Presidency University
Non-Parametric Methods
STAT05C11

Spandan Ghoshal, Ritwick Mondal, Kalpesh Chatterjee, Niranjan Dey
Roll No. : STAT01, STAT10, STAT26, STAT29

December 16, 2020

Kruskal-Wallis Test

Introduction

Kruskal wallis is a non parametric test that can be used to determine whether two or more independent samples were selected from populations having the same distribution. Also this way it can be thought as an extension of the Mann-Whitney U test, which can be used for only two groups.

It is also known as Kruskal-Wallis H test since William Kruskal and W.Allen Wallis first published this method in the year 1952.

Important points

- Kruskal wallis test is equivalent to the one-way ANOVA.
- An extension of the Mann-Whitney U test.
- Sometimes we call it the one way ANOVA on ranks.
- The kruskal wallis test will tell us if there is a significant difference between groups.
- We use the sums of the ranks of the different samples to compare the distributions.
- A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample.

Assumptions

- Each sample must be randomly selected.
- The size of the each sample must be at least 5.
- Observations should be independent.
- Variables should be measured on an ordinal scale or a continuous scale.

K-W one way ANOVA test and multiple comparison

The Kruskal-Wallis test is the natural extension of the wilcoxon test for location with two independent samples to the situation of k mutually independent samples from continuous populations.The null hypothesis is that the k populations are same. But when we assume the location model this hypothesis can be written in terms of the respective location parameters as :-

$$\mathscr{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$
$$\mathscr{H}_1 : \text{At least two } \theta's \text{ differ}$$

To perform the test all $n_1 + n_2 + ...n_k = N$ observations are pooled into a single array and ranked from 1 to N.

Method

Since under $\mathscr{H}_0$ we have essentially a single sample of size N from the common population,combine the N observations into a single ordered sequence from smallest to largest and assign the ranks $1, 2, .., N$ to the sequence. If adjacent ranks are well distributed among the k samples, the total sum of ranks $\sum\limits_{i=1}^{N} i = \frac{N(N+1)}{2}$, would be divided proportionally according to sample size among the k samples and will be denoted by,

$$R_i = \sum_{j=1}^{n_i} r_{ij}$$

For the i th sample which contains n_i observations , under null hypothesis$\mathscr{H}_0$ the expected sum of ranks would be :-

$$E\left(R_i\right) = E\left(\sum_{j=1}^{n_i} r_{ij}\right) = \sum_{j=1}^{n_i} E\left(r_{ij}\right) = \sum_{j=1}^{n_i} \frac{N+1}{2} = \frac{n_i\left(N+1\right)}{2}$$

which can also be thought alternatively as the proportion of total rank sum for each sample of size n_i i.e,

$$\frac{n_i}{N}\frac{N(N+1)}{2} = \frac{n_i(N+1)}{2}$$

Since the deviation for each group from its expected rank sum i.e., $R_i - \frac{n_i(N+1)}{2}$ can be thought as a measure of deviation from the null assumption, a reasonable test statistic could be based on a function of the all these deviations. Since deviations in either direction indicate disparity between the samples and absolute ($|.|$) values are not particularly tractable mathematically, the sum of squares of these deviations can be employed as,

$$S = \sum_{i=1}^{k}\left[R_i - \frac{n_i(N+1)}{2}\right]^2$$

Hence, the null hypotheis should be rejected for large value of S.

Null Distribution of S (no tie case)

In order to determine the null probability distribution of S, we first consider all the possible arrangements of ranks $1, 2, \ldots, N$ into k groups of size n_i each. This can be done in $\frac{N!}{\prod_{i=1}^{k} n_i!}$. Then for each of these arrangements, we calculate the value of the S statistic and let us denote by $t(s)$ number of arrangements for which $S = s$, then the corresponding probability of S taking the value s is,

$$f(s) = \frac{t(s)}{\frac{N!}{\prod_{i=1}^{k} n_i!}} = t(s)\frac{\prod_{i=1}^{k} n_i!}{N!}$$

Drawbacks of S Statistic

- First of all the calculation for S becomes very tedious for even $n_i \geq 5$ as the number of such arrangements rapidly increase with increasing values of n_i's.
- Also there is no standard asymptotic distribution for S which can be used for large sample tests.
- S only consider the sum of square of deviations of R_i from its mean but it do not standarize the observations R_i.

