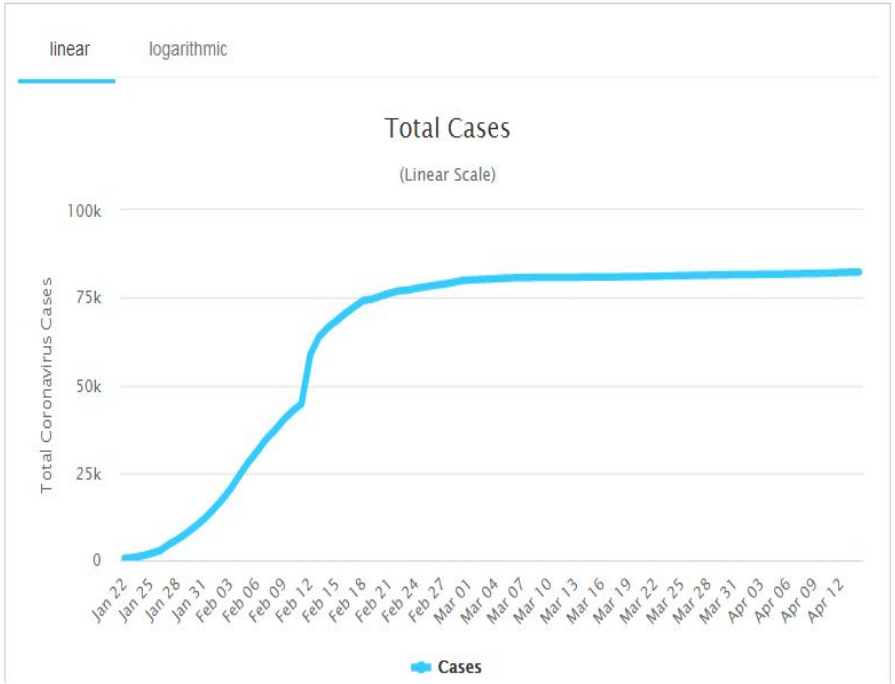
NORMALIZATION

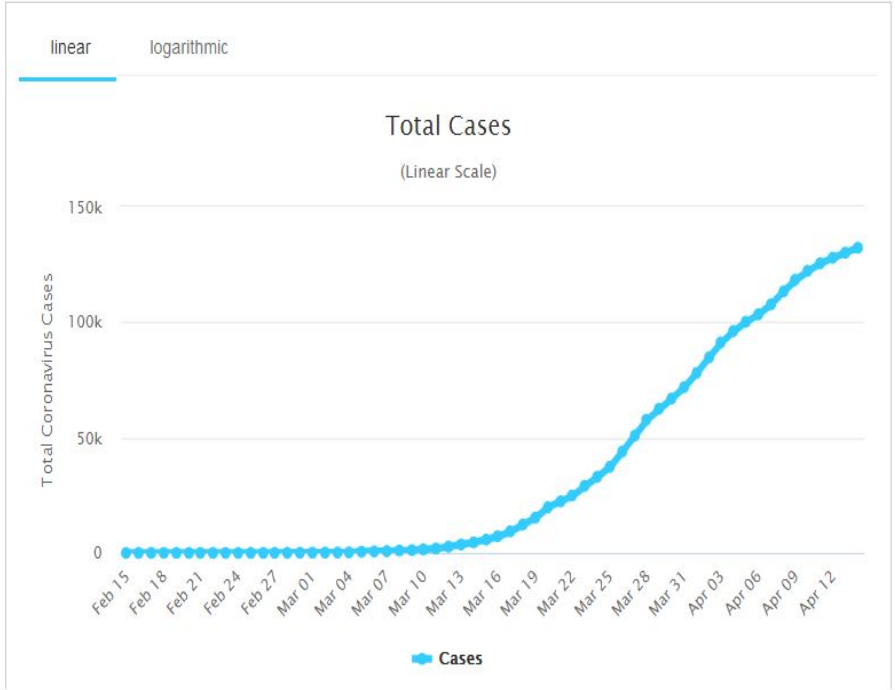
Elimination of Units of Measurement for Easier Comparison of Data from Different Places

To Normalize or Not To Normalize??

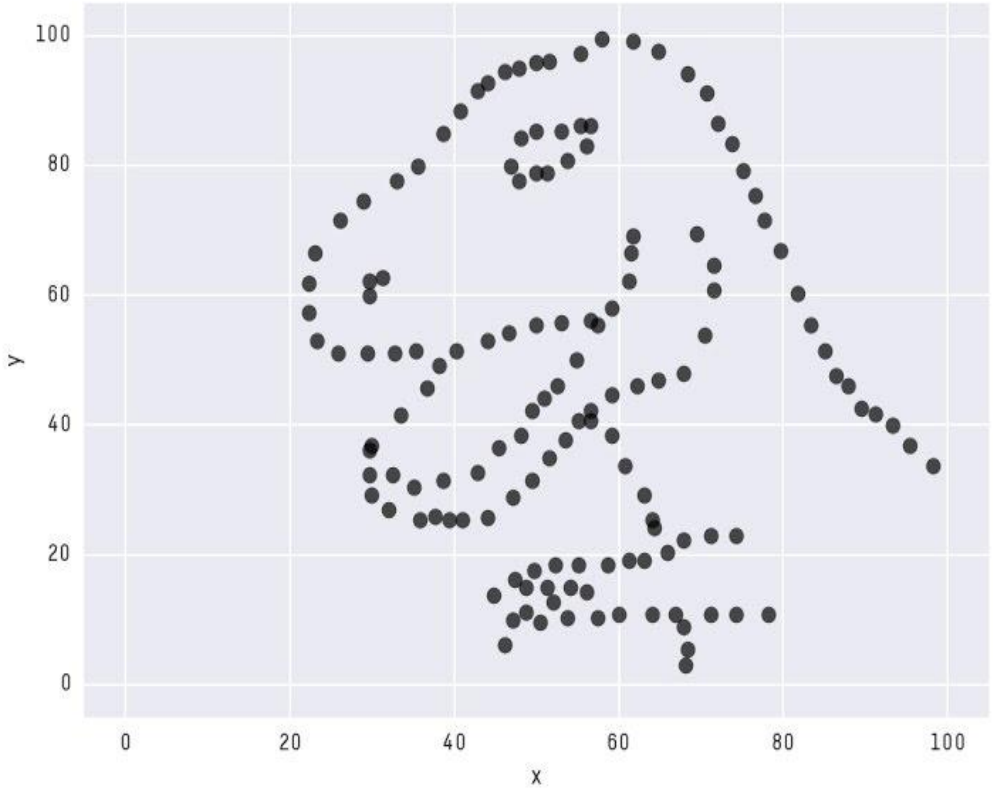
Total Coronavirus Cases in China



Total Coronavirus Cases in Germany



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



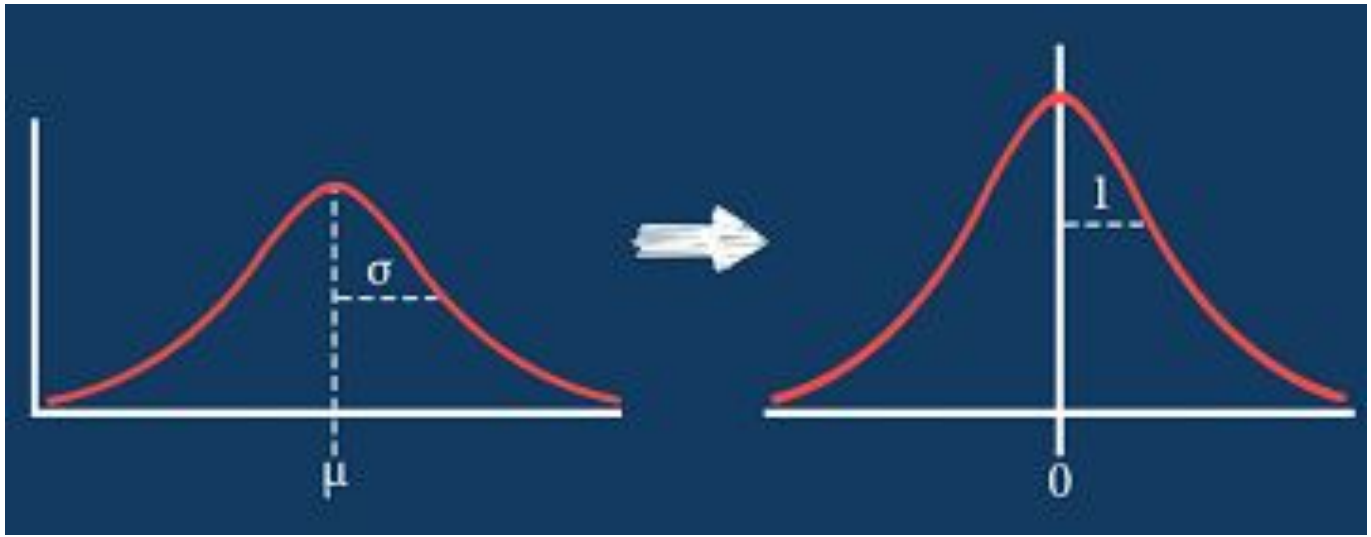
X Mean: 54.2632025
Y Mean: 47.8315781
X SD : 16.7650109
Y SD : 26.9353144
Corr. : -0.0645195

Benefits :

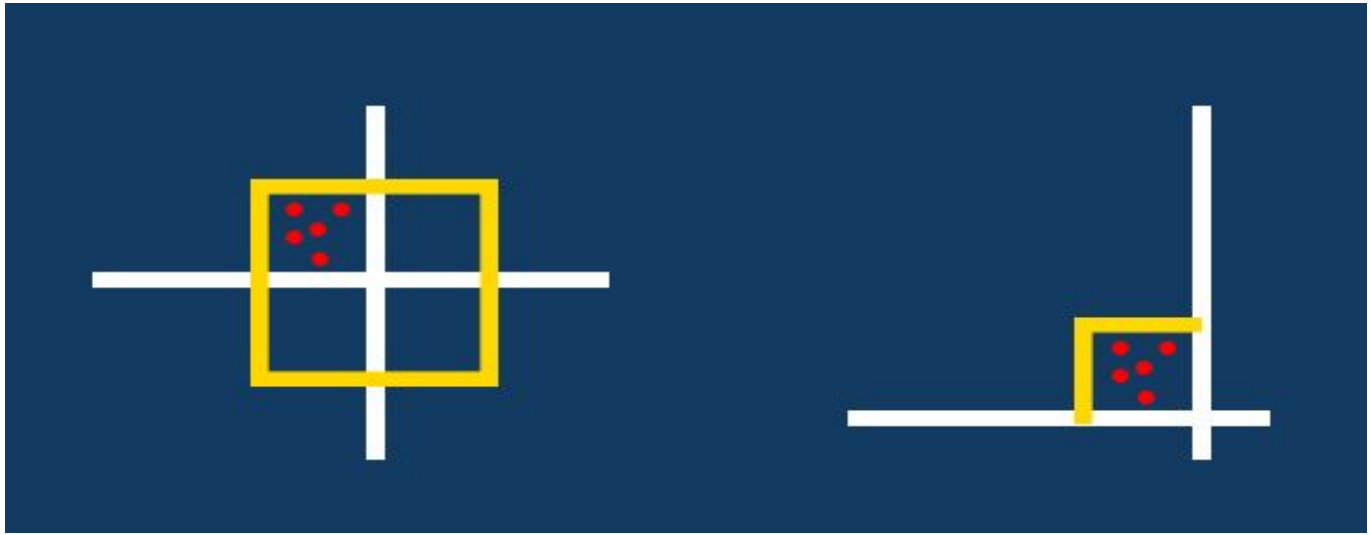
- All input variables have same treatment in the model
- Easier comparison of data
- In terms of database, produces smaller tables with smaller rows
  - Efficient
  - Fewer null values & less redundant data

