Homework 1

Question 1:

The distribution I developed is similar to the chi-squared but calculate the sum of absolute values instead of squared values. The definition is : If xi~N(0, 1) , then y=sum(abs(xi))~Dong(n).

Parameters for this function are:

n	number of observations
df	degree of freedom

Table 1. Mean and Variance for Dong distribution

n	df	mean	variance
10000	10	7.948539	3.59024
10000	100	79.807136	36.53036
10000	1000	797.633324	352.85265
10000	10000	7979.188234	3572.22830

Table 1 shows the mean and variance for ten thousand observations with different df. Approximately, the expectation=0.8*df, variance=0.36*df.

Question 2:

Ten thousand observations were sampled with df=5, their properties were graphed in Fig 1. Mean values for these observations is 3.973651 and the variance is 1.832585.

Potentially, this distribution may be used where chi-squared distribution were previously used, such as test the independence for categorical data.

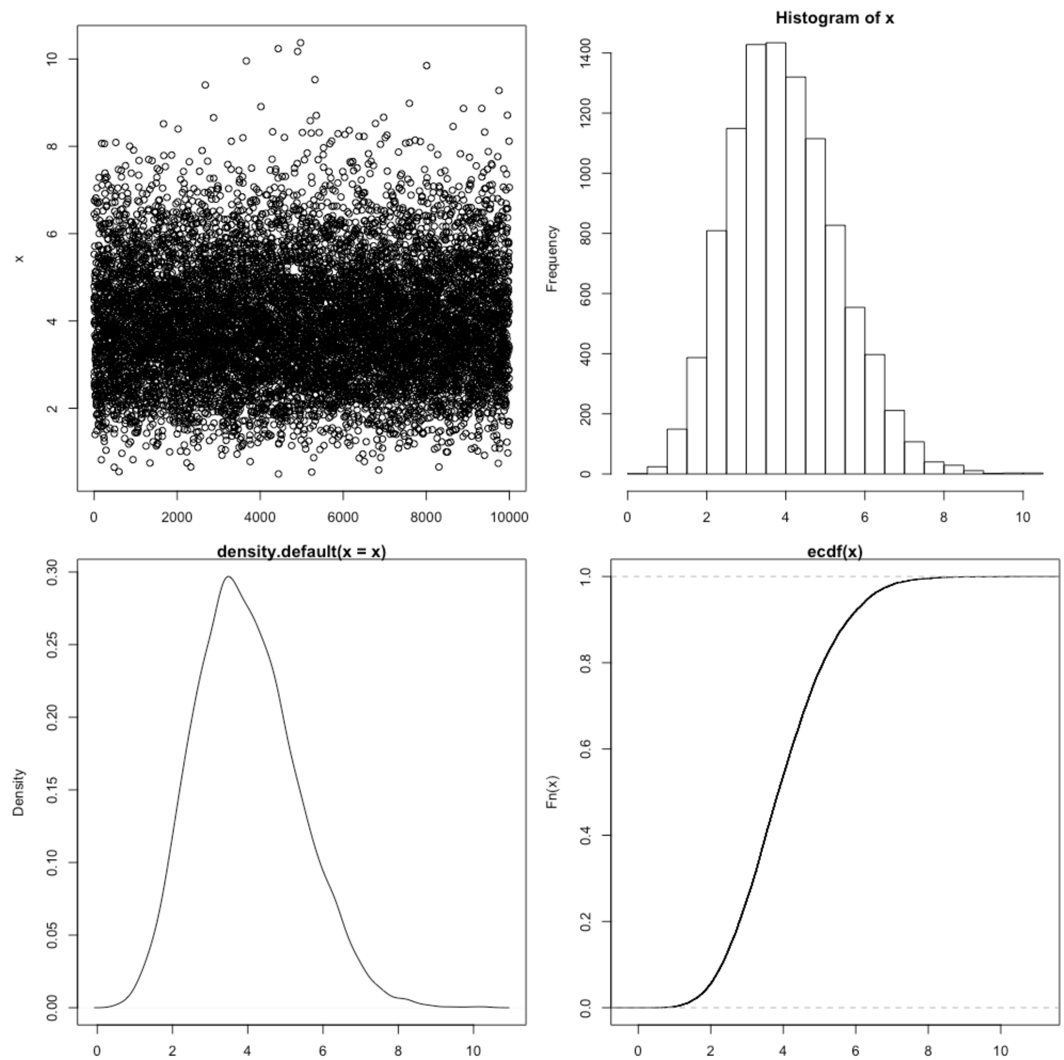

Fig 1. Ten thousand observations following the Dong distribution

Question 3:

The function “rf2(n, df1, df2)” (see the R code) was developed following these steps:

1) generate df1 and df2 observations following normal distribution by using rnorm function

2) calculate the sum of squares, then $U \sim X^2(df1)$, $V \sim X^2(df2)$

3) $F=(U/df1)/(V/df2) \sim F(df1, df2)$

A comparison showed it works almost the same with the default “df” function (Fig 2).

