Statistica
Prof. Giommi

01/10/19

Introduzione al campionamento da popolazioni finite:

Indagine → strumento tramite cui si acquisiscono informazioni su fenomeni attinenti a una popolazione. Può essere di 4 tipi:

1) Completa (censimento) → quando vengono osservate tutte le unità componenti la popolazione (n = N); è impossibile se la popolazione è infinita o se l'osservazione è distruttiva. E' semplice sul piano teorico ma complessa in pratica, perché se N è numerosa a livello economico è costosa e occorre molto tempo;

2) Parziale (= indagine campionaria) → indaga solo una parte della popolazione ($n \leq N$). E' più facile da mettere in pratica del censimento perché ha costi limitati, tempi ridotti, permette di raccogliere un elevato numero di informazioni e le rilevazioni risultano essere molto accurate. In questa indagine vi sono due problemi teorici: il modo in cui deve essere scelto il campione e i procedimenti da usare per estendere l'evidenza campionaria alla popolazione;

3) Sperimentale (es. in medicina) → è lo studio dell'effetto sui soggetti di indagine dei diversi valori di una variabile tenendo per quanto possibile sotto controllo altre variabili rilevanti;

4) Osservazionale → è l'osservazione di una o più variabili di studio sui soggetti di indagine senza possibilità di controllo sperimentale sugli stessi.

Aspetti quantitativi di un'indagine:
-Dimensione della popolazione = N
-Dimensione del campione = n

L'indagine è costituita da un insieme di fasi interrelate relative alla selezione del campione e alla stima dei parametri, complessivamente definite piano di indagine, costituito da:
-Definizione della popolazione obiettivo (quella oggetto di studio);
-Scelta dei caratteri (= variabili) da studiare, modo di definirli e osservarli;
-Scelta di livelli spaziali e temporali di indagine;
-Definizione del metodo di raccolta, codifica, elaborazione dei dati;
-Definizione dei costi, dei livelli di precisione e di accuratezza desiderati;
-Stime e altre analisi statistiche;
-Metodologie di calcolo degli errori campionari;
-Metodi di controllo e rilevazione e correzione degli errori non campionari;
-Presentazione e diffusione dei risultati.

Quando si fanno indagini nella realtà non si ha un unico obiettivo: sono tante le variabili che studiamo, sono tanti i parametri che vogliamo stimare.

Popolazione (N):

Insieme finito o non finito di unità che non interessano prese singolarmente ma per il contributo che danno all'insieme di appartenenza; le caratteristiche della popolazione sono dette "*parametri*".
Esiste anche una tipologia particolare di popolazione, nota come "**popolazione obiettivo**", che riguarda la popolazione che voglio indagare; è costituita da elementi componenti, estensione spaziale ed estensione temporale, tutti caratteri che posso definire con esattezza.

Campione (n):

Qualsiasi sottoinsieme della popolazione.
I campioni si distinguono in probabilistici e non probabilistici in base a specifiche caratteristiche:
Campioni probabilistici (sono inclusi quelli con e senza reimmissione):

1) Deve essere possibile enumerare tutti i campioni estraibili (almeno a livello teorico, a livello pratico è impossibile farlo con tutti); per avere tutti i possibili campioni si applica la formula: $2^N - 1$ → comprende anche un sottoinsieme vuoto, ovvero che non ha neanche un campione, per questo si sottrae uno.
2) Deve essere possibile **assegnare una probabilità** ad ogni campione; la probabilità è un valore che varia **tra 0** (sicuramente non verrà selezionato) **e 1** (certamente sarà selezionato), quindi deve essere possibile assegnare a ogni unità della popolazione una probabilità strettamente positiva. Non ci possono essere singole unità con probabilità zero.
3) Deve esserci la possibilità di disporre di un meccanismo pratico per selezionare con le probabilità assegnate → la procedura consente di selezionare un campione con probabilità corrispondente a quella che gli era stata associata a priori.
Altre classificazioni dei campioni probabilistici:
-Campioni con replicazione (reimmissione, ripetizione) → una stessa unità statistica si può presentare più di una volta.
-Campioni senza replicazione.
Campioni non probabilistici: quelli che non rispettano le precedenti condizioni, ovvero:
1) Campioni di volontari o formati in modo spontaneo → es. sondaggio in una trasmissione televisiva, risponde chi vuole rispondere spontaneamente;
2) Campioni formati in modo fortuito → assomigliano ai campioni casuali → vengono presi in maniera casuale (es. riso preso da una risai per avere dei campioni);
3) Campioni a scelta ragionata: si suddividono in:
-Campioni per quota → vengono formati tenendo conto che nella popolazione ci sono categorie di persone che possono essere importanti da inserire → es. campione rappresentativo degli abitanti del comune di Firenze: se c'è il 50% degli uomini e delle donne il campione deve rispettare questa proporzione, così come se c'è una maggioranza superiore a un'età. Il campione per quota considera le proporzioni (chiamate quote) che sono presenti nella popolazione e cerca di rispettarle.
-Campioni di unità tipo;
-Aree barometro → aree territoriali che si avvicinano alla media delle cose che si vogliono studiare. Es. elezioni politiche → c'è una proporzione di voti che vanno a ogni formazione: ci sono aree territoriali dove la proporzione che riceve ciascuna compagine è identica o molto vicina a quelle generali. Sono dette aree barometro perché se cambia in quelle zone è probabile che cambi anche il risultato finale;
-Campioni a valanga → vengono formati quando vogliamo raccogliere unità statistiche che sono rare nel contesto della popolazione (es. le attività che svolgono; si cercano unità caratteristiche e si chiede se ne conoscono altre che svolgono la stessa cosa e così via).
I campioni non probabilistici sono fondamentali nella produzione delle statistiche ufficiali (Istat) e sono i campioni prevalenti nell'ambito del SISTAN.

