CS306: Data Analysis and Visualization
Assignment: Report

Rajdeep Pinge 201401103

Problem description:

Parametric testing based on t-test requires three assumptions:

1. Assumption of normality

2. Homogeneity of variance

3. Data independence

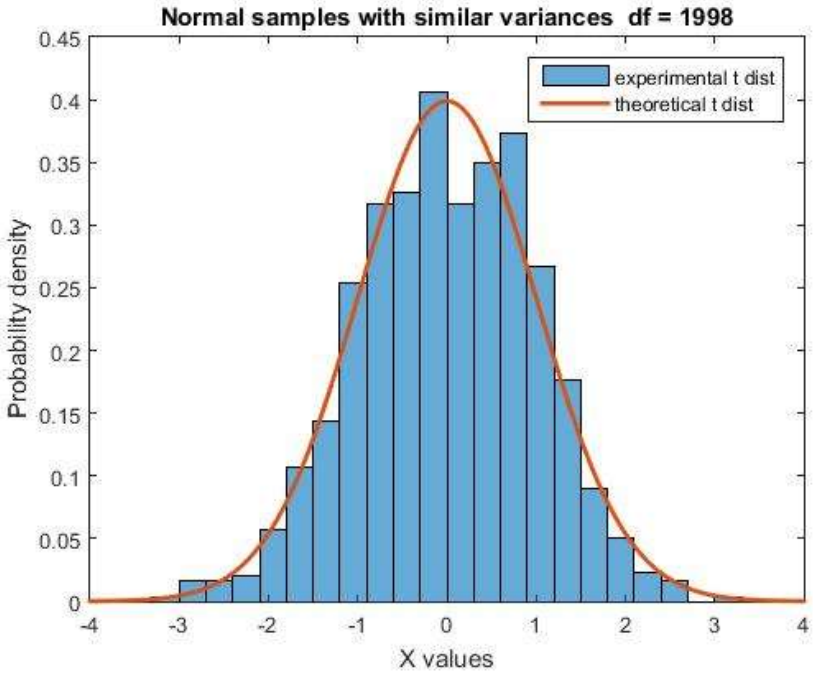
These are required so that the sampling distribution of t follows the theoretical t-distribution with the corresponding degree of freedom. The goal is to verify the role of first two assumptions. To that end, obtain the sampling distribution of t in four cases, and analyze the role of the said assumptions:

(a) Normal samples with similar variances

Here to generate normal sample, first a standard normal sample is generated in matlab and then it is shifted and scaled to the desired mean and standard deviation so as to get a general normal distribution.

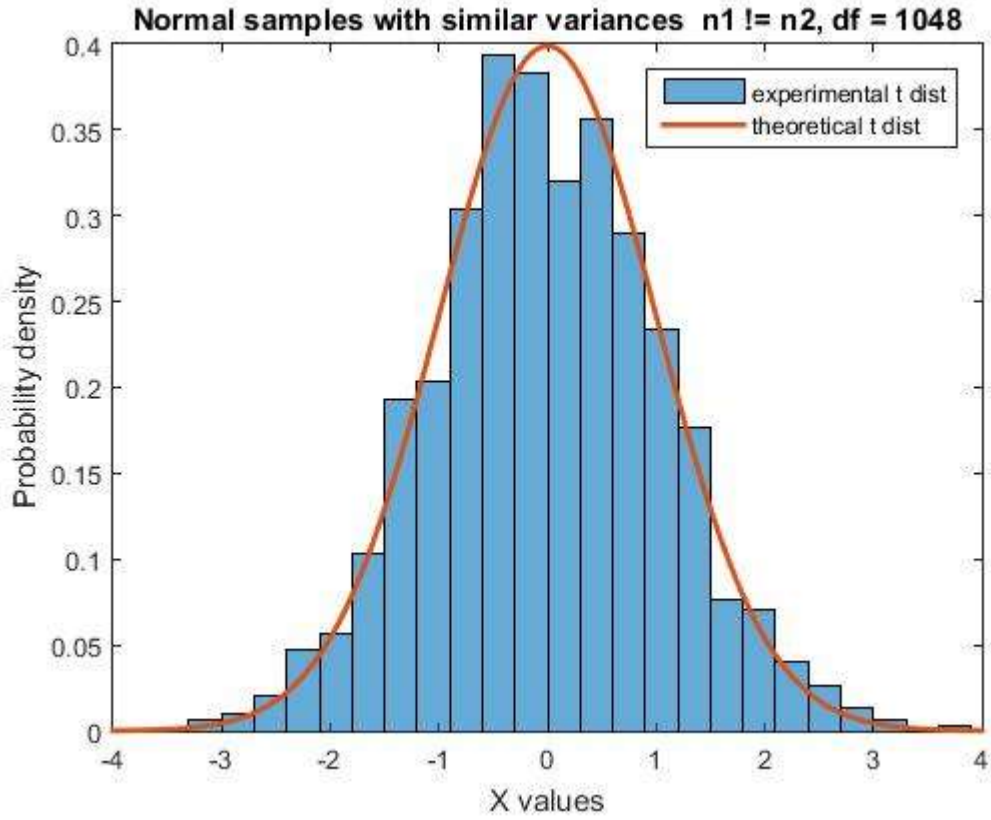
The variances are similar meaning (sigma1*sigma1 < 2*sigma2*sigma2)

