

# Lecture consigliate

- S. Brasini, M. Freo, F. Tassinari, G. Tassinari, *Statistica aziendale e analisi di mercato*, 2002, Manuali, Il Mulino, Bologna  
**Capitoli 3 e 5**
- M. Barisone, R. Mannheim, *I sondaggi*, 1999, Il Mulino, Bologna
- M. Chiaro, *I sondaggi telefonici*, 1996, CISU, Roma

# Scopo: ottenere informazioni per ...

- analizzare bisogni, preferenze e comportamenti di acquisto dei consumatori
- valutare la spesa sostenuta mensilmente o annualmente dalle famiglie in riferimento a uno o più prodotti
- stimare la quota di mercato di una marca
- ...

**Strumento: indagini campionarie**

# Rilevazioni parziali o incomplete

Osservazione limitata a un sottoinsieme della popolazione oggetto di interesse

Le rilevazioni incomplete non sempre possono essere definite rilevazioni campionarie.

Mentre una qualunque rilevazione parziale può investire acriticamente una parte delle unità di una popolazione, per trarre informazioni su quest'ultima, la cosiddetta indagine campionaria considera sì solo una parte, ma selezionata con procedure che garantiscano la rappresentatività della popolazione da cui è estratto. Detta rappresentatività permette di estendere alla popolazione i risultati ottenuti con il campione.

# Indagini campionarie

- Inchieste
- Sondaggi di opinione

# Rilevazione campionaria vs Rilevazione censuaria

## Vantaggi della rilevazione campionaria

- modalità organizzative più flessibili
- contenimento costi di indagine sul campo
- maggiore rapidità raccolta ed elaborazione dati
- possibilità di approfondire l'analisi
- possibilità di garantire una maggiore accuratezza
- opportunità di concentrare le risorse sul controllo della qualità della rilevazione

# Rilevazione campionaria vs Rilevazione censuaria

## Svantaggi della rilevazione campionaria

- l'indagine censuaria restituisce il **valore vero dei parametri di interesse** (proporzioni, percentuali, medie, totali, ...)  
l'indagine campionaria ne fornisce solo una **stima**

Mediante i dati rilevati è possibile ottenere un **valore approssimato** dei parametri, al quale è associabile un grado di fiducia quantificabile se e solo se la formazione del campione risponde a criteri di tipo probabilistico.

# Modalità di selezione del campione

- scelta di comodo (*convenience sampling*)
- scelta ragionata (*judgmental sampling*)
- scelta casuale (*random sampling*)

Nella pratica quotidiana dei sondaggi di opinione e delle ricerche di mercato tutti e tre gli approcci sono ampiamente utilizzati

# Campionamento statistico

Da un insieme di unità costituenti l'oggetto di studio si estrae un **numero ridotto di casi**, con criteri che garantiscano la **rappresentatività della popolazione** da cui il campione è estratto.

Tale rappresentatività permette la **generalizzazione** dei risultati ottenuti con il campione **all'intera popolazione**.



# Campionamento di comodo (1 di 2)

La formazione di un campione di comodo prescinde completamente dalla selezione di un sottoinsieme di unità da una lista posta in corrispondenza con la popolazione obiettivo.

In genere sono gli intervistatori, se previsti dal piano di rilevazione, a decidere arbitrariamente chi avvicinare tra quanti condividono una medesima condizione, come ad esempio il trovarsi contemporaneamente in un dato luogo pubblico (bar, cinema, ...)

# Campionamento di comodo (2 di 2)

Nessuna particolare attenzione viene dedicata a evitare i rifiuti di collaborazione

I tempi rapidi dell'indagine privilegiano  
le **quantità** delle risposte  
alla  
**qualità** dei rispondenti.

# Campionamento presso esercizi commerciali

## Fonti distorsive (1 di 2)

- Punto vendita
  - gli intervistati risiedono spesso nello stesso quartiere
  - i frequentatori più assidui hanno una maggiore opportunità di esser intervistati
  - il tipo di negozio può non rappresentare l'intero universo degli utenti del sistema distributivo

# Campionamento presso esercizi commerciali

## Fonti distorsive (2 di 2)

- **Collocazione degli intervistatori**  
il luogo fisico dove avviene l'intervista (entrata, uscita, interno del negozio) influisce sulla disponibilità a collaborare e sulla qualità delle risposte
- **Tempo**  
il giorno e l'orario dell'intervista incidono anch'essi sull'andamento dell'indagine

Campioni di comodo = Campioni per caso

I risultati valgono in sé e per sé  
senza possibilità di generalizzazione  
alla popolazione  
o specificazione  
di una misura di attendibilità.

