UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

ELEMENTI DI PROBABILITA' E STATISTICA: LEZ.5

Prof. Roberto Capone
A.A. 2023/24
Corso di Laurea in Scienze Biologiche

La raccolta dei dati

Quando la media non basta

La media non basta riassumere ragionevolmente i dati; ci serve anche una misura di quanto la media sia rappresentativa, cioè di quanto i dati si accumulano vicino alla media o di quanto invece sono sparsi tra tutti i possibili valori. In altre parole, ci serve una misura della dispersione dei dati

Indice di variabilità

Una prima misura di dispersione e l'intervallo di variabilità dato dalla differenza tra il dato massimo e il dato minimo

$$IV = x_{max} - x_{min}$$

L'intervallo di variabilità ci dice quanto sono sparsi i dati ma non quanto sono dispersi rispetto alla media.
Inoltre è una misura molto grossolana; per questo motivo viene utilizzato raramente

La raccolta dei dati

Scarto quadratico

Siano $x_1, x_2, \ldots, x_n$ dati di media $\bar{x}$. Il valore della media rappresenta il dato i-esimo a meno di un errore o scarto:

$$\bar{x} - x_i$$

E ragionevole pensare che la scelta migliore della media sia quella che minimizza gli errori. Tali errori vanno minimizzati nel loro complesso e possono essere sia negativi che positivi. Tuttavia il segno dell'errore è irrilevante, quindi ha senso usare come misura dell'errore quello che chiamiamo scarto quadratico

$$(\bar{x} - x_i)^2$$

Varianza

Definiamo Varianza il numero

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = \frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \cdots + (\bar{x} - x_n)^2}{n}$$

La raccolta dei dati

Varianza

Definiamo Varianza il numero

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = \frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \cdots + (\bar{x} - x_n)^2}{n}$$

L'idea intuitiva è che una varianza grande significa che i dati sono molto dispersi rispetto alla media o, in altre parole, che la media riassume male i dati.
Viceversa una varianza piccola significa che i dati sono concentrati intorno alla media e la media riassume bene i dati.
Tuttavia può essere difficile dire la varianza è piccola o grande in assoluto; dobbiamo confrontarla con qualcosa.
Il termine naturale di paragone potrebbe essere la media stessa. Solo che la varianza contiene dei quadrati per cui non è direttamente confrontabile con la media.

La raccolta dei dati

Varianza e Deviazione Standard

Definiamo Varianza il numero

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = \frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \cdots + (\bar{x} - x_n)^2}{n}$$

Ovviare a questa difficoltà si introduce la deviazione standard che la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2}$$

La deviazione standard può essere confrontata direttamente con la media. Tanto più è piccola la deviazione standard rispetto alla media tanto più i dati sono concentrati intorno alla media (cioè tanto meglio la media riassume i dati).

La raccolta dei dati

La deviazione standard può essere confrontata direttamente con la media. Tanto più è piccola la deviazione standard rispetto alla media tanto più i dati sono concentrati intorno alla media (cioè tanto meglio la media riassume i dati).

Per esprimere quantitativamente questa osservazione si introduce il coefficiente di variazione:

$$CV(x_i) = \frac{\sigma}{\bar{x}}$$

Si tratta di una misura della dispersione dei dati che non dipende dalle unità di misura usata e che permette di confrontare la dispersione di dati diversi. In particolare, un coefficiente di variazione piccolo (per esempio inferiore a 1/2), vuol dire davvero che la media riassume bene i dati e che i dati hanno una dispersione piccola intorno alla media

La raccolta dei dati

Ancora sulla varianza

C'è un'altra formula per il calcolo della varianza che spesso è utile:

$$\sigma_i^2 \quad = Media\left(x_i^2\right) - Media^2(x_i)$$

Cioè la varianza è uguale alla differenza fra la media dei quadrati e il quadrato della media.

