*Seconda parte*

# *Teoria dei campioni*

## INTRODUZIONE

Nell'inferenza che abbiamo trattato nella statistica di base, per studiare un fenomeno partivamo da un campione costituito da *n* possibili realizzazioni della variabile casuale, e ad ogni estrazione si generava una variabile casuale distribuita come la variabile casuale del fenomeno da investigare (rilevazione peso, misurazioni della temperatura, etc.), in pratica potevamo estrarre un campione da qualcosa di virtualmente infinito, ossia era possibile misurare infinite volte. In base a questo potevamo infierire sulle proprietà della popolazione sfruttando le proprietà, teoremi e definizioni asintotiche della variabile sotto investigazione, approssimando ad una gaussiana, potevamo definire una funzione di massima verosimiglianza per i parametri che qui erano ignoti ma fissi.

Ora non possiamo più sfruttare le definizioni asintotiche sulla distribuzione della nostra variabile casuale, perché le realizzazioni sono valori, numeri, non conosciamo la distribuzione della nostra popolazione, ogni campione che prendiamo è diverso dall'altro ed a seconda del campione che prendiamo potremmo avere un diverso risultato. Tra i due tipi di aleatorietà, una dovuta al possibile errore di misura insito nelle misure che ci vengono offerte e l'altra derivante dal campionamento (la prima la subiamo mentre la seconda la possiamo controllare) introduciamo il secondo tipo di aleatorietà nel momento in cui effettuiamo la scelta delle nostre *n* unità, l'aleatorietà la introduciamo noi con il nostro piano di campionamento, con il nostro disegno, quindi anche la stima e gli stimatori saranno legati a questo tipo di aleatorietà; dunque la varianza della stima, dello stimatore, in pratica significa studiare la varianza del nostro metodo di selezione.

Stiamo parlando in pratica di *errori campionari* è in di *errori extra-campionari*, i primi sono connessi alla variabilità statistica dei risultati che derivano dalla natura dell'inferenza, dovuti principalmente alla fase di selezione e quindi al nostro piano di campionamento, e possono essere analizzati con metodi statistici, mentre i secondi sono connessi a tutte le circostanze per cui il dato rilevato non corrisponde a quello vero, per una serie complessa di fattori che intervengono nelle fasi di un'indagine statistica. Ora ci occupiamo del primo tipo di errori e dedichiamo poco tempo alle problematiche extra-campionarie. A differenza dell'approccio classico sull'inferenza dove la popolazione è stata ipotizzata illimitata, in questo capitolo considereremo estrazioni da una

popolazione le cui unità sono in numero finito, realizzando così un campionamento da popolazioni finite. Il settore della statistica che si occupa di questi aspetti è *teoria dei campioni*. Vediamo alcune differenze tra l'approccio classico all'inferenza e la teoria del campionamento da popolazioni finite[6]:

- In una popolazione finita le unità statistiche esistono e sono elencabili, mentre in una popolazione infinita esse sono per lo più ipotetiche;
- Nel campionamento da popolazioni finite, è più rilevante il valore $y_i$ assunto nella i-esima unità statistica dalla variabile *Y*, che la sua intera distribuzione;
- Come vedremo, sia la procedura di selezione delle unità per formare il campione che gli stimatori dei parametri utilizzano spesso "informazioni aggiuntive" su una variabile ausiliaria X di agevole reperimento e che sia funzione di *Y,* oggetto principale dell'indagine campionaria;
- La funzione di verosimiglianza in una popolazione finita è costante per tutti i valori $y_1, y_2, ....., y_n$, per cui non è di per sé informativa circa i parametri di interesse; in altre parole, se θ è un parametro definito in funzione di *Y* che nel nostro caso rappresenta la variabile aleatoria, l'inferenza della stessa, qualunque sia la sua funzione di probabilità, non ci fornisce alcuna informazione rispetto al nostro parametro, dato che la verosimiglianza è costante rispetto a θ;
- Non esistono estimatori lineari non distorti ed efficienti, estimatori sufficienti, ne sono sempre applicabili teoremi e definizioni asintotiche, proprio perché N è finito;

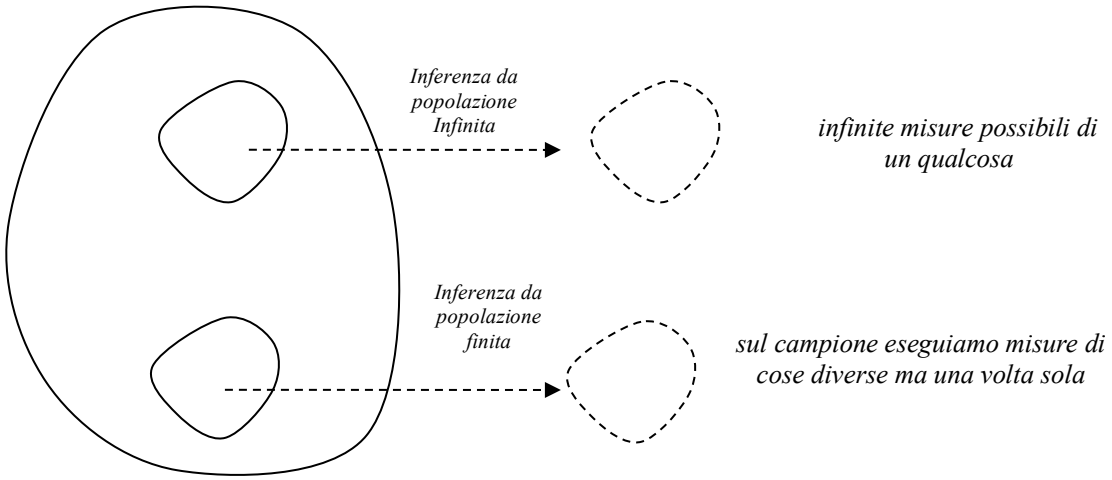
Tali differenze hanno imposto una trattazione autonoma e, per molti aspetti, peculiare nella teoria dell'inferenza per popolazioni finite, che richiede un notevole impegno formale. Nel seguito, ci limiteremo alla presentazione degli elementi essenziali in relazione agli schemi campionari più diffusi, discutendo della stima dei parametri di maggiore rilevanza.

[6] Campionamento da popolazioni finite, statistica, Piccolo D., *IL MULINO*, 2000

Riassumendo, in statistica si distinguono due teorie dell'inferenza statistica: *per popolazioni finite* e *popolazioni infinite*. Le popolazioni finite sono popolazioni reali

potenzialmente etichettabili, mentre le popolazioni infinite sono ipotetiche, riguardano immaginarie ripetizioni di prove, ottenute sotto certe condizioni teoriche. Su questa distinzione si può essere d'accordo, sebbene dal punto di vista del calcolo delle probabilità, siamo sempre di fronte ad un processo di generazione di dati. Cioè, in un caso abbiamo una variabile aleatoria che è generata dal campionamento da una popolazione finita costituita da *N* elementi, nell'altro abbiamo una variabile aleatoria che è generata sotto delle condizioni definite in maniera più astratta. Un tentativo di unificazione delle due teorie si basa sulla nozione di superpopolazione, in cui si assume che una popolazione finita si possa considerare come un campione estratto da una popolazione infinita, in pratica consideriamo una situazione più complessa, prendendo come riferimento entrambe le fonti di aleatorietà, quella legata al campionamento e quella legata all'errore di misura, in tal caso si parlerà di *superpolazione* (in cui abbiamo infinite misure); è il classico caso in cui, chiedendo il reddito, ci vengono date una serie di risposte diverse (e che dunque consideriamo variabili aleatorie) per un determinato campione. Si comprende dunque come l'aleatorietà sia duplice.

