Statistica sociale - 1

Prof. Antonio Mussino

a. a. 2022-23

SAPIENZA
UNIVERSITÀ DI ROMA

Il programma in sintesi - 1

- Introduzione al corso
- Richiami sulle definizioni statistiche
  - Le matrici di dati
  - Le variabili
  - Le unità
- Alcuni richiami sulle fonti statistiche nel sociale
  - Le fonti
  - Le fonti ufficiali in Italia
  - Le fonti ufficiali nell'Unione europea
  - Le fonti internazionali
  - Le principali indagini per la produzione di statistiche sociali

Il programma in sintesi - 2

- La metodologia delle indagini campionarie
  - Il campionamento
  - Il questionario
  - Le modalità di intervista
- Indici e indicatori
  - Attendibilità e validità
  - La sintesi degli indicatori
- L'elaborazione dei dati
  - Codifica e input
  - Il caso univariato
  - Il caso bivariato
  - Il caso multivariato(logica esplorativa e logica inferenziale)

Introduzione

- La diffusione prima e la crisi negli ultimi anni del *welfare* hanno stimolato la necessità di quantificare i bisogni sociali, programmare gli interventi e valutare i risultati delle azioni e delle prestazioni.
- La disciplina che più ha fornito contributi in tale contesto è stata la Statistica, che ha sviluppato metodi e tecniche specifici per i problemi emersi: questa particolare area della Statistica è classificata accademicamente come **Statistica sociale**.

La Statistica sociale

- Nella Statistica sociale convivono due anime, che sono specifiche dell'area di studio:
- quella il cui fine è esplorare, descrivere e sintetizzare il sociale, che fornisce strumenti utili per svolgere indagini, costruire banche dati, definire indici e indicatori (la ***Statistica del sistema sociale***);
- quella più metodologica, che invece ha sviluppato metodi e tecniche specifiche per analizzare al meglio i dati risultanti dalla prima (la ***Statistica per la ricerca sociale***).

Il corso - 1

- Questo corso si rivolge a studenti che possiedono già i primi rudimenti della Statistica, descrittiva e inferenziale, quindi della Probabilità, dell'Economia, della Demografia e della Sociologia.

- L'obiettivo è analizzare operativamente un sistema sociale o una delle sue componenti.

Il corso - 2

- Una prima parte tratterà delle fonti e delle banche dati (amministrativi o tratti da indagini campionarie);
- una seconda degli aspetti metodologici delle indagini campionarie;
- una terza degli indici e degli indicatori calcolabili sulle banche dati;
- una quarta introdurrà le strategie multidimensionali di sintesi di tali indicatori.
- Alla parte istituzionale si aggiungeranno, eventualmente, temi specifici con seminari «ad hoc».

Alcuni richiami sulle definizioni statistiche

Le matrici dei dati - 1

- Qualunque sia l'approccio statistico scelto, le elaborazioni e le analisi partono sempre dalla raccolta dei dati, che potrà essere:
  - una rilevazione presso archivi pubblici e/o privati (***dati secondari***) oppure
  - sarà il risultato di una rilevazione diretta *sul campo*, ad esempio con interviste tramite questionario (***dati primari***).

Le matrici dei dati - 2

- I dati così raccolti sono registrati fisicamente su di un *tabellone*:
- una volta un grande foglio *quadrettato*, oggi su di un *foglio di calcolo* (Excel, Access, SPSS e così via).

- La logica che sta dietro questa operazione è quella della ***matrice dei dati***, dalla quale partiremo per definire tutti gli elementi che potremmo analizzare.

Le matrici dei dati - 3

- Una matrice di dati è, genericamente, un insieme di numeri o di codici disposti ordinatamente per riga e per colonna.

- Dal punto di vista statistico è un insieme di informazioni ottenuto dall'intersezione di due insiemi omogenei al loro interno **I** e **J**, uno organizzato per riga e l'altro per colonna.

