CS306: Data Analysis and Visualization
Lab 9: Report

Rajdeep Pinge 201401103
Aditya Joglekar 201401086

Objective:

To draw statistical inferences on given data, and compare parametric and nonparametric methods.

Experiment 1: Load data_lab9.mat. The vectors 'new' and 'traditional' represent the degree of reading power (DRP) scores for two groups of students who followed a new reading method and the traditional method (higher DRP implies better reading power). Your goal is to answer if the new reading method improves reading ability of elementary school students (as compared to traditional method), as measured by DRP scores. Your analysis may also include a comparison of parametric and nonparametric tests.

Parametric method- 2-sample unpaired pooled t-test

Since the test statistic is mean therefore, by CLT we can get that the sampling distribution of mean gives normal distribution. Therefore first assumption of t-test is satisfied.

Standard Deviation of Traditional method = 9.4651

Standard Deviation of New method = 11.0074

Therefore homogeneity of variance is satisfied as the difference in variance is small.

Independence of data samples is assumed which satisfies third assumption of t-test.

Hence all the 3 assumptions of t-test are satisfied. Therefore we will apply the t-test to check if there is any statistically significant difference between the two populations.

Hypothesis:

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$

Alternate hypothesis H_a: $\mu_1 - \mu_2 \neq 0$

Where μ_1: mean of the population traditional data set

μ_2: mean of the population of new data set

t-score = -2.4076

Degrees of Freedom (df) = 23 + 21 - 2 = 42

Using the t-score calculator we get

P-value = 0.0103

If we take t-score to be positive, then we get the p-value on the opposite tail.

Therefore p-value = 1 - 0.0103 = 0.9897

If we take significance level alpha to be = 0.05 and let it be a two tailed test

Then then we take 0.025 to be the critical condition

Since here 0.9897 > 0.975 which is the significance level, the t-test rejects the null hypothesis and concludes that there is significant difference between the two populations.

Here we assume that 0.05 significance level is sufficient. We also assume that the samples are representative and hence the results obtained based on these samples are actually applicable to the entire populations.

Non Parametric Method:

Non parametric tests are the other type of statistical tests. They can be used for any test statistic and not just the mean and do not make any assumption about the underlying data distribution or the distribution of test statistic. So we can use them provided their assumptions of i.i.d data samples and homogeneity of variance is satisfied. We have used two **non parametric method**s to test if there is any **statistically significant** difference between the two populations. They are:

1) Bootstrap resampling
2) Permutation tests

1) Bootstrap resampling:

We have been provided with just two samples from s**upposedly 'different' populations.** Our job is to find out if the new reading method has any positive treatment effect over the traditional method.

Here, we do not make any assumptions regarding the distribution of the test statistic. Instead we will use **repeated resampling with replacement** from the two samples and calculate a test statistic which is useful measure a possible difference between the two groups.

Null Hypothesis: is that the new method does not have any **statistically significant effect** compared to the **traditional method** and hence the two groups are in fact from the same population distribution. Thus, we are justified in combining the two samples and resampling from the "pooled sample"