Il piano di indagine (o di campionamento):

Attualmente nella teoria del campionamento da popolazioni finite si definisce:
-Schema di campionamento (schema di selezione) → l'insieme delle regole da seguire nella formazione del campione;
-Piano di campionamento → la distribuzione di probabilità sull'universo dei possibili campioni S. Questa distribuzione è indicata come p(s) nella quale s è un generico campione appartenente all'insieme S. Questa distribuzione è compresa tra 0 e 1 (inclusi) e la somma di tutte le distribuzioni dà come risultato 1.

03/10/19

Esempi di campioni probabilistici:
1) Rilevazione continua delle forze di lavoro (Istat);
2) L'indagine sui consumi delle famiglie (Istat);
3) L'indagine multiscopo (Istat) → salute, aspetti della vita quotidiana, sicurezza, uso del tempo libero; di anno in anno può cambiare il suo obiettivo.

Come si forma un campione probabilistico (cioè casuale)?
1) Schema dell'urna (teorico);
2) Tavole dei numeri casuali (la storia);
3) Algoritmi matematici tradotti in routine informatiche.
N.B: Si può dire che un **campione è rappresentativo quando è casuale**.
Rappresentativo è sinonimo di probabilistico (cioè casuale).

Stima:
La stima è il procedimento mediante il quale **un valore (stima)** ricavato come funzione (stimatore) delle osservazioni campionarie, **viene assunto a rappresentare il valore incognito di una grandezza** caratteristica (parametro) della popolazione obiettivo.
Gli indici di precisione nella stima sono:
-**Varianza (V($\bar{Y}$))** → V($\bar{Y}$) = media degli scarti quadratici delle possibili stime dal vero valore da stimare;
-**Errore standard della stima (ES($\bar{Y}$))** = $\sqrt{V(\bar{Y})}$
N.B.= V($\bar{Y}$) e ES($\bar{Y}$) non possono essere calcolati nella pratica perché dipendono dai valori incogniti della popolazione, però possiamo stimare la varianza:
$\mathbf{v(\bar{Y})} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$ dove $\left(1 - \frac{n}{N}\right)$ è il fattore di correzione e si omette ogni volta che N >> n

$s^2 = \frac{\Sigma(y - \bar{y})^2}{n-1}$ dove y = valori dei campioni e $\bar{y}$ = medie dei valori dei campioni

Parametri di maggior interesse nelle indagini osservazionali:
-Totali (occupati, forza lavoro);
-Medie (reddito pro-capite);
-Rapporti (tasso di disoccupazione);
} per caratteri quantitativi
-Proporzioni (tasso di occupazione/attività) → rapporto fra parametro e totale, si può riferire a qualsiasi attributo anche di natura qualitativa (es. proporzione di uomini sul totale della classe, gli uomini sono qualitativi). Si usano le proporzioni o le percentuali per i caratteri qualitativi dicotomici, per quelli non dicotomici sono oggetto di stima le distribuzioni di frequenza assolute e/o relative nella popolazione.
-Indici di variabilità (strumentali per la precisione degli stimatori).

Proprietà degli stimatori:
Proprietà: che siano più prossimi possibile al valore numerico del parametro (incognito) o coincidesse con il parametro; normalmente non coinciderà con il valore del parametro che vogliamo stimare perché è stata effettuata su un campione. In qualsiasi indagine campionaria i parametri che stimiamo sono sempre incogniti eccetto un unico caso, quello delle elezioni.
Questa proprietà viene espressa come differenza tra stimatore ed errore di stima: consiste quindi nella possibilità di ridurre ai minimi termini l'errore di stima.

T → Stimatore
t → Stima
θ (Theta) → Parametro (incognito)
D (differenza) = **t – θ → errore di stima** (o errore di campionamento)
D = 0 → situazione ideale. D non può mai essere zero nell'indagine campionaria, eccetto nei censimenti; avrà sempre un valore, che sia positivo o negativo (se è inferiore a θ). Per noi è importante che sia più piccolo possibile. Per ridurlo nell'indagine occorre ridurre la dimensione campionaria, oppure considerare un piano di campionamento che possa garantirci una riduzione di D. Gli errori di stima si identificano attraverso gli indici di variabilità.

Precisione dello stimatore:
-La qualità dello stimatore è data dalla sua precisione: maggiore è la varianza (errore standard), minore è la precisione e viceversa;
-La precisione è data dall'**errore standard** dello stimatore, che corrisponde al reciproco della varianza dello stimatore o della radice quadrata della varianza.

