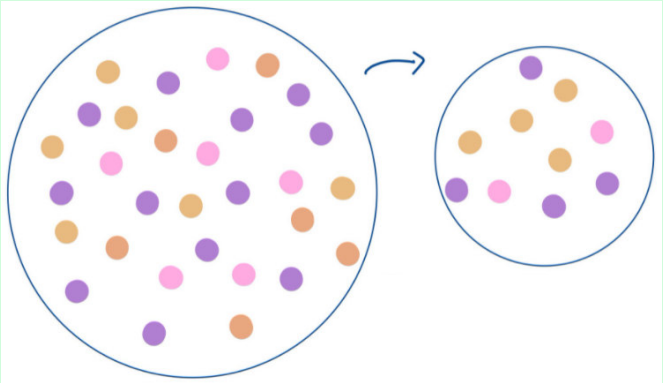
Campionamento

Con il termine **popolazione** in statistica si intende un insieme finito o infinito di tutte le unità statistiche di cui si vuole indagare una certa caratteristica che le individua come omogenee.

Per determinare le caratteristiche fondamentali di una popolazione statistica non è necessario analizzare tutte le unità statistiche della popolazione d'interesse, ma è sufficiente una parte di esse → **CAMPIONE**



Statistica inferenziale

Statistica inferenziale è volta all'induzione probabilistica delle caratteristiche ignote della popolazione.

Osservazioni svolte su un campione di unità rappresentative di tutta la popolazione, selezionate con date procedure, entro dati livelli di errore consentono di ottenere conclusioni che possono essere generalizzate all'intera popolazione.

Le 4 fasi principali sono:
1- estrazione di un campione della popolazione
2- calcolo delle statistiche campionarie, cioè dei valori di sintesi relativi ai dati del campione
3- stima dei parametri della popolazione in base ai risultati del campione
4- verifica dei risultati raggiunti

Rappresentatività del campione

Il procedimento di **S. I.** conduce a risultati esatti solo se il campione è perfettamente **rappresentativo** della popolazione.

Significa che il campione dovrebbe rispecchiare e riprodurre le caratteristiche essenziali e la stessa distribuzione della popolazione.

La **rappresentatività** è garantita dalla **casualità** della selezione delle unità statistiche del campione.

Campioni probabilistici

I campioni probabilistici sono caratterizzati dalla **casualità**, ciascuna unità della popolazione ha la stessa probabilità, diversa da zero, di essere estratta.

Consentono inferenza, ossia la **generalizzazione** dei risultati ottenuti alla popolazione intera, con scarti non significativi imputabili al caso.

Se estraiamo un campione di studenti a sorte tra quelli presenti in qualsiasi giorno in università non è un campione probabilistico:

- I non frequentanti hanno probabilità nulla di essere estratti
- Le matricole hanno una probabilità più alta di essere estratte
- Gli studenti fuori corso hanno una minore probabilità di essere estratti.

Statistica campionaria

Per **parametro** della popolazione si intende quel valore numerico utilizzato come misura di una delle caratteristiche della popolazione di riferimento (ad es. la media, la varianza etc.).

Nell'inferenza statistica la **statistica campionaria** rappresenta lo stimatore puntuale del corrispondente parametro della popolazione.

La **stima puntuale** è il valore di uno stimatore puntuale utilizzato per stimare un parametro della popolazione.

Statistica campionaria

	Parametro della popolazione	Statistica campionaria
Media	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Deviazione standard	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

La **statistica campionaria** è una caratteristica del campione come la media campionaria, la deviazione standard campionaria etc.

Tipo di campionamento

Campionamento casuale semplice: è la procedura di scelta casuale più semplice. Ogni unità ha la stessa probabilità di far parte del campione.

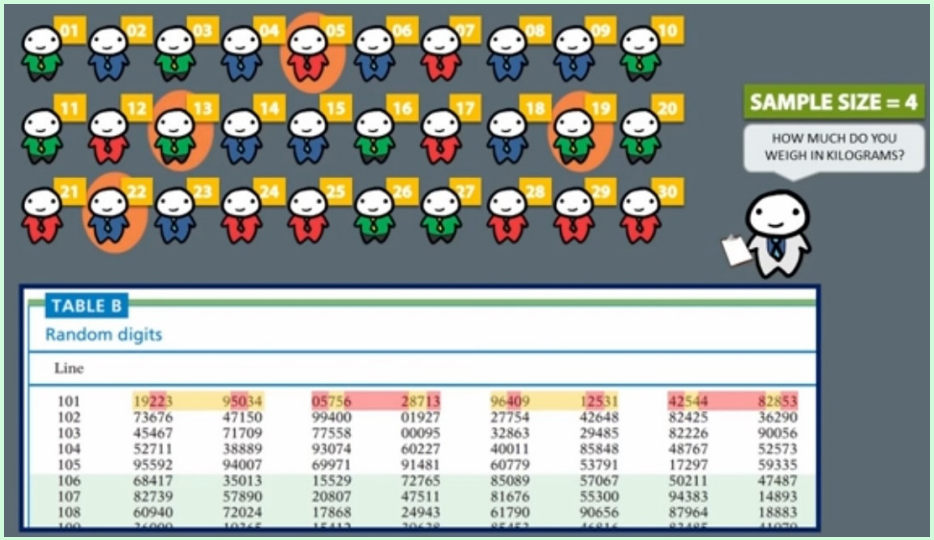
La casualità viene ottenuta estraendo numeri a partire da un elenco (detto "lista di campionamento") in cui sono presenti tutti gli individui della popolazione da studiare.