When Sample Sizes are same.



In this graph, the histogram indicates the probability distribution of the experimental t distribution while the red graph shows the theoretical t distribution. It can be observed that the both the distributions closely match indicating that when both the assumptions are satisfied, the t test is successful and distribution of t follows the theoretical t-distribution with the corresponding degree of freedom.

When sample sizes are not same.



Even when sample sizes are not same, if both the above assumptions are satisfied, then the distribution of t follows the theoretical t-distribution with the corresponding degree of freedom.

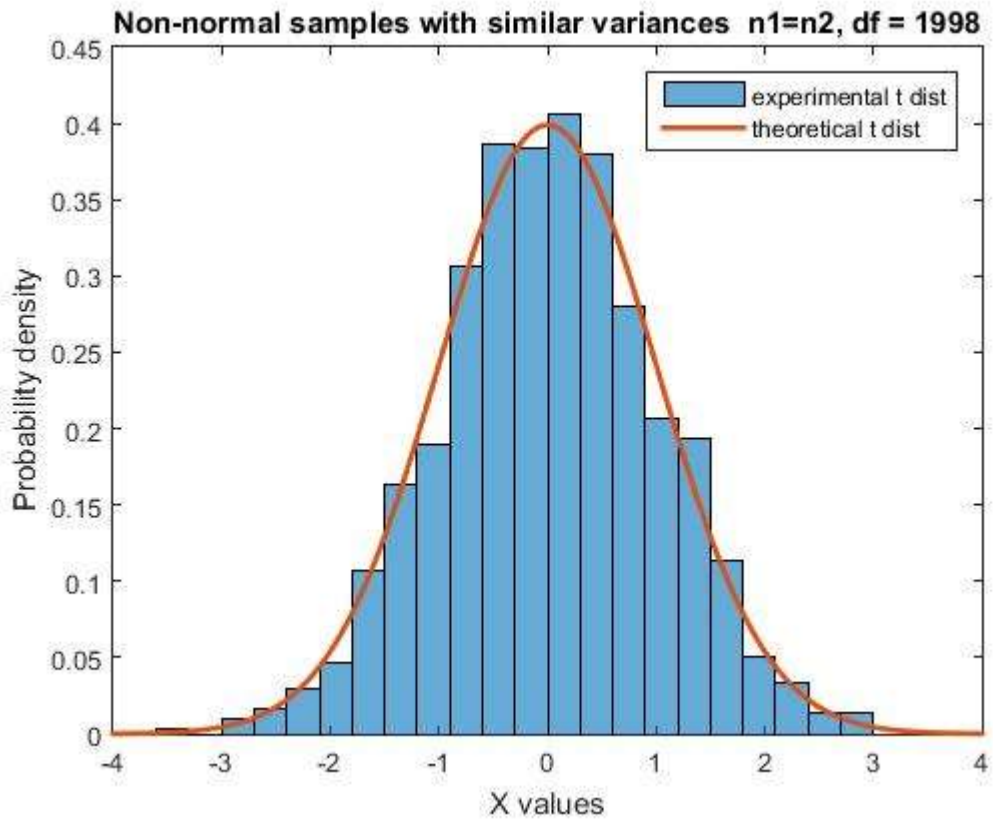
Hence if both the assumptions are satisfied, then the t test does not depend on the sample sizes.

(b) Non-normal samples with similar variances

Here we have generated Poisson distribution sample as the non normal sample.

