LEZIONE 3 — DISTRIBUZIONI DI FREQUENZA

- LA MATRICE di DATI ⇒ è una tabella che presenta i dati raccolti per ogni unità statistica della popolazione relativamente a ciascuna variabile rilevata

RIGHE: dati relativi a una particolare unità statistica

COLONNE: (eccetto la prima se si riporta il numero o l'etichetta dell'unità statistica) rappresentano una variabile

Dalla matrice si possono ottenere
(*) DIVERSE RAPPRESENTAZIONI DEI DATI, la più usata è la TABELLA di FREQUENZA

NB: Molto spesso le variabili QUALITATIVE vengono trasformate in un numero opportuno di VARIABILI NUMERICHE ⇒ si chiama in una variabile indicatore di ciascuna delle modalità possibili

⇓

es: Abbiamo una variabile qualitative con K modalità

⇓

da questa v. qualitativa si generano K-1 VARIABILI DUMMY (variabili indicatore)

⇓

Abbiamo così trasformato a una variabile qualitativa in variabile numerica

- vale 1 se si osserva una certa modalità
- vale 0 altrimenti

vedi slides per ESEMPI

Distribuzioni di frequenze (parte 1)

- I dati raccolti rappresentano l'informazione disponibile su una variabile X
- Come ottenere una visione di insieme del fenomeno?

⇓

Serve aggregare le unità statistiche che presentino modalità identiche

⇓

OVVERO: quante volte ogni singola modalità è presente nella DISTRIBUZIONE DISAGGREGATA

⇓

si deve ristrutturare i dati per avere una visione d'insieme del fenomeno ⇒ costruire una distribuzione che non sia disaggregata ma che abbia un certo tipo di aggregazione

⇓

Useremo le DISTRIBUZIONI di FREQUENZA

OPERATORE SOMMATORIA

$x_1, x_2, x_3, \dots, x_n$ la sommatoria permette di scrivere la somma: $x_1 + x_2 + x_3 + \dots + x_n$ in modo + compatto

$\sum_{i=1}^{N} x_i$ → N → dove finisce; x_i → ARGOMENTO; $i=1$ → da dove parte la sommatoria

ESEMPI

- $\sum_{i=1}^{n} 3x_i = 3 \cdot \sum_{i=1}^{n} x_i$ ⇒ la costante è indipendente (c'è ·)
- $\sum_{i=1}^{n} (2 + x_i)$ = spezziamo la sommatoria dove c'è + = $\underbrace{\sum_{i=1}^{n} 2}_{\text{si somma 2 per n volte}} + \sum_{i=1}^{n} x_i = 2n + \sum_{i=1}^{n} x_i$

NB: $2N = 2+2+2+2+\dots+2 \neq 2^n$

- $\sum_{i=1}^{n} \frac{4}{x_i} = 4 \sum_{i=1}^{n} \frac{1}{x_i} \neq \frac{4}{\sum_{i=1}^{n} x_i}$
- $\sum_{i=1}^{n} \log x_i \neq \log \sum_{i=1}^{n} x_i$

SOMMATORIA DOPPIA

- $\sum_{i=1}^{3} \sum_{j=1}^{4} (2i + 4j) = \sum_{i=1}^{3} \sum_{j=1}^{4} 2i + \sum_{i=1}^{3} \sum_{j=1}^{4} 4j = 2 \sum_{i=1}^{3} \sum_{j=1}^{4} i + 4 \sum_{i=1}^{3} \sum_{j=1}^{4} j =$

($i \neq j$) ⇒ i è come se fosse una costante

$= 2 \sum_{i=1}^{3} (i + i + i + i) - 4 \sum_{i=1}^{3} (1+2+3+4) = 2 \sum_{i=1}^{3} 4i - 4 \sum_{i=1}^{3} 10 =$

$= 8 \sum_{i=1}^{3} i - 4 \cdot 3 \cdot 10 = 8 \cdot (1+2+3) - 120 = 8 \cdot 6 - 120 = -72$

