Statistics for Data Science - CSIT528_01

Project 1

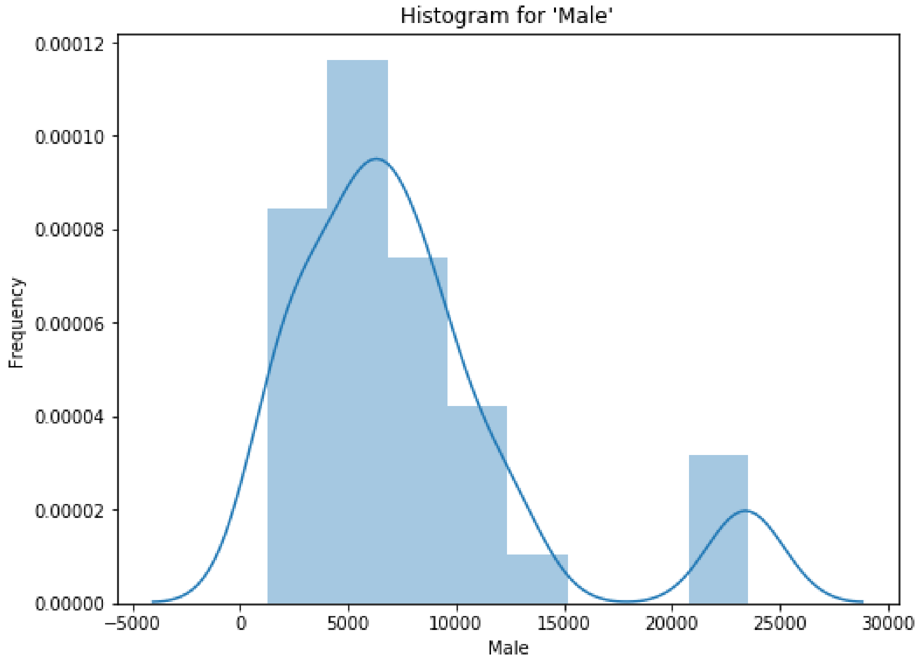
Name: Nandu Voore **E-Mail: vooren1@montclair.edu**

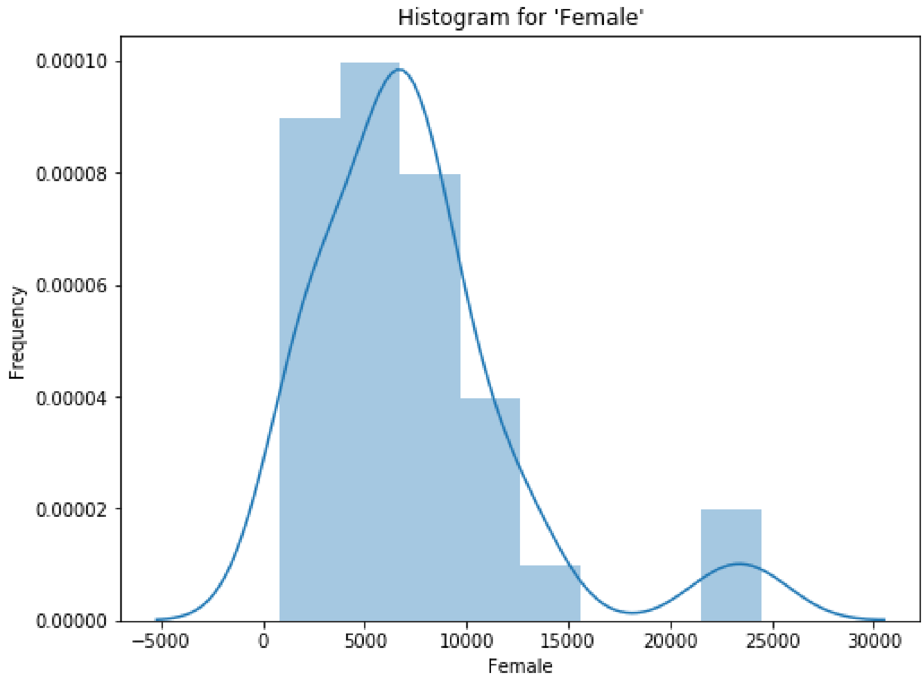
Initial Hypothesis:

Initial assumption is made that females tend to have a greater number of photos stored on their mobile phones compared to males.

