EMAT30007 Applied Statistics

Lab 5: Nonparametric and permutation methods

This lab focusses on the material covered in Lecture 5: hypothesis testing using nonparametric and permutation methods

1. Kolmogorov-Smirnov test for one sample

(Download the functions for the Kolmogorov distribution from Blackboard - they are from https://uk.mathworks.com/matlabcentral/fileexchange/4369-kolmogorov-distribution-functions)

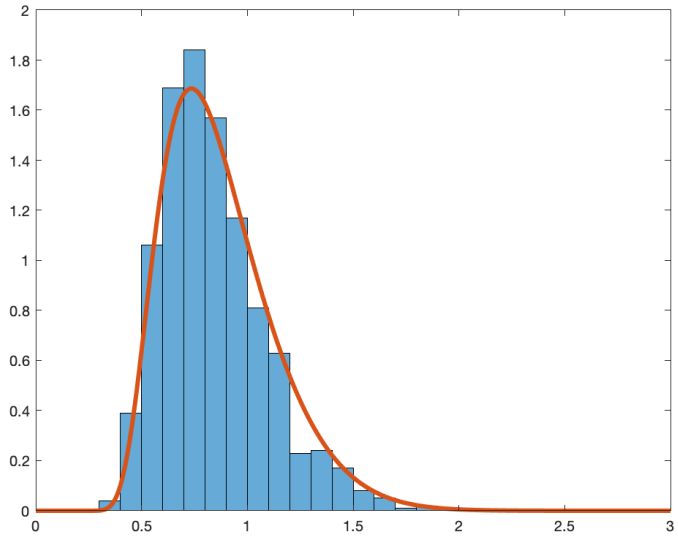
Generate the Kolmogorov distribution using sampling

1. Draw $r = 1000$ random samples of size $n = 100$ from an Exponential distribution and for each sample compute the Kolmogorov test statistic D.
2. Compute the empirical PDF (histogram) of the Random Variable $K = \sqrt{n}\, D$ and compare it to the theoretical Kolmogorov distribution using `kolmpdf`.
3. Repeat the steps of point 2. using the formula $K = D\sqrt{n} + \frac{1}{6\sqrt{n}} + \frac{D\sqrt{n}-1}{4n}$ and check it is a better approximation of the Kolmogorov distribution.
4. Repeat steps 1. and 3. with samples drawn from other continuous distributions, for example Normal and Uniform, and verify that all histograms are close to the Kolmogorov distribution.

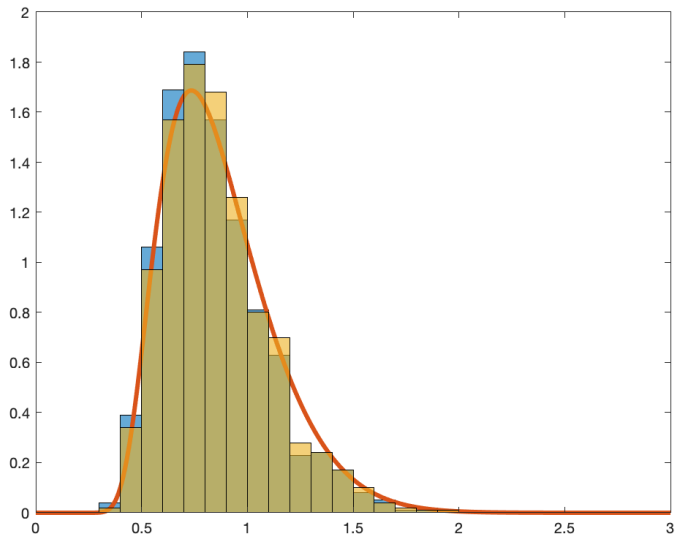
Solution

```
% 1.
pd = makedist('Exponential', 'mu', 1);

% 4.
% pd = makedist('Normal');

n = 100;
r = 1000;

% compute the K-S test statistic
Ds = zeros(1,r);
for i = 1:r
    x = random(pd, n, 1);
    [f, x_values] = ecdf(x);
    % K-S test statistic: sup of distances
    Ds(i) = max( max(abs(f - cdf(pd, x_values))), max(abs(f(1:end-1,:) - cdf(pd, x_values(2:end,:)) )) );
end
```

```
% 2.
K = Ds * n^0.5;

% plot Kolmogorov PDF
clf
histogram(K, 'normalization', 'pdf')
hold on
k = 0:0.01:3;
plot(k, kolmpdf(k), '-', 'Linewidth', 3);
```




```
% 3.
K_accurate = Ds * n^0.5 + 1/(6 * n^0.5) + (Ds * n^0.5 - 1)/4/n;

% plot Kolmogorov PDF
% clf
histogram(K_accurate, 'normalization', 'pdf')
hold on
```




```
% k = 0:0.01:3;
% plot(k, kolmpdf(k), '-', 'Linewidth', 3);
```

