Probabilità e Statistica

Isadora Antoniano-Villalobos
isadora.antoniano@unive.it

Laurea in Informatica
(Data science/ Tecnologie e scienze dell'informazione)
Università Ca' Foscari di Venezia

Anno academico 2024/2025

Costanti Caratteristiche
Testo Capitoli 4 e 5

Costanti caratteristiche

Una **costante caratteristica** o indice è un numero associato ad una variabile aleatoria o alla sua distribuzione di probabilità e sintetizza informazione d'interesse sul fenomeno rappresentato dalla variabile.

- Conoscere il valore di un indice significa avere informazione sulla distribuzione stessa.

Definiremo ora alcune importanti costanti caratteristiche:

1. Il **valore atteso**, che è un indice di posizione
2. La **varianza**, che è un indice di dispersione
3. I quantili di una variabile aleatoria, che contengono informazione sia sulla posizione che sulla forma di una distribuzione

Valore atteso di una variabile aleatoria discreta

Valore atteso

Se X è una v.a. discreta con valori $\{x_1, x_2, \ldots\}$ e funzione di probabilità $P(x_i) = p_i$, allora il **valore atteso** o **media** di X è

$$\mathbb{E}[X] = \sum_i x_i p_i = x_1 p_1 + x_2 p_2 + \ldots \ .$$

Esercizi:

1. Qual è il valore atteso di una variabile di Bernoulli con parametro p?
2. Qual è il valore atteso di una variabile Geometrica con parametro p?
3. Qual è il valore atteso di una variabile di Poisson con parametro λ?

Valore atteso di una variabile aleatoria discreta

- **Esempio:** Se A è un evento qualsiasi, $\mathbf{1}_A$ è una variabile aleatoria che vale 1 con probabilità $\mathbb{P}[A]$ e 0 con probabilità $\mathbb{P}\left[\bar{A}\right]$.
  - Il suo valore atteso è:
$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A]$$
- **Esempio:** Torniamo alla solita urna composta da quattro palle bianche numerate da 1 a 4 e tre palle nere numerate da 1 a 3, dalla quale si campiona casualmente una palla, registrando $X =$ numero estratto dall'urna.
  - Per la variabile $X=$ numero estratto, si ha:

$$P_X(x) = \begin{cases} 2/7 & \text{se } x = 1, 2, 3; \\ 1/7 & \text{se } x = 4; \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\mathbb{E}[X] = 1\frac{2}{7} + 2\frac{2}{7} + 3\frac{2}{7} + 4\frac{1}{7} = \frac{16}{7}$

Valore atteso di una variabile aleatoria discreta

- Per la variabile casuale che rappresenta l'importo vinto nel gioco (vincere 30 euro se la pallina estratta è bianca e perderne 20 se è nera), abbiamo

$$P_Y(y) = \begin{cases} 4/7 & \text{se } y = 30, \\ 3/7 & \text{se } y = -20, \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\mathbb{E}[Y] = 30\frac{4}{7} - 20\frac{3}{7} = \frac{60}{7}$

☞ **Esercizio:** Ricordiamo che $S =$ somma dei due numeri sulle palline estratte senza reinserimento. $\mathbb{E}[S] =?$

Valore atteso di una variabile aleatoria discreta

- **Esempio:** Consideriamo la seguente scommessa. Tu scegli un prezzo p da pagare per giocare, io lo pago e poi guadagno 1 euro se l'evento A si verifica o 0 euro se A non si verifica. Qual è il prezzo giusto della scommessa?
  - Sia X è il mio guadagno (aleatorio) nella scommessa. Allora,

$$X(\omega) = -p + 1 \cdot \mathbf{1}_A(\omega) + 0 \cdot \mathbf{1}_{\bar{A}}(\omega).$$

$$P_X(x) = \begin{cases} \mathbb{P}[A] & \text{se } x = 1 - p, \\ \mathbb{P}\left[\bar{A}\right] & \text{se } x = -p, \\ 0 & \text{altrimenti} \end{cases}$$

Un prezzo ragionevole per giocare è quello che mi dà un guadagno atteso nullo, $\mathbb{E}[X] = 0$ (altrimenti nessuno accetterebbe la scommessa). Cioè

$$(1 - p)\,\mathbb{P}[A] - p\,\mathbb{P}\left[\bar{A}\right] = 0 \Leftrightarrow p = \mathbb{P}[A].$$

Valore atteso di $g(X)$

Sia $Y = g(X)$ una v.a. ottenuta trasformando la v.a. X tramite la funzione $g : \mathbb{R} \to \mathbb{R}$. Il valore atteso di Y si può calcolare anche senza conoscere direttamente la distribuzione di probabilità di Y:

- Se $Y = g(X)$ per una variabile X discreta:

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_i g(x_i) p_i$$

.

