UNIVERSITÁ DEGLI STUDI DI TERAMO

SCIENZE DELLA COMUNICAZIONE

CORSO PER LAVORATORI

IL CAMPIONAMENTO

FABRIZIO ANTOLINI
fantolini@unite.it

IL CAMPIONAMENTO PROBABILISTICO E RAGIONATO

Lo scopo di un'indagine statistica è conoscere o comprendere meglio un determinato aspetto della realtà che ci circonda. Per raggiungere tale obiettivo, ci si può basare su dati già raccolti da altri oppure eseguire una nuova rilevazione statistica.

Nel caso si intenda perseguire la seguente opzione, bisogna decidere se la raccolta dei dati avverrà su tutta la popolazione (realizzando un censimento) o solo su una parte di essa, tramite un indagine campionaria.

Spesso la rilevazione su tutta la popolazione non è conveniente per questioni di tempi di realizzazione molto lunghi e di costi difficilmente sostenibili. In alcuni casi poi è proprio impossibile.

Pertanto, per effettuare la maggior parte delle indagini è necessario ricorrere ad un campionamento delle unità statistiche.

A tal fine è importante comprendere le differenze tra popolazione finita ed infinita e i loro parametri.

IL CAMPIONAMENTO PROBABILISTICO E RAGIONATO

Popolazione finita:

- è un insieme di unità su cui si può osservare un certo carattere (ad esempio, gli investimenti annui di tutte le aziende di un paese, il numero di figli di ogni famiglia italiana, etc..)
- i parametri della popolazione sono delle costanti che descrivono aspetti caratteristici della distribuzione del carattere nella popolazione stessa:

Media della popolazione: $\boldsymbol{\mu = \frac{1}{N}\sum_{i=1}^{n} x_i}$

Varianza della popolazione: $\boldsymbol{\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$

Popolazione infinita:

- è composta da tutte le unità potenzialmente osservabili e non necessariamente già esistenti. Il carattere d'interesse può essere rappresentato da una variabile casuale con una certa distribuzione di probabilità. In questo caso si indicherà con "popolazione ***X***" la variabile casuale ***X***.

Media della popolazione: $\boldsymbol{\mu = E(x)}$

Varianza della popolazione: $\boldsymbol{\sigma^2 = E[X - E(x)]^2}$

IL CAMPIONAMENTO PROBABILISTICO E RAGIONATO

BREVI DEFINIZIONI

Per **campione** si intende una porzione della popolazione obiettivo dell'indagine che si analizza con la speranza di ottenere delle informazioni estendibili all'intera popolazione.

Infatti, anche nel caso di rilevazioni effettuate su un campione, l'obiettivo di una ricerca resta sempre quello di migliorare la conoscenza della popolazione obiettivo.

Per **campionamento** si intende invece il procedimento pratico attraverso cui vengono selezionate, all'interno della popolazione obiettivo, le unità statistiche che andranno a far parte del campione.

IL CAMPIONAMENTO PROBABILISTICO E RAGIONATO

IL CAMPIONE PROBABILISTICO

Se la scelta delle unità da inserire nel campione è casuale si parla di campionamento probabilistico. In questo tipo di campionamento, ogni unità statistica della popolazione oggetto di indagine avrà una probabilità maggiore di zero ed individuabile a priori di essere inclusa nel campione. Quindi, se la regola di selezione del campione è di tipo probabilistico, l'estrazione del campione avviene in accordo con qualche specifica distribuzione di probabilità.

Per poter effettuare un campionamento probabilistico è quindi indispensabile disporre di una lista contenente l'elenco di tutte le unità che costituiscono la popolazione che si vuole analizzare. Per una selezione probabilistica è necessario individuare:

- lo spazio campionario Ω, formato da tutti i possibili campioni estraibili con una medesima tecnica da una popolazione.
- la probabilità di ogni campione c in Ω di essere estratto

La coppia $\{\Omega, c\}$ è detta piano di campionamento.

N.B: Quando la scelta delle unità da includere nel campione è svolta in modo non casuale e si parla di campionamento non probabilistico. I campionamenti non probabilistici hanno in genere tempi e costi di rilevazione molto più bassi di quelli probabilistici. Per questo motivo molte ricerche sono in realtà svolte ricorrendo a questo tipo di campionamento (ad esempio: campionamento a scelta ragionata, il campionamento per quote, il campionamento a valanga, etc..)

