Statistica Inferenziale

Stefania Bartoletti

10 Maggio 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 $\overline{\mathsf{X}}_n$ è una statistica in quanto

$$\overline{\mathsf{X}}_n = \frac{1}{n}\sum_{j=1}^{n}\mathsf{X}_j$$

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?

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?

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

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?

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

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

Esempio

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

Esempio

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

Esempio

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

Esempio

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

Diseguaglianza di Chebychev

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

$$\mathbb{P}\{|\mathsf{Y} - \mathbb{E}\{\mathsf{Y}\}| \geqslant a\} \leqslant \frac{1}{a^2}\mathbb{V}\{\mathsf{Y}\}$$

Diseguaglianza di Chebychev

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

$$\mathbb{P}\{|\mathsf{Y} - \mu| \geqslant k\sigma\} \leqslant \frac{1}{k^2\sigma^2}\sigma^2$$

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

Legge dei grandi numeri

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

$$\mathbb{P}\left\{|\overline{X}_n - \mu| \geqslant \epsilon\right\} \leqslant \frac{1}{\epsilon^2}\frac{\sigma^2}{n} = \frac{\sigma^2}{n\epsilon^2}$$

Pertanto,

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{X}_n - \mu| \geqslant \epsilon\right\} = 0$$

Legge dei grandi numeri e densità di probabilità

Supponiamo che anzichè essere interessati alla media, vogliamo calcolare la probabilità di un evento, ad esempio $p = \mathbb{P}\{\mathsf{X} \in \mathcal{C}\}$, dove $\mathcal{C} = (a, b]$ con $a < b$. Come possiamo calcolarla a partire dal campione?

Legge dei grandi numeri e densità di probabilità

Supponiamo che anzichè essere interessati alla media, vogliamo calcolare la probabilità di un evento, ad esempio $p = \mathbb{P}\{\mathsf{X} \in \mathcal{C}\}$, dove $\mathcal{C} = (a, b]$ con $a < b$. Come possiamo calcolarla a partire dal campione?

In realtà lo abbiamo già visto in alcune esercitazioni, in cui abbiamo usato la frequenza relativa. Ovvero, abbiamo contato quante volte occorre nel campione l'evento $\mathsf{X}_i \in \mathcal{C}$.

Legge dei grandi numeri e densità di probabilità

Per ogni elemento del campione X_i, definisco Y_i come una nuova variabile aleatoria che assume valore 1 se $X_i \in \mathcal{C}$ e 0 altrimenti.

Questa variabile aleatoria si chiama indicatore. Qual è il suo valore atteso?

Legge dei grandi numeri e densità di probabilità

Per ogni elemento del campione X_i, definisco Y_i come una nuova variabile aleatoria che assume valore 1 se $\mathsf{X}_i \in \mathcal{C}$ e 0 altrimenti.

Questa variabile aleatoria si chiama indicatore. Qual è il suo valore atteso?

$$\mathbb{E}\left\{\mathsf{Y}_i\right\} = 1 \cdot \mathbb{P}\left\{\mathsf{X}_i \in \mathcal{C}\right\} + 0 \cdot \mathbb{P}\left\{\mathsf{X}_i \notin \mathcal{C}\right\} = p.$$

Queste sono variabili independenti essendo indipendenti X_i. La frequenza relativa è $\frac{1}{n}\sum_{i=1}^{n} \mathsf{Y}_i = \overline{\mathsf{Y}}_n$.

Legge dei grandi numeri e densità di probabilità

Applicando la legge dei grandi numeri a $\overline{\mathsf{Y}}_n$ con $\mu = p$, si ha

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mathsf{Y}}_n - p| \geqslant \epsilon\right\} = 0$$

Come poteva essere intuitivo, acquisendo una sequenza abbastanza lunga di realizzazioni, possiamo ottenere una stima accurata della probabilità p.

