

Statistica Inferenziale

Stefania Bartoletti

4 Giugno 2021

Statistica

All'inizio del corso abbiamo definito la statistica descrittiva, che descrive un insieme molto grande di oggetti, detto popolazione, a cui sono associate delle quantità misurabili. Abbiamo visto che in un'indagine si seleziona un sottoinsieme ridotto di oggetti, che viene detto campione.

Statistica

All'inizio del corso abbiamo definito la statistica descrittiva, che descrive un insieme molto grande di oggetti, detto popolazione, a cui sono associate delle quantità misurabili. Abbiamo visto che in un'indagine si seleziona un sottoinsieme ridotto di oggetti, che viene detto campione.

Oltre a descriverlo, la statistica ha lo scopo di analizzarlo per trarre da esso delle conclusioni valide per la popolazione nel suo insieme. In particolare, la **statistica inferenziale** è la scienza che si occupa di trarre conclusioni dai dati sperimentali.

Lo studio delle variabili aleatorie, le loro distribuzioni, le statistiche di sintesi e le loro proprietà hanno proprio lo scopo di modellare la distribuzione di probabilità della popolazione, in maniera tale che dato un campione estratto in maniera casuale, le quantità numeriche loro associate possono essere pensate come variabili aleatorie **indipendenti e identicamente distribuite (i.i.d.)**.

Lo studio delle variabili aleatorie, le loro distribuzioni, le statistiche di sintesi e le loro proprietà hanno proprio lo scopo di modellare la distribuzione di probabilità della popolazione, in maniera tale che dato un campione estratto in maniera casuale, le quantità numeriche loro associate possono essere pensate come variabili aleatorie **indipendenti e identicamente distribuite (i.i.d.)**.

Per esempio, per molti esperimenti e osservazioni che riguardano fenomeni naturali, come la misura di un dato, ci rendiamo conto che se ripetiamo l'esperimento di misura più volte anche nelle stesse condizioni ci ritroviamo ad avere risultati diversi, a causa di fattori aleatori o non controllabili.

Inferenza

In pratica la distribuzione del campione non è mai completamente nota e vogliamo usare i dati per fare dell'inferenza sulla distribuzione. A volte potrebbe essere noto il tipo di distribuzione ma non i suoi parametri, a volte potremmo sapere solo se è discreta o continua, a volte non sappiamo assolutamente nulla.

Inferenza

In pratica la distribuzione del campione non è mai completamente nota e vogliamo usare i dati per fare dell'inferenza sulla distribuzione. A volte potrebbe essere noto il tipo di distribuzione ma non i suoi parametri, a volte potremmo sapere solo se è discreta o continua, a volte non sappiamo assolutamente nulla.

Si parla di **inferenza parametrica** quando la distribuzione è nota a meno di un insieme di parametri e **inferenza non parametrica** quando non si sa nulla sulla distribuzione del campione.

Ripetizione dell'esperimento

Uno degli approcci per fare inferenza sulla distribuzione, è quello di ripetere l'esperimento più volte. Osservando una sequenza abbastanza lunga di dati, l'intuito ci dice che possiamo conoscere la distribuzione. Lo abbiamo fatto in diversi esercizi in R. Questa intuizione è dimostrabile attraverso **la legge dei grandi numeri**.

Il termine **statistica** indica anche una variabile aleatoria che è una funzione dei dati di un campione.

Il termine **statistica** indica anche una variabile aleatoria che è una funzione dei dati di un campione.

Per esempio, per misurare un dato, posso effettuare molteplici misure e calcolare la media, per *mediare* rispetto all'aleatorietà della misura, all'errore di misura.

Il termine **statistica** indica anche una variabile aleatoria che è una funzione dei dati di un campione.

Per esempio, per misurare un dato, posso effettuare molteplici misure e calcolare la media, per *mediare* rispetto all'aleatorietà della misura, all'errore di misura.

In questi casi, ogni misura si può considerare una realizzazione di una variabile aleatoria e dunque n misure sono una sequenza di variabili aleatorie indipendenti e identicamente distribuite, ovvero, si considera che seguano la stessa distribuzione.

La media campionaria

Dato un campione di n variabili aleatorie indipendenti e identicamente distribuite X_i , la media campionaria \bar{X}_n è una statistica in quanto

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

di fatto, si tratta di una funzione di variabili aleatorie che rappresentano il campione.