- $\sum_{i=1}^{4} \sum_{j=1}^{3} 2ij = 2 \sum_{i=1}^{4} \left(\sum_{j=1}^{3} ij \right) = 2 \sum_{i=1}^{4} i \sum_{j=1}^{3} j = 2 \sum_{i=1}^{4} i \,(1+2+3) = 2 \sum_{i=1}^{4} i \cdot 6 = 12 \sum_{i=1}^{4} i =$

$= 12 \cdot (1+2+3+4) = 12 \cdot 10 = 120$

_ DISTRIBUZIONI di FREQUENZA _ (parte 2)

- I dati raccolti rappresentano l'informazione disponibile su una variabile X
- Come ottenere una visione di insieme del fenomeno?

⇓ aggregare le unità statistiche in modo che vengano resi disponibili in modo sintetico

↳ aggregare le unità statistiche che presentano modalità identiche

⇓ si conta ogni quante volte ogni singola modalità si presenta nella **DISTRIBUZIONE DISAGGREGATA** (nella nostra grande matrice di dati che raccoglie tutte le informazioni così come sono rilevate)

⇓ si costruiscono a questo scopo le **distribuzioni di frequenze** ⇒ si passa da una disp. disaggr. ad una distribuzione in cui c'è un'aggregazione dei dati in base alle modalità identiche

⇓ da qui si parte con le analisi

INDICHIAMO

- X è una variabile qualitativa / quantitativa
- $x_1, x_2, \dots x_K$ le modalità della variabile ← supponiamo che siano K modalità distinte

Distribuzione di frequenza assoluta

(poco informativa)

- N osservazioni di X
- Modalità $x_1, \dots, x_K$ che si presentano ciascuna con frequenza $n_1, \dots, n_K$

⇓

x_1 si presenta n_1 volte
x_2 si presenta n_2 volte
...
x_K si presenta n_K volte

NB: l'insieme delle coppie costituite da $(x_1, n_1), (x_2, n_2), \dots, (x_K, n_K)$

- x_i: ogni modalità
- n_i: il numero di volte in cui quella modalità viene rilevata

⇓

costituisce la **DISTRIBUZIONE di FREQUENZA ASSOLUTA**

- si scrive: $\{(x_1, n_1), (x_2, n_2), \dots, (x_K, n_K)\}$
- si rappresenta:

X	freq. ass. n_i
x_1	n_1
x_2	n_2
⋮	⋮
x_K	n_K
	N

i: indice che dice che va da 1 a K modalità

la somma di quello che osserviamo deve essere = N

→ N dati che sono stati osservati

modalità della variabile X

corrispondenti valori della freq. assoluta

NB: $N = \sum_{i=1}^{K} n_i$ OSSERVIAMO

Distribuzione di frequenza (relativa e percentuale)

- **FREQ. RELATIVA**: insieme di coppie e le quantità di ciascuna coppia sono $(x_1, f_1) \dots (x_K, f_K)$

permettono di fare confronti

K = numero delle x

x_1 → modalità; f_1 → frequenza relativa

$$f_i = \frac{n_i}{N} = \frac{\text{frequenza assoluta corrispondente}}{N}$$

= quanto è frequente quella particolare modalità

NB: la somma delle frequenze relative:

$$\sum_{i=1}^{K} f_i = \sum_{i=1}^{K} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{K} n_i = \frac{N}{N} = 1$$

$\frac{1}{N}$ costante rispetto alla sommatoria; $\sum_{i=1}^{K} n_i = N$

- **FREQ. PERCENTUALE**: a volte la freq. relativa viene data in termini percentuali

$\{(x_1, f_1\%), \dots, (x_K, f_K\%)\}$

x_1 → modalità; $f_1\%$ → frequenza della modalità in percentuale

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

NB: la somma deve fare 100%:

$$\sum_{i=1}^{K} f_i\% = 100$$

tabella:

X	freq. relativa f_i	freq. percentuale $f_i\%$
x_1	f_1	$f_1\%$
⋮	⋮	⋮
x_K	f_K	$f_K\%$
-	1	100

ATTENZIONE: Fare attenzione agli arrotondamenti nella freq. relativa!

