PERCORSO AGENZIA DELLE ENTRATE

Variabili aleatorie e Modelli probabilistici

Introduzione

Benvenuti!

In questa lezione continueremo a parlare di variabili aleatorie e modelli probabilistici.

In particolare, andremo ad approfondire:

- il campionamento con ripetizione
- la definizione delle statistiche campionarie e le proprietà degli stimatori puntuali
- la stima intervallare, con particolare attenzione alla stima della media della popolazione e alla stima delle differenza tra due medie

Bene, non ci resta che cominciare…

Concetti base – Introduzione al campionamento

Si vuole studiare una popolazione con riferimento a particolari caratteristiche di interesse. La popolazione viene esaminata in modo parziale, considerando un campione di unità statistiche, cioè un aggregato di unità, appartenenti alla popolazione di riferimento, selezionate mediante l'esperimento di campionamento.

- La statistica inferenziale fornisce strumenti e metodi per ricavare dai dati campionari/informazioni sulla popolazione
- L'inferenza statistica studia l'analisi dei dati che costituiscono un campione casuale, cioè selezionato mediante un esperimento casuale (aleatorio)

Il campionamento in breve

Se l'obiettivo è quello di ottenere informazioni sulla popolazione utilizzando un suo sottoinsieme, il campione, allora si pone il problema di come estrarre le unità statistiche che entrano a far parte del campione medesimo.

- Il campione deve essere rappresentativo
- La dimensione campionaria viene indicata con *n*
- Regola di selezione di tipo probabilistico

Esempio: il campionamento casuale semplice

- È necessario disporre di una lista di campionamento che contiene tutte le unità statistiche della popolazione
- Alle unità della lista viene associato un numero (etichetta)
- Dall'insieme dei numeri che identificano le unità statistiche ne vengono estratti casualmente n, con ripetizione

Le *n* unità sono estratte in modo tale che:

- ogni unità della popolazione ha la stessa probabilità di essere estratta
- le *n* estrazioni sono effettuate ognuna indipendentemente dall'altra

Esistono ovviamente altri metodi di campionamento.

E' semplice dimostrare che l'operazione di estrazione di un campione casuale semplice dalla popolazione X dà luogo ad una n-pla $(X_1, X_2, \dots, X_n)$ a componenti indipendenti ed identicamente distribuite a X.

Esempio di campionamento

Consideriamo una popolazione di 4 individui di età: 18, 20, 22, 24. Quali sono i possibili campioni di numerosità 2?
Ovviamente usiamo una popolazione piccola come esempio.
Otteniamo 16 diversi campioni di numerosità 2.
Le distribuzioni marginali delle componenti del campione sono identiche tra loro e con la distribuzione nella popolazione (identicamente distribuite).
Le distribuzioni delle singole componenti del campione sono indipendenti. La distribuzione congiunta è pari al prodotto delle marginali.

Statistiche campionarie

Una volta estratto il campione, calcoliamo su questo una o più statistiche (media campionaria, somma campionaria, varianza campionaria, ecc.) utili per fare inferenza.

- Le statistiche assumono valori diversi su campioni diversi
- La probabilità che una statistica assuma un determinato valore dipende dalla probabilità di avere un campione con determinate caratteristiche
- Ogni statistica è una variabile aleatoria
- La loro distribuzione è detta campionaria in quanto è generata considerando l'universo di tutti i possibili campioni

Esempio: la distribuzione della media campionaria

$n = 2$

$\bar{x}$	$\mathbf{P}(\bar{x})$	$\bar{x}\mathbf{P}(\bar{x})$	$(\bar{x} - \mu_{\bar{x}})^2\mathbf{P}(\bar{x})$
18	1/16	18/16	9/16
19	2/16	38/16	8/16
20	3/16	60/16	3/16
21	4/16	84/16	0
22	3/16	66/16	3/16
23	2/16	46/16	8/16
24	1/16	24/16	9/16
		336/16	40/16

$E(\bar{x}) = 21, Var(\bar{x}) = 2.5$

- Per ogni campione precedentemente estratto, calcoliamo la media.
- A ciascuna media è associata la rispettiva probabilità data dalla probabilità di uno specifico campione di essere estratto
- Costruiamo così la distribuzione della (variabile aleatoria) media campionaria
- Ricordiamo che la media della popolazione è 21
- A seconda del campione estratto, abbiamo medie campionarie diverse

Ricordiamo che uno stimatore è una funzione dei dati campionari.
Possiamo considerare la media campionaria come stimatore della media popolazione, anche se al variare del campione otteniamo medie campionarie diverse e lontane da 21?
La risposta è sì. Ma spieghiamo perché. Lo stimatore media campionaria è uno stimatore non distorto della media della popolazione. Cosa vuol dire?