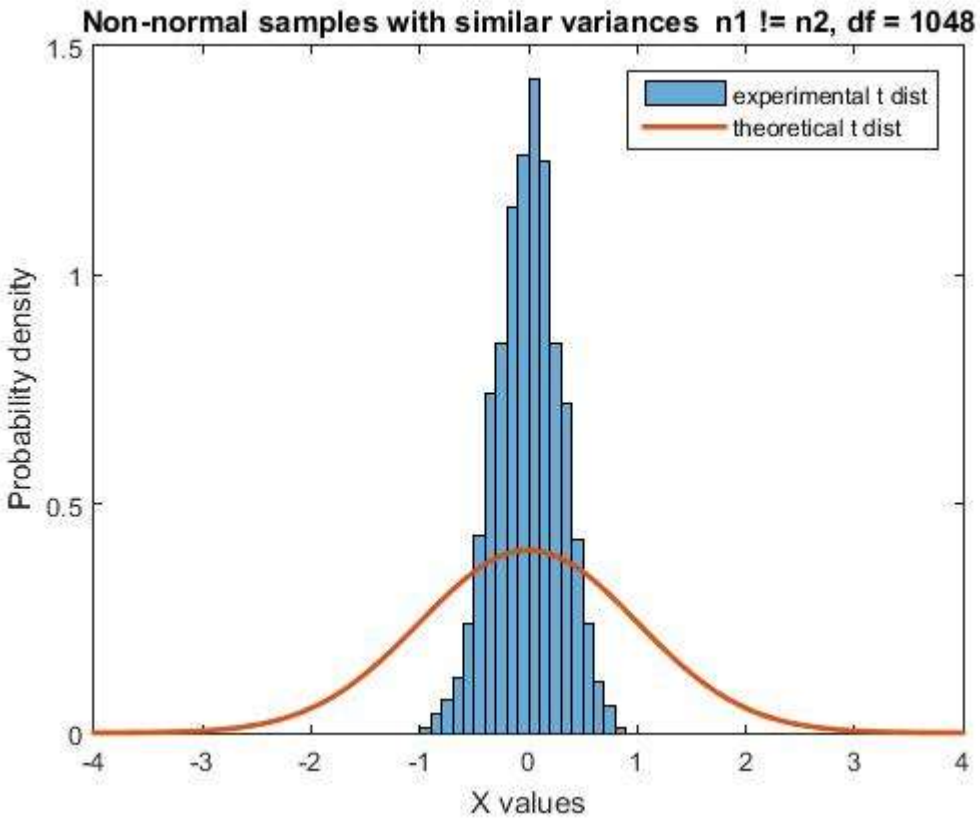
The variances are similar meaning (sigma1*sigma1 < 2*sigma2*sigma2)

When Sample Sizes are same.



Here also the experimental and theoretical distributions closely match indicating that if the sample sizes match, then there is very little effect of the non-normality of the samples.

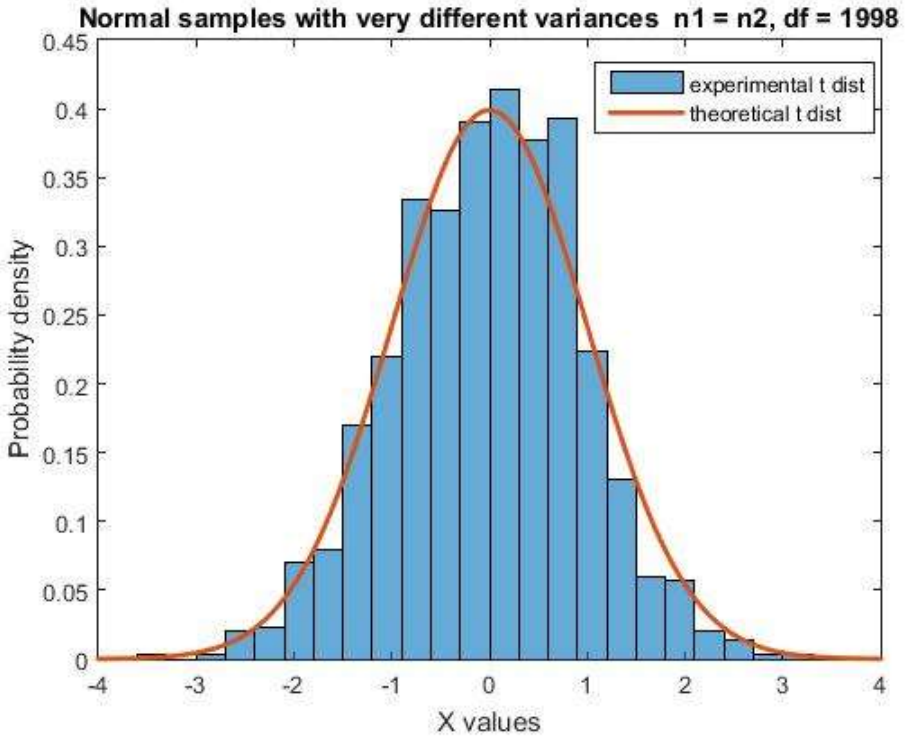
When sample sizes are not same.