Generating plots:

Histogram plots:-





Median and IQR for Male Data:
Median: 6435.0
IQR: 4552.25

Median and IQR for Female Data:
Median: 6489.5
IQR: 5192.0

Histogram plot analysis:

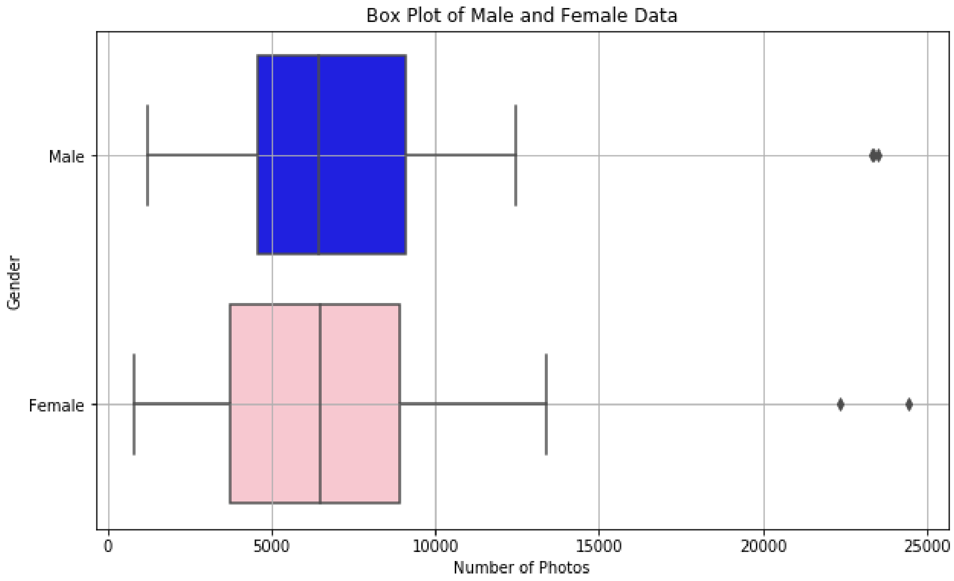
Male Data: The distribution appears to be right-skewed. The histogram shows a higher concentration of bars on the left side, with a longer tail extending towards the right side. This suggests that most males have a lower image count, with a few having a significantly higher number of images.

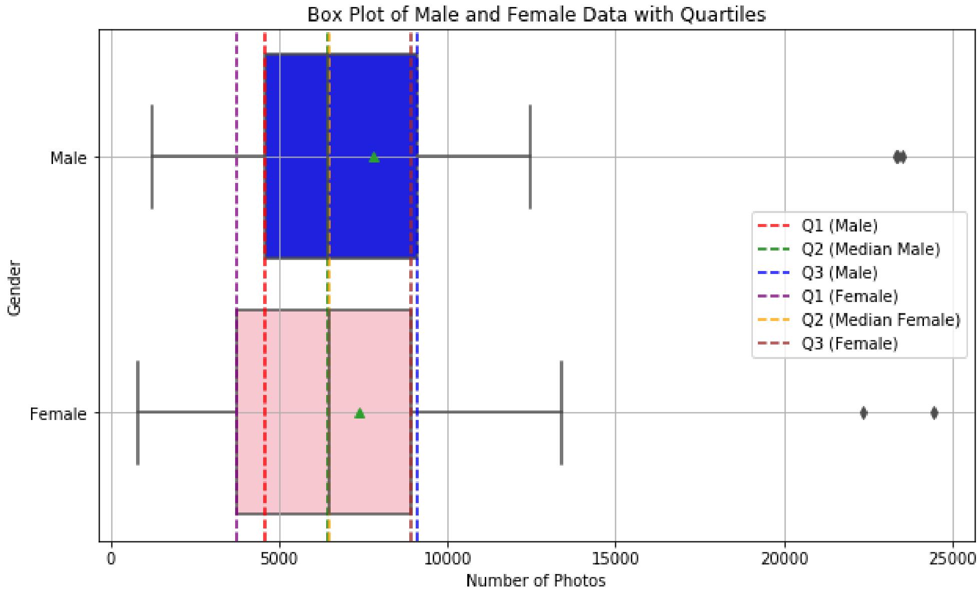
Female Data: Female data distribution is almost closer to symmetric with a slight left skew. From the histogram plot, it has a peak in the centre which means a concentration of data points around the average value. However the tail on the left side seems slightly longer than the tail on the right side. This indicates there may be few more females with lower image counts than those with higher counts.

