Distribuzioni campionarie

Statistica descrittiva→ raccogliere, presentare e descrivere i dati

Statistica inferenziale→ trarre conclusioni e prendere decisioni riguardanti una popolazione sulla base dei dati campionari

Popolazione→ insieme di tutte le unità o individui oggetto di studio

Campione→ sottoinsieme della popolazione, il campione viene utilizzato perché:

- consente di ottenere risultati statistici con precisione sufficientemente alta
- presenta notevoli vantaggi rispetto ad un censimento
    - organizzazione più semplice
    - minore spesa
    - tempi più brevi

campionamento→ modalità di estrazione del campione dalla popolazione

inferenza→ processo di generalizzazione per il quale i risultati ottenuti su un campione vengono estesi alla popolazione

inferenza statistica→è un procedimento di induzione di tipo quantitativo, per cui l'incertezza del procedimento viene quantificata. L'incertezza è dovuta a:

- variabilità campionaria→ tutti i possibili campioni sono diversi
- errori di misurazione→ misurando più volte la stessa entità si ottengono valori diversi

facciamo inferenza sulla popolazione esaminando i risultati campionari

- statistiche campionarie (note)
- parametri della popolazioni (non sono noti, ma possono essere stimati usando il campione)

<u>deduzione:</u>

- dal generale al particolare
- tipica della logica e della matematica
- conclusioni certe

<u>induzione:</u>

- dal particolare al generale
- tipica delle discipline scientifiche
- conclusioni incerte

campionamento

si estrae un'unità a caso→ il valore che si osserverà è una v.a. X con

- supporto =modalità
- probabilità = frequenza relativa

la distribuzione di probabilità coincide con la distribuzione delle frequenze relative del carattere X.

Popolazione finita→ esiste un collettivo di N unità da cui se ne estraggono casualmente n<N, possono essere utilizzate strategie di campionamento diverse:

- semplice (con equiprobabilità) o complesso (con probabilità diverse)
- con ripetizione (reimbussolamento) o senza (In blocco)

popolazione infinita→ la popolazione è infinita ogni volta in cui non è opportunamente delimitata, cioè non è concettualmente possibile elencare i suoi membri.

<u>Parametri</u>

- Tipicamente l0inferenza riguarda alcuni parametri
- Poiché il campionamento casuale genera una v.a. X con la stessa distribuzione del carattere x, si chiamano parametri anche gli indici della distribuzione della variabile aleatoria X
- Sono quantità:
    - Fisse
    - incognite

<u>campionamento casuale</u>

- la teoria statistica di base si basa sulla nozione di campione casuale
- indichiamo con X la variabile aleatoria che descrive la distribuzione del carattere nella popolazione e supponiamo di estrarre un campione di dimensione n
- prima di effettuare l'estrazione, il valore mostrerà la i-ma unità estratta è ignoto, p una variabile aleatoria che indichiamo con xi
- il campione è un vettore di n variabili aleatorie x1, x2, x3 … xn
- il campionamento si dice casuale quando le n variabili sono indipendenti e identicamente distribuite come X
- il metodo di campionamento senza ripetizione o in blocco non può generare un campione casuale poiché induce correlazione tra gli elementi campionari
- il metodo di campionamento con ripetizione può generare un campione casuale, ma non è detto che ciò accada:
    - se i valori sono correlati nella popolazione, lo sono anche nel campione, qualunque metodo di campionamento si usi

distribuzione campionaria della media

- nozioni di stimatore→ gli stimatori sono popolazione e campione, ad ogni quantità della popolazione ha un suo analogo nel campione. Al parametro corrisponde la statistica, è naturale quindi cercare di stimare un parametro di interesse con la corrispondente statistica. Quando una statistica viene usata a fini inferenziali per stimare un parametro viene detta stimatore.
- esempio con una piccola popolazione
- proprietà di non distorsione
- errore standard

inferenza sulla media

tipicamente i parametro di interesse primario è la media della popolazione. Disponendo di un campione, lo stimatore naturale della media della popolazione è la media campionaria. Prima di estrarre il campione, la media campionaria è una variabile aleatoria.

I due principiali indici di sintesi di una variabile aleatoria sono il valore atteso e la varianza. La media campionaria è uno stimatore della media della popolazione. Una volta estratto il campione, lo stimatore produce una stima, questa stima è un processo inferenziale e quindi è soggetto ad errore, questo errore va corretto.

