Cluster sampling example slideshare

In cluster sampling, the researchers divide a population into smaller groups called clusters. They then randomly select from these clusters to form a sample. Cluster sampling is a probability sampling method often used to study large populations, especially those that are widely dispersed geographically. Researchers typically use pre-existing devices such as schools or cities as their clusters. How cluster samples The simplest form of cluster sampling is single-step sampling of clusters. That means 4 important steps. Research exampleYou are interested in the average reading level for all seventh graders in your city. It would be very difficult to get a list of all seventh graders and collect data from a random sample spread across the city. However, you can easily get a list of all schools and collect data from a subset of these. Thus, you decide to use the cluster sampling method. Step 1: Define your population As with other forms of sampling, you must first begin by clearly defining the population you want to study. PopulationIn your reading study program, your population is all seventh graders in your city. Step 2: Divide your sample into clusters This is the most important part of the process. The quality of your clusters and how well they represent the larger population determines the validity of your results. Ideally, you would like your clusters to meet the following criteria: Each cluster's population should be as diverse as possible. You want each potential property of the entire population to be represented in each cluster. Each cluster should have a similar distribution of characteristics to the distribution of the population as a whole. All in all, the clusters should cover the entire population. There is no overlap between clusters (i.e. the same people or devices do not appear in more than one cluster). Ideally, each cluster should be a mini-representation of the entire population. In practice, however, clusters often do not fully represent the characteristics of the population, which is why this method provides less statistical certainty than simple random sampling. Since clusters are usually naturally occurring groups, such as schools, cities or households, they are often more homogeneous than the population as a whole. You should be aware of this when performing your study, as it may affect its validity. ClustersYou cluster seventh graders of the school they attend. To cover the entire population, you need to include each school in the city. There is no overlap because each student attends only one school. Step 3: Select random clusters to use as your sample If each cluster is itself a mini-representation of the larger population, randomly selecting and sampling from the clusters allows you to imitate simple random sampling, which in turn supports the validity of your results. if the clusters are not representative, then random sampling allows you to collect data on a diverse range of clusters, which should still give you an overview of the population as a whole. SampleYou assign a number to each school and use a random number generator to select a random selection. You select the number of clusters based on how large you want your sample size to be. This in turn is based on the estimated size of the entire seventh-grade population, your desired confidence interval, and confidence level, and your best guess of the standard deviation (a measure of how dispersed the values in a population are) of the reading levels in seventh graders. You then use a sample size calculator to estimate the required sample size. Step 4: Collect data from the sample You then complete your study and collect data from each device in the selected clusters. Data collectionYou test the reading levels of each seventh grader in the schools randomly selected for your selection. Multi-step cluster sampling In multistage clusters, instead of collecting data from each device in the selected clusters, randomly select individual units from within the cluster to use as your sample. You can then collect data from each of these individual devices — this is called double-step sampling. You can also continue this procedure, with progressively smaller and less random samples, commonly referred to as multi-step sampling. You should use this method when it is unpersent or too expensive to test the entire cluster. Example: Multi-step sampling Instead of collecting data from each seventh grader in the selected