Qualità delle statistiche e l'utilizzo di fonti non-statistiche a fini statistici

Carlo Filippucci
carlo.filippucci@unibo.it

DI COSA TRATTO

- L'idea di ***qualità*** delle statistiche
- Il rilievo delle fonti non-statistiche (amministrative e gestionali) nei sistemi statistici moderni – big data-
- Il controllo dell'affidabilità – qualità - di tali fonti e forse di qualunque altra informazione

Cosa intendiamo per QUALITÀ di una statistica

Categoria **complessa** non limitata al solo errore statistico

- **accuratezza** (errore campionario e non-campionario)
- **tempestività**
- **coerenza**
- **confrontabilità** nel tempo e nello spazio,
- **accessibilità**
- **costi**

Errore nelle statistiche

L'osservazione e la misura sono soggette a errore

Inevitabile

L'errore non è una anomalia da rifuggire ma va riconosciuto come intrinseco a tutte le forme di attività e conoscenza

Heisenberg, principio di indeterminazione

Popper, principio di falsificazione delle teorie

Errore nelle statistiche e paradigma dell'inertezza

L'INCERTEZZA – categoria fondamentale della conoscenza moderna

Consente di superar la logica deterministica ottocentesca che pretendeva di determinare esattamente una conoscenzaa partire da un numero finito di cause.

L'errore per secoli è stato demonizzato per fare spazio alla cultura della certezza, a quella conoscenza che non si fa domande ma che dispone di risposte assolute, di fedi.

La certezza e non il dubbio è lo strattagemma di Satana (Acconcio, 1563).

Errore nelle statistiche

La forza della rappresentazione statistica sta nel cercare di rendere oggettivamente comprensibile e valutabile da tutti gli utilizzatori del dato l'entità del suo errore

quindi i limiti entro i quali l'analisi sostanziale di un fenomeno ha valore.

Il dato statistico è dunque il frutto di un processo mentale e operativo che viene esplicitato in tutte le sue fasi ed è pertanto valutabile *oggettivamente*.

Errore nelle statistiche

La questione è dunque

- CONOSCERE L'ENTITA' E LA TIPOLOGIA DELL'ERRORE
- ESPLICITARLO
- INDIVIDUARE MODELLI E STRATEGIE PER RIMUOVERLO E PREVENIRLO

ERRORI NELLE STATISTICHE

CAMPIONARIO

- INDAGINI CAMPIONARIE

N ON-CAMPIONARIO

- INDAGINI CAMPIONARIE E CENSUARIE E FONTI NON STATISTICHE

ERRORE CAMPIONARIO – per molto tempo unico errore considerato

ERRORE DIPENDE :

- Dimensione del campione
- Variabilità della popolazione e del campione
- Precisione delle stime
- Livello di confidenza

vincoli sugli errori relativi percentuali imposti a livello di provincia, regione e Italia
vincoli sugli errori relativi percentuali imposti a livello di provincia, regione e Italia

	Forze di lavoro	Occupati	Persone in cerca occupazione	Persone in cerca I occupazione	Disoccupati	Altri in cerca di occupazione	5% popolazione in età lavorativa
Provincia	8%	-	25%	-	-	-	-
Regione	-	-	12%	-	-	-	11,5%
Italia	0,5%	0,5%	1,96%	2,7%	2,7%	3,5%	1,65%

vincoli sugli errori relativi percentuali imposti a livello di regione e Italia

	Forze di lavoro	Occupati	Persone in cerca occupazione	Persone in cerca I occupazione	Disoccupati	Altri in cerca occupazione	5% popolazione in età lavorativa
Regione	-	-	11%	-	-	-	11,5%
Italia	0,5%	0,5%	1,96%	2,7%	2,7%	3,5%	1,65%

ERRORE NON CAMPIONARIO

- Errori di identificazione della popolazione mancata pertinenza dei dati
- Errori di copertura (sovra copertura o sottocopertura)

  dipendono dalla qualità della lista

  implicano distorsioni nei dati
- Errori dovuti mancata risposta (totale e parziale)

  implicano riduzione numerosità e aumento variabilità

  distorsione

ERRORE NON CAMPIONARIO

- Errori di misura

  -tecnica di indagine (PAPI, CAPI, CATI, CAWI, altre)

