Analisi dei dati

Un goliardico riassunto

Ollari Dmitri

13 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

La legge debole dei grandi numeri afferma che, se si considera una sequenza di variabili aleatorie indipendenti ed identicamente distribuite $X_1, X_2, \ldots, X_n$, la media campionaria $\bar{X_n}$ converge in probabilità al valore atteso della variabile aleatoria, ovvero:

$$\lim_{n\to\infty} P(|\bar{X_n} - \mu| < \epsilon) = 1$$

dove μ è il valore atteso di X_i e ϵ è un valore positivo arbitrario. In altre parole, la probabilità che la media campionaria si discosti dal valore atteso di più di un valore prefissato ϵ diventa sempre più piccola all'aumentare del numero di osservazioni n.

Questa legge fornisce una garanzia formale dell'andamento della media campionaria al crescere del numero di osservazioni e rappresenta un importante risultato in teoria della probabilità e nelle applicazioni pratiche dell'analisi dei dati.

Capitolo 6

Modelli di variabili aleatorie

6.1 Variabili aleatorie di Bernoulli

Una variabile aleatoria di Bernoulli è una variabile aleatoria discreta che assume valore 1 con probabilità p e valore 0 con probabilità $1-p$, dove $0 \leq p \leq 1$. Ad esempio, la variabile aleatoria che rappresenta il lancio di una moneta equilibrata è di tipo Bernoulli, in quanto assume valore 1 se il risultato del lancio è testa e 0 se è croce.

$$P(X = 0) = 1 - p \tag{6.1}$$

$$P(X = 1) = p \tag{6.2}$$

6.2 Variabili aleatorie binomiale

Una variabile aleatoria binomiale è una variabile aleatoria discreta che rappresenta il numero di successi in una sequenza di n prove indipendenti, ciascuna con probabilità di successo p. La variabile aleatoria binomiale viene indicata con $X \sim B(n, p)$ e assume i valori $0, 1, 2, \ldots, n$. La sua funzione di probabilità è data da:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

dove $\binom{n}{k}$ è il coefficiente binomiale, che rappresenta il numero di modi diversi in cui è possibile ottenere k successi in n prove. In altre parole, la variabile aleatoria binomiale conta il numero di volte che si ottiene un certo evento in una serie di prove ripetute indipendenti.

Il valore atteso di una variabile aleatoria binomiale è dato da:

$$E[X] = np$$

mentre la sua varianza è:

$$Var(X) = np(1-p)$$

6.2.1 Calcolo esplicitp della distribuzione binomiale

Supponendo X una binomiale di parametri (n, p), la funzione di ripartizione è:

$$P(X \leq i) = \sum_{k=0}^{i} \binom{n}{i} p^k (1-p)^{n-k} \tag{6.3}$$

Per calcolare la funzione di massa:

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \tag{6.4}$$

6.3 Variabile aleatoria di Poisson

La variabile aleatoria di Poisson è una variabile aleatoria discreta che descrive il numero di eventi rari che si verificano in un intervallo di tempo o di spazio, dato il tasso di occorrenza di tali eventi. Ad esempio, il numero di chiamate che un centralino telefonico riceve in un minuto, il numero di incidenti stradali in un'ora, il numero di particelle radioattive che decadono in un secondo.

La variabile aleatoria di Poisson è indicata con X e si esprime attraverso un parametro λ che rappresenta il tasso di occorrenza degli eventi. La sua funzione di probabilità è data dalla seguente formula:

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

Mentre la media e la varianza sono:

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

Una caratteristica interessante della Poissoniana è che può approssimare una binomiale con parametri (n, p) quando n è molto grande e p molto piccolo, ponendo $\lambda = np$:

$$P(X = i) \approx \frac{\lambda^i}{i!}e^{-\lambda} \tag{6.5}$$

6.3.1 Calcolo esplicito della distribuzione di Poisson

Il calcolo esplicito della distribuzione di Poisson si basa sulla seguente formula per la funzione di probabilità:

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

dove k è il numero di eventi che si verificano in un certo intervallo di tempo o di spazio, λ è il parametro della distribuzione che rappresenta il numero medio di eventi in quell'intervallo.

Per calcolare la probabilità di ottenere un certo numero di eventi k, basta sostituire i valori di k e λ nella formula e fare i calcoli. Ad esempio, supponiamo di voler calcolare la probabilità di osservare esattamente 2 eventi in un intervallo di tempo di un'ora, sapendo che in media accadono 3 eventi in un'ora. In questo caso, la formula diventa:

$$P(X = 2) = \frac{e^{-3}3^2}{2!} \approx 0.224$$

Quindi la probabilità di osservare esattamente 2 eventi in un'ora, con una media di 3 eventi in un'ora, è di circa il 22,4

6.4 Variabili aleatorie ipergeometriche

Le variabili aleatorie ipergeometriche sono utilizzate per modellare situazioni in cui si effettuano estrazioni senza reinserimento da un insieme finito di elementi che si distinguono in due categorie (ad esempio, in un lotto di componenti elettronici, quelli funzionanti e quelli difettosi).

In particolare, si consideri un insieme di N elementi, dei quali K appartengono alla categoria 1 e $N - K$ alla categoria 2. Si effettuano n estrazioni senza reinserimento e si vuole determinare la probabilità che k estrazioni diano un esito di categoria 1. La variabile aleatoria X che misura il numero di successi (cioè il numero di elementi estratti appartenenti alla categoria 1) è una variabile aleatoria ipergeometrica.

La funzione di probabilità di una variabile aleatoria ipergeometrica è data dalla seguente formula:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

dove $\binom{a}{b}$ rappresenta il coefficiente binomiale che indica il numero di modi in cui si possono scegliere b elementi da un insieme di a elementi.

Il valore atteso di una variabile aleatoria ipergeometrica è:

$$E(X) = n\frac{K}{N}$$

mentre la varianza è:

$$Var(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$$

Le variabili aleatorie ipergeometriche sono utili, ad esempio, nella valutazione della qualità di un lotto di componenti elettronici, in cui si vuole stimare la proporzione di componenti funzionanti effettuando un campionamento senza reinserimento.

6.5 Variabili aleatorie uniformi

Una variabile aleatoria uniforme è una distribuzione di probabilità in cui ogni valore possibile dell'intervallo di una variabile casuale ha la stessa probabilità di essere scelto. In altre parole, la probabilità di ottenere un valore in un certo intervallo è proporzionale alla lunghezza dell'intervallo.

Esistono due tipi di variabili aleatorie uniformi: la variabile aleatoria uniforme continua e la variabile aleatoria uniforme discreta.

La variabile aleatoria uniforme continua è caratterizzata dalla funzione di densità di probabilità (pdf):

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{per } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$

dove a e b sono i limiti inferiori e superiori dell'intervallo di valori possibili per la variabile casuale. La funzione di distribuzione cumulativa (cdf) è data da:

$$F(x) = \begin{cases} 0 & \text{per } x < a \\ \frac{x-a}{b-a} & \text{per } a \leq x \leq b \\ 1 & \text{per } x > b \end{cases}$$

La variabile aleatoria uniforme discreta è caratterizzata dalla funzione di massa di probabilità (pmf):

$$P(X = k) = \begin{cases} \frac{1}{n} & \text{per } k = 1, 2, \ldots, n \\ 0 & \text{altrimenti} \end{cases}$$

dove n è il numero di valori possibili per la variabile casuale. La funzione di distribuzione cumulativa (cdf) è data da:

$$F(k) = \begin{cases} 0 & \text{per } k < 1 \\ \frac{k-1}{n-1} & \text{per } 1 \leq k \leq n \\ 1 & \text{per } k > n \end{cases}$$

In entrambi i casi, il valore atteso della variabile casuale è dato dalla formula:

$$E[X] = \frac{a+b}{2} \text{ o } \frac{n+1}{2}$$

a seconda che si tratti di una variabile aleatoria uniforme continua o discreta. La varianza della variabile casuale è invece data da:

$$Var(X) = \frac{(b-a)^2}{12} \text{ o } \frac{n^2-1}{12}$$

a seconda che si tratti di una variabile aleatoria uniforme continua o discreta.

