Analisi dei Dati

Introduzione e concetti di base

Domenico De Stefano

a.a. 2024/2025

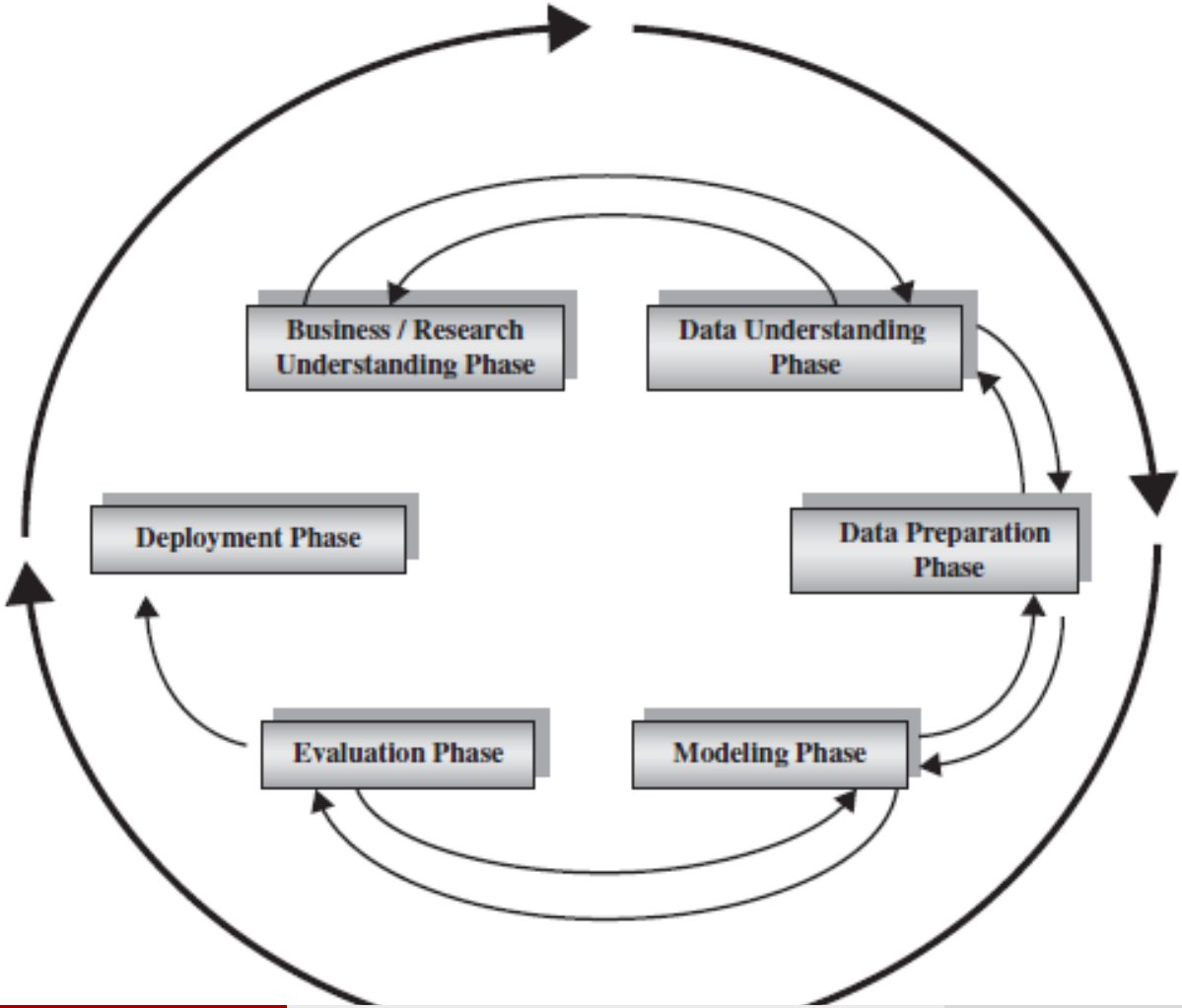
Rosso di sera...