Possibili piani di campionamento:
1) Campionamento casule semplice (CCS) (senza reimmissione)
-Nessuna media campionaria è uguale alla media della popolazione;
-La media di tutte le medie dei campioni (ovvero la media campionaria) è lo stimatore corretto. Lo stimatore corretto non garantisce che la stima del campione osservato sia prossima al valore da stimare; sarebbe importante che non ci fossero campioni che producono stime distanti dal vero valore da stimare. Proprietà della media campionaria → media campionaria = media incognita della popolazione.
-La qualità dello stimatore è data dalla sua precisione;
-La dimensione dei campioni è determinata dalla casualità dell'estrazione;
-Nella pratica l'estrazione avviene sempre senza reimmissione.
Es. 4 unità (U1, U2, U3, U4) N = 4, n = 2 (campioni)
Possibili campioni {s}: (U1 U2) (U1 U3) (U1 U4) (U2 U3) (U2 U4) (U3 U4) Non conta niente l'ordine.
Numero campioni (cardinalità S): $C_{4,2} = 6$ → CCS = p(s) = 1/6
Se da una popolazione N si vuole estrarre un campione n, questo piano attribuisce la stessa probabilità di selezione a ciascun campione. Occorre stimare la media incognita della popolazione calcolando la media sul campione; il singolo campione può produrre una stima anche abbastanza diversa dal vero valore da stimare (infatti nessuna media campionaria è uguale alla media della popolazione).
Stima di una proporzione (P, p): una proporzione è una parte della popolazione avente uno specifico attributo che vogliamo indagare; è considerata come una media calcolata su un carattere, e può assumere solo due valori:
1 = possesso dell'attributo
0 = mancanza dell'attributo
P è la proporzione di unità che hanno un certo attributo nella popolazione ed è uguale alla media del carattere ($P = \bar{Y}$); stessa cosa vale per p, che rappresenta la proporzione campionaria ($p = \bar{y}$).
2) Campionamento casuale stratificato
-La popolazione è suddivisa in sottopopolazioni (**strati**) → le sottopopolazioni rappresentano una partizione della popolazione (sono individuate in modo che ciascuna unità della popolazione appartenga soltanto ad una sottopopolazione. Es. presenti nell'aula tra uomini e donne, o uno è uomo o è donna);

-Estrazione di campioni casuali **indipendenti** da ogni strato → da ciascuna sottopopolazione cioè da ciascuno strato anziché selezionarli dai campioni. L'indipendenza ci consente delle semplificazioni quando parliamo di varianze. Il campione complessivo è il risultante della somma di tutti i campioni selezionati in ciascuno strato ($n = n_1 + n_2 + \ldots + n_k$).

Obiettivi:

-Ottenere stimatori con elevata precisione, ovvero ottenere una varianza della media più piccola;

-Avere stime che riguardano i domini di studio (voglio sapere la stima della media di un voto di un esame per uomini e poi per le donne separatamente).

Indice h per indicare la numerazione

H = numero di sottopopolazioni che facciamo

h = n° di strati

N = ΣNh

n = Σnh

ΣWh = 1

Wh = Nh /N = Proporzione della popolazione nello strato h → dimensione strato (Nh) / dimensione complessiva della relazione (N).

Stratificazione proporzionale:

Si ha quando la dimensione dei campioni è formata applicando in tutti gli strati la stessa frazione di campionamento, per cui essa risulta proporzionale a quella dello strato di provenienza.

Frazione di campionamento costante (o frazione stratificata proporzionale):

fh = nh/Nh = n/N = f (h = 1….H) → è costante in ogni strato.

Vantaggi:

1) Per la popolazione generale, stime (stime medie complessive) più precise rispetto al CCS. Vuol dire che hanno un errore standard più basso a parità di dimensione campionaria;

2) Facilità di applicazione (con un numero limitato di strati).

Svantaggi:

1) Precisione diversa per strati di diversa dimensione → a numerosità maggiore corrisponde una precisione migliore;

2) Difficoltà di applicazione con molti strati.

La stratificazione proporzionale è buona per la stima media della popolazione; se voglio le stime strato per strato ha dei grossi difetti perché ha campioni grandi negli strati grandi e piccoli in quelli piccoli, e quest'ultimi potrebbero darmi delle stime poco precise. Nel caso in cui si ipotizzi che da uno strato all'altro la variabilità sia uguale, si può prendere in ogni strato la stessa identica dimensione. In questo campionamento c'è sempre un vantaggio a dividere la popolazione in strati e fare un campione proporzionale a meno che in tutti gli strati che abbiamo fatto non ci sia la stessa media.

Quando il campionamento è stratificato proporzionale:

Wh = Nh/N

-**Stimatore della media** (è uno stimatore corretto): $\bar{y}_{stp} = \Sigma W_h \bar{y}_h$

-**Varianza dello stimatore**: $V(\bar{y}_{stp}) = \frac{(1-f)}{n} \Sigma W_h {S^2}_h = \frac{(1-f)}{n} {S^2}_w$

con ${S^2}_w$ = media ponderata della varianza di strato.