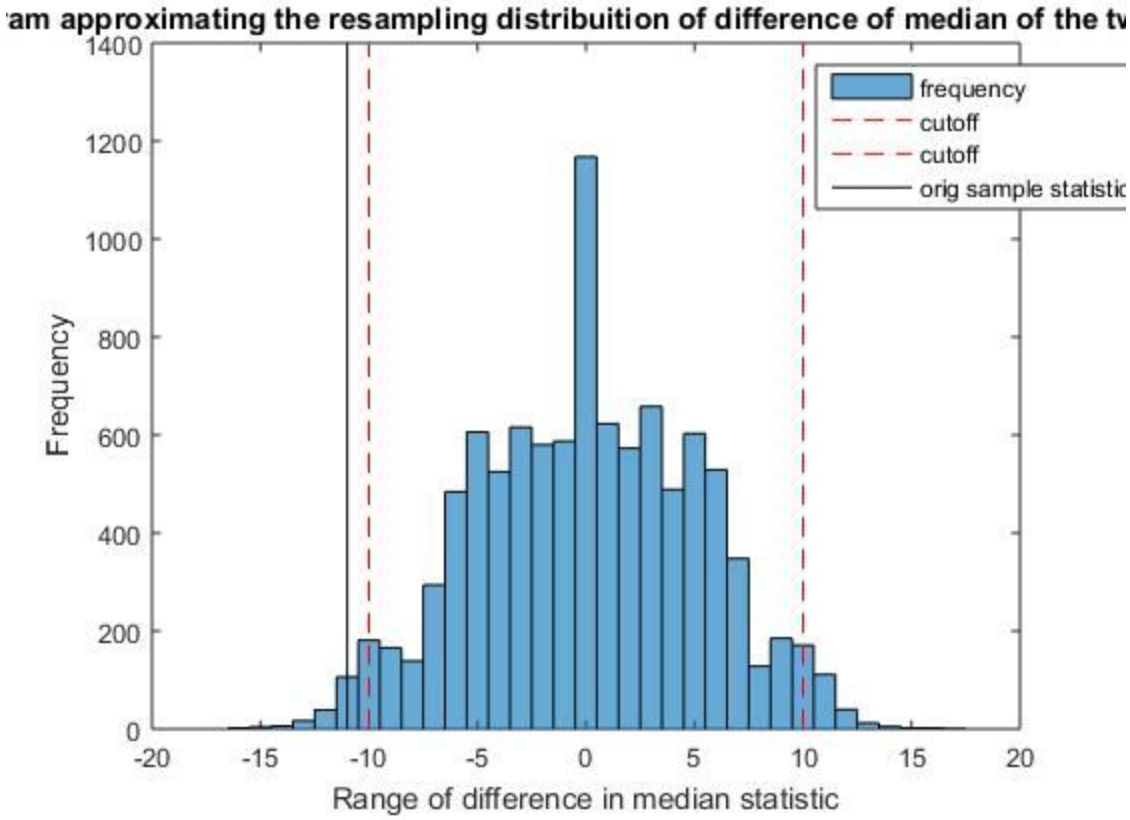
Test statistic used: Median: For every iteration, we will sample with replacement from the "pooled sample", distribute the data items into X and Y and calculate the difference of medians of medians of the two groups.

Often there are **a few,exceptional students in the class** who might **skew** the mean. But the median will not be affected. Thus, here it is a good idea to **use median to compare** the two populations.

We will thus get the **bootstrap resampling distribution** of the "difference of median" test assuming that **null hypothesis** is true.

Then, we locate the difference of **mean of the actual samples on the distribution** and check the the likelihood of getting a value that is at least as extreme(that is greater than or equal to) to our current answer, under the assumption that the **Null hypothesi**s is true. If this likelihood (which

is similar to p-value) is **small then our result** is not likely to be because of chance but because of **statistically significant difference.**



We obtained the following graph:

We have marked the Bootstrap percentile CI at 2.5th and 97.5th percentile. Here the difference of median in the actual sample lies outside the CI. This indicates that the likelihood of this **difference being because of chance is very less**. This corresponds to the p-value being very **low** in a parametric test. We thus decide to **reject the null hypothesis.** There is a significant effect of the new method over the **traditional method**. This is because the difference **median(traditional) - median(new)** is negative indicating the **superiority of the new method**.