DISEGNI DI CAMPIONAMENTO PROBABILISTICI E NON PROBABILISTICI

I CAMPIONAMENTI CASUALI

Un campione probabilistico è un campione scelto in modo casuale.

I campioni possono essere estratti casualmente dalla popolazione:

- con ripetizione: una volta estratta, l'unità viene rimessa dentro la popolazione e quindi potrebbe essere nuovamente estratta
- senza ripetizione: una volta estratta un'unità, questa viene messa da parte e quindi non può essere più ri-estratta.

N.B: Due campioni non ordinati di uguale numerosità sono diversi tra loro se almeno un'unità del primo campione non è contenuta nel secondo campione. Nei campioni ordinati conta invece anche l'ordine con cui si presentano le diverse unità.

CAMPIONAMENTO CASUALE SEMPLICE

Nel campionamento casuale semplice i campioni di uguale dimensione hanno tutti stessa probabilità di essere estratti.

Le unità statistiche sono estratte a sorte (come se ogni unità statistica fosse un numero del Lotto da estrarre in modo casuale da un'urna) e non è possibile l'autoselezione tra chi deve rispondere.

- si devono conoscere le unità della popolazione
- tutte le unità devono essere reperibili
- si deve procedere all'estrazione casuale delle unità.

Il campione casuale ottenuto con estrazioni senza ripetizione è composto da ***n*** variabili casuali che hanno marginalmente stessa distribuzione di probabilità ma non sono indipendenti.

La distribuzione di probabilità della generica X_i è uguale a quella del carattere X nella popolazione.

CAMPIONAMENTO CASUALE SEMPLICE

Esempio: popolazione composta da 4 grandi aziende (N=4) e Carattere = ”Fatturato annuo”

X1 = 52; X2 = 49; X3 = 65; X4 = 74

c_1 =	52	49	65	c_7 =	49	65	74	c_{13} =	65	74	52	c_{19} =	74	52	49
c_2 =	52	65	49	c_8 =	49	74	65	c_{14} =	65	52	74	c_{20} =	74	49	52
c_3 =	49	52	65	c_9 =	65	49	74	c_{15} =	74	52	65	c_{21} =	52	74	49
c_4 =	49	65	52	c_{10} =	65	74	49	c_{16} =	74	65	52	c_{22} =	52	49	74
c_5 =	65	52	49	c_{11} =	74	49	65	c_{17} =	52	65	74	c_{23} =	49	52	74
c_6 =	65	49	52	c_{12} =	74	65	49	c_{18} =	52	74	65	c_{24} =	49	74	52

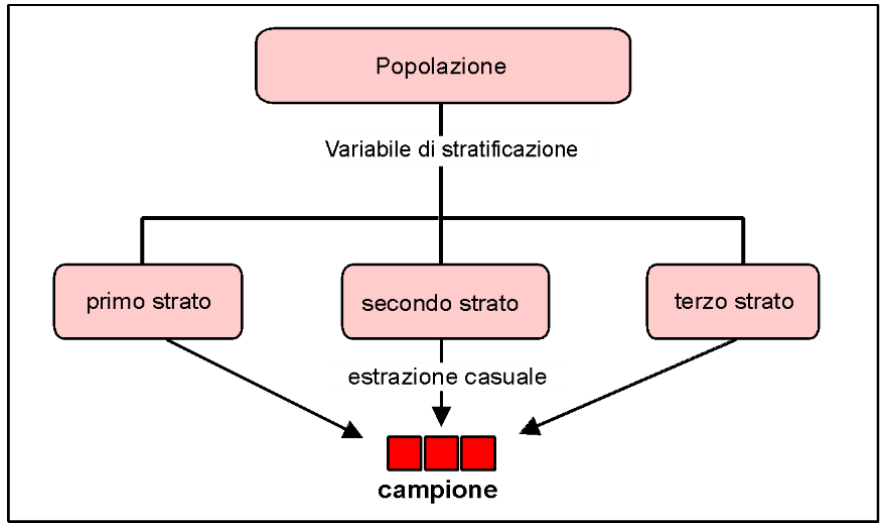
- Spazio campionario Ω, costituito dai campioni ordinati di dimensione 3, estratti senza ripetizione.
- Ogni campione ha uguale probabilità (c) di essere estratto, pari a 1/24.

