Chapter 271

Cluster Randomization – Create Cluster Proportions Dataset

Introduction

A *cluster randomization trial* occurs when whole groups or *clusters* of individuals are treated together. Examples of such clusters are clinics, hospitals, cities, schools, or neighborhoods. In the two-group case, each cluster is randomized to receive a particular treatment. That is, all individuals in a cluster receive the same treatment. One way to analyze the data from such a design is to form the means of each cluster and then analyze those means using a two-sample t-test, an unequal-variance two-sample t-test, or a regression analysis. When the endpoint is a binary variable coded as a zero or one, the mean is the proportion.

This procedure creates a new dataset containing the cluster proportions from an original dataset containing information on individuals. This summarized dataset can then be analyzed further using t-tests or regression analysis.

Cluster-randomized trials are covered in several texts, including Hayes and Moulton (2017), Campbell and Walters (2014), Eldridge and Kerry (2012), Donner and Klar (2000), and Murray (1998).

Data Structure

A dataset analyzed by this procedure requires three variables: a categorical cluster variable, a categorical group variable, and a binary data variable that is coded with either a 0 (no) or a one (yes).

Here is an example of a dataset that can be successfully manipulated with this procedure. The Cluster column gives the cluster identification number. The Group column gives an identification number of the group to which each cluster belongs. All group values in a given cluster should be equal. A Data column (Outcome) gives the endpoint value for each individual. This example dataset is called **ClusRandProps**.

ClusRandProps dataset (subset)

Cluster	Group	Outcome
1	1	0
1	1	0
1	1	0
1	1	0
1	1	1
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	1
1	1	0
1	1	1
.	.	.
.	.	.
.	.	.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

The options on this screen control the variables that are used in the analysis.

Cluster Variables

Cluster Variable(s)

The cluster variable defines how the data will be summarized. Each category in this variable will result in a separate row of the new, summarized database.

Usually, you will specify one Cluster variable and one Treatment Group variable, but you can specify up to a total of seven cluster variables and one treatment group variable.

The categories in these variables may be text (e.g. “Low, Med, High”) or numeric (e.g. “1, 2, 3”).

Statistics are computed for each combination of values from all variables entered here.

Category Order

The data values in each variable will be sorted alpha-numerically before being listed. If you want the values to be displayed in a different order, specify a custom value order for the data columns entered here using the Column Info Table on the Data Window.

Treatment Group Variable

The treatment group variable defines the treatment group associated with each cluster. The program assumes that all rows in a particular cluster have the same treatment group. The function of this variable is to define the treatment groups on the summarized dataset.

Only one Treatment Group Variable may be specified. The categories in this variable may be text (e.g. “Low, Med, High”) or numeric (e.g. “1, 2, 3”).

Numeric Variables to be Summarized for each Cluster

Binary Endpoint Variable(s)

The binary endpoint is sometimes called the response or dependent variable. In the case its values must be zero or one so that it's mean will equal the proportion.

Specify one or more variables whose count, mean, and sum are to be calculated for each cluster. These statistics will be computed for each combination of the values in the cluster variables that you have selected.

Text values will be skipped in the calculations.

Covariate Variable(s)

Covariates are additional variables that may be useful, so their values are carried forward to the summary dataset.

Specify one or more variables whose counts, means, and sums are to be calculated for each cluster. These statistics will be computed for each combination of the values in the cluster variables that you have selected.

The data in these variables must be numeric. Text values will be skipped in the calculations.

Frequency (Count) Variable

Frequency Variable

Specify an optional frequency (count) variable. This data column contains integers that represent the number of observations (frequency) associated with each row of the dataset.

When Omitted

If this option is left blank, each dataset row has a frequency of one. This variable lets you modify that frequency. This may be useful when your data are tabulated and you want to enter counts.

Cluster Statistical Summary Storage

Store the Cluster Summaries in a New NCSS Data File…

When this box is checked, the current dataset will close and the resulting cluster summaries will be written to a new data table.

You will be prompted to save any changes to the current dataset before continuing. Any unsaved data will be lost!

Output File

The new data table with the cluster summaries will automatically be saved in the output file entered below. If no output file is specified, no file will be saved. You can manually save the output data file to any file you wish.

Output File Name

Enter an output file name in which to store the data. Double-click the box or click on the file selection button to browse for a file or folder.

The selected output file will be overwritten and the current database will be closed, so make sure to save all data before continuing.

No Output File Name

If no output file is specified, the summary output data will not be automatically saved. You can manually save the output data file to any file you wish. If you have chosen to reopen the current dataset without entering an output file name, then the summary output data will be lost.