☞ **Esercizio:** Se il raggio R di una sfera è una variabile casuale uniforme discreta in $\{1, 2, 3\}$, qual è il valore atteso del volume?

- In generale, se X e Y sono due variabili discrete:

$$\mathbb{E}[g(X, Y)] = \sum_i \sum_j g(x_i, y_j) p_{X,Y}(x_i, y_j)$$

.

Valore atteso di XY

Se X e Y sono due variabili casuali, il valore atteso del prodotto $W = XY$ si può calcolare conoscendo la distribuzione congiunta $p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y]$:

- Se X e Y sono discrete:

$$\mathbb{E}[W] = \mathbb{E}[XY] = \sum_i \sum_j x_i y_j \, p_{X,Y}(x_i, y_j)$$

.

☞ **Esercizio:** Se X e Y sono indipendenti con $\mathbb{E}[X] = \mu_X$ e $\mathbb{E}[Y] = \mu_Y$, trovare il valore atteso di $V = g(X)h(Y)$ per due funzioni qualsiasi g e h.

Proprietà del valore atteso

Il valore atteso ha le seguenti proprietà, per X e Y variabili casuali, a, b e c costanti, g e h funzioni:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- In particolare $\mathbb{E}[X - \mathbb{E}[X]] = 0$. La variabile $X - \mathbb{E}[X]$ si dice **centrata**
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- In generale $\mathbb{E}[g(X) + h(Y)] = \mathbb{E}[g(X)] + \mathbb{E}[h(Y)]$
- Ancora più in generale $\mathbb{E}[a\,g(X) + b\,h(Y) + c] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(Y)] + c$
- Un modo alternativo per calcolare la media è

$$\mathbb{E}[X] = \sum_x \mathbb{P}[X > x] = \sum_x \bar{F}(x)$$

Valore atteso condizionato

Valore atteso

Sia X una v.a. discreta con funzione di probabilità $p(x)$. Sia A un evento con $\mathbb{P}[A] > 0$ e sia $p_{X|A}(x)$ la funzione di probabilità condizionata di X dato A. Allora il **valore atteso condizionato** o **media condizionata** di X dato A è

$$\mu_X = \mathbb{E}[X] = \sum_x x p_{X|A}(x) = \sum_x x \frac{\mathbb{P}[X = x, A]}{\mathbb{P}[A]}$$

➔ Ricordiamo che la probabilità condizionata soddisfa tutte le proprietà della probabilità. Allo stesso modo, la funzione di probabilità condizionata soddisfa tutte le proprietà delle funzioni di probabilità e, di conseguenza, il valore atteso condizionato soddisfa tutte le proprietà del valore atteso.

Valore atteso condizionato

➔ **Esempio:** Torniamo alla solita urna e consideriamo l'evento A il numero estratto dal'urna è 1 o 2.

- Formalmente, $A = [X \leq 2]$ e

$$\mathbb{P}\left[A\right] = \mathbb{P}\left[X \leq 2\right] = p_x(1) + p_x(2) = \frac{4}{7}$$

- La funzione di probabilità condizionata di X dato A è:

$$P_{X|A}(x) = \begin{cases} \frac{2/7}{4/7} = 1/2 & \text{se } x = 1, 2; \\ 0 & \text{altrimenti} \end{cases}$$

- Allora,

$$\mathbb{E}\left[X\right] = 1\frac{1}{2} + 2\frac{1}{2} = \frac{3}{2}$$

Valore atteso condizionato

Legge delle probabilità totali per il valore atteso

Sia $C_1, C_2, \ldots$ un apartizione di Ω e sia X una v.a. Allora

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|C_i]\,\mathbb{P}[C_i]$$

In particolare, se Y è una variabile casuale con supporto $\{y_1, y_2, \ldots\}$ e funzione di probabilità p_Y, allora

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|Y = y_i]\,p_Y(y_i)$$

☞ Per qualsiasi funzione g,

$$\mathbb{E}[g(X)] = \sum_i \mathbb{E}[g(X)|Y = y_i]\,p_Y(y_i)$$

Varianza di una variabile aleatoria discreta

Varianza e scarto quadratico medio

Se X è una v.a. discreta con valori $\{x_1, x_2, \ldots\}$ e funzione di probabilità $P(x_i) = p_i$, allora la **varianza** di X è la media del errore quadratico di X rispetto al suo valore atteso:

$$\sigma_X^2 = \mathrm{Var}\left[X\right] = \sum_i (x_i - \mathbb{E}\left[X\right])^2 p_i$$

Da non confondere con lo **scarto quadratico medio** o **deviazione standard**:

$$\sigma_X = \mathrm{std}[X] = \sqrt{\mathrm{Var}\left[X\right]}$$

Esercizo:

- Qual è la varianza di una variabile di Bernoulli con parametro p?