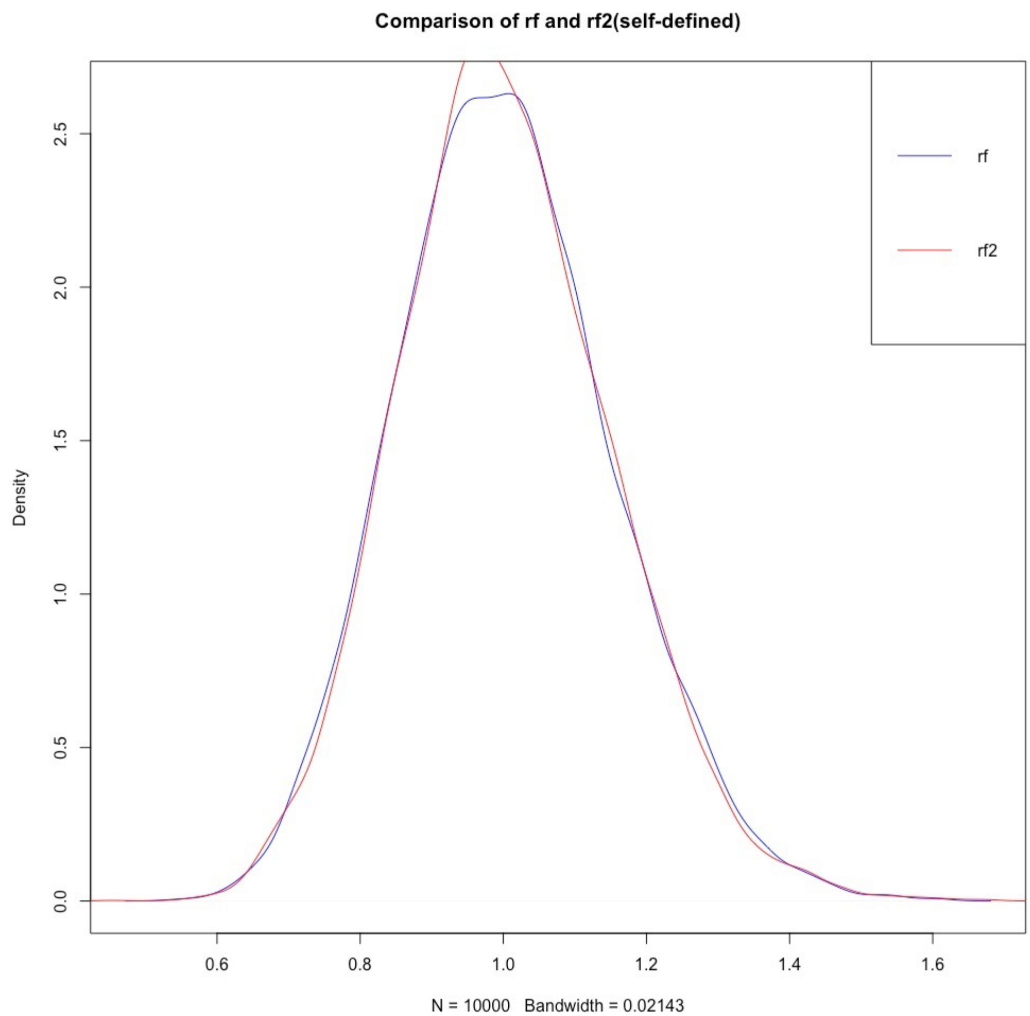

Fig 2. Comparison of rf (default) and rf2 (self-defined) functions

Question 4:

Set the df1=100, df2=1000, and 10, 100, 1000 and 100000 F distributed variables were sampled, H0: All the samples were from a distribution with mean of 1.002004.

Table 2. T-test for mean

n	mean	expected mean	t-test *p*	H0 (5% threshold)
10	1.0091091	1.002004	0.09595051	accept
100	0.9985443		0.93392793	accept
1000	1.0018997		0.56051767	accept
100000	1.0020603		0.21102045	accept

According to the t-test p values in table 2, under 5% of threshold, accept the hypothesis that all the 4 samples were from a distribution with mean of 1.002004.

Question 5:

Set the df1=100, df2=1000, and sampled 10, 100, 1000 and 100000 F distributed variables, H0: All the samples were from a distribution with variance of 0.02213665.

Table 3. Chi-squared test for variance

n	variance	expected variance	chi-squared test *p*	H0 (5% threshold)
10	0.02339999	2213665	0.39127282	accept
100	0.02844433		0.02956745	reject
1000	0.02286461		0.22906609	accept
100000	0.02223917		0.15022465	accept

According to the chi-squared test p values, under 5% of the threshold, accept the hypothesis that the 1st, 3rd and 4th samples with 10, 1000, 100000 variables were from a distribution with variance of 0.02213665, reject the hypothesis that the 2nd sample with 100 variables was from a distribution with variance of 0.02213665.