6.6 Variabili aleatorie normali

Le variabili aleatorie normali (o gaussiane) sono molto importanti nella teoria della probabilità e nelle applicazioni pratiche, poiché molti fenomeni naturali e sociali seguono una distribuzione normale.

Una variabile aleatoria normale, indicata con X, è caratterizzata da due parametri: la media μ e la deviazione standard σ. La sua funzione di densità di probabilità (pdf) è data dalla seguente formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

dove e è il numero di Nepero (costante matematica approssimativamente pari a 2,71828), μ è il rapporto tra la circonferenza e il diametro di un cerchio (approssimativamente pari a 3,14159), μ rappresenta il valore atteso della variabile e σ la sua deviazione standard.

La funzione di distribuzione cumulativa (cdf) di X è invece data dalla seguente formula:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

La curva di una distribuzione normale è a forma di campana, simmetrica rispetto alla media μ. Il suo valore atteso, la mediana e la moda coincidono, ovvero:

$$E[X] = \text{mediana}(X) = \text{moda}(X) = \mu$$

Inoltre, il 68% dei dati si trova entro un intervallo di una deviazione standard dalla media, il 95% entro due deviazioni standard e il $99,7\%$ entro tre deviazioni standard. Questa proprietà è nota come la regola empirica o regola del 68-95-99,7.

Le variabili aleatorie normali standardizzate, indicate con Z, sono ottenute dalla trasformazione:

$$Z = \frac{X - \mu}{\sigma}$$

e hanno media zero e deviazione standard pari a uno. La funzione di distribuzione cumulativa di Z è indicata con $\Phi(z)$ e viene chiamata funzione di distribuzione cumulativa standard normale. Esistono diverse tabelle o calcolatori online che permettono di calcolare le probabilità associate a $\Phi(z)$ per diversi valori di z.

6.7 Variabili aleatorie esponenziali

Le variabili aleatorie esponenziali sono un tipo di distribuzione di probabilità continua che descrive il tempo tra due eventi successivi in un processo di Poisson. Ad esempio, se si stanno monitorando gli arrivi di clienti in un negozio e si vuole sapere quanto tempo passerà tra un cliente e il successivo, la distribuzione esponenziale può essere usata per modellare il tempo di attesa.

La funzione densità di probabilità (pdf) della distribuzione esponenziale è:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

dove λ è il parametro di scala della distribuzione. Il valore atteso della distribuzione esponenziale è dato da:

$$E[X] = \frac{1}{\lambda}$$

e la deviazione standard è:

$$Var(X) = \frac{1}{\lambda^2}$$

Inoltre, la distribuzione esponenziale soddisfa la proprietà di mancanza di memoria, che significa che la probabilità che un evento si verifichi dopo un certo intervallo di tempo dipende solo dal tempo trascorso e non dal tempo trascorso finora. Questa proprietà la rende utile per modellare eventi rari e imprevisti, come guasti a macchinari o incidenti.

6.7.1 Processo di Poisson

Il processo di Poisson è un processo stocastico che descrive il numero di eventi che si verificano in un intervallo di tempo, assumendo che questi eventi si verifichino in modo casuale e indipendente nel tempo.

In forma matematica, il processo di Poisson è definito come:

$$N(t) \sim \text{Pois}(\lambda t)$$

dove $N(t)$ è la variabile aleatoria che rappresenta il numero di eventi che si verificano nell'intervallo di tempo $[0, t]$ e λ è il parametro di intensità del processo, che rappresenta il numero medio di eventi che si verificano in unità di tempo.

La funzione di probabilità di $N(t)$ è data dalla distribuzione di Poisson:

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

dove k è il numero di eventi che si verificano nell'intervallo di tempo $[0, t]$.

Il processo di Poisson gode delle proprietà di stazionarietà e indipendenza dei tratti, il che significa che il numero di eventi che si verificano in un intervallo di tempo dipende solo dalla lunghezza dell'intervallo e non dalla posizione dell'intervallo nel tempo, e che gli eventi che si verificano in diversi intervalli di tempo sono indipendenti tra loro.

6.8 Variabili aleatorie di tipo gamma

La variabile aleatoria di tipo gamma è una generalizzazione della distribuzione esponenziale e della distribuzione di Poisson. Una variabile aleatoria X si dice distribuita secondo una legge di tipo gamma con parametri α e λ se la sua funzione di densità di probabilità è data da:

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

dove $\Gamma(\alpha)$ è la funzione Gamma, definita come:

$$\Gamma(\alpha) = (n-1)!$$

Il parametro α è un intero positivo e rappresenta il numero di eventi che si verificano in un certo intervallo di tempo, mentre il parametro λ rappresenta la frequenza con cui questi eventi si verificano.

La variabile aleatoria di tipo gamma è particolarmente utile per modellare processi di conteggio con tassi di eventi variabili nel tempo. In particolare, se X_i è la durata del i-esimo intervallo di tempo tra due eventi consecutivi, allora la somma $Y_n = X_1 + X_2 + \cdots + X_n$ segue una distribuzione di tipo gamma con parametri $\alpha = n$ e λ uguale alla frequenza media degli eventi.

Il valore atteso di una variabile aleatoria di tipo gamma è pari a $\dfrac{\alpha}{\lambda}$, mentre la varianza è pari a $\dfrac{\alpha}{\lambda^2}$.

6.8.1 Relazione tra chi-quadro e gamma

La relazione tra la distribuzione chi-quadrato e la distribuzione gamma è che la distribuzione chi-quadrato è un caso particolare della distribuzione gamma.

In particolare, se una variabile aleatoria Z segue una distribuzione normale standard, allora la somma dei quadrati di k campioni estratti da Z segue una distribuzione chi-quadrato con k gradi di libertà.

Inoltre, se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti e identicamente distribuite con distribuzione normale standard, allora la somma dei loro quadrati segue una distribuzione chi-quadrato con n gradi di libertà.

La densità di probabilità della distribuzione gamma è data da:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad x \geq 0$$

dove α e β sono i parametri della distribuzione, e $\Gamma(\alpha)$ è la funzione Gamma.

La distribuzione chi-quadrato con k gradi di libertà è una distribuzione gamma con $\alpha = k/2$ e $\beta = 1/2$. In particolare, la densità di probabilità della distribuzione chi-quadrato con k gradi di libertà è:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad x \geq 0$$

Quindi la distribuzione chi-quadrato è un caso particolare della distribuzione gamma, dove i parametri della distribuzione gamma sono determinati dal numero di gradi di libertà della distribuzione chi-quadrato.

6.9 Le distribuzioni T

Le distribuzioni T sono una famiglia di distribuzioni di probabilità continue che hanno una forma simile alla distribuzione normale standard, ma che dipendono anche dal parametro chiamato "gradi di libertà".

In particolare, la distribuzione T con k gradi di libertà è definita come la distribuzione della variabile casuale:

$$T = \frac{Z}{\sqrt{V/k}}$$

dove Z è una variabile casuale standard normale, V è una variabile casuale chi-quadro con k gradi di libertà e Z e V sono indipendenti.

La distribuzione T è spesso utilizzata quando la deviazione standard della popolazione non è nota e deve essere stimata dalla deviazione standard campionaria. In questo caso, la statistica del test T è definita come:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

dove $\bar{X}$ è la media campionaria, μ è la media della popolazione, S è la deviazione standard campionaria e n è la dimensione del campione. La distribuzione T con n-1 gradi di libertà viene utilizzata per calcolare la probabilità di ottenere una statistica T data una determinata ipotesi sulla media della popolazione.

Le distribuzioni T sono utilizzate in diverse applicazioni statistiche, come ad esempio nell'analisi dei test t, nella regressione lineare, nella valutazione delle differenze tra gruppi e in altri tipi di analisi.

Il valore atteso di una distribuzione T non è zero in generale, ma dipende dai gradi di libertà della distribuzione. In particolare, se la distribuzione T ha n gradi di libertà, il valore atteso è zero se $n > 1$, mentre se $n = 1$ il valore atteso non esiste.

La varianza di una distribuzione T con n gradi di libertà è pari a $\frac{n}{n-2}$ se $n > 2$, mentre se $n \leq 2$ la varianza non esiste.

