# Lecture 4: Sampling

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

November 27, 2023

# Plan

Plan

1) Sampling

2) Imbalanced sample

3) Matching

# Sampling - design and procedures



**Sampling** is a key component of any research design. **Sampling design** involves several basic questions.

– Should a sample be taken?

– If so, what process should be followed?

– What kind of sample should be taken?

– How large should it be?

– What can be done to control and adjust for non-response errors?

# The sampling design process



1) Define the target population

2) Determine the sampling frame

3) Select a sampling technique

4) Determine the sample size

5) Execute the sampling process

6) Validate the sample

# The sampling design process



We have already considered sample size determination from a **statistical perspective**.

However, for now we consider important qualitative factors in determining the sample size. These are:

– The importance of the decision,

– the nature of the research,

– the number of variables,

– the nature of the analysis,

– sample sizes used in similar studies,

– incidence rates,

– completion rates,

– resource constraints.

# The sampling design process

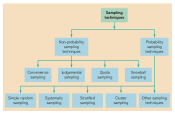


| Type of study | Minimum size | Typical range |
|---|---|---|
| Problem identification research | 500 | 1,000-2,500 (e.g. market potential) |
| Problem-solving research | 200 | 300-500 (e.g. pricing) |
| Product tests | 200 | 300-500 |
| Test marketing studies | 200 | 300-500 |
| TV, radio, print or online advertising | 150 | 200-300 (per advertisement tested) |
| Test-market audits | 10 stores | 10-20 stores |
| Focus groups | 6 groups | 6-12 groups |

Figure:

# Classification of sampling techniques





# Classification of sampling techniques



Features of **non-probability sampling** include the following.

– Some units in the population have a zero probability of selection.

– Individual units in populations have an unknown probability of being selected.

– Inability to measure sampling error.

Features of **probability sampling** include the following.

– Every population element has a known, non-zero probability of being selected in the sample.

– Probability sampling makes it possible to estimate the margins of sampling error.

# Non-probability sampling techniques

1) **Convenience sampling**

attempts to obtain a sample of convenient elements. Often, participants are selected because they happen to be in the right place at the right time.

Examples:

– the use of students and members of social organisations

– street interviews without qualifying the participants

– 'people-on-the-street' interviews.

2) **Judgemental sampling**

is a form of convenience sampling in which the population elements are selected based on the judgement of the researcher.

Examples:

– test markets

– purchase engineers selected in business-to-business (B2B) market research

– expert witnesses used in court.

# Non-probability sampling techniques

3) **Quota sampling**

may be viewed as two-stage restricted judgemental sampling.

The first stage consists of developing control categories, or quota controls, of population elements.

In the second stage, sample elements are selected based on convenience or judgement.

4) In **snowball sampling**,

an initial group of participants is selected, usually at random.

After being interviewed, these participants are asked to identify others who belong to the target population of interest.

Subsequent participants are selected based on the *referrals*.

# Probability sampling techniques



1) In **simple random sampling (SRS)**,

each element in the population has a known and equal probability of selection.

Each possible sample of a given size, $n$, has a known and equal probability of being the sample actually selected. This implies that every element is selected independently of every other element.

2) In **systematic sampling**,

the sample is chosen by selecting a random starting point and then picking every $i$-th element in succession from the sampling frame.

The sampling interval, $i$, is determined by dividing the population size, $N$, by the sample size, $n$, and rounding to the nearest integer.

When the ordering of the elements is related to the characteristic of interest, systematic sampling increases the representativeness of the sample.

If the ordering of the elements produces a cyclical pattern, systematic sampling may decrease the representativeness of the sample.

# Probability sampling techniques

3) **Stratified sampling**

is a two-step process in which the population is partitioned into subpopulations, or strata.

The strata should be mutually exclusive and collectively exhaustive in that every population element should be assigned to one, and only one, stratum and no population elements should be omitted.

Next, elements are selected from each stratum by a random procedure, usually SRS.

A major objective of stratified sampling is to increase precision without increasing cost.

The elements within a stratum should be as *homogeneous* as possible, but the elements in different strata should be as heterogeneous as possible.

The stratification factors should also be closely related to the characteristic of interest.

Finally, the factors (variables) should decrease the cost of the stratification process by being easy to measure and apply.