# Televoto

Rientra tra i campioni di comodo

- Partecipa chi vuole
- I dati ottenuti non sono riferibili a tutti gli spettatori, né tanto meno rappresentano l'opinione dell'intera popolazione
- Derivano da un **processo di autoselezione spontanea** e non dall'applicazione di un criterio di scelta di natura ragionata o probabilistica

# Campionamento ragionato

- L'individuazione degli intervistati è demandata completamente a chi predispose il piano di rilevazione dell'indagine
- Il loro apporto è ritenuto imprescindibile per lo svolgimento dell'indagine (testimoni privilegiati o *opinion leaders*)
- La dimensione del campione viene fissata di norma in base a criteri di pura convenienza
- Tipo di campionamento **molto usato perché** la selezione del campione è snella, ha tempi di realizzazione molto rapidi ed è perciò **poco costoso**

# Campionamento per quote (1 di 2)

Particolare campionamento a scelta ragionata, realizzato con modalità organizzative del campionamento di comodo

- si prescinde completamente dalla disponibilità di una lista di appartenenti alla popolazione obiettivo
- non vi è dunque alcuna selezione campionaria che imponga ai rilevatori di contattare determinate unità statistiche
- gli intervistatori decidono chi avvicinare
- **unico vincolo**: campione e popolazione devono condividere la medesima composizione relativa rispetto a caratteri quali:
  - sesso ed età
  - zona di residenza
  - livello di istruzione
  - condizione professionale

Grande importanza hanno l'aggiornamento e l'attendibilità delle fonti statistiche attraverso le quali vengono determinate le quote



# Campionamento per quote (2 di 2)

## Definizione delle quote prefissate

- **marginali**: ognuna delle assegnazioni è indipendente
- **associate**: le fonti consentono di dare indicazioni su due o più caratteri simultaneamente

L'idea è che, a parità di composizione strutturale, il campione possa ritenersi rappresentativo della popolazione di riferimento.

Si tratta di una **rappresentatività presunta** e non verificata in concreto.

L'intervista a persone più disponibili o più facilmente reperibili, e quindi più simili tra loro come atteggiamento e abitudini (situazione tipica del campionamento per quote), potrebbe comportare una **sottostima sistematica della variabilità** esistente nella popolazione.

# Campionamento casuale

Equivale a un' **estrazione di palline numerate**, di forma e peso uguali, **da un'urna nel quadro di un piano di campionamento** (*sampling frame*) che assegna una probabilità nota a priori a ogni campione appartenente a un certo insieme (universo dei campioni)

Se il campione è scelto casualmente è possibile:

- calcolare il **rischio d'errore** a cui ci si espone nella stima delle caratteristiche di interesse (**errore casuale di campionamento**)
- estendere mediante processi di **inferenza induttiva inversa** i risultati all'intera popolazione

# Vincolo di bilancio

E' della massima importanza **offrire stime affidabili** delle caratteristiche della popolazione da cui proviene il campione **rispettando l'ordine di grandezza della spesa** che viene generalmente **preventivata**.

Si tratta del cosiddetto **vincolo di bilancio**

# Campionamento “ragionato” o “probabilistico”

- Il campione ragionato ha la pretesa, almeno nell'intenzione di chi lo seleziona, di essere rappresentativo
- Il campione probabilistico è, invece, solo uno fra i tanti campioni possibili, ma si ha la garanzia che l'operatore non inserisca elementi di distorsione nel campione selezionato
- Il campione probabilistico elimina il soggettivismo pericoloso e difficilmente accettabile, tipico del campionamento ragionato
- Il campione probabilistico consente l'uso di un modello matematico che permetta la valutazione dell'attendibilità dei risultati
- Solo se il campione è casuale è possibile valutare il rischio di errore a cui si va incontro nella stima delle caratteristiche oggetto di interesse

## Aspetti delicati delle ricerche campionarie

Cosa è indispensabile conoscere per leggere e interpretare correttamente i risultati di un'indagine campionaria?

