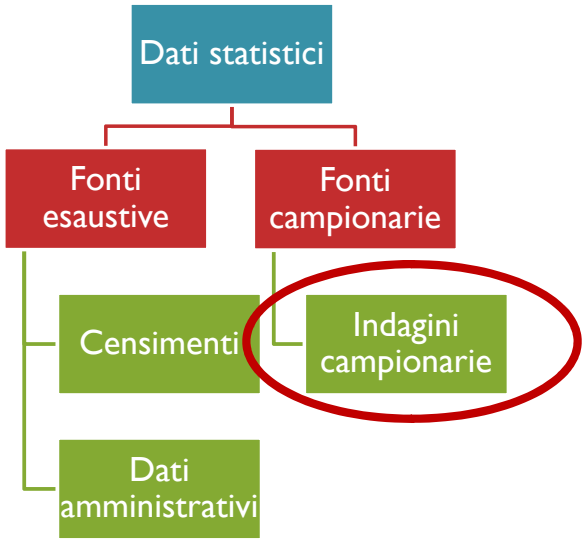
La costruzione dei dati e la rappresentazione di informazione statistica

Giovanna Boccuzzo
Dipartimento di Scienze
Statistiche Università di Padova

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Da dove provengono i dati statistici?



Ci sono i sondaggi d'opinione…

Referendum, il "no" avanti: sondaggi amari per Renzi

Sondaggi Brexit: in testa il NO il giorno del referendum. Ultimi risultati

Brexit, cresce l'onda del Sì. Juncker: "La Ue ce la farà comunque"

" Il sondaggista: "Col salva banche rischia alle amministrative"

Summer School: la matematica incontra il mondo
San Pellegrino Terme, 5 - 6 - 7 Settembre 2016

Ci sono le stime "ufficiali"…

Pil: crescita zero nel secondo trimestre, + 0,8% il dato annuo. Istat: confermate le s congiunturali, ritoccata leggermente verso l'alto la stima di crescita.

PREVISIONI

La Banca d'Italia gela gli ottimisti: «Il Paese crescerà meno del previsto»

Istat: disoccupazione scende all'11,4% ma aumenta quella giovanile, prosegue deflazione

ECONOMIA | 14 LUGLIO 2016

Quanti sono i poveri in Italia

E cosa vuol dire, da un punto di vista economico e statistico, essere "poveri": i nuovi dati dell'ISTAT

E ci sono anche i sondaggi retribuiti….

Sondaggi Retribuiti

SondaggiPagati.Net

sondaggi retribuiti
come guadagnare esprimendo opinioni!

SONDAGGI ONLINE RETRIBUITI

pagasondaggi

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Il più famoso esempio di sondaggio sbagliato

WHY THE 1936 *LITERARY DIGEST* POLL FAILED

Literary Digest
2,3 milioni rispondenti
Landon: 55% Roosevelt: 41%

Gallup
50mila rispondenti
Landon: 44% Roosevelt: 56%

Come andò:
Landon: 37% Roosevelt: 62%

Cosa andò storto?

Lista non completa e non rappresentativa:
lista telefonica + iscritti club + registro automobilistico = distorsione da selezione

Non risposte:
Tasso di non risposta pari al 77%!

La lista iniziale prevedeva 10 milioni di unità

Non risposte non casuali:
Distorsione da non risposta, i sostenitori di Landon votarono in più

