Lezione 10
Ripercorriamo il programma

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo parlato di statistica inferenziale, ovvero dell'insieme di tecniche statistiche che consentono di generalizzare i risultati ottenuti dai dati raccolti su un campione alla popolazione da cui è stato estratto.

Abbiamo visto come possiamo calcolare la numerosità di un campione e come possiamo definire l'intervallo di confidenza entro il quale ricadrà la nostra stima.

Infine abbiamo visto le principali tecniche di campionamento, suddividendole in due grandi categorie (probabilistiche e non probabilistiche.

Ci sono dei dubbi?

I fenomeni sociali. Come rilevarli e trattarli in modo statistico

Che cos’è il dato?

L’unità statistica

La popolazione (o collettivo)

La rilevazione dei dati

Le fasi di un’indagine statistica

Le variabili statistiche

- Variabili qualitative (classificazione)
  - Sconnesse
  - Ordinate
- Variabili quantitative
  - Discrete (esito di un conteggio)
  - Continue (esito di una misurazione)

Rappresentazione delle variabili

Distribuzione unitaria

Distribuzione di frequenza (assolute, relative e percentuali)

Distribuzioni cumulate

Distribuzioni aggregate

Rappresentazione delle variabili

Rappresentazioni grafiche

- Pittogrammi
- Ortogrammi
- Aerogrammi
- Istogrammi
- Cartogrammi

Sintetizzare le distribuzioni di frequenze

I valori centrali

Valori centrali non analitici

Moda

Mediana

Quantili

Valori centrali analitici

Media arimetica

Media geometrica

Media armonica

Sintetizzare le distribuzioni di frequenze

I valori di disuguaglianza

Omogeneità e eterogeneità

Dispersione

Variabilità rispetto a un centro (devianza, varianza, scarto quadratico medio, coefficiente di variazione)

La forma di una distribuzione (simmetria e curtosi)

I rapporti statistici

Le serie storiche e territoriali

I numeri indice

Analisi delle relazioni tra due caratteri

La tabella di contingenza (o a doppia entrata)

Frequenze marginali e condizionate

Indipendenza in distribuzione (Chi quadrato, Phi quadrato, V di Cramer)

Relazioni tra variabili quantitative (Rho)

Relazione lineare tra due variabili (retta di regressione)

L'indice di determinazione R^2

Inferenza statistica

L’errore campionario

Intervallo di confidenza

Tecniche di campionamento

L’errore campionario

- Campionamento casuale semplice
- Campionamento sistematico
- Campionamento stratificato
- Campionamento a stadi
- Campionamento a grappoli
- Campionamento per aree

L’errore non campionario

- Campionamento per quote
- Campionamento a valanga
- Campionamento a scelta ragionata

