

Chapter 119

Confidence Intervals for One Mean in a Cluster-Randomized Design

Introduction

This procedure calculates sample size and half-width for confidence intervals of a mean from a cluster-randomized design in which the outcome variable is continuous. It uses the results from Ahn, Heo, and Zang (2015), Lohr (2019), and Campbell and Walters (2014).

Suppose that the mean of a continuous outcome variable of a sample from a population of subjects (or items) is to be estimated with a confidence interval. Further suppose that the population is separated into small groups, called *clusters*. These clusters may contain different numbers of items.

This procedure allows you to determine the appropriate number of clusters to be sampled so that the width of a confidence interval of the mean may be guaranteed at a certain confidence level.

Technical Details

The following discussion summarizes the results in Ahn *et al.* (2015), pages 24 - 27.

Suppose you are interested in estimating the outcome variable in a population that is made up of a large number of clusters. It may be possible to improve estimation accuracy for a given budget by sampling clusters rather than individuals.

Mean and Variance

In this design, assume that a simple random sample is drawn from each cluster. Let X_{ki} indicate a continuous outcome variable of the i^{th} subject in cluster k . Denote the number of subjects sampled from this cluster as M_k . Let the number of clusters be denoted by K . The average cluster size is M . The estimate of the population mean is calculated from the cluster means as follows.

$$\bar{x} = \frac{\sum_{k=1}^K M_k \bar{x}_k}{\sum_{k=1}^K M_k}$$

Conditional on the empirical distribution of the M_k 's, this mean has a normal distribution with mean and variance as shown below.

$$E(\bar{x}) = \mu$$

$$V(\bar{x}) = \frac{\sigma^2 \sum_{k=1}^K M_k \{1 + (M_k - 1)\rho\}}{(\sum_{k=1}^K M_k)^2}$$

Confidence Intervals for One Mean in a Cluster-Randomized Design

where σ^2 is the variance of X and ρ is the correlation of observations within a cluster (often called the intraclass correlation coefficient). The definition of ρ is

$$\rho = \text{corr}(X_{ki}, X_{k'i'}) \text{ for } i \neq i'$$

This value is assumed to be independent of the number of observations in the cluster. It may be estimated using the ANOVA method which can be written as follows

$$\hat{\rho} = \frac{MSC - MSW}{MSC + (M - 1)MSW}$$

$$MSC = \frac{M}{K - 1} \sum_{k=1}^K (\bar{X}_k - \bar{X})^2$$

$$MSW = \frac{1}{K(M - 1)} \sum_{k=1}^K \sum_{i=1}^M (X_{ki} - \bar{X}_k)^2$$

Now, if it is assumed that the M_k are distributed randomly with expectation M and variance τ^2 , $V(\bar{x})$ can be approximated with

$$\hat{V}(\bar{x}) = \frac{\sigma^2}{K} \left\{ \frac{(1 - \rho)}{M} + \rho + \rho C^2 \right\}$$

where $C = \tau/M$ is the coefficient of variation of the cluster sizes.

Therefore, a confidence interval for μ can be constructed as follows.

$$CI(\mu) = \bar{x} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\bar{x})}$$

The half width of this interval, which we call d , is therefore

$$d = |z_{1-\alpha/2}| \sqrt{\hat{V}(\bar{x})}$$

This can be rearranged to solve for the number of clusters, K , as follows

$$K = \left(\frac{\sigma z_{1-\alpha/2}}{d} \right)^2 \left\{ \frac{(1 - \rho)}{M} + \rho + \rho C^2 \right\}$$

Note that this method is advised in Lohr (2019) page 311.

where $d = (UCL_\mu - LCL_\mu)/2$ which is the *half width* of the confidence interval.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, refer to the Procedure Window chapter.

Design Tab

The Design tab contain most of the parameters and options of interest for this procedure.

