Analisi dei dati

Un goliardico riassunto

Ollari Dmitri

9 maggio 2023

Indice

Elenco delle figure

Capitolo 1

Statistica descrittiva

1.1 Grandezze che sintetizano i dati

1.1.1 Media

Avendo un'insieme di dati $x_1, x_2, \ldots, x_n$, la **media campionaria** è definita da:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.1}$$

1.1.2 Mediana

Avendo un'insieme di dati **ordinati(crescente)** con ampiezza n, la **mediana** è il valore che **occupa la posizione**:

- $\frac{n+1}{2}$ se n è **dispari**
- Media tra il valore $\frac{n}{2}$ e $\frac{n}{2} + 1$ se n è **pari**

La mediana campionaria (o mediana campionaria ordinata) è un indice di tendenza centrale utilizzato per descrivere il valore centrale di un insieme di dati numerici.

1.1.3 Moda

Unico valore che ha frequenza massima.

1.1.4 Varianza

La varianza è una misura della dispersione o della variabilità dei dati all'interno di un insieme di dati numerici. In altre parole, la varianza indica quanto i dati si distribuiscono intorno alla media.

Dato un'insieme di dati $x_1, x_2, \ldots, x_n$, la **varianza campionaria (s^2)** è:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.2}$$

1.1.5 Deviazione standard

La deviazione standard campionaria (o deviazione standard campionaria corretta) è una misura della dispersione dei dati all'interno di un campione di dati numerici. Come la varianza, la deviazione standard è un'importante misura di dispersione dei dati, ma a differenza della varianza, ha la stessa unità di misura degli elementi dell'insieme di dati e quindi è più facilmente interpretabile.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1.3}$$

1.1.6 Percentili campionari

I percentili campionari sono una misura statistica utilizzata per suddividere un insieme di dati in 100 parti uguali. In altre parole, i percentili campionari suddividono l'insieme di dati in modo tale da poter analizzare la distribuzione dei dati in modo più dettagliato.

Per calcolare i percentili campionari, si segue il seguente procedimento:

1. Si ordina l'insieme di dati in ordine crescente.
2. Si moltiplica il numero totale di elementi nell'insieme di dati per la percentuale desiderata (ad esempio, se si vuole trovare il 25° percentile, si moltiplica il numero totale di elementi per 0,25).
3. Si arrotonda il risultato ottenuto al punto 2 all'intero successivo, se il risultato non è già un intero.
4. Si individua l'elemento nell'insieme di dati che si trova all'indice calcolato al punto 3. Questo valore corrisponde al percentile desiderato.

1.1.7 Box plotting

Il diagramma a scatola si presenta come un rettangolo orientato verticalmente, che ha come estremi inferiori e superiori il primo quartile (Q1) e il terzo quartile (Q3) della distribuzione. All'interno del rettangolo si trova una linea orizzontale che rappresenta la mediana (o secondo quartile, Q2), mentre le linee che si estendono dalla parte superiore e inferiore del rettangolo (i cosiddetti baffi) rappresentano la variazione massima e minima dei dati.

1.2 Disuguaglianza di Chebyschev

La Disuguaglianza di Chebyshev è un importante teorema della teoria della probabilità che fornisce una stima superiore sulla percentuale di dati che si discostano dalla media, per ogni distribuzione di probabilità.

La disuguaglianza afferma che, per qualsiasi distribuzione di probabilità con media μ e varianza finita σ^2, la percentuale di dati che si discostano dalla media per più di k volte la deviazione standard (cioè la distanza dalla media in termini di deviazioni standard) è inferiore o uguale a $1/k^2$, ovvero:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

dove X è una variabile aleatoria con media μ e varianza σ^2, e k è un qualsiasi valore positivo.

Questa disuguaglianza implica che, per qualsiasi distribuzione di probabilità, almeno il 75% dei dati si trova entro due deviazioni standard dalla media, e almeno il 89% dei dati si trova entro tre deviazioni standard dalla media. In altre parole, la disuguaglianza di Chebyshev fornisce un limite superiore alla dispersione dei dati, indipendentemente dalla loro distribuzione.

La disuguaglianza di Chebyshev è utile per stimare la probabilità di osservare un valore estremo in un insieme di dati, senza conoscere la loro distribuzione. Ad esempio, se si vuole stimare la percentuale di studenti di una scuola che hanno un QI inferiore a 130, sapendo solo la media e la deviazione standard del QI degli studenti, la disuguaglianza di Chebyshev può essere utilizzata per fornire una stima superiore sulla percentuale di studenti che hanno un QI inferiore a 130.