In *proportionate stratified sampling*, the size of the sample drawn from each stratum is proportionate to the relative size of that stratum in the total population.

In *disproportionate (optimal) stratified sampling*, the size of the sample from each stratum is proportionate to the relative size of that stratum and to the standard deviation of the distribution of the characteristic of interest among all the elements in that stratum.

# Probability sampling techniques

4) In **cluster sampling**,

the target population is first divided into mutually exclusive and collectively exhaustive subpopulations, or clusters.

Next, a random sample of clusters is selected, based on a probability sampling technique such as SRS. For each selected cluster, either all the elements are included in the sample (one-stage) or a sample of elements is drawn probabilistically (two-stage).

Elements within a cluster should be as *heterogeneous* as possible, but clusters themselves should be as homogeneousas possible.

Ideally, each cluster should be a *small-scale representation of the population*.

In probability proportionate to size sampling, the clusters are sampled with probability proportional to size.

In the second stage, the probability of selecting a sampling unit in a selected cluster varies inversely with the size of the cluster.

# Stratified sampling vs Cluster sampling



| **Factor** | **Stratified sampling** | **Cluster sampling (one-stage)** |
|---|---|---|
| Objective | Increase precision | Decrease cost |
| Subpopulations | All strata are included | A sample of clusters is chosen |
| Within subpopulations | East stratum should be homogeneous | Each cluster should be heterogeneous |
| Across subpopulations | Strata should be heterogeneous | Clusters should be homogeneous |
| Sampling frame | Needed for the entire population | Needed only for the selected clusters |
| Selection of elements | Elements selected from each stratum randomly | All elements from each selected cluster are included |

# Graphical illustation



**A graphical illustration of non-probability techniques**

**1 Convenience sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Group D happens to assemble at a convenient time and place. So all the elements in this group are selected. The resulting sample consists of elements 16, 17, 18, 19 and 20. Note that no elements are selected from groups A, B, C or E

**2 Judgemental sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

The researcher considers groups B, C and E to be typical and convenient. Within each of these groups one or two elements are selected based on typicality and convenience. The resulting sample consists of elements 8, 10, 11, 13 and 24. Note that no elements are selected from groups A and D

**3 Quota sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

A quota of one element from each group, A to E, is imposed. Within each group, one element is selected based on judgement or convenience. The resulting sample consists of elements 3, 6, 13, 20 and 22. Note that one element is selected from each column or group

**4 Snowball sampling**

Random Selection — Referrals

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Elements 2 and 9 are selected randomly from groups A and B. Element 2 refers elements 12 and 13. Element 9 refers element 18. The resulting sample consists of elements 2, 9, 12, 13 and 18. Note that no element is selected from group E

**A graphical illustration of probability sampling techniques**

**1 Simple random sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Select five random numbers from 1 to 25. The resulting sample consists of population elements 3, 7, 9, 16 and 24. Note that there is no element from group C

**2 Systematic sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Select a random number between 1 and 5, say 2. The resulting sample consists of a population 2, (2 + 5) = 7, (2 + 5 x 2) = 12, (2 + 5 x 3) = 17 and (2 + 5 x 4) = 22. Note that all the elements are selected from a single row

**3 Stratified sampling**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Randomly select a number from 1 to 5 from each stratum, A to E. The resulting sample consists of population elements 4, 7, 13, 19 and 21. Note that one element is selected from each column

**4 Cluster sampling (two-stage)**

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 6 | 11 | 16 | 21 |
| 2 | 7 | 12 | 17 | 22 |
| 3 | 8 | 13 | 18 | 23 |
| 4 | 9 | 14 | 19 | 24 |
| 5 | 10 | 15 | 20 | 25 |

Randomly select three clusters, B, D and E. Within each cluster, randomly select one or two elements. The resulting sample consists of population elements 7, 18, 20, 21 and 23. Note that no elements are selected from clusters A and C

