Jennifer Chun
MATH 123 Final Project
Meshkat
6/10/20

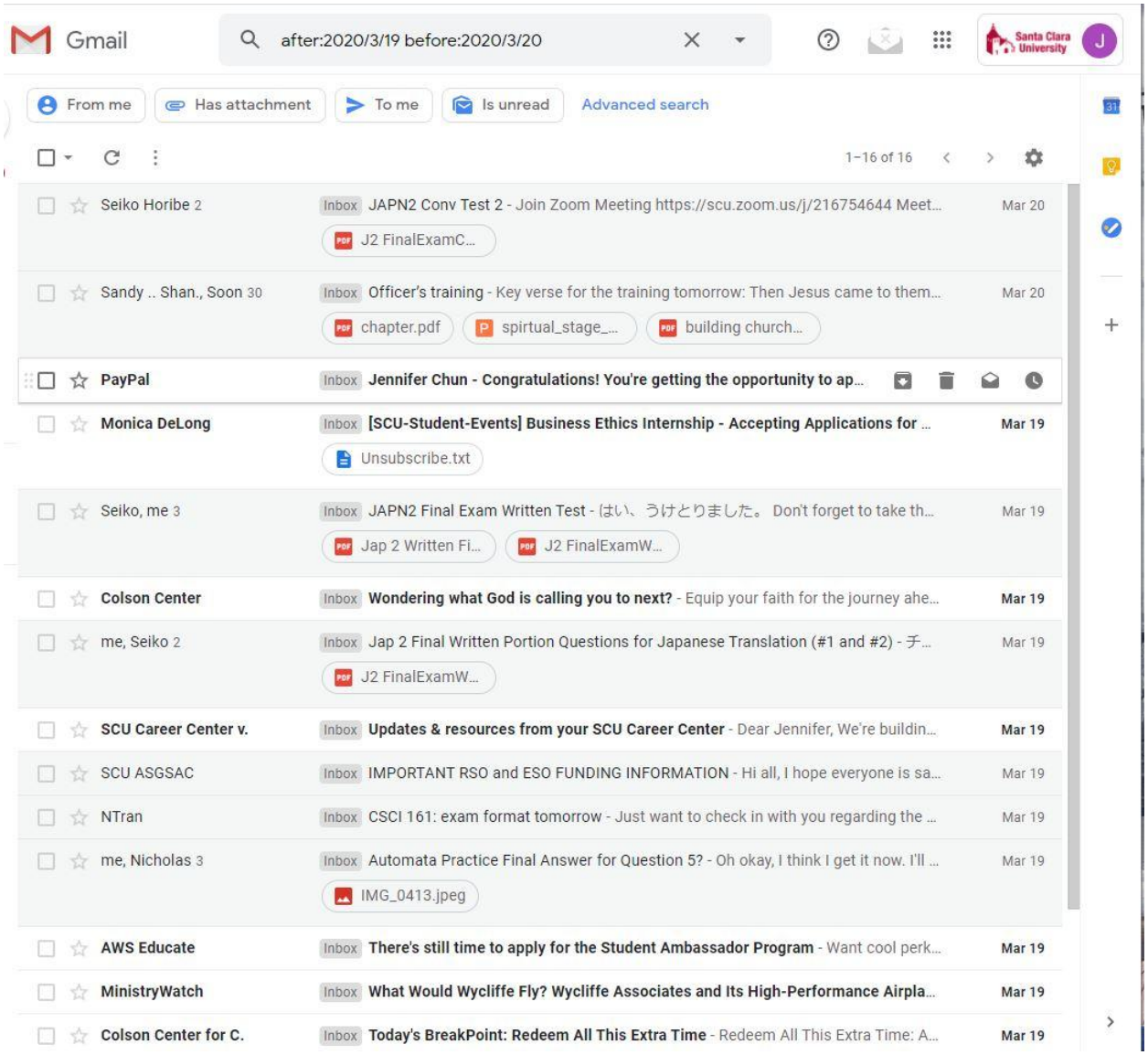
Types And Totals Of Emails Before And After Santa Clara County's COVID -19 Shelter-in-Place Orders

Note: I utilized GNU Octave (a free, open-source Matlab software clone, found at https://www.gnu.org/software/octave/) and its statistics package (found at https://octave.sourceforge.io/statistics/) to help me with this project. Octave was used primarily to help with the tedious work (i.e. calculating the mean of the number of school emails I received over a time period) and with calculating the values of the Least-Squares test through the *plsregress(X,Y,NCOMP)* function. However, I did all the *Z*-test hypothesis testing calculations by hand.

For my statistics final project, I wanted to test whether the emails that I have received before and after the Shelter-in-Place orders in Santa Clara County have changed.

