Statistica

Esercitazione 12

Alfonso Iodice D'Enza

iodicede@unicas.it

Università degli studi di Cassino

Outline

1. disuguaglianza di Markov

Outline

Outline

Outline

Outline

Importanti disuguaglianze

Variabili casuali con distribuzioni non note

Le disuguaglianze di Markov e Chebyshev sono due risultati importanti perchè consentono di porre una 'soglia' superiore alle probabilità di eventi rari che riguardano variabili casuale di cui non si conosce la distribuzione, ma solo valore atteso oppure valore atteso e varianza.

disuguaglianza di Markov

Sia X una variabile casuale mai negativa, allora per qualunque valore $a > 0$

$$P(X \geq a) \leq \frac{E[X]}{a}$$

disuguaglianza di Chebyshev

Sia X una variabile casuale di cui si conoscono solo la media μ e varianza σ^2, allora dato un qualunque valore $k > 0$, vale la seguente relazione

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

Importanti disuguaglianze

disuguaglianza di Markov

Sia X una variabile casuale mai negativa, allora per qualunque valore $a > 0$

$$P(X \geq a) \leq \frac{E[X]}{a}$$

dimostrazione

Si supponga che la v.c. X si distribuisca secondo una funzione di densità incognita f

$$E[X] = \int_0^\infty x f(x)dx = \underbrace{\int_0^a x f(x)dx}_{\geq 0} + \underbrace{\int_a^\infty x f(x)dx}_{\geq 0} \geq$$

$$\geq \underbrace{\int_a^\infty x f(x)dx \geq \int_a^\infty a f(x)dx}_{\text{perchè } x \in [a, \infty] \text{ quindi } x \geq a} = \underbrace{a \int_a^\infty f(x)dx}_{a \text{ è una costante}} = a\ P(X \geq a)$$

dunque

$$E[X] \geq aP(X \geq a) \rightarrow P(X \geq a) \leq \frac{E[X]}{a}$$

disuguaglianza di Chebyshev

Sia X una variabile casuale di cui si conoscono solo la media μ e varianza σ^2, allora dato un qualunque valore $k > 0$, vale la seguente relazione $P(\mid X - \mu \mid \geq k) \leq \frac{\sigma^2}{k^2}$

dimostrazione

Si consideri l'evento per cui vale $\mid X - \mu \mid \geq k$: i valori di X, μ e k per cui vale la disuguaglianza sono gli stessi per cui vale la disuguaglianza $(X - \mu)^2 \geq k^2$. La probabilità che si verifichi una delle disuguaglianze precedenti è dunque la stessa. Inoltre la variabile casuale $(X - \mu)^2$ è non negativa (essendo un quadrato), dunque si può applicare la disuguaglianza di Markov, con $a = k^2$, quindi

$$P(\mid X - \mu \mid \geq k) = \underbrace{P((X - \mu)^2 \geq k^2) \leq \frac{\overbrace{E\left[(X - \mu)^2\right]}^{\text{notare che } = \sigma^2}}{k^2}}_{\text{disuguaglianza di Markov}}$$

che verifica la disuguaglianza di Chebyshev

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

Esercizio

Il numero di automobili prodotte da una fabbrica in una settimana si distribuisce secondo una variabile casuale X con media pari a 50.

- Qual'è la probabilità che la produzione superi occasionalmente le 75 auto?
- Qual'è la probabilità che la produzione sia compresa tra 40 e 60 pezzi, sapendo che la varianza della distribuzione è pari a 25?

svolgimento

- Poichè l'unica conoscenza della distribuzione di X è che $E[X] = 50$, per calcolare $P(X \geq 50)$ si ricorre alla disuguaglianza di Markov.

$$P(X \geq a) \leq \frac{E[X]}{a} \rightarrow P(X \geq 75) \leq \frac{50}{75} = 0.67$$

- In questo caso, oltre alla media $\mu = 50$, è nota anche la varianza $\sigma^2 = 25$. Si vuole la probabilità che $40 \leq X \geq 60$, quindi $\mid 60 - 50 \mid = \mid 40 - 50 \mid = 10 = k$, dunque $P(\mid X - 50 \mid \geq 10)$ rappresenta la probabilità che la produzione si discosti di più di 10 unità dalla media.

$$P(\mid X - 50 \mid \geq 10) \leq \frac{25}{10^2} = 0.25$$

dunque la probabilità che la produzione si discosti di meno di 10 unità dalla media è $P(40 \leq X \geq 60)$.

