

COESIONE  
ITALIA



*A Scuola di*  
**OPENCOESIONE**

# *Conoscere e preparare un'indagine di statistica ufficiale*

Rita Lima, Servizio Metodi, qualità e metadati, Direzione Centrale per la Metodologia e il Disegno dei Processi Statistici

18/12/2024



# INDICE

## ***Contenuti:***

- Le fasi di una indagine statistica
- La progettazione dell'indagine
- L'elaborazione dei dati raccolti
- La diffusione dei risultati
- La qualità dei dati
- Esempi di indagine statistica

## ***Competenze acquisite:***

- Raccolta, analisi, rappresentazione e descrizione dei dati
- Realizzazione e applicazione dei principali indicatori statistici

## ***Rafforzamento competenze generali in tema di:***

- Impostazione di una ricerca con uso di dati quantitativi
- Tecniche di rappresentazione dati per un'efficace presentazione di risultati analitici

Per **INDAGINE STATISTICA** si intende un insieme di attività finalizzate ad approfondire la conoscenza di un fenomeno mediante alcuni passaggi (**FASI**):

- ❑ **TRADUZIONE** di un concetto teorico (**CARATTERE DI INTERESSE**) in un concetto operativo (**CARATTERE MISURABILE**)
- ❑ **RACCOLTA** informazioni sul comportamento del singolo componente (**UNITÀ STATISTICA**) dell'insieme che si vuole indagare (**POPOLAZIONE O FENOMENO COLLETTIVO**)
- ❑ **TRATTAMENTO** delle informazioni raccolte:
  - **ELABORAZIONE** dei risultati dell'**OSSERVAZIONE** di uno o più **CARATTERI** del **FENOMENO COLLETTIVO**
  - **ANALISI** dei dati raccolti (**QUANTITATIVI** e **QUALITATIVI**) e **DIFFUSIONE** dei risultati ottenuti

Esempio

|                             |  |
|-----------------------------|--|
| <b>Fenomeno collettivo</b>  | Situazione occupazionale in Sardegna   |
| <b>Popolazione</b>          | Residenti in Sardegna in un certo momento  |
| <b>Unità statistiche</b>    | Singoli individui  |
| <b>Caratteri</b>            | Sesso, età, stato civile, condizione professionale, ...  |
| <b>Osservazione</b>         | Registrazione delle risposte ad un questionario  |
| <b>Aspetti del fenomeno</b> | <ul style="list-style-type: none"><li>• Quota disoccupati</li><li>• Quota disoccupazione giovanile</li><li>• Durata della disoccupazione</li><li>• Relazione tra occupazione, disoccupazione e sesso</li><li>• ...</li></ul> |

I **DATI** sono una raccolta di informazioni (esprese in forma numerica).

La **VARIABILE** (o **CARATTERE**) è una caratteristica di interesse rilevata sulle unità statistiche (ad esempio, età, peso, reddito, ...). La scelta della variabili dipende dal concetto teorico: forma di classificazione della realtà che consente di raggruppare fatti ed oggetti; una lettura della realtà (es. capacità relazionale; età; genere; ecc.) per controllarla e misurarla = **Definizione operativa di un concetto**

Una variabile può essere:

- **QUALITATIVA** o **CATEGORIALE** quando le sue modalità sono espresse in forma verbale (sesso, livello di istruzione, gusto del gelato, ...). Se le modalità sono solo due si parla di variabili **DICOTOMICHE O BINARIE** (M/F, presenza/assenza, ecc.).

A sua volta una variabile qualitativa può essere:

- **SCONNESSA** o **NOMINALE** se non esiste nessun ordinamento tra le modalità.

Esempi: la variabile sesso con modalità M e F; la variabile corso di laurea frequentato con modalità Statistica, Economia, Architettura, ecc....

- **ORDINALE** se è possibile individuare un ordinamento naturale delle modalità.

Esempi: la variabile livello di istruzione con modalità elementare, media inferiore, media superiore, .... ; la variabile giudizio con modalità insufficiente, sufficiente, discreto, ottimo.



Oppure, una variabile può essere:

- **QUANTITATIVA** (o **NUMERICA**) quando le modalità sono espresse da numeri (età, peso, ..) si identificano tramite numeri.

A sua volta una variabile quantitativa può essere:

- **CARDINALE DISCRETA** quando l'insieme delle modalità è finito o numerabile ed è ottenuto tramite un'operazione di conteggio (classe dei numeri naturali).

Esempi: la variabile numero di figli con modalità 0, 1, 2, 3...; la variabile numero di biglie con modalità 0,1,2.....



- **CARDINALE CONTINUA** quando l'insieme delle modalità è un intervallo (ossia un sottoinsieme) ottenuto tramite un'operazione di misurazione (classe dei numeri reali).

Esempi: la variabile peso (in kg), la variabile altezza (in cm),.....



### VARIABILI QUALITATIVE vs QUANTITATIVE (Traduzione di espressioni verbali in numeri)

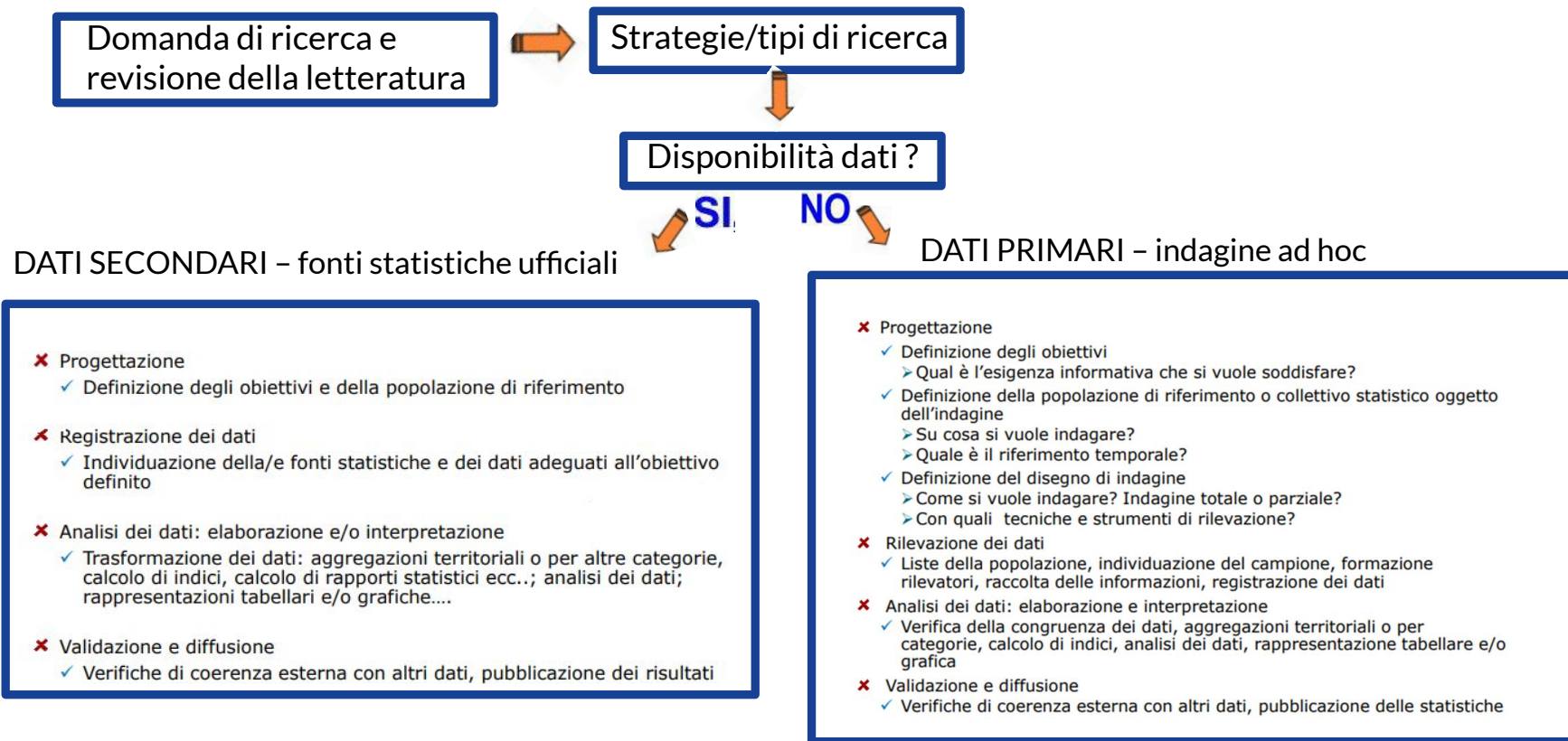
- A seconda del tipo di variabili osservate, sono possibili diverse analisi statistiche.
- Ci sono degli strumenti statistici appositi per studiare tipi diversi di variabili.
- Tra i vari tipi di dati è implicita una gerarchia (le variabili quantitative possono essere discretizzate, le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali, quelle ordinali possono essere considerate nominali).

| Stati della proprietà   | Procedura       | Tipo di variabile | Operazioni possibili      |
|-------------------------|-----------------|-------------------|---------------------------|
| Discreti non ordinabili | Classificazione | Nominale          | $= \neq$                  |
| Discreti ordinabili     | Ordinamento     | Ordinale          | $= \neq > <$              |
| Discreti enumerabili    | Conteggio       | Cardinale         | $= \neq > < + - \times ;$ |
| Continui                | Misurazione     | Cardinale         | $= \neq > < + - \times ;$ |

### DATI UNIVARIATI vs MULTIVARIATI

- L'analisi univariata considera una sola variabile rilevata sulle unità.; l'analisi bivariata considera due variabili contemporaneamente; .....lo studio congiunto di più variabili è detto analisi multivariata ed è per estensione, ogni forma di esplorazione e concettualizzazione dei dati raccolti.

## IL CONCETTO TEORICO, IL CONCETTO OPERATIVO E LE FASI DELL'INDAGINE STATISTICA





## I DATI SECONDARI – fonti di statistiche ufficiali

Sono quelli ottenuti dalle **statistiche ufficiali** ovvero prodotti da enti (prevalentemente) pubblici secondo canoni standard di qualità per soddisfare esigenze informative di pubblica utilità.