Kruskal-Wallis Test Statistic

Due to all the drawbacks of the S statistic, a better statistic could be a weighted sum of squares of deviations with the reciprocals sample size used as weights,then the test will be more useful and significant.This test statistic,due to Kruskal and Wallis (Kruskal-Wallis H Statistic) is defined as :-

$$\begin{aligned}
H &= \frac{12}{N\left(N+1\right)}\sum_{i=1}^{k}\frac{1}{n_i}\left[R_i-\frac{n_i\left(N+1\right)}{2}\right]^2 \\
&= \frac{12}{N\left(N+1\right)}\sum_{i=1}^{k}\frac{1}{n_i}\left[n_i\overline{R}_i-\frac{n_i\left(N+1\right)}{2}\right]^2 \\
&= \frac{12}{N\left(N+1\right)}\sum_{i=1}^{k}n_i\left[\overline{R}_i-\frac{\left(N+1\right)}{2}\right]^2
\end{aligned}$$

Here by R_i, we denote the i^{th} average rank sum $\overline{R}_i = {}^{R_i}/_{n_i}, i = 1\left(1\right)k$.

H can also be written as $H = \frac{12}{N(N+1)}\sum\limits_{i=1}^{k}\frac{R_i^2}{n_i} - 3\left(N+1\right)$.

Mean and Variance of $\overline{R}_i$

Under null assumption, the k groups can be thought as SRSWOR samples of size n_i each from a $U\left\{1,2,\ldots,N\right\}$ population. Hence,

$$\begin{aligned}
\text{Population Mean :}\mu &= \frac{N+1}{2} \\
\text{Population Variance :}\sigma^2 &= \frac{N^2-1}{12}
\end{aligned}$$

Similarly, $\overline{R}_i$ can be thought as the sample mean of a SRSWOR sample of size n_i hence,

$$\begin{aligned}
E\left(\overline{R}_i\right) =&\mu = \frac{N+1}{2} \\
V\left(\overline{R}_i\right) =&\frac{\sigma^2}{n_i}\frac{N-n_i}{N-1} = \frac{N^2-1}{12n_i}\frac{N-n_i}{N-1} = \frac{\left(N+1\right)\left(N-n_i\right)}{12n_i}
\end{aligned}$$

Asymptotic Distribution of H

If n_i is large, the standarized random variable

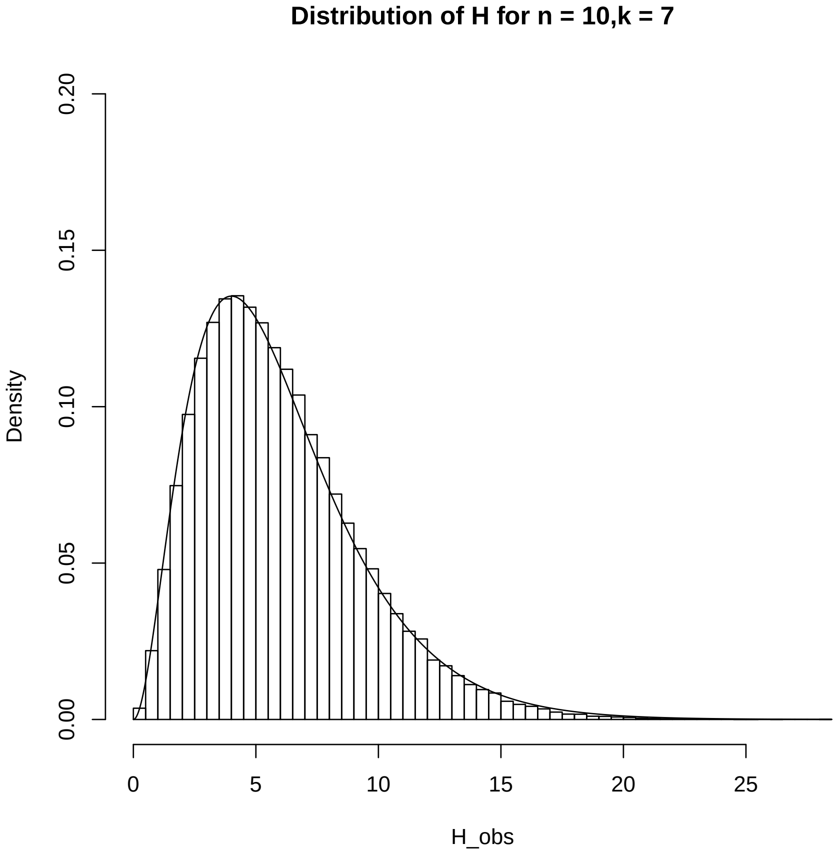
$$z_i = \frac{\overline{R}_i - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n_i}}} \overset{d}{\sim} N\left(0,1\right)$$