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

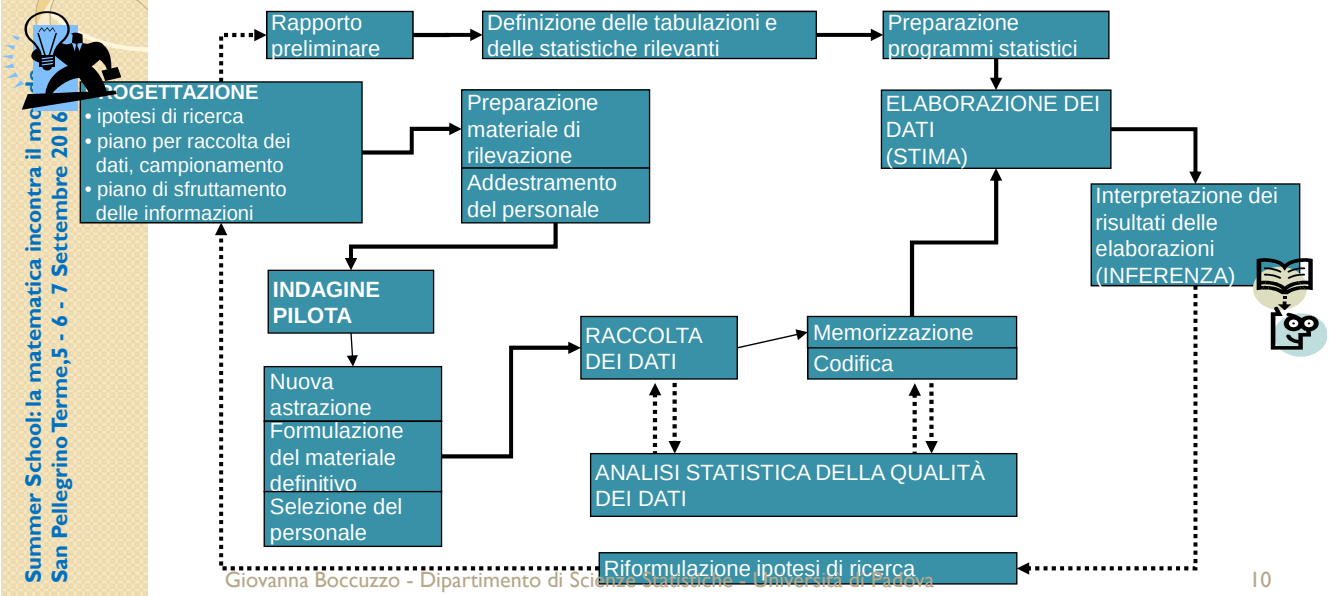
La sfiducia verso la statistica…

È giustificata?

La costruzione del dato mediante campionamento

- La realizzazione di un'indagine campionaria può essere molto onerosa in termini di tempo e risorse
- L'indagine sulle forze di lavoro è la più rilevante in Italia; rappresenta la principale fonte di informazione statistica sul mercato del lavoro italiano
- Considera ogni anno un campione di 250mila famiglie (circa 600mila persone) in 1100 comuni, le stime hanno un dettaglio provinciale.
- Le famiglie sono intervistate 4 volte in 15 mesi, la prima faccia a faccia (CAPI), poi per telefono (CATI)