6.10 Le distribuzioni F

La distribuzione F è una distribuzione di probabilità continua che compare comunemente nell'ambito dell'analisi della varianza (ANOVA) e dei test statistici basati sul rapporto di due varianze.

La distribuzione F ha due gradi di libertà, uno per il numeratore e uno per il denominatore. La distribuzione F è quindi definita da due parametri, noti come gradi di libertà del numeratore e del denominatore.

Il valore atteso della distribuzione F dipende dai gradi di libertà del numeratore e del denominatore ed è definito come:

$$E(X) = \begin{cases} \dfrac{d_2}{d_2 - 2} & \text{se } d_2 > 2 \\ \text{undefined} & \text{se } d_2 \leq 2 \end{cases}$$

dove d_2 rappresenta i gradi di libertà del denominatore.

La varianza della distribuzione F dipende anch'essa dai gradi di libertà del numeratore e del denominatore ed è definita come:

$$Var(X) = \begin{cases} \dfrac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)} & \text{se } d_2 > 4 \\ \text{undefined} & \text{se } d_2 \leq 4 \end{cases}$$

dove d_1 rappresenta i gradi di libertà del numeratore.

La distribuzione F è simmetrica rispetto al valore 1 e assume valori positivi. La sua forma dipende dai gradi di libertà del numeratore e del denominatore e ha una coda lunga sulla destra quando il denominatore è piccolo rispetto al numeratore. La distribuzione F viene solitamente utilizzata per confrontare la varianza di due campioni indipendenti.

6.11 Distribuzione logistica

La distribuzione logistica è una distribuzione di probabilità continua utilizzata spesso in statistica e modellistica per descrivere l'evoluzione di un processo nel tempo.

La funzione di densità di probabilità (PDF) della distribuzione logistica è data da:

$$f(x) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}$$

dove μ è il parametro di posizione che indica il valore atteso della distribuzione e s è il parametro di scala. La funzione di distribuzione cumulativa (CDF) può essere ottenuta integrando la PDF:

$$F(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Il parametro di scala s controlla l'inclinazione della curva della distribuzione logistica. Più grande è il valore di s, più stretta sarà la curva e più concentrati saranno i dati attorno al valore atteso μ.

La distribuzione logistica può essere utilizzata per modellare diverse variabili aleatorie, come ad esempio la distribuzione dei punteggi dei test standardizzati o la distribuzione della crescita di popolazioni biologiche.

Capitolo 7

La distribuzione delle statistiche campionarie

La distribuzione delle statistiche campionarie si riferisce alla distribuzione di probabilità di una qualsiasi quantità calcolata da un campione casuale, come la media campionaria, la varianza campionaria o la proporzione campionaria. Questa distribuzione è importante perché ci consente di fare inferenze sulla popolazione a partire dai dati del campione.

In generale, la distribuzione di una statistica campionaria dipende dalla distribuzione della variabile casuale sottostante nella popolazione, dalla dimensione del campione e dal tipo di statistica. Ad esempio, se la variabile casuale sottostante è distribuita normalmente, allora la media campionaria segue anche una distribuzione normale, indipendentemente dalla dimensione del campione, mentre la somma campionaria segue una distribuzione normale solo se il campione è grande abbastanza da soddisfare le condizioni del teorema del limite centrale.

Inoltre, la distribuzione delle statistiche campionarie può essere utilizzata per costruire intervalli di confidenza per i parametri della popolazione e per testare le ipotesi sulla popolazione. Ad esempio, se vogliamo testare se la media della popolazione è uguale a un valore specifico, possiamo calcolare la statistica t come rapporto tra la differenza tra la media campionaria e il valore specificato e l'errore standard della media campionaria, e utilizzare la distribuzione t di Student per calcolare la probabilità di osservare un valore della statistica t almeno estremo come quello che abbiamo osservato.

7.1 La media campionaria

La media campionaria è una statistica campionaria che viene utilizzata per stimare il valore medio della popolazione da cui è stata estratta la campione. Essa rappresenta la media aritmetica dei valori osservati del campione e viene calcolata attraverso la seguente formula:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

dove $\bar{X}$ rappresenta la media campionaria, n rappresenta la dimensione del campione e X_i rappresenta il valore della i-esima osservazione nel campione.

La media campionaria è uno stimatore non distorto del valore medio della popolazione, il che significa che, in media, la media campionaria si avvicina al valore medio della popolazione. Inoltre, la distribuzione della media campionaria segue la distribuzione normale (o gaussiana) se la dimensione del campione è sufficientemente grande, grazie al teorema del limite centrale. Questo permette di utilizzare la media campionaria per effettuare inferenze sulla popolazione, come ad esempio stimare un intervallo di confidenza per il valore medio della popolazione o testare un'ipotesi riguardante il valore medio della popolazione.

Il valore atteso della media campionaria corrisponde al valore medio della popolazione, ovvero:

$$E(\bar{X}) = \mu$$

dove μ rappresenta il valore medio della popolazione.

La varianza della media campionaria dipende dalla varianza della popolazione e dalla dimensione del campione, ed è data dalla seguente formula:

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

dove σ^2 rappresenta la varianza della popolazione e n rappresenta la dimensione del campione.

Questa formula indica che la varianza della media campionaria diminuisce all'aumentare della dimensione del campione. In altre parole, aumentare la dimensione del campione riduce la variabilità della media campionaria e la rende quindi più affidabile come stima del valore medio della popolazione.

7.2 Il teorema del limite centrale

Il teorema del limite centrale afferma che, se consideriamo una sequenza di variabili aleatorie indipendenti e identicamente distribuite con media μ e varianza σ^2, allora la somma di queste variabili aleatorie si avvicina sempre di più alla distribuzione normale all'aumentare della dimensione del campione.

In particolare, il teorema afferma che la media campionaria $\bar{X}$ di un grande numero di osservazioni estratte da una popolazione con media μ e varianza σ^2 si avvicina sempre di più alla distribuzione normale standard ($\mu = 0$, $\sigma = 1$) all'aumentare della dimensione del campione.

Questo significa che, anche se la distribuzione della popolazione non è normale, la distribuzione della media campionaria diventa sempre più simile a una distribuzione normale all'aumentare della dimensione del campione. In altre parole, la distribuzione della media campionaria diventa sempre più simmetrica e approssimativamente normale, indipendentemente dalla forma della distribuzione della popolazione.

Questo teorema ha importanti implicazioni pratiche in ambito statistico, perché consente di utilizzare la distribuzione normale per fare inferenze sulla media della popolazione anche quando la distribuzione della popolazione non è nota o non è normale, a patto che il campione sia abbastanza grande.

7.3 La varianza campionaria

La varianza campionaria è una stima della varianza della popolazione sulla base di un campione. La formula per la varianza campionaria è data da:

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

dove n è la dimensione del campione, x_i è il valore della i-esima osservazione nel campione, e $\bar{x}$ è la media campionaria, definita come:

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

La varianza campionaria è una stima non distorta della varianza della popolazione, nel senso che il valore atteso della varianza campionaria è uguale alla varianza della popolazione. Tuttavia, la varianza campionaria tende a sottostimare la varianza della popolazione, in quanto è basata su un campione e non sull'intera popolazione.

Inoltre, la varianza campionaria è un'importante misura di dispersione dei dati all'interno del campione. Più è grande la varianza campionaria, maggiore è la dispersione dei dati all'interno del campione.

7.4 Le distribuzioni delle statistiche di popolazioni normali

La distribuzione delle statistiche delle popolazioni normali è anch'essa una distribuzione normale. Ci sono due importanti statistiche delle popolazioni normali, ovvero la media campionaria e la varianza campionaria.

La media campionaria di un campione di dimensione n da una popolazione normale con media μ e varianza σ^2 è data da:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

dove X_i sono i valori del campione.

La media campionaria segue una distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$, ovvero:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

La varianza campionaria S^2 è un'altra importante statistica delle popolazioni normali. La varianza campionaria è definita come:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

dove X_i sono i valori del campione e $\bar{X}$ è la media campionaria.

La varianza campionaria segue una distribuzione chi-quadrato con $n-1$ gradi di libertà, ovvero:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

dove σ^2 è la varianza della popolazione.