In definitiva tutte queste considerazioni conducono a ritenere che la differenza tra popolazioni finite ed infinite, sia in pratica da riassumere dicendo che la differenza è tra *popolazioni di distribuzione nota* o *non nota* (o non riconducibili ad una forma nota).




La statistica che conosciamo noi  riguarda solo esclusivamente il primo caso, cioè infinite misure possibili di un qualcosa, mentre quella che tratteremo è la seconda.

Perché è completamente diverso? Perché le probabilità sono diverse, tra l'altro sia meccanicamente sia fisicamente i due processi sono completamente diversi, in quello di sopra possa avere un errore di misura che se vogliamo possiamo considerare come errore derivante da un processo fisico, inseriamo una gaussiana e lo studiamo, possiamo avere infiniti risultati, nel caso di sotto invece non possiamo avere infiniti campioni. Le due cose sono diverse, sono alternative, o possono essere utilizzate entrambe? Possiamo anche ragionare con ambedue, estraendo un certo numero di elementi e facendo su ciascuno un'analisi; ora teniamo presente che siamo in una facoltà di economia e quindi le unità statistiche di cui parliamo sono imprese e famiglie per cui se vogliamo possiamo estendere la nostra analisi. Ad esempio supponiamo di avere una popolazione di topolini, abbiamo *1000* topolini bianchi, ne estraiamo un campione, ad esempio *100*, e somministriamo il nostro farmaco, e vediamo se muoiono o restano in vita. Avremo prima un problema di casualità di estrazione dei 100 topolini, poi un problema di casualità legato alla misura. Questi due problemi quindi sono completamente diversi, e ciò che li rende completamente diversi è il fatto che  se l'errore di misura non lo possiamo controllare, il secondo errore invece lo possiamo controllare, siamo noi che introduciamo il caso estraendo i *100* topolini tra i *1000*, e come li estraiamo, quanti e come, lo decidiamo noi, detta in modo poco scientifico il primo è un caso naturale, mentre il secondo è un caso che noi consciamente e scientificamente decidiamo d'introdurre.

Il secondo effetto casuale lo introduciamo perché non possiamo guardare tutte le unità, quindi un problema di costi, sia di tempo sia di soldi. Il motivo per cui si esegue una selezione casuale, lasciata al caso e non ad una selezione ragionata, sta nel fatto che una selezione ragionata si basa su una forte soggettività. Il problema è che anche quando siamo convinti di essere stati totalmente obiettivi, se non ci affidiamo al caso ma selezionano le persone in modo ragionato, involontariamente possiamo introdurre delle forti distorsioni. Prendiamo il caso classico degli intervistatori che ci fermano agli angoli delle strade, perché questo non è affatto casuale ma è distorto? Perché la persona che si decide di intervistare è scelta sul momento dall'intervistatore, ad esempio scegliendo una bella donna, quindi la nostra indagine quantomeno sul sesso sarà distorta, ed anche sulla scelta del posto, sceglieremo i posti più frequentati e più popolari.

Concludendo quindi, la teoria che mette insieme l'inferenza da *popolazione finite* ed *infinite* si chiama "*approccio da superpopolazione*", non solo abbiamo una casualità legata al campionamento ma ciascuna unità non la consideriamo un numero ma una possibile realizzazione di una variabile aleatoria.

Per la teoria dei campioni quanto viene rilevato non è affetto da errore, è un numero, ovvero estraiamo dal nostro archivio gli individui, consegniamo il questionario, chiediamo qualcosa di quantitativo come ad esempio il reddito, e ci risponderanno con un numero, ma l'aleatorietà è dovuta solo al fatto di aver selezionato quella unità statistica piuttosto che un'altra, e quindi la teoria dei campioni tratta solo questo tipo di aleatorietà, ovvero la selezione, quello che facevamo a statistica di base dipendeva solo dal fatto che quello che avevamo rilevato non era un numero ma una realizzazione di una variabile casuale. Ambedue i tipi di casualità li vedremo alla fine del corso.

Ora per analizzare tutti i passi della teoria dei campioni da cosa dobbiamo partire? Prima di tutto dobbiamo fare *il disegno campionario*, cosa che non facevamo a statistica di base perché l'aleatorietà non la potevamo controllare noi, qui invece sì, successivamente dobbiamo estrarre il campione e quindi parleremo di *selezione campionaria*; solo da questo punto in poi gli argomenti sembrano simili (stime, stime di varianza, etc.). Sempre tornando al concetto d'inferenza dobbiamo fare una stima puntuale e successivamente fare una stima di varianza per costruire gli intervalli di confidenza, ad esempio in una stima del reddito pescarese calcolare la sua varianza, ovvero quanto si discosta dalla stima. Riepilogando dobbiamo fare il *disegno campionario*, *la selezione*, *i criteri di stima*, ed i *criteri di stima dell'incertezza della varianza campionaria.*

# 6. Il campione

## 6. Il disegno campionario

Per arrivare al disegno campionario definiamo come:

$$S = \{S_1, S_2, ....., S_n, P\}$$

L’insieme di tutti i possibili sottoinsiemi della popolazione che chiamiamo S; ciascuno dei campioni all’interno delle parentesi, $S_1, S_2, ....., S_n, P$ è un sottoinsieme della popolazione. Ora se a ciascun sottoinsieme della popolazione assegniamo una probabilità di essere selezionato, avremo:

$$P(S) = P(S_1) = P(S_2) = ..... = P(S_I) = P(P)$$

dove tutti i campioni hanno la stessa probabilità di essere estratti. Sarebbe la distribuzione di probabilità su tutti gli insiemi di *S.* Questa distribuzione di probabilità si chiama disegno campionario, ovvero tutti i nostri criteri di selezione, qualsiasi criterio adottiamo, si traduce di fatto nel dare una probabilità ad un qualsiasi sottoinsieme, probabilità che tra l'altro può essere anche zero. Ad esempio tornando ai topolini, se prendiamo un campione di *100* topolini non facciamo altro che assegnare una probabilità zero a tutti campioni maggiori di *100*; fare un campione casuale semplice significa che in pratica estraiamo i sottoinsiemi non solo a caso ma tutti con la stessa probabilità è senza reinserimento, il campione casuale semplice altro non è che porre tutte le probabilità uguali per tutti i sottoinsiemi, ad esempio:

$$\#(S_1) = 100$$

che sta ad indicare tutti i sottoinsiemi hanno cardinalità *100*, devono avere tutti la stessa probabilità di estrazione. Per quanto grande sia la popolazione noi possiamo elencare tutti i pezzi, elenchiamo tutte le possibili parti in cui è possibile dividerla.
Ad esempio quando facciamo un censimento diamo probabilità uguale ad uno al nostro campione, naturalmente di numerosità *N*, e zero a tutti gli altri, quello che vogliamo dire è che di tutte queste parti né dobbiamo osservare solo una, fare un disegno campionario vuol dire che quale parte vogliamo osservare non possiamo definirlo in modo deterministico ma, qualsiasi criterio di scelta noi adottiamo, si traduce nel fissare una probabilità di estrazione, ad esempio fare un campione di 1000 imprese su una popolazione di 4 milioni significa dare una probabilità uguale per tutti i campioni composti da 1000 e zero a tutti gli altri, qualsiasi altro campione non composto da 1000

elementi per noi è assolutamente indifferente, ha la stessa probabilità. Nel prossimo paragrafo vengono elencati i criteri più ragionevoli con cui si possono elencare le unità statistiche.

