An Overview of The Kruskal Wallis Test

Spandan Ghoshal Ritwick Mondal
Kalpesh Chatterjee Niranjan Dey

January 13, 2021

Origins of the test

- The **Kruskal–Wallis test** by ranks or **Kruskal–Wallis H test** named after **William Kruskal** and **W. Allen Wallis**, or **one-way ANOVA on ranks** is a non-parametric method for testing whether samples originate from the same distribution.

Origins of the test

- The **Kruskal–Wallis test** by ranks or **Kruskal–Wallis H test** named after **William Kruskal** and **W. Allen Wallis**, or **one-way ANOVA on ranks** is a non-parametric method for testing whether samples originate from the same distribution.
- It is used for comparing two or more independent samples of equal or different sample sizes. It extends the **Mann–Whitney U test**, which is used for comparing **only two** groups. The **parametric equivalent** of the Kruskal–Wallis test is the **one-way analysis of variance (ANOVA)**.

What can we conclude using this test?

- A significant Kruskal–Wallis test indicates that **at least one sample stochastically dominates one other sample**.

What can we conclude using this test?

- A significant Kruskal–Wallis test indicates that **at least one sample stochastically dominates one other sample**.
- However the test **does not identify where** this **stochastic dominance** occurs or for **how many pairs of groups** stochastic dominance obtains.

What can we conclude using this test?

- A significant Kruskal–Wallis test indicates that **at least one sample stochastically dominates one other sample**.
- However the test **does not identify where** this **stochastic dominance** occurs or for **how many pairs of groups** stochastic dominance obtains.
- For this we can compare pairwise groups using **pairwise Mann–Whitney** or using asymptotic distributions.

Basic Assumptions of the test

- Since it is a nonparametric method, the Kruskal–Wallis test does not assume a **normal distribution** of the **residuals**, unlike the analogous one-way analysis of variance.

Basic Assumptions of the test

- Since it is a nonparametric method, the Kruskal–Wallis test does not assume a **normal distribution** of the **residuals**, unlike the analogous one-way analysis of variance.
- Here we only assume of an **identically shaped** and **scaled** distribution **for all groups**, **except** for any **difference** in **medians**.

Hypothesis for the Test

- Here we consider k mutually independent samples with n_i observations in each sample $i = 1, \ldots, k$ from continuous populations with the assumption that they are **identically shaped** and **scaled**. Let the median of the i^{th} sample be $\theta_i, i = 1, \ldots, k$.

Hypothesis for the Test

- Here we consider k mutually independent samples with n_i observations in each sample $i = 1, \ldots, k$ from continuous populations with the assumption that they are **identically shaped** and **scaled**. Let the median of the i^{th} sample be $\theta_i, i = 1, \ldots, k$.
- Then the null hypothesis will be all the medians for individual groups are equal which we can write as :-

$$\mathscr{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$

Hypothesis for the Test

- Here we consider k mutually independent samples with n_i observations in each sample $i = 1, \ldots, k$ from continuous populations with the assumption that they are **identically shaped** and **scaled**. Let the median of the i^{th} sample be $\theta_i, i = 1, \ldots, k$.
- Then the null hypothesis will be all the medians for individual groups are equal which we can write as :-

$$\mathscr{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$

- And naturally, the alternate hypothesis will be $\mathscr{H}_1$: At least two $\theta_i's$ differ.

Testing Procedure

- Since under $\mathscr{H}_0$ we have essentially a single sample of size $N = \sum n_i$ from the common population, combine the N observations into a single ordered sequence from smallest to largest and assign the ranks $1, 2, .., N$.

Testing Procedure

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

Testing Procedure

- Since under $\mathscr{H}_0$ we have essentially a single sample of size $N = \sum n_i$ from the common population, combine the N observations into a single ordered sequence from smallest to largest and assign the ranks $1, 2, .., N$.
- If adjacent ranks are well distributed among the k samples, the total sum of ranks $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$, would be divided proportionally according to sample size.
- Let us denote the rank of the j^{th} observation from the i^{th} sample as R_{ij}.

