PERCORSO AGENZIA DELLE ENTRATE

I principali Indici di associazione

Introduzione

Benvenuti!

In questa lezione parleremo dei principali indici di associazione.

In particolare, andremo ad approfondire:

- la dipendenza tra due caratteri
- le coppie di caratteri quantitativi

Bene, non ci resta che cominciare…

La dipendenza tra due caratteri

Nell'analizzare coppie di carattere congiuntamente, una caratteristica fondamentale di ogni analisi dei dati è l'analisi della dipendenza tra i due caratteri, cioè il possibile legame tre i due caratteri.

Introduciamo due concetti "estremi" legati all'associazione:

- massima dipendenza
- indipendenza

Due caratteri sono massimamente dipendenti quando le distribuzioni di uno, condizionate alle modalità dell'altro, sono massimamente diverse. In altre parole, Y dipende perfettamente da X se conoscendo le modalità di X posso *predire* con certezza le modalità di Y.

Dal punto di vista pratico:

- la dipendenza perfetta implica che in una tabella doppia per ogni riga *i* c'è un solo una colona *j* per la quale le frequenze congiunte $n_{ij} > 0$
- Tra due caratteri sussiste interdipendenza perfetta se ad ogni modalità di uno dei due caratteri corrisponde una ed una sola modalità dell'altro carattere e viceversa

Due caratteri sono detti indipendenti se le distribuzioni di uno, condizionate alle modalità dell'altro, hanno le stesse frequenze relative o percentuali.

Dal punto di vista pratico:

- due caratteri sono indipendenti se e solo se la generica frequenza assoluta congiunta n_{ij} corrispondente alla i-esima modalità di X e alla j-esima di Y è uguale al prodotto della frequenza marginale della riga *i* e della colonna *j* diviso per la numerosità del collettivo

In tutti gli altri casi i due caratteri saranno detti dipendenti, avranno cioè un certo livello di associazione o dipendenza non nullo.

Se nessuno dei due casi estremi appena citati viene osservato, allora abbiamo la necessità di misurare il livello di associazione/dipendenza.

I principali indici di associazione sono:

- il chi-quadrato
- la V di Cramer

Il **chi-quadrato**:

- è un indice assoluto
- misura la distanza che c'è tra le frequenze congiunte osservate e quelle teoriche sotto il caso di indipendenza, cioè ci dice quanto ci scostiamo dal caso di indipendenza
- assume valore minimo, pari a 0, nel caso di indipendenza
- assume valore massimo, pari alla numerosità del collettivo per il minimo tra il numero di righe meno uno e il numero di colonne meno 1, nel caso di massima dipendenza
- dipende dalla numerosità del collettivo e dal numero di modalità osservate per i due caratteri

La **V di Cramer**:

- è un indice relativo
- è funzione del chi-quadrato, ed è pari alla radice quadrata del chi-quadrato diviso per il suo massimo
- assume valore minimo, pari a 0, nel caso di indipendenza
- assume valore massimo, pari a 1, nel caso di massima dipendenza

Entrambi gli indici si basano solamente sulle frequenze (assolute) congiunte, non tenendo conto in alcun modo delle modalità. Come abbiamo appreso in precedenza, questo è un limite quando si analizzano coppie di caratteri quantitativi.

Le coppie di caratteri quantitativi

Per coppie di caratteri quantitativi, ci concentriamo ora su distribuzioni doppie unitarie per semplicità di trattazione.

I principali indici di associazione/dipendenza per coppie di caratteri quantitativi sono:

- la covarianza
- la correlazione lineare

In particolare, la **covarianza**:

- misura il legame lineare tra due caratteri quantitativi X e Y
- è pari alla media aritmetica del prodotto degli scarti di due caratteri dalle loro rispettive medie
- può essere calcolato, ed è più facile da ricordare, come la media dei prodotti meno il prodotto tra le medie
- è un indice assoluto
- varia tra meno il prodotto tra gli scarti quadratici medi e più il prodotto tra gli scarti quadratici medi
- quando scarti positivi (negativi) del carattere X tendono ad associarsi a scarti positivi (negativi) del carattere Y, allora i loro prodotti saranno positivi e la covarianza risulterà positiva; quando scarti positivi del carattere X tendono ad associarsi a scarti negativi del carattere Y (o viceversa), allora i loro prodotti saranno negativi e la covarianza risulterà negativa