# Pros and Cons



| *Technique* | *Strengths* | *Weaknesses* |
|---|---|---|
| ***Non-probability sampling*** | | |
| **Convenience sampling** | Least expensive, least time-consuming, most convenient | Selection bias, sample not representative, not recommended for descriptive or causal research |
| **Judgemental sampling** | Low cost, convenient, not time-consuming ideal for exploratory research designs | Does not allow generalisation, subjective |
| **Quota sampling** | Sample can be controlled for certain characteristics | Selection bias, no assurance of representativeness |
| **Snowball sampling** | Can estimate rare characteristics | Time-consuming |
| ***Probability sampling*** | | |
| **Simple random sampling (SRS)** | Easily understood, results projectable | Difficult to construct sampling frame, expensive, lower precision, no assurance of representativeness |
| **Systematic sampling** | Can increase representativeness, easier to implement than SRS, sampling frame not always necessary | Can decrease representativeness depending upon 'order' in the sampling frame |
| **Stratified sampling** | Includes all important subpopulations, precision | Difficult to select relevant stratification variables, not feasible to stratify on many variables, expensive |
| **Cluster sampling** | Easy to implement, cost effective | Imprecise, difficult to compute and interpret results |

# Imbalanced samples



Problem:

– Data imbalance is predominant and inherent in the real world.

– Data often demonstrates skewed distributions with a long tail.

– However, most of algorithms are designed around the assumption of a uniform distribution over each target category (classification).

One way the imbalance may affect the algorithm is when algorithm completely ignores the **minority class**.

The reason this is an issue is because the minority class is often the class that we are most interested in.

E.g.

1) a classifier to classify fraudulent and non-fraudulent transactions from various observations

2) probability of bankruptcies withing the industry

# Imbalanced samples

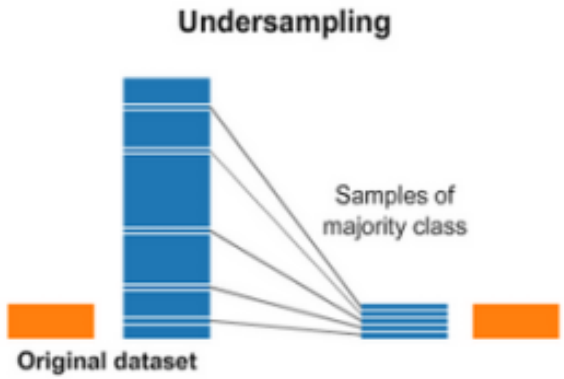


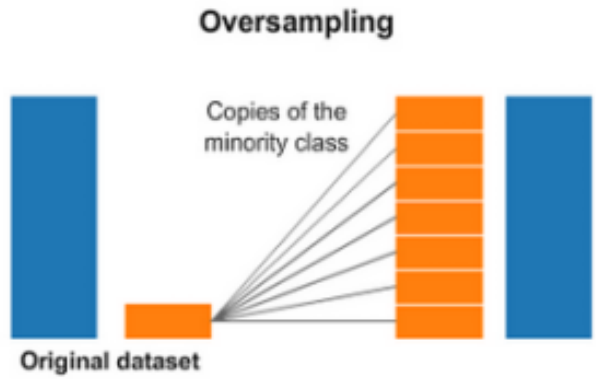
There are two main approaches to random resampling for imbalanced classification; they are oversampling and undersampling.

– Random Oversampling: Randomly duplicate examples in the minority class.

– Random Undersampling: Randomly delete examples in the majority class.

Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. Random undersampling involves randomly selecting examples from the majority class and deleting them from the training dataset.



# Undersampling



1) Random sampler

It is the easiest and fastest way to balance the data by randomly selecting a few samples from the majority class.

2) NearMiss

Adds some common sense rules to the selected samples by implementing different heuristics.

(*) NearMiss-2: Majority class examples with a minimum average distance to three furthest minority class examples.

3) Tomek Links.

Tomek links are pairs of examples of opposite classes in close vicinity.

In this algorithm, we end up removing the majority element from the Tomek link which provides a better decision boundary for a classifier.

# Undersampling

3) Tomek Links.

Tomek links are pairs of examples of opposite classes in close vicinity.

In this algorithm, we end up removing the majority element from the Tomek link which provides a better decision boundary for a classifier.

# Undersampling



The Concept:

Tomek Links is one of a modification from Condensed Nearest Neighbors (CNN) undersampling technique that is developed by Tomek (1976).

