Comp 4905
Project Report

Professor:

Robert Biddle, M. Math, PhD

Student:

Cole Douglas
100988503

Table of Contents

Acknowledgements

I am enormously grateful to Professor Robert Biddle for his help and guidance on this project.

I would like to thanks Carleton University Computer Science Department for their expertise and dedication. I would also like to thank the Undergraduate Advisor Edina Storfer for always helping students when they need her. I'm glad I choose Carleton University. I have successfully accomplished my academic studies and feel adequately prepared to join the work force.

Introduction

There is a need in empirical study to statistically compare two data samples for significant difference. The most common tools, such as t-tests, assume continuous data with central tendencies (approx. normality), and we can show such data differences with histograms. However where the data is not continuous or non normal, we must use ordinal tests, such as the Mann-Whitney-Wilcoxen test, but there is no way to show such data that illustrates the differences. Current data visualization techniques are inadequate for ordinal data.

Ordinal data can be considered contiguous or even nominal, but it isn't. When you convert it, the structure changes to favour the type of data to which you converted it.

For example, take a bar chart. If you assign value to ranks, you create the mistaken impression that the ranks are equally spaced apart. What you need is a visualization that shows the data structure without falsely implementing a sense of value.

Ideally, I would create and implement a new visualization which allows a user to analyze ordinal data in it's pure form.

I will use my time to follow a basic three step waterfall method.

Step 1: Design a prototype

- Figure out what works and what doesn't

Step 2: Test it on some users

- Have some people try the prototype to see if it has a friendly UX

Step 3: Collate results

- See what worked and what didn't

Technical Information

Tools

For this project, I used the R programming language. I used R because it is industry standard for this type of problem.

I assumed the user would be storing their data on a CSV file. R allows easy data manipulation if they do not.

Installation Instructions

An unzipped copy of the files can be found at https://github.com/ColeDouglas/Comp4905Backup

To run, import the "OrdinalViewGraph.R" package, in your preferred manner. You may then use the ordinal_view_graph() function as you need. See the Man Page for more details. The package also comes with a Testing() function, or you can see Appendix C for the my last set of results.

To install manually:

Step 1. Locate your local R/*xxx*-library folder.

- If windows, then xxx is windows, etc

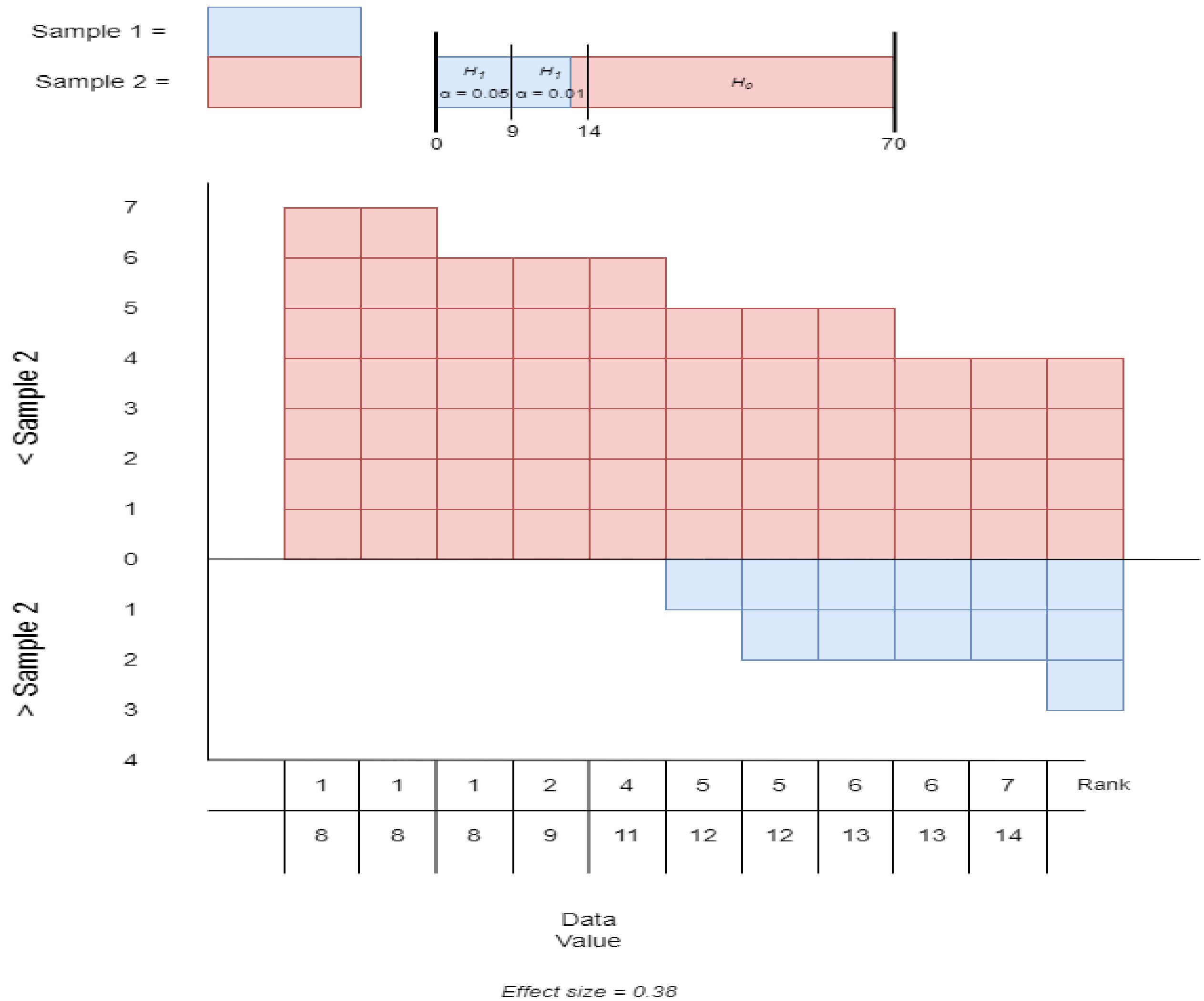
Step 2. insert the OrdinalViewGraph folder into your current version of R

To install using a package manager:

Step 1. Open OrdinalViewGraph.R

Step 2. Use the install command.

Solution
The Ordinal View Graph

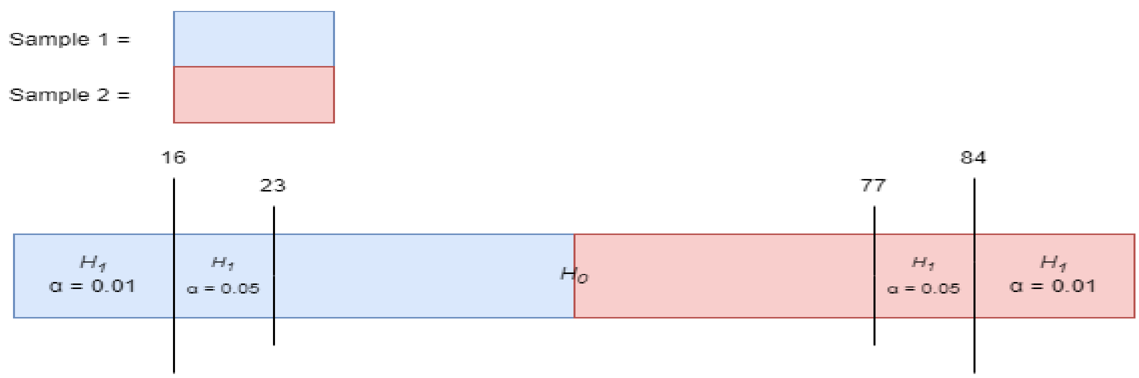


The ordinal view graph solves all the mentioned problem in one intuitive visualization.
The running bar chart lets the user see the data spread without giving a false sense of value.
The test statistic in the upper left corner let's the user see whether or not to reject the null hypothesis at several values of α.
Together the user has a complete understanding of the structure of the data, without being misinformed about it's nature.

Unused Prototypes

During the evolution of the project, several other prototypes were explored.

Prototype Alpha:



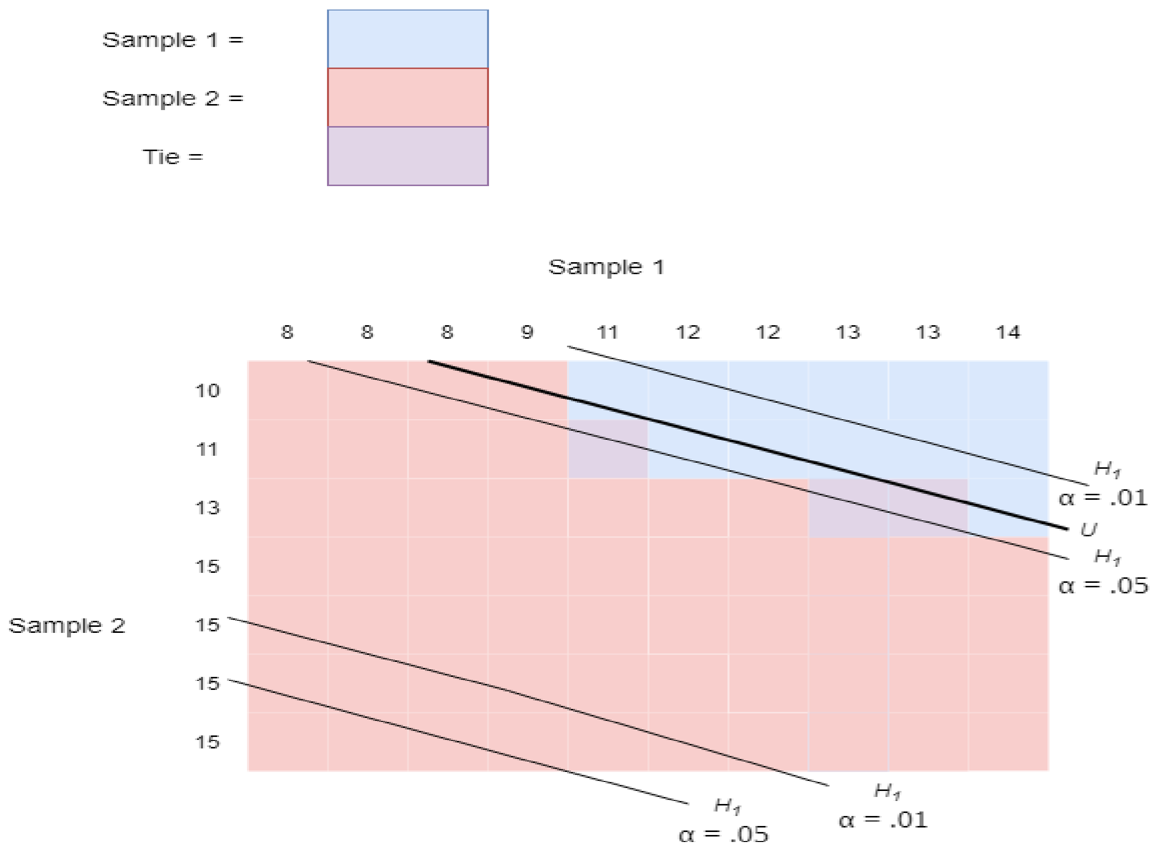
This was the immediate obvious solution. While it does provide the test statistic easily, it contains no additional information on the shape of the data.

User Feedback:

Users found it immediately helpful to know whether or not to reject the null hypothesis. They were unable to identify what this means datawise however. They were also confused about why there is two sets of critical values.

Using this feedback, I made sure to display some data structure the user can interpret. I also tried to emphasis that the test statistic is a measure of contrast between to samples.

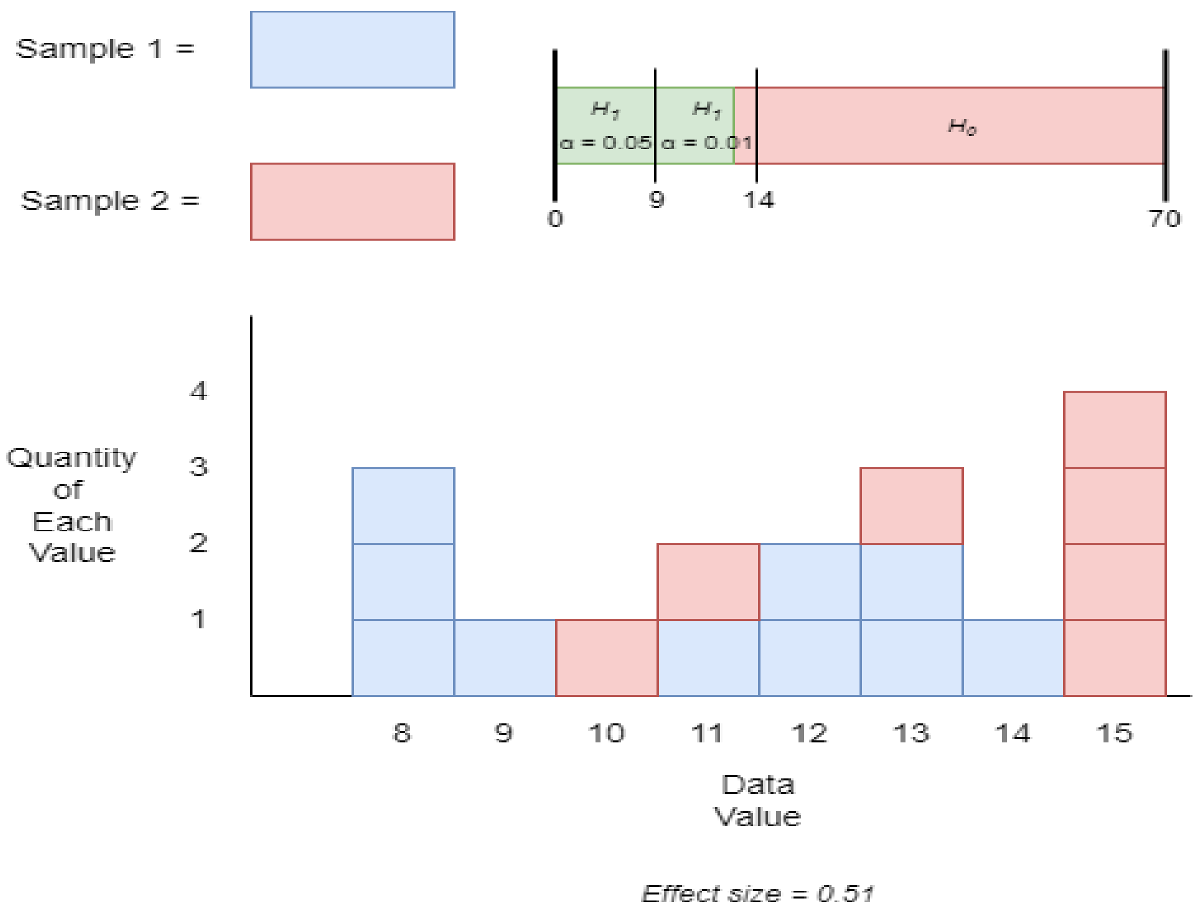
Prototype Beta:



This is the next logical step to solving the problem. All the information is present. Unfortunately, the data was proven to be unintuitive to understand.

User Feedback:
Users were less inclined to understand this chart despite the fact it contains all relevant data. I predict information overload. Also it is not comparable to familiar data structures so the user didn't have any precedent experience from which to draw.

Prototype Gamma:



This solved several major issues with Prototype logical step to solving the problem. All the information is present. Unfortunately, instances with a large number of ties created ambiguity. A solution which removed ties would have to be found.

User Feedback:
Users understand the shape of the data. It is not any graph they were familiar with, but they intuitively understood how each sample was ranked compared to the other.

Examples

Example 1:

Edge Case - Complete Overlap

In this example, you can see two hypothetical satisfaction surveys done on a group of students before and after a change was made to their curriculum.

U_1 = 50,

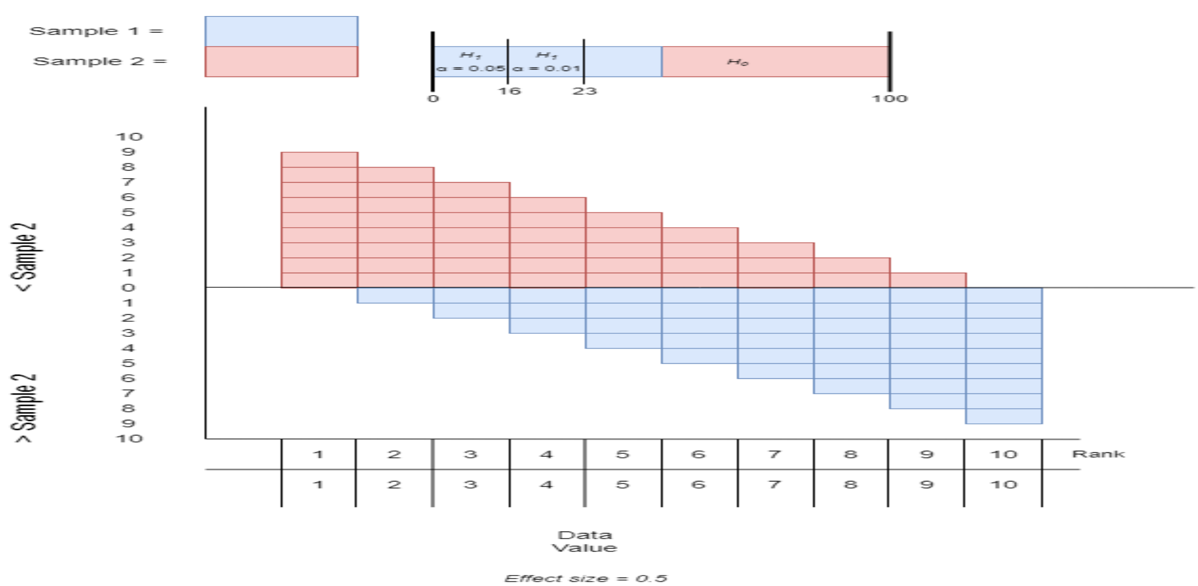
This means the likelihood of the pre-change survey value outranking the post change survey value is fifty over one hundred

U_2 = 50,

This means the likelihood of the post change survey value outranking the pre-change survey value is fifty over one hundred

Critical value = 23.

In order to reject the null hypothesis, one of the test statistics would have to be less than twenty three.



Since neither of the samples is in the H_1 range, we cannot reject the null hypothesis. There is insufficient evidence to assume these samples are from different populations. Meaning that it's unlikely the curriculum change effected the student's satisfaction.