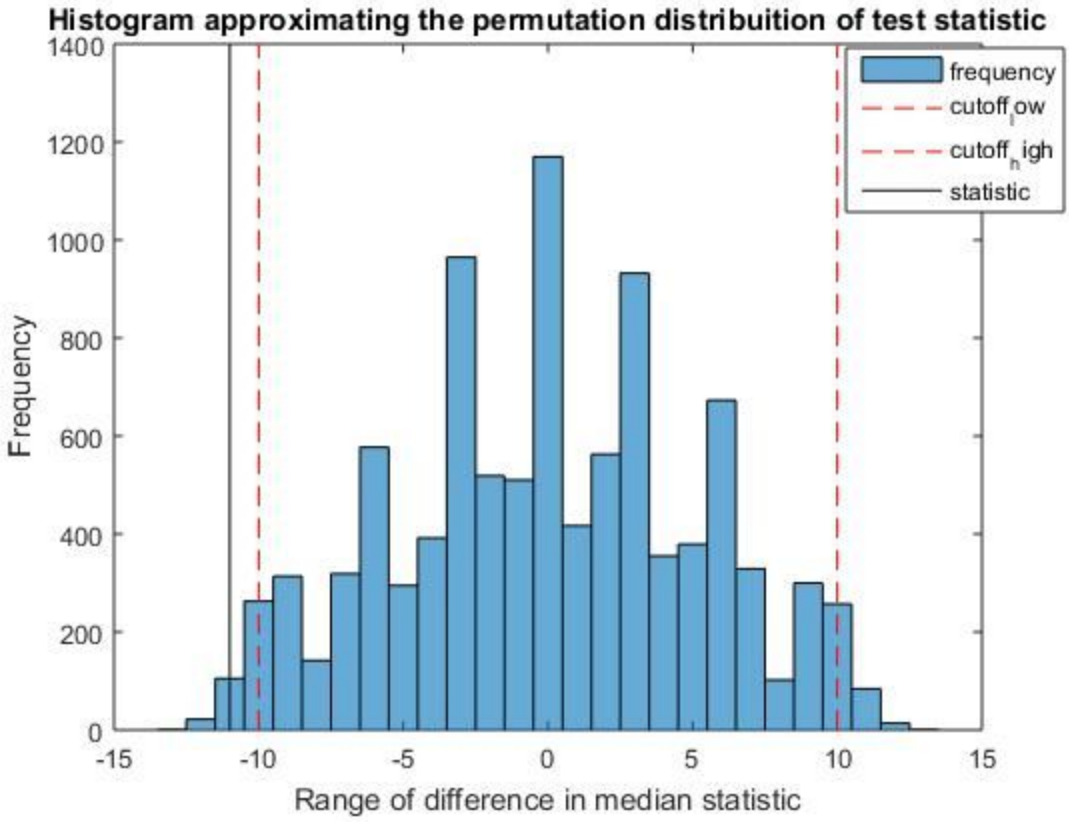
2) Using permutation test

This is similar to the bootstrap in that it assumes the **same null hypothesis.** In this test when we assume the **null hypothesis**, we basically say that there is no relation b**etween the group labels and the values** and this is in fact why we combine **the two samples and shuffle the labels.**

We recompute the **test statistic after this shuffling** and obtain a **resampling distribution** from the permuted samples.

The rest is similar to the **process of the bootstrap**. We will **reject the null hypothesis if the actual test statistic does not lie within the CI** because it is **not very likely to happen due to chance.**

The graph obtained is as follows:



