# Tecniche di rilevazione statistica

Il disegno di ricerca

Indagini censuarie e campionarie

Introduzione al campionamento

Indagini campionarie basate su questionario

# Il disegno di ricerca

**Con il termine "disegno di ricerca" ci si riferisce a tutti gli aspetti inerenti alla pianificazione e conduzione di una ricerca, comprese le fasi di analisi dei dati raccolti e di presentazione l'interpretazione finale dei risultati.**

Le principali fasi di una indagine statistica sono:

- **Definizione degli scopi della ricerca**
- **Definizione della popolazione e selezione del campione**
- **Definizione delle variabili di interesse**
- **Costruzione dello strumento e delle modalità di raccolta dati**
- **Raccolta dati**
- **Registrazione e pulitura dei dati**
- **Analisi statistica**
- **Presentazione e interpretazione dei risultati**

# Risorse e vincoli nel disegno di ricerca

Nella progettazione e conduzione di una indagine statistica sono da considerare risorse e vincoli di diverso tipo che condizionano le varie fasi della ricerca, tra i quali:

- **Risorse e vincoli finanziari della ricerca:** riguardano le risorse necessarie per condurre ciascuna delle varie fasi di ricerca, da quelle iniziali (es. ricerca pilota), alla preparazione del materiale necessario (es. stampa questionari), alla scelta dell'ampiezza del campione, alla fase di analisi dei dati.

- **Vincoli temporali:** possono riguardare i contenuti scientifici della ricerca (ad esempio se i dati non vengono raccolti entro un certo intervallo di tempo, cambia il significato o la rilevanza dei risultati), o possono riguardare aspetti tecnici (ad esempio le risorse umane e finanziare sono disponibili solo in un certo intervallo di tempo)

## Gli scopi della ricerca

**Le indagini statistiche si distinguono in relazione alle finalità della ricerca che possono essere di tipo:**

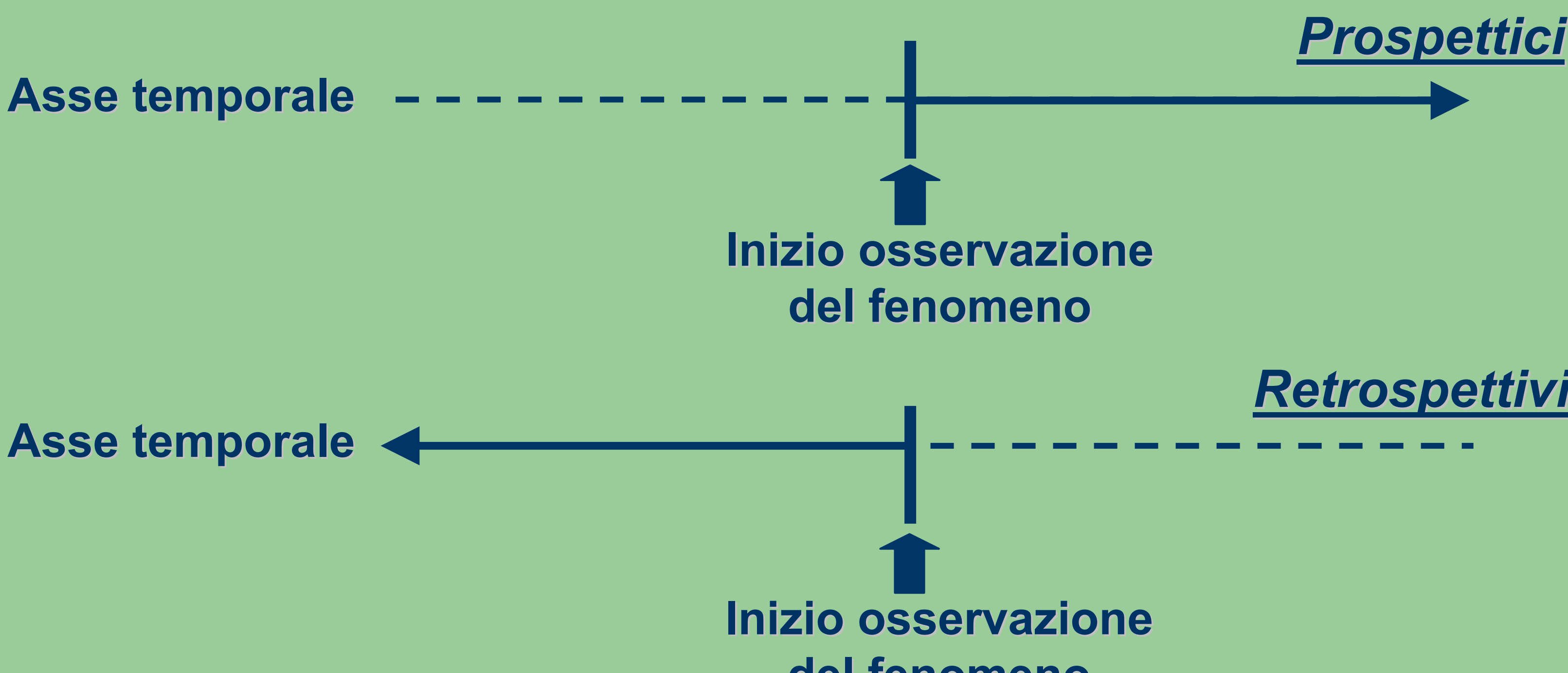
- **Descrittivo o esplorativo**: l'interesse del ricercatore si limita alla descrizione del fenomeno senza formulare e verificare ipotesi su quanto osservato (es. sondaggi, indagini pilota, …)

- **Esplicativo o analitico:** l'interesse del ricercatore va oltre la descrizione del fenomeno nel campione selezionato, e ricerca possibili relazioni tra variabili, generalizzabili all'intera popolazione.

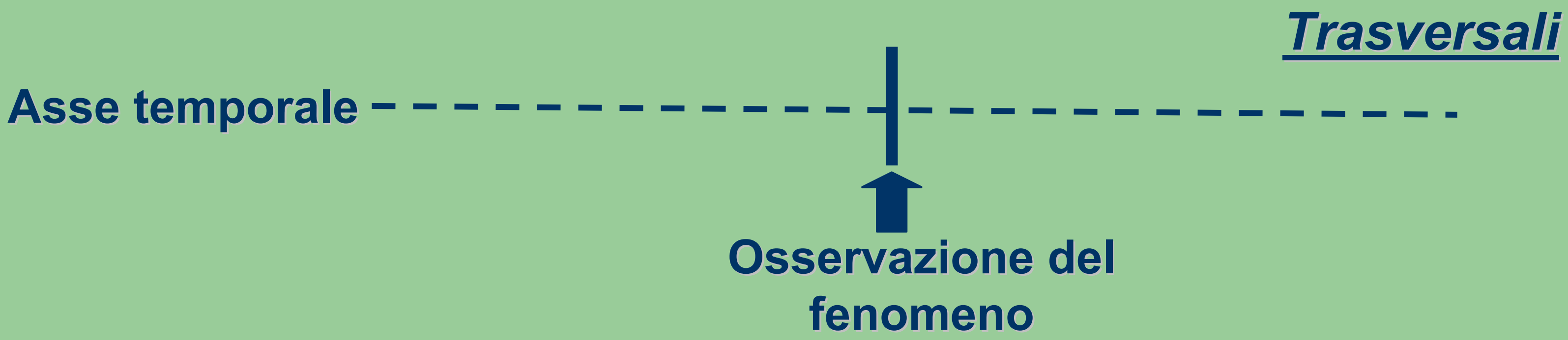
**Nel definire il disegno di ricerca, il ricercatore deve valutare qual è la tipologia di studio più idonea rispetto allo scopo della ricerca e rispetto alle risorse e ai vincoli esistenti. Gli studi si distinguono in:**

- **Studi longitudinali (es. indagini panel)**: si prevedono per ogni unità di rilevazione più misurazioni nel tempo (misure ripetute). Si misura ad esempio come variano nel tempo le preferenze dei consumatori o gli atteggiamenti rispetto ad una data tematica, …).