Definizioni

- L'analisi dei dati è il processo di pulizia, trasformazione, esplorazione e modellazione di dati con il fine di evidenziare informazioni che suggeriscano conclusioni e supportino le decisioni
- il Data mining è il processo mediante il quale è possibile scoprire regolarità e ottenere informazioni a partire da un insieme spesso molto ampio di dati e spesso per fini previsivi
- molte tecniche coincidono nei due ambiti. Le tecniche utilizzate sono tecniche e metodologie statistiche

Analisi dei dati come processo (iterativo)

Cross-Industry Standard Process for Data Mining (CRISP-DM) sviluppato da Daimler-Chrysler, SPSS e NCR.



Analisi dei dati come processo (iterativo) I

1. Business/Research Understanding Phase
   - Definire gli obiettivi del progetto di ricerca (tipicamente è la risposta a una domanda di ricera: cosa vogliamo comprendere? che vincoli o restrizioni ci sono?) e le risorse necessarie per implementarlo (in quante unità lavoreranno? costi? tempo?)
   - Trasformare questi obiettivi in termini statistici
   - Preparare una strategia per raggiungere tali obiettivi (tempistiche/fasi del progetto di ricerca)
2. Data Understanding Phase
   - Raccogliere i dati
   - analisi esplorativa dei dati raccolti per inziare a comprendere le loro caratteristiche e avere delle prime informazioni in merito (strumenti di analisi statistica descrittiva univariata/bivariata come tabelle, grafici, indici). Suggerisce le possibili tecniche o modelli statistici che potrebbero essere usati nelle fasi successive
   - Valutare la qualità dei dati

Analisi dei dati come processo (iterativo) II

- (facoltativo) individuare dei sottoinsiemi dei dati che contengano informazioni di interesse e concentrare l'attenzione su di esso

3. Data Preparation Phase
   - lavoro di preparazione del dataset che contenga una versione pulita e sistematizzata dei cosiddetti dati grezzi raccolti (tipicamente struttura dati casi$\times$variabili)
   - Selezionare casi e variabili che si vogliono analizzare e che siano appropriati per raggiungere gli obiettivi conoscitivi del progetto
   - Se necessario, operare opportune trasformazioni di variabili
   - Organizzare ulteriormente i dati per usare determinate tecniche o modelli statistici
4. Modeling/Analysis Phase
   - Individuare e applicare le possibili tecniche o modelli appropriati per raggiungere gli obiettivi conoscitivi del progetto (ad es. valutare se le assunzioni del modello sono rispettate, ecc.)

Analisi dei dati come processo (iterativo) III

- spesso è possibile applicare diverse tecniche per raggiungere un determinato obiettivo
- eventualmente ritornare alla fase di preparazione dei dati per riorganizzarli e renderli processabili rispetto ad una particolare tecnica statistica

5. Evaluation Phase
   - la fase di analisi e modellizzazione ha prodotto risultati da tecniche diverse. Tali risultati devono essere valutati in merito alla loro qualità e la loro efficacia prima di essere utilizzati
   - Valutare se i risultati dell'analisi/modello raggiungono gli obiettivi specificati nella fase 1
   - Stabilire quali sono i limiti dell'analisi e quali problemi non è stato possibile affrontare
   - Decidere in che modo tali risultati possono essere usati

Analisi dei dati come processo (iterativo) IV

6. Deployment Phase
   - Specificare le modalità di uso del modello/tecnica adottati (spesso i risultati ottenuti nella fase 3 non rappresentano la fine del progetto ma sono il punto di inizio di un uso più concreto del modello stesso, ad esempio in campo aziendale)
   - Realizzare un report