The data source for my sample was my SCU email inbox. I never delete any emails, so the number of emails I counted matches the number of emails that I truly received on those days. The Santa Clara County Shelter-in-Place order (utilized to "flatten the curve" of the total number of COVID-19 cases) officially started on March 16th at 12 AM. So, I decided to have my first data set (which represented the emails I received before the Shelter-in-Place order) go up to March 15th and to have my second data set (which represented the emails I received after the Shelter-in-Place order) start on March 16th. I decided to have both data sets contain 70 days: January 4th - March 15th ("Before") and March 16th - May 25th ("After"). To ensure that the data would truly be of a random sample, I utilized the *sort(randperm(n,k))* function from GNU Octave. This function served to create a random permutation of n non-repeating numbers from k total numbers and sort the permutation of numbers from smallest to largest. I chose to have my random sample be 50 days, so I ran *sort(randperm(50,70))* twice, once for each sample group. Then, I numbered the days in each data set by Day 1 to Day 70. For example, in my "Before" data set, Day 1 was January 4th and Day 70 was March 15th. After, I found the exact dates corresponding to each day number. From there, I meticulously used Gmail's advanced search bar to help me find all the emails from each random sample day. I glanced through all the email senders and email headers and grouped the emails into five categories: total number of emails, school emails (anything related to SCU, including club emails), Handshake emails (since I noticed that I was receiving a lot of emails from them), spam emails (anything that tried to sell me something or convince me to sign up for something I didn't want), and other miscellaneous emails (daily Christianity subscription emails, etc.). I put the total number of each group of emails into a Google Doc.

Example: Finding the number of each type of email for March 19th 2020



The numbers on my Google Doc corresponded with the total number of emails, school emails, Handshake emails, spam emails and miscellaneous emails, respectively.

- 19-Mar 16 9 0 3 4

Afterwards, I compiled all the numbers from each category into their own array in Octave. Since I had five categories for each data sample, I created ten arrays.

<u>B : Before March 16th 12 AM</u>

B_total = [32 33 23 38 25 25 17 29 27 15 17 16 9 18 9 13 16 24 18 24 13 7 20 26 8 27 10 13 27 18 23 10 10 34 8 38 17 20 23 29 9 32 20 10 19 19 23 26 27 9 16] ;

B_school = [12 17 15 21 15 14 11 24 15 7 11 3 4 10 3 6 8 12 9 14 6 3 13 15 4 13 6 6 12 13 9 6 6 24 5 23 10 13 15 19 5 16 7 6 11 10 15 20 11 3 11];

B_handshake = [7 7 3 6 1 3 1 0 2 1 1 4 0 1 0 0 1 2 2 2 0 0 1 1 0 5 0 0 5 1 4 0 0 2 0 6 0 0 1 2 2 7 3 0 2 2 2 1 7 0 1];

B_spam = [5 3 1 4 3 3 3 3 6 2 2 4 0 2 0 1 2 4 4 4 2 1 2 2 2 4 1 2 4 1 4 0 0 3 1 3 3 2 2 2 5 5 1 3 3 2 1 5 2 1];

B_other = [8 6 4 7 6 5 2 2 4 5 3 5 5 5 6 8 5 6 3 4 5 3 5 8 2 5 3 5 6 3 6 4 4 5 2 6 4 5 5 6 2 4 5 3 3 4 4 4 4 4 3];

A : After March 16th 12 AM
A_total = [20 22 14 16 14 8 3 16 12 11 12 24 19 21 7 19 8 19 16 29 7 25 29 22 18 7 29 23 21 20 25 7 18 18 20 24 22 11 18 26 35 22 20 16 17 15 18 24 24 10];

A_school = [12 12 8 9 7 3 2 6 2 4 6 14 10 9 3 8 2 8 8 16 3 15 17 12 8 2 16 11 12 9 15 3 10 13 11 7 12 7 11 15 28 10 9 11 12 7 11 10 13 4];

A_handshake = [2 1 0 0 0 0 0 2 2 0 0 1 1 2 0 2 1 2 1 2 0 2 1 2 1 0 2 2 1 1 4 0 4 2 0 2 1 0 0 2 0 0 3 1 0 1 1 1 1 0];

A_spam = [3 2 3 3 3 3 0 3 5 5 3 5 5 9 2 5 2 5 5 4 2 5 4 6 3 2 5 4 3 5 2 1 1 1 4 9 5 3 3 5 3 10 5 2 4 3 4 10 8 5];

A_other = [3 7 3 4 4 2 1 5 3 2 3 4 3 1 2 4 3 4 2 7 2 3 7 2 6 3 6 6 5 5 4 3 3 2 5 6 4 1 4 4 4 2 3 1 1 4 2 3 2 1];

N : Number of day (only applies to After data)
N = [1 2 3 4 5 6 7 8 9 0 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50]

The data of the total number of emails is summarized as follows (values calculated by an “if loop” for each array to total up all the array values):

Total Emails	Before	After
School	556	473
Handshake	98	54
Spam	125	202

Other	228	171
Total	1003	901

The data of the mean number of emails of each category is summarized as follows (values calculated by the *nanmean()* function for each array):

