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, $V\left(\hat{P}\right)$ 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.

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

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

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

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

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

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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Reports

Numeric Results

Solve For: K (Number of clusters)

Confidence Interval Half-Width d	Number of Clusters K	Average Cluster Size M	Coefficient of Variation of Cluster Sizes COV	Total Sample Size N	Proportion P	Intracluster Correlation (ICC) ρ	Confidence Level CL
0.05	151	3	0.3	453	0.4	0.1	0.95
0.05	107	5	0.3	535	0.4	0.1	0.95
0.05	74	10	0.3	740	0.4	0.1	0.95
0.05	63	15	0.3	945	0.4	0.1	0.95
0.05	57	20	0.3	1140	0.4	0.1	0.95
0.10	38	3	0.3	114	0.4	0.1	0.95
0.10	27	5	0.3	135	0.4	0.1	0.95
0.10	19	10	0.3	190	0.4	0.1	0.95
0.10	16	15	0.3	240	0.4	0.1	0.95
0.10	15	20	0.3	300	0.4	0.1	0.95

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

Summary Statements

A cluster-randomized design will be used to obtain a two-sided 95% confidence interval for a single proportion. The average response proportion is assumed to be 0.4 and the intracluster correlation coefficient of subjects within a cluster is assumed to be 0.1. With an average cluster size of 3, to produce a confidence interval with a half-width of no more than 0.05, a total of 151 clusters will be needed. The total sample size needed is 453.

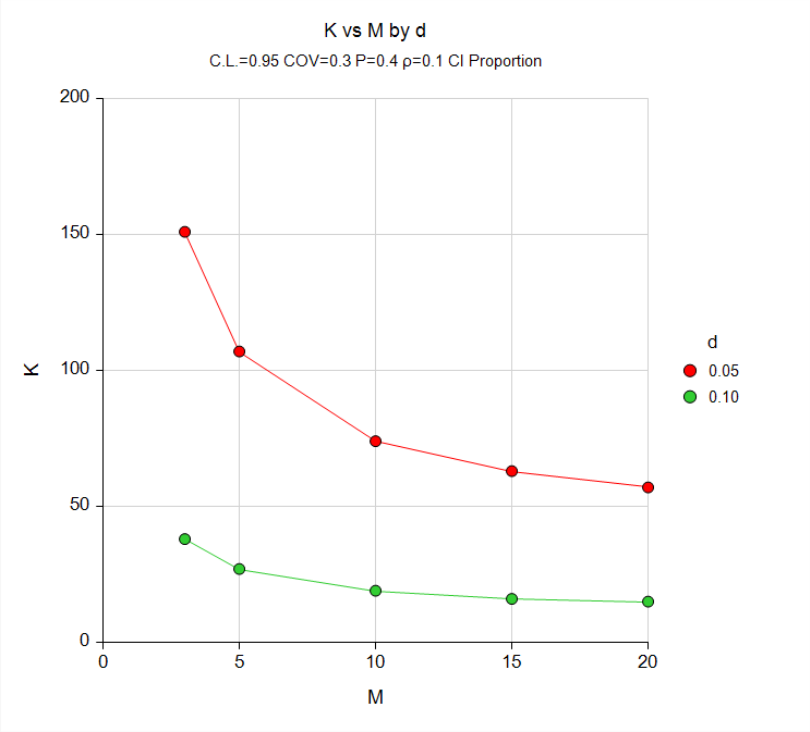
References

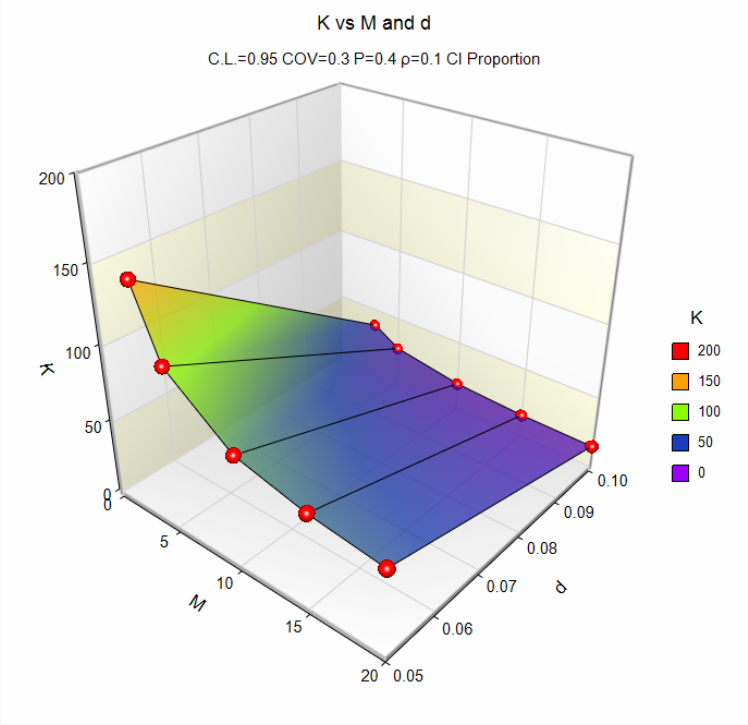
Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley. New York.
Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.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

This report gives the results for each of the scenarios.

Plots Section

Plots





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.

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

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

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

$$= 245.86\{0.266\}$$

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

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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

Solve For: K (Number of clusters)

Confidence Interval Half-Width d	Number of Clusters K	Average Cluster Size M	Coefficient of Variation of Cluster Sizes COV	Total Sample Size N	Proportion P	Intracluster Correlation (ICC) ρ	Confidence Level CL
0.05	66	6	0.4	396	0.2	0.1	0.95

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