Il campionamento per randomizzazione semplice viene agevolmente applicato quando si dispone di una popolazione già numerata, preferibilmente composta di un numero non elevato di unità.

In questa procedura di selezione casuale si distinguono due modalità di estrazione dei campioni: con ripetizione (o bernulliani) e senza ripetizione (o in blocco); a seconda che vi sia reimmissione o no delle unità estratte.

Tavole dei numeri casuali

Esistono delle apposite tavole costituite da un insieme di numeri ricavati mediante algoritmi matematici, in modo che nel lungo andare ogni cifra, ogni coppia di cifre, ogni terna, ecc. abbia la stessa frequenza di ogni altra. Si individua casualmente un punto di partenza dal quale procedere ordinatamente per riga o per colonna.

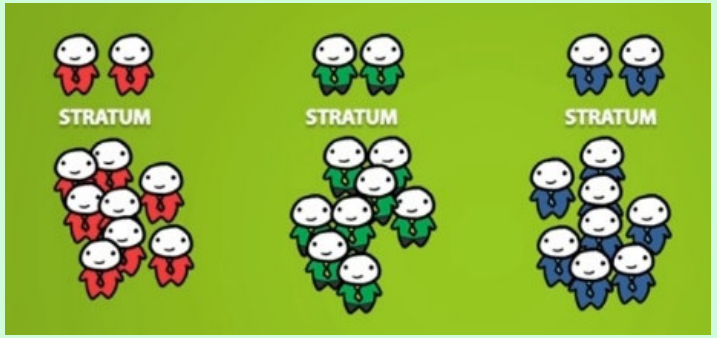


Tipo di campionamento

Campionamento stratificato: è una procedura di campionamento con la quale inizialmente si procede nella suddivisione della popolazione in un numero determinato di strati o classi il più possibile omogenei al loro interno, rispetto al carattere indagato e successivamente nell'estrazione di un campione casuale semplice di numerosità prefissata da ciascuno strato.

Campione estratto

Popolazione divisa in strati



Tipo di campionamento

Campionamento stratificato: la base per la creazione degli strati è a discrezione del soggetto che disegna il campione.

Es. area geografica, età, settore di attività economica.

Il campionamento stratificato funziona meglio quando la varianza tra gli elementi di ciascuno strato è relativamente piccola.

L'efficienza dipende, quindi, da quanto sono omogenei gli elementi all'interno di ciascuno strato.

Se gli strati sono omogenei, il campionamento stratificato fornisce risultati precisi quanto il campionamento casuale semplice, ma utilizzando una dimensione campionaria più piccola.

Tipo di campionamento

Campionamento multistadio: è una procedura di campionamento composito che presuppone l'individuazione di una struttura gerarchica della popolazione, in cui le unità finali sono incluse in insiemi di livello via via più elevato.

Esempio
1.estrazione casuale di un campione di comuni (unità di primo stadio)

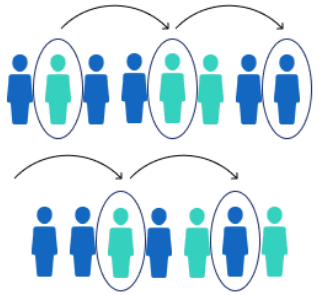
2. estrazione di un campione casuale di famiglie (unità di secondo stadio) da ciascuna lista anagrafica per ogni comune selezionato

Tipo di campionamento

Campionamento sistematico: differisce dal campionamento casuale semplice soprattutto dal punto di vista della tecnica di estrazione dei soggetti: si estraggono unità da una lista a intervalli regolari nella stessa. La prima unità va scelta casualmente poi si sceglie un'unità ogni intervallo k=N/n.

Si seleziona un punto di partenza casuale dalla popolazione e poi si procede selezionando le altre unità ad es. con una progressione aritmetica di ragione 15 fino all'esaurimento della lista.

Il campionamento sistematico è stato ideato per ridurre il lavoro sulle tavole dei numeri casuali ed è tutt'oggi ancora molto utilizzato.



