La statistica descrittiva seconda parte

a cura della *prof.ssa Anna Rita Valente*

INDICI DI POSIZIONE CENTRALE

Sono dei valori che descrivono in modo sintetico una serie di dati raccolti.
I più semplici ma anche i più significativi sono:

- **MEDIA**

ARITMETICA SEMPLICE (n valori)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

ARITMETICA PONDERATA (n valori, f frequenze)

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n}{n}$$

- **MODA**

È il valore a cui corrisponde la frequenza massima

- **MEDIANA**

è la modalità che occupa la posizione centrale nella distribuzione ***ordinata*** della variabile

Se n è dispari allora è il valore centrale; Se n è pari allora è la media aritmetica dei due valori centrali

ESEMPI

N 1 – Calcola la media aritmetica dei seguenti dati:

74 – 87 – 92 – 81 -83 - 75 – 69 – 90 – 72 - 83

N 2 – Dopo aver costruito la tabella della distribuzione di frequenze assolute dei seguenti dati, calcola la media ponderata.

23 – 15 – 18 – 22 – 13 – 19 – 24 – 18 – 20 – 16 – 20 – 21
16 – 19- 18 – 21 – 22 – 24 - 16 – 20 – 17 – 19 – 20 - 16

N 3 – Determina la media, la moda e la mediana dei seguenti valori assunti da una variabile statistica x:

15 18 16 17 15 15 19 18 20 15 17 19 20

INDICI DI VARIABILITA’

▶ CAMPO DI VARIAZIONE

E’ la differenza tra il valore massimo e quello minimo

$$Campo\ di\ variazione = x_{max} - x_{min}$$

▶ SCARTO SEMPLICE MEDIO

E’ la media aritmetica dei valori assoluti delle differenze tra ogni singolo valore della sequenza e la media aritmetica M della sequenza

$$sm = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ ...}{n}$$

$\bar{x} = \text{è}\ la\ media\ dei\ dati\ , n = numero\ di\ dati$

MISURA DI QUANTO LE MODALITA’ DI UN CARATTERE SI DISCOSTANO DALLA LORO MEDIA

prof.ssa ANNA RITA VALENTE

VARIANZA

E’ la media aritmetica dei quadrati della differenza tra ogni singolo valore della sequenza e la media aritmetica della sequenza:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ \dots}{n}$$

La varianza è una misura di variabilità di una distribuzione, cioè quanta "diversità" c'è tra le modalità del fenomeno che stiamo studiando.

Se si ha una **distribuzione di frequenze** la formula è:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 \cdot f_1 + (x_2 - \bar{x})^2 \cdot f_2 + \ \dots}{f_1 + f_2 + \ \dots . + f_n}$$

La varianza è sempre maggiore o uguale a zero, e non può mai essere negativa!

▶ SCARTO QUADRATICO MEDIO (o DEVIAZIONE STANDARD)

E' la radice quadrata della VARIANZA

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ \dots}{n}}$$

E' un indice più utilizzato rispetto allo scarto semplice medio sm, perché più sensibile alle piccole variazioni delle distribuzioni: è più preciso nel determinare quanto i valori sono dispersi rispetto al valore centrale. Come per la media, se abbiamo una **distribuzione di frequenze** la formula è:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \bar{x})^2 \cdot f_1 + (x_2 - \bar{x})^2 \cdot f_2 + \ \dots}{f_1 + f_2 + \ \dots . + f_n}}$$

ESEMPIO

N 1 – Uno studente possiede una piccola biblioteca composta da 250 libri; se si suddividono i libri rispetto al numero di pagine, si ottiene la seguente tabella:

NUMERO PAGINE	FREQUENZA LIBRI
0-50	8
51-100	10
101-150	73
151-200	80
201-250	33
251-300	30
301-350	16

Calcola qual è il numero medio di pagine dei libri della biblioteca e lo scarto quadratico medio.

RISOLUZIONE

Il numero medio di pagine dei libri si ottiene come media ponderata considerando come rappresentante di ciascuna delle sette classi di frequenza il suo valore centrale (dato dalla semisomma dei due valori estremi della classe)

$$M = \frac{8 \cdot 25 + 10 \cdot 75{,}5 + 73 \cdot 125{,}5 + \cdots + 16 \cdot 325{,}5}{8 + 10 + 73 + 80 + 33 + 30 + 16} = 180{,}3$$

Per calcolare lo scarto quadratico medio, occorre calcolare la varianza applicando la formula tenendo in considerazione anche le frequenze assolute:

$$\sigma = \sqrt{\frac{(25 - 180{,}3)^2 \cdot 8 + (75{,}5 - 180{,}3)^2 \cdot 10 + \cdots}{250}} = 69{,}29$$

Possiamo dunque concludere che i libri della biblioteca hanno un numero di pagine che, in media, si discosta di circa 69 pagine dal valor medio.

