# Applied Statistics

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

Statistics, Statistical Learning, Computing and Data Analytics
February, 17, 2022

# The presentation at a glance

Simple Probability Samples

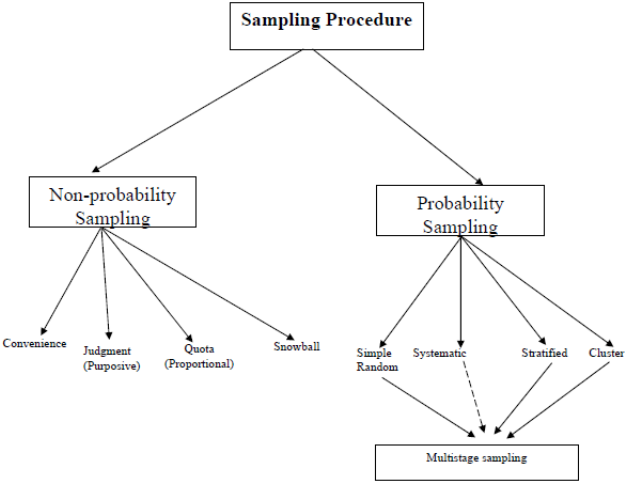
Estimators

Simple Random Sampling

Stratified Sampling

# Simple Probability Samples

# Types of Samples



# Simple Probability Sample

- In a probability sample each unit in the population has a **known probability of selection**, and a random number table or a randomization mechanism is used to choose the specific units to be included in the sample.
- If a probability sampling design is implemented well, an investigator can use a relatively small sample to make inferences about an arbitrarily large population.
- Sample space $S_0$: all possible ordered samples $s$ of size $n(s)$ that can be constructed with the $N$ labels that form the population $U$.

# Basic types of Probability Samples - Simple Random Sampling

- Simple Random Sampling (SRS) is the simplest form of probability sample.
  - An SRS of size $n$ is taken when every possible subset of $n$ units in the population has the same probability of being part of the sample.
  - SRSs are at the basis of more complex designs.
  - In taking a random sample the investigator is mixing every member of the population before selecting $n$ units.
  - The investigator does not need to examine every member of the population.

# Stratified random sample

- In the Stratified random sample the population is divided into subgroups called strata.
- Then a SRS is selected from each stratum.
- The SRSs in the strata are selected independently.
  - The strata are often subgroups of interest to the investigator (to the survey).
  - Elements in the same stratum often tend to be more similar than randomly selected elements from the whole population.
  - As a consequence stratification increases precision.

## Cluster sample

- In a Cluster sample observation units in the population are aggregated into larger sampling units called clusters.
- Cluster sampling is used when natural groups are present in a population.
- The investigator takes a SRS of the clusters (e.g. schools) and then subsample all or some members of the clusters. They then randomly select among these clusters to form a sample.
- *Pay attention*: In stratified sampling, individuals are randomly selected from all strata to make up the sample. On the other hand cluster sampling, the sample is formed when all individuals are taken from randomly selected clusters.

- In stratified sampling, there is homogeneity within the group, while in the case of cluster sampling the homogeneity is found between groups.
- Heterogeneity occurs between groups in stratified sampling. In contrast, group members are heterogeneous in cluster sampling.
- When the sampling method adopted by the researcher is stratified, then the categories are imposed by him. Otherwise, categories are groups that already exist in cluster sampling.
- *Stratified sampling aims to improve accuracy and representation. Unlike cluster sampling, the goal of which is to improve cost effectiveness and operational efficiency*.

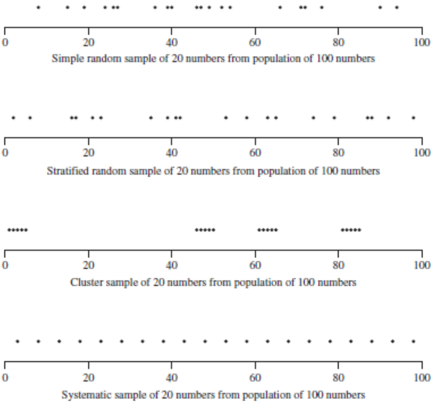
# Systematic sample

- In a Systematic sample a starting point is chosen from alist of population members using a random numeer.
- That unit, and every $k$-th unit thereafter, is chosen in the sample.
- A systematic sample thus consists of units that are equally spaced in the list.

**Figure 1:** Source: S. Lohr (2019). Sampling Design and Analysis. CRC press. (pg. 27)

FIGURE 2.1

Examples of a simple random sample, stratified random sample, cluster sample, and systematic sample of 20 integers from the population {1, 2, . . . , 100}.



