Concetti chiave e regole

Il teorema di Bayes

Il **teorema di Bayes** esprime la probabilità $p(A_i|B)$ che un evento B sia stato causato da una fra le n possibili ipotesi A_i :

$$p(A_i|B) = \frac{p(B|A_i) \cdot p(A_i)}{p(B|A_1) \cdot p(A_1) + p(B|A_2) \cdot p(A_2) + ... + p(B|A_n) \cdot p(A_n)}$$

dove $p(B|A_i)$ è la probabilità che si verifichi l'evento B, supposto che si sia verificato A_i.

Popolazione e campione

Lo studio di un fenomeno statistico si basa sull'analisi di un **campione** della popolazione; è poi l'inferenza statistica che si occupa di stabilire le regole in base alle quali estendere le osservazioni fatte sull'intera popolazione.
Il rapporto tra la numerosità del campione e l'intera popolazione si chiama **tasso di campionamento** e si esprime di solito in forma percentuale.
Un campione è di solito di tipo casuale e si parla di **campionamento casuale semplice** se:

- ogni elemento della popolazione ha la stessa probabilità di essere estratto
- ogni campione ha la stessa probabilità di essere formato.

Se la popolazione ha N elementi ed il campione estratto ne ha n, nell'ambito di un campionamento casuale semplice possiamo avere

- il **campionamento bernoulliano** (estrazioni con reimmissione): numero di campioni $= N^n$
- il **campionamento in blocco** (estrazioni senza reimmissione): numero di campioni $= \binom{N}{n}$

Le variabili campionarie

L'esito della i-esima estrazione di un elemento del campione è una variabile aleatoria X_i che può assumere un qualsiasi valore della popolazione. Ogni campione di ampiezza n è un **vettore aleatorio** $(X_1, X_2, ..., X_n)$ le cui componenti sono le variabili aleatorie X_i; il suo spazio campionario è l'insieme Ω di tutte le realizzazioni campionarie di ampiezza n estraibili dalla popolazione.

Parametri e stimatori

Un **parametro** di una popolazione è un qualsiasi indice che caratterizza in modo sintetico una popolazione. Di solito i parametri di una popolazione non sono noti, di essi si può però fare una stima mediante il corrispondente parametro del campione; il parametro del campione prende il nome di **stimatore**. Uno stimatore si dice:

- **corretto** se il suo valore atteso è uguale al parametro ϑ che deve stimare
- **consistente** se, al crescere dell'ampiezza n del campione, la sua varianza tende a zero
- **più efficiente** di un altro stimatore se la sua varianza è minore della varianza dell'altro stimatore.

I principali stimatori di una popolazione X di media μ e varianza σ^2 sono i seguenti.

- La **media campionaria** è la variabile aleatoria $\overline{X_n}$ che descrive le medie di tutti i possibili campioni di ampiezza n che si possono estrarre dalla popolazione; si verifica che:

  - per un campionamento bernoulliano: $E(\overline{X_n}) = \mu \qquad V(\overline{X_n}) = \frac{\sigma^2}{n}$

  - per un campionamento in blocco: $E(\overline{X_n}) = \mu \qquad V(\overline{X_n}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

  La media campionaria è uno stimatore corretto e consistente della media della popolazione.

© Istituto Italiano Edizioni Atlas

- La **varianza campionaria** è la variabile aleatoria S_n^2 che descrive le varianze di tutti i possibili campioni di ampiezza n che si possono estrarre dalla popolazione; si verifica che:

  - per un campionamento bernoulliano: $E(S_n^2) = \sigma^2 \cdot \frac{n-1}{n}$

  - per un campionamento in blocco: $E(S_n^2) = \sigma^2 \cdot \frac{n-1}{n} \cdot \frac{N}{N-1}$

  La varianza campionaria non è quindi uno stimatore corretto della varianza della popolazione.
  Per avere uno stimatore corretto occorre considerare un opportuno fattore di correzione e si verifica che lo stimatore corretto della varianza della popolazione è:

  - per un campionamento bernoulliano: $\widehat{S_n^2} = S_n^2 \cdot \frac{n}{n-1}$

  - per un campionamento in blocco: $\widehat{S_n^2} = S_n^2 \cdot \frac{n}{n-1} \cdot \frac{N}{N-1}$

- La **frequenza campionaria** è la variabile aleatoria F_n che descrive le varianze di tutti i possibili campioni di ampiezza n che si possono estrarre dalla popolazione; si verifica che:

  - per un campionamento bernoulliano: $E(F_n) = p \qquad V(F_n) = \frac{p(1-p)}{n}$

  - per un campionamento in blocco: $E(F_n) = p \qquad V(F_n) = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}$

  La frequenza campionaria è uno stimatore corretto della frequenza della popolazione.

