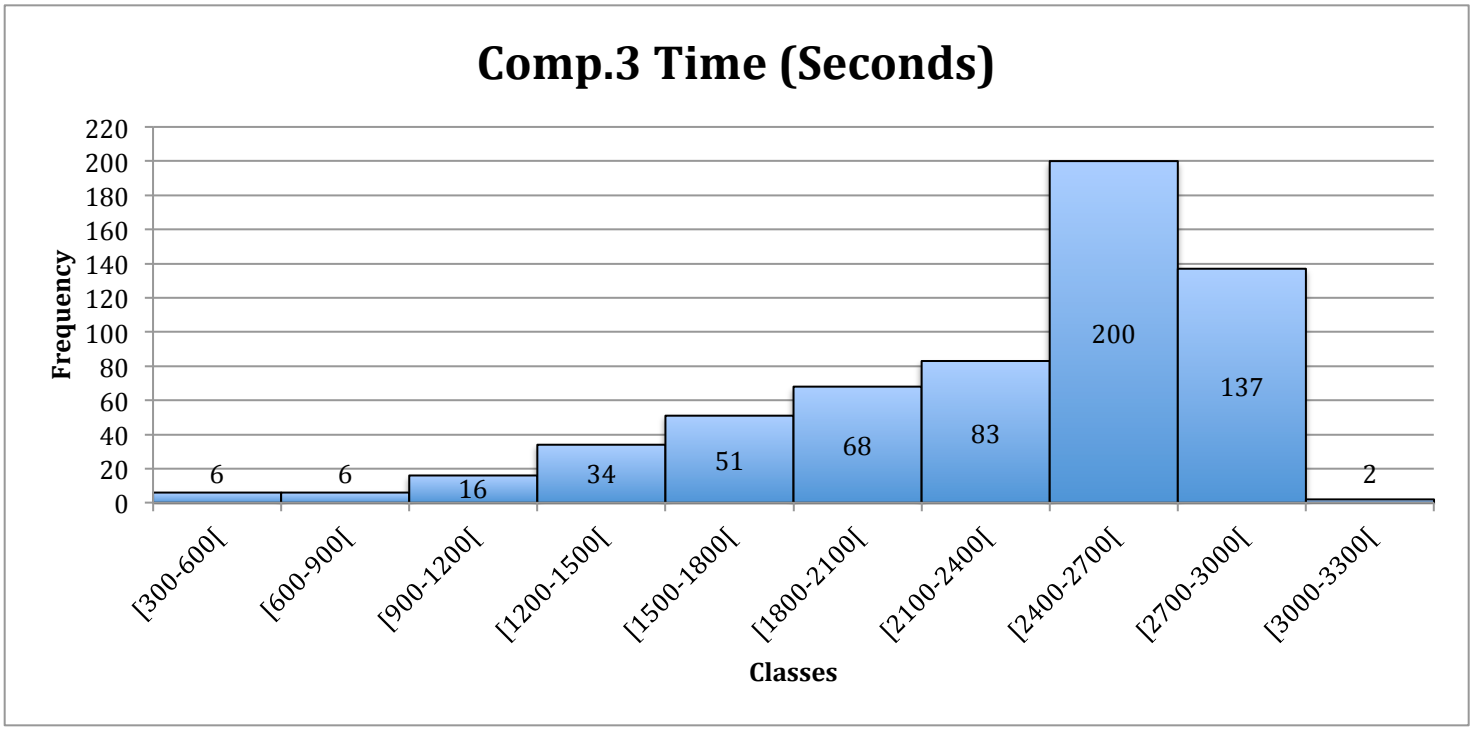




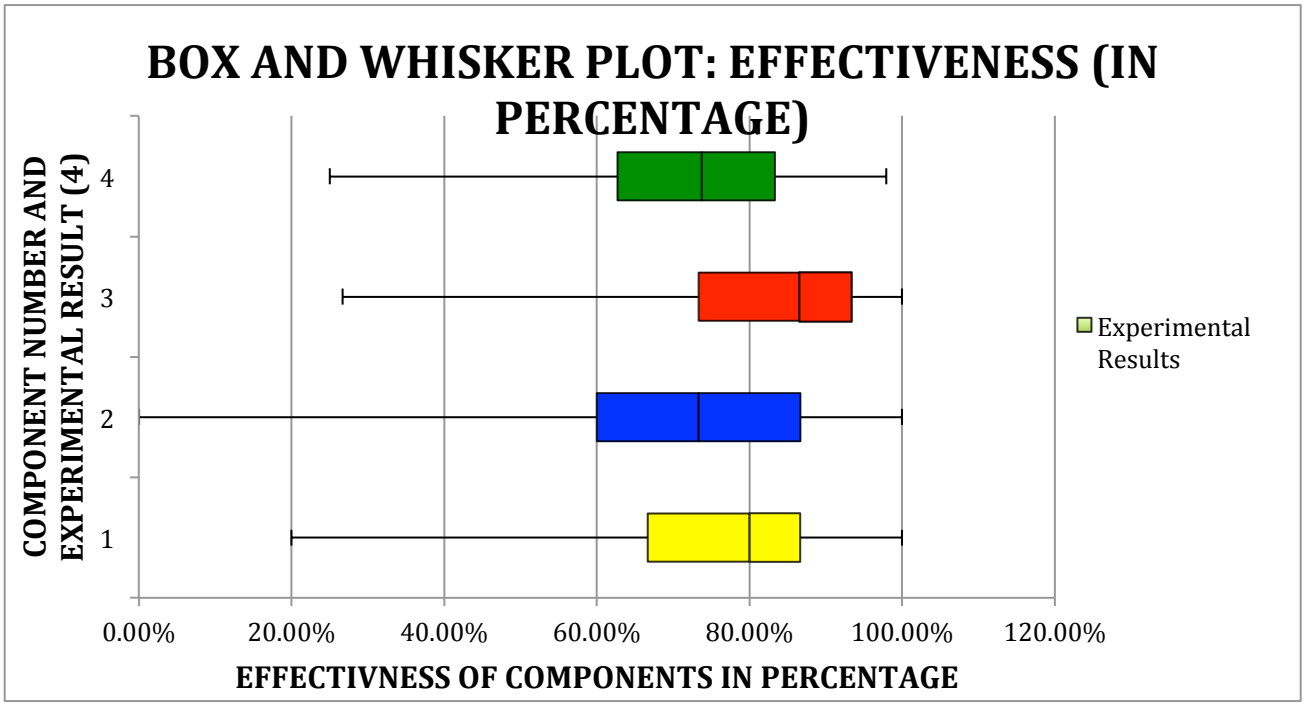
Appendix 3: The Range of time of each drug component's effectiveness



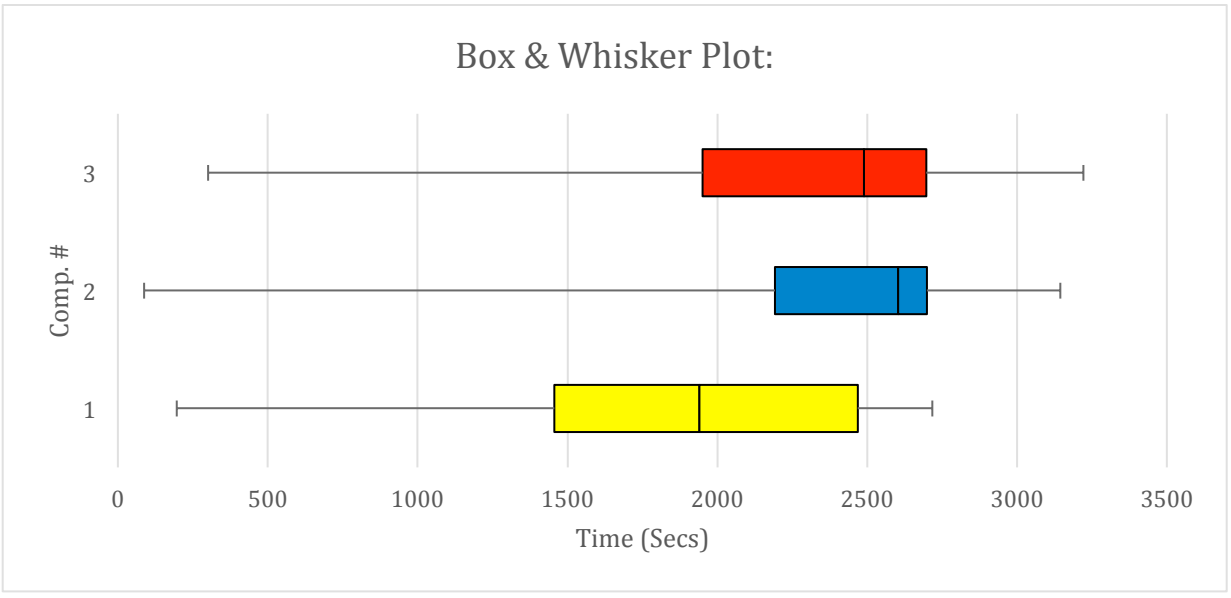

Comp.2 Time (Seconds)
Frequency
260
240
220
200
180
160
140
120
100
80
60
40
20
0
1
2
2
9
22
37
57
75
255
142
1
[0-300[
[300-600[
[600-900[
[900-1200[
[1200-1500[
[1500-1800[
[1800-2100[
[2100-2400[
[2400-2700[
[2700-3000[
[3000-3300[
Classes