Variabili aleatorie

L'idea ora è quella di interpretare una misura come una funzione a valori reali definita su uno spazio degli eventi. Ci proponiamo cioè di formalizzare il concetto di probabilità che una certa misura dia un certo risultato con una certa approssimazione.
Poi i nostri strumenti di misura hanno una precisione finita, di solito il risultato di una misura non è un numero reale esatto con tutte le sue infinite cifre decimali ma è un numero reale indicato con un certo errore.
In altre parole, la misura ci dice solo che il valore vero appartiene ad un determinato intervallo.
Quindi non siamo interessati alla probabilità che la misura dia un determinato valore ma alla probabilità che il valore della misura cada in un certo intervallo.

Variabili aleatorie

Ciò può essere formalizzato introducendo il concetto di variabile aleatoria. Sia U uno spazio degli eventi. Indichiamo con A la famiglia dei sottoinsiemi di U di cui possiamo calcolare la probabilità. Allora una variabile aleatoria a valori reali è una funzione

$$X: U \to R$$

Tale che per ogni intervallo (aperto o chiuso) o) I di R, l'insieme appartenga ad A:

$$X^{-1}(I) = \{a \in U | X(a) \in I\}$$

In altre parole stiamo dicendo che una variabile aleatoria è semplicemente una funzione dall'insieme universo U all'insieme dei numeri reali R per cui siamo in grado di calcolare la probabilità che il valore X cada in un certo intervallo I.

Diremo che un insieme $D \subset R$ è discreto se esiste $\varepsilon > 0$ tale che due elementi diversi di D distano sempre almeno ε:

Se a e b sono diversi, si ha:

$$|a - b| \geq \varepsilon$$

Diremo allora che una variabile aleatoria $X: U \to R$ è discreta se la sua immagine $X(U) \subset R$ È un'insieme discreto

Variabili aleatorie

Per le variabili aleatorie valgono tutte le definizioni introdotte nelle elezioni precedenti.
Ad esempio analogamente alla definizione di eventi indipendenti, diremo che due variabili aleatorie $X_1, X_2 : U \to R$ definite sullo stesso spazio degli eventi sono indipendenti se:

$$P(\{X_1 \in I_1\} \cap \{X_2 \in I_2\} = P(X_1 \in I_1) \cdot P(X_2 \in I_2)$$

Ovvero,
$X_1 \, e \, X_2$ sono indipendenti se gli eventi $\{X_1 \in I_1\}$ e $\{X_2 \in I_2\}$ sono indipendenti quali che siano gli intervalli o semirette I_1 e I_2

Dato uno spazio degli eventi finito $U = \{v_1, v_2 \ldots, v_n\}$ con distribuzione di probabilità uniforme e una variabile aleatoria $X : U \to R$, possiamo definire la media:

$$\bar{x} = \frac{X(v_1) + \cdots + X(v_n)}{n}$$

Variabili aleatorie

Sia

$$X: U \to R$$

una variabile aleatoria discreta. Il valore atteso(o valore medio o medio o speranza) di X è il numero $E(X)$ che indichiamo con μ_X è dato da

$$E(X) = \mu_X = \sum P(X = x_j)x_j$$

La varianza di una variabile aleatoria discreta è :

$$\sigma_X^2 = \sum p_j \left(x_j - \mu_X\right)^2$$

La deviazione standard

$$\sigma_X = \sqrt{\sigma_X^2}$$

Variabili aleatorie

ESEMPIO 1

Sia $U = \{1,2,\ldots,6\}$ lo spazio degli eventi del lancio di un dado a sei facce non truccato e indichiamo con $X: U \to R$ la variabile aleatoria banale data dal valore del lancio del dado:

$$X(j) = j \text{ per } j = 1,\ldots,6.$$

Siccome il dado non è truccato, su U abbiamo la distribuzione uniforme di probabilità per cui

$$p_j = p(X = j) = \frac{1}{6} \text{ per } j = 1,\ldots,6$$

$$E(X) = \mu_X = \sum P(X = x_j)x_j = p_1 \cdot 1 + p_2 \cdot 2 + \cdots + p_6 \cdot 6 = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

La varianza è :

$$\sigma_X^2 = \sum p_j\left(x_j - \mu_X\right)^2 = p_1(1 - 3.5)^2 + p_2(2 - 3.5)^2 + \cdots + p_6(6 - 3.5)^2 = 2.916$$

La deviazione standard

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{2.916} = 1.71$$

Variabili aleatorie

ESEMPIO 2

Calcoliamo il valore atteso e la varianza della somma X del lancio di due dadi non truccati.