## 6.2 I metodi di selezione campionaria

Il metodo più banale che probabilmente già conosciamo è il campione casuale semplice; prendiamo *n* unità da *N* senza reinserimento poiché con il reinserimento abbiamo un livello di applicabilità pari a zero.

Quali altri criteri introduciamo? Facciamo l'esempio di un'indagine condotta sui consumatori di birra; come sappiamo la birra è un consumo tipicamente maschile, quindi sarebbe opportuno stratificare la nostra indagine rispetto al sesso, poiché se non facciamo la stratificazione e capitano all'interno del campione solo donne la nostra sottostima del consumo di birra sarà notevole, è molto più ragionevole selezionare ad esempio cinque donne e cinque uomini. La stessa cosa può dirsi ad esempio per un campione d'imprese, sarebbe ragionevole prenderne una parte per ogni settore di attività economica, oppure se dobbiamo fare una stima sul costo del lavoro, se i questionari li mandiamo totalmente a caso potremmo correre il rischio che finiscano tutti in un solo settore di attività economica, ad esempio in quello dei servizi in cui vige una particolare legislazione; o ancora se facciamo un'indagine sui consumi delle famiglie senza tener conto delle dimensioni della famiglia, potremmo avere un campione alquanto distorto.

Questo sistema implica praticamente di dover prendere la popolazione e tagliarla a fette, spezzettarla in sottopopolazioni e fare tanti campioni completamente indipendenti l'uno dall'altro per ciascuna sottopopolazione; le sottopopolazioni si chiamano strati. Il campione ottenuto da questo procedimento prende il nome di *campione casuale stratificato*, riguardo al numero di campioni da estrarre in ciascun strato si utilizzano altri criteri. Il criterio di stratificazione può e deve basarsi esclusivamente su variabili contenute nell'archivio.

La numerosità campionaria di ciascun strato, la frequenza campionaria di ciascun strato nulla ci vieta che possa cambiare, in alcuni strati possiamo addirittura fare un censimento, ad esempio nell'indagine sulle imprese, se avessimo uno strato composto da sole tre imprese le prenderemo ovviamente tutte, non faremo un campione;

solitamente gli strati per le imprese si basano sul settore di attività economica mentre l'altra variabile tipica di stratificazione è il codice geografico (*NUTS*).

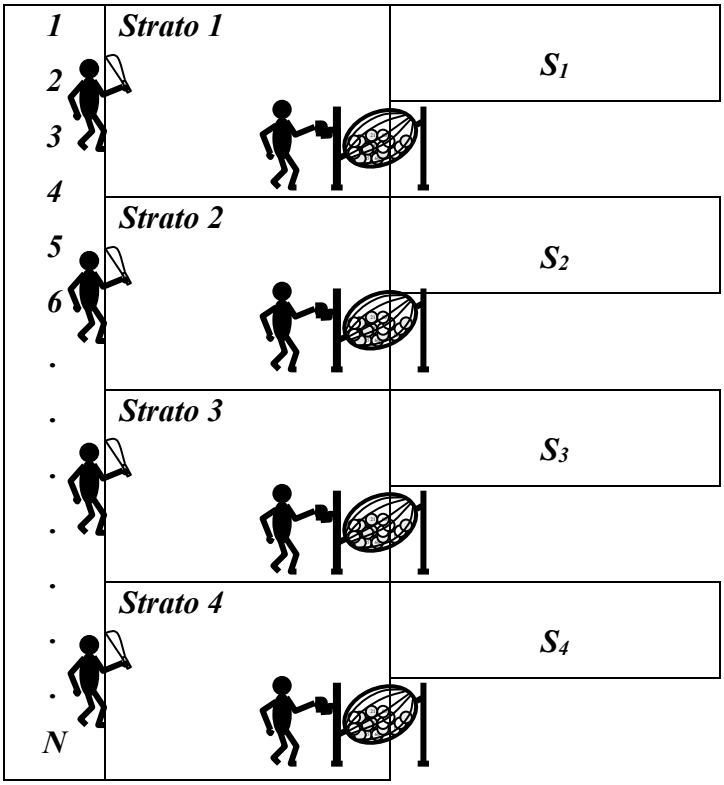



*Fig.6.1 – Stratificazione della popolazione*

In ciascuno strato osserviamo una parte, e ciascuna parte può avere una diversa probabilità di selezione, al punto che niente ci vieta che in alcuni strati possiamo selezionare tutte le unità. Come altro metodo di campionamento possiamo avere il campionamento a più strati, ossia dopo aver stratificato la popolazione potremmo fare anche un'altra cosa, potremmo fare il contrario, ovvero prendiamo a caso delle fette di popolazione , ed all'interno di queste osserviamo tutto. Quindi se dividiamo prima la popolazione in pezzi sulla base di variabili d'archivio e prendiamo delle unità a caso stiamo facendo un *campionamento stratificato* mentre se prendiamo a caso solo alcuni pezzi della popolazione e nei pezzi osserviamo tutte le unità statistiche, in questo secondo caso stiamo eseguendo un *campionamento a grappoli*.

Possiamo anche selezionare alcuni gruppi a caso ed all'interno di questi estrarre alcune unità e quindi altri campioni, questo è il caso del *campionamento a due stadi*;

naturalmente gli stadi possono partire da uno per il campionamento *casuale semplice*, fino al *campionamento più stadi*; ad esempio da una nazione possiamo estrarre le regioni per poi dalle regioni passare ad estrarre le province e dalle province estrarre i comuni ed infine unità. Ma perché mai dovremmo estrarre dei campioni e poi osservare tutte le unità? Mentre il due stadi o il campionamento a più stadi vengono fatti esclusivamente perché non si dispone di una lista delle unità statistiche (tra l'altro sono totalmente inefficienti), il campionamento a grappoli viene utilizzato quando non è disponibile un archivio all'ultimo stadio delle unità. In ogni caso se nello stratificato possiamo rendere più efficiente la stima, in questo caso utilizziamo i grappoli solo perché non disponiamo di una lista delle unità.