This graph is very different from the above graphs. It shows that if the sample sizes are not normal, then there is significant effect of the non-normality of the samples. That is the assumption of normality is necessary if the sample sizes are not normal.

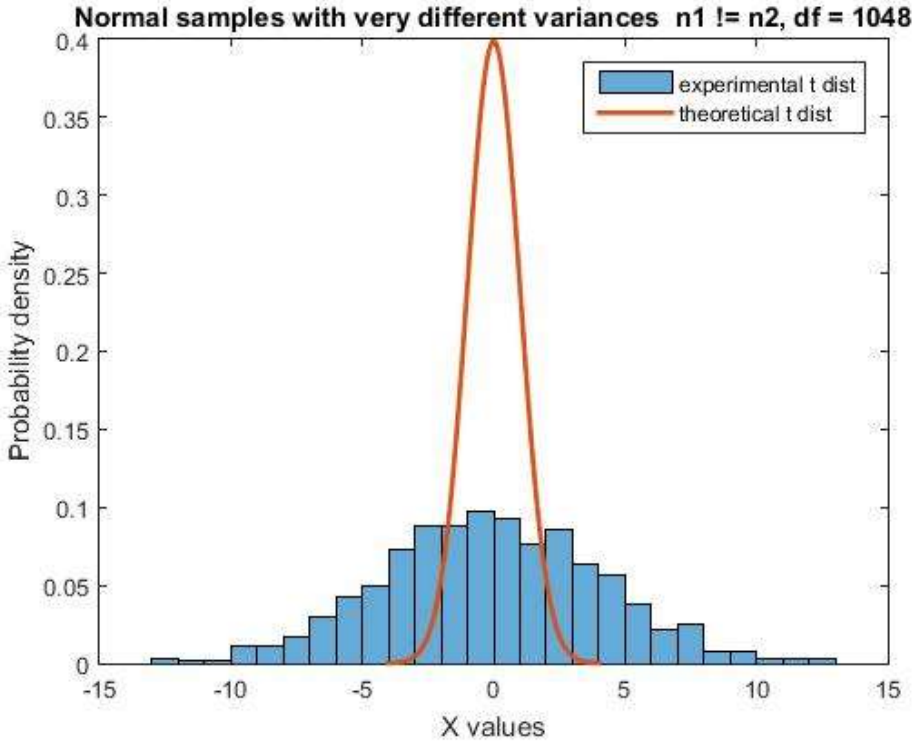
(c) Normal samples with very different variances

When Sample Sizes are same.



For the above graph, we see that difference in variance does not affect the experimental t distribution. The condition being that the sample size must be same.

