Group Comparisons w.r.t different Tests

Last week leftover

Problem in multi grouping? (last-week)

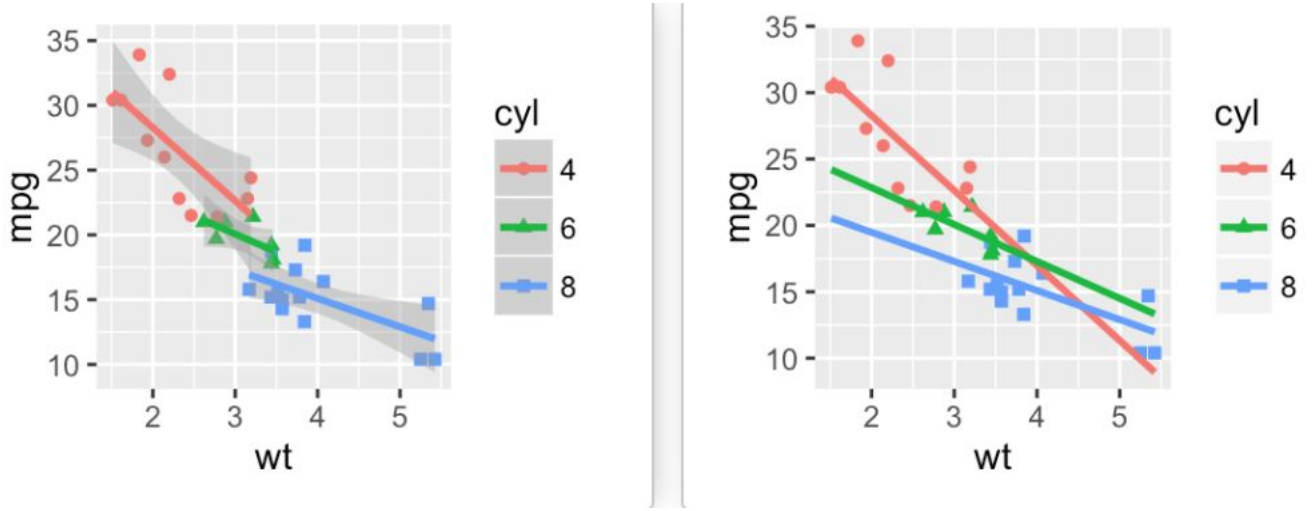
Add regression lines

Add regression lines

- ggplot(mtcars, aes(x=wt, y=mpg, color=factor(cyl), shape=factor(cyl))) + geom_point() + geom_smooth(method=lm)

Remove confidence intervals and Extend the regression lines

- ggplot(mtcars, aes(x=wt, y=mpg, color=factor(cyl), shape=factor(cyl))) + geom_point() + geom_s mooth(method=lm, se=FALSE, fullrange=TRUE)