Somma	*Coppie ordinate*	*Probabilità*
2	(1,1)	1/36
3	(1,2), (2,1)	1/18
4	(1,3), (2,2), (3,1)	1/12
5	(1,4), (2,3), (3,2), (4,1)	1/9
6	(1,5), (2,4), (3,3), (4,2), (5,1)	5/36
7	(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)	1/6
8	(2,6), (3,5), (4,4), (5,3),(6,2)	5/36
9	(3,6), (4,5), (5,4), (6,3)	1/9
10	(4,6), (5,5), (6,4)	1/12
11	(5,6), (6,5)	1/18
12	(6,6)	1/36

$$p(X=2)=p(X=12)=\frac{1}{36}$$

$$p(X=3)=p(X=11)=\frac{1}{18}$$

$$p(X=4)=p(X=10)=\frac{1}{12}$$

$$p(X=5)=p(X=9)=\frac{1}{9}$$

$$p(X=6)=p(X=8)=\frac{5}{36}$$

$$p(X=7)=\frac{1}{6}$$

$$E(X)=\mu_X=\sum P(X=x_j)x_j=\frac{1}{36}\cdot 2+\frac{1}{18}\cdot 3+\frac{1}{12}\cdot 4+\frac{1}{9}\cdot 5+\frac{5}{36}\cdot 6+\frac{1}{6}\cdot 7+\frac{5}{36}\cdot 8+\frac{1}{9}\cdot 9+\frac{1}{12}\cdot 10+\frac{1}{18}\cdot 11+\frac{1}{36}\cdot 12$$

$$=\frac{532}{36}=7$$

Variabili aleatorie

ESEMPIO 2

La varianza è data da:

$$\sigma_X^2 = \sum p_j \left(x_j - \mu_X\right)^2 = p_1(2-7)^2 + p_2(3-7)^2 + \cdots + p_{11}(12-7)^2 = 5.83$$

La deviazione standard

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{5.83} = 2.42$$

Variabili aleatorie bernoulliane

Ci troviamo di fronte alla seguente situazione: lo spazio degli eventi U consiste in tutti i possibili risultati di n sperimenti indipendenti, ad esempio n lanci di un dado o le nascite di n figli.
In ciascun esperimento un evento può avvenire con probabilità p.
Consideriamo allora la variabile aleatoria $X: U \to R$ che conta il numero di volte che l'evento E è effettivamente accaduto negli n esperimenti

Esso soddisfa la legge

$$p(X = k) = \binom{n}{k} p^k q^{n-k}$$

Per $k = 0, \dots n,$ dove $q = 1 - p$.
Qualsiasi variabile aleatoria discreta che soddisfa questa legge viene detta variabile aleatoria bernoulliana di tipo (n, p).
Ci proponiamo di calcolare valore atteso e varianza per una variabile aleatoria bernoulliana

Variabili aleatorie bernoulliane

Per quanto riguarda il valore atteso: se un evento E in un esperimento bernoulliano può accadere con probabilità p, in n esperimenti, ci aspettiamo che l'evento accada np volte.
Quindi la nostra intuizione ci suggerisce che il valore atteso di una variabile aleatoria bernoulliana X di tipo (n, p), si ottiene:

$$E(X) = np$$

La varianza si ottiene dalla formula:

$$\sigma_X^2 = np(1-p)$$

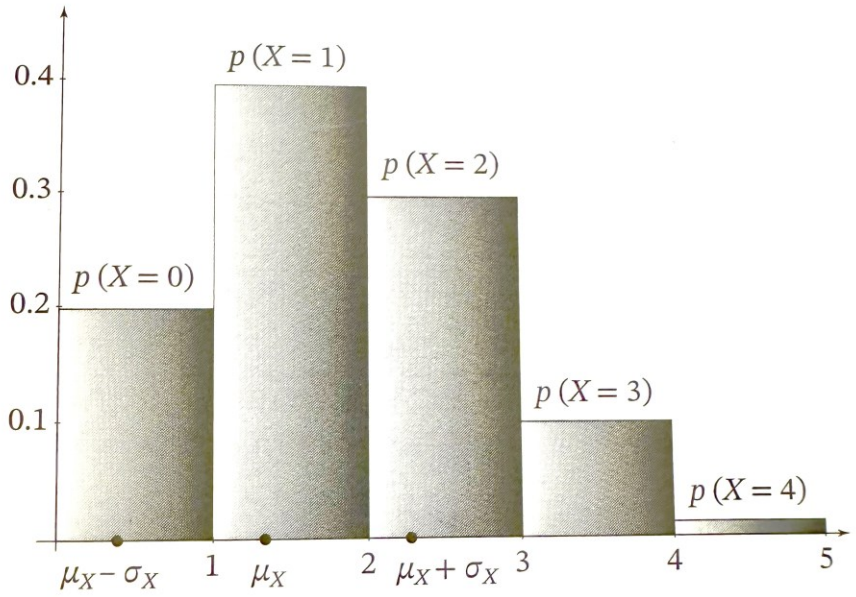
La deviazione standard è data dalla formula

$$\sigma = \sqrt{np(1-p)}$$

Variabili aleatorie bernoulliane

ESEMPIO

Rappresentiamo con un istogramma la distribuzione di probabilità di una variabile aleatoria bernoulliana di tipo $(4,1/3)$.
Con sopra l'intervallo $[i, i+1]$ ha altezza pari a $P(X=i)$.
Sull'asse delle ascisse sono indicati il valore atteso μ_X e i valori $\mu_X - \sigma_X$ e $\mu_X + \sigma_X$



Come si può vedere il valore atteso cade nell'intervallo in cui la probabilità è massima.
Inoltre l'intervallo $[\mu_X - \sigma_X;\ \mu_X + \sigma_X]$ comprende una buona metà dei valori di X.
E si potrebbe facilmente verificare che l'intervallo $[\mu_X - 2\sigma_X;\ \mu_X + 2\sigma_X]$ comprende quasi tutti i valori di X

Variabili aleatorie bernoulliane

ESEMPIO

Supponiamo che la probabilità che un paziente sottoposto ad una data operazione muoia entro un mese sia del 12%. Se il 1 Aprile sono state effettuate quattro operazioni, quanti pazienti ci aspettiamo che siano ancora vivi il 1 maggio?

Se la probabilità che un paziente muoia entro un mese è

$$q = 12/100$$

La probabilità che sopravviva per almeno un mese dopo l'operazione è

$$p = 1 - q = 88/100.$$

Se il 1 Aprile sono state effettuate quattro operazioni, il numero di pazienti che sopravvivono fino al 1 maggio è una variabile aleatoria bernoulliana di tipo $(4, p)$.
Vogliamo trovare il valore atteso

$$E(X) = np = 4p = 4 \cdot \frac{88}{100} = \frac{88}{25} = 3.52$$

Quindi ci aspettiamo che un mese dopo l'operazione siano ancora vivi fra i 3 e i 4 pazienti.

$$\sigma^2 = 4p(1-p) = 4 \cdot \frac{88}{100} \cdot \frac{12}{100} = \frac{264}{625} = 0.4224$$

$$\sigma = 0.65$$

Distribuzione di Poisson

L'uso della distribuzione binomiale richiede di conoscere a priori il numero massimo n di eventi possibili.
In molti casi non è detto che lo si sappia.
Inoltre la distribuzione binomiale necessità di sapere a priori la probabilità p che l'evento accada.