Methods :

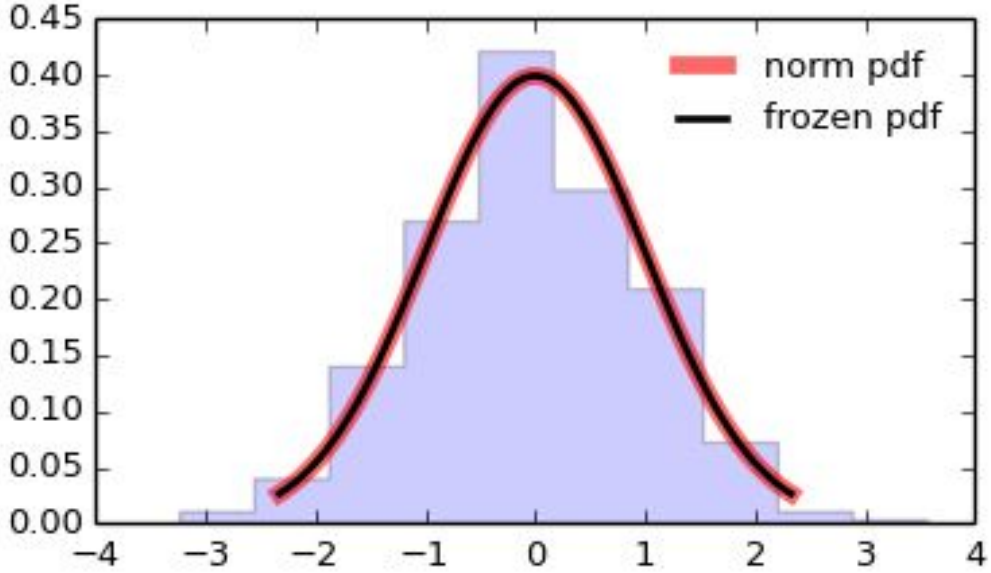
- **Standardization** - Transforming data into a z-score or t-score i.e. transform data to have a mean of 0 and standard deviation of 1



- **Feature Scaling** - Rescaling data to have values between 0 & 1



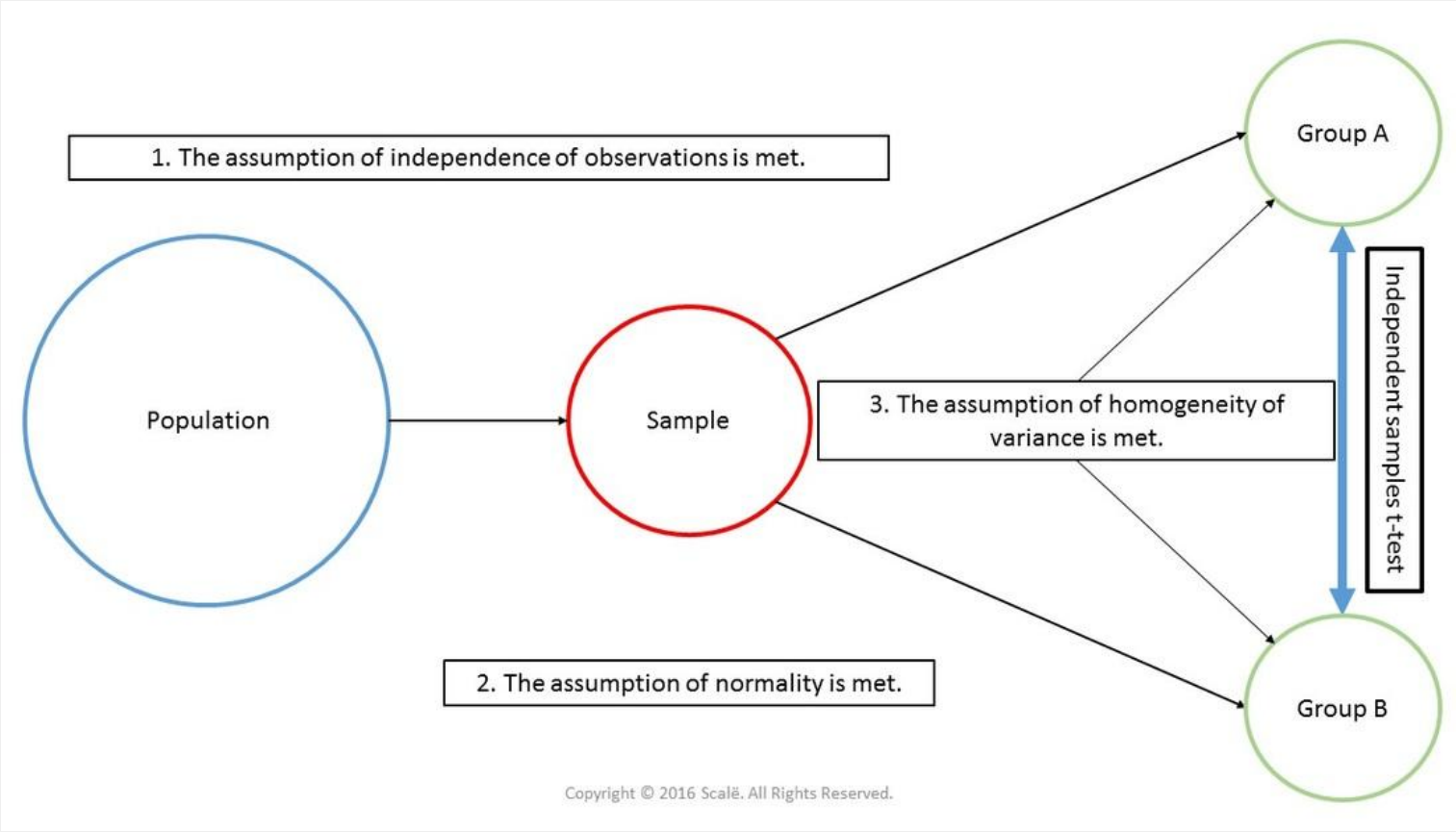
- **Normalizing Moments** - using the formula μ/σ



- **Normalizing Vectors** - Transforming a vector so that it has a length of 1

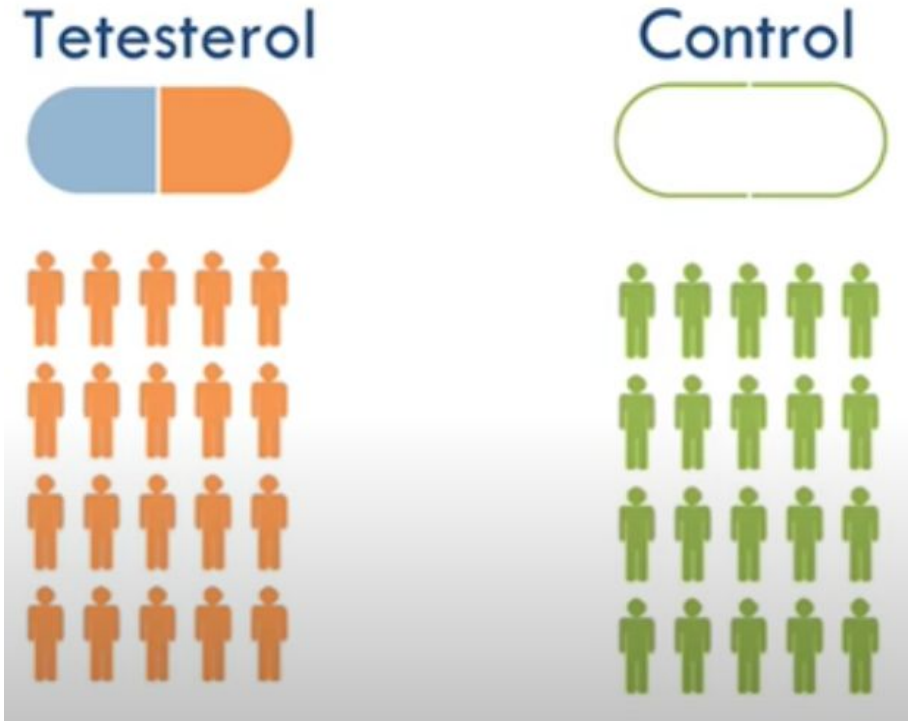
Group Comparison of Variables within 2 Groups

T-Test"

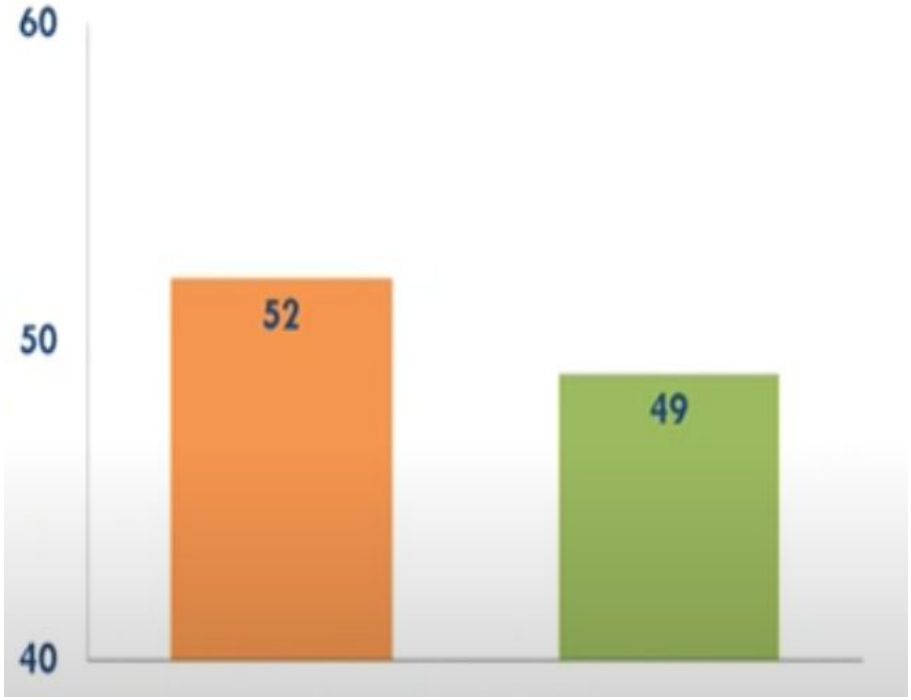


Let us imagine a scenario…

" A cholesterol control pill is being tested on 2 groups of people where one group gets the real drug while the other group receives an inactive drug (placebo) "