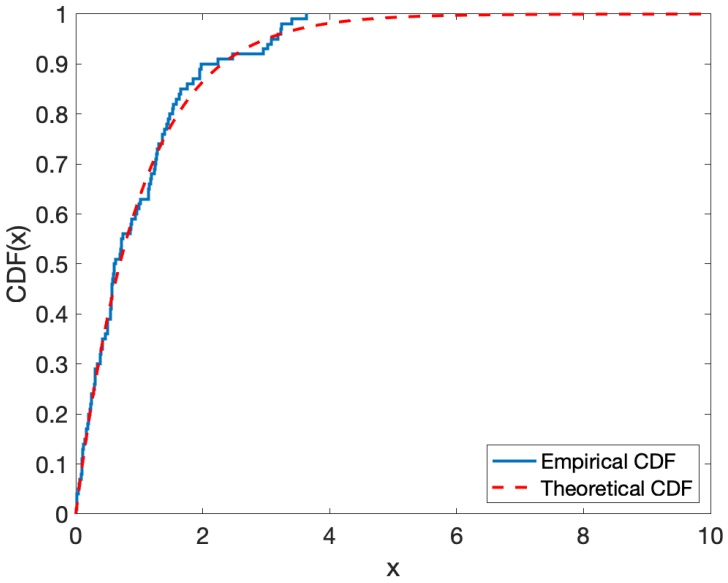
One-sample K-S test on a Exponential distribution

1. Generate a sample of size $n = 100$ from an Exponential distribution with parameter $\mu = 1$ and plot the empirical and theoretical CDFs.
2. Compute the K-S statistic D under the hypothesis that the data comes from an Exponential distribution with parameter $\mu = 1$
3. Compute the p-values of $K = D\sqrt{n}$ and $K = D\sqrt{n} + \frac{1}{6\sqrt{n}} + \frac{D\sqrt{n} - 1}{4n}$ using `kolmcdf`
4. Use Matlab's `kstest` to check your results
5. Repeat points 1.-4. under the hypothesis that the data comes from an Exponential distribution with parameter $\mu = 1.2$. Do you reject the null hypothesis at a significance level of $\alpha = 0.01$ (make a few attempts generating different samples)? If not, increase n until you find that the null hypothesis is rejected at level 0.01.

Solution

```
% 1.
n = 100;
pd = makedist('Exponential', 'mu', 1);
x = random(pd, n, 1);

% % 5.
```

```
% pd2 = makedist('Exponential', 'mu', 1.2);
% x = random(pd2, n, 1);

% plot
[f, x_values] = ecdf(x);

clf
J = stairs(x_values,f);
hold on;
xx = 0:0.1:10;
K = plot(xx, cdf(pd, xx),'r--');

set(J,'LineWidth',2);
set(K,'LineWidth',2);
legend([J K],'Empirical CDF','Theoretical CDF','Location','SE');
set(gca, 'FontSize', 16.0);
xlabel('x')
ylabel('CDF(x)')
```




```
% 2.
% K-S test statistic: supremum of distances
D = max( max(abs(f - cdf(pd, x_values))), max(abs(f(1:end-1,:) - cdf(pd, x_values(2:end,:)) )) )
```

```
D = 0.0504
```

```
% 3.
% p-value
1 - kolmcdf(D * n^0.5)
```

```
ans = 0.9610
```

```
% better approximation of the p-value
1 - kolmcdf(D * n^0.5 + 1/(6 * n^0.5) + (D * n^0.5 - 1)/4/n)
```

```
ans = 0.9498
```

```
%4.
% check
[h, p, ksstat] = kstest(x, 'CDF', pd, 'alpha', 0.01)
```

```
h = logical
   0
p = 0.9498
ksstat = 0.0504
```

With $n = 1000$ or higher we reject the null hypothesis at significance level 0.01.