Unlike the CNN method that are only randomly select the samples with its $k$ nearest neighbors from the majority class that wants to be removed, the Tomek Links method uses the rule to selects the pair of observation (say, a and b) that are fulfilled these properties:

1. The observation **a**'s nearest neighbor is **b**.

2. The observation **b**'s nearest neighbor is **a**.

3. Observation **a** and **b** belong to **a** different class. That is, **a** and **b** belong to the minority and majority class (or vice versa), respectively.

# Undersampling

Mathematically, it can be expressed as follows.

Let $d\left(x_i, x_j\right)$ denotes the Euclidean distance between $x_i$ and $x_j$, where $x_i$ denotes sample that belongs to the minority class and $x_j$ denotes sample that belongs to the majority class.

If there is no sample $x_k$ satisfies the following condition:

1. $d\left(x_i, x_k\right) < d\left(x_i, x_j\right)$, or

2. $d\left(x_j, x_k\right) < d\left(x_i, x_j\right)$

then the pair of $\left(x_i, x_j\right)$ is a Tomek Link.

# Undersampling



Advantages

– Reduce the risk of their analysis or machine learning algorithm skewing toward the majority.

Without resampling, scientists might come up with *the accuracy paradox* where they run a classification model with 90% accuracy. On closer inspection, though, they will find the results are heavily within the majority class.

– Fewer storage requirements and better run times for analyses. Less data means you or your business needs less storage and time to gain valuable insights.

Disadvantages

– Removing enough majority examples to make the majority class the same or similar size to the minority class results in a significant loss of data.

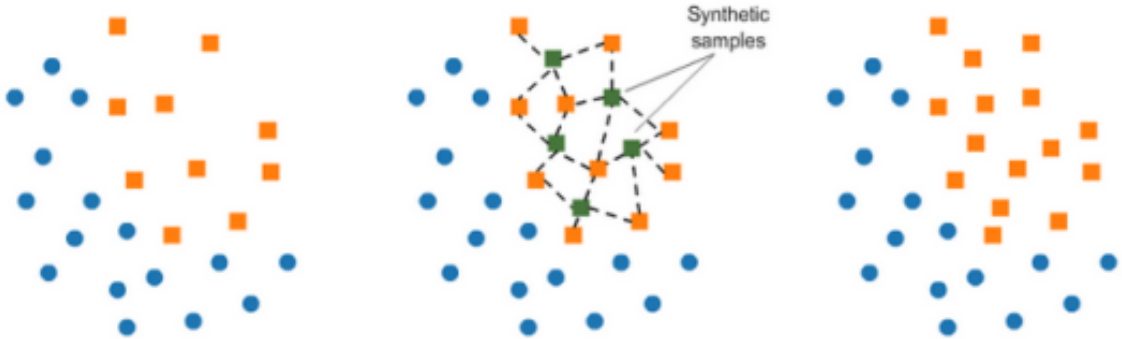
– The sample of the majority class chosen could be biased, meaning, it might not accurately represent the real world, and the result of the analysis may be inaccurate. Therefore, it can cause the classifier to perform poorly on real unseen data.

# Oversampling



1) SMOTE

In SMOTE (Synthetic Minority Oversampling Technique) we synthesize elements for the minority class, in the vicinity of already existing elements.




2) Generative Adversarial Neural Networks

# Oversampling



SMOTE is one of the most popular oversampling techniques that is developed by Chawla et al. (2002). Unlike random oversampling that only duplicates some random examples from the minority class, SMOTE generates examples based on the distance of each data (usually using Euclidean distance) and the minority class nearest neighbors, so the generated examples are different from the original minority class.

In short, the process to generate the synthetic samples are as follows.

1. Choose random data from the minority class.

2. Calculate the Euclidean distance between the random data and its $k$ nearest neighbors.

3. Multiply the difference with a random number between 0 and 1 , then add the result to the minority class as a synthetic sample.

4. Repeat the procedure until the desired proportion of minority class is met.

This method is effective because the synthetic data that are generated are relatively close with the feature space on the minority class, thus adding new "information" on the data, unlike the original oversampling method.

# Oversampling



Advantages

– It improves the overfitting caused by random oversampling as synthetic examples are generated rather than a copy of existing examples.

–No loss of information.

– It's simple.