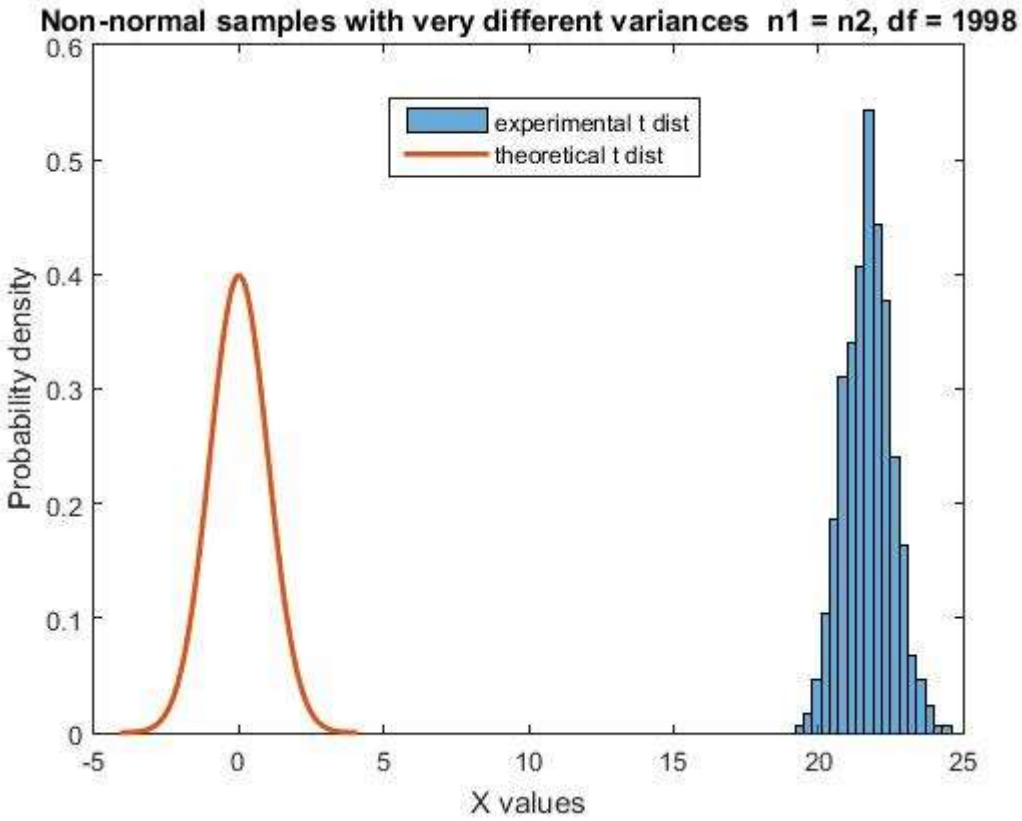
When sample sizes are not same.



Similar to the part (b), here also, the assumption of homogeneity of variances is necessary for the t distributions to match if the sample sizes are different.

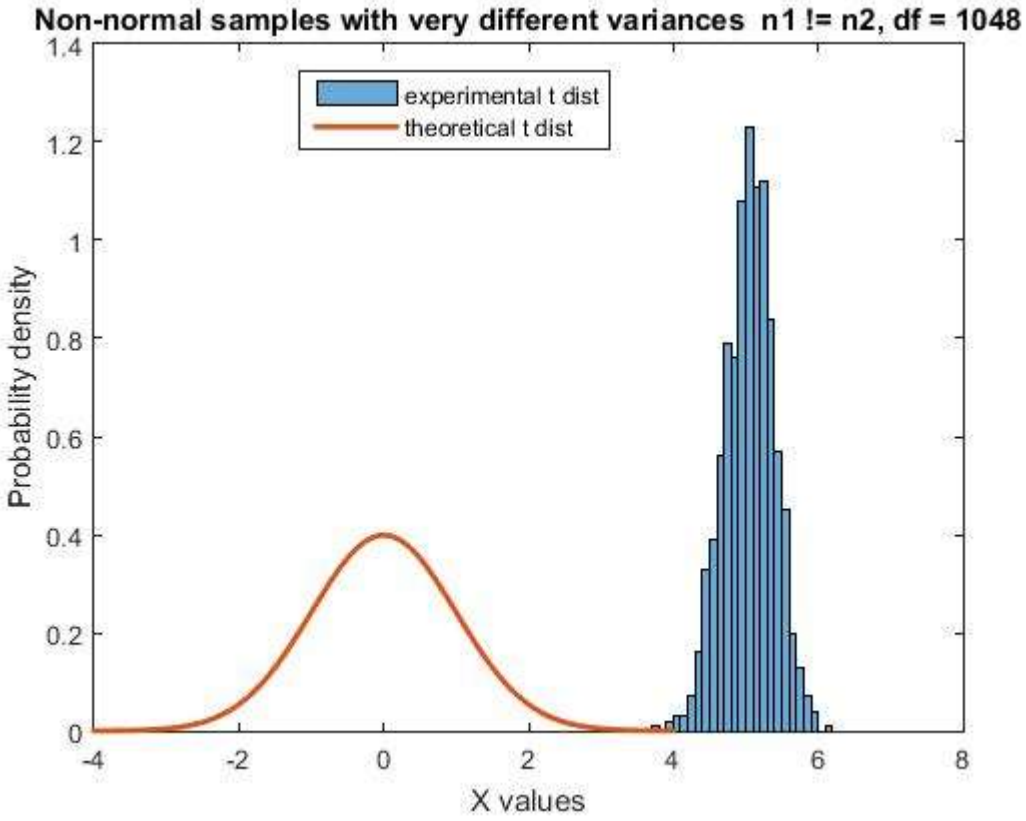
(d) Non-normal samples with very different variances

When Sample Sizes are same.



