Statistica inferenziale

Problema

- Nello studio delle distribuzioni teoriche di probabilità si suppone di conoscere i principali parametri della popolazione che esaminiamo (ad esempio la media, varianza).

- Nelle applicazioni, i valori di questi parametri non sono noti. È troppo costoso o addirittura impossibile conoscere tutti i dati, individuo per individuo.

- Occorre perciò risalire a questi parametri utilizzando le informazioni contenute in un campione di osservazioni. Il processo attraverso il quale si traggono conclusioni su un'intera popolazione in base ad un campione si chiama **inferenza statistica**.

Stima intervallare

- Se per esempio se fossimo intressati al peso dei maschi in una certa regione, potremmo annotarne il perso $x_1, \ldots, x_n$ di n individui scelti a caso e utilizzare la *media campionaria*

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

  come **stima puntuale** per la media μ della popolazione.

- Però, da solo, questo numero non ha alcun significato: anche nelle migliori delle condizioni $\bar{x}$ non sarà mai esattamente uguale a μ.

- Al più potremmo fare una **stima intervallare** di μ. Ovvero trovare un intervallo I, per esempio $\left[\bar{x} - \varepsilon, \bar{x} + \varepsilon\right]$, che con (buona) probabilità p contenga μ.

- Questo intervallo I si chiama **intervallo di confidenza** e la probabilità p si chiama **livello di confidenza**.

Stima intervallare

- L'**intervallo di confidenza** ed il **livello di confidenza** non sono variabili indipendenti.
- Più alto il livello di confidenza più grande dovrà essere l'intervallo di confidenza (quindi più imprecisa la misura). E viceversa.
- Possiamo prima fissare l'ampiezza dell'intervallo di confidenza e ottenere un livello di confidenza, o più comunemente fissare un livello di confidenza e ottenere la minima l'ampiezza dell'intervallo di confidenza.

Stimatori

Vediamo ora precisamente la relazione tra intervallo di confidenza e livello di fiducia.

- Consideriamo $\bar{x}$ come il risultato di una variabile aleatoria. Se il campione è scelto casualmente, la media ottenuta da diversi campioni varierà con una legge che ora cercheremo di determinare.

- Un campione di n individui viene considerato come un insieme di n variabili aleatorie indipendenti: $X_1, \ldots, X_n$ identicamente distribuite. La **variabile aleatoria media campionaria** $\bar{X}$ è

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

- Per esempio, se siamo interessati al peso degli individui di una popolazione e consideriamo un campione di 10 individui, X_2 è il peso del secondo individuo considerato.

Stimatori

La distribuzione di probabilità di $\bar{X}$ si chiama **distribuzione della media campionaria** con campioni di dimensione n. Si dice che $\bar{X}$ è uno **stimatore** del parametro μ. Il numero n si chiama **dimensione** o **taglia** del campione, in inglese **size**.

Siamo interessati alle seguenti due questioni:

- Fissamo un ε vogliamo calcolare
$$P\left(\bar{X} - \varepsilon \leq \mu \leq \bar{X} + \varepsilon\right).$$
Ovvero fissati un intervallo di cofidenza determinare il livello di confidenza.

- Fissamo un s vogliamo calcolare ε tale che
$$P\left(\bar{X} - \varepsilon \leq \mu \leq \bar{X} + \varepsilon\right) = s.$$
Ovvero fissato un livello di confidenza determinare un intervallo di confidenza.

Digressione: la scelta del campione

- La scelta del campione è una procedura importante da compiere con attenzione. È un problema molto difficile di cui non ci occuperemo.

- Ad esempio nel campionare le intenzioni di voto potrebbe essere un errore ricavare i loro nomi da un elenco telefonico, perché verrebbero ad essere mal rappresentate le persone che non dispongono del telefono o che hanno optato per non comparire nell'elenco.

- Per esempio, se vogliamo valutare la pressione diastolica media per una determinata popolazione non possiamo accettare volontari per comporre il nostro campione.

La distribuzione di $\bar{X}$

- Supponiamo che v.a.i. $X_1, \ldots, X_n$ abbiano la stessa distribuzione con media μ e deviazione standard σ. Allora è facile verificare che

$$\bar{X} \text{ è una v.a. con media } \mu \text{ e deviazione standard } \frac{\sigma}{\sqrt{n}}.$$

- Quindi $\mathrm{Var}(\bar{X}) \to 0$ per $n \to \infty$. Ovvero con buona probabilità il valore di $\bar{X}$ non è distante da μ.