TABELLE A DOPPIA ENTRATA

La statistica di cui ci siamo occupati finora è quella univariata, perché si è occupata di studiare dati provenienti dalle rilevazioni di un solo carattere su una data popolazione; adesso ci occupiamo di vedere come si estendono le nozioni studiate quando i caratteri da analizzare sono due per esempio X e Y. L'obiettivo sarà quello di scoprire e mettere in relazione i dati di X e quelli di Y (ad esempio è possibile studiare il legame che esiste tra il peso e l'altezza di un gruppo di persone,…)

Distribuzioni congiunte e marginali

Su una popolazione formata da 5 amici si sono rilevati due caratteri: l'età (X) e la città di nascita (Y). I dati grezzi possono essere organizzati nella seguente tabella:

NOME	ETA'	CITTA' DI NASCITA
ALBERTO	30	MILANO
MARIA	35	TORINO
GIOVANNI	32	MILANO
PAOLA	30	MILANO
ALESSANDRO	32	ROMA

Costruiamo una **tabella** delle frequenze ma questa volta a **DOPPIA ENTRATA**, che riporti le frequenze con cui si manifestano le varie coppie di modalità osservate. Nella prima colonna riportiamo le modalità di X e nella prima riga le modalità di Y. Riassumendo:

X\Y	Y1	Y2	Y3	....	YN	TOT
X1						F(x1)
X2						F(x2)
X3			F(x3,y3)			F(x3)
...						...
XN						F(xn)
TOT	f(y1)	f(y2)	f(y3)	...	F(yn)	n

FREQUENZA ASSOLUTA DI X E Y (CONGIUNTA)

DISTRIBUZIONE MARGINALE DI X

DISTRIBUZIONE MARGINALE DI Y

NUMERO COMPLESSIVO DI UNITA'

RIPRENDIAMO L’ESEMPIO PRECEDENTE E COSTRUIAMO LA TABELLA A DOPPIA ENTRATA

X\Y	MILANO	TORINO	ROMA	TOTALE
30	2	0	0	2
32	1	0	1	2
35	0	1	0	1
TOTALE	3	1	1	5

DISTRIBUZIONE MARGINALE DI X

DISTRIBUZIONE MARGINALE DI Y

NUMERO COMPLESSIVO DI UNITA’

SE FISSIAMO L’ATTENZIONE SULLA PRIMA RIGA (O COLONNA) DELLA TABELLA POSSIAMO DIRE CHE CIASCUNA DI ESSE, SINGOLARMENTE PRESA, RAPPRESENTA UNA **DISTRIBUZIONE CONDIZIONATA DI Y RISPETTO ALLA MODALITA’ DI X.**

prof.ssa ANNA RITA VALENTE

E’ anche possibile costruire le distribuzioni condizionate relative, ponendo a rapporto le frequenze congiunte con i rispettivi totali di riga o colonna. Esempio, in riferimento all’esempio precedente, la distribuzione condizionata di X rispetto a y1 è:

X	X|y1	Distribuzione condizionata relativa
30	2	2/3
32	1	1/3
35	0	0/3
totale	3	1

PROVA TU …

SONO STATI INTERVISTATI I 20 STUDENTI DI UNA CLASSE E SU CIASCUNO SONO STATI RILEVATI CONGIUNTAMENTE I DUE CARATTERI X: IL SESSO E Y: IL N. DI ORE CHE DEDICANO MEDIAMENTE ALLO STUDIO IN UNA GIORNATA. SI SONO OTTENUTI I DATI NELLA SEGUENTE TABELLA:

X	M	M	F	M	M	F	M	F	M	F	F	M	F	M	F	M	F	M	M	F
Y	1	2	2	4	3	1	2	3	3	4	2	3	4	2	1	2	3	3	3	4

1. COSTRUISCI UNA TABELLA A DOPPIA ENTRATA CHE ORGANIZZI I DATI GREZZI E INDIVIDUA LE DISTRIBUZIONI MARGINALI DI X E Y.
2. DETERMINA LA DISTRIBUZIONE DI X, CONDIZIONATA ALLA MODALITA' «3 ORE DI STUDIO» E LA CORRISPONDENTE DISTRIBUZIONE CONDIZIONATA RELATIVA.
3. DETERMINA LA DISTRIBUZIONE DI Y, CONDIZIONATA ALLA MODALITA' «FEMMINA» E LA CORRISPONDENTE DISTRIBUZIONE CONDIZIONATA RELATIVA.

prof.ssa ANNA RITA VALENTE

Fine seconda parte

Grazie per l'attenzione