Indice

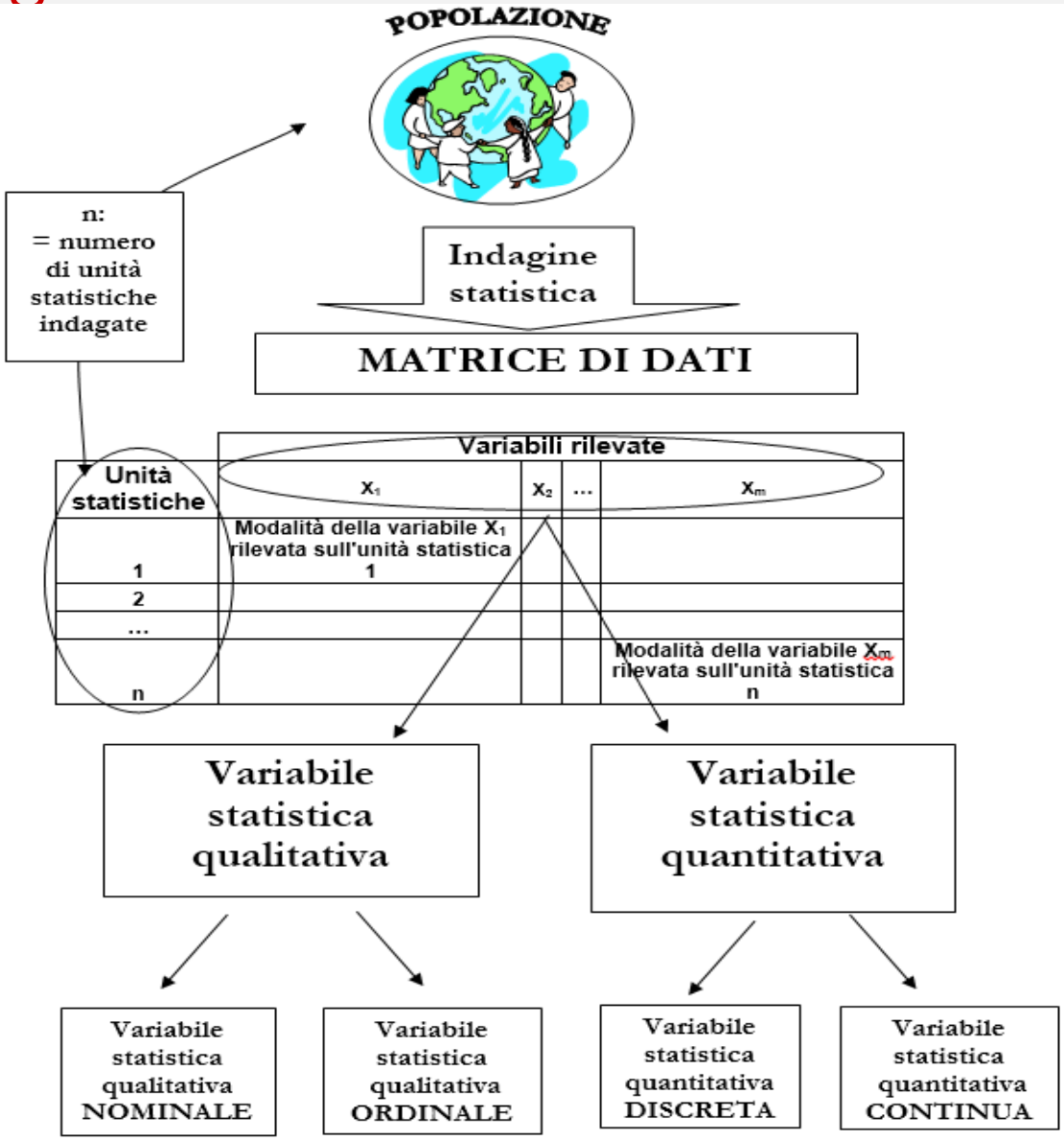
Indice

Terminologia elementare

Un dato statistico è il risultato della rilevazione (misurazione/osservazione/esperimento) di un qualche carattere su un'unità statistica appartenente a una popolazione.