Varianza di una variabile aleatoria discreta

➔ **Esempio:** Per la variabile $X =$ numero estratto dalla solita urna, abbiamo:

$$P_X(x) = \begin{cases} 2/7 & \text{se } x = 1, 2, 3; \\ 1/7 & \text{se } x = 4; \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\text{Var}[X] = 1\frac{2}{7} + 2^2\frac{2}{7} + 3^2\frac{2}{7} + 4^2\frac{1}{7} - \left(\frac{16}{7}\right)^2$.

☞ **Esercizio:** calcolare la varianza per le altre due variabili dell'esempio.

Proprietà della varianza

La varianza, ha le seguenti proprietà, per X e Y variabili casuali, a, b e c costanti, g e h funzioni:

- $\mathrm{Var}\left[a\right] = 0$
- $\mathrm{Var}\left[aX + b\right] = a^2\mathrm{Var}\left[X\right]$
- In particolare $\mathrm{Var}\left[X/\mathrm{std}[X]\right] = 1.$ La variabile $X - \mathbb{E}\left[X\right]$ si dice **riscallata**
- Se X e Y sono indipendenti, $\mathrm{Var}\left[g(X) + h(Y)\right] = \mathrm{Var}\left[g(X)\right] + \mathrm{Var}\left[h(Y)\right]$
- In generale, se X e Y sono indipendenti, $\mathrm{Var}\left[a\, g(X) + b\, h(Y) + c\right] = a^2\mathrm{Var}\left[g(X)\right] + b^2\mathrm{Var}\left[h(Y)\right]$
- Un modo alternativo per calcolare la varianza è

$$\sigma_X^2 = \mathrm{Var}\left[X\right] = \sum_i x_i^2 p_i - \left[\mathbb{E}\left[X\right]\right]^2 = \mathbb{E}\left[X^2\right] - \left[\mathbb{E}\left[X\right]\right]^2 .$$

Media e varianza per alcune distribuzioni note

- **Distribuzione ipergeometrica:** Se $X \sim \text{Ig}\,(m, n, k)$,

$$\mathbb{E}\left[X\right] = k\frac{m}{m+n} \quad \text{Var}\left[X\right] = k\frac{m}{m+n}\frac{n}{m+n}\frac{m+n-k}{m+n-1}$$

- **Distribuzione di Bernoulli:** Se $X \sim \text{Ber}\,(p)$,

$$\mathbb{E}\left[X\right] = p \quad \text{Var}\left[X\right] = p(1-p)$$

- **Distribuzione Binomiale:** Se $X \sim \text{Binom}\,(n, p)$,

$$\mathbb{E}\left[X\right] = np \quad \text{Var}\left[X\right] = np(1-p)$$

- **Distribuzione Geometrica:** Se $X \sim \text{Geom}\,(p)$,

$$\mathbb{E}\left[X\right] = \frac{1}{p} \quad \text{Var}\left[X\right] = \frac{(1-p)}{p^2}$$

- **Distribuzione di Poisson:** Se $X \sim \text{Pois}\,(\lambda)$,

$$\mathbb{E}\left[X\right] = \lambda \quad \text{Var}\left[X\right] = \lambda$$

La covarianza

La covarianza

La **covarianza** fra due v.a. X e Y è il valore atteso di una particolare trasformazione $g(x, y) = (x - \mathbb{E}[X])(y - \mathbb{E}[Y])$:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] .$$

Sviluppando il prodotto dentro il valore atteso e semplificando alcuni termini, otteniamo una formula pratica per il calcolo della covarianza:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] .$$

➔ Se $\mathrm{Cov}[X, Y] = 0$ diciamo che X e Y sono **incorrelate**.

Indipendenza e covarianza

Se X e Y sono indipendenti allora Cov $[X, Y] = 0$.

☞ Il viceversa non è vero!

Esempio: Consideriamo

$$X = \begin{cases} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{cases} \quad \text{e} \quad Y = \begin{cases} 0 & X \neq 0 \\ 1 & X = 0 \end{cases}$$

☞ Ovviamente, le due variabili non sono indipendenti, ma si dimostra facilmente che Cov $[X, Y] = 0$ e perciò che X e Y sono incorrelate.