Errore di stima

Ogni campione è caratterizzato da un errore di stima, di fatto questo errore è ignoto quindi non si può valutare se una sia stima sia accurata oppure no. Per questo motivo si valutano le proprietà dello stimatore, uno stimatore si dice non distorto quando il valore atteso dell'errore di stima è nullo.

Se E (X- media x) = o

La media campionaria è uno stimatore non distorto. Uno stimatore è non distorto quando coincide con il parametro id interesse.

Variabilità dello stimatore

- Supponiamo che lo stimatore in questione sia non distorto. Questa è una buona proprietà, che tuttavia non garantisce l'accuratezza della stima.
- Ci si pongono delle domande del tipo:
- Qual è l'ordine di grandezza degli errori di stima
- Quanto è probabile incorrere in un errore di stima più grande di un certo valore prefissato?
- Occorre dunque quantificare il livello di incertezza associato allo stimatore, cioè quanto le stime variano da campione a campione→ varianza campionaria ed errore standard

Errore standard della media

La varianza campionaria della media campionaria è $\frac{\text{varianza di x nella popolazione}}{\text{dimensione del campione}}$. L'errore standard della media campionaria è la deviazione standard e descrive la variabilità di X intorno alla media di x

La deviazione standard della media campionaria si trova facendo $\frac{deviazione\ standard\ di\ X\ nella\ popolazione}{\text{radice quadrata della dimensione della popolazione}}$

L'errore standard della media campionaria è:

- Direttamente proporzionale alla deviazione standard del carattere nella popolazione— quanto più il carattere varia nella popolazione, tanto più la media varia da campione a campione
- Inversamente proporzionale alla radice quadrata della dimensione del campione→ quanto più grande è il campione, tanto meno la media varia da campione a campione

La media campionaria è una statistica relativa all'osservazione di un campione composto da n unità.

Nel calcolare la media i valori grandi e piccoli si compensano→ la media è meno variabile delle singole osservazioni.

Campionamento senza ripetizione

- Non produce un campione casuale
- La media campionaria è uno stimatore non distorto in entrambi i casi
- Tuttavia, la sua varianza è inferiore se il campionamento è senza ripetizione
- Nel caso di campionamento senza ripetizione la varianza è minore a causa di un fattore chiamato fattore di correzione per popolazioni finite → $\frac{N-n}{N-1} \approx 1-f$

Confronto campionamento con e senza ripetizione

- Il buon senso suggerisce di evitare il campionamento con ripetizione perché ammette la possibilità che una unità compaia più volte
- Il campionamento con ripetizione causa un aumento della varianza campionaria, quindi una perdita di efficienza.
- Il campionamento senza ripetizione genera dipendenza tra le osservazioni→ il campione non può essere casuale
- Il grado di dipendenza indotto dal campionamento senza ripetizione è funzione della frazione di campionamento f= $\frac{numerosità\ del\ campione}{\text{numerosità della popolzione}}$
- La dipendenza indotta è tanto più debole quanto più la frazione di campionamento f è piccola
- Se la frazione di campionamento a ' gli elementi campionari sono approssimativamente indipendenti e quindi i due tipi di estrazione sono quasi del tutto equivalenti

- Per la maggior parte delle finalità quando la frazione di campionamento è inferiore al 5% la dipendenza indotta al campionamento senza ripetizione è trascurabile
- Quando la popolazione è infinita la distinzione fra campionamento con e senza ripetizione svanisce del tutto

Correzione per popolazioni finite

Va applicata se:

- La popolazione è finita
- Il campionamento è in blocco
- Il campione è ampio rispetto alla popolazione, quindi n è maggiore del 5% di N

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \quad \text{oppure} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Distribuzione campionaria della media

La media campionaria è uno stimatore della media della popolazione. La media di un campione è uno stimatore non distorto con errore standard. La medi di un campione casuale ha sempre valore atteso media del campione= media della popolazione ed errore standard pari all'errore standard.

Questi dati non sono sufficienti salvo nei casi in cui la distribuzione appartenga ad una famiglia i cui parametri sono completamente identificati da valore atteso e deviazione standard. Un esempio di questo è se la media campionaria ha distribuzione normale, allora

$$\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

Ricorrendo alle tavole della standardizzata si può rispondere alle domande. In quali casi la media campionaria ha distribuzione esattamente o approssimativamente normale?

