Applied Statistics

Gaia Bertarelli, Sant'Anna School of Advanced Studies (Italy)

Statistics, Statistical learning, Computing and Data Analytics
February, 15, 2022

The presentation at a glance

Stratified Sampling

Cluster sampling

Non-probability sampling techniques

Stratified Sampling

Stratified Sampling

- If the variable of interest takes on different mean values in different subpopulations, we may be able to obtain more precise estimates of population quantities by taking a **stratified** random sample.
- We divide the populationi in H subpopulation. Each subpopulation is called Stratum. Strata do not overlap.
- We draw an independent probability sample from each stratum, **then pool the information to obtain overall population estimates**.

Why Stratified Sampling?

- Protected from the possibility of vary bad sample.
- Known precision for subgroups of the population.
- More convenient to administer.
- Stratified sampling often gives more precise (lower variance) estimates of population means and totals.

Notation

- H: number of strata.
- N: population size.
- N_h: population size in stratum h.
- $N = N_1 + N_2 + N_3 + \cdots + N_H$
- In the stratified random sampling we independently take an SRS from each stratum: n_h units are randomly selected from the N_h population units in stratum h.
- The total sample size is $n = n_1 + n_2 + n_3 + \cdots + n_H$

Notation for Stratification: The population quantities are:

$$y_{hj} = \text{value of } j\text{th unit in stratum } h$$

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{population total in stratum } h$$

$$t = \sum_{h=1}^{H} t_h = \text{population total}$$

$$\bar{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} = \text{population mean in stratum } h$$

$$\bar{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}}{N} = \text{overall population mean}$$

$$S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1} = \text{population variance in stratum } h$$

Estimates in each Stratum

- Mean:
$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj}$$
- Total:
$$\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} = N_h \bar{y}_h$$
- Sampling variance:
$$s_h^2 = \sum_{j \in S_h} \frac{(y_{hj} - \bar{y}_h)^2}{(n_h - 1)}$$

Estimates in Stratifies Sampling

- Population total estimator:

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}$$

- Population mean estimator:

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}$$

- $\frac{N_h}{N}$: the proportion of the population units in stratum h.
- Population proportion estimator:

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$

- The properties of these estimators follow the properties of SRS estimators (unbiased, definiton of Variance, Standard Errors and Confidence intervals)

- **Unbiasedness.** $\bar{y}_{\text{str}}$ and $\hat{t}_{\text{str}}$ are unbiased estimators of $\bar{y}_U$ and t. An SRS is taken in each stratum, so (2.30) implies that $E[\bar{y}_h] = \bar{y}_{hU}$ and consequently

$$E\left[\sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N} E[\bar{y}_h] = \sum_{h=1}^{H} \frac{N_h}{N}\, \bar{y}_{hU} = \bar{y}_U.$$

- **Variance of the estimators.** Since we are sampling independently from the strata, and we know $V(\hat{t}_h)$ from the SRS theory, the properties of expected value in Section A.2 and (2.16) imply that

$$V(\hat{t}_{\text{str}}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h}. \tag{3.3}$$

- **Standard errors for stratified samples.** We can obtain an unbiased estimator of $V(\hat{t}_{str})$ by substituting the sample estimators s_h^2 for the population parameters S_h^2. Note that in order to estimate the variances, we need to sample at least two units from each stratum.

$$\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h} \tag{3.4}$$

$$\hat{V}(\bar{y}_{str}) = \frac{1}{N^2}\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}. \tag{3.5}$$

As always, the standard error of an estimator is the square root of the estimated variance: $\mathrm{SE}(\bar{y}_{str}) = \sqrt{\hat{V}(\bar{y}_{str})}$.

- **Confidence intervals for stratified samples.** If either (1) the sample sizes within each stratum are large, or (2) the sampling design has a large number of strata, an approximate $100(1-\alpha)\%$ confidence interval (CI) for the population mean $\bar{y}_U$ is

$$\bar{y}_{\text{str}} \pm z_{\alpha/2}\, \text{SE}\,(\bar{y}_{\text{str}}).$$

  The central limit theorem used for constructing this CI is stated in Krewski and Rao (1981). Some survey software packages use the percentile of a t distribution with $n - H$ degrees of freedom (df) rather than the percentile of the normal distribution.