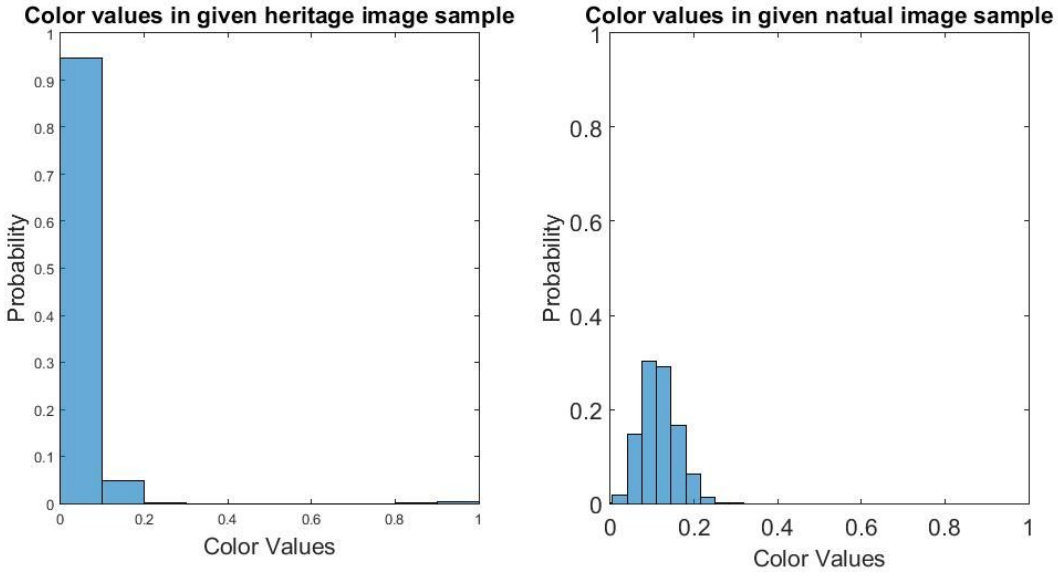
We will thus reject the Null Hypothesis.

This reinforces the earlier conclusion that the new method is indeed a significant improvement over the traditional method.

Experiment 2: Development of content specific methods is important but challenging in several video processing applications. One such example is that of heritage and natural content. Heritage content represents specific type of data which is different from natural scene data. We can quantify this in terms of color. Heritage images in general have lower color because the structures represented in them are old and tend to be less colorful.

Load data_lab9.mat. The vectors 'natural' and 'heritage' represent color values of natural and heritage pictures. You need to analyze this data, and answer if natural content is different from heritage content in terms of the color measure. Your analysis should be end-to-end (justify any choices that you make: choice of test statistic, parametric or nonparametric, effect of sample size etc.).

Since we have to do end-to-end analysis of the given samples, we will first look at the distribution of both the samples in order to get intuitive idea about the nature of the two populations. Then we will move to statistical inference.



Histogram for the **heritage image colours is very skewed** and has some outliers while the histogram for the **colour values of the natural image sample** 'looks' symmetric.

Since the **histograms of sample are very different in nature**, it is highly likely that the **two populations are significantly different**. But since we don't have the information about the underlying populations, we need to **perform statistical tests to confirm this** 'gut-feeling'.

Choosing appropriate test statistic:

Since heritage images **have some outliers, "mean" would be affected by them**. Therefore median is a **better test statistic** as it is robust to outliers.

Choosing appropriate statistical test:

- Since we are using median as test statistic**, t-test cannot be applied**. This is because we cannot **apply CLT to median** and therefore cannot guarantee that the sampling distribution of test statistic (Here - median) would be normal. Hence, the **basic assumption of t-test is not true**.
- Looking at the sample size, since both samples are of **size 9968**, which means **degrees of freedom are of the order of 10^4,** the **t-test would not be accurate enoug**h and might give **statistical significance just because of the increased degrees of freedom**. This is another drawback of **performing t-test**. Hence **we won't be applying t-test**.

- Therefore we go for **non-parametric test.**

Non-Parametric testing

Test Statistic Used - **Median**

Null Hypothesis:

H_0 = Median of natural images (population) = Median of heritage images (population)

Alternate Hypothesis

H_a = They are not equal.

