PERCORSO AGENZIA DELLE ENTRATE

Variabili aleatorie e Modelli probabilistici

Introduzione

Benvenuti!

In questa lezione parleremo di variabili aleatorie e modelli probabilistici.

In particolare, andremo ad approfondire:

- la definizione di variabili aleatorie discrete e continue
- la definizione del valore atteso e della varianza di una variabile aleatoria, e della funzione di ripartizione
- i modelli probabilistici per variabili aleatorie discrete: Bernoulli, Binomiale, Poisson
- i modelli probabilistici per variabili aleatorie continue: Normale (o Gaussiana)

Bene, non ci resta che cominciare…

Le variabili aleatorie

Formalmente, una variabile aleatoria è una funzione definita sullo spazio campionario S che associa un numero reale ad ogni evento elementare.
Possiamo però rendere questa definizione più intuitiva e di facile comprensione. Infatti, possiamo definire una variabile aleatoria:

- una variabile che assume determinati valori (modalità o realizzazioni) con una certa probabilità

Le variabili aleatorie, così come i caratteri quantitativi precedentemente studiati nella parte di statistica descrittiva, possono essere divise in due tipologie:

- Variabili aleatorie discrete
- Variabili aleatorie continue

Le variabili aleatorie sono definite discrete se:

- l'insieme delle possibili realizzazioni è finito o numerabile

Le variabili aleatorie sono definite continue se:

- l'insieme delle possibili realizzazioni è un intervallo della retta reale

Esempio: lancio una moneta due volte

I possibili risultati sono (TT,CT,TC,CC), ciascuno con probabilità 1/4

Definiamo la variabile aleatoria X = (numero di volte che esce testa).
Le possibili realizzazioni della variabile aleatoria sono:

- 0 con probabilità 1/4
- 1 con probabilità 2/4
- 2 con probabilità 1/4

Abbiamo così ottenuto la distribuzione di probabilità di una variabile aleatoria discreta.
Cogliamo l'occasione per ribadire che la somma delle probabilità deve essere pari a 1.

Valore atteso e varianza di una variabile aleatoria discreta

Così come per i caratteri quantitativi visti nella parte di statistica descrittiva, introduciamo semplici misure riassuntive delle variabili aleatoria. In particolare, facciamo riferimento a:

- Valore atteso
- Varianza

Il valore atteso è:

- la media aritmetica delle realizzazioni pesata con le rispettive probabilità

La varianza è:

- la media aritmetica degli scarti al quadrato pesata con le rispettive probabilità

Hanno lo stesso nome, significato e proprietà degli indici visti in statistica descrittiva. L'unica differenza è che ora il sistema di pesi non è dato dalle frequenze, ma dalle probabilità.

Il valore atteso e la varianza di una variabile aleatoria godono delle stesse proprietà viste per la media e la varianza in statistica descrittiva.
Qui di seguito riportiamo le proprietà per trasformazioni lineare. Sono del tutto identiche a quelle che abbiamo già visto in statistica descrittiva:

$$E(a+bX) = a + bE(X), \quad V(a+bX) = b^2V(X)$$

Un'importante trasformazione lineare è la standardizzazione.
Standardizzare una variabile equivale a sottrarre dai valori osservati il valore atteso della variabile aleatoria e dividere per la radice della varianza. La variabile standardizzata, convenzionalmente indicata con la lettera Z, ha valore atteso 0 e varianza pari a 1:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - \mu}{\sigma}$$
$$E(Z) = \frac{E(X)}{\sigma} - \frac{\mu}{\sigma} = 0$$
$$V(Z) = \frac{V(X)}{\sigma^2} = 1$$

Le variabili aleatorie continue: specificità

- Per le v.a. continue le probabilità sono assegnate ad intervalli e non a singoli punti.

Pertanto, la probabilità di un intervallo compreso tra c e d è pari all'area sotto la funzione di densità f(x) della variabile aleatoria. Tale area si calcola attraverso un integrale, che, a volte, può essere complesso da risolvere. Vari software ci vengono in aiuto.
Ricordiamo che la probabilità in un singolo punto è zero, pari all'integrale calcolato in un singolo punto.

Per calcolare la probabilità di un intervallo, in pratica, ci affidiamo alla funzione di ripartizione.

- La funzione di ripartizione di una variabile aleatoria X è definita come:

$$F(x_0) = P(X < x_0)$$

La funzione di ripartizione ricorda le frequenze relative cumulate. Ci dice quanta probabilità c'è fino ad un certo punto x_0

- Se conosciamo la funzione di ripartizione possiamo calcolare in modo molto semplice la probabilità di un qualsiasi intervallo:

$$P(a < X < b) = F(b) - F(a)$$

Esempio: calcolo di probabilità di un intervallo

Nel nostro esempio vogliamo calcolare la probabilità evidenziata in rosso corrispondente a:

$$P(a < X < b) = F(b) - F(a)$$

Questa può essere ottenuta come differenza tra la funzione di ripartizione calcolata nell'estremo superiore dell'intervallo (rosa) e la funzione di ripartizione calcolata nell'estremo inferiore dell'intervallo (azzurro).