2. Kolmogorov-Smirnov test for two samples

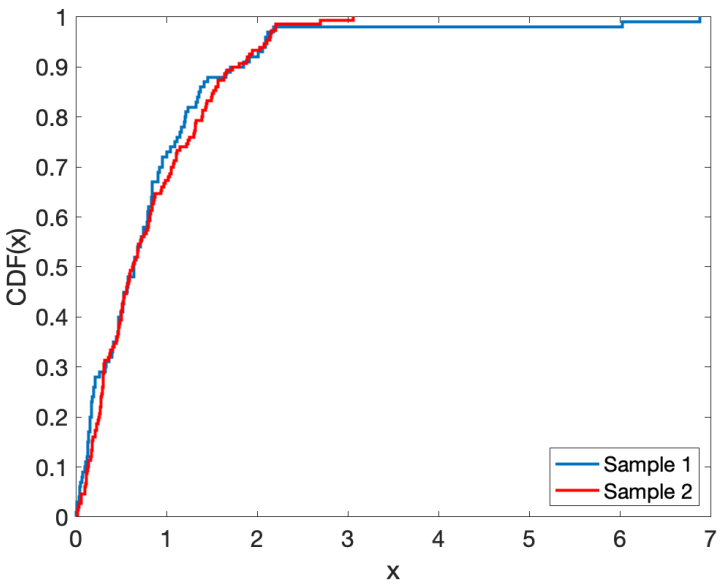
Two-sample K-S test on a Exponential distribution

1. Generate a sample of size $n = 100$ from an Exponential distribution with parameter $\mu = 1$
2. Generate a sample of size $m = 50$ from an Exponential distribution with parameter $\mu = 1$
3. Use Matlab's `kstest2` to test the hypothesis that the two data samples come from the same population using the significance threshold $\alpha = 0.01$
4. Repeat points 1-3 for $r = 10000$ times: With a significance threshold $\alpha = 0.01$, what is the probability of type I errors?

5. Repeat points 1-3 for $r = 10000$ times and with the second sample drawn from an Exponential distribution with parameter $\mu = 2$. With a significance threshold $\alpha = 0.01$, what is the probability of type II errors?

Solution

```
% 1.

pd = makedist('Exponential', 'mu', 1);

n = 100;
x = random(pd, n, 1);
[fx, x_values] = ecdf(x);

% 2.
m = 150;
y = random(pd, m, 1);
[fy, y_values] = ecdf(y);

% plot
clf
J = stairs(x_values,fx);
hold on;
K = stairs(y_values,fy, '-r');

set(J,'LineWidth',2);
set(K,'LineWidth',2);
legend([J K],'Sample 1','Sample 2','Location','SE');
set(gca, 'FontSize', 16.0);
xlabel('x')
ylabel('CDF(x)')
```




```
% 3.
[h,p,ks2stat] = kstest2(x, y, 'alpha', 0.01)
```

```
h = logical
   0
p = 0.3329
ks2stat = 0.1200
```

```
% compare to the p-value from the formula for large samples given in the lecture
pval = 1 - kolmcdf(ks2stat * (n*m/(n+m))^0.5)
```

```
pval = 0.3533
```

```
% 4.
r = 10000;
n = 100;
m = 150;

n_rejections = 0;
for i = 1:r
    x = random(pd, n, 1);
    y = random(pd, m, 1);
    [h,p,ks2stat] = kstest2(x, y, 'alpha', 0.01);
    n_rejections = n_rejections + h;
```

```
end

% the probability of type I errors is equal to the fraction of times we
% rejected the null hypothesis (which is true)
typeIerr = n_rejections / r
```

```
typeIerr = 0.0114
```

```
% 5.
pd2 = makedist('Exponential', 'mu', 2);

r = 10000;
n = 100;
m = 150;

n_rejections = 0;
for i = 1:r
    x = random(pd, n, 1);
    y = random(pd2, m, 1);
    [h,p,ks2stat] = kstest2(x, y, 'alpha', 0.01);
    n_rejections = n_rejections + h;
end

% the probability of type II errors is equal to the fraction of times we
% didn't reject the null hypothesis (which is false)
typeIIerr = (r - n_rejections) / r
```

```
typeIIerr = 0.0389
```

3. Permutation test for equal means of two populations (A/B test)

A streaming media service wants to test if the new version of its website is more engaging than the old one. Two groups of random users are selected:

- for each user in the **first group**, the number of minutes spent on the **old version** of the website have been recorded and saved in the file `m_old.csv`
- for each user in the **second group**, the number of minutes spent on the **new version** of the website have been recorded and saved in the file `m_new.csv`.