Ora se in un'indagine utilizziamo un campione casuale semplice d'imprese - ci sembra ragionevole che un tabaccaio abbia la stessa probabilità di essere inserito nel campione della Fiat? Certamente no; come sappiamo il campione casuale semplice assegna la stessa probabilità alle unità statistiche. In alternativa potremmo usare il campione stratificato, ma non è comunque il migliore. Invece di cercare una soluzione tra i metodi di campionamento enunciati potremmo modificare le probabilità di inclusione in questo modo: potremmo usare come variabili di ponderazione delle variabili continue contenute nell'archivio, nel caso delle imprese abbiamo il numero di addetti, quindi - perché dobbiamo ridurci a soli tre codici quando possiamo avere un continuo di probabilità per le nostre unità? Quindi avremo una probabilità d'inclusione per ogni unità, ossia qualsiasi variabile oggetto di studio che noi ipotizziamo essere correlata positivamente con la probabilità d'inclusione del campione è facilitata dall'avere una probabilità proporzionale, prendere variabili proporzionali agli addetti facilità la stima di variabili correlate positivamente con gli addetti, fissare delle probabilità proporzionali, correlate positivamente con gli addetti facilita enormemente la stima di qualsiasi variabile, ossia la correlazione nel caso delle imprese si ha con la variabile dimensionale. Quindi dipende dalla variabili che stiamo indagando, se mettiamo delle probabilità di essere estratti proporzionali per una qualsiasi variabile che ha una correlazione positiva con gli addetti, rendiamo il campione molto più efficiente perché proprio nel criterio di selezione il campione già ne tiene conto in qualche modo. Di solito tutte le variabili economiche hanno una correlazione con il numero di addetti dallo 0,9 in su; con questo procedimento in pratica discriminiamo non solo tra blocchi

ma tra un'unità e l'altra, questo tipo di campionamento si chiama *P.P.S.* ( Proportional Probability Sampling) ovvero *campionamento casuale proporzionale.* Naturalmente nulla ci vieta di dividere prima in fette la popolazione (settore di attività economica, regioni, province, etc.), quindi stratificare, e successivamente invece di usare il campionamento casuale semplice per ogni strato utilizzare il campionamento casuale proporzionale.

Ora, se ordiniamo la popolazione secondo un qualche criterio, e diamo un passo a questo criterio, un passo fisso di estrazione, ad esempio un'estrazione ogni 10 unità - è ragionevole? Sì, questo è il caso del *campionamento sistematico*, ed è ragionevole a seconda del comportamento della nostra variabile; questo tipo di selezione o questo criterio non deve influire sulla variabile, poiché potremmo avere dei problemi. Ad esempio se intendiamo contare il numero di pecore su un territorio, estraiamo un mese a caso e successivamente stabiliamo un passo di campionamento di 12 mesi commettiamo un grave errore perché a seconda del mese troveremo una diversa condizione della popolazione, in pratica il nostro passo di campionamento non deve avere lo stesso andamento della variabile, si potrebbe generare un andamento ciclico; ma perché dovremmo usare questo tipo di campionamento?

Per le variabili affette da stagionalità è sicuramente il meno adatto, il suo utilizzo è consigliato quando c'è la totale e completa impossibilità di estrarre un numero casuale da una popolazione; inoltre il campione sistematico è molto utilizzato quando l'oggetto di rilevazione stessa è il territorio, ad esempio per selezionare un campione al fine di contare gli alberi su un territorio e contarne il numero ogni *3 km*, in pratica il campionamento sistematico ha il pregio di spargere bene, è un modo per cercare ovunque un qualcosa mentre con il campione casuale semplice le unità potrebbero trovarsi tutte raggruppate.

Da una domanda di uno studente: che differenza c'è tra *campione a grappolo* ed *a più stadi*? Nel campione a grappoli estrarremo a caso un certo numero di pezzi, ed intervisteremo tutte le unità all'interno di ogni pezzo, mentre in un campione a più stadi non facciamo il censimento per ogni grappolo ma un'estrazione a caso e di nuovo estrazione a caso tra le unità, così per un qualsiasi numero di stadi.

Precedentemente abbiamo detto che fare un disegno campionario significa elencare tutti i possibili sottoinsiemi di una popolazione e poi per ciascuno di questi dire che probabilità ha di essere estratto, ossia prima definiamo l'insieme:

$$S = \{S_1, S_2, ....., S_n, P\}$$

e la relativa probabilità, poi associamo ad ognuno dei sottoinsiemi possibili una probabilità di essere estratto:

$$P(S) = P(S_1) = P(S_2) = ..... = P(S_I) = P(P)$$

Questo si chiama disegno campionario ma è un costrutto assolutamente teorico e non ci serve sapere la probabilità di tutto l'insieme ma la probabilità che ha una singola unità di entrare nel campione, ci serve quella che chiamiamo *probabilità d'inclusione del 1° ordine* ovvero per ogni unità dell'archivio e quindi della popolazione dobbiamo stabilire che probabilità ha ogni singola unità di essere estratta,

$$\pi_i \qquad i = 1, 2, ..., N$$

e ci serve sapere la *probabilità di inclusione del 2° ordine* ovvero per ogni coppia di unità $i$ e $j$ che probabilità hanno ambedue di essere estratte.

$$\pi_{ij} \qquad i = 1, 2, ..., N \qquad j = 1, 2, ..., N$$

Se da un punto di vista teorico il criterio di selezione è legato a doppia mandata con la distribuzione di probabilità, ovvero c'è una corrispondenza biunivoca fra criterio di selezione e distribuzione di probabilità, conoscendo la distribuzione di probabilità possiamo conoscere il criterio di selezione $P(S) = P(S_1) = P(S_2) = ..... = P(S_I) = P(P)$ che le ha generate. Questo discorso con le probabilità del primo e del secondo ordine non è più vero, perché se noi conosciamo le probabilità delle singole unità e delle coppie non riusciamo a costruire la probabilità dell'insieme, per fare questo dovremmo

avere le probabilità delle unità singole, delle coppie, delle terne etc. Quindi limitarci a queste due significa che la corrispondenza biunivoca tra criterio di selezione e distribuzione di probabilità dell'insieme viene meno, vuol dire che dove i criteri di selezione si posso costruire in modo da avere le stesse probabilità del primo e del secondo ordine, magari potranno essere diverse le probabilità del terzo ordine ma a noi non interessano, quindi le probabilità di selezione del primo e del secondo ordine non identificano più univocamente un criterio di selezione. Ci interessano solo le probabilità del primo e del secondo ordine perché dobbiamo calcolare dei parametri lineari, come la media, ossia il momento primo, oppure parametri quadratici, come la varianza ossia il momento secondo.

Vediamo ora la probabilità del primo e del secondo ordine del campione casuale semplice da *N,* che individua la popolazione e quella del secondo ordine ossia la probabilità che hanno due unità di entrare nel campione, naturalmente tutto senza reinserimento. Ora queste $\pi$ per i piani di campionamento semplici sono semplici da calcolare, quando andiamo in piani di campionamento a più stadi diventano più complicate. Useremo le due probabilità per fare stime puntuali (probabilità del primo ordine) e per stime di varianze (probabilità del secondo ordine).