Mean Number of Daily Emails	Before	After
School	11.118	9.4600
Handshake	1.9412	1.0800
Spam	2.5000	4.0400
Other	4.5294	3.4200
Total	19.980	18.020

The data of the standard deviation for each group of emails is summarized as follows (values calculated by the *std()* function for each array):

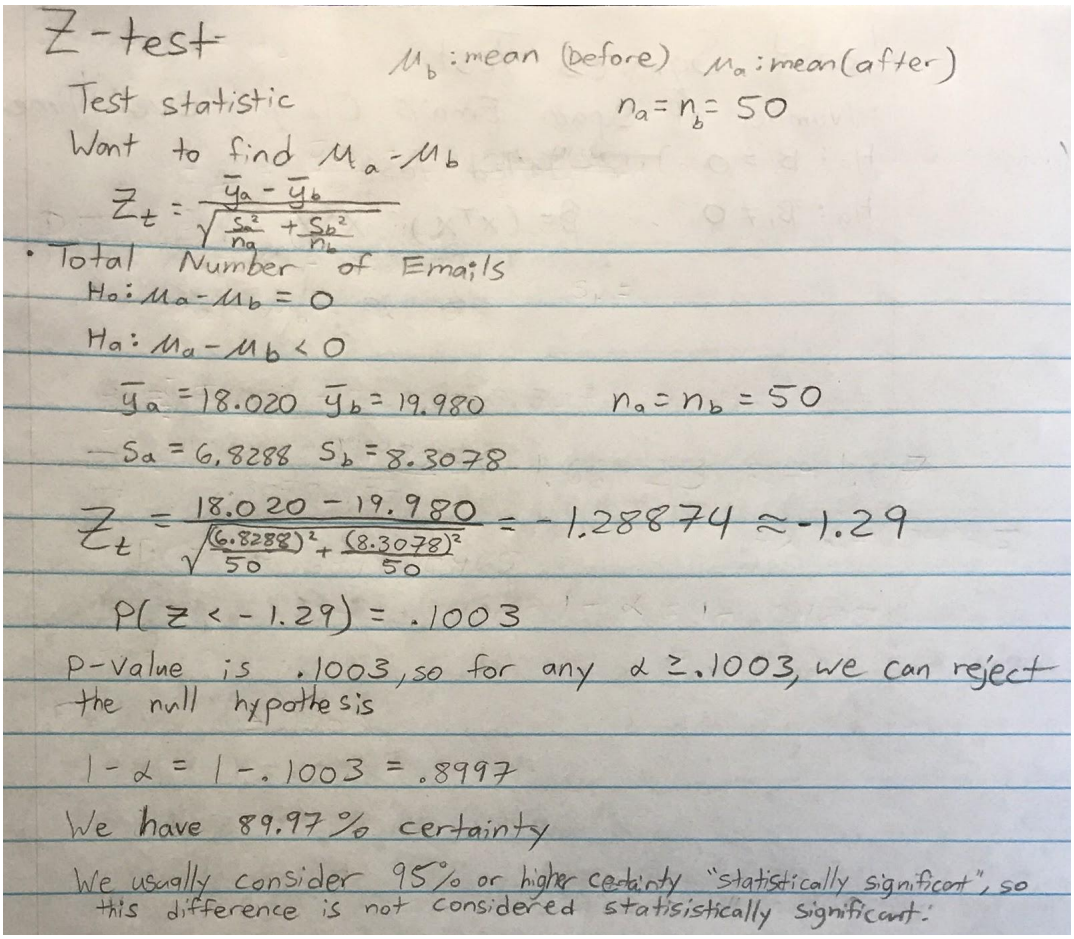
Standard Deviation of Daily Emails	Before	After
School	5.6094	4.9165
Handshake	2.1763	1.0467
Spam	1.4743	4.0400
Other	1.5278	3.4200
Total	8.3078	6.8288

For this project, I decided to utilize the Z-test to compare the mean of emails before and after the order by using the u_a - u_b test. For this, I wanted to see if the difference was statistically significant and showed that the number of emails I received After was less than or greater than the ones Before. I utilized the 1-tail test for the Z-test because from the mean number of emails chart above, we can clearly compare the Before and After means to check which one is larger. I also decided to utilize the Least Squares test to see if the total number of certain categories of emails has been changing since the order started. If b_1 (the slope) $\neq 0$, then the number of daily emails in those email categories have changed over the course of the Shelter-in-Place.