Modality: Both the distributions are Unimodal. We can observe a single peak in the histogram, indicating one main cluster of data points for male and female.

Outliers: Outliers are data points that fall far away from the main body of the distribution. In both the male and female histogram plots we can observe data points away from the main cluster. On the right most side there is a single bar significantly higher than most other bars, these are the outliers.

Box plot analysis:-





- ➔ We can identify outliers in a box plot by using Inter quartile range (IQR). If the data extends 1.5 time the IQR from the upper and lower quartiles (the box ends), any data points beyond these are considered outliers.
- ➔ We can clearly observe that there are 2 outliers in Male data and two outliers in the female data.
- ➔ The second plot contains the markings of the quartiles in colours.

Observations and decisions for measures to use :

Symmetry and Skewness are very important when choosing appropriate statistical measures for analysing datasets. In symmetric distributions, the left and right sides mirror each other, here the measures of central tendency like the mean and median can be used interchangeably and also the standard deviation and the IQR provide accurate descriptions of spread. However our data sets are skewed distributions which have long tails on one side, in this case we will proceed using **Median** and **IQR** as measures of centre and standard deviation. Median is not influenced by the outliers (extreme values), whereas the mean is influenced. Similarly the IQR is very robust over standard deviation while considering the measure of spread as IQR provides robustness to data even when outliers are present.

Observations from Male data set:

- ➔ The male data set contains outliers which caused skewness as observed in histogram plot, this could have impact when mean and standard deviation are used as a measure of spread, hence proceeding with median and IQR.
- ➔ The median is less influenced by outliers and provides us a better estimate of the typical value in the data set. Similarly the IQR is robust to outliers and gives us a better indication of the variability in the data.

Observations from Female data set :

- ➔ The female data set also contains outliers which caused skewness as observed in the histogram plot.
- ➔ To minimize the impact of outliers and ensure consistency in the analysis its preferred to use Median and Standard Deviation for female data set also.
- ➔ Using IQR and Median allows to get more accurate comparison considering the skewness in the data.

In Summary, with the presence of outliers in both the data sets its more appropriate to use Median and IQR as the measure of centre and spread. This decision ensures robustness to the outliers and provides us more accurate representation of the data sets and its characteristics.

Discussion of Samples Using Chosen Centre and Spread Measurement

Measure of Centre:

We have chosen median as the measure of centre because the data set exhibits skewness as it has outliers. Comparing the medians, male data has a median of 6435 and female data has a median of 6489.5, therefore the female group has a slightly higher measure of centre.

Measure of Spread:

We chose the interquartile range (IQR) as the measure of spread because it is robust to outliers, which are present in both data sets. Comparing the IQRs, the male data has an IQR of 4552.25, while the female data has an IQR of 5192. Therefore, the female group has a greater measure of spread.

Comparison:

The higher value of median value observed in the female group implies that, on average, females tend to have a greater number of photos stored on their mobile phones than males. Furthermore, the larger interquartile range (IQR) for the female group indicates a wider variability in the number of photos stored among females compared to males. These observations suggest disparities between the two underlying populations concerning the quantity of photos stored on their mobile phones.

Hypothesis Test:

This is a two-sample t-test because we are comparing the means of two independent groups male and female. The null hypothesis (H0) states that there is no difference between the population means, while the alternative hypothesis (H1) states that there is a difference between the population means.