CAMPIONAMENTO CASUALE STRATIFICATO

Si utilizza quando le unità statistiche possono essere suddivise in gruppi distinti (detti ***strati***) sulla base delle conoscenze che si hanno a priori sulla popolazione (ad esempio, gli strati possono essere utilizzati per suddividere le unità tra aree rurali ed aree urbane oppure per fasce di età).

Da ogni strato si estraggono poi a sorte le unità statistiche con un campionamento casuale semplice. Si ottengono così tanti campioni quanti sono gli strati della popolazione. Questi campioni sono poi riuniti tutti insieme in modo da formare un unico grande campione che sarà quello su cui si effettuerà l'analisi statistica.

Schema

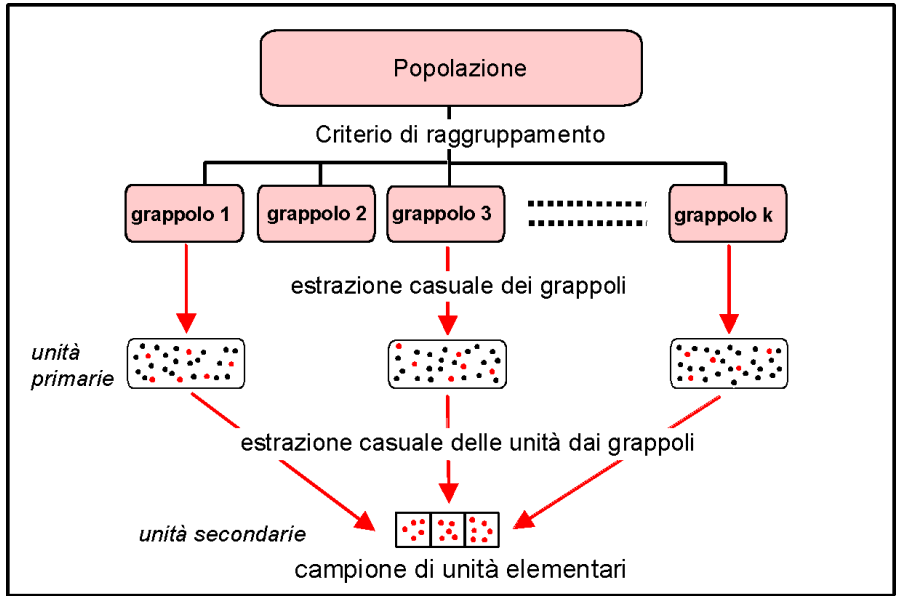


- a parità di numerosità campionaria, il campionamento casuale stratificato può produrre risultati più precisi rispetto ad un campionamento casuale semplice, in quanto tiene conto della somiglianza delle unità che fanno parte dello stesso strato
- è preferibile quando si vogliono ridurre i tempi ed i costi della fase di raccolta dati.

CAMPIONAMENTO A GRAPPOLI E A STADI

- Nel campionamento casuale a grappoli la popolazione è suddivisa in gruppi/sottoinsiemi (detti ***grappoli***). Si selezionano, con un'estrazione casuale senza ripetizione, un certo numero di grappoli e si prendono come unità campionarie tutte le unità appartenenti ai grappoli estratti.
- Nel campionamento casuale a due stadi la popolazione viene suddivisa in un certo numero di grappoli. Il primo ***stadio*** è uguale a quello del campionamento a grappoli. Successivamente, da ogni grappolo selezionato è estratto un campione casuale di unità statistiche di secondo stadio e così via per il numero di stadi prefissati.

Schema



LA DISTRIBUZIONE DEI CAMPIONI

Una distribuzione di campionamento è una distribuzione di probabilità di una statistica ottenuta attraverso un gran numero di campioni prelevati da una popolazione specifica.

La distribuzione di campionamento di una data popolazione è la distribuzione delle frequenze di una gamma di risultati diversi che potrebbero eventualmente verificarsi per una statistica della popolazione.

La **statistica campionaria è una variabile casuale** a cui è associata una distribuzione di probabilità detta distribuzione campionaria.