Sampling weights in Stratified Sampling

- In stratified sampling it is possible to have different inclusion probabilities in different strata $\rightarrow$ weights may be unequal in different strata.
- The sampling weight is the number of units in the population represented by the sample member y_{hj}.
- the sampling weight is

$$w_{hj} = \frac{N_h}{n_h}$$

Major issues

- Stratified sampling has three major design issues:
    1. Defining the strata.
    2. Choosing the total sample size.
    3. Allocating the observations to the defined strata.

Allocating Observations to Strata

- Now we assume that the strata have already been fixed and we study methods of allocating observations to the strata.
- Different types of allocation:
  1. Equal Allocation ($n_1 = n_2 = n_3 = \cdots = n_H$)
  2. Proportional Allocation
  3. Optimal Allocation
  4. Allocation for specified precision within strata
- then it is necessary to define the sample size.

Proportional Allocation

- **Goal: obtain a sample that reflects the population with respect to the stratification variable.** The final sample is a miniature version of the population $\rightarrow$ the number of sampled units in each stratum is **proportional to the size of the stratum**$\rightarrow \frac{n_h}{n} = \frac{N_h}{N} \rightarrow n_h = \left(n \frac{N_h}{N} \right)$
- The inclusion probability $\pi_{hj} = \frac{n_h}{N_h} = \frac{n}{N}$ is **the same for all strata** as in SRS but in SRS more *bad* samples are possible.
- Each unit in the sample represent the same number of units in the population.

Proportional Allocation (ii)

- The stratified sampling estimator of the population mean (e.g.) is simply the average of all the observation.
- When the strata are large enough, the variance of the stratified estimator (e.g.mean) under proportional allocation is usually at most as large as the variance of the estimator under the SRS with the same number of observation.
- The more unequal the stratum estimates are, the more precision the researcher will gain by using proportional allocation.
- If the variances are more or less equal across the strata proportional allocation is the best option tu increase precision.

Optimal Allocation

- **Idea: to sample more in larger strata and where variability of y is greater**.
- Optimal allocation works well for sampling units such as corporations, cities, hospitals which vary a lot in size.
- General optimal allocation works well when some strata are much more expensive to sample than others.
- When calculating n_h^{opt} we must take S_h as known. If they are not, it is necessary to approximate.
- It can happen that $n_h^{opt} \geq N_h$ for some h. In this case we can decide to census the stratum, e apply the optimal allocation algorithm to the remaining strata.

Optimal Allocation (ii)

- Optimal Allocation

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) n$$

where c_h represent the cost of taking an observation in stratum h.

- Sampling heavily within a stratum if:
    1. The stratum accounts for a large part of the population;
    2. The variance within the stratum is large: more sample size to compensate for the heterogeneity;
    3. Sampling in the stratum is not expensive.

Neyman allocation

- Neyman allocation is a special case of optimal allocation.
- It is uses when the cost in the strata (but not the variances) are apprx. equal.
- $n_h \propto N_h S_h$
- The formula is.

$$n_h = \left(\frac{N_h S_h}{\sum_{l=1}^{H} N_l S_l} \right) n$$

- If S_h are well allocated, Neyman allocation will give an estimator with smaller variance than the proportional allocation.

Lessons learned

- Stratified sampling is more efficient of simple random sampling. *We can allocate the sample efficiently if we can locate the strata where the variability of the study variable is concentrated*.
- When we have positive **asymmetric populations** (many small values, few large values) is worthwhile try to intensively sample the right tail of the distribution.
- The optimal allocation is based on the consideration of a single variable as target.