Null Hypothesis (H0):

The null hypothesis is the default assumption that there is no significant difference between the population means of the two groups. In this context, the null hypothesis (H0) states that there is no difference between the average number of photos stored by males and females.

Alternative Hypothesis (H1):

The alternative hypothesis is the opposite of the null hypothesis. It suggests that there is a significant difference between the population means of the two groups. In this context, the alternative hypothesis (H1) states that there is a difference between the average number of photos stored by males and females.

Test Statistic (T-Statistic):

The test statistic (t) is a numerical value calculated from the sample data. It measures how much the sample means differ relative to the variation observed within the samples.

A larger absolute value of the t-statistic suggests a greater difference between the sample means.

The formula for calculating the above parameters are

$$t = \frac{\text{difference in sample means-what null hypothesis says the difference is)}}{SE_{est}}$$

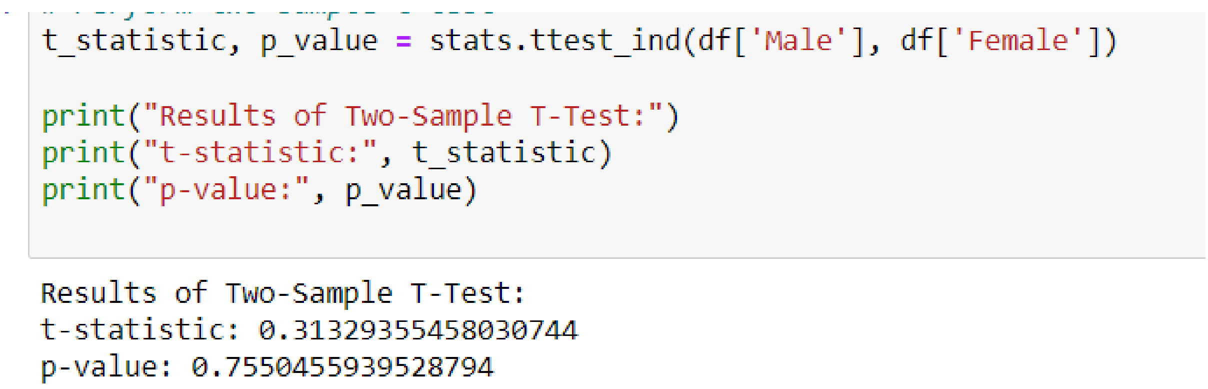
On performing the t-test we get the results as

t-statistic : 0.313293554580307444
p-value   : 0.7550455939528794

Calculating 95% confidence interval for difference between median gives us

95% Confidence Interval for Difference in Medians: (-2375.547487605669, 2266.547487605669)

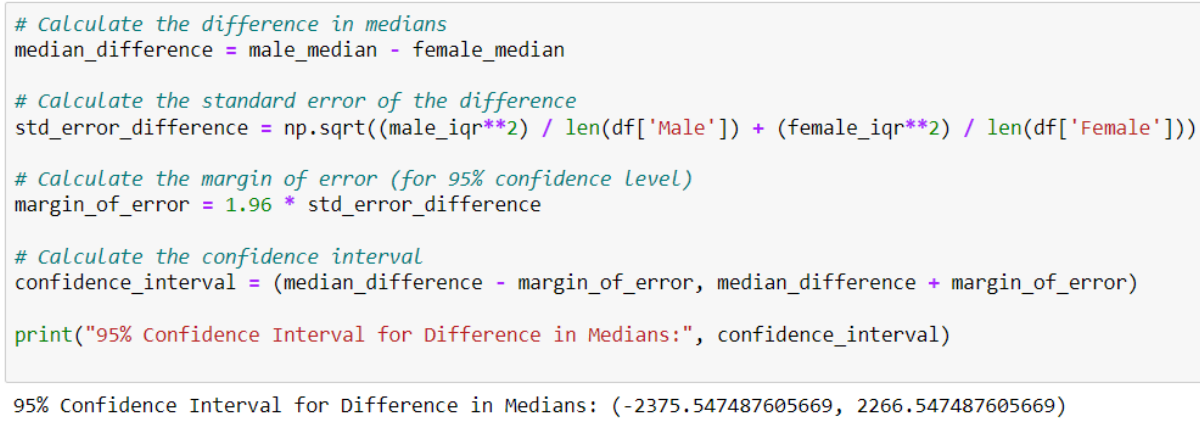
Code snippets:



```
t_statistic, p_value = stats.ttest_ind(df['Male'], df['Female'])

print("Results of Two-Sample T-Test:")
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

