Experimental data and survey data

An **experiment** involves the collection of measurements or observations about populations that are treated or controlled by the experimenter.

A **survey** is an examination of a system in operation in which the investigator does not have an opportunity to assign different conditions to the objects of the study.

We might use a survey to compare two countries with different types of economic systems. If there is a significant difference in some economic measure, such as per-capita income, it does not mean that the economic system of one country is superior to the other.

The survey takes conditions as they are and cannot control other **variables** that may affect the economic measure, such as comparative richness of natural resources, population health, or level of literacy.

All that can be concluded is that at this particular time a significant difference exists in the economic measure.

Unfortunately, surveys of this type are frequently misinterpreted.
Surveys in general do not give answers that are as clear-cut as those of experiments. If an experiment is possible, it is preferred. For example, in order to determine which of two methods of teaching reading is more effective, we might conduct a survey of two schools that are each using a different one of the methods. But the results would be more reliable if we could conduct an experiment and set up two balanced groups within one school, teaching each group by a different method.

Quantitative research in many areas of social, biological, and behavioural science would be impossible without surveys. However, in surveys we must be alert to the possibility that our measurements may be affected by variables that are not of primary concern.

Different data selection procedures
Three desirable criteria for a study

**Realism**

variables are chosen basing on feasibility (how to do treatments, how to evaluate results, short, medium or long period, which variables, tests, opinion? Etc.)

**Randomness**

is it possible to completely randomize data selection? Or to study groups? Is it possible that results are objective? Etc.

**Representativeness**

what is the population? how to choose sample statistical units? national or local level? use of benchmark, etc.
Three big classes of statistical studies

**Realism** ⇒ observational studies

**Randomness** ⇒ experimental studies

**Representativeness** ⇒ surveys

A trade-off between ideal situation is always needed; most often the best design results by combining different approaches
The trade-off

**Experimental study**

ideal experiment, all controlled

loss of representativeness and realism

**Observational study**

collection of data without randomness and representativeness

**Survey**

probability samples, where each element in the population has a known positive probability to be included in the sample

Loss of randomness and realism, introduction of control is needed in order to avoid bias
The role of variables

1. **Esplicative variables** (E)
   
   The **aim of the research**, what we want to measure
   
   - dependent /predictant (Y)
   - independent /predictor (X)

   They derive from a priori knowledge about the context (theoretical and practical hypotheses and implications)

2. **Control variables** (C)

   That can be controlled by the research design, aiming at reduce random error (R) or at reduce bias coming from noise variables (D)
3. **Noise variables** (D)

cannot be controlled and may be confounded with the explicative (E); the researcher aims at removing them and at including them to the group (C) or (R); ask for the use of non experimental designs;

4. **Randomized variables** (R)

Not controlled, treated as “random error”.

In experimental design variables are randomized by the researcher, in surveys variables are assumed to be random
Sampling

- Sampling procedures depend on:
  - Phenomenon to be studied
  - External factors

- Census vs sample analysis

**WHY SAMPLE:**

- Quicker and cheaper than alternative (census)
- As accurate (can be) as census
- Destructive sampling
- Infinite population
- Finite population but impossible to observe each unit
Basic vocabulary

Sample
Representative subgroup of the population

The general aim is to make inferences about the population for purposes of:
Hypothesis testing
Estimation
Description

Population or Universe
Set of all statistical units to be studied, or set of all statistical information describing phenomenon

Random variable $X$, with density $f(x; \theta)$

Selection of $n$ statistical units from $X$: sequence of r.v. $X_1, \ldots, X_n$

Observed sample: realization $x_1, \ldots, x_n$

Random sample
Collection of iid r.v. $X_1, \ldots, X_n$ from $X$, each with density $f(x; \theta)$
Inference is based on repeated sampling

\( x_1, ..., x_n \) is one of all possible samples that can be selected from \( X \)

Inference is made by a function \( T(X_1, ..., X_n) \) called statistic, that is a synthesis of sample or an estimator \( T_n(X_1, ..., X_n) \).