Comp.3 Time (Seconds)
Frequency
220
200
180
160
140
120
100
80
60
40
20
0
6
6
16
34
51
68
83
200
137
2
[300-600[
[600-900[
[900-1200[
[1200-1500[
[1500-1800[
[1800-2100[
[2100-2400[
[2400-2700[
[2700-3000[
[3000-3300[
Classes

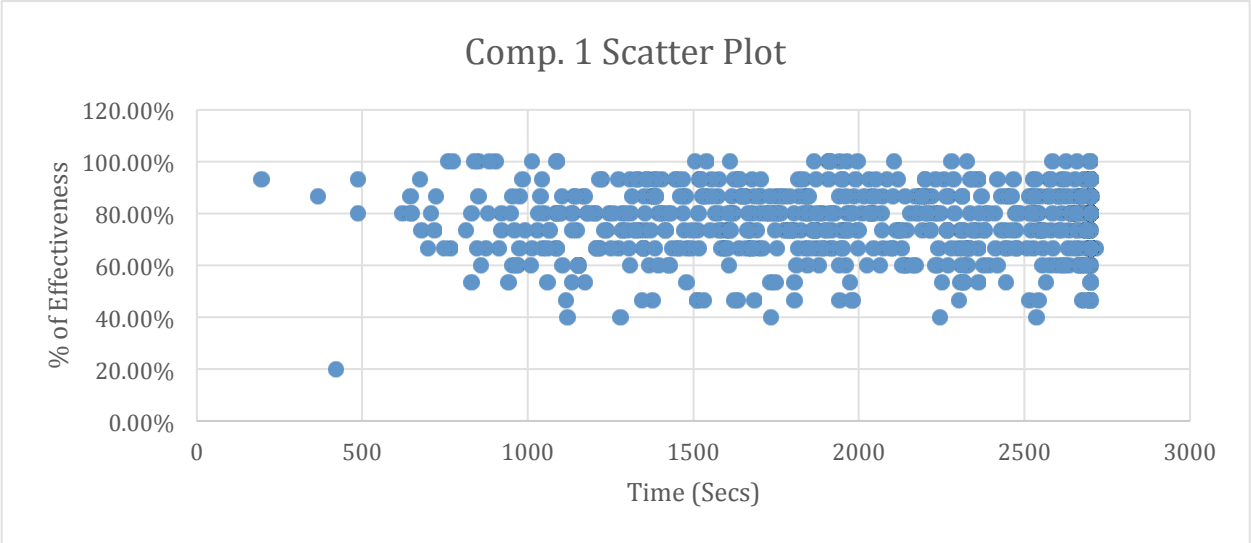
Appendix 4: Box and Whisker plot comparing the range of effectiveness of each drug component

