Lezione 19/10/2022

Correlazione lineare tra due caratteri (esercizio)

In 4 supermercati (che chiamo A, B, C e D) di una nota catena sono stati rilevati la superficie di esposizione, in migliaia di metri quadrati, e il fatturato settimanale, in migliaia di euro. Sono stati ottenuti i seguenti dati.

	A	B	C	D
x_i (superficie)	0.2	0.5	0.8	1
y_i (fatturato)	50	120	150	200

Studia la correlazione lineare tra i due caratteri.

Il primo metodo per capire se esiste una correlazione lineare tra i caratteri $X = superficie$ e $Y = fatturato$ è di disegnare lo scatter plot o diagramma di dispersione. Ogni supermercato viene rappresentato nel piano cartesiano da un punto le cui coordinate sono le modalità associate ad esso:

$$A : (0.2, 0.5),\ B : (0.5, 120),\ C : (0.8, 150) \text{ e } D : (1, 200).$$

I punti si rappresentano nel piano cartesiano e si osserva se essi si distribuiscono lungo una linea retta (vedi il grafico negli appunti). Una volta osservato che esiste una corrispondenza lineare si determina se è crescente (o diretta) oppure decrescente (inversa). Nel caso di questo esercizio, la corrispondenza è crescente.

Ci chiediamo: Quanto è forte la corrispondenza? Ad occhio (guardando il grafico) mi accorgo che la corrispondenza è forte, ma per quantificala matematicamente uso la covarianza oppure il coefficiente di correlazione lineare.

$$\sigma_{xy} = 16, 25.$$

Il fatto che σ_{xy} è un numero positivo, conferma che tra X e Y c'è una correlazione lineare crescente. Per sapere quanto la correlazione è forte devo sapere quanto σ_{xy} si avvicina al prodotto delle deviazioni standard σ_x e σ_y.

Calcolo $\sigma_x \cdot \sigma_y = 0.30 \cdot 54 = 16, 29$. Dato che σ_{xy} è molto vicina a $16, 29$ allora posso dire che la correlazione lineare tra X e Y è molto forte.

Il coefficiente di correlazione lineare è un altro indice, alternativo alla covarianza, che consente di capire se esiste, quanto è forte e di che tipo è la correlazione lineare tra X e Y.

$$R = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \sim 0.99$$

Il fatto che R è un numero positivo, conferma che tra X e Y c'è una correlazione lineare crescente. Dato che R è molto vicino a 1, la correlazione lineare è molto forte.

Dopo aver capito che tra X e Y esiste una correlazione lineare forte di tipo crescente, voglio calcolare la retta di regressione che è la retta che meglio si avvicina a tutti i punti A, B, C e D.

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x + \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x} = 176,87x + 19,46.$$

Osservazione: il segno del coefficiente angolare della retta di regressione ($\frac{\sigma_{xy}}{\sigma_x^2}$) è uguale al segno della covarianza ($\sigma_{x,y}$). Dalla geometria analitica, sappiamo inoltre che

- **se il coefficiente angolare della retta è positivo allora la retta è crescente**
- **se il coefficiente angolare della retta è negativo allora la retta è**
  decrescente.

Questo è coerente con il significato della covarianza (covarianza positiva=retta crescente e covarianza negativa=retta decrescente).

Variabili aleatorie

Definizione. *Sia Ω lo spazio campionario di un esperimento casuale, una* variabile aleatoria (o casuale) *è una funzione $X : \Omega \to R$ che associa un numero reale a ogni possibile risultato dell'esperimento.*

Esempio. *Esperimento: lancio di due dadi. Considero la variabile aleatoria che associa a ogni risultato (x, y) la somma di x e y. Dunque, $X((1,1)) = 2$, $X((2,1)) = 3$, $X((3,4)) = 7$, $X((1,6)) = 7$, ...*

Esempio. *Esperimento: lancio due monete. Considero la variabile aleatoria che associa a ogni risultato (x, y) il numero di teste. Dunque, $X((T,T)) = 2$, $X((T,C)) = 1$, $X((C,T)) = 1$, $X((C,C)) = 0$.*

Chiamiamo $X(\Omega)$ l'immagine di X.
Negli esempi precedenti $X(\Omega) = \{2, \ldots, 12\}$ e $X(\Omega) = \{0, 1, 2\}$.

Una variabile aleatoria si può rappresentare con una tabella.

Esempio. *Esperimento: lancio due monete. Considero la variabile aleatoria che associa a ogni risultato (x, y) il numero di teste.*

Ω	(T,T)	(T,C)	(C,T)	(C,C)
X	2	1	1	0

Le variabili aleatorie possono essere continue o discrete.
Una

Definizione. *variabile aleatoria che può assumere solo un numero finito di valori o un'infinità numerabile di valori è detta* ***variabile aleatoria discreta****, mentre una variabile aleatoria che assume un'infinità non numerabile di valori è detta* ***continua****.*

In parole semplici, se $X(A)$ dove $A \subseteq \Omega$ è sempre un numero naturale o 0 allora X è discreta, se $X(A)$ può essere un qualsiasi numero reale (-1, 1,2, $\sqrt{3}$, ...) allora X è una variabile continua.