Le matrici dei dati - 4

- L'insieme di riga è, generalmente, costituito dalle ***unità statistiche*** di riferimento (o *individui*), quello per colonna dai ***caratteri*** su di esse rilevati. L'elemento generico all'interno della matrice $\mathbf{x_{ij}}$ individua la determinazione che il carattere j-esimo assume nell'unità statistica i-esima.

Le matrici dei dati - 5

- In sintesi, organizzare una matrice di dati vuol dire scegliere i caratteri rispetto ai quali raccogliere le informazioni (i dati), stabilire come riportarle sulla matrice stessa (la codifica) e decidere su quali unità statistiche raccoglierle.

Le variabili - 1

- Riprendiamo le singole componenti della matrice: il termine *carattere* è tipico della scuola statistica italiana; decisamente più usato è il termine *variabile*, che segnala come per un'analisi statistica ci sia bisogno di una *variabilità* nell'insieme dei dati.

- Le variabili possono presentarsi in forme diverse e assumere determinazioni differenti: a seconda della loro natura hanno proprietà logico-formali diverse e le loro determinazioni possono essere trattate algebricamente in modo diverso.

Le variabili - 2

- Una classificazione operativa molto efficace è quella che individua variabili ***qualitative*** e ***quantitative***.

- Nel primo caso le determinazioni vengono definite ***modalità*** e non hanno una valenza numerica (es. genere, stato civile, sport praticato, titolo di studio); sono rappresentabili mediante codici informativi di tipo alfanumerico: ovvero, anche se sono rappresentate mediante numeri, questi non ne hanno la valenza.

Le variabili - 3

- Nel secondo caso le determinazioni sono numeri, che hanno una valenza numerica effettiva e possono essere usati per misurare le differenze (es. peso, statura, numero di fratelli, reddito).
- È pertanto possibile effettuare su questi valori tutte le operazioni algebriche, necessarie per calcolare tutti gli indici e i coefficienti statistici, cosa che non è possibile fare con le variabili qualitative.

Le variabili - 4

- Nell'ambito delle variabili qualitative è possibile individuarne un sottoinsieme in cui le modalità possono essere gerarchizzate (es. titolo di studio, in cui un titolo elementare è inferiore a uno medio superiore, questo a laurea e così via) individuando così una tipologia che è definita ***ordinale***.

Le variabili - 5

- Anche le quantitative sono differenziabili in ***continue*** (se i numeri utilizzabili sono quelli reali, ovvero se è possibile prevedere decimali), oppure ***discrete*** (se i numeri utilizzabili sono quelli interi, ovvero non è possibile prevedere decimali);
- questa suddivisione non è, in effetti, utile operativamente, perché entrambe queste tipologie sono elaborate come se esse fossero continue: è, infatti, possibile calcolare come numero medio di figli per donna un numero con più decimali (es. 1,29).

Le variabili - 6

- A queste tipologie ne vanno aggiunte altre due che hanno una natura completamente diversa, ma sono entrambe molto usate nella Statistica sociale: le variabili ***booleane*** e quelle ***lessicali***.

- Le variabili booleane (o binarie, o *dummy* nella terminologia anglosassone) possono assumere solamente due valori: "0" in caso di assenza del carattere considerato e "1" in caso di presenza. Si tratta di una variabile usata strumentalmente per elaborare quantitativamente le modalità delle variabili qualitative.

Le variabili - 7

- Le variabili lessicali possono assumere sequenze alfabetiche qualsiasi nell'ambito di testi letterali, ad esempio risposte aperte a domande di questionari, articoli di rivista, capitoli o paragrafi di libri, messaggi sui *social network*.

- I metodi per elaborarle fanno parte della ***Statistica testuale***, che è ampiamente usata nell'ambito della ricerca sociale.

Le variabili - 8

- Un’altra classificazione molto utilizzata nelle scienze sociali e umane è quella di Stevens (**NOIR**), in cui per tutte le variabili è definita una *scala di misurazione*, che però è diversa a seconda delle loro caratteristiche; ovvero può essere: Nominale, Ordinale, Intervallare, di Rapporti.

