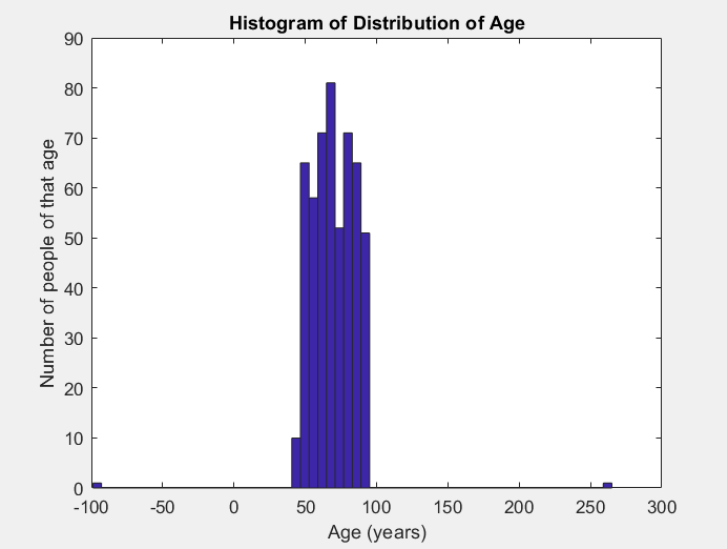
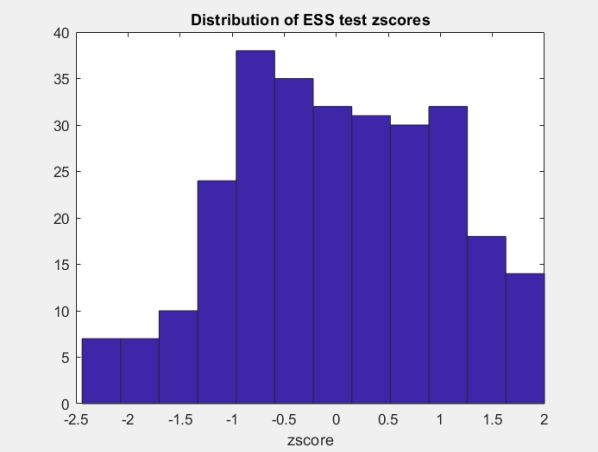
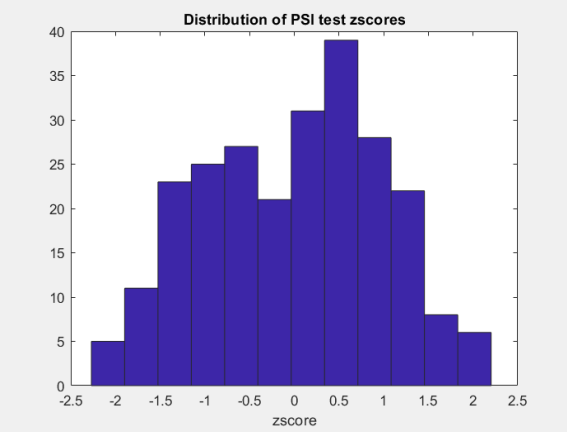
Katie Foster

Data Analysis

1. Summarize the age data with basic summary statistics and a plot. What does the difference between the mean and median suggest about the distribution?
   a. 

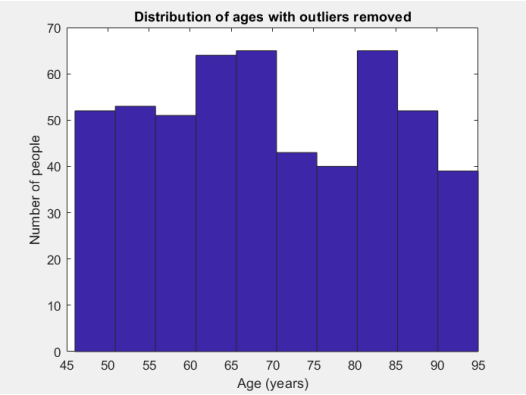
   b. mean = 69.8, median = 68, max = 265, min = -99, std = 17.9636, variance = 322.7
   c. The mean age is 69.8 and the median age is 68. This data suggests that there are slightly more people on the younger side of the mean than on the older side.
2. Since our data has two different measures of postural stability, we'll need a way to appropriately compare these values. In a new variable (let's call it Common Stability Score or CSS), represent the stability scores of each participant on the all on the same scale. This scale can be ES, PSI, or a shared scale such as z-score.
   a. I used zscores, look at the attached matlab code to see my process
3. Plot and comment on the CSS distribution.