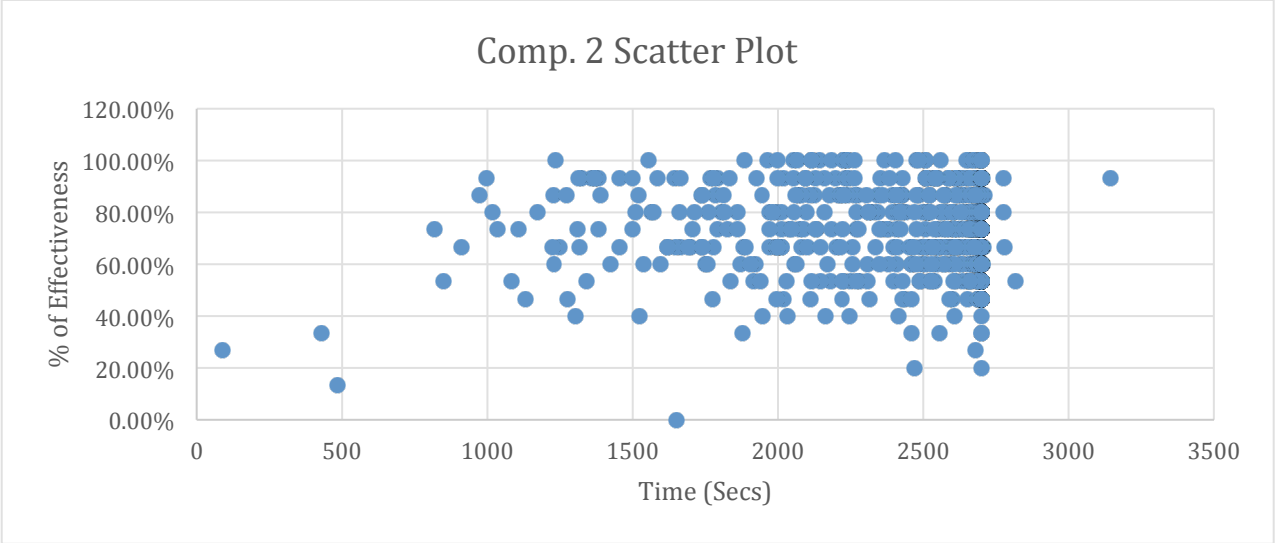


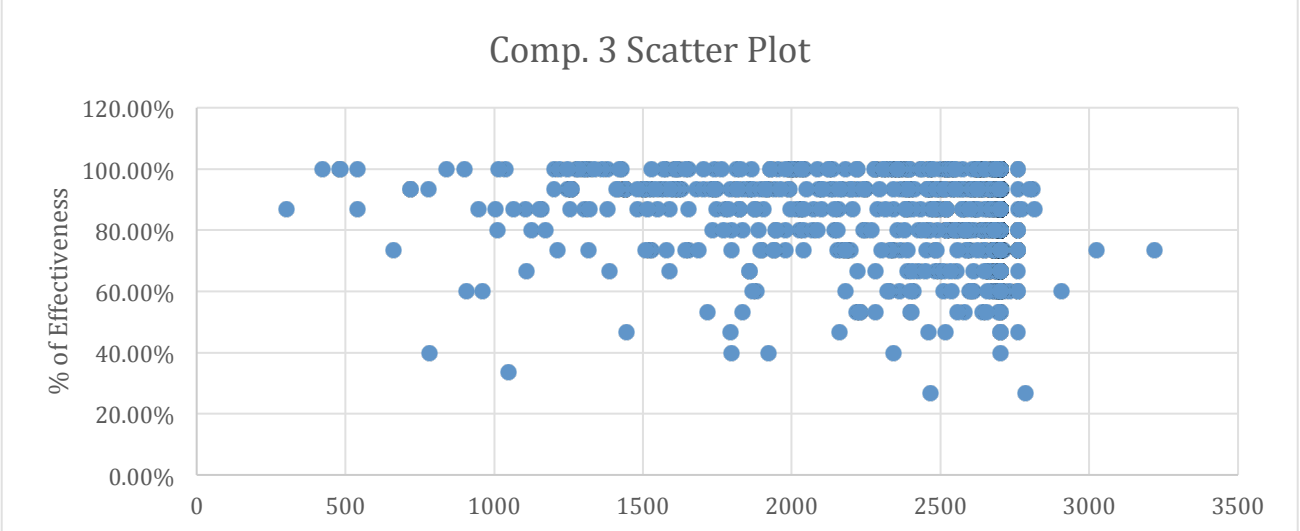
Appendix 5: Box and Whisker Plot Comparing the Range of Time of Each Drug Component's Effectiveness