  -questionario (formulazione delle domanda, lunghezza)
  -rilevatore

  -rispondente

  implicano distorsione nei dati

- Errori di processo (registrazione, codifica e classificazione, correzioni e imputazioni, editing)

  implicano distorsione nei dati

ERRORI NELLE STATISTICHE e QUALITA' DEI DATI

Fondamentale conoscere:

- Indicatori relativi a tutte le problematiche ricordate
- Strategie di trattamento adottate
- Effetti di tali strategie
- Calcolo dell'errore totale di misura

ma anche molto difficile intervenire a posteriori

PREVENIRE

Controllo del processo di produzione in tutte le sue fasi

Total quality management

Utilizzo fonti non-statistiche a fini statistici

L'utilizzo di tali archivi nell'ambito della produzione statistica non è una novità

- *Almeno fino al secondo dopoguerra, questo tipo di informazioni è stato, insieme ai censimenti, la principale fonte per la statistica*

- Fine anni '80 nuova attenzione verso le fonti non statistiche grazie a: evoluzione della tecnica, aumento attività burocratica gestionale, fabbisogno di informazioni

Utilizzo fonti amministrative in Italia

La presidenza Rey realizzò una svolta sostanziale nella statistica italiana - Rapporto Moser del 1983 –

puntava a sviluppare:

- il coordinamento e l'integrazione di basi di dati,
- l'impiego delle metodologie e delle tecniche statistiche,
- l' utilizzo di dati amministrativi per fini statistici;
- Uno sviluppo e coordinamento dell'attività statistica svolta dall'Istat e dagli altri enti della PA.

Il contesto

A. **crescente disponibilità di archivi derivanti** da attività burocratico-amministrative e gestionali per numerosi ambiti fenomenici e soggetti. Diversi soggetti istituzionali e molti privati emergono come "nuovi produttori" di dati

Enti e soggetti privati che raccolgono informazioni, ESEMPI:

Privati

- Catene commerciali
- Social network e più in generale il *web*
- Banche, assicurazioni e istituti finanziari (per analisi rischiosità); indicatori macroeconomici e previsioni di interesse generale

PA ed Enti locali

- dati sulla popolazione, enormi basi di dati fiscali, sanitarie, previdenziali e assicurative, relative alle costruzioni

Il contesto

B. interesse crescente a sfruttare questi archivi da parte delle agenzie di statistica

*le indagini sono solo una delle fonti della statistica e un contributo sempre maggiore viene dall'utilizzo di fonti amministrative e DALL'**INTEGRAZIONE** tra queste e le indagini realizzando un sistema statistico capace di avvalersi del "contributo di diversi soggetti"*

Utilizzi a fini statistici ESEMPI:

- Censimenti – utilizzo anagrafi e per la lista delle imprese una combinazione di fonti (INPS, Camere di commercio, Enel)
- PIL (integrazione dati mancanti con dati fiscali)
- statistiche sui consumi e sanitarie
- statistiche giudiziarie penali : Ministero della giustizia, dalle Procure della Repubblica e dal Casellario giudiziale centrale;
- Statistiche sulla delittuosità: Ministero dell'Interno –
- Statistiche penitenziarie :Dipartimento dell'amministrazione penitenziaria e dal Dipartimento per la giustizia minorile.

Statistiche ufficiali sui beni confiscati e sequestrati Ministero della Giustizia

Raccolti tramite :

Cancellerie , uffici del registro, agenzia del Demanio, Prefetture, Questure, comuni

- **I dati sono raccolti su schede cartacee e poi inserite in un supporto informatico**
- **Problemi:**

classificazioni non corrette, incongruenze nella indicazione delle diverse tipologie di destinazione, diffusa incompletezza dei dati, ecc.