Sample sizes

- If you are interested in the estimates in the strata: sample size coming from SSR.
- The different methods of allocating observations to strata give the relative samle size $\frac{n_h}{n}$.
- After the construction of strata and after that observations are allocated to strata:

$$V(\bar{y}_{\text{str}}) \leq \frac{1}{n} \sum_{h=1}^{H} \frac{n}{n_h} \left(\frac{N_h}{N} \right)^2 S_h^2 = \frac{v}{n},$$

where $v = \sum_{h=1}^{H} (n/n_h)(N_h/N)^2 S_h^2$. Thus, if the fpcs can be ignored and if the normal approximation is valid, an approximate 95% CI for the population mean will be $\bar{y}_{\text{str}} \pm z_{\alpha/2}\sqrt{v/n}$. Set $n = z_{\alpha/2}^2 v/e^2$ to achieve a desired margin of error e.

Please read Section 3.5 *Defining Strata* in your textbook Lohr (2019).

Cluster sampling

Cluster sampling

Suppose we want to find out how many bicycles are owned by residents in a community of 10,000 households. We could take a simple random sample (SRS) of 400 households, or we could divide the community into blocks of about 20 households each and sample every household (or subsample some of the households) in each of 20 blocks selected at random from the 500 blocks in the community. The latter plan is an example of cluster sampling. The blocks are the **primary sampling units** (psus), or **clusters**. (In this chapter, we use the terms cluster and psu interchangeably.) The households are the **secondary sampling units** (ssus); often the ssus are the elements in the population.

Cluster sampling

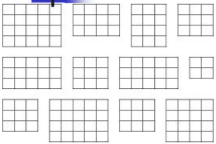
- In cluster sampling individual elements of the population are allowed in the sample only in they belong to a cluster (psu) that is included in the sample.
- The sampling unit (psu) is not the same as the observation unit (ssu)

Glossary

- Cluster sampling: A probability sampling design in which observations are grouped into clusters (psu). A probability sample of psus is selected from the populationof psus.
- Primary sampling unit (psu): the unit that is sampled from the popualtion.
- Secondary sampling unit: a subunit that is subsampled from the selected psus.
- One-stage cluster sampling: a cluster sampling design in which al ssus in selected psus are observed.
- Two-stage cluster sampling:a cluster sampling design in which the ssus in selected psus are subsampled.

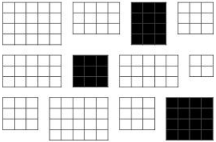
Sample of 40 elements

1-stage CS



A block of cells is a cluster

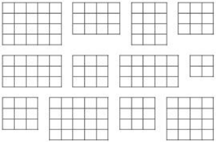
Take an SRS of clusters; observe all elements within the clusters in the sample:



SU is a cluster

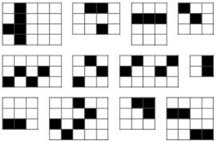
Don't sample from every cluster

ST



A block of cells is a stratum

Take an SRS from ever stratum:

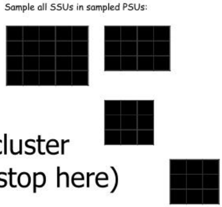


SU is an element (or OU)

Sample from every stratum

4

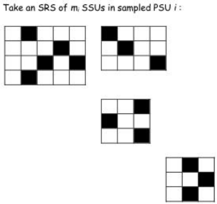
1-stage vs. 2-stage cluster sampling



1-stage cluster sample (stop here)

OR

Stage **1** of 2-stage cluster sample (select PSUs)



Stage **2** of 2-stage cluster sample (select SSUs w/in PSUs)

7

Notation Cluster Sampling

- The universe U is the population on N psus.
- N here is the number od psus not of observation units.
- S is the sample of psus.
- S_i is the sample of ssus chosen from the i^{th} psu.
- y_{ij} measurement for j^{th} element in i^{th} osu