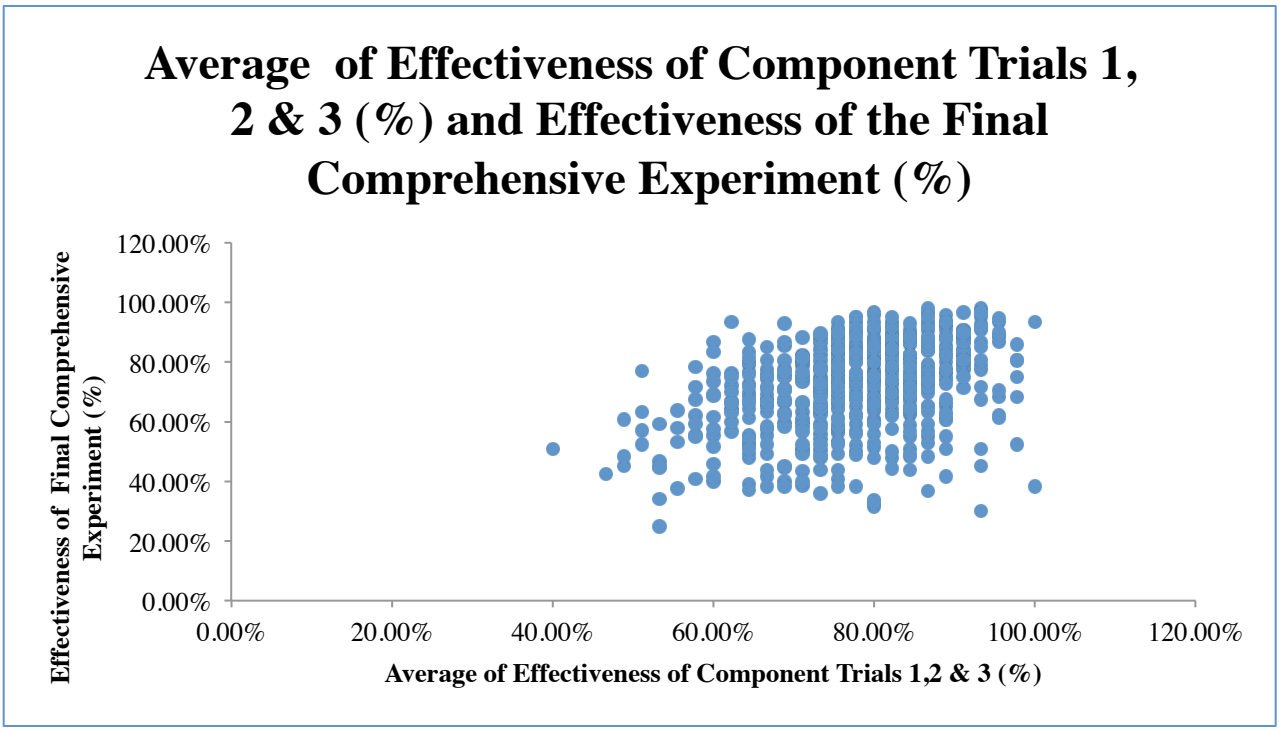
Appendix 6: Scatter Plot Comparing the Time it took Each Drug Component to Become Affective to its Level of Effectiveness







Appendix 7: Scatter Plot comparing the average effectiveness of each drug component trial to the effectiveness of the final comprehensive experiment



Appendix 8: Analysis of the relationship between the time it took each drug component to become affective to its level of effectiveness

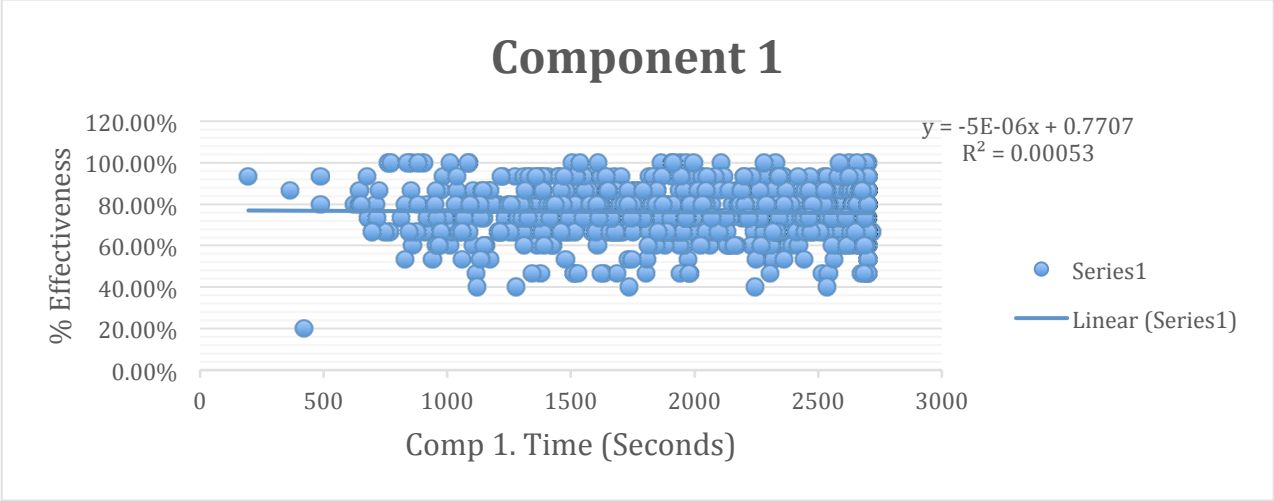
The predictive equation "**y = b0 + bixi**", is the equation involving the relationship between the time it took each drug component to become affective (x) to its level (%) of effectiveness (y).