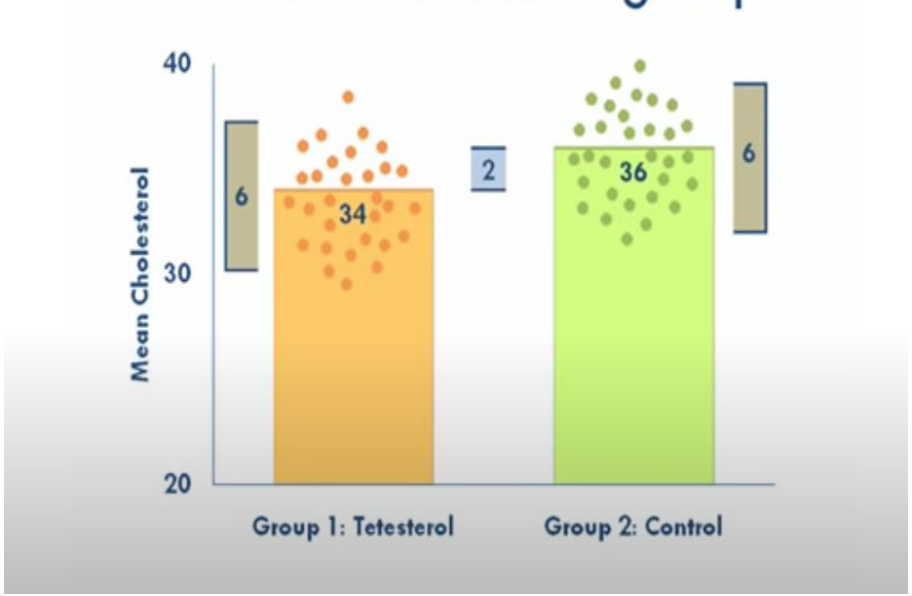


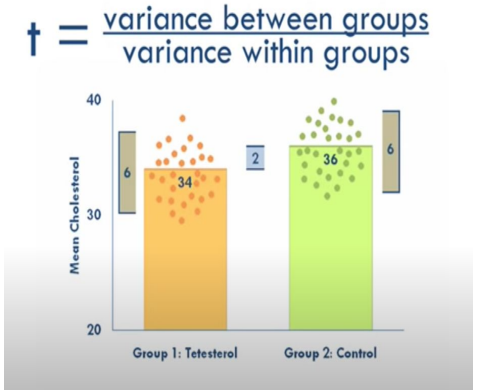
Even when the mean cholesterol levels of both groups are different, we cannot satisfactorily deduce whether the drug is reliably working or not



And that is when **"T-Test"** comes in handy which checks, **if two means/averages (of two groups) are reliably different from each other**

$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$





- Where difference between groups is 2 and difference within groups is 6
  So, t=> 2/6 = 0.3
- To, check the reliability of test *p-value* is used. Whereas, p-value tells
  that there is a real difference between two groups or its just a fluke.
- Usually if p-value is less the 0.05 (means 5%) or less
  then hen the effect is real otherwise not.
- But do not forget p-value depends upon Sample size. Bigger the sample size better the accuracy.

There are 3 main types of T-Test :

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

Limitations in T-test :

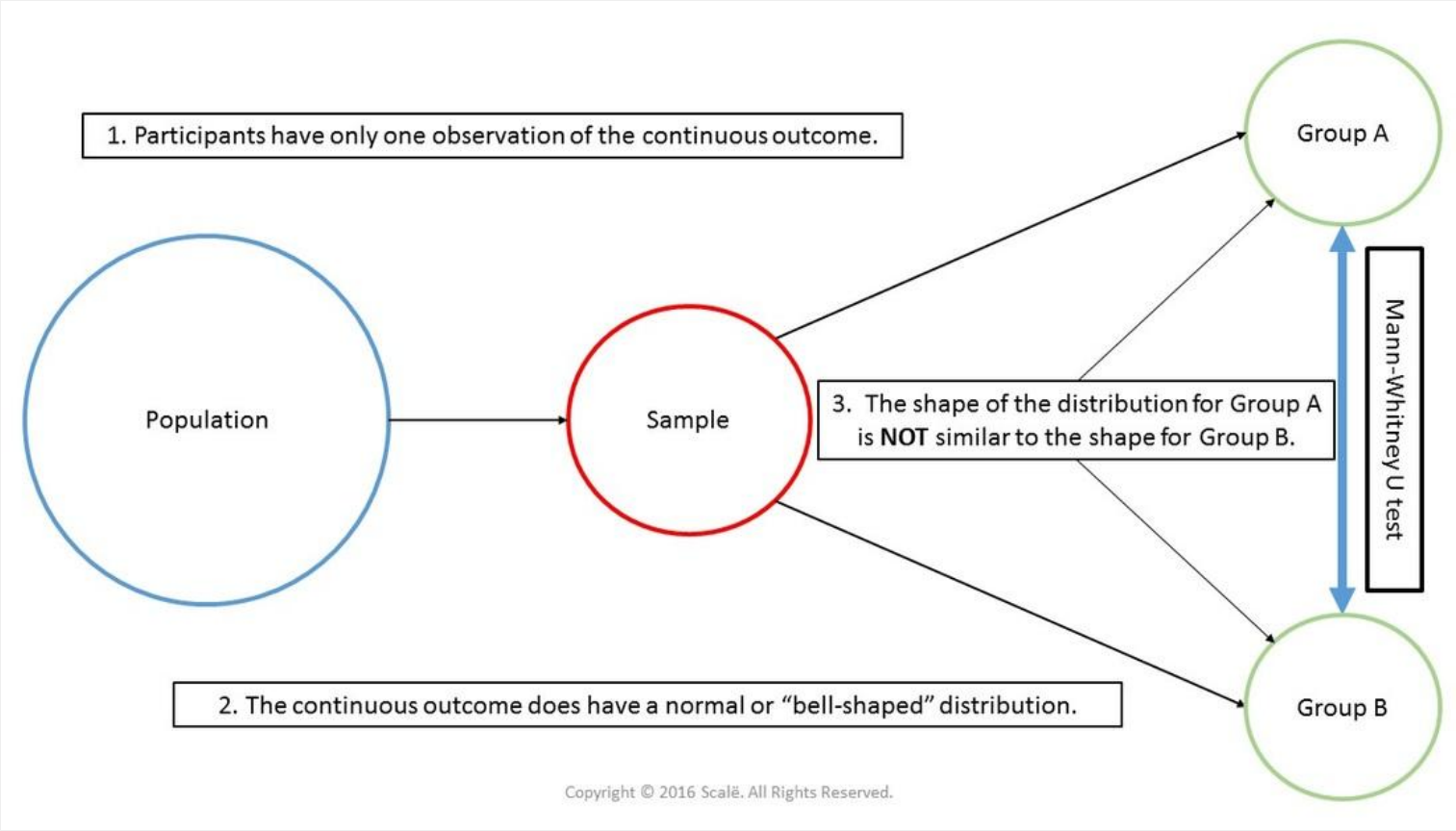
- Results can only be applied to population that resembles the sample.
  e.g. Cholesterol drug test was conducted for adults, So i can not be true for children.
- Sample and Population should be roughly normal in distribution.
- Each group should have same number of data points. Otherwise there will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

Overcoming T-Test

Non-parametric tests like Mann-Whitney U-Test which performs the same job as T-Test, but it can work with normal distribution and ordered level data.

Group Comparison of Variables within 2 Groups

"Mann-Whitney U-Test"



The **U-Statistic** provides **degree of overlap in ranks between two groups of data.**

Let us take a closer look at a set of data containing salaries of various sales person and managers in a company. Note that each data instance has been ranked with 1 being assigned to the lowest and 7 to the highest salary in the dataset

Salary	Rank	Position
$39,907.00	2	Sales
$44,330.00	6	Manager
$40,324.00	4	Sales
$52,404.00	7	Manager
$41,034.00	3	Manager
$30,221.00	1	Sales
$42,198.00	5	Sales

Calculating U :

1. Identify groups with smaller summed ranks
2. For each data point in smaller group, add up how many data points in other group are smaller than the rank
3. Compare U-Statistic to distribution tables to get significance level

Step 1

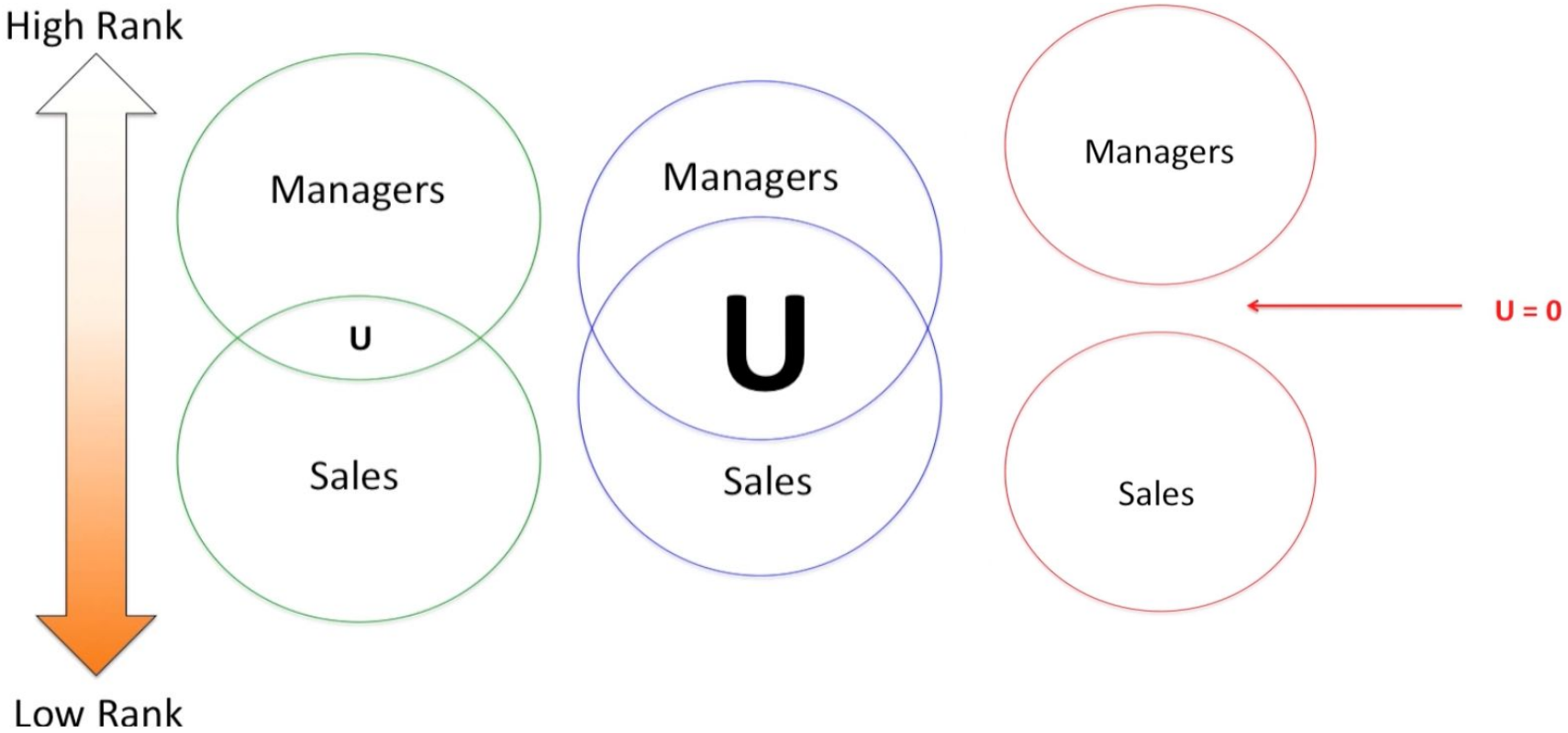
Salary	Rank (low to high)	Position
$30,221.00	1	Sales
$39,907.00	2	Sales
$40,324.00	3	Sales
$41,034.00	4	Manager
$42,198.00	5	Sales
$44,330.00	6	Manager
$52,404.00	7	Manager

