PERCORSO AGENZIA DELLE ENTRATE

Statistica descrittiva: La Variabilità

Introduzione

Benvenuti!

In questa lezione procederemo nella nostra trattazione, in quanto gli indici di posizione e analitici introdotti finora non sono sufficienti a sintetizzare le caratteristiche principali contenute in una distribuzione di frequenza. Una caratteristica molto importante a questo proposito è rappresentata dalla variabilità.

In particolare, andremo ad approfondire:

- la variabilità per caratteri quantitativi
- le proprietà della varianza
- e, infine, le distribuzioni doppie

Bene, non ci resta che cominciare…

Il concetto di variabilità

Con il concetto di variabilità si intende l'attitudine del carattere ad assumere modalità diverse su unità diverse.

Per misurare la variabilità di un carattere facciamo inizialmente riferimento agli indici di:

- omogeneità
- eterogeneità

Definiamo il concetto di omogeneità:

- Un collettivo è omogeneo rispetto ad un carattere, od anche che le sue unità sono tutte omogenee tra di loro, se tutte le unità presentano la stessa modalità del carattere.
- L'omogeneità più bassa si ha quando le frequenze relative sono uguali tra di loro.
- Gli indici di omogeneità devono assumere il massimo nel caso di omogeneità ed il minimo nel caso di eterogeneità (vale il viceversa per gli indici di eterogeneità).

Esistono vari indici per calcolare l'omogeneità di un carattere:

- Gli indici di omogeneità richiedono il calcolo delle frequenze relative
- Gli indici relativi di omogeneità consento il confronto tra popolazioni diverse
- Gli indici relativi di omogeneità sono pari all'indice assoluto diviso per il suo massimo
- Gli indici di omogeneità possono essere calcolati per tutte le tipologie di caratteri

Per quanto riguarda il concetto di eterogeneità, invece:

- Un collettivo è eterogeneo rispetto ad un carattere se le unità NON presentano tutte la stessa modalità del carattere
- L'eterogeneità più bassa si ha quando le frequenze relative sono uguali tra di loro e pari a 0, tranne una che sarà pari a 1
- Gli indici di eterogeneità devono assumere il massimo nel caso in cui le frequenze relative sono tutte uguali tra loro ed il minimo nel caso di omogeneità

Esistono vari indici per calcolare l'eterogeneità di un carattere:

- Gli indici di eterogeneità richiedono il calcolo delle frequenze relative
- Gli indici relativi di eterogeneità consento il confronto tra popolazioni diverse
- Gli indici relativi di eterogeneità sono pari all'indice assoluto diviso per il suo massimo
- Il più utilizzato indice di eterogeneità è l'entropia (relativa)

Gli indici di eterogeneità possono essere calcolati per tutte le tipologie di caratteri:

- Basandosi solo sulle frequenze relative, non tengono conto di tutte le informazioni contenute nei caratteri quantitativi

La variabilità per caratteri quantitativi

Gli indici di variabilità per caratteri quantitativi danno una misura della dispersione dei termini della distribuzione rispetto ad una media o di quanto differiscono tra loro le modalità presenti nelle unità, basandosi su misure delle diversità tra due modalità.

- Gli scostamenti medi sono indici di variabilità rispetto ad un indice di posizione o ad un indice di analitico:
    - lo scostamento semplice dalla mediana non è nient'altro che la media aritmetica degli scarti tra le modalità e il valore mediano in valore assoluto
    - lo scostamento semplice medio dalla media aritmetica è la media aritmetica degli scarti tra le osservazioni e la media

Fate attenzione, che la proprietà della media aritmetica ci dice che la somma degli scarti dalla media aritmetica fa zero, ma in questo caso noi abbiamo il valore assoluto degli scarti e quindi sappiamo che questo sarà sicuramente un valore maggiore o uguale a zero.

Entrambi gli scostamenti medi assumono il loro valore minimo, cioè zero, in corrispondenza di assenza di variabilità, cioè nel caso in cui tutte le unità avessero la stessa modalità, che ovviamente sarebbe pari alla media e anche alla mediana.