- popolazione: l'insieme di casi su cui si vuole studiare un fenomeno di interesse (ad esempio la domanda di ricerca che ci si pone a inizio del progetto)
- unità statistica: il caso individuale componente del collettivo statistico (popolazione o campione);
- variabile (o carattere): ogni aspetto elementare oggetto di rilevazione sulle unità statistiche del collettivo;
- modalità di una variabile: i diversi modi con cui questa si presenta per le n unità statistiche del collettivo
- supporto: insieme (teorico) delle modalità di una variabile

Schema indagine



Variabili qualitative

- Una variabile è qualitativa o categorica se le modalità che si presentano sono espresse in forma verbale (qundi non numerica!);
  - una variabile qualitativa è nominale se le sue modalità non implicano una graduazione (in alcuni vecchi testi è detta anche sconnessa);
  - una variabile qualitativa è ordinale se le sue modalità implicano una graduazione;
- le modalità possono essere predefinite a priori;
- a volte, nelle indagini, le modalità vengono desunte a posteriori dalla descrizione dettagliata che il rilevatore fa dello stato della singola unità relativamente alla variabile in questione.

Esempio: qualitativa nominale /1

Ti è piaciuta l'ultima edizione del Festival di Sanremo?

- L'ho visto e mi è piaciuto
- L'ho visto e non mi è piaciuto
- Non l'ho visto

Esempio: qualitativa nominale /2

Qual è il tuo genere letterario preferito?

- Comico/umoristico
- Fantascienza
- Fantasy
- Giallo/noir/thriller
- Psicologico
- Romantico
- Storico
- Altro

Esempio: qualitativa ordinale /1

Quanto frequentemente bevi birra?

- Mai
- Raramente
- Qualche volta
- Spesso
- Ogni giorno
- Più volte al giorno

Esempio: qualitativa ordinale /2

Qual è il tuo titolo di studio?

- Licenza elementare
- Licenza media
- Diploma di scuola secondaria superiore
- Laurea

Variabili quantitative

- Una variabile è quantitativa se le modalità che si presentano sono espresse in forma numerica;
  - una variabile quantitativa è discreta se l'insieme delle sue modalità è finito oppure numerabile (detto in altri termini, se la quantità che rappresenta varia "a salti"). Spesso la loro rilevazione è frutto di un **conteggio**;
  - una variabile quantitativa è continua se l'insieme delle sue modalità è un intervallo, limitato o illimitato. Spesso la loro rilevazione è frutto di una **misurazione**;

Esempi: variabili quantitative

- Quante volte sei stato al cinema negli ultimi tre mesi?
- Qual è la tua altezza (in centimetri)?
- In che anno sei nato?
- Quante ore dedichi settimanalmente allo studio?

Esercizio

Che tipo di carattere è il prefisso telefonico?

(a) numerica, continua
(b) numerica, discreta
(c) qualitativa, nominale
(d) qualitativa, ordinale

Matrice dei dati

Esempio di dati raccolti mediante questionario su studenti del corso di statistica a scienze politiche (dello scorso A.A.):

variabile ↓

Stu.	genere	Sanremo	···	ore di sonno	ore di studio
1	maschio	Non l'ho visto	···	8	2
2	femmina	L'ho visto e mi è piaciuto	···	6	30
3	maschio	Non l'ho visto	···	9	5
4	femmina	Non l'ho visto	···	8	25
⋮	⋮	⋮	⋮	⋮	⋮
52	femmina	Non l'ho visto	···	8	20

Dalla matrice dei dati alla sintesi: produrre conoscenza dai dati

Una volta raccolti i dati e disposti nella matrice dei dati questi devono essere organizzati e sintetizzati/analizzati attraverso gli strumenti della statistica descrittiva

La sintesi/analisi dei dati si ottiene principalmente attraverso 3 tipi di strumenti:

- Tabelle (distribuzioni di frequenza)
- Grafici
- Indici numerici

Oggetto delle indagini statistiche: la popolazione

L'obiettivo è rispondere ad una certa domanda di ricerca su di una popolazione attraverso lo studio di alcune caratteristiche.
Una popolazione è un collettivo (un insieme di oggetti o individui)

- i componenti del collettivo sono detti unità statistiche, sono esempi
  - la popolazione degli italiani di sesso maschile con oltre 18 anni al 01/01/2020;
  - le famiglie italiane al 01/01/2020;
  - i clienti di un negozio.
- La popolazione può essere finita (ad es. la popolazione italiana) o infinita (ad es. tutte le persone iscritte a Scienze Politiche, oggi o in futuro).