In dettaglio

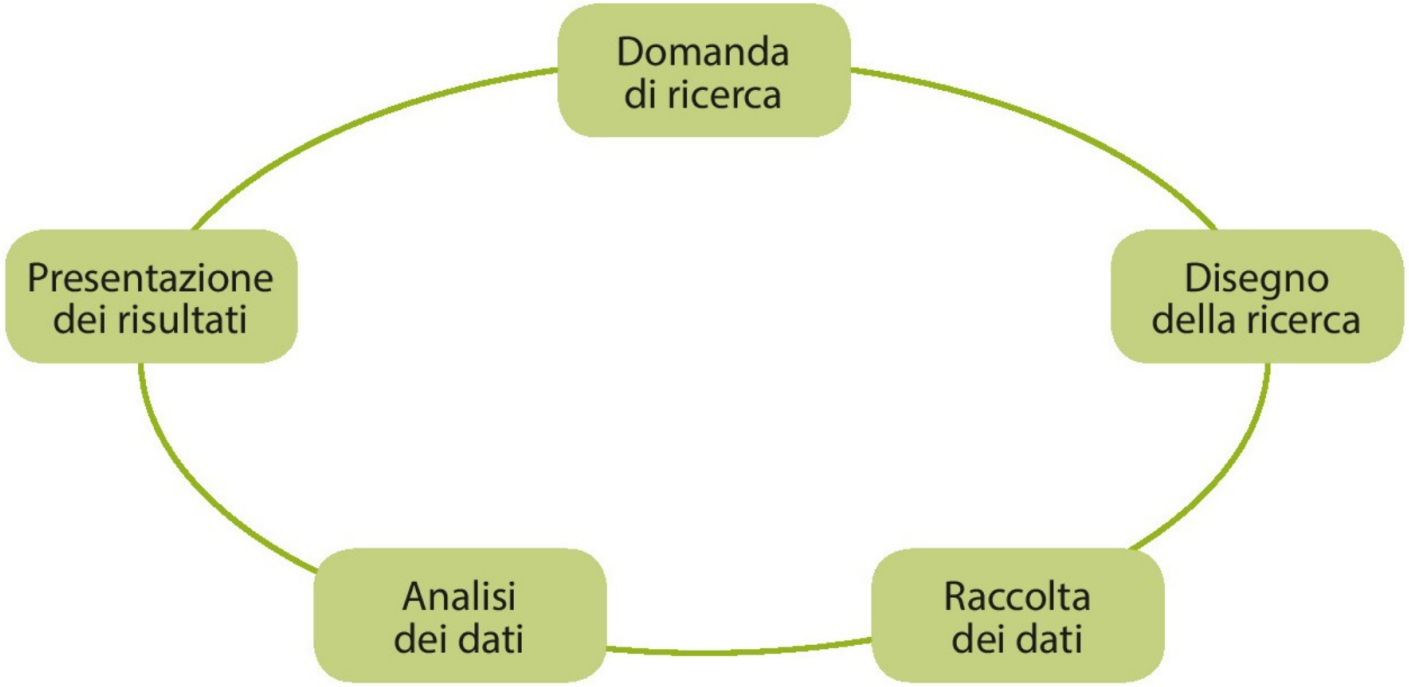
I fenomeni sociali e la statistica

La ricerca:

- Qualitativa
- Quantitativa

Il dato e i tre livelli del dato:

- Microdati
- Macrodati
- Metadati.



I fenomeni sociali e la statistica

L'unità statistica:

- Unità di raccolta (o rilevamento)
- Unità di analisi (o di riferimento)

La popolazione (o collettivo):

- Popolazione o collettivo di stato
- Popolazione o collettivo di movimento

- Popolazione empirica.
- Popolazione teorica

I fenomeni sociali e la statistica

La rilevazione dei dati:

- Indagine totale (censimento)
- Indagine parziale (campionaria)

Le tecniche di raccolta dei dati:

- Questionario
- Fonti statistiche secondarie (ad es. dati di fonte amministrativa)

I fenomeni sociali e la statistica

Le caratteristiche di un'unità vengono chiamate variabili statistiche e possono assumere diverse modalità:

- Variabili qualitative (classificazione)
  - Sconnesse
  - Ordinate
- Variabili quantitative
  - Discrete (esito di un conteggio)
  - Continue (esito di una misurazione)

Rappresentazione delle variabili

Matrice dei dati (distribuzione unitaria semplice o multipla)

Prima sintesi: per ogni modalità rappresento il numero di unità che la possiedono.

Distribuzione di frequenza

- Frequenze assolute (totale = N)
- Frequenze relative (totale = 1)
- Frequenze percentuali (totale = 100)

Distribuzioni cumulate

Distribuzioni aggregate (quando riduciamo il numero di modalità delle variabili analizzate)

Rappresentazioni grafiche

Pittogrammi (figure o simboli)

Ortogrammi (per variabili qualitative sconnesse)

Aerogrammi (diagrammi a torta, ad anello, a radar)

Istogrammi (per variabili quantitative)

Cartogrammi

I valori centrali

Indici di posizione o valori centrali non analitici

- Moda (tutte le variabili)
- Mediana (da variabili qualitative ordinate)
- Quantili

Valori centrali analitici.

- Media aritmetica
- Media geometrica
- Media armonica

I valori di disuguaglianza

Omogeneità/eterogeneità

Tutti i casi sono stati assegnati alla stessa modalità/sono divisi equamente tra tutte le modalità

Indice di eterogeneità di Gini.

$$E = 1 - \sum_{i=1}^{k} f_i^2$$

Dove $f_i = \frac{n_i}{N}$

Minimo = 0 massima omogeneità (una modalità ha frequenza relativa pari a 1 e le altre 0.