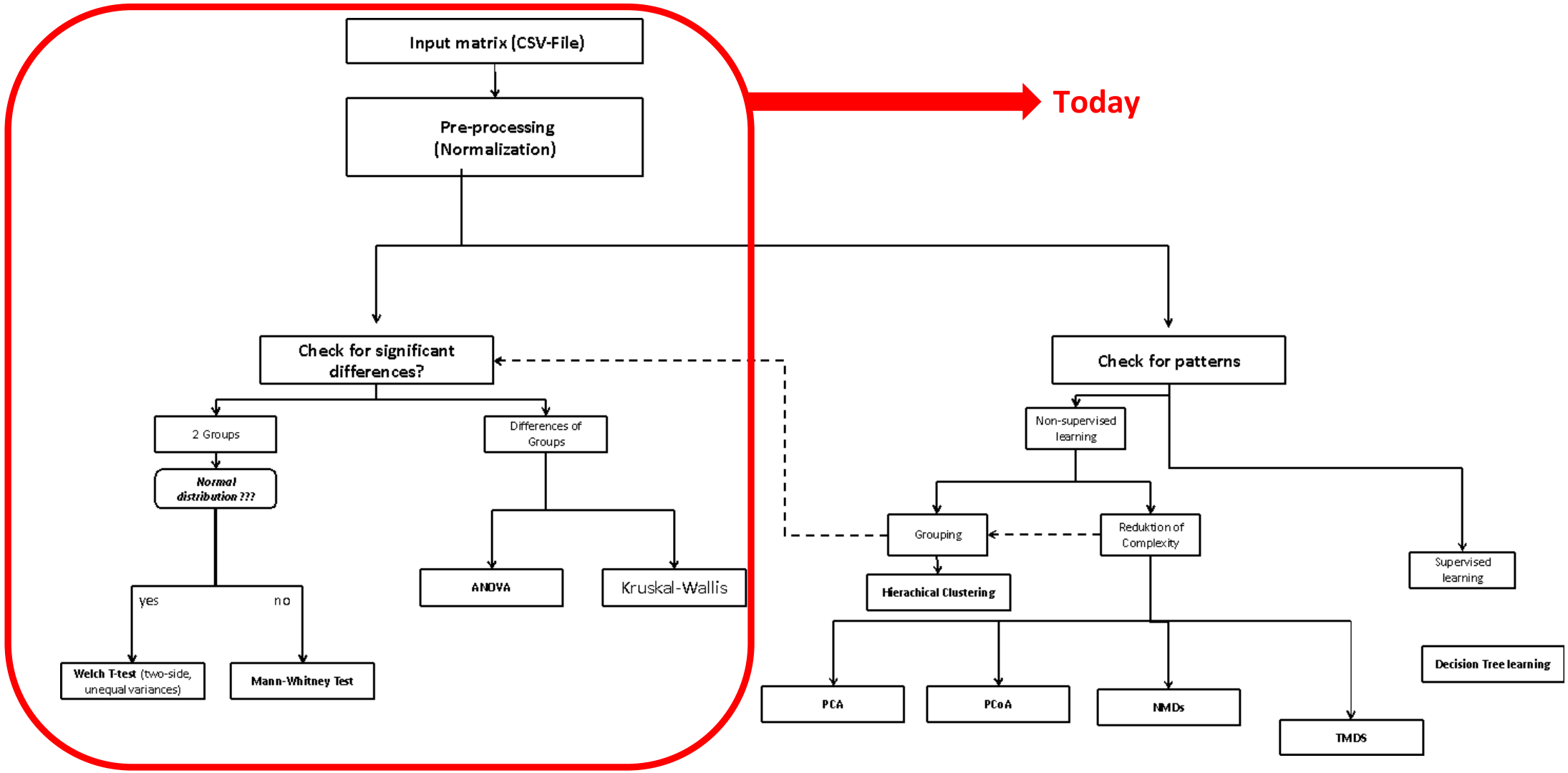
TASK for Practice

- TASK: First take data=USA and put rownames= state (as we did in tutorial). Now using gather command combine all types of crime into one column and name it ‘crime’ and their values in column ‘rate’ and save in variable ‘p’.
- Afterwards, create a scatterplot of data=p and put column ‘rate’ on x axis and ‘state’ on y axis. And color them with respect to column ‘crime. And this is how it should look like.

Agenda

1. Normalization of Data in R
    - Re-scaling and Standardization
2. Group comparison of variables within Two groups
    - Realization of Normality
    - T-Test and U-Test
3. Group comparison of variables within multiple groups
    - Realization of Normality
    - Anova and Kruskal-wallis TEST

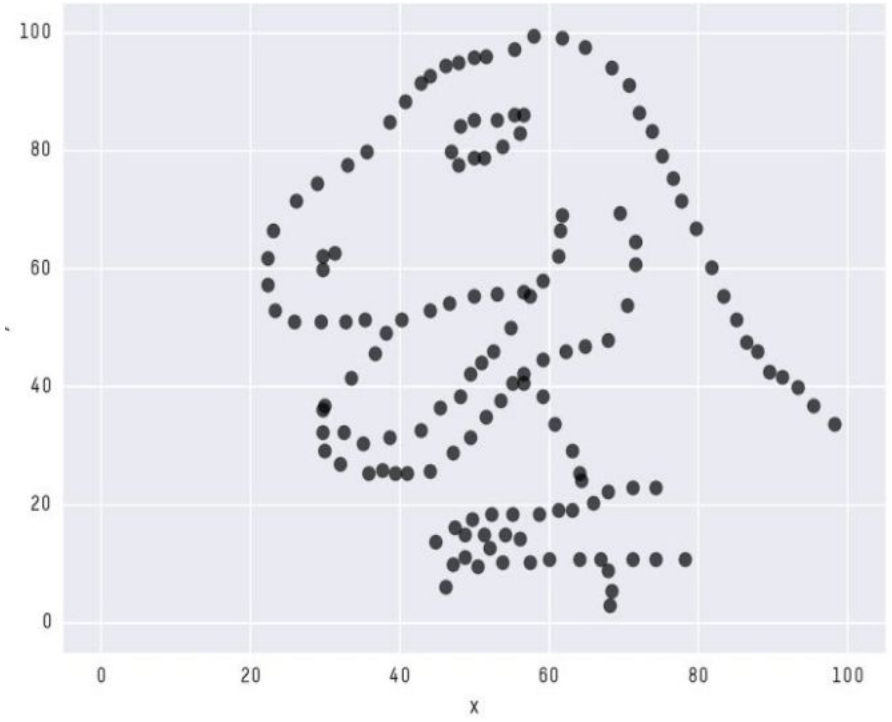
Overview of Project road-map



Normalization

Elimination of Units of Measurement for Easier Comparison of Data from Different Places