Solve For

Solve For

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are the number of clusters, the half width of the interval, or the confidence level. Select the *number of clusters* when you want to calculate the sample size needed. Select the *half width* when you want to investigate the precision of a certain sample size.

Confidence

Confidence Level

Enter the confidence level (or confidence coefficient). This is the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that contain the population mean.

The practical range is between 0.5 and 1. Common values are 0.95 and 0.99. Use 0.9973 if you want z to be 3.0 or 0.977249 if you want z to be 2.0.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Precision

d (Precision, Half-Width)

Enter d , the precision, margin of error, or confidence interval half-width. This is half the distance between the lower and upper confidence limits of the mean. It is also the distance from the mean to either confidence limit.

The formula is $d = |UCL(\mu) - LCL(\mu)|/2$.

The range is $0 < d$.

You can enter a single value or a list of values.

Sample Size – Number of Clusters and Cluster Size

K (Number of Clusters)

This is the number of clusters in the sample.

This value must be a positive number.

You can use a list of values such as “10 20 30”.

M (Average Cluster Size)

This is the average number of items (subjects) per cluster.

This value must be a positive number that is at least 1. It can be a decimal (fractional) number such as 2.7.

You can use a list of values such as “5 8 10”.

Confidence Intervals for One Mean in a Cluster-Randomized Design

COV of Cluster Sizes

Enter the coefficient of variation of the cluster sizes (number of subjects per cluster). This value must be zero or a positive number. You can use a list of values such as “0.4 0.6 0.8”.

Coefficient of Variation

The COV of X is defined as the standard deviation of X divided by the mean of X .

Campbell and Walters (2014) page 71 give guidance on the possible values of COV. They indicate that as the average cluster size increases, COV tends toward 0.65. They say that typical values of COV range from 0.4 to 0.9.

Standard Deviation

The standard deviation, calculated by the sample formula (divide by $M-1$), is a measure of the variability. When no other information is available, Campbell and Walters (2014) page 71 suggest using $(\text{Maximum Cluster Size} - \text{Minimum Cluster Size}) / 4$.

All Cluster Sizes Equal

When all cluster sizes are equal, $\text{COV} = 0$.

Effect Size

This section lets you enter the settings for the standard deviation of the data.

σ (Standard Deviation)

Enter the subject-to-subject standard deviation.

Note that σ must be a positive number.

You can enter a single value such as 5, or a series of values such as 1 3 5 7 9, or 1 to 10 by 1.

When a series of values is entered, PASS will generate a separate calculation result for each value of the series.

ρ (Intraclass Correlation, ICC)

This is the value of the intraclass (or intraclass) correlation coefficient. It may be interpreted as the correlation between any two observations in the same cluster. It may also be thought of as the proportion of the variation in response that can be accounted for by the between-cluster variation.

Possible values are from 0 to just below 1. Typical values are between 0.0001 and 0.3.

You may enter a single value or a list of values.

Confidence Intervals for One Mean in a Cluster-Randomized Design

Example 1 – Finding the Number of Clusters

A study using a cluster design is being planned to estimate the effectiveness of a certain drug in treating high blood pressure. The clusters will be doctor's practices. A sample of patients within the practice will receive the drug and their blood pressure will be measured. The researchers want to compare the number of clusters required when the number of patients measured is 3, 5, 10, 15, 20. The COV of the actual number of patients per cluster is estimated at 0.3.