**Gli studi longitudinali possono essere di due tipologie**:



- **Studi trasversali:** viene effettuata una unica misurazione nel tempo. Rappresentano gli studi più frequenti.



## Gli studi si distinguono anche in:

### Studi osservazionali

**Osservazione passiva di una realtà**

**Es. Si rilevano le opinioni di un campione di consumatori circa due diverse campagne pubblicitarie.**

**Ridotta formalizzazione metodologica**

**Problemi:**
-problematicità dell'applicazione di procedure statistiche inferenziali

**Vantaggi:**
-possibilità di studiare problematiche non esaminabili tramite studi sperimentali.

### Studi sperimentali

**Intervento sperimentale del ricercator**

**Es. Per valutare l'efficacia di due divers campagne pubblicitarie, si selezionan casualmente due campioni c consumatori a ciascuno dei quali vien somministrata in modo random un campagna.**

**Approccio metodologico rigoroso**

**Vantaggi:**
-diretta applicazione delle procedure statistiche inferenziali

**Problemi:**
-fattibilità dello studio / appropriatezza de metodo.

# Introduzione alle indagini campionarie

**La conoscenza e la valutazione statistica di un fenomeno di interesse può essere derivata da:**

- **Rilevazioni complete** (es. censimenti)

  Non sempre attuabili per problemi di costo, tempi non rapidi di ottenimento dati, distruzione delle unità osservate (es. controlli di qualità), …

- **Rilevazioni parziali** (indagini campionarie)

**Sia le rilevazioni complete che quelle parziali possono fornire informazioni distorte della realtà a causa di:**

- Distorsioni delle informazioni rilevate legate a:
  - ✓ Comportamento del rilevatore (es. quesiti proposti con orientamento alla risposta, …)
  - ✓ Comportamento del rispondente (es. ritrosia nelle risposte, scarso scrupolo nel fornire le risposte, …)
- Copertura non totale delle unità o presenza di dati mancanti
- Errori nella fase di registrazione, codifica, elaborazione dei dati

**Per le rilevazioni parziali si aggiunge l'errore di campionamento.**

Tale errore può essere stimato o meno a seconda che il campione sia **casuale** o **non casuale**.

# Introduzione al campionamento

## Campioni casuali o probabilistici

**Consentono la stima dell'errore di campionamento e la "bontà" dei risultati.**

**Le unità dello studio sono selezionate con procedimenti equivalenti alla scelta di elementi da un insieme omogeneo, secondo un criterio che non comporti la selezione preferenziale di un particolare insieme di unità.**

**La selezione casuale può essere effettuata in "blocco", ossia senza ripetizione delle unità che comporranno il campione, oppure con ripetizione.**

**Per applicare un procedimento di estrazione casuale occorre che sia disponibile una lista completa di tutte le unità della popolazione o collettivo da cui si seleziona il campione. Ogni unità del collettivo ha quindi una probabilità nota di essere inclusa nel campione.**

## Alcune tipologie di campionamento casuale:

- **Campionamento casuale semplice**
- **Campionamento casuale stratificato**
- **Campionamento a cluster o a grappoli**
- **Campionamento sistematico**

## Campione casuale semplice

Un campione casuale semplice di ampiezza *n* da una popolazione finita di ampiezza *N* è un campione selezionato in modo tale che ciascuna unità ha la stessa probabilità di essere estratta.

Dalla lista degli elementi della popolazione la procedura di selezione si basa sull'uso dei numeri casuali, ad esempio mediante le tavole dei numeri casuali.

Tali tavole contengono sequenze di numeri ottenuti con una selezione casuale ad esempio mediante elaboratore elettronico.

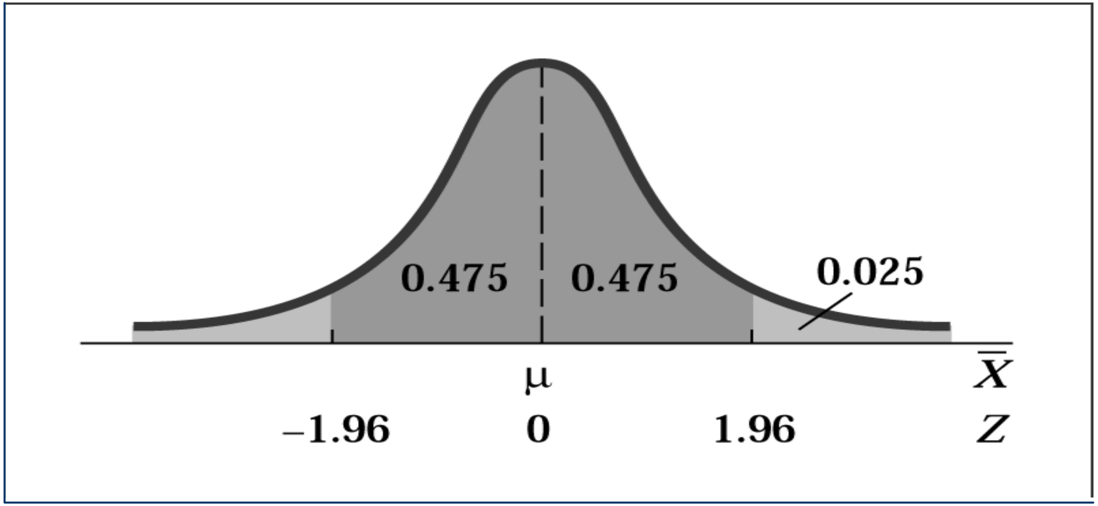
## Campione casuale semplice:

### Stima della media della popolazione

Se il campione casuale semplice selezionato è sufficientemente ampio ($n \geq 30$), il Teorema del Limite Centrale consente di concludere che la distribuzione campionaria della media $\bar{x}$ (stimatore puntuale della media della popolazione $\mu$), può essere approssimata da una distribuzione Normale di media $\mu$ e deviazione standard $\sigma_{\bar{x}}$ (errore standard della media).

**Intervallo di confidenza per $\mu$ al (1-$\alpha$)%:** $\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$

***Es*. Valori critici $\pm z_{\alpha/2}$ per un intervallo di confidenza al 95%**



# Campione casuale semplice :

## Stima della media della popolazione

Se selezioniamo un campione casuale semplice di ampiezza *n* da una popolazione finita di ampiezza *N*, uno stimatore dell’errore standard della media $\sigma_{\bar{x}}$ è:

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

Dove $s$ è la deviazione standard campionaria.

**Utilizzando $s_{\bar{x}}$ come stimatore di $\sigma_{\bar{x}}$ un intervallo di confidenza per la media della popolazione $\mu$ al (1-$\alpha$)% diventa:**

$$\bar{x} \pm z_{\alpha/2} s_{\bar{x}}$$

## Campione casuale semplice :

### Stima della media della popolazione

***Esempio***: Una rivista specializzata sulle navi e la pesca ha 8000 abbonati. In un campione casuale semplice di $n$=484 abbonati il reddito medio annuo risulta pari a 30.500 $ con una deviazione standard di 7.040 $.

Una stima puntuale non distorta del reddito medio annuale di tutti gli abbonati è data quindi dalla media campionaria $\bar{x}$= 30.500 $.