- The word “order” will be used to refer to the Shelter-in-Place order that started on March 16th at 12 AM.
- The word “sheltering” will be used to refer to the Sheltering in Place (the time after the start of the Shelter-in-Place order)

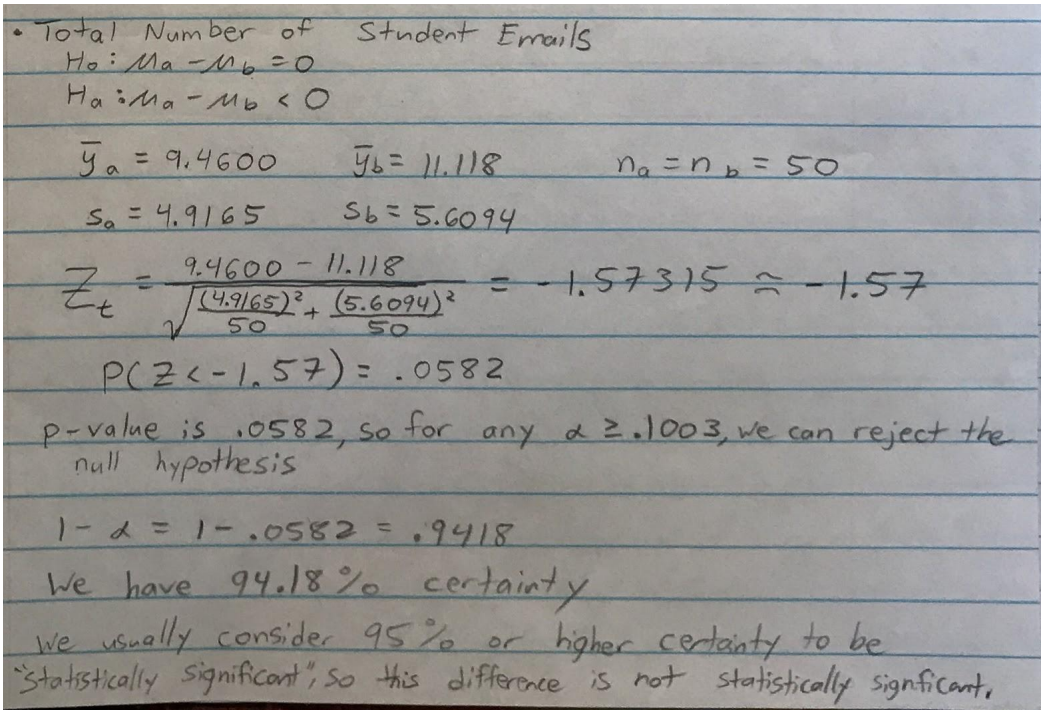
Hypothesis Testing

- Z-tests (u_a - u_b)
    - Is the mean of the total number of emails received after the order different than the mean of the total number of emails received before the order? (1-tailed interval)
        - H_o: The means are the same because the difference between them = 0.
        - H_a: The mean of the After emails is different from the mean of the Before emails because the difference is negative (since the experimental mean of After is less than that of Below).

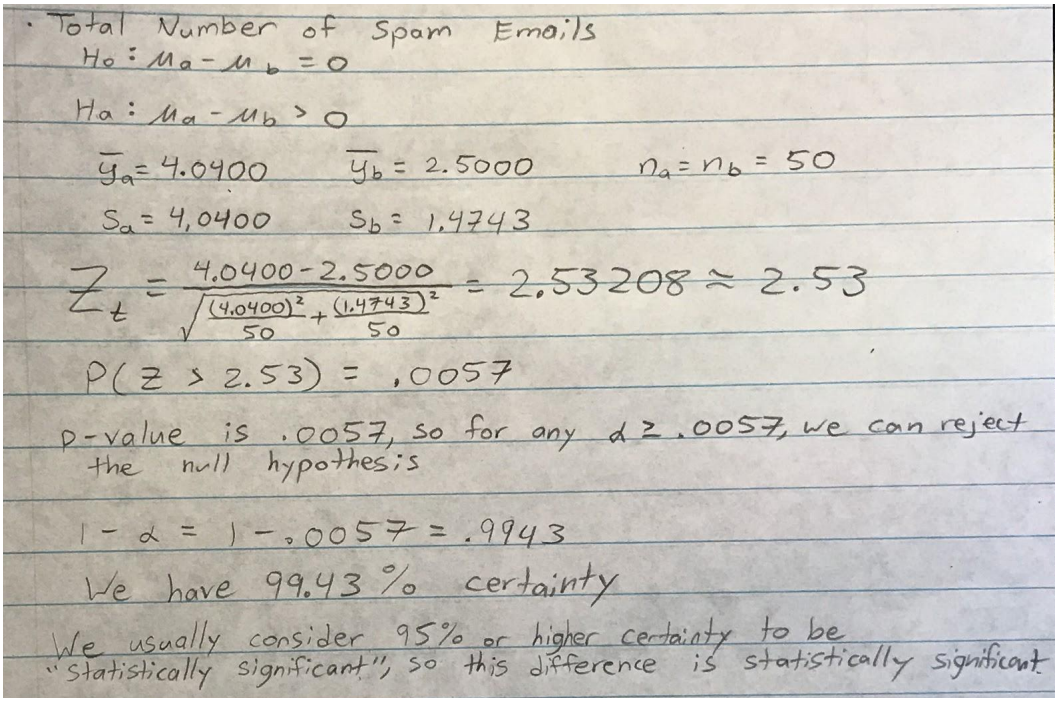


    - Is the mean of the total number of school emails received after the order different than the mean of the total number of school emails received before the order? (1-tailed interval)
        - H_o: The means are the same because the difference between them = 0.

