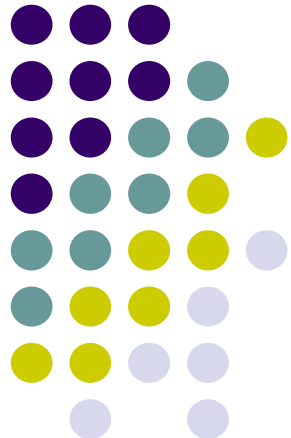
Levine, Krehbiel, Berenson

Statistica

Casa editrice: Pearson

Capitolo 7

Distribuzioni campionarie



Insegnamento: Statistica
Corsi di Laurea Triennale in Economia

Dipartimento di Economia e Management, Università di Ferrara
Docenti: Prof. Stefano Bonnini, Dott.ssa Valentina Mini

Argomenti

- La distribuzioni campionarie
- La distribuzione della media campionaria
  - Non distorsione della media campionaria
  - Errore standard della media campionaria
  - Campionamento da popolazione normale
  - Campionamento da popolazione non normale: il teorema del limite centrale
- La distribuzione della proporzione (frequenza) campionaria
- Metodi di campionamento
- Valutare l'adeguatezza delle indagini campionarie

Distribuzioni campionarie

- L'interesse dell'**inferenza statistica** è di trarre conclusioni sulla popolazione e su alcuni sui parametri e non sul solo campione
- A questo scopo si utilizzano delle statistiche calcolate sulla base del campione casuale per **stimare** i valori dei corrispondenti parametri dell'intera popolazione
- La **media campionaria** è una statistica utilizzata per stimare la media di una caratteristica di interesse (ad es. il peso dell'unità di prodotto) riferita all'intera popolazione (un dato processo produttivo)
- La **proporzione campionaria** è una statistica utilizzata per stimare la proporzione di unità (ad es. la quota di voti elettorali) in una popolazione (gli elettori) che hanno una certa caratteristica (danno la preferenza ad un certo candidato)

Distribuzioni campionarie

- 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
- Gli elementi da includere nel campione sono scelti mediante l'uso di un generatore di numeri casuali
- Supponendo di procedere all'estrazione di tutti i possibili campioni, la distribuzione di tutti i risultati ottenuti si dice **distribuzione campionaria**

La distribuzione della media campionaria

- Nel Capitolo 3 abbiamo introdotto diverse misure di posizione, tra le quali quella indubbiamente più utilizzata è la media aritmetica
- La media campionaria – la media degli elementi di un campione – viene utilizzata per stimare la media della popolazione
- La **distribuzione della media campionaria** è la distribuzione di tutte le possibili medie che osserveremmo se procedessimo all'estrazione di tutti i possibile campioni di una certa ampiezza
- La media campionaria è **non distorta** per la media della popolazione, cioè la media di tutte le possibili medie campionarie (calcolate a partire campioni di uguale ampiezza *n*)

La distribuzione della media campionaria

La proprietà di non distorsione (o correttezza) può essere dimostrata in via empirica mediante il seguente esempio: consideriamo una popolazione costituita dai 4 dattilografi, ciascuno dei quali commette un numero di errori di battitura sotto riportato

Dattilografo	Numero di errori
Ann	$X_1 = 3$
Bob	$X_2 = 2$
Carla	$X_3 = 1$
Dave	$X_4 = 4$

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

$$\mu = \frac{3 + 2 + 1 + 4}{4} = 2.5 \text{ errori}$$

$$\sigma = \sqrt{\frac{(3 - 2.5)^2 + (2 - 2.5)^2 + (1 - 2.5)^2 + (4 - 2.5)^2}{4}} = 1.12 \text{ errori}$$

La distribuzione della media campionaria

Supponiamo ora di estrarre con reimmissione dalla popolazione un campione di n=2 dattilografi. 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 ($\mu_{\bar{X}}$) è proprio uguale alla media della popolazione μ, quindi la media campionaria è uno **stimatore non distorto** della media della popolazione.