In this case, even when the sample sizes are same the distributions are not same. This shows that at least one assumptions is necessary along with the equality of the sample sizes so that the experimental distribution matches with the theoretical distribution.

When sample sizes are not same.



As expected, the experimental distribution is nowhere close to the theoretical distribution. Here neither of the assumptions are satisfied nor is the sample size same. Therefore the experimental distributions differs vastly from the theoretical distribution.

Your analysis should provide answers for the following:

1.Is sample normality or population normality required for t-test?

For the t test to work with any samples having different sample sizes, sample normality is important. If the sample sizes are same and the second assumptions violates, then in that case also sample normality is required.

2. Is homogeneity of variance necessary?

For the t test to work with any samples having different sample sizes, homogeneity of variance is essential. If the sample sizes are same and the samples are not normally distributed, then also this condition is necessary for the t test to give theoretical t distribution.

3. Does your answer to previous question depend on whether the sample sizes are equal or not?

Yes, the answers to the previous questions do depend on whether the sample sizes are equal or not. This may be because large difference in mean and/or effect significant difference in variance which affects the t distribution if one of the assumptions is not satisfied.

4. What are the implications if one performs t-test and one or both assumptions are violated? (hint: observe the experimental and theoretical t-distribution and see how the deviations between the two affects the decision about null hypothesis)

If both the assumptions are violated, then the experimental distribution varies to a large extent from the theoretical distribution. Hence the t scores would not be accurate and the decision about the null hypothesis would be wrong.

If one of the assumptions is violated then the extent of deviation between distributions would depend on the sizes of the samples taken for the experiment. If the sizes are different, then also the experimental t distribution would change.

In case of non normal distribution sample the range of t values would be much smaller than the theoretical ones.

In case of non homogeneity of variances, the range of t values would be much larger than the theoretical ones.

In both the cases, the calculated t values would not represent accurately if the samples have significant difference or not i.e. any comment regarding acceptance or rejection of null hypothesis would be incorrect

Codes:

A part

```
% Author - Rajdeep Pinge
% Date - 03-03-2017

% Code to verify the role of two assumptions:
% 1. Assumption of Normality
% 2. Homogeneity of Variance
% required by the t-test so that the sampling distribution of t follows
% the theoretical t-distribution with the corresponding degree of freedom.

% Here we use normal samples having similar variance

clear;
close all;

iterations = 1e3;

t_arr = zeros(iterations, 1);        % array to score t values of each t
test

for i = 1:iterations

    % generate first normally distributed sample
    n1 = 1000;
    norm1 = normrnd(zeros(n1, 1), ones(n1, 1));

    mu1 = 10;
    sigma1 = 10;

    norm_dist_1 = norm1 * sigma1 + mu1;


    % generate second normally distributed sample
    n2 = 50;
    norm2 = normrnd(zeros(n2, 1), ones(n2, 1));

    mu2 = 10;
    sigma2 = 10;
    norm_dist_2 = norm2 * sigma2 + mu2;


    % perform t test
    t_arr(i) = calculate_t(mean(norm_dist_1), mean(norm_dist_2),
std(norm_dist_1), std(norm_dist_2), n1, n2);

end


histogram(t_arr, 'Normalization', 'pdf')
hold on
plot((-4:4e-3:4), tpdf((-4:4e-3:4), n1+n2-2), 'LineWidth', 2);
title(['Normal samples with similar variances  n1 != n2, df = '
num2str(n1+n2-2)])
xlabel('X values')
```

```
ylabel('Probability density')
legend('experimental t dist', 'theoretical t dist')
```