by Lindeberg-Levy CLT.

Since, there is a linear dependence between the quantities $\overline{R}_i$ as, the total of all rank sums $\sum\limits_{i=1}^{k} n_i\bar{R}_i = \frac{N(N+1)}{2}$, so all the k Z_i's can't be independently distributed (atmost $k-1$ of them can). So it can be shown if no n_i is very small then, under $\mathscr{H}_0$, $H = \sum\limits_{i=1}^{k} \frac{N-n_i}{N} Z_i^2$ is approximately distributed as a Chi-Squared Distribution with $k-1$ degrees of freedom $\left(\chi^2_{k-1}\right)$. Hence, we reject $\mathscr{H}_0$ at α level of significance if $H_{obs} \geq \chi^2_{\alpha;k-1}$.

Sampling Distribution of H

Using R we plot histogram of 100000 observed values of H for $n = 10, k = 7$:-



Tie Case

If ties to the extent t are present and are handled by the midrank method, the variance of the finite population is,

$$\sigma^2 = \frac{N^2-1}{12} - \frac{\sum t\left(t^2-1\right)}{12}$$

then the Kruskal-Wallis Statistic becomes,

$$
\begin{aligned}
H^{'} &= \sum_{i=1}^{k} \frac{N-n_i}{N} \left\{ \frac{\left[\overline{R}_i - \frac{N+1}{2}\right]^2}{\frac{(N+1)(N-n_i)}{12n_i} - \frac{N-n_i}{n_i(N-1)} \frac{\sum t(t^2-1)}{12}} \right\} \\
&= \sum_{i=1}^{k} \frac{\left[\overline{R}_i - \frac{N+1}{2}\right]^2}{\frac{N(N+1)}{12n_i}\left[1 - \frac{\sum t(t^2-1)}{N(N^2-1)}\right]} = \frac{H}{1 - \frac{\sum t(t^2-1)}{N(N^2-1)}}
\end{aligned}
$$

So, we just need to divide the original H statistic by the factor $1 - \frac{\sum t(t^2-1)}{N(N^2-1)}$.

Pairwise Comparison

If the null hypotheis is rejected, one may naturally want to compare different groups pairwise to check if their location parameters are equal or not.

From the asymptotic normal distribution of Z_i, we can easily make groupwise comparisons using the statistic $Z_{ij}, 1 \leq i < j \leq k$,

$$
Z_{ij} = \frac{R_i - R_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \overset{d}{\sim} N(0,1)
$$

We reject our null hypothesis $\mathscr{H}_0^{ij} : \theta_i = \theta_j$ at α^* level of significance if,

$$
\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k(k-1)}
$$

since, we are comparing $^{k(k-1)}/_{2}$ many pairs,

$$
\begin{aligned}
P\left(\mathscr{H}_0^{ij} \text{ accepted } \forall i,j\right) &= P\left(\bigcap_{1 \leq i < j \leq k} \mathscr{H}_0^{ij} \text{ accepted}\right) \\
&\geq \sum_{1 \leq i < j \leq k} P\left(\mathscr{H}_0^{ij} \text{ accepted}\right) - \frac{k(k-1)}{2} \\
&= \sum_{1 \leq i < j \leq k} (1 - \alpha^*) - \frac{k(k-1)}{2} = 1 - \alpha
\end{aligned}
$$

In words, the probability that all the statements are correct or all the pairs have equal location parameters, is atleast $1 - \alpha$. Hence, we take, $\alpha \geq 0.20$ because we are making such large number of statements.