1.3 Insieme di dati bivariati

In statistica, un insieme di dati bivariati è un insieme di dati che contiene informazioni su due variabili per ogni osservazione nel campione. In altre parole, ogni osservazione nel campione è rappresentata da una coppia ordinata di numeri, uno per ogni variabile. Ad esempio, un insieme di dati bivariati potrebbe contenere informazioni sul reddito e sull'età di un gruppo di persone.

Un insieme di dati bivariati può essere rappresentato graficamente utilizzando un grafico di dispersione, dove i valori delle due variabili sono rappresentati sui due assi cartesiani. Ciò consente di visualizzare la relazione tra le due variabili e identificare eventuali pattern o tendenze.

1.4 Coefficiente di correlazione campionaria

Il coefficiente di correlazione campionaria è una misura di quanto due variabili aleatorie siano linearmente correlate tra loro in un campione di dati. Il coefficiente di correlazione può assumere valori compresi tra -1 e 1, dove un valore di 1 indica una correlazione positiva perfetta, un valore di -1 indica una correlazione negativa perfetta, e un valore di 0 indica assenza di correlazione.

Il coefficiente di correlazione campionaria si calcola come il rapporto tra la covarianza campionaria delle due variabili e il prodotto delle deviazioni standard campionarie delle due variabili. In formule:

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

dove n è la dimensione del campione, x_i e y_i sono i valori osservati delle due variabili nel campione, $\bar{x}$ e $\bar{y}$ sono le medie campionarie delle due variabili, e $\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ e $\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ sono le deviazioni standard campionarie delle due variabili.

Un valore di r vicino a 1 indica una forte correlazione positiva tra le due variabili, mentre un valore di r vicino a -1 indica una forte correlazione negativa tra le due variabili. Un valore di r vicino a 0 indica che le due variabili non sono correlate linearmente.

Capitolo 2

Elementi di probabilità

2.1 Spazio degli esiti ed eventi

Lo spazio degli esiti è l'insieme di tutti gli esiti possibili di un esperimento.

2.2 Spazio degli esiti equiprobabili

Lo spazio degli esiti equiprobabili è un concetto fondamentale della probabilità e si riferisce all'insieme di tutti gli esiti possibili di un esperimento casuale in cui ogni esito ha la stessa probabilità di occorrere. In altre parole, è uno spazio campionario in cui ogni esito ha la stessa probabilità di accadere.

La probabilità che accada 1 evento dello spazio degli esiti equiprobabili è:

$$P(E) = \frac{1}{N} \tag{2.1}$$

Dove N è il numero di esiti possibili.

2.3 Assiomi della probabilità

- la probabilità di un evento è compresa tra 0 e 1
- la probabilità dello spazio degli esiti è 1
- unendo due eventi la loro probabilità si somma

2.4 Coefficiente binomiale

Il coefficiente binomiale è il numero di sottoinsiemi di k elementi che si possono formare a partire da un insieme di n elementi.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.2}$$

2.5 Probabilità condizionata

La probabilità condizionata è una misura di probabilità che tiene conto di una certa informazione o condizione nota. In altre parole, la probabilità di un evento A, dato che un evento B si è verificato, viene calcolata in base alla conoscenza del verificarsi di B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.3}$$

2.6 Fattorizzazione di un evento e formula di Bayes

La fattorizzazione di un evento è un metodo che consente di scrivere la probabilità di un evento composto come prodotto delle probabilità di eventi più semplici. In altre parole, si tratta di scomporre un evento complesso in eventi più elementari, al fine di semplificare il calcolo della probabilità complessiva.

La formula di Bayes, invece, è un teorema della teoria della probabilità che permette di calcolare la probabilità condizionata di un evento A, data la conoscenza di un evento B. In particolare, la formula di Bayes afferma che:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.4}$$

dove $P(A|B)$ rappresenta la probabilità di A dato B, $P(B|A)$ rappresenta la probabilità di B dato A, $P(A)$ rappresenta la probabilità di A, e $P(B)$ rappresenta la probabilità di B.

2.7 Eventi indipendenti

Due eventi A e B si dicono **indipendenti** se la probabilità di A non viene influenzata dalla conoscenza dell'avvenimento B e viceversa, ovvero se:

$$P(A|B) = P(A) \tag{2.5}$$

$$P(B|A) = P(B) \tag{2.6}$$

In altre parole, la conoscenza di uno degli eventi non fornisce alcuna informazione utile per determinare la probabilità dell'altro evento. Ad esempio, se si lancia una moneta equilibrata due volte, il risultato del primo lancio non influisce sulla probabilità di ottenere testa o croce al secondo lancio.