7.5 Campionamento da insiemi finiti

Il campionamento da insiemi finiti è una tecnica utilizzata in statistica per selezionare un sottoinsieme casuale di elementi da una popolazione finita. Questo processo di campionamento viene utilizzato per studiare le proprietà della popolazione sulla base delle informazioni raccolte dal campione selezionato.

In pratica, si sceglie un campione casuale di dimensione n dalla popolazione finita di N elementi in modo tale che ogni elemento della popolazione abbia la stessa probabilità di essere selezionato. Si possono utilizzare diversi metodi per selezionare il campione, ad esempio il campionamento casuale semplice o il campionamento sistematico.

Il campionamento da insiemi finiti permette di stimare la media, la varianza, la proporzione e altre proprietà della popolazione sulla base delle statistiche calcolate sul campione selezionato. In particolare, la media campionaria, la varianza campionaria e la proporzione campionaria sono stime utilizzate per stimare rispettivamente la media, la varianza e la proporzione della popolazione.

L'errore standard della media campionaria è definito come la deviazione standard della distribuzione campionaria della media e rappresenta l'incertezza associata alla stima della media della popolazione sulla base del campione selezionato. L'errore standard della media campionaria diminuisce all'aumentare della dimensione del campione selezionato.

In sintesi, il campionamento da insiemi finiti è un metodo comune per studiare le proprietà delle popolazioni finite e per stimare le statistiche della popolazione sulla base del campione selezionato.

Capitolo 8

Stima parametrica

La stima parametrica è un metodo utilizzato in statistica inferenziale per stimare i parametri di una distribuzione di probabilità che descrive i dati di una popolazione. In questo metodo, si suppone che la distribuzione di probabilità della popolazione sia nota a priori e si cerca di stimare i suoi parametri.

Il processo di stima parametrica coinvolge due fasi: la scelta di una funzione di probabilità appropriata per descrivere la popolazione e la stima dei parametri di questa funzione sulla base dei dati campionati.

Una volta scelta la funzione di probabilità, il passo successivo consiste nell'utilizzare i dati campionati per stimare i parametri della distribuzione. Ci sono diversi metodi di stima dei parametri, tra cui il metodo dei momenti, il metodo della massima verosimiglianza e il metodo di Bayes.

La stima parametrica può essere utile quando si ha una buona conoscenza della distribuzione di probabilità della popolazione, ma non si ha accesso all'intera popolazione. Tuttavia, se la distribuzione di probabilità della popolazione è sconosciuta o non si può assumere, allora si può utilizzare la stima non parametrica.

8.1 Stima di massima verosimilarità

La stima di massima verosimiglianza (maximum likelihood estimation, in inglese) è un metodo di stima parametrica che cerca di ottenere il valore del parametro di una distribuzione di probabilità che rende più probabile l'osservazione dei dati campionari.

In pratica, si suppone che i dati osservati siano stati generati da una distribuzione di probabilità nota, ma con un valore ignoto del parametro. L'obiettivo è di trovare il valore del parametro che massimizza la probabilità (o verosimiglianza) di ottenere i dati osservati.

Ad esempio, supponiamo di avere un campione di n osservazioni estratto da una distribuzione normale con media μ e varianza σ^2 ignote. La stima di massima verosimiglianza cerca di trovare i valori di μ e σ^2 che massimizzano la probabilità di ottenere il campione osservato.

La funzione di verosimiglianza è data dal prodotto delle densità di probabilità dei singoli dati:

$$L(\mu, \sigma^2|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2)$$

dove $\mathbf{x}$ è il vettore delle osservazioni, $f(x_i|\mu, \sigma^2)$ è la densità di probabilità della distribuzione normale con media μ e varianza σ^2 valutata in x_i.

Per trovare i valori di μ e σ^2 che massimizzano $L(\mu, \sigma^2|\mathbf{x})$, si calcola la derivata logaritmica della funzione di verosimiglianza rispetto ai parametri:

$$\frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu} = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \sigma^2} = 0$$

Risolvendo queste equazioni, si ottengono le stime di massima verosimiglianza per μ e σ^2. Queste stime sono note per essere non polarizzate e asintoticamente efficienti, il che significa che per campioni sufficientemente grandi, si avvicinano al valore vero del parametro con una precisione elevata.

8.1.1 Stima dei tempi di vita

La stima della distribuzione dei tempi di vita di una popolazione è un problema comune nell'analisi dei dati. In generale, i tempi di vita possono essere modellati da diverse distribuzioni di probabilità, tra cui la distribuzione esponenziale, la distribuzione di Weibull e la distribuzione di log-normale.

Una tecnica comune per stimare la distribuzione dei tempi di vita è l'analisi della sopravvivenza, che è basata sulla stima della funzione di sopravvivenza della popolazione. La funzione di sopravvivenza descrive la probabilità che un individuo sopravviva fino a un certo momento.

La stima della funzione di sopravvivenza può essere fatta utilizzando la tecnica di Kaplan-Meier, che è basata sull'analisi dei dati di sopravvivenza censurati. La censura si verifica quando la durata dell'evento non è osservata completamente, ad esempio perché il soggetto esce dallo studio prima che l'evento si verifichi o perché lo studio termina prima che tutti i soggetti abbiano sperimentato l'evento.

In alternativa, la distribuzione dei tempi di vita può essere stimata utilizzando la regressione di Cox, che è un metodo di analisi della sopravvivenza basato sulla modellizzazione della funzione di rischio, che è il tasso istantaneo di fallimento in un dato momento. La regressione di Cox permette di modellare l'effetto di variabili predittive sulla sopravvivenza, come ad esempio l'età o il sesso.

In generale, la scelta della tecnica di stima della distribuzione dei tempi di vita dipende dalle caratteristiche dei dati e dagli obiettivi dell'analisi.

8.2 Intervalli di confidenza

Gli intervalli di confidenza sono una tecnica statistica che viene utilizzata per stimare un parametro di una popolazione sconosciuta basandosi sui dati raccolti da un campione. L'obiettivo è quello di fornire un intervallo di valori entro il quale si ritiene che il parametro della popolazione si trovi con una certa probabilità.

In generale, l'intervallo di confidenza si costruisce utilizzando la stima del parametro ottenuta dal campione e l'errore standard della stima. L'errore standard è una misura della variabilità delle stime del parametro che si otterrebbero se si ripetesse il campionamento più volte.

La formula generale per calcolare un intervallo di confidenza per una stima di un parametro è:

intervallo di confidenza = stima del parametro ± margine di errore

dove il margine di errore dipende dal livello di confidenza desiderato e dall'errore standard della stima.

Il livello di confidenza rappresenta la probabilità che l'intervallo di confidenza contenga effettivamente il parametro della popolazione. In generale, i livelli di confidenza più comuni sono il 90%, il 95% e il 99%.

In sintesi, gli intervalli di confidenza forniscono un modo per quantificare l'incertezza nelle stime dei parametri della popolazione e per comunicare la precisione delle stime ai lettori.

Poiché la variabile aleatoria X ha una distribuzione normale standard e una varianza nota, possiamo utilizzare la distribuzione normale standardizzata per calcolare l'intervallo di confidenza.

L'intervallo di confidenza al 95% per la media della variabile aleatoria X è dato da:

$$(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

Dove $\bar{X}$ è la media campionaria, σ è la deviazione standard nota della variabile aleatoria X, e n è la dimensione del campione.

Nota che il valore 1.96 è il quantile corrispondente al livello di confidenza del 95% della distribuzione normale standardizzata.

Sostituendo i valori noti, otteniamo l'intervallo di confidenza al 95%:

$$(\bar{X} - 1.96\frac{1}{\sqrt{n}}, \bar{X} + 1.96\frac{1}{\sqrt{n}})$$

8.2.1 Intervalli di confidenza per la media di una normale con varianza incognita

Gli intervalli di confidenza per la media di una distribuzione normale con varianza incognita si basano sulla distribuzione t di Student.

L'intervallo di confidenza per la media di una distribuzione normale con varianza incognita e una dimensione del campione n inferiore a 30 è dato dalla seguente formula:

$$\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

dove $\bar{X}$ è la media campionaria, s è la deviazione standard campionaria, n è la dimensione del campione e $t_{n-1,\alpha/2}$ è il valore t corrispondente ai gradi di libertà $n-1$ e al livello di confidenza desiderato α.