Campione	Dattilografi	Risultati campionari	Media campionari
1	Ann, Ann	3,3	$\bar{X}$ =3
2	Ann, Bob	3,2	$\bar{X}$ =2.5
3	Ann, Carla	3,1	$\bar{X}$ =2
4	Ann, Dave	3,4	$\bar{X}$ =3.5
5	Bob, Ann	2,3	$\bar{X}$ =2.5
6	Bob, Bob	2,2	$\bar{X}$ =2
7	Bob, Carla	2,1	$\bar{X}$ =1.5
8	Bob, Dave	2,4	$\bar{X}$ =3
9	Carla, Ann	1,3	$\bar{X}$ =2
10	Carla, Bob	1,2	$\bar{X}$ =1.5
11	Carla, Carla	1,1	$\bar{X}$ =1
12	Carla, Dave	1,4	$\bar{X}$ =2.5
13	Dave, Ann	4,3	$\bar{X}$ =3.5
14	Dave, Bob	4,2	$\bar{X}$ =3
15	Dave, Carla	4,1	$\bar{X}$ =2.5
16	Dave, Dave	4,4	$\bar{X}$ =4
			$\mu_{\bar{X}}$ =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 μ.

La distribuzione della media campionaria

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:

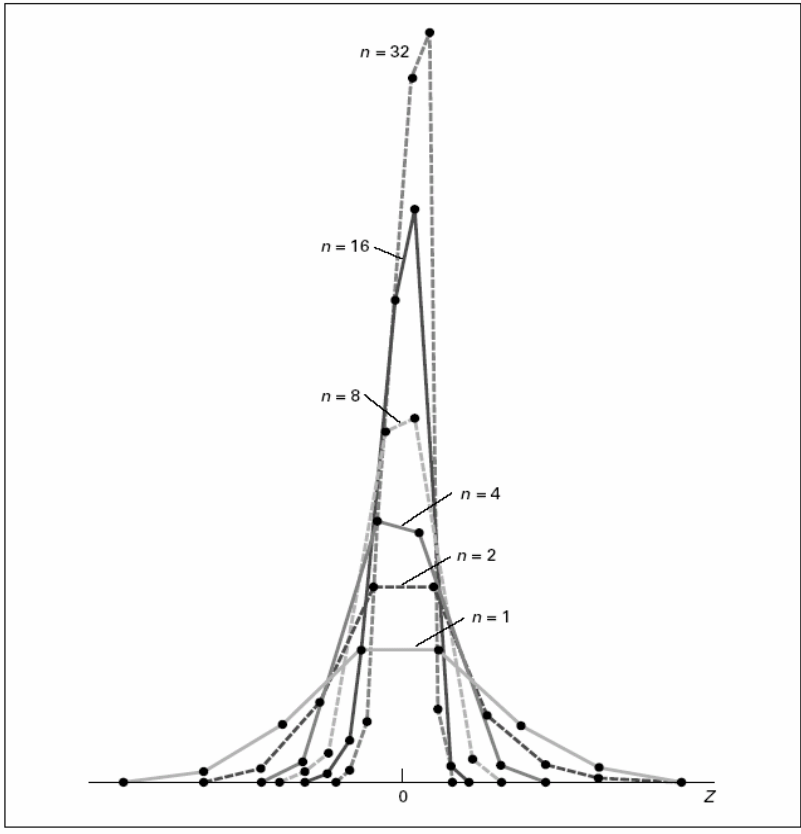
L'errore standard della media

$$\sigma_{\bar{X}} = \sigma / \sqrt{n} \qquad \textbf{(7.3)}$$

L'errore standard della media è uguale allo scarto quadratico medio della popolazione diviso $\sqrt{n}$.

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 $\mu_{\overline{X}} = \mu$ e scarto quadratico medio pari all'errore standard $\sigma_{\overline{X}}$. 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.

La distribuzione della media campionaria

Per cogliere a fondo il concetto di distribuzione della media campionaria, torniamo al caso del macchinario per il riempimento delle scatole di cereali è supponiamo che sia predisposto in maniera tale che la quantità di cereali in una scatola abbia una distribuzione normale con media 368, con lo scarto quadratico medio della popolazione per questo processo di riempimento pari a 15 grammi.

