Guida delle formule utilizzate

I VALORI CENTRALI

GLI INDICI DI POSIZIONE (pag. 63)

Sono quelli che operano sulle frequenze della distribuzione: moda e mediana.

LA MODA (pag. 64)

È la modalità di una distribuzione che presenta la frequenza (assoluta o relativa), più alta.

Esempio

Num. Auto possedute	Frequenza assoluta
0	15
1	35
2	30
3	15
4	5
Totale	100

La moda è 1, ovvero quella modalità col maggior numero di casi

LA MEDIANA (pag. 67)

La mediana è quella modalità che si colloca a metà della serie ordinata di valori, in modo che metà delle unità presentino valori uguali o inferiori alla mediana e metà valori uguali o superiori.

N è il numero di osservazioni

Bisogna ordinare le osservazioni in modo crescente

Diversi casi:

1. *N é un numero pari*
   Me = la semisomma dei valori che si trovano nella posizione N/2 e N/2 + 1
   Esempio:
   Abbiamo 4 voti: 7, 9, 8, 10 e vigliamo calcolare la mediana
   Ordiniamo i valori 7, 8, 9, 10
   N/2 =4/2 e N72 + 1 = 4/2 + 1.
   La mediana è la semisomma dei valori che corrispondono alla posizione 2 e 3, ovvero 8 e 9. Me = 8,5
2. *N è un numero dispari*
   Me = il valore che si trova nella posizione (N+1)/2
   Esempio:
   Abbiamo 5 voti: 8, 7, 9, 8, 10 e vigliamo calcolare la mediana
   Ordiniamo i valori 7, 8, 8, 9, 10
   N+1/2 =6/2
   La mediana è il valore che corrisponde alla posizione 3. Me = 8
3. *La distribuzione quantitativa è in classi*

Me = $l_m + (0{,}5 - F_{m-1})/(F_m - F_{m-1})\Delta_m$
Dove:
l_m = limite inferiore della classe mediana
F_{m-1} = Frequenza relativa cumulata fino alla classe precedente a quella mediana
F_m = Frequenza relativa cumulata fino alla classe mediana
Δ_m = ampiezza della classe mediana
Esempio

Voto	Frequenze assolute	Frequenze relative	Frequenze relative cumulate
18-20	5	0,17	0,17
21-23	8	0,27	0,43
24-26	10	0,33	0,77
27-29	4	0,13	0,90
30-30 e lode	3	0,10	1,00
Totale	30	1,00	

La classe mediana è 24-26 (dove ricade lo 0,50 come frequenza relativa cumulata)
l_m = 24
F_{m-1} = 0,43
F_m = 0,77
Δ_m = 2
Me = 24 + (0,5 - 0,43) / (0,77 – 0,43) * 2
Me = 24 + (0,07/0,34) * 2 = 24,4

GLI INDICI ANALITICI (pag. 72)

Sono quelli che operano sui valori di una variabile quantitativa attraverso operazioni algebriche: media aritmetica, media geometrica e media armonica

LA MEDIA ARITMETICA (pag. 73)

La media aritmetica è la somma dei valori osservati della variabile X, diviso per il numero di osservazioni.

Formula generica

$$M_x = \frac{x_1 + x_2 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Se ho una distribuzione di frequenze

$$M_x = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

k = numero di classi
N = numero di osservazioni
Esempio

Num. Auto possedute	Frequenza assoluta
0	15
1	35
2	30
3	15
4	5
Totale	100

$$M_x = \frac{(0 * 15) + (1 * 35) + (2 * 30) + (3 * 15) + (4 * 5)}{100} = \frac{160}{100} = 1{,}6$$

Frequenze in classi (pag. 75)

Se la distribuzione di frequenze è aggregata in classi

$$M_x = \frac{\sum_{i=1}^{k} c_i n_i}{N}$$

Dove

c_i = il valore centrale della classe i-esima

Esempio

Classi di voto	Frequenza assoluta
18-22	10
23-26	20
27-30	30
Totale	60

$$M_x = \frac{(20 * 10) + (24{,}5 * 20) + (28{,}5 * 30)}{100} = \frac{1545}{60} = 25{,}75$$

Media aritmetica ponderata (pag. 77)

$$M_x = \frac{\sum_{i=1}^{k} x_i p_i}{\sum_{i=1}^{k} p_i}$$

LA MEDIA GEOMETRICA (pag. 80)

È il valore che sostituito a tutti gli altri valori della distribuzione ne lascia inalterato il prodotto

$$M_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

LA MEDIA ARMONICA (pag. 80)

È la media aritmetica degli inversi dei valori della distribuzione

$$M_{ar} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

I VALORI DI DISUGUAGLIANZA

INDICE DI ETEROGENEITÀ DI GINI (pag. 93)

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

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

Assume valori tra 0 e (k-1)/k.
Assume valore 0 quando una modalità ha frequenza relativa 1 e tutte le altre 0 (massima eterogeneità)
Assume valore (k-1)/k quando tutte le frequenze relative sono uguali a 1/k, dove k è il numero di modalità.