For Example: In Machine Learning, NORMALIZATION is only required when features have widely varying ranges

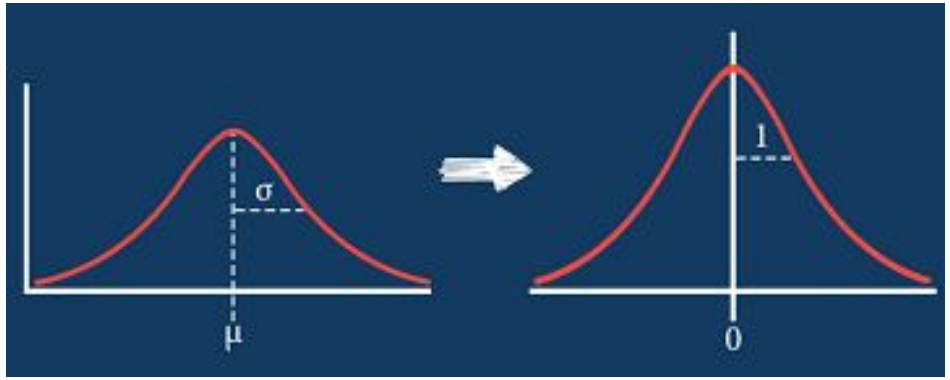


X Mean: 54.2632025
Y Mean: 47.8315781
X SD : 16.7650109
Y SD : 26.9353144
Corr. : -0.0645195

Methods of Normalization

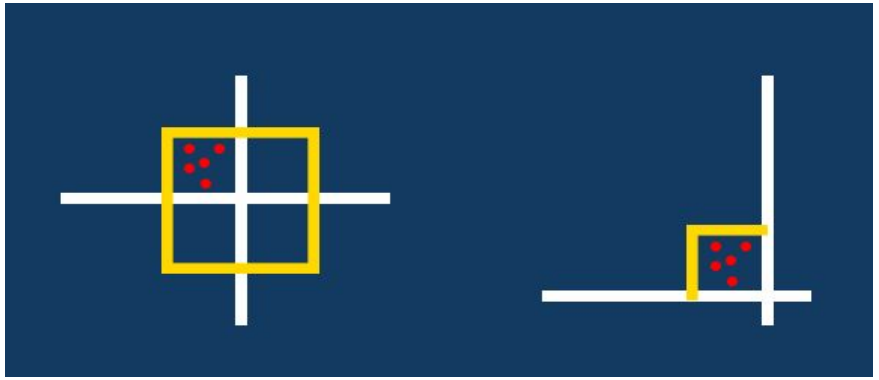
Standardization:

Transforming data into a z-score or t-score i.e. transform data to have a mean of 0 and standard



Feature Scaling:

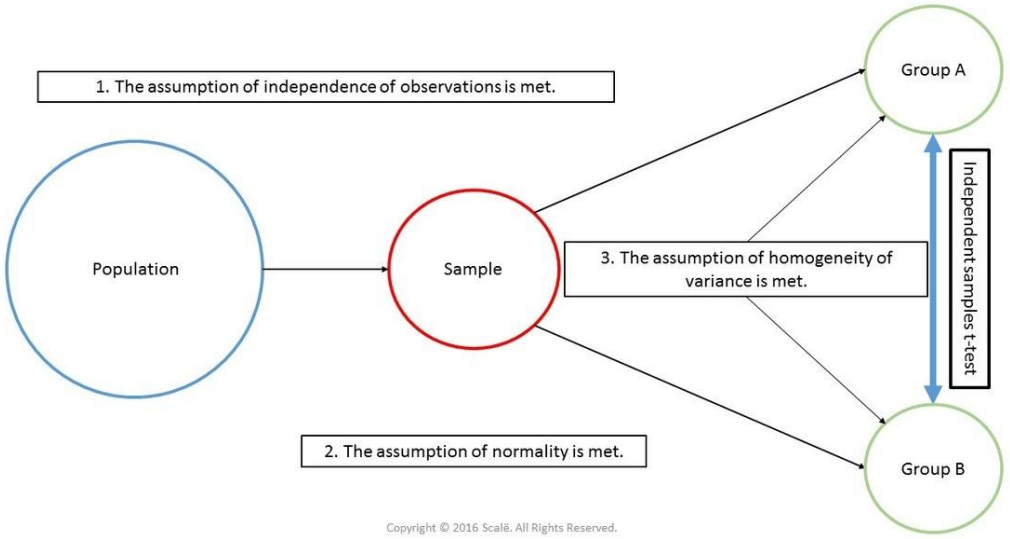
Rescaling data to have values between 0 & 1