Una stima dell'errore standard della media sarà pari a :

$$s_{\bar{x}} = \sqrt{\frac{8000-484}{8000}}\left(\frac{7040}{\sqrt{484}}\right) = 310$$

**Utilizzando $s_{\bar{x}}$ come stimatore di $\sigma_{\bar{x}}$ un intervallo di confidenza per il reddito annuale medio degli abbonati alla rivista $\mu$ al 95% è pari a:**

$$30.500 \pm 1{,}96 \times 310 = 30.500 \pm 608 = (29.892\ \$, 31.108\ \$)$$

**Il valore 1,96x310=608 rappresenta il limite dell'errore di campionamento**

## Campione casuale semplice :

### Stima del totale della popolazione

Se selezioniamo un campione casuale semplice di ampiezza *n* da una popolazione finita di ampiezza $N$, indicando con $X$ il totale della popolazione, uno stimatore puntuale di $X$ è dato da:

$$\hat{X} = N\,\bar{x}$$

Uno stimatore dell'errore standard della stima puntuale è dato da:

$$s_{\hat{X}} = N s_{\bar{x}} \qquad \text{dove} \qquad s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

**Un intervallo di confidenza per il totale della popolazione $X$ al (1-$\alpha$)% è dato da:**

$$N\,\bar{x} \pm z_{\alpha/2} s_{\hat{X}}$$

## Campione casuale semplice :

### Stima del totale della popolazione

***Esempio***: Da una popolazione di 500 scuole è selezionato un campione casuale semplice di $n$=50 scuole pubbliche. La superficie media risulta pari nel campione a 22.000 piedi al quadrato con una deviazione standard di 4.000 piedi al quadrato.

Una stima puntuale della superficie totale di tutte le scuole è data da:

$$\hat{X} = 500 \times 22.000 = 11.000.000$$

Una stima dell'errore standard di $\hat{X}$ sarà pari a :

$$s_{\hat{X}} = 500 \times \left( \sqrt{\frac{500-50}{500}} \left( \frac{4000}{\sqrt{50}} \right) \right) =$$

$$s_{\hat{X}} = 500 \times 536{,}66 = 268.330$$

**Un intervallo di confidenza per la superficie totale di tutte le scuole $X$ al 95% è pari a:**

$$11.000.000 \pm 1{,}96 \times 268.330 = 11.000.000 \pm 525.927$$

$= (10.474.073, 11.525.927)$ **piedi al quadrato**

## Campione casuale semplice :

### Stima di una proporzione della popolazione

Se selezioniamo un campione casuale semplice di ampiezza *n* da una popolazione finita di ampiezza $N$, indicando con $p$ la frazione degli elementi della popolazione che hanno una determinata caratteristica, uno stimatore puntuale non distorto di $p$ è dato dalla proporzione campionaria $\overline{p}$ .

Uno stimatore dell'errore standard della proporzione è dato da:

$$s_{\overline{p}} = \sqrt{\frac{N-n}{N}\left(\frac{\overline{p}(1-\overline{p})}{n-1}\right)}$$

**Un intervallo di confidenza per la proporzione della popolazione $p$ al (1-$\alpha$)% è dato da:**

$$\overline{p} \pm z_{\alpha/2} s_{\overline{p}}$$

## Campione casuale semplice :

### Stima di una proporzione della popolazione

***Esempio***: Da una popolazione di 500 scuole è selezionato un campione casuale semplice di $n$=50 scuole pubbliche. Nel campione, delle 50 scuole, 35 indicano l'uso di gas naturale per le cucine.

Una stima puntuale della proporzione di tutte le scuole che usano gas naturale è data da:

$$\overline{p} = 35/50 = 0{,}70$$

Una stima dell'errore standard di $\overline{p}$ sarà pari a:

$$s_{\overline{p}} = \sqrt{\frac{500-50}{500}\left(\frac{0{,}7(1-0{,}7)}{50-1}\right)} = 0{,}0621$$

**Un intervallo di confidenza per la proporzione $p$ di scuole che usano gas naturale nell'insieme di tutte le scuole, al 95% è pari a:**

$$0{,}7 \pm 1{,}96 \times 0{,}0621 = 0{,}7 \pm 0{,}1217$$

$$= (0{,}5783; 0{,}8217)$$

## Campione casuale semplice :

### Calcolo dell'ampiezza campionaria per la stima della media

Un approccio usuale prevede la specificazione della precisione desiderata per le stime dei parametri di interesse.

Ad esempio, riprendendo l'espressione per la stima dell'errore standard della media:

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

Indichiamo con $B$ il limite dell'errore di campionamento:

$$B = z_{\alpha/2} s_{\bar{x}} = z_{\alpha/2}\sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

Risolvendo rispetto ad $n$:

$$n = \frac{Ns^2}{N\left(\frac{B^2}{2z_{\alpha/2}}\right) + s^2}$$

**Scegliendo quindi un valore per $B$ e stimando $s^2$ da precedenti indagini o da uno studio pilota, è possibile quindi calcolare l'ampiezza campionaria $n$.**

## Campione casuale semplice :

### Calcolo dell'ampiezza campionaria per la stima della media

***Esempio***: Per una popolazione di 5000 laureati di una università, si vuole stimare la media del salario iniziale, calcolando un intervallo di confidenza al 95% ponendo $B$ = 500 \$. Da uno studio condotto l'anno precedente si ottiene la stima di $s^2$ pari a 3.000 \$. Si vuole determinare l'ampiezza del campione.

$$n = \frac{5000 \times (3000)^2}{5000 \times \left( \frac{(500)^2}{2 \times 1{,}96} \right) + (3000)^2} = 137{,}2$$

**Arrotondando per eccesso il valore ottenuto, si ottiene che un campione casuale di ampiezza pari a 140 laureati, fornirà un intervallo di confidenza al 95% della media del salario iniziale, con una precisione (ampiezza dell'intervallo di confidenza: 2$B$) pari a 1000 \$. Per cautelarsi ulteriormente si può arrotondare il risultato a 150 laureati.**

## Campione casuale semplice :

**Calcolo dell'ampiezza campionaria per la stima del totale**
L'espressione per il calcolo della ampiezza campionaria necessaria per stimare il totale della popolazione con una data precisione, è data da:

$$n = \frac{Ns^2}{\left(\frac{B^2}{2z_{\alpha/2}N}\right) + s^2}$$

***Esempio***: Riprendendo l'esempio precedente, per la popolazione di 5000 laureati di una università, si vuole stimare il totale del salario iniziale, calcolando un intervallo di confidenza al 95% ponendo $B$ = 2.000.000 $. Si vuole determinare l'ampiezza del campione.

$$n = \frac{5000 \times (3000)^2}{\left(\frac{(2.000.000)^2}{2 \times 1{,}96 \times 5000}\right) + (3000)^2} 211{,}2$$

**Arrotondando per eccesso il valore ottenuto, si ottiene che un campione casuale di ampiezza pari a 215 laureati, fornirà un intervallo di confidenza al 95% del totale del salario, con una precisione (ampiezza dell'intervallo di confidenza: 2*B*) pari a 4.000.000 $.**

## <u>Campione casuale semplice</u> :

**Calcolo dell'ampiezza campionaria per la stima della proporzione**

L'espressione per il calcolo della ampiezza campionaria necessaria per stimare la proporzione della popolazione con una data precisione, è data da:

$$n = \frac{N\,\bar{p}(1-\bar{p})}{N\left(\dfrac{B^2}{2z_{\alpha/2}}\right) + \bar{p}(1-\bar{p})}$$

**Se non si hanno stime per la proporzione campionaria, è possibile utilizzare $\bar{p}$ = 0,5.**

# Campione casuale stratificato

La popolazione è inizialmente divisa in $H$ gruppi detti strati. Da ciascuno strato $h$ è quindi selezionato un campione casuale di ampiezza $n_h$.

Se la variabilità all'interno degli strati è minore della variabilità tra gli strati, il campionamento stratificato può condurre ad una precisione maggiore delle stime.