Demonstration with Example

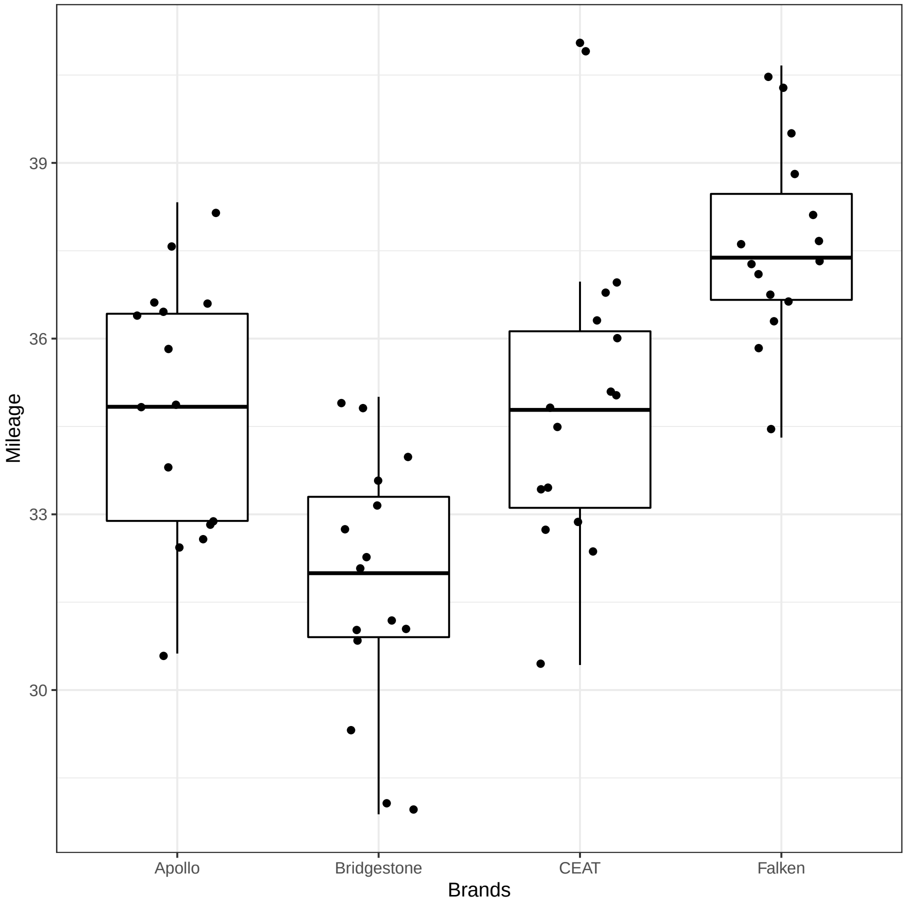
For demonstrating purpose, we choose a dataset consisting of Mileage of 60 car models from 4 different manufacturing companies. Namely, we have the four companies **Apollo, Bridgestone, CEAT** and **Falken**.

Here is a glimpse of the raw dataset we have in R :-

```
   Brands  Mileage
44   CEAT 32.16845
33   CEAT 33.41499
35   CEAT 36.97277
6  Apollo 35.91500
48 Falken 36.12400
2  Apollo 36.43500
55 Falken 37.38200
11 Apollo 36.43000
5  Apollo 36.30400
47 Falken 38.93700
```

So for exploratory data analysis, we first plot the mileage values for the four different companies(factor levels) :-