$$\pi_i = \frac{n}{N}$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

### 6.2.2 In pratica

Vediamo ora il *campionamento casuale semplice*, da una popolazione di $N$ unità vogliamo estrarre $n$ unità - come facciamo? Possiamo utilizzare un computer, ma la selezione non sarà assolutamente casuale, le funzioni comunque sono deterministiche, la matematica non potrà essere mai casuale, anche se alla fine questi numeri vengono calcolati in modo talmente strano da assomigliare ad una sequenza casuale, e vengono detti numeri *pseudo - casuali*, vi è una regola deterministica che si comporta in modo simile al caso. Queste funzioni generano una distribuzione di numeri uniforme tra zero

ed uno. Tutti i numeri compresi tra zero ed uno hanno la stessa probabilità di essere estratti – ma cosa ci facciamo con la funzione, ossia con i numeri? Dobbiamo sempre estrarre *n* imprese da *N*, ma se i numeri che abbiamo sono compresi tra zero ed uno, poiché abbiamo, ad esempio, 100 imprese nell'archivio, abbiamo bisogno di numeri casuali compresi tra zero e cento - allora cosa facciamo? Moltiplichiamo semplicemente per cento, per farli diventare tra uno ed *N* li moltiplichiamo per *N* ed aggiungiamo uno, ovvero basta fare una trasformata semplice, supponiamo che il numero generato sia r, moltiplichiamo per *(N-1)* ed aggiungiamo *1*, e di quello che otteniamo ne facciamo la parte intera ovvero $r(N-1)\oplus 1=(parte\,\mathrm{int}\,era)$. Teoricamente questo procedimento ci dice il numero dell'impresa che dovremmo prendere, ad esempio la numero *3.456.789-esima.*

Quindi come vediamo nessuno impedirebbe a questo procedimento di prendere due volte la stessa impresa, quindi non va bene - allora come si fa a fare un'estrazione senza reinserimento? Esiste una strada molto semplice e veloce per evitare la laboriosità del metodo precedentemente descritto; di questi numeri compresi tra *0* ed *1*, di quelli di cui abbiamo parlato prima, ne generiamo uno per ogni unità della popolazione, tutti i numeri casuali tra *0* ed *1*; creiamo in pratica una nuova variabile di numeri casuali indipendenti compresi tra *0* ed *1*, un numero per ogni unità, e basta ordinare tutto rispetto alla popolazione in modo crescente in base alla nuova variabile, ossia rispetto ad *r,* e prendere le nostre prime *n* unità. In pratica associare ad ogni unità della popolazione un numero casuale compreso tra *0* ed *1* significa rimischiare l'archivio e prendere le prime *n* unità.

E come si fa un *campione stratificato*? Supponendo di avere già stratificato la popolazione in tre gruppi, piccole, medie e grandi imprese, certamente non ripeterò per *H* strati lo stesso procedimento, ordiniamo prima rispetto al codice di strato (facoltativo), ed all'interno dello strato ordiniamo rispetto al numero casuale, in questo caso ordiniamo rispetto a due variabili, quindi prendiamo le prime $n_h$ unità. Se vogliamo invece che le probabilità di finire tra le prime unità siano proporzionali con la variabile dimensionale introduciamo il *campionamento P.P.S.*, praticamente estraiamo un campione proporzionale alla variabile, e la variabile deve essere continua; la prima cosa da fare è trasformare la nostra variabile addetti $x_i$ in un qualcosa che si comporti in

modo simile alla probabilità, dividendo la nostra $x_i$ per la somma delle $x_i$ e moltiplicando per *n*, ed otterremo un qualcosa cosa che varia tra 0 ed 1 *n*-volte e fissiamo le probabilità del primo ordine proporzionale a $x_i$.

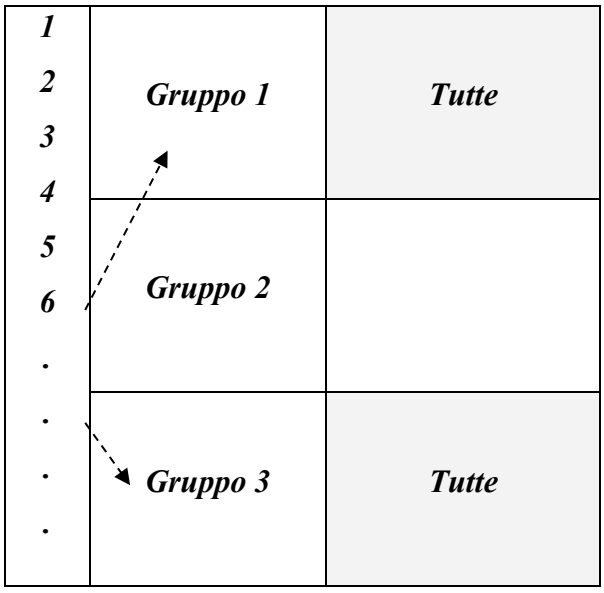
$$\pi_i = \frac{x_i}{\sum x_i} \cdot n$$

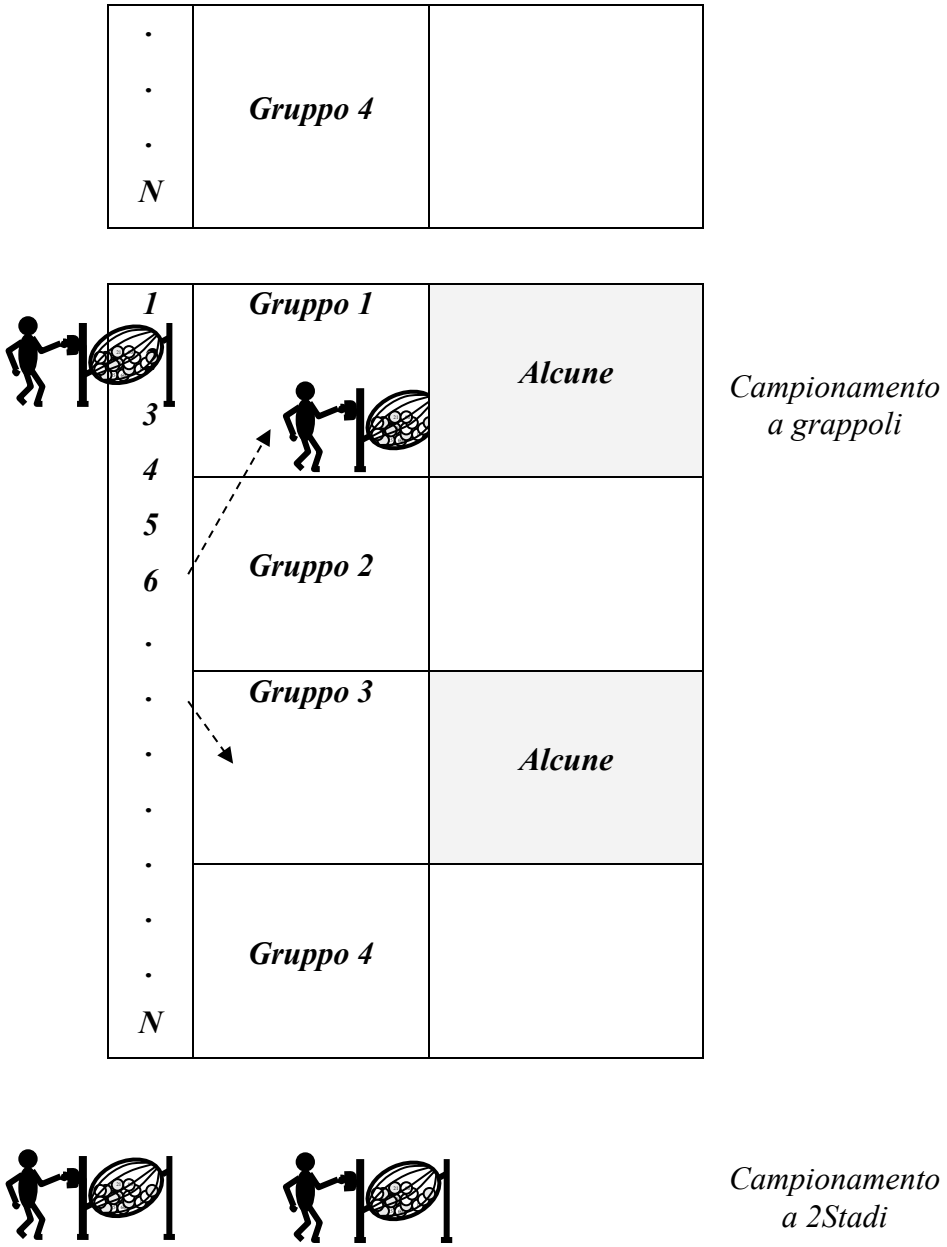
Come estraiamo un campione con queste probabilità? Creiamo questa terza variabile, ordiniamo in modo esclusivamente decrescente perché più è alto $\pi_i$ e più l'unità deve avere probabilità di cadere nel nostro campione e prendiamo le prime *n* unità; quindi ora il campione è proporzionale a $\pi_i$; questo sistema di selezione prende il nome di *Sequential Poisson Sampling* perché fare questa operazione non da una proporzionalità certa ma una proporzionalità molto approssimata.