Esempi di caratteristiche

Le caratteristiche (variabili) da rilevare potrebbero essere:

- per la popolazione degli italiani di sesso maschile con oltre 18 anni al 01/01/2020;
  - l'abitudine al fumo, il comportamento elettorale, il numero di film visti in un mese
- per le famiglie italiane al 01/01/2020;
  - il reddito familiare, la spesa per consumi, il numero di figli;
- per i clienti di un negozio.
  - la spesa per acquisti, pagamento con carta di credito, l'età, ...

La componente elementare che costituisce una popolazione à detta unità statistica

Dati questi esempi di popolazione quali sono le unità statistiche?

Indice

Censimento

- La prima forma di raccolta dati si attua osservando tutti gli individui di una popolazione
  - Questo è un censimento (esattamente quello che conduce l'ISTAT ogni 10 anni sull'**intera popolazione residente in Italia** `http://www.istat.it/it/censimento-popolazione`).
- Ci sono problemi nel condurre un censimento:
  - Può essere difficile: ci sono sempre individui difficili da localizzare. E questi individui potrebbero avere caratteristiche che li distinguono dal resto della popolazione.
  - Le popolazioni sono in movimento.
  - Fare un censimento è costoso.
  - L'elaborazione di tutti i dati è lunga e complessa

Campionamento

Quando non possiamo (vogliamo) osservare l'intera popolazione facciamo ricorso al campionamento.

Campionamento e inferenza

Osserviamo una parte della popolazione, il campione, e generalizziamo all'intera popolazione quanto osservato sul campione.

- Campionare è naturale: lo si fa anche in cucina.
- Immaginiamo di cucinare una zuppa: per avere un'idea della possibile riuscita, si fa un assaggio (il censimento non lascerebbe zuppa per la cena).
- Quando si assaggia un cucchiaio di zuppa e si decide che il **contenuto del cucchiaio** non è abbastanza salato, si sta facendo analisi esplorativa (ossia usiamo solo strumenti di statistica descrittiva).
- Se si conclude che tutta la zuppa è insipida, si fa inferenza.

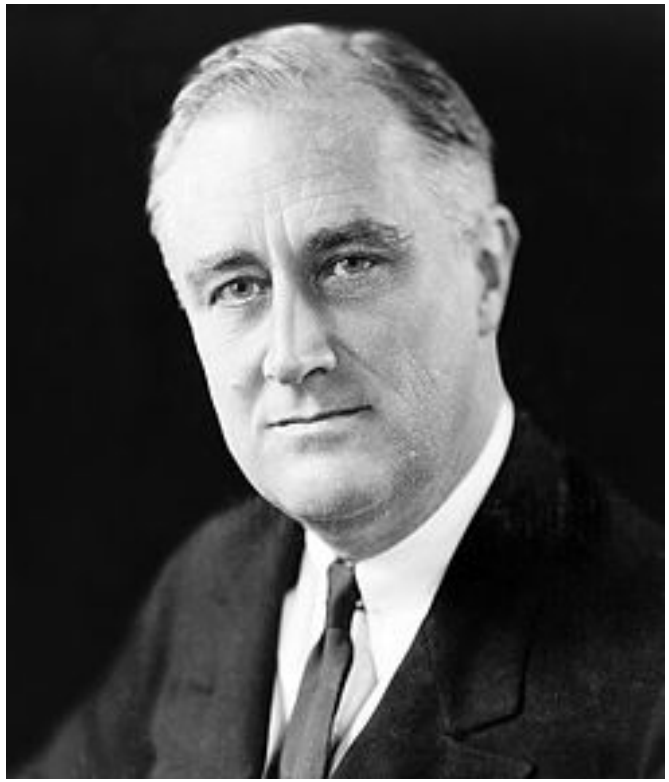
Campionamento e rappresentatività

- Perché l'*inferenza* sia valida, l'assaggio deve essere rappresentativo dell'intera preparazione.
- Se buttiamo prima il sale, poi tutti gli ingredienti, non mescoliamo mai e assaggiamo la zuppa in superficie, probabilmente non abbiamo un assaggio "rappresentativo".
- Se buttiamo prima il sale, poi tutti gli ingredienti, poi mescoliamo bene tutti gli ingredienti prima dell'assaggio, probabilmente la rappresentatività dell'assaggio migliorerà.