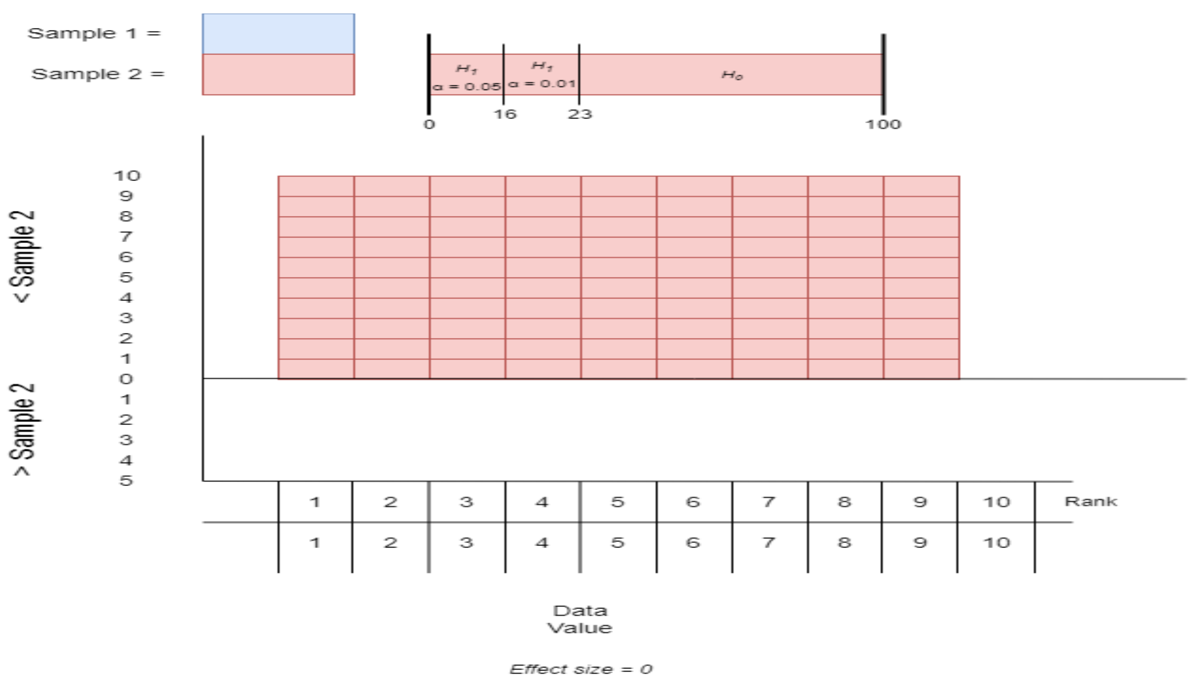
Example 2:
Edge Case - Complete Separation

In this example, you can see two sets of running times done on a group of athletes before and after a change was made to their foot-ware.

$U_1 = 0$,
This means the likelihood of the new shoe time outranking the old shoe time is zero over one hundred

$U_2 = 100$,
This means the likelihood of the old shoe time outranking the new shoe time is one hundred over one hundred

Critical value = 23.
In order to reject the null hypothesis, one of the test statistics would have to be less than twenty-three.



Since one of the samples is in the H_1 range, we can reject the null hypothesis. There is sufficient evidence to assume these samples are from different populations. Meaning that it's likely that changing the athletes shoes effected their speed.

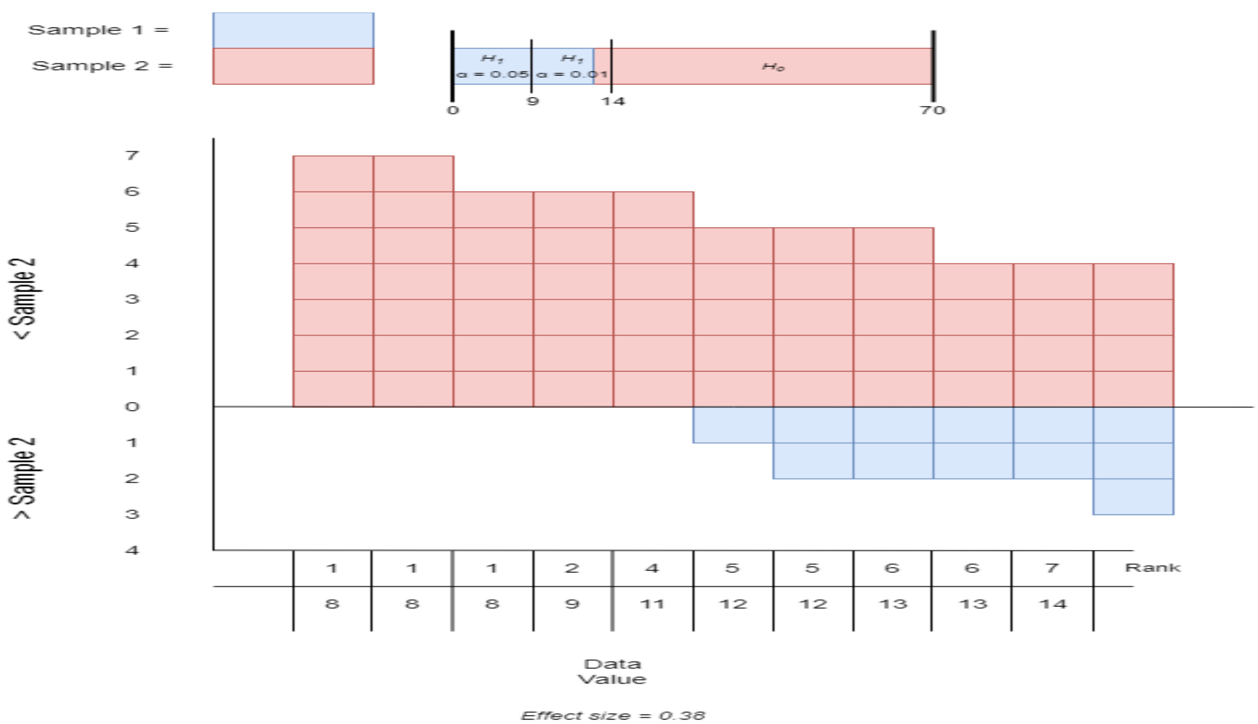
Example 3:
Edge Case - Reject the null hypothesis
at α = 0.05 only.

In this example, you can see two sets of responses on an eight-point Likert scale done on a group of customers before and after a change was made to a product.

$U_1 = 13.5$,
This means the likelihood of the old product response outranking the new product response is thirteen and a half over seventy.

$U_2 = 56.5$,
This means the likelihood of the new product response outranking the old product response is fifty-six and a half over seventy.

Critical value = 14.
In order to reject the null hypothesis, one of the test statistics would have to be less than fourteen.



Since one of the samples is in the H_1 range where α = 0.05, we can reject the null hypothesis. It's safe to assume the new product is preferred to the old one.
However it is only safe to assume so at a probability $p < .05$.

Conclusions

There is no current methods to properly visualize the aforementioned data. I have determined this visualization to be the most user-friendly way to present this data

The ordinal view graph is a new and efficient data visualization which solves all the mentioned problems. It can be implemented immediately. When I ran a few test trials with known and educated associates of mine, they found the chart intuitive albeit busy. I solved all the major problems I set out to accomplish.

In the future, making the graph also output a weighted vector could be useful for recommender systems

Appendices

Appendix A:
Wilcoxon/Mann-Whitney U Test

Explanation:

The Wilcoxon/Mann-Whitney U test is used to test differences in distributions of data from two independent samples. The basic idea is that you would want to know if the two samples come from the same population, or if a variable has separated them.

First, you put the two independent samples together. You rank them appropriately. Then you compare the median of the relative ranks using the test statistic (sort of, but not really).

The test statistic is comparing two probabilities. If you were to pull one result x from treatment X and compare it to one result y from treatment Y, there will be a probability P_1 that $x > y$ and a probability P_2 that $y > x$. The test statistic is how likely that $P_1 = P_2$.

Requirements:

Similarly shaped data, to prevent false positives.
Independent
Ranked, or continuous.

Formulas:

The Test Statistic

The smaller of:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Example:

There are two treatments. The researcher is interested in if they are different.

H_0: The two treatments come from the same population.
H_1: One of the treatments has created a new population.

The samples are:

	Day 1	Day 2	Day 3	Day 4	Day 5
Treatment 1	15	6	9	12	5
Treatment 2	12	10	6	5	6

Step 1: Add them together

Treatment	1	2	1	2	1	2	1	2	1	2
Result	15	12	6	10	9	6	12	5	5	6

Step 2: Rank them

Ordered Results		Ranks	
Treatment 1	Treatment 2	Treatment 1	Treatment 2
5	5	1.5	1.5
6	6	4	4
	6		4
9		6	
	10		7
12	12	8.5	8.5
15		10	

Step 4: Calculate the Test Statistic

Treatment 1 Total Rank = (1.5 + 4 + 6 + 8.5 + 10) = 30
Treatment 2 Total Rank = (1.5 + 4 + 4 + 7 + 8.5) = 25

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_2$$

$$U_1 = (5)(5) + \frac{5(6)}{2} - 30$$

$$U_1 = 10$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_1$$

$$U_2 = (5)(5) + \frac{5(6)}{2} - 25$$

$$U_2 = 15$$

What does this mean?

For the Wilcoxon/Mann-Whitney test, the test statistic is the number of results from one treatment which exceed a possible number from the other. See the following graph for assistance.

If I were to pull one random result from treatment 1 and compare it to one random result from treatment 2, here are the 25 possibilities.

		Treatment 2				
Treatment 1		5	6	6	10	12
	5	Equal	Treatment 1	Treatment 1	Treatment 1	Treatment 1
	6	Treatment 2	Equal	Equal	Treatment 1	Treatment 1
	9	Treatment 2	Treatment 2	Treatment 2	Treatment 1	Treatment 1
	12	Treatment 2	Treatment 2	Treatment 2	Treatment 2	Equal
	15	Treatment 2	Treatment 2	Treatment 2	Treatment 2	Treatment 2

There are 8 outcomes where treatment 1 is less than treatment 2, so $U_1 = 8$
There are 13 outcomes where treatment 2 is less than treatment 1, so $U_2 = 13$
There are 4 outcomes where treatment 2 equals treatment 1, so both U_1 and U_2 gain 4(.5) = 2

$U_1 = 8 + 2 = 10$
$U_2 = 13 + 2 = 15$

∴ $U_1 = 10$ and $U_2 = 15$

If $U_n = 0$, that means every result of treatment *n* is smaller than every other result of the other. A complete split in samples.
The closer to 50% U is, the more similar the samples.
Since $U_1 < U_2$, the value of the test statistic is 10.

Step 5. Compare the test statistic to a critical value.

The test statistic is 10. This is a two-tailed test with 5 results for each sample. Which means the critical value equals 2. Since 10 > 2, we cannot reject the null hypothesis and should conclude that these two samples are from the same population.

Appendix B:
Bibliography

Understanding R
https://www.tutorialspoint.com/r

Understanding of the Mann Whitney U-Test:
https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/BS704_Nonparametric4.html

https://en.wikipedia.org/wiki/Mann–Whitney_U_test

Appendix C:
Test Suite

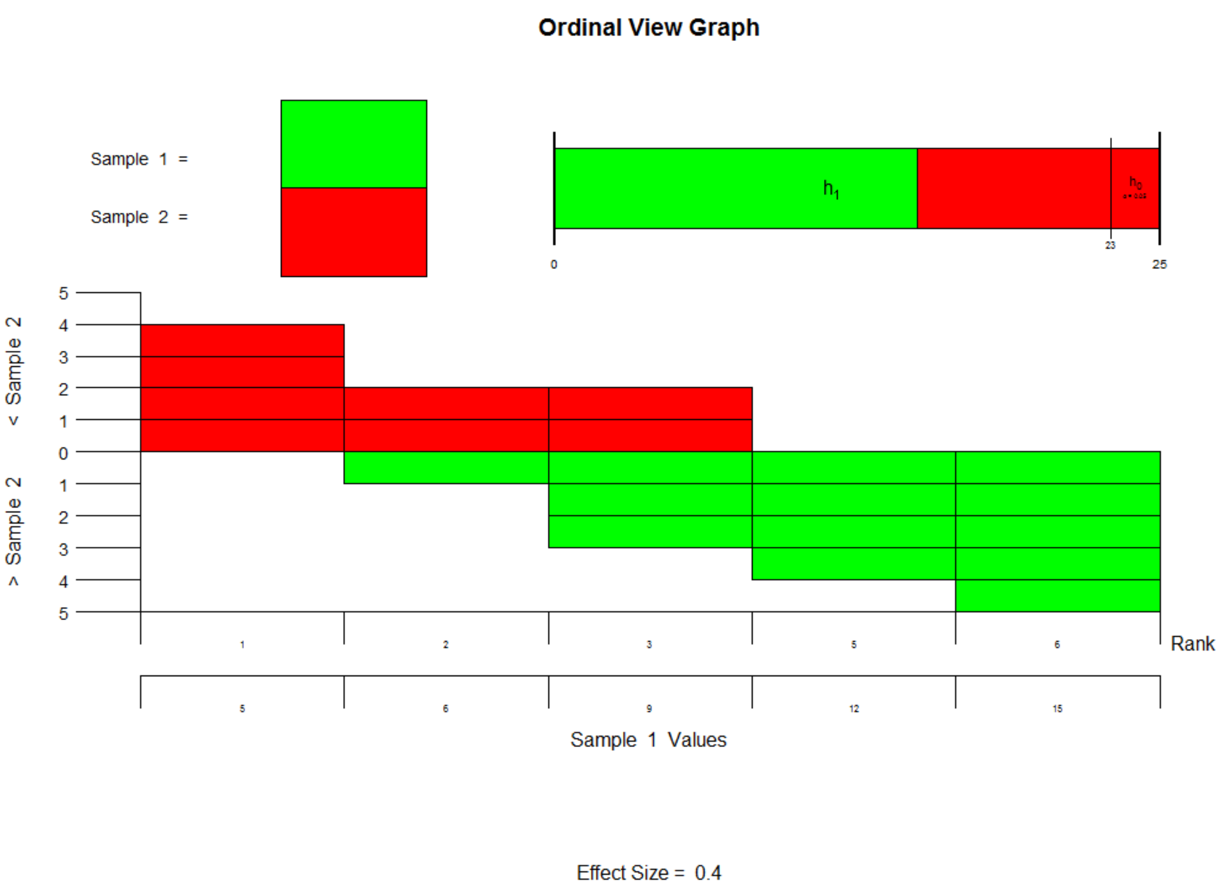
Test Number	Description
Examples from the Report	
1	Appendix B
2	Example 1 - Complete Overlap
3	Example 2 - Complete Separation
4	Example 3 - Reject the Null hypothesis
Randomized Data	
5	Same Population
6	Different Population
7	50/50 either way
Large Data Sets	
8	Large - Same Population
9	Large - Different Population
10	Large - 50/50 either way
Different P-values	
11	Two tailed esoteric P-value
12	One tailed esoteric P-value
13	Two tailed random P-value
14	One tailed random P-value
Categorical Interpretations	
15	blue < green < red < yellow
16	green < yellow < blue < red
Pairwise Analysis	
17	Three Data Sets
18	Four Data Sets
19	Five Data Sets

Test 1:
Example from Appendix B

Data:
Sample 1: (15, 6, 9, 12, 5)
Sample 2: (12, 10, 6, 5, 6)

Parameters:
ordinal_view_graph(Data)

Output:

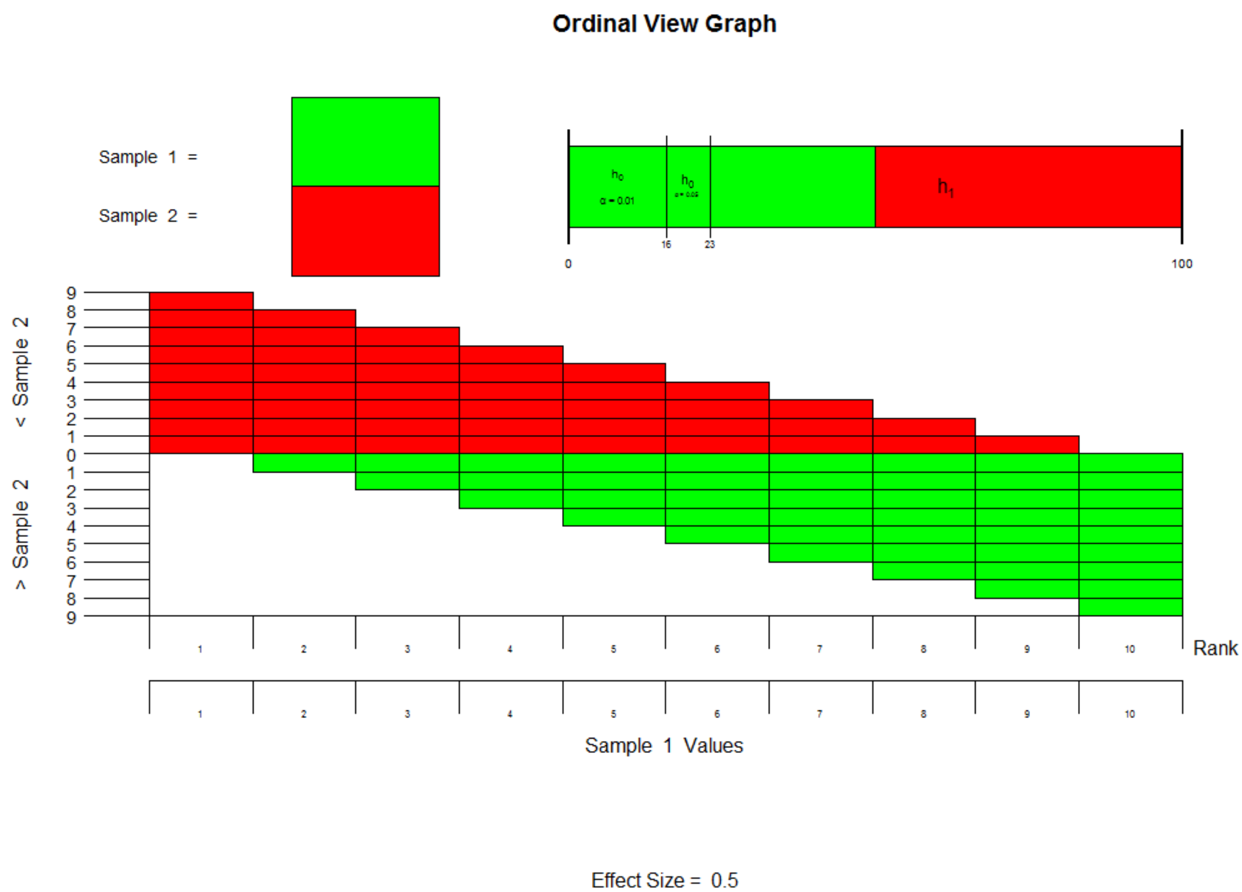


Test 2: Example 1
Complete Overlap

Data:
Sample 1: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 2: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Parameters:
ordinal_view_graph(Data)

Output:

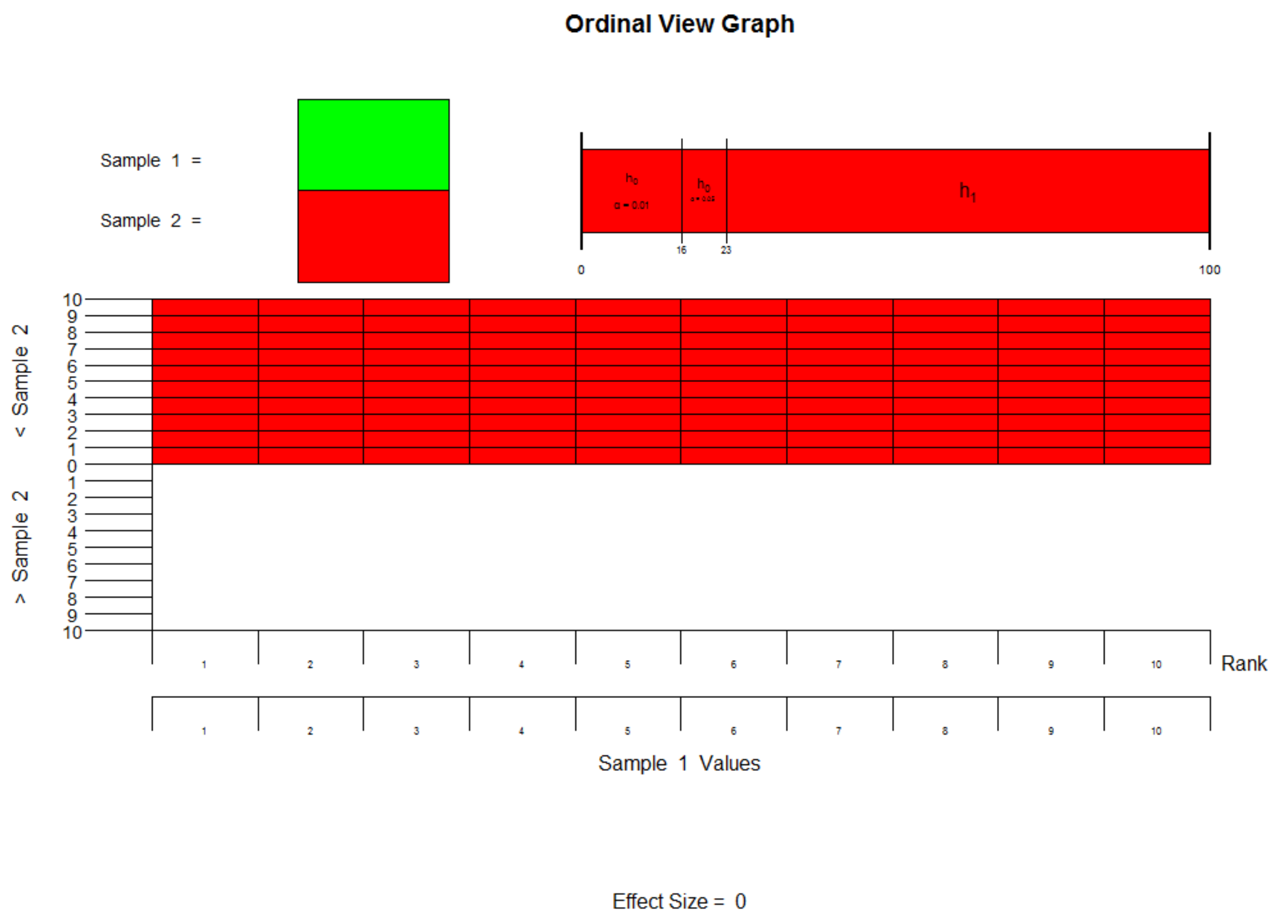


Test 3:
Complete Separation

Data:
Sample 1: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 2: (11, 12, 13, 14, 15, 16, 17, 18, 19, 20)

Parameters:
ordinal_view_graph(Data)

Output:

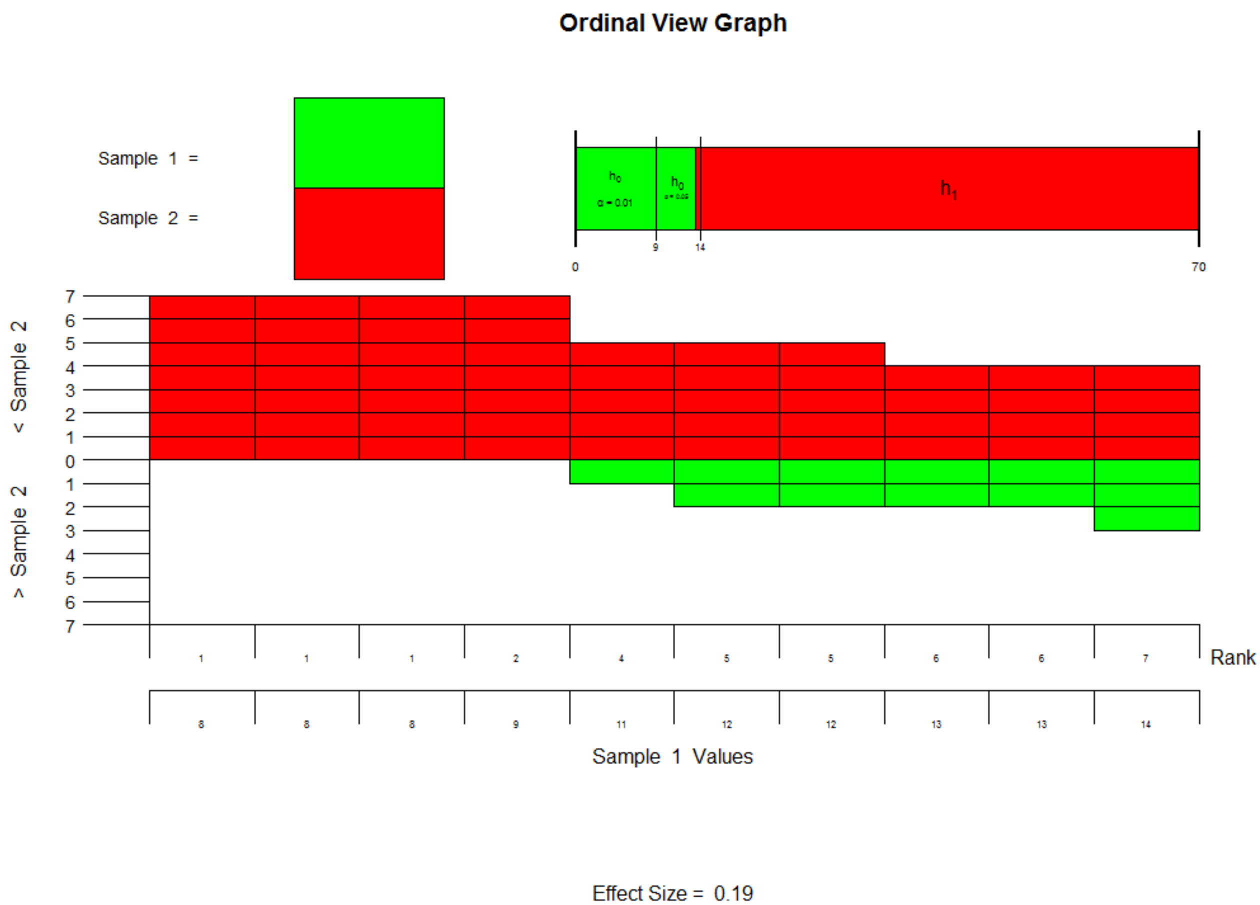


Test 4:
Reject the null hypothesis

Data:
Sample 1: (8, 8, 8, 9, 11, 12, 12, 13, 13, 14)
Sample 2: (10, 11, 13, 15, 15, 15, 15)

Parameters:
ordinal_view_graph(Data)

Output:

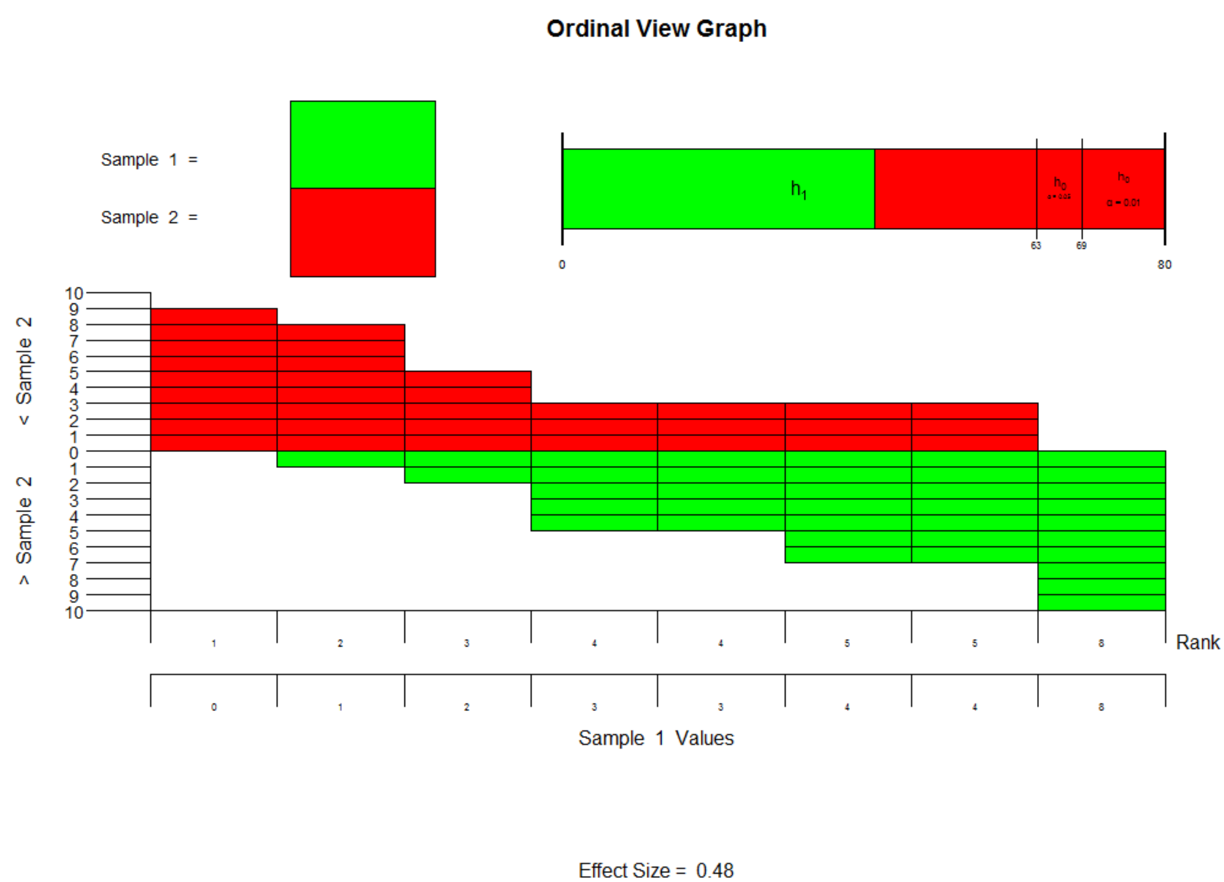


Test 5:
Random - Same Population

Data:
Sample 1: A set of 5 to 15 values between 1 and 20
Sample 2: A set of 5 to 15 values from the same population as Sample 1

Parameters:
ordinal_view_graph(Data)

Output:

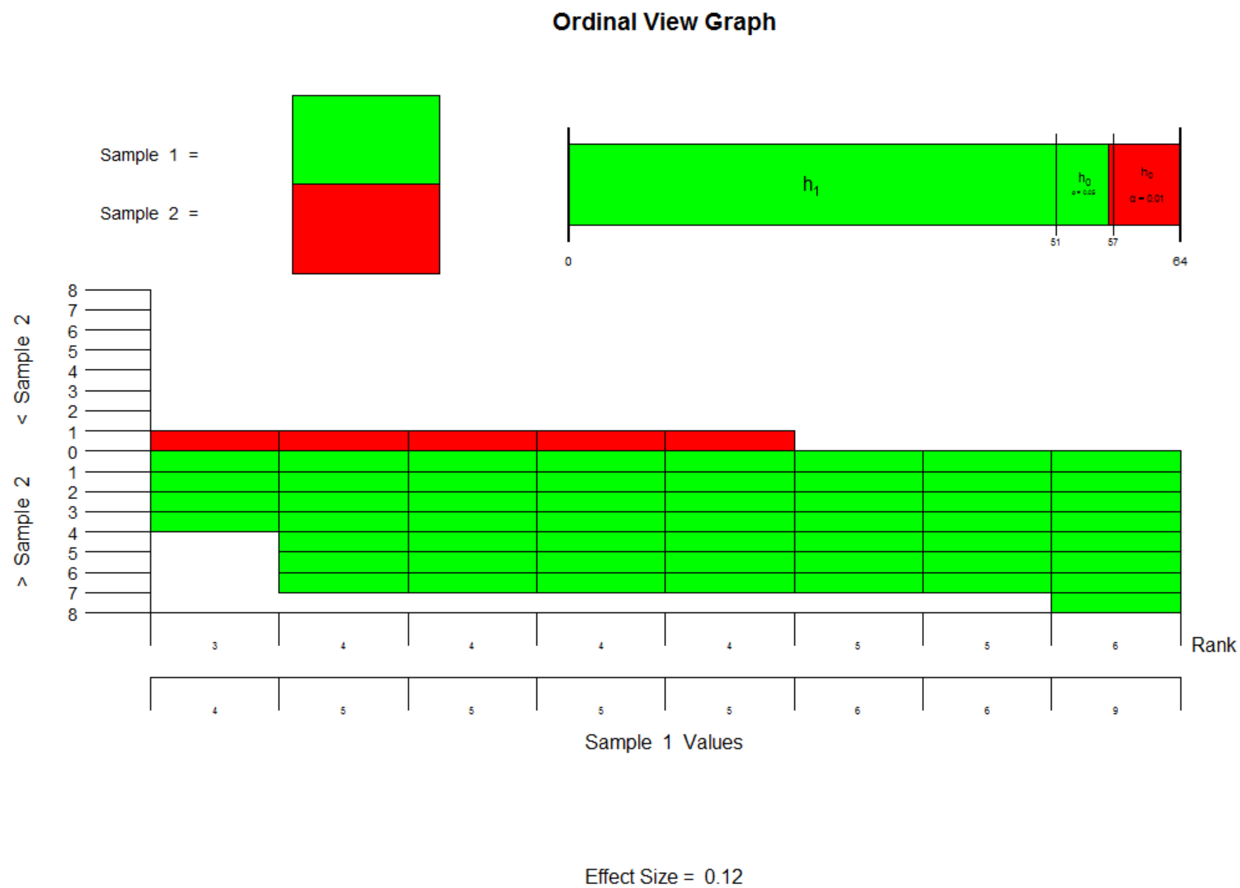


Test 6:
Random - Different Population

Data:
Sample 1: A set of 5 to 15 values between 1 and 20
Sample 2: A set of 5 to 15 values from the definitely separate population as Sample 1

Parameters:
ordinal_view_graph(Data)

Output:

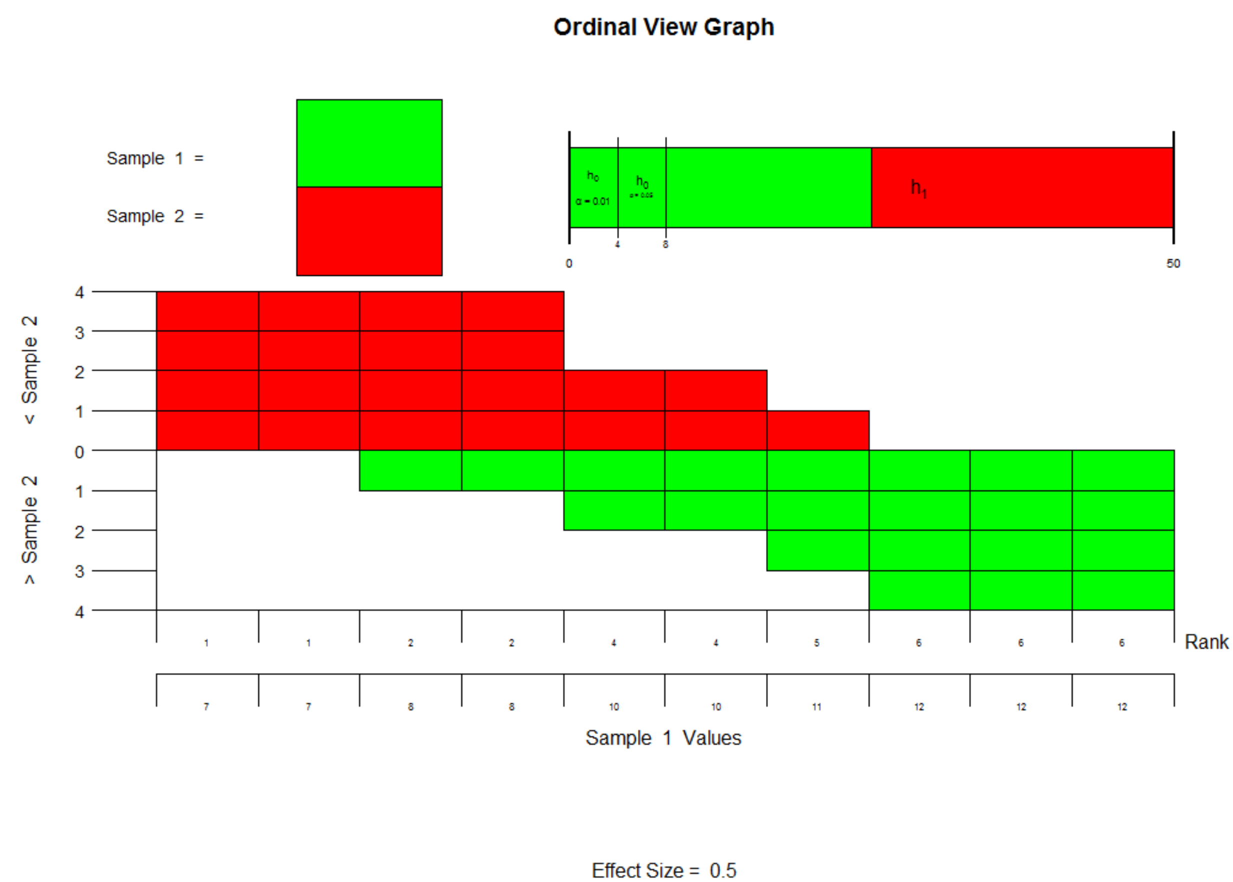


Test 7:
Random - 50/50 percent chance either way

Data:
  Sample 1: A set of 5 to 15 values between 1 and 20
  Sample 2: A set of 5 to 15 values which may or may not be the same population as sample 1

Parameters:
 ordinal_view_graph(Data)

Output:

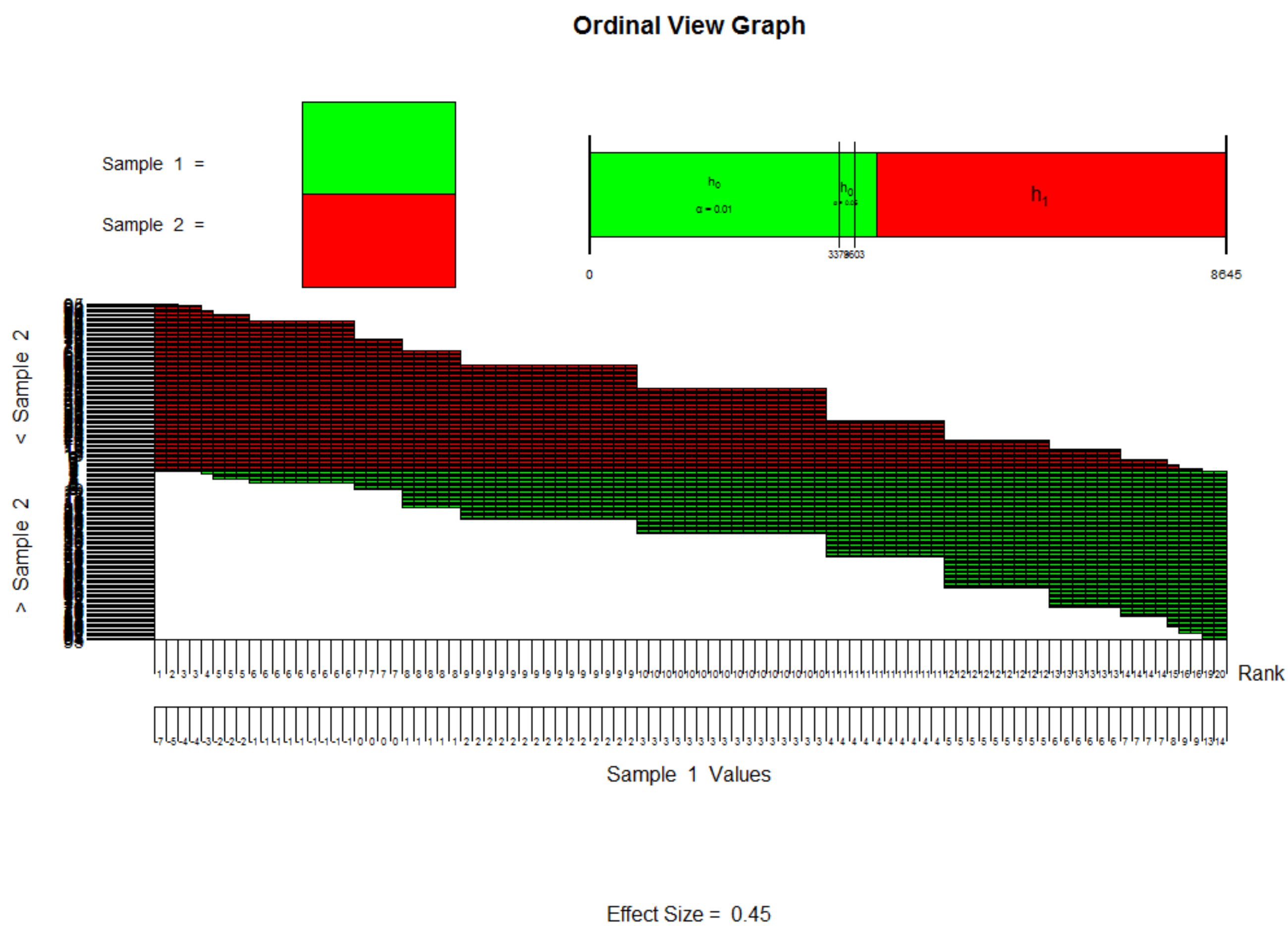


Test 8:
Large Sample Size - Same Population

Data:
Sample 1: A set of 50 to 100 values between 1 and 20
Sample 2: A set of 50 to 100 values from the same population as Sample 1

Parameters:
ordinal_view_graph(Data)

Output:

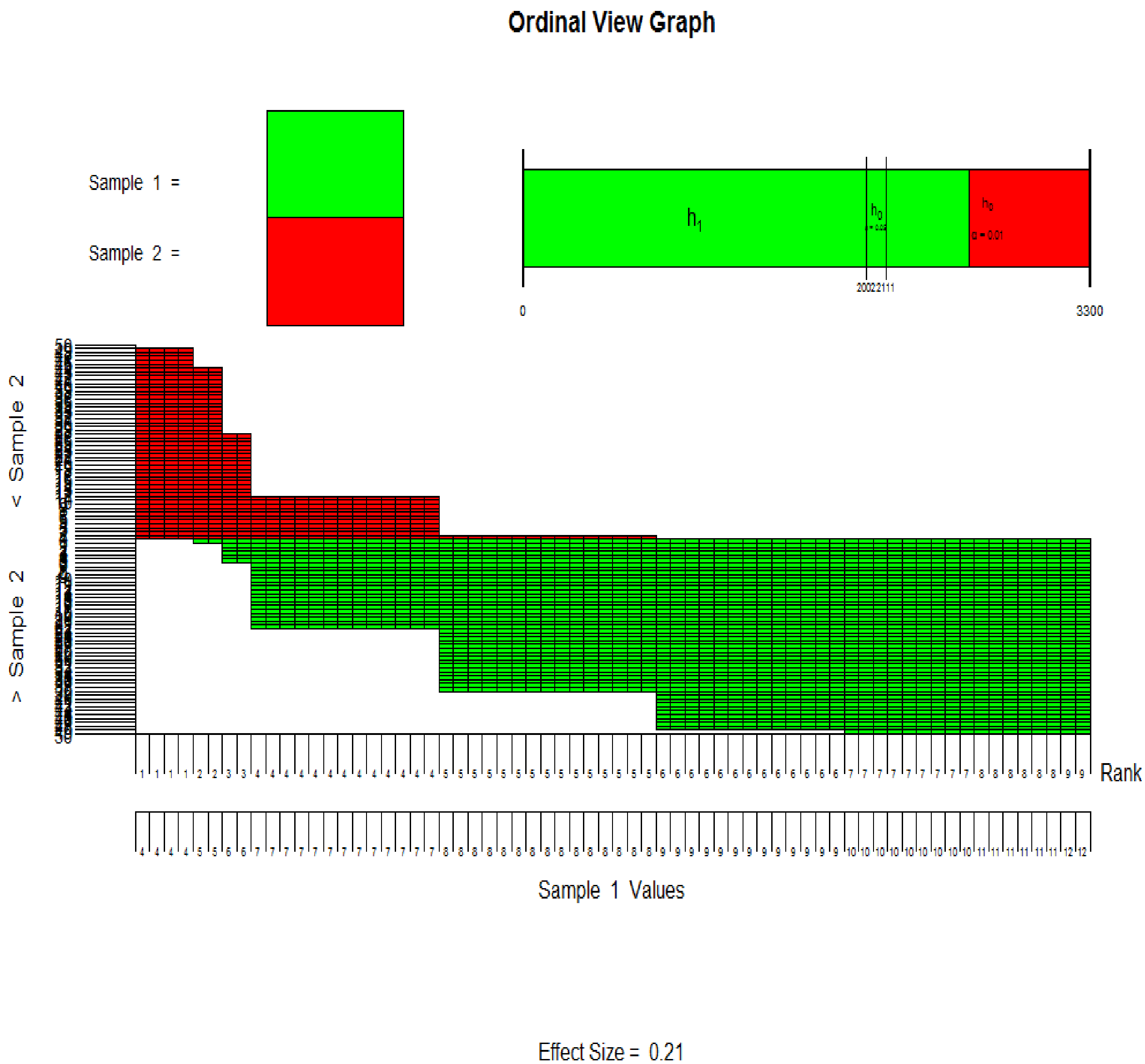


Test 9:
Large Sample Size - Different Population

Data:
Sample 1: A set of 50 to 100 values between 1 and 20
Sample 2: A set of 50 to 100 values from the definitely separate population as Sample 1

Parameters:
ordinal_view_graph(Data)

Output:

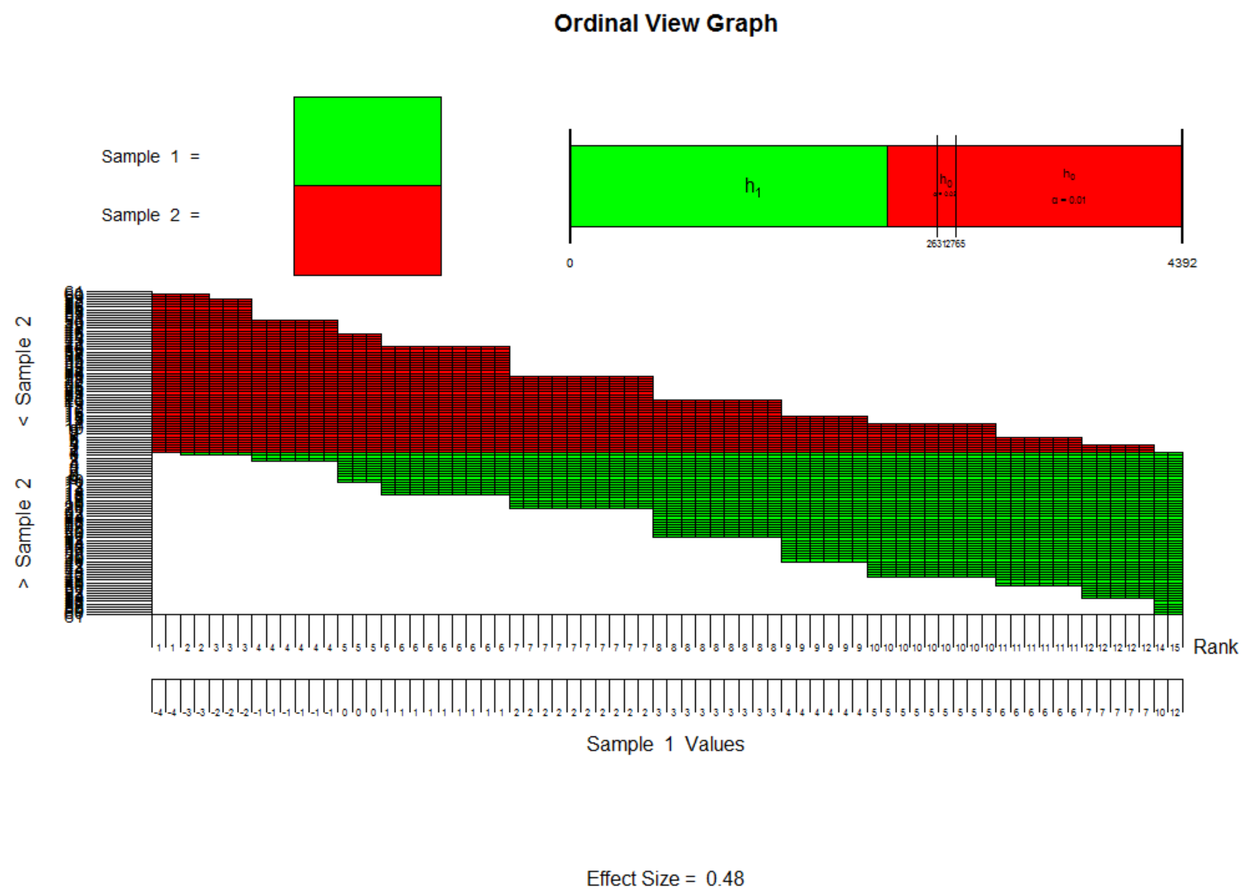


Test 10:
Large Sample Size - 50/50 percent chance either way

Data:
  Sample 1: A set of 50 to 100 values between 1 and 20
  Sample 2: A set of 50 to 100 values which may or may not be the same population as sample 1

Parameters:
 ordinal_view_graph(Data)

Output:

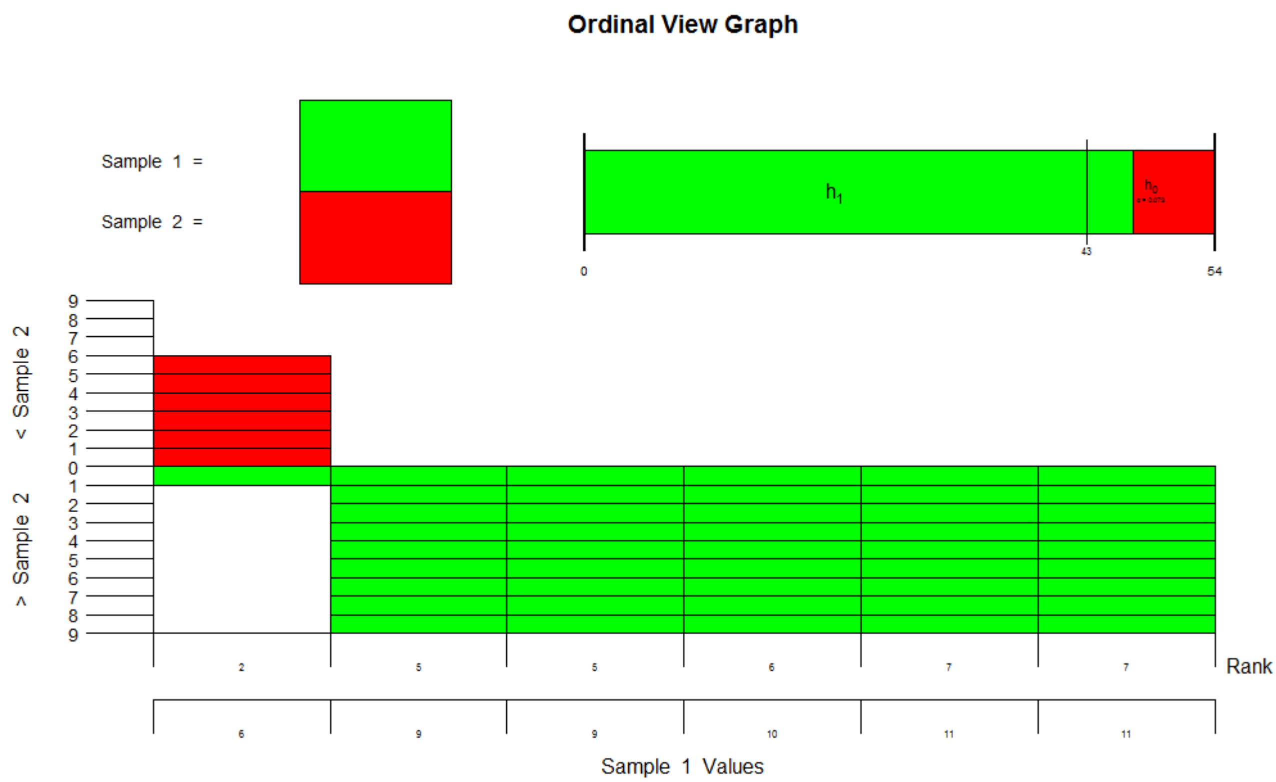


Test 11:
Two tailed esoteric P-value

Data:
  Sample 1: A set of 5 to 15 values between 1 and 20
  Sample 2: A set of 5 to 15 values which may or may not be the same population as sample 1

Parameters:
  ordinal_view_graph(Data, 0.073)

Output:



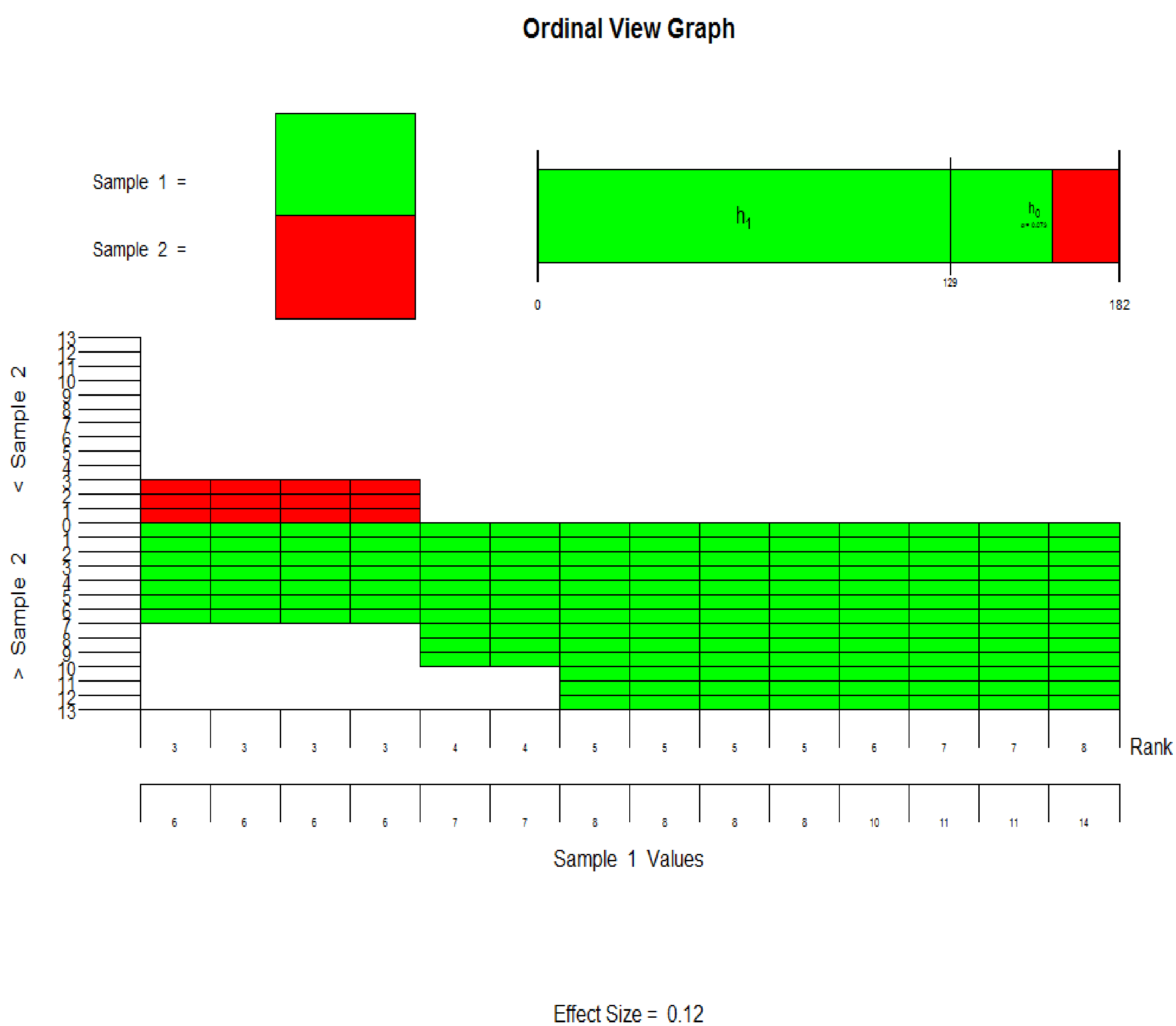
Effect Size = 0.13

Test 12:
One tailed esoteric P-value

Data:
Sample 1: A set of 5 to 15 values between 1 and 20
Sample 2: A set of 5 to 15 values which may or may not be the same population as sample 1

Parameters:
ordinal_view_graph(Data, 0.073, 1)

Output:



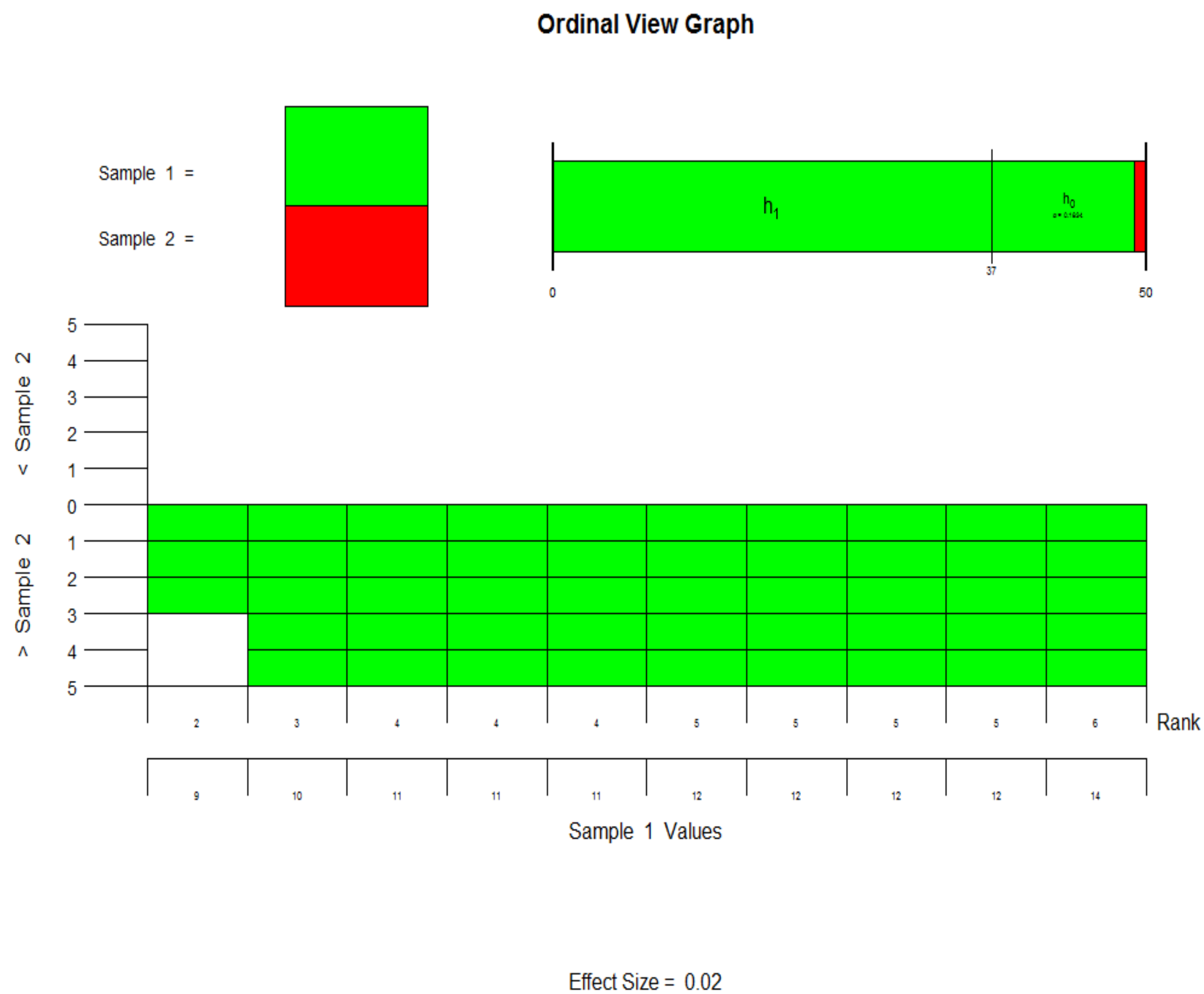
Test 13:
Two tailed random P-value

Data:
Sample 1: A set of 5 to 15 values between 1 and 20
Sample 2: A set of 5 to 15 values which may or may not be the same population as sample 1

Random Probability: a number between 0.01 and 0.25

Parameters:
ordinal_view_graph(Data, Random Probability)

Output:



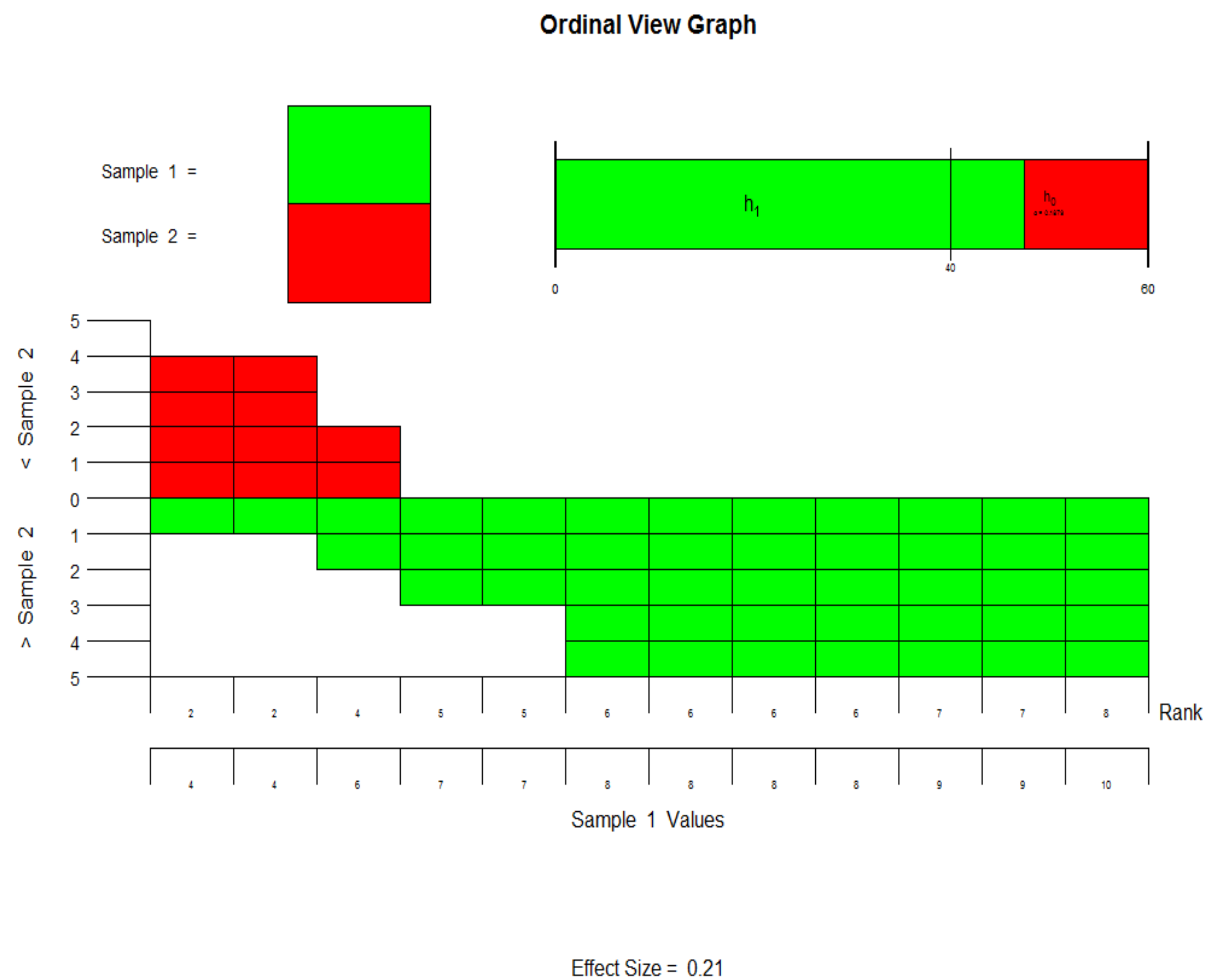
Test 14:
One tailed random P-value

Data:
Sample 1: A set of 5 to 15 values between 1 and 15
Sample 2: A set of 5 to 15 values which may or may not be the same population as sample 1

Random Probability:
a number between 0.01 and 0.25

Parameters:
ordinal_view_graph(Data, Random Probability, 1)

Output:



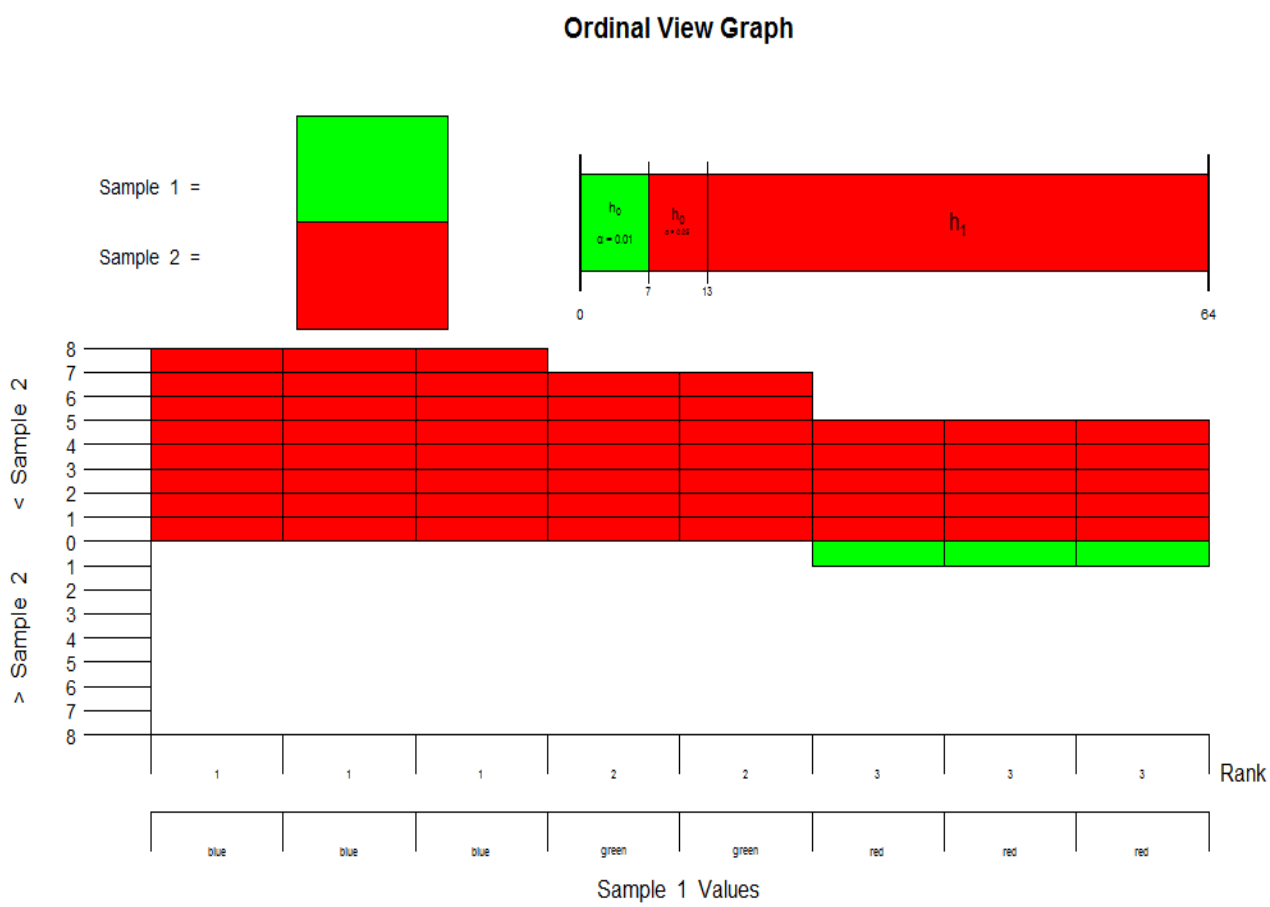
Test 15:
blue < green < red < yellow

Data:
Sample 1: (green, blue, red, red, blue, red, green, blue)
Sample 2: (yellow, yellow, red, yellow, red, yellow, yellow, green)

Rank Vector:
(Blue, Green, Red, Yellow)

Parameters:
ordinal_view_graph(Data, 0.05, 2, Rank Vector)

Output:



Effect Size = 0.11

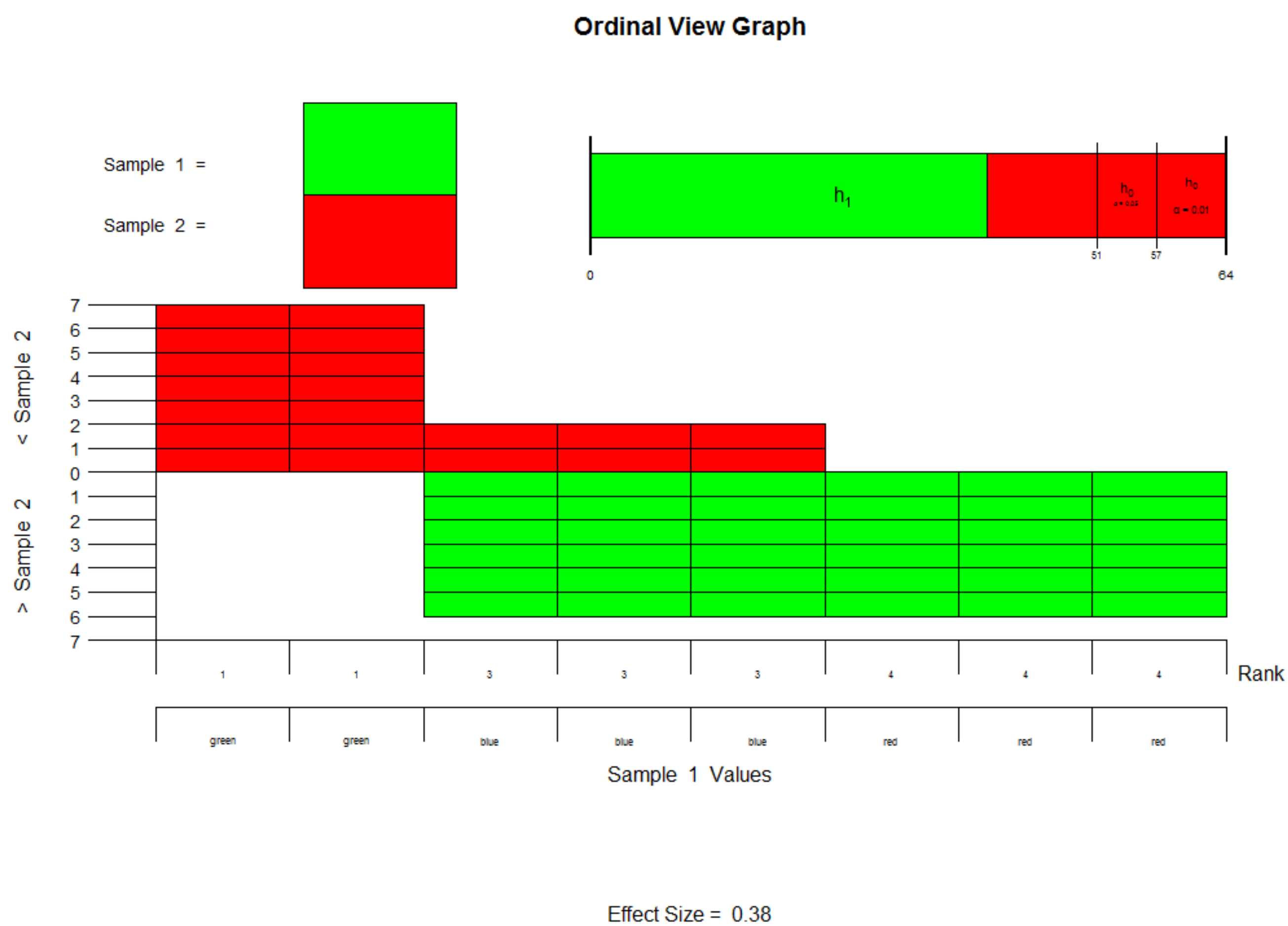
Test 16:
green < yellow < blue < red

Data:
Sample 1: (green, blue, red, red, blue, red, green, blue)
Sample 2: (yellow, yellow, red, yellow, red, yellow, yellow, green)

Rank Vector:
(Green, Yellow, Blue, Red)

Parameters:
ordinal_view_graph(Data, 0.05, 2, Rank Vector)

Output:

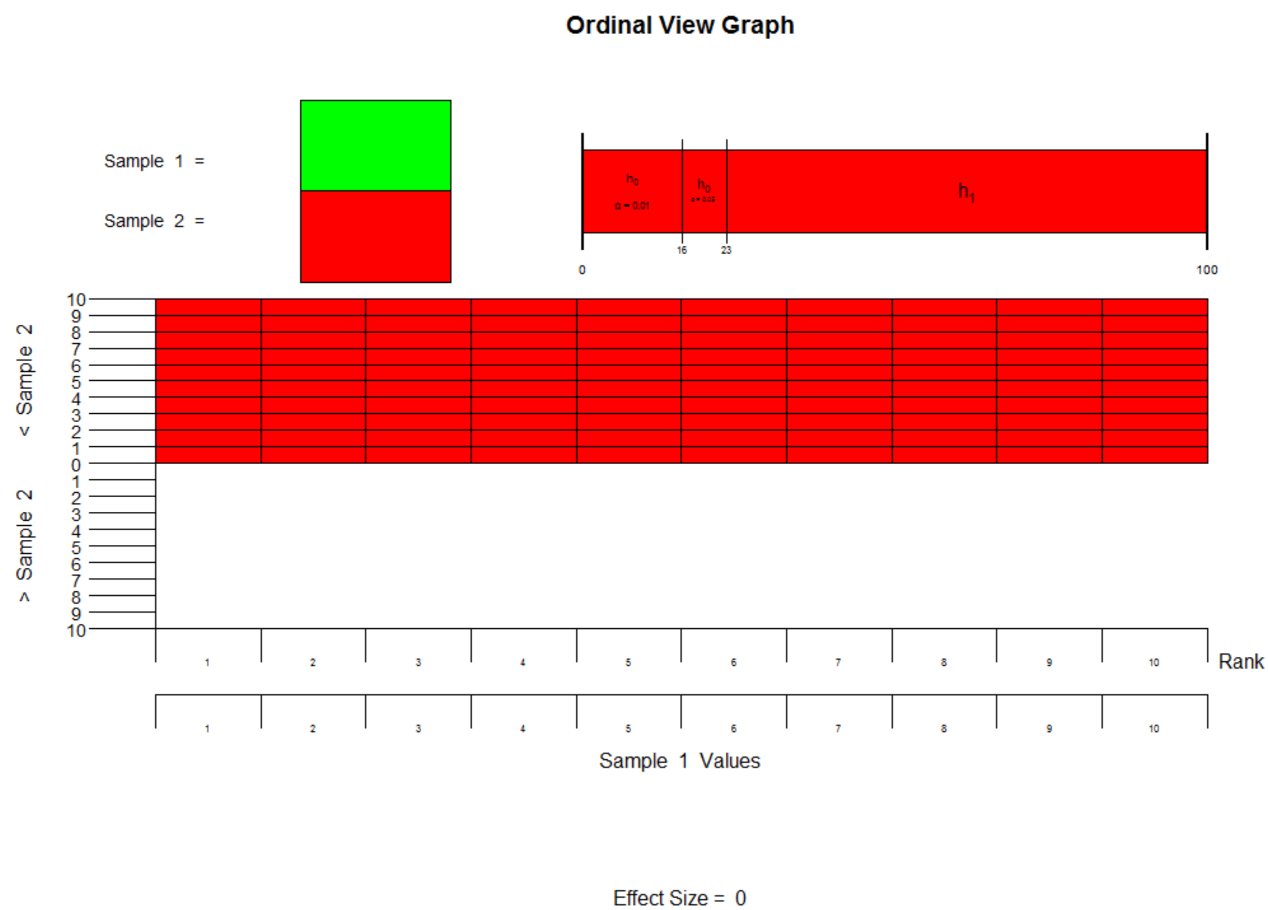


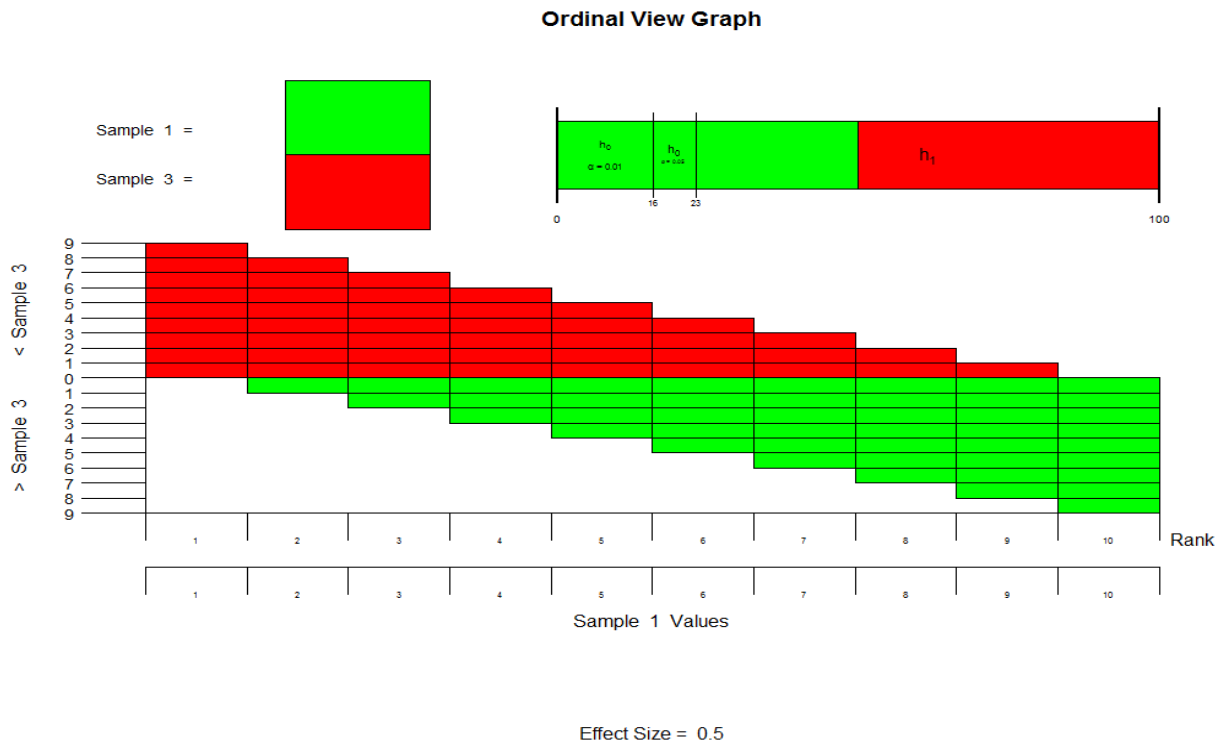
Test 17:
Three Data Sets

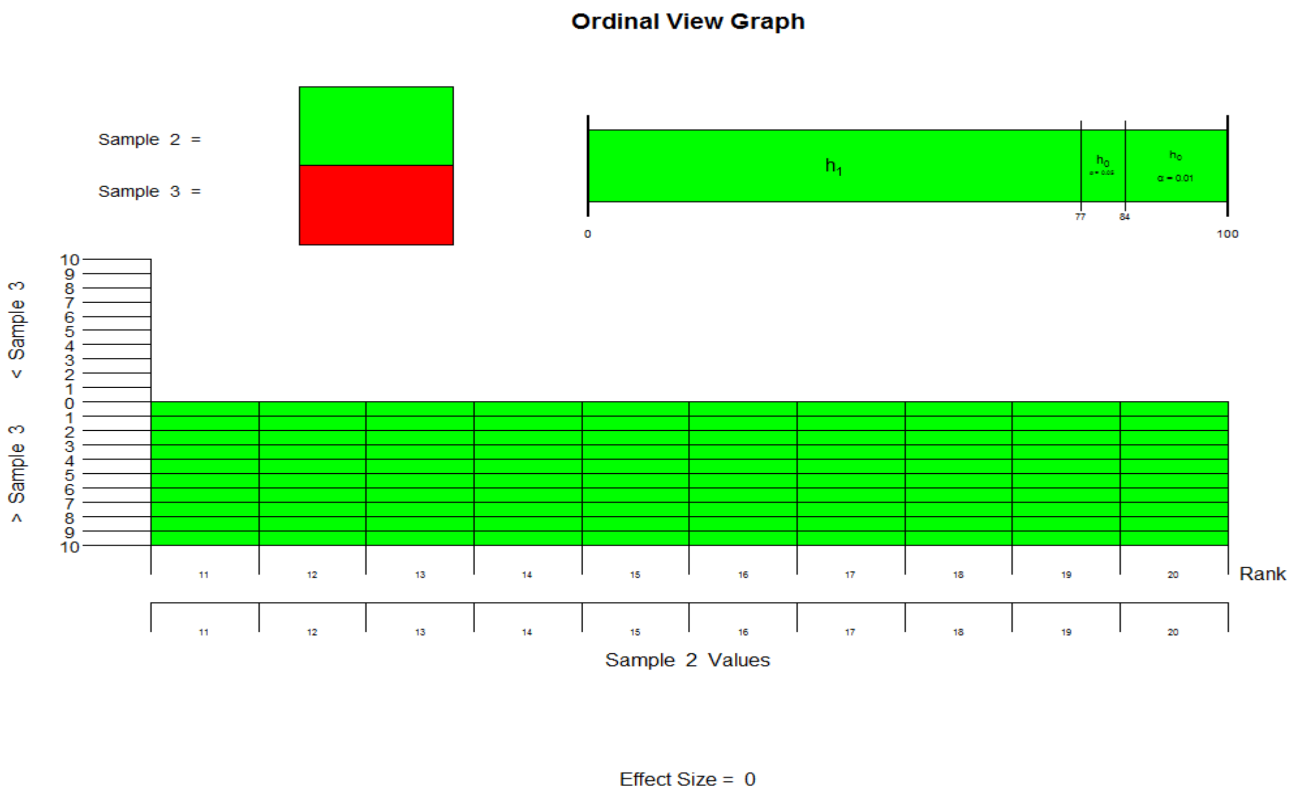
Data:
Sample 1: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 2: (11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
Sample 3: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Parameters:
ordinal_view_graph(Data)

Output:




Ordinal View Graph
Sample 1 =
Sample 3 =
h0
α = 0.01
h0
h1
0
16
23
100
< Sample 3
> Sample 3
Rank
Sample 1 Values
Effect Size = 0.5

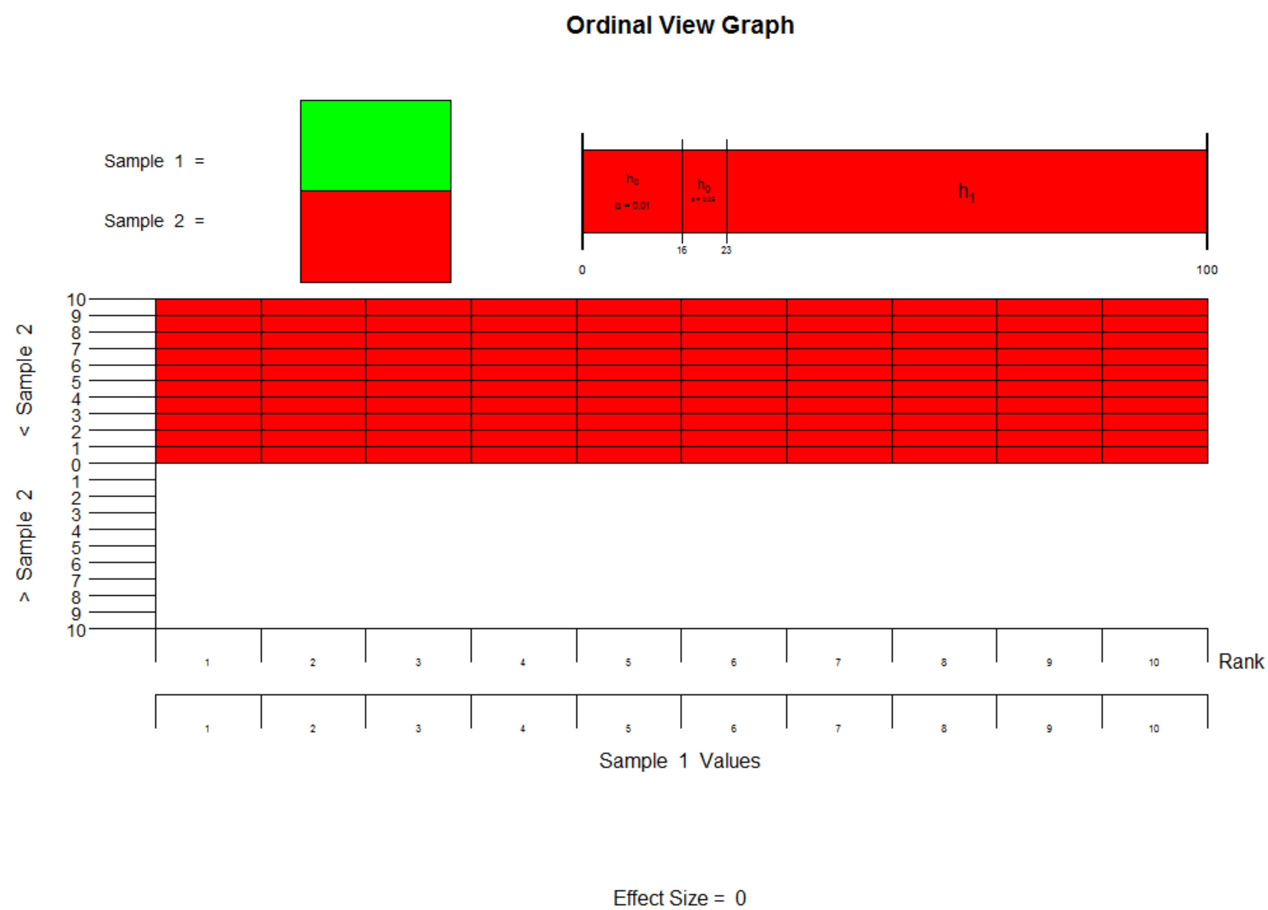

Ordinal View Graph
Sample 2 =
Sample 3 =
h1
h0
h0
α = 0.01
0
77
84
100
< Sample 3
> Sample 3
Rank
Sample 2 Values
Effect Size = 0

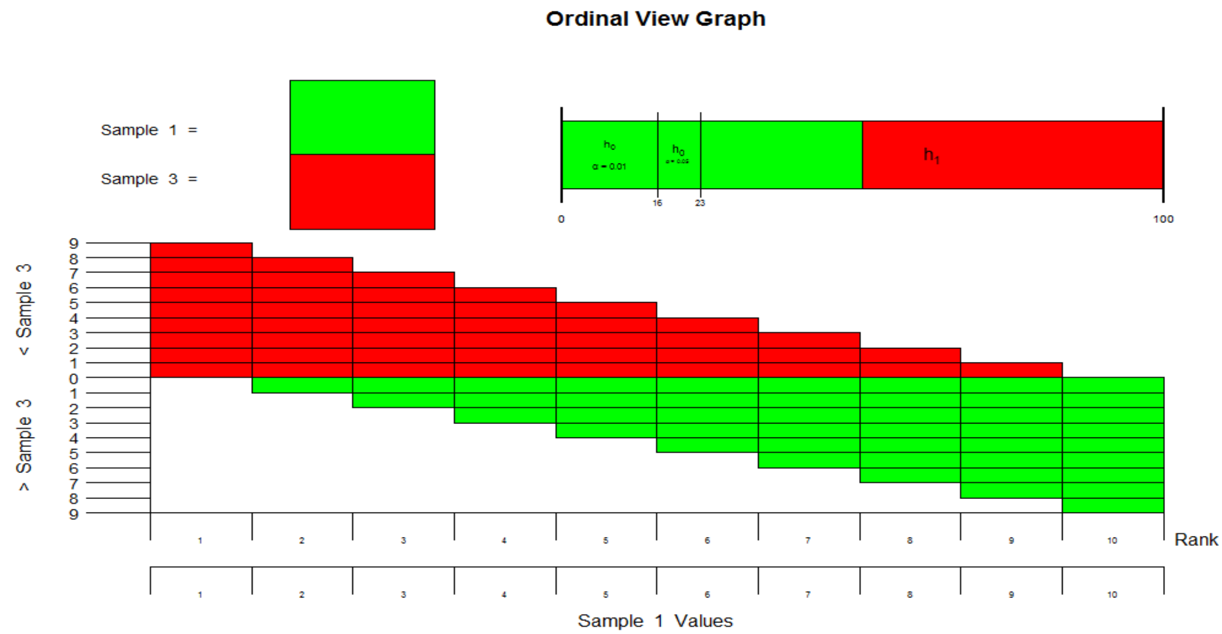
Test 18:
Four Data Sets

Data:
Sample 1: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 2: (11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
Sample 3: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 4: (21, 22, 23, 24, 25, 26, 27, 28, 29, 30)

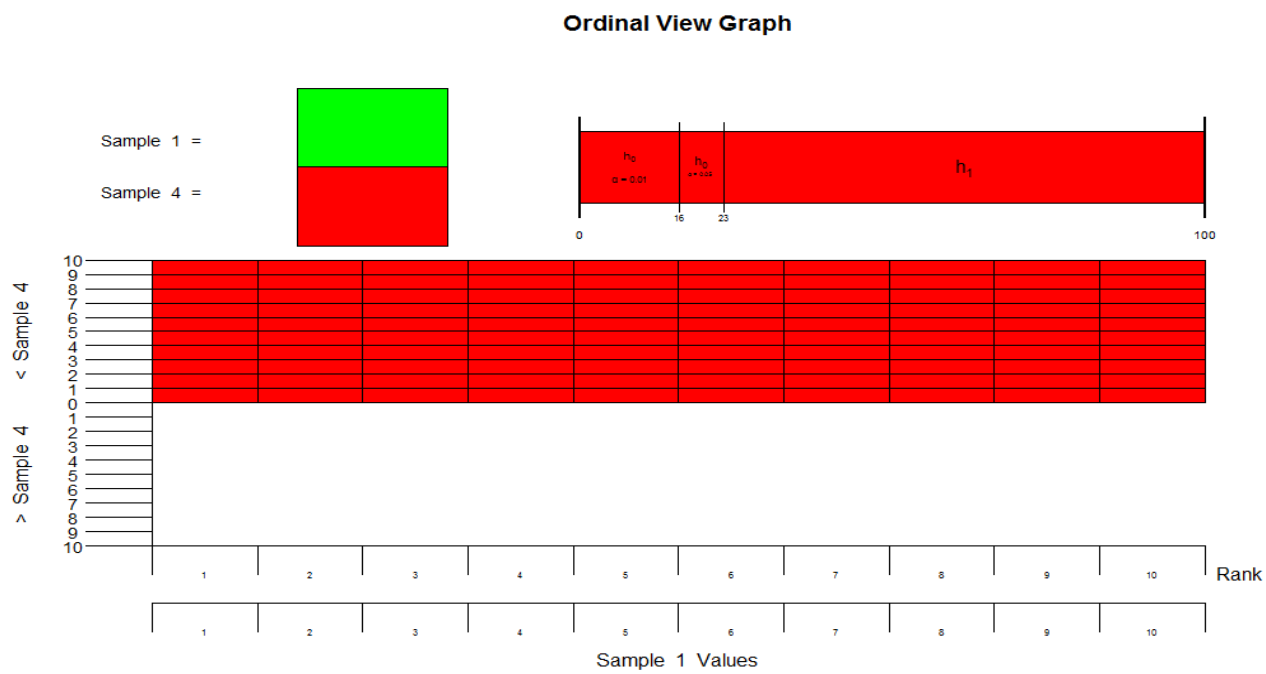
Parameters:
ordinal_view_graph(Data)

Output:

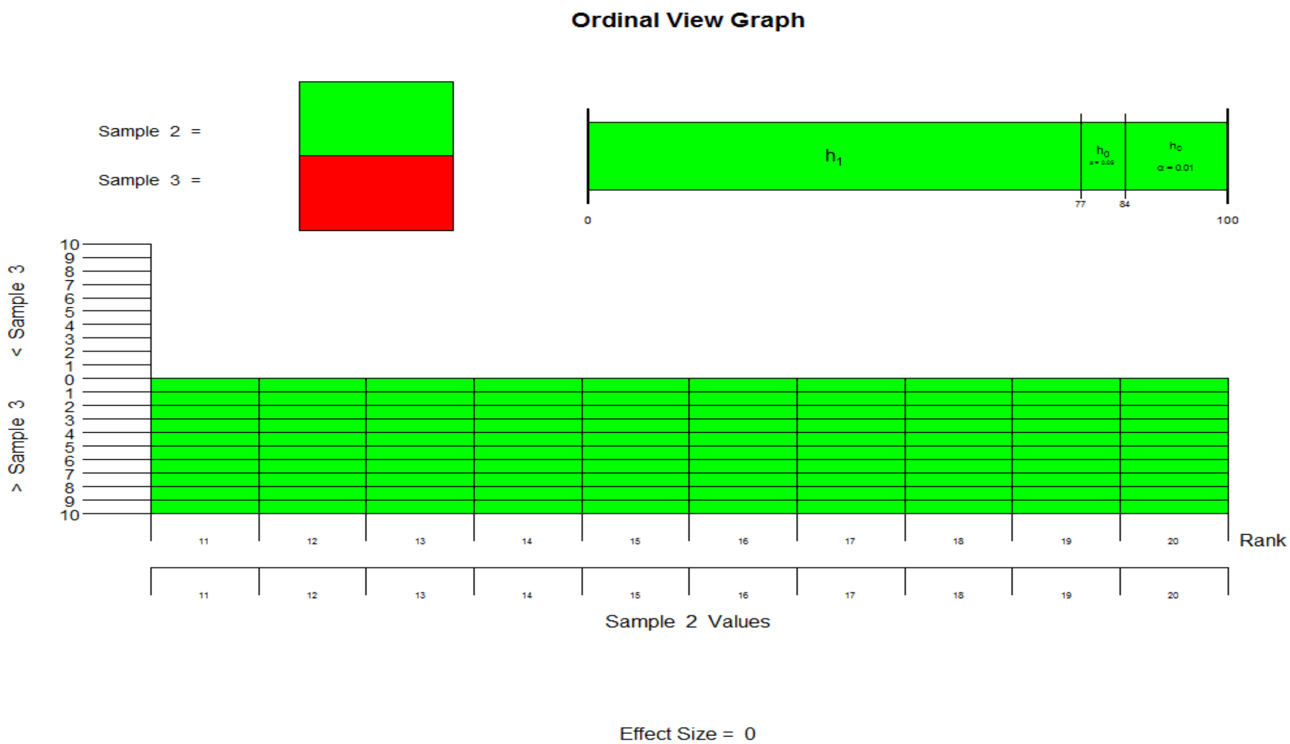


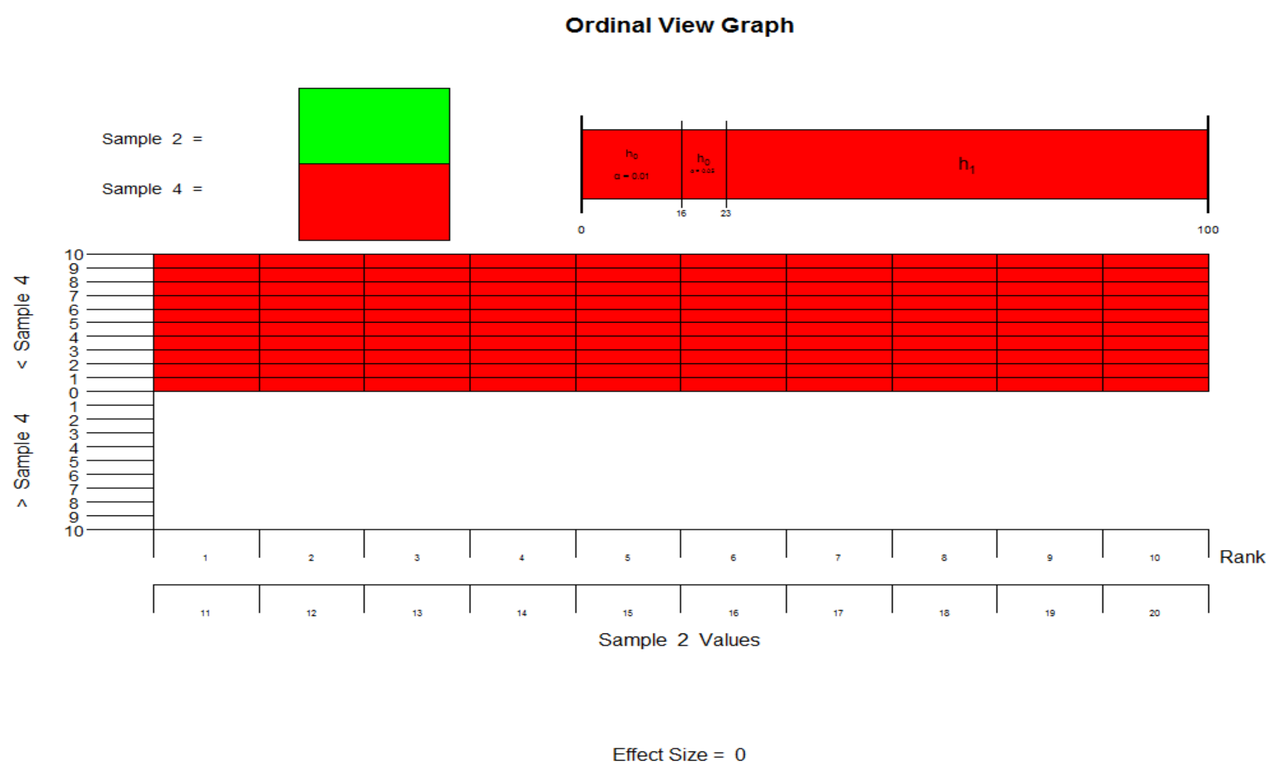


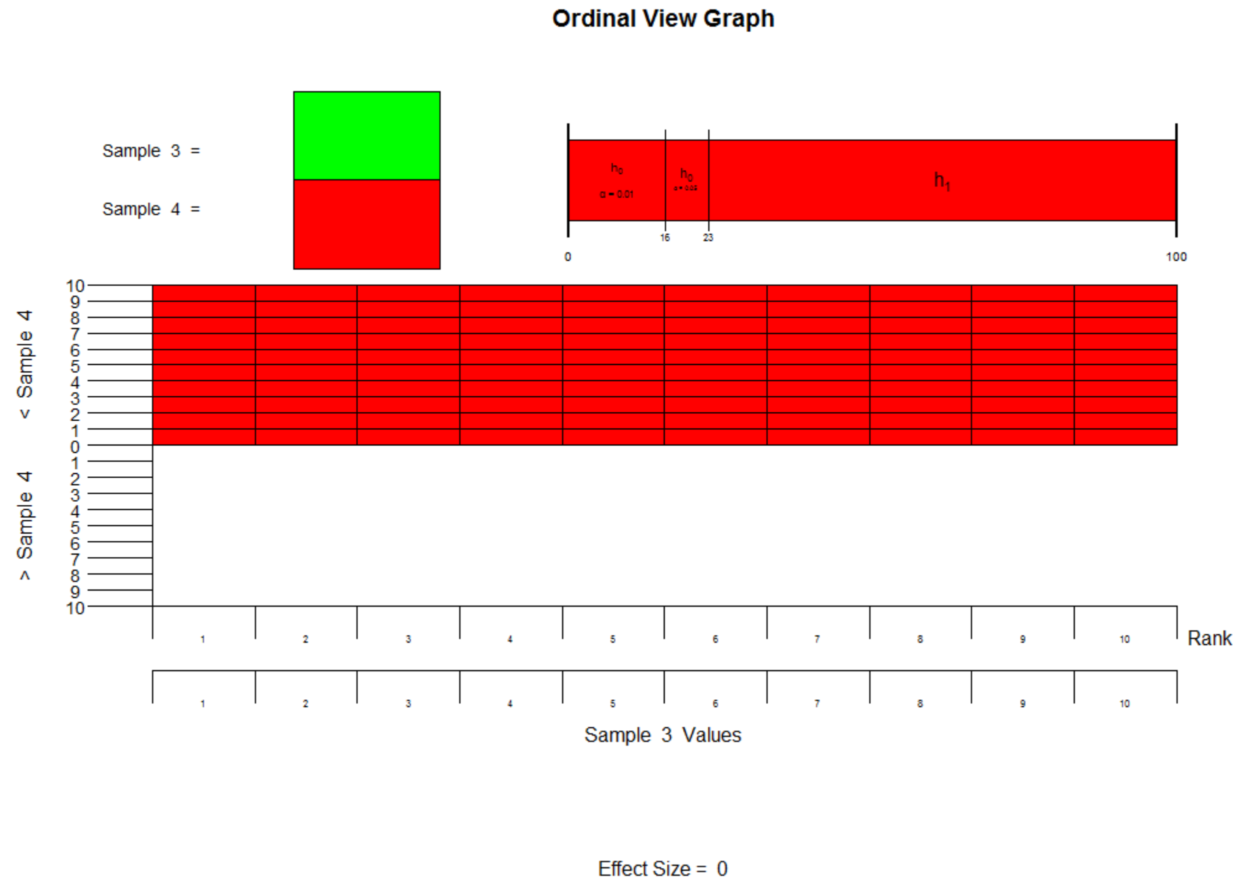
Effect Size = 0.5



Effect Size = 0


Ordinal View Graph
Sample 2 =
Sample 3 =
h1
h0
h0
α = 0.01
0
77
84
100
< Sample 3
> Sample 3
Rank
Sample 2 Values
Effect Size = 0


Ordinal View Graph
Sample 2 =
Sample 4 =
h0
α = 0.01
h0
h1
0
16
23
100
< Sample 4
> Sample 4
Rank
Sample 2 Values
Effect Size = 0

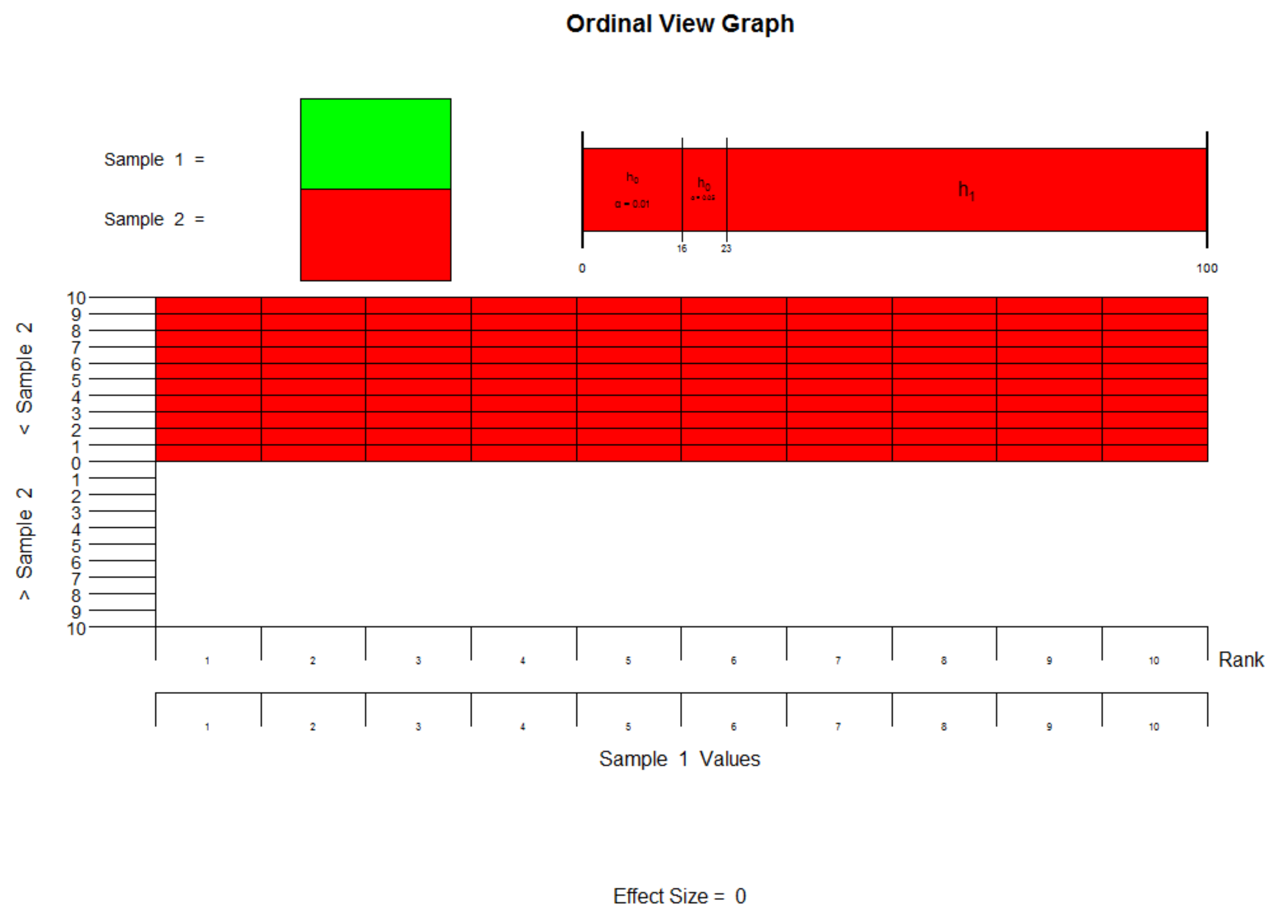

Ordinal View Graph
Sample 3 =
Sample 4 =
h0
α = 0.01
h0
h1
0
16
23
100
< Sample 4
> Sample 4
Rank
Sample 3 Values
Effect Size = 0

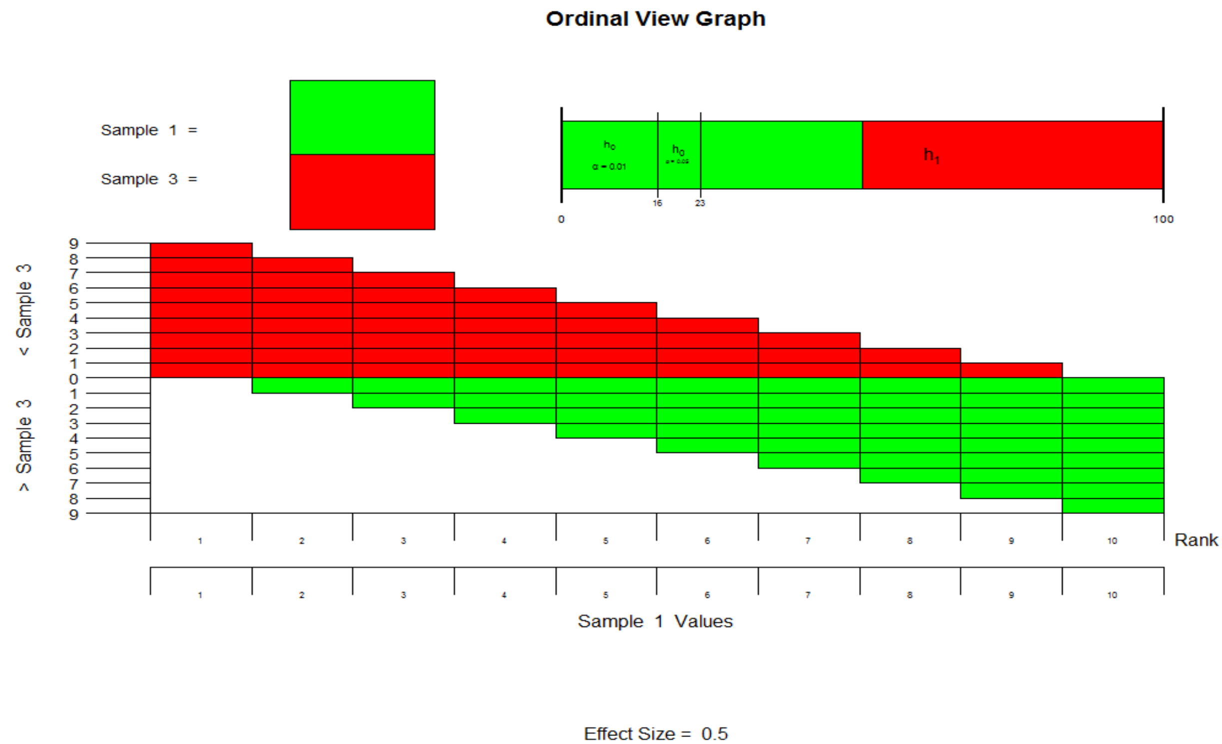
Test 19:
Five Data Sets

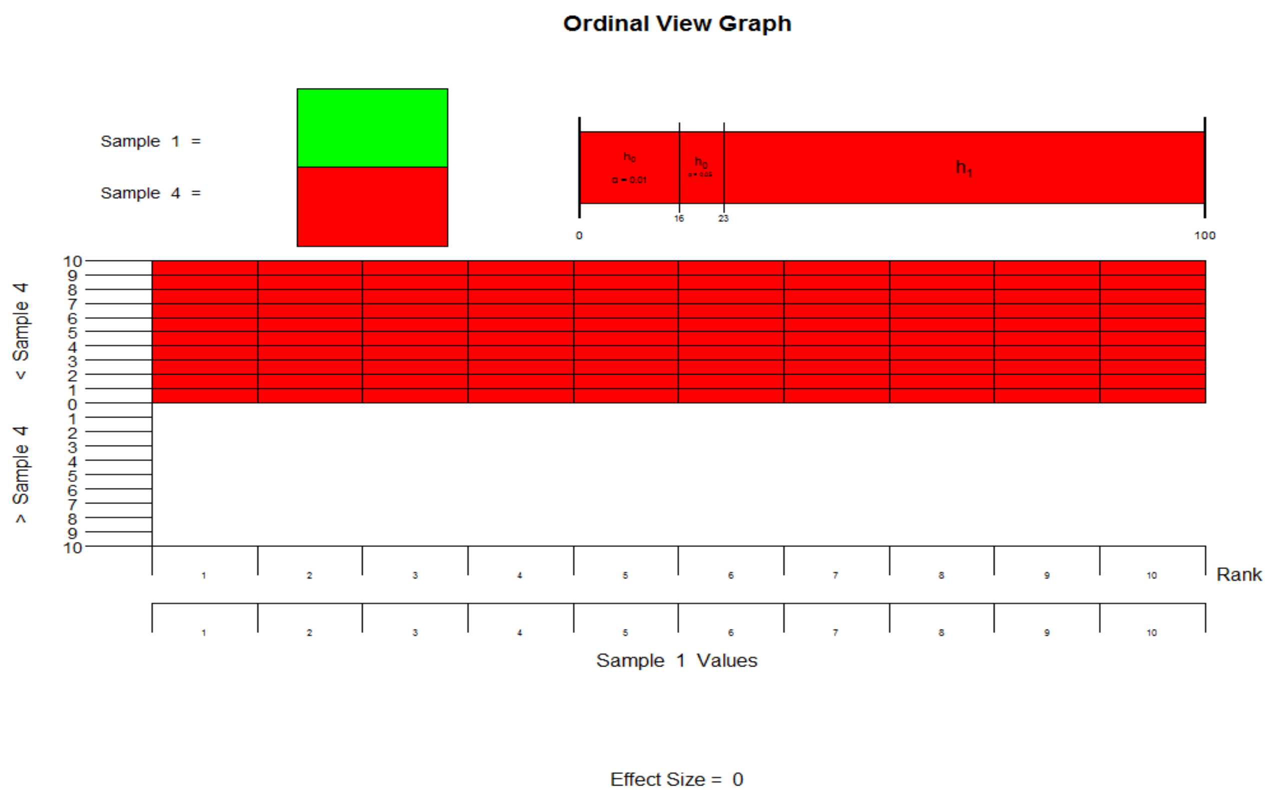
Data:
Sample 1: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 2: (11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
Sample 3: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Sample 4: (21, 22, 23, 24, 25, 26, 27, 28, 29, 30)
Sample 5: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

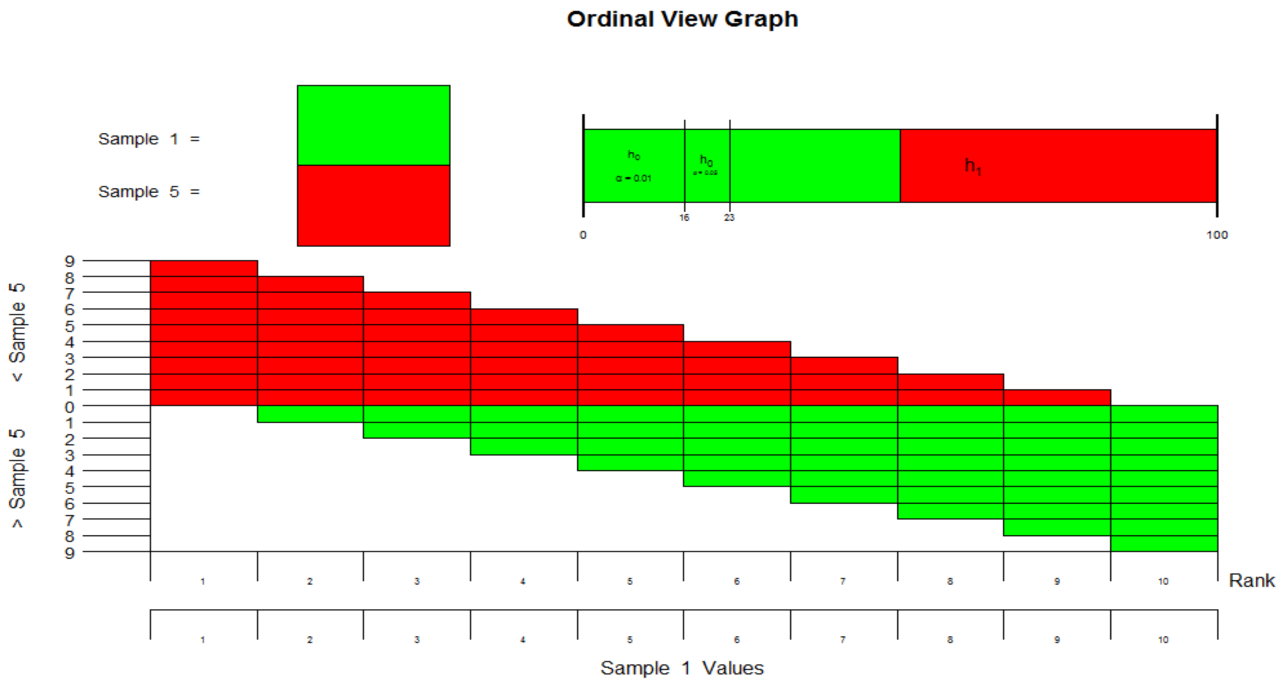
Parameters:
ordinal_view_graph(Data)