Se si arrotonda bisogna cercare di arrotondare in modo che la somma delle freq. relative faccia 1

- PERCHÈ si ricorre a queste frequenze?
  - x creare una distribuzione aggregata (modo + compatto per partire con i dati)
  - effettuare confronti circa il comportamento della variabile X
- COME scegliere quale frequenza usare? (ASSOLUTA, RELATIVA o PERCENTUALE?)

VANTAGGI e SVANTAGGI

ATTENZIONE:
- si perde la percezione dell'ordine di grandezza di X

Solitamente si usa quella relativa
- permette di valutare la composizione interna delle popolazioni velocemente
- permette di effettuare confronti (riguardo alla distribuzione della variabile x o il comportamento in ≠ gruppi che hanno numerosità ≠)

Distribuzione di frequenza assoluta cumulata

- è costruita cumulando le osservazioni

si parte dalle modalità distinte x_i

per ciascuna modalità x_i si costruisce N_i (somma di tutte le unità statistiche che ci sono fino a quella particolare modalità)

N_i: tutte quelle modalità statistiche che hanno un valore osservato della X minore o uguale rispetto a quella particolare modalità x_i

ci deve essere un certo ordine

infatti: la distribuzione di frequenza di tipo cumulato si può costruire per:
- le variabili QUANTITATIVE (quando le modalità sono dei numeri)
- le variabili ORDINALI (qualitative ordinali)

ATTENZIONE! se è SCALA SCONNESSA NON è ordinabile!

- la freq. assoluta cumulata si costruisce così:

quando si osserva la prima modalità (x_1), si osserva la sua distribuzione di frequenza assoluta (n_1) che è uguale alla frequenza assoluta cumulata (N_1)

$x_1 \longrightarrow N_1 = n_1$

$x_2 \longrightarrow N_2 = n_1 + n_2$

- - - - - - -

$x_K \longrightarrow N_K = n_1 + n_2 + \dots + n_K = N$ totale delle freq. cumulate = tot delle osservazioni N

Distribuzione di frequenze relative cumulate

$F_i = \frac{N_i}{N}$ cioè

(*) Uguale alla freq. assoluta cumulata, basta dividere per N

$F_i = f_1 + f_2 + \dots + f_i$ → anche %

- $x_1 \longrightarrow F_1 = f_1$; $x_1 \longrightarrow F_1\% = f_1\%$
- $x_2 \longrightarrow F_2 = f_1 + f_2$; $x_2 \longrightarrow F_2\% = f_2\%$
- $x_k \longrightarrow F_k = f_1 + f_2 + \dots + f_k = 1$; $x_k \longrightarrow F_k\% = f_k\% = 100$

Da tenere a mente
- la dist. di freq. accumulata si calcola SOLO per variabili almeno ordinali (le qualitative su scale ordinale e per le quantitative)
- l'ultima modalità ha associata la distribuzione (come dist. di freq relativa cumulata) l'ultimo valore che è pari a 1 se relative cumulata e N se la distr. è assoluta cumulata
- le X devono essere ordinate nella tabella in modo crescente
- l'ultima riga che costruiamo nella tabella la costruiamo per avere un'idea dell'ammontare delle unità statistiche

⇓

Nelle colonne di distribuzione cumulate NON si sommano (come la prima colonna)

Rilevazione delle variabili in classi

è molto frequente x variabili come età, PIL, fatturato, investimenti, reddito, tempo, ...

CONSIDERIAMO variabili quantitative di tipo continuo

⇓

chiamiamo CAMPO di VARIAZIONE l'intervallo iniziale di rilevazione

⇓

= l'intervallo costituito dai valori che vanno dal più piccolo valore osservato per le modalità della variabile fino al più grande

⇓

SUDDIVIDIAMO in SOTTOINTERVALLI

e assegnamo le unità statistiche a ciascun intervallo a cui appartengono ⇒ si OTTIENE la rilevazione /una suddivisione in classi della variabile che si chiama SERIAZIONE in INTERVALLI

