

# Statistica sociale – 5

*Prof. Antonio Mussino*

**a. a. 2016-2017**



**SAPIENZA**  
UNIVERSITÀ DI ROMA

# Campionamento non probabilistico

# Il campionamento per quote - 1

- Nell'ambito della ricerca sociale e, in particolare, in quello delle ricerche di mercato, si utilizza una strategia molto efficace, che parte comunque dalla stratificazione della popolazione: il ***campionamento per quote***.
- Una volta che lo schema generale di stratificazione è stato definito e il numero di interviste da effettuare in ogni strato è stato fissato, la scelta di quali unità ***effettivamente*** intervistare è lasciata agli **intervistatori**.

# Il campionamento per quote - 2

- Siamo in presenza di un campionamento a più stadi, in cui nell'ultimo stadio la scelta non è aleatoria!
- Non si potrebbe applicare l'inferenza, anche se un'inferenza puntuale viene comunque fatta.
- Questo metodo è preferibile perché meno costoso e più facile da gestire.

# Il campionamento per quote - 3

- Come funziona il **quota sampling**?
- Un piano di campionamento in cui la popolazione è stratificata per genere, facoltà di appartenenza (Lettere, Scienze, Ingegneria e Statistica) e tipologia del corso di studi (laurea triennale e magistrale):
  - si vuole stimare la percentuale di studenti che praticano attività sportiva presso il CUS (Centro Sportivo Universitario).

# Il campionamento per quote - 4

- Il campione è di **1000** interviste, che devono essere così ripartite per facoltà: Lettere **340**; Scienze **300**; Ingegneria **290**; Statistica **70**.
- Complessivamente si vogliono intervistare **480** studenti e **520** studentesse; **650** in corsi di laurea triennali e **350** in corsi di laurea magistrale.

# Il campionamento per quote - 5

- All'intervistatore responsabile della Facoltà di Lettere viene chiesto di intervistare **240** studentesse e **110** studenti; **221** della triennale e **119** della magistrale;
- per Scienze i numeri sono **150** (f) e **150** (m) per genere e **196** (t) e **104** (m) per corso;
- per Ingegneria **90** (f) e **200** (m) per genere e **193** (t) e **97** (m) per corso;
- per Statistica, infine, **40** (f) e **30** (m) per genere e **40** (t) e **30** (m) per corso.

# Il campionamento per quote - 6

- L'intervistatore sceglierà quali studenti intervistare, evidentemente ottimizzando il loro reperimento, in modo da rispettare le *quote* prima definite.
- È evidente che lo schema di stratificazione deve essere il più accurato possibile in questa strategia, proprio per non lasciare troppa libertà, che potrebbe portare a situazioni paradossali:
  - ad esempio, le 40 studentesse di statistica potrebbero essere tutte di un corso triennale e l'intervistatore, pur rispettando i vincoli a lui assegnati, introdurrebbe una distorsione sistematica nei risultati (tutte le studentesse potrebbero essere più giovani degli studenti!).

# Il campionamento per quote - 7

- Questo si eviterebbe se le quote fossero definite tenendo conto contemporaneamente (schema di assegnazione ***interrelated*** – quote **associate**) e non separatamente (schema di assegnazione ***independent*** – quote **marginali**) dei due criteri di stratificazione, così come suggerito nella tabella che segue:

# Il campionamento per quote - 8

Maschi	Lettere	Scienze	Ingegneria	Statistica	Totale
Triennale	65	98	131	18	312
Magistrale	35	52	69	12	168
Femmine	Lettere	Scienze	Ingegneria	Statistica	Totale
Triennale	156	98	62	22	338
Magistrale	84	52	28	18	182
<b>Totale</b>	<b>340</b>	<b>300</b>	<b>290</b>	<b>70</b>	<b>1000</b>

# Il campionamento per quote - 9

- Questa strategia migliora i risultati, ma non garantisce da rischi di errore sistematico, dovuti alla particolare composizione della popolazione;
- l'esempio ci aiuta per esemplificare quanto detto: un intervistatore cercherà di ottimizzare il suo tempo cercando gli studenti laddove sono più presenti, ovvero nella sede della rispettiva facoltà, dove seguono le lezioni e studiano.

# Il campionamento per quote - 10

- Così un *quota sampling* sottostimerebbe gli studenti ***non frequentanti***, la cui partecipazione sportiva è probabilmente diversa da quella dei ***frequentanti***.
- Al contrario, in un campionamento probabilistico, ad esempio campionando dalle liste di segreteria, la probabilità di essere intervistati è uguale per chi frequenta e per chi non frequenta!

# Il campionamento per quote - 11

- Importante:
- controllare la variabilità delle risposte per intervistatore, ossia verificare che questa non sia dovuta alle modalità di scelta degli intervistati, ma anche a quelle di somministrazione del questionario!

# Il campionamento per quote - 12

- Svantaggi del **quota sampling**?
- impossibilità di stimare l'errore di campionamento nell'effettuare stime.
- non sempre è possibile considerare molte caratteristiche per stratificare per la ridotta numerosità dei campioni.

# Il campionamento per quote - 13

- Vantaggi del **quota sampling**
  - il minor costo, vantaggio che si riduce proporzionalmente al crescere del dettaglio nella stratificazione;
  - la facilità di organizzare la rilevazione, senza andare a cercare, e ricercare se non reperibile, l'unità da intervistare;
  - la possibilità di campionare in assenza della lista delle unità della popolazione e di caratteristiche note rispetto alle quali stratificarla.