Legge dei grandi numeri e densità di probabilità

In termini di densità di probabilità, sappiamo che se scegliamo $\mathcal{C} = (a - h, a + h)$, con h piccolo

Legge dei grandi numeri e densità di probabilità

In termini di densità di probabilità, sappiamo che se scegliamo $\mathcal{C} = (a - h, a + h)$, con h piccolo

$$\overline{\mathsf{Y}}_n \simeq p = \int_{a-h}^{a+h} f_{\mathsf{X}}(x)dx \simeq f_{\mathsf{X}}(a)2h$$

Legge dei grandi numeri e densità di probabilità

In termini di densità di probabilità, sappiamo che se scegliamo $\mathcal{C} = (a-h, a+h)$, con h piccolo

$$\overline{\mathsf{Y}}_n \simeq p = \int_{a-h}^{a+h} f_{\mathsf{X}}(x)dx \simeq f_{\mathsf{X}}(a)2h$$

Segue che

$$f_{\mathsf{X}}(a) \simeq \frac{\overline{\mathsf{Y}}_n}{2h} = \frac{\text{quante volte } x_i \in \mathcal{C}/n}{\text{larghezza di } \mathcal{C}}$$

Legge dei grandi numeri e densità di probabilità

In termini di densità di probabilità, sappiamo che se scegliamo $\mathcal{C} = (a-h, a+h)$, con h piccolo

$$\overline{\mathsf{Y}}_n \simeq p = \int_{a-h}^{a+h} f_{\mathsf{X}}(x)dx \simeq f_{\mathsf{X}}(a)2h$$

Segue che

$$f_{\mathsf{X}}(a) \simeq \frac{\overline{\mathsf{Y}}_n}{2h} = \frac{\text{quante volte } x_i \in \mathcal{C}/n}{\text{larghezza di } \mathcal{C}}$$

Andiamo a fare qualche prova su R...

Teorema del Limite Centrale

- Consideriamo n variabili aleatorie $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ indipendenti e identicamente distribuite con media μ e deviazione standard σ.

Teorema del Limite Centrale

- Consideriamo n variabili aleatorie $X_1, X_2, \ldots, X_n$ indipendenti e identicamente distribuite con media μ e deviazione standard σ.
- Sia s_n la somma di tali variabili aleatorie e $\overline{X}_n$ la loro media campionaria

$$s_n = \sum_{i=1}^{n} X_n$$

$$\overline{X}_n = \frac{s_n}{n}$$

Teorema del Limite Centrale

- Consideriamo n variabili aleatorie $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ indipendenti e identicamente distribuite con media μ e deviazione standard σ.
- Sia s_n la somma di tali variabili aleatorie e $\overline{\mathsf{X}}_n$ la loro media campionaria

$$\mathsf{s}_n = \sum_{i=1}^{n} \mathsf{X}_n$$

$$\overline{\mathsf{X}}_n = \frac{\mathsf{s}_n}{n}$$

- Dalle proprietà della media e della varianza sappiamo che

$$\mathbb{E}\left\{\overline{\mathsf{X}}_n\right\} = \mu$$

$$\mathbb{V}\left\{\overline{\mathsf{X}}_n\right\} = \frac{\sigma^2}{n}$$

Standardizzazione

Data una qualsiasi variabile aleatoria X con media μ e dev. standard σ, la variabile aleatoria

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

è una nuova variabile aleatoria con media 0 e dev. standard 1.

Standardizzazione

Data una qualsiasi variabile aleatoria X con media μ e dev. standard σ, la variabile aleatoria

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

è una nuova variabile aleatoria con media 0 e dev. standard 1.
Cosa succede se applico la standardizzazione a $\overline{X}_n$ e s_n?

Standardizzazione

Data una qualsiasi variabile aleatoria X con media μ e dev. standard σ, la variabile aleatoria

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

è una nuova variabile aleatoria con media 0 e dev. standard 1.
Cosa succede se applico la standardizzazione a $\overline{X}_n$ e s_n?
Ottengo in entrambi i casi

$$Z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$$

Teorema del Limite Centrale

Per n abbastanza grande

$$\overline{\mathsf{X}}_n \approx \mathcal{N}(\mu, \sigma^2/n)$$

E quindi

$$\mathsf{Z}_n \approx \mathcal{N}(0,1), \quad \mathsf{s}_n \approx \mathcal{N}(n\mu, n\sigma^2)$$

Ovvero $\overline{\mathsf{X}}_n$ si può approssimare con una variabile aleatoria normale con media μ (come X_i) e varianza σ^2/n (minore della varianza di X_i). Inoltre Z_n è una variabile aleatoria normale standard.