Valore atteso e varianza di una variabile aleatoria continua

Hanno lo stesso nome, significato e proprietà degli indici visti per le v.a. discrete. Questo perché l'integrale non è altro che una somma sul continuo.

Data la variabile aleatoria continua X e la probabilità *p*, chiameremo quantile di X associato a p il valore x_p tale che:

- $P(X < x_p) = p$

Modelli probabilistici

Approfondiamo ora variabili aleatorie con funzione di probabilità o densità di probabilità espressa da una funzione analitica.
In particolare, discuteremo le v.a. Bernoulliana, Binomiale, Poisson e Normale (Gaussiana).
Ricordiamo che ne esistono molte altre, come la v.a. Uniforme o la t di Student (che introdurremo nell'ultima VL).

Bernoulli

- Due soli possibili risultati: successo/insuccesso

Denotiamo con π la probabilità di successo, la probabilità di insuccesso è 1- π.
La funzione di probabilità può essere scritta come la probabilità di successo elevata al successo per (uno meno la probabilità di successo), pari alla probabilità di insuccesso, elevata all'insuccesso.
Abbiamo così:

$$P(0) = 1-\pi \text{ e } P(1) = \pi$$

$$E(X) = 0 \times (1 - \pi) + 1 \times \pi =$$
$$V(X) = (0 - \pi)^2(1 - \pi) + (1 - \pi)^2\pi = \pi(1 - \pi)(\pi + 1 - \pi) = \pi(1 - \pi)$$

Binomiale

- Definiamo la variabile aleatoria X: numero di successi in n prove indipendenti
- In ogni prova abbiamo successo-insuccesso
- Denotiamo con π la probabilità di successo in una singola prova

$$P(x) = \frac{n!}{x!\,(n-x)!} \pi^x (1-\pi)^{n-x}$$

La prima parte della formula a destra dell'uguale si chiama coefficiente binomiale e tiene conto delle possibili combinazioni. Il ! indica il fattoriale di un numero.

Esempio: Binomiale

Qual è la probabilità di avere una volta 3 lanciando due volte un dado?

- Soluzione 1: calcoliamo questa probabilità sfruttando la definizione di probabilità, casi favorevoli su casi possibili (purché ugualmente possibili). Abbiamo quindi 10/36.
- Soluzione 2: usiamo la v.a. Binomiale e definiamo
  - n = 2 prove indipendenti
  - in ogni prova esce 3 (successo)
  - $\pi = 1/6$ è la probabilità di successo
  - x = 1 numero di successi di cui vogliamo conoscere la probabilità

$$P(X=1) = \frac{2!}{1!\,(2-1)!} \left(\frac{1}{6}\right)^1 \left(1-\frac{1}{6}\right)^{2-1} = \frac{10}{36}$$

Per la v.a. Binomiale

$$E(X) = n\pi$$
$$V(X) = n\pi(1-\pi)$$

Poisson

Nel caso di variabili aleatorie di tipo conteggio (es. quante volte vado dal dottore), la v.a. di riferimento è la Poisson.

$$X \sim Poisson(\lambda), \quad x \geq 0; 0 < \lambda < +\infty$$
$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Equidispersione: $E(X) = V(X) = \lambda$

Ancora su Binomiale e Poisson

- Binomiale:
  - Il valore atteso e la varianza crescono al crescere di n
  - La distribuzione è simmetrica per $\pi = 0.5$
  - In ogni caso, per $n \to \infty$, la distribuzione tende a essere simmetrica rispetto al valore atteso

- Poisson
  - La variabile casuale Binomiale, al crescere di n e al diminuire di π, tende a una variabile casuale di Poisson con parametro $\lambda = n\pi$