Come se ne possono valutare qualità e limiti?

Vanno individuati i punti nodali a cui possa riferirsi chi debba trarne indicazioni per prendere decisioni o anche soltanto consolidare o rimuovere un'opinione

Conseguire gli obiettivi indicati  
significa, in primo luogo, discutere ...

... i problemi di formazione dei campioni

# Fasi di progettazione dell'indagine

- identificazione della popolazione di riferimento dalla quale selezionare il campione
- definizione degli obiettivi conoscitivi della ricerca
- individuazione delle informazioni da raccogliere
- scelta dei criteri di selezione del campione (piano di campionamento)
- scelta della metodologia di stima dei parametri di interesse della popolazione (se si è selezionato un campione con criterio casuale)
- determinazione della dimensione del campione da realizzare
- scelta della procedura di rilevazione
- messa a punto del questionario
- valutazione dei costi di progettazione e di esecuzione dell'intera ricerca

## Fasi successive

- fasi operative di rilevazione
- memorizzazione su supporto informatico dei dati
- elaborazione dei dati
- analisi dei dati



# Formazione di un campione probabilistico disponendo di una lista

- campionamento casuale semplice senza ripetizione o estrazione in blocco
- campionamento casuale semplice con ripetizione (quasi mai utilizzato)
- selezione con passo sistematico (la lista deve essere ordinata casualmente; se c'è qualche sistematicità nell'ordine rispetto alla variabile di interesse il sistema non è più equivalente al campionamento casuale semplice)
- stratificazione: gli strati sono costruiti generalmente a partire da una o più variabili correlate al carattere oggetto di interesse, ipotizzando che unità omogenee rispetto a queste lo siano anche per il carattere d'interesse
- campionamento a grappoli
- campionamento a due o più stadi

## Dimensione del campione (1 di 2)

In base a quali criteri si fissa la dimensione del campione?

Si sceglie l'ampiezza più piccola che soddisfi certi requisiti di precisione delle stime stabiliti a priori

Indicare la precisione delle stime significa fissare il margine di errore che si è disposti a tollerare

Per ogni tipo di campionamento e per i singoli parametri da stimare la teoria dei campioni fornisce le formule di calcolo della numerosità campionaria

## Dimensione del campione (2 di 2)

La dimensione  $n$  del campione può essere determinata una volta fissati:

- piano di campionamento
- dimensione della popolazione
- grandezza dell'errore che si è disposti a tollerare
- grado di fiducia (95% o 99% = pratica certezza)

Se, invece, a causa, per esempio, di un budget limitato, vengono fissati:

- modalità di campionamento
- numerosità del campione

possono essere calcolati gli errori massimi corrispondenti ai vari gradi di fiducia

# Calcolo della dimensione di un campione

**Problema:** stima di una proporzione o percentuale con campionamento casuale semplice

$N$  = dimensione della popolazione

$P$  = proporzione incognita

$\theta$  = margine d'errore ammesso

$1 - \alpha$  = grado di fiducia

$z$  = variabile standardizzata con media 0 e varianza 1

$$n = \frac{z_{\alpha/2}^2 N}{\frac{(N-1)\theta^2}{P(1-P)} + z_{\alpha/2}^2}$$

$$n = \frac{z_{\alpha/2}^2 N}{4(N-1)\theta^2 + z_{\alpha/2}^2}$$

# Considerazioni sulla dimensione del campione

- Fissati il margine d'errore della stima e il grado di fiducia, all'aumentare della dimensione della popolazione corrispondono incrementi meno che proporzionali della numerosità del campione
- Fissata la dimensione della popolazione, la diminuzione del margine di errore richiede aumenti più che proporzionali della numerosità campionaria
- Per un dato margine d'errore, e quando la dimensione della popolazione supera  $N=5000$ , la scelta del grado di fiducia del 99% richiede incrementi della numerosità campionaria di entità non trascurabile rispetto a quelli necessari per ottenere stime a un grado di fiducia al 95%