schools, limit your selection in two additional steps: From each school, randomly select a sample of seventh grades. From within these classes, randomly select a sample of students. The resulting sample is much smaller and therefore easier to collect data from. Professional editors proofread and edit your paper by focusing on: Academic style Vague Sentences Grammar Style texture See an example Pros and cons Cluster sampling is commonly used for its practical advantages, but it has some drawbacks in terms of statistical validity. Benefits Cluster sampling is time- and cost-effective, especially for samples that are widely dispersed and would be difficult to properly sample by other means. Because cluster sampling uses randomization, if the population is clustered correctly, your study will have high external validity because your sample will reflect the characteristics of the larger population. Cons Internal validity is less strong than with simple random selection, especially as you use more stages of clustering. If your clusters are not a good mini-representation of the population as a whole, then it is harder to rely on your sample Results. Cluster sampling is much more complex to plan than other forms of sampling. Cluster sampling FAQs What are the types of cluster sampling? There are three types of cluster sampling: clusters in single-step, double-stage, and multi-stage clusters. In all three types, first divide the population into clusters, then select random cluster for use in your example. In single-step sampling, you collect data from each device within the selected clusters. In double-step sampling, select a random sample of units from within the clusters. In multi-step sampling, repeat the random sampling element procedure from within the clusters until you reach a manageable sample size. Slideshare uses cookies to improve functionality and performance, and to provide you with relevant advertising. If you continue browsing the website, you agree to the use of cookies on this website. See our user agreement and privacy policy. Slideshare uses cookies to improve functionality and performance, and to provide you with relevant advertising. If you continue browsing the website, you agree to the use of cookies on this website. See our privacy policy and user agreement for details. 1. Stat-3203: Sampling Technique-II (Chapter-2: Clusters and Multistage Sampling) Md. Menhazul Abedin Lecturer Statistics Discipline Khulna University, Khulna-9208 E-mail: menhaz70@gmail.com 2. Target and Disposition ⏶Single stage cluster sampling ⏶Cluster sampling with equal and unequal sample size ⏶Properties ⏶Advantages and Disadvantages ⏶Multistage Cluster Sampling (Two Steps) 3. Acknowledgment • Daroga Singh & F. S. Chaudhary • M. Nurul Islam • Ravindra Singh & Naurang Singh Mangat 4. Background... • SRS • Stratified • Systematic 5. Cluster • A cluster is a collection or group, consisting of several (nonhomogeneuos) population elements 6. Intuition... • Study variable: Income/ Awarness/ health status, etc. • Ghatbhogh, Rupsa, Naihati • PSU: Primary sampling unit • Single-stage sampling sample collects information from all individual 7. Intuition... • Upazila Union • Two Stages Sampling PSU SSU 8. Intuition... • Study variable: Income/Awarness/Sunt etc• Multistage sampling division district UpazilaUnion by Households 9. Why cluster sampling? • Feasibility: No collective framework needed • Economy: Cost reduction • Flexibility of clustering: Cluster size manipulation possible (such as political breakdown, administrative breakdown, commercial capital) 10. Cons... • Loss of precision: • Problems in the analysis: • Do you think any other drawbacks...? Please insert here... 11. Cluster Sampling and Other • Cluster Sampling and SRS • Cluster Sampling and Stratified • Cluster Sampling and Systematic 12. Applications 13. Cluster Sampling Cluster-1 Cluster-3Cluster-2 Cluster-4 Cluster-5 sample 14. Definition... • Cluster sampling is a method of sampling, consisting of first selecting, on random groups, called clusters of elements from the population, and then selecting all elements within each cluster to make up the sample. (M. Nurul Islam) 15. Stratified sampling Strata-1 N1 Strata-2 N2 Strata-3 N2 Strata-4 N2 n1 n3n2 n4 N1+N2+ N3+N4= N n1+n2+n 3+n4=n 16. Single-step cluster sampling (equal) Cluster Element 1 2 3 ... I.... N 1 y11 y21 y31 ... yi1 ... y N1 2 y12 y22 y32 ... yi2 ... y N2 ... ... j y1j y2j y3j ... yij ... y Nj ... ... M y1M y2M