- H_a: The mean of the After emails is different from the mean of the Before emails because the difference is negative (since the experimental mean of After is less than that of Below).

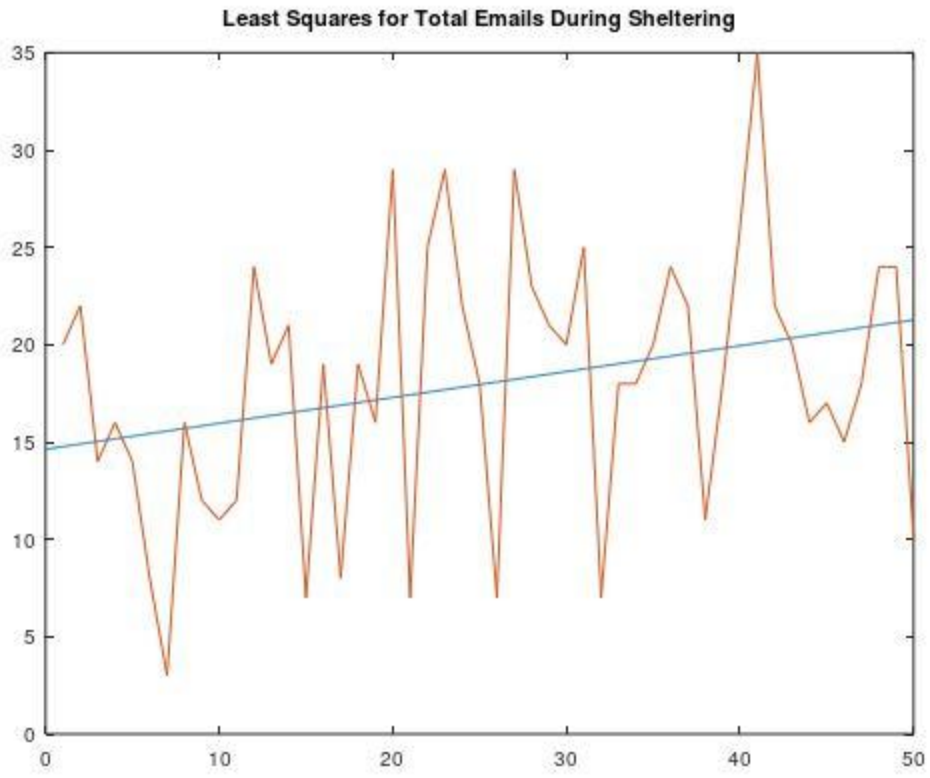


- Is the mean of the total number of spam emails received after the order different than the mean of the total number of spam school emails received before the order? (1-tailed interval)
  - H_o: The means are the same because the difference between them = 0.
  - H_a: The mean of the After emails is different from the mean of the Before emails because the difference is positive (since the experimental mean of After is more than that of Below).