VALORI CARATTERISTICI BASATI SUGLI SCARTI AL QUADRATO (pag. 98)

DEVIANZA (pag. 98)

La devianza è la somma del quadrato degli scarti dalla media

$$Dev(x) = \sum_{i=1}^{N} (x_i - M(x))$$

VARIANZA (pag. 98)

Fisher propose di suddividere la devianza per il numero di osservazioni N, questo valore prese il nome di varianza.

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

SCARTO QUADRATICO MEDIO (pag. 98)

Pearson propose di trasformare la varianza in una grandezza lineare, estraendo la radice quadrata della varianza.

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

INDICE DI ETEROGENEITA' DI GINI (pag. 92)

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

Dove k è il numero di categorie e $f_i = \frac{n_i}{N}$

CAMPO DI VARIAZIONE (pag. 100)

Il campo di variazione, o range, è la differenza tra il valore massimo e il valore minimo di una distribuzione.
Range(x) = max(x) – min(x)

FORMA DI UNA DISTRIBUZIONE (pag. 101)

ASIMMETRIA (pag. 101)

Si parla di asimmetria quando non è possibile individuare un asse verticale che suddivida una distribuzione in due parti specularmente uguali.
Simmetria → Moda = Mediana = Media
Asimmetria positiva → Moda < Mediana < Media
Asimmetria negativa → Moda > Mediana > Media

$$M_3 = \frac{\sum_{i=1}^{N}(x_i - M(x))^3}{N}$$

Se M_3 > 0 si ha asimmetria positiva
Se M_3 = 0 si ha simmetria
Se M_3 < 0 si ha asimmetria negativa

CURTOSI (pag. 101)

$$B^2 = \frac{\sum_{i=1}^{N}(x_i - M(x)/\sigma)^4}{N}$$

-Se è pari a 3 la distribuzione assume una forma NORMALE
-Se è maggiore di 3 la distribuzione ha una forma appuntita rispetto alla normale (distribuzione LEPTOCURTICA)
-Se è inferiore a 3 la distribuzione ha una forma appiattita rispetto alla normale (distribuzione PLATICURTICA)

RAPPORTO DI CONCENTRAZIONE DI GINI (pag. 104)

Misura la disuguaglianza nella distribuzione di una proprietà trasferibile (ad es. reddito, consumi, ecc.)

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

Dove:
p_i = proporzione dei casi sul numero di frequenze totale
q_i = proporzione della quantità posseduta sul totale

STANDARDIZZAZIONE (pag. 105)

$$z_i = \frac{x_i - M(x)}{\sigma}$$

La nuova variabile ottenuta avrà media 0 e varianza 1 e media, moda e mediana coincidenti.

STATISTICA BIVARIATA

INDICE DI ASSOCIAZIONE DEL CHI QUADRATO DI PEARSON (pag. 140)

È un indice per stabilire se ci sia connessione o meno tra due caratteri statistici X e Y qualitativi, ponendo a confronto le frequenze osservate nelle distribuzioni dei due caratteri con le corrispondenti frequenze teoriche che si avrebbero nel caso di loro assoluta indipendenza.

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

L'indice può essere calcolato anche nel seguente modo

$$\chi^2 = (\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1)N$$

L'indice varia tra 0 e, a parità di associazione cresce al crescere di N.

PHI QUADRO O CONTIGENZA QUADRATICA MEDIA (pag. 140)

Pearson ha proposto un indice che non dipenda da N e che prende il nome di contingenza quadratica media.

$$\Phi^2 = \frac{\chi^2}{N}$$

Questo indice si può calcolare così:

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

L'indice Φ^2 varia tra 0, nel caso di indipendenza tra X e Y (associazione nulla) e il valore più piccolo tra il numero delle righe di una tabella meno 1 e il numero di colonne meno 1.