Prior studies have shown a standard deviation of 35 in the patient-by-patient blood pressure measurement. The intraclass correlation within a practice was shown to be about 0.01. The confidence level is set to 0.95 and d is set to two values: 1 and 1.5.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	K (Number of clusters)
Confidence Level.....	0.95
d (Precision, Half-Width).....	1 1.5
M (Average Cluster Size)	3 5 10 15 20
COV of Cluster Sizes.....	0.3
σ (Standard Deviation)	35
ρ (Intraclass Correlation, ICC).....	0.01

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results								
Solve for: K (number of clusters)								
C.I.	Number	Average	COV of	Total	Std	Intra-	Conf	
Half-	of	Cluster	Cluster	Sample	Dev	cluster	Level	
Width	Clusters	Size	Sizes	Size		Coef	CL	
d	K	M	COV	N	σ	ρ , ICC		
1.0000	1605	3	0.3000	4815	35.0000	0.0100	0.9500	
1.0000	984	5	0.3000	4920	35.0000	0.0100	0.9500	
1.0000	518	10	0.3000	5180	35.0000	0.0100	0.9500	
1.0000	362	15	0.3000	5430	35.0000	0.0100	0.9500	
1.0000	285	20	0.3000	5700	35.0000	0.0100	0.9500	
1.5000	713	3	0.3000	2139	35.0000	0.0100	0.9500	
1.5000	437	5	0.3000	2185	35.0000	0.0100	0.9500	
1.5000	230	10	0.3000	2300	35.0000	0.0100	0.9500	
1.5000	161	15	0.3000	2415	35.0000	0.0100	0.9500	
1.5000	127	20	0.3000	2540	35.0000	0.0100	0.9500	

Confidence Intervals for One Mean in a Cluster-Randomized Design

References

- Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley. New York.
- Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.
- Lohr, Sharon L. 2019. Sampling. Design and Analysis. CRC Press. Boca Raton, FL.

Report Definitions

d is the half-width of the confidence interval of the mean. The formula is $d = (UCL - LCL)/2$.

K is the number of clusters in that are selected in the sample.

M is the average (sample) size of the clusters.

COV is the coefficient of variation of the cluster sizes. If it is zero, all cluster sizes are equal.

N is the combined sample size of all clusters. The formula is $N = K \times M$.

σ is the standard deviation of the subject responses in the whole population.

ρ (ICC) is the intraclass correlation among the responses within a cluster.

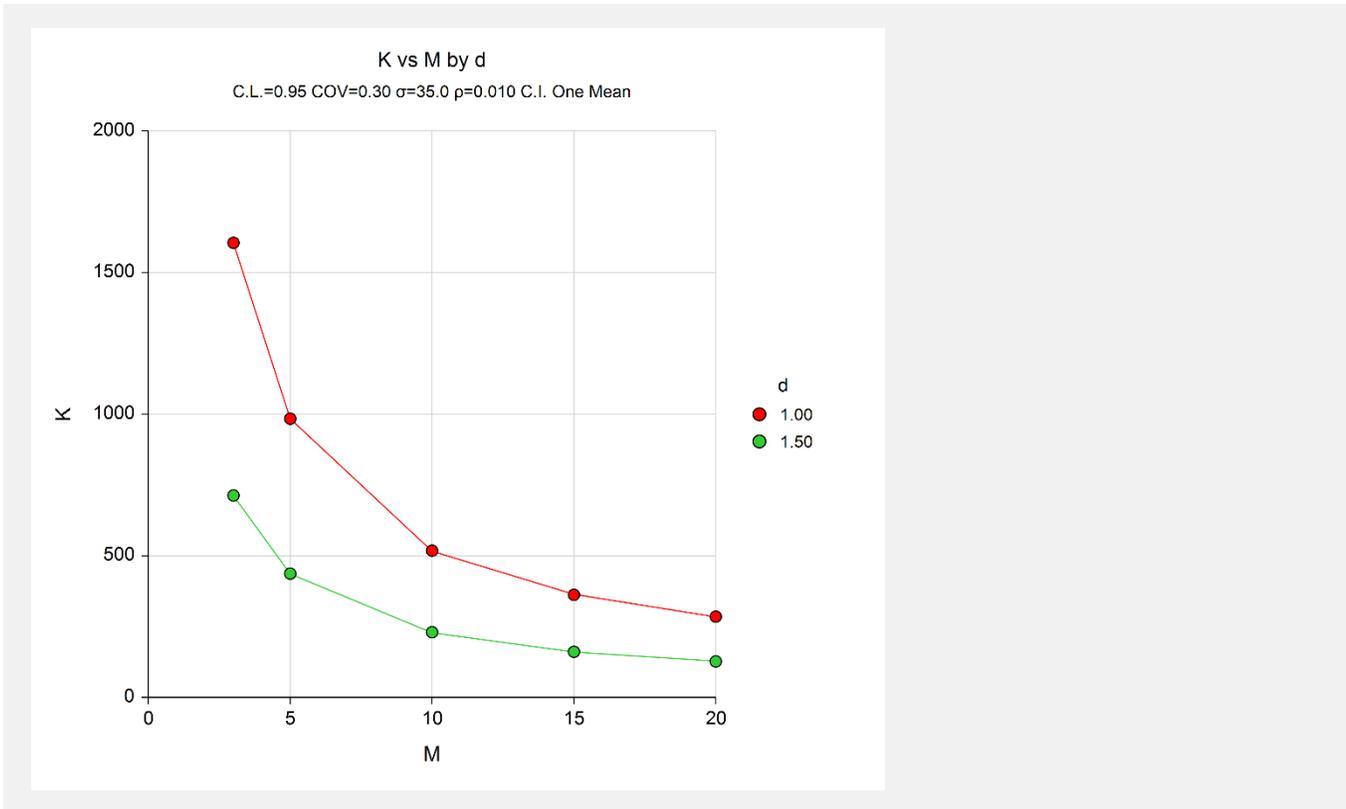
CL is the confidence level of the confidence interval.