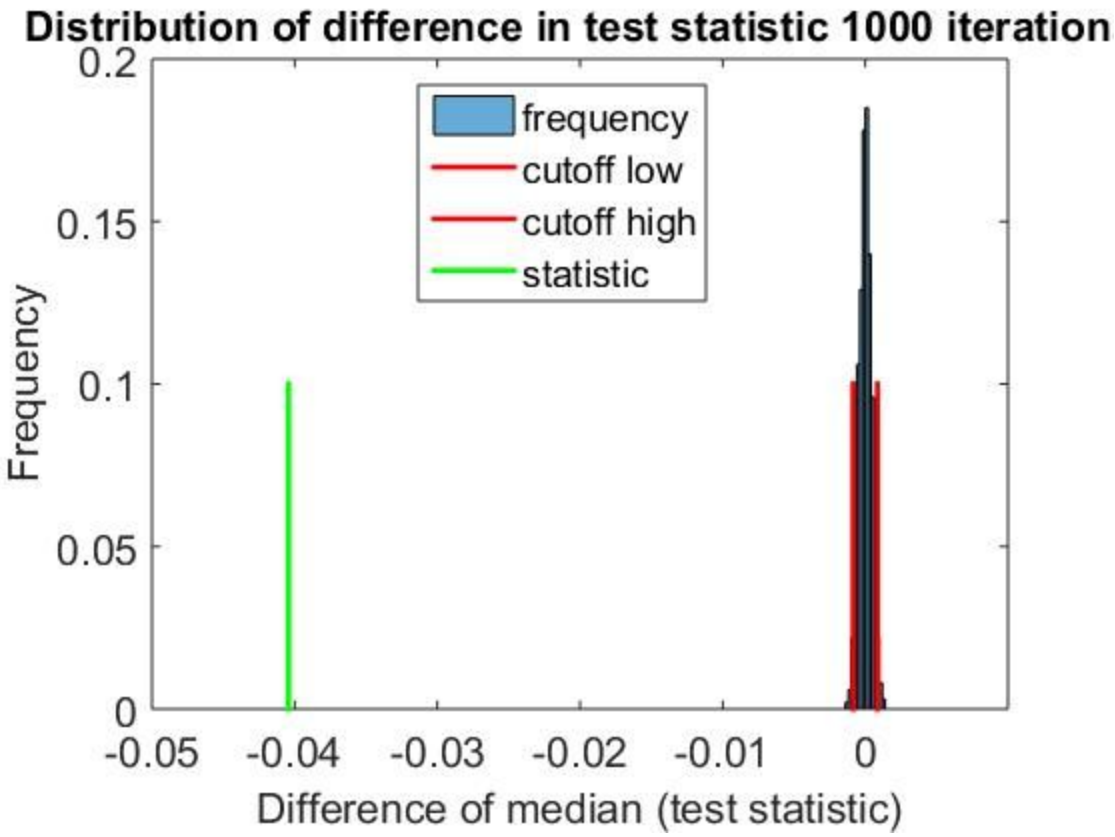
I. Bootstrap resampling method

We follow a **similar approach as in question 1.** Merge two samples and then resample the data to take samples of the same size.

The process is as follows:

a. Find out their test statistic i.e. median in this case and take difference of median. Perform this a large number of times. Here it is performed 1000 times.
b. Plot distribution of difference of median.
c. Find its bootstrap Confidence Interval by finding 2.5 percentile and 97.5 percentile as its limits with alpha = 0.05
d. Check if the difference of mean of the original samples lies within the bootstrap CI, if not then reject the null hypothesis.