-**Stimatore campionario della varianza**: $V(\bar{y}_{stp}) = \frac{(1-f)}{n} \Sigma W_h {s^2}_h$

Ricordando che:
$\bar{Y}_h$ = vera media nello strato h;
$\bar{y}_h$ = media campionaria nello strato h;
$s^2{}_h$ = varianza elementare (per il carattere Y) nello strato h.
A parità di dimensione campionaria la varianza della media campionaria nella stratificazione proporzionale non è mai superiore a quella nel CCS.
A parità di dimensione campionaria e a parità di costi la stratificazione proporzionale ci consente di migliorare le stime che avremmo utilizzando un CCS.
Se vogliamo ottenere l'efficienza degli stimatori, gli strati devono essere il più possibile omogenei al loro interno. Più numerose sono le variabili, più è difficile che la stratificazione sia efficace per ciascuna di esse; non è conveniente formare un alto numero di strati (nonostante questo aumenti l'omogeneità interna a ciascuno strato) perché diminuisce la dimensione campionaria in ciascuno strato, ma aumenta la variabilità delle stime.
Per ottenere un campionamento stratificato occorre:
1) Conoscere la proporzione di popolazione (W_h) negli strati che si vogliono formare;
2) Ogni unità della popolazione deve essere attribuibile soltanto ad uno strato;
3) Se si usano stimatori corretti, occorre selezionare almeno una unità da ogni strato. Occorre invece selezionare almeno 2 unità se si vuole stimare correttamente da campione la varianza degli stimatori.

Stratificazione ottimale:

Più è alto il costo, minore è la frazione di campionamento → $f_h \propto \frac{S_h}{\sqrt{c_h}}$ dove:

S_h = deviazione standard; se non si approssima adeguatamente si perde precisione rispetto al CCS;
c_h = costo di osservazione di una unità nello strato h.
Negli strati più piccoli per avere sufficiente precisione degli stimatori occorre applicare una frazione di campionamento maggiore.

3) Campionamento sistematico:
Richiede di estrarre casualmente solo la 1^ unità del campione, che è formato prendendo un'unità ogni k presenti nella lista (cioè in N) a partire dalla prima estratta. In questo campionamento come nel CCS, ogni unità della popolazione ha la stessa probabilità di entrare a far parte del campione.
Se prima si applica l'intervallo di selezione il campionamento sistematico può essere assimilato al CCS.
Intervallo di selezione → **k = N/n** è il reciproco della frazione di campionamento.
Es. N = 1500 n = 100 k = 1500/1000 = 15
Per formare il campione scelgo un numero casuale compreso tra 1 e 15, che mi permette di individuare la 1^ unità estratta, le altre vengono selezionate ogni 15 fino all'esaurimento della lista.
Supponiamo di aver estratto 6; campione sistematico:
6, 6+15, 6+2x15…..6+99x15 → 6, 21, 36…..1491.

4) Campionamento a grappoli e a più stadi
-**Grappoli** → campionamento di aggregati (unità di rilevazione); tutte le unità componenti l'aggregato (unità di studio) entrano a far parte del campione. I grappoli sono aggregazioni preesistenti nella popolazione e sono eterogenei al loro interno: l'omogeneità aumenta quando diminuisce la loro dimensione ed è l'ideale che siano eterogenei al loro interno perché in questo modo risultano più simili tra loro, dato che mettendo nel campione solo alcuni di essi perdiamo un po' di variabilità. E' il contrario di quello che avviene con la stratificazione perché in questo caso devono essere omogenei al loro interno ed eterogenei tra loro.
Es. nelle indagini Istat sulle popolazioni l'aggregato è la famiglia (che può anche essere composta

da una sola persona).
-**Stadi** → campionamento di grappoli da grappoli di livello gerarchico superiore.
Es. Quando un aggregato ha al suo interno altri aggregati → indagine regionale, considero i comuni e poi dai comuni considero le famiglie.
Problemi che ci impediscono di selezionare direttamente l'unità che vogliamo studiare:
1) Indisponibilità della lista delle unità di studio;
2) Costi.
N.B: Gli **strati** sono **omogenei** al loro interno, gli **stadi** sono **eterogenei** al loro interno; più eterogenei sono, meno è la probabilità che ci sia variabilità tra uno stadio e l'altro; è indifferente quale stadio vado a selezionare.
I grappoli in genere sono precostituiti: si trovano così già in natura.

Confronto: Grappoli vs CCS
-Nella pratica operativa: grappoli omogenei al loro interno (es: famiglie, scuole, classi scolastiche, ecc.);
-Grappoli: la dimensione del campione risulta variabile in termini di unità di studio;
-Grappoli vs CCS: a parità di dimensione il CCS fornisce stime più precise, ma ad un costo ben superiore;
-A parità di costi (risorse) il campionamento a grappoli può avere dimensione mediamente maggiore e conseguentemente produrre stimatori più precisi rispetto al CCS.

Nomenclatura:
Y → carattere di studio nella popolazione;
$\bar{Y}$ → media del carattere Y nella popolazione;
$\bar{y}$ → stima campionaria di $\bar{Y}$
Valori nella popolazione.
$Y_1 = 20$, $Y_2 = 40$, $Y_3=36$, $Y_4=48$; $Y= 36$
A ciascuna media corrisponde un errore di stima.
Nessuna media campionaria è uguale alla media della popolazione; medie campionarie = 30, 28, 34, 38, 44, 42.
Frazione di campionamento → n/N → rapporto tra la dimensione del campione con quello della popolazione; è indicato anche come **f**. Rappresenta inoltre la probabilità che ciascuna unità della popolazione ha di entrare a far parte del campione (probabilità di essere inclusa nel campione = probabilità di inclusione).

04/10/19

Distribuzione campionaria completa:
Dai valori della popolazione tiro fuori le medie da ciascun campione estraibile dalla popolazione, ma nella realtà non conosco la popolazione.
-**Varianza elementare** → varianza degli elementi che compongono la popolazione.
Se N è molto grande N – 1 è irrilevante;
-**Varianza dello stimatore** → Nella pratica la vera varianza dello stimatore non può essere calcolata, dato che non è nota la varianza elementare $\mathbf{S^2}$, tuttavia è possibile stimarla dal campione inserendo nella sua formula la stima da campione della varianza elementare. E' una varianza della somma di variabili;
-**Vera varianza dello stimatore** → quella che praticamente non siamo in grado di calcolare.