Output:

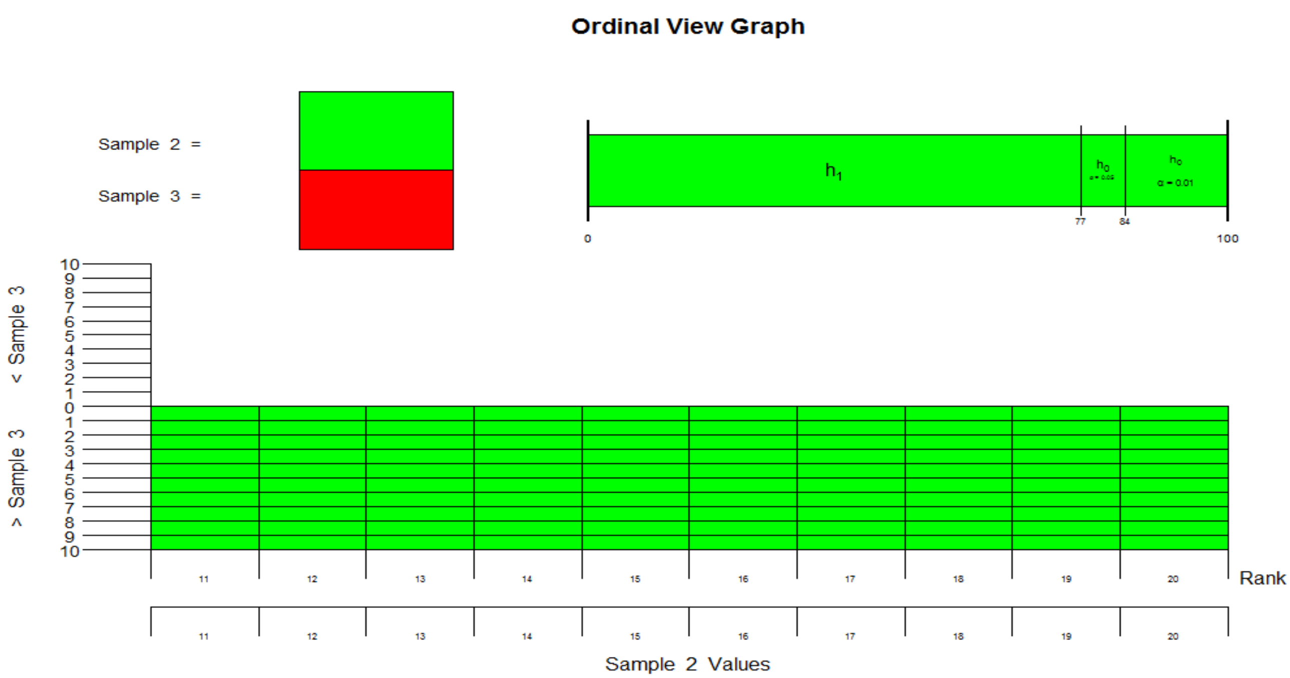



Ordinal View Graph
Sample 1 =
Sample 3 =
h0
h0
h1
0
16
23
100
< Sample 3
> Sample 3
Rank
Sample 1 Values
Effect Size = 0.5

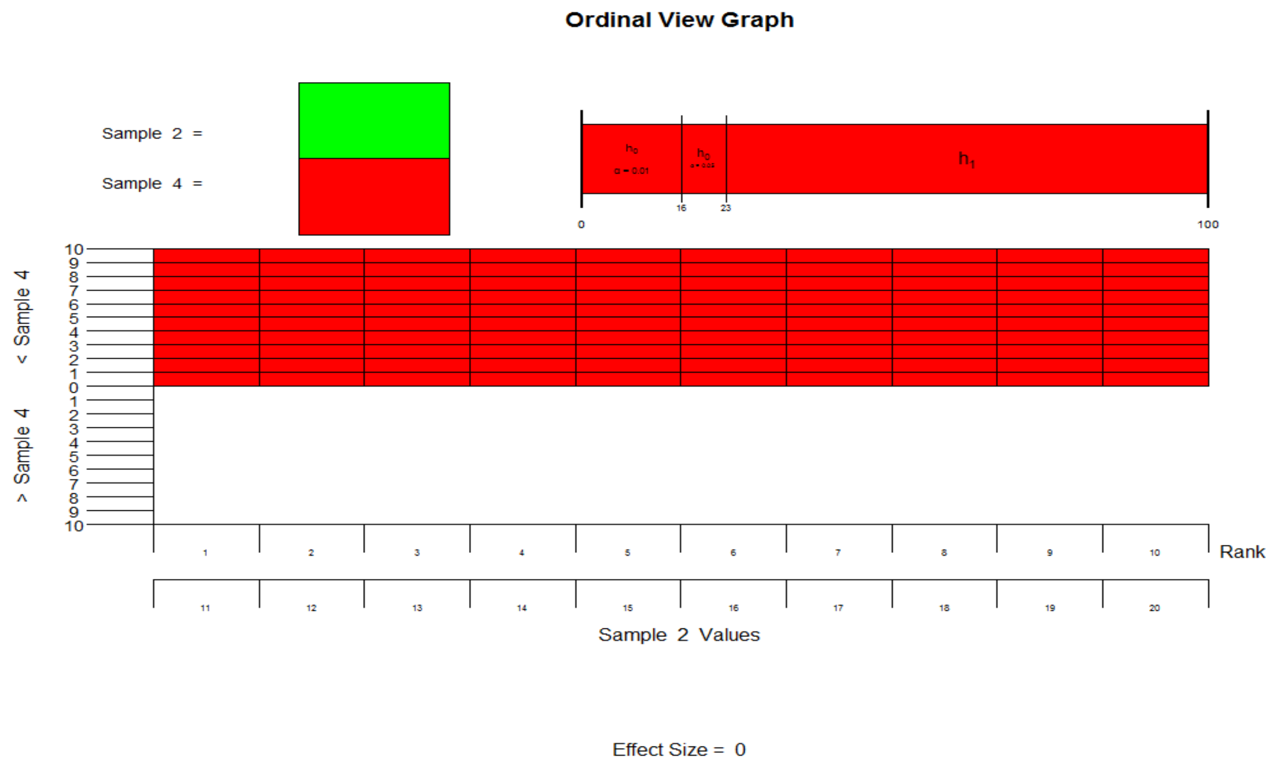

Ordinal View Graph
Sample 1 =
Sample 4 =
h0
h0
h1
0
16
23
100
< Sample 4
> Sample 4
Rank
Sample 1 Values
Effect Size = 0

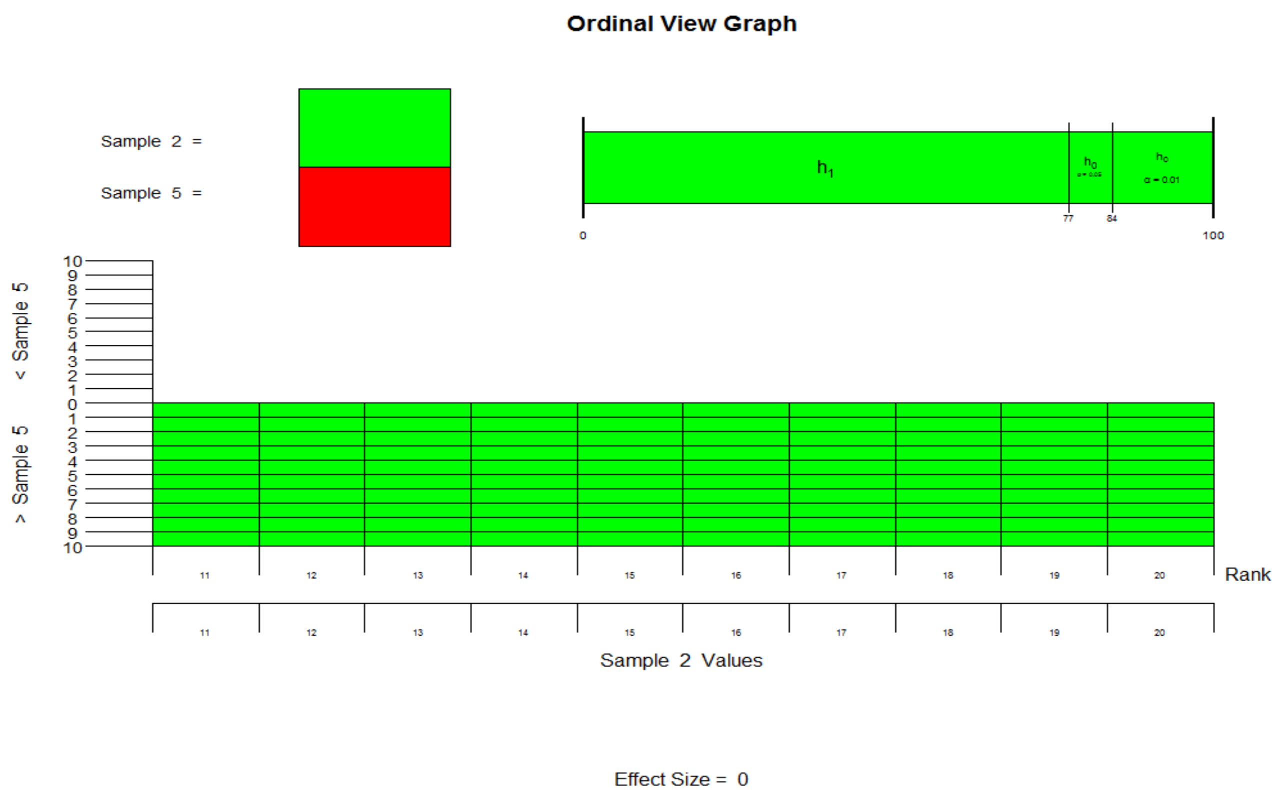


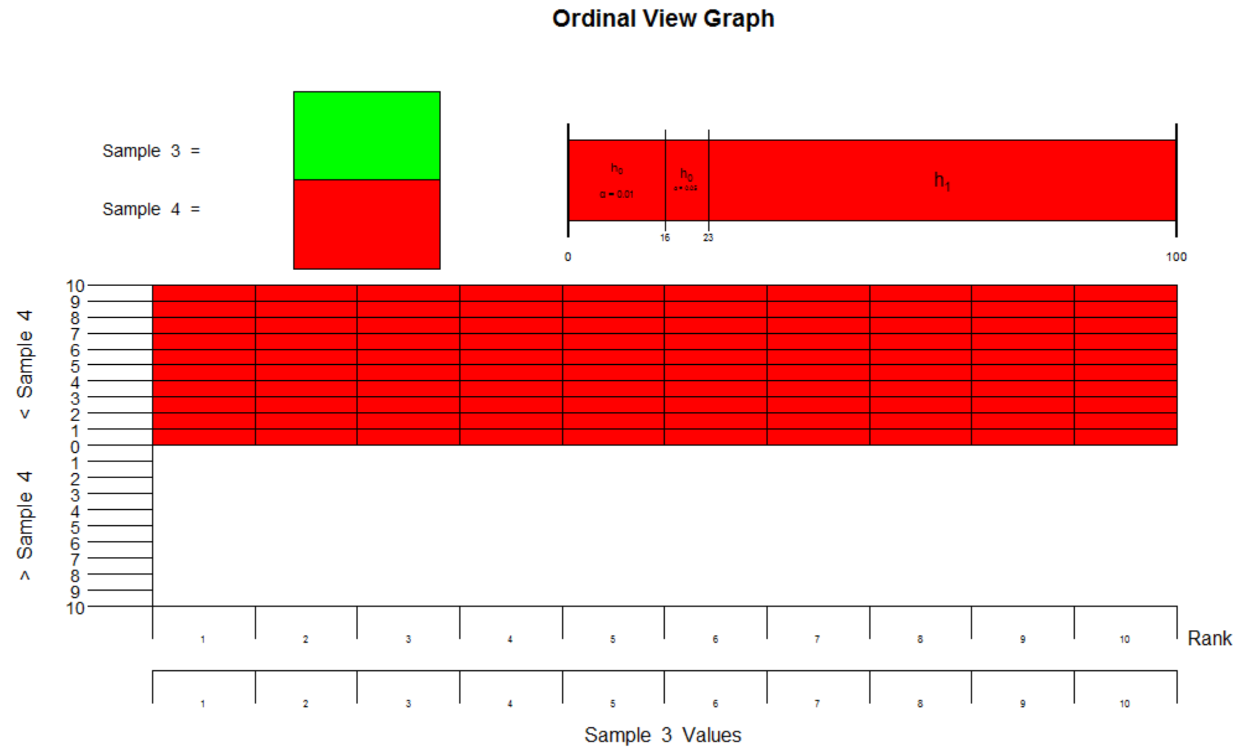
Effect Size = 0.5



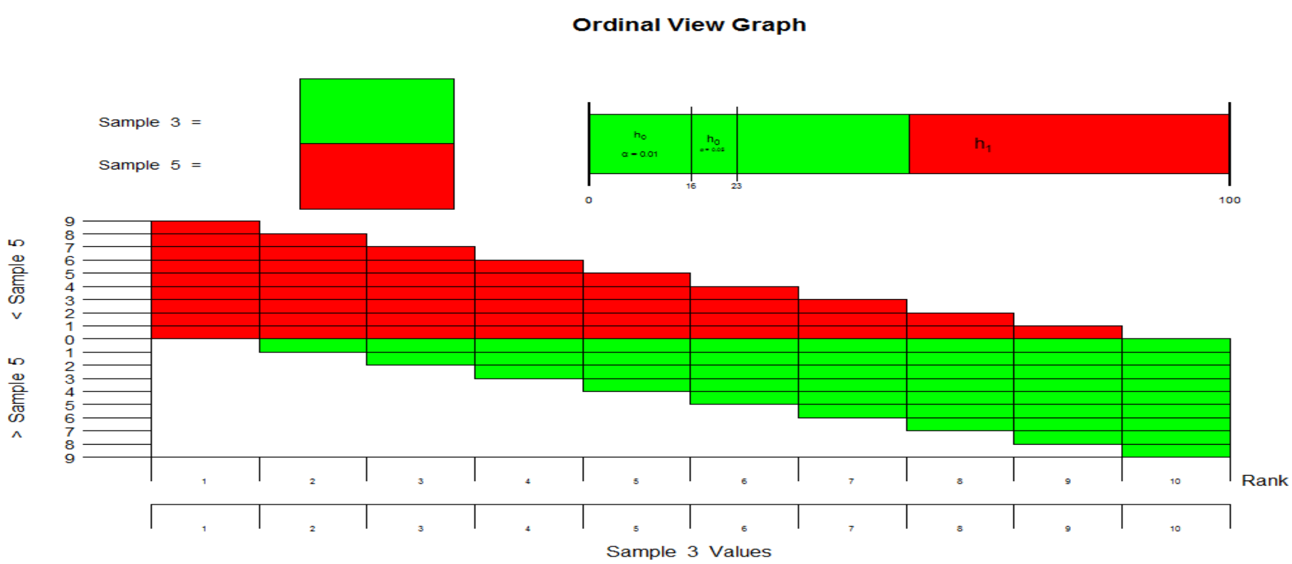
Effect Size = 0


Ordinal View Graph
Sample 2 =
Sample 4 =
h1
0
16
23
100
< Sample 4
> Sample 4
Rank
Sample 2 Values
Effect Size = 0

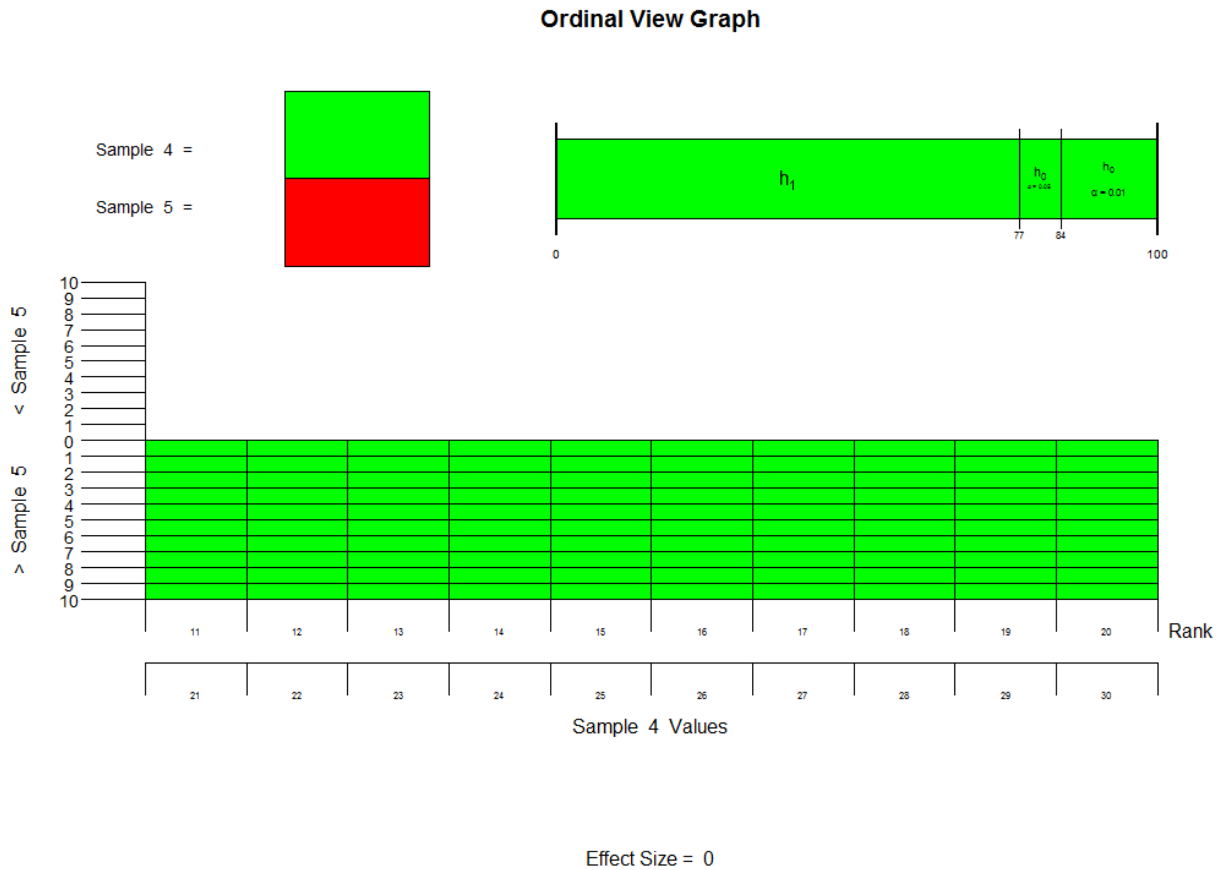

Ordinal View Graph
Sample 2 =
Sample 5 =
h1
0
77
84
100
< Sample 5
> Sample 5
Rank
Sample 2 Values
Effect Size = 0



Effect Size = 0



Effect Size = 0.5


Ordinal View Graph
Sample 4 =
Sample 5 =
h1
h0
h0
α = 0.01
0
77
84
100
< Sample 5
> Sample 5
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Rank
21
22
23
24
25
26
27
28
29
30
Sample 4 Values
Effect Size = 0