# Dimensione del campione per livello di confidenza e precisione della stima

	Livello di confidenza									
	95%					99%				
	Margine d'errore %									
	1	2	3	5	10	1	2	3	5	10
<b>1.000</b>	906	706	516	278	88	943	806	649	400	143
<b>5.000</b>	3.288	1.622	880	357	94	3.845	2.271	1.350	588	161
<b>10.000</b>	4.899	1.936	964	370	95	6.247	2.938	1.561	624	164
<b>25.000</b>	6.939	2.191	1.023	378	96	9.991	3.567	1.722	648	165
<b>50.000</b>	8.057	2.291	1.045	381	96	12.486	3.841	1.783	657	166
<b>100.000</b>	8.763	2.345	1.056	383	96	14.267	3.994	1.815	661	166
<b>200.000</b>	9.164	2.373	1.061	383	96	15.363	4.075	1.832	663	166
<b>500.000</b>	9.423	2.390	1.065	384	96	16.105	4.126	1.842	665	166
<b>1.000.000</b>	9.513	2.395	1.066	384	96	16.369	4.143	1.846	665	166

# Calcolo della dimensione di un campione

**Problema:** stima del valor medio di un carattere con campionamento casuale semplice

$$n = \frac{z_{\alpha/2}^2 \sigma^2 N}{(N-1)\theta^2 + z_{\alpha/2}^2 \sigma^2}$$

**Problema:** stima del totale di un carattere con campionamento casuale semplice

$$n = \frac{z_{\alpha/2}^2 \sigma^2 N}{\frac{(N-1)\theta^2}{N^2} + z_{\alpha/2}^2 \sigma^2}$$

## Determinazione della varianza $\sigma^2$

Per usare la formula del calcolo della numerosità campionaria è necessario fornire una valutazione della varianza  $\sigma^2$  tramite:

- una stima preliminare della variabilità del carattere, proveniente, per esempio, da conoscenze a priori sulla popolazione in questione
- una stima ottenuta nella fase di pre-test del questionario, sottoposto al vaglio di un piccolo campione ragionato di intervistati,
- la relazione nota tra lo scarto quadratico medio  $\sigma$  e il campo di variazione (*range*)

$$\sigma \leq \frac{\text{range}}{2} \Rightarrow \sigma^2 \leq \frac{\text{range}^2}{4}$$

che riduce il problema alla stima delle intensità massima e minima sulla base di ragionevoli congetture



# Progettazione di un sondaggio

Fissare gli obiettivi di un sondaggio significa stabilire cosa si vuole conoscere e, soprattutto, con quale approssimazione.

Successivamente occorre identificare:

- la popolazione di riferimento
- l'unità di osservazione
- il tipo e le caratteristiche del piano di campionamento
- scegliere gli stimatori (campionamento probabilistico)

Nella fase di progettazione è, inoltre, importante definire:

- la modalità di raccolta dei dati
- la preparazione del questionario
- valutazione dei costi di indagine

# Popolazione oggetto di studio

Prima ancora di scegliere i criteri da seguire per la selezione delle unità campionarie, è fondamentale definire in modo chiaro la popolazione oggetto di studio, cioè identificare quali unità ne fanno parte e quali invece ne sono escluse.

**Popolazione:** insieme di unità nelle quali sono presenti le caratteristiche oggetto di analisi

Il problema è individuare a priori le unità che costituiscono il collettivo di riferimento.

Costruire ex-novo una lista può essere molto oneroso in termini di tempo e di risorse necessari.

## Le liste

Per selezionare campioni di famiglie o individui sono disponibili liste di vario tipo:

- anagrafe della popolazione
- liste elettorali
- elenchi degli abbonati del telefono
- elenchi delle utenze domestiche (ENEL, TELECOM, ...)

L'uso dei registri anagrafici è consentito tuttavia solo agli enti pubblici e agli organismi privati con finalità pubbliche.

# Liste elettorali

Le liste elettorali sono consultabili da chiunque (rispetto *privacy*)

Organizzate per sezioni di circa 500-600 elettori e per elenchi generali ordinati alfabeticamente e distinti per maschi e femmine.