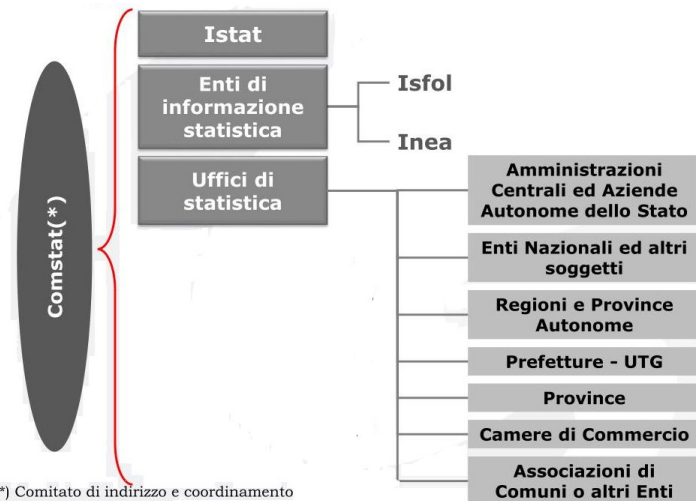
La statistica ufficiale è un **bene pubblico** (rappresentazione di fenomeni di rilevanza pubblica).

### LA FASE DELLA PROGETTAZIONE

- ❑ **WHO**= Chi è il produttore? È riconosciuto ed autorevole a livello pubblico?
- ❑ **WHY**= Per quali obiettivi viene prodotta l'informazione statistica: per servire il paese e i decisori politici o per obiettivi commerciali?
- ❑ **WHAT**= La produzione statistica è riportata in programmi statistici normati? Quali fenomeni sono investigati? Le statistiche sono confrontabili?
- ❑ **HOW**= Si adottano metodologie solide, classificazioni standard, procedure e strumenti standardizzati, sistemi di controllo della qualità?



## I PRODUTTORI DI STATISTICHE UFFICIALI E IL SISTAN



ALTRI ORGANISMI E ENTI NAZIONALI E INTERNAZIONALI Eurostat, OCSE, FMI, Bankitalia

ALTRI PRODUTTORI DI INFORMAZIONI STATISTICHE Censis, Università, Centri Studi, Fondazioni, Istituti privati di ricerca e di ricerche di mercato

Fattori istituzionali e organizzativi: l'indipendenza professionale, il mandato per la raccolta dei dati, l'adeguatezza delle risorse, l'impegno in favore della qualità, la riservatezza statistica e l'imparzialità e obiettività.

L'Istituto nazionale di statistica (ISTAT), attivo dal 1926, è il maggiore produttore di dati economici e sociali del Paese e dal 2016 è un ente pubblico di ricerca, rivolto alla produzione e all'analisi di dati.

Provvede a:

- ❑ **INDIRIZZARE** e **COORDINARE** le attività statistiche degli altri soggetti del SISTAN
- ❑ **ASSISTERE** sugli aspetti tecnici enti e uffici del SISTAN



[www.istat.it](http://www.istat.it) fonte d'informazione **OPEN** fruibile gratuitamente e riutilizzabile. Dati rilasciati sotto la licenza Creative Commons, che prevede la possibilità di riprodurre, distribuire, trasmettere e adattare liberamente dati e analisi dell'Istituto, a condizione che sia citata la fonte (Istat, 2019).

Banche Dati – Istat pagina Banche dati e sistemi informativi del sito istituzionale (elenco completo).

#### Microdati:

| IDQUESTIONARIO     | MODELLO | TIPALL | NOCC | NFAM | TITGOD | SUPERF |     |
|--------------------|---------|--------|------|------|--------|--------|-----|
| 111111111111111111 | L       |        | 1    | 5    | 1      | 4      | 200 |

collezioni di dati elementari/individuali

#### Macrodati:

dati aggregati, riferiti a gruppi omogenei di dati elementari

Popolazione residente al 1° Gennaio 2014  
per età, sesso e stato civile - Italia

| Eta'      | Totale Maschi | Totale Femmine | Maschi + Femmine |
|-----------|---------------|----------------|------------------|
| 100 e più | 2.993         | 14.891         | 17.884           |

#### Metadati:

- descrizioni
  - definizioni e informazioni
- per interpretare correttamente i dati**



## DOMANDA DELLA RICERCA: TRA I GIOVANI DI 14-19 ANNI CHI FUMA DI PIÙ?

Collettivo: giovani di 14-19 anni

Informazioni sul processo statistico:

<https://www.istat.it/it/archivio/91926>

Estrazione dati: <http://dati.istat.it/>

ENGLISH HOME

POPOLAZIONE E FAMIGLIE | SOCIETÀ E ISTITUZIONI | ISTRUZIONE E LAVORO | **ECONOMIA** | AMBIENTE E TERRITORIO

CERCA NEL SITO | Statistiche A-Z | Glossario

**INFORMAZIONI SULLA RILEVAZIONE**  
**INDAGINE MULTISCOPO SULLE FAMIGLIE: ASPETTI DELLA VITA QUOTIDIANA**

**Che cosa è**  
 L'indagine campionaria "Aspetti della vita quotidiana" fa parte di un sistema integrato di indagini sociali - le Indagini Multiscopo sulle famiglie e rileva informazioni fondamentali relative alla vita quotidiana degli individui e delle famiglie. A partire dal 1993, l'indagine viene svolta ogni anno. Le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno e se sono soddisfatti del funzionamento di quei servizi di pubblica utilità che dovrebbero contribuire al miglioramento della qualità della vita. Scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita sono i temi indagati. L'indagine rientra nel Programma statistico nazionale che comprende l'insieme delle rilevazioni statistiche necessarie al Paese. Il Programma statistico nazionale è in vigore e consultabile sul sito internet dell'Istat alla sezione [Normative](#).

**Chi risponde**  
 L'indagine è eseguita su un campione di circa 25.000 famiglie distribuite in circa 800 comuni italiani di diversa ampiezza demografica. Sono intervistati tutti gli individui appartenenti alle famiglie rientranti nel campione. Se uno di essi

PERIODO DI RIFERIMENTO: ANNO 2020  
 DATA DI PUBBLICAZIONE: 26 MARZO 2020

**ALLEGATI**  
 FACSIMILE IMF ISTAT 7\_R\_20  
 FACSIMILE IMF ISTAT 7\_A\_20  
 LETTERA INFORMATIVA ITALIANO  
 LETTERA INFORMATIVA PROVINCIA AUTONOMA DI BOLZANO  
 LETTERA INFORMATIVA SLOVENO

**Esplora Temi**

Cerca nei temi  Annulla

Tutti i temi

Censimento agricoltura 2010  
 Censimenti permanenti imprese, istituzioni pubbliche e nonprofit  
 Censimento popolazione e abitazioni 2011

- Ambiente ed energia
- Caratteristiche del territorio
- Popolazione e famiglie
- Condizioni economiche delle famiglie e disuguaglianze
- Salute e sanità
  - Stili di vita e fattori di rischio
    - Abitudini al fumo
      - **Abitudine al fumo - età dettaglio**
      - Abitudine al fumo - età, titolo di studio
      - Abitudine al fumo - posizione nella professione
      - Abitudine al fumo -

**Aspetti della vita quotidiana** : *Abitudine al fumo - età dettaglio*

Personalizza | Esporta | Grafici | La tua interrogazione

Misura: per 100 persone con le stesse caratteristiche

Sesso: totale

| Tipo dato         | persone di 14 anni e più per abitudine al fumo |             | persone di 14 anni e più fumatori che fumano sigarette |          | persone di 14 anni e più fumatori per sigarette fumate |                     |                      |                    | numero medio di sigarette al giorno |
|-------------------|--|-------------|--|----------|--|---------------------|----------------------|--------------------|-------------------------------------|
|                   | fumatori                                       | ex fumatori | non fumatori   | fumatori | fino a 5 sigarette                                     | da 6 a 10 sigarette | da 11 a 20 sigarette | oltre 20 sigarette |                                     |
| Seleziona periodo |  |             |  |          |  |                     |                      |                    |                                     |
| 14-17 anni        | 6.3  | 2.7         | 89.6   |          | 100  | 56                  | 28.6                 | 15.5               | 0                                   |
| 18-19 anni        | 19   | 6.2         | 73.6   |          | 99.6   | 42.9                | 41.1                 | 13.7               | 2.4                                 |
| 20-24 anni        | 27.7   | 10.1        | 60.7   |          | 99.9   | 38.5                | 38.3                 | 22.7               | 0.6                                 |
| 25-34 anni        | 25.1   | 16.2        | 57.8   |          | 99.3   | 28.5                | 37.7                 | 31.5               | 2.2                                 |
| 35-44 anni        | 24   | 22          | 53.6   |          | 98.6   | 27.1                | 35.5                 | 33.7               | 3.7                                 |

Dati estratti il 07 Aug 2020, 12h43 UTC (GMT) da I.Stat.

**Metadati**  
 Aspetti della vita quotidiana

Sorgente

■ Fonte(i) dei dati usata (e)

**Multiscopo sulle famiglie: aspetti della vita quotidiana - parte generale:** L'indagine campionaria "Aspetti della vita quotidiana" fa parte di un sistema integrato di indagini sociali - le Indagini Multiscopo sulle famiglie - e rileva le informazioni fondamentali relative alla vita quotidiana degli individui e delle famiglie. Dal 1993 al 2003 l'indagine è stata condotta ogni anno nel mese di novembre. Nel 2004 l'indagine non è stata effettuata e dal 2005 viene condotta ogni anno nel mese di febbraio. Le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno. Area tematica su aspetti sociali diversi si susseguono nei questionari, permettendo di capire come vivono gli individui e quanto sono soddisfatti delle loro condizioni, della situazione economica, della zona in cui vivono, del funzionamento dei servizi di pubblica utilità che dovrebbero contribuire al miglioramento della qualità della vita. Scuola, lavoro, vita familiare e di relazione, tempo libero, partecipazione politica e sociale, salute, stili di vita, accesso ai servizi sono indagati in un'ottica in cui oggettività dei comportamenti e soggettività delle aspettative, delle motivazioni, dei giudizi contribuiscono a

## RISPOSTA

In Italia nel 2018 fuma:

- ✓ il 6,3% dei ragazzi tra 14 e 17 anni
- ✓ il 19,0% dei ragazzi tra 18 e 19 anni

Fonte: Istat, Indagine multiscopo "Aspetti della vita quotidiana"

## I DATI SECONDARI – fonti di dati amministrativi

Si definiscono **fonti amministrative** le informazioni raccolte e conservate da istituzioni pubbliche (Anagrafi, Archivi Ministeri, ecc.) per la registrazione di uno stato di fatto riguardanti persone fisiche o giuridiche del territorio di competenza per finalità fiscali, pensionistiche, giuridiche o anagrafiche.

esempi: iscritti in anagrafe per nascita, matrimoni, delitti denunciati, separazioni civili, ecc.