y3M ... Yim... y NM Cluster total y1 y2 y3 ... Yi... y N Cluster means y1 y2 y3 ... Yi... y N Layout of NM popn element inclusters 17. Single-step cluster sampling (equal) Cluster Element 1 2 3 ... I.... n 1 y11 y21 y31 ... yi1 ... y n1 2 y12 y22 y32 ... yi2 ... y n2 ... ... j y1j y2j y3j ... yij ... y nj ... ... M y1M y2M y3M ... Yim... y nM Cluster total y1 y2 y3 ... Yi... y n Cluster mean y1 y2 y3 ... Yi... y n Layout of nM test element incluster 18. Single-stage cluster sampling (equal) • Indivisual cluster mean • yi = 1 M yi1 + + + yiM = y i M = 1 M j=1 M yij • n cluster mean (sample mean) • yn = 1 n i=1 n yi • Sample mean 1y = y nM = 1 nM i=1 n j=1 M yij = 1 nM i=1 n yi = 1 nM i=1 n M yi = 1 n i=1 n yi = yn= n cluster mean = n cluster mean 19. Single-stage sampling of clusters (equal) • N cluster mean YN = 1 N i=1 N yi • Population mean Y = Y NM = 1 NM i=1 N j=1 M yij = 1 NM i=1 N yi = 1 NM i=1 N M yi = 1 N i=1 N yi = Yn = N = N cluster mean population = N cluster mean 20. Single-stage cluster sampling (equal) • Variance calculation: V yn = N − n N 1 n 1 M2 i=1 N yi − i=1 N yi N 2 N − 1 V yn = N − n N 1 n i=1 N yi − Y 2 N − 1 = 1−f n S b 2 • Replace S b 2 with s b 2 = i=1 n y i− y n 2 n−1 • Estimator of V yn is v yn = 1−f n s b 2 21. Single step sampling of clusters (equal) • Theorem 8.1: defined mean is unbiased and estimate the variance of the mean. (Need correlation is discussed next image) • V yn = (1−f)(NM−1) nM2(N−1) S2 [1 + (M − 1)ρ] Or V yn ≈ 1−f nM S2 [1 + (M − 1)ρ] 22. Correlation between clusters • The similarity between observations within a cluster can be quantified using intracer Correlation Coefficient (ICC), sometimes also called intraclass correlation coefficient. • This is very similar to the well-known Pearson correlation coefficient; just that we do not simultaneously look at observations of two variables on the same object, but at the same time we look at two values of the same variable, but taken at two different objects. • Calculation as Auto-correlation (discussed) 23. Intra-cluster correlation • Mean squared between elements in population S2 = i=1 N j=1 M y ij− Y 2 NM−1 • Intra cluster correlation ρ = E(yij − Y)(yjk − Y) E yij − Y 2 = 2 i=1 N j=1<k M (yij − Y)(yjk − Y) (M − 1)(NM − 1)S2 24. Variance in terms of ρ • V yn = N−n N 1 n i=1 N y i− Y 2 N−1 • Expand the square term and relate with ρ • V yn = (1−f)(NM−1) nM2(N−1) S2 [1 + (M − 1)ρ] • About N large NM − 1 ≈ NM and N − 1 ≈ N • V yn ≈ 1−f nM S2 [1 + (M − 1)ρ] • V yn = 1−f nM S2 [1 + (M − 1)ρ] [simplicity] 25. Design effect • Variance for Cluster sampling • V yn = 1−f nM S2 [1 + (M − 1)ρ] • Variance for Simple random sampling • V y nM = NM−nM NM S2 nM = 1−f nM S2 • Dividing V y nM = 1 + M − 1 ρ = definitive • What is inter pretation of Design effect? It's simple, you can find it. Try your best. 26. Ratio of ρ, definitely and M • Deff = 1 + M − 1 ρ – See its property when – ρ = 1 [definitely =M all M values in a cluster are equal] – M = 1 [SRS = cluster sampling] – ρ = 0 [void cluster – Deff = 0 or +1 find range for intra-cluster correlation 27. Cluster sampling efficiency • V y nM V n = 1 1+ M−1 ρ = 1 Deff • Consider its properties when – ρ > 0 Cluster sampling less effective compared to SRS – ρ < 0 Cluster sampling more efficiently compared to SRS 28. Single-step cluster sampling (Equal) • Find Optimum n and M subject to restriction cost. – Ignore the provisionally 29. Example • Example: 8.2 • Example: 8.3 30. Single-step sampling of Unequal Cluster clusters 31. Single-stage cluster sampling (Unequal) ClusterElements 1 2 3 ... I.... N 1 y11 y21 y31 ... yi1 ... y N1 2 y12 y22 y32 ... yi2 ... y N2 ... ... j y1j y2j y3j ... yij ... y Nj ... ... Mi y1M1 y2M2 y3M3 ... yiM I ... y NM N Cluster total y1 y2 y3 ... Yi... y N Cluster means y1 y2 y3 ... Yi... y N 32. • Total number of elements M0 = i=1 N Mi • Total number of elements in each cluster yi = j=1 M in yij • Average number of elements per cluster M = i=1 N Mi N = M0 N Single-stage cluster sampling (Unequal) 33. Single-stage cluster sampling (Unequal) • Population mean (1) Y = i=1 N j=1 M i yij i=1 N Mi = i=1 N Mi yi i=1 N Mi = i=1 N Mi yi M0 • Population mean (2) YN = i=1 N yi N • Are they the same? 34. Single-stage cluster sampling (Unequal) • Sample mean (1) yn = i=1 n y n n N Bias for Y but unbiased for YN • Sample mean (2) • yn = N nM0 i=1 n Mi yi = 1 n i=1 n (M i y i M) This is unbiased for Y 35. Single-step Cluster Sampling (Unequal) Cluster Elements 1 2 3 ... I.... N 1 y11 y21 