B part

```
% Author - Rajdeep Pinge
% Date - 03-03-2017

% Code to verify the role of two assumptions:
% 1. Assumption of Normality
% 2. Homogeneity of Variance
% required by the t-test so that the sampling distribution of t follows
% the theoretical t-distribution with the corresponding degree of freedom.

% Here we use non normal samples having similar variance

clear;
close all;

iterations = 1e3;

t_arr = zeros(iterations, 1);        % array to score t values of each t
test

for i = 1:iterations

    % generate first poisson distributed sample
    n1 = 1000;
    norm1 = poissrnd(ones(n1, 1));

    lambda1 = 2;
    norm_dist_1 = norm1 * lambda1;


    % generate second poisson distributed sample
    n2 = 50;
    norm2 = poissrnd(ones(n1, 1));

    lambda2 = 2;
    norm_dist_2 = norm2 * lambda2;


    % perform t test
    t_arr(i) = calculate_t(mean(norm_dist_1), mean(norm_dist_2),
std(norm_dist_1), std(norm_dist_2), n1, n2);

end


histogram(t_arr, 'Normalization', 'pdf')
hold on
plot((-4:4e-3:4), tpdf((-4:4e-3:4), n1+n2-2), 'LineWidth', 2);
```

```
title(['Non-normal samples with similar variances  n1 != n2, df = '
num2str(n1+n2-2)])
xlabel('X values')
ylabel('Probability density')
legend('experimental t dist', 'theoretical t dist')
```

C part

```
% Author - Rajdeep Pinge
% Date - 03-03-2017

% Code to verify the role of two assumptions:
% 1. Assumption of Normality
% 2. Homogeneity of Variance
% required by the t-test so that the sampling distribution of t follows
% the theoretical t-distribution with the corresponding degree of freedom.

% Here we use normal samples having differences in variance (non
% homogenious variance)

clear;
close all;

iterations = 1e3;

t_arr = zeros(iterations, 1);        % array to score t values of each t
test

for i = 1:iterations

    % generate first normally distributed sample
    n1 = 1000;
    norm1 = normrnd(zeros(n1, 1), ones(n1, 1));

    mu1 = 2;
    sigma1 = 3;

    norm_dist_1 = norm1 * sigma1 + mu1;


    % generate second normally distributed sample
    n2 = 50;
    norm2 = normrnd(zeros(n2, 1), ones(n2, 1));

    mu2 = 2;
    sigma2 = 30;
    norm_dist_2 = norm2 * sigma2 + mu2;


    % perform t test
    t_arr(i) = calculate_t(mean(norm_dist_1), mean(norm_dist_2),
std(norm_dist_1), std(norm_dist_2), n1, n2);
```

```
end


histogram(t_arr, 'Normalization', 'pdf')
hold on
plot((-4:4e-3:4), tpdf((-4:4e-3:4), n1+n2-2), 'LineWidth', 2);
title(['Normal samples with very different variances  n1 != n2, df = '
num2str(n1+n2-2)])
xlabel('X values')
ylabel('Probability density')
legend('experimental t dist', 'theoretical t dist')
```

D part

```
% Author - Rajdeep Pinge
% Date - 03-03-2017

% Code to verify the role of two assumptions:
% 1. Assumption of Normality
% 2. Homogeneity of Variance
% required by the t-test so that the sampling distribution of t follows
% the theoretical t-distribution with the corresponding degree of freedom.

% Here we use non normal samples having difference in variance (non
% homogenious variance)

clear;
close all;

iterations = 1e3;

t_arr = zeros(iterations, 1);        % array to score t values of each t
test

for i = 1:iterations

    % generate first poisson distributed sample
    n1 = 1000;
    norm1 = poissrnd(ones(n1, 1));

    lambda1 = 7;
    norm_dist_1 = norm1 * lambda1;


    % generate second poisson distributed sample
    n2 = 1000;
    norm2 = poissrnd(ones(n2, 1));

    lambda2 = 2;
    norm_dist_2 = norm2 * lambda2;
```

```
    % perform t test
    t_arr(i) = calculate_t(mean(norm_dist_1), mean(norm_dist_2),
std(norm_dist_1), std(norm_dist_2), n1, n2);

end


histogram(t_arr, 'Normalization', 'pdf')
hold on
plot((-4:4e-3:4), tpdf((-4:4e-3:4), n1+n2-2), 'LineWidth', 2);
title(['Non-normal samples with very different variances  n1 = n2, df = '
num2str(n1+n2-2)])
xlabel('X values')
ylabel('Probability density')
legend('experimental t dist', 'theoretical t dist')
```

Function to perform t test

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

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

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

% return values
    % 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
```