La media campionaria

Siano X_i variabili aleatorie con media μ e varianza σ^2 . Qual è il valore atteso della variabile aleatoria rappresentata dalla media campionaria?

$$\frac{1}{n} \sum_i X_i$$

La media campionaria

Siano X_i variabili aleatorie con media μ e varianza σ^2 . Qual è il valore atteso della variabile aleatoria rappresentata dalla media campionaria?

$$\mathbb{E} \{ \bar{X}_n \} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \{ X_i \} = \mu$$

La media campionaria

Siano X_i variabili aleatorie con media μ e varianza σ^2 . Qual è il valore atteso della variabile aleatoria rappresentata dalla media campionaria?

$$\mathbb{E} \{ \bar{X}_n \} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \{ X_i \} = \mu$$

Intuitivamente, al crescere di n , la probabilità che \bar{X}_n si avvicini a μ cresce.

Esempio

- ▶ Siano $X_i \sim \text{Ber}(p)$ con $\mu = p = 0.5$ e $i = 1, 2, \dots, n$

Esempio

- ▶ Siano $X_i \sim \text{Ber}(p)$ con $\mu = p = 0.5$ e $i = 1, 2, \dots, n$
- ▶ La media campionaria rappresenta la proporzione del numero di teste su n lanci.

Esempio

- ▶ Siano $X_i \sim \text{Ber}(p)$ con $\mu = p = 0.5$ e $i = 1, 2, \dots, n$
- ▶ La media campionaria rappresenta la proporzione del numero di teste su n lanci.
- ▶ Intuitivamente, al crescere di n , la probabilità che \bar{X}_n si avvicini a μ cresce.

Esempio

- ▶ Siano $X_i \sim \text{Ber}(p)$ con $\mu = p = 0.5$ e $i = 1, 2, \dots, n$
- ▶ La media campionaria rappresenta la proporzione del numero di teste su n lanci.
- ▶ Intuitivamente, al crescere di n , la probabilità che \bar{X}_n si avvicini a μ cresce.
- ▶ $\mathbb{P} \{ |\bar{X}_n - \mu| < 0.1 \}$ al variare di n

Esempio

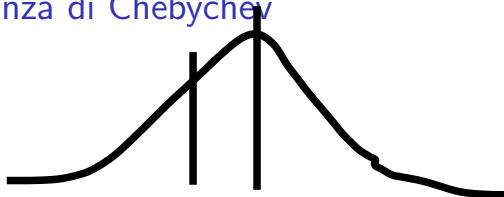
- ▶ Siano $X_i \sim \text{Ber}(p)$ con $\mu = p = 0.5$ e $i = 1, 2, \dots, n$
- ▶ La media campionaria rappresenta la proporzione del numero di teste su n lanci.
- ▶ Intuitivamente, al crescere di n , la probabilità che \bar{X}_n si avvicini a μ cresce.
- ▶ $\mathbb{P} \{ |\bar{X}_n - \mu| < 0.1 \}$ al variare di n
- ▶ Proviamo su R...

Diseguaglianza di Chebychev

Per una variabile aleatoria arbitraria Y e qualsiasi $a > 0$

$$\mathbb{P}\{|Y - \mathbb{E}\{Y\}| \geq a\} \leq \frac{1}{a^2} \mathbb{V}\{Y\}$$

Diseguaglianza di Chebychev



Se denotiamo $\sigma^2 = \mathbb{V}\{Y\}$ e poniamo $a = k\sigma$

$$\mathbb{P}\{|Y - \mu| \geq k\sigma\} \leq \frac{1}{k^2\sigma^2}\sigma^2$$

$$\mathbb{P}\{|Y - \mu| < k\sigma\} \geq 1 - \frac{1}{k^2}$$

Legge dei grandi numeri

Siano X_i variabili aleatorie con media μ e varianza σ^2 . Ricordiamo che $\mathbb{E}\{\bar{X}_n\} = \mu$ e $\mathbb{V}\{\bar{X}_n\} = \frac{\sigma^2}{n}$. Se applichiamo la disuguaglianza di Chebychev alla media campionaria, con $\epsilon > 0$ otteniamo

$$\mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} \leq \frac{1}{\epsilon^2} \frac{\sigma^2}{n} = \frac{\sigma^2}{n\epsilon^2}$$

Pertanto,

$$\lim_{n \rightarrow \infty} \mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} = 0$$