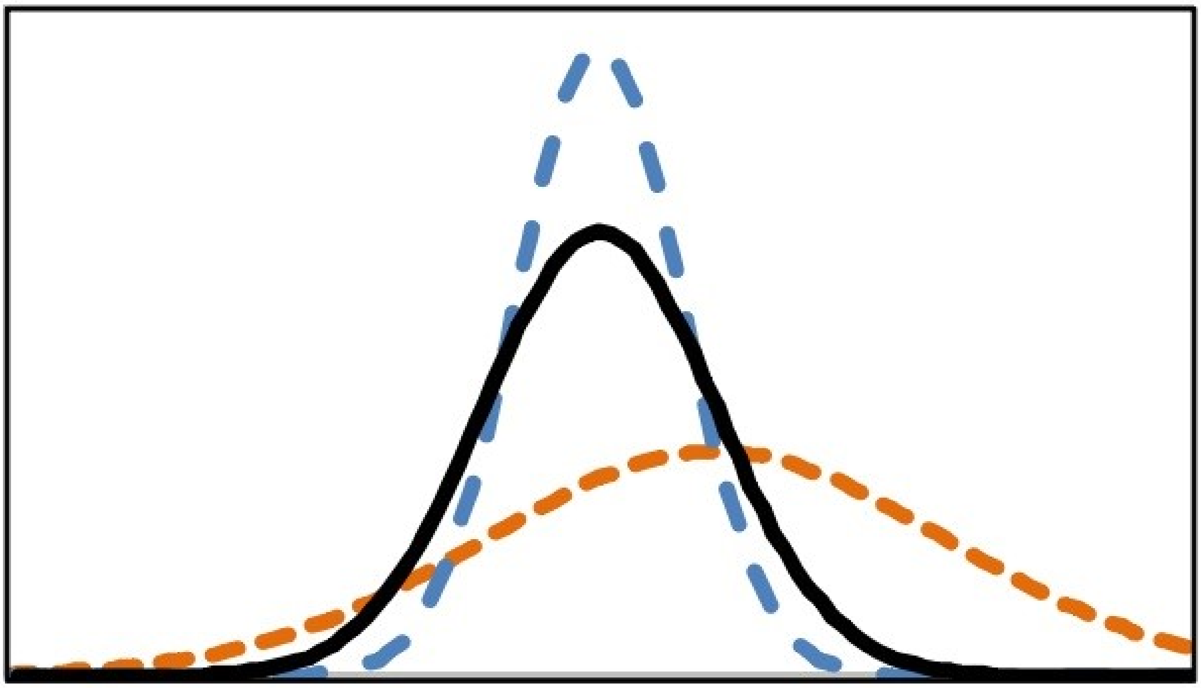
La variabile aleatoria Normale

È la v.a. continua più importante. La sua funzione di densità dipende da due parametri, la media e la varianza

- ha una forma a campana
- è simmetrica
- media, moda e mediana coincidono
- la locazione è data dalla media
- la dispersione è data dalla varianza σ^2

- ha un range infinito $(-\infty, +\infty)$

Nel nostro esempio, confrontiamo tre diverse Normali. Che differenze notate? Quale ha varianza maggiore tra la blu e la nera? E tra tutte tre? Ricordiamo che la varianza è una misura di variabilità. La Normale arancione ha sia media che varianza maggiori rispetto alle altre due.
La blu e la nera hanno la stessa media, ma la nera ha varianza maggiore.



Per il calcolo, senza un adeguato software, delle probabilità di un intervallo di una v.a. aleatoria Normale, si utilizzano le cosiddette tavole della Normale standardizzata. Ricordiamo che la v.a. Z è una v.a. con media zero e varianza 1.

Visto che la media e la mediana coincidono, la probabilità di osservare un valore inferiore alla media è pari a quella di osservare un valore maggiore della media, ed entrambe sono pari a 0.5.

Esempio: Normale

Supponiamo X sia una v.a. Normale con media 8.0 e deviazione standard 5.0. Calcolare $P(X > 8.6)$.
Questa probabilità equivale a calcolare $P(Z>0.12)$, dove 0.12 è il valore ottenuto tramite standardizzazione. Tale probabilità è pari a 1 meno la funzione di ripartizione calcolata in 0.12.

Intervalli notevoli

La probabilità di un intervallo simmetrico, centrato sulla media e di raggio pari a 1, 2, o 3 deviazioni standard è nota.

Quantili di una v.a. Normale

Partendo da una probabilità nota, o conoscendo la funzione di ripartizione, è possibile trovare il corrispondente quantile della distribuzione. Rendiamo questo concetto più immediato attraverso un esempio.

Supponiamo che la distribuzione della spesa per cliente in un grande supermercato sia ben approssimata da una normale con media $\mu = 50$ e deviazione standard $\sigma = 10$. Determinare il livello di spesa sotto il quale abbiamo solo una probabilità del 20% di trovare un cliente.

In una v.a. Normale standardizzata, utilizzando le tavole, tale valore è pari a -0.84 (è ovviamente negativo). Ma quale è il valore riferito alla v.a. Normale originale? Basta risolvere una semplice equazione. Ricordiamo la formula della standardizzazione:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - \mu}{\sigma}$$

Da cui otteniamo X = $\mu + z * \sigma$ = 50 + (-0.84)*10

Conclusioni

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

Ti ricordo che abbiamo visto:

- le variabili aleatorie discrete e continue
- i modelli probabilistici per variabili aleatorie discrete: Bernoulli, Binomiale, Poisson
- i modelli probabilistici per variabili aleatorie continue: Normale (o Gaussiana)

Grazie per l'attenzione!