Cluster Statistics Storage

Select how the cluster statistics will be stored in the output file. This option does not affect calculations.

The options are

- **Store as Rows**
  Data variable names and associated summary statistics are stored row-by-row in the output data table.

Variable	Group	Count	Mean
Var 1	A	12	15.7
Var 1	B	37	12.6
Var 2	A	12	27.5
Var 2	B	37	33.6

- **Store as Columns (Recommended)**
  Data variable names and associated summary statistics are stored column-by-column in the output data table.

Group	Var 1 Count	Var 1 Mean	Var 2 Count	Var 2 Mean
A	12	15.7	12	27.5
B	37	12.6	37	33.6

In both cases, the summary statistics for the cluster variables are listed row-by-row.

Automatically Reopen the Current Dataset…

The storage operation requires the current dataset to close while the summary data is written to the new data table. When this box is checked, the current dataset will automatically reopen once the process is complete.

No Output File Name

If no output file is specified, the summary output data will not be automatically saved. If you have chosen to reopen the current dataset without entering an output file name, then the summary output data will be lost.

Unsaved Data Warning

If you have not saved the current dataset to a file, the data will not reopen and will be lost!

Missing Values Tab

This panel lets you specify up to five missing values (besides the default of blank). For example, ‘0’, ‘9’, or ‘NA’ may be missing values in your database.

Missing Value Inclusion

Specifies whether to include observations with missing values in the tables.

Delete All indicates that you want the missing values totally ignored.

Include in All indicates that you want the missing values treated just like any other category.

Missing Values

Specify up to five individual missing values here, one per box.

Reports Options Tab

The options on this screen control the appearance of the reports.

Report Options

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

Value Labels are used to make reports more legible by assigning meaningful labels to numbers and codes.

- **Data Values**
  All data are displayed in their original format, regardless of whether a value label has been set or not.
- **Value Labels**
  All values of variables that have a value label variable designated are converted to their corresponding value label when they are output. This does not modify their value during computation.
- **Both**
  Both data value and value label are displayed.

Summary Table Formatting

Column Justification

Specify whether data columns in the tables will be left or right justified.

Column Widths

Specify how the widths of columns in the contingency tables will be determined.

The options are

- **Autosize to Minimum Widths**
  Each data column is individually resized to the smallest width required to display the data in the column. This usually results in columns with different widths. This option produces the most compact table possible, displaying the most data per page.
- **Autosize to Equal Minimum Width**
  The smallest width of each data column is calculated and then all columns are resized to the width of the widest column. This results in the most compact table possible where all data columns have the same with. This is the default setting.
- **Custom (User-Specified)**
  Specify the widths (in inches) of the columns directly instead of having the software calculate them for you.

Custom Widths

Enter one or more values for the widths (in inches) of columns in the contingency tables.

- **Single Value**
  If you enter a single value, that value will be used as the width for all data columns in the table.

- **List of Values**

  Enter a list of values separated by spaces corresponding to the widths of each column. The first value is used for the width of the first data column, the second for the width of the second data column, and so forth. Extra values will be ignored. If you enter fewer values than the number of columns, the last value in your list will be used for the remaining columns.

  Type the word "Autosize" for any column to cause the program to calculate it's width for you. For example, enter "1 Autosize 0.7" to make column 1 be 1 inch wide, column 2 be sized by the program, and column 3 be 0.7 inches wide.

Wrap Column Headings onto Two Lines

Check this option to make column headings wrap onto two lines. Use this option to condense your table when your data are spaced too far apart because of long column headings.

Use Short Statistical Names on Reports and Plots

Normally, the names of the statistical items in the reports and plots are complete names, such as "Standard Deviation." Checking this option causes a shorter name, such as "SD", to be used instead so that more columns can be displayed together in tables and so that plot titles and labels are not so long. A maximum of 13 columns can be displayed on a single row.

Decimal Places

Item Decimal Places

These decimal options allow the user to specify the number of decimal places for items in the output. Your choice here will not affect calculations; it will only affect the format of the output.

- **Auto**

  If one of the "Auto" options is selected, the ending zero digits are not shown. For example, if "Auto (0 to 7)" is chosen,

  0.0500 is displayed as 0.05
  1.314583689 is displayed as 1.314584

  The output formatting system is not designed to accommodate "Auto (0 to 13)", and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

Plots Tab

The options on this panel control the appearance of the plots that are displayed. Click the plot format button to change the plot settings.

Show Statistic Plots

Check to display a separate plot for each statistic in each table. This may result in several plots being created for each table. Plots are created with a table's row item on the group axis. If there is only one row in a table, then no plot is output.