a. 

b. 

c. The PSI test scores seem to follow more of a standard distribution and have fewer scores at the extremes. The ESS scores have more scores at the high and low ends of the distribution and peaks around a z-score of -1.2.

4. Identify any extreme outliers or errors in the data. Decide how to deal with these and justify your decision. Use this revised data for the rest of your analysis.
    a. There are two extreme outliers for age, which are -99 and 265. These entire data points should be thrown away. It is a good decision to get rid of these data points because there are only two points and they are both very far away from the age distribution of the other points. Also there are no humans currently alive that are -99 or 265 years old, so they are clearly faulty points.

b. 

5. Determine the counts of the true positive, true negative, false positive, and false negatives for your screening test.
    a. true_pos = 252
    b. true_neg = 239
    c. false_pos = 24
    d. false_neg = 11
6. If you randomly selected someone from your study group, what is the probability that your tests will flag them as having Parkinson's?
    a. 52.47%
7. If you randomly select someone from your study group who you know does not have Parkinson's, what is the probability that your test will incorrectly diagnose them with the disease?
    a. 9.16%
8. If you randomly select someone from your study group, what the probability that your test will diagnose them correctly?
    a. 93.35%
9. Pose at least 3 additional questions that you could try to answer with these data. For example, how do the age distributions compare across diagnoses? A wide variety of questions are welcome, this is to help you think about what could be answered and what could not, given a particular data set.
    a. Are people more likely to have a false negative/false positive on one test over the other?
    b. Is one of the tests used more on different age populations? (And different populations in general, but we can't figure that out without more data)
    c. Are people of certain ages more likely to get false results on the screen?
10. Attempt to answer 1 of your 3 additional questions.
    a. Question a: The false negative/positive distribution is as follows:
        i. ES score data
            1. true_pos = 129
            2. true_neg = 135
            3. false_pos = 11

4. false_neg = 3
5. Normalized (divided by the total number of this kind of score):
   0.4640 0.4856 0.0396 0.0108

ii. PSI score
1. true_pos = 122
2. true_neg = 103
3. false_pos = 13
4. false_neg = 8
5. Normalized (divided by the total number of this kind of score):
   0.4959 0.4187 0.0528 0.0325

iii. Conclusion: The PSI score produces more false negatives and false positives, which implies that it may be the slightly worse test

Week 1

```
data = readtable('Parkinsons.txt');
```

```
mean(data.age)
```

```
ans = 69.8498
```

```
median(data.age)
```

```
ans = 68
```

```
max(data.age)
```

```
ans = 265
```

```
min(data.age)
```

```
ans = -99
```