Il campione si configura come una rappresentazione in piccolo della popolazione, per cui se la popolazione ha una distribuzione normale, ci aspettiamo che i valori del campione siano approssimativamente distribuiti come una normale Quindi, se la media della popolazione è 368 grammi, la media del campione ha una buona probabilità di essere vicina a 368 grammi.

La distribuzione della media campionaria

Per approfondire ulteriormente questa analisi, potremmo calcolare la probabilità che il campione di 25 scatole abbia una media inferiore a 365 grammi. Ma come calcolare tale probabilità? Dal paragrafo 6.2 sappiamo che, se la distribuzione è normale, per calcolare una probabilità dobbiamo innanzitutto procedere con la standardizzazione, in questo caso riferendoci alla media campionaria : $\overline{X}$

Standardizzazione della media campionaria

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \qquad \textbf{(7.4)}$$

Quindi il valore *Z* corrispondente ad una media campionaria pari a 365 grammi, per un campione di 25 scatole è dato da

$$Z = (365 - 368) / (15 / \sqrt{25}) = -3/3 = -1.00$$

La distribuzione della media campionaria

Dalla Tavola E.2 si ricava che l'area cumulata fino Z=−1.00 è 0.1587, pertanto il 15.87% di tutti i possibili campioni di ampiezza 25 ha una media campionaria al di sotto di 365 grammi.

Questo non equivale a dire che 0.1587 è la probabilità che una **singola** scatola contenga meno di 365 grammi. Infatti, per una singola scatola si ha:

$$Z = (X - \mu)/\sigma = (365 - 368)/15 = -3/5 = -0.20$$

e l'area cumulata fino a Z=−0.20 è 0.4207, pertanto ci aspettiamo che il 42.07% delle singole scatole contenga meno di 365 grammi.

In generale, la probabilità che la media del campione sia lontana dalla media della popolazione è inferiore alla probabilità che la **singola osservazione** lo sia.

La distribuzione della media campionaria

In alcuni casi può essere utile determinare un intervallo in cui cade una proporzione prefissata delle medie campionarie. Si tratta di determinare un intervallo centrato sulla media della popolazione tale che l'area sottesa alla curva normale su tale intervallo abbia uno specifico valore.

Risolvendo l'equazione (7.4) rispetto a $\overline{X}$ si ottiene:

$$\overline{X} = \mu + Z\frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.5)}$$

Quindi, l'estremo inferiore $\overline{X}_L$ e l'estremo superiore $\overline{X}_U$ dell'intervallo centrato sulla media della popolazione che contiene il 95% delle medie campionarie e dato da

$$\overline{X}_L = 368 + (-1.96)(15)/\sqrt{25} = 368 - 5.88 = 362.12$$

$$\overline{X}_U = 368 + (1.96)(15)/\sqrt{25} = 368 + 5.88 = 373.88$$

Campionamento da popolazione non 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 non abbia distribuzione normale.

Il teorema del limite centrale

Quando l'ampiezza del campione casuale diventa sufficientemente grande, la distribuzione della media campionaria può essere approssimata dalla distribuzione normale. E questo indipendentemente dalla forma della distribuzione dei singoli valori della popolazione.