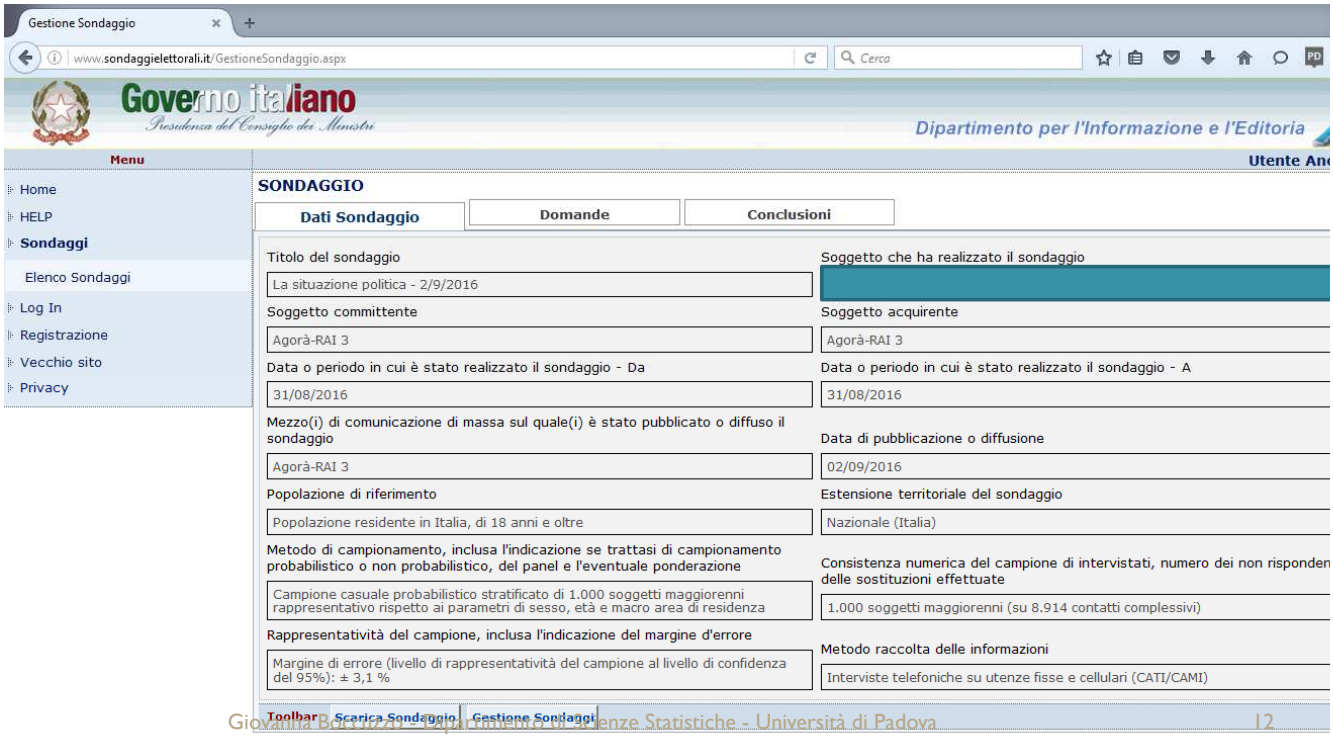
Fasi dell'indagine statistica



Come si svolgono i sondaggi?

dal Mattino di Padova del 29/5/2016

XXX ha realizzato questa indagine che si è svolta a livello nazionale dal 22 marzo al 4 aprile 2016 su un **campione rappresentativo** della popolazione residente in Italia, con età superiore ai 18 anni. I rispondenti totali sono stati **1997 (su 13.287 contatti)**, l'analisi dei dati è stata **riproporzionata** sulla base del genere, del territorio, delle classi d'età, della condizione professionale e del titolo di studio. Il **margine di errore** è pari a +- 2.2%. La rilevazione è avvenuta con un'indagine attraverso i principali social network e con un **campione casuale** raggiungibile con i metodo **CAWI e CATI**.



Analizziamo punto per punto….

Campione casuale (o probabilistico): Ogni unità della popolazione ha probabilità non nulla di essere selezionata: $0<p_i \leq 1, \sum p_i=1$.

Per estrarre un campione casuale serve una LISTA ESAUSTIVA* della popolazione

* a meno di campionamento a stadi

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

La lista

- *Popolazione residente in Italia, con età superiore ai 18 anni.*

Possibili liste da cui trarre il campione:

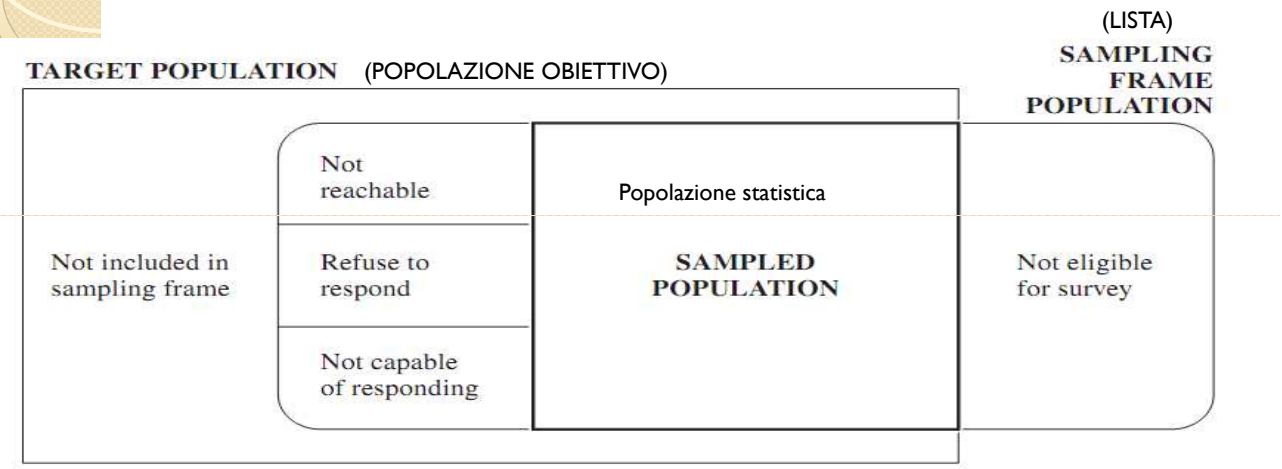
- Anagrafi della popolazione
- Liste elettorali
- Censimento

- Non disponibili per sondaggi elettorali, d'opinione, ricerche di marketing …

  → sono acquistate e utilizzate altre liste (di numeri telefonici e indirizzi mail), non esaustive e presumibilmente non rappresentative

Distorsione da selezione

Popolazione obiettivo e popolazione statistica



Disegno campionario stratificato vs. per quote

Suddivisione della popolazione per sesso età e titolo di studio

Strati

M <20 Alto	M 20-45 Alto	M >45 Alto
M <20 Medio	M 20-45 Medio	M >45 Medio
M <20 Basso	M 20-45 Basso	M >45 Basso
F <20 Alto	F 20-45 Alto	F >45 Alto
F <20 Medio	F 20-45 Medio	F >45 Medio
F <20 Basso	F 20-45 Basso	F >45 Basso

Stratificato: la popolazione è suddivisa in strati, da ogni strato è estratto un campione

≠

Quote: è nota la quota della popolazione negli strati, e si cercano rispondenti finché non si riempiono tutti gli strati

Distorsione da non risposta

- *I rispondenti totali sono stati **1997 (su 13.287 contatti)***
- Il tasso di non risposta è altissimo (85%), chi ha risposto (e riempito le quote) molto probabilmente non è rappresentativo di chi non ha risposto (lavoro diverso, orari diversi, rifiuto, …)

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Tecnica di rilevazione

- *La rilevazione è avvenuta con un'indagine attraverso i principali social network e con un **campione casuale** raggiungibile con i metodo **CAWI e CATI***
  - CATI: Computer Assisted Telephone Interview;
  - CAWI: Computer Assisted Web Interview
- Rilevazione attraverso social network: pochissimi volontari, nulla di casuale

L'errore

*Il **margine di errore** è pari a +- 2.2%.*

$$n \cong \frac{z^2_{\alpha/2}\frac{s^2}{D^2}}{1+\frac{z^2_{\alpha/2}s^2}{D^2 N}} = \frac{z^2_{\alpha/2}s^2}{D^2+\frac{z^2_{\alpha/2}s^2}{N}}$$

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Ingredienti per la costruzione di un buon dato statistico:

- Lista/e di partenza
- Tempo
- Disponibilità economica
- Accuratezza di rilevazione
- Competenza in tutte le fasi dell'indagine

Le risorse impiegate dipendono in buona parte dagli obiettivi di conoscenza e dall'uso che se ne fa delle statistiche prodotte.
L'importante è non fare di tutta l'erba un fascio

Dal dato all'informazione statistica

Disporre di dati statistici non significa disporre automaticamente di informazione statistica effettivamente utile.