### VANTAGGI

- Riducono il costo della rilevazione
- Riducono il disturbo statistico

### SVANTAGGI

- Concetti, definizioni, classificazioni adottate possono non coincidere
- La qualità con cui le informazioni sono raccolte nei dati amministrativi può non essere sufficiente
- Non sempre è garantita la disponibilità dei dati in tempi utili alle necessità di informazione statistica

## I DATI SECONDARI – fonti di dati privati

- l'obbligo per i **soggetti privati titolari** dei dati di mettere a disposizione i dati necessari allo sviluppo, la produzione e la diffusione di statistiche europee;
- l'istituzione di un sistema di condivisione dei dati all'interno del SSE allo scopo di sviluppare e produrre statistiche di elevata qualità.

## DOMANDA DELLA RICERCA: IN QUALE RIPARTIZIONE TERRITORIALE CI SONO PIÙ LAUREATI?

Collettivo: popolazione di 9 anni e più per grado di istruzione e ripartizione

Informazioni sul processo statistico: dati:

<https://www4.istat.it/it/censimenti-permanenti/popolazione-e-abitazioni>

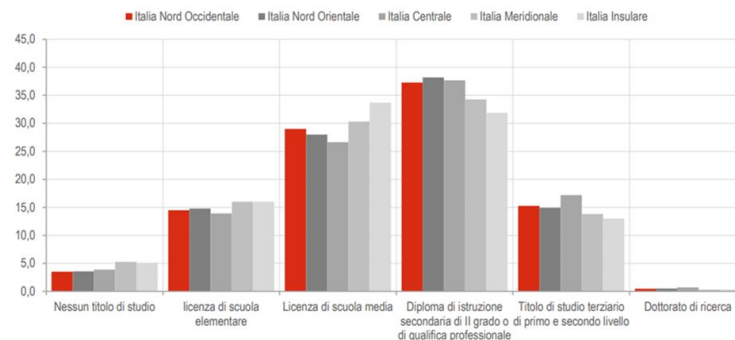
Estrazione dati: <http://dati-censimentipermanenti.istat.it/>

### Registri e principali statistiche prodotte



### Strategia del Censimento Permanente:

**Censimento = Attività che integra dati amministrativi e dati da indagare**



Fonte: Istat, Censimento e dinamica demografica, 2021

**RISPOSTA** In Italia nel 2021 i laureati sono: il 17,2% al Centro, il 15,3% al Nord-ovest, il 14,9% al Nord-est, il 13,8% nel Meridione e il 13% nelle Isole.

## I DATI PRIMARI - indagini ad hoc

Sono informazioni non esistenti, generate dal processo di ricerca (definizione): sono quelli raccolti attraverso attività in prima persona - come osservazione, registrazione, misurazione, monitoraggio di persone (e loro comportamenti), controllo di oggetti o eventi - per raggiungere obiettivi di ricerca specifici e generare nuove conoscenze. Quando vengono raccolti sistematicamente e messi a disposizione di terzi, i dati primari diventano dati secondari (approfondimento).

## LA FASE DELLA PROGETTAZIONE (1)

- ❑ **WHAT**= Fissare gli obiettivi e individuare il **fenomeno collettivo** di interesse dell'indagine (nota che più ampio l'arco degli argomenti trattati, maggiori le complessità da affrontare sul piano concettuale, statistico ed operativo);
- ❑ **WHO**= Scegliere/definire la **popolazione** (campo di osservazione) le **unità statistiche** e i **caratteri o variabili statistiche** da indagare (es. qualitative ordinali o sconnesse, quantitative discrete o continue);
- ❑ **HOW**= Definire se interessa descrivere un fenomeno nella sua **componente statica o dinamica** e se si vuole confrontare i risultati con informazioni relative ad **altre realtà territoriali**;
- ❑ **WHEN - WHERE**= Delimitare il **periodo di riferimento** e l'**area geografica** a cui ci si riferisce (limiti spazio-temporali).



## LA FASE DELLA PROGETTAZIONE (2)

Come si vuole indagare? Indagine totale o campionaria? Con quale tecniche e strumenti di rilevazione?

- ❑ Definizione del **disegno di indagine** ossia specificazione del tipo di indagine per la raccolta delle informazioni e la produzione delle relative statistiche
- ❑ Individuazione della popolazione e della lista delle unità statistica
  - ❖ **rilevazione totale o esaustiva (censimento)** si ha la conoscenza esatta del fenomeno studiato: viene rilevata tutta la popolazione di interesse.
  - ❖ **rilevazione parziale (campione)** si perviene a una **stima** (più o meno precisa) del fenomeno studiato: viene osservata una parte o campione della popolazione di interesse.
    - Definizione del piano di campionamento (lista, schema, errori: correttezza, efficienza, consistenza)

### Importante distinzione:

#### Campioni probabilistici:

- è noto l'insieme dei possibili campioni;
- è nota la probabilità di selezione di ciascun campione

#### Campioni non probabilistici

- Tutti gli altri

## TIPI DI RILEVAZIONE A CONFRONTO

### INDAGINE TOTALE

#### VANTAGGI

- misura reale (non affetta da errore campionario) della popolazione
- informazioni rilevate base di riferimento per studi successivi
- elevato dettaglio di analisi

#### SVANTAGGI

- tempi lunghi per rilascio delle informazioni
- costi elevati sia in termini di risorse che economici
- errore non campionario

### INDAGINE CAMPIONARIA

#### VANTAGGI

- riduzione dei costi
- maggiore rapidità
- maggiore accuratezza (<dimensioni)
- assenza di alternative (popolazioni illimitate)

#### SVANTAGGI

- errore campionario
- discriminazioni all'interno della popolazione
- eccessiva dimensione del campione in caso di eventi rari

## Indicatori di qualità' sullo Stato della lista e sulla Rilevazione

Indicatori calcolati a partire dalle informazioni acquisite durante le prime fasi di rilevazione; forniscono una misura indiretta della qualità delle liste di riferimento.

### liste e rilevazione: definizioni

#### TOTALE UNITÀ = NUMERO COMPLESSIVO DELLE UNITÀ OGGETTO DI INDAGINE

Approfondimenti e Esempi

Per le indagini campionarie coincide con il numero di unità campionate.

Numero totale di famiglie residenti in Italia

Numero totale di studenti iscritti all'anagrafe MIUR

Numero totale di imprese iscritte alla Camera di Commercio

.....

## Indicatori di qualità sulla lista

### TASSO DI UNITÀ RISOLTE

Capacità dell'indagine di identificare le unità eleggibili, serve come denominatore per gli altri indicatori, dipende dalla tecnica

### TASSO DI ERRORI DI LISTA

Qualità della lista in termini di errate inclusioni (unità non eleggibili) e errate informazioni (errori di lista che pregiudicano il contatto)

### TASSO DI UNITÀ NON ELEGGIBILI

- Tasso di unità non più esistenti
- Tasso di unità con variazioni di stato
- Tasso di unità fuori target
- Tasso di unità non contattate per errore di lista

## Indicatori di qualità sulla Rilevazione

### TASSO DI MANCATA RISPOSTA TOTALE (MRT)

- al lordo delle unità non risolte
- al netto degli errori di lista che pregiudicano il contatto
- riferito alle unità eleggibili accertate e correttamente incluse nel campione/popolazione

### TASSO DI RIFIUTO

#### TASSO DI MANCATO CONTATTO

- per errori di lista
- per altri motivi

#### TASSO DI MANCATA RISPOSTA PER ALTRI MOTIVI

MRT dovuto a motivi diversi dal rifiuto e mancato contatto

#### • TASSO DI RISPOSTA

Complementare del Tasso MRT

## LA FASE DELLA PROGETTAZIONE (3)

- ❑ Promuovere la partecipazione alla rilevazione delle unità di rilevazione (famiglia, impresa,...)
- ❑ Scelta della **tecnica di indagine** (contatto) più idonea a raccogliere le informazioni
- ❑ Formazione dei **rilevatori**
- ❑ Formulazione del **questionario** e pre - test
- ❑ Concettualizzazione (relazione e gerarchie), individuazione aree e sotto-aree, contenuti, formulazione e sequenza domande
- ❑ Predisposizione di meccanismi di **controllo**

- Interviste telefoniche – **CATI (Computer Assisted Telephone Interview)**
- Interviste face to face – **PAPI (Paper And Pen Interview)**
- Interviste personali con PC – **CAPI (Computer Assisted Personal Interview)**
- Questionari auto-compilati su PC – **CASI (Computer Assisted Self Interview)**
- Interviste on-line – **CAWI (Computer Assisted Web Interview)**

+  
↑  
↓  
-

| Strutturazione  | Standardizzazione  | Direttività  |
|---|--|--|
| Gli argomenti sono specificati in maniera dettagliata | Le stesse domande sono proposte nello stesso ordine a tutti gli intervistati | L'intervistato sceglie tra alternative di risposta prefissate            |
| Gli argomenti sono definiti in maniera generale       | Gli argomenti sono proposti a seconda dell'andamento della conversazione     | L'intervistato è libero di formulare la propria risposta come preferisce |

FONTE: MERAVIGLIA (2005: 159)

Spesso sono usati due o più modi di raccolta dei dati per: risparmiare, aumentare il tasso di risposta, ridurre gli errori di misura, ecc.

| Fasi          | Operazioni              | Fonti errore                                 | Tipo errore                        |
|---------------|-------------------------|--|------------------------------------|
| Progettazione | scelta unità, variabili | modello concettuale                          | rilevanza teorica                  |
|               | redazione questionario  | struttura del vocabolario, quesiti, codifica | errori di misura                   |
|               | diffusione dei dati     |  | rilevanza effettiva<br>trasparenza |

**Sistema di controlli dell'errore affiancati alla fase di progettazione:** invio di lettera preavviso ai rispondenti; istituzione di un numero verde per chiarimenti; selezione e formazione rilevatori; analisi completezza e ridondanza delle liste utilizzate; monitoraggio mancate risposte, indagini ad hoc su non rispondenti, ect.