- Uno stimatore è non distorto se il valore atteso dello stimatore è pari al parametro; quindi, la media campionaria in media (cioè il suo valore atteso) restituisce la media della popolazione.
- La media campionaria è una variabile aleatoria. Oltre al valore atteso, possiamo calcolare la varianza. I passaggi sono riportati nella tabella.
- In generale, si dimostra che la media e la varianza della media campionaria sono legate ai corrispondenti valori della popolazione nel modo seguente

$$E(\bar{X}) = \mu$$
$$V(\bar{X}) = \frac{\sigma^2}{n}$$

- All'aumentare della numerosità campionaria, migliora la precisione della stima puntuale.
- Se la distribuzione della popolazione è di tipo normale, allora è possibile dimostrare che anche la media campionaria ha una distribuzione di tipo normale.

Stima per intervallo

- Un intervallo di confidenza (denominato anche stima intervallare) fornisce un intervallo di valori e un valore di probabilità che rappresenta la verosimiglianza che un intervallo includa veramente il valore del parametro della popolazione (sconosciuto)
- Costruire un intervallo di valori plausibili per il parametro di interesse
- Date le due statistiche L_1 e L_2, tali che $L_1 < L_2$, l'intervallo aleatorio $[L_1, L_2]$ è detto intervallo di confidenza per θ al livello $1-\alpha$ se $\Pr(L_1 < \theta < L_2) = 1-\alpha$
- La probabilità che l'intervallo aleatorio $[L_1, L_2]$ copra il vero valore di θ è $1-\alpha$
- Sbagliato: la probabilità che θ cada nell'intervallo $[L_1, L_2]$ è $1-\alpha$

Intervalli di confidenza per la media di una popolazione

Ci concentriamo su due diversi casi: quando la popolazione ha una distribuzione Normale o quando siamo in presenza di grandi campioni.

- Vedremo come costruire intervalli di confidenza per la media di una popolazione Normale, quando lo scarto quadratico medio è noto o incognito (e quindi da stimare)
- Vedremo come costruire intervalli di confidenza per la media di una popolazione per campioni di numerosità superiore a 30
- In entrambi i casi, l'intervallo di confidenza sarà simmetrico, centrato sulla stima puntuale, cioè sulla media campionaria. Quindi, l'estremo inferiore (superiore) è pari alla media campionaria meno (più) una quantità
- La differenza tra estremo superiore e inferiore dell'intervallo viene detta ampiezza dell'intervallo

Popolazione Normale, scarto quadratico medio noto

Consideriamo una popolazione normale con media μ (incognita) e assumiamo che sia nota la varianza σ^2.

- Ricordiamo la standardizzazione: la media campionaria meno la sua media diviso per il suo scarto quadratico medio. Siccome la popolazione è Normale, anche la distribuzione della media campionaria sarà Normale.
- Prendiamo un quantile della Normale standardizzata $z_{\alpha/2}$ che lasci una probabilità nella coda parti ad $alpha/2$. Ricordiamo che la Normale è simmetrica. La probabilità di osservare un valore compreso in un intervallo simmetrico riferito al quantile $z_{\alpha/2}$ è $1-\alpha$
- Risolvendo la disequazione, lasciando al centro la quantità ignota, cioè la media della popolazione otteniamo le formule per definire gli estremi dell'intervallo di confidenza
- L'estremo inferiore (superiore) è pari alla media campionaria meno (più) il quantile della Normale standardizzata che lascia nella coda una probabilità pari ad alpha/2 moltiplicata per lo scarto quadratico medio della media campionaria
- L'ampiezza dell'intervallo di confidenza dipende dalla numerosità campionaria, dal livello di confidenza e dalla varianza della popolazione:
    - all'aumentare della numerosità campionaria, l'ampiezza dell'intervallo si riduce
    - all'aumentare del livello di confidenza, l'ampiezza aumenta
    - all'aumentare della varianza della popolazione, l'ampiezza aumenta