Il **dato statistico** è il prodotto finale della singola indagine o rilevazione statistica

Es: il numero di morti per tumore nel 1972 in Italia è pari a 105.093 unità

L' **Informazione statistica** è la contestualizzazione del dato statistico, lo sfruttamento contemporaneo di più dati/indicatori per produrre informazione

Es: il numero di morti per tumore nel 1972 in Italia è pari a 105.093 unità

Il n° di morti per tumore nel 1992 è pari a 151.162 unità

Il tasso di mortalità per i maschi è passato da 17,2 per 10mila nel 1971-73 a 19,2 nel 1989-91. Per le donne da 10,5 a 10,08

Costruire una tabella (banale…)

TITOLO: cosa, descritto con quale tipo di dato, secondo quali variabili. Luogo, anno.

	Variabile testata	
Variabile fiancata	Corpo tabella	Totali di colonna
	Totali di riga	Totale

Tabella di percentuali

La tabella può contenere:

- Percentuali di riga, o
- Percentuali di colonna, o
- Percentuali di cella

	Variabile testata	
Variabile fiancata	xx xx xx xx xx	100
	xx xx xx xx xx	100
	xx xx xx xx xx	100
	xx xx xx xx xx	100

	Variabile testata	
Variabile fiancata	xx xx xx xx	xx
	xx xx xx xx	xx
	xx xx xx xx	xx
	100 100 100 100	100

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Scelta delle percentuali

	Fino elementari	Media inferiore	Media superiore	Laurea o titoli superiori	Totale
Fumatori	13	95	31	6	145
Non Fumatori	4	40	47	14	105
Totale	17	135	78	20	250

Fumatori di 18 anni e più secondo il titolo di studio

	Fino elementari	Media	Media	Laurea o	Totale
% riga					

% colonna	76,5	70,4			58,0

La scelta della % sbagliata porta a conclusioni errate!

Esempi di rappresentazioni inappropriate

"Si nota come le prime indicazioni sembrino supportare l'ipotesi secondo la quale gli italiani usufruiscono in misura maggiore di strutture private rispetto agli stranieri."

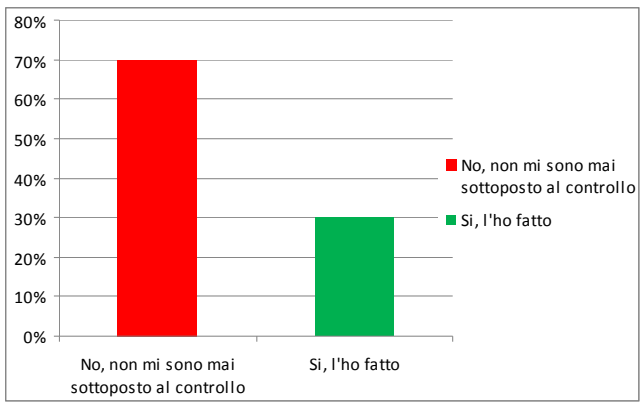
	Italiano	Straniero	Totale
Pubblica	69.7%	30.3%	100%
Privata	83.3%	16.7%	100%

Tabella 6.1: Relazione tra tipo di struttura e cittadinanza

Summer School: la matematica incontra il mondo
San Pellegrino Terme, 5 - 6 - 7 Settembre 2016

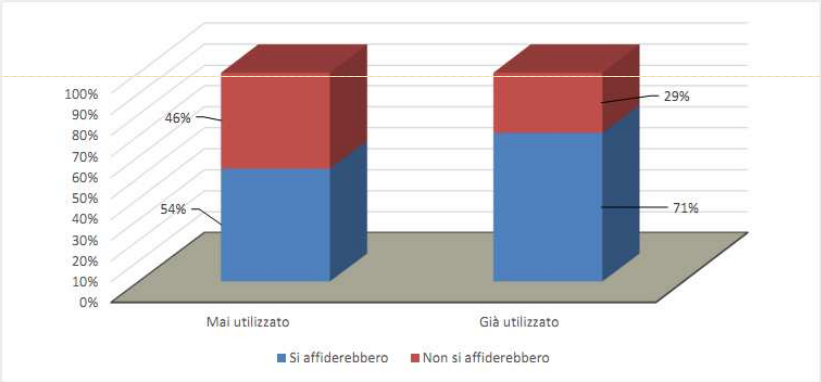
Esempi di rappresentazioni inappropriate