## LA FASE DI RACCOLTA: RILEVAZIONE E REGISTRAZIONE DEI DATI

- ❑ **Raccogliere**, senza influenzare il rispondente, le informazioni utilizzando le diverse modalità
- ❑ **Registrare** i dati riportando le risposte dei questionari (strumento di osservazione) su supporto informatico:
  - in alcuni casi lettura ottica dei questionari
  - ricodifica quesiti aperti in codici standard (es. ATECO)
  - non è necessaria quando la rilevazione è in CATI, CAPI, CAWI

**Sistema di controlli dell'errore affiancati alla fase di rilevazione** : buona impressione per contatti futuri, uso programmi per registrazione controllata dei dati; applicazione tecniche di identificazione automatica di incoerenze nei dati (es. un professionista con la sola licenza elementare); correzioni dati con valori accettabili mediante re intervista; calcolo indicatore di qualità (tassi di risposta), ect.

| Fasi                  | Operazioni                  | Fonti errore                            | Tipo errore                                       |
|-----------------------|-----------------------------|---|---|
| Rilevazione sul campo | formazione elenchi di unità | liste supervisor<br>rilevatori          | errori di misura (selez. da mancata copertura)    |
|                       | raccolta dati               | supervisor<br>rilevatori<br>rispondenti | errori di misura (mancate risposte, incongruenze) |
|                       | registrazione               | operatori                               | errori di misura                                  |

## DALLA MATRICE DEI DATI .....

I dati raccolti con le varie tecniche di rilevazione devono essere organizzati in una forma che ne permetta un'agevole analisi. A tal fine è pressoché indispensabile usare la matrice dei dati, definita anche matrice “**casi per variabili**” perché nella tabella ogni **riga** rappresenta un caso, ogni **colonna** una variabile (una proprietà) e ogni **cella** il valore rilevato per ogni caso su ogni variabile.

Uno, pochi, molti casi.....la matrice si complica quanto più grande è il numero n di osservazioni.

## .....ALLA TASSONOMIA DEL CONCETTO

- impiego della matrice dei dati;
- presenza di definizioni operative dei “modi” della matrice dei dati (perlopiù casi e variabili);
- impiego della statistica e dell'analisi dei dati.

Qualitativa nominale

Qualitativa ordinale

| ID | CORSO LAUREA | SESSO | MEDIA VOTI | CREDITI | RENDIMENTO |
|----|--------------|-------|------------|---------|------------|
| 1  | SAM          | M     | 22         | 6       | discreto   |
| 2  | SAM          | F     | 24         | 71      | buono      |
| 3  | SAM          | M     | 21         | 19      | discreto   |
| 4  | SAM          | F     | 26         | 27      | buono      |
| 5  | SAM          | F     | 27         | 9       | ottimo     |
| 6  | SAM          | M     | 26         | 10      | buono      |
| 7  | SAM          | F     | 25         | 18      | buono      |
| 8  | SAM          | M     | 24         | 27      | buono      |
| 9  | SAM          | F     | 27         | 10      | ottimo     |
| 10 | SAM          | F     | 24         | 17      | buono      |
| 11 | SAM          | M     | 26         | 18      | buono      |
| 12 | SAM          | M     | 30         | 18      | ottimo     |
| 13 | SAM          | F     | 29         | 84      | ottimo     |
| 14 | SPO          | M     | 27         | 27      | ottimo     |
| 15 | SPO          | F     | 23         | 9       | discreto   |
| 16 | SPO          | F     | 27         | 30      | ottimo     |
| 17 | SPO          | M     | 28         | 33      | ottimo     |
| 18 | SPO          | M     | 29         | 30      | ottimo     |
| 19 | SPO          | F     | 28         | 48      | ottimo     |
| 20 | ORU          | F     | 26         | 66      | buono      |

Quantitativa continua  
(è una media!)

Quantitativa discreta (deriva  
da un conteggio)

## LA FASE DI ELABORAZIONE DEI RISULTATI

Nella fase di **elaborazione** si applicano strumenti propri dell'analisi statistica (metodi inferenziali-induttivi) al fine di ottenere una sintesi e una descrizione dei dati sotto forma di:

- ❑ **Frequenze** (assolute, relative, percentuali, cumulate) e **tabelle** (semplici e a doppia entrata)
- ❑ **Grafici** (a torta, a barre, istogramma)
- ❑ **Indicatori** (media, mediana e moda)



La scelta del metodo da usare dipende dal tipo di fenomeno osservato (quantitativo o qualitativo).

**Sistema di controlli dell'errore affiancati alla fase:** analisi descrittive e esplorative per individuare incongruenze (valori anomali ed errori); ricerca di relazioni; analisi confermative o verifica di ipotesi; indicatori di qualità (varianza), ect.

| Fase              | Operazioni   | Fonti errore                 | Tipo errore                                 |
|-------------------|--------------|------------------------------|---|
| Elaborazione dati | elaborazione | programmi<br>base statistica | errori di calcolo<br>rilevanza<br>effettiva |
|                   | diffusione   |                              | tempestività                                |

## FREQUENZE E TABELLE

### Distribuzioni di frequenza semplice

Una tabella di frequenza associa alle modalità della variabile X, qualitativa o quantitativa, il numero di volte ni in cui ciascuna modalità si manifesta n1, ..., nk

n = numero delle unità statistiche rilevate

X = carattere oggetto di studio

K = numero totale dei diversi valori assunti dalla variabile X (modalità)

xi = modalità i-esima del carattere X i=1, ..., k

ni = frequenze assolute

Per caratteri quantitativi continui:

- Si raggruppano i valori in intervalli (classi);
- Gli intervalli non devono essere troppi né troppo pochi;
- L'ampiezza degli intervalli è preferibile che sia uguale per poter facilitare il confronto tra classi.

| MODALITA' | FREQUENZE ASSOLUTE | FREQUENZE RELATIVE | FREQUENZE PERCENTUALI | FREQUENZE CUMULATE |
|-----------|--------------------|--------------------|-----------------------|--------------------|
| $x_i$     | $n_i$              | $f_i$              | $p_i$                 | $N_i$              |
| $x_1$     | $n_1$              | $n_1/n=f_1$        | $f_1*100$             | $n_1$              |
| $x_2$     | $n_2$              | $n_2/n=f_2$        | $f_2*100$             | $n_1+n_2$          |
| $x_3$     | $n_3$              | $n_3/n=f_3$        | $f_3*100$             | $n_1+n_2+n_3=n$    |
|           | $n$                | 1                  | 100                   |                    |

$N_i = \sum_{j=1}^i n_j$   
 $\left\{ \begin{array}{l} N_1 = n_1 \\ N_k = n \\ N_i - N_{i-1} = n_i \end{array} \right.$

## FREQUENZE IN SINTESI

### FREQUENZA ASSOLUTA ( $n_i$ )

numero di osservazioni corrispondente ai diversi valori (modalità/intervalli di classe) della variabile

### FREQUENZA RELATIVA: ( $p_i = n_i / n$ )

rapporto tra il numero di osservazioni corrispondente ai diversi valori (modalità/intervalli di classe) della variabile e il totale delle osservazioni. Vantaggio: confrontare distribuzioni di frequenza basate su numeri differenti di unità statistiche.

### FREQUENZA RELATIVA PERCENTUALE: ( $p_i \% = n_i / n * 100$ )

quanto volte un fenomeno si manifesta su 100 osservazioni

### FREQUENZA CUMULATA ASSOLUTA E RELATIVA: ( $F_i$ ); $P_i = F_i / n$ ; $P_i \% = F_i / n * 100\%$ )

numero di osservazioni il cui valore (o la cui %) è inferiore o uguale ad una data modalità o a un dato valore  $x_i$

$$0 \leq n_i \leq n$$

$$\sum_{i=1}^K n_i = n_1 + n_2 + \dots + n_K = n$$

$n$  = numero totale delle osservazioni

$K$  = numero dei valori/modalità/classi della variabile

$$0 \leq p_i \leq 1$$

$$\sum_{i=1}^K p_i = p_1 + p_2 + \dots + p_K = 1$$

$$0 \leq p_i \% \leq 100\%$$

$$\sum_{i=1}^K p_i \% = p_1\% + p_2\% + \dots + p_K\% = 100\%$$

## TABELLE DI FREQUENZA PER UNA VARIABILE DISCRETA

N.ro di libri letti in un mese da 45 studenti

Passi da seguire

- identificare il valore minimo e quello massimo (nell'esempio 0 e 9 libri)
- contare quante volte compare ogni valore (modalità) cioè quante sono le  $n_i$  con uguale numero di  $x_i$

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 6 | 3 | 4 | 7 | 2 | 3 | 2 | 3 | 2 | 6 | 4 | 3 | 9 | 3 |
| 2 | 0 | 3 | 3 | 4 | 6 | 5 | 4 | 2 | 3 | 6 | 7 | 3 | 4 | 2 |
| 5 | 1 | 3 | 4 | 3 | 7 | 0 | 2 | 1 | 3 | 1 | 5 | 0 | 4 | 5 |

$n=45$       $X = (x_1, x_2, \dots, x_{45}) = (5, 6, 3, 4, \dots, 5, 0, 4, 5)$

| X             | Frequenza assoluta $n$ | Frequenza relativa | Frequenza relativa cumulata |
|---------------|------------------------|--------------------|-----------------------------|
| 0             | 3                      | 6,7%               | 6,7%                        |
| 1             | 3                      | 6,7%               | 13,3%                       |
| 2             | 7                      | 15,6%              | 28,9%                       |
| 3             | 12                     | 26,7%              | 55,6%                       |
| 4             | 7                      | 15,6%              | 71,1%                       |
| 5             | 5                      | 11,1%              | 82,2%                       |
| 6             | 4                      | 8,9%               | 91,1%                       |
| 7             | 3                      | 6,7%               | 97,8%                       |
| 8             | 0                      | 0,0%               | 97,8%                       |
| 9             | 1                      | 2,2%               | 100,0%                      |
| <b>Totale</b> | <b>45</b>              | <b>100,0%</b>      |                             |

### PERCHÉ USARE LE FREQUENZE RELATIVE?

Per il confronto della distribuzione di una variabile in campioni di dimensioni diverse.

### PERCHÉ USARE LE FREQUENZE CUMULATE?

Per misurare il numero totale di osservazioni inferiore (o superiore) ad un valore prefissato (ad es.: il 71% degli studenti legge meno di 5 libri in un mese ; il 56% al massimo 3).