Notation Cluster Sampling (ii)

psu Level—Population Quantities

$$N = \text{ number of psus in the population}$$

$$M_i = \text{ number of ssus in psu } i$$

$$M_0 = \sum_{i=1}^{N} M_i = \text{ total number of ssus in the population}$$

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{ total in psu } i$$

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{ population total}$$

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2 = \text{ population variance of the psu totals}$$

Notation Cluster Sampling (iii)

ssu Level—Population Quantities

$$\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0} = \text{ population mean}$$

$$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i} = \text{ population mean in psu } i$$

$$\mathcal{S}^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1} = \text{ population variance (per ssu)}$$

$$\mathcal{S}_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1} = \text{ population variance within psu } i$$

Notation Cluster Sampling (iv)

Sample Quantities

$$n = \text{ number of psus in the sample}$$

$$m_i = \text{ number of ssus in the sample from psu } i$$

$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i} = \text{ sample mean (per ssu) for psu } i$$

$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = \text{ estimated total for psu } i$$

$$\hat{t}_{\text{unb}} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i = \text{ unbiased estimator of population total}$$

$$s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2$$

Notation Cluster Sampling (v)

$$s_i^2 = \sum_{j \in \mathcal{S}_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1} = \text{ sample variance within psu } i$$

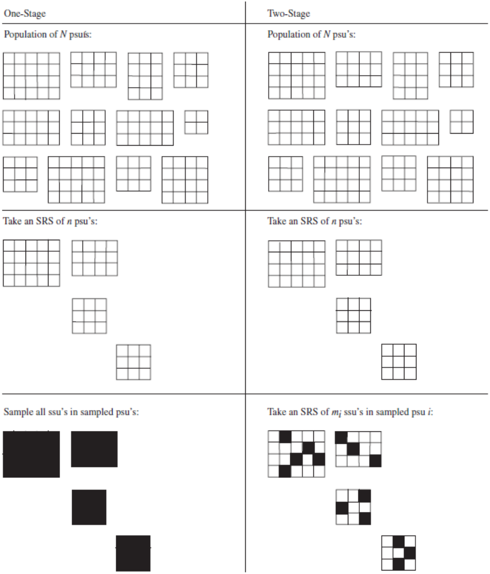
$$w_{ij} = \text{ sampling weight for ssu } j \text{ in psu } i$$

One-stage cluster sampling

- One-stage cluster sampling is used in many surveys in which **the cost of sampling ssus is small compared to the cost of sampling psus**.
- In the simplest design, we tahe an SRS of n psus from the population and measure our variable of interest in **every element in the sampled psus** $\rightarrow M_i = m_i$
- One-stage cluster sampling with an SRS of psus produces a self-weighted sample: a sample in which all probabilities of inclusion π_i are equal, so that al sampling weights w_i are the same.
- $w_{ij} = \frac{N}{n}$

Two-stage cluster sampling

- It may be expensive to measure ssus relative to the cost of sampling psus.
- Two-stage cluster sampling:
  1. Select an SRS S OF n psus from the population of N psus.
  2. Select an SRS of ssus from each selected psus. The SRS of m_i elements from the i^{th} psu denote S_i.


One-Stage
Two-Stage
Population of N psu's:
Population of N psu's:
Take an SRS of n psu's:
Take an SRS of n psu's:
Sample all ssu's in sampled psu's:
Take an SRS of m_i ssu's in sampled psu i:

Inclusion probability and weight

- $$\pi_{ij} = P(j^{th} \text{ ssu in } i^{th} \text{ psu is selected}) =$$
  $$= P(i^{th} \text{ psu selected }) \times P(j^{th} \text{ ssu selected}|i^{th} \text{ psu selected}) =$$
  $$= \frac{n}{N}\frac{m_i}{M_i}$$
- $$w_{ij} = \frac{1}{\pi_{ij}} = \frac{NM_i}{nm_i}$$
- ssu j in psu i represents $\frac{1}{\pi_{ij}} = \frac{NM_i}{nm_i}$ ssus in the population: itself and $\frac{1}{\pi_{ij}} = \frac{NM_i}{nm_i} - 1$ units which are not sampled.