La **correlazione lineare**, invece:

- è un indice che misura la relazione lineare tra due caratteri quantitativi X e Y
- è espresso dal rapporto tra la covariaza tra i due caratteri X e Y ed il prodotto dei rispettivi scarti quadratici medi, cioè la covarianza divisa per il suo massimo
- il coefficiente di correlazione è compreso tra -1 e 1
- se pari a zero, allora non vi è relazione di tipo lineare tra i due caratteri
- si noti che l'incorrelazione tra due caratteri implica correlazione nulla, ma non è vero il contrario
- se pari a 1, allora esiste un legame lineare perfetto positivo
- se pari a -1, allora esiste un legame lineare perfetto negativo
- è invariante per trasformazioni lineari, a meno del segno

Fissiamo meglio le idee. Date due variabili quantitative, diremo che sono:

- correlate positivamente se variano in modo concorde, ossia all'aumentare [diminuire] dell'una aumenta [diminuisce] anche l'altra
- correlate negativamente se variano in modo discorde, ossia all'aumentare [diminuire] dell'una, l'altra diminuisce [aumenta]

Osserviamo che due caratteri risultano concordi se gli scarti dalla media tendono ad essere dello stesso segno, mentre risultano discordi se tali scarti tendono ad essere di segno opposto.

Passiamo ora a studiare la **regressione**. Obiettivo dell'analisi di regressione è studiare il legame di causa-effetto che intercorre tra due variabili quantitative X (detta variabile indipendente) e Y (detta variabile dipendente).
Il legame tra due variabili viene espresso mediante una funzione del tipo lineare:

- y = beta0 + beta1*x
    - beta0: valore di y per x=0
    - beta1: variazione di y per un aumento unitario di x

Nella realtà difficilmente due variabili sono legate da una relazione esatta. Per ovviare a questo inconveniente adottiamo il modello:

- y_i = beta0 + beta1*x_i + eps_i
    - beta0 = intercetta
    - beta1 = coefficiente di regressione (pendenza della retta)
    - eps_i $ = residuo o errore (riflette le imperfezioni della relazione lineare ed eventuali variabili esplicative omesse)

Per capire la relazione di causa-effetto che lega y ed x, è necessario stimare i parametri beta0 e beta1. Le loro stime si ottengono attraverso il metodo dei minimi quadrati, cioè minimizzando la parte residuale (al quadrato).
Dal metodo dei minimi quadrati si ottiene che:

- beta1 è pari alla covarianza divisa per la varianza della variabile indipendente
- beta0 è pari alla media della variabile dipendente (y) meno il coefficiente di regressione (beta1) moltiplicato per la media della variabile indipendente (x)

Una volta stimati i parametri del modello, vogliamo sapere quanto bene il modello approssimi i dati. Per farlo, partiamo dalla formula di scomposizione della **varianza**:

- la varianza totale si può scomporre come la somma di varianza spiegata dal modello e varianza residua
- la varianza spiegata dal modello è la varianza dei valori stimati dal modello
- la varianza residua è la varianza non spiegata dal modello, cioè la variabilità dei residui (le differenze tra i valori stimati e quelli osservati)

Per avere un indice sintetico della bontà di adattamento del modello ai dati, calcoliamo il rapporto tra variabilità spiegata dalla regressione e variabilità totale che prende il nome di **coefficiente di determinazione:**

- assume valore minimo (pessimo adattamento) a zero
- assume valore massimo (perfetto adattamento, i punti sono allineati lungo una retta) a 1
- è anche pari al quadrato del coefficiente di correlazione lineare

Conclusioni

Bene, con questo siamo giunti alla fine anche di questa video lezione.

Ti ricordo che abbiamo introdotto:

- il chi-quadrato
- la V di Cramer
- la covarianza
- le correlazioni lineari
- la regressione
- la varianza
- e, infine, il coefficiente di determinazione

Grazie per l'attenzione!