## TABELLA DI FREQUENZA PER UNA VARIABILE CONTINUA

Peso in kg di 40 studenti

Passi da seguire

- identificare il valore minimo e quello massimo (nell'esempio 60 e 200 kg)
- raggruppare in classi (arbitrario nel numero e nell'ampiezza) che comprendano più modalità

campo di variazione =  $\text{Range} = X_{\max} - X_{\min} = 140$

numero degli intervalli =  $k = 7$

ampiezza degli intervalli =  $\delta_i = \text{Range} / k = 140/7 = 20$

- contare quante volte compare ogni valore (intervalli di classe) (cioè quante sono le  $n_i$  con uguale classe)

**N.B.** ogni unità statistica deve essere assegnata ad un'unica modalità o intervallo di classe

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 107 | 83  | 100 | 128 | 143 | 127 | 117 | 125 | 64  | 119 |
| 98  | 111 | 119 | 130 | 170 | 143 | 156 | 126 | 113 | 127 |
| 130 | 120 | 108 | 95  | 192 | 124 | 129 | 143 | 198 | 131 |
| 163 | 152 | 104 | 119 | 161 | 178 | 135 | 146 | 158 | 176 |

$n=40$

$X = (x_1, x_2, \dots, x_{40}) = (107, 83, 100, \dots, 146, 158, 176)$

| $X_i$     | $n_i$ | $f_i$ | $F_i$ |
|-----------|-------|-------|-------|
| [60-80)   | 1     | 2.5   | 2.5   |
| [80-100)  | 3     | 7.5   | 10.0  |
| [100-120) | 10    | 25.5  | 35.0  |
| [120-140) | 12    | 30.0  | 65.0  |
| [140-160) | 7     | 17.5  | 82.5  |
| [160-180) | 5     | 12.5  | 95.0  |
| [180-200) | 2     | 5.0   | 100.0 |

## FREQUENZE E TABELLE A DOPPIA ENTRATA

Si parla di Analisi Bivariata quando su ogni unità statistica, appartenente ad una determinata popolazione, si rilevano due caratteri X e Y.

Distribuzione bivariata doppia o congiunta di X e Y (frequenze assolute)

| X\Y   | $d_1$    | $d_2$    | ... | $d_l$    | ... | $d_h$    |          |
|-------|----------|----------|-----|----------|-----|----------|----------|
| $c_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1l}$ | ... | $n_{1h}$ | $n_{1.}$ |
| $c_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2l}$ | ... | $n_{2h}$ | $n_{2.}$ |
| ·     | ·        | ·        | ·   | ·        | ·   | ·        | ·        |
| $c_j$ | $n_{j1}$ | $n_{j2}$ | ... | $n_{jl}$ | ... | $n_{jh}$ | $n_{j.}$ |
| ·     | ·        | ·        | ·   | ·        | ·   | ·        | ·        |
| $c_k$ | $n_{k1}$ | $n_{k2}$ | ... | $n_{kl}$ | ... | $n_{kh}$ | $n_{k.}$ |
|       | $n_{.1}$ | $n_{.2}$ | ... | $n_{.l}$ | ... | $n_{.h}$ | $n$      |

$$n_{.1} = \sum_{j=1}^k n_{j1}, \quad \dots, \quad n_{.l} = \sum_{j=1}^k n_{jl}, \quad \dots, \quad n_{.h} = \sum_{j=1}^k n_{jh}$$

$$n_{1.} = \sum_{l=1}^h n_{1l}, \quad \dots, \quad n_{j.} = \sum_{l=1}^h n_{jl}, \quad \dots, \quad n_{k.} = \sum_{l=1}^h n_{kl}$$

$$n = \sum_{j=1}^k n_{j.} = \sum_{l=1}^h n_{.l} = \sum_{j=1}^k \sum_{l=1}^h n_{jl}$$

- la distribuzione di frequenza congiunta delle due variabili;
- le due distribuzioni marginali da X e della Y;
- le k distribuzioni condizionate della Y|c<sub>j</sub> (con j = 1, 2, ..., k);
- le h distribuzioni condizionate della X|d<sub>l</sub> (con l = 1, 2, ..., h).

## FREQUENZE IN SINTESI

### FREQUENZA CONGIUNTA ASSOLUTA

numero delle volte con cui la coppia di modalità ( $x_i$ ,  $y_j$ ) si presenta, ovvero la frequenza con la quale, su di un'unità statistica, il carattere X assume la modalità  $x_i$  e contemporaneamente il carattere Y assume la modalità  $y_j$ .

$$0 \leq n_{ij} \leq n$$

$$\sum_{ij} n_{ij} = n$$

$n$  = numero totale delle osservazioni

### FREQUENZA MARGINALE per riga: (: riferita alla riga $i$ -ma)

frequenza della modalità  $i$ -ma del carattere X per riga, senza tener conto delle modalità dell'altro carattere Y.

somma lungo le righe

### FREQUENZA MARGINALE per colonna: (: riferita alla colonna $j$ -ma)

frequenza della modalità  $j$ -ma del carattere Y per colonna, senza tener conto delle modalità dell'altro carattere X.

somma lungo le colonne

### FREQUENZA CONDIZIONATA (carattere CONDIZIONATO $X/Y=y_j$ e carattere CONDIZIONATO $Y/X=x_i$ )

Frequenza della modalità  $i$ -ma del carattere X ( $j$ -ma del carattere Y) condizionata ad una modalità dell'altro carattere Y (X). Cioè come varia X data una modalità di Y (e viceversa).

## TABELLE DI FREQUENZA A DOPPIA ENTRATA (variabili discrete)

Nella tabella sono riportati i dati di 1000 famiglie secondo la variabile  $X$  = numero di auto possedute dalla famiglia e  $Y$  = numero di componenti della famiglia.

| X | Y   |     |     |     |     |      |
|---|-----|-----|-----|-----|-----|------|
|   | 1   | 2   | 3   | 4   | 5   |      |
| 0 | 10  | 20  | 20  | 150 | 50  | 250  |
| 1 | 85  | 85  | 330 | 50  | 50  | 600  |
| 2 | 5   | 85  | 10  | 0   | 0   | 100  |
| 3 | 0   | 10  | 40  | 0   | 0   | 50   |
|   | 100 | 200 | 400 | 200 | 100 | 1000 |

Distribuzioni condizionate

| X/Y=2 | frequenze |          |
|-------|-----------|----------|
|       | assolute  | relative |
| 0     | 20        | 0,100    |
| 1     | 85        | 0,425    |
| 2     | 85        | 0,425    |
| 3     | 10        | 0,050    |
|       | 200       | 1,000    |

| Y/X=1 | frequenze |          |
|-------|-----------|----------|
|       | assolute  | relative |
| 1     | 85        | 0,14167  |
| 2     | 85        | 0,14167  |
| 3     | 330       | 0,55000  |
| 4     | 50        | 0,08333  |
| 5     | 50        | 0,08333  |
|       | 600       | 1        |

Distribuzioni univariate:

| X | $f(x)$ |
|---|--------|
| 0 | 250    |
| 1 | 600    |
| 2 | 100    |
| 3 | 50     |
|   | 1000   |

| Y | $f(y)$ |
|---|--------|
| 1 | 100    |
| 2 | 200    |
| 3 | 400    |
| 4 | 200    |
| 5 | 100    |
|   | 1000   |

Distribuzioni univariate relative:

| X | $f^R(x)$ |
|---|----------|
| 0 | 0,25     |
| 1 | 0,60     |
| 2 | 0,10     |
| 3 | 0,05     |
|   | 1,00     |

| Y | $f^R(y)$ |
|---|----------|
| 1 | 0,1      |
| 2 | 0,2      |
| 3 | 0,4      |
| 4 | 0,2      |
| 5 | 0,1      |
|   | 1,0      |

## TABELLE DI FREQUENZA PER DUE VARIABILI (variabili continua/discreta)

In un collettivo di 12.268.000 giovani si è osservato la frequenza di pratica sportiva per classi di età ottenendo la seguente distribuzione di frequenze relative percentuali

Quanti sono i giovani che praticano sport in modo saltuario e che hanno un'età superiore a 19 anni e non superiore a 24 anni?

| Classe di età | 2021                    |                      |                                 |  |
|---------------|-------------------------|----------------------|---------------------------------|--|
|               | praticano sport         |                      |                                 | non praticano sport,<br>né attività fisica |
|               | in modo<br>continuativo | in modo<br>saltuario | solo qualche<br>attività fisica |  |
| 15-17 anni    | 42,3                    | 12,9                 | 25                              | 19,9                                       |
| 18-19 anni    | 37,4                    | 16,2                 | 25                              | 21,4                                       |
| 20-24 anni    | 38,2                    | 16,1                 | 24,3                            | 21,4                                       |
| 25-34 anni    | 32,8                    | 15,3                 | 29                              | 22,8                                       |

## DIAGRAMMA A TORTA: (dati qualitativi o nominali)

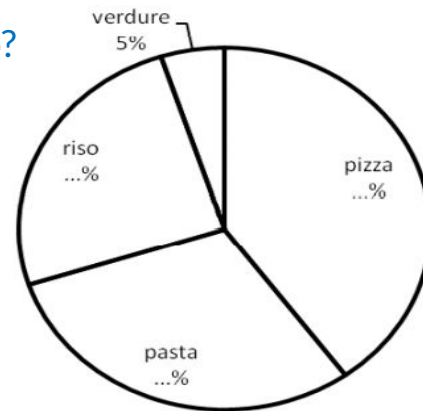
E' un cerchio diviso in tanti settori circolari (spicchi) di dimensioni proporzionali ai valori che esse rappresentano: intero rappresenta il 100% in un angolo giro di 360°. Le diverse modalità sono rappresentate da uno spicchio della torta. L'angolo al centro è proporzionale alla frequenza relativa di quella modalità:

$$angolo = 360 \cdot \text{frequenza assoluta}/n$$

Quanti sono gli studenti che hanno indicato la pizza come cibo preferito?

**ESEMPIO** Su un gruppo di studenti intervistato sul cibo preferito, le verdure è risultato il cibo preferito dal 5% degli intervistati

| Cibo preferito | Studenti |
|----------------|----------|
| Pizza          | .....    |
| Pane           | 79       |
| Riso           | 43       |
| Verdura        | 12       |



Se il totale degli studenti intervistati è  $5\%:12=100\%:x$  ossia  $12 \times 100 : 5 = 240$ , quelli che preferiscono la pizza sono  $240 - 79 - 43 - 12 = 106$ .