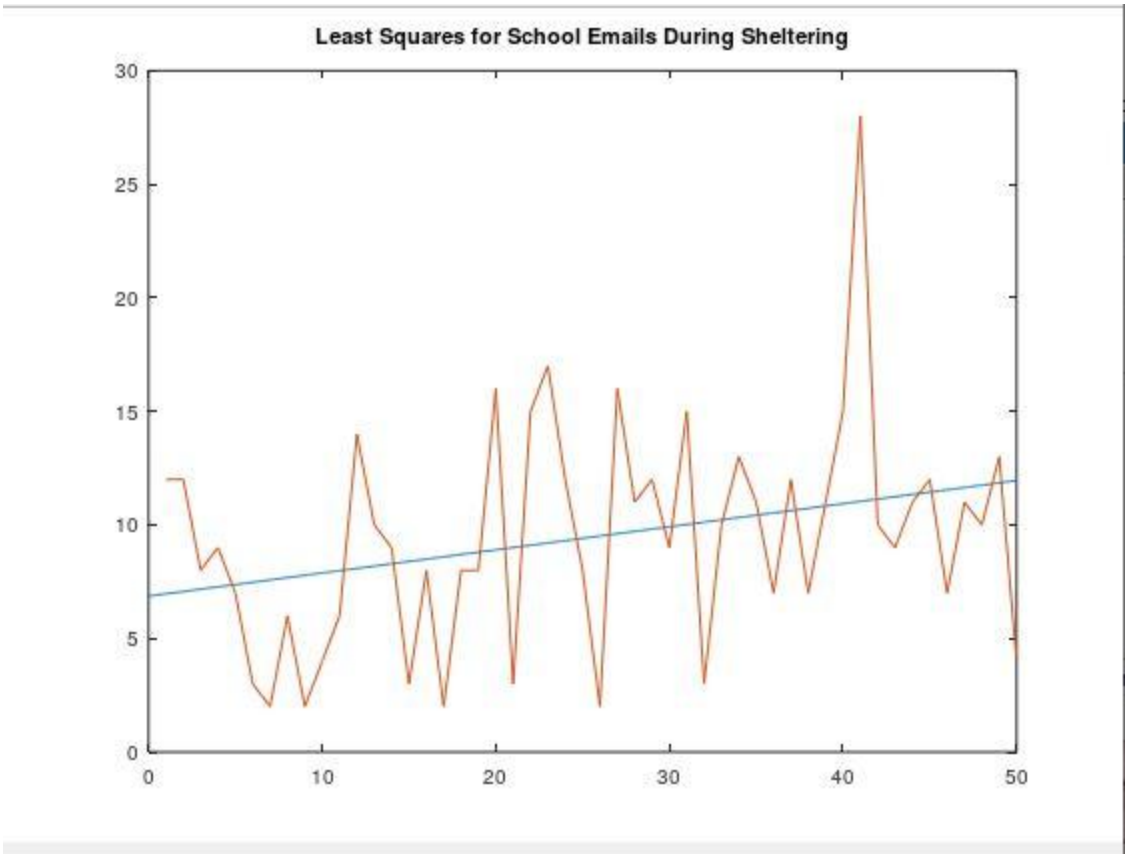
- 

Least Squares Regression

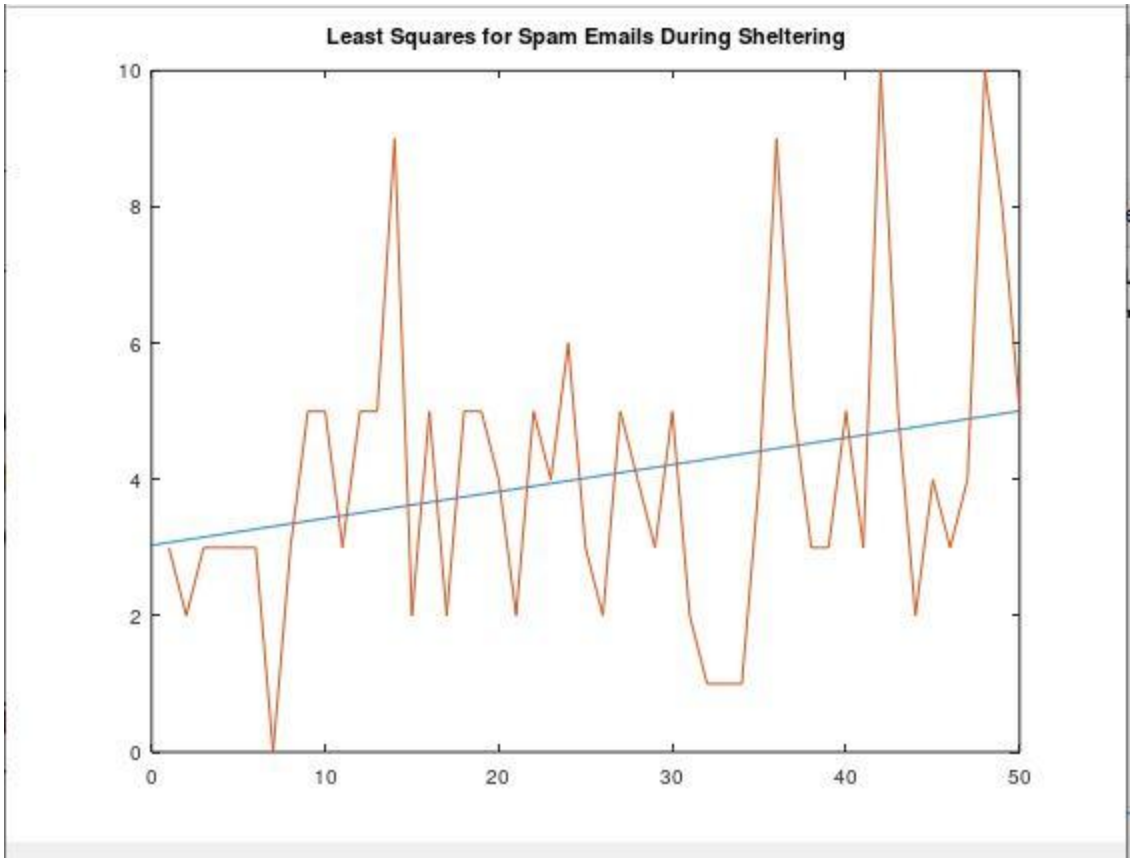
- Orange Graphs are from the number of daily emails in the selected category, and the blue line represents the Least Squares Regression Line.
- Has the number of daily total emails been increasing over time after Shelter-in-Place order started?
    - H_o: The number of daily total emails has stayed the same over the course of sheltering because the slope (b_1) = 0.
    - H_a: The number of daily total emails has changed over the course of sheltering because the slope (b_1) ≠ 0.
        - B_1 = 0.13278