- Un teorema che non dimostreremo: *la somma di v.a. normali indipendenti ha a sua volta distribuzione normale*.

  Segue che se $X_1, \ldots, X_n$ hanno distribuzione $N(\mu, \sigma^2)$ allora

$$\bar{X} \text{ ha distribuzione } N\left(\mu, \frac{\sigma^2}{n}\right).$$

Teorema del Limite Centrale

In generale quando $X_1, \ldots, X_n$ sono v.a. indipendenti equidistribuite (ma non necessariamente normali) il seguente teorema ci da informazioni sulla distribuzione di $\bar{X}$.

- Il **Teorema del Limite Centrale** afferma che quando n è abbastanza grande

  $\bar{X}$ è una v.a. con distribuzione approssimativamente $N\left(\mu, \frac{\sigma^2}{n}\right)$.

- Equivalentemente: la v.a.

  $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  ha approssimativamente distribuzione $N(0, 1)$.

In questo corso non è possibile introdurre le nozioni che permenttono di enunciare il teorema in maniera precisa. Una regola empirica è ritenere che l'approssimazione "sufficientemente" buona per $n \geq 30$

Esempio 1a

Consideriamo la distribuzione dei livelli di colesterolo in individui di età compresa fra i 20 ei 74 anni. La media della popolazione è $\mu = 212\,\mathrm{mg/dL}$ e la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$. Selezioniamo campioni ripetuti di dimensione $n = 49$ dalla popolazione.

- Qual è la probabilità che la media campionaria sia superiore a $230\,\mathrm{mg/dL}$?

Ci interessa $P(\bar{X} \geq 230)$. Dal teorema del limite centrale $\bar{X}$ ha distribuzione (approssimativamente) normale con media $\mu = 212\,\mathrm{mg/dL}$ ed errore standard $\sigma/\sqrt{n} = 56/7 = 8$. La variabile $Z = (\bar{X} - 212)/8$ è una normale standardizzata.

Dunque $P(\bar{X} \geq 230) = P(Z \geq (230 - 212)/8) = P(Z \geq 2.25)$. Dalla tabella della distribuzione normale, l'area a destra di $z = 2.25$ è 0.01. Quindi circa l'1% dei campioni di dimensione 49 avrà una media maggiore o uguale a $230\,\mathrm{mg/dL}$.

Esempio 1b

Consideriamo la distribuzione dei livelli di colesterolo in individui di età compresa fra i 20 ei 74 anni. La media della popolazione è $\mu = 212\,\mathrm{mg/dL}$ e la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$. Selezioniamo campioni ripetuti di dimensione $n = 49$ dalla popolazione.

- Trovare il minimo ε tale che la media campionaria appartiene all'intervallo $\big[\mu - \varepsilon, \mu + \varepsilon\big]$ con probabilità del 95%.

Vogliamo trovare ε tale che $P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) = 0.95$ ovvero tale che $P(\varepsilon/8 \leq Z \leq \varepsilon/8) = 0.95$. Dalla tavola della distribuzione $N(0,1)$ sappiamo che $0.95 = P\big(-1.96 \leq Z \leq 1.96\big)$ quindi $\varepsilon = 8 \cdot 1.96 = 15.68$

Esempio 1c

Consideriamo la distribuzione dei livelli di colesterolo in individui di età compresa fra i 20 ei 74 anni. La media della popolazione è $\mu = 212\,\mathrm{mg/dL}$ e la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$.

- Qual è la minima dimensione del campione affinché la media campionaria sia compresa nell'intervallo $\left[\mu - 5, \mu + 5\right]$ con una probabilità del 95%?

Dobbiamo trovare una n per cui $P(\mu - 5 \leq \bar{X} \leq \mu + 5) = 0.95$ ovvero tale che

$$P\left(-\frac{5}{\sigma/\sqrt{n}} \leq Z \leq \frac{5}{\sigma/\sqrt{n}}\right) = 0.95$$

Dalla tabella troviamo $1.96 = \dfrac{5}{\sigma/\sqrt{n}}$ e risolvendo otteniamo $n = 369$.