V DI CRAMER (pag. 141)

Per ottenere un valore che vari tra 0 e 1 si può rapportare Φ^2 al suo valore massimo.

L'indice normalizzato V di Cramer si ottiene dalla radice quadrata del rapporto tra l'indice Φ^2e il massimo, ovvero min[(k-1);(h-1)]

$$V = \sqrt{\frac{\Phi^2}{min[(k-1);(h-1)]}}$$

L'indice assume valore 0 in assenza di associazione ovvero perfetta indipendenza tra X e Y e 1 in caso di perfetta dipendenza.

Esempio di calcolo degli indici di associazione

Frequenze osservate

Genere/Investimento	Fondi azionari	Obbligazioni	Azioni	Titoli di stato	Totale
Maschio	4	5	2	1	12
Femmina	1	3	2	2	8
Totale	5	8	4	3	20

Frequenze teoriche

Genere/Investimento	Fondi azionari	Obbligazioni	Azioni	Titoli di stato	Totale
Maschio	3	4,8	2,4	1,8	12
Femmina	2	3,2	1,6	1,2	8
Totale	5	8	4	3	20

Contingenze (valori osservati - valori teorici)

Genere/Investimento	Fondi azionari	Obbligazioni	Azioni	Titoli di stato	Totale
Maschio	1	0,2	-0,4	-0,8	0
Femmina	-1	-0,2	0,4	0,8	0
Totale	0	0	0	0	0

Contingenze al quadrato diviso per il valore teorico

Genere/Investimento	Fondi azionari	Obbligazioni	Azioni	Titoli di stato	Totale
Maschio	0,33	0,01	0,07	0,36	0,76
Femmina	0,50	0,01	0,10	0,53	1,15
Totale	0,83	0,02	0,17	0,89	1,91

Chi quadrato = 1,91

Phi quadrato = 1,91/20 = 0,1

V di Cramer = 0,31

COEFFICIENTE DI CORRELAZIONE LINEARE DI BRAVAIS PEARSON ρ (pag. 154)

Il coefficiente lineare di Bravais-Pearson è un indice della relazione lineare tra X e Y e misura l'interdipendenza lineare tra le due variabili.

$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$ ovvero $\rho = \frac{Codev(X,Y)}{\sqrt{Dev(X)Dev(Y)}}$

L'indice varia tra -1 e 1 e assume i seguenti valori:

ρ =+1 concordanza perfetta tra X e Y

ρ =0 indipendenza lineare tra X e Y

ρ =-1 discordanza perfetta tra X e Y

Esempio di calcolo

Ho i seguenti valori di X (test di ingresso) e Y (voto finale di matematica) di 8 studenti.

Studente	Test	Voto finale
1	12	8
2	10	7
3	14	8
4	9	5
5	9	6
6	13	9
7	11	7
8	8	5

Calcolo il coefficiente di correlazione lineare $\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$

	Test	Voto
Scarto quad. medio	1,98	1,36
Covarianza	2,47	
Rho	0,91	

MODELLO DI REGRESSIONE (pag. 157)

L'equazione di regressione può essere scritta nei seguenti modi:

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$

$Y = \hat{\alpha} + \hat{\beta}X + \varepsilon$

Dove $\hat{\alpha}$ è l'intercetta e $\hat{\beta}$ è il coefficiente angolare della retta.

La prima è la formula della retta stimata, che è uguale alla retta osservata più un certo valore erratico.

$Y = \hat{Y} + \varepsilon$ quindi $\varepsilon = Y - \hat{Y}$

METODO DEI MINIMI QUADRATI (pag. 161)

Si chiama Metodo dei minimi quadrati (Ordinary Least Squares criterion – OLS) il metodo che rende minima la sommatoria dei quadrati degli scarti tra i valori osservati e quelli teorici di Y. Queste sono le formule per calcolare i due parametri $\hat{\alpha}$ *e* $\hat{\beta}$.

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

$\hat{\alpha} = M(Y) - \hat{\beta}(X)$

L'INDICE DI DETERMINAZIONE R^2 (pag. 163)

L'indice di determinazione serve per misurare la bontà di un modello di regressione lineare ed è pari al quadrato del coefficiente di correlazione tra Y e $\hat{Y}$.

$$R^2 = \frac{[Codev(X,Y)]^2}{Dev(X)Dev(Y)}$$

L'indice di determinazione varia tra 0 e 1.

È pari a 1 quando valori teorici e valori osservati sono uguali, Y e $\hat{Y}$ coincidono. Il modello di regressione è quindi il migliore.