Disadvantages

– While generating synthetic examples, SMOTE does not take into consideration neighboring examples that can be from other classes. This can increase the overlapping of classes and can introduce additional noise.

– SMOTE is not very practical for high-dimensional data.

# Matching



Causal inference "balancing technique"

Often used with "Diff-in-diff" to minimize sampling bias.

**The idea** is to find, from a large group of nonparticipants, individuals who are observationally similar to participants in terms of characteristics not affected by the program.

- When a treatment cannot be randomized, the next best thing to do is to try to mimic randomization—that is, try to have an observational analogue of a randomized experiment.

- Matching is not the only way to eliminate bias (e.g. regressions with control variables and/or instruments).

- Matching is a non-parametric method. You do not need to assume any functional form for the causal relationship being investigated.

# Matching



1) **Simple matching** compares groups of observations corresponding to the same discrete values $\boldsymbol{x}$.

$$\Delta^M = \sum_k w_k * \left(\bar{y}_{1,k} - \bar{y}_{0,k}\right)$$

$\bar{y}_{1,k}$ — average result for the treatment group

$\bar{y}_{0,k}$ — average result for control group

$w_k$ is the proportion of observations belonging to the $k$-th group among the entire sample

Alternative form (Dehejia and Wahba, 2002)

$$\Delta^M = \frac{1}{N_T} \sum_{i=1}^{N_T} \left( y_i - \frac{1}{N_{C,i}} \sum_{j \in \{D=0\}} y_j \right)$$

$N_T$ - number of observations in the treatment group

$N_{C,i}$ — number of observations in that part of the control group of observations that is matched with the $i$-th object from the group exposed to the influence

# Matching

2) Nearest-neighbor matching

The $i$-th observation from the experimental group is compared with the set of closest observations from the control group.

Euclidean distance is used as a measure of proximity.

$$A_i = \left\{ j \mid \min_j \left\| x_i - x_j \right\| \right\}$$

$A_i -$ set of objects from the control group that are compared with the $i$-th observation and the experimental group

# Matching



3) Propensity score matching

If the vector of explanatory variables has a large dimension, or if there are continuous variables among the variables, then an exact comparison is not entirely convenient.

In this case, a propensity score is used - the conditional probability that an object will be affected given the given values of the regressors.

The propensity score is usually estimated using a logit or probit model.

$$P\left(D_i = 1 \mid x_i^{(1)}, x_i^{(2)} \ldots, x_i^{(k)}\right) = \\ = p\left(x_i^{(1)}, x_i^{(2)} \ldots, x_i^{(k)}\right)$$

Thus, matching is carried out in two stages:

1. For each observation, the value of the propensity measure is estimated (for example, based on a logit model)

2. Then a comparison of objects with similar values of the propensity measure is carried out using

– Nearest neighbor matching

– Comparison with stratification

– Radial matching

# Matching

Example of policy evaluation employing PSM

– comparing UK Tax Returns of Foreign Multinationals to Matched Domestic Firms (Bilicka 2019)

– the role of direct government interactions with start-ups in comparison of alliances with governments, firms, research organizations, and not-for-profit organizations (Doblinger et al. 2019).

– the short-term and long-term effects of business incubators on the performance of innovative start-ups in terms of sales revenues and job creation. Here we could compare incubated and not incubated start-ups (Lukeš et al. 2019).

– the role of public R&D subsidy facilitate firms' exploratory innovation (Gao et al. 2021)

# Matching

Assumptions for PSM

– PSM is a useful approach when only **observed characteristics** are believed to affect program participation.

Whether this belief is actually the case **depends on the unique features** of the program itself, in terms of targeting as well as individual take-up of the program.

– Assuming **selection on observed characteristics** is **sufficiently strong** to determine program participation, baseline data on a wide range of pre-program characteristics will allow the probability of participation based on observed characteristics to be specified more **precisely**.

– Some tests can be conducted to **assess the degree of selection bias** or participation on unobserved characteristics.