Stima di una proporzione:

La proporzione (π) di unità che hanno un particolare attributo nella popolazione, può essere vista come la media di un carattere dicotomico (attributo si = 1, attributo no = 0). Di conseguenza la proporzione campionaria che indichiamo con p o π, ha le stesse proprietà della media campionaria (è la proprietà di una media campionaria che assume solo due possibili valori: 1,0). In particolare la proporzione campionaria è uno stimatore corretto della proporzione nella popolazione.

Proprietà formali dello stimatore media nel CCS:

Correttezza = $E(\bar{y}) = \mu$ con μ = media del carattere Y nella popolazione.
Efficienza = $V(\bar{y}) \leq V(T)$ con T = costante diversa dalla media aritmetica (es. mediana del campione)
Consistenza = $V(\bar{y}) \rightarrow 0$ al crescere della dimensione campionaria.

Proprietà della media:

-Internalità → quella più importante;
-Proprietà associatività della media → se prendo un gruppo, lo divido in sottogruppi e calcolo la media di ogni sottogruppo, la media complessiva la posso ricavare dalle medie dei sottogruppi purché faccia una media ponderata dalle medie dei gruppi;
-Proprietà della correttezza → la media di tutte le medie corrisponde alla media nella popolazione.

15/10/19

Intervalli di confidenza:

Sono intervalli di numeri concentrati sulla stima puntuale che con un fissato livello di probabilità contiene il vero valore del parametro. La probabilità che l'intervallo di confidenza contenga il vero valore del parametro è detto livello di confidenza o fiducia. In genere si scelgono livelli di confidenza **prossimi a 1** (0.9, 0.95, 0.99), ma non 1 perché altrimenti abbiamo un intervallo così ampio che non ci è di nessuna utilità. Più alto è il livello di confidenza più ampia è la probabilità e più alto è l'intervallo. Ha tanto più significato quanto più è piccolo, ma più è piccolo meno è la probabilità che gli si può assegnare; dà un'alta precisione della stima, ma ha un livello di confidenza ridotto quindi aumenta la probabilità che il vero parametro non ci sia.
Consideriamo un intervallo di stime al quale sia associato un livello di confidenza; voglio associarci una probabilità che questo intervallo contenga un vero valore incognito. Per costruire una stima per intervallo si seguono delle regole fisse.
Scopo: determinare un intervallo numerico (che sarà intorno alla stima puntuale, la quale sarà il punto centrale di questo intervallo) e che ci aspettiamo contenga con un certo livello di fiducia, il valore del parametro.
Fissato livello di probabilità → siamo sempre in una situazione di incertezza. Ho fiducia che… per un tot, tuttavia c'è sempre la possibilità che l'intervallo non la contenga. La stima puntuale utilizza le osservazioni di un campione casuale per ottenere una stima del parametro tramite un singolo valore numerico. Tale approccio possiede un punto di debolezza: la stima ottenuta sul campione osservato potrebbe differire molto dal valore del parametro nella popolazione. E' dunque opportuno che l'inferenza su un certo parametro si basi non solo sulla stima puntuale, ma dia informazioni anche su quanto precisa sia la stima, ossia su quanto è probabile che sia vicina al vero valore del parametro. A tal fine si considera oltre alla stima puntuale, un intervallo di stime plausibili al quale sia associato un fissato livello di confidenza.

Intervallo di confidenza per un parametro:

Obiettivo: determinare due statistiche campionarie:
$L_I = L_I(Y_1,...,Yn)$ e $L_S = L_S(Y1,...,Yn)$;
-$L_I \leq L_S$ per ogni possibile campione;
-L'intervallo $[L_I;L_S]$ contiene il parametro θ con probabilità $1-\alpha$ → $P(L_I \leq \theta \leq L_S) = 1-\alpha$;
-$1-\alpha$ è detto livello di fiducia o livello di confidenza;
-Una volta estratto il campione si ottiene l'intervallo di confidenza stimato;
-Non è possibile sapere se l'intervallo stimato contenga o meno il valore vero del parametro;
-La chiave per costruire un intervallo di confidenza è la distribuzione campionaria dello stimatore utilizzato per ottenere la stima puntuale;
-La distribuzione campionaria dello stimatore permette di determinare la probabilità che lo stimatore produca una stima che cade entro una certa distanza dal parametro.

Statistica campionaria:

E' il frutto dell'elaborazione dei valori del campione.
Li → funzione dei valori del campione.
Variabili aleatorie → lettera maiuscola
variabili osservate → lettera minuscola
Se la distribuzione campionaria dello stimatore è normale (anche approssimativamente), allora:
-Con probabilità di circa il 95% lo stimatore produrrà una stima del parametro che ricade a 2 errori standard dal parametro;
-Con probabilità di circa il 99.7% lo stimatore produrrà una stima del parametro che ricade a 3 errori standard dal parametro;
-Minore è l'errore standard, maggiore è la precisione dello stimatore.
Un intervallo di confidenza si può dunque costruire aggiungendo e sottraendo dalla stima puntuale un multiplo dell'errore standard dello stimatore, detto "***margine di errore***" e dipende dalla variabilità (errore standard) della distribuzione campionaria dello stimatore.
Forma tipica degli intervalli di confidenza: **stima puntuale ± margine di errore**

Intervallo di confidenza per una proporzione: campioni di dimensione elevata

I campioni per noi sono sempre di dimensione elevata; supponendo che la variabile di interesse Y sia binaria, la distribuzione di Y nella popolazione è Bernoulliana con probabilità di successo π: $Y \sim Bernoulli(\pi)$ e π è il parametro di interesse.