We get the following graph

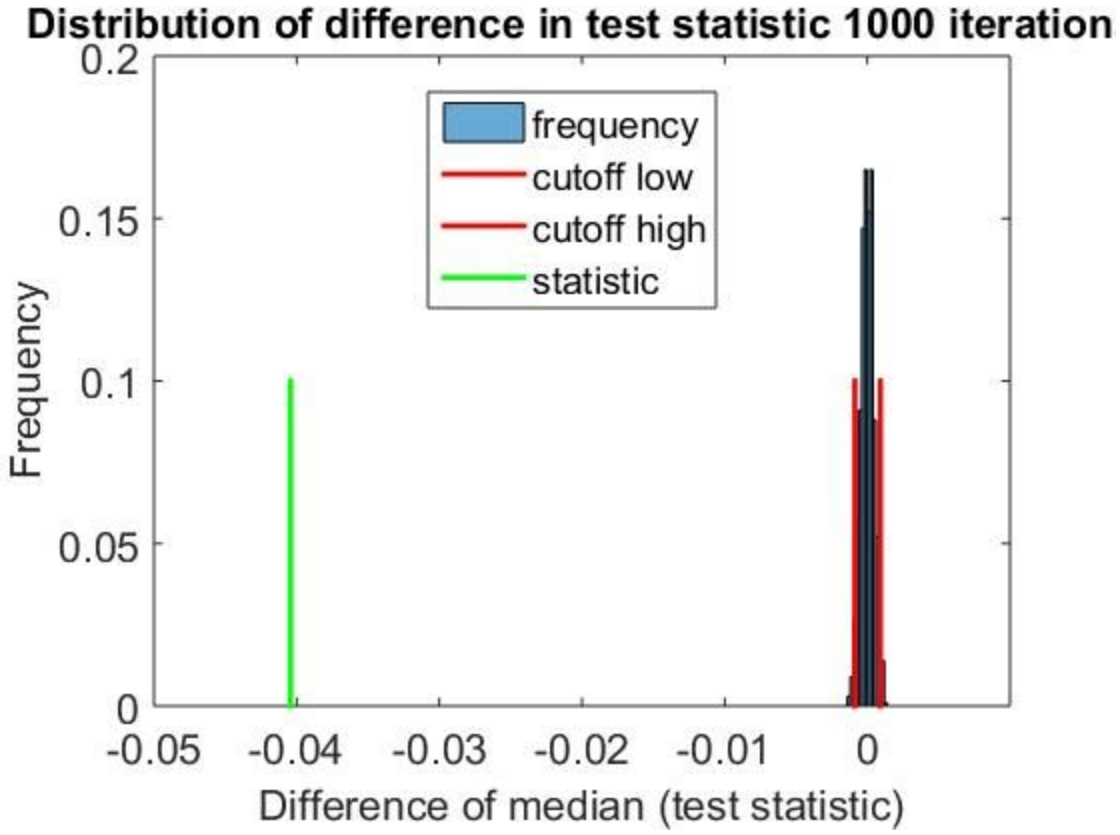


Here the **difference in mean of the original sample is way out of the bootstrap CI** and hence we **reject null hypothesis** i.e. the **two populations from which the given two samples** have been obtained have **statistically significant difference**. So we can b**e confident that the hull hypothesis is to be rejected.**

This result validates our initial guess that there has to be a significant difference between images of historical structures and images of nature.

II. Permutation test:

Here we get the following graph



So even here, we will reject the null hypothesis as the actual statistic lies far outside the assumed distribution generated by label permutation.

This emphasizes our earlier conclusion that there is significant difference between images of historical structures and images of nature.

Codes:

Q1

t-test and Bootstrap

```
% Author - Aditya Joglekar

% Date - 11th April, 2017


clear;

close all;


load('data_lab9.mat')

%% performing t test

std(traditional)

std(new)

[t3, df3] = calculate_t(mean(traditional), mean(new), std(traditional),
std(new), numel(traditional), numel(new))


%% non parametric test - bootstrapping

combined= [traditional new];


% now for each itera resample, assign to x and y, compute the stat diff.

ite=10000;

n1= size(traditional,2); % columns of trad

n2= size(new,2);


stat_arr= zeros(ite,1);


for i=1:ite
```

```
      % sample with replace n1+n2, permute then n1 to n2
      sample_with_rep= datasample(combined,n1+n2); % samples data with
REPLACEMENT


      sample_with_rep( randperm( length(sample_with_rep) ) ); % shuffled


      X= sample_with_rep(1:n1);
      Y= sample_with_rep((n1+1):(n1+n2));


      stat_arr(i)= median(X)- median(Y);
end


histogram(stat_arr);
hold on
title('Histogram approximating the resampling distribuition of difference of
median of the two groups')
xlabel('Range of difference in median statistic');
ylabel('Frequency');
% compute 2.5 and 97.5 th percentile


ci_lower=prctile(stat_arr,2.5)
ci_higher= prctile(stat_arr,97.5)
orig_median_diff= median(traditional)- median(new)


line([ci_lower ci_lower], [0 1400], 'Color','red','LineStyle','--')
line([ci_higher ci_higher], [0 1400], 'Color','red','LineStyle','--')
line([orig_median_diff orig_median_diff ], [0 1400], 'Color','black')
```