Position	Mean Rank	Sum of Ranks
Sales	2.75	11
Managers	5.67	17

Step 2

Salary	Rank (low to high)	Position
$30,221.00	1	Sales
$39,907.00	2	Sales
$40,324.00	3	Sales
$41,034.00	4	Manager
$42,198.00	5	Sales
$44,330.00	6	Manager
$52,404.00	7	Manager

$$U = 0 + 0 + 0 + 1 = 1$$



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

Step 3

$n_1 = 4, n_2 = 3$ $U = 1.00$

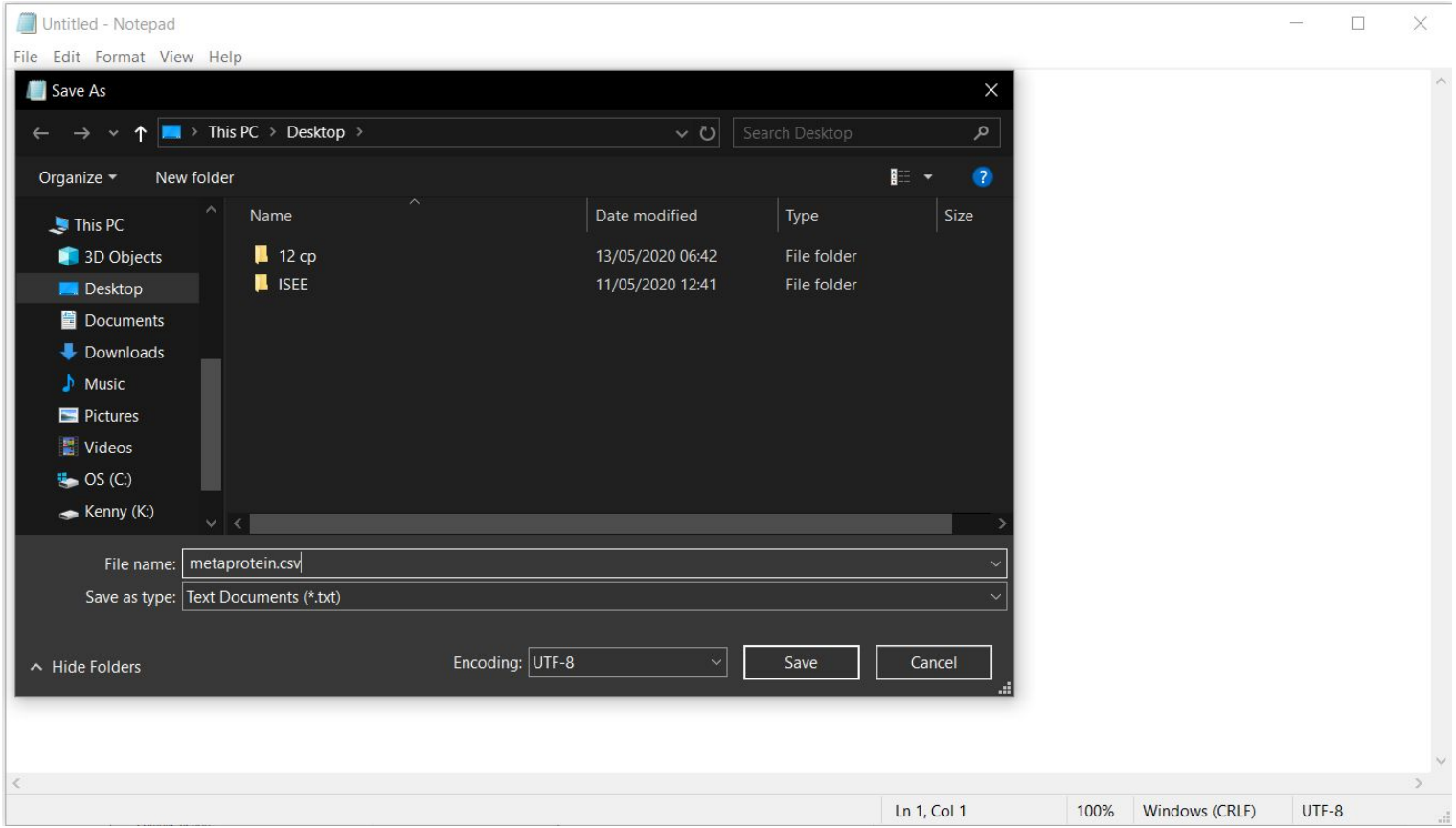
p > .05; n.s.

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42

Group Comparison of Variables within 2 Groups

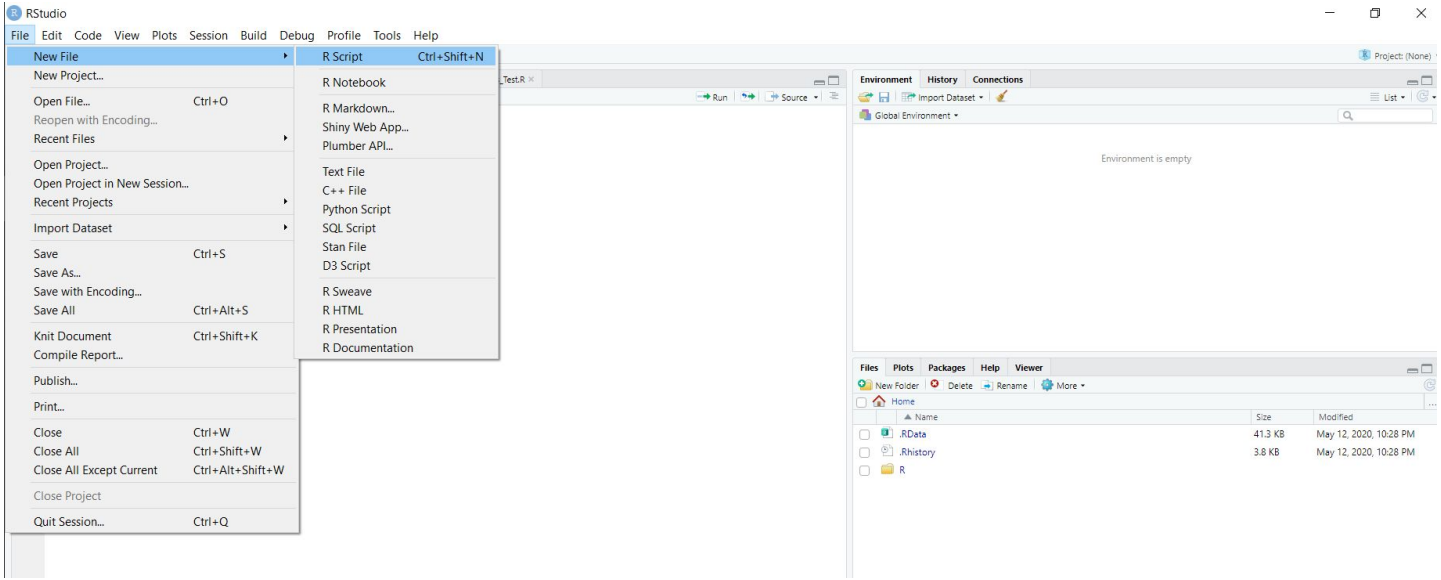
"Implementation of T-Test and Mann-Whitney U-Test in R"

We have implemented the T-Test and U-Test on a database of metaproteins. You can view the data file from the following link. Copy the contents of the file in a notepad and save as a .csv file



Steps for Implementation

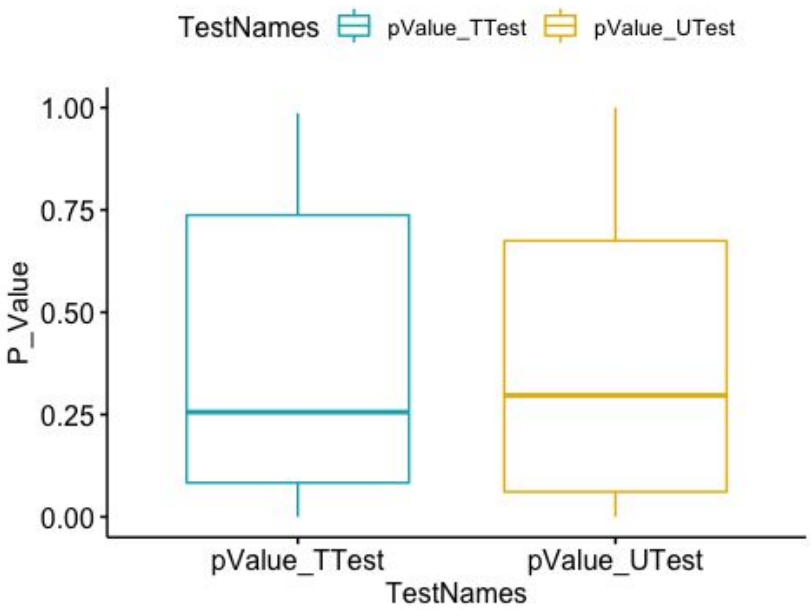
1. Install R and RStudio
2. After you have set up your RStudio, you can get our R script for T-Test and U-Test from here
3. Open a new script in RStudio

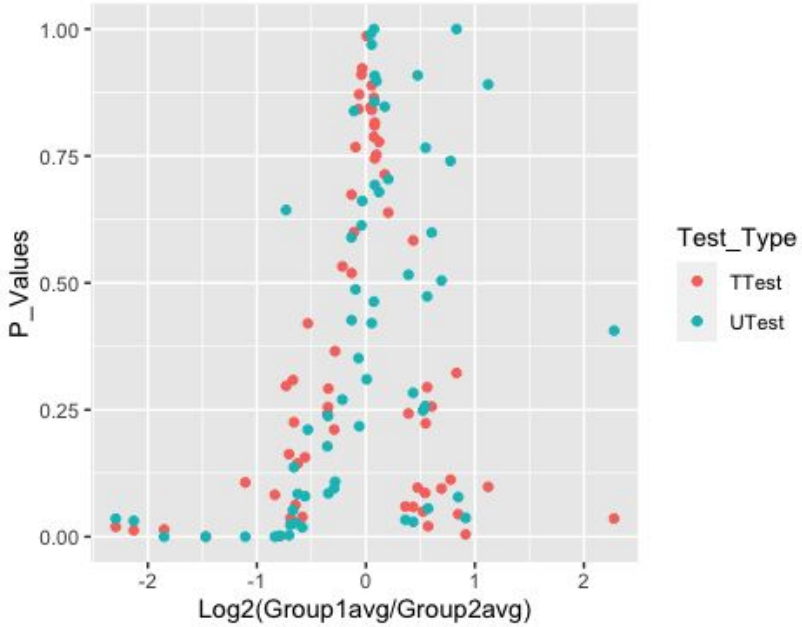


4. From the R script that we have provided, copy the whole code and paste it in the new R script that you have created.
5. Hit Run button

Visualization

Box-Plot and Scatter-Plot of our results





Group comparison of Multivariate Data

"Analysis of Group Similarities (ANOSIM)"

ANOSIM is a nonparametric procedure for testing the hypothesis of no difference between two or more groups of entities based on permutation test of among- and within-group similarities

Steps :

1. Calculate dissimilarity matrix.
2. Calculate rank dissimilarities (smallest dissimilarity is given a rank of 1).
3. Calculate mean among- and withingroup rank dissimilarities.
4. Calculate test statistic R (an index of relative within-group dissimilarity).