Stimatore puntuale di π:

Proporzione di successi campionaria = media campionaria

$\hat{\pi} = \bar{Y} = \frac{y_{1+\cdots+}\, y_n}{n}$ Si ricorda: $E(\hat{\pi}) = \mu_{\hat{\pi}} = \pi$ e $V(\hat{\pi}) = \sigma^2_{\hat{\pi}} = \frac{\pi\,(1-\pi)}{n}$

Errore standard della proporzione campionaria:

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi\,(1-\pi)}{n}}$$

La proporzione campionaria π è una media campionaria, quindi per campioni di dimensioni sufficientemente elevate la sua distribuzione campionaria si può approssimare con una distribuzione normale per il teorema del limite centrale. Formalmente, per il teorema del limite centrale, se n è sufficientemente grande:

$\hat{\pi} \approx N\left(\pi, \frac{\pi\,(1-\pi)}{n}\right)$ e quindi $\frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi\,(1-\pi)}{n}}} \approx N\,(0,1)$

Distribuzione di Bernoulli:

Variabile binaria → valore 1 con probabilità di successo π o 0, con probabilità di successo 1 – π.
Per stimare una proporzione (es. quanti uomini e donne ci sono in aula) prendo un campione di 10 studenti, trovo 4 uomini (0.4) e 6 donne (0.6): 0.4 e 0.6 sono le stime puntuali, sono anche le medie delle variabili se dò valore 1 a donna e 0 a uomo.
Nella distribuzione normale il 95% delle osservazioni è compreso entro 1.96 deviazioni standard dalla media. Utilizzando l'approssimazione normale si ha che la proporzione campionaria, come stimatore, assumerà valori che si trovano entro $1.96 \cdot \sigma\pi$ unità dal parametro π con probabilità pari a 0.95, dove 1.96 è il valore che nella normale standard lascia alla sua destra un'area pari a 0.025.
Fissato $1-\alpha$, utilizzando l'approssimazione normale si ha che la proporzione campionaria, come stimatore, assumerà valori che si trovano entro $z_{\alpha/2} \cdot \sigma\pi$ unità dal parametro π con probabilità $1-\alpha$, dove $z_{\alpha/2}$ è il valore che nella normale standard lascia alla sua destra un area pari a $\alpha/2$.
Una volta osservato il campione, $y_1,...,y_n$, si ha un solo valore dello stimatore $\pi = (y_1+...+y_n)/n$ e non è noto se tale valore si trova entro $z_{\alpha/2} \cdot \sigma\pi$ unità da π.
Se π si trova entro $z_{\alpha/2} \cdot \sigma\pi$ unità da π allora l'intervallo di estremi $\pi \pm z_{\alpha/2} \cdot \sigma\pi$ contiene π, altrimenti tale intervallo non contiene π.

Errore standard nella proporzione campionaria: $\boldsymbol{\sigma\pi = \sqrt{(\pi \cdot (1-\pi))/n}}$
Non è noto perché dipende dal parametro π ignoto che interessa stimare.
Intervallo di confidenza al livello di confidenza $1-\alpha$ per la proporzione:

$$\mathbf{IC_{1-\alpha}(\pi)} = \left[\hat{\pi} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \; ; \; \hat{\pi} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Dove $z_{\alpha/2}$ è il valore che nella normale standard lascia alla sua destra un'area pari a $\alpha/2$:
$P(Z > z_{\alpha/2}) = \alpha/2$.
Con un livello di confidenza pari a $1-\alpha$ si ha una probabilità pari a α che il metodo produca un IC che non contiene il vero valore del parametro.
Con un livello di confidenza pari a $1-\alpha = 0.95$ si ha una probabilità pari a $\alpha = 0.05$ che il metodo produca un IC che non contiene il vero valore del parametro.
Ampiezza dell'intervallo di confidenza di livello di confidenza $1-\alpha$:

$$\text{Estremo superiore – estremo inferiore} = \hat{\pi} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} - \left(\hat{\pi} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

$$\text{Ossia} \rightarrow \text{ampiezza} = 2 \text{ Margine di errore} = 2z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Il M.E. è uguale al prodotto tra il valore $z_{\alpha/2}$ e l'errore standard.
-Maggiore è il livello di confidenza, maggiore sarà la possibilità che l'intervallo di confidenza contenga il vero valore del parametro (perché l'intervallo è più ampio, è maggiore il M.E.);
-Maggiore è il livello di confidenza, maggiore è l'ampiezza dell'intervallo di confidenza (ossia maggiore è il margine di errore) e quindi minore la precisione (accuratezza) della stima;
-Con un livello di confidenza $1-\alpha = 1$, l'intervallo di confidenza per la proporzione sarebbe [0,1], che non è di alcun aiuto perché include tutti i possibili valori per π;
-La scelta del livello di confidenza è il risultato di un compromesso tra il desiderio che l'inferenza sia corretta e la precisione della stima: al migliorare di un aspetto l'altro peggiora e viceversa;
-Valori tipici del livello di confidenza: $1-\alpha = 0.90, 0.95, 0.99$;
-Il margine di errore è inversamente proporzionale alla √n;
-L'errore standard è massimo per $\pi = 0.5$;