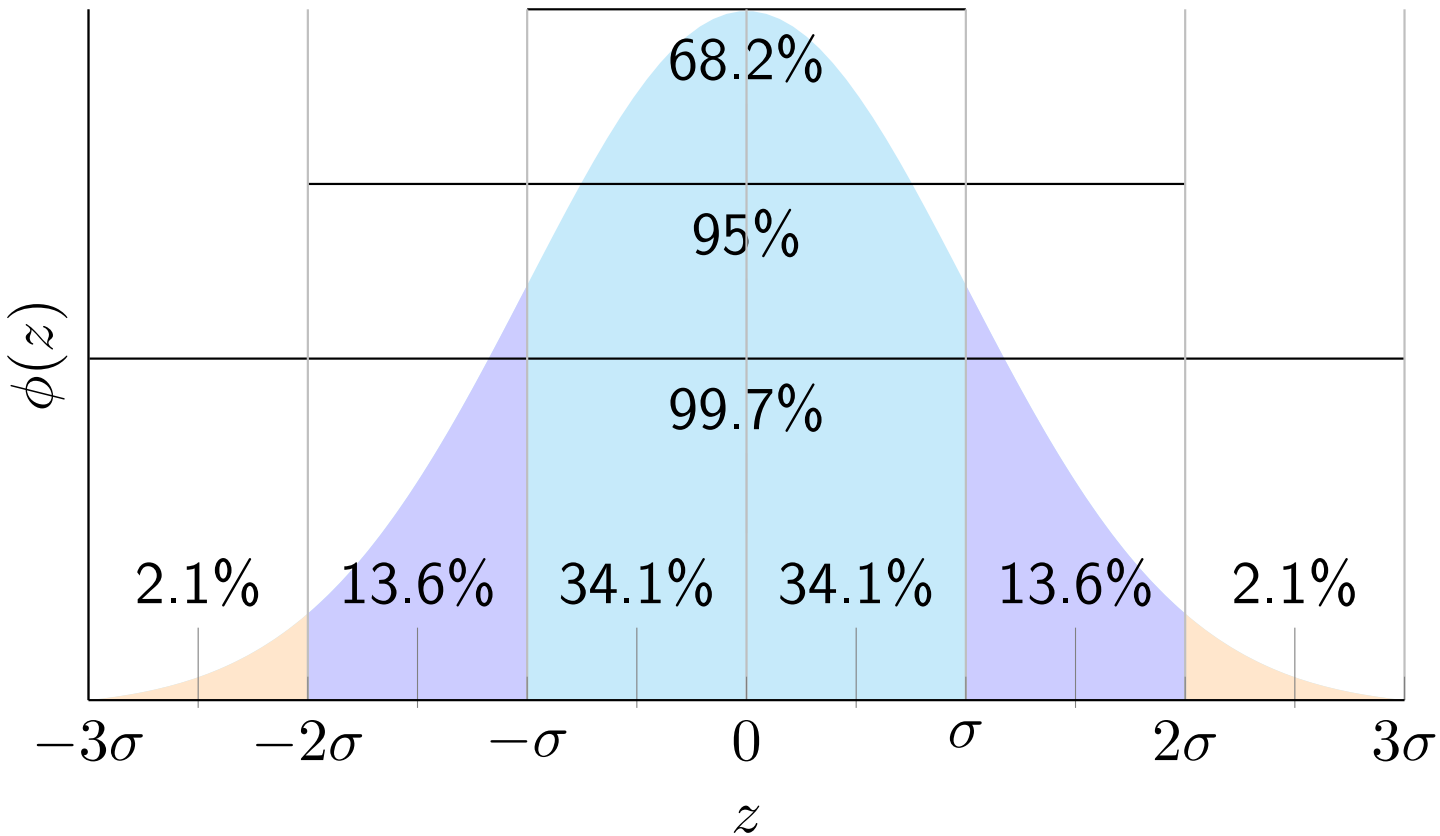
Teorema del Limite Centrale

Una definizione più precisa passa per la cdf della variabile aleatoria, che ci aiuta a capire cosa significa che una variabile aleatoria segue *approssimativamente* una certa distribuzione, ovvero

$$\lim_{n \to \infty} F_{\mathsf{Z}_n}(z) = \Phi(z)$$

Regola Empirica

Si dimostra che $\mathbb{P}\{-1 \leqslant \mathsf{Z} \leqslant 1\} \simeq 68\%$, $\mathbb{P}\{-2 \leqslant \mathsf{Z} \leqslant 2\} \simeq 95\%$, e $\mathbb{P}\{-3 \leqslant \mathsf{Z} \leqslant 3\} \simeq 99\%$



Abbastanza grande?

Proviamo a vedere qualche esempio..