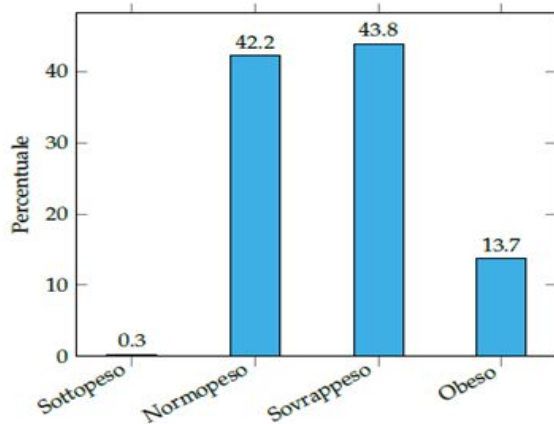
### DIAGRAMMA A BARRE O NASTRI: (dati qualitativi o quantitativi discreti)

E' costituito da tanti rettangoli (**barre**) quante sono le modalità del carattere, ponendo sull'asse delle ascisse le modalità della variabile X (**le basi uguali**) e sulle ordinate le frequenze (assolute o relative) corrispondenti ad ogni modalità (**le altezze**).

- L'impiego delle frequenze assolute o relative non cambia la forma della distribuzione.
- Il ricorso alle frequenze relative è necessario se si vogliono confrontare due diverse distribuzioni.

**ESEMPIO** grafico delle frequenze relative percentuali ( $pi \% = ni / n * 100$ ) dello stato di salute della popolazione italiana classificata in quattro gruppi: sottopeso, normopeso, sovrappeso e obeso.

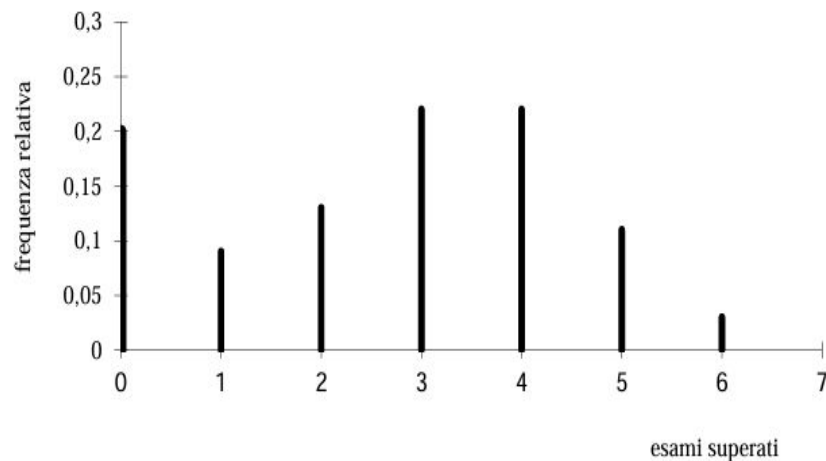
Quante sono le persone in sovrappeso?



Fonte prove Invalsi 2011

Grafico del numero di esami superati dagli iscritti al primo anno di un certo corso di laurea nel momento dell'iscrizione all'anno successivo

| Esami superati | Frequenza relativa |
|----------------|--------------------|
| 0              | 0.20               |
| 1              | 0.09               |
| 2              | 0.13               |
| 3              | 0.22               |
| 4              | 0.22               |
| 5              | 0.11               |
| 6              | 0.03               |
|                | 1.00               |



## ISTOGRAMMA: (dati quantitativi continui)

E' composto da un insieme di rettangoli tra loro adiacenti.

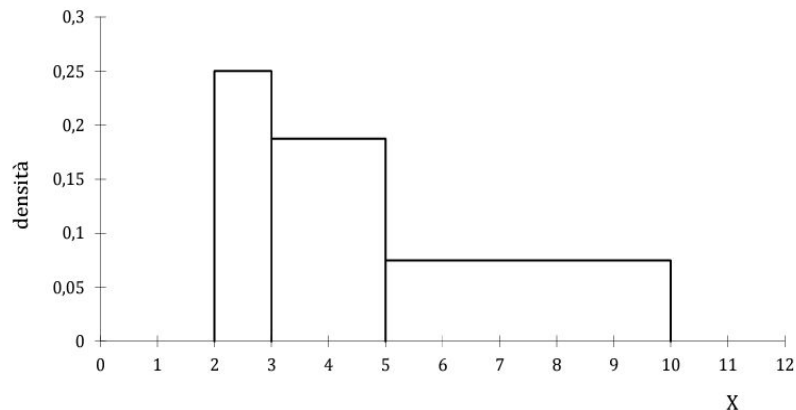
Il numero dei rettangoli corrisponde al numero di classi in cui è stata suddivisa la variabile.

Le basi dei rettangoli affiancate devono coprire l'intera gamma dei valori della variabile.

La larghezza della base dei rettangoli dipende dall'ampiezza di tali classi (differenza tra l'estremo superiore e l'estremo inferiore della stessa).

L'altezza di ogni rettangolo indica la densità dei casi presenti in ogni classe ed è data dal rapporto tra la frequenza (assoluta o relativa) e l'ampiezza della classe. L'area di ciascun rettangolo è proporzionale alla frequenza.

| X      | Frequenza | Frequenza relativa | Ampiezza | Densità |
|--------|-----------|--------------------|----------|---------|
| 2 - 3  | 4         | 0.250              | 1        | 0.2500  |
| 3 - 5  | 6         | 0.375              | 2        | 0.1875  |
| 5 - 10 | 6         | 0.375              | 5        | 0.0750  |
|        | 16        | 1.000              |          |         |



### GRAFICI IN SINTESI

Le forme che possono assumere i grafici sono molto diverse fra loro e variano a seconda della natura della variabile considerata, nel senso che alcune rappresentazioni grafiche sono idonee per certi tipi di variabile ma non per altri.

La rappresentazione più usata è

- il diagramma a torta o diagramma a barre per variabili categoriali;
- il diagramma a barre per variabili discrete;
- l'istogramma per variabili continue.

Le rappresentazioni grafiche delle distribuzioni di frequenza forniscono:

- una immagine della distribuzione dei dati (barre più alte rappresentano modalità più frequenti);
- aiutano a descrivere la forma della distribuzione dei dati;
- sono fortemente comunicative.

## ❑ **Moda o valore modale (variabile qualitativa sconnessa o ordinabile)**

**DEF.** è la modalità della variabile alla quale è associata la **maggior frequenza**, cioè quella che si manifesta più volte in sede di rilevazione. Può essere calcolata per qualsiasi tipo di variabile. In una distribuzione è possibile individuare un solo valore modale (unimodale); possono esistere due valori che compaiono entrambi con la frequenza massima (distribuzione bimodale); ecc.

In una distribuzione relativa a una variabile quantitativa continua si ha la classe modale corrispondente all'intervallo che presenta la densità di frequenza più elevata.

## ❑ **Mediana (variabile quantitativa discreta o continua)**

**DEF.** è la modalità (valore) che occupa la **posizione centrale** nella distribuzione ordinata della variabile (preceduto e seguito dallo stesso numero di osservazioni). Può essere calcolata per variabili almeno su scala ordinale (non su variabili sconnesse).

Per calcolare la mediana di  $n$  dati:

- si ordinano gli  $n$  di dati in ordine crescente o decrescente;
- se il numero di dati è dispari la mediana corrisponde al valore centrale, ovvero al valore che occupa la posizione  $(n + 1) / 2$ .
- se il numero  $n$  di dati è pari, la mediana è stimata utilizzando i due valori che occupano le posizioni  $(n / 2)$  e  $(n / 2 + 1)$  (generalmente si sceglie la semisomma dei due valori mediani).

## Media (variabile quantitativa discreta o continua)

**DEF.** è la somma dei valori osservati divisi per il numero complessivo di valori (solo per variabili quantitative).

Se le modalità  $x_i$  si presentano ciascuna con frequenze assolute  $n_i$ , si calcola come il rapporto tra la somma dei prodotti di ciascun valore della variabile  $X$  per la rispettiva frequenza e la somma totale delle frequenze.

$$M(X) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

$$M(X) = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k x_i n_i}{\sum_{i=1}^k n_i}$$

Se la distribuzione è in classi, è necessario sintetizzare ciascuna classe mediante il suo valore centrale; si calcola la media aritmetica come nei casi precedenti, utilizzando il valore centrale della classe.

### Alcune proprietà della media

□ Sintetizza il carattere di un collettivo statistico lasciando invariato l'ammontare totale del carattere stesso, vale a dire che può essere sostituito ai dati osservati senza farne variare la somma.

$$x_1 + x_2 + \dots + x_n = M(x) + M(x) + \dots + M(x)$$

$$x_1 n_1 + x_2 n_2 + \dots + x_n n_n = M(x) n_1 + M(x) n_2 + \dots + M(x) n_n$$

□ E' compreso tra il più piccolo e il più grande valore dei dati osservati.

$$x_1 \leq M(x) \leq x_n$$

□ La somma algebrica degli scarti è uguale a 0.

$$\sum (x_i - M(x)) = 0$$

La tabella riporta la distribuzione delle frequenze congiunte degli occupati (15 anni e oltre), relative alla provincia autonoma di Trento, per settore di attività economica e posizione nella professione (anno 2020):

Qual è l'indice in grado di rappresentare il “settore di attività economica” degli occupati?

| Settore di attività economica     | 2020           |               | Totale         |
|-----------------------------------|----------------|---------------|----------------|
|                                   | Posizione      |               |                |
|                                   | Dipendenti     | Indipendenti  |                |
| Agricoltura, silvicoltura e pesca | 2.666          | 6.217         | 8.883          |
| Industria                         | 46.511         | 12.535        | 59.046         |
| Servizi                           | 141.402        | 27.187        | 168.589        |
| <b>Totale</b>                     | <b>190.579</b> | <b>45.939</b> | <b>236.518</b> |

E' il valore che compare più frequentemente (la moda) della distribuzione di frequenza “settore di attività economica” che è “Servizi”, in quanto è il settore con più occupati.