Example 1 – Creating a Summarized Dataset from the ClusRandProps Data

This section presents an example of how to analyze the data contained in the ClusRandProps dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of this procedure window.

1 Open the ClusRandProps dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **ClusRandProps.NCSS**.
- Click **Open**.

2 Open the Cluster Randomization – Create Cluster Proportions Dataset window.

- Using the Analysis menu, the Tools Menu, or the Procedure Navigator, find and select the **Cluster Randomization – Create Cluster Proportions Dataset** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **Cluster Variable(s)** to **Cluster**.
- Set the **Treatment Group Variable** to **Group**.
- Set the **Binary Endpoint Variable(s)** to **Outcome**.
- Check the **Store the Cluster Summaries to a New NCSS Data File box**.
- Set **Output File Name** to **%mydocs_NCSS%\Data\Cluster Proportions.NCSS**.
- Set **Cluster Statistics Storage** to **Store as Column**.
- Check **Automatically Reopen the Current Dataset after the Save Operation Completes**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Summary List Storage Information

Output Data File Name:	{NCSS Documents Folder}\Cluster Proportions.NCSS
Original Raw Data File:	{Example Data Folder}\ClusRandProps.NCSS
Data Variable(s):	(1) Outcome
Group Variable(s):	(2) Cluster, Group
Summary Statistic(s):	(4) Count, Mean, Sum

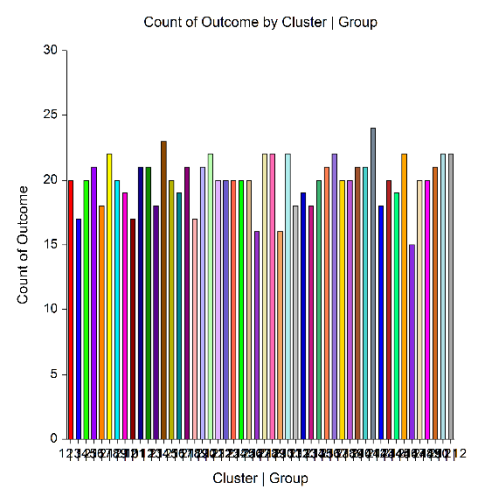
This report shows where the new, summarized file is stored.

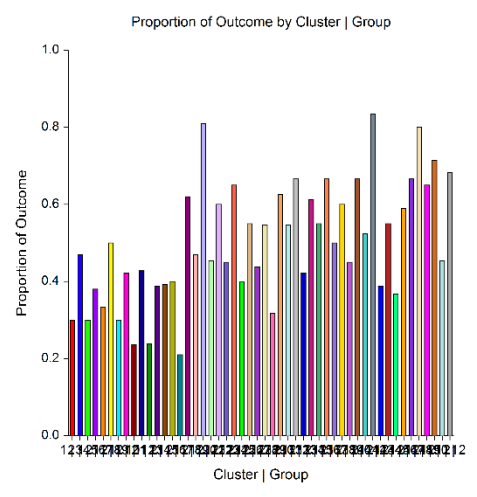
Summary List of Outcome

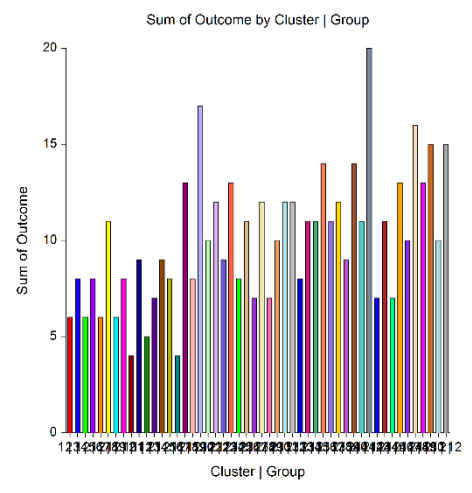
Cluster | Group	Statistics for Outcome		
	Count	Proportion	Sum
1 | 1	20	0.3	6
2 | 2	17	0.4705882	8
3 | 1	20	0.3	6
4 | 2	21	0.3809524	8
5 | 1	18	0.3333333	6
6 | 2	22	0.5	11
7 | 1	20	0.3	6
8 | 2	19	0.4210526	8
.	.	.	.
.	.	.	.
.	.	.	.

This report displays count, proportion (mean), and sum of the Outcome variable for each cluster.

Plots of each Statistic for Outcome







This report displays the statistics cluster by cluster.

New Cluster Proportions Dataset

You can open the new Cluster Proportions dataset by using the File menu on the Data Window. The following dataset will appear.