In molti casi invece questa probabilità non è nota.

L'osservazione del fenomeno può fornire solo il numero medio μ di eventi in un dato intervallo di tempo. Vogliamo vedere se usando solo questo numero medio μ riusciamo lo stesso a calcolare la probabilità che si verifichino k eventi nel dato intervallo

Distribuzione di Poisson

Supponiamo quindi di essere nella seguente situazione:

- Abbiamo fissato un dato intervallo di tempo (o una data regione di spazio)
- in un qualsiasi istante di tempo di questo intervallo può accadere un evento specifico e noi non siamo in grado di predire quando
- l'accadere o meno dell'evento in un dato istante e indipendente dalla l'accadere o meno dell'evento in un altro istante per cui la distribuzione degli eventi è casuale
- sappiamo che in media accadono μ eventi nell'intervallo di tempo dove l'intervallo è stato scelto sufficientemente ampio da poter a priori consentire che vi avvenga un numero di eventi anche molto maggiore di μ.

Un fenomeno che soddisfa queste condizioni è detto fenomeno di Poisson di media μ

Distribuzione di Poisson

I seguenti fenomeni sono tutti esempi di fenomeni di Poisson:

- il numero di automobili che passano in un dato punto di una strada in un fissato periodo di tempo;
- il numero di errori di battuta commessi scrivendo una pagina di testo;
- il numero di telefonate che un call center riceve in un minuto;
- il numero di mutazioni in una fissata sequenza di DNA sottoposto ad una data quantità di radiazione
- il numero di decadimenti radioattivi in una fissata quantità di sostanza radioattiva in un dato intervallo di tempo
- il numero di stelle in un dato volume di spazio
- il numero di virus che possono infettare una data cellula in uno specificato intervallo di tempo in una coltura cellulare fissata

Distribuzione di Poisson

Senza entrare nei dettagli della dimostrazione, diciamo che la probabilità che in un fenomeno di Poisson di media μ nel dato intervallo di tempo avvengano k eventi è data da

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$$

Questa formula rappresenta la distribuzione di Poisson e ogni variabile avvalori in N che soddisfa questa relazione viene detta variabile aleatoria di Poisson di media μ

ESEMPIO 1

Sapendo che in media in un anno 12 cavalleggeri prussiani vengono uccisi dal calcio di un cavallo qual è la probabilità che nel 1861 ne siano stati uccisi solo 7?

Siccome il numero di cavalleggeri uccisi in un anno dal calcio di cavallo è una variabile aleatoria di Poisson di media 12, vale la relazione

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$$

sostituendo i valori numerici otteniamo

$$P(X = 7) = \frac{12^7}{7!} e^{-12} = 0.044$$

Distribuzione di Poisson

ESEMPIO 2

Qual è la probabilità che in un call center che riceve in media 20 telefonate al minuto non arrivi alcuna telefonata tra le 09:15 e le 09:16 del 9 dicembre 2022?

Dobbiamo calcolare la probabilità $P(X = 0)$ di zero eventi in un fenomeno di Poisson di media 20

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$$

sostituendo i valori numerici otteniamo

$$P(X = 0) = \frac{20^0}{0!} e^{-20} = 2.6 \cdot 10^{-9}$$

La probabilità che i centralinisti di questo call center possano riposarsi per un minuto è veramente piccola

Distribuzione di Poisson

ESEMPIO 3

Sapendo che una cellula in una data coltura cellulare viene infettata in media da due virus al giorno, qual è la probabilità che il 20 maggio 2023 una data cellula sia infettata da almeno un virus?

Indicando con X la variabile di Poisson, che conta il numero di virus che infettano una data cellula, abbiamo

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{2^0}{0!} e^{-2} = 0.86$$

Quindi c'è una probabilità di circa l'86% che il 20 maggio 2023 una data cellula si è infettata

Ci limitiamo solo a fare questa **osservazione** senza la dimostrazione:
la varianza di una variabile aleatoria di Poisson è uguale alla sua media