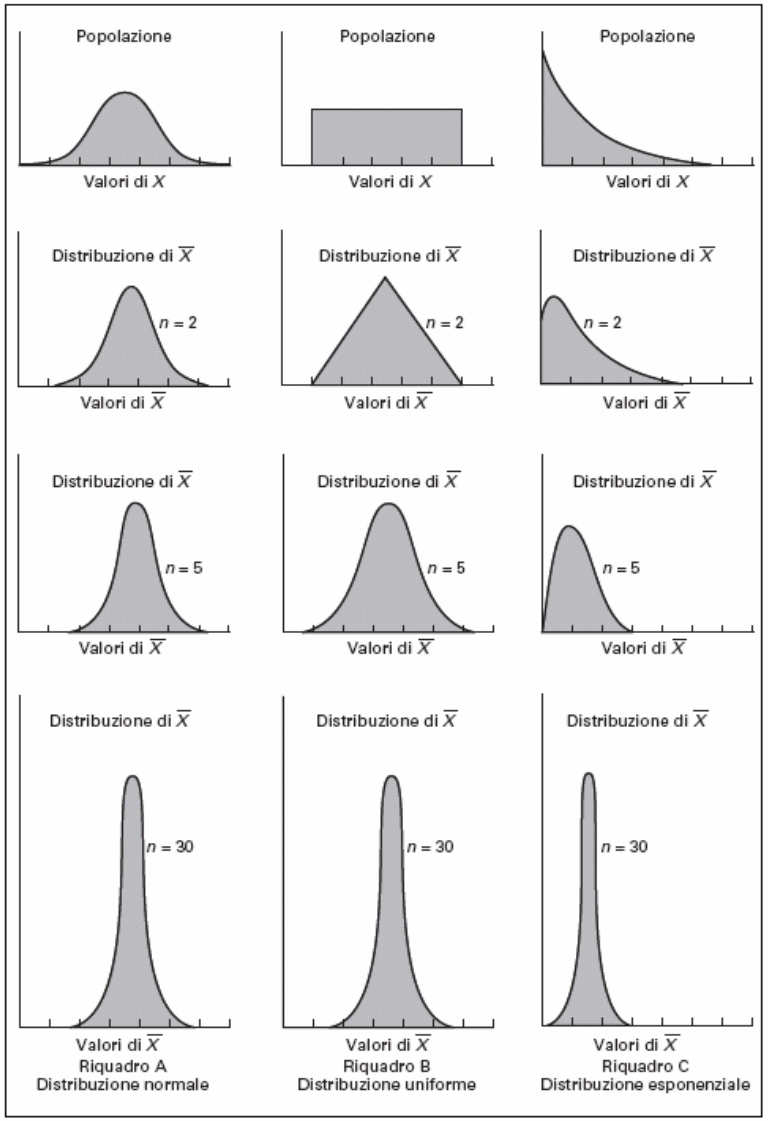
Campionamento da popolazione non normale: il teorema del limite centrale

Si tratta, allora, di stabilire cosa si intende per "sufficientemente grande", problema ampiamente affrontato dagli statistici. Come regola di carattere generale, molti sono concordi nell'affermare che quando il campione raggiunge un'ampiezza pari almeno a 30, la distribuzione della media campionaria è approssimativamente normale. Tuttavia, il teorema del limite centrale può essere applicato anche con campioni di ampiezza inferiore se si sa che la distribuzione della popolazione ha alcune caratteristiche che la avvicinano di per se stessa alla normale (ad esempio, quando è simmetrica).

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 popolazione.

Campionamento da popolazione non normale: il teorema del limite centrale

Ciascuna delle distribuzioni campionarie riportate è ottenuta estraendo 500 campioni diversi dalle rispettive popolazioni. Sono state considerate diverse ampiezze campionarie (n = 2, 5, 30). Nella seconda colonna è riportata la distribuzione della media campionaria nel caso di una popolazione la cui distribuzione (uniforme o rettangolare) è simmetrica e nella terza si considera una popolazione con distribuzione obliqua a destra (esponenziale).



Campionamento da popolazione non normale: il teorema del limite centrale

Sulla base dei risultati ottenuti per le distribuzioni note (la normale, l'uniforme l'esponenziale) possiamo trarre alcune conclusioni in merito al teorema del limite centrale:

- Per la maggior parte delle popolazioni, indipendentemente dalla forma della loro distribuzione, la distribuzione della media campionaria è approssimativamente normale, purché si considerino campioni di almeno 30 osservazioni.
- Se la distribuzione della popolazione è abbastanza simmetrica, la distribuzione della media campionaria è approssimativamente una normale, purché si considerino campioni di almeno 5 osservazioni.
- Se la popolazione ha una distribuzione normale, la media campionaria è distribuita secondo la legge normale, indipendentemente dall'ampiezza del campione.