Grafico 3: Predisposizione a controlli andrologici, da parte degli studenti (maschi) dell'Università di Padova, inerente al mese di Maggio 2016. Istogramma.

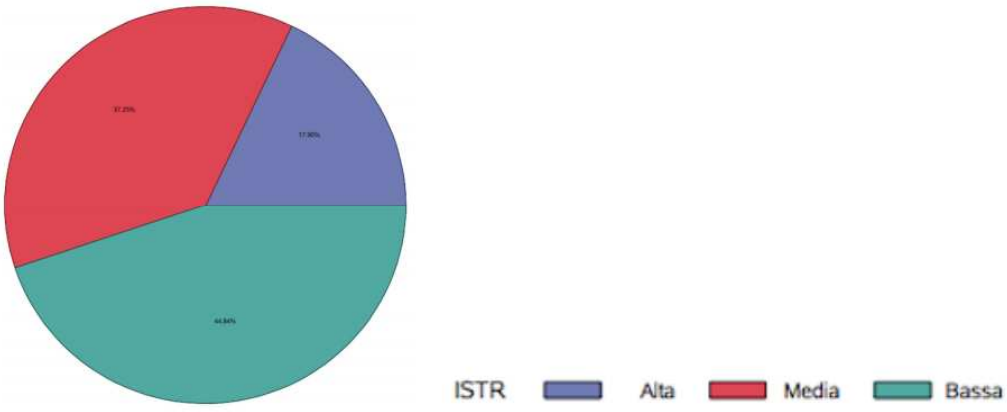


Esempi di rappresentazioni inappropriate

GRAFICO 6: Studenti che affiderebbero l'organizzazione di un viaggio ad un Tour Operator, non avendo mai usufruito del servizio; studenti che affiderebbero ancora l'organizzazione di un viaggio ad un Tour Operator, avendo già usufruito del servizio. Quoziente per 100 persone.



Esempi di rappresentazioni inappropriate



(a) *Istruzione*

Figura 5: Distribuzione dei felici nelle Condizioni Socio Economiche e Demografiche

Non solo percentuali…

- Tipologie di rapporti statistici:
  - **Rapporti di composizione (Percentuali)**
  - **Rapporti di coesistenza**
  - **Rapporti di derivazione (Tassi)**
  - **Rapporti medi (Densità)**

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Il tasso

Quoziente che si ottiene dal rapporto fra l'intensità di un certo fenomeno e l'intensità di un altro che ne costituisca il presupposto necessario.

Tasso di laurea regolare (triennale)= Laureati in corso/immatricolati 3 anni prima

Non è sempre ovvio costruire un tasso, per via della definizione del denominatore (popolazione "a rischio" di generare il numeratore)

Un esempio: tasso o rapporto di abortività volontaria?

Tasso di abortività volontaria: IVG/donne età feconda

1994	RAV x 1000	TAV x 1000
Veneto	157.7	5.4
Sicilia	134.7	6.9

Rapporto abortività volontaria: IVG/gravidanze, spesso calcolato come IVG/Nati

1994	RAV x 1000	TAV x 1000
ITALIA NORD-OCCIDENTALE	288.9	9.7
ITALIA MERID. E INSULARE	220.0	9.8

Indicatore poco specifico

Summer School: la matematica incontra il mondo
San Pellegrino Terme,5 - 6 - 7 Settembre 2016

Definire indicatori appropriati

- Finalizzati
- Sensibili e specifici
- Accurati e precisi