Testing Procedure

- Let us denote the corresponding rank sums for each sample by R_i.

Testing Procedure

- Let us denote the corresponding rank sums for each sample by R_i.
- Then $R_i = \sum_{j=1}^{n_i} R_{ij}$ and $E\left(R_i\right) = E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E\left(r_{ij}\right)$ since each of the ranks $R_{ij} \sim U\left\{1, \ldots, N\right\}$ hence, $E\left(R_{ij}\right) = \frac{N+1}{2}$ finally, $E\left(R_i\right) = \frac{n_i(N+1)}{2}$.

Testing Procedure

- Let us denote the corresponding rank sums for each sample by R_i.
- Then $R_i = \sum_{j=1}^{n_i} R_{ij}$ and $E(R_i) = E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E(r_{ij})$ since each of the ranks $R_{ij} \sim U\{1, \ldots, N\}$ hence, $E(R_{ij}) = \frac{N+1}{2}$ finally, $E(R_i) = \frac{n_i(N+1)}{2}$.
- This can also be thought as $E(R_i) = \left(\frac{n_i}{N}\right) \times \frac{N(N+1)}{2}$.

Testing Procedure

- Let us denote the corresponding rank sums for each sample by R_i.
- Then $R_i = \sum_{j=1}^{n_i} R_{ij}$ and $E\left(R_i\right) = E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E\left(r_{ij}\right)$ since each of the ranks $R_{ij} \sim U\left\{1, \ldots, N\right\}$ hence, $E\left(R_{ij}\right) = \frac{N+1}{2}$ finally, $E\left(R_i\right) = \frac{n_i(N+1)}{2}$.
- This can also be thought as $E\left(R_i\right) = \left(\frac{n_i}{N}\right) \times \frac{N(N+1)}{2}$.
- The deviation for each observed group rank sum from its expected value i.e., $R_i - \frac{n_i(N+1)}{2}$ can be thought as a measure of deviation from the null assumption.

Testing Procedure

- Let us denote the corresponding rank sums for each sample by R_i.
- Then $R_i = \sum_{j=1}^{n_i} R_{ij}$ and $E\left(R_i\right) = E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E\left(r_{ij}\right)$ since each of the ranks $R_{ij} \sim U\left\{1, \ldots, N\right\}$ hence, $E\left(R_{ij}\right) = \frac{N+1}{2}$ finally, $E\left(R_i\right) = \frac{n_i(N+1)}{2}$.
- This can also be thought as $E\left(R_i\right) = \left(\frac{n_i}{N}\right) \times \frac{N(N+1)}{2}$.
- The deviation for each observed group rank sum from its expected value i.e., $R_i - \frac{n_i(N+1)}{2}$ can be thought as a measure of deviation from the null assumption.
- Hence, a reasonable test statistic could be based on a function of the all these deviations.

The S statistic

- Since deviations in either direction indicate disparity between the samples and absolute $(|.|)$ values are not mathematically friendly, the sum of squares of these deviations can be a good choice for constructing the test statistic.

The S statistic

- Since deviations in either direction indicate disparity between the samples and absolute $(|.|)$ values are not mathematically friendly, the sum of squares of these deviations can be a good choice for constructing the test statistic.
- So we construct the statisitc $S = \sum_{i=1}^{k} \left[R_i - \frac{n_i(N+1)}{2} \right]^2$.

The S statistic

- Since deviations in either direction indicate disparity between the samples and absolute ($|.|$) values are not mathematically friendly, the sum of squares of these deviations can be a good choice for constructing the test statistic.
- So we construct the statisitc $S = \sum_{i=1}^{k} \left[R_i - \frac{n_i(N+1)}{2}\right]^2$.
- Hence, the null hypotheis $\mathscr{H}_0$ 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 $\frac{N!}{\prod_{i=1}^{k} n_i!}$.

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 $S = s$.