1. Use a permutation test to determine if there has been a significant increase in the time spent on the new version of the website, using a significance level of 0.1%.

(Download the files `m_old.csv` and `m_new.csv` from Blackboard, and place them in Matlab's current folder, which is displayed in the left column, in order to load them using `readmatrix`)

Solution

```
g2 = readmatrix("m_new.csv");
g1 = readmatrix("m_old.csv");

% observed test statistic
observation = mean(g2) - mean(g1)
```

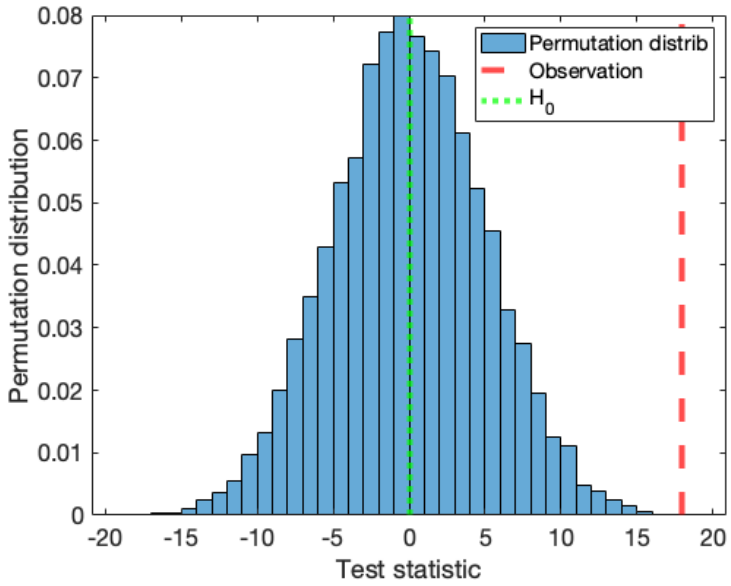
```
observation = 17.9639
```

```
% permutation test
% randomly reassing individuals to the two groups
n1 = length(g1);
n2 = length(g2);
g12 = cat(2, g1, g2);

S = 10000;
tstat = zeros(1, S);
for i = 1:S
    r = randperm(n1 + n2);
    gr = g12(r);
    gr1 = gr(1:n1);
    gr2 = gr(n1+1:n1+n2);
    m = mean(gr2) -  mean(gr1);
    tstat(i) = m;
end

% right-tailed p-value
pval = mean(tstat >= observation)
```

```
pval = 1.0000e-04
```

Since $p < \alpha = 0.001$ we reject the null hypothesis that the time spent on the website has not increased, at 0.1% significance level.

```
clf
histogram(tstat, 'normalization', 'pdf');
hold on
xline(observation, '--r', 'LineWidth', 4);
hold on
xline(0.0, ':g', 'LineWidth', 4);
legend('Permutation distrib', 'Observation', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Permutation distribution')
```



4. Permutation test for matched pairs

A streaming media service wants to test if the new version of its website is more engaging than the old one. A group of $n = 200$ random users is selected and for each user the minutes spent on the old and new versions of the website have been recorded and saved in the files `m_before.csv` and `m_after.csv`.

1. Use a permutation test to determine if there has been a significant increase in the time spent on the new version of the website, using a significance level of 0.1%.

(Download the files `m_before.csv` and `m_after.csv` from Blackboard, and place them in Matlab's current folder, which is displayed in the left column, in order to load them using `readmatrix`)

Solution

```
g1 = transpose(readmatrix("m_before.csv"));
g2 = transpose(readmatrix("m_after.csv"));
Dm = g2 - g1;
observation = mean(Dm)
```