Simple random sample of 20 numbers from population of 100 numbers

Stratified random sample of 20 numbers from population of 100 numbers

Cluster sample of 20 numbers from population of 100 numbers

Systematic sample of 20 numbers from population of 100 numbers

# Framework

- We need to be able to list the $N$ units in the finite population.
- The finite population (Universe) of $N$ units is denoted by the index set $U = \{1, 2, \ldots, N\}$
- Out of this population we can choose different samples, which are subsets of $U$
- Suppose $U = \{1, 2, 3, 4\}$ we can select 6 samples from this finite population $S_1 = \{1, 2\}$, $S_2 = \{1, 3\}$, $S_3 = \{1, 4\}$, $S_4 = \{2, 3\}$, $S_5 = \{2, 4\}$, $S_6 = \{3, 4\}$.

- Each possible sample $S$ from the population has a know probability $P(S)$ of been selected (and the posisble probabilities sum to 1).
- In a probability sample, since each possible sample has a known probability of being the chosen sample, each unit in the population has a known probability of appearing in our selected sample. It is called Inclusion probability

$$\pi_i = P(\text{ unit } i \text{ in sample})$$

- The Sampling Weight, for any sampling design, is the reciprocal of the inclusion probability

$$w_i = \frac{1}{\pi_i}$$

The sampling weight of unit $i$ is the number of population units represented by unit $i$.

# Estimators

# Estimators

- An Estimator is a Statistic (a random variable (rv)) whose calculated value is used to estimate a population parameter $\theta$.
- An Estimate A particular realization of an estimator $\hat{\theta}$.
- **Type of Estimators**:
  1. point estimate: single number that can be regarded as the most plausible value of $\theta$.
  2. interval estimate: a range of numbers, called a confidence interval indicating, can be regarded as likely containing the true value of $\theta$
- In the Frequentist world view parameters are fixed, statistics are rv and vary from sample to sample (i.e., have an associated sampling distribution).
- In theory, there are many potential estimators for a population parameter.

# Estimators (ii)

- Good Estimators Are:
  1. Consistent: as the sample size increases $\hat{\theta}$ gets closer to $\theta$.
  2. Unbiased: $E(\hat{\theta}) = \theta$ and we call Bias the quantity:

     $$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

  3. Precise: Sampling distribution of $\hat{\theta}$ should have small variance

     $$V(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$$

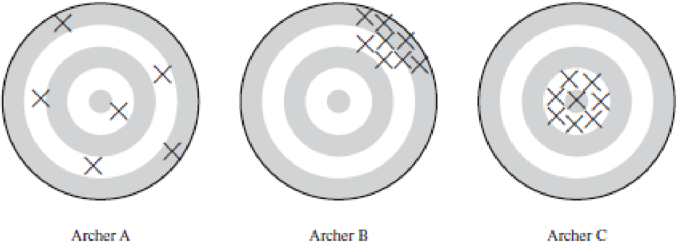
  4. Accurate: Mean Squared Error is small

     $$MSE(\hat{\theta}) = V(\hat{\theta}) + (Bias(\hat{\theta}))^2$$

     (we have to work with MSE because sometimes we work with biased estimators)

# Properties of Estimators

**Figure 2:** A: unbiased; B: precised but not unbiased; C: accurate



# Estimators (iii)

- The finite population $U = \{1, 2, \ldots, N\}$ has as measured values $\{y_1, y_2, \ldots, y_N\}$.
- It is possible to select a sample $S$ of $n$ units from $U$ using the probabilities of selection who defined a sampling design.
- $y_i$ fixed but unknwon unless that unit $i$ appears in the selected sample $S$.
- Without any other statistical assumptions, the only information we have about the set of $y_i^{\prime}s$ in the population in the set $\{y_i : i \in S\}$.

# Population Quantity

- Population Total:

$$t = \sum_{i=1}^{N} y_i$$

- Population Mean:

$$\bar{y}_U = \frac{t}{N}$$

- Population Variance:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2$$

- Population Proportion:

$$p = \frac{\text{Number of units with the characteristic of interest in the population}}{N}$$

# Horvitz-Thompson estimator

- $$\hat{Y}_\pi = \sum_{i \in s} \frac{y_i}{\pi_j}$$

  where $i \in s$ are the unit in the sample

- $$E(\hat{Y}_\pi) = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E[I_i(s)] = \sum_{j=1}^{N} y_j = Y$$

  → unbiased

- The basic design-consistent Horvitz-Thompson estimator is the most natural estimator to use if no auxiliary information available at the estimation stage.

- $$V(\hat{Y}_\pi) = \sum_{i \in s} \left(\frac{1-\pi_i}{\pi_i^2}\right) y_i^2 + \sum_{i \in s} \sum_{i \neq i'} \left(\frac{\pi_{ii'} - \pi_i \pi_{i'}}{\pi_i \pi_{i'}}\right) \frac{1}{\pi_{ij}} y_i y_j$$

# Simple Random Sampling

# Simple Random Sampling

- **Two** types of Simple random sampling:
  1. Simple random sampling with replacement (SRSWR): the same unit may be included more than one in the sample.
  2. Simple random sampling without replacement (SRS): all units in the sample are distinct.
- In SRSWR a sample of size $n$ from a population of $N$ units can be seen as $n$ independent sample of size 1. One unit is randomly selected from the population to be ALWAYS the first sampled unit with probability $\frac{1}{N}$.