Proprietà della covarianza

1. $\text{Cov}[X, Y] = \text{Cov}[Y, X]$
2. $\text{Cov}[X, X] = \text{Var}[X]$
3. $\text{Cov}[aX, Y] = a\text{Cov}[X, Y]$
4. $\text{Cov}[X, a] = 0$
5. $\text{Cov}\left[\sum_i X_i, \sum_j Y_j\right] = \sum_i \sum_j \text{Cov}[X_i, Y_j]$

Varianza di una somma

Dalle proprietà della covarianza, discende un risultato molto utile nelle applicazioni statistiche:

$$\begin{aligned}
\mathrm{Var}\left[\sum_i X_i\right] &= \mathrm{Cov}\left[\sum_i X_i, \sum_j X_j\right] = \sum_i \sum_j \mathrm{Cov}\left[X_i, X_j\right] \\
&= \sum_i \mathrm{Var}\left[X_i\right] + \sum_i \sum_{j \neq i} \mathrm{Cov}\left[X_i, X_j\right] \\
&= \sum_i \mathrm{Var}\left[X_i\right] + 2 \sum_i \sum_{j > i} \mathrm{Cov}\left[X_i, X_j\right].
\end{aligned}$$

☞ In particolare, se le X_i sono **a due a due indipendenti**, allora $\mathrm{Cov}\left[X_i, X_j\right] = 0$ per $i \neq j$ e

$$\mathrm{Var}\left[\sum_i X_i\right] = \sum_i \mathrm{Var}\left[X_i\right].$$

La correlazione

La correlazione

La **correlazione** fra due v.a. X e Y è definita come:

$$\text{Cor}\left[X, Y\right] = \frac{\text{Cov}\left[X, Y\right]}{\sqrt{\text{Var}\left[X\right] \cdot \text{Var}\left[Y\right]}},$$

sempre che $\text{Var}\left[X\right]$ e $\text{Var}\left[Y\right]$ siano diverse da 0.

Si ha che

$$-1 \leq \text{Cor}\left[X, Y\right] \leq 1.$$

Inoltre, se $Y = a + bX$ allora $\text{Cor}\left[X, Y\right] = \pm 1$, a seconda del segno di b.

➔ Per questa ragione si dice che $\text{Cor}\left[X, Y\right]$ misura l'intensità del legame lineare fra X e Y.

La correlazione

Esempio: Tornando all'esempio del numero di errori nei due moduli di un programma.

👉 Abbiamo

$$\begin{aligned}\text{Cov}\left[X,Y\right] &= \mathbb{E}\left[XY\right]-\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\\ &= 1\cdot 1\cdot 0.1+1\cdot 2\cdot 0.1+1\cdot 3\cdot 0.1-0.5\cdot 1.05\\ &= 0.6-0.525=0.075\end{aligned}$$

e, essendo $\text{Var}\left[X\right]=0.5^2$ e $\text{Var}\left[Y\right]=1.0712^2$,

$$\text{Cor}\left[X,Y\right]=\frac{0.075}{0.5\cdot 1.0712}=0.14.$$

Mediana di una variabile aleatoria

Mediana

La **mediana** di una variabile aleatoria X è il minimo valore m per cui

$$F(m) = \mathbb{P}\left[X \leq m\right] \geq \frac{1}{2}.$$

- **Esempio:** La variabile X dell'esempio dell'urna ha mediana in 2, mentre la Y ce l'ha in 30.
  - **Esercizio:** E la S? Disegnate i grafici delle relative f.r. e verificate quanto appena detto.

Quantili di una variabile aleatoria

Quantili

Fissato un valore $\alpha \in (0, 1)$, il **quantile di livello α** di una variabile aleatoria X è il minimo valore q_α per cui

$$F(q_\alpha) = \mathbb{P}\left[X \leq q_\alpha\right] \geq \alpha.$$

- I quantili sono **indici di posizione** che generalizzano il concetto di mediana di una distribuzione.
- La **mediana** non è altro che il quantile di livello $1/2$.
- I **percentili** sono i quantili di livello $k/100$, con $k = 1, 2, \ldots, 99$.
- I **decili** sono i quantili di livello $k/10$, con $k = 1, 2, \ldots, 9$.
- I **quartili** sono i quantili di livello 0.25 (primo quartile), 0.5 (mediana) e 0.75 (terzo quartile).

☞ **Esercizio:** Calcolare i quartili delle variabili considerate nei diversi esempi.