Calculation of R:

- Calculate dissimilarity matrix
- Calculate rank dissimilarities (smallest dissimilarity is given a rank of 1)
- Calculate mean among- and withingroup rank dissimilarities.
- Calculate test statistic R (an index of relative within-group dissimilarity)

In addition to the one-way ANOSIM shown in **Figure 1**, two-way ANOSIM as well as crossed and nested designs are possible.

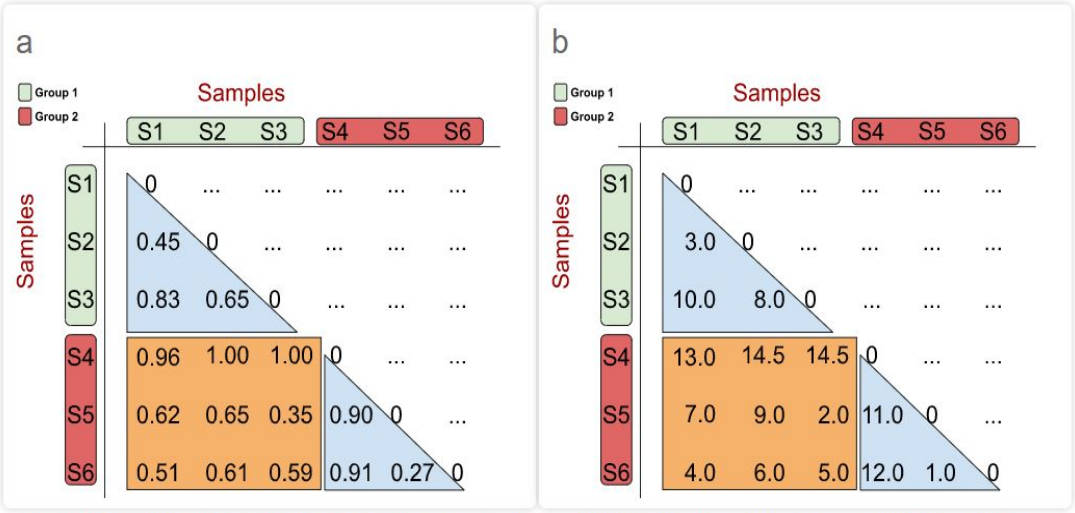


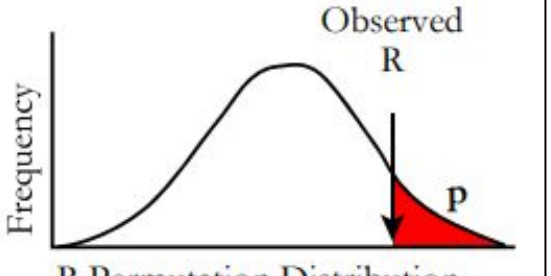
Figure 1: Grouping in a dissimilarity matrix. Given a dissimilarity matrix with defined groups (a) a ranked dissimilarity matrix may be calculated (b) from which ANOSIM may compare the mean rank within groups (blue triangles) to the mean rank between groups (orange rectangle). Note that ANOSIM does not strictly require ranked data; however, if ANOSIM is to be coupled with NMDS, this is recommended. For more than two groups, the mean ranks of all within- and between-group sub-matrices are considered simultaneously when computing the ANOSIM R statistic.

Key Assumption : The ranges of (ranked) dissimilarities within groups are equal, or at least very similar.

Test Statistic "R"

$$R = \frac{\bar{r}_A - \bar{r}_W}{M/2}$$

M = N(N-1)/2
= number of sample pairs



$-1 \leq R \leq 1$

R is interpreted like a correlation coefficient and is a measure of 'effect size' :

- R = 1 when all pairs of samples within groups are more similar than to any pair of samples from different groups.
- R = 0 expected value under the null model that among-and within group dissimilarities are the same on average.
- R < 0 numerically possible but ecologically unlikely

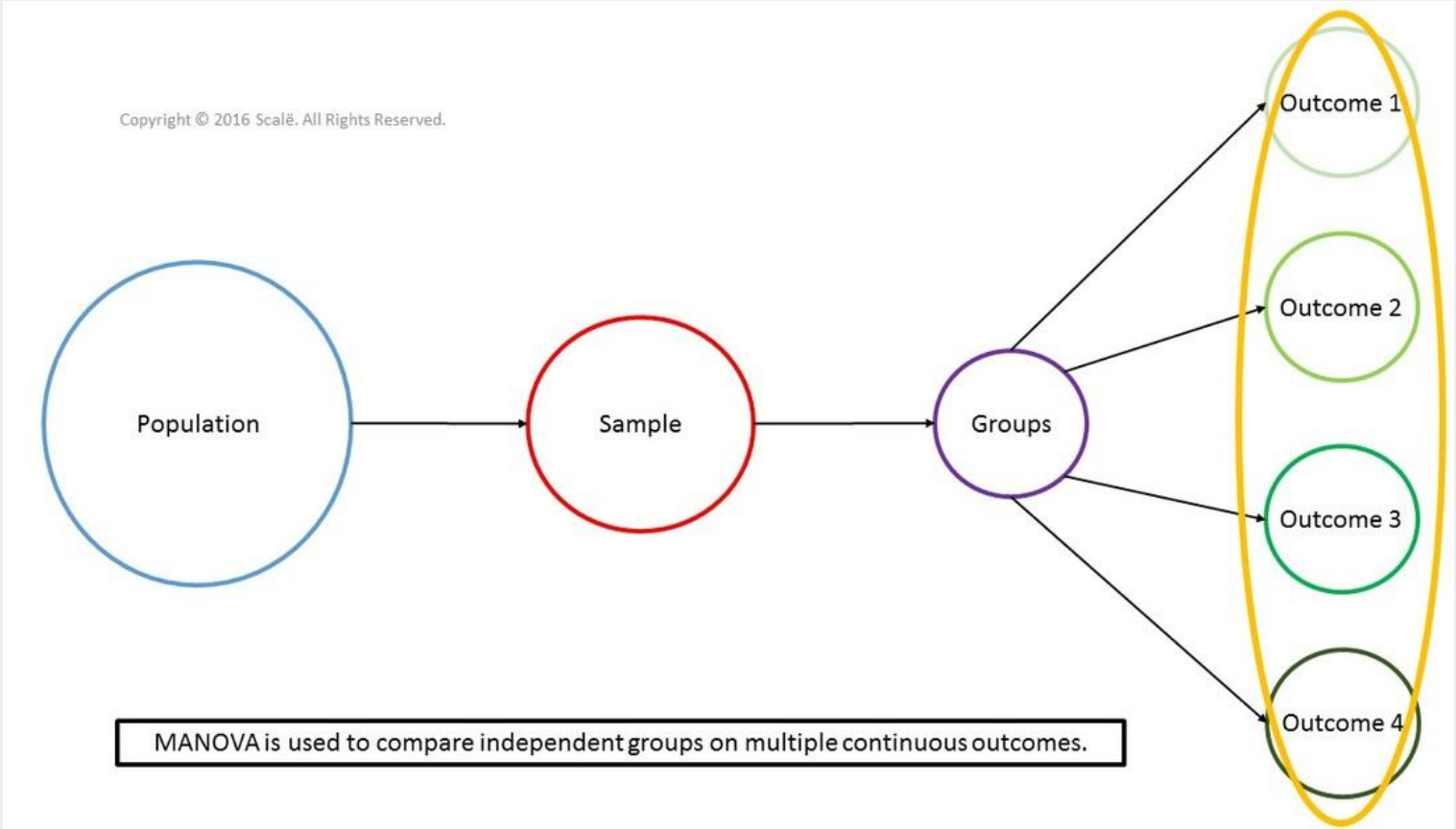
The significance test is simply the fraction of permuted R's that are greater than the observed R

Points to consider :

- Do not assign group membership based on the results of clustering (or a similar exploratory method) applied to the same data set and then treat a significant ANOSIM result as meaningful. This is an example of data dredging.
- Running ANOSIM on groups with very different dispersions can lead to unreliable results. Groups with very different dispersions may produce high R values, even if there's no real difference in their centroids. If differences in group dispersion are as meaningful to your analysis as differences in group centre, this may not be an issue
- Criticism of this and other (dis)similarity-based methods should be considered (e.g. Warton et al., 2012).

Group comparison of Multivariate Data

"Permutational Multivariate Analysis of Variance (PERMANOVA) "



ANOSIM tests whether distances between groups are greater than within groups. PERMANOVA tests whether distances differ between groups. Both tests are sensitive to unbalanced designs and differences in dispersion (variance) within groups (e.g. not good when your groups have different variability).

Key Assumptions:

According to Anderson (2001), the only assumption of PERMANOVA is that the objects in the data set are exchangeable under the null hypothesis. That further implies:

- exchangeable objects (sites, samples, observations, etc.) are **independent**
- exchangeable objects have **similar multivariate dispersion** (i.e. each group has a similar degree of **multivariate scatter**. See Anderson, 2001 and 2006)

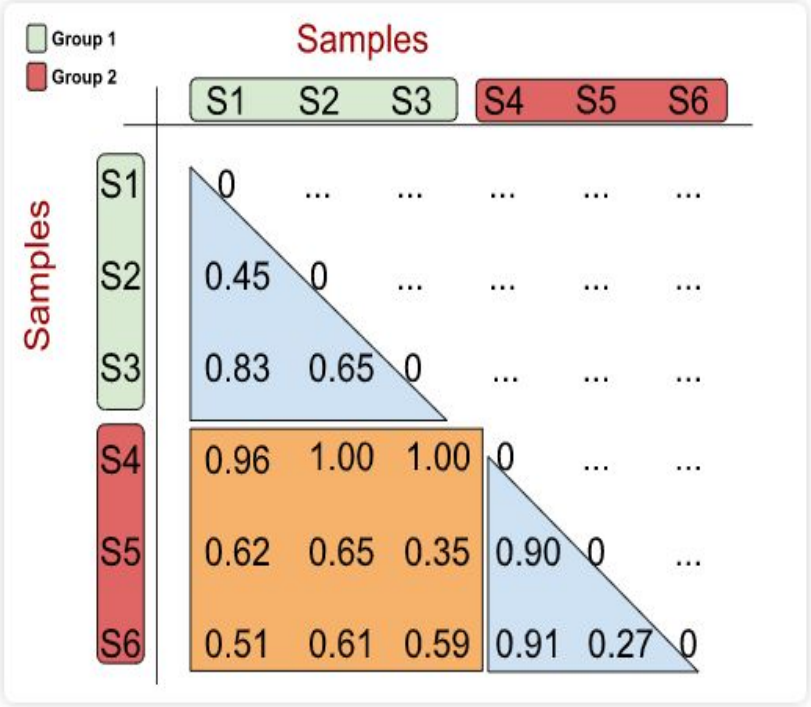


Figure 1: A grouped Bray-Curtis dissimilarity matrix. Note that the matrix is symmetrical about its diagonal. PERMANOVA will compare the within-group dissimilarities (blue triangles) to the between group dissimilarities (orange square) through a pseudo *F*-ratio (**Equation 1**)

The F-Ratio:

$$F = \frac{SS_A \div (a - 1)}{SS_W \div (N - a)}$$

Equation 1: The *F*-ratio used in standard PERMANOVA similar to the traditional *F*-ratio used in ANOVA, however, does not share the same distribution. SS_W is the sum of squared dissimilarities within groups, SS_A is the sum of squared dissimilarities among (between) groups, *a* is the number of groups, and *N* is the total number of objects. The terms (*a*-1) and (*N*-*a*) are the degrees of freedom associated with the explanatory factor (the grouping variable) and the residuals. See Anderson (2001) for discussion and formulae for SS_W and SS_A for simple and more complex designs.