\( T \) is a r.v. and its density is called sample distribution
Independence of $X_i$ \[ \Pr(x_i) = \Pr(x_i \mid x_{i-1}) \]

**Continuous r.v.**

Independence is always respected

\[ \Pr(x_i) = \Pr(x_i \mid x_{i-1}) = 0 \]

**Discrete r.v.**

Independence is assured only for Bernoulli trial (extraction with replacement)

- $N = \infty$: independence also in sampling without replacement
- $n$ small with respect to $N$: quasi-independence in sampling w.r.
- Finite $N$: “simple” random sample
Sample theory

**Notation**
- Population P of N statistical units $U_1, ..., U_N$
- Sample C of n s.u. $x_1, ..., x_n$

Density that generates $x_i$ is no more relevant
$x_i$ value is relevant

- Sample space $\Omega_n(P)$: number of all possible sample of size n that can be extracted from a population of size N
- Sample rate $n/N$: proportion of s.u. included in C
- Sampling process: procedure for selection of s.u.
Sampling and non sampling error

The accuracy of sample results (estimation) is generally affected by two types of error: sampling error and non sampling (measurement) error.

Estimation of population mean:

\[
\mu = \bar{x} \pm \varepsilon_s \pm \varepsilon_{ns}
\]

**Sampling error**

results when the sample selected is not perfectly representative of the population.

**Non sampling (measurement) error**

Includes all factors other than sampling error that may cause inaccuracy or bias in survey results.
Potential sampling errors

- Random Sampling Error: due to chance, cannot be avoided

- Systematic Error

- Respondent Error
  - Non Response Error
  - Response Error
  - Deliberate
  - Honest
    (Likely causes: enumerator, social desirability)

- Administrative Error: flows in the design or execution of the sample that cause it to not be representative
  - Sample Selection
  - Administrative Error
  - Enumerator Cheating
  - Data Processing Error
Probability or non probability sampling methods

**Probability sampling**

Every element of the population has a known and equal likelihood of being included in the sample (probability of selection)

The inference process is possible

It allows to control sampling error (by controlling the variance of the estimator)

**Non-probability sampling**

Any sample that does not meet the requirements of a probability sample

The inference process is not possible
The aim of using sampling designs

To reduce variance of the estimator (= to improve precision of the estimation)
Non probability sampling designs

**Convenience sample**

Used for reasons of convenience.

*E.g.*: companies often use their own employees for preliminary tests of new products

**Judgemental sample**

The selection criteria are based on researcher’s personal judgement about what constitutes a representative sample

*E.g.*: experts

**Quota sample**

Quotas, based on demographic or classification factors selected by the researcher, are established for population subgroups

**Snowball sample (rare populations)**

Additional respondents are selected based on referrals from initial ones.

*E.g.*: irregular immigration
Probability sampling designs

**Simple random sampling (SRS) without replacement**

Is the purest form of probability sampling.

The known and equal probability of selection for a s.u. is computed as follows:

\[ \Pr(U_i \in C) = \frac{n}{N} \]

Probability of selection = Sample rate

Sample space: \[ \binom{N}{n} = \frac{N!}{n!(N-n)!} \]
possible samples

- The selection follows the Cochran procedure
- When \( N \) is large and \( n \) is small probability of inclusion tends to SRS with replacement
**Simple random sampling (SRS) with replacement**

**Sample space:** \( N^n \) possible samples

**Probability of selection:** 
\[
\Pr(U_i \in C) = 1 - \left(1 - \frac{1}{N}\right)^n
\]

- **Procedure of selection**
  
  If a sampling frame is available:
  1. Assign a number to each \( U_i \) in \( P \)
  2. Select \( n \) numbers in \([1; N]\) from a random numbers table
**Systematic sampling**

Selection without replacement from lists built according to some criterion (alphabetical, temporal, etc) independent from $X$

- **Procedure**
  1. Determine the **skip interval** $k = \frac{N}{n}$
  2. Extract a random number $d \in \{1, 2, \ldots, k\}$
  3. Find units $U_i$ such that $i = d + k(j - 1)$ for $j = 1, 2, \ldots, n$

- **Advantages over SRS:**
  - simpler, less time consuming, less expensive

- **Disadvantage over SRS:**
  - hidden patterns within population list may be pulled into the sample
Stratified sampling

- Procedure of selection
  1. The population is divided into \( S \) mutually exclusive and exhaustive subsets (e.g. male and female), called **strata**

  2. The number of elements selected from each stratum is fixed

    - **Proportional allocation**: \( n_s (s = 1, \ldots, S) \) is directly proportional to the size of the stratum \( s \) in \( P \)
    
    - **Disproportional (or optimal) allocation**: \( n_s \) is proportional to relative size of the stratum \( s \) in \( P \), i.e. it is determined basing on size and is directly proportional to the *within variance* of \( s \) (the lower the variance, the fewer s.u. are needed)
      
