HUDM6026 Homework_07

Chenguang Pan

Mar 17, 2023

Q1:

Do some research on the two-sample Kolmogorov-Smirnov (K-S) test for equality of distributions. Describe the null and alternative hypotheses and discuss how the test statistic is computed

MY SOLUTION:

[Part of my answer for this question refered to ChatGPT's responses, crosschecked with Wikipedia]

1.1 Two sample K-S test's concept

The two-sample Kolmogorov-Smirnov(K-S) test is a nonparametric test used to test whether two datasets are from the same population. K-S test works by comparing the cumulative distribution functions (CDF) of the two datasets.

The null hypothesis is that two datasets are drawn from the same population or that they have the same underlying probability distributions. The alternative hypothesis is that they are not from the same population or that they do not have the same underlying probability distributions. That is,

$$H_0 : F_1(x) = F_2(x) \quad \text{v.s.} \quad H_1 : F_1(x) \neq F_2(x)$$

, where $F_1(x)$ and $F_2(x)$ are the distributions of two sets of random variables X_1 and X_2.

1.2 How to compute the two-sample K-S test

The K-S test compares the largest vertical distance between the two empirical cumulative distribution functions(ECDF). This is also called the Kolmogorov-Smirnov statistic (D). The test statistic is defined as:

$$D = max|F_1(x) - F_2(x)|$$

, where $F_1(x)$ and $F_2(x)$ are the ECDF of two datasets being compared.
The maximum distance (i.e., the D) between the two CDFs is then compared to a critical value

$$c(\alpha) \cdot \sqrt{\frac{m+n}{m \cdot n}}$$

, where $c(\alpha) = \sqrt{-0.5ln(0.5\alpha)}$, depending on the sample size and significance level chosen.

If the test statistic is greater than the critical value, the null hypothesis is rejected, indicating that the two sets of data are not drawn from the same population.

Q2:

Install package MatchIt and load it. Then call data(lalonde). Examine the help on lalonde and describe the meaning of the treat and re78 variables.

MY SOLUTION:

```
> install.packages("MatchIt")
```

```
> # extract the basic information of the given dataset
> library(MatchIt)
> data(lalonde)
> head(lalonde)
     treat age educ   race married nodegree re74 re75       re78
NSW1     1  37   11  black       1        1    0    0  9930.0460
NSW2     1  22    9 hispan       0        1    0    0  3595.8940
NSW3     1  30   12  black       0        0    0    0 24909.4500
NSW4     1  27   11  black       0        1    0    0  7506.1460
NSW5     1  33    8  black       0        1    0    0   289.7899
NSW6     1  22    9  black       0        1    0    0  4056.4940
> dim(lalonde)
[1] 614   9
```

The help file shows that `lalonde` has one treated group from NSW with the size of 185 and one comparison group from PSID with the size of 429. The variable `treat` is the treatment assignment(1= treated, 0=control). The `re78` is income in 1978, in US dollars.

Q3:

Run the two-sample K-S test to test if participant income in 1978 is identically distributed across treatment group assignment or not...

MY SOLUTION:

```
> # subset the treatment and control groups
> x <- lalonde$re78[lalonde$treat ==1]
> y <- lalonde$re78[lalonde$treat ==0]
>
> # run the two sample k-s test
> ks.test(x, y)

    Asymptotic two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.098608, p-value = 0.1619
alternative hypothesis: two-sided
```

The test statistic D is .0986. Since the p-value is .162 and it is greater than the significant level .05, we fail to reject the null hypothesis. That is, there is no statistically significant difference between the treatment and comparison groups.

Q4:

Create a function that takes two arguments, x and y, each a vector of values, and outputs the value of the two-sample K-S statistic D

MY SOLUTION:

There is a function `ecdf()` built in R. I plan to use it directly and won't write the `ecdf` function from the scratch.

```
> ks_d <- function(x, y){
+   # get the maximize vertical distance between the two ecdfs
+   d <- max(abs(ecdf(x)(x)-ecdf(y)(x)))
+   return(d)
+ }
```

Q5:

Run an approximate permutation test with $B = 1000$ permutation replications to determine the estimated ASL for testing the null hypothesis that participant incomes are identically distributed across treatment arms. Use $a = 0.05$