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 $S = s$.
- Finally, we can write, $P[S = s] = t(s) \frac{\prod_{i=1}^{k} n_i!}{N!}$.

Drawbacks of S Statistic

- First of all the calculation for exact distribution of 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.

Drawbacks of S Statistic

- First of all the calculation for exact distribution of 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.
- Here is a table of no of cases for different values of n_i's :-

n	${(n_1+n_2+n_3)!}/{n_1!n_2!n_3!}$	n	${(n_1+n_2+n_3)!}/{n_1!n_2!n_3!}$
2	90	9	227, 873, 431, 500
3	1, 680	10	5, 550, 996, 791, 340
4	34, 650	11	136, 526, 995, 463, 040
5	756, 756	12	3, 384, 731, 762, 521, 200
6	17, 153, 136	13	84, 478, 098, 072, 866, 400
7	399, 072, 960	14	2, 120, 572, 665, 910, 728, 000
8	9, 465, 511, 770	15	53, 494, 979, 785, 374, 631, 680

Drawbacks of S Statistic

- First of all the calculation for exact distribution of 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.
- Here is a table of no of cases for different values of n_i's :-

n	${}^{(n_1+n_2+n_3)!}\!/\!{}_{n_1!n_2!n_3!}$	n	${}^{(n_1+n_2+n_3)!}\!/\!{}_{n_1!n_2!n_3!}$
2	90	9	227,873,431,500
3	1,680	10	5,550,996,791,340
4	34,650	11	136,526,995,463,040
5	756,756	12	3,384,731,762,521,200
6	17,153,136	13	84,478,098,072,866,400
7	399,072,960	14	2,120,572,665,910,728,000
8	9,465,511,770	15	53,494,979,785,374,631,680

- Also there is no standard asymptotic distribution for S which can be used for large sample tests.

Drawbacks of S Statistic

- First of all the calculation for exact distribution of 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.
- Here is a table of no of cases for different values of n_i's :-

n	$(n_1+n_2+n_3)!/n_1!n_2!n_3!$	n	$(n_1+n_2+n_3)!/n_1!n_2!n_3!$
2	90	9	227,873,431,500
3	1,680	10	5,550,996,791,340
4	34,650	11	136,526,995,463,040
5	756,756	12	3,384,731,762,521,200
6	17,153,136	13	84,478,098,072,866,400
7	399,072,960	14	2,120,572,665,910,728,000
8	9,465,511,770	15	53,494,979,785,374,631,680

- Also there is no standard asymptotic distribution for S which can be used for large sample tests.
- Lasltly, 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 H Statistic

- Due to all the drawbacks of the S statistic, William H. Kruskal & W. Allen Wallis (1952) proposed the following statistic for testing $\mathscr{H}_0$:-

$$H = \frac{12}{N\left(N+1\right)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3\left(N+1\right)$$

Kruskal-Wallis H Statistic

- Due to all the drawbacks of the S statistic, William H. Kruskal & W. Allen Wallis (1952) proposed the following statistic for testing $\mathscr{H}_0$:-

$$H = \frac{12}{N\left(N+1\right)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\left(N+1\right)$$

- Here,

$$\begin{aligned}
k &= \text{the number of samples} \\
n_i &= \text{the number of observations in the } i^{th}\text{sample} \\
N &= \sum_i n_i \text{ total number of observations in all samples} \\
R_i &= \text{ the sum of the ranks of the } i^{th} \text{ sample}
\end{aligned}$$

Different Representations of H

1. Firstly, the conventional H described before $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(N+1)$ which is easiest for computation purposes.

Different Representations of H

1. Firstly, the conventional H described before $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(N+1)$ which is easiest for computation purposes.
2. Secondly, $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2}\right]^2 = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left[\overline{R}_i - \frac{(N+1)}{2}\right]^2$ which is useful in further discussions since it uses the average rank sums $\overline{R}_i$.