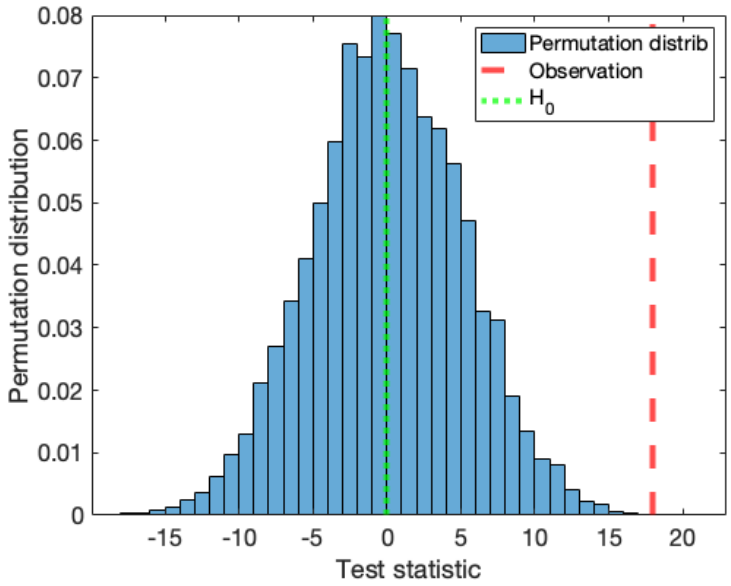
```
observation = 17.9639
```

```
% permutation test
% randomly swap each user's visit times to the new and old versions
n1 = length(g1);
n2 = length(g2);
g12 = cat(1, g1, g2);

S = 10000;
tstat = zeros(1, S);
for i = 1:S
    r = randi(2, 1, n1) - 1;
    % rm is a random matrix where each row can be either [-1 1] or [1 -1]
    % with equal probability
    rm = (cat(1, r, 1 - r) - 0.5) * 2;
    gr = g12 .* rm;
    m = mean(gr(2, :) + gr(1, :));
    tstat(i) = m;
end

% right-tailed p-value
pval = mean(tstat >= observation)
```

```
pval = 5.0000e-04
```

Since $p < \alpha = 0.001$ we reject the null hypothesis that the time spent on the website has not increased, at 0.1% significance level.

```
clf
histogram(tstat, 'normalization', 'pdf');
hold on
xline(observation, '--r', 'LineWidth', 4);
hold on
xline(0.0, ':g', 'LineWidth', 4);
legend('Permutation distrib', 'Observation', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Permutation distribution')
```



5. Permutation test for a relationship

We want to establish if temperature has been increasing over the years. To this end, we consider the monthly time series of midrange temperatures, that is the arithmetic mean of the maximum and minimum temperature in each month, recorded in Oxford since 1853.

1. Use a permutation test to determine if there has been a significant increase of the midrange temperatures over time, using a significance level of 0.1%.

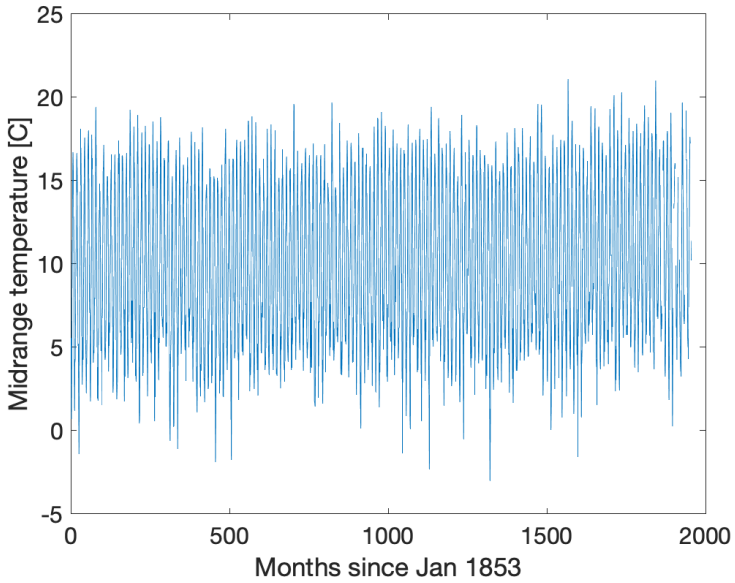
Download the file `oxford_data.txt` from Blackboard, place it in Matlab's current folder, which is displayed in the left column, and load it using `readmatrix`. The columns in the file are:

- Year
- Month
- Mean daily maximum temperature (tmax)
- Mean daily minimum temperature (tmin)
- Days of air frost (af)
- Total rainfall (rain)

(The original historical data is from https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data)