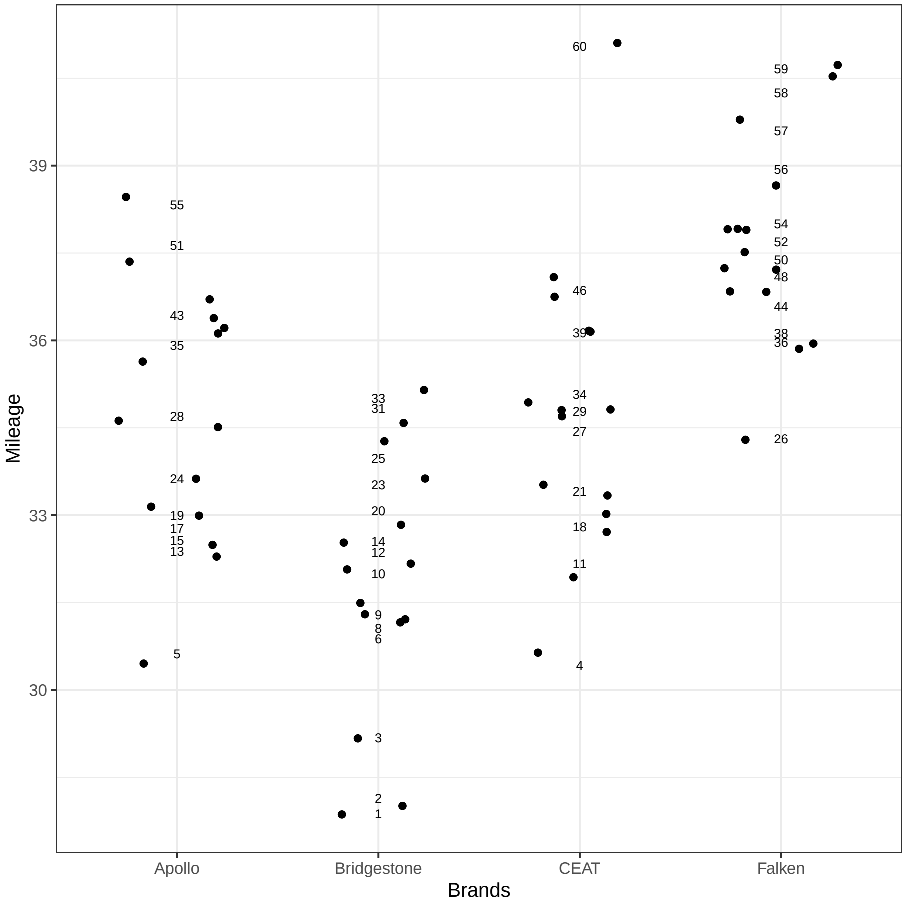
```
Loading required package:  ggplot2
Loading required package:  magrittr
```



Clearly, from the boxplot, we can easily suspect that the mean mileage for different companies are not equal. Only Apollo and CEAT has quiet close mean values. We shall verify our claims using the Kruskal-Wallis Statistic. For this we firstly, assign rank for each observed mileage in the data set and after assigning, the data would look like :-

```
        Brands  Mileage rank
3       Apollo 32.77700   17
42        CEAT 36.11675   37
50      Falken 36.58600   44
54      Falken 40.25200   58
43        CEAT 41.05000   60
37        CEAT 34.95412   32
52      Falken 36.73700   45
14      Apollo 30.62300    5
25 Bridgestone 30.88100    6
26 Bridgestone 28.14400    2
```

For a better understanding, we plot the mileages along with the assigned rank for each observation in the following plot :-



Now, we calculate the rank sum values for each level as :-

```
     Apollo Bridgestone          CEAT       Falken
        458             204            443          725
```

Which means that $R_1 = 458, R_2 = 204, R_3 = 443, R_4 = 725$ and $n_1 = n_2 = n_3 = n_4 = 15$ since we have 15 observations from each brand(company) so $N = \sum n_i = 60$.

Hence, the observed value of the Kruskal-Wallis H Statistic is :-

$$H_{obs} = \frac{12}{N\left(N+1\right)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\left(N+1\right) = \frac{12}{60 \times 61} 64883.6 - 3 \times 61 = 29.73311$$

and since, $H_{obs} = 29.73311 > \chi^2_{4-1;0.05} = 7.814728$, so we can reject the null hypothesis :-

$$\mathscr{H}_0 : \mu_{\text{Apollo}} = \mu_{\text{Bridgestone}} = \mu_{\text{CEAT}} = \mu_{\text{Falken}}$$

at 5% level of significance.

