4. La distribuzione delle statistiche campionarie

Indice

- 4. La distribuzione delle statistiche campionarie
  - La media campionaria
  - Il teorema del limite centrale
  - La varianza campionaria
  - Le distribuzioni delle statistiche di popolazioni normali
  - Campionamento da insiemi finiti

La **statistica** è la scienza che si occupa di trarre conclusioni dai dati sperimentali. Una situazione tipica riguarda lo studio di un insieme molto grande, detto **popolazione**, composto da oggetti a cui sono associate quantità misurabili. L'approccio statistico consiste nel selezionare un sottoinsieme ridotto di oggetti, chiamato **campione**, e analizzarlo per trarre conclusioni valide per l'intera popolazione.

Per poter effettuare inferenze sulla popolazione basandosi sui dati del campione, è necessario assumere alcune condizioni sulle relazioni che legano questi due insiemi. Una ipotesi fondamentale è che esista una distribuzione di probabilità nella popolazione, nel senso che, se si estraggono oggetti in modo casuale, le quantità numeriche a essi associate possono essere pensate come variabili aleatorie indipendenti e identicamente distribuite secondo una certa distribuzione F. Se il campione viene selezionato in modo casuale, è ragionevole supporre che i suoi dati siano valori indipendenti provenienti da tale distribuzione.

Un insieme di variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) $X_1, X_2, \ldots, X_n$, tutte con la stessa distribuzione F, si dice **campione aleatorio** estratto dalla distribuzione F.

In pratica, la distribuzione F non è mai completamente nota, ma è possibile utilizzare i dati per fare **inferenza** su di essa. In alcuni casi, F può essere nota a meno di alcuni parametri incogniti; in altri, potremmo non sapere nulla su F. I problemi in cui la distribuzione è nota eccetto che per un insieme di parametri incogniti sono detti problemi di inferenza **parametrica**, mentre quelli in cui non si sa nulla sulla distribuzione sono problemi di inferenza **non parametrica**.

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

La media campionaria

Data una **popolazione** di elementi con una quantità misurabile associata a ciascuno, consideriamo un **campione** aleatorio di dati $X_1, X_2, \ldots, X_n$ estratto da questa popolazione. Denotiamo con μ e σ^2 la media e la varianza della popolazione. La **media campionaria** è definita come:

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

Poiché $\bar{X}$ è una funzione delle variabili aleatorie X_i, essa stessa è una variabile aleatoria e una statistica.

Calcoliamo l'aspettazione della media campionaria:

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{n\mu}{n} = \mu$$

Quindi, la media campionaria è uno **stimatore non distorto** della media μ della popolazione.

La varianza della media campionaria è:

$$\operatorname{Var}(\bar{X}) = \operatorname{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Pertanto, la varianza di $\bar{X}$ diminuisce all'aumentare di n. Ciò significa che la media campionaria ha la stessa media μ della popolazione, ma la sua variabilità si riduce con l'aumentare della dimensione del campione.

Il teorema del limite centrale

Il **teorema del limite centrale** afferma che la somma di un gran numero di variabili aleatorie indipendenti e identicamente distribuite tende ad avere una distribuzione approssimativamente normale, indipendentemente dalla distribuzione originale delle variabili.

Formalmente, siano $X_1, X_2, \ldots, X_n$ variabili aleatorie i.i.d. con media μ e varianza σ^2. Allora, per n sufficientemente grande, la somma $S_n = X_1 + X_2 + \cdots + X_n$ è approssimativamente normale con media $n\mu$ e varianza $n\sigma^2$:

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2)$$

Normalizzando la somma, otteniamo una distribuzione normale standard:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx \mathcal{N}(0,1)$$

Questo risultato implica che anche la media campionaria $\bar{X}$ è approssimativamente normale per grandi n:

$$\bar{X} = \frac{S_n}{n} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Quindi, la variabile standardizzata:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1)$$

Applicazione al caso binomiale: Se X è una variabile aleatoria binomiale con parametri n e p, può essere vista come la somma di n variabili di Bernoulli indipendenti X_i, dove:

$$X_i = \begin{cases} 1 & \text{con probabilità } p \\ 0 & \text{con probabilità } 1-p \end{cases}$$

Poiché $E[X_i] = p$ e $\mathrm{Var}(X_i) = p(1-p)$, per n grande, il teorema del limite centrale ci permette di approssimare la distribuzione binomiale con una normale:

$$\frac{X - np}{\sqrt{np(1-p)}} \approx \mathcal{N}(0,1)$$

La varianza campionaria

Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio proveniente da una distribuzione con media μ e varianza σ^2. La **varianza campionaria** S^2 è definita come:

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

dove $\bar{X}$ è la media campionaria. La radice quadrata di S^2, denotata con S, è la **deviazione standard campionaria**.