Ad esempio, una università vuole condurre una indagine sul salario iniziale dei propri laureati. Si vuol tener conto del fatto che sono attivi cinque indirizzi di studi: Accounting, Finance, Information system, Marketing e Operations management. La popolazione dei laureati sarà quindi divisa in cinque strati in corrispondenza dei diversi indirizzi e da ciascuno strato verrà selezionato un campione casuale di laureati.

# Campione casuale stratificato:

## Stima della media della popolazione

Uno stimatore non distorto della media della popolazione $\mu$ è ottenuto mediante una media ponderata delle medie campionarie dei singoli strati con pesi pari alle frazioni della popolazione in ciascuno strato:

$$\bar{x}_{st} = \sum_{h=1}^{H} \left( \frac{N_h}{N} \right) \bar{x}_h$$

$H =$ **Numero di strati**

$\bar{x}_h =$ **Media campionaria per lo strato *h***

$N_h =$ **Numero di elementi nella popolazione nello strato *h***

$N =$ **Numero totale di elementi nella popolazione:** $N = N_1 + \ldots + N_h + \ldots + N$

## Campione casuale stratificato:

### Stima della media della popolazione

Uno stimatore dell'errore standard della media è dato da:

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h\left(N_h - n_h\right)\frac{s_h^2}{n_h}}$$

Dove $s_h$ è la deviazione standard campionaria per lo strato $h$, ed $n_h$ è la numerosità campionaria dello strato $h$.

**Un intervallo di confidenza per la media della popolazione $\mu$ al (1-$\alpha$)% è dato da:**

$$\bar{x} \pm z_{\alpha/2} s_{\bar{x}_{st}}$$

## Campione casuale stratificato:

### Stima della media della popolazione

***Esempio***: Per una popolazione di 1500 laureati di un college universitario, si vuole stimare la media del salario iniziale. Nell'ambito del college sono previsti 5 indirizzi diversi (strati). Si seleziona un campione stratificato di $n$ =180 studenti, ottenendo i seguenti risultati:

| Strati | $\bar{x}_h$ | $s_h$ | $N_h$ | $n_h$ |
|---|---|---|---|---|
| Accounting | 35.000 $ | 2.000 $ | 500 | 45 |
| Finance | 33.500 $ | 1.700 $ | 350 | 40 |
| Information system | 41.500 $ | 2.300 $ | 200 | 30 |
| Marketing | 32.000 $ | 1.600 $ | 300 | 35 |
| Operations management | 36.000 $ | 2.250 $ | 150 | 30 |

## Campione casuale stratificato:

### Stima della media della popolazione

***Esempio***: per la stima puntuale della media campionaria si procede come segue:

$$\bar{x}_{st} = \sum_{h=1}^{H} \left( \frac{N_h}{N} \right) \bar{x}_h$$

$$\bar{x}_{st} = \frac{500}{1.500} \times 35.000 + \frac{350}{1.500} \times 33.500 + \frac{200}{1.500} \times 41.500 +$$

$$+ \frac{300}{1.500} \times 32.000 + \frac{150}{1.500} \times 36.000 = 35.017$$

# Campione casuale stratificato:

## Stima della media della popolazione

***Esempio***: per la stima dell'errore standard, si procede come segue:

| Strati | *h* | $N_h(N_h - n_h)\dfrac{s_h^2}{n_h}$ |
|---|---|---|
| **Accounting** | **1** | $500(500-45)\dfrac{(2000)^2}{45} = 20.222.222.222$ |
| **Finance** | **2** | $350(350-40)\dfrac{(1700)^2}{40} = 7.839.125.000$ |
| **Information system** | **3** | $200(200-30)\dfrac{(2300)^2}{30} = 5.995.333.333$ |
| **Marketing** | **4** | $300(300-35)\dfrac{(1600)^2}{35} = 5.814.857.143$ |
| **Operations management** | **5** | $150(150-30)\dfrac{(2250)^2}{30} = 3.037.500.000$ |

# Campione casuale stratificato:

## Stima della media della popolazione

***Esempio***:

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h (N_h - n_h)\frac{s_h^2}{n_h}}$$

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{(1500)^2} \times 42.909.037.698} = 138$$

**L'intervallo di confidenza per il salario medio iniziale $\mu$ della popolazione degli studenti del college al 95% è dato da:**

$$35.017 \pm 1{,}96 \times 138 = 35.017 \pm 270$$

$$= (34.747\ \$,\ 35.287\ \$)$$

## Campione casuale stratificato :

### Stima del totale della popolazione

Indicando con $X$ il totale della popolazione, uno stimatore puntuale di $X$ è dato da:

$$\hat{X} = N\,\bar{x}_{st}$$

Uno stimatore dell'errore standard della stima puntuale è dato da:

$$s_{\hat{X}} = N s_{\bar{x}_{st}} \qquad \text{dove} \qquad s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h\left(N_h - n_h\right)\frac{s_h^2}{n_h}}$$

**Un intervallo di confidenza per il totale della popolazione $X$ al (1-$\alpha$)% è dato da:**

$$N\,\bar{x}_{st} \pm z_{\alpha/2}\, s_{\hat{X}}$$

## Campione casuale stratificato :

### Stima del totale della popolazione

***Esempio***: Riprendendo l'esempio precedente, per la popolazione di 1500 laureati di un college universitario, si vuole stimare il totale del salario iniziale:

Una stima puntuale del salario iniziale di tutti i laureati è data da:

$$\hat{X} = 1500 \times 35.017 = 52.525.500\,\$$$

Una stima dell'errore standard di $\hat{X}$ sarà pari a :

$$s_{\hat{X}} = 1500 \times 138 = 207.000$$

**L'intervallo di confidenza per il salario iniziale totale dei laureati $X$ al 95% è pari a:**

$$52.525.500 \pm 1{,}96 \times 207.000 = 52.525.500 \pm 405.720$$

$$= (52.119.780\,\$,\ 52.931.220\,\$)$$

## Campione casuale stratificato :

### Stima di una proporzione della popolazione

Uno stimatore non distorto di una proporzione $p$ della popolazione è ottenuta mediante una media ponderata delle proporzioni campionarie dei singoli strati con pesi pari alle frazioni della popolazione in ciascuno strato:

$$\bar{p}_{st} = \sum_{h=1}^{H} \left( \frac{N_h}{N} \right) \bar{p}_h$$

$H =$ **Numero di strati**

$\bar{p}_h =$ **Proporzione campionaria per lo strato** $h$

$N_h =$ **Numero di elementi nella popolazione nello strato** $h$

$N =$ **Numero totale di elementi nella popolazione:** $N = N_1 + \ldots + N_h + \ldots + N$

## Campione casuale stratificato :

### Stima di una proporzione della popolazione

Uno stimatore dell'errore standard della proporzione è dato da:

$$s_{\overline{p}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\overline{p}_h\left(1-\overline{p}_h\right)}{n_h - 1}\right]}$$

**Un intervallo di confidenza per la proporzione della popolazione $p$ al (1-$\alpha$)% è dato da:**

$$\overline{p}_{st} \pm z_{\alpha/2} s_{\overline{p}_{st}}$$

## Campione casuale stratificato :

### Stima di una proporzione della popolazione

***Esempio***: Riprendendo l'esempio precedente, per la popolazione di 1500 laureati di un college universitario, si vuole stimare la proporzione di laureati che ricevono un salario iniziale maggiore o uguale a 36.000 $:

Dal campione di 180 laureati, la stima puntuale dei laureati con salario iniziale maggiore o uguale di 36.000 $ è ottenuta procedendo come segue:

| Strati | Laureati con salario ≥ 36.000 $ nel campione | $N_h$ | $n_h$ |
|---|---|---|---|
| Accounting | 16 | 500 | 45 |
| Finance | 3 | 350 | 40 |
| Information system | 29 | 200 | 30 |
| Marketing | 0 | 300 | 35 |
| Operations management | 15 | 150 | 30 |