Solution

```
% load temperature data
oxford = readmatrix("oxford_data.txt", "Delimiter", '|', 'Range', 'C:D');

% midrange temperatures
temp = mean(oxford, 2);

% numer of months since January 1853
time = transpose(1:length(temp));

% plot midrange temperature time series
clf
plot(time, temp);
set(gca, 'FontSize', 16.0);
xlabel('Months since Jan 1853')
ylabel('Midrange temperature [C]')
```




```
% use Pearson's correlation coefficient as test statistic
cc = corrcoef(time, temp, 'Rows', 'complete');
observation = cc(1, 2)
```

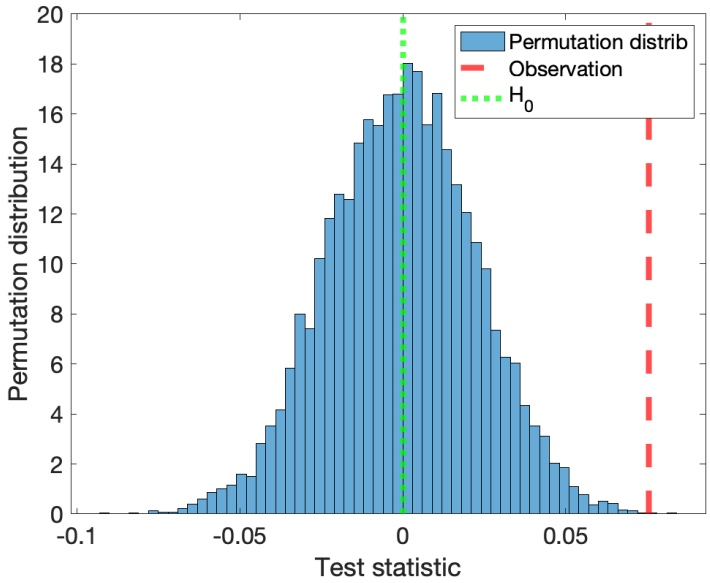
```
observation = 0.0755
```

```
% permutation test
% randomly shuffle the time variable
S = 10000;
tstat = zeros(1, S);
for i = 1:S
    time = randperm(length(temp));
    cc = corrcoef(time, temp, 'Rows','complete');
    tstat(i) = cc(1, 2);
end

% right-tailed p-value
pval = mean(tstat >= observation)
```

```
pval = 2.0000e-04
```

Since $p < \alpha = 0.001$ we reject the null hypothesis that temperature and time are not correlated at 0.1% significance level.

```
% plot
clf
histogram(tstat, 'normalization', 'pdf');
hold on
xline(observation, '--r', 'LineWidth', 4);
hold on
xline(0.0, ':g', 'LineWidth', 4);
legend('Permutation distrib', 'Observation', 'H_0')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Permutation distribution')
```



6. Constrained permutation test

We want to test the hypothesis that male and female students are equally likely to pass the summer exam. Last year's data is reported in the table below:

	Male	Female	
Pass	18	17	35
Fail	12	3	15
	30	20	50

1. Use a permutation test to determine if the probability to pass the exam does not depend on the student's gender, using a significance level of 5%.

Solution

```
% let's create the data reported in the table

% vector of student genders (20 females, 30 males)
% 1 female, 0 male
gender = [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

% total number of students
n = length(gender)
```

```
n = 50
```

```
% number of female students
n_females = sum(gender)
```

```
n_females = 20
```

```
% vector of exam outcomes (35 pass, 10 fail)
% 1 fail, 0 pass
result = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

% number of failures
n_fail = sum(result)
```

```
n_fail = 15
```

```
% number of females who failed the exam
observation = sum(gender .* result)
```

```
observation = 3
```

```
% permutation test
% randomly reassign the exam outcomes
S = 10000;
tstat = zeros(1, S);
for i = 1:S
    % number of females who failed the exam
    tstat(i) = sum(gender(randperm(n)) .* result);
end

% two-tailed p-value
```

```
2 * mean(tstat <= observation)
```

```
ans = 0.1102
```

Since $p > \alpha = 0.05$ we do not reject the null hypothesis that passing the exam does not depend on the student's gender at 5% significance level.

```
% plot
clf
histogram(tstat, 'normalization', 'pdf');
hold on
xline(observation, '--r', 'LineWidth', 4);
legend('Permutation distrib', 'Observation')
set(gca, 'FontSize', 16.0);
xlabel('Test statistic')
ylabel('Permutation distribution')
```