# Matching



The validity of PSM depends on two conditions:

*A1: conditional independence (namely, that unobserved factors do not affect participation)*

*A2: sizable common support or overlap in propensity scores across the participant and nonparticipant samples.*

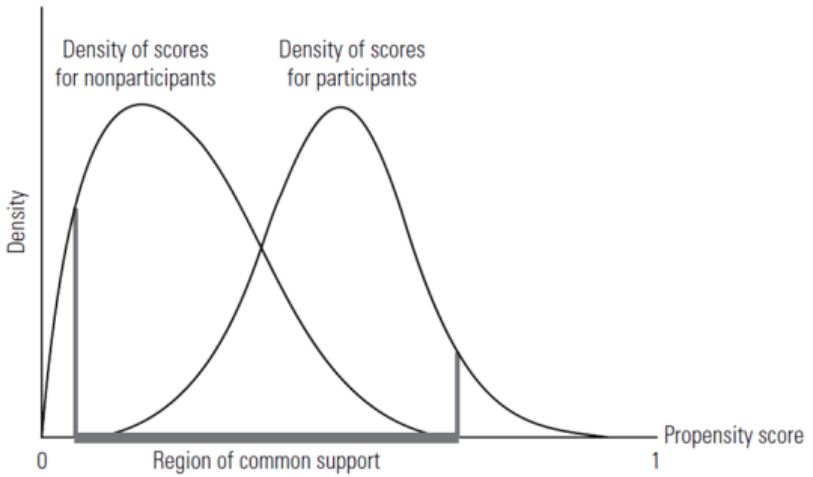
If one assumes that **differences in participation** are based solely on **differences in observed characteristics**, and if enough nonparticipants are available to match with participants, the corresponding treatment effect can be measured even if treatment is not random.

*A3: Independent observations ensure that the outcome and treatment for one individual has no effect on the outcome or treatment for any other individual.*

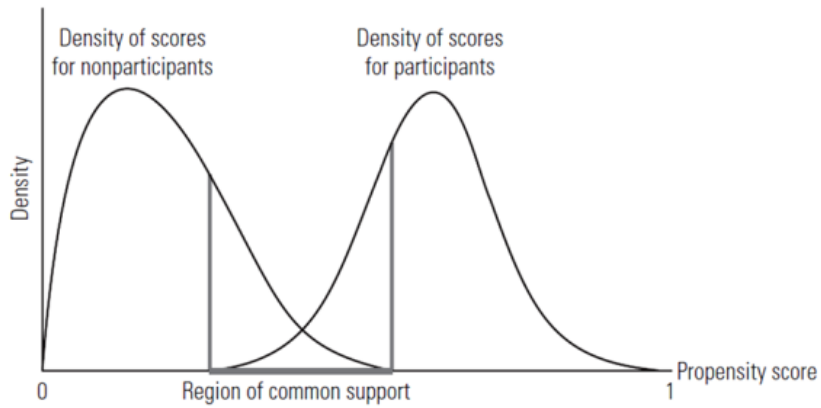
# Matching



**Example of Common Support**




**Example of Poor Balancing and Weak Common Support**



# Matching

**Counterfactual**

With matching methods, one tries to develop a **counterfactual or control group** that is as similar to the treatment group as possible in terms of observed characteristics.

Each participant is matched with an observationally similar nonparticipant, and then the average difference in outcomes across the two groups is compared to get the program treatment effect

The problem is to **credibly identify groups** that look alike.

# Matching



Example. **The role of incubators for start-ups**

*Do business incubators really enhance entrepreneurial growth? Evidence from a large sample of innovative Italian start-ups (Lukeš et al. 2019)*

Innovative start-ups are often considered to be a **key source of innovation and job creation**. As such they are the subject of several types of supportive public policies.

This study examines the short-term and long-term effects of business incubators on the performance of innovative start-ups in terms of **sales revenues and job creation**.

A large sample of N = 2544 innovative Italian start-ups, of which 606 were incubated, was followed over a period of up to 6 years.

Treatment: the incubated indicator (a dummy variable).

Outcome indicators: sales revenue (log of revenue) and job creation.

Tobit and Poisson regressions and PSM analyses are applied.

Data source: the Companies Register in the Italian Chamber of Commerce and the AIDA database (2016) provided by Bureau van Dijk, which include annual data on registered innovative start-ups created between 2009 and 2014.