- Has the number of daily school emails been increasing over time after Shelter-in-Place order started?
    - H_o: The number of daily school emails has stayed the same over the course of sheltering because the slope (b_1) = 0.
    - H_a: The number of daily school emails has changed over the course of sheltering because the slope (b_1) ≠ 0.
        - B_1 = 0.10176



- Has the number of daily spam emails been increasing over time after Shelter-in-Place order started?
    - H_o: The number of daily spam emails has stayed the same over the course of sheltering because the slope (b_1) = 0.
    - H_a: The number of daily spam emails has changed over the course of sheltering because the slope (b_1) ≠ 0.
        - B_1 = 0.039407

- 

Conclusion

Since the difference in the calculated mean of the spam email category was the only hypothesis test that produced a p-value that was “statistically significant”, this means that we can safely make the conclusion that the number of spam emails before and after the Shelter-in-Place are different. Since the mean number of spam emails in After were greater than Before, we can conclude that I did receive more spam emails during the sheltering than before the sheltering. Since the other two hypothesis tests (total emails and school emails) did not produce a p-value that was “statistically significant”, we can’t say with absolute certainty that the number of emails for their Before and After groupings were different. In calculating the Least Squares Regression of the total, school and spam emails over shelters and finding that $b_1 > 0$ for all cases, we can conclude that there was a trend of the number of emails in each category increasing over the course of my tested sheltering period.