Una proprietà importante è che S^2 è uno **stimatore non distorto** della varianza σ^2 della popolazione:

$$E[S^2] = \sigma^2$$

Ciò significa che, in media, la varianza campionaria S^2 coincide con la varianza della popolazione.

Le distribuzioni delle statistiche di popolazioni normali

Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una distribuzione normale $\mathcal{N}(\mu, \sigma^2)$, con le X_i indipendenti tra loro. La media e la varianza campionarie sono:

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

Poiché la somma di variabili normali indipendenti è ancora normale, la media campionaria $\bar{X}$ segue una distribuzione normale con media μ e varianza σ^2/n:

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

Pertanto, la variabile standardizzata:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Inoltre, nel caso di campioni da una popolazione normale, la varianza campionaria S^2 (opportunamente scalata) segue una distribuzione **chi quadrato** con $n-1$ gradi di libertà:

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

Un aspetto fondamentale è che $\bar{X}$ e S^2 sono variabili aleatorie **indipendenti** nel caso normale. Questa proprietà consente di utilizzare la distribuzione t di Student per costruire intervalli di confidenza e test statistici. In particolare, la variabile:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

segue una distribuzione t di Student con $n-1$ gradi di libertà.

Campionamento da insiemi finiti

Data una popolazione finita di N elementi, un **campione aleatorio** di dimensione n è un sottoinsieme di n elementi scelto in modo tale che tutti i $\binom{N}{n}$ possibili sottoinsiemi abbiano la stessa probabilità di essere selezionati.

Supponiamo che una frazione p degli elementi della popolazione possieda una certa caratteristica. Allora, ci sono pN elementi con la caratteristica e $(1-p)N$ senza. Selezionando un campione casuale di dimensione n, definiamo:

$$X_i = \begin{cases} 1 & \text{se l'} \ i\text{-esimo elemento possiede la caratteristica} \\ 0 & \text{altrimenti} \end{cases}$$

La somma $X = X_1 + X_2 + \cdots + X_n$ rappresenta il numero di elementi nel campione che possiedono la caratteristica. La media campionaria è quindi:

$$\bar{X} = \frac{X}{n}$$

Notiamo che:

$$P(X_i = 1) = p$$

Tuttavia, le variabili X_i non sono indipendenti perché la selezione è fatta senza reinserimento. La probabilità condizionata dipende dagli esiti precedenti:

$$P(X_j = 1 \mid X_i = 1) = \frac{pN - 1}{N - 1}, \quad P(X_j = 1 \mid X_i = 0) = \frac{pN}{N - 1}$$

Quando N è molto grande rispetto a n, questa dipendenza è trascurabile, e le X_i possono essere considerate **approssimativamente indipendenti**.

La media e la varianza di X sono:

$$E[X] = np$$

$$\mathrm{Var}(X) = np(1 - p)\left(\frac{N - n}{N - 1}\right)$$

Il fattore $\dfrac{N - n}{N - 1}$ è noto come **fattore di correzione per popolazioni finite**. Per $N \gg n$, questo fattore è circa 1, e le formule si riducono a quelle per il campionamento con reinserimento o per popolazioni infinite.

Per la media campionaria $\bar{X}$, otteniamo:

$$E[\bar{X}] = p$$

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n^2} = \frac{p(1 - p)}{n}\left(\frac{N - n}{N - 1}\right)$$

Questo risultato mostra che la media campionaria è uno **stimatore non distorto** della proporzione p nella popolazione, e la sua varianza tiene conto della dimensione finita della popolazione attraverso il fattore di correzione.

Nota. In tutti i casi trattati, l'aumento della dimensione del campione n porta a una riduzione della varianza degli stimatori, migliorando la precisione delle inferenze statistiche.