MY SOLUTION:
The sample sizes for treatment and control groups are 185 and 429.

```
> data(lalonde)
> set.seed(666)
> # subset the treatment and control group
> treat_group <- lalonde$re78[lalonde$treat ==1]
> control_group <- lalonde$re78[lalonde$treat ==0]
> # get the observed the K-S test's D
> d_observed <- ks_d(treat_group, control_group)
> # get the sample sizes of treatment and control groups
> size_treat <- length(treat_group); size_control <- length(control_group)
> # define the repeating times
> B <- 1000
> # create a empty array to load the test statistic D from different permutations
> ks_ds <- rep(NA,B)
> # creat a variable to count how many times the re-sampling Ds is greater than the
> # observed one.
> d_bigger <- 0
>
> # randomly sample data from combined dataset in the sizes of 185 and 429
> for (i in 1:B) {
+   # randomly sample data from the whole re78 data in the size of 185+429
+   dat <- sample(lalonde$re78)
+   # split the random sample into 185 and 429 sample sizes
+   x <- dat[1: size_treat]
+   y <- dat[(size_treat+1): (size_treat+size_control)]
+   # plug the x and y into the D function above
+   d_temp <- ks_d(x,y)
+   # write the d_temp to the ks_ds
+   ks_ds[i] <- d_temp
+   # compare the d_temp with the observed KS test's D
```

```
+   if (d_temp >= d_observed){
+     d_bigger <- d_bigger + 1
+   }
+ }
> # get the p value
> p_value <- round(d_bigger/B,3)
> p_value
[1] 0.116
> print(paste0("The Ds from permuated data are ",
+              d_bigger,
+              " times greater than the observed D ",
+              round(d_observed,3),"."))
[1] "The Ds from permuated data are 116 times greater than the observed D 0.099."
> print(paste0("The P value is ",
+              p_value, "."))
[1] "The P value is 0.116."
```

The result shows that the re-sampling-based p value (.116) is greater than .05. We fail to reject the null hypothesis. This calculated p value is close to p value .162 obtained from theoretical null distribution in Question 3.

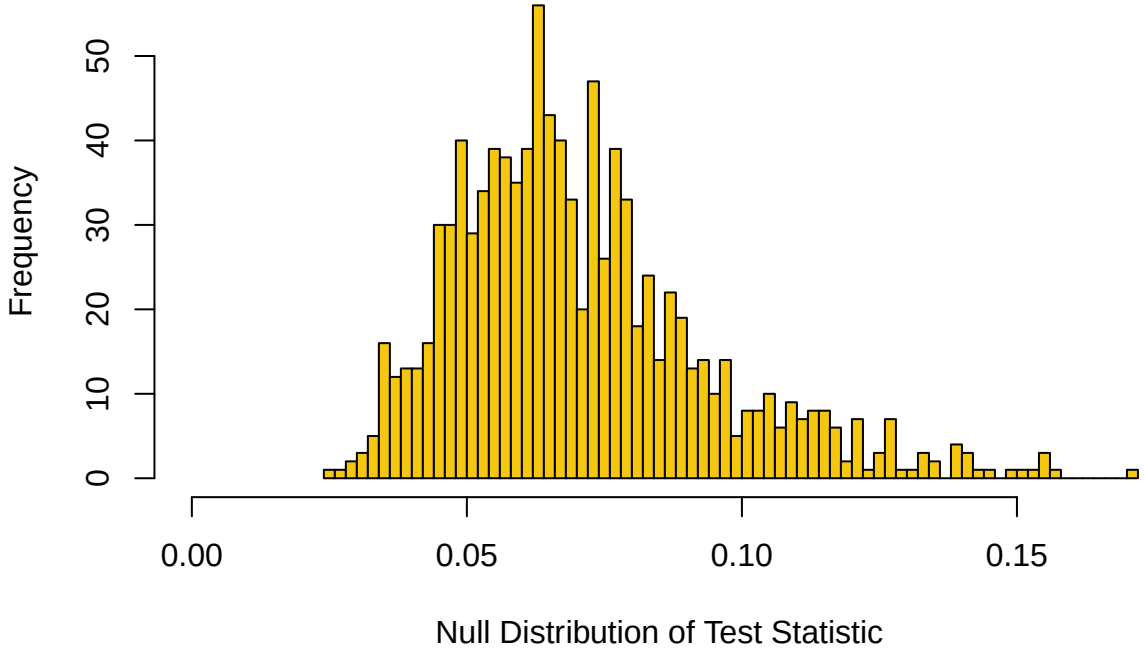
Q6:

Plot a histogram of the permutation distribution created by applying the K-S statistic to the $B = 1000$ permutation replications.

MY SOLUTION:

```
> hist(ks_ds, breaks =100, xlim=c(0,0.17),
+      main="Figure 1. A histogram of the re-sampling-based test statistics",
+      xlab="Null Distribution of Test Statistic", col=7)
```

Figure 1. A histogram of the re−sampling−based test statistics



Reference

Keller, B.(2023). *HUDM 6026 Computational Statistics: 07 Permutation Test*[Lecture notes].

In Wikipedia.(2023). *Kolmogorov–Smirnov test.* https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *An introduction to statistical learning.*

Rizzo, M. L. (2019). *Statistical computing with R.* Chapman and Hall/CRC.