Ad esempio, per calcolare un intervallo di confidenza al 95% per la media di una distribuzione normale con una dimensione del campione di 25, una media campionaria di 10 e una deviazione standard campionaria di 2, si ha:

$$10 \pm t_{24,0.025}\frac{2}{\sqrt{25}}$$

Il valore t corrispondente ai gradi di libertà 24 e al livello di confidenza del 95% è di circa 2,064. Quindi, l'intervallo di confidenza al 95% per la media è:

$$10 \pm 0.826 = [9.174, 10.826]$$

Ciò significa che con una confidenza del 95%, la media reale della popolazione si trova all'interno dell'intervallo di [9.174, 10.826].

8.2.2 Intervalli di confidenza per la varianza di una distribuzione normale

Gli intervalli di confidenza per la varianza di una distribuzione normale sono definiti in base alla distribuzione campionaria della statistica test $\frac{(n-1)S^2}{\sigma^2}$, dove S^2 è la varianza campionaria e σ^2 è la varianza della popolazione.

Se consideriamo un campione di n osservazioni estratte da una popolazione normale con varianza σ^2 incognita, l'intervallo di confidenza per la varianza al livello di confidenza $1-\alpha$ può essere espresso come:

$$(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}})$$

dove $\chi^2_{\alpha/2,n-1}$ e $\chi^2_{1-\alpha/2,n-1}$ sono i quantili della distribuzione chi-quadro con $n-1$ gradi di libertà al livello di significatività $\alpha/2$ e $1-\alpha/2$, rispettivamente.

In questo modo, l'intervallo di confidenza può essere interpretato come la regione in cui ci aspettiamo che cada la vera varianza della popolazione con una probabilità di $1-\alpha$.

8.3 Stime per la differenza tra le medie di due normali

Per l'intervallo di confidenza per la differenza tra le medie con varianze note e uguali:

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Per l'intervallo di confidenza per la differenza tra le medie con varianze note e diverse:

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Per l'intervallo di confidenza per la differenza tra le medie con varianze sconosciute e uguali:

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2,\, n_1+n_2-2} \cdot \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Per l'intervallo di confidenza per la differenza tra le medie con varianze sconosciute e diverse:

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2,\, \nu} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

dove $\bar{X}_1$ e $\bar{X}_2$ sono le medie campionarie delle due normali, σ_1^2 e σ_2^2 sono le varianze delle due normali, s_1^2 e s_2^2 sono le varianze campionarie delle due normali, n_1 e n_2 sono le dimensioni dei campioni, α è il livello di significatività, $z_{\alpha/2}$ e $t_{\alpha/2,\,\nu}$ sono i valori critici delle distribuzioni normale e t di Student, rispettivamente, e ν è il numero di gradi di libertà, dato da $\nu = n_1 + n_2 - 2$.

8.4 Intervalli di confidenza approssimati per la media di una Bernoulliana

Gli intervalli di confidenza per la media di una variabile aleatoria di Bernoulli possono essere approssimati utilizzando la distribuzione normale.

Supponiamo di avere un campione di dimensione n da una popolazione di Bernoulli con una probabilità di successo incognita p. Sia $\hat{p}$ la proporzione campionaria di successi. Allora, la media campionaria $\overline{X}$ ha una distribuzione normale approssimata con media $\mu_{\overline{X}} = p$ e varianza $\sigma^2_{\overline{X}} = \frac{p(1-p)}{n}$.

Per calcolare l'intervallo di confidenza per la media della popolazione, possiamo utilizzare la stessa formula dell'intervallo di confidenza per la media di una normale con varianza incognita, sostituendo la deviazione standard campionaria con la deviazione standard campionaria approssimata:

$$\overline{X} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})/n}}{\sqrt{n}}$$

dove $z_{\alpha/2}$ è il valore critico della distribuzione normale standard per un livello di confidenza del $(1-\alpha)\%$. Notare che l'approssimazione con la distribuzione normale è valida solo se $np \geq 5$ e $n(1-p) \geq 5$.

8.5 Intervalli di confidenza per la media di una distribuzione esponenziale

Gli intervalli di confidenza per la media di una distribuzione esponenziale dipendono dalla specifica distribuzione dei dati e dal livello di confidenza desiderato. Tuttavia, quando il numero di campioni è grande, si può utilizzare l'intervallo di confidenza approssimato della media:

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

dove $\bar{X}$ è la media campionaria, s è la deviazione standard campionaria, n è il numero di campioni e $z_{\alpha/2}$ è il valore critico corrispondente al livello di confidenza desiderato α.

Nel caso della distribuzione esponenziale, la media è $\frac{1}{\lambda}$ e la varianza è $\frac{1}{\lambda^2}$. Pertanto, la deviazione standard è $\frac{1}{\lambda}\sqrt{n}$ e l'intervallo di confidenza approssimato diventa:

$$\frac{1}{\bar{X}} \pm z_{\alpha/2} \frac{1}{\bar{X}\sqrt{n}}$$

dove $\bar{X}$ è la media campionaria e n è il numero di campioni. Questo intervallo di confidenza approssimato può essere utilizzato solo se il numero di campioni è abbastanza grande (solitamente $n > 30$). Se il numero di campioni è piccolo, si deve utilizzare una distribuzione t-student invece di una distribuzione normale per calcolare il valore critico.

8.6 Valutare l'efficienza degli stimatori puntuali

Per valutare la correttezza di uno stimatore mediante il bias, è necessario calcolare la differenza tra il valore atteso dello stimatore e il valore del parametro che si sta stimando. Se tale differenza è uguale a zero, lo stimatore è corretto.

Formalmente, il bias di uno stimatore $\hat{\theta}$ del parametro θ è definito come:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

dove $E(\hat{\theta})$ rappresenta il valore atteso dello stimatore $\hat{\theta}$ e θ rappresenta il valore vero del parametro.

Se $\text{Bias}(\hat{\theta}) = 0$, lo stimatore è detto non distorto o corretto. Se invece $\text{Bias}(\hat{\theta}) \neq 0$, lo stimatore è distorto e l'entità del bias indica la direzione e l'entità dello scostamento dall'effettivo valore del parametro.

Inoltre, se $\text{Bias}(\hat{\theta}) \geq 0$ per ogni possibile valore del parametro θ, lo stimatore è detto non negativamente distorto.

In sintesi, il bias rappresenta la differenza tra il valore atteso dello stimatore e il valore vero del parametro. Se tale differenza è uguale a zero, lo stimatore è corretto.

8.7 Stimatori Bayesiani

Gli stimatori bayesiani sono un tipo di stimatore utilizzato nell'ambito della statistica bayesiana. In questo approccio, si parte da una distribuzione di probabilità a priori per il parametro che si vuole stimare, detta distribuzione a priori. Successivamente, si utilizza la legge di Bayes per ottenere la distribuzione a posteriori, ovvero la distribuzione di probabilità del parametro dato il campione osservato. Infine, la stima del parametro viene ottenuta come valore atteso della distribuzione a posteriori.

Più in dettaglio, se θ rappresenta il parametro da stimare e $X_1, \ldots, X_n$ rappresentano i dati osservati, la distribuzione a posteriori è data da:

$$p(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)p(\theta)}{p(X_1, \ldots, X_n)}$$

dove $p(X_1, \ldots, X_n|\theta)$ rappresenta la verosimiglianza dei dati, $p(\theta)$ rappresenta la distribuzione a priori del parametro e $p(X_1, \ldots, X_n)$ rappresenta la probabilità marginale dei dati.

La stima bayesiana del parametro è quindi data da:

$$\hat{\theta}_{\text{Bayes}} = \int \theta p(\theta|X_1, \ldots, X_n) d\theta$$

ovvero il valore atteso della distribuzione a posteriori.

Gli stimatori bayesiani presentano alcune caratteristiche interessanti rispetto agli stimatori classici. Ad esempio, permettono di incorporare conoscenze a priori sul parametro da stimare e di tenere conto della variabilità del parametro stesso. Inoltre, consentono di ottenere stime puntuali e intervalli di credibilità, che rappresentano una generalizzazione degli intervalli di confidenza classici. Tuttavia, la loro applicazione richiede un certo livello di competenza nella specifica della distribuzione a priori e la scelta di una distribuzione a priori inadeguata può portare a stime poco accurate.