Component 1: y= -5E-06x + 0,7707

Component 2: y = 1E-05x + 0,7034

Component 3: y = -4E-05x + 0,9388

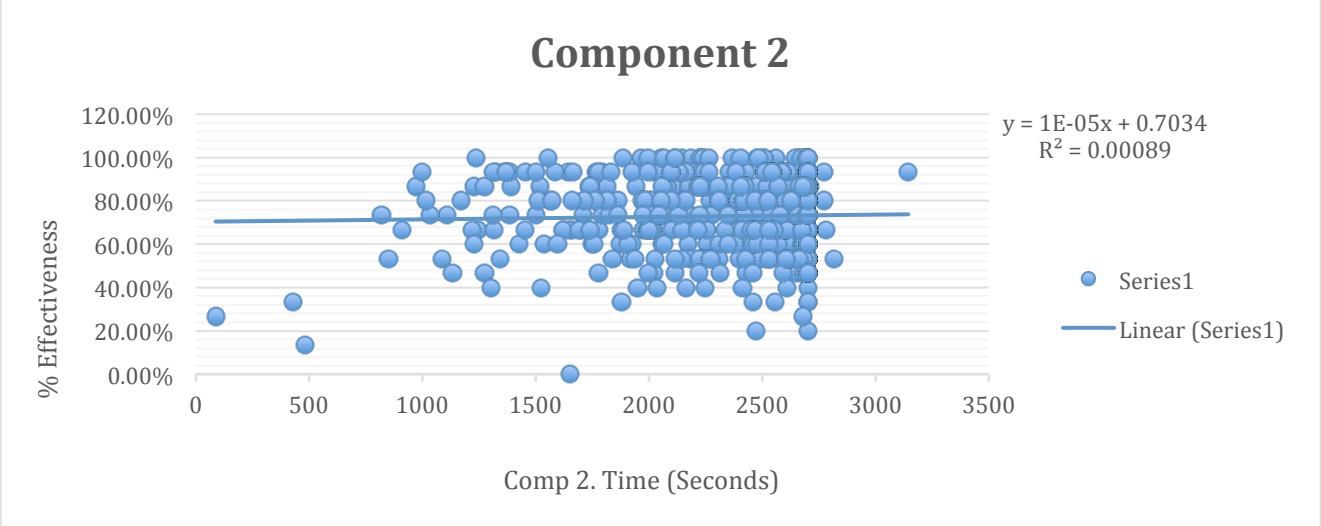
See below the graph and interpretation of the following equations;



R squared: Only .05% of the movement of Y is explained by x, this is extremely low, suggesting almost no relationship.

Slope: Using the graph and the slope function, we conclude that the slope is negative and incredibly small. **Slope is equal to** 5,19919E-06. It shows low almost (+) or (-) relationships between both time and success. The P-value given by the summary output of the linear regression is 57.13%, this is larger than any general alpha used (10%, 5%, or 1%) thus we can conclude that the slope is almost irrelevant (or that there is not a significant linear relationship between x and y variables).

Only at an 88% confidence level does the slope become significant.



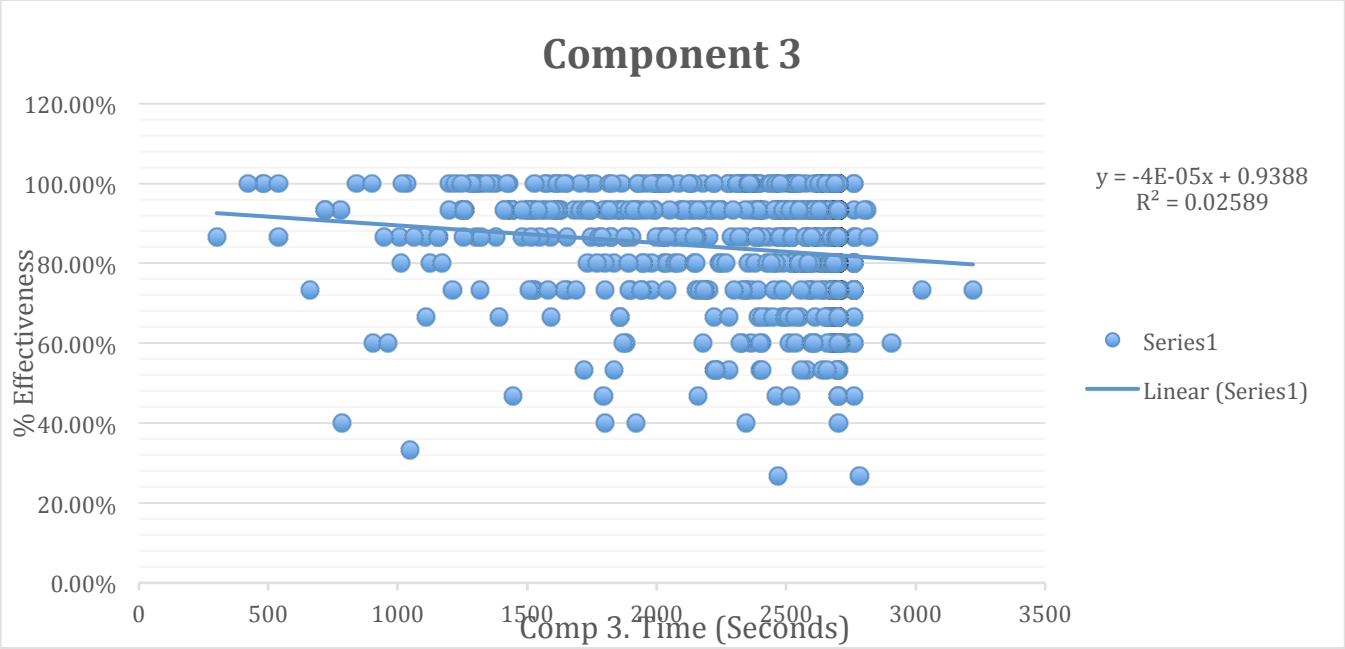
Interpretation of the predictive equation, its graph, and over all meaning for Component 2;