Campionamento non probabilistico

Campionamento di convenienza: è una tecnica di campionamento non probabilistico, in cui gli elementi sono selezionati in base alla convenienza.

Es. le interviste volontarie per ricerche di mercato

Hanno il vantaggio che la selezione campionaria e la raccolta dei dati sono relativamente facili.
Tuttavia è impossibile valutare il campione in termini di rappresentatività della popolazione, per cui bisogna prestare molta attenzione nell'interpretazione dei risultati.

La **distorsione del campionamento (bias di selezione)** determina una differenza costante tra i risultati del campione e i risultati teorici dell'intera popolazione. Poiché il metodo di campionamento è arbitrario, la rappresentazione demografica della popolazione è quasi sempre distorta.

Esempi

Un ricercatore ha un elenco di tutti i residenti di una città di 500.000 abitanti e decide di generare un campione casuale individuando un individuo ogni 100 dall'elenco.
Tipo di campionamento? **Sistematico**

Un ricercatore vuole analizzare le caratteristiche delle persone appartenenti a diverse fasce di reddito annuale, decide quindi di creare dei gruppi in base al reddito familiare annuale.
Tipo di campionamento? Stratificato

Estrazione della lotteria.
Tipo di campionamento? Casuale semplice

Esempi

Il nuovo censimento dell'ISTAT
Tipo di campionamento? Multistadio (primo stadio estrazione dei Comuni, secondo stadio estrazione delle famiglie)

Volontari per la sperimentazione di un farmaco.
Tipo di campionamento? Convenienza

Margine di errore

La stima del parametro è probabilistica, essa comporta, cioè, un errore dovuto all'impossibilità di determinare con esattezza il parametro.

Il margine d'errore nell'inferenza statistica è fisiologico trattandosi di stime.

L'effettiva estrazione di tutti i campioni possibili (di pari ampiezza) da una popolazione è l'unico modo per capire quanto le statistiche calcolate su un campione possano discostarsi dai dati ricavati dall'analisi di tutti gli altri campioni.

Per evitare tale operazione esistono degli strumenti in grado di valutare la robustezza delle stime effettuate.

Distribuzione della media campionaria

La **media campionaria** è la variabile aleatoria Xn che descrive le medie di tutti i possibili campioni di ampiezza n che si possono estrarre dalla popolazione. Ha una propria distribuzione, dispersione etc.

Se si hanno a disposizione i valori medi di tutti i campioni possibili della stessa popolazione di riferimento si può generare la **distribuzione della media campionaria**, ossia quella distribuzione che raccoglie tutti i valori possibili che la media campionaria può assumere nei vari campioni estratti da una popolazione.

Conoscere i valori più frequenti della media campionaria vuol dire anche sapere quali sono i più probabili.

Distribuzione della media campionaria

Medie dei campioni estratti

935, 867, 743, 654, 194, 234, 236, 704, 560, 350, 353,
760, 705, 413, 520, 712, 340, 342, 813, 403, 204, 378,
800, 788, 675, 456, 657, 702, 456, 412, 506, 675, 645,
430, 506, 348, 344, 514, 488, 605, 320, 500, 501, 653,
621, 704, 706, 566, 534, 390, 400, 333, 475, 506, 721,
711, 509, 430, 322, 421

x_1=...
x_2=...
x_3=...
.
.
x_k=...

Il valore medio della distribuzione della media campionaria è dato dalla media aritmetica delle medie di ogni campione.

Si sommano tutte le medie e si divide per il numero di osservazioni, ossia i campioni. Si ottiene così la media della distribuzione.

Distribuzione campionaria

La **distribuzione campionaria** è una distribuzione di probabilità che consiste di tutti i valori possibili che una statistica campionaria può assumere e delle probabilità associate di ciascun valore.

Poiché campioni casuali differenti forniscono valori differenti per gli stimatori puntuali, questi ultimi sono considerati variabili casuali.

Tale distribuzione è generata teoricamente prendendo infiniti campioni di dimensione n e calcolando i valori della statistica per ogni campione.

Le più utilizzate sono le distribuzioni della media campionaria e della varianza campionaria.

Esempio

Popolazione di 4 unità con peso medio 73,25kg.

Unità della popolazione	Peso (kg)
A	78
B	65
C	68
D	82

Esempio

Consideriamo tutti i campioni di 2 unità estraibili e calcoliamo la media per ciascun campione:

Campioni	Media campionaria
AB	(78+65/2)= 71,5
AC	(78+68/2)= 73
AD	(78+82/2)= 80
BA	71,5
BC	66,5
BD	73,5
CA	73
CB	66,5
CD	75
DA	80
DB	73,5
DC	75

Media delle medie campionarie = 73,25 (nessuno dei campioni estratti presenta esattamente la media di 73,25).

La media delle medie campionarie è proprio uguale alla media della popolazione, quindi la media campionaria è uno stimatore non distorto della media della popolazione.

