Chapter 114

Confidence Intervals for One Proportion in a Cluster-Randomized Design

Introduction

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

Suppose that the proportion of a binary 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 proportion may be guaranteed at a certain confidence level.

Technical Details

The following discussion summarizes the results in Ahn *et al.* (2015), pages 24 - 33.

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 binary (0 or 1) 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 proportion is calculated from the cluster proportions as follows.

$$\hat{P} = \frac{\sum_{k=1}^{K} M_k \hat{P}_k}{\sum_{k=1}^{K} M_k}$$

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

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

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

where σ^2 is the variance of X which is equal to $P(1-P)$ and ρ is the correlation of observations within a cluster (often called the intracluster correlation coefficient). The definition of ρ is

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

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

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

$$MSC = \frac{1}{K-1}\sum_{k=1}^{K} M_k\left(\hat{P}_k - \hat{P}\right)^2$$

$$MSW = \frac{1}{(M_T - K)}\sum_{k=1}^{K} x_k\left(1 - \hat{P}_k\right)$$

$$x_k = \sum_{i=1}^{M_k} X_{ki}$$

$$M_T = \sum_{k=1}^{K} M_k$$

$$M_A = \frac{\left(M_T - \sum_{k=1}^{K} M_k^2/M_T\right)}{K-1}$$

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

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

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

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

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

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

$$d = \left|z_{1-\alpha/2}\right|\sqrt{\hat{V}(\hat{P})}$$

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

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

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

where $d = (UCL_P - LCL_P)/2$ 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 proportion being estimated.

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 proportion. It is also the distance from the proportion to either confidence limit.

The formula is d = |UCL(P) - LCL(P)|/2.

The range is 0 < d < 1. However, usually d < 0.5.

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”.

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 (Maximum Cluster Size - Minimum Cluster Size) / 4.

All Cluster Sizes Equal

When all cluster sizes are equal, COV = 0.

Effect Size

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

P (Proportion)

Enter an estimate of the proportion of the event of interest.

Note that $0 < P < 1$.

You can enter a single value such as *0.1*, or a series of values such as *0.1 0.2 0.3*, or *0.1 to 0.5 by 0.1*.

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

ρ (Intracluster Correlation, ICC)

This is the value of the intracluster (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. You may enter a single value or a list of values.

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 evaluated to determine if it is greater than a certain threshold. A binary response will be recorded. 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 an event proportion of 0.4. The intracluster correlation within a practice was shown to be about 0.1. The confidence level is set to 0.95 and *d* is set to two values: 0.05 and 0.1.

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**.

Option	Value
Design Tab	
Solve For	**K (Number of clusters)**
Confidence Level	**0.95**
d (Precision, Half-Width)	**0.05 0.1**
M (Average Cluster Size)	**3 5 10 15 20**
COV of Cluster Sizes	**0.3**
P (Proportion)	**0.4**
ρ (Intracluster Correlation, ICC)	**0.1**

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. Half-Width d	Number of Clusters K	Average Cluster Size M	COV of Cluster Sizes COV	Total Sample Size N	Prop P	Intra-cluster Corr Coef ρ, ICC	Conf Level CL
0.0500	151	3	0.3000	453	0.4000	0.1000	0.9500
0.0500	107	5	0.3000	535	0.4000	0.1000	0.9500
0.0500	74	10	0.3000	740	0.4000	0.1000	0.9500
0.0500	63	15	0.3000	945	0.4000	0.1000	0.9500
0.0500	57	20	0.3000	1140	0.4000	0.1000	0.9500
0.1000	38	3	0.3000	114	0.4000	0.1000	0.9500
0.1000	27	5	0.3000	135	0.4000	0.1000	0.9500
0.1000	19	10	0.3000	190	0.4000	0.1000	0.9500
0.1000	16	15	0.3000	240	0.4000	0.1000	0.9500
0.1000	15	20	0.3000	300	0.4000	0.1000	0.9500