## Campione casuale stratificato :

### Stima di una proporzione della popolazione

***Esempio***:

$$\overline{p}_{st} = \left(\frac{500}{1500} \times \frac{16}{45}\right) + \left(\frac{350}{1500} \times \frac{3}{40}\right) + \left(\frac{200}{1500} \times \frac{29}{30}\right) + $$

$$+ \left(\frac{300}{1500} \times \frac{0}{35}\right) + \left(\frac{150}{1500} \times \frac{15}{30}\right) = 0{,}3149$$

# Campione casuale stratificato :

## Stima di una proporzione della popolazione

***Esempio***: per la stima dell'errore standard, si procede come segue:

| Strati | *h* | $N_h(N_h - n_h)\left[\dfrac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]$ |
|---|---|---|
| **Accounting** | **1** | $500(500-45)\dfrac{(16/45)(29/45)}{45-1} = 1184{,}7363$ |
| **Finance** | **2** | $350(350-40)\dfrac{(3/40)(37/40)}{40-1} = 193{,}0048$ |
| **Information system** | **3** | $200(200-30)\dfrac{(29/30)(1/30)}{30-1} = 37{,}7778$ |
| **Marketing** | **4** | $300(300-35)\dfrac{(0/35)(35/35)}{35-1} = 0$ |
| **Operations management** | **5** | $150(150-30)\dfrac{(15/30)(15/30)}{30-1} = 155{,}1724$ |

## Campione casuale stratificato :

### Stima di una proporzione della popolazione

***Esempio***:

$$s_{\overline{p}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\overline{p}_h(1-\overline{p}_h)}{n_h - 1}\right]}$$

$$s_{\overline{p}_{st}} = \sqrt{\frac{1}{(1500)^2}\times 1570{,}6913} = 0{,}0264$$

**L'intervallo di confidenza per la proporzione della popolazione $p$ al 95% è dato da:**

$$0{,}3149 \pm 1{,}96 \times 0{,}0264$$

$$= (0{,}2632\,,\,0{,}3666)$$

## Campione casuale stratificato :

### Calcolo dell'ampiezza campionaria

Un approccio prevede la scelta dell'ampiezza campionaria globale *n* e successivamente si decide come assegnare le unità campionarie ai vari strati.

Alternativamente si può procedere fissando la numerosità campionaria di ciascuno strato e ottenendo per somma sugli strati, la numerosità globale.

Nel primo caso, l'ampiezza campionaria globale *n* è scelta in modo tale da ottenere con una data precisione, le stime dei parametri di interesse. L'allocazione successiva delle unità campionarie ai vari strati può essere guidata dai seguenti fattori:

- Il numero di elementi della popolazione in ciascuno strato
- La varianza degli elementi in ciascuno strato
- Il costo della selezione degli elementi in ciascuno strato

## Campione casuale stratificato :
## Calcolo dell'ampiezza campionaria

Ponendo che il costo di selezione delle unità per i vari strati sia approssimativamente lo stesso, possiamo applicare il metodo dell'allocazione di Neyman:

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right)$$

Dove $n$ è stato scelto in modo da ottenere un prefissato livello di precisione ($2B$=ampiezza dell'intervallo di confidenza) nella stima dei parametri di interesse. Ad esempio:

**Per la stima della media della popolazione**

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{2z_{\alpha/2}}\right) + \sum_{h=1}^{H} N_h s_h^2}$$

**Per la stima del totale della popolazione**

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{2z_{\alpha/2}} + \sum_{h=1}^{H} N_h s_h^2}$$

**Per la stima di una proporzio della popolazione**

$$n = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{\bar{p}_h(1-\bar{p}_h)}\right)^2}{N^2\left(\frac{B^2}{2z_{\alpha/2}}\right) + \sum_{h=1}^{H} N_h \bar{p}_h (1-\bar{p}}$$

## Campione casuale stratificato :

### Calcolo dell'ampiezza campionaria per la media della popolazione

***Esempio***: un rivenditore di Chevrolet vuole stimare il reddito medio mensile dei clienti che hanno acquistato una Corvette, una Geo Prizm o una Cavalier (strati), mediante un intervallo di confidenza al 95%, con una precisione tale che $B$=100 \$. I dati a disposizione e le deviazioni standard campionarie stimate da uno studio pilota sono riportati di seguito:

| Strati | $s_h$ | $N_h$ |
|---|---|---|
| Corvette | 1300 \$ | 500 |
| Geo Prizm | 900 \$ | 350 |
| Cavalier | 500 \$ | 200 |

Il primo passo è il calcolo dell'ampiezza campionaria *n* globale per la stima del reddito medio mensile con la precisione richiesta.

## Campione casuale stratificato :

### Calcolo dell'ampiezza campionaria per la media della popolazione

***Esempio***:

$$\sum_{h=1}^{3} N_h s_h = 100 \times 1300 + 200 \times 900 + 300 \times 500 = 460.000$$

$$\sum_{h=1}^{3} N_h s_h^2 = 100 \times (1300)^2 + 200 \times (900)^2 + 300 \times (500)^2 = 406.000.000$$

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2 \left(\frac{B^2}{2 z_{\alpha/2}}\right) + \sum_{h=1}^{H} N_h s_h^2} = \frac{(406.000)^2}{(600)^2 \left(\frac{(100)^2}{2 \times 1{,}96}\right) + 406.000.000} = 160$$

## Campione casuale stratificato :

### Calcolo dell'ampiezza campionaria per la media della popolazione

***Esempio***:

A questo punto le unità campionarie vengono allocate nei vari strati nel modo seguente:

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right)$$

**Per lo strato dei clienti proprietari di una Corvette**

$$n_1 = 160\left(\frac{100 \times 1300}{460.000}\right)$$

$$n_1 = 45$$

**Per lo strato dei clienti proprietari di una Geo Prizm**

$$n_2 = 160\left(\frac{200 \times 900}{460.000}\right)$$

$$n_2 = 63$$

**Per lo strato dei clienti proprietari di una Cavalie**

$$n_3 = 160\left(\frac{300 \times 500}{460.000}\right)$$

$$n_3 = 52$$

## Campione casuale stratificato :

### Calcolo dell'ampiezza campionaria

Ponendo che sia il costo di selezione delle unità per i vari strati sia la varianza degli elementi nei vari strati, siano approssimativamente uguali, possiamo applicare il metodo dell'allocazione proporzionale:

$$n_h = n\left(\frac{N_h}{N}\right)$$

Dove *n* è stato scelto in modo da ottenere un prefissato livello di precisione ($2B$=ampiezza dell'intervallo di confidenza) nella stima dei parametri di interesse.

## Campionamento a cluster o grappoli

La popolazione è inizialmente divisa in *N* gruppi detti cluster che rappresentano delle unità ausiliarie che contengono le unità da rilevare (es. la popolazione di elettori viene divisa in base alle province di residenza che costituiscono i cluster).

Viene quindi selezionato un campione casuale di *n* cluster tra gli *N*. A questo punto si può procedere come segue:

**Campionamento ad uno stadio**: Tutte le unità di rilevazione contenute nel campione casuale di cluster costituiranno il campione finale.

**Campionamento a due stadi**: si selezionerà un campione casuale di unità di rilevazione da ogni cluster.

## Campionamento a cluster :

***Esempio***:

Una società di 1000 aziende certificatrici vuole condurre una indagine sui propri dipendenti che svolgono il ruolo di certificatori.

Il numero totale di dipendenti certificatori di tutte le aziende è pari a 12.000.

Per minimizzare i costi e i tempi dell'indagine, la società decide di applicare un campionamento a cluster ad uno stadio.

Vengono quindi selezionate casualmente 10 aziende. I dipendenti certificatori di tali aziende costituiranno il campione finale per l'indagine.