# SRS

- A SRS of size $n$ is selected so that **every possible subset** of $n$ distinct units in the population has the same probability of being selected as the sample.
- There are $\binom{N}{n}$ possible samples.
- The probability of selecting any individual sample $s$ of $n$ units is

$$p(S) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!} = \frac{1}{N(N-1)\dots(N-n+1)}$$

- Remember:
  - $k! = k(k-1)(k-2)\dots 1$
  - $0! = 1$
  - $\binom{N}{n} = \frac{N}{n!(N-n)!}$

# Sampling weights is SRS

- In SRS each unit has a first order inclusion probability equal to $\pi_i = \frac{n}{N}$
- The second order inclusion probability is equal to

$$\pi_{i,i'} = \frac{n(n-1)}{N(N-1)}$$

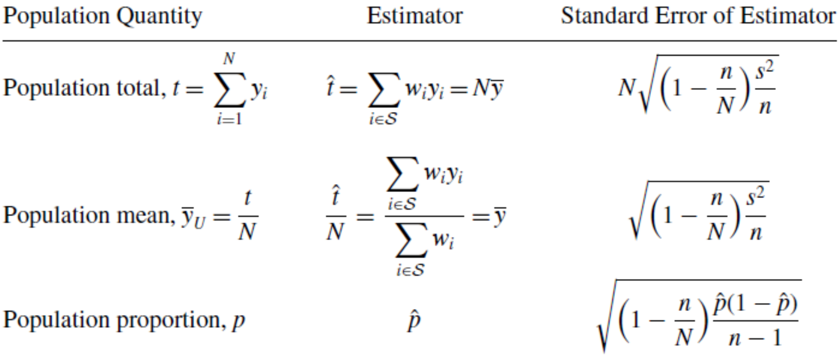
- The design is self-weighting and measurable.
- As a consequence each unit has the same weight (or factor of expansion)

$$w_i = \frac{N}{n}$$

.

- Every unit in the sample represents itself plus $\frac{N}{n} - 1$ unsampled units in the population.

# Estimators SRS



| Population Quantity | Estimator | Standard Error of Estimator |
|---|---|---|
| Population total, $t = \sum_{i=1}^{N} y_i$ | $\hat{t} = \sum_{i \in \mathcal{S}} w_i y_i = N\bar{y}$ | $N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$ |
| Population mean, $\bar{y}_U = \frac{t}{N}$ | $\frac{\hat{t}}{N} = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i} = \bar{y}$ | $\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$ |
| Population proportion, $p$ | $\hat{p}$ | $\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}$ |

# Finite Population Correction

- $\left(1 - \frac{n}{N}\right)$ it is called Finite Population Correction(fpc): this correction is made because, if we have a small population, the greater our sampling fraction $\frac{n}{N}$ is, the more information we have about the population and thus the smaller is the variance.
- For most samples coming from large populations fpc is generally equal to 1 (when $n$ is very small relative to $N$, the finite population correction is almost equal to one).
- Variance is estimated from the sample, but the fpc it is used to assess the error in estimation, which is due to the fact that not all data from the finite population are observed.
- Practically it is used when you sample without replacement from more than 5% of a finite population.

# Finite Population Correction (ii)

- Some formulas used to compute standard errors are based on the idea that (1) samples are selected from an infinite population or (2) samples are selected with replacement.
- This does not present much of a problem when the sample size ($n$) is small relative to the population size ($N$); that is, when the sample is less than 5% of the population.
- If you are using a SRSWR you do not have to use the fpc.

# Confidence Intervals

- When reporting the results of a survey it is necessary to give an idea of how accurate the estimates are.
- Confidence Intervals (CI) are used to indicate the accuracy of an estimate.
- We do not know the values of the statistics from all possible samples so we can not compute the exact confidence interval → Asymptotic results. Our population is supposed to be part of a **superpopulation**.
- For big enough sample the CI is given by

$$\text{estimate } \pm z_{\alpha/2} SE(\text{estimate})$$

- The margin of error of an estimate is the half-width of a confidence interval, i.e. $z_{\alpha/2} SE(\text{estimate})$