# Il campionamento per quote - 14

- In sintesi nella ricerca sociale in senso ampio, in particolare per le fonti ufficiali, questo tipo di campionamento presenta <<scarsa scientificità>> (Corbetta, 2003),
- ma nelle **ricerche di mercato** e nei **sondaggi di opinione** il risparmio in termini di **budget** lo fa preferire, tenendo conto che non vale la pena sostenere il costo di un campione probabilistico, dato che le **fonti di errore** più rilevanti in questi casi sono di altra natura.

# Altri tipi di campionamento non probabilistici

- Già abbiamo visto che il ***campionamento sistematico*** e quello ***per quote*** non si possono tecnicamente considerare probabilistici, ma vengono ampiamente utilizzati nella ricerca sociale come se lo fossero, utilizzando al meglio alcune loro peculiarità.
- Nella ricerca sociale, comunque, si utilizzano anche altre forme di campionamento non probabilistico.

# Il campionamento a valanga - 1

- Quando nulla si può conoscere della popolazione di riferimento, si può usare il ***campionamento a valanga***:
  - si pensi, ad esempio, a una indagine sugli immigrati illegali, popolazione al di fuori di ogni controllo anagrafico, censuario o comunque legale.
- In questo caso le unità da inserire nel campione sono individuate a partire da quelle già intervistate.

# Il campionamento a valanga - 2

- Si inizia contattando, con scelta ragionata ovvero con informazioni a priori in possesso del ricercatore, un gruppo di unità che faccia parte della popolazione in questione e si chiede a queste stesse unità di fornire i nominativi o i contatti con altre unità, ampliando così il campione che cresce come sotto l'effetto di una valanga.
- Il rischio è, ovviamente, di indirizzarsi solo verso una parte della popolazione sconosciuta, o comunque di contattare gli individui più attivi e disponibili.

# Il campionamento ragionato

- La prima fase di contatto precedentemente descritta corrisponde a una scelta che si può definire ***campionamento ragionato***, in quanto le unità sono scelte sulla base di alcune loro caratteristiche che sono quelle che ci interessa tenere sotto osservazione nella ricerca.
- Così quando in un'indagine sul disagio sociale si delimita l'area di riferimento a quartieri periferici o con alto tasso di criminalità.

# Il problema delle *non risposte* - 1

- Tornando al caso prioritario del campionamento probabilistico, se noi abbiamo una lista precisa di unità da contattare e qualcuna di queste non vuole rispondere o collaborare all'indagine, ci troviamo di fronte al problema delle ***non risposte***.
- È come se, effettuando l'estrazione di palline da un'urna, qualcuna delle palline non volesse essere estratta (Moser, Kalton, 1979)!

# Il problema delle *non risposte* - 2

- Si tratta di un problema molto serio, perché può introdurre una distorsione sistematica nei risultati.
- Le cause per cui una unità statistica non partecipa all'indagine sono varie:
  - rifiuto di essere coinvolto, intervistato;
  - cambio di residenza;
  - irreperibilità;
  - orario non adeguato e così via.

# Il problema delle *non risposte* - 3

- Le cause **oggettive** non dovrebbero distorcere la stima finale, tranne che in alcune indagini particolari, e comunque si possono prevedere liste di unità a cui attingere per la sostituzione.
- Il problema è più delicato nel caso di rifiuto **soggettivo** di collaborare o di rispondere a uno specifico quesito: questo può creare problemi molto consistenti che dovrebbero essere affrontati con una diversa metodologia di stima.

# Il problema delle *non risposte* - 4

- Consideriamo il caso dei sondaggi pre-elettorali: è opinione comune che gli elettori di centro-destra abbiano una maggiore ritrosia a dichiarare il loro voto di quelli di centro-sinistra.
- Ipotizziamo di aver individuato un campione di **1000** unità e di avere ricevuto il rifiuto a rispondere da parte di **300** intervistati.
- Tra i **700** rispondenti, **280** si sono dichiarati per il centrosinistra, **220** per il centrodestra, **100** per altre liste e **100** indecisi.

# Il problema delle *non risposte* - 5

- Nel commentare i risultati diremmo che il **40%** si è dichiarato favorevole al centrosinistra e il **31%** al centrodestra, con una differenza, quindi, di **9 punti percentuali**.
- Se non vi fosse alcuna differenza di comportamento elettorale nel gruppo dei rifiuti, questa stima, con tutte le cautele del caso, sarebbe il risultato finale della nostra analisi.
- In realtà, come detto, la maggiore difficoltà a dichiarare il proprio orientamento politico si ha tra gli elettori di centro destra.

# Il problema delle *non risposte* - 6

- Quindi si può pensare che le proporzioni nel sottogruppo dei rifiuti siano almeno inverse a quelle del sottogruppo di chi si è espresso: qui il centro sinistra avrebbe **93** preferenze e il centrodestra **120**:
  - in totale su **1000** unità **373** si orienterebbero per il centro sinistra e **340** per il centro destra e la stima puntuale della differenza sarebbe di soli **3,3 punti percentuali!**
- N.B. In questo caso si è utilizzata una ***stima puntuale***, anche se, come detto, è più utile fornire una stima intervallare, con la misurazione dell'errore commesso, perché l'intervento di correzione sui dati ha fatto perdere la possibilità di usare metodi inferenziali.

# Il problema delle *non risposte* - 7

- N.B. In questo caso si è utilizzata una **stima puntuale**, anche se, come detto, è più utile fornire una stima intervallare, con la misurazione dell'errore commesso, perché l'intervento di correzione sui dati ha fatto perdere la possibilità di usare metodi inferenziali!!!!!!!