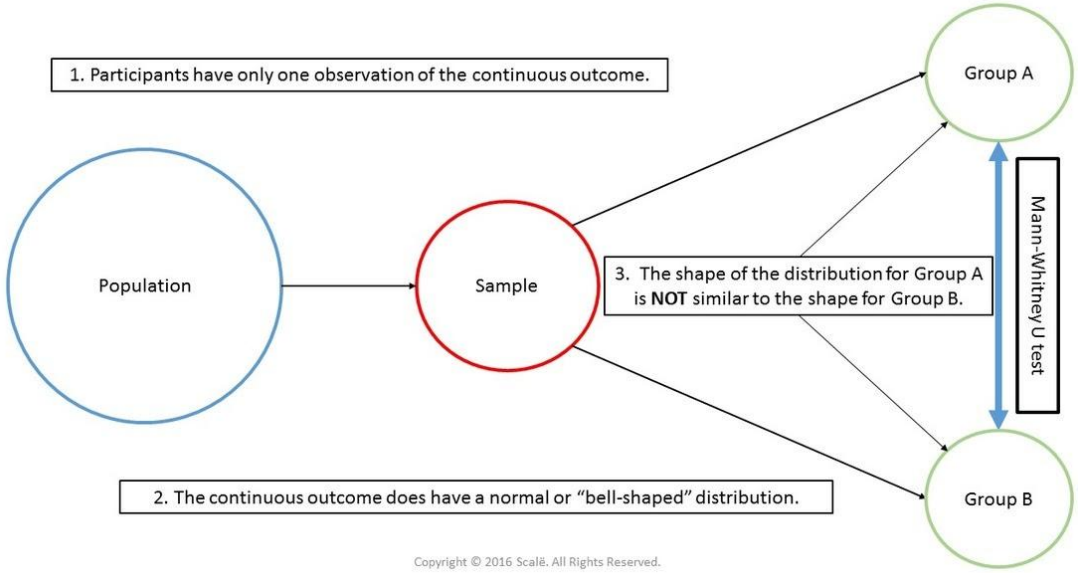
Move to Normalization.R Script

(Take from GITHUB)

Group Comparison of Variables within 2 Groups



"T-TEST"