Cluster	Group	Outcome_Count	Outcome_Proportion	Outcome_Sum
1	1	20	0.3	6
2	2	17	0.470588235294118	8
3	1	20	0.3	6
4	2	21	0.380952380952381	8
5	1	18	0.333333333333333	6
6	2	22	0.5	11
7	1	20	0.3	6
8	2	19	0.421052631578947	8
9	1	17	0.235294117647059	4
10	2	21	0.428571428571429	9
11	1	21	0.238095238095238	5
12	2	18	0.388888888888889	7
13	1	23	0.391304347826087	9
14	2	20	0.4	8
15	1	19	0.210526315789474	4
16	2	21	0.619047619047619	13
17	1	17	0.470588235294118	8
18	2	21	0.80952380952381	17
19	1	22	0.454545454545455	10
20	2	20	0.6	12
21	1	20	0.45	9
22	2	20	0.65	13
23	1	20	0.4	8
24	2	20	0.55	11
25	1	16	0.4375	7
26	2	22	0.545454545454545	12
27	1	22	0.318181818181818	7
28	2	16	0.625	10
29	1	22	0.545454545454545	12
30	2	18	0.666666666666667	12
31	1	19	0.421052631578947	8
32	2	18	0.611111111111111	11
33	1	20	0.55	11
34	2	21	0.666666666666667	14
35	1	22	0.5	11
36	2	20	0.6	12
37	1	20	0.45	9
38	2	21	0.666666666666667	14
39	1	21	0.523809523809524	11
40	2	24	0.833333333333333	20
41	1	18	0.388888888888889	7
42	2	20	0.55	11
43	1	19	0.368421052631579	7
44	2	22	0.590909090909091	13
45	1	15	0.666666666666667	10
46	2	20	0.8	16
47	1	20	0.65	13
48	2	21	0.714285714285714	15
49	1	22	0.454545454545455	10
50	2	22	0.681818181818182	15

This dataset can now be analyzed using the Two-Sample T-Test procedure in which the two groups are defined by the Group column and the Response is the Outcome_Proportion column. We suggest that the Randomization test, the Mann-Whitney U test, and/or the Aspin-Welch Unequal-Variance T-Test be used to test for significance.

Example 1b – Analyzing the Summarized Dataset

This section continues the analysis begun with Example 1 by analyzing the summarized dataset, **Cluster Proportions**, using the Two-Sample T-Test procedure.

You may follow along here by making the appropriate entries or load the completed template **Example 1b** by clicking on Open Example Template from the File menu of the Two-Sample T-Test procedure window.

1 Open the Cluster Proportions dataset that you just created in Example 1.

- From the File menu of the NCSS Data window, select **Cluster Proportions** in the list of recent datasets.

or

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Cluster Proportions.NCSS**.
- Click **Open**.

2 Open the Two-Sample T-Test window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Two-Sample T-Test** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **Data Input Type** to **Response Variable(s) and Group Variable(s)**.
- Set the **Response Variable(s)** to **Outcome_Proportion**.
- Set the **Group Variables** to **Group**.

4 Select the reports.

- Select the **Reports tab**.
- Check the **Descriptive Statistics and Confidence Intervals** report.
- Check the **Confidence Interval of μ1 – μ2** report.
- Check the **Equal-Variance T-Test** report.
- Check the **Unequal-Variance T-Test** report.
- Check the **Randomization Test** report.
- Check the **Mann-Whitney U Test** report.
- Check the **Exact Test** report.
- Check the **Normal Approximation Test** report.
- Check the **Normal Approximation Test with Continuity Correction** report.
- Check the **Tests of Assumptions** report.

5 Select the plots.

- Select the **Plots tab**.
- Check the **Probability Plot** and **Box Plot** report.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Two-Sample Test Report

Descriptive Statistics

Variable	Count	Mean	Standard Deviation of Data	Standard Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
Group=1	25	0.4143283	0.1202122	0.02404244	2.0639	0.3647071	0.4639495
Group=2	25	0.5908215	0.1293741	0.02587481	2.0639	0.5374185	0.6442245

Descriptive Statistics for the Median

Variable	Count	Median	95% LCL of Median	95% UCL of Median
Group=1	25	0.4210526	0.3333333	0.4545455
Group=2	25	0.6	0.5454546	0.6666667

Two-Sided Confidence Interval for μ1 - μ2

Variance Assumption	DF	Mean Difference	Standard Deviation	Standard Error	T*	95% C. I. of μ1 - μ2 Lower Limit	95% C. I. of μ1 - μ2 Upper Limit
Equal	48	-0.1764932	0.1248772	0.0353206	2.0106	-0.24751	-0.1054763
Unequal	47.74	-0.1764932	0.176603	0.0353206	2.0109	-0.2475199	-0.1054665