Le variabili - 9

- Nel primo caso (**Nominale**) l'unica forma possibile di misurazione è quella di assegnare un nome diverso a ogni modalità diversa e gli unici operatori applicabili sono quelli logici di = o #.

- Nel secondo caso (**Ordinale**) c'è la possibilità di ordinare le modalità e, quindi, si aggiungono gli operatori > e <.

Le variabili - 10

- Nel terzo caso (**Intervallare**) si può misurare la differenza fra due modalità, partendo da una posizione (origine) e con un'unità di misura convenzionali: è così possibile aggiungere gli operatori algebrici + e -.

- Il fatto che l'origine sia in questo caso convenzionale ci impedisce di effettuare anche ulteriori operazioni algebriche (a cominciare da * e /), che sono riservate al gradino più alto della scala, ovvero quello di **Rapporti** nel quale è possibile anche misurare le proporzioni fra le grandezze delle modalità.

Le variabili - 11

- Per meglio chiarire quest'ultima differenza possiamo mettere a confronto una variabile che rappresenti il punteggio a una scala di ansia con il peso:
  - nel primo caso il valore "0" non è effettivo, ovvero non misura l'assenza di ansia, ma solo il valore più basso della scala stessa;
  - nel secondo il valore "0" rappresenta effettivamente l'assenza di peso (definito *zero assoluto*).
- Quindi potrò dire che chi ha ansia pari a 80 ha un valore più grande di chi ne ha 40, ma non possiamo percepire questa differenza come una situazione in cui il primo individuo abbia il "doppio" di ansia, come invece è possibile per il peso.

Le unità - 1

- Torniamo ora all'altro insieme che definisce la matrice dei dati: quello delle ***unità statistiche***.
- Queste possono essere ***elementari***, se su di esse sono stati raccolti i dati (ovvero *unità di rilevazione*), o ***composte***, se rispetto ad esse i dati sono stati aggregati (ovvero *unità di analisi*); le unità di rilevazione possono anche essere direttamente unità di analisi.
- Unità elementare può essere un intervistato, un paziente di un ospedale, un atleta e così via; unità composta un ospedale, un comune, una squadra e così via.

Le unità - 2

- Le unità possono essere tutti i membri di un collettivo di riferimento e in questo caso usiamo per questo aggregato il termine ***popolazione***;
- oppure una parte del collettivo, selezionata *ad hoc* e in questo caso usiamo il termine ***campione***.
- Se ci troviamo di fronte a un campione selezionato in modo *adeguato* le informazioni raccolte saranno utili per stimare non solo la situazione in questo sottogruppo, bensì per estendere la stima (*inferire*) alla popolazione di riferimento da cui il campione è stato estratto (**Inferenza statistica**).

Le matrice dei dati - 6

- Per ogni variabile della matrice, presa singolarmente, è possibile effettuare operazioni di sintesi in termini di calcolo di frequenze, di valori medi (misure di tendenza centrale), di misure della variabilità (dispersione) e così via (analisi ***univariata***).
- Considerando poi le variabili a coppie è possibile, in base alla loro natura, effettuare operazioni che consentano di misurare l'esistenza, o meno, la direzione e l'intensità della eventuale relazione fra di loro, considerando eventualmente anche l'effetto su tale relazione di ulteriori variabili (in genere una o due) (analisi ***bivariata***).

Le matrice dei dati - 7

- Obiettivo ultimo è quello di ottenere la rappresentazione sintetica delle relazioni fra tutte le variabili che ci interessano, ma anche fra tutte le unità, presenti nella matrice dei dati (analisi ***multivariata***).
- La trattazione della parte metodologica di questa casistica è data per acquisita ai fini della nostra presentazione; il nostro obiettivo è, invece, quello di proporre le più comuni elaborazioni che vengono effettuate al riguardo quando si elaborano i microdati di un questionario o un insieme di indicatori desunti da indagini di Fonti ufficiali.