$$\frac{\pi_i}{r}$$

Gli altri piani di campionamento sono più o meno una combinazione di quelli che sono stati trattati.

***Appendice .1***






**7. Le stime**

Vediamo ora come si fanno le stime; fingiamo di non conoscere la statistica, facciamo conto di avere *1000* imprese e di estrarne tramite un campione casuale semplice *100*; vogliamo stimare il totale del valore aggiunto sull'intera popolazione, mandiamo il questionario, riceviamo le risposte e constatiamo che il totale all'interno delle imprese campionate è 50.000.000, ora quanto è il totale della popolazione? Non dobbiamo fare altro che:

$$50.000.000 \cdot \frac{1000}{100} \qquad \Rightarrow \qquad \frac{N}{n}$$

questo $\frac{1000}{100}$ non è altro che l'inverso della probabilità d'inclusione, ovvero definiamo:

$$W_i = \frac{1}{\pi_i}$$

che rappresenta il *peso dell'importo all'universo*, ossia quante unità della popolazione sono rappresentate dalla nostra unità estratta o nel caso in cui venisse estratta. Significa quindi che un totale, se calcolato in questo modo rappresenta una stima del totale della popolazione, ottenuto tramite i pesi di riporto all'universo. E successivamente definiamo:

$$\hat{t}_{HT} = \sum_{i=1}^{n} w_i y_i$$

Un totale calcolato cosi è un totale pesato e rappresenta una stima del totale della popolazione che si chiama stimatore di *Horvitz-Thompson*, indicato con la sigla *HT;* questo stimatore, ottenuto per semplice ponderazione dei dati ha la proprietà di essere corretto ed è il più efficiente tra gli estimatori del totale; essere corretto significa:

$$E(\hat{t}_{HT}) = t$$

ossia il valore atteso dello stimatore è uguale al totale stimato della popolazione. Ovvero facendo una stima per ogni campione che può essere estratto (cioè al variare dei campioni estratti) e successivamente calcolando la media di tali stime si giunge a tale conclusione: in media lo stimatore di *Horvitz-Thompson* è uguale al totale vero della popolazione. Per quanto riguarda l'efficienza si ha che la varianza dello stimatore deve essere la minore tra le varianze di tutti i possibili stimatori di un totale a condizione però che sia corretto, ossia la larghezza della distribuzione campionaria di questo stimatore è la più stretta tra tutte le possibili distribuzioni, ossia la più vicina al valore vero.

$$V(\hat{t}_{HT}) < V(\hat{t})$$

## 7.1 La stima del totale in pratica

Supponiamo di avere un archivio e di voler soddisfare i seguenti punti:

*1) Generare un campione casuale semplice, SRS , di numerosità n=5;*

*2) Stratificare per codice di attività economica (agricoltura, industria, servizi) dove $n_1=1$, $n_2=2$, $n_3=2$ ed estrarre il campione;*

*3) Generare un P.P.S. n=5*

*4) Rilevare per ciascuna impresa le spese di ricerca e sviluppo, R&S; le imprese intervistate ci forniscono i cinque valori in migliaia di euro, 150, 110, 50, 0, 200; fare la stima $\hat{t}_{HT}$ per tutti e tre i piani di campionamento precedenti;*

| *Imprese* | *Attività Economica* | *n. addetti* | *n. Random* | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 10 | 0,85 | | | | |
| 2 | | 15 | 0,75 | | | | |
| 3 | | 5 | 0,10 | | | | |
| 4 | | 10 | 0,05 | | | | |
| 5 | | 15 | 0,35 | | | | |
| 6 | | 20 | 0,40 | | | | |
| 7 | | 30 | 0,75 | | | | |
| 8 | | 40 | 0,60 | | | | |
| 9 | | 10 | 0,20 | | | | |
| 10 | | 20 | 0,30 | | | | |
| | | 175 | | | | | |

*1 Agricoltura*

*2 Industria*

*3 Servizi*

*Naturalmente abbiamo tralasciato tutte le altre cose che dovevamo mettere all'interno dell'archivio, è molto semplificato*

***1. Svolgimento***

| *Imprese* | *Attività Economica* | *n. addetti* | *n. Random* | *Riordiniamo i numeri Random* | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 10 | 0,85 | 0,05 *(4)* | | | |
| 2 | | 15 | 0,75 | 0,10 *(3)* | | | |
| 3 | | 5 | 0,10 | 0,20 *(9)* | | | |
| 4 | | 10 | 0,05 | 0,30 *(10)* | | | |
| 5 | | 15 | 0,35 | 0,35 *(5)* | | | |
| 6 | | 20 | 0,40 | 0,40 | | | |
| 7 | | 30 | 0,55 | 0,55 | | | |
| 8 | | 40 | 0,60 | 0,60 | | | |
| 9 | | 10 | 0,20 | 0,75 | | | |
| 10 | | 20 | 0,30 | 0,85 | | | |
| | | 175 | | | | | |

In questo caso abbiamo riordinato i numeri random in modo crescente.
Nel compito naturalmente verrà specificato se ordinare in modo crescente o decrescente.

***2. Svolgimento***

Supponiamo che le unità siano già ordinate rispetto al codice di attività economica, ma potevano anche essere disordinate, mischiate, inoltre dall'esercizio precedente abbiamo già ordinato i numeri Random, quindi non ci rimane altro che prendere le nostre imprese.

| *Imprese* | *Attività Economica* | *n. addetti* | *n. Random* | *Riordiniamo i numeri Random* | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 0,85 | 0,05 *(4)* | | | |
| 2 | 1 | 15 | 0,75 | 0,10 *(3)* | | | |
| 3 | 2 | 5 | 0,10 | 0,20 *(9)* | | | |
| 4 | 2 | 10 | 0,05 | 0,30 *(10)* | | | |
| 5 | 2 | 15 | 0,35 | 0,35 | | | |
| 6 | 2 | 20 | 0,40 | 0,40 | | | |
| 7 | 3 | 30 | 0,55 | 0,55 | | | |
| 8 | 3 | 40 | 0,60 | 0,60 | | | |
| 9 | 3 | 10 | 0,20 | 0,75 *(2)* | | | |
| 10 | 3 | 20 | 0,30 | 0,85 | | | |
| | | 175 | | | | | |