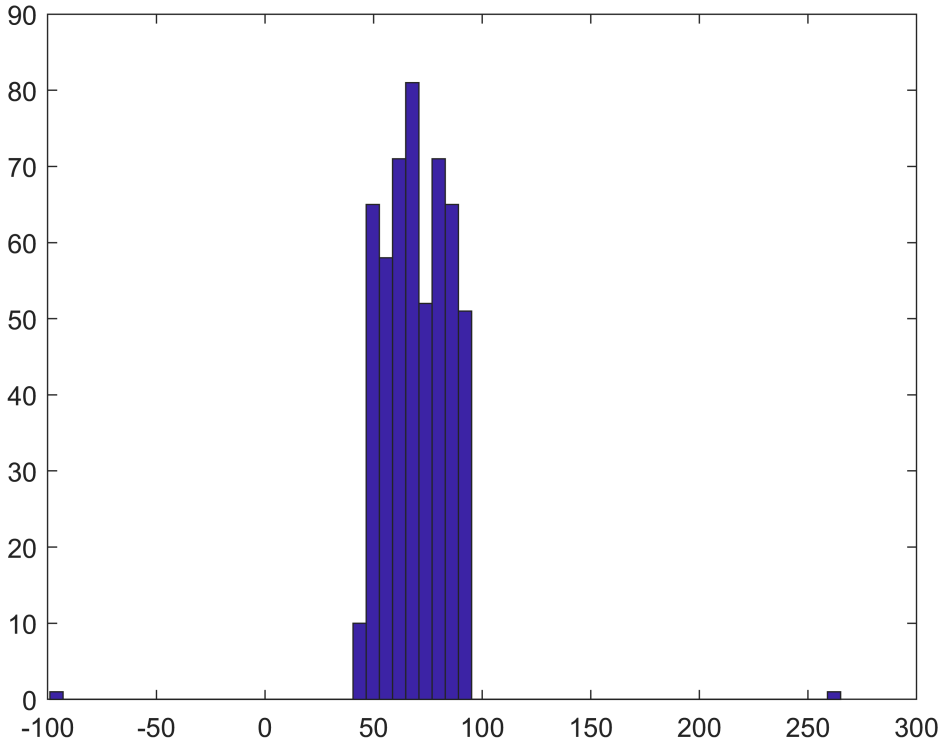
```
var(data.age)
```

```
ans = 322.6917
```

```
std(data.age)
```

```
ans = 17.9636
```

```
hist(data.age, 60)
```



This is where I removed the outliers in the data using:

```
% data([find(data.age==min(age))],:) = [];
```

I used code from someone else, but modified it a bit

Jason Joseph Rebello (2020). True Positives, False Positives, True Negatives, False Negatives from 2 Matrices (https://www.mathworks.com/matlabcentral/fileexchange/47364-true-positives-false-positives-true-negatives-false-negatives-from-2-matrices), MATLAB Central File Exchange. Retrieved January 28, 2020.

```
true_vals = data.actual + data.screen;
true_pos = length(find(true_vals == 2))
```

```
true_pos = 252
```

```
true_neg = length(find(true_vals == 0))
```

```
true_neg = 239
```

```
false_vals = data.actual - data.screen;
false_pos = length(find(false_vals == -1))
```

```
false_pos = 24
```

```
false_neg = length(find(false_vals == 1))
```

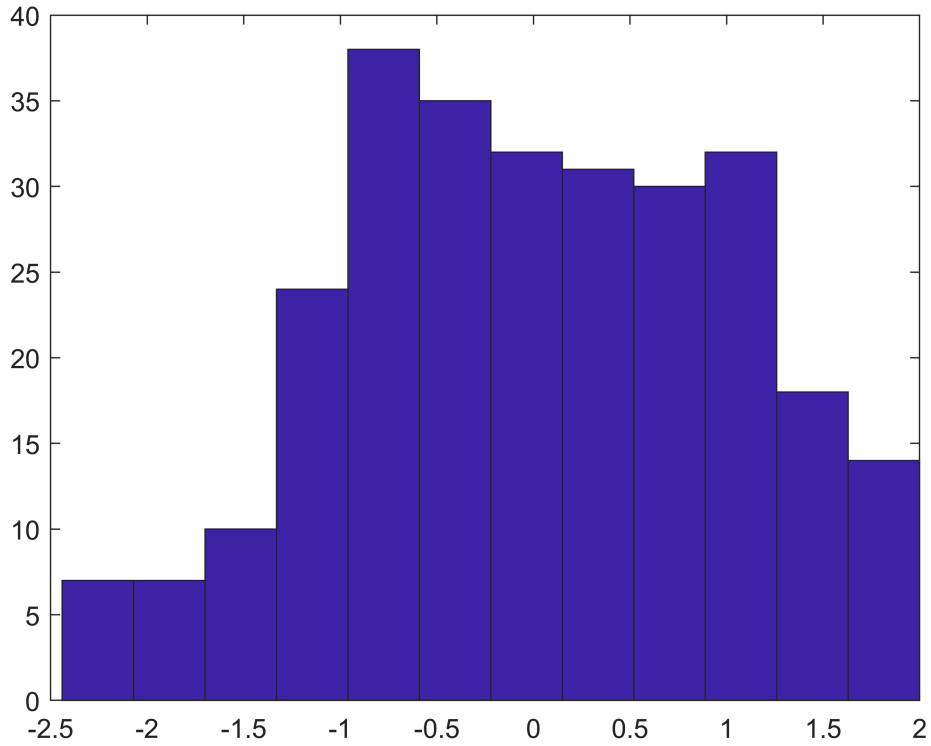
```
false_neg = 11
```

```
length(find(data.screen))/length(data.screen)
```