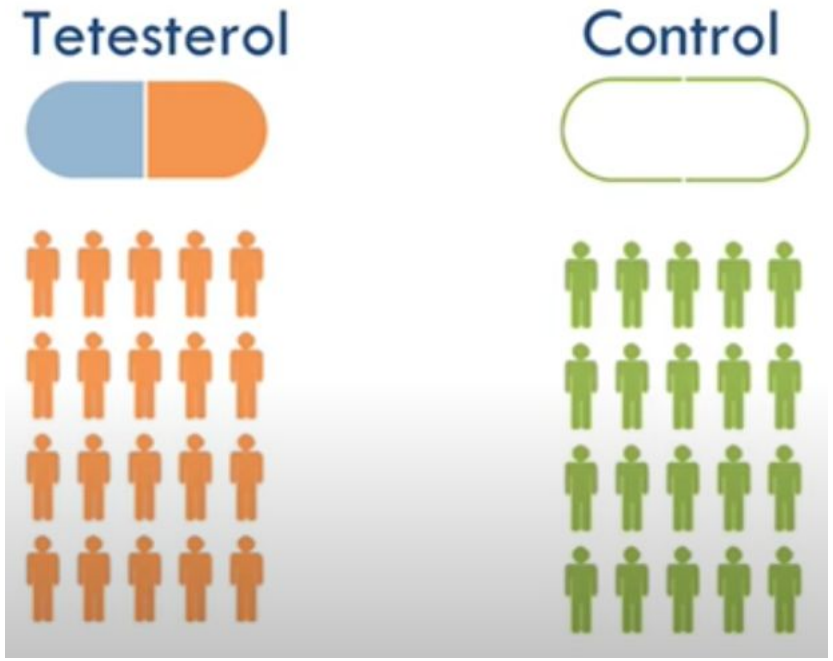


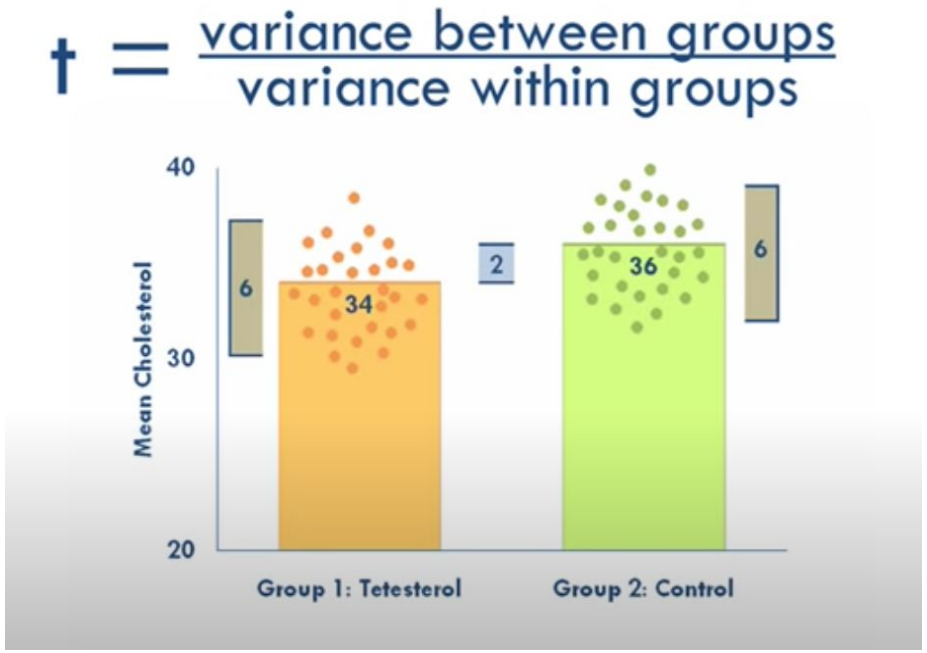
"Mann-Whitney U-Test"

T-Test types for Different Groups

Independent sample test	Paired sample test	One sample test
Tests the mean of two different groups	Tests the mean of one group twice	Tests the mean of one group against a set mean
e.g. Testing the average quality of two different batches of beer	e.g. Testing balance of people before and after drinking alcohol	e.g. Testing IQ of group of people against a standard value 100