Cluster sampling: issues

1. What overall precision is needed? $\rightarrow$ common in all the survey designs
2. What size should psus be? $\rightarrow$ often natural.
3. How many ssus should be sampled in each psu selected for the sample?
4. How many psus should be sampled?

- Intraclass correlation coefficient (ICC): the pearson correlation coefficient of all pairs of unit within the same cluster.

Choosing Subsampling sample size

- Goal: Most information possible for the least cost and inconvenience.
- Suppose that we want to conduct a two-stage cluster survey when all the psus have the same number, equal to M,of ssus.
- Consider the simple total cost function

$$C = c_1 n + c_2 nm$$

  where

  - c_1 cost per psu (not including c_2
  - c_2 cost of measuring each ssu

Choosing Subsampling sample size (ii)

- The optimal m and n are the values that minimize the variance for the fixed total cost C.
- We have

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$$

and

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

- R_a^2 is a measure of homogeneity defined as

$$R_a^2 = 1 - \frac{MSW}{S^2}$$

and MSW is the pooled value of the within-cluster variances.

Although we discussed only designs where all M_i's are equal, we can use these methods with unequal M_i's as well: just substitute $\bar{M}$ for M in the above work, and decide the average subsample size $\bar{m}$ to take. Then either take $\bar{m}$ observations in every cluster, or allocate observations so that

$$\frac{m_i}{M_i} = \text{constant}.$$

As long as the M_i's do not vary too much, this should produce a reasonable design.

Number of psus (choosing the sample size)

After the psu size is determined and the subsampling fraction set, we then look at the number of psus to sample, n. Like any survey design, design of a cluster sample is an iterative process: (1) Determine a desired precision, (2) choose the psu and subsample sizes, (3) conjecture the variance that will be achieved with that design, (4) set n to achieve the precision, and (5) iterate (adding stratification and auxiliary variables to use in ratio estimation) until the cost of the survey is within your budget.

If clusters are of equal size and we ignore the psu-level fpc, (5.30) implies that

$$V(\hat{\bar{y}}_{\text{unb}}) \leq \frac{1}{n}\left[\frac{\text{MSB}}{M} + \left(1 - \frac{m}{M}\right)\frac{\text{MSW}}{m}\right] = \frac{1}{n}v.$$

Number of psus (choosing the sample size)

An approximate $100(1-\alpha)\%$ CI will be

$$\hat{\bar{y}}_{\text{unb}} \pm z_{\alpha/2}\sqrt{\frac{1}{n}v}.$$

Thus, to achieve a desired CI half-width e, set $n = z_{\alpha/2}^2 v / e^2$

Non-probability sampling

Types of Non-Probability samples

In probability sampling, units are selected randomly from a population with a predetermined selection probability. This feature of the sample design ensures that a probability sample with full response has no selection bias. When accurate data are obtained from every unit selected for a probability sample, the only type of error is sampling error, which can be estimated using the methods in Chapter 9.

Nonprobability samples are not guaranteed to be free of selection bias, but some types may have less bias than others. In addition, even if they have selection bias, nonprobability samples may be useful for answering questions of interest about a population, and they are often easier and less expensive to collect than probability samples. A medical study may be conducted on volunteers, but if an experimental treatment reduces migraines in the study patients randomly assigned to that treatment, it may also be effective for persons not in the study.

Non-probability sampling

- Non-probability techniques, relying on the judgment of the researcher or on accident, cannot generally be used to make generalizations about the whole population.
- Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying.
- A core characteristic of non-probability sampling techniques is that samples are selected based on the **subjective judgement** of the researcher, rather than random selection (i.e., probabilistic methods)

Principles of non-probability sampling

- Non-probability sampling represents a valuable group of sampling techniques that can be used in research that follows
  - qualitative,
  - mixed methods,
  - and even quantitative research designs.
- Researchers following a quantitative research design often feel that they are forced to use non-probability sampling techniques because of some inability to use probability sampling (e.g., the lack of access to a list of the population being studied). However, this is not the case for researchers following a **qualitative research design**.