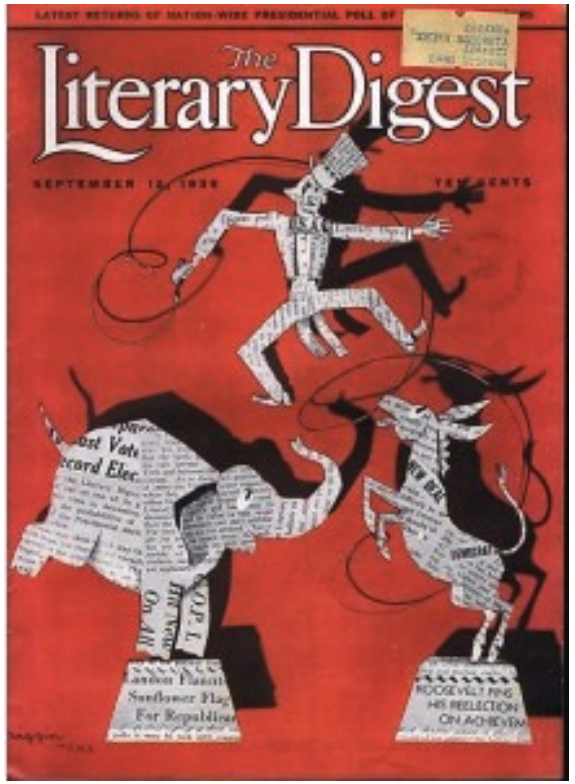
Esempio: Landon vs. Roosvelt

Un esempio storico di campione non rappresentativo:

Nel 1936, Alf Landon si propose come candidato repubblicano alla presidenza opponendosi a Franklin Roosevelt, candidato democratico.



L'indagine del *The Literary Digest*



- *The Literary Digest* fu un settimanale pubblicato negli USA dal 1890 al 1938.
- È noto soprattutto per il clamoroso fallimento nel prevedere il risultato delle elezioni presidenziali del 1936.
- *The Literary Digest* contattò circa 10 milioni di americani, ricevendo 2.4 milioni di risposte.
- L'indagine predisse che Landon avrebbe stravinto e che il partito democratico avrebbe avuto solo il 43% dei voti.
- Risultati: il partito democratico vinse, con il 62% dei voti.
- Il giornale fu totalmente screditato e cessò le pubblicazioni dopo poco.

The Literary Digest Poll – cosa andò storto?

- Il "campione" era enorme (2.4 milioni di persone),
- però il giornale contattò
  - i suoi lettori,
  - i possessori di automobili
  - gli utenti telefonici
- Questi gruppi avevano un reddito ben superiore alla media nazionale (era il periodo della grande depressione),
- cioè era un gruppo di elettori molto più probabilmente sostenitori dei repubblicani.
- In altre parole, il campione non era rappresentativo dell'intera popolazione.

Campioni grandi sono preferibili, ma..

- The Literary Digest aveva un campione di 2.4 milioni di persone, che è enorme, ma siccome era distorto, non produsse previsioni accurate.
- In termini culinari: se la zuppa non è ben mescolata, non importa quanto grande è l'assaggio....

Come dev'essere il campione

Quando diciamo che il campione sono n individui selezionati nella popolazione, questo non vuol dire che qualunque gruppo di n individui vada bene.

Campione "rappresentativo"

Un campione "rappresentativo" è un sottoinsieme della popolazione che ne riflette le caratteristiche. (Una versione in miniatura della popolazione.)

È il fatto che il campione è rappresentativo che consente di generalizzare i risultati che si ottengono sulla base di calcoli fatti sul campione, alla popolazione.