T-Test Example

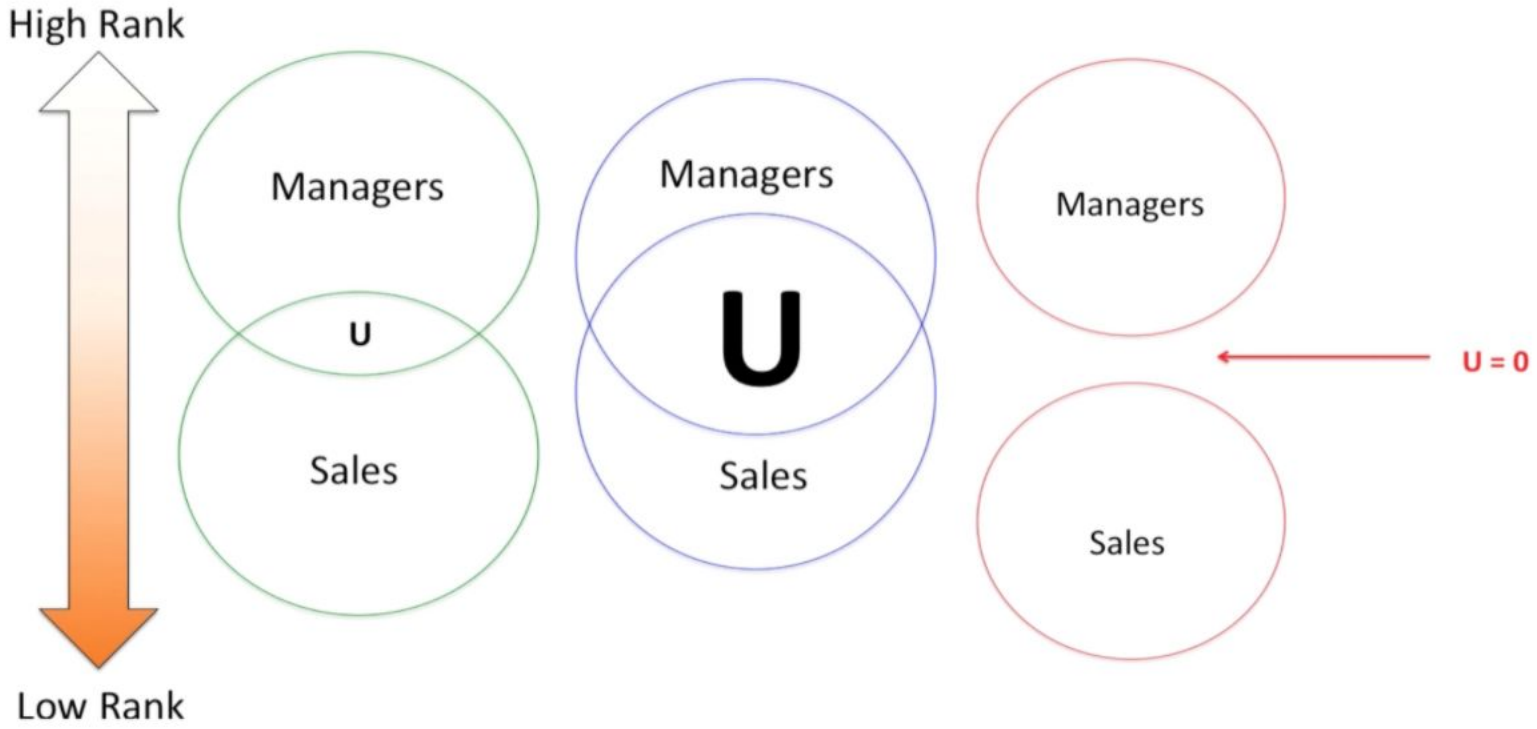




Limitations in T-test

- Results can only be applied to population that resembles the sample. e.g. Cholesterol drug test was conducted for adults, So i can not be true for children.
- Sample and Population should be roughly normal in distribution.
- Each group should have same number of data points. Otherwise there will be inaccurate results.
- All data should be independent.

U-Test Example

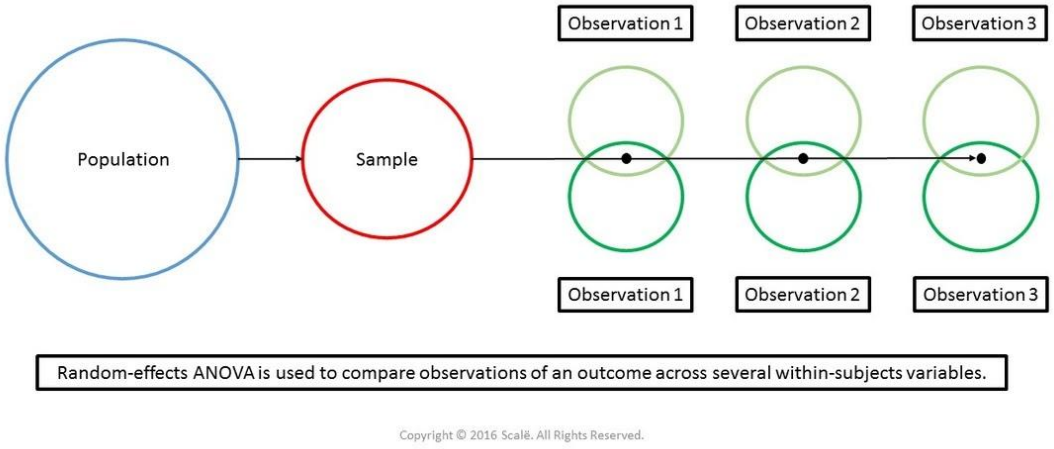


- Smaller U, bigger difference between groups
- Larger U, smaller differences between groups
- U = 0, no overlapping, completely different