Errore standard

Occorre considerare la volatilità della media campionaria.

L'**errore standard** (o scarto quadratico medio della media campionaria) indica lo scostamento medio delle singole medie campionarie rispetto al valore medio della popolazione, quindi la variabilità dalla media della popolazione. È una stima di quanto la media del campione si avvicini alla media della popolazione.

È una misura dell'errore che ci si può aspettare quando scegliamo un campione di una certa ampiezza dalla popolazione.

Calcolo dell'errore standard

L'errore standard dei valori della distribuzione della media campionaria è uguale alla deviazione standard della popolazione divisa per la radice quadrata dell'ampiezza del campione:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

n = numerosità del campione

In questo caso $6{,}97/\sqrt{2} = 4{,}93$

Errore standard nel caso di σ non nota

Nel caso in cui non sia nota la deviazione standard della popolazione si può utilizzare la deviazione standard del campione.

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Errore standard

L’entità dell’errore standard dipende da:

1- dimensione del campione.
Maggiore è l’ampiezza, minore è l’errore standard. Un campione più ampio che si avvicina alla numerosità della popolazione determina una stima più precisa perché si basa su un numero maggiore di osservazioni.

2- deviazione standard della popolazione. Maggiore è la volatilità della distribuzione dei dati nella popolazione, maggiore sarà l’errore standard.

Errore standard

Se la deviazione standard della popolazione esprime quanto variano i singoli valori rispetto alla media, l'errore standard indica la variazione della media campionaria tra i diversi campioni.

La dispersione della media campionaria è inferiore rispetto a quella della popolazione.

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.

Deviazione standard del campione

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Bessel's correction

In un campione la dispersione intorno alla media è minore perché è composto da meno elementi rispetto alla popolazione.
Pertanto, la dispersione statistica intorno al valore medio nel campione è naturalmente inferiore rispetto a quella dell'intera popolazione. Per attenuare questo effetto, ossia la tendenza a sottostimare i parametri della popolazione, nel caso dei campioni viene usata la formula della deviazione standard campionaria con il fattore di correzione Bessel.

Esempio

xi	$xi - \bar{x}$	$(xi - \bar{x})^2$
11	0,74	0,55
8,5	-1,76	3,10
7,8	-2,46	6,05
7,4	-2,86	8,18
11,4	1,14	1,30
9	-1,26	1,59
9,5	-0,76	0,58
13,2	2,94	8,64
14,3	4,04	16,32
10,5	0,24	0,06
$\bar{X}$ 10,3		46,39

Campione di dimensione n = 10

$$S = \left(\frac{46,39}{10-1}\right)^{\wedge(1/2)} = 2,27$$

Se si calcolasse la dispersione usando la formula della deviazione standard per l'intera popolazione la variabilità risulterebbe inferiore e pari a 2,15.

Deviazione standard ≠ errore standard

La **deviazione standard** misura la dispersione di un set di dati di un campione o una popolazione rispetto alla media.

L'**errore standard** misura quanta discrepanza è probabile che ci sia ad es. nella media di un campione rispetto alla media della popolazione.
Descrive l'incertezza nella stima di un valore statistico (es. media, proporzione ecc.).
L'errore standard viene utilizzato per misurare l'accuratezza statistica di una stima. Es. l'errore standard fornisce l'accuratezza di una media campionaria misurando la variabilità da campione a campione delle medie campionarie. Descrive, quindi quanto sia precisa la media del campione come stima della media reale della popolazione.

Esercizio

L'errore standard è la differenza tra il valore di una statistica campionaria ed il corrispondente valore del parametro nella popolazione? Vero
L'errore standard è sempre più piccolo della deviazione standard? Vero

Calcolare la media e la deviazione standard campionaria per il campione selezionato.
Qual è l'errore standard?
Se il campione fosse di 20 unità come varierebbe l'errore std?

Reddito mensile
1.100
1.150
1.130
1.200
1.350
1.330
1.220
1.400
1.240
1.350
1.600
1.620
1.430
1.520
1.540
1.630

Soluzione

Reddito mensile	xi - x	$(xi - x)^2$
1.100	-263	69.235
1.150	-213	45.422
1.130	-233	54.347
1.200	-163	26.610
1.350	-13	172
1.330	-33	1.097
1.220	-143	20.485
1.400	37	1.360
1.240	-123	15.160
1.350	-13	172
1.600	237	56.110
1.620	257	65.985
1.430	67	4.472
1.520	157	24.610
1.540	177	31.285
1.630	267	71.222
$\bar{X}$ = 1.363		487.744

$S = (487.744/(16-1))^{(1/2)} = 180$

$SE(\bar{x}) = 180/\sqrt{16} = 45$

$SE(\bar{x}) = 180/\sqrt{20} = 40$