# Sample Size Estimation

- **Specify the tolerable error**: the precision needed can be defined as

$$P(|\bar{y} - \bar{y}_U| \leq e) = 1 - \alpha$$

- $e$ is the **margin of error**.
- $e$ and $\alpha$ are fixed by the researcher.
- If the relative precision must be fixed and and it is preferred to check the CV with respect to the absolute error

$$P\Big(|\frac{\bar{y} - \bar{y}_U|}{\bar{y}_U} \leq r\Big) = 1 - \alpha$$

# Sample size estimation (ii)

- SRSWR

$$n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2$$

- SRS (absolute error)

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{z^2_{\alpha/2} S^2}{e^2 + \frac{z^2_{\alpha/2} S^2}{N}}$$

- SRS (relative precision)

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{z^2_{\alpha/2} S^2}{r\bar{y}^2_U + \frac{z^2_{\alpha/2} S^2}{N}}$$

# 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:

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

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

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

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

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

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

## 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:

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

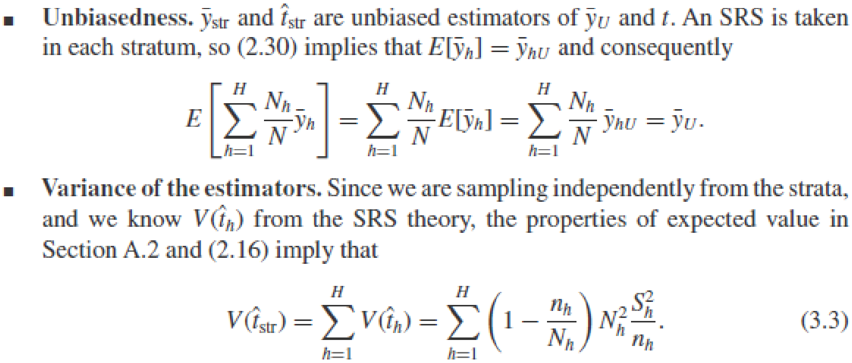
- Population mean estimator:

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

- $\frac{N_h}{N}$: 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)

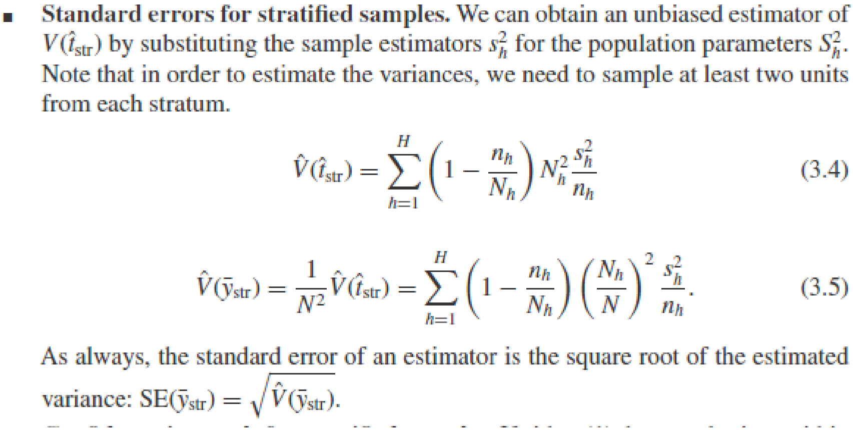


- **Unbiasedness.** $\bar{y}_{\text{str}}$ and $\hat{t}_{\text{str}}$ are unbiased estimators of $\bar{y}_U$ and $t$. An SRS is taken in each stratum, so (2.30) implies that $E[\bar{y}_h] = \bar{y}_{hU}$ and consequently

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

- **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

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

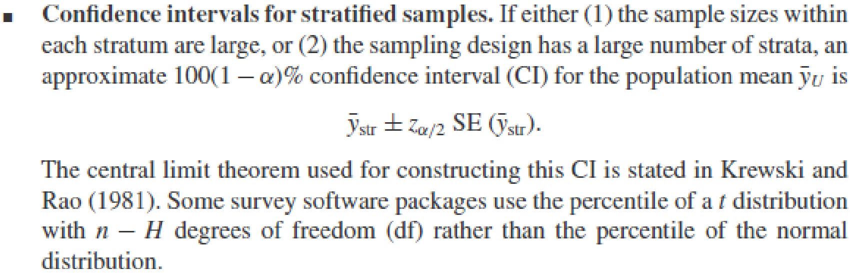


- **Standard errors for stratified samples.** We can obtain an unbiased estimator of $V(\hat{t}_{\text{str}})$ by substituting the sample estimators $s_h^2$ for the population parameters $S_h^2$. Note that in order to estimate the variances, we need to sample at least two units from each stratum.

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

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

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



- **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 → 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.

# Gaia Bertarelli

## Department of Economics and Management in the Era of Data Science

Department
of Excellence
2018 - 2022

EMbeDS
Economics and Management
in the era of Data Science

ISTITUTO
DI MANAGEMENT
Sant'Anna
Scuola Universitaria Superiore Pisa

management e sanità

gaia.bertarelli@santannapisa.it