Different Representations of H

1. Firstly, the conventional H described before $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(N+1)$ which is easiest for computation purposes.
2. Secondly, $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2}\right]^2 = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left[\overline{R}_i - \frac{(N+1)}{2}\right]^2$ which is useful in further discussions since it uses the average rank sums $\overline{R}_i$.
3. For understanding the similarity of this test with the conventional ANOVA test we write H in the following form :-

$$H = (N-1)\,\frac{\sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{R}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}\right)^2}$$

where $\overline{R} = \frac{1}{N} \sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij} = \frac{N+1}{2}$.

Kruskal-Wallis H Statistic

- For understanding the nature of H, a better formulation would be :-

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2}{\left(N^2-1\right)/12}$$

where $\overline{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$ is the mean of n_i ranks in the i^{th} sample.

Kruskal-Wallis H Statistic

- For understanding the nature of H, a better formulation would be :-

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2}{\left(N^2-1\right)/12}$$

  where $\overline{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$ is the mean of n_i ranks in the i^{th} sample.

- If we ignore the factor $\frac{N-1}{N}$ and note that $\frac{1}{2}(N+1)$ is the mean and $\frac{1}{12}\left(N^2-1\right)$ is the variance of the uniform distribution over the first N integers, we see that H is essentially a sum of squared standardized deviations of random variables from their population mean.

Kruskal-Wallis H Statistic

- For understanding the nature of H, a better formulation would be :-

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2}{\left(N^2-1 \right) / 12}$$

where $\overline{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$ is the mean of n_i ranks in the i^{th} sample.

- If we ignore the factor $\frac{N-1}{N}$ and note that $\frac{1}{2}(N+1)$ is the mean and $\frac{1}{12}\left(N^2-1\right)$ is the variance of the uniform distribution over the first N integers, we see that H is essentially a sum of squared standardized deviations of random variables from their population mean.
- In this respect H is similar to a χ^2 variate which is defined as a sum of square of standardized normal variates, subject to certain conditions on the relations among the terms of the sum.

Kruskal-Wallis H Statistic

- For understanding the nature of H, a better formulation would be :-

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2}{\left(N^2-1\right)/12}$$

where $\overline{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$ is the mean of n_i ranks in the i^{th} sample.

- If we ignore the factor $\frac{N-1}{N}$ and note that $\frac{1}{2}(N+1)$ is the mean and $\frac{1}{12}\left(N^2-1\right)$ is the variance of the uniform distribution over the first N integers, we see that H is essentially a sum of squared standardized deviations of random variables from their population mean.
- In this respect H is similar to a χ^2 variate which is defined as a sum of square of standardized normal variates, subject to certain conditions on the relations among the terms of the sum.
- If the n_i's are not too small, the $\overline{R}_i$ jointly will be approximately normally distributed and the relations among them will meet the χ^2 conditions.

Kruskal-Wallis H Statistic

- For understanding the nature of H, a better formulation would be :-

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2}{\left(N^2-1\right)/12}$$

where $\overline{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$ is the mean of n_i ranks in the i^{th} sample.
- If we ignore the factor $\frac{N-1}{N}$ and note that $\frac{1}{2}\left(N+1\right)$ is the mean and $\frac{1}{12}\left(N^2-1\right)$ is the variance of the uniform distribution over the first N integers, we see that H is essentially a sum of squared standardized deviations of random variables from their population mean.
- In this respect H is similar to a χ^2 variate which is defined as a sum of square of standardized normal variates, subject to certain conditions on the relations among the terms of the sum.
- If the n_i's are not too small, the $\overline{R}_i$ jointly will be approximately normally distributed and the relations among them will meet the χ^2 conditions.
- We further investigate approximate distribution of H with more rigourous mathematical arguments.

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 $U\{1, 2, \ldots, N\}$ population.

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 $U\{1, 2, \ldots, N\}$ population.
- Hence,

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

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 $U\{1, 2, \ldots, N\}$ population.
- Hence,

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

- And the mean rank of the i^{th} sample can be thought as a SRSWOR sample of size n_i hence,

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

Asymptotic Distribution of $\overline{R}_i$

- If n_i is large, the standarized random variables,

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

by Lindeberg-Levy CLT.