      Borderline case: If \( \text{Var}(s) = 0 \) only one observation is necessary

  3. **SRS within each stratum**

- The stratified sampling is more efficient (i.e. the variance of the estimator is lower) than SRS when there exists a salient classification factor correlated with the behaviour of interest
Example

Problem

We want to foresee the audience for a new left-wing talk show
We are convinced that political opinion discriminates the preference

Possible solution

We could proceed at regional level

1. Each region is a stratum
2. SRS within each region

- Sample size in each region could be inversely proportional to the percentage of votes polled by the left-wing parties on last elections (proportional allocation)
Cluster sampling

- Procedure of selection
  1. The population is divided into $G$ mutually exclusive and exhaustive subsets called clusters
  2. A SRS of the clusters is selected

- It is efficient when subgroups have very high within variance and very low between variance (have similar behaviour)

  Selected cluster can represent $P$

Two-stages cluster sampling

1. Cluster sample (to select clusters)
2. SRS within clusters selected
**Example**

**Problem**

The direction of the school Y wants to measure the satisfaction of last year students about the laboratories service

**Possible solution**

They could proceed at class level

1. Each class is a cluster
2. SRS of classes

Hypothesis: each class can be considered as a “replication” of the others (very low between variance) and then of the whole population
Effectiveness and complexity of sample designs

Convenience of a sampling method is measured by the Design effect index:

$$D_{Eff} (T_n | \pi) = \frac{\text{MSE}(T_n | \text{SRS})}{\text{MSE}(T_n | \pi)}$$

If the design effect is greater than 1, we had better adopt \( \pi \) (as it reduces MSE)

If \( T_n \) is unbiased, the \( D_{Eff} \) is the ratio between the two variances in SRS and in \( \pi \)
Sampling managing designs

Randomized response sampling
It rules how to manage questions-answers (not how to select statistical units) when respondents may be reticent or embarrassed

Sensitive personal question

The respondent asks YES or NO to 1 of 2 possible questions:

Q1: are you homosexual?
Q2: are you heterosexual?

He/she chooses between Q1 and Q2 according to an event E such that:
- Only he/she may know if E is true
- The probability of E is known

E = your mother was born on December
where \( p(E) = \frac{1}{12} \)
He/she answers to:
Q1 if E is true
Q2 if E is false

\[ f_1 = \text{proportion of YES to Q1} \]
\[ f_2 = \text{proportion of NO to Q2} \]

He/she is homosexual

The proportion \( \pi \) of homosexual people is estimated as:

\[ \pi = f_1 \times p(E) + f_2 \times [1 - p(E)] \]
Mixture sampling

It rules how to manage the **experiment** (not how to select statistical units) when the phenomenon is **very rare**

Typical example: looking for a rare virus in a large number of blood tubes

- To analyze all tubes
  - or
- To mix a small quantity of blood from **groups** of tubes, restricting the number of analyses to do

**Example**

100 blood tubes (at worst 100 tests)

10 groups of 10 tubes (mix of blood from 10 tubes):

  10 tests, found 1 group positive to virus, look for virus within 10 tubes (at worst 20 tests)
Capture and recapture sampling

It rules how to manage the experiment (not how to select statistical units) when the population size \( N \) is unknown.

**Typical example: estimation of number \( N \) of fishes in a lake**

1. First sample of size \( n_1 \): fishes are labelled and are put back in the lake.
2. Second sample of size \( n_2 \): \( g \) is the number of fishes labelled.

We suppose that the following proportion is true:

\[
\frac{n_1}{N} = \frac{g}{n_2}
\]

\[
N = \frac{n_1 \times n_2}{g}
\]
Sample size

Trade-off between costs and accuracy

\( \delta \) = allowable error (accuracy)

\( \alpha \) = probability level

From confidence interval theory:

\[ \Pr \left( \left| T_n - \theta \right| \leq \delta \right) = 1 - \alpha \]

Under Normal distribution assumption or, anyway, from Central Limit Theorem:

\[ \delta = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]

Then:

\[ n \geq \left( \frac{\delta}{z_{\alpha/2} \sigma} \right)^2 \]

For Simple Random Sampling
2 problems

1. $\sigma$ is usually unknown

   Pilot survey for estimating $\sigma$ by $s$

   $\sigma$ is given the worst value basing on the known relation:

   $$0 \leq \sigma \leq \frac{(x_{\max} - x_{\min})^2}{9}$$

   for all unimodal distributions

2. Sample theory develops for inference on one parameter, whereas surveys are made for investigate several aspects