Capitolo 9

Verifica delle ipotesi

La verifica delle ipotesi (o test di ipotesi) è un processo statistico che consiste nel valutare se i dati osservati sono compatibili con una determinata ipotesi sulla popolazione di riferimento.

Il processo di verifica delle ipotesi comporta l'individuazione di una statistica di prova (test statistic) che viene calcolata dai dati del campione. Questa statistica di prova viene confrontata con una distribuzione di probabilità nota, generalmente derivata dalla teoria statistica e comunemente chiamata distribuzione di riferimento o distribuzione campionaria sotto l'ipotesi nulla. L'ipotesi nulla è l'ipotesi che vogliamo testare, mentre l'ipotesi alternativa è l'ipotesi che vogliamo accettare se l'ipotesi nulla viene rifiutata.

In base al confronto tra la statistica di prova e la distribuzione di riferimento, si calcola il valore p (p-value) che rappresenta la probabilità di ottenere una statistica di prova uguale o più estrema di quella osservata, sotto l'ipotesi nulla. Se il valore p è inferiore a una soglia di significatività prestabilita (tipicamente 0.05 o 0.01), allora si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa. Al contrario, se il valore p è maggiore della soglia di significatività, si accetta l'ipotesi nulla.

Va notato che il valore p non rappresenta la probabilità che l'ipotesi nulla sia vera o falsa, ma rappresenta solo la probabilità di ottenere un'osservazione almeno altrettanto estrema di quella osservata, supponendo che l'ipotesi nulla sia vera. In altre parole, il valore p ci indica quanto è "strana" o "rara" l'osservazione rispetto all'ipotesi nulla.

La verifica delle ipotesi può essere un'operazione complessa e va effettuata con attenzione, evitando di trarre conclusioni affrettate o erronee. È importante specificare chiaramente l'ipotesi nulla e l'ipotesi alternativa, scegliere una statistica di prova appropriata e una distribuzione di riferimento adeguata, e definire una soglia di significatività congrua con il contesto e con il rischio di commettere errori di tipo I (rifiutare erroneamente l'ipotesi nulla) e di tipo II (accettare erroneamente l'ipotesi nulla).

9.1 Livelli di significatività

I livelli di significatività sono un parametro importante nella verifica delle ipotesi. In particolare, rappresentano la probabilità massima di commettere un errore di tipo I (ovvero respingere l'ipotesi nulla quando in realtà è vera) durante il processo di verifica dell'ipotesi.

Solitamente, si sceglie un livello di significatività α (spesso 0.05 o 0.01) prima di eseguire il test di ipotesi. Questo significa che, se il valore p (probabilità di ottenere il risultato osservato o uno più estremo assumendo che l'ipotesi nulla sia vera) è inferiore a α, allora l'ipotesi nulla viene respinta. In altre parole, se il valore p è molto basso (inferiore al livello di significatività scelto), allora è improbabile che i dati siano stati generati dall'ipotesi nulla e si conclude che questa è falsa.

Tuttavia, è importante notare che la scelta del livello di significatività è soggettiva e può influenzare la decisione finale. Inoltre, respingere l'ipotesi nulla non significa necessariamente che l'ipotesi alternativa sia vera, ma solo che ci sono evidenze statistiche sufficienti per sospettare che l'ipotesi nulla non sia vera.

9.1.1 Tipologie di errori

Gli errori che si possono commettere durante la verifica di un'ipotesi sono di due tipi:

- Errore di tipo I: si rifiuta un'ipotesi vera (falso positivo). Il livello di significatività α corrisponde alla probabilità di commettere un errore di tipo I.
- Errore di tipo II: si accetta un'ipotesi falsa (falso negativo). La potenza del test $(1 - \beta)$ corrisponde alla probabilità di non commettere un errore di tipo II.

Si noti che α e β sono in genere valori fissati a priori, mentre la potenza del test dipende dal valore dell'effetto da rilevare e dal campione utilizzato per la verifica dell'ipotesi.

9.2 La verifica delle ipotesi sulla media di una popolazione normale

9.2.1 Varianza nota

Supponiamo di avere un campione casuale di dimensione n estratto da una popolazione normale con media μ e varianza σ^2, nota. Vogliamo testare l'ipotesi che la media della popolazione sia uguale a un certo valore prefissato μ_0, ovvero:

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

Calcoliamo la statistica del test:

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

dove $\overline{X}$ è la media campionaria del campione estratto. La statistica z segue una distribuzione normale standard.

Calcoliamo il valore p del test, ovvero la probabilità di ottenere una statistica del test almeno tanto estrema quanto quella osservata, sotto l'ipotesi nulla H_0:

$$p = 2(1 - \Phi(|z|))$$

dove Φ è la funzione di distribuzione cumulativa standard della distribuzione normale.

Confrontiamo il valore p con il livello di significatività α scelto. Se il valore p è inferiore al livello di significatività scelto, ovvero se $p < \alpha$, allora rigettiamo l'ipotesi nulla H_0 e accettiamo l'ipotesi alternativa H_1. Altrimenti, non rigettiamo l'ipotesi nulla H_0.

Notare che se il livello di significatività scelto è del 5%, allora rigettiamo H_0 se il valore p è inferiore a 0.05.

Inoltre, se il valore di z è positivo, allora la media campionaria $\overline{X}$ è maggiore di μ_0, mentre se è negativo, allora $\overline{X}$ è minore di μ_0. Se $|z|$ è grande, allora il valore di $\overline{X}$ è molto lontano da μ_0, il che indica che l'ipotesi nulla è improbabile.

È importante notare che il test di ipotesi z richiede la conoscenza della varianza della popolazione, che spesso non è nota e deve essere stimata dalla varianza campionaria s^2. In tal caso, si utilizza il test di ipotesi t invece di quello di ipotesi z.

9.2.2 I test unilaterali

I test unilaterali sono una tipologia di test statistici utilizzati per verificare un'ipotesi riguardo alla direzione di un effetto o di un cambiamento in una variabile. A differenza dei test bilaterali, che verificano se un effetto è presente in entrambe le direzioni, i test unilaterali si concentrano solo su una delle due direzioni.

Ad esempio, supponiamo di avere un'ipotesi che afferma che un nuovo farmaco aumenta la pressione sanguigna. Se volessimo utilizzare un test bilaterale, dovremmo verificare se il nuovo farmaco aumenta o diminuisce la pressione sanguigna. Invece, se volessimo utilizzare un test unilaterale, dovremmo verificare solo se il nuovo farmaco aumenta la pressione sanguigna.

I test unilaterali possono essere di due tipi: test unilaterali destri e test unilaterali sinistri. Un test unilaterale destro viene utilizzato quando si vuole verificare se il valore di una variabile è maggiore di un certo valore, mentre un test unilaterale sinistro viene utilizzato quando si vuole verificare se il valore di una variabile è inferiore a un certo valore.

In generale, il processo di verifica delle ipotesi con un test unilaterale prevede la definizione di un livello di significatività (generalmente 0,05 o 0,01), la scelta del tipo di test (unilaterale destro o sinistro), il calcolo della statistica del test (ad esempio il t-score o lo z-score) e la comparazione della statistica del test con il valore critico corrispondente al livello di significatività e al grado di libertà del test.

Se la statistica del test è maggiore (per il test unilaterale destro) o minore (per il test unilaterale sinistro) del valore critico, l'ipotesi nulla viene rigettata e si conclude che vi è evidenza statistica a favore dell'ipotesi alternativa. Al contrario, se la statistica del test è minore (per il test unilaterale destro) o maggiore (per il test unilaterale sinistro) del valore critico, l'ipotesi nulla viene accettata e si conclude che non vi è evidenza statistica a favore dell'ipotesi alternativa.

9.2.3 Quando la varianza non è nota

Ecco un esempio di verifica delle ipotesi sulla media di una popolazione normale nel caso in cui la varianza non è nota in LaTeX:

Sia $X_1, X_2, ..., X_n$ un campione casuale di dimensione n da una popolazione normale con media μ e varianza σ^2 sconosciuta. Supponiamo di voler verificare l'ipotesi nulla $H_0 : \mu = \mu_0$ contro l'ipotesi alternativa $H_1 : \mu \neq \mu_0$ al livello di significatività α.