VANTAGGI
- utile quando la variabile si presenta con molte modalità (+ pratico)
- la sua rappresentazione permette di presentare i dati in modo più compatto (x le rappresentazioni grafiche)

SVANTAGGI
- perdita di informazioni che si ha in quanto i dati non vengono riportati esattamente come sono rilevati, ma si rileva solo la classe a cui appartiene
- "confusione" nella chiusura delle classi (la costruzione delle classi in modo che sia ben definito il passaggio da una classe all'altra)

Quando si decide di rilevare una variabile in classi

⇓

Bisogna scegliere il numero K di classi (è una scelta di tipo pratico

NB° Non dobbiamo dividere le osservazioni in troppe classi, perché c'è il rischio di avere delle classi piccole con frequenza nulla oppure con frequenze estremamente basse

MA non si devono raggruppare le modalità in poche classi perché significa creare delle classi decisamente piene di osservazioni e quindi avere una grande perdita di informazioni

⇓

- perdita di informazioni deve essere ridotta al minimo
- serve persuasione del numero di classi ed efficace presentazione dei dati (poche classi)

⇓

regoletta pratica: $K = \log_2(N) + 1$ per la scelta del numero di classi

- Costruzione della distribuzione a classi
  - Dove serve, le classi possono essere di ampiezza diversa
  - le classi non devono sovrapporsi ⇒ sono una di seguito all'altra
  - Ogni unità statistica entra in una sola classe
  - ATTENZIONE: alla CHIUSURA DELLE CLASSI

tipica scelta: $(c_{i-1}, c_i]$, $c_{i-1} \dashv c_i$ classe aperta a sinistra

$[c_{i-1}, c_i)$, $c_{i-1} \vdash c_i$ classe aperta a destra

$[c_{i-1}, c_i]$

~~$c_{i-1} \dashv c_i$~~ // ~~(c_{i-1}, c_i)~~ non sono mai compresi!

Quantità che si costruiscono per poter continuare le analisi (utili)

- VALORE CENTRALE DELLA CLASSE: valore definito come la semisomma degli estremi

$$\frac{\text{estremo superiore} + \text{estremo inferiore}}{2}$$

$$\bar{m}_i = \frac{c_{i-1} + c_i}{2}$$

- è un valore rappresentativo di quella classe

⇓

la nostra variabile X viene rappresentata tramite delle modalità che sono adesso delle classi

- AMPIEZZA DELLA CLASSE: differenza tra gli estremi

$$d_i = c_i - (c_{i-1})$$

- ci permette di avere un'idea della differenza tra le classi in base a come sono state costruite

- FREQUENZA ASSOLUTA: stabilite le unità statistiche all'interno di ciascuna classe

n_i

NB: Questa va PRIMA di tutte le altre quantità

PRECISAZIONE: Nel momento in cui noi decidiamo di scegliere un'elemento rappresentativo della classe (es valore centrale della classe), molto spesso facciamo un'assunzione

↓

uniforme distribuzione delle classi

SE Non si sa nulla sulla distribuzione di x dentro la classe, allora non c'è altra scelta ⇒

assumiamo che le unità statistiche siano distribuite EQUAMENTE

⇓

Il valore centrale di classe sarà un buon rappresentante della classe

Assunzione: * Supponiamo di dividere la classe $(c_{i-1}, c_i]$ di freq. ass n_i e ampiezza d_i in n_i intervallini uguali di ampiezza $\frac{d_i}{n_i}$

* Posizioniamo una unità statistica al centro di ogni intervallino
* Le rimanenti unità statistiche saranno collocate nell'intervallo tutte alla stessa distanza l'una dall'altra

NB: Come confrontare le classi di ≠ ampiezza?

↓

NON sarà possibile lavorare con le freq. assolute o relative

↓

i risultati sarebbero sbagliati

↓

BISOGNA: considerare insieme alla freq. e l'ampiezza di classe su cui le unità statistiche insistono

⇓

Densità di frequenza di ogni classe $(c_{i-1}, c_i]$

$$f_i = \frac{n_i}{d_i}$$

freq. assoluta corretta per l'ampiezza di classe

oppure

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

(Vedi esercizi ultima slide)