- Nel caso delle differenze medie invece andiamo a misurare la variabilità in intesa come differenza tra le modalità assunte tra le diverse unità

Possiamo calcolare quindi per ogni unità quanto essa assuma un valore diverso da ciascuna delle altre unità del collettivo e successivamente ne facciamo una media. Come risultato otteniamo le differenze medie che possono essere:

- senza ripetizione, laddove la differenza tra i valori appartenenti alla stessa modalità non sono considerati
- con ripetizione, quando il numero di differenze considerate tiene conto anche della differenza con se stessi
- le due differenze differiscono solamente per il denominatore

Tra gli indici di variabilità che confrontano le modalità osservate con la media aritmetica, il più noto è senza dubbio:

- la varianza che, generalmente, si indica con il simbolo sigma quadro
- la varianza è la media dei quadrati degli scarti dalla media aritmetica
- la varianza può essere anche definita come la media dei quadrati meno il quadrato della media
- la varianza è zero laddove fossimo in presenza di assenza di variabilità, cioè quando tutte le modalità sono uguali tra loro e pertanto uguale alla media

La varianza gode di alcune proprietà:

- la varianza è sempre non negativa, in quanto media di somma di quadrati, quindi numeri sempre positivi
- il suo valore cresce all'aumentare della variabilità
- il suo valore si annulla in caso di assenza di variabilità
- gode di una proprietà particolare:
    - nel caso di trasformazioni lineari, la nuova varianza dipende solamente dal parametro relativo al cambiamento dell'unità di misura, quello che nella relazione y = a + bx è rappresentato dal termine b

Esistono altri indici di variabilità:

- il campo di variazione, cioè la differenza tra la modalità massima in modalità minima, anche noto come range
- la differenza interquartile, cioè la differenza tra il terzo e il primo quartile

Rimarchiamo che tutti questi indici, implicando operazioni matematiche, è possibile calcolarli solamente per caratteri di tipo quantitativo!

Statistica descrittiva: Distribuzioni doppie

Finora abbiamo analizzato e sintetizzato le informazioni di un carattere alla volta.

Andiamo ora a vedere come rappresentare e sintetizzare le informazioni di due caratteri analizzati simultaneamente.

Il punto di partenza è dato dalla distribuzione doppia di frequenza.

In una distribuzione doppia di frequenza:

- le modalità del carattere X (Sesso, nel nostro esempio) sono elencate sulle righe

- le modalità del carattere Y (Occupazione, nel nostro esempio) sono elencate sulle colonne

Esistono vari tipi di “distribuzioni” a caratterizzare una distribuzione doppia di frequenza.

- La distribuzione congiunta ci fornisce informazioni sul verificarsi simultaneamente di una coppia di modalità, una per X e una per Y.
- Le distribuzioni marginali (una per ogni carattere) forniscono informazioni su un carattere alla volta, esattamente come discusso in precedenza.
- Le distribuzioni condizionate (di riga e di colonna) forniscono informazioni su di un carattere fissando, cioè condizionatamente, ad una specifica modalità dell’altro carattere

Indichiamo con n_{ij} le frequenze assolute congiunte, cioè il numero di volte che una coppia di modalità si presenta nella nostra popolazione

- Le frequenze assolute per il carattere X, cioè la sua distribuzione marginale, sono date dalla somma delle frequenze congiunte su tutte le colonne
- Le frequenze assolute per il carattere Y, cioè la sua distribuzione marginale, sono date dalla somma delle frequenze congiunte su tutte le righe

Per i caratteri quantitativi, accanto alle medie e alle varianze marginali (cioè per ogni singolo carattere), è possibile calcolare:

- Medie condizionate, ad esempio la media del carattere Y condizionata ad una specifica modalità di X
- Varianze condizionate, ad esempio la varianza del carattere Y condizionata ad una specifica modalità di X, che esprimono la variabilità intorno alle medie delle unità delle distribuzioni condizionate

Così come visto in precedenza, le medie e le varianze condizionate godono di alcune proprietà:

- La media marginale può essere ottenuta sfruttando la proprietà associativa della media aritmetica, come media delle medie, in cui i pesi delle medie condizionate sono dati dalle frequenze
- La varianza marginale NON può essere ottenuta come media delle varianze condizionate. La proprietà associativa non vale per la varianza
- La varianza marginale è pari alla somma di due quantità:
    - Varianza within, cioè la media delle varianze condizionate
    - Varianza between, cioè la “varianza” tra le medie condizionate o, più precisamente, la media degli scarti al quadrato tra la media marginale e le medie condizionate, anche in questo caso pesato con le frequenze

Conclusioni

Con questo abbiamo concluso anche questa seconda video lezione.

Ti ricordo che abbiamo parlato di:

- variabilità e dei suoi caratteri quantitativi
- ed infine di statistica descrittiva, introducendo le distribuzioni doppie

Alla prossima video lezione!