Massimo = (k-1)/k

I valori di disuguaglianza

Misure di dispersione

Differenza interquartile

Differenza interquartile = $Q_3 - Q_1$

Variabili qualitative ordinate

Range = $x_{max} - x_{min}$

Per variabili quantitative

I valori di disuguaglianza

Variabilità rispetto a un centro

Scostamento Semplice Medio

$$SSM = \frac{\sum_{i=1}^{N} |x_i - m(x)|}{N}$$

Devianza = $\sum_{i=1}^{n} (x_i - m(x))^2$

Varianza $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - m(x))^2}{N}$

Scarto quadratico medio (deviazione standard)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m(x))^2}{N}}$$

I valori di disuguaglianza

Variabilità rispetto a un centro

Coefficiente di variazione

$$C_v = \frac{\sigma}{m(x)}$$

I valori di disuguaglianza

La forma di una distribuzione

Asimmetria e curtosi

La nozione di asimmetria ha senso se un carattere è almeno ordinabile.
Simmetria -> Moda = Mediana = Media
Asimmetria positiva -> Moda < Mediana < Media
Asimmetria negativa -> Media < Mediana < Moda

Indice di asimmetria di Fisher

$M_3 = \frac{\sum_{i=1}^{n}(x_i - m(x))^3}{N}$ Se $M_3 > 0$ asimmetria positiva, se $M_3 < 0$ asimmetria negativa

Indice di curtosi

$B^2 = \frac{\sum_{i=1}^{n}(x_i - m(x))/\sigma)^4}{N}$ B^2 = 3 mesocurtiche, B^2 > 3 distribuzioni più appuntite (leptocurtiche)
B^2 < 3 distribuzioni più piatte (platicurtiche)

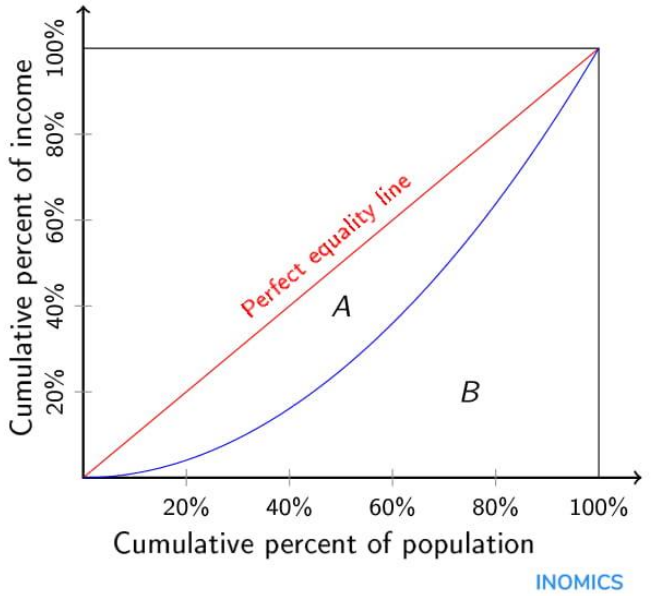
Concentrazione di una variabile trasferibile

Rapporto di concentrazione di Gini

$$R = 1 - \frac{\sum_{i=0}^{N-1} q_i}{\sum_{i=0}^{N-1} p_i}$$

è pari a **0** in presenza di **equidistribuzione** del reddito, cioè tutte le persone hanno la stessa ricchezza.

è pari a **1** in presenza di **massima concentrazione** del reddito, cioè solo una persona detiene tutto la ricchezza.



Rapporti statistici

Distinguiamo 4 tipologie di rapporti:

- **rapporto di composizione**,
- **rapporto di coesistenza**,
- **rapporto di derivazione**,
- **rapporto di densità**.

Relazioni tra due caratteri

Tabella di contingenza (o a doppia entrata) mette in relazione due variabili.