Correlation: The correlation coefficient is 0.02983 which is very weak, suggesting no relationship.

R Squared: Only 0.09 % of the movement of Y can be explained by X. This is so incrementally small that we can simply say that none of the movement of y is explained by x. Thus the movement in results is not explained by time.

Slope: The slope is equal to 1,09735E-05. This is very small, and proves that the change in Y is not caused so much by a change of X, so we think the slope is not significant. A P-value of 46.54% is given in the summary output of the regression. This is much larger than any of three alpha's usually used (10%, 5%, 1%), thus we can conclude that there is no significant linear relationship between x and y variables.

Only at around a 75% confidence level does the slope become significant.



Interpretation of the predictive equation, its graph, and over all meaning for Component 3;

Correlation: The correlation coefficient is -0.1609 which is weak, but is still more than component 1 and 2. Therefore, the relationship between component 3 and it's results is stronger than for component 1 or 2. However, let's not forget that the relationship is negative, making it the worst because it is negative and stronger than component 1 and 2.

R Squared: 2.5% of the change in Y is explained by X. Thus for component 3, about 2.5% of the change in effectiveness results can be explained by time.

Slope: A negative relationship can be easily determined on the graph. The P value of 0.64% indicates that even though the slope is extremely small it is significant. This is because we can reject the null hypothesis that the slope is equal to 0 considering a P-value of 0.64% is smaller than any of the 3 general alpha levels used (10%, 5%, 1%).

95% confidence interval around the mean time for each drug components effectiveness

- **Steps:**
  - Uses the descriptive statistic tool in the Data Analysis menu,
  - Select regions,
  - Select 95% interval

Component 1

1864 seconds <---------- 1911.51 seconds ----------> 1959.02 seconds

Component 2

2346.84 seconds <---------- 2382.46 seconds ----------> 2418.09 seconds

Component 3

2227.40 seconds <---------- 2269,61 seconds ----------> 2311.8 seconds

For interpretation purposes, transform seconds into minutes;

Component 1	31.07 min --- 31.86 min --- 32.65 min	(thus + or – .79 minutes)
Component 2	39.11 min --- 39.70 min --- 40.30 min	(thus + or - .47 minutes)
Component 3	37.12 min --- 37.83 min --- 38.53 min	(thus + or - .71 minute)

8 - C

Hypothesis test for the mean effectiveness of each drug component (with a 5% alpha)

Ho: u = 70%

Ha: u < 70%

Z-critical = 1,647392953

Component 1 = 78,93170433 z-observed

Component 2 = 131,3024341 z-observed

Component 3 = 105,6609081 z-observed

Rejection rule; If z-observed is smaller than z-critical reject Ho

Results: The Z-observed are biger for each component of the drug, then the Z-Critical which is 1.6474. Therefore, we **do not reject Ho.** In addition, this correlates with the fact that each P values are less than the alpha of .05%. It clarifies the significance difference.

*There is sufficient evidence to not reject Ho at an alpha of 5% (95% confidence level), thus we assume that the mean effectiveness of each component is less than 70%

COMPONENT 1 DESCRIPTIVE STATISTIC

- (95% CONFIDENCE)

Component 1: Time to become Effective	
Mean	1911,509121
Standard Error	24,19252264
Median	1940
Mode	2700
Standard Deviation	594,0729967
Sample Variance	352922,7254
Kurtosis	-0,88050741
Skewness	-0,331741638
Range	2521
Minimum	196
Maximum	2717
Sum	1152640
Count	603
Confidence Level(95,0%)	47,51199603

COMPONENT 2 DESCRIPTIVE STATISTIC

- (95% CONFIDENCE)