***3. Svolgimento***

| *Imprese* | *Attività Economica* | *n. addetti* | *n. Random* | *Riord. i numeri Random* | $\pi_i = \frac{x_i}{\sum x_i} \cdot n$ *sono probabilità d'inclusione* | $\frac{\pi_i}{r}$ | *Ordin.* $\frac{\pi_i}{r}$ *decres.* |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 0,85 | 0,05 *(4)* | 0,29 | 0,34 | 5,8*(4)* |
| 2 | 1 | 15 | 0,75 | 0,10 | 0,43 | 0,57 | 1,9*(7)* |
| 3 | 2 | 5 | 0,10 | 0,20 *(9)* | 0,14 | 1,4 | 1,9*(8)* |
| 4 | 2 | 10 | 0,05 | 0,30*(10)* | 0,29 | 5,8 | 1,56*(9)* |
| 5 | 2 | 15 | 0,35 | 0,35 | 0,43 | 1,23 | 1,45*(10)* |
| 6 | 2 | 20 | 0,40 | 0,40 | 0,57 | 1,42 | 1,42 |
| 7 | 3 | 30 | 0,55 | 0,55 *(7)* | 0,86 | 1,56 | 1,4 |
| 8 | 3 | 40 | 0,60 | 0,60 *(8)* | 1,14 | 1,9 | 1,23 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 3 | 10 | 0,20 | 0,75 | 0,29 | 1,45 | 0,57 |
| 10 | 3 | 20 | 0,30 | 0,85 | 0,57 | 1,9 | 0,34 |
| | | 175 | | | | | |

Quando costruiamo la nuova variabile possiamo anche procedere in questo modo: $\frac{\sum x_i}{n}\pi_i = x_i$ per poi dividere $x_i$ per un solo numero $\frac{\sum x_i}{n}\pi_i = \frac{x_i}{\frac{\sum x_i}{n}}$ e velocizziamo i calcoli. Inoltre se guardiamo la casella colorata quella unità la prendiamo sicuramente, è un problema di *worming,* non è un errore, se ha una probabilità uguale ad *1* o maggiore quella unità la prendiamo sicuramente; quando moltiplichiamo per *n,* in popolazioni in cui la variabile dimensionale è molto concentrata, otteniamo moltissime unità che hanno probabilità maggiore di *1.* Se la popolazione è molto concentrata, ribadiamo che l'unità la prendiamo sicuramente, facciamo finta di non vederla ed andiamo avanti, senza fare nessun rapporto con i numeri Random; questa unità sta nel campione per principio, è un problema che è inutile porci. Nei campioni d'imprese significa che quando c'è forte concentrazione avremo probabilità maggiori di uno.

### *4. Svolgimento (per il campione casuale semplice estratto, R&S)*

Passiamo alle stime basate sui campioni che abbiamo calcolato fino ad ora, e prendiamo in considerazione come risposta per le spese di ricerca e sviluppo le cifre che sono nel testo dell'esercizio ed in ogni tabella le unità colorate di verde sono quelle che abbiamo estratto, inoltre poiché è un campione casuale semplice le mie unità hanno tutte la stessa probabilità di essere estratte $\pi_i = \frac{n}{N}$ quindi $W_i = \frac{N}{n}$ ossia $W_i = \frac{1}{\pi_i}$ ed $Y_i$ è la variabile che riguarda le spese di ricerca e sviluppo. Nella tabella abbiamo tralasciato le colonne dei numeri Random e delle altre variabili perché non ci servono più, prendiamo solo le $Y_i$ delle risposte.

| *Imprese* | *Attività Economica* | $Y_i$ | $W_i = \frac{I}{\pi_i}$ *peso di riporto all'universo* | $\hat{t}_{HT} = \sum_{i=1}^{n} W_i Y_i$ | |
|---|---|---|---|---|---|
| 1 | 1 | | | | |
| 2 | 1 | | | | |
| 3 | 2 | 150 | 2 | 300 | |
| 4 | 2 | 110 | 2 | 220 | |
| 5 | 2 | 50 | 2 | 100 | |
| 6 | 2 | | | | |
| 7 | 3 | | | | |
| 8 | 3 | | | | |
| 9 | 3 | 0 | 2 | 0** | |
| 10 | 3 | 200 | 2 | 400 | |
| | | | | 1020 | |

$$\hat{t}_{HT} = \sum_{i=1}^{n} W_i Y_i = 1020$$

** Attenzione che uno zero ed un trattino sono due cose totalmente diverse, qui ci hanno risposto ma non spendono nulla, quando invece il dato è mancante si mette il trattino.

### *4. Svolgimento (per il campione casuale stratificato estratto R&S)*

Attenzione che le risposte rimangono le stesse ma di solito cambiano, quelle riguardanti la

variabile, anche qui $\pi_i = \frac{n}{N}$ quindi $W_i = \frac{N}{n}$ ossia $W_i = \frac{1}{\pi_i}$;

| *Imprese* | *Attività Economica* | $Y_i$ | $W_i = \frac{I}{\pi_i}$ *peso di riporto all'universo, in* | $\hat{t}_{HT} = \sum_{i=1}^{n} W_i Y_i$ |
|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| | | | *questo caso per ogni strato,* $w_{i,h} = \frac{n_h}{N_h}$ *quindi* $w_{i,h} = \frac{1}{\pi_{i,h}}$ *dello strato* | | |
| 1 | 1 | | | | |
| 2 | 1 | 150 | 2 | 300 | |
| 3 | 2 | 50 | 2 | 100 | |
| 4 | 2 | 110 | 2 | 220 | |
| 5 | 2 | | | | |
| 6 | 2 | | | | |
| 7 | 3 | | | | |
| 8 | 3 | | | | |
| 9 | 3 | 0 | 2 | 0 | |
| 10 | 3 | 200 | 2 | 400 | |
| | | | | 1020 | |

Come vediamo anche se cambiano le probabilità di inclusione i pesi di riporto all'universo rimangono uguali, questo per dire che due piani di campionamento diversi possono generare le stesse probabilità di inclusione, stessi pesi di riporto all'universo.

***4. Svolgimento (per il campione con probabilità proporzionale PPS, R&S)***

| *Imprese* | $Y_i$ *(R&S)* | $W_i = \frac{I}{\pi_i}$ *peso di riporto all'universo* | $\hat{t}_{HT} = \sum_{i=1}^{n} W_i Y_i$ | |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | 150 | 3,45 | 517,5 | |
| 5 | | | | |
| 6 | | | | |
| 7 | 0 | 1,16 | 0 | |
| 8 | 110 | 2,33 | 256,3 | |
| 9 | 200 | 3,45 | 690 | |
| 10 | 50 | 7,14 | 357 | |
| | | | 1820,8 | |

Ora come calcoliamo il peso di riporto all'universo del P.P.S.? Praticamente avremo una diversa probabilità d'inclusione per ogni unità e dunque un diverso peso di riporto per ogni unità, inoltre il worming che avevamo trovato 1,14 avrà peso di riporto uguale ad 1 logicamente, perché dato che l'unità l'abbiamo messa dentro forzatamente non può che rappresentare se stessa, ogni unità che viene censita non può che rappresentare se stessa.