Post Analysis:

As in ANOVA, a significant result indicates that there is a significant difference between the groups defined; however, there is no way of knowing which groups are significantly separated. A posteriori testing, using NPMANOVA, of each pair of groups can be performed after a significant result to determine this. As these are pairwise comparisons, the test statistic involved is the non-parametric, multivariate analogue of the t-statistic, with significance determined by permutation, as above.

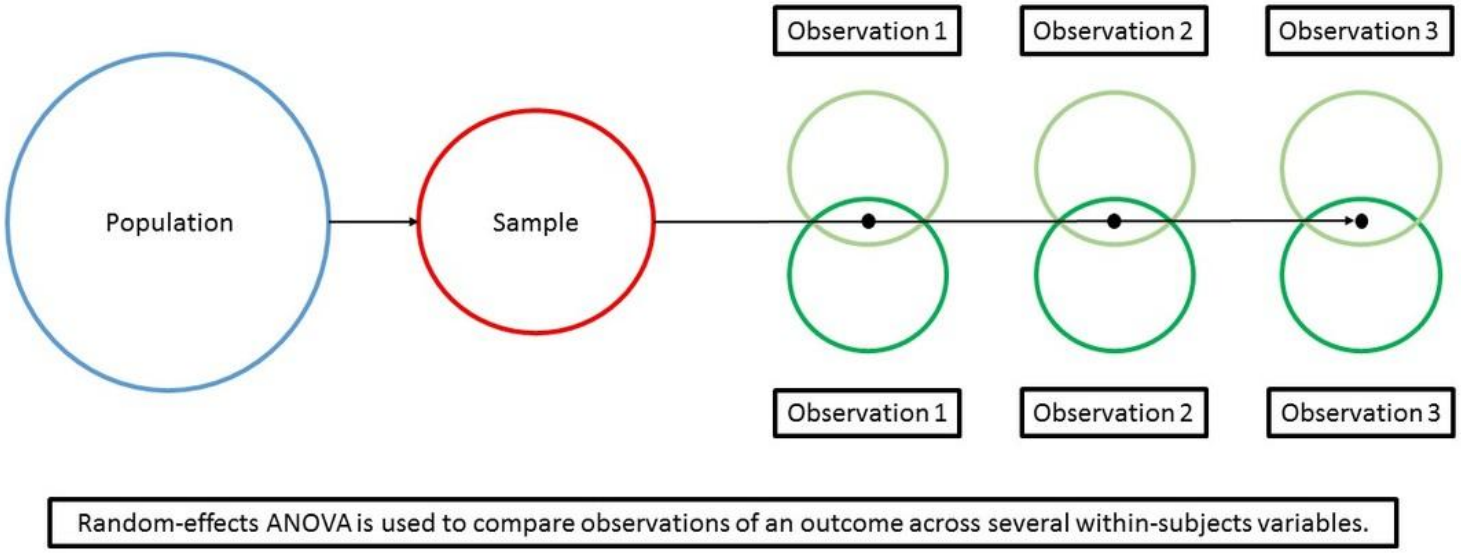
As this involves multiple testing, an appropriate correction should be applied.

Points to Consider:

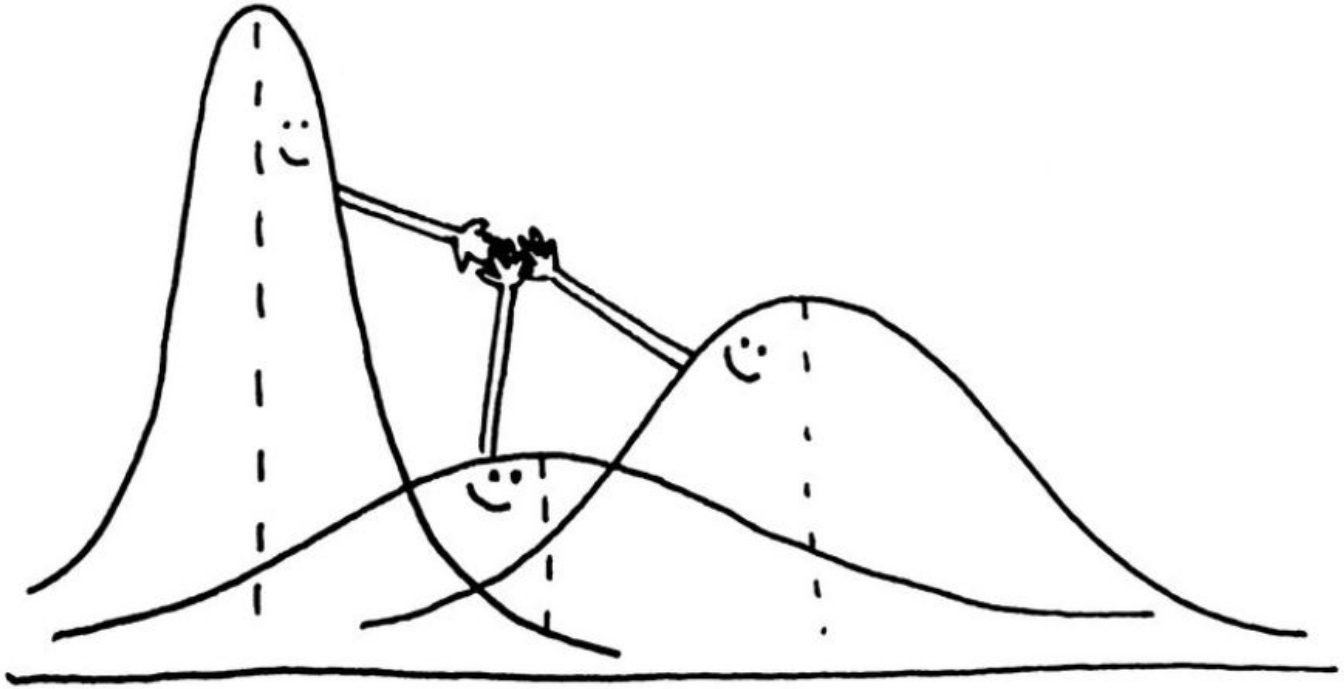
- PERMANOVA takes no account of correlations between variables and any hypothesis that depends on detecting such relationships will not be addressed.
- Nested or hierarchical designs require an appropriate permutational scheme, carefully understanding which objects are truly exchangeable under the null hypothesis. Most importantly, the analyst must define "strata" within which to restrict permutations.
- This method generally assumes balanced designs, however, unbalanced designs can be handled (see McArdle and Anderson, 2001).
- Anderson (2001) warns that groups of objects with different dispersions, yet no significant differences in centres (centres are similar to means, but may be non-Euclidean), may result in misleadingly low P-values. It is thus recommended that the dispersion be evaluated and considered when interpreting the results of PERMANOVA. See Anderson (2006) for a discussion on tests of multivariate dispersion.
- Criticisms of this and other (dis)similarity-based methods should be taken into account (e.g. Warton et al. 2012).

Group comparison of Multiple Groups

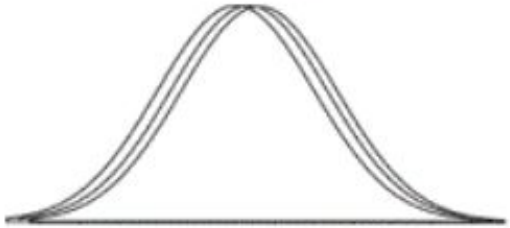
"Analysis of Variance (ANOVA)"



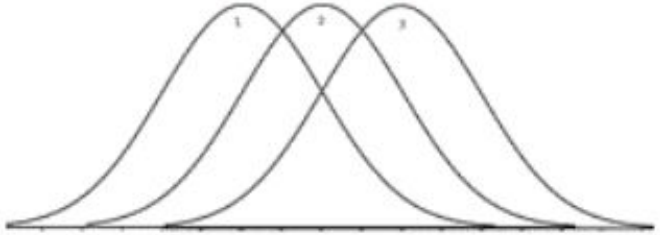
Analysis of Variance (ANOVA) is a parametric statistical technique used to compare datasets. It is similar in application to techniques such as t-test and z-test, in that it is used to compare means and the relative variance between them. However, analysis of variance (ANOVA) is best applied where more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



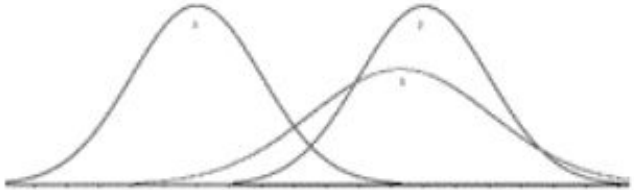
Little discrimination



Some Discrimination



Discrimination between Two Groups, but not the third



Large Discrimination

Source: Psychstat – Missouri State

Assumptions:

1. **Independence of case:** Independence of case assumption means that the case of the dependent variable should be independent of the sample should be selected randomly. There should not be any pattern in the selection of the sample.
2. **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.
3. **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.

Steps for Calculation:

1. Calculate the **sample means** for each of our samples as well as the mean for all of the sample data.
2. Calculate the **sum of squares of error**. The sum of all of the squared deviations is the sum of squares of error, abbreviated SSE.
3. Calculate the **sum of squares of treatment**. We square the deviation of each sample mean from the overall mean. The sum of all of these squared deviations is multiplied by one less than the number of samples we have. This number is the sum of squares of treatment, abbreviated SST.
4. Calculate the **degrees of freedom**. The overall number of degrees of freedom is one less than the total number of data points in our sample, or n - 1. The number of degrees of freedom of treatment is one less than the number of samples used, or m - 1. The number of degrees of freedom of error is the total number of data points, minus the number of samples, or n - m.
5. Calculate the **mean square of error**. This is denoted MSE = SSE/(n - m).
6. Calculate the **mean square of treatment**. This is denoted MST = SST/m - `1.
7. Calculate the **F statistic**. This is the ratio of the two mean squares that we calculated. So F = MST/MSE.