Asymptotic Distribution of $\overline{R}_i$

- If n_i is large, the standarized random variables,

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

  by Lindeberg-Levy CLT.

- Hence, we can say,

$$Z_i^2 = \frac{12n_i}{\left(N+1\right)\left(N-n_i\right)}\left(\overline{R}_i - \frac{N+1}{2}\right)^2 \overset{d}{\sim} \chi_1^2$$

Asymptotic Distribution of H

- Since there is a linear dependence between the quantities $\overline{R}_i$ as, the sum of all ranks is

$$\sum_{i=1}^{k} n_i \overline{R}_i = \sum_{i=1}^{k} R_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} r_{ij} = \frac{N(N+1)}{2}$$

so all the k variates $Z_i, i = 1, \ldots, k$ can't be independent (Any $k-1$ of them are independent.)

Asymptotic Distribution of H

- Since there is a linear dependence between the quantities $\overline{R}_i$ as, the sum of all ranks is

$$\sum_{i=1}^{k} n_i \overline{R}_i = \sum_{i=1}^{k} R_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} r_{ij} = \frac{N(N+1)}{2}$$

so all the k variates $Z_i, i = 1, \ldots, k$ can't be independent (Any $k-1$ of them are independent.)

- Kruskal (1952) showed that under $\mathscr{H}_0$, if no n_i is very small, the random variable :-

$$H = \sum_{i=1}^{k} \frac{N - n_i}{N} Z_i^2 = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2$$

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 $H_{obs} \geq \chi^2_{\alpha;k-1}$.

Tie Case

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

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

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,

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

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

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.

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 $Z_{ij}, 1 \leq i < j \leq k$,

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

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 $Z_{ij}, 1 \leq i < j \leq k$,

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

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

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

Pairwise Comparison

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

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

Pairwise Comparison

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

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

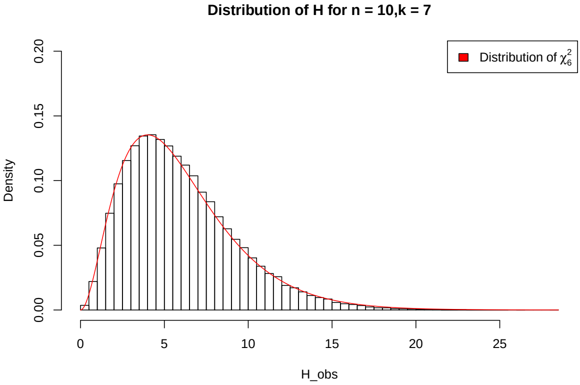
- In words, the probability that all the statements are correct or all the pairs have equal location parameters, is atleast $1-\alpha$. Hence, we take, $\alpha \ge 0.20$ because we are making such large number of statements.

Sampling Distribution of H using Simulation

- Using R we plot the approximate sampling distribution (histogram) of 10^5 realized values of H for $n_i = 10 \, \forall \, i$ and $k = 7$.

Sampling Distribution of H using Simulation

- Using R we plot the approximate sampling distribution (histogram) of 10^5 realized values of H for $n_i = 10 \, \forall \, i$ and $k = 7$.
- We can see how accurately the asymptotic distribution fits the actual density function of χ^2_6:-



Demonstration with Example

- For demonstration purposes, 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**.