### Notazione:

$N =$ **Numero di cluster nella popolazione**

$n =$ **Numero di cluster selezionati nel campione casuale**

$M =$ **Numero totale di elementi nella popolazione**

$\overline{M} = M / N$ **Numero medio di elementi per cluster**

### Esempio:

$N = 1000$ **aziende**

$n = 10$ **aziende**

$M = 12.000$ **dipendenti**

$\overline{M} = 12.000 / 1000$

$\overline{M} = 12$ **dipendenti**

## Campionamento a cluster :

### Stima della media della popolazione:

Uno stimatore della media della popolazione $\mu$ è dato da:

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i}$$

$n =$ **Numero di cluster selezionati nel campione casuale**

$M_i =$ **Numero di elementi nel cluster *i***

$x_i =$ **Totale dei valori delle osservazioni nel cluster *i***

## Campionamento a cluster :

## Stima della media della popolazione:

Uno stimatore dell'errore standard della media è dato da:

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_c M_i\right)^2}{n-1}}$$

$N =$ **Numero di cluster nella popolazione**

$n =$ **Numero di cluster selezionati nel campione casuale**

$M_i =$ **Numero di elementi nel cluster *i***

$M =$ **Numero totale di elementi nella popolazione:** $M = M_1 + \ldots + M_i + \ldots + M_N$

$x_i =$ **Totale delle osservazioni nel cluster *i***

$\bar{M} = M / N$ **Numero medio di elementi per cluster**

**Un intervallo di confidenza per la media della popolazione $\mu$ al (1-$\alpha$)% è dato da:**

$$\bar{x}_c \pm z_{\alpha/2} s_{\bar{x}_c}$$

## Campionamento a cluster :

## Stima della media della popolazione:

***Esempio***: riprendendo l'esempio precedente, si vuole stimare il salario medio dei dipendenti utilizzando il campione a cluster definito precedentemente.

| Aziende ($i$) | Dipendenti ($M_i$) | Salario totale ($x_i$) (in1.000 $) | $\left(x_i - \bar{x}_c M_i\right)^2$ |
|---|---|---|---|
| 1 | 8 | 384 | $[384 – 51{,}250(8)]^2 = 676{,}000$ |
| 2 | 25 | 1.350 | $[1350 - 51{,}250(25)]^2 = 4.726{,}563$ |
| 3 | 4 | 148 | $[148 - 51{,}250(4)]^2 = 3.249{,}000$ |
| 4 | 17 | 857 | $[857 - 51{,}250(17)]^2 = 203{,}063$ |
| 5 | 7 | 296 | $[296 - 51{,}250(7)]^2 = 3.937{,}563$ |
| 6 | 3 | 131 | $[131 - 51{,}250(3)]^2 = 517{,}563$ |
| 7 | 15 | 761 | $[761 - 51{,}250(15)]^2 = 60{,}063$ |
| 8 | 4 | 176 | $[176 - 51{,}250(4)]^2 = 841{,}000$ |
| 9 | 12 | 577 | $[577 - 51{,}250(12)]^2 = 1.444{,}000$ |
| 10 | 33 | 1.880 | $[1880 - 51{,}250(33)]^2 = 35.626{,}563$ |
| **Totale** | **128** | **6.560** | **51.281,378** |

## Campionamento a cluster :
## Stima della media della popolazione:

***Esempio***:

La stima del salario medio del totale dei dipendenti $\mu$ è data da:

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} = \frac{6560}{128} = 51{,}250 \text{ migliaia di dollari}$$

La stima dell'errore standard della media è dato da:

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_c M_i\right)^2}{n-1}} = \sqrt{\left(\frac{1000-10}{1000\times 10\times 12^2}\right)\frac{51.281{,}378}{10-1}} = 1{,}98$$

**L'intervallo di confidenza per il salario medio al (1-$\alpha$)% è dato da:**

$$\bar{x}_c \pm z_{\alpha/2} s_{\bar{x}_c} = 51{,}25 \pm 1{,}96 \times 1{,}98 = 51{,}25 \pm 3{,}88$$

$$= (47{,}37 \text{ migliaia di \$; } 55{,}13 \text{ migliaia di \$})$$

## Campionamento a cluster :

### Stima del totale della popolazione:

Uno stimatore del totale della popolazione $X$ è dato da:

$$\hat{X} = M\bar{x}_c$$

$M$ = **Numero totale di elementi nella popolazione:** $M = M_1 + \ldots + M_i + \ldots + M_N$

$\bar{x}_c$ = **Stimatore della media**

Uno stimatore dell'errore standard della stima è dato da:

$$s_{\hat{X}} = Ms_{\bar{x}_c}$$

**Un intervallo di confidenza per il totale della popolazione** $X$ **al (1-**$\alpha$**)% è dato da:**

$$\hat{X} \pm z_{\alpha/2} s_{\hat{X}}$$

## Campionamento a cluster :

### Stima del totale della popolazione:

***Esempio***: riprendendo l'esempio precedente, si vuole stimare il salario totale dei dipendenti utilizzando il campione a cluster definito precedentemente.

La stima del salario totale di tutti i dipendenti $X$ è data da:

$$\hat{X} = M\overline{x}_c = 12.000 \times 51{,}250 = 615.000 \text{ migliaia di \$}$$

La stima dell'errore standard è data da:

$$s_{\hat{X}} = M s_{\overline{x}_c} = 12.000 \times 1{,}979 = 23.748 \text{ migliaia di \$}$$

**L'intervallo di confidenza per il salario totale dei dipendenti $X$ al 95% è dato da:**

$$\hat{X} \pm z_{\alpha/2} s_{\hat{X}} = 615.000 \pm 1{,}96 \times 23.748 = 615.000 \pm 46.546{,}08$$

$$= (568.453{,}92; 661.546{,}08) \text{ migliaia di \$}$$

## Campionamento a cluster :

### Stima di una proporzione della popolazione:

Uno stimatore di una proporzione $p$ della popolazione è ottenuta nel modo seguente:

$$\overline{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i}$$

$n =$ **Numero di cluster nella selezionate nel campione casuale**

$M_i =$ **Numero di elementi nel cluster *i***

$a_i =$ **Numero di elementi nel cluster *i* con una data caratteristica di interesse**

## Campionamento a cluster :
## Stima di una proporzione della popolazione:

Uno stimatore dell'errore standard per la proporzione è dato da:

$$s_{\overline{p}_c} = \sqrt{\left(\frac{N-n}{Nn\overline{M}^2}\right)\frac{\sum_{i=1}^{n}\left(a_i - \overline{p}_c M_i\right)^2}{n-1}}$$

$N =$ **Numero di cluster nella popolazione**

$n =$ **Numero di cluster selezionati nel campione casuale**

$M_i =$ **Numero di elementi nel cluster *i***

$M =$ **Numero totale di elementi nella popolazione:** $M = M_1 + \ldots + M_i + \ldots + M_N$

$a_i =$ **Numero di elementi nel cluster *i* con una data caratteristica di interesse**

$\overline{M} = M / N$ **Numero medio di elementi per cluster**

**Un intervallo di confidenza per la proporzione della popolazione** $p$ **al (1-**$\alpha$**)% è dato da:**

$$\overline{p}_c \pm z_{\alpha/2} s_{\overline{p}_c}$$

## Campionamento a cluster :
## Stima di una proporzione della popolazione:

***Esempio***: riprendendo l'esempio precedente, si vuole stimare la proporzione di dipendenti donne utilizzando il campione a cluster definito precedentemente.