Distribuzione delle statistiche campionarie

Popolazione e campione

La popolazione è un insieme molto grande di oggetti a cui sono associate delle quantità misurabili. Il campione è un sottoinsieme ridotto della popolazione. L'obiettivo dell'approccio statistico è analizzare il campione per trarre da esso informazioni circa la popolazione.

campione casuale

Per effettuare inferenze sulla popolazione in base al campione, si assume che vi sia la popolazione segua una distribuzione di probabilità F. Estraendo casualmente degli oggetti dalla popolazione per formare il campione, si assume che ciascun valore ad essi associato sia una variabile casuale caratterizzata dalla distribuzione F della popolazione.

definizione

Un insieme di $X_1, X_2, \ldots, X_n$ di variabili aleatorie indipendenti e distribuite secondo una distribuzione F, si definisce campione casuale della distribuzione F.

Distribuzione delle statistiche campionarie

Inferenza parametrica e non parametrica

La distribuzione F della popolazione è non nota. In alcuni casi è tuttavia possibile che si conosca la famiglia di distribuzioni di variabili casuali a cui F appartiene, e dunque si utilizza il campione per fare inferenza sui parametri che identificano F: è il caso dell'inferenza parametrica. In altri casi non si ha alcuna informazione su F: in questi casi si fa ricorso a tecniche di inferenza non parametrica.

La statistica

Uno stimatore è una funzione dei dati campionari. Poichè le osservazioni campionarie sono v.c., e poichè una funzione di v.c. è a sua volta una v.c., allora lo stimatore è una v.c. funzione dei dati campionari. Una statistica è uno stimatore che utilizza i dati campionari per ottenere la stima di un parametro della distribuzione F.

Distribuzione della media campionaria

Si consideri ad esempio una popolazione - ad esempio i lavoratori dipendenti - su cui sia misurata una quantità numerica - ad esempio il reddito annuo percepito -; il campione casuale estratto da tale popolazione è $X_1, X_2, \ldots, X_n$, i valori associati agli elementi del campione sono v.c. indipendenti e identicamente distribuite (i.i.d.), tutte caratterizzate dalla stessa distribuzione F i cui parametri sono μ e σ^2 (media e varianza).

La statistica media campionaria

$$\bar{X} := \frac{X_1 + X_2 + \ldots + X_n}{n}$$

funzione delle v.c. $X_1, X_2, \ldots, X_n$ del campione: si tratta di una variabile casuale.

Distribuzione della media campionaria

La statistica media campionaria

$$\bar{X} := \frac{X_1 + X_2 + \ldots + X_n}{n}$$

funzione delle v.c. $X_1, X_2, \ldots, X_n$ del campione: si tratta di una variabile casuale.

valore atteso di $\bar{X}$

$$E\left[\bar{X}\right] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] =$$
$$= \frac{E\left[X_1\right] + E\left[X_2\right] + \ldots + E\left[X_n\right]}{n} =$$
$$= \frac{\mu + \mu + \ldots + \mu}{n} = \frac{n\mu}{n} = \mu$$

varianza di $\bar{X}$

$$var\left(\bar{X}\right) = var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$
$$= var\left(\frac{X_1}{n}\right) + var\left(\frac{X_2}{n}\right) + \ldots + var\left(\frac{X_n}{n}\right) =$$
$$= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \ldots + \frac{\sigma^2}{n^2} = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}$$

nota

La distribuzione di $\bar{X}$ risulta quindi centrata su μ, mentre la sua varianza diminuisce all'aumentare di n.

Teorema del limite centrale

Teorema del limite centrale (TLC)

Tale teorema è un risultato molto importante della teoria della probabilità: esso afferma che la somma di un numero elevato di v.c. indipendenti si distribuisce approssimativamente secondo una normale.

TLC:

Se si considerano le v.c. $X_1, X_2, \ldots, X_n$ indipendenti e identicamente distribuite, tutte con media μ e varianza σ^2, allora

$$X_1 + X_2 + \ldots + X_n \quad \sim \quad N(n\mu, n\sigma^2)$$

Se alla somma in questione si sottrae la media $n\mu$ e si divide per lo scarto quadratico medio $\sqrt{n\sigma^2} = \sigma\sqrt{n}$ si ottiene la relazione precedente in versione standardizzata

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \quad \sim \quad N(0,1)$$

Esercizio 2

Una compagnia di assicurazione ha 25000 polizze attive. Ciascun assicurato percepisce un risarcimento annuo che rappresenta una v.c. che si distribuisce con media pari a 320 euro e scarto quadratico medio pari a 540 euro. Qual'è la probabilità che la compagnia paghi complessivamente 8300000 euro?

