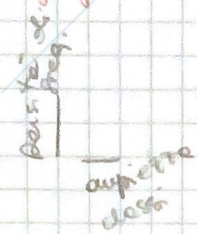
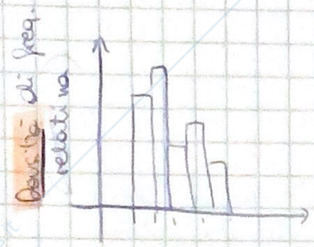
LEZIONE 6 Rapp. grafiche (parte 2)

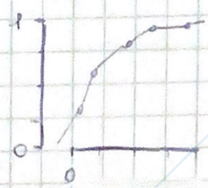
Rappresentazioni grafiche di distribuzioni per variabili in classi

Istogramma





Funzione di ripartizione empirica



Sale tramite segmenti di retta (incremento del fenomeno)

MAX = 1

(useremo molto questo)

• INTRODUZIONE INDICE di POSIZIONE

Primo inx di indici che servono a sintetizzare l'info di una variabile x

In particolare indica la posizione della variabile all'interno dell'inx di riferimento

SCOPO
- sintetizzare (in un unico valore l'informazione che proviene dalla variabile)
- localizzarla nell'inx dei valori di riferimento

Il valore che "sintetizza" l'informazione dà subito un'idea dell'ordine di grandezza di un fenomeno

distingueremo 2 tipi di medie

Media lasca

una delle modalità privilegiata o perché frequente, o perché in una particolare posizione (la sintetizza meglio)

Media analitica

risultato di una manipolazione di dati

può non coincidere con le modalità osservate

NB: Ci sono dei minimi requisiti che una media deve soddisfare x essere tale

$x_1, \dots, x_N$ distribuzione disaggregata per X variabile quantitativa

x_1 → valore minimo (non necessariamente il primo numero)

x_N → valore massimo

M: è una generica media

⇓

deve soddisfare 3 requisiti:

① Principio di Cauchy: $x_1 \le M \le x_N$ deve essere compresa o uguale tra il valore + piccolo osservato e quello + grande

⇓

xk è un valore che deve sintetizzare l'informazione della variabile

② Principio di Chisini (o di invarianza): $f(M, M, \dots, M) = f(x_1, x_2, \dots, x_N)$

la funzione applicata sui dati deve essere uguale (coincide) con la funzione quando a ciascun dato osservato sostituiamo la media

③ Criterio della minima perdita: g (funzione generica) sintetizza la perdita di informazione che abbiamo passando da tutti i dati ad una singola media

M minimizza la perdita g

$\sum_{i=1}^{N} (x_i - M)^{(2)}$ ← le sono nel piano cartesiano

MEDIE LASCHE — 1° gruppo di indici di posizione

Tre tipi
- Moda
- Mediana
- Quartili e quantili

MODA:
- modalità che si presenta + spesso ← caratteristica + imp.
- calcolabile x tutte le variabili
- può non essere unica ... ⇒ poco informativa

NB: la moda corrisponde alla modalità che si presenta con maggior frequenza

- Quando ci sono più mode si ha una distribuzione bimodale (meno informativa)

NB: Nella distribuzione in classi la moda corrisponde alla modalità che si presenta con più densità! $\left(\frac{n_i}{N}\right) \frac{f_i}{d_i}$ e si prende il valore di riferimento della classe

MEDIANA:
- è il valore m associato all'unità statistica che occupa la posizione centrale della distribuzione [prima bisogna ordinare le unità statistiche in ordine crescente rispetto alle modalità]
- le variabili devono essere ordinabili (le variabili qualitative su scala sconnessa NON sono considerate)
- principio di Cauchy $(x_1 \le m \le x_n)$
- la somma dei valori assoluti degli scarti da un valore c è minima se c = mediana

$$\boxed{\sum_{i=1}^{N} |x_i - m| = \min_c \sum_{i=1}^{N} |x_i - c|}$$

(*) Per distribuzioni di frequenza : N DISPARI (tot. osservazioni)

1. ordinare le modalità in senso crescente
2. si cerca l'unità che sta in posizione $(N+1)/2$ ↓ lascia alla sua sx e dx lo stesso numero di unità : $(N-1)/2$
3. la mediana è la modalità posseduta da quella unità ↑

(*) Per distribuzioni di frequenza : N PARI

a) ordinare le modalità in senso crescente

si cercano le unità in posizione $N/2$ e $N/2+1$ che lasciano la prima alla sua sx e la seconda alla sua dx lo stesso numero di ~~modalità~~ unità $N/2-1$

caso di variabile qualitativa ordinale
- se la modalità posseduta dalle unità statistiche è la stessa, quella è la MEDIANA
- se la modalità posseduta dalle unità NON è la stessa la mediana non esiste

caso di variabile quantitativa discreta
- se la modalità posseduta dalle unità è la stessa, quella è la MEDIANA
- se la modalità posseduta dalle unità NON è la stessa, ogni valore compreso tra le due modalità è MEDIANA, di solito si sceglie il valore centrale

(*) Per distribuzioni in classi :
- individuazione della classe mediana $[c_{i-1}, c_i)$
- individuazione della mediana all'interno della classe mediana $[c_{i-1}, c_i)$