Popolazione Normale, scarto quadratico medio NON noto

Nelle applicazioni pratiche spesso la varianza della popolazione è non nota e quindi stimata calcolando la varianza campionaria corretta

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

- La varianza campionaria corretta è uno stimatore non distorto della varianza della popolazione
- La varianza campionaria $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ è uno stimatore distorto della varianza della popolazione. In media NON restituisce il vero valore della varianza della popolazione
- Se standardizziamo la media campionaria utilizzando $\tilde{S}$ invece di σ, la variabile aleatoria

$$\frac{\bar{x} - \mu}{\frac{\tilde{S}}{\sqrt{n}}}$$

non ha una distribuzione di tipo normale standard

$$\frac{\bar{x} - \mu}{\frac{\tilde{S}}{\sqrt{n}}} \sim t_\nu$$

- La distribuzione t di Student dipende da un solo parametro $\nu(\nu > 1)$, detto gradi di libertà (gdl), ha media 0 (come la normale standard) e varianza $\nu/(\nu-1)$ (maggiore di 1 e quindi maggiore della varianza della normale standard)
- All'aumentare dei gradi di libertà, la distribuzione t di Student è sempre meglio approssimata da una normale standard. In formule $\nu \to \infty \Rightarrow t_\nu \to Z$

Per costruire l'intervallo di confidenza per la media della popolazione, ragioniamo come fatto in precedenza.

- Prendiamo un quantile della t di Student t che lasci una probabilità nella coda parti ad $alpha/2$. Ricordiamo che la t di Student è simmetrica. La probabilità di osservare un valore compreso in un intervallo simmetrico riferito al quantile $t_{\alpha/2}$ è $1-\alpha$
- Risolvendo la disequazione, lasciando al centro la quantità ignota, cioè la media della popolazione otteniamo le formule per definire gli estremi dell'intervallo di confidenza
- L'estremo inferiore (superiore) è pari alla media campionaria meno (più) il quantile della t di Student che lascia nella coda una probabilità pari ad alpha/2 moltiplicata per lo scarto quadratico medio campionario diviso per la numerosità del campione
- L'ampiezza dell'intervallo di confidenza dipende dalla numerosità campionaria, dal livello di confidenza e dalla varianza campionaria corretta:
    - all'aumentare della numerosità campionaria, l'ampiezza dell'intervallo si riduce
    - all'aumentare del livello di confidenza, l'ampiezza aumenta

- all'aumentare della varianza campionaria corretta, l'ampiezza aumenta

Grandi campioni

- Consideriamo una popolazione normale con media μ (incognita) e supponiamo che la numerosità sia sufficientemente ampia da poter assumere

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

- Vale tutto ciò già detto nel caso di popolazione Normale con scarto quadratico medio noto.

Intervalli di confidenza per la differenza tra due medie, scarto quadratico medio incognito

Qualora volessimo costruire un intervallo di confidenza tra due medie con scarto quadratico medio incognito, dobbiamo assumere che $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

- Uno stimatore non distorto di σ^2 è dato da

$$\widehat{\sigma^2} = \frac{(n_x - 1)\widetilde{s_x^2} + (n_y - 1)\widetilde{s_y^2}}{n_x + n_y - 2}$$

- Di conseguenza

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\widehat{\sigma^2}}{n_x} + \frac{\widehat{\sigma^2}}{n_y}}} \sim t_{n_x + n_y - 2}$$

- Si può poi procedere a costruire l'intervallo come descritto nel caso di popolazione normale con scarto quadratico medio incognito

Conclusioni

Bene, questo è tutto.

Ti ricordo che abbiamo parlato delle variabili aleatorie e dei modelli probabilistici. In particolare abbiamo visto:

- il campionamento con ripetizione
- la definizione delle statistiche campionarie e le proprietà degli stimatori puntuali
- infine, la stima intervallare

Grazie per l'attenzione!