Kruskal Wallis Test for Differences in Distribution

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: November 01, 2017

PDF file location: http://www.murraylax.org/rtutorials/kw.pdf

HTML file location: http://www.murraylax.org/rtutorials/kw.html

---

Note on required packages: The following code requires the package `psych` to perform statistics related to the median. The code also requires a feature in the `tidyverse` package. If you have not already done so, download, install, and load the library with the following code:

```
install.packages("psych")
```

```
install.packages("tidyverse")
```

```
library("psych")
```

```
library("tidyverse")
```

---

1. Introduction

Kruskal Wallis Test is a statistical method to determine if there is a difference in distributions between two or more *independent* groups, where the groups are defined by the outcomes for a single categorical variable.

Kruskal Wallis Test can be assumed to test the hypothesis that the *center* (as measured by the median or interpolated median) of the distribution for the outcome variable is different for different categories of the explanatory variable, if it can be assumed that the *shape of the distribution* is the same across categories.

In its purpose, it is essentially an non-parametric alternative of the one-way ANOVA for a difference in means for two or more groups.

Because the Kruskal Wallis test is a non-parametric test using ranks of the observations in its calculations, it can be performed on *ordinal* data or interval/ratio data.

2. Hypotheses

The null hypothesis says that the center of the distribution for the outcome variable all the groups are equal to each other. The alternative hypothesis says that *at least one* of the groups has a different measure of center than the others.

Let $K \geq 2$ denote the number of groups for the explanatory variable. Let M_i denote the population interpolated median for group i. The null and alternative hypotheses for the Kruskal Wallis Test are given by,

Null: $M_1 = M_2 = ... = M_K$ (i.e. all the interpolated medians are equal)

Alternative: In at least one pair-wise comparison, $M_i \neq M_j$, for some $i \neq j$ (i.e. at least one interpolated median is different than another).

3. Assumptions

The following assumptions are necessary for the Kruskal Wallis test:

- **Random assignment:** The group that an observations belongs to is determined independently from the outcome variable.
- **Independent groups:** Observations in one group are independent of observations in all other groups. There must therefore be different sampling units in each group.
- **Homogeneity in distribution shape:** The shape of the distribution of the outcome variable for the different groups are the same, even though the population medians or interpolated medians may be different.

4. Example: Current Population Survey

The data set below includes data from more than 52,000 individuals over the age of 25 years that participated in the 2016 Current Population Survey. The variables include the following:

- `HOURS`: Usual weekly hours worked over all jobs (ratio)
- `SEX`: Male or female (categorical)
- `RACE`: Racial identity (categorical)
- `EDUC`: Education (categorical)
- `EDUNUM`: A numerical code for the education categories in `EDUC` where larger numbers indicate a higher level of educational attainment
- `MARRIED`: Marital status (categorical)

The following code downloads the data, opens it, and saves it in a data frame called `wdat`.

```
load(url("http://murraylax.org/datasets/cpshours.RData"))
```

We can use this data to determine if educational attainment, an ordinal variable, is different for people in different categories for one or more of other the categorical variables.

5. Computing Medians / Interpolated Medians

Let us calculate the median and interpolated median level of education for the different racial groups. The code below uses the pipe operator (`%>%`) with the `group_by()` and `summarise()` functions, to compute and report the median level of education by group.

```
wdat %>%
  group_by(RACE) %>%
  summarise(MedianEdu = median(EDUCNUM))
```

```
## # A tibble: 5 x 2
##                            RACE MedianEdu
##                           <fctr>     <dbl>
## 1 American Indian/Aleut/Eskimo          2
## 2       Asian/Pacific Islander          3
## 3                         Black         2
## 4                         Other         2
## 5                         White         3
```

We can do the same thing to summarize the interpolated median by racial category.

```
wdat %>%
  group_by(RACE) %>%
  summarise(MedianEdu = interp.median(EDUCNUM))
```

```
## # A tibble: 5 x 2
##                            RACE MedianEdu
##                          <fctr>     <dbl>
## 1 American Indian/Aleut/Eskimo  2.140299
## 2       Asian/Pacific Islander  3.021533
## 3                         Black  2.353535
## 4                         Other  2.392857
## 5                         White  2.570952
```

We can see there are differences in the sample medians and interpolated medians for level of education between different races. In the next section, we formally test for a difference in medians.

6. Conducting the Kruskal Wallis Hypothesis Test

The code below calls the `kruskal.test()` function to test the hypothesis that average educational level is different for people in different races.

```
kt <- kruskal.test(EDUCNUM ~ RACE, data=wdat)
```

The *formula*, `EDUCNUM ~ RACE` tells the function to examine the outcome variable `EDUCNUM` as it depends on `RACE`. The second parameter, `data=wdat`, specifies the data frame to find these variables.

We can review the outcome of the statistical test by viewing the return object `kt`.

```
kt
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  EDUCNUM by RACE
## Kruskal-Wallis chi-squared = 783.79, df = 4, p-value < 2.2e-16
```

With a p-value equal to 2.2×10^{-16}, we have sufficient statistical evidence that the average level of educational attainment is different for different racial groups.