Come NON dev'essere il campione

NON si ottiene un campione rappresentativo

- prendendo le persone presenti in quest'aula,
- prendendo gli amici/parenti/conoscenti,
- ponendo una domanda in una trasmissione televisiva e invitando il pubblico a rispondere via telefono o sms o internet.

questi gruppi di persone hanno caratteristiche peculiari, non possiamo escludere che queste siano legate alle caratteristiche che stiamo indagando, quindi introdurremmo delle distorsioni.
Per grande che sia, un campione non rappresentativo non consente generalizzazioni.

Campione autoselezionato

In particolare, riportare risulati basati sul **porre una domanda in una trasmissione televisiva e invitare il pubblico a rispondere via telefono o SMS o internet.** è abbastanza usuale.

I risultati vanno visti con molta diffidenza per varie ragioni

- risponderà più facilmente chi ha più a cuore il problema, ovvero ha un'opinione "forte" su esso;
- i rispondenti sono tutti spettatori di quella particolare trasmissione (oltreché spettatori televisivi ecc.).

Come ottengo un campione rappresentativo?

L'idea è di selezionare le unità da includere nella popolazione in modo casuale, poi ci sono diversi metodi

- Il modo più semplice è scegliere n individui in modo che ciascun individuo della popolazione abbia la stessa probabilità di essere estratto.
- Altre opzioni sono spesso usate allo scopo di
  - migliorare la rappresentatività,
  - semplificare la procedura (risparmiare quattrini);

  tra queste
  - campione stratificato,
  - campione a grappoli,
  - campione a più stadi;

  tutti possono essere estratti, le probabilità possono variare.

Indice

Esempio 1

Per studiare un nuovo farmaco, viene organizzata una prova clinica.

Esempio 1

Per studiare un nuovo farmaco, viene organizzata una prova clinica.

Sono reclutati 50 pazienti. A 25 di questi, estratti a caso, viene somministrato il nuovo farmaco; ai rimanenti 25 un placebo (una sostanza inerte che viene somministrata per far credere di aver ricevuto un farmaco).

Esempio 1

Per studiare un nuovo farmaco, viene organizzata una prova clinica.

Sono reclutati 50 pazienti. A 25 di questi, estratti a caso, viene somministrato il nuovo farmaco; ai rimanenti 25 un placebo (una sostanza inerte che viene somministrata per far credere di aver ricevuto un farmaco).

Dopo un periodo di tempo, i due gruppi sono confrontati per vedere se il gruppo che ha ricevuto il trattamento mostra effetti positivi.

Esempio 2

Per studiare gli effetti dell'inquinamento atmosferico sulla salute, nel 2004 lo studio MISA (metanalisi italiana degli studi sugli effetti a breve termine dell'inquinamento atmosferico) ha studiato le relazioni tra l'inquinamento e le morti per per cause respiratorie e per cause cardiovascolari nel periodo 1996-2002 in 15 città italiane, scelte tra i principali centri urbani del paese. Complessivamente, lo studio ha coinvolto un totale di 9 milioni e centomila abitanti al censimento 2001.

Definizioni

- Esempio 1 $\longrightarrow$ esperimento.
  Il ricercatore assegna un "trattamento" ad alcuni individui scelti attraverso un meccanismo casuale. Il punto cruciale è che c'è un trattamento che viene somministrato e un meccanismo casuale per somministrarlo.

- Esempio 2 $\longrightarrow$ studio osservazionale.
  Il ricercatore "osserva" semplicemente, senza intervenire sui soggetti. Può essere di tipo retrospettivo, condotto sulla base di documentazione raccolta in passato e, quindi, già esistente prima della decisione di iniziare lo studio.

Riassumendo...

Classi di indagine/raccolta dati

- Censimento
- Indagini campionarie
- (ce ne sarebbe una terza) analisi di dati già raccolti per altri fini (es. dati amministrativi)
- (e una quarta) big data

Tipologia

- Esperimenti (si interviene sulle u.s.).
- Studi osservazionali

Lista non esaustiva!

Ogni tipo di indagine differisce in termini di ammontare di risorse richieste e "forza" dell'inferenza che può essere condotta.