Points to Consider:

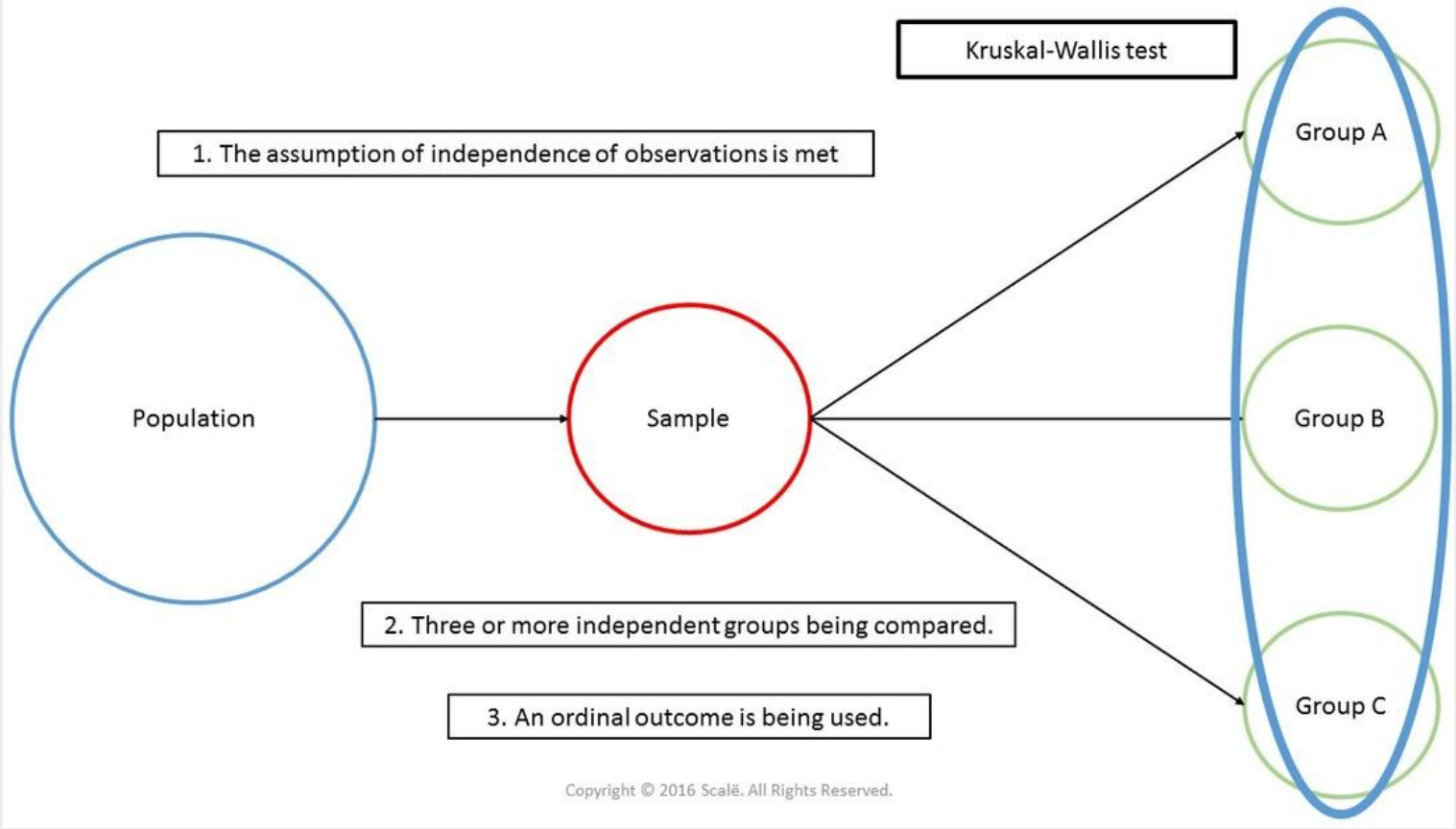
Balanced experiments (those with an equal sample size for each treatment) are relatively easy to interpret; Unbalanced experiments offer more complexity. For single-factor (one-way) ANOVA, the adjustment for unbalanced data is easy, but the unbalanced analysis lacks both robustness and power. The simplest techniques for handling unbalanced data restore balance by either throwing out data or by synthesizing missing data. More complex techniques use regression. ANOVA is (in part) a test of statistical significance. The American Psychological Association (and many other organizations) holds the view that simply reporting statistical significance is insufficient and that reporting confidence bounds is preferred.

While ANOVA is conservative (in maintaining a significance level) against multiple comparisons in one dimension, it is not conservative against comparisons in multiple dimensions.

A common mistake is to use an ANOVA (or Kruskal–Wallis) for analysis of ordered groups, e.g. in time sequence (changes over months), in disease severity (mild, moderate, severe), or in distance from a set point (10 km, 25 km, 50 km). Data in three or more ordered groups that are defined by the researcher should be analyzed by Linear Trend Estimation.

Group comparison of Multiple Groups

"Kruskal-Wallis-Test (H-Test)"



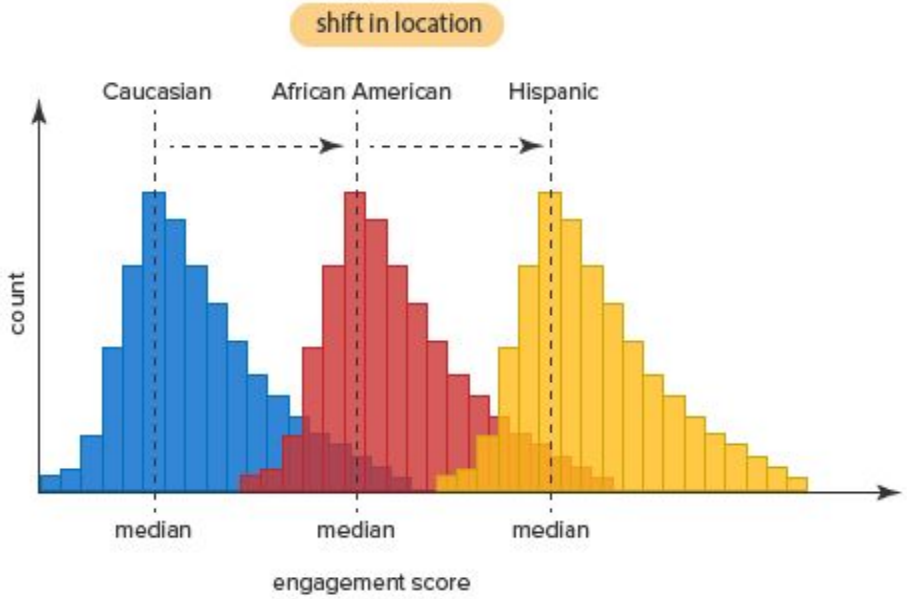
The Kruskal-Wallis H test (sometimes also called the "one-way ANOVA on ranks") is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. It is 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.

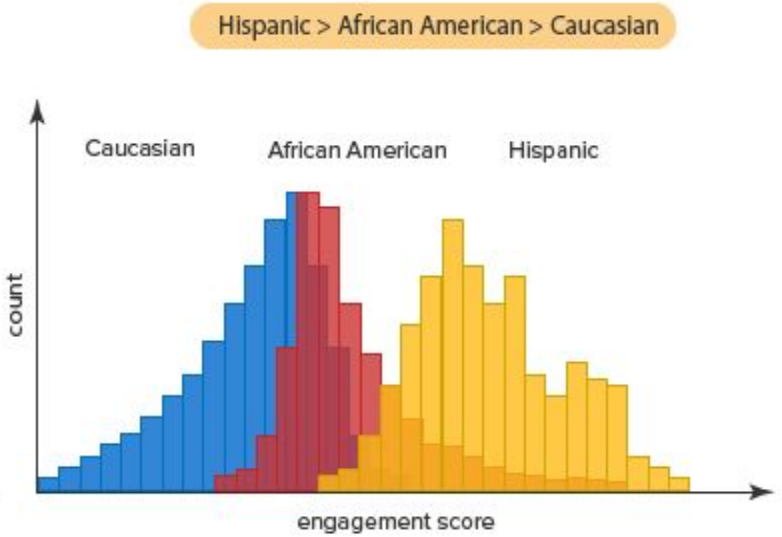
For example, you could use a Kruskal-Wallis H test to understand whether exam performance, measured on a continuous scale from 0-100, differed based on test anxiety levels (i.e., your dependent variable would be "exam performance" and your independent variable would be "test anxiety level", which has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

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.

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
4. In order to know how to interpret the results from a Kruskal-Wallis H test, you have to determine **whether the distributions in each group (i.e., the distribution of scores for each group of the independent variable) have the same shape (which also means the same variability)**. To understand what this means, take a look at the diagram in the next slide:





In the diagram on the left above, the distribution of scores for the "Caucasian", "African American" and "Hispanic" groups have the same shape. On the other hand, in the diagram on the right above, the distribution of scores for each group are not identical (i.e., they have different shapes and variabilities).

Steps for Calculation:

Sample question: A shoe company wants to know if three groups of workers have different salaries:

Women: 23K, 41K, 54K, 66K, 78K.
Men: 45K, 55K, 60K, 70K, 72K
Minorities: 18K, 30K, 34K, 40K, 44K.

1. Sort the data for all groups/samples into ascending order in one combined set. 20K 23K 30K 34K 40K 41K 44K 45K 54K 55K 60K 66K 70K 72K 90K
2. Assign ranks to the sorted data points. Give tied values the average rank. (20K 1) (23K 2) (30K 3) (34K 4) (40K 5) (41K 6) (44K 7) (45K 8) (54K 9) (55K 10) (60K 11) (66K 12) (70K 13) (72K 14) (90K 15)
3. Add up the different ranks for each group/sample.

Women: 23K, 41K, 54K, 66K, 90K = 2 + 6 + 9 + 12 + 15 = 44.
Men: 45K, 55K, 60K, 70K, 72K = 8 + 10 + 11 + 13 + 14 = 56.
Minorities: 20K, 30K, 34K, 40K, 44K = 1 + 3 + 4 + 5 + 7 = 20.

4. Calculate the H statistic. In this case H = 6.72
5. Find the critical chi-square value, with c-1 degrees of freedom. For 3 – 1 degree of freedom and an alpha level of .05, the critical chi-square value is 5.9915.
6. Compare the H value from Step 4 to the critical chi-square value from Step 5. If the critical chi-square value is less than the H statistic, reject the null hypothesis that the medians are equal. If the chi-square value is not less than the H statistic, there is not enough evidence to suggest that the medians are unequal.

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

$$H = \left[\frac{12}{15(15+1)} \left[\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5} \right] \right] - 3(15+1)$$

In this case, 5.9915 is less than 6.72, so you can reject the null hypothesis.