Summary Statements

A total sample of 4815 subjects were obtained by sampling 1605 clusters with an average of 3 subjects each. This design achieves a precision (measured by the half-width of the confidence interval of the mean) of 1.0000. The standard deviation of subjects is 35.0000. The intraclass correlation coefficient of subjects within a cluster is 0.0100. The coefficient of variation of cluster sizes is 0.3000. The confidence level of the confidence interval is 0.9500.

This report gives the results for each of the scenarios.

Plots Section



The values from the Numeric Results report are displayed in this plot. Note the large change in K from $M = 3$ to $M = 10$, and the relatively small change in K from $M = 10$ to $M = 20$.

Example 2 – Validation using Hand Calculations

We could not find a published example to use for validating this procedure. Therefore, we will show the calculation of the first row of Example 1. In this example, $CL = 0.95$, $d = 1$, $M = 3$, $COV = 0.3$, $\sigma = 35$, and $\rho = 0.01$.

The calculation of K proceeds as follows.

$$\begin{aligned} K &= \left(\frac{\sigma Z_{1-\frac{\alpha}{2}}}{d} \right)^2 \left\{ \frac{(1-\rho)}{M} + \rho + \rho C^2 \right\} \\ &= \left(\frac{35 \times 1.96}{1} \right)^2 \left\{ \frac{(1-0.01)}{3} + 0.01 + 0.01 \times 0.3^2 \right\} \\ &= 4705.96 \{0.33 + 0.01 + 0.01 \times 0.09\} \\ &= 4705.96 \{0.3409\} \\ &= 1604.26 \text{ or } 1605 \end{aligned}$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	K (Number of clusters)
Confidence Level	0.95
d (Precision, Half-Width)	1
M (Average Cluster Size)	3
COV of Cluster Sizes	0.3
σ (Standard Deviation)	35
ρ (Intracluster Correlation, ICC)	0.01

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results							
Solve for: K (number of clusters)							
C.I. Half-Width d	Number of Clusters K	Average Cluster Size M	COV of Cluster Sizes COV	Total Sample Size N	Std Dev σ	Intra-cluster Corr Coef ρ , ICC	Conf Level CL
1.0000	1605	3	0.3000	4815	35.0000	0.0100	0.9500

PASS also obtains a K of 1605 which validates the procedure.