POICHÉ:

$$(m - c_{i-1}) : (N/2 - N_{i-1}) = (c_i - c_{i-1}) : n_i$$

SI HA:

$$m = c_{i-1} + \frac{0,5 - F_{i-1}}{f_i} d_i$$

prendo dove supera F_i 0,5

Intervallo generico $[c_{i-1}, c_i)$

m è da qualche parte

c_{i-1} c_i

.... N_{i-1} .. n_i unità stat. .. N_i (freq. assolute cumulate)

$\frac{N}{2}$

⇓

relazione: l'ampiezza di questo intervallo $[c_{i-1}, \frac{N}{2})$ ne un'ampiezza di questo tipo: $(m - c_{i-1})$ sta a tutte le unità statistiche che ci sono in quell'intervallo, come l'ampiezza di tutto l'intervallo sta a tutte le u.s. n_i

$$(m - c_{i-1}) : \left(\frac{N}{2} - N_{i-1}\right) = c_i - c_{i-1} : n_i$$

unica quantità non nota

⇓

$$m - c_{i-1} = \frac{\left(\frac{N}{2} - N_{i-1}\right)(c_i - c_{i-1})}{n_i}$$

⇓

$0{,}5 = \frac{1}{2}$

freq. relativa cumulata $= F_{i-1}$

ampiezza della classe $= d_i$

$$m = \frac{\left(\frac{N}{2} - \frac{N_{i-1}}{N}\right)(c_i - c_{i-1})}{\frac{n_i}{N}} + c_{i-1}$$

$\frac{n_i}{N}$ → freq. relativa: f_i

⇓

$$\boxed{m = c_{i-1} + \frac{(0{,}5 - F_{i-1})}{f_i} \cdot d_i}$$

NB: la mediana è un valore interno alla classe mediana

⇓

controllare sempre che il p.to sia all'interno della classe mediana

QUANTILE

- (estensione della mediana)
- indice associato a particolari posizioni nella distribuzione ordinata di x
- si parla di quantile di ordine q ($0 < q < 1$)

⇓

valore del supporto di x che suddivide la distribuzione in 2 parti

sinistra: q% delle unità stat.

destra: (1-q)% delle unità stat.

⇒ • il quantile prende un nome particolare a seconda del valore di q del suo ordine che si considera

⇓

$\boxed{\text{QUARTILI}}$ (+ famosi) quantili con valori di $q = 1/4, 1/2, 3/4$ (dett. anche primo, secondo, terzo quartile)

secondo = MEDIANA (50%)

- si parla anche di **PERCENTILI** (q multiplo di 100)

o

DECILI (q multiplo di 10)

NB • Dobbiamo avere delle VARIABILI almeno ORDINABILI x poter individuare i quantili

* Quantili x variabili quantitative discrete

- X con valori crescenti: $x_1, \dots, x_N$
- $h = \frac{\ell}{4} N$, $\ell = 1, 2, 3$

ℓ: 1, 2, 3 quantità che identifica l'ordine del quantile ($1/4, 2/4, 3/4$)

N: tot. osservazioni

$q_\ell =$ semisomma di x in posizione h e x in posizione h+1 se h intero

$$\boxed{\frac{x_h + x_{h+1}}{2} \quad \text{per } h \text{ intero}}$$

oppure

$$\boxed{x_{[h]+1} \quad \text{per } h \text{ non intero}}$$

- $[h]$ = parte intera di h

Generale

- Per trovare un **DECILE**: si suddivide la distribuzione in 10 parti e si cerca h

$$\boxed{h = \frac{\ell}{10} \cdot N \quad \ell = 1, \dots, 9}$$ poi si procede come per i quartili

- Per trovare un **CENTILE** si suddivide la distribuzione in 100 parti e si cerca h

$$\boxed{h = \left(\frac{\ell}{100}\right) \cdot N \quad \ell = 1, \dots, 99}$$ poi si procede come per i quartili

Quantili per distribuzioni in classi

- logica = per trovare la MEDIANA

questo valore cambierà x i quantili

$$m = c_{i-1} + \left(\frac{0,5 - F_{i-1}}{f_i}\right) d_i$$

per trovare la mediana

ES. cerchiamo l'ottavo decile (sx 80% US dx 20% U.S.)
= quantile di ordine 0,8

1. Si cerca sulla colonna delle freq. relative cumulate (F_i) dove superiamo l'80% delle U.S., prendiamo la propria CLASSE
2. Si trova l'ottavo decile (d_8)

$$d_8 = \underbrace{c_{i-1}}_{\text{estremo inf.}} + \left(\frac{\overbrace{0,8}^{\text{valore del quantile che stiamo cercando}} - F_{i-1}\ (F_i \text{ della classe prec.})}{f_i\ (\text{della classe del nostro decile})}\right) \cdot \underbrace{d_i}_{\text{ampiezza classe}}$$

$\Downarrow$

$$\boxed{d_n = c_{i-1} + \left(\frac{vd_n - F_{i-1}}{f_i}\right) \cdot d_i}$$

NB: il risultato deve rientrare nella classe da cui siamo partiti

ES 2 cerchiamo il primo quartile

$$\boxed{q_n = c_{i-1} + \left(\frac{vq_n - F_{i-1}}{f_i}\right) \cdot d_i}$$