- **Distribuzioni marginali (di riga e di colonna)**
- **Distribuzioni condizionate (di riga e di colonna)**

Relazioni tra due caratteri

Date due variabili statistiche X e Y, **l'indice di correlazione di Pearson** è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili.

$$\rho = \frac{Cov(XY)}{\sigma_x \, \sigma_y}$$

Se $\rho > 0$ si ha correlazione positiva tra X e Y

Se $\rho = 0$ non si ha correlazione tra X e Y

Se $\rho < 0$ si ha correlazione negativa tra X e Y

Se $0 < |\rho_{xy}| < 0{,}3$ si ha correlazione debole tra X e Y

Se $0{,}3 < |\rho_{xy}| < 0{,}7$ si ha correlazione moderata tra X e Y

Se $|\rho_{xy}| > 0{,}7$ si ha correlazione forte tra X e Y

Relazioni tra due caratteri

Indipendenza in distribuzione (a partire da una tabella a doppia entrata).

- **Indice di associazione del Chi quadrato di Pearson** facendo ricorso ai quadrati delle contingenze, divise per la frequenza teorica.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{c_{ij}^2}{n_{ij}^*}$$

Relazioni tra due caratteri

Indipendenza in distribuzione (a partire da una tabella a doppia entrata).

Sempre Karl Pearson ha proposto un altro indice Φ^2 (Phi quadro), o **contingenza quadratica media**.

L'indice è: $\Phi^2 = \frac{\chi^2}{N}$

Che può essere calcolato anche nel seguente modo, che non richiede il calcolo delle frequenze teoriche.

$$\Phi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1$$

Il valore massimo che può assumere è pari al valore più piccolo tra il numero di righe della tabella -1 (k – 1) e il numero di colonne della tabella – 1 (h – 1).

$\max \Phi^2 = \min [(k - 1); (h - 1)]$

Retta di regressione

La regressione è un indicatore statistico che indica l'esistenza o meno di una relazione significativa tra due (analisi bivariata) o più variabili (analisi multivariata) quantitative.

La formula che approssima la relazione tra X e Y con una linea retta è la seguente:

$\hat{Y}_i = \alpha + \beta X_i$

Dove:

$$\beta = \frac{\sum_{i=1}^{N}(x_i - M(X))(y_i - M(Y))}{\sum_{i=1}^{N}(x_i - M(X))^2} = \frac{Codev\ (X,Y)}{Dev\ (X)}$$

$\alpha = M(Y) - \beta\ M(X)$

.

L’indice di determinazione

Dal punto di vista teorico possiamo dire che un modello di regressione è tanto migliore quanto i valori della Y e quelli ottenuti con la retta di regressione hanno una correlazione vicina a 1.

L’indice di determinazione, detto anche coefficiente **R^2** è il quadrato del coefficiente di correlazione lineare fra Y e $\hat{Y}$.

La statistica inferenziale

Definiamo **statistica inferenziale** l'insieme delle teorie e delle tecniche che ci consentono di estendere i risultati di un'indagine di tipo campionario all'intera popolazione di riferimento.

Errore campionario $e = z\sigma(\hat{m}) = z\frac{s}{\sqrt{n}}\sqrt{1-f}$

1 – f è un fattore di correzione per popolazioni finite, dove f è la frazione di campionamento n/N (numerosità campionaria/ampiezza della popolazione).

Se la popolazione è infinita, o comunque quando il campione è inferiore al 5% della popolazione, il fattore di correzione si approssima a 1 e si può trascurare.

La statistica inferenziale

L'errore è tanto più grande:

- Quanto è più elevato il livello di fiducia che vogliamo avere. Se 95% z = 1,96, se 99% z = 2,58, ecc. e = $\mathbf{z}\frac{s}{\sqrt{n}}\sqrt{1-f}$
- Quanto più è elevata la variabilità del fenomeno oggetto di studio -> e = $z\frac{\boldsymbol{s}}{\sqrt{n}}\sqrt{1-f}$
- Quanto più piccola è la numerosità campionaria -> e = $z\frac{s}{\sqrt{\boldsymbol{n}}}\sqrt{1-f}$

La statistica inferenziale

Disegni di campionamento **probabilistici**:

- Campionamento casuale semplice
- Campionamento sistematico
- Campionamento stratificato
- Campionamento a stadi
- Campionamento a grappoli
- Campionamento per aree

Disegni di campionamento **non probabilistici**:

- Campionamento per quote.
- Campionamento a valanga.
- Campionamento a scelta ragionata