Minorenni esclusi: la copertura completa della popolazione è limitata

La formazione di campioni casuali non è agevole perché occorre predeterminare la composizione del campione per sesso

Campioni di famiglie: la probabilità di estrazione di una famiglia è proporzionale al numero di elettori in essa presenti

Altre informazioni, come titolo di studio e posizione professionale, potrebbero non essere aggiornate

## Elenchi telefonici

Generalmente si effettua un campionamento sistematico, scegliendo un numero di individui maggiore di quello prefissato per sostituire quelli inutilizzabili

E' possibile anche campionare sistematicamente le pagine e poi subcampionare all'interno di esse

# Registro delle imprese commerciali

Per formare un campione di imprese o di esercizi commerciali, la principale lista di riferimento è il

**Registro delle imprese tenuto dalla Camera di Commercio**

La lista contiene tutte le imprese con sede legale nella provincia, a esclusione dei professionisti e dei coltivatori.

Sono disponibili:

- indirizzo
- data di costituzione
- forma giuridica
- attività esercitata
- numero degli addetti

Non essendo previste forme di aggiornamento automatiche  
la qualità dei dati è variabile

# Tecniche di raccolta delle informazioni

- interviste personali (*face-to-face interviews*)
- interviste telefoniche (*telephone interviews*)
- sondaggi postali (*mail surveys*)
- sondaggi via Internet (*web survey*)
  
- indagini miste (*mixed mode sureys*)

## Interviste personali (PAPI e CAPI)

Vengono condotte abitualmente presso l'abitazione delle unità statistiche se il campione è selezionato con criterio casuale

Vengono effettuate presso luoghi pubblici ed esercizi commerciali nel caso di campionamento di comodo o per quote

**PAPI:** *Paper And Pencil Interview*; l'intervistatore compila manualmente un questionario cartaceo

**CAPI:** *Computer Assisted Personal Interview*; l'intervistatore compila un questionario in formato elettronico direttamente con un computer portatile



# Pro e contro delle interviste personali

- + E' possibile approfondire determinati argomenti nell'ambito di indagini su temi molto articolati.
- + L'intervistatore può garantire una buona qualità del dato rilevato potendo fornire eventuali chiarimenti sulle domande.
- Costi elevati rispetto alle altre tecniche di rilevazione
- Tempi lunghi di realizzazione delle interviste
- Potenziale influenza dell'intervistatore con conseguente distorsione nelle risposte (*response bias*)

# Interviste telefoniche

- + economicità rispetto alle interviste personali
- + tempestività della rilevazione
- + possibilità di controllo dell'operato degli intervistatori durante l'intervista e non solo a posteriori
- + contatto più agevole con gli intervistati
- + possibilità di usare la metodologia **CATI** (*Computer Assisted Telephone Interview*) che gestisce automaticamente le fasi dell'intervista e controlla il suo svolgimento, migliorando la qualità dei dati
- copertura telefonica parziale e incompleta della popolazione (in Italia 95% di famiglie presenti)
- presenza di utenze riservate, cioè di famiglie che richiedono volontariamente di non apparire sugli elenchi (5% di utenze riservate)

# Metodologia CATI

Utilizzata dalla maggior parte degli istituti di ricerca come supporto per:

- definizione e selezione delle unità di campionamento
- inserimento dei dati
- codifica delle risposte
- tabulazione e analisi dei risultati

# Rilevazione CATI

## Sistema interattivo tra intervistatore e PC

- Il questionario è memorizzato nel PC e le domande scorrono sul video.
- Le risposte vengono digitate tramite tastiera e direttamente memorizzate su database.
- Il software effettua in tempo reale controlli di completezza e coerenza riducendo al minimo le possibilità di errore per il rilevatore.
- Il sistema può gestire automaticamente le telefonate

Il sistema CATI presuppone una struttura telefonica centralizzata con la presenza di supervisori che controllano le interviste

# Fasi di un sondaggio telefonico con metodologia CATI

1. **Estrazione del campione.** Il sistema dispone della lista completa e provvede a distribuire i nominativi agli intervistatori in base a quote o altri parametri prestabiliti
2. **Rilevazione dei dati.** Il software compone il numero da chiamare. L'intervistatore legge le domande e digita le risposte. Il sistema verifica le risposte e seleziona automaticamente la sequenza di domande da porre. Viene registrato ogni evento (abbandoni, mancate risposte) e i tempi di svolgimento (data, orario, durata e tentativi effettuati)
3. **Field telefonico.** Il software permette il controllo istantaneo dello svolgimento dell'indagine (risposte ottenute per singole domande, quote di campionamento, ...)
4. **Elaborazione dei risultati.** Il sistema è in grado di effettuare il controllo e l'editing dei dati rilevati. Può, inoltre, generare tavole di spoglio e di consultazione dei risultati in maniera guidata