y31 ... yi1 ... y N1 2 y12 y22 y32 ... yi2 ... y N2 ... ... j y1j y2j y3j ... yij ... y Nj ... ... M y1M1 y2M2 y3M3 ... yiM I ... y NM N Cluster total y1 y2 y3 ... Yi... y N Cluster means y1 y2 y3 ... Yi... y N 36. • Take an example Single-stage cluster sampling (Unequal) 37. • Additional study - Cluster sampling with PPS sampling (No need at present) Single-stage cluster sampling (Unequal) 38. Background... • A device may contain too many elements to get a measurement on each • A device may contain elements that are almost the same. Multi-step sampling of clusters (Two-step) 39. Background... • V y nM V y n = 1 1+ M−1 ρ or V Cluster V SRS = 1 1+ M−1 ρ – What will happen when M increases?????? • Less efficient cluster sampling • Large cluster draws small sample Two-stage cluster sampling 40. • Partial sampling (two-step sampling) • A two-stage cluster is one, obtained by first selecting a selection of clusters and then again selecting a selection of elements from each sample cluster. • By → Household (partial sample) Multi-stage cluster sampling (Two-stage) 41. Multistage sampling (Two Steps) Cluster Mi Population Element Total Cluster Mean 1 M1 y11, y12, ... , y1j, ... , y1M1 Y1 = j=1 M 1 y 1j Y1 = Y1 M1 2 M2 y21, y22, ... , y2M2 Y2 = j =1 M 2 y 2j Y2 = Y2 M2 ... ... in Mi yi1, yi2, ... , yij, ... , yiM i Yi = j =1 M i yij Yi = Yi Mi ... ... N M N y N1, y N2, ... , y Nj, ... , y NM N YN = j=1 M N y Nj YN = YN M N 42. • Y = i=1 N Yi = i=1 N j=j M i yij • M0 = i=1 N Mi • Yi = j=j M i y ij M i = Y i M i • Population mean Y = i=1 N j=j M i yij i=1 N Mi = Y M0 = i=1 N Yi M0 = i=1 N Yi M0 = i=1 N Mi Yi M0 • Population pooled mean Yi = i=1 N Yi N = j=j M i yij N = i= i= 1 N Mi Yi N Multi-step cluster sampling (Two-step) Red and blue mean are different. Red is individual cluster mean but blue is polled mean 43. Multi-stage cluster sampling (Two-stage) Unit Mi mi Sample observation Total Cluster mean 1 M1 m1 y11, y12, ... , y1j, ... , y1m1 y1 = j =1 m 1 y 1j y1 = y1 m1 2 M2 m2 y21, y22, ... i Mi mi yi1, yi2, ... , yij, ... , yim i yi = j =1 m i yij yi = yi mi ... ... n M n y n1, y n2, ... , y nj, ... , ynm n y N = j=1 m n y nj yn = yn m n 44. • y = i=1 n yi = i=1 n j=j m i yij • m0 = i=1 n mi , m = m0 n • Average value per second stage unit • yi = j = j m i y ij m i = y i m i , y = y m0 • Average value per first step unit yn = y n = i=1 n j=j m i y ij n Multistage

Cluster Sampling (Two Stage) 45. • Number of estimators is defined (You can define more with good properties as a researcher) • surface(1) = i=1 n y n regular mean based on the first stage unit mean. • surface(2) = i=1 n M i y i n M = N M0 i=1 n M i y i n based on M0 = i=1 n M i y i i=1 n M i = surface Known as ratio estimator • YR = i=1 n M in y i i=1 n M in total multistage cluster sampling estimator (Two-stage) replaces M0by M0 = N i=1 n M in n 46. Why such Graffiti Features? • in th cluster total= Mi yi • Estimates of total Y over selected n clusters i=1 n Mi yi • Average value for Y per cluster is i=1 n M i y i n • Estimator of total Y over N cluster N n i =1 n Mi yi • Total= Total value of Y per cluster is i=1 n M i y i n • Estimator of total Y above N cluster N n i=1 n Mi yi • Total= Total= Total frequency × mean 47. Multi-stage cluster sampling (Two-stage) Unit Mi mi Sample observation Total Cluster mean 1 M1 m1 y11, y12, ... , y1j, ... , y1m1 y1 = j =1 m 1 y 1j y1 = y1 m1 2 M2 m2 y21, y22, ... i Mi mi yi1, yi2, ... , yij, ... , yim i yi = j =1 m i yij yi = yi mi ... ... n M n y n1, y n2, ... , y nj, ... , ynm n y N = j=1 m n y nj yn = yn m n 48. Why such Graffiti Features? • N n i=1 n Mi yi = Y = M0 × mean = M0 N M0 i=1 n M i y i n • mean = Y M0 [Estimator for Y] • Thus surface(2) = N M0 i=1 n M i y i n 49. Unbiased... • Theorem 9.1: Estimation souths(2) is unconditional and its variance is given by V y ts 2 = 1 − f1 1 M2 S b 2 n + 1 nN M2 i=1 N Mi 2 1 − f2i Si 2 mi Where f1 = n N , f2i = m in M in Condition is given next figure 50. Conditional expectation • E X = E[E X Y] • E u = E1 E2 u b∗ = j p b∗ = B∗ (E u B∗) • E1 is unconditional in our context the expectation of first stage selection • E2 conditional expectation in our context expectations of second stage selection from a given set of first stage units. • V x = V E X Y + E V X Y • V y ts 2 = V1 E2 y ts 2 n + E1 V2 y ts 2 n 51. Advantages • Flexible than one step • Quality control purpose • Large survey • Less cost & more convenience over stratified sampling of the same size 52. • Study examples examples