Progetto SIPPI

"sistema Informativo Prefetture e Procure dell'Italia meridionale"

finalizzato alla creazione di una banca dati centralizzata automatizzata per la gestione di tutti i dati e le informazioni relative

Il contesto

… ma anche:

C. Crescenti necessità informative per il governo e lo studio dei fenomeni a livello locale

trasferimento a livello locale di responsabilità di governo e

non adeguata disponibilità di statistiche territoriali

*Molte statistiche per **PICCOLE AREE** vengono prodotte indirettamente con procedure che ricorrono a dati amministrativi*

D. disponibilità di informazioni a livello micro

individuare, esplorare nuovi fenomeni, comportamenti e strategie delle famiglie e delle imprese, trattare nuove problematiche arricchendo la comprensione dei fenomeni che si può ottenere dalle statistiche macroeconomiche

Arricchire le conoscenze che risultano troppo condizionate da dati statistici derivati da modelli concettuali indeboliti dall'evoluzione della società moderna

Il contesto

- rivoluzione nella produzione di dati
- nuovi problemi e sfide alla statistica ufficiale
- necessità di affrontare in modo nuovo il tema della qualità delle statistiche

QUALCHE PROBLEMA

l’informazione è abbondante, accessibile, sovente utilizzata per produrre statistiche MA

il problema dell’utilizzatore non è solo quello di scegliere ma anche quello di non esserne condizionato o anche fuorviato

Esempi:

Non contrapporre PIL a presenze nei ristoranti; forze lavoro a dati INPS

QUALCHE PROBLEMA

Disporre di molte informazioni non implica che queste siano tutte egualmente pertinenti rispetto alle nostre esigenze conoscitive

tanto meno assicura la produzione di statistiche che **rispettino i fondamentali principi di imparzialità, obbiettività ed affidabilità** che il codice della statistica europea pone alla base dell'autorevolezza ed efficacia di un sistema statistico

Una digressione da ricordare: specificità del dato statistico

una qualunque INFORMAZIONE QUANTITATIVA NON È UN DATO STATISTICO non è ancora una conoscenza sostanziale

Il DATO STATISTICO scaturisce da : un preciso obbiettivo conoscitivo, un sistema di ipotesi, definizioni e criteri classificatori, una strategia ed un processo di misura e di validazione controllati che consentano di misurarne l'affidabilità attraverso indicatori chiari, obbiettivi e comprensibili a tutti---

Una digressione da ricordare: specificità del dato statistico

… e dunque è

il fondamento dell'istanza galileiana della verifica empirica…

Il limite ad una conoscenza basata sia sui luoghi comuni, sulle impressioni soggettive, sulle conoscenze parziali e sulle generalizzazioni di casi specifici

Una digressione da ricordare: specificità del dato statistico cnt.

…Ma attenzione

Le statistiche rispondono a finalità conoscitive specifiche e la cui attendibilità non dipende esclusivamente da fattori tecnico-statistici

Una digressione da ricordare: specificità del dato statistico cnt.

il valore euristico dell'evidenza empirica fornita dalle statistiche è legittimata in quanto:

- rende esplicite le scelte e le assunzioni alla base dei dati e tutti gli aspetti del processo di misura
- consente di valutare se tali ipotesi e scelte corrispondono alle istanze conoscitive del utilizzatore
- Fornisce agli utilizzatori dati elementari affidabili per l'elaborazione autonoma

…PER CONTRO…

Fonti non statistiche derivanti da finalità burocratiche e/o gestionali
LIMITI

- rispondono a finalità conoscitive autonome
- rispondono a criteri di affidabilità, quanto ci sono, radicalmente diversi da quelli dell'indagine statistica
- appaiono scoordinate, occasionali, realizzate con metodologie diverse, soggette a cambiamenti frequenti
- Presentano una scarsa attenzione ad una definizione di qualità ben specificata e alla misura dell errore,
- Non hanno documentazione adeguata del procedimento di rilevazione e del processo che ha portato al dato fornito.

…dunque per le fonti non statistiche: DUE RISCHI

Rilevanti per una società che deve effettuare scelte consapevoli e sempre più fondate sulla conoscenza fattuale

- utilizzo acritico di informazioni la cui portata conoscitiva nel caso migliore non è chiara ma può anche essere sbagliata (se non volutamente fuorviante).
- contrapposizione di dati contrastanti che porterebbe ad una perdita di credibilità della statistica quale *modus intelligendi* della scienza e della conoscenza moderne.