-Massimo valore del margine di errore per $1-\alpha$ fissato
$z_{\alpha/2} \cdot \sqrt{(0.5 \cdot 0.5)/n} = z_{\alpha/2} \cdot 1/\sqrt{(4 \cdot n)} = z_{\alpha/2} \cdot 1/(2 \cdot \sqrt{n})$
-La dimensione del campione deve essere quadruplicata per ottenere un intervallo di confidenza di ampiezza dimezzata (ossia per ottenere una doppia precisione).
In sintesi:
-L'ampiezza dell'intervallo di confidenza per la proporzione cresce al crescere del livello di confidenza;
-L'ampiezza dell'IC per la proporzione decresce al crescere della dimensione del campione;
-Queste proprietà valgono per tutti gli intervalli di confidenza.

Interpretazione del livello di confidenza:
-Il livello di confidenza ci dice come si comporta il metodo, utilizzato per la costruzione dell'IC, quando venga applicato ripetutamente a differenti campioni casuali;
-Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un IC al livello di confidenza $1-\alpha$ allora circa il $(1-\alpha)\%$ degli intervalli conterrebbe π;
-Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un IC al livello di confidenza $1-\alpha = 0.90$ allora circa il $(1-\alpha)\% = 10\%$ degli intervalli conterrebbe π.

17/10/19

Intervallo di confidenza per una proporzione:
1) In pratica viene selezionato un solo campione di dimensione prestabilita e si costruisce un unico IC utilizzando le osservazioni dell'unico campione selezionato;
2) Non si può sapere se l'intervallo di confidenza contenga o meno il vero valore del parametro, π;
3) Il livello di confidenza è una quantità che è relativa alle proprietà del metodo utilizzato per costruire l'IC.
Importanza di avere campioni di dimensione elevata:
-La probabilità che l'intervallo di confidenza contenga il vero valore del parametro, π, è approssimativamente uguale al livello di confidenza: l'approssimazione migliora con campioni di grandi dimensioni;
-Per il teorema del limite centrale, per n sufficientemente grande, la distribuzione campionaria della proporzione campionaria è approssimativamente normale;
-L'approssimazione normale è in generale adeguata se si hanno almeno 15 osservazioni per categoria;
-Al crescere della dimensione del campione, l'errore standard stimato della proporzione campionaria tende a assumere valori prossimi al vero errore standard.

Dimensione del campione e accuratezza della stima:
Campioni di dimensioni maggiori danno intervalli più stretti quindi minor margine di errore ($z_{\alpha/2}$ per l'errore standard) quindi maggior precisione a parità di livello di confidenza.
Qaudruplicando la dimensione, dimezzo il M.E. e quindi l'intervallo.

Intervallo di confidenza per la media:
Supponiamo che il carattere di interesse Y sia quantitativo con media μ nella popolazione; l'IC per la media ha la forma: **stima puntuale ± margine di errore** con il M.E. multiplo dell'errore standard dello stimatore.
Si distinguono tre casi:
1) Intervallo di confidenza per la media di una popolazione normale con varianza nota;

2) Intervallo di confidenza per la media di una popolazione normale con varianza non nota →
E' la situazione pratica più comune: prendiamo il campione, troviamo la media campionaria e la stima, poi stimiamo la varianza. La varianza della media la scriviamo con s quadro anziché δ^2 perché non la conosco, la standardizzazione non è più una distribuzione normale ma una distribuzione t.
L'intervallo di confidenza per la media a livello di confidenza 1−α è dato da:

$$\left[\bar{Y} - t_{(n-1),\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \; ; \; \bar{Y} + t_{(n-1),\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$

dove $t_{(n-1),\alpha/2}$ è il valore che nella distribuzione t−Student con n −1 gdl lascia alla sua destra un'area pari a α/2: $P(T_{n-1} > t_{(n-1),\alpha/2}) = \alpha/2$
M.E = t· Errore standard = t· s/√n
Ampiezza dell'intervallo di confidenza al livello di confidenza 1−α:
Estremo superiore - Estremo inferiore
$(Y + t_{(n-1),\alpha/2}\cdot S/\sqrt{n}) - (Y - t_{(n-1),\alpha/2}\cdot S/\sqrt{n}) = 2\cdot t_{(n-1),\alpha/2}\cdot S/\sqrt{n}$ → ossia:
Ampiezza = 2·Margine di errore = $2\cdot t_{(n-1),\alpha/2}\cdot S/\sqrt{n}$
Fissata la dimensione del campione e fissato il livello di confidenza, al variare dei campioni estratti, la lunghezza degli intervalli corrispondenti non rimane costante poiché varia il valore di S.
3) Intervallo di confidenza per la media (μ) di una popolazione non normale per campioni di dimensione elevata → è importante la scelta della dimensione campionaria perché incide sulla precisione dei risultati inferenziali. Un criterio si basa sull'ampiezza dell'IC per il parametro di interesse: formalmente, si cerca il valore di n per il quale un IC per il parametro di interesse ha un margine di errore corrispondente a un certo valore. Gli elementi chiave che influiscono sulla determinazione dell'ampiezza campionaria sono:
-M.E. → dipende direttamente dall'errore standard della distribuzione campionaria dello stimatore puntuale;
-L'errore standard dello stimatore, il quale dipende dalla dimensione campionaria.