Component 2 : Time to become affective	
Mean	2382,461028
Standard Error	18,13947658
Median	2603
Mode	2700
Standard Deviation	445,4340446

Sample Variance	198411,4881
Kurtosis	2,941019428
Skewness	-1,719951609
Range	3056
Minimum	88
Maximum	3144
Sum	1436624
Count	603
Confidence Level(95,0%)	35,6243436

COMPONENT 3 DESCRIPTIVE STATISTIC

- (95% CONFIDENCE)

Column1	
Mean	2269,613599
Standard Error	21,4712782
Median	2489
Mode	2700
Standard Deviation	527,2499596
Sample Variance	277992,5199
Kurtosis	0,916504339
Skewness	-1,253682139
Range	2920
Minimum	301
Maximum	3221
Sum	1368577
Count	603
Confidence Level(95,0%)	42,16771024

Component 1

t-Test: Paired Two Sample for Means

	1406	*0,933333333*
Mean	1912,348837	0,760465116
Variance	353084,0545	0,017861273
Observations	602	602
Pearson Correlation	-0,021313259	

Hypothesized Mean Difference	0
df	601
t Stat	78,93170433
P(T<=t) one-tail	0
t Critical one-tail	1,647392953
P(T<=t) two-tail	0
t Critical two-tail	1,963919017

Component 2

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	2382,461028	0,72957435
Variance	198411,4881	0,026947825
Observations	603	603
Pearson Correlation	0,029775935	
Hypothesized Mean Difference	0	
df	602	
t Stat	131,3024341	
P(T<=t) one-tail	0	
t Critical one-tail	1,647388728	
P(T<=t) two-tail	0	
t Critical two-tail	1,963912434	

Component 3

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	2269,613599	0,839027087
Variance	277992,5199	0,020748258
Observations	603	603
Pearson Correlation	-0,160917466	
Hypothesized Mean Difference	0	
df	602	
t Stat	105,6609081	
P(T<=t) one-tail	0	

t Critical one-tail	1,647388728
P(T<=t) two-tail	0
t Critical two-tail	1,963912434

Appendix 9: Analysis of the relationship between the average effectiveness of each drug component trial and the effectiveness of the final comprehensive experiment

a) The following equation is the predictive equation that describes the relationship between the average effectiveness of each drug component trial (x) and the effectiveness of the final comprehensive experiment (y)

$y = b0 + bixi$

$y = 0.289 + 0.557x$

Correlation: The results on the scatter plot (see Appendix 7) determine a correlation coefficient of 0.394 which shows that the relationship between the two variables are linear and positive however not very strong.

R Square: The correlation coefficient lead us to the coefficient of determination which in this case is 15.53%. It shows that 15.53% of the total movement in (y) is explained by its relationship with (x). Thus it can be concluded that approximately 16% of variation in the experiment results can be explained by the average effectiveness of the three components.

b) 99% confidence interval around the mean for the effectiveness of the final comprehensive element :

Results:

Final experiment -

70.22% ——————————— 71.77% ———————————73.32%

Interpretation: Based on the results, with 99% confidence level, this range will contain the true average effectiveness of the final experiment. We are 99% confident that the final experiment is effective within a range that is 1.55% higher or lower than the population mean.

c) Test hypothesis for the mean effectiveness of the final comprehensive experiment (1% alpha)

$H_0: \mu > 70\%$

$H_a: \mu \leq 70\%$

Results:

z critical= 2.575

z observed= 2.94

Interpretation: Since z observed is greater than the z critical, the rejection rule says that the Ho cannot be rejected. Thus, it can concluded that the mean effectiveness of the final comprehensive experiment is greater than 70%.

Appendix 10: Multiple Regression Equation

Regression Statistics	
Multiple R	0.39932256
R Square	0.15945851
Adjusted R S	0.15524878
Standard Err	0.13566642
Observations	603

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2.09151052	0.69717017	37.8786156	2.0033E-22
Residual	599	11.0248204	0.01840538		
Total	602	13.1163309			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	0.28909501	0.04377012	6.60484825	8.7897E-11	0.20313345	0.37505656	0.1759903	0.40219971
Comp. 1 Resul	0.15384474	0.04259966	3.61140742	0.00033007	0.07018189	0.2375076	0.04376457	0.26392491
Comp.2 Result	0.2428489	0.03506579	6.92552114	1.1231E-11	0.17398206	0.31171575	0.15223672	0.33346109
Comp. 2 Resul	0.16015221	0.04007182	3.99662971	7.2262E-05	0.08145388	0.23885054	0.05660416	0.26370026

Multiple Regression Model

Y(hat) = 0.289095005824418 + 0.153844742381727 x+ 0.242848901466937x+0.160152207037704x