Il test statistic è dato da:

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

dove $\overline{X}$ è la media campionaria e S è la deviazione standard campionaria.

Sotto l'ipotesi nulla H_0, la statistica del test segue una distribuzione t di Student con $n-1$ gradi di libertà. Possiamo quindi calcolare il valore critico $t_{\alpha/2,n-1}$ tale che:

$$P(t_{n-1} < t_{\alpha/2,n-1}) = \frac{\alpha}{2}$$

Il valore critico a due code è quindi $-t_{\alpha/2,n-1}$ e $t_{\alpha/2,n-1}$.

Se t cade nella regione critica, cioè se $t < -t_{\alpha/2,n-1}$ o $t > t_{\alpha/2,n-1}$, allora si rifiuta l'ipotesi nulla H_0 a favore dell'ipotesi alternativa H_1 al livello di significatività α. In caso contrario, si accetta l'ipotesi nulla H_0.

Quindi, la regola di decisione per la verifica delle ipotesi sulla media di una popolazione normale nel caso in cui la varianza non è nota al livello di significatività α è la seguente:

- Rifiutiamo H_0 se $t < -t_{\alpha/2,n-1}$ o $t > t_{\alpha/2,n-1}$.
- Accettiamo H_0 se $-t_{\alpha/2,n-1} \leq t \leq t_{\alpha/2,n-1}$.

9.3 Verifica se due popolazioni normali hanno la stessa media

9.3.1 Varianze note

La verifica se due popolazioni normali hanno la stessa media con le varianze note può essere effettuata tramite il test Z di Student. L'ipotesi nulla è che le due popolazioni abbiano la stessa media ($H_0 : \mu_1 = \mu_2$), mentre l'ipotesi alternativa è che abbiano medie diverse ($H_1 : \mu_1 \neq \mu_2$).

La statistica del test è data da:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dove $\bar{X}_1$ e $\bar{X}_2$ sono le medie campionarie delle due popolazioni, σ_1 e σ_2 sono le rispettive deviazioni standard note, n_1 e n_2 sono le dimensioni campionarie delle due popolazioni e μ_1 e μ_2 sono le medie delle due popolazioni.

Sotto l'ipotesi nulla, la statistica del test segue una distribuzione normale standard $Z \sim N(0,1)$. Si può quindi calcolare il valore di Z e confrontarlo con il valore critico ottenuto dalla distribuzione normale standard per il livello di significatività scelto.

Ad esempio, per un test bilaterale con livello di significatività $\alpha = 0.05$, il valore critico è dato da $z_{\alpha/2} = 1.96$ (perché la distribuzione normale standard è simmetrica intorno allo zero). Se il valore di Z calcolato è maggiore di 1.96 o minore di -1.96, allora si può rifiutare l'ipotesi nulla a favore dell'ipotesi alternativa.

9.3.2 Varianze non note ma supposte uguali

Ecco un esempio di verifica delle ipotesi per testare se due popolazioni normali hanno la stessa media, supponendo che le varianze siano incognite ma uguali, in LaTeX:

Dati due campioni $X_1, X_2, ..., X_n$ e $Y_1, Y_2, ..., Y_m$ estratti da due popolazioni normali con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2 sconosciute ma uguali, si considera la seguente ipotesi nulla $H_0 : \mu_1 = \mu_2$ e l'ipotesi alternativa $H_1 : \mu_1 \neq \mu_2$. Si definisce la statistica di test Z come:

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

dove $\overline{X}$ e $\overline{Y}$ sono le medie campionarie dei due campioni, S_1^2 e S_2^2 sono le varianze campionarie non corrette e n e m sono le dimensioni dei due campioni. Si assume che Z segua una distribuzione normale standard sotto l'ipotesi nulla.

A questo punto, si calcola il valore critico di Z corrispondente al livello di significatività $\alpha/2$ (dove α è il livello di significatività scelto) utilizzando la tabella della distribuzione normale standard o un software statistico. L'intervallo di accettazione sarà quindi dato da $-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$.

Infine, si calcola il valore della statistica di test Z a partire dai dati del campione e si confronta con il valore critico calcolato precedentemente. Se il valore di Z rientra nell'intervallo di accettazione, si accetta l'ipotesi nulla H_0 al livello di significatività scelto; altrimenti, si rigetta l'ipotesi nulla e si accetta l'ipotesi alternativa H_1.

9.3.3 Varianze non note e diverse

La verifica delle ipotesi per confrontare le medie di due popolazioni normali con varianze incognite e diverse può essere effettuata utilizzando il test t di Student.

L'ipotesi nulla è che le medie delle due popolazioni siano uguali, mentre l'ipotesi alternativa è che le medie siano diverse.

La statistica del test t è data da:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

dove $\bar{x}_1$ e $\bar{x}_2$ sono le medie campionarie delle due popolazioni, n_1 e n_2 sono le dimensioni dei campioni, s_p è la stima comune della deviazione standard delle due popolazioni, data da:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

dove s_1^2 e s_2^2 sono le varianze campionarie delle due popolazioni.

La statistica t segue una distribuzione t di Student con $n_1 + n_2 - 2$ gradi di libertà. L'intervallo di confidenza bilaterale per la differenza tra le medie delle due popolazioni con un livello di confidenza del 95

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

dove $t_{\alpha/2,n_1+n_2-2}$ è il valore critico della distribuzione t di Student con $\alpha/2$ di probabilità di coda a sinistra e $n_1 + n_2 - 2$ gradi di libertà.

Per accettare o rifiutare l'ipotesi nulla, confrontiamo il valore della statistica t con il valore critico della distribuzione t di Student. Se il valore assoluto della statistica t è maggiore del valore critico, allora rifiutiamo l'ipotesi nulla e concludiamo che le medie delle due popolazioni sono significativamente diverse. Altrimenti, accettiamo l'ipotesi nulla e concludiamo che non c'è sufficiente evidenza statistica per concludere che le medie delle due popolazioni siano diverse.

9.4 Verifica di ipotesi sulla varianza di una popolazione normale

Per verificare l'ipotesi H_0 sulla varianza di una popolazione normale, è possibile utilizzare la distribuzione del test chi-quadro.

L'idea è di confrontare la somiglianza tra la varianza del campione e la varianza nota della popolazione σ_0 ipotizzata nell'ipotesi nulla.

L'ipotesi nulla afferma che la varianza del campione è uguale alla varianza nota della popolazione σ_0, mentre l'ipotesi alternativa afferma che la varianza del campione è diversa dalla varianza nota della popolazione σ_0.

Per calcolare il test chi-quadro, occorre calcolare la statistica del test, che è data dalla formula:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

dove n è la dimensione del campione, s^2 è la varianza campionaria, e σ_0^2 è la varianza nota della popolazione.

La statistica del test segue una distribuzione chi-quadro con $n-1$ gradi di libertà. L'intervallo di accettazione dell'ipotesi nulla dipende dal livello di significatività del test e dal numero di gradi di libertà.

In generale, se la statistica del test χ^2 è maggiore del valore critico della distribuzione chi-quadro con $n-1$ gradi di libertà al livello di significatività scelto, l'ipotesi nulla viene rigettata a favore dell'ipotesi alternativa. Altrimenti, l'ipotesi nulla viene accettata.

In sintesi, il test chi-quadro permette di verificare se la varianza del campione è uguale alla varianza nota della popolazione, e di determinare se ci sono prove sufficienti per rigettare l'ipotesi nulla a favore dell'ipotesi alternativa.

Accetto l'ipotesi H_0 se:

$$\chi^2_{1-\frac{\alpha}{2},n-1} \leq (n-1)\frac{S^2}{\sigma_0^2} \leq \chi^2_{\frac{\alpha}{2},n-1} \tag{9.1}$$

9.4.1 Verifica se hanno la stessa varianza - media e varianza incognite DA FIXARE

Per verificare se due popolazioni hanno la stessa varianza, con media incognite ma varianza incognita ma uguale, si può utilizzare un test statistico basato sulla distribuzione di Fisher-Snedecor.