18/10/19

Distribuzione t di Student:
-Dipende da un parametro a valori interi positivi detto "*gradi di libertà*" (gdl);
-Ha una forma campanulare, simmetrica intorno alla media uguale a zero;
-Presenta un'ampiezza leggermente diversa per ciascun differente valore dei gdl;
-Presenta aree sulle code più grandi (più pesanti) ed è più dispersa rispetto alla distribuzione normale standard;
-La deviazione standard della distribuzione t di Student è leggermente più grande di 1 (il valore esatto dipende dai gdl);
-Quanto più elevato è il valore dei gdl tanto più la distribuzione tenderà a assomigliare a una distribuzione normale standard;
-Anche al crescere della dimensione del campione, la distribuzione t−Student diventa sempre meno dispersa e assomiglia sempre di più alla distribuzione normale → quando abbiamo campioni di dimensione troppo elevata se abbiamo **gdl > 100 ricorriamo alla tavola della normale** poiché le differenze tra le due distribuzioni sono molto piccole;
-Si ricordi che se $Y \sim N(\mu,\sigma^2)$ allora $Y \sim N(\mu,\sigma^2/n)$;
-La distribuzione t−Student viene introdotta per tener conto dell'incertezza sulla varianza σ^2, che, se non nota, deve essere stimata con S^2;
-La t−Student è più dispersa della distribuzione normale standard;
-Se la dimensione del campione n è sufficientemente grande:

$IC_{1-\alpha}(\mu) = y \pm t_{n-1,\alpha/2} \cdot s/\sqrt{n} \approx y \pm z_{\alpha/2} \cdot s/\sqrt{n}$

-Lo stimatore S^2 è uno stimatore consistente. Al crescere della dimensione del campione S^2 si avvicina sempre più al vero valore della varianza σ^2 e, di conseguenza, l'errore standard stimato della media campionaria, $s/\sqrt{n}$, approssima sempre meglio il vero errore standard, $\sigma/\sqrt{n}$.

Scelta della dimensione del campione per stimare una proporzione:

Carattere di interesse: $Y \sim Bernoulli(\pi)$

Obiettivo: Stimare la probabilità di successo π;

Al fine di determinare la dimensione del campione che garantisca una precisione della stima desiderata si deve:

1) specificare il margine di errore;

2) specificare la probabilità con la quale si ottiene quel margine di errore (cioè quella che vogliamo associare all'IC);

Esempio → Sondaggio di opinione in un certo paese: Favorevoli versus contrari all'eutanasia:

Obiettivo: determinare n tale che la proporzione dei favorevoli nella popolazione si trovi entro 0.04 punti dal vero valori con probabilità 0.95.

Si tratta di determinare la numerosità campionaria necessaria per garantire un M.E. del 4% e un livello di confidenza del 95%. In altri termini, dobbiamo determinare n in modo tale che l'IC al livello di confidenza del 95% sia pari a $\pi \pm 0.04$.

Quando stimiamo la dimensione del campione per ottenere un IC che abbia un certo M.E. prefissato, poiché il valore del multiplo che ci fa ottenere IC dipende dal campione, non posso usare la distribuzione t di Student, ma uso quella normale.

Si assuma che la distribuzione campionaria della proporzione campionaria sia ben approssimata da una distribuzione normale → allora la proporzione campionaria assumerà un valore che si trova entro $z_{\alpha/2}$ errori standard da π con probabilità $1-\alpha$.

Margine di errore = semi-lunghezza dell'intervallo di confidenza

$$\text{M.E.} = z_{\alpha/2}\sigma_{\hat{\pi}} = z_{\alpha/2}\sqrt{\frac{\pi\,(1-\pi)}{n}}$$

Quindi la proporzione campionaria assumerà un valore che si trova entro il M.E. (ossia entro $z_{\alpha/2}\sigma_{\hat{\pi}}$) da π con probabilità $1 - \alpha$.

Problema: la formula che permette di determinare la dimensione del campione n per la stima di una proporzione π dipende dal parametro che interessa stimare.

E' necessario ipotizzare un valore per π: se non si hanno informazioni sul possibile valore di π si può prendere un approccio "prudenziale". Il massimo valore dell'errore standard della proporzione campionaria si ha per $\pi = 0.5$. Si determina la dimensione campionaria necessaria per ottenere a un livello di confidenza fissato, $1-\alpha$ il margine di errore desiderato ponendo $\pi = 0.5$

$$n = z_{\alpha/2}{}^2\,\frac{0{,}5\,(1-0{,}5)}{M^2}$$

Questo approccio garantisce che al livello di confidenza $1-\alpha$ il margine di errore non sarà superiore al valore M fissato, qualunque sia il vero valore di π.

Quando abbiamo un carattere quantitativo:

Valore max – Valore min. → è compreso tra la media - 3δ e media + 3δ. E' quindi compreso in un'ampiezza di 6δ.

Es. Se stimo il n° medio di esami dati dagli studenti di Psicologia e ho un certo IC di 0,95 con un M.E. prefissato (quindi non voglio superarlo) i voti devono variare da 0 (nessun esame dato) a 21 (massimo numero); 21/6 = 3,5.