Function to calculate t-score

```
% function to calculate degree of freedom and t value for given set of data


function [t, df] = calculate_t(x1, x2, s1, s2, n1, n2)


% parameters
      % x1 : mean of first sample
      % x2 : mean of second sample
      % s1 : standard deviation of first sample
      % s2 : standard deviation of second sample
      % n1 : number of sample points of first sample
      % n2 : number of sample points of second sample


% return values
      % df : degrees of freedom
      % t : t value of the data




      % calculating degrees of freedom
      df = n1 + n2 - 2;


      % calculating t value
      t = (x1 - x2) / ( sqrt( ((s1*s1*(n1-1) + s2*s2*(n2-1))/df) * (1/n1 +
1/n2) ) );
end
```

Permutation test

```
% Author - Aditya Joglekar
% Date - 11th April, 2017

clear;
close all;

load('data_lab9.mat')

combined= [traditional new];



% now for each itera resample, assign to x and y, compute the stat diff.
ite=10000;
n1= size(traditional,2); % columns of trad
n2= size(new,2);

stat_arr= zeros(ite,1);

for i=1:ite

      % sample with replace n1+n2, permute then n1 to n2
      %sample_with_rep= datasample(combined,n1+n2);

      % shuffled, doubt
   combined= combined( randperm( length(combined) ) );
```

```
        X= combined(1:n1);
        Y= combined((n1+1):(n1+n2));

        stat_arr(i)= median(X)- median(Y);
end

histogram(stat_arr);
hold on
title('Histogram approximating the permutation distribuition of difference of median of the two groups')
xlabel('Range of difference in median statistic');
ylabel('Frequency');

% compute 2.5 and 97.5 th percentile

ci_lower=prctile(stat_arr,2.5)
ci_higher= prctile(stat_arr,97.5)
orig_median_diff= median(traditional)- median(new)

line([ci_lower ci_lower], [0 1400], 'Color','red','LineStyle','--')
line([ci_higher ci_higher], [0 1400], 'Color','red','LineStyle','--')
line([orig_median_diff orig_median_diff ], [0 1400], 'Color','black')
```