Theoretical reasons

- Unlike probability sampling, **the goal is not** to achieve objectivity in the selection of samples, or necessarily attempt **to make generalisations** (i.e., statistical inferences) from the sample being studied to the wider population of interest.
- Instead, researchers following a qualitative research design tend to be interested in the **intricacies** of the sample being studied. Whilst making generalisations from the sample to the population under study may be desirable, it is more often a secondary consideration.
- Even whether this is desired, there are additional problems of bias and transferability (or validity).

Practical reasons

- Non-probability sampling is often used because the procedures used to select units for inclusion in a sample are much easier, quicker and cheaper when compared with probability sampling.
- To sample hidden or hard-to-reach population where a list of the population simply does not exist.
- Non-probability sampling can also be particularly useful in exploratory research where the aim is to find out if a problem or issue even exists in a quick and inexpensive way.

Main types of non-probability sampling

- Quota sampling
- Convenience sampling
- Purposive sampling
- Self-selection sampling
- Snowball sampling

Basic ideas: Quota sampling

- Quota sampling is a sampling methodology wherein data is collected from a homogeneous group.
- It involves a two-step process where two variables can be used to filter information from the population.
- It can easily be administered and helps in quick comparison.
- Many persons confuse quota sampling with stratified sampling. In quota sampling, quota classes are formed that serve the role of strata, but **the survey taker uses a nonprobability sampling method such as convenience sampling to reach the desired sample size** in each quota class.

Basic ideas: convenience sampling

- A convenience sample is simply one where the units that are selected for inclusion in the sample are the **easiest to access**.

Basic ideas: purposive sampling

- Purposive sampling, also known as judgmental, selective, or subjective sampling, is a form of non-probability sampling in which researchers **rely on their own judgment** when choosing members of the population to participate in their surveys.
- This survey sampling method requires researchers to have **prior knowledge** about the purpose of their studies so that they can properly choose and approach eligible participants for surveys

Purpose vs Convenience sampling

- The terms purposive sampling and convenience sampling are often used interchangeably, but they do not mean the same thing.
- Convenience sampling is when researchers leverage individuals that can be identified and approached with as little effort as possible. These are often individuals that are geographically close to the researchers or those who have previously completed an online survey.
- Purposive sampling is when researchers thoroughly think through how they will establish a sample population, even if it is not statistically representative of the greater population at hand. As the name suggests, researchers went to this community on purpose because they think that these individuals fit the profile of the people that they need to reach.

Purpose vs Convenience sampling

- While the findings from purposive sampling do not always have to be statistically representative of the greater population of interest, they are qualitatively generalizable.
- The more prior information that researchers have about their particular communities of interest, the better the sample that they're going to select.

Basic ideas: Self-selection sampling

- Self-selection sampling is appropriate when we want to allow units or cases, whether individuals or organisations, to choose to take part in research on their own accord.
- The key component is that research subjects (or organisations) volunteer to take part in the research rather than being approached by the researcher directly.
- A sample is self-selected when the inclusion or exclusion of sampling units is determined by whether the units themselves agree or decline to participate in the sample, either explicitly or implicitly.

Basic ideas: Snowball sampling

- Snowball sampling is particularly appropriate when the population you are interested in is **hidden and/or hard-to-reach** (drug addicts, homeless people, individuals with AIDS/HIV, prostitutes . . .)
- This is a sampling technique, in which **existing subjects provide referrals to recruit samples** required for a research study.
- This sampling method involves a **primary data source nominating other potential data sources** that will be able to participate in the research studies.
- Snowball sampling consists of two steps:
  1. Identify potential subjects in the population. Often, only one or two subjects can be found initially.
  2. Ask those subjects to recruit other people (and then ask those people to recruit. Participants should be made aware that they do not have to provide any other names.

Gaia Bertarelli

Department of Economics and Management in the Era of Data Science

gaia.bertarelli@santannapisa.it