```
Results of Two-Sample T-Test:
t-statistic: 0.31329355458030744
p-value: 0.7550455939528794
```



```
# Calculate the difference in medians
median_difference = male_median - female_median

# Calculate the standard error of the difference
std_error_difference = np.sqrt((male_iqr**2) / len(df['Male']) + (female_iqr**2) / len(df['Female']))

# Calculate the margin of error (for 95% confidence level)
margin_of_error = 1.96 * std_error_difference

# Calculate the confidence interval
confidence_interval = (median_difference - margin_of_error, median_difference + margin_of_error)

print("95% Confidence Interval for Difference in Medians:", confidence_interval)
```

```
95% Confidence Interval for Difference in Medians: (-2375.547487605669, 2266.547487605669)
```

Discussion of Hypothesis Test Results:

In a two-tailed test, we examine both ends of a distribution without assuming a specific direction. My initial hypothesis is that females have more photos stored on their mobile phones than males. In this context, the null hypothesis states that there's no difference in the average number of photos stored between males and females. Conversely, the alternative hypothesis proposes that there is indeed a difference between the population means, but it doesn't specify the direction of this difference.

Following the two-sample t-test, we obtain a t-statistic and a p-value. If the p-value falls below our chosen significance level α = 0.05, we reject the null hypothesis. Conversely, if the p-value exceeds α, we fail to reject the null hypothesis.

A t-statistic of 0.3132 indicates a very small difference between the means of the male and female data. The high p-value (0.7550) is greater than the commonly used significance level of 0.05.

Interpretation:

With a high p-value, we **fail to reject the null hypothesis**. This suggests that there is not enough evidence to conclude a statistically significant difference between the average number of photos for males and females based on this data and the chosen significance level.

Confidence Interval:

95% Confidence Interval for Difference in Medians: (-2375.55, 2266.55)

The confidence interval includes zero. This further supports the idea that the true difference in medians between the two populations might be close to zero, aligning with the t-test results.

Overall Comparison:

Considering both the t-test and confidence interval, we can't claim a statistically significant difference between the centre values (means or medians) of the number of photos for males and females in this data set. It's possible that there might be a very small difference, but the current sample size or variability in the data might not be sufficient to detect it with high confidence.

Analysing Confidence Interval

The confidence interval provides a range within which we are reasonably certain that the true difference in the median number of photos stored between males and females lies. In our case, the 95% confidence interval for the difference in medians is calculated as (-2375.55, 2266.55). This means that we can be 95% confident that the actual disparity in medians falls somewhere within this interval.

This confidence interval suggests that there might not be a significant difference in the median number of photos stored between males and females. However, due to the wide range of the confidence interval, there is uncertainty regarding the true difference in medians.

While the values within the confidence interval are theoretically plausible, the wide range may stem from the variability in our sample data and the relatively small sample size. Given this uncertainty, it's important to recognize that the actual difference in medians may lie anywhere within this range.

My understanding:

The confidence interval (-2375.55, 2266.55) suggests a wide range of possible differences in the median number of photos stored between males and females. Although theoretically plausible, the interval's width raises concerns about precision, indicating uncertainty due to sample variability or small sample size limitations.

Comparison of Hypothesis Test and CI:

The findings from both the hypothesis test and confidence interval lead to a coherent conclusion regarding the distinction between population medians. They suggest a lack of substantial difference, implying that the population means are probably similar. Consequently, both analyses corroborate the notion that there is no significant gap in the median number of stored photos between males and females.

Conclusion:-

My initial assumption of considering that females have more photos than makes is incorrect. Both the hypothesis test and confidence interval suggested that there is not much significant difference between the medians of the number of photos stored by males and females.