L'indagine Istat su viaggi e vacanze ha rilevato che nel 2018 il numero di persone (in migliaia) che non sono andate in vacanza per motivi economici segue la seguente distribuzione per età (in anni):

Qual è la classe di età mediana?

| Classe di età | Frequenze | Frequenze cumulate |
|---------------|-----------|--------------------|
| 15-24         | 1502      | 1502               |
| 25-34         | 1735      | 3237               |
| 35-44         | 2134      | 5371               |
| 45-54         | 2678      | 8049               |
| 55-64         | 2221      | 10270              |
| 65 e più      | 2889      | 13159              |

L'ultimo valore della colonna della frequenza cumulata è la somma delle frequenze di tutte le classi di età che è pari a 13159. Una volta trovata la somma delle frequenze (13159) si divide per due per ottenere la frequenza mediana ( $13159/2 = 6579,5$ ). Poi si cerca nella tabella in quale frequenza cumulata è compresa la frequenza mediana.

In questo caso, la frequenza mediana (6579,5) si trova nella terza classe dato che 6579,5 è un valore compreso tra le frequenze cumulate 5371 e 8049.

Di seguito i voti in matematica di una classe di 20 alunni: 5,3,6,6,7,7,8,5,8,7,6,4,4,5,3,3,7,7,8,6

## Qual è il voto medio?

- Procedimento 1       $M(x) = \frac{5+3+6+6+7+7+8+5+8+7+6+4+4+5+3+3+7+7+8+6}{20} = 5,75$
- Procedimento 2      Se ordiniamo i valori in una tabella:

$$M(x) = \frac{115}{20} = 5,75$$

| Voti   | Alunni |           |
|--------|--------|-----------|
| $x_i$  | $n_i$  | $x_i n_i$ |
| 3      | 3      | 9         |
| 4      | 2      | 8         |
| 5      | 3      | 15        |
| 6      | 4      | 24        |
| 7      | 5      | 35        |
| 8      | 3      | 24        |
| Totale | 20     | 115       |

## LA FASE DI REVISIONE E VALIDAZIONE

La **revisione** consiste nel:

- ❑ valutare la congruità delle informazioni raccolte con le finalità dell'indagine;
- ❑ individuare le **fonti di errore** più rilevanti analizzando i dati e correggendo gli errori;
- ❑ predisporre modifiche al processo di produzione in modo da ridurre gli effetti degli errori in successive occasioni di indagine;
- ❑ validare i dati corretti (se la **qualità dei dati** è buona (es. tempestività ai fini della diffusione) dell'informazione agli utenti.



**Sistema di controlli dell'errore affiancati alla fase:** procedure manuali, automatiche e/o interattive per controlli logico/formali (regole di compatibilità) su campo di variazione variabili, relazioni fra variabili e norme di compilazione dei questionari, ect.

| Fase      | Operazioni    | Fonti errore                        | Tipo errore                                  |
|-----------|---------------|-------------------------------------|--|
| Revisione | registrazione | operatori                           | errori di misura                             |
|           | revisione     | revisori<br>programmi<br>automatici | errori di misura<br>(possibile<br>selezione) |

## Revisione: esempio di casi possibili

**VALORI IMPUTATI = VALORI SU CUI SI È PROCEDUTO CON REVISIONE MEDIANTE REGOLE DI IMPUTAZIONE (VALORI CANCELLATI, DA BLANK A NON BLANK, ...)**

**VALORI NON IMPUTATI = VALORI NON TRASFORMATI DALLE PROCEDURE DI IMPUTAZIONE**

## Indicatori di qualità sulla revisione

Derivano dal confronto tra dati grezzi e dati puliti. Forniscono una misura di quanto i dati sono stati imputati, ma non di come e di quali effetti ciò abbia avuto sulle distribuzioni.

- Tasso di imputazione (Misura della quantità di interventi sui dati)

Approfondimenti sul tipo di intervento: da codice a codice, da blank a codice, da codice a blank, valori non blank, numero di variabili con tasso di imputazione superiore al 2% e al 5%, ect.

Nota: Il tasso di valori non blank imputati rappresenta la quantità di informazione di buona qualità a livello di dato grezzo

## LA FASE DI DIFFUSIONE

È la fase conclusiva della rilevazione, grazie alla quale si rendono disponibili (**accessibili e comprensibili**) i dati raccolti (**le statistiche e l'analisi statistica**).

Fa percepire all' «**esterno**» la specificità e l'**utilità** del lavoro svolto (**promozione**).

- adeguata offerta di informazioni
- massima facilità di accesso ai dati e ai metadati
- tempo ridotto tra raccolta, elaborazione e diffusione
- forme efficienti di diffusione e analisi del grado di utilizzo delle statistiche
  - caratteristiche degli utenti
  - canali utilizzati
- valutare se la qualità dei dati è adatta alla diffusione
- valutare la necessità di modificare il processo di produzione

Alle attività di produzione e di diffusione dei dati sono collegate strategie di comunicazione mirate e differenziate per segmenti di utenti, con lo scopo di trasmettere l'informazione statistica.

<https://www.spreaker.com/episode/trailer-dati-alla-mano--53425780>

Un importante ruolo nelle attività di comunicazione dei dati è rivestito dai canali social.

<https://www.youtube.com/watch?v=bpCaoFaLiU>

## TEMPI: definizioni

### INSERIMENTO DELLE DATE RELATIVAMENTE A:

- data di pubblicazione effettiva dei dati definitivi
- data di riferimento dei dati dell'indagine
- data di pubblicazione dei dati provvisori (se applicabile)
- data di pubblicazione programmata dei dati definitivi
- data della pubblicazione cartacea

## Indicatori di qualità' sul TEMPI

- Tempestività: Indica la diffusione dei risultati secondo le necessità degli utenti. Viene riferita alla distanza di tempo tra la data di diffusione dei risultati definitivi e la loro data di riferimento
- Anticipazione dati provvisori: Per le indagini che diffondono dati provvisori, quanto tempo prima questi sono diffusi rispetto a quelli definitivi
- Ritardo della pubblicazione cartacea: Differenza tra data della pubblicazione cartacea e primo rilascio dei dati
- Puntualità: La puntualità si riferisce alla diffusione dei dati secondo il calendario prestabilito

## Costi per l'indagine

Non rappresentano un requisito della qualità ma un vincolo.

Esempio

Per indagini con periodicità inferiore all'anno, costo per replicazione (ottenuto come media sul costo annuale)

## Indicatori di qualità sui costi

Personale impiegato nell'indagine (anni/persona)

- per livello e tipologia di contratto (di ruolo/non di ruolo)

Costi per raccolta e registrazione

- stampa questionari
- spedizione questionari
- raccolta dati
- registrazione in service
- .....

## Rilevazione Istat Indagine Viaggi e vacanze - Diffusione



10 febbraio 2020

VIAGGI E VACANZE IN ITALIA E ALL'ESTERO | ANNO 2019

### Rallenta il turismo dei residenti: -8,8% i viaggi, -5,0% i pernottamenti

➔ I viaggi dei residenti in Italia nel 2019 sono 71 milioni e 883 mila (411 milioni e 155 mila pernottamenti) con una flessione sull'anno precedente che interrompe la ripresa iniziata nel 2016. In calo sia le vacanze (-8,4%) sia i viaggi di lavoro (-12%).

In estate, il 37,8% della popolazione fa almeno una vacanza.

Il 76,2% dei viaggi ha come destinazione una località italiana (-12,8% sul 2018), il 23,8% è diretto all'estero.

# 89%

La quota di viaggi per motivi di vacanza.

Rappresentano il 93,4% dei pernottamenti.

# 5,7 notti

Durata media dei viaggi

# 63%

Percentuale dei viaggi estivi occasione di partecipazione ad attività culturali

## Nota metodologica

### Riferimenti normativi

La rilevazione di informazioni riguardanti il turismo è prevista dal Programma statistico nazionale, che raccoglie l'insieme delle rilevazioni statistiche necessarie al Paese. Inoltre, essa viene svolta in conformità alle definizioni concettuali e metodologiche espresse dal Regolamento per le Statistiche del Turismo 692/2011, che ha sostituito la precedente Direttiva 95/57/CE.

### Obiettivi conoscitivi e quadro di riferimento

"Viaggi e vacanze" è un focus inserito nell'intervista finale dell'[indagine sulle Spese delle famiglie](#) a partire dal 2014, e consente di rilevare informazioni sui movimenti turistici dei residenti in Italia. Tali informazioni erano rilevate precedentemente dall'indagine trimestrale [Viaggi, vacanze e vita quotidiana](#), condotta dal 1997 al 2013.

Il focus ha la finalità di ottenere informazioni sui [movimenti turistici](#) della popolazione (domanda turistica). Le stime prodotte riguardano il numero di turisti, viaggi, pernottamenti in viaggio e escursioni sul territorio nazionale o all'estero.

Il quadro normativo della rilevazione ha come riferimento il [Regolamento per le Statistiche del Turismo 692/2011, nell'ambito del framework](#) concettuale e metodologico delle [International Recommendations for Tourism Statistics 2008 \(IRTS 2008\)](#). Il turismo è definito come l'insieme delle attività e dei servizi riguardanti le persone che si spostano al di fuori del loro "ambiente abituale" per vacanza o per motivi di lavoro. Rientrano pertanto nei flussi turistici tutti gli spostamenti non abituali, con pernottamento (viaggi) o senza (escursioni). L'individuazione dell'ambiente abituale di una persona permette di distinguere correttamente il fenomeno turistico dalla mobilità, che non rientra nel campo di osservazione della domanda turistica.

Ad esempio, i viaggi e le escursioni abituali, quelli cioè effettuati settimanalmente nella stessa località, diversa dal luogo in cui si vive, sono comunque assimilabili all'ambiente abituale e non rientrano nei flussi turistici; si presuppone, infatti, che tali spostamenti siano riconducibili alla vita quotidiana e alle abitudini dell'individuo. Sono altresì esclusi dalla definizione di "turista" le persone che si spostano giornalmente o settimanalmente per lavoro, per studio o per motivi personali, quando cioè lo spostamento rientra nell'ambito di attività di *routine*.

I viaggi turistici (non abituali) sono classificati, secondo gli standard internazionali, distinguendo i viaggi per motivi di lavoro da quelli per motivi di vacanza e le vacanze "brevi" (da 1 a 3 notti) da quelle "lunghe" (più di 3 notti). Tra le vacanze rientrano i viaggi per svago, piacere, relax, per visitare parenti e religiosi.