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

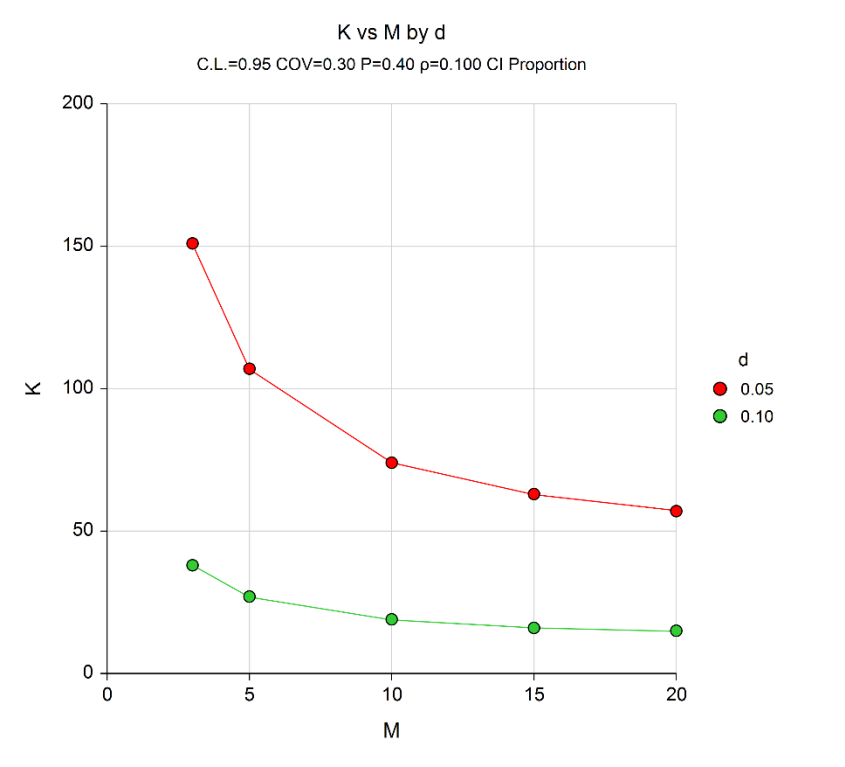
d is the half-width of the confidence interval of the proportion. The formula is d = (UCL - LCL)/2.
K is the number of clusters in that are selected in the sample.
M is the average (sample) size of the clusters.
COV is the coefficient of variation of the cluster sizes. If it is zero, all cluster sizes are equal.
N is the combined sample size of all clusters. The formula is N = K × M.
P is the proportion in the population of subjects that exhibit the response.
ρ (ICC) is the intracluster correlation among the responses within a cluster.
CL is the confidence level of the confidence interval.

Summary Statements

A total sample of 453 subjects were obtained by sampling 151 clusters with an average of 3 subjects each. This design achieves a precision (measured by the half-width of the confidence interval of the proportion) of 0.0500. The response proportion is 0.4000. The intracluster correlation coefficient of subjects within a cluster is 0.1000. 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 = 0.05, M = 6, COV = 0.4, P = 0.2, and ρ = 0.1.

The calculation of K proceeds as follows.

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

$$= \left(\frac{1.96\sqrt{0.2 \times 0.8}}{0.05}\right)^2 \left\{\frac{(1-0.1)}{6} + 0.1 + 0.1 \times 0.4^2\right\}$$

$$= 245.86\{0.15 + 0.1 + 0.1 \times 0.16\}$$

$$= 245.86\{0.266\}$$

$$= 65.4 \text{ or } 66$$

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**.

Option	Value
Design Tab	
Solve For	**K (Number of clusters)**
Confidence Level	**0.95**
d (Precision, Half-Width)	**0.05**
M (Average Cluster Size)	**6**
COV of Cluster Sizes	**0.4**
P (Proportion)	**0.2**
ρ (Intracluster Correlation, ICC)	**0.1**

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	Prop P	Intra-cluster Corr Coef ρ, ICC	Conf Level CL
0.0500	**66**	6	0.4000	396	0.2000	0.1000	0.9500

PASS also obtains a K of 66 which validates the procedure.