Also, using R, we can calculate the p-value of the test as :-

```
[1] 1.570466e-06
```

which is significant upto $\alpha = 0.001$ which shows that the difference between the means are highly significant.

Now, since the null hypotheis is rejected, our natural tendency would be to compare the pairwise mean values for different levels as :-

$$Z_{ij} = \frac{R_i - R_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

In R, this can be evaluated simultaneously as :-

```
                Apollo Bridgestone       CEAT    Falken
Apollo        0.000000    39.83038   2.352188 -41.86894
Bridgestone -39.830378     0.00000 -37.478190 -81.69932
CEAT         -2.352188    37.47819   0.000000 -44.22113
Falken       41.868941    81.69932  44.221128   0.00000
```

Now, we consider two groups significantly different if

$$\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k\,(k-1)}$$

here we have the $\left|Z_{ij(obs)}\right|$ values as :-

```
               Apollo Bridgestone      CEAT   Falken
Apollo       0.000000    39.83038  2.352188 41.86894
Bridgestone 39.830378     0.00000 37.478190 81.69932
CEAT         2.352188    37.47819  0.000000 44.22113
Falken      41.868941    81.69932 44.221128  0.00000
```

and the cut off point τ_{α^*} as :-

```
[1] 2.638257
```

So, now we compare if the $\left|Z_{ij(obs)}\right|$ values exceed τ_{α^*} and thus get the following TRUE FALSE matrix :-

```
            Apollo Bridgestone  CEAT Falken
Apollo       FALSE        TRUE FALSE   TRUE
Bridgestone   TRUE       FALSE  TRUE   TRUE
CEAT         FALSE        TRUE FALSE   TRUE
Falken        TRUE        TRUE  TRUE  FALSE
```

Since, only for the group $(\text{Apollo}, \text{CEAT})$ the outcome is FALSE hence their mean differences are not significant at $\alpha = 5\%$ and we can conclude that all other groups have significantly different mean mileage values and this can also be verified from the boxplot given before.

Approximations to the exact distribution of the kruskal-wallis test statistic

1. **(Wallace Approximation)** Given by **Wallace (1959)**, this approximation is very similar to the F statistic we use in ordinary analysis of variance that can be written by :-

   $$F = \frac{H/_{k-1}}{(N-H-1)/_{N-k}} = \frac{(N-k)\,H}{(k-1)\,(N-H-1)}$$

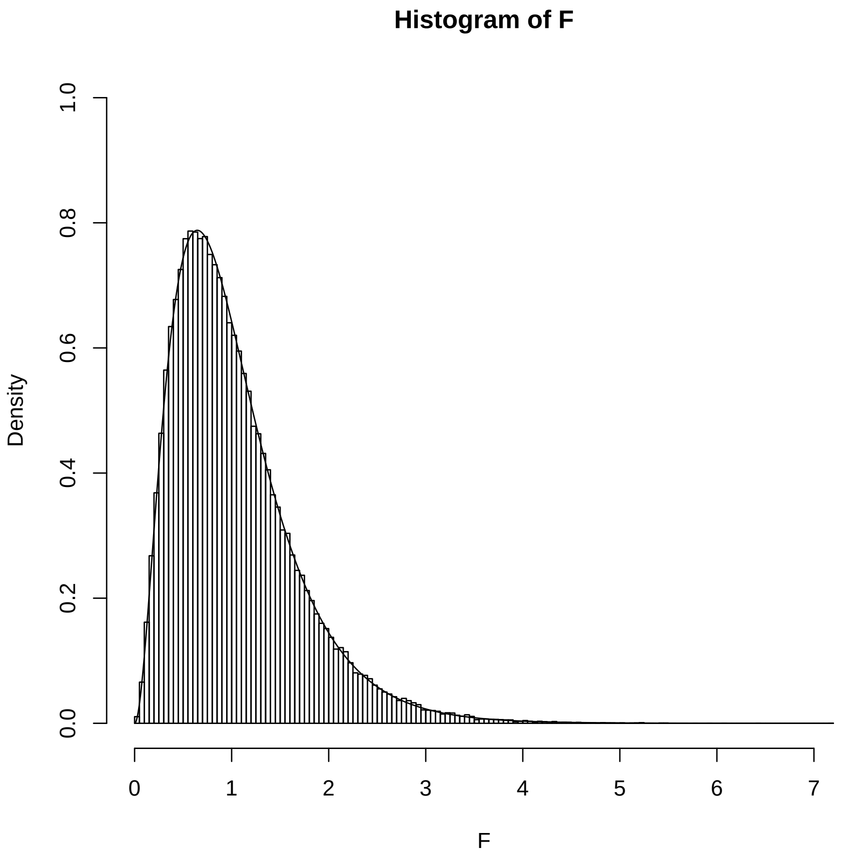
   which approximately follows a $F_{k-1,N-k}$ distribution where H is the ordinary Kruskal-Wallis Statistic.