Se due eventi sono indipendenti, allora la probabilità dell'evento composto A e B è data dal prodotto delle probabilità di A e B:

$$P(A \cap B) = P(A) \cdot P(B) \tag{2.7}$$

Capitolo 3

Variabili aleatorie e valore atteso

In teoria delle probabilità, una variabile aleatoria è una funzione che assegna un valore numerico a ciascun possibile esito di un esperimento casuale. In altre parole, una variabile aleatoria è una quantità che può assumere diversi valori a seconda dell'esito dell'esperimento.

Ad esempio, se si lancia un dado equilibrato, il numero che esce è una variabile aleatoria, che può assumere i valori da 1 a 6 con uguale probabilità.

Il valore atteso di una variabile aleatoria è il valore medio che ci si aspetta di ottenere se si ripete l'esperimento un gran numero di volte. In altre parole, è una sorta di media ponderata dei possibili valori che la variabile aleatoria può assumere, in cui i pesi sono le rispettive probabilità.

Il valore atteso di una variabile aleatoria X si indica con $E(X)$ e si calcola come:

$$E(X) = \sum x \times P(X = x) \tag{3.1}$$

dove la somma è estesa a tutti i possibili valori x che X può assumere, e $P(X = x)$ è la probabilità che X assuma il valore x.

Ad esempio, se si considera la variabile aleatoria che indica il numero uscito lanciando un dado equilibrato, il valore atteso è:

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 + = 3.5 \tag{3.2}$$

Questo significa che se si ripete il lancio del dado un gran numero di volte, ci si può aspettare che la media dei risultati si avvicini a 3.5.

Il valore atteso è una misura importante della "centralità" di una variabile aleatoria, e consente di fare previsioni sul comportamento medio della variabile stessa. Ad esempio, se si conosce il valore atteso di una variabile aleatoria X, è possibile stimare la probabilità che X assuma valori superiori o inferiori a una determinata soglia.

3.1 Funzione di ripartizione

La funzione di ripartizione di una variabile aleatoria X si indica con $F(X)$ e si definisce come:

$$F(X) = P(X \leq x)$$

dove x è un valore qualsiasi, e $P(X \leq x)$ rappresenta la probabilità che la variabile aleatoria X assuma un valore minore o uguale a x.

La funzione di ripartizione ha le seguenti proprietà:

1. $F(X)$ è una funzione non-decrescente, ovvero per ogni valore di x_1 e x_2 tali che $x_1 \leq x_2$, si ha $F(x_1) \leq F(x_2)$.
2. $F(X)$ è limitata superiormente da 1 e inferiormente da 0, ovvero $F(-\infty) = 0$ e $F(+\infty) = 1$.
3. La probabilità che la variabile aleatoria X assuma un valore compreso tra due valori x_1 e x_2 si può calcolare come la differenza tra le rispettive probabilità cumulative, ovvero:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

La funzione di ripartizione è una delle proprietà fondamentali di una variabile aleatoria, in quanto consente di calcolare molte altre quantità importanti, come la media e la varianza della variabile stessa. Inoltre, permette di effettuare test di ipotesi e di costruire intervalli di confidenza sulla base dei dati osservati.

3.2 Variabili aleatorie discrete e continue

Una variabile aleatoria si dice discreta se può assumere solo un numero finito o numerabile di valori distinti. Ad esempio, il numero di facce che esce dal lancio di un dado è una variabile aleatoria discreta, in quanto può assumere solo i valori 1, 2, 3, 4, 5, o 6.

Una variabile aleatoria si dice continua se può assumere un numero infinito di valori in un intervallo. Ad esempio, la lunghezza di un filo può essere una variabile aleatoria continua, in quanto può assumere qualunque valore in un intervallo continuo di valori.

- Per le variabili aleatorie discrete, la funzione di probabilità assegna una probabilità a ciascun valore possibile che può assumere la variabile. Questa funzione è spesso rappresentata da un grafico a barre, in cui l'altezza di ogni barra corrisponde alla probabilità associata a un valore particolare.
- Per le variabili aleatorie continue, la funzione di probabilità assume una forma di densità di probabilità, che descrive la probabilità di trovare la variabile in un intervallo di valori. Questa funzione può essere rappresentata da un grafico a linea, in cui l'area sottostante alla curva corrisponde alla probabilità di trovare la variabile in un intervallo specifico.

3.3 Coppie e vettori di variabili aleatorie

La funzione di ripartizione congiunta di X e Y è:

$$F(x, y) := P(X \leq x, Y \leq y) \tag{3.3}$$

Dove la **virgola** denota l'**intersezione** degli eventi.