Esempi di variabili aleatorie discrete:

- X= Numero di teste nel lancio di due monete;
- X= somma dei risultati dei lanci di due dadi;
- X= voto di uno studente estratto a caso in una classe;
- X= numero di palline rosse estratte da un'urna in 8 estrazioni con rimpiazzo;
- X= numero di lanci di una moneta da effettuare affinchè esca per la prima volta testa.

Esempi di variabili aleatorie continue:

- X= altezza di uno studente estratto a caso in una classe ($X \in [140cm, 200cm]$);
- X= temperatura misurata a Varese in un momento a caso ($X \in [8\ gradi, 30\ gradi]$);
- X=peso di un pacco di biscotti prodotto da una fabbrica ($X \in [500g, 1000g]$).

In generale, se X è una variabile aleatoria, si usano notazioni del tipo seguente

Evento "X assume il valore a": $X = a$;

Evento "X assume valori compresi nell'intervallo (a, b)": $a < X < b$;

Evento "X assume valori minori o uguali a c": $X \leq c$.

Variabili aleatorie discrete

D'ora in poi ci occupiamo di variabili aleatorie discrete.

Definizione. *Sia $X : \Omega \to R$ una variabile aleatoria, si chiama* distribuzione di probabilità *di X la funzione $p : R \to [0, +\infty[$ tale che*

$$p(k) = P(X = k)$$

per ogni $k \in R$.

Esempio. *Se considero la variabile aleatoria X che conta il numero di teste nel lancio di due monete, allora $p(0) = p(2) = \frac{1}{4}$, $p(1) = \frac{1}{2}$ e $p(k) = 0$ per ogni $k \notin \{0, 1, 2\}$.*

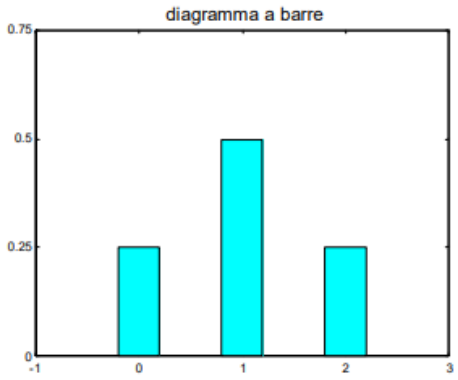
La distribuzione di probabilità di una variabile aleatoria discreta si può rappresentare mediante una tabella, un diagramma a barre o a un istogramma.

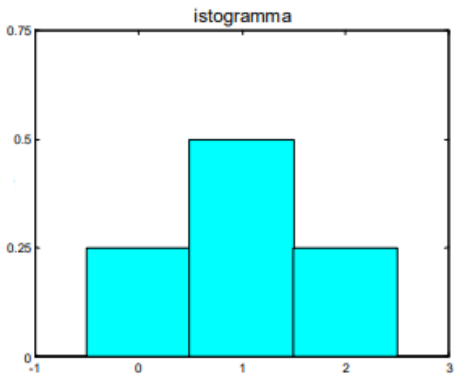
Esempio. *Consideriamo la variabile aleatoria che esprime il numero di teste nel lancio di due monete.*

La tabella riporta nella prima riga i valori di $X(\Omega)$ e nella seconda riga le rispettive probabilità.

k	0	1	2
$p(k)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Il diagramma a barre e l'istogramma riportano sull'asse orizzontale i valori di $X(\Omega)$ a cui corrisponde un rettangolo. Nel caso del diagramma a barre l'altezza del rettangolo associato a k è uguale alla probabilità $P(X = k)$. Nel caso dell'istogramma l'area del rettangolo di k è uguale a $P(X = k)$. Se inoltre supponiamo che le basi dei rettangoli dell'istogramma misurino 1 allora $P(X = k)$ è anche l'atezza dei rettangoli.





Proprietà delle variabili aleatorie discrete

1. $P(X = k) = 0$ se $k \notin X(\Omega)$;
2. Se $X(\Omega) = \{k_1, \ldots, k_n\}$, allora $P(X = k_1) + \ldots + P(X = k_n) = 1$.

Esempio. *Consideriamo la variabile aleatoria X che esprime il numero di teste nel lancio di due monete, allora $X(\Omega) = \{0, 1, 2\}$.*

$P(X = -3) = P(X = 1, 2) = 0$ e $P(X = k) = 0$ per ogni k diverso da 0, 1 e 2.

Inlotre, $P(X = 0) + P(X = 1) + P(X = 2) = \dfrac{1}{4} + \dfrac{1}{2} + \dfrac{1}{4}$.