Equal-Variance T-Test

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
μ1 - μ2 ≠ 0	-0.1764932	0.0353206	-4.9969	48	0.00001	Yes

Aspin-Welch Unequal-Variance T-Test **(This is a key report)**

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
μ1 - μ2 ≠ 0	-0.1764932	0.0353206	-4.9969	47.74	0.00001	Yes

Randomization Tests

Alternative Hypothesis: |μ1 - μ2| ≠ 0. This is a Two-Sided Test.
Number of Monte Carlo samples: 10000

Variance Assumption	Prob Level	Reject H0 at α = 0.050?
Equal Variance	0.00010	Yes
Unequal Variance	0.00010	Yes

Mann-Whitney U or Wilcoxon Rank-Sum Test for Difference in Location **(This is another key report)**

Variable	Mann-Whitney U	Sum of Ranks (W)	Mean of W	Std Dev of W
Group=1	105	430	637.5	51.50416
Group=2	520	845	637.5	51.50416

Number of Sets of Ties = 13, Multiplicity Factor = 168

Test Type	Alternative Hypothesis	Z-Value	Prob Level	Reject H0 at α = 0.050?
Exact*	Location Diff. ≠ 0			
Normal Approximation	Location Diff. ≠ 0	-4.0288	0.00006	Yes
Normal Approx. with C.C.	Location Diff. ≠ 0	-4.0191	0.00006	Yes

* The Exact Test is provided only when there are no ties and the sample size is ≤ 20 in both groups.

Cluster Randomization – Create Cluster Proportions Dataset

Tests of the Normality Assumption for Group=1

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9718	0.68992	No
Skewness	0.6258	0.53143	No
Kurtosis	-0.0891	0.92904	No
Omnibus (Skewness or Kurtosis)	0.3996	0.81890	No

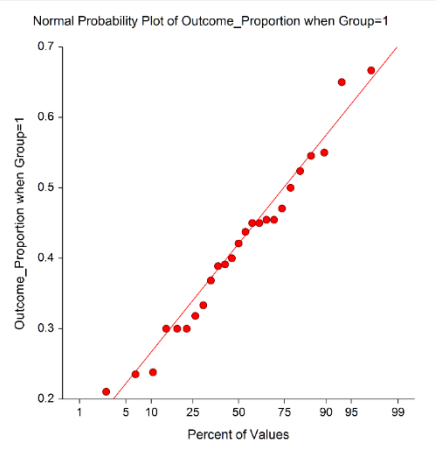
Tests of the Normality Assumption for Group=2

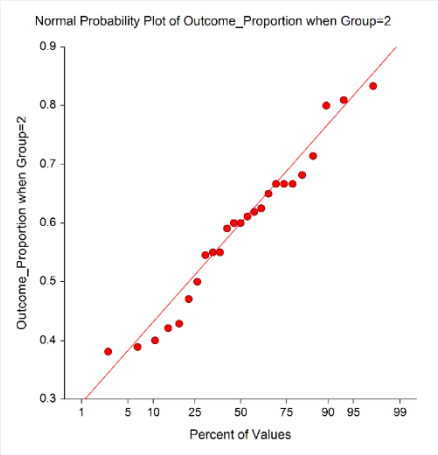
Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9577	0.37051	No
Skewness	0.0860	0.93150	No
Kurtosis	-0.6137	0.53939	No
Omnibus (Skewness or Kurtosis)	0.3841	0.82528	No

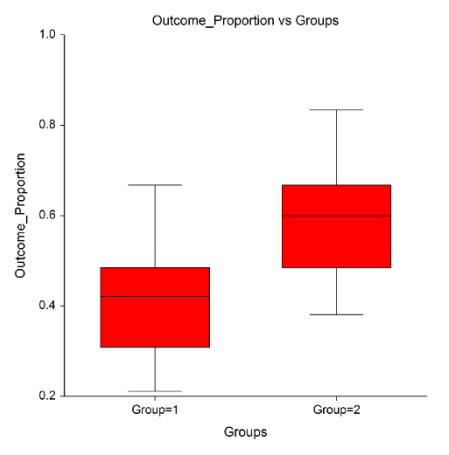
Tests of the Equal Variance Assumption

Equal-Variance Test	Test Statistic	Prob Level	Reject H0 of Equal Variances at α = 0.050?
Variance-Ratio	1.1582	0.72189	No
Modified-Levene	0.0775	0.78193	No

Plots Section







This report displays the results of the various tests. The probability plots let you assess the validity of the normality assumptions. The box plots show the separation between the groups.