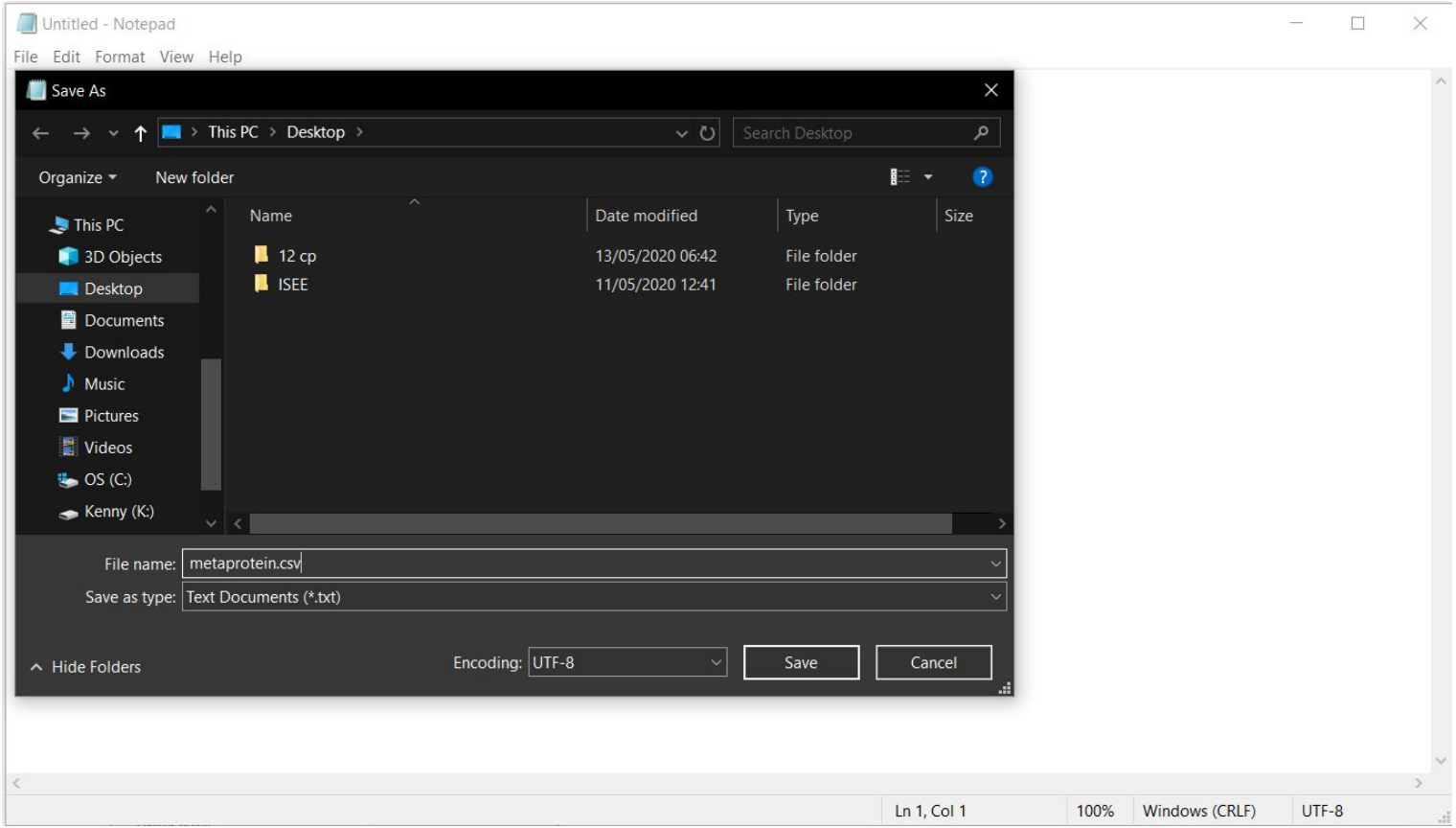
Points to Consider:

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

Group comparison of variables within Multiple Groups

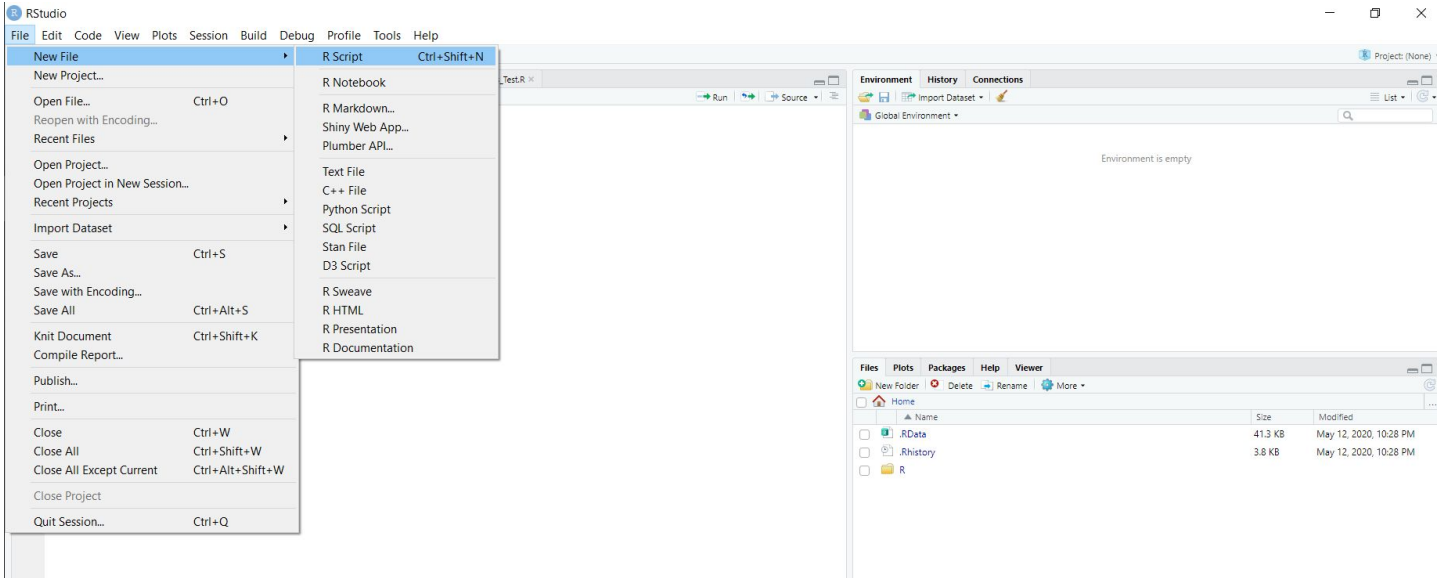
"Implementation of T-Test and Mann-Whitney U-Test in R"

We have implemented the ANOVA and Kruskal-Wallis-Test on a database of metaproteins. You can view the data file from the following link. Copy the contents of the file in a notepad and save as a .csv file



Steps for Implementation

1. Install R and RStudio
2. After you have set up your RStudio, you can get our R script for ANOVA and Kruskal-Wallis Test from here
3. Open a new script in RStudio



4. From the R script that we have provided, copy the whole code and paste it in the new R script that you have created.
5. Hit Run button

Visualization

Tabular Output of our results

	A	B	C	D	E	F	G	H	I
1		BE_03	BE_04	BE_06	BE_08	BE_09	BE_10	BE_40	BE_07
2	Meta-Protein 14920	0	0.00192215	0	0	0.00182017	0	0	0.00161626
3	Meta-Protein 18939	0.00283019	0.01585776	0.01396825	0.01029513	0.0098289	0.01756955	0.01464226	0.00184715
4	Meta-Protein 37176	0.00424528	0.00576646	0	0.00343171	0.00618857	0	0	0
5	Meta-Protein 21610	0.00141509	0.00576646	0	0.00068634	0.00728067	0	0	0
6	Meta-Protein 29454	0.04198113	0.00048054	0.00126984	0.00068634	0.02220604	0.00073206	0	0.00992842
7	Meta-Protein 13169	0.0009434	0.00192215	0	0.00068634	0.00618857	0	0	0.01870238
8	Meta-Protein 25412	0	0.00096108	0	0.00137268	0.00291227	0	0	0
9	Meta-Protein 14408	0	0	0	0	0	0	0	0
10	Meta-Protein 3095	0.00188679	0.00528592	0.01015873	0.00617708	0.00364034	0.01098097	0.0109817	0.00046179
11	Meta-Protein 4786	0.00235849	0.00672754	0.0031746	0.00754976	0.00582454	0.00439239	0.00399334	0.00069268
12	Meta-Protein 9821	0	0	0	0	0	0	0	0.00138536
13	Meta-Protein 14576	0	0	0	0	0	0	0	0.00946664
14	Meta-Protein 4031	0	0	0	0	0	0	0	0
15	Meta-Protein 27644	0	0	0	0	0.00145613	0	0.00366057	0.00253983
16	Meta-Protein 22148	0.0004717	0.00288323	0	0	0.00473244	0	0	0.00623413
17	Meta-Protein 745	0.0009434	0.00240269	0	0	0.0032763	0	0	0
18	Meta-Protein 16846	0.0004717	0.00096108	0	0	0.00182017	0	0	0
19	Meta-Protein 35206	0	0	0	0	0.00036403	0	0	0.00392519
20	Meta-Protein 37487	0.00235849	0.00432484	0.00380952	0.00411805	0.00145613	0.00805271	0.00232945	0
21	Meta-Protein 31477	0	0.00144162	0	0	0.00254823	0	0	0.00046179
22	Meta-Protein 9431	0	0	0	0	0	0	0	0
23	Meta-Protein 19602	0	0	0	0	0	0	0	0.00184715

BE_64	space1	description	space2	p-value (Anc	space3	type 2-1 p-v	type 3-1 p-v	type 2-3 p-v	space4	mean type 1	mean type 2	mean type 3
0.00019945		Metaprotein		0.00455408		0.00304279	0.05450932	0.60033619		0.00044655	0.0060041	0.00466912
0.00329095		Metaprotein		0.00028665		0.00096028	0.0003778	0.69151661		0.01009067	0.0054136	0.00455188
0.00598355		Metaprotein		0.00739123		0.01563371	0.00839671	0.80206496		0.0016905	0.00439175	0.0049018
0.00772875		Metaprotein		0.02514782		0.05070118	0.02501869	0.78826954		0.00165464	0.00420065	0.00479707
0.00344054		Metaprotein		0.04383059		0.16839227	0.995635	0.06866407		0.00701265	0.0031268	0.00722123
0.0038893		Metaprotein		0.3025108		0.65668581	0.27727904	0.598837		0.00237024	0.00373792	0.00502556
0.00129644		Metaprotein		0.1382958		0.14265162	0.18852159	0.99914173		0.00082943	0.00299853	0.00303659
0.00279232		Metaprotein		0.04822316		0.09983898	0.92666462	0.12720687		0	0.00467057	0.00091576
0.00254301		Metaprotein		0.00644933		0.01084349	0.00900273	0.89668202		0.00597849	0.00345073	0.00313046
0.00034904		Metaprotein		0.0949015		0.24573027	0.07816652	0.60871859		0.0042128	0.00326954	0.00279593
0.00553478		Metaprotein		0.00136778		0.04076734	0.00086484	0.11216772		0.00020262	0.00242892	0.00398176
0.00014959		Metaprotein		0.08957683		0.0804068	0.1714752	0.97408927		0.00078889	0.00287622	0.00269914
4.99E-05		Metaprotein		0.00508271		0.00545479	0.33125481	0.13485404		0	0.00309879	0.0015117
0.00029918		Metaprotein		0.00037908		0.01001144	0.00022927	0.14164882		0.00087688	0.00224522	0.00298552
0		Metaprotein		0.14734137		0.1605245	0.18487618	0.99142614		0.00123704	0.00242805	0.00249642
0.00398903		Metaprotein		0.00774699		0.00546965	0.11983558	0.45574181		0.00060637	0.00242625	0.00184927
0.00428821		Metaprotein		0.00215825		0.00301778	0.00446215	0.96775338		0.00027108	0.00198718	0.00209109
0.00344054		Metaprotein		0.04437846		0.03487311	0.2708096	0.59514795		0.00035744	0.00231195	0.00167509
0.0021441		Metaprotein		0.06435827		0.05087355	0.22277594	0.7976161		0.00377889	0.00196794	0.00238173
0.00019945		Metaprotein		0.00601208		0.00508934	0.02266234	0.95008667		0.00037097	0.00198349	0.00185491
0.00074794		Metaprotein		0.00313787		0.00399682	0.37261944	0.08416366		0	0.00238398	0.00106354
0.00114685		Metaprotein		3.29E-06		0.00827829	2.13E-06	0.00244969		0.00015393	0.00150897	0.00281657