| Aziende ($i$) | Dipendenti ($M_i$) | Dipendenti donne ($a_i$) | $(a_i - \overline{p}_c M_i)^2$ |
|---|---|---|---|
| 1 | 8 | 2 | $[2 – 0,2734(8)]^2$ =0,0350 |
| 2 | 25 | 8 | $[8 - 0,2734(25)]^2$ = 1,3572 |
| 3 | 4 | 0 | $[0 - 0,2734(4)]^2$ =1,1960 |
| 4 | 17 | 6 | $[6 - 0,2734(17)]^2$ = 1,8284 |
| 5 | 7 | 1 | $[1 - 0,2734(7)]^2$ = 0,8350 |
| 6 | 3 | 2 | $[2 - 0,2734(3)]^2$ = 1,3919 |
| 7 | 15 | 2 | $[2 - 0,2734(15)]^2$ = 4,4142 |
| 8 | 4 | 0 | $[0 - 0,2734(4)]^2$ = 1,1960 |
| 9 | 12 | 5 | $[5 - 0,2734(12)]^2$ = 2,9556 |
| 10 | 33 | 9 | $[9 - 0,2734(33)]^2$ = 0,0005 |
| **Totale** | **128** | **35** | **15,2098** |

## Campionamento a cluster :

## Stima di una proporzione della popolazione:

***Esempio***:

La stima della proporzione di dipendenti donne $p$ è data da:

$$\bar{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} = \frac{2+8+\ldots+9}{8+25+\ldots+33} = \frac{35}{128} = 0{,}2734$$

La stima dell'errore standard è data da:

$$s_{\bar{p}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}\left(a_i - \bar{p}_c M_i\right)^2}{n-1}} = \sqrt{\left(\frac{1000-10}{1000\times 10\times 12^2}\right)\frac{15{,}2098}{10-1}} = 0{,}0341$$

**L'intervallo di confidenza per la proporzione di dipendenti donne al 95% è dato da:**

$$\bar{p}_c \pm z_{\alpha/2} s_{\bar{p}_c} = 0{,}2734 \pm 1{,}96 \times 0{,}0341 = 0{,}2734 \pm 0{,}0668$$

$$= (0{,}2066\,;\,0{,}3402)$$

## Campionamento a cluster :

### Calcolo dell'ampiezza campionaria

Una volta formati i cluster, il primo ordine di problema è la scelta del numero *n* di cluster da selezionare. La procedura per il campionamento a cluster è simile a quella precedentemente descritta per gli altri metodi di campionamento. Viene difatti deciso un appropriato livello di precisione per le stime dei parametri di interesse, scegliendo un valore per *B* che rappresenta il limite dell'errore di campionamento (2*B*=ampiezza dell'intervallo di confidenza).

L'ampiezza dei cluster nella popolazione e la varianza tra i cluster sono due fattori chiave per decidere quanti cluster includere nel campione casuale. Se i cluster sono simili, la varianza tra questi sarà bassa, per cui il numero di cluster da selezionare può essere ridotto.

Inoltre se l'ampiezza dei cluster è elevata, il numero dei cluster da selezionare può essere ridotto.

## Campionamento sistematico

Il campionamento sistematico consiste nello scegliere casualmente le unità dalla popolazione tramite una selezione che segue un modulo fisso, predeterminato a priori.

In genere, scelto casualmente il termine iniziale $r$ nell'intervallo $(1,N/n)$, si selezioneranno le successive unità applicando la procedura $r+ik$, con $i=0,1,\ldots,n-1$.

Ad esempio dovendo selezionare un campione di $n=50$ unità da una popolazione di $N=5000$ unità, verrà scelta casualmente la prima unità tra le prime $k=5000/50=100$ della lista, ad esempio la decima per cui $r=10$, si procederà a selezionare la seconda unità che sarà la $r+k=10+100=110^a$, la terza unità che sarà la $r+2xk=10+200=210^a$, e così via.

# Campioni non casuali o non probabilistici

**Non consentono la stima dell'errore di campionamento.**
**Tra gli schemi più frequenti:**

- Campioni ragionati

  Sono utilizzati ampiamente nelle ricerche di mercato finalizzate ad esempio alla valutazione delle potenzialità di un nuovo prodotto o di una campagna pubblicitaria in settori particolari della popolazione o in aree territoriali di particolare interesse.

- Campioni per quote

  Si seleziona la popolazione oggetto dell'indagine secondo alcune variabili strutturali (es. sesso, classi di età, professione, ecc.) rispettando delle "quote" proporzionali.

# Indagini campionarie basate su questionario

**Molte ricerche in vari campi di indagine, si prefiggono di individuare, valutare e spiegare alcune caratteristiche degli individui o di altre unità statistiche di interesse, utilizzando il *questionario* come modalità di raccolta dei dati.**

Si distinguono diverse tipologie di questionario:

- Questionario cartaceo
- Questionario informatizzato

  (Indagini CAPI – Computer Assisted Personal Interview)

  (Indagini CATI – Computer Assisted Telephone Interview)

Si distinguono diverse tipologie di compilazione del questionario:

- Autocompilazione o autosomministrazione (es. questionario postale)
- Compilazione mediante intervista diretta o telefonica
- Somministrazione mista mediante intervista e autocompilazione

# Profilo delle principali tecniche per le ricerche di mercato

## Intervista diretta

| Pregi | Difetti | Tendenza |
|---|---|---|
| Permette di:<br>-instaurare il miglior rapporto con gli intervistati;<br>-Utilizzare domande aperte e di controllo;<br>-Utilizzare materiale di supporto;<br>-Trattare molteplici temi nella stessa intervista;<br>-Indagare sulle cosidette “popolazioni clandestine” | Costi elevati.<br>Richiede affidabilità dei rilevatori.<br>Tempi lunghi di realizzazione. | Decremento |

## Sondaggio postale

| Pregi | Difetti | Tendenza |
|---|---|---|
| Consente di intervistare persone irraggiungibili in altro modo. | Autoselezione di coloro che restituiscono il questionario.<br>Tasso elevato di non risposta.<br>Tempi lunghi di realizzazione. | Notevole decremento |

## Sondaggio telefonico

| Pregi | Difetti | Tendenza |
|---|---|---|
| Massima rapidità.<br>Costi contenuti.<br>Consente di intervistare persone irraggiungibili in altro modo. | Esclusione dei non abbonati o di quelli con nominativo assente dagli elenchi.<br>Tasso elevato di non risposta. | Notevole incremento |

**Panel** (indagini condotte utilizzando campioni permanenti o continui, costituiti dalle medesime unità)

| **Pregi** | **Difetti** | **Tendenza** |
|---|---|---|
| Permette di:<br>-Stimare variazioni nel tempo;<br>-Ridurre le distorsioni legate al ricordo. | Costi elevati<br>Difficoltà di mantenere stabile il campione.<br>Chi vi partecipa può essere soggetto a condizionamenti. | Notevole incremento |

## Fonti di variazione o errore nella qualità dei dati raccolti

**Per il buon esito della rilevazione, la formulazione del questionario deve essere tale da dissipare resistenze, ambiguità, timori, incertezze interpretative e condizionamenti di ogni genere.**

Le principali fonti di variazione o errore sono riconducibili in generale a

- **Caratteristiche del questionario e del contesto in cui avviene la raccolta dati** (inadeguatezza delle domande, mancanza di privacy, presenza di rumore, …)
- **Caratteristiche del rispondente** (livello di istruzione, personalità, grado di motivazione, …)
- **Caratteristiche del rilevatore o dell’intervistatore** (professionalità, caratteristiche individuali, …)

# Fonti di variazione o errore nella qualità dei dati raccolti

I principali tipi di errore o distorsioni nelle risposte sono:

- **Errori motivazionali**: il rispondente non è sufficientemente motivato ad assumere il ruolo di "informatore", per cui tende a dare risposte incomplete o non accurate, rispondendo ad esempio "*non so*" anche quando egli possiede di fatto la risposta