Fonti non-statistiche derivanti da finalità burocratiche e/o gestionali

pongono quindi problemi non semplici che tuttavia si possono affrontare con strumenti e attenzioni specifiche

Le problematiche relative ad errore non campionario ed alla qualità dei dati sono molto vicine quando non si identificano con

L’utilizzo a fini statistici di fonti non statistiche

L’utilizzo di fonti non statistiche: tre questioni generali

- **assicurare la *privacy*,**

volta a raggiungere un consenso informato dei possessori dell’informazione

- **assicurare la *confidentiality*,**

evitare usi impropri dell’informazione contenuta negli archivi una volta che questa sia resa disponibile a fini statistici

- **assicurare la qualità delle informazioni**

Occorrono modelli concettuali, metodologie e procedure operative che tengano conto di approcci culturali diversi (giuridico, amministrativo, gestionale, statistico) al fine di garantire la produzione di dati affidabili, coerenti con principi e standard condivisi e **confrontabili nel tempo e nello spazio e quindi capaci di rispondere a principi qualitativi condivisi**

l'utilizzo a fini statistici di una fonte non-statistica

deve essere valutato attraverso una sorta di analisi dei benefici che essa apporta ma anche dei costi che questa implica e dei limiti che presenta valutando:

- qualità della fonte
- qualità del processo di produzione dei dati

qualità di una fonte

- ***Legittimità***, conformità della fonte ad un "codice etico" relativo alle modalità di raccolta delle informazioni.
- ***autorevolezza***, il prestigio, l'indipendenza e la professionalità del produttore della fonte
- ***credibilità-attendibilità***, esistenza di requisiti che assicurino il rispetto dei principi generali e delle pratiche adottate per la raccolta di dati e l'esistenza di standard di raccolta dell'informazione
- ***trasparenza***, documentazione relativa alla fonte
- ***fruibilità***, possibilità di ottenere l'informazione con sufficiente semplicità in formato elettronico
- ***stabilità***, permanenza e continuità della fonte, delle definizioni e delle informazioni rilevate

Qualità delle informazioni

- ***attinenza***, corrispondenza tra le definizioni (quindi obbiettivi) statistiche e quelle utilizzate nelle fonti non statistiche
- ***precisione***, aderenza dell'informazione raccolta al fenomeno considerato e valutazione degli errori di misura
- ***tempestività***, distanza tra il momento a cui si manifestano gli eventi e quello in cui l'informazione viene raccolta
- ***puntualità***, distanza tra il momento in cui l'informazione viene resa accessibile e il riferimento eventi
- ***conformità***, consistenza tra stesse informazioni raccolte da fonti diverse
- ***comparabilità***, coerenza tra le informazioni prodotte nel tempo, nello spazio e per i domini rilevanti

errori di misura da valutare

Gli *errori di misura tipici fonte statistica*

- difficoltà ad esprimere concetti e definizioni misurabili,
- disponibilità/capacità di fornire l'informazione da parte di chi la detiene,
- processo di trattamento ed *editing* dei dati
- metodo e strumenti di misura adottati

errori di misura da valutare cnt

Gli *errori di misura tipici fonte non statistica*

- **trattamento dei dati in seguito ai controlli** in fase di acquisizione dell'informazione, dall'applicazione di specifiche regole di correzione e trattamento in sede di utilizzo statistico e dalla trasformazione delle variabili amministrative in variabili statistiche.
- **diversa "qualità" delle diverse variabili** osservate – si presta maggior attenzione alle variabili rilevanti per l'ente
- **differenza temporale** tra il momento in cui si chiede la registrazione degli eventi e quello in cui si sono verificati.

errori di rappresentazione

- lista/copertura,
- mancata risposta,
- Linking mancanti – nel caso di incroci tra più fonti-

Per concludere

- Alle fonti non-statistiche dobbiamo fare ricorso
- Occorre fare tutte le verifiche necessarie per giudicare e informare sulla sua qualità e su quella del dato statistico che si può ricavare
- Se il dato statistico è sempre viziato da una componente di errore - che possiamo esplicitare e calcolare – anche sulle informazioni non statistiche dobbiamo effettuare controlli adeguati e documentarli