Demonstration with Example

- For demonstration purposes, 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 :-

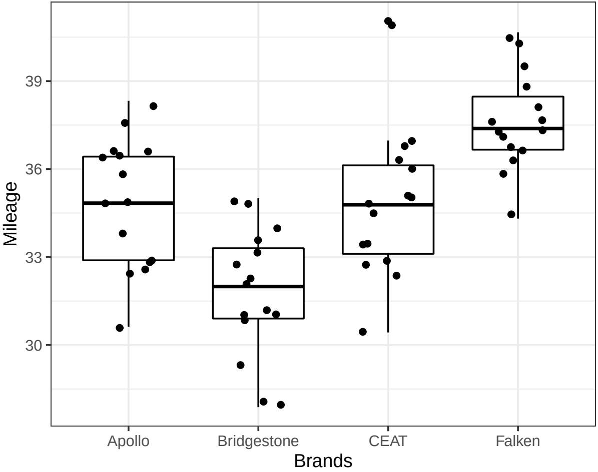
```
   Brands  Mileage
44   CEAT 32.16845
33   CEAT 33.41499
35   CEAT 36.97277
6  Apollo 35.91500
48 Falken 36.12400
2  Apollo 36.43500
55 Falken 37.38200
11 Apollo 36.43000
5  Apollo 36.30400
47 Falken 38.93700
```

Demonstration with Example

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

Demonstration with Example

- So for exploratory data analysis, we first plot the mileage values for the four different companies(factor levels) :-
- Since here we have one factor variable with four levels (brands) hence we use boxplot for demonstration :-



Demonstration with Example

- Clearly, from the boxplot, we can easily suspect that the mean mileage for different companies are not equal. Only Apollo and CEAT seem to have "close" median values.

Demonstration with Example

- Clearly, from the boxplot, we can easily suspect that the mean mileage for different companies are not equal. Only Apollo and CEAT seem to have “close” median values.
- We shall verify our claims using the Kruskal-Wallis testing procedures stated so far.

Demonstration with Example

- For this we firstly, assign rank for each observed mileage in the data set.

Demonstration with Example

- For this we firstly, assign rank for each observed mileage in the data set.
- After assigning, the data would look like :-

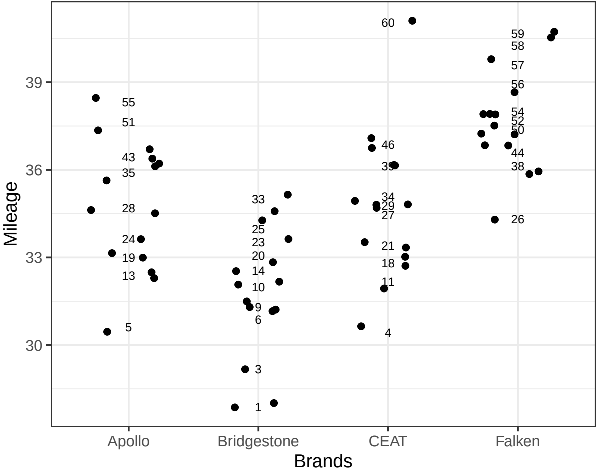
```
         Brands  Mileage rank
3        Apollo 32.77700   17
42         CEAT 36.11675   37
50       Falken 36.58600   44
54       Falken 40.25200   58
43         CEAT 41.05000   60
37         CEAT 34.95412   32
52       Falken 36.73700   45
14       Apollo 30.62300    5
25 Bridgestone 30.88100   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

Demonstration with Example

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

Demonstration with Example

- For a better understanding, we plot the mileages along with the assigned rank for each observation in the following plot :-
- Here we use the jittered plot for different car brands along different vertical axes and also corresponding rank in the pooled sample :-



Demonstration with Example

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

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

Demonstration with Example

- 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 $N = \sum n_i = 60$.

Demonstration with Example

- 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 :-

$$H_{obs} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$
$$= \frac{12}{60 \times 61} 64883 \cdot 6 - 3 \times 61 = 29 \cdot 73311$$

Demonstration with Example

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

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

at 5% level of significance.

Demonstration with Example

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

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

  at 5% level of significance.
- Also, using R, we can calculate the p-value of the test as :-

$$P_{\mathscr{H}_0} \left(H \geq H_{obs} \right)$$

Demonstration with Example

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

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

  at 5% level of significance.
- Also, using R, we can calculate the p-value of the test as :-

$$P_{\mathscr{H}_0}\left(H \geq H_{obs}\right)$$

- Which in this case comes out to be :-

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

Demonstration with Example

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

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

at 5% level of significance.

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

$$P_{\mathscr{H}_0}\left(H \geq H_{obs}\right)$$

- Which in this case comes out to be :-

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

- which is significant upto $\alpha = 0.001$ or in other words the difference between the means are highly significant.

Demonstration with Example

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

$$Z_{ij} = \frac{\overline{R}_i - \overline{R}_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Demonstration with Example

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

$$Z_{ij} = \frac{\overline{R}_i - \overline{R}_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- Here we get the average rank sums as
$\overline{R}_1 = 30 \cdot 533, \overline{R}_2 = 13 \cdot 6, \overline{R}_3 = 29 \cdot 533, \overline{R}_4 = 48 \cdot 33$

Demonstration with Example

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

$$Z_{ij} = \frac{\overline{R}_i - \overline{R}_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- Here we get the average rank sums as $\overline{R}_1 = 30 \cdot 533, \overline{R}_2 = 13 \cdot 6, \overline{R}_3 = 29 \cdot 533, \overline{R}_4 = 48 \cdot 33$
- In R, all these pairwise differences can be evaluated simultaneously as :-