- Campione casuale da una popolazione normale-→ la media campionaria ha distribuzione esattamente normale
- Campione casuale da una popolazione in cui la distribuzione del carattere non è normale→ la media campionaria ha approssimativamente distribuzione normale. Questo fatto è una conseguenza del teorema del limite centrale.

Il teorema del limite centrale è un risultato asintotico, cioè indica quello che accade quando n tende all'infinito. Questo teorema dice che se sei interessato alla distribuzione della media campionaria e disponi di un campione abbastanza ampio, non ti preoccupare di qual ò la distribuzione del carattere nella popolazione perché ciò è irrilevante_ infatti, qualunque essa sia, la media campionaria ha distribuzione approssimativamente normale.

Al crescere della dimensione campionaria n l'approssimazione diventa sempre migliore. Quanto dev'essere la dimensione campionaria n affinché l'approssimazione sia buona? Nella maggior

parte dei casi è sufficiente un campione di ampiezza 25/30, ma in alcuni casi favorevoli possono bastare anche 5 elementi.

Standardizzazione della media campionaria

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Se si usa la correzione per popolazioni finite allora

$$Z = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Per studiare il comportamento della media campionaria nell'insieme dei possibili campioni è necessario disporre di media e deviazione standard della popolazione.

Di solito tali valori non sono noti e bisogna ragionare in modo deduttivo. In molti casi la distribuzione della media è (almeno approssimativamente) normale.

- Intervallo di accettazione al 95%--> intervallo centrato sulla media della popolazione che contiene il 95% delle medie campionarie. Per passare dalla scala standard alla scala originale delle medie si effettua la trasformazione→ media di x+ errore standard*Z
- Intervallo di accettazione al livello (1-α) %--> è un intervallo centrato sulla media della popolazione che contiene (1-α) % delle medie campionarie, se si estrae un campione casuale, si ha una probabilità di (1-α) che la media del campione sia compresa in tale intervallo

Distribuzione campionaria della proporzione

- In molte applicazioni il carattere di interesse è qualitativo con sue modalità, si dice anche che i dati sono binari o dicotomici.
- In tal caso la distribuzione del carattere nella popolazione è necessariamente bernoulli (successo/insuccesso)
- L'unico parametro è p= probabilità di successo
- Popolazione finita→ p = proporzione di successi
- Lo stimatore naturale della proporzione nella popolazione, p, è il corrispondente nel campione, cioè la proporzione campionaria P= $\frac{X}{n}$→numero di casi che presentano la caratteristica di interesse/numero di prove (ampiezza campionaria)
- La proporzione campionaria è un tipo di media campionaria→ valgono tutte le proprietà viste in generale per la media campionaria. Media=P
- Deviazione standard = $\sqrt{\frac{p(1-p)}{n}}$
    - La deviazione standard bernoulli è limitata superiormente
    - Quindi l'errore standard della proporzione campionaria è limitato superiormente
- Varianza= deviazione standard^2

Approssimazione della normale

- La proporzione campionaria ha una distribuzione le cui probabilità si calcolano facilmente da quelle binomiale
- Quando l'ampiezza campionaria n è grande il calcolo delle probabilità binomiali è complesso
- Tuttavia, quando n è grande, per il teorema del limite centrale la distribuzione della proporzione campionaria è ben approssimata dalla normale.
- Criterio per giudicare se la distribuzione della proporzione campionaria è ben approssimata dalla normale→ n * P (1-P) > 9. n è l'ampiezza del campione, mentre P(1-P) è il grado di simmetria della bernoulli.

La proporzione campionaria standardizzata è:

$$Z_{\hat{P}} = \frac{\hat{P} - \text{valore atteso}}{\text{errore standard}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

campioni affetti da distorsione da selezione→ un campione è distorto quando la probabilità di inclusione nel campione di individui appartenenti alla popolazione dipende dalle caratteristiche della popolazione oggetto di studio

un campione distorto fornisce una stima falsata delle caratteristiche della popolazione oggetto dell'inferenza. In linea di massima meglio un campione piccolo casuale che un grande campione distorto. La dimensione campionaria n determina l'errore campionaria, misurato dall'errore standard (deviazione standard/ $\sqrt{n}$) → campione più grande= stima più precisa.

Se il campione è distorto una dimensione maggiore non riduce la distorsione.