Ad esempio:

Media campionaria: $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i$

Varianza campionaria corretta: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)^2$

LA DISTRIBUZIONE DEI CAMPIONI

Le statistiche campionarie sono esempi di statistiche che possono essere utilizzate per stimare i corrispondenti parametri della popolazione.

Si può, infatti, dimostrare che:

Valore atteso della media campionaria: $\boldsymbol{E[\bar{X}] = \mu}$

Valore atteso della varianza campionaria corretta: $\boldsymbol{var[\bar{X}] = \frac{1}{n}\sigma^2}$

- Se il valore atteso della media campionaria è uguale alla media della popolazione significa che questa è uguale al parametro media da stimare e che la distribuzione è centrata intorno alla media
- Il fatto, invece, che il valore atteso della varianza campionaria sia uguale alla varianza della popolazione fratto la sua numerosità, significa che la dispersione dei valori intorno alla media è piccola quando l'ampiezza del campione è grande.
- Questa significa che se l'ampiezza del campione è grande, i valori di media campionaria, usati per stimare la media della popolazione tendono ad essere più concentrati intorno alla stessa media della popolazione rispetto a quanto lo sarebbero se l'ampiezza fosse piccola.

LA DISTRIBUZIONE DELLA MEDIA CAMPIONARIA

L'interesse dell'inferenza statistica è di trarre conclusioni sulla popolazione e su alcuni sui parametri e non sul solo campione. In via ipotetica, per usare le statistiche campionarie con lo scopo di stimare i parametri della popolazione, dovremmo analizzare tutti i campioni che possono essere estratti da questa. Nella pratica, da una popolazione viene estratto a caso un solo campione, di ampiezza prestabilita.

- ***La distribuzione della media campionaria è la distribuzione di tutte le possibili medie che osserveremmo se procedessimo all'estrazione di tutti i possibili campioni di una certa ampiezza***

Esempio

Consideriamo una popolazione costituita da 4 individui ciascuno dei quali pubblica un numero di libri in un anno:

$$X_1 = 3 \quad X_2 = 2 \quad X_3 = 1 \quad X_4 = 4$$

Considerando l'intera popolazione, la media e lo scarto quadratico medio possono essere calcolati:

$$\mu = 2{,}5$$

$$\sigma = 1{,}12$$

LA DISTRIBUZIONE DELLA MEDIA CAMPIONARIA

Supponiamo ora di estrarre con re immissione dalla popolazione un campione di ***n***=2 individui. In totale potremo estrarre ($N^n = 4^2 = 16$) campioni, riportati in tabella con le rispettive medie campionarie. Notiamo che la media delle medie campionarie ($\boldsymbol{\mu}$) è proprio uguale alla media della popolazione quindi la media campionaria è uno stimatore non distorto della media della popolazione.

Campione	Individui	Risultati campionari	Media campionaria
1	1;1	3;3	$\overline{X} = 3$
2	1;2	3;2	$X = 2{,}5$
3	1;3	3;1	$X = 2$
4	1;4	3;4	$X = 3{,}5$
5	2;1	2;3	$X = 2{,}5$
6	2;2	2;2	$X = 2$
7	2;3	2;1	$X = 1{,}5$
8	2;4	2;4	$X = 3$
9	3;1	1;3	$X = 2$
10	3;2	1;2	$X = 1{,}5$
11	3;3	1;1	$X = 1$
12	3;4	1;4	$X = 2{,}5$
13	4;1	4;3	$X = 3{,}5$
14	4;2	4;2	$X = 3$
15	4;3	4;1	$X = 2{,}5$
16	4;4	4;4	$X = 4$
			$\mu_X = 2{,}5$