L'ipotesi nulla (H0) è che le due popolazioni abbiano la stessa varianza, mentre l'ipotesi alternativa (H1) è che le varianze siano diverse.

Il test di verifica si basa sulla costruzione della statistica di test data da:
$F = \frac{s_1^2}{s_2^2}$
dove s_1^2 e s_2^2 sono le varianze campionarie delle due popolazioni.

La statistica di test F segue una distribuzione di Fisher-Snedecor con $n_1 - 1$ gradi di libertà al numeratore e $n_2 - 1$ gradi di libertà al denominatore, dove n_1 e n_2 sono le dimensioni dei due campioni.

A questo punto si può calcolare il valore p associato alla statistica di test F, e confrontarlo con il livello di significatività α scelto a priori. Se il valore p è maggiore di α, non si rifiuta l'ipotesi nulla H0, altrimenti si rifiuta e si accetta l'ipotesi alternativa H1.

Il test può essere a una coda (se l'ipotesi alternativa prevede che una varianza sia maggiore dell'altra) o a due code (se l'ipotesi alternativa prevede che le varianze siano diverse).

In alternativa, è possibile utilizzare un intervallo di confidenza per la differenza delle varianze campionarie, che consente di stabilire se le due varianze sono significativamente diverse a un livello di confidenza specificato.

Accetto l'ipotesi H_0 se:

$$F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq F_{\frac{\alpha}{2},n-1,m-1} \tag{9.2}$$

9.5 La verifica di ipotesi su ua popolazione di Bernoulli

Supponiamo di avere una popolazione di Bernoulli con parametro p, dove vogliamo verificare l'ipotesi $H_0 : p = p_0$ contro l'ipotesi alternativa $H_1 : p \neq p_0$, con un livello di significatività α.

Sia X la variabile casuale che rappresenta il numero di successi in n prove indipendenti con probabilità di successo p. La media campionaria $\hat{p} = X/n$ segue approssimativamente una distribuzione normale con media p e deviazione standard $\sqrt{\frac{p(1-p)}{n}}$.

Definiamo la statistica di test Z come:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Se H_0 è vera, allora Z segue una distribuzione normale standard. Possiamo quindi calcolare il valore p come:

$$p = P(|Z| > |z_{\alpha/2}|)$$

dove $z_{\alpha/2}$ è il valore critico corrispondente al livello di significatività α. Se $p \leq \alpha$, allora rifiutiamo H_0; altrimenti, non abbiamo sufficienti evidenze per rifiutare H_0.

9.5.1 Verificare se due popolazioni di Bernoulli hanno lo stesso parametro

Per verificare se due popolazioni di Bernoulli hanno lo stesso parametro, possiamo utilizzare il test di ipotesi per la differenza di proporzioni. Supponiamo di avere due campioni indipendenti, con n_1 e n_2 osservazioni rispettivamente, e vogliamo testare l'ipotesi nulla $H_0 : p_1 = p_2$ contro l'ipotesi alternativa $H_1 : p_1 \neq p_2$, dove p_1 e p_2 sono le proporzioni delle due popolazioni.

Il test si basa sulla statistica del test Z data da:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

dove $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ è la stima della proporzione combinata delle due popolazioni, e $\hat{p}_1$ e $\hat{p}_2$ sono le stime delle proporzioni nei due campioni.

La statistica del test Z segue approssimativamente la distribuzione normale standard se H_0 è vera. Quindi, possiamo calcolare il valore p come l'area della coda della distribuzione normale standard a destra del valore assoluto di Z:

$$p = 2\Phi(-|Z|)$$

dove Φ è la funzione di distribuzione cumulativa della distribuzione normale standard. Se p è minore di un livello di significatività scelto α, possiamo rigettare l'ipotesi nulla H_0 e concludere che ci sono prove sufficienti

per affermare che le proporzioni delle due popolazioni sono diverse. Altrimenti, non abbiamo prove sufficienti per rigettare l'ipotesi nulla e dobbiamo accettarla.

È importante notare che il test di ipotesi per la differenza di proporzioni assume che le due popolazioni siano indipendenti e che i successi e i fallimenti siano distribuiti come una distribuzione di Bernoulli. Inoltre, la statistica del test Z può essere utilizzata solo quando il conteggio atteso dei successi e dei fallimenti è maggiore di 5 in entrambi i campioni.

9.6 La verifica di ipotesi sulla media di una distribuzione di Poisson

La verifica di ipotesi sulla media di una distribuzione di Poisson viene effettuata quando vogliamo stabilire se la media λ di una popolazione di distribuzione di Poisson segue un certo valore λ_0 o meno.

Le ipotesi sono: - H_0: la media della popolazione di distribuzione di Poisson è λ_0 - H_1: la media della popolazione di distribuzione di Poisson non è λ_0

La statistica del test è data da:

$$Z = \frac{\bar{X} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}}$$

dove $\bar{X}$ è la media campionaria, n è la dimensione del campione e λ_0 è il valore che vogliamo testare.

Assumendo H_0, la statistica del test segue una distribuzione normale standard.

La regione di accettazione e la regione critica dipendono dal livello di significatività α del test e dal tipo di test (a una o due code). In generale, la regione critica si trova in corrispondenza dei valori di Z tali che la probabilità cumulativa della distribuzione normale standard è inferiore a α.

Se il valore di Z calcolato dal campione rientra nella regione di accettazione, si accetta H_0. Altrimenti, si rifiuta H_0 e si accetta H_1.

Capitolo 10

Regressione

La regressione è una tecnica statistica utilizzata per studiare la relazione tra una variabile dipendente e una o più variabili indipendenti. La regressione lineare è il tipo più comune di regressione, in cui si cerca di modellare la relazione tra una variabile dipendente continua e una o più variabili indipendenti tramite una funzione lineare.

L'obiettivo della regressione è di trovare il modello di regressione che fornisce la miglior previsione della variabile dipendente data la conoscenza delle variabili indipendenti. Per fare ciò, si cerca di minimizzare la differenza tra i valori osservati della variabile dipendente e i valori previsti dal modello di regressione.

La regressione può essere utilizzata per scopi diversi, ad esempio per analizzare i dati di un esperimento scientifico, per fare previsioni di vendita, per identificare i fattori che influenzano il prezzo delle case, ecc.

In generale, la regressione può essere eseguita in due fasi principali: la fase di addestramento, in cui viene costruito il modello di regressione, e la fase di test, in cui il modello viene utilizzato per fare previsioni su nuovi dati.

La regressione lineare si basa sull'equazione:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

dove y è la variabile dipendente, $x_1, x_2, ..., x_p$ sono le variabili indipendenti, $\beta_0, \beta_1, \beta_2, ..., \beta_p$ sono i coefficienti del modello e ϵ è l'errore casuale.

La stima dei coefficienti può essere effettuata utilizzando il metodo dei minimi quadrati, che cerca di minimizzare la somma dei quadrati degli errori tra i valori osservati e quelli previsti dal modello.

Esistono diversi tipi di regressione, come la regressione logistica per le variabili binarie o la regressione polinomiale per le relazioni non lineari. Inoltre, esistono anche tecniche di regressione non parametriche, come la regressione spline, che non richiedono l'assunzione di una forma funzionale specifica per il modello di regressione.

10.1 Stima dei parametri di regressione

La stima dei parametri di regressione viene effettuata attraverso il metodo dei minimi quadrati. Supponiamo di avere un modello di regressione lineare semplice:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

dove Y_i è la variabile dipendente, X_i è la variabile indipendente, ϵ_i è l'errore casuale, β_0 e β_1 sono i parametri da stimare. Il metodo dei minimi quadrati consiste nel trovare i valori di β_0 e β_1 che minimizzano la somma dei quadrati degli scarti tra i valori osservati di Y_i e quelli predetti dal modello:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

La soluzione a questo problema può essere trovata attraverso le derivate parziali e l'uguaglianza a zero:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

dove $\bar{X}$ e $\bar{Y}$ sono le medie campionarie di X e Y. La stima della varianza degli errori può essere calcolata come:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n - 2}.$$

Questo valore è utilizzato per calcolare gli errori standard delle stime dei parametri:

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

$$\text{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

Queste quantità possono essere utilizzate per costruire gli intervalli di confidenza e test di ipotesi sui parametri di regressione.