Maggiori informazioni sono disponibili sul sito dell'Istat all'indirizzo: [I](#)

### Fonti di dati

La fonte informativa è rappresentata dall'indagine sulle Spese delle famiglie e vacanze". L'indagine è campionaria e continua (è svolta tutti i mesi e su base trimestrale, e a due stadi di cui il primo è stratificato: le unità secondo stadio sono le famiglie.

### Classificazioni

Nella rilevazione sono utilizzate le classificazioni territoriali Istat di Comuni, Province e Regioni, le classificazioni Istat degli Stati Esteri e *Nomenclature of Territorial Units for Statistics - NUTS*, la classificazione dell'attività economica Ateco 2007 (Nace Rev.2), la classificazione ISCED dei titoli di studio. Per alcune caratteristiche del viaggio, tra cui tipo di alloggio, motivo e tipo di destinazione, si utilizzano le classificazioni dei metadati di Eurostat, consultabili all'indirizzo: [eurostat's metadata server-ramon](#).

### Diffusione

Tra febbraio e marzo di ogni anno la Statistica Report "Viaggi e vacanze in Italia e all'estero" diffonde le stime provvisorie riferite all'anno precedente.

Le stime definitive sono consultabili, a partire dal mese di luglio, nel datawarehouse dell'Istituto [I.Stat](#), sotto il tema:



Definizione della popolazione, della tipologia di studio, della tecnica di indagine, ...

## CALENDARIO DELLE DIFFUSIONI E DEGLI EVENTI

<https://www.youtube.com/watch?v=w0-BhhYqN2A>

## DATI ANALISI E PRODOTTI

## METODI E STRUMENTI

## INFORMAZIONI E SERVIZI

## PER GLI UTENTI

Sportelli sul territorio  
European data support  
Biblioteca  
Archivio storico

## PER I GIORNALISTI

Appuntamenti  
- Calendario diffusionsi ed

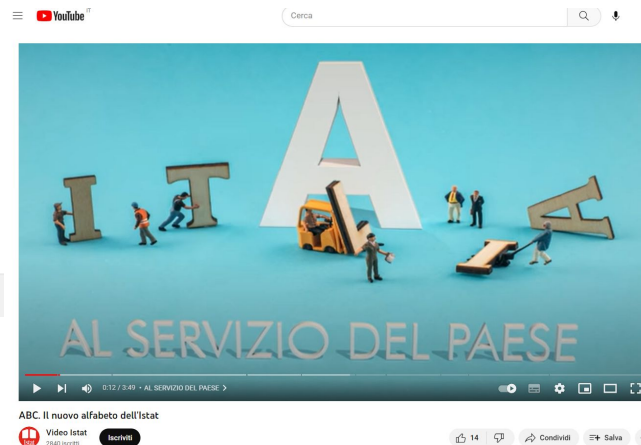
Il calendario delle diffusionsi e degli eventi, già completo dei comunicati stampa i cui **rilasci sono fissati per l'intero anno**, viene aggiornato con cadenza settimanale con le date di tutte le diffusionsi alla stampa, della pubblicazione dei prodotti editoriali, degli eventi e degli aggiornamenti delle banche dati.

📄 [Calendario annuale dei comunicati stampa 2022](#) (pdf)

📄 [Calendario annuale dei comunicati stampa 2023](#) (pdf)

< 17 - 23 APRILE 2023 >

Lunedì 17 Aprile 2023



## Diffusione e qualità

La fase della diffusione ha ripercussioni sulla qualità dei dati in termini di:

- **ACCESSIBILITÀ** ossia possibilità per gli utilizzatori di entrare in possesso dei dati
- **CONFRONTABILITÀ** ossia possibilità di paragonare nel tempo e nello spazio le statistiche riguardanti il fenomeno di interesse.

## DATI PRIMARI – INDAGINE AD HOC

$$\text{Valore osservato (variabile)} = \text{Valore vero (concetto)} + \text{Errore}$$



Qual è l'aspetto caratterizzante dell'ufficialità?

Il rispetto

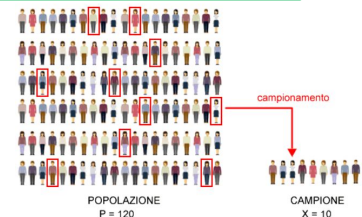
- di regole metodologiche condivise,
- di principi etico-professionali,
- dell'autonomia scientifica.

### Errore non campionario

- \* errori di lista delle unità da osservare
- \* errori nel questionario
- \* errori dovuti alla tecnica di indagine
- \* errori dei rilevatori
- \* errori dovuti ai rispondenti
- \* mancate risposte (totali e parziali)



### Errore campionario

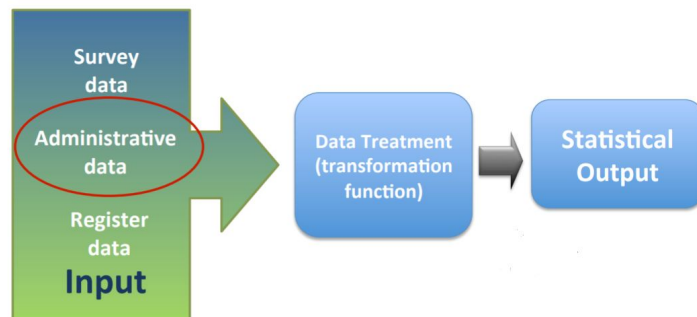


L'obiettivo finale di generare **VALORE PUBBLICO** e impatti migliorativi per la collettività non può prescindere da un'attenzione costante verso la **QUALITÀ**.

## DATI SECONDARI – FONTI STATISTICHE UFFICIALI

- Nuove fonti, nuovi metodi e tecniche (accessibilità, tempestività, granularità, ...) e nuovi produttori di statistiche che “danno i numeri”

**CAUTELA NELL'USO**  
(verifica della rispondenza  
rispetto agli obiettivi  
conoscitivi)



**LA CAPACITÀ DI ESTRARRE  
VALORE DAI DATI È LEGATA  
ALLA CAPACITÀ DI INTEGRARE  
DATI CHE PROVENGONO DA  
FONTI DIFFERENTI!**

- Nuove forme di comunicazione dei dati (video, spot di pubblica utilità, tutorial, podcast,...)  
Data Comedy Show (Gli anziani lo fanno meno di tutti - 16/11/2021)  
<https://www.youtube.com/watch?v=NN3ujhVzrk>

## LE QUESTIONI APERTE E IL VALORE DEI DATI

- Un mondo di dati e di produttori di **STATISTICHE NON TUTTE UGUALI** per esigenze conoscitive sempre più diverse.
  - **AUTOREVOLEZZA DELLA FONTE** (credenziali e reputazione di imparzialità, indipendenza scientifica, impianto organizzativo pubblico)
  - **QUALITÀ DEL DATO PRODOTTO** (affidabilità, pertinenza, trasparenza, riservatezza, tempestività etc.) tanto maggiore è la qualità tanto minore è l'errore connesso in ogni fase
  
- L'informare come «dare senso alla realtà» richiede quantomeno informazioni **CORRETTE**.
  
- I dati statistici - asset strategico per conoscere la realtà, .....ma si traducono in **INFORMAZIONE STATISTICA** quanto più si questa si diffonde e viene utilizzata.

| Criteria di qualità              | Descrizione  |
|----------------------------------|--|
| <b>Pertinenza</b>                | Il grado in cui le statistiche rispondono alle esigenze attuali e potenziali degli utilizzatori.   |
| <b>Accuratezza</b>               | Il grado di corrispondenza fra le stime e i valori veri ignoti.  |
| <b>Comparabilità</b>             | La misurazione dell'impatto delle differenze tra i concetti statistici applicati, gli strumenti e le procedure di misurazione quando le statistiche si comparano per aree geografiche, ambiti settoriali o periodi di tempo. |
| <b>Coerenza</b>                  | L'idoneità dei dati a essere attendibilmente combinati in modi diversi e a vari scopi.   |
| <b>Tempestività</b>              | L'intervallo di tempo intercorrente fra il momento della diffusione dell'informazione e l'evento o il fenomeno da essa descritto.  |
| <b>Puntualità</b>                | L'intervallo di tempo intercorrente fra la data di diffusione dei dati e il termine previsto per la loro diffusione.   |
| <b>Accessibilità e chiarezza</b> | Le modalità e le condizioni alle quali gli utilizzatori possono acquisire, utilizzare e interpretare i dati.   |

Fonte: Corte dei conti europea, sulla base dell'articolo 12 del regolamento (CE) n. 223/2009.

## ***BIBLIOGRAFIA ESSENZIALE***

AA.VV., Manuale di tecniche di indagine, Vol. 1-6, ISTAT (1989)

[https://lipari.istat.it/digibib/Metodi%20e%20norme/Manuale\\_e\\_tecniche\\_di\\_indagine/](https://lipari.istat.it/digibib/Metodi%20e%20norme/Manuale_e_tecniche_di_indagine/)

H.M. Blalock H.M. (1969), Statistica per la ricerca sociale, Il Mulino (riedizioni) [ed. orig.: Social Statistics, McGraw Hill, New York, 1960].

G.Cicchitelli, P.D'Urso, M.Minozzo, Statistica. Principi e Metodi, Mylab (2022).

P.Corbetta, Metodologia e tecniche della ricerca sociale, Il Mulino (2014).

G.Leti, Statistica Descrittiva, Il Mulino. (1983 e riedizioni)

D.Piccolo, Statistica, Il Mulino. (2010 e riedizioni).

L.Ricolfi, Manuale di analisi dei dati. Fondamenti, Laterza (2024).



COESIONE  
ITALIA



*A Scuola di*  
**OPENCOESIONE**

Rita Lima  
[lima@istat.it](mailto:lima@istat.it)  
[www.istat.it](http://www.istat.it)

*Grazie*



Istat





**Presidenza del Consiglio dei Ministri**  
Dipartimento per le politiche di coesione e per il sud

In collaborazione con



**MIM**  
Ministero dell'Istruzione  
e del Merito



Progetto finanziato con il sostegno di



**UNIONE EUROPEA**  
Fondo Sociale Europeo  
Fondo Europeo di Sviluppo Regionale



**PROGRAMMA  
OPERATIVO  
COMPLEMENTARE**



**GOVERNANCE  
E CAPACITÀ  
ISTITUZIONALE  
2014-2020**