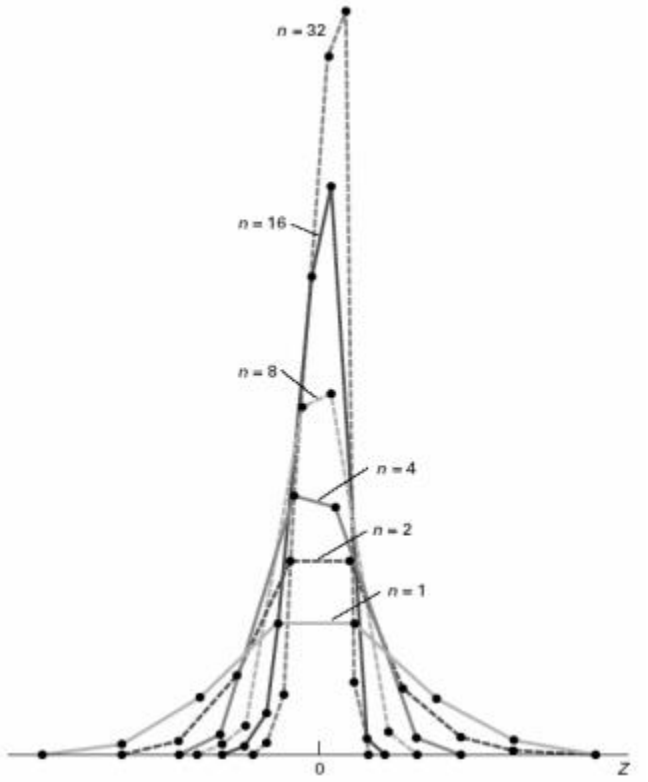
Quindi, anche se non sappiamo quanto la media di un dato campione sia vicina alla medie della popolazione, siamo sicuri che la media delle medie di tutti i campioni che potremmo selezionare coincide con la media della popolazione $\boldsymbol{\mu}$.

- Mentre le osservazioni nella popolazione assumono anche valori estremamente piccoli o estremamente grandi, la media campionaria è caratterizzata da una minore variabilità rispetto ai dati originali
- Le medie campionarie saranno quindi caratterizzate, in generale, da valori meno dispersi rispetto a quelli che si osservano nella popolazione. Lo scarto quadratico medio della media campionaria, detto errore standard della media, quantifica la variazione della media campionaria da campione a campione.

LA DISTRIBUZIONE DELLA MEDIA CAMPIONARIA

Introdotta l'idea di distribuzione campionaria e definito l'errore standard della media, bisogna stabilire quale sia la distribuzione della media campionaria:
Se un campione è estratto da una popolazione normale con media μ e scarto quadratico medio σ, la media campionaria ha distribuzione normale indipendentemente dall'ampiezza campionaria n, ed è caratterizzata da valore atteso $\mu_X = \mu$ e scarto quadratico medio pari all'errore standard σ_X .

Esempio



In figura sono riportate le distribuzioni delle medie campionarie di 500 campioni di ampiezza 1,2,4,8,16 e 32 estratti da una popolazione normale

IL TEOREMA DEL LIMITE CENTRALE

Sinora abbiamo analizzato la distribuzione della media campionaria nel caso di una popolazione con distribuzione normale.

Tuttavia, si presenteranno spesso casi in cui la distribuzione della popolazione non è normale. In questi casi è utile riferirsi ad un importante teorema della statistica, ***il teorema del limite centrale***, che consente di dire qualcosa sulla distribuzione della media campionaria anche nel caso in cui una popolazione che non abbia distribuzione normale.

Sia $X_1, X_2, X_3, \ldots, X_n$ una successione di variabili casuali indipendenti e identicamente distribuite ***(i.i.d),*** con media $\boldsymbol{\mu}$ e varianza $\boldsymbol{\sigma^2}$ finite, posto:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Si ha che la v.c converge in distribuzione, per n tendente all'infinito, alla v.c Normale standardizzata:

$$Z_n = \frac{(\overline{X_n} - \mu)\sqrt{n}}{\mu}$$

IL TEOREMA DEL LIMITE CENTRALE

Ciò spiega l'importanza che la funzione gaussiana assume nelle branche matematiche della statistica e della teoria della probabilità in particolare.

Il teorema del limite centrale svolge un ***ruolo cruciale in ambito inferenziale***, in quanto consente di fare inferenza sulla media della popolazione senza dover conoscere la forma specifica della distribuzione della ***p*** o popolazione.

Proprietà del teorema del limite centrale

- Il teorema è soddisfatto quando si ha a che fare con campioni di grande dimensione. Questo assicura che la somma delle medie campionarie è uguale a una distribuzione normale. Secondo il teorema del limite centrale si ritiene che un campione sia significativo quando supera più di 30 unità statistiche. In tal caso la distribuzione della media campionaria tenderà ad una distribuzione gaussiana.

Questa affermazione è valida per qualsiasi tipo di distribuzione.