# Matching



Looking for a counterfactual

To ensure comparability firms must meet the following requirements:

(1) to have been operational for less than 5 years,

(2) to have a yearly turnover of less than 5 million Euros,

(3) to keep their profits,

(4) not to be the result of a merger, split, or sell-off of a company or branch, and

(5) to be of innovative character, i.e., having high R&D expenses, a highly educated workforce, perhaps with patents or industrial properties.

6 categories of industries: ICT (NACE section J, codes 5811–6399), professional, scientific and technical activities (M, 6910–7500), manufacturing (C, 1010–3320), wholesale and retail (G, 4510– 4799), administrative and support (N, 7710–8299), other.

# Matching

## Balancing tests

Descriptive statistics for quantitative variables. Raw data

| | Incubated | | | | | Not incubated | | | | | Diff |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **SD** | **Min** | **Max** | **N** | **Mean** | **SD** | **Min** | **Max** | ***p*** |
| **Revenue (000s EUR)** | 1392 | **118.85** | 299.44 | 0 | 4207 | 3745 | **176.42** | 430.29 | 0 | 6253 | 0.000 |
| **No. of employees** | 1279 | **2.02** | 3.85 | 0 | 50 | 3519 | **2.17** | 3.92 | 0 | 58 | 0.249 |
| **Start-up age** | 1646 | **2.17** | 1.20 | 1 | 6 | 4573 | **1.99** | 1.10 | 1 | 6 | 0.000 |
| **Initial capital (000s EUR)** | 1280 | **35.42** | 86.06 | 0 | 1900 | 3446 | **55.37** | 433.29 | 0 | 14,003 | 0.010 |
| **No. of peers in incubator** | 1646 | **12.83** | 15.21 | 0 | 58 | 4573 | **0.00** | 0.00 | 0 | 0 | 0.000 |
| **Regional GDP (bln. EUR)** | 1646 | **59.89** | 59.74 | 0.5 | 162 | 4572 | **53.26** | 58.65 | 0.5 | 162 | 0.000 |
| **City size** | 1646 | **1611.89** | 1339.78 | 127 | 4074 | 4573 | **1408.28** | 1213.68 | 87 | 4074 | 0.000 |

# Matching



Results of propensity-score matching regressions

| | Start-up age | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| ***Log of revenue*** | | | | |
| **Incubation effect** | − 0.3026* | − 0.2039 | 0.1083 | 0.0591 |
| | (0.1217) | (0.1583) | (0.1974) | (0.2695) |
| **N of Incubated start-ups** | 518 | 397 | 179 | 78 |
| **N of non-incubated start-ups** | 1538 | 1027 | 404 | 153 |
| **Observations** | 2056 | 1424 | 583 | 231 |
| ***Number of employees*** | | | | |
| **Incubation effect** | − 0.2285 | − 0.5151 | − 0.2269 | 0.3745 |
| | (0.1863) | (0.2810) | (0.4067) | (0.6402) |
| **N of Incubated start-ups** | 518 | 421 | 181 | 86 |
| **N of non-incubated start-ups** | 1575 | 1116 | 434 | 173 |
| **Observations** | 2093 | 1537 | 615 | 259 |

Note: * $p < 0.05$, ** $p < 0.01$.

# Matching



Findings

– Incubator tenancy had a negative short-term effect on start-ups' sales revenue.

– Incubation seems to speed up sales revenue growth in the long run.

– Incubator tenancy had a negligible effect on start-ups' job creation.

– Evidence justifying public spending on business incubators was not found.

Possible heterogeneity of the effect

– The effects of incubator characteristics, in terms of ownership, certification, and size on the growth of tenant start-ups were further analysed, but these effects were found to be negligible.

# Summary



1) Non-probability vs Probability sampling

| *Factors* | *Conditions favouring the use of:* | |
|---|---|---|
| | *Non-probability sampling* | *Probability sampling* |
| Nature of research | Exploratory | Conclusive |
| Relative magnitude of sampling and non-sampling errors | Non-sampling errors are larger | Sampling errors are larger |
| Variability in the population | Homogeneous (low) | Heterogeneous (high) |
| Statistical considerations | Unfavourable | Favourable |
| Operational considerations | Favourable | Unfavourable |

2) Oversampling vs undersampling to improve accuracy

3) Matching to remove sampling bias