The Entire Octave Code:

```
%MATH 123 (Probability and Statistics II) Final Project
%Jennifer Chun
%6/11/20
```

```
%Objective: Check numerical work and generate graphs for my final project (applying different Hypothesis Tests)
%->Hypothesis Test: Choosing a null hypothesis (a specific value) and using p-values to determine whether the data is with the null hypothesis; if not, it is in the alternative hypothesis (a broad range of values that aren't in the null hypothesis)
%Data: Sample sets of the number of emails before and after the official start of the Santa Clara County Shelter-in-Place (March 16th, 2020 at 12 AM)
%Emails sorted in five categories:
%-Total number of emails
%-School emails (student body, school administration, club emails)
%-Handshake (student-job networking website) emails
%-"Spam"
%-All other emails (daily prayer emails, etc.)

%Program Installation Requirements:
%-GNU Octave (GUI)
%-Octave Statistics Package, found at https://octave.sourceforge.io/statistics/

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%All sample days randomly selected by a number generator created by Octave on a separate Octave filesort (randperm(n,k))
%First random sample: Jan 4th - March 15th 2020
%Second random sample: March 16th - May 25th 2020
%Both random samples have 50 days within the time intervals


%B : Before March 16th 12 AM
B_total = [32 33 23 38 25 25 17 29 27 15 17 16 9 18 9 13 16 24 18 24 13 7 20 26 8 27 10 13 27 18 23 10 10 34 8 38 17 20 23 29 9 32 20 10 19 19 23 26 27 9 16]  ;
B_school = [12 17 15 21 15 14 11 24 15 7 11 3 4 10 3 6 8 12 9 14 6 3 13 15 4 13 6 6 12 13 9 6 6 24 5 23 10 13 15 19 5 16 7 6 11 10 15 20 11 3 11];
B_handshake = [7 7 3 6 1 3 1 0 2 1 1 4 0 1 0 0 1 2 2 2 0 0 1 1 0 5 0 0 5 1 4 0 0 2 0 6 0 0 1 2 2 7 3 0 2 2 2 1 7 0 1];
B_spam = [5 3 1 4 3 3 3 3 6 2 2 4 0 2 0 1 2 4 4 4 2 1 2 2 2 4 1 2 4 1 4 0 0 3 1 3 3 2 2 2 5 5 1 3 3 2 1 5 2 1];
B_other = [8 6 4 7 6 5 2 2 4 5 3 5 5 5 6 8 5 6 3 4 5 3 5 8 2 5 3 5 6 3 6 4 4 5 2 6 4 5 5 6 2 4 5 3 3 4 4 4 4 4 3];

%A : After March 16th 12 AM
A_total = [20 22 14 16 14 8 3 16 12 11 12 24 19 21 7 19 8 19 16 29 7 25 29 22 18 7 29 23 21 20 25 7 18 18 20 24 22 11 18 26 35 22 20 16 17 15 18 24 24 10];
```

```
A_school = [12 12 8 9 7 3 2 6 2 4 6 14 10 9 3 8 2 8 8 16 3 15 17 12 8 2 16 11 12 9 15 3 10 13
11 7 12 7 11 15 28 10 9 11 12 7 11 10 13 4];
A_handshake = [2 1 0 0 0 0 0 2 2 0 0 1 1 2 0 2 1 2 1 2 0 2 1 2 1 0 2 2 1 1 4 0 4 2 0 2 1 0 0 2 0 0
3 1 0 1 1 1 1 0];
A_spam = [3 2 3 3 3 3 0 3 5 5 3 5 5 9 2 5 2 5 5 4 2 5 4 6 3 2 5 4 3 5 2 1 1 1 4 9 5 3 3 5 3 10 5 2
4 3 4 10 8 5];
A_other = [3 7 3 4 4 2 1 5 3 2 3 4 3 1 2 4 3 4 2 7 2 3 7 2 6 3 6 6 5 5 4 3 3 2 5 6 4 1 4 4 4 2 3 1 1
4 2 3 2 1];

%N : Number of day (only applies to after data)
N = [1 2 3 4 5 6 7 8 9 0 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50]
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%Initial Calculations
%(Sums)
 B_total_sum = 0;
 B_school_sum = 0;
 B_handshake_sum = 0;
 B_spam_sum = 0;
 B_other_sum = 0;
 A_total_sum = 0;
 A_school_sum = 0;
 A_handshake_sum = 0;
 A_spam_sum = 0;
 A_other_sum = 0;

 for i = 1:50,
    B_total_sum += B_total(i);
    B_school_sum += B_school(i);
    B_handshake_sum += B_handshake(i);
    B_spam_sum += B_spam(i);
    B_other_sum += B_other(i);
    A_total_sum += A_total(i);
    A_school_sum += A_school(i);
    A_handshake_sum += A_handshake(i);
    A_spam_sum += A_spam(i);
    A_other_sum += A_other(i);
 endfor
 %Outputting all values
B_total_sum
B_school_sum
B_handshake_sum
```

```
B_spam_sum
B_other_sum
A_total_sum
A_school_sum
A_handshake_sum
A_spam_sum
A_other_sum

%(Mean Values)
B_total_mean = nanmean(B_total)
B_school_mean = nanmean(B_school)
B_handshake_mean = nanmean(B_handshake)
B_spam_mean = nanmean(B_spam)
B_other_mean = nanmean(B_other)
A_total_mean = nanmean(A_total)
A_school_mean = nanmean(A_school)
A_handshake_mean = nanmean(A_handshake)
A_spam_mean = nanmean(A_spam)
A_other_mean = nanmean(A_other)

%(Standard Deviation)
B_total_std = std(B_total)
B_school_std = std(B_school)
B_handshake_std = std(B_handshake)
B_spam_std = std(B_spam)
B_other_std = std(B_other)
A_total_std = std(A_total)
A_school_std = std(A_school)
A_handshake_std = std(A_handshake)
A_spam_mean = nanmean(A_spam)
A_other_mean

%Hypothesis Tests
%Least Squares Test

%Total Emails
O = ones(50,1) ;
P = N';
Q =  A_total';
fprintf("Testing Least Squares Test For Total Emails:  ");
X = [O P];
Y =  Q;
NCOMP = 1;
```

```
[XLOADINGS,YLOADINGS,XSCORES,YSCORES,COEFFICIENTS,FITTED] =
plsregress(X,Y,NCOMP);
b0 = COEFFICIENTS(1)
b1 = COEFFICIENTS(2)
COEFFICIENTS

figure(1)
x = 0:0.1:50;
plot(x, (b1*-(51/2) + A_total_mean + b1*x), A_total);
title("Least Squares for Total Emails During Sheltering");

%School Emails

O = ones(50,1) ;
P = N';
Q =  A_school';
fprintf("Testing for Least Squares Test For School Emails:  ");
X = [O P];
Y =  Q;
NCOMP = 1;
[XLOADINGS,YLOADINGS,XSCORES,YSCORES,COEFFICIENTS,FITTED] =
plsregress(X,Y,NCOMP);
b0 = COEFFICIENTS(1)
b1 = COEFFICIENTS(2)
COEFFICIENTS

figure(2)
x = 0:0.1:50;
plot(x, (b1*-(51/2) + A_school_mean + b1*x), A_school);
title("Least Squares for School Emails During Sheltering");
%plot(N,A_school,A_handshake,A_spam,A_other, A_total)
%plot(N,A_total)



%Spam Emails
O = ones(50,1) ;
P = N';
Q =  A_spam';
fprintf("Testing for Least Squares Test For Spam Emails:  ");
X = [O P];
Y =  Q;
NCOMP = 1;
```

```
[XLOADINGS,YLOADINGS,XSCORES,YSCORES,COEFFICIENTS,FITTED] =
plsregress(X,Y,NCOMP);
b0 = COEFFICIENTS(1)
b1 = COEFFICIENTS(2)
COEFFICIENTS

figure(3)
x = 0:0.1:50;
plot(x, (b1*-(51/2) + A_spam_mean + b1*x), A_spam);
title("Least Squares for Spam Emails During Sheltering");
%plot(N,A_spam)
```