Final Report

Sogol Haddadi

Project Objectives

This project intends to answer the following questions using simulation:

1. Is the Kolmogorov Smirnov (KS) test an asymptotically exact test for testing the means of two distributions.
2. Is the Kolmogorov Smirnov (KS) test an asymptotically consistent test for testing the means of two distributions.

Kolmogorov Smirnov test (Brief Description)

The KS test is used to compare the underlying population of two samples. The null hypothesis is two samples are coming from the same populations. The test statistic is the maximum distance between two empirical CDF's and asymptotically follows a Kolmogorov distribution. The null hypothesis is rejected for the large values of the test statistic.

$H_0 : F_Y = F_X$

test statistic $= D = sup_y|\hat{F}_Y(y) - \hat{F}_X(y)|$

$\sqrt{mn/(m+n)}D \rightarrow^d K$

Methodology

To answer these questions, I simulated 1000 samples from populations with sample sizes of 10, 50, and 100. Then, I found the rejection rates. In this regard, a function `sample_anyDistribution()` was created and saved in the R directory. This function returns 1000 samples of size `n` using the built-in density functions in R. To perform the KS test and find the rejection rates, `ks_test_rejection()` function was developed. Also, I created another function called `dist_comb()` that creates a data frame of the two by two combinations of the distributions. There is more information about these functions in the `man` folder.

Here are the null and alternative hypothesis:

The null hypothesis is:

- mean of the first population = mean of the second population

The alternative hypothesis is:

- mean of the first population $\neq$ mean of the second population

Question 1

The distributions I have chosen to answer the first question are as follows. All the distributions have the same population means (0.5).

1. $Normal(\mu = 1/2, \sigma^2 = 1)$
2. $exponential(\lambda = 2)$
3. $beta(\alpha = 2, \beta = 2)$
4. $gamma(\alpha = 2, \beta = 1/4)$

Here is the result of my simulations:

rejection_rate_10	rejection_rate_50	rejection_rate_100	dist1	dist2
0.042	0.885	1	normal	exponential
0.089	0.999	1	normal	beta
0.063	0.951	1	normal	gamma
0.049	0.658	0.951	exponential	beta
0.015	0.218	0.425	exponential	gamma
0.021	0.199	0.366	beta	gamma

Since the null hypothesis is true (population means are equal), I would expect to reject the null hypothesis 5% of the times. The results show that the KS test is not asymptotically exact for testing the population means since as the sample size increased, the rejection rate did not get close to 5%.

Question 2

The distributions I have chosen to answer the second question are as follows:

1. $Normal(\mu = 1/2, \sigma^2 = 1)$,with the population mean of 1/2
2. $exponential(\lambda = 1)$,with the population mean of 1
3. $beta(\alpha = 2, \beta = 1)$,with the population mean of 2/3
4. $gamma(\alpha = 2, \beta = 1)$,with the population mean of 2

These populations have different means. The means are shown next to the populations above. If the KS test is asymptotically consistent, we would expect to reject the null hypothesis 100% of the times as sample size increases. Here are the results of the simulations:

rejection_rate_10	rejection_rate_50	rejection_rate_100	dist1	dist2
0.037	0.777	0.997	normal	exponential
0.095	0.998	1	normal	beta
0.353	1	1	normal	gamma
0.066	0.98	1	exponential	beta
0.215	0.984	1	exponential	gamma
0.816	1	1	beta	gamma

As the sample size increases, the rejection rate gets close to 1. Therefore, we can say that the KS test is asymptotically consistent for testing the population means.

Techniques I learned in the class and I used in my project:

To speed up the simulations, I included vectorization in my project. I tried to include parallelization but the functions ran slower. I created documentation for the functions I created. The documentation can be found in the `man` folder in my project repository.

Conclusion

Based on the results of my simulations, we can conclude:

- The KS test is not asymptotically exact for testing the population means of two samples
- The KS test is asymptotically consistent for testing the population means of two samples

Reference

- http://st551.cwick.co.nz/lecture/lecture_26/