Q2

Bootstrap

```
% Author - Rajdeep Pinge

% Date - 11th April, 2017


clear;

close all;


load('data_lab9');


heritage_images = heritage;

natural_images = natural;



mean(heritage_images)

std(heritage_images)

subplot(1, 2, 1), histogram(heritage_images, 10, 'Normalization',
'probability');

axis([0 1 0 1])

title('Color values in given heritage image sample')

xlabel('Color Values')

ylabel('Probability')



mean(natural_images)

std(natural_images)

subplot(1, 2, 2), histogram(natural_images, 10, 'Normalization',
'probability');

axis([0 1 0 1])

title('Color values in given natual image sample')
```

```
xlabel('Color Values')
ylabel('Probability')


set(gca,'FontSize',16)
set(findall(gcf,'type','text'),'FontSize',16)
print('q2_sample_histogram','-djpeg')

%% non-parametric testing - bootstraping, resampling method

combined = [heritage_images natural_images];


% now for each iteration resample, assign to x and y, compute the statistical
difference.
ite = 1000;
n1 = numel(natural_images); % number of elements in each sample
n2 = numel(heritage_images);

% array to store distribution of test statistic
stat_arr = zeros(ite,1);

for i=1:ite

      % sample with replace n1+n2, permute then n1 to n2
      sample_with_rep = datasample(combined, n1+n2);

      sample_with_rep( randperm( length(sample_with_rep) ) ); % shuffled

      X = sample_with_rep(1:n1);
```

```
        Y = sample_with_rep((n1+1):(n1+n2));

        % store difference of medians
        stat_arr(i)= median(X)- median(Y);
end

figure
histogram(stat_arr, 'Normalization', 'probability');
title('Distribution of difference in test statistic 1000 iterations')
xlabel('Difference of median (test statistic)')
ylabel('Frequency')

set(gca,'FontSize',16)
set(findall(gcf,'type','text'),'FontSize',16)
print('q2_bootstrap','-djpeg')



% compute 2.5 and 97.5 th percentile

ci_lower = prctile(stat_arr,2.5)
ci_higher = prctile(stat_arr,97.5)

orig_median_diff = median(heritage_images)- median(natural_images)

hold on
line([ci_lower,
ci_lower],[0,0.1],[0,0],'LineStyle','-','Color','r','LineWidth', 2);

hold on
```

```
line([ci_higher,
ci_higher],[0,0.1],[0,0],'LineStyle','-','Color','r','LineWidth', 2);


hold on

line([orig_median_diff,
orig_median_diff],[0,0.1],[0,0],'LineStyle','-','Color','g','LineWidth', 2);



%% non-parametric testing - permutation method



median(heritage_images)

median(natural_images)
```

Permutation Method

```
% Author - Rajdeep Pinge

% Date - 11th April, 2017


clear;

close all;


load('data_lab9');
```

```
heritage_images = heritage;

natural_images = natural;



mean(heritage_images)

std(heritage_images)

subplot(1, 2, 1), histogram(heritage_images, 10, 'Normalization',
'probability');

axis([0 1 0 1])

title('Color values in given heritage image sample')

xlabel('Color Values')

ylabel('Probability')



mean(natural_images)

std(natural_images)

subplot(1, 2, 2), histogram(natural_images, 10, 'Normalization',
'probability');

axis([0 1 0 1])

title('Color values in given natual image sample')

xlabel('Color Values')

ylabel('Probability')



set(gca,'FontSize',16)

set(findall(gcf,'type','text'),'FontSize',16)

print('q2_sample_histogram','-djpeg')

%% non-parametric testing - permutation method
```

```
combined = [heritage_images natural_images];



% now for each iteration resample, assign to x and y, compute the statistical
difference.
ite = 1000;
n1 = numel(natural_images); % number of elements in each sample
n2 = numel(heritage_images);

% array to store distribution of test statistic
stat_arr = zeros(ite,1);

for i=1:ite

      % sample with replace n1+n2, permute then n1 to n2
      %sample_with_rep = datasample(combined, n1+n2);

      combined= combined( randperm( length(combined) ) );



      X= combined(1:n1);
      Y= combined((n1+1):(n1+n2));

      % store difference of medians
      stat_arr(i)= median(X)- median(Y);
end

figure
histogram(stat_arr, 'Normalization', 'probability');
title('Distribution of difference in test statistic 1000 iterations')
```

```
xlabel('Difference of median (test statistic)')

ylabel('Frequency')


set(gca,'FontSize',16)

set(findall(gcf,'type','text'),'FontSize',16)

print('q2_bootstrap','-djpeg')



% compute 2.5 and 97.5 th percentile


ci_lower = prctile(stat_arr,2.5)

ci_higher = prctile(stat_arr,97.5)


orig_median_diff = median(heritage_images)- median(natural_images)


hold on

line([ci_lower,
ci_lower],[0,0.1],[0,0],'LineStyle','-','Color','r','LineWidth', 2);


hold on

line([ci_higher,
ci_higher],[0,0.1],[0,0],'LineStyle','-','Color','r','LineWidth', 2);


hold on

line([orig_median_diff,
orig_median_diff],[0,0.1],[0,0],'LineStyle','-','Color','g','LineWidth', 2);
```