Svolgimento

Il risarcimento di ciascun cliente è X_i con $i = 1, \ldots, n$ ed $n = 25000$. La richiesta complessiva di risarcimento da parte di tutti i clienti è $X = \sum_{i=1}^{n} X_i$. Poichè X è la somma delle v.c. X_i che sono i.i.d., per il teorema del limite centrale risulta che X si distribuisce come una normale con media $n\mu = 25000 \times 320 = 8000000$. e scarto quadratico medio $\sigma\sqrt{25000} = 85381$. Si vuole dunque $P(X > 8300000)$. In unità standard, il valore in questione è

$$Z = \frac{X - n\mu}{\sigma\sqrt{n}} = \frac{8300000 - 8000000}{85381} = 3.51$$

dunque

$$P(X > 8300000) = P(Z > 3.51) \approx 0.$$

Esercizio 3

Il numero ideale di studenti di un corso del primo anno di università è 150. Il management didattico dell'università sa che, in base agli anni precedenti, solo il 30% degli iscritti frequenta effettivamente i corsi, dunque decide di accettare fino a 450 nuove iscrizioni. Qual'è la probabilità che il numero di studenti frequentanti sia superiore a 150?

Svolgimento

Si definisca la v.c. X come il numero di studenti che frequentano, ciascuno studente iscritto corrisponde ad una prova Bernulliana il cui esito può essere frequenta o non frequenta. X si distribuisce pertanto secondo una distribuzione binomiale di parametri $n = 450$ e la probabilità di successo (lo studente iscritto frequenta) è $p = 0.3$.

Per il teorema del limite centrale, poichè X è la somma di n v.c. Bernoulliane X_i, ciascuna con media $E[X_i] = p$ e varianza pari a $var(X_i) = p(1-p)$, allora X si distribuisce approssimativamente come una normale con media $\mu = np$ e varianza pari a $\sigma^2 = np(1-p)$. (Si tratta dell'approssimazione della binomiale alla normale).

La probabilità cercata è dunque $P(X > 150.5)$ (lo '+0.5' in aggiunta è dovuto alla correzione di continuità). Standardizzando il problema si ha

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{150.5 - 450 \times 0.3}{\sqrt{450 \times .3 \times .7}} = \frac{150.5 - 135}{9.72} = 1.59$$

da cui

$$P(X > 150.5) = P(Z > 1.59) \approx 0.06.$$

Varianza campionaria

Varianza campionaria

Dato un campione casuale $X_1, X_2, \ldots, X_n$ proveniente da una distribuzione con media μ e varianza σ^2. Sia $\bar{X}$ la media campionaria. La statistica varianza campionaria è

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

valore atteso della varianza campionaria

Ricordando la relazione per la quale $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$, dove $\bar{x} = \sum_{i=1}^{n} x_i/n$ dunque

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \quad \text{da cui}$$

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

andando ad effettuare il valore atteso di entrambi i lati dell'equazione si ha

$$(n-1)E\left[S^2\right] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE\left[\bar{X}^2\right]$$

Varianza campionaria

valore atteso della varianza campionaria (seconda parte)

$$(n-1)E\left[S^2\right] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE\left[\bar{X}^2\right] = nE\left[X_1^2\right] - nE\left[\bar{X}^2\right]$$

poichè per qualunque v.c. Y vale la relazione $E\left[Y^2\right] = var(Y) + E\left[Y\right]^2$ allora

$$(n-1)E\left[S^2\right] = nE\left[X_1^2\right] - nE\left[\bar{X}^2\right] = \underbrace{n\ var(X_1) + E\left[X_1\right]^2}_{nE\left[X_1^2\right]} - \underbrace{\left(n\ var(\bar{X}) + E\left[\bar{X}\right]^2\right)}_{nE\left[\bar{X}^2\right]}$$

poichè sappiamo che $E\left[X_1\right] = \mu$, $var(X_1) = \sigma^2$, $E\left[\bar{X}\right] = \mu$, $var(\bar{X}) = \frac{\sigma^2}{n}$, quindi

$$(n-1)E\left[S^2\right] = n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 = n\sigma^2 - \sigma^2 = \sigma^2(n-1) \Rightarrow E\left[S^2\right] = \sigma^2$$

il valore atteso della varianza campionaria è uguale alla varianza della popolazione
(nota: ecco perchè il denominatore di S^2 è $(n-1)$ e non n...)