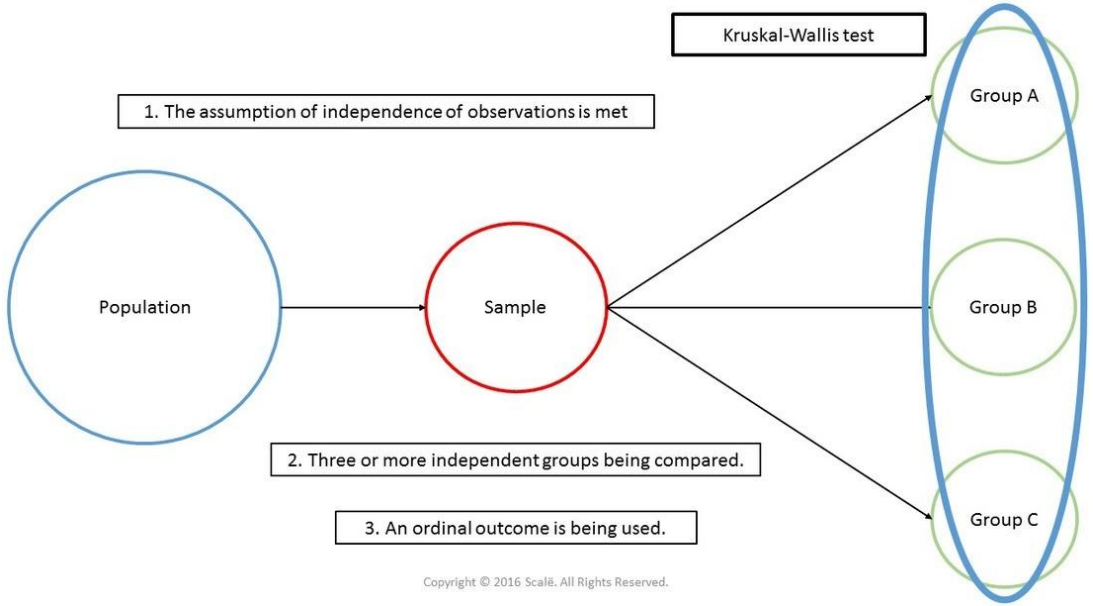
Move to T-Test and U-Test.R Script

(Take from GITHUB)

Group comparison of Multiple Groups



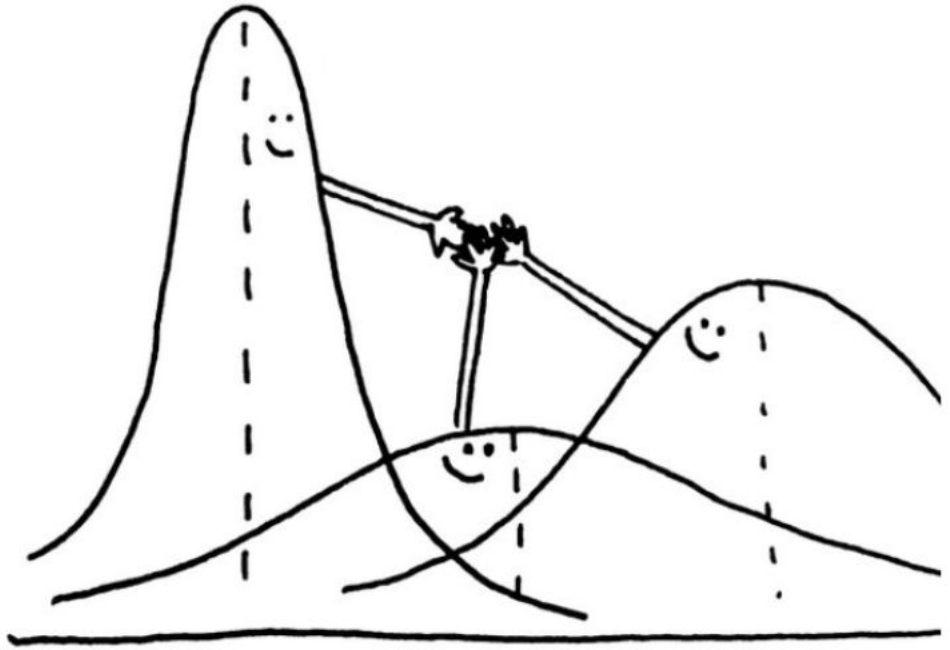
“Analysis of Variance (ANOVA) ”