- **Errori sociali**: il rispondente fornisce risposte alla scopo di fornire anche inconsapevolmente una certa immagine di sé

- **Errori di memoria**: il rispondente ha difficoltà a recuperare in memoria le informazioni richieste

- **Errori di comunicazione**: il rispondente ad esempio non ha compreso appieno la domanda o la tematica oggetto di indagine, o il rilevatore non ha compreso adeguatamente la risposta

- **Errori strutturali**: il questionario o l'intervista è troppo lunga o troppo breve, le categorie previste per la risposta sono inadeguate, l'ordine delle domande è inadeguato, …

# Contenuti di un questionario nelle indagini di mercato:

- **Fatti**

  (es. possesso di specifiche caratteristiche del rispondente, situazioni, avvenimenti, …)
- **Conoscenze**

  (informazioni che il rispondente possiede su un dato argomento)
- **Opinioni**

  (modo di pensare del rispondente su un dato argomento)
- **Atteggiamenti**

  (es. stili di vita)
- **Motivazioni**

  (informazioni che spiegano il comportamento e l’atteggiamento del rispondente)

**La difficoltà di reperire informazioni aumenta generalmente passando dai fatti alle motivazioni**

**La classificazione delle domande può essere intesa secondo diversi criteri:**

1)
- **Dirette o personali** (chiamano in causa direttamente il rispondente)
- **Indirette o impersonali** (informazioni che il rispondente fornisce su terzi)

2)
- **Semplici** (domande che prevedono una sola risposta)
- **Multiple** (domande che prevedono più risposte)

3)
- **Primarie o domande filtro** (domande che non dipendono da altre domande)
- **Secondarie** (domande che dipendono da altre domande)

4)
- **Aperte** (domande a risposta libera, non precodificata)
- **Chiuse** (domande a risposta preformulata)
- **Semichiuse** (prevedono la modalità "altro: specificare")

5)
- **Di controllo** (domande poste per verificare l'attendibilità di risposte fornite in precedenza)

**Le scale di misura sono gli strumenti per rilevare le modalità con le quali una variabile rilevata da una domanda può manifestarsi:**

- **Scala nominale** (la variabile si manifesta secondo due o più modalità qualitative non ordinabili)
- **Scala ordinale** (la variabile si manifesta secondo due o più modalità qualitative ordinabili)
- **Scala numerica** (la variabile si manifesta secondo modalità quantitative)

**Le modalità scelte come categorie di risposta ad una data domanda, devono essere mutuamente escludentesi ed esaustive**

**Le scale di valutazione sono gli strumenti per indagare su orientamenti ideologici, opinioni, atteggiamenti, su un dato argomento.**

Sono costituite da una batterie di asserzioni o domande (item) nei confronti delle quali gli intervistati devono precisare il proprio punto di vista in termini di accordo/disaccordo, soddisfazione/insoddisfazione, accettazione/rifiuto, etc.

Tra le più comuni:

**Scala ordinale di Likert che prevede che per ogni item siano rilevati diversi livelli di accordo/disaccordo secondo ad esempio il seguente schema:**

| Modalità | Equivalente numerico |
|---|---|
| Molto accordo | 5 |
| Abbastanza accordo | 4 |
| Né d'accordo né contrario | 3 |
| Abbastanza contrario | 2 |
| Molto contrario | 1 |

**Tali scale possono prevedere un numero diverso di modalità, in genere variabile da quattro a sette.**

La graduazione dei livelli può essere realizzata mediante opportuna verbalizzazione, o tramite simboli o in maniera grafica tramite una scala graduata sulla quale il rispondente riporterà la propria posizione e che sarà in seguito convertita in termini numerici:

**D'accordo**

**contrario**

**Esempio: Scala di Likert per la valutazione dell'atteggiamento della popolazione italiana nei confronti degli extracomunitari**

| *Items* | *Approvo decisamente* | *Approvo* | *Sono indeciso* | *Disapprovo* | *Disapprovo decisamente* |
|---|---|---|---|---|---|
| Gli extracomunitari tolgono lavoro agli italiani | 5 | 4 | 3 | 2 | 1 |
| Gli extracomunitari contribuiscono a peggiorare la qualità di vita degli italiani | 5 | 4 | 3 | 2 | 1 |
| Gli extracomunitari dovrebbero abitare in zone separate dagli italiani | 5 | 4 | 3 | 2 | 1 |

…

N.B. Barrare la casella che corrisponde al proprio atteggiamento

# Organizzazione dei dati raccolti

## La corretta organizzazione dei dati in banche dati (database)

**Esempio**: sono stati raccolti i dati relativi alla performance (1Yr$Ret=rendimento percentuale a un anno) di un campione di 194 fondi di investimento, suddivisi in 59 a capitalizzazione integrale (Object=1) e 135 misti (Object=2). Per una corretta ed efficace analisi statistica dei dati, essi devono essere strutturati secondo il seguente schema:

| N | Fund | 1Yr$Ret | Object |
|---|---|---|---|
| 1 | Alliance Capital A GrowInc | 30.8 | 2 |
| 2 | Berger SmCoGrow | 29.9 | 1 |
| 3 | Jurika & Voyles Kaufmann | 28.9 | 1 |
| 4 | Baron Funds BanRosSC | 35.5 | 2 |
| ... | .... | ... | ... |
| 192 | MainStay Inst MainPwrGr | 36.1 | 2 |
| 193 | Vanguard Index Inst | 30.9 | 2 |
| 194 | Vanguard Index 500 | 30.8 | 2 |

**Nome Variabili**

**Unità statistica**

Non devono esserci né righe né colonne completamente vuote. Se ci sono dei dati mancati essi vanno codificati in maniera appropriata (es. cella vuota).

# Regole di costruzione dei database: la codifica delle variabili

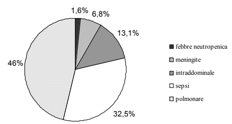
| Codice_cliente | Età | Sesso |
|---|---|---|
| 1 | 85 | Maschio |
| 2 | 72 anni | m |
| 3 | 81 | Femmina |
| 4 | 72 | maschio |
| 5 | ANNI 77 | F |
| 6 | 78 | 2 |
| 7 | 70 e 5 mesi | M |
| 8 | 74 | Femmina |
| … | … | … |

| Codice_cliente | Età | Sesso |
|---|---|---|
| 1 | 85 | Maschio |
| 2 | 72 | Maschio |
| 3 | 81 | Femmina |
| 4 | 72 | Maschio |
| 5 | 77 | Femmina |
| 6 | 78 | Femmina |
| 7 | 70 | Maschio |
| 8 | 74 | Femmina |
| … | … | … |

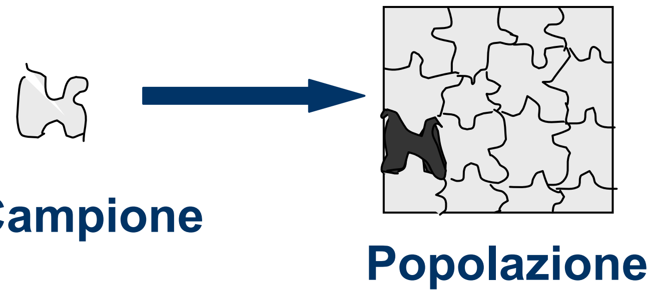
# PERCORSO DI ANALISI STATISTICA

## Analisi descrittiva del campione

| N. fattori predisponenti | N. pazienti |
|---|---|
| 1 | 273 |
| 2 | 188 |
| 3 | 106 |
| 4 | 56 |
| **Totale** | **623** |




## Analisi inferenziale




**Calcolo di statistiche descrittive: Frequenze, valori medi, indicatori di variabilità, di forma, …**

**Problemi di stima dei parametri, verifica di ipotesi sui parametri, modellistica, …**