```
                Apollo Bridgestone       CEAT    Falken
Apollo       0.0000000    2.655359  0.1568125 -2.791263
Bridgestone -2.6553585    0.000000 -2.4985460 -5.446621
CEAT        -0.1568125    2.498546  0.0000000 -2.948075
Falken       2.7912627    5.446621  2.9480752  0.000000
```

Demonstration with Example

- Now, we consider two groups significantly different if

$$\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k\,(k-1)} = \frac{0 \cdot 20}{7 \times 6} \approx 0 \cdot 0049$$

Demonstration with Example

- Now, we consider two groups significantly different if

$$\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k\,(k-1)} = \frac{0\cdot 20}{7\times 6} \approx 0\cdot 0049$$

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

	Apollo	Bridgestone	CEAT	Falken
Apollo	0.0000000	2.655359	0.1568125	2.791263
Bridgestone	2.6553585	0.000000	2.4985460	5.446621
CEAT	0.1568125	2.498546	0.0000000	2.948075
Falken	2.7912627	5.446621	2.9480752	0.000000

Demonstration with Example

- Now, we consider two groups significantly different if

$$\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k\left(k-1\right)} = \frac{0\cdot 20}{7\times 6} \approx 0\cdot 0049$$

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

```
               Apollo Bridgestone         CEAT    Falken
Apollo      0.0000000    2.655359 0.1568125 2.791263
Bridgestone 2.6553585    0.000000 2.4985460 5.446621
CEAT        0.1568125    2.498546 0.0000000 2.948075
Falken      2.7912627    5.446621 2.9480752 0.000000
```

- And the cut off point τ_{α^*} as :-

```
[1] 2.128045
```

Demonstration with Example

- 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

Demonstration with Example

- 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 (Apollo, CEAT) the outcome is FALSE hence their mean differences are not significant at $\alpha = 20\%$ 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.

Some other approximations to the exact distribution of the kruskal-wallis test statistic

- **(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.

Some other approximations to the exact distribution of the kruskal-wallis test statistic

- **(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.

- **(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,

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

Some other approximations to the exact distribution of the kruskal-wallis test statistic

- **(Satterthwaite Approximation)** A more powerful test using the concept of Welch–Satterthwaite equation where the degrees of freedom of the F statistic previously stated is approximated in other words :-

$$F = \frac{H/k-1}{(N-H-1)/N-k} = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{R}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}_i\right)^2} \overset{a}{\sim} F_{k-1,\widehat{f}}$$

where $\widehat{f} = \frac{\left(\sum_{i=1}^{k} (n_i-1)v_i\right)^2}{\sum_{i=1}^{k} (n_i-1)v_i^2}$ and $v_i = \frac{1}{n_i-1} \sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}_i\right)^2$.