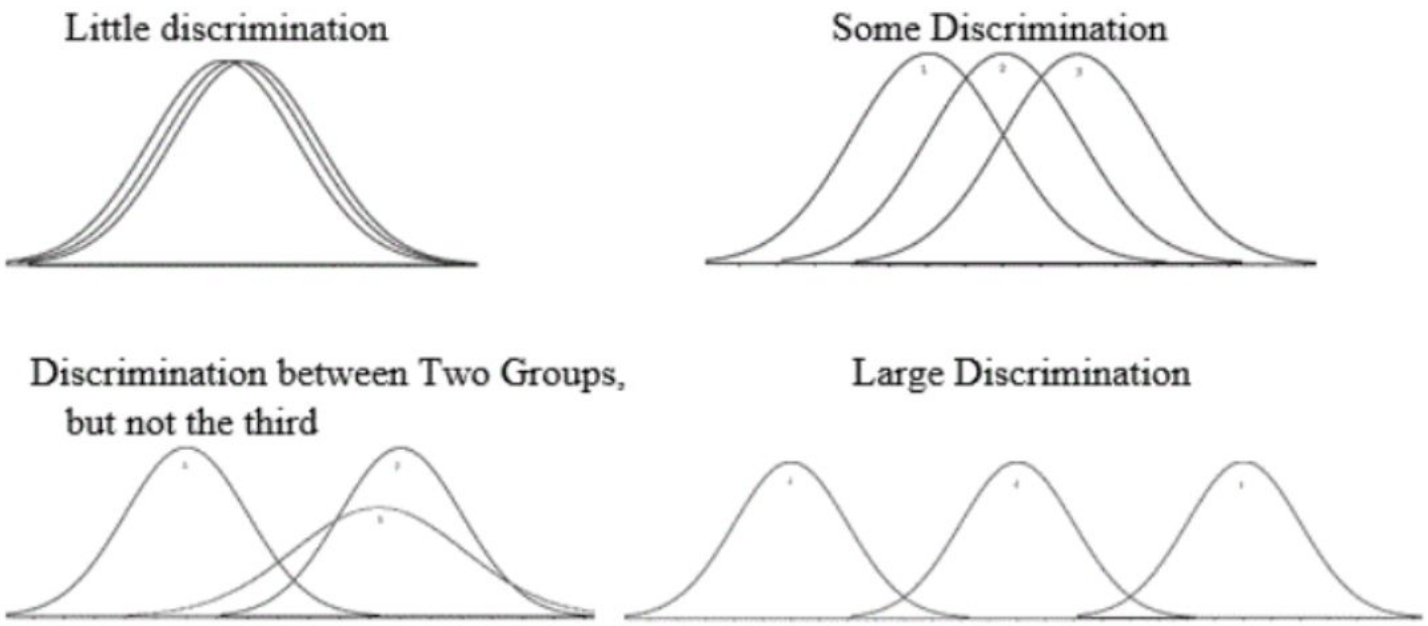
“Kruskal-Wallis-Test (H-Test)”

Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a parametric statistical technique used to compare datasets. It is similar in application to techniques such as t-test, in that it is used to compare means and the relative variance between them. However, analysis of variance (ANOVA) is best applied where more than 2 populations or samples are meant to be compared.



Group Discrimination



Source: Psychstat – Missouri State

Assumptions

- **Independence of case:** Independence of case assumption means that the case of the dependent variable should be independent of the sample. There should not be any pattern in the selection of the sample.
- **Normality:** Distribution of each group should be normal. The Kolmogorov-Smirnov or the Shapiro-Wilk test may be used to confirm the normality of the group.
- **Homogeneity:** Homogeneity means variance between the groups should be the same. Levene's test is used to test the homogeneity between groups.

If particular data follows the above assumptions, then the analysis of variance (ANOVA) is the best technique to compare the means of two, or more, populations.

Kruskal-Wallis H-test

- (sometimes also called the "one-way ANOVA on ranks") is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. It is considered the nonparametric alternative to the one-way ANOVA, and an extension of the Mann-Whitney U test to allow the comparison of more than two independent groups.

It is important to realize that the Kruskal-Wallis H test is an omnibus test statistic and cannot tell you which specific groups of your independent variable are statistically significantly different from each other; it only tells you that at least two groups were different. Since you may have three, four, five or more groups in your study design, determining which of these groups differ from each other is important.

Example

you could use a Kruskal-Wallis H test to understand whether exam performance, measured on a continuous scale from 0-100, differed based on test anxiety levels (i.e., your dependent variable would be "exam performance" and your independent variable would be "test anxiety level", which has three independent groups: students with "low", "medium" and "high" test anxiety levels).

Assumptions

1. Your **dependent variable** should be measured at the **ordinal or continuous level** (i.e., interval or ratio). Examples of ordinal variables include Likert scales (e.g., a 7-point scale from "strongly agree" through to "strongly disagree")
2. Your **independent variable** should consist of **two or more categorical, independent groups**. Typically, a Kruskal-Wallis H test is used when you have three or more categorical, independent groups, but it can be used for just two groups (i.e., a Mann-Whitney U test is more commonly used for two groups)
3. You should have **independence of observations**, which means that **there is no relationship between the observations in each group or between the groups themselves**. For example, there must be different participants in each group with no participant being in more than one group

Move to Anova & Krusskal-wallis .R Script

(Take from GITHUB)