La distribuzione della proporzione (frequenza) campionaria

Consideriamo una variabile qualitativa che assume solo due modalità, a seconda che presenti o meno una certa caratteristica. Ad esempio, consideriamo se un consumatore scelto a caso preferisce il nostro prodotto o quello della concorrenza. Siamo interessati alla proporzione nella popolazione, indicata con π, che viene stimata dalla proporzione campionaria, indicata con *p*.

Proporzione campionaria

$$p = X/n = \quad \textbf{(7.6)}$$

$$= \frac{\text{numero di casi nel campione che presentano la caratteristica di interesse}}{\text{ampiezza campionaria}}$$

La proporzione assume valori compresi tra 0 ed 1, estremi inclusi.

La distribuzione della proporzione (frequenza) campionaria

La proporzione campionaria è uno **stimatore non distorto** della proporzione nella popolazione (così come la media campionaria è uno stimatore non distorto per la media della popolazione). L'**errore standard della proporzione campionaria**, misura la dispersione delle proporzioni campionarie (osservate in tutti i possibili campioni) attorno alla proporzione della popolazione:

Errore standard della proporzione campionaria

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad \textbf{(7.7)}$$

Se consideriamo tutti i possibili campioni di una certa ampiezza, la distribuzione di tutte proporzioni campionarie si dice **distribuzione campionaria della proporzione**

La distribuzione della proporzione (frequenza) campionaria

Quando il campionamento è effettuato con reimmissione (da una popolazione di ampiezza finita), la distribuzione della proporzione è legata alla binomiale. Quando $n\pi$ e $n(1-\pi)$ sono entrambi almeno uguali a 5, la distribuzione binomiale può essere approssimata con la distribuzione normale. Quindi per valutare alcune probabilità relative alla proporzione campionaria, possiamo standardizzare ed utilizzare la distribuzione normale:

Standardizzazione della proporzione campionaria

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \qquad \textbf{(7.8)}$$

Metodi di campionamento

Il campione viene definito come la parte di una popolazione che si seleziona per l'analisi. Piuttosto che ricorrere a un censimento completo dell'intera popolazione, le procedure di campionamento statistico si concentrano su un piccolo gruppo, rappresentativo della popolazione. Il campione che ne risulta fornisce le informazioni che possono essere utilizzate per stimare le caratteristiche della popolazione nel suo insieme.

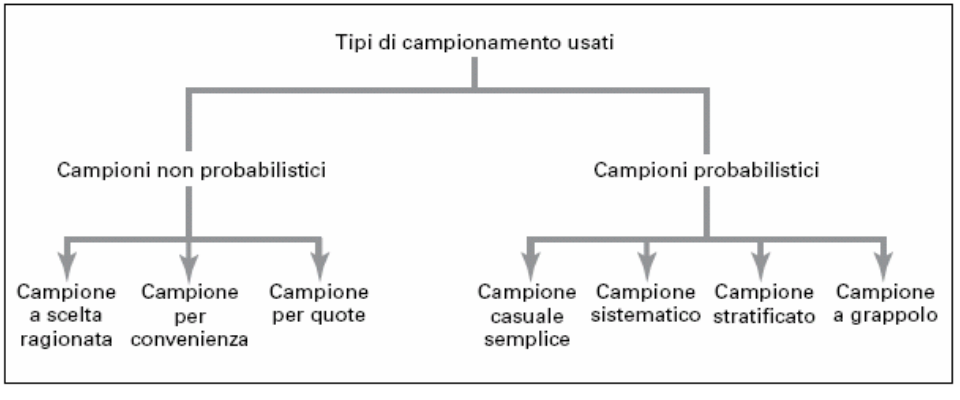
Ci sono tre motivi principali per utilizzare un campione:

- L'estrazione di un campione richiede meno tempo di un censimento
- Un campione è meno costoso di un censimento
- Un campione è più pratico da gestire di un censimento della popolazione considerata

Metodi di campionamento

Ci sono fondamentalmente due tipi di campioni:

- **campioni non probabilistici**: le unità campionarie sono selezionate senza tenere conto della loro probabilità di appartenere al campione perciò non può essere applicata la teoria sviluppata per il campionamento probabilistico
- **campioni probabilistici**: è nota la probabilità che una certa unità della popolazione faccia parte del campione. Il campionamento probabilistico dovrebbe essere usato ogni volta sia possibile, perché è il solo metodo che consente di ottenere inferenze corrette sulla base di un campione



Metodi di campionamento

Nel **campionamento casuale semplice** si estrae un campione in cui ogni individuo o oggetto della popolazione ha la stessa probabilità di essere selezionato. Inoltre, campioni della medesima dimensione hanno tutti la stessa probabilità di essere selezionati. Il campionamento casuale semplice è la più semplice tecnica di selezione del campione. Ci sono due metodi fondamentali per la selezione del campione:

- **con reimmissione**: le unità, una volta selezionate, vengono rimesse nella lista, da cui hanno la stessa probabilità di essere selezionate di nuovo
- **senza reimmissione**: una persona o un oggetto, una volta selezionati, non sono rimessi nella lista e pertanto non possono essere scelti di nuovo

Metodi di campionamento

Nel **campionamento sistematico** gli *N* elementi della lista della popolazione sono ripartiti in *n* gruppi costituiti da *k* elementi: $k = N / n$

dove *k* è arrotondato all'intero più vicino. Per ottenere un campione sistematico, si sceglie a caso un elemento tra i *k* elementi del il primo gruppo. Gli altri elementi del campione del campione si ottengono scegliendo da quel punto in poi ogni *k*-esimo elemento successivo dell'intera lista della popolazione.

Sebbene più facili da usare, il campionamento casuale e il campionamento sistematico sono in genere meno efficienti di altri schemi di campionamento più sofisticati: i dati raccolti potrebbero dare una rappresentazione non buona delle caratteristiche sottostanti della popolazione (i parametri).

Metodi di campionamento

In un **campionamento stratificato**, gli *N* elementi della lista della popolazione sono suddivisi in distinte sottopopolazioni, o strati, sulla base di una caratteristica comune. Si conduce un campionamento casuale semplice in ogni strato e si combinano.

Questo metodo di campionamento è più efficiente sia del campionamento casuale semplice che del campionamento sistematico, perché assicura che gli individui o oggetti della popolazione siano rappresentati adeguatamente nel campione, e questo garantisce una maggiore precisione delle stime dei parametri sottostanti alla popolazione. È l'omogeneità degli individui o oggetti all'interno di ogni strato che, quando combinata attraverso gli strati, fornisce la precisione.

Metodi di campionamento

Nel **campionamento a grappolo**, gli *N* elementi della lista sono suddivisi in molti gruppi, detti grappoli (sottopopolazioni), in maniera tale che ogni grappolo sia rappresentativo dell'intera popolazione. Si estrae poi un campione casuale di grappoli e tutti gli individui o oggetti di ciascuno dei grappoli selezionati sono inclusi nel campione. I grappoli possono essere definiti sulla base di raggruppamenti naturali, come quelli determinati dalle regioni, dalle città, dalle circoscrizioni elettorali, dai quartieri urbani, dagli edifici o dalle famiglie.

Il campionamento a grappolo tende a essere meno efficiente sia del campionamento casuale semplice che del campionamento stratificato, e si rende necessaria una dimensione complessiva del campione più grande per ottenere risultati precisi come quelli che si ottengono da procedure più efficienti.

Valutare l'adeguatezza delle indagini campionarie

Nella valutazione dell'adeguatezza di un'indagine bisogna considerare se essa è basata su un campionamento probabilistico o non probabilistico: il solo modo per fare inferenze statistiche corrette da un campione all'intera popolazione è far l'uso di un campione probabilistico. Indagini che fanno ricorso al campionamento non probabilistico sono soggette a serie distorsioni che potrebbero rendere i risultati privi di ogni significato.

Anche quando fanno uso del campionamento probabilistico, i sondaggi sono soggetti a possibili errori:

1. Errore di copertura o distorsione nella selezione
2. Errore o distorsione da mancata risposta
3. Errore di campionamento
4. Errore di misurazione