3.3.1 Distribuzione congiunta - variabili aleatorie discrete

$$p(x_i, y_i) := P(X = x_i, Y = y_j) \tag{3.4}$$

È la **funzione di massa di probabilità congiunta**.

Le funzioni di massa individuali si possono ottenere:

$$p_X(x_i) := P(X = x_i) = \sum_j p(x_i, y_j) \tag{3.5}$$

3.3.2 Distribuzione congiunta - variabili aleatorie continue

$$P(X \in A, Y \in B) = \int_B \int_A f(x, y) dx dy \tag{3.6}$$

È la **densità congiunta**.

Per ricavare le individuali:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \tag{3.7}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx \tag{3.8}$$

3.3.3 Variabili aleatorie indipendenti

Due variabili aleatorie sono indipendenti se tutti gli eventi relativi alla prima sono indipendenti dalal seconda e viceversa.

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{3.9}$$

3.3.4 Distribuzioni condizionali

Le distribuzioni condizionali di variabili aleatorie si riferiscono alla distribuzione di una variabile aleatoria data un'informazione o un vincolo su un'altra variabile aleatoria. In altre parole, quando abbiamo informazioni su una variabile aleatoria, possiamo utilizzare tali informazioni per stimare la distribuzione di un'altra variabile aleatoria.

La distribuzione condizionale di X data Y = y è definita come:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

3.4 Valore atteso

Il valore atteso (o media) di una variabile aleatoria è una misura della tendenza centrale dei suoi possibili valori. In altre parole, il valore atteso rappresenta il valore medio che ci si aspetta di ottenere da una variabile aleatoria se viene ripetutamente campionata.

Il valore atteso di una variabile aleatoria discreta X è definito come:

$$E[X] = \sum_i x_i P(X = x_i) \tag{3.10}$$

per una variabile aleatoria continua, il valore atteso è definito come:

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx \tag{3.11}$$

3.5 Proprietà del valore atteso

- Linearità: il valore atteso di una somma di variabili aleatorie è la somma dei loro valori attesi. In altre parole, se X e Y sono variabili aleatorie e a e b sono costanti, allora:

$$E[aX + bY] = aE[X] + bE[Y]$$

- Additività: il valore atteso di una funzione di una variabile aleatoria è la somma dei valori attesi della funzione per ciascun valore della variabile. In altre parole, se $g(X)$ è una funzione di X, allora:

$$E[g(X)] = \sum_x g(x) P(X = x)$$

- Monotonia: se X e Y sono variabili aleatorie tali che $X \leq Y$, allora:

$$E[X] \leq E[Y]$$

- Indipendenza: se X e Y sono variabili aleatorie indipendenti, allora:

$$E[XY] = E[X]E[Y]$$

- Variance: la varianza di una variabile aleatoria X è definita come:

$$Var[X] = E[(X - E[X])^2]$$

  e può essere espressa come:

$$Var[X] = E[X^2] - (E[X])^2$$

3.6 Covarianza e varianza della somma di variabili aleatorie

La covarianza tra due variabili aleatorie X e Y è definita come:

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$$

La varianza della somma di due variabili aleatorie è data da:

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

Se le variabili sono indipendenti, la covarianza è zero e l'equazione si riduce a:

$$Var[X + Y] = Var[X] + Var[Y]$$

Capitolo 4

Funzione generatrice dei momenti

La funzione generatrice dei momenti è utile perché fornisce una rappresentazione compatta delle informazioni sui momenti di una variabile aleatoria. Inoltre, se due variabili aleatorie hanno la stessa funzione generatrice dei momenti, allora hanno gli stessi momenti e quindi la stessa distribuzione di probabilità.

Infine, la funzione generatrice dei momenti è particolarmente utile per variabili aleatorie discrete, poiché consente di calcolare facilmente le probabilità e le statistiche associate a queste variabili.

Nello specifico quando si tratta di **variabili aleatorie discrete**:

$$\Phi(t) = E[e^{tX}] = \sum_x e^{tx} p(x) \tag{4.1}$$

Nello specifico quando si tratta di **variabili aleatorie continua**:

$$\Phi(t) = E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f(x) dx \tag{4.2}$$

Derivando una volta il moto nell'origine si ottiene:

$$\Phi'(0) = E[X] \tag{4.3}$$

In generale si ha che:

$$\Phi^n(0) = E[X^n] \tag{4.4}$$

Se X e Y sono variabili aleatorie **indipendenti** con funzioni generatrici Φ_X e Φ_Y, e se Φ_{X+Y} è la funzione generatrice dei momenti di $X + Y$, allora:

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t) \tag{4.5}$$

Capitolo 5

La legge debole dei grandi numeri