- Date due popolazioni X e Y rispettivamente di ampiezza N_1 e N_2, medie μ_1 e μ_2, varianze σ_1^2 e σ_2^2, la **differenza tra le medie** è la variabile aleatoria $\overline{X} - \overline{Y}$ che si costruisce considerando le differenze tra le medie di tutti i campioni di ampiezza n_1 e n_2 estratti dalle due popolazioni; si verifica che:

  - per un campionamento bernoulliano: $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \qquad V(\overline{X_n}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

  - per un campionamento in blocco: $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \qquad V(\overline{X_n}) = \frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}$

  La differenza tra le medie è uno stimatore corretto della differenza tra le medie di due popolazioni.

Il caso della distribuzione normale

Se una popolazione ha distribuzione normale con media μ e varianza σ^2, anche la variabile aleatoria media campionaria si distribuisce normalmente con la stessa media μ e varianza $\frac{\sigma^2}{n}$. Inoltre, il teorema del limite centrale garantisce che questo comportamento della media campionaria vale anche se la popolazione non ha distribuzione normale (purché abbia media μ e varianza σ^2 entrambe finite). Se però il campione è sufficientemente grande ($n > 30$), al crescere dell'ampiezza n del campione, comunque sia distribuita la popolazione di media μ e varianza σ^2 entrambe finite, la media campionaria ha distribuzione normale con la stessa media della popolazione e varianza $\frac{\sigma^2}{n}$.

La stima dei parametri

La stima di un parametro ϑ può essere:

- **puntuale**, e in questo caso restituisce un valore preciso per il parametro della popolazione con una precisione prefissata
- **per intervallo**, e in questo caso restituisce un intervallo che, con un fissato margine di errore, contiene il valore vero del parametro.

© Istituto Italiano Edizioni Atlas

La stima puntuale

Nel caso in cui il parametro da stimare sia la media, la frequenza e la differenza tra le medie, la stima viene fatta attribuendo a ϑ il corrispondente parametro del campione (tutti questi stimatori sono corretti); l'errore viene valutato con la deviazione standard della distribuzione campionaria di quel parametro.
Nel caso in cui il parametro da stimare sia la varianza, si deve ricorrere alla varianza corretta del campione.
L'errore viene valutato dalle seguenti espressioni:

- se è nota la varianza della popolazione:

	Media	Differenza tra le medie
Campionamento bernoulliano	$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
Campionamento in blocco	$\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$	$\sqrt{\frac{\sigma_1^2}{n_1}\cdot\frac{N_1-n_1}{N_1-1}+\frac{\sigma_2^2}{n_2}\cdot\frac{N_2-n_2}{N_2-1}}$

- se non è nota la varianza della popolazione:

	Media	Differenza tra le medie	Frequenza
Campionamento bernoulliano	$\frac{s}{\sqrt{n-1}}$	$\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}$	$\sqrt{\frac{f(1-f)}{n}}$
Campionamento in blocco	$\frac{s}{\sqrt{n-1}}\cdot\sqrt{1-\frac{n}{N}}$	$\sqrt{\frac{s_1^2}{n_1-1}\cdot\left(1-\frac{n_1}{N_1}\right)+\frac{s_2^2}{n_2-1}\cdot\left(1-\frac{n_2}{N_2}\right)}$	$\sqrt{\frac{f(1-f)}{n}}\cdot\sqrt{\frac{N-n}{N-1}}$

La stima per intervallo

Per determinare l'intervallo di confidenza a livello $1-\alpha$ che contiene il valore vero del parametro si ricorre:

- alla variabile standardizzata z se il campione è un grande campione $(n > 30)$ oppure se, pur trattandosi di un piccolo campione $(n \leq 30)$, è nota la varianza σ^2 oppure la popolazione ha distribuzione normale
- alla variabile t di Student se il campione è piccolo e σ^2 non è nota.

La verifica di ipotesi

Una ipotesi statistica è una affermazione che viene fatta in merito ad un parametro ϑ di una variabile aleatoria X.
Si chiama **ipotesi nulla** l'affermazione $H_0 : \vartheta = \vartheta_0$; una qualunque ipotesi H_1 diversa da quella nulla viene detta **ipotesi alternativa**.
Tra le ipotesi alternative, le più significative sono:

- $H_1 : \vartheta \neq \vartheta_0$ che dà origine a un test a due code
- $H_1 : \vartheta > \vartheta_0$ oppure $H_1 : \vartheta < \vartheta_0$ che danno origine a un test a una coda

© Istituto Italiano Edizioni Atlas