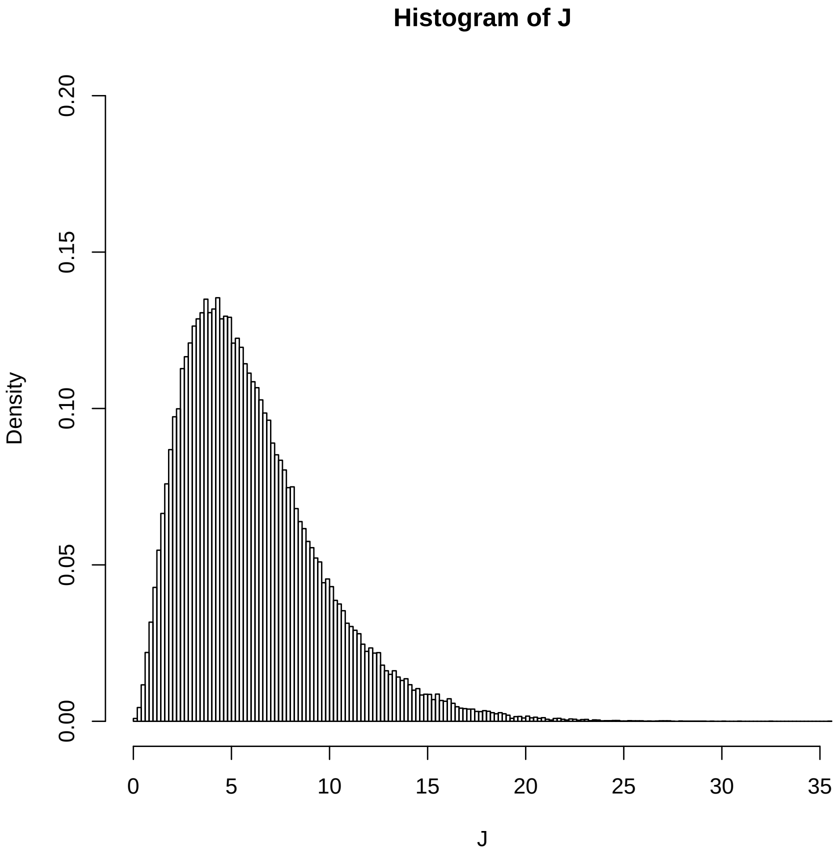
2. **(Iman Approximation)** This interesting approximation is based on techniques given by **Iman (1974,1976)** where a test statistic is formed by the linear combination of the χ^2 and F approximations already stated as :-

   $$J = \frac{(k-1)\,F + H}{2} = \frac{H}{2}\left(\frac{N-k}{N-H-1} + 1\right)$$

   The approximate critical values are given by,

   $$J_\alpha \approx \frac{(k-1)\,F_{k-1,N-k;\alpha} + \chi^2_{k-1;\alpha}}{2}.$$


Histogram of F
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
F


Histogram of J
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
35
J