# Pro e contro della metodologia CATI

- + Rapidità nella realizzazione della ricerca
- + Riduzione dei tempi e dei costi di elaborazione
- + Eliminazione delle fasi di realizzazione e compilazione del questionario, di *data entry* e di controllo dei dati in termini di completezza e coerenza
- + Possibilità di disporre di report in tempo reale
- + Elevato livello di qualità dei dati
- Necessità di dotazione elevata di hardware e software
- Tempi più lunghi nella fase di addestramento degli intervistatori

# Limiti della metodologia CATI

- I costi di predisposizione del software necessario allo svolgimento di un'indagine CATI diventano accettabili con non meno di 700-1000 interviste.
- Il metodo CATI può risultare poco conveniente per ricerche che comportano numerose domande aperte
- è necessario limitare la durata dell'intervista a 10-15 minuti per evitare stanchezza o disinteresse dell'intervistato a causa dell'assenza di un confronto faccia a faccia tra intervistato e intervistatore

# I sondaggi postali

Tradizionale alternativa alle interviste personali e telefoniche

- + Costi contenuti
- + Riduzione dei tempi e dei costi di elaborazione
- + Si elimina l'effetto dell'intervistatore
- + Indicato per argomenti delicati per i quali la presenza dell'intervistatore condiziona il rispondente
- Risultati di qualità inferiore: tassi di risposta pari alla metà di quelli che si ottengono con le altre tecniche di intervista
- Non c'è la possibilità di migliorare la qualità del dato (aiuto dell'intervistatore)



# Fasi di un sondaggio postale

- scelta delle caratteristiche del questionario: lunghezza, layout, ...
- definizione della lettera di accompagnamento e presentazione della ricerca
- pianificazione di un sistema di solleciti per far fronte alla mancata collaborazione dei partecipanti
- scelta del tipo di busta di ritorno, affrancatura e modalità di spedizione
- scelta degli incentivi per gli intervistati

# Limiti del sondaggio postale

- Elevata quota di mancate risposte
- I questionari spesso risultano incompleti
- Non è detto che il rispondente sia il reale destinatario della lettera

# Le autointerviste: metodologia CASI

L'intervistato può rispondere al questionario ...

- utilizzando il proprio PC; il questionario viene memorizzato su dischetto e poi inviato via posta oppure spedito in allegato tramite la posta elettronica
- facendo uso della tastiera di un telefono multifrequenza collegato a un computer
- fornendo a voce le proprie valutazioni dopo aver contattato un numero telefonico dedicato (generalmente un numero verde) gestito da un computer che provvede a effettuare l'intervista

Il software messo a punto in tale contesto è denominato  
**CASI** (*Computer Assisted Self Interview*)

# Sondaggi via Internet: metodologia CAWI

Il sistema CAWI (*Computer Assisted Web Interview*) può essere visto come la naturale evoluzione tecnologica dei tradizionali sondaggi postali

## Fasi di un sondaggio CAWI

- spedizione di un messaggio di posta elettronica alle unità statistiche, assimilabile alla lettera di presentazione della ricerca, dove viene indicato il sito a cui collegarsi per compilare il questionario
- l'intervistato compila il form i cui dati vengono automaticamente memorizzati in un database

# Pro e contro della metodologia CAWI

- + Abbattimento dei costi; con la posta elettronica si possono inviare simultaneamente migliaia di messaggi
- + Aumento della velocità di spedizione dei messaggi
- + Elevata probabilità che il messaggio sia letto dal destinatario
- + Possibilità di pianificare meticolosamente i solleciti
- + Possibilità di analizzare i *log files* di accesso al sito dove risiede il questionario
- Si fa riferimento ai soli utenti della rete
- Distorsione dei risultati dovuti alle possibili differenze nella struttura socio-demografica rispetto alla popolazione