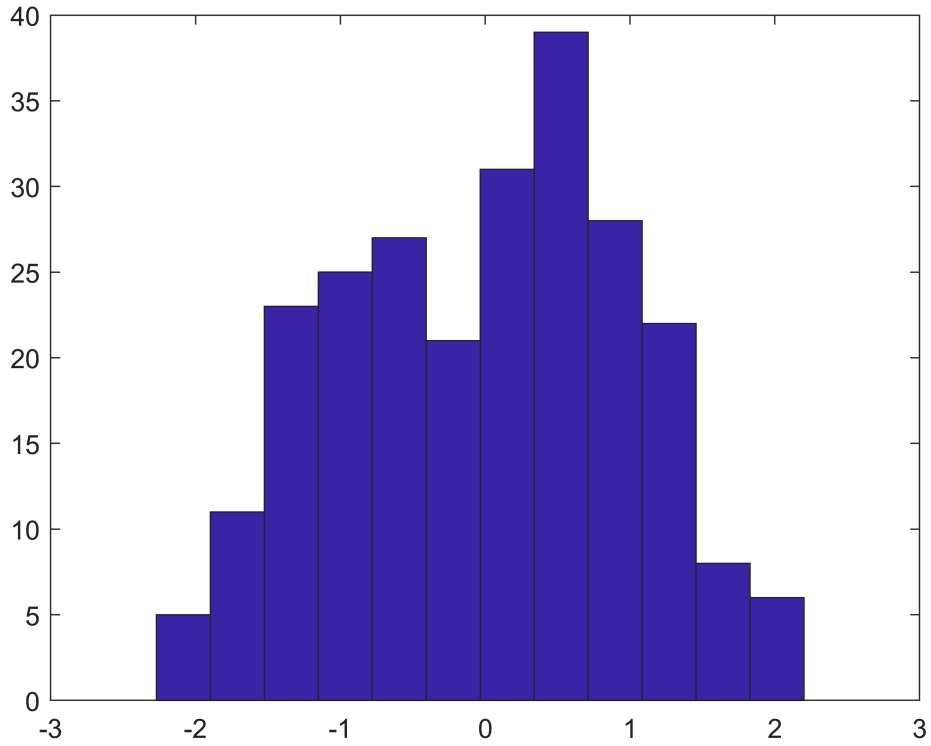
```
ans = 0.5247
```

```
ESdata = data(strcmp(data.PStype, 'ES'), :);
PSIdata = data(strcmp(data.PStype, 'PSI'), :);
zscoreESS = zscore(ESdata.PSS);
zscorePSI = zscore(PSIdata.PSS);
```

```
hist(zscoreESS, 12)
```




```
hist(zscorePSI, 12)
```



```
true_vals = ESdata.actual + ESdata.screen;
true_pos = length(find(true_vals == 2))
```

```
true_pos = 129
```

```
true_neg = length(find(true_vals == 0))
```

```
true_neg = 135
```

```
false_vals = ESdata.actual - ESdata.screen;
false_pos = length(find(false_vals == -1))
```

```
false_pos = 11
```

```
false_neg = length(find(false_vals == 1))
```

```
false_neg = 3
```