Nota bene che a livello pratico quando andiamo ad effettuare la stima dopo aver estratto il campione non c'interessa in che ordine sommiamo le risposte delle unità statiche (i valori della *Y*), cosa fondamentale è che alle stesse unità corrispondano le relative risposte. Noi per convenzione rispetteremo l'ordinamento ottenuto dall'estrazione, all'esame sarà questo il modo di procedere; dall'ultimo svolgimento avremo *8,4,10,7,9*.

**7.2 Osservazioni**

Sappiamo che il codice identificativo dell'impresa va messo anche sul questionario per riagganciarla all'archivio, quindi se volevamo fare stime di valor medio invece che stime di totale dovevamo calcolare il valore medio delle spese in ricerca e sviluppo e la prima cosa che viene in mente, poiché abbiamo stimato un totale, è di dividere per N; non è detto che questa sia una soluzione sbagliata, ma sicuramente per essere giusta vuol dire che le unità devono avere le stesse probabilità di inclusione o sul totale della popolazione o su dei pezzi, quindi dovremmo avere o un campione casuale semplice o stratificato, e quindi calcoleremo la stima della media con

$$\hat{\mu}_{HT} = \frac{\hat{t}_{HT}}{N}$$

o comunque è indifferente fare anche nell'altro modo

$$\hat{\mu}_{HT} = \frac{\sum y_i \cdot w_i}{\sum w_i}.$$

Nel caso del campione con probabilità proporzionale invece la cosa cambia poiché non tutte le unità hanno la stessa probabilità di inclusione e la prima ipotesi che viene in mente è che la stima della media sia in un certo senso ponderata sulla probabilità d'inclusione della nostra variabile, quindi calcoleremo la stima della media con

$$\hat{\mu}_{HT} = \frac{\sum y_i \cdot w_i}{\sum w_i}$$

Ora, negli altri piani di campionamento il peso di riporto all'universo dipende solo dal numero di unità che ci sono o all'interno di un gruppo o nella popolazione quindi rispetta solo questo. Invece nel PPS i pesi di riporto sono calcolati non sulla base di quante unità ci sono ma sul numero di addetti che ci sono, in modo da rispettare il numero di addetti della popolazione. Quindi ci siamo agganciati ad un'altra variabile. Inoltre rispettare il numero di addetti è un bene perché il numero di addetti è correlato

con la nostra variabile, e solo in questo caso, abbiamo progettato il nostro campione sulla base di un'altra cosa. Alla fine del corso vedremo come calcolare stimatori che rispettano più vincoli di coerenza.
Questa caratteristica si indica con l'espressione "*campione basato sul disegno*" ossia che rispetta una sola cosa, solo il disegno campionario.

## 8. Richiami d'inferenza

Supponiamo di avere la nostra solita popolazione, e costruiamo tutti i possibili sottoinsiemi

$$S = \{S_1, S_2, ....., S_n, P\}$$

sappiamo dire che probabilità c'è per il campione di essere selezionato, ossia definiamo il disegno campionario,

$$P(S_1), P(S_2), ....., P(S_I), P(P)$$

Poi abbiamo detto che la stima del totale si fa in questo modo, ossia la stima del totale di tutte le unità contenute nel campione

$$\hat{t}_{HT} = \sum_{i \subset s} w_i \cdot y_i$$

e dove

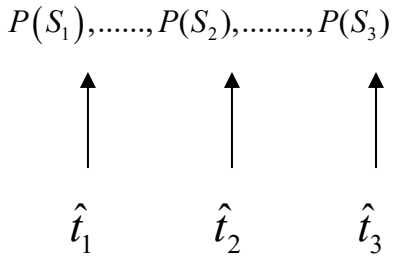
$$w_i = \frac{1}{\pi_i}$$

è l'inverso delle probabilità di inclusione che riscritto in modo più formale diventa

$$\pi_i = P(i \subset s)$$

ossia la probabilità dell'unità di *i* di essere inclusa nel campione selezionato.

Per ciascun sottoinsieme abbiamo un valore dello stimatore, ovvero se prendiamo un campione piuttosto che un altro abbiamo una stima diversa




È vero che nella realtà calcoliamo una sola stima ma ne può venir fuori una qualsiasi, a meno che non abbiamo stabilito per quei campioni una probabilità uguale a zero.

Il $\hat{t}_s$ esprime la nostra stima, ad esempio la stima delle spese in ricerca e sviluppo, e la probabilità di ottenere ciascuna di queste stime è data dalla probabilità del campione di essere generato, quindi, rappresentando su un grafico i possibili valori di questa stima, sulle *X*, e in ordinata le probabilità di essere estratto, probabilità che nascono dal campione, otteniamo una figura simile

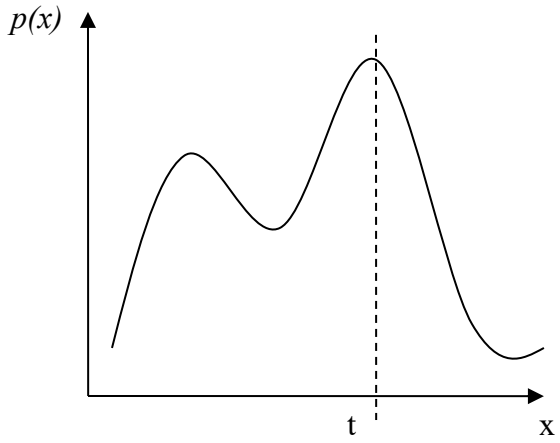



*Fig. 8.1 – Densità di probabilità dell'Horvitz-Thompson*

dove

$$E(\hat{t}_i) = t = \sum \hat{t}_i \cdot p(S_i)$$

Questa funzione prende il nome di distribuzione campionaria dello stimatore $\hat{t}_{HT}$. Questa distribuzione campionaria è come se fosse una distribuzione di frequenza, e di questa distribuzione possiamo calcolare il valore medio, la varianza, o qualsiasi altra stima su una distribuzione di dati, quindi avrà un suo valore medio che è, all'estrarre dei campioni, quanto mediamente ci attendiamo che sia il valore del nostro stimatore. Questo valore medio non è altro che la somma di tutti possibili valori di $\hat{t}_s$ per le probabilità di essere osservato, è una media ponderata $E(\hat{t}_s)$. Quindi supponendo di estrarre tutti possibili campioni dalla popolazione -cosa assolutamente teorica- ed osservarli, lo stimatore di *HT* che valore assume in media? Se noi osservassimo infiniti campioni, e su ciascuno di questi facessimo contemporaneamente la stima, in media non ci allontaneremmo dal valore $t$, quello che c'interessa del criterio di selezione è la stima, anche se facciamo la stima una volta sola, in media non sbagliamo.

Questa proprietà degli stimatori si chiama *correttezza*, ossia in media rispettare il valore vero della popolazione; significa che una distribuzione come quella del grafico precedente è incentrata sul valore vero della popolazione. In pratica il nostro stimatore $\hat{t}_{HT}$ a qualsiasi variabile lo applichiamo ha questa proprietà, può capitare che sia più alto o più basso ma questa casualità non è sistematica.

Ovviamente se al posto dell'uguale ci fosse stato $E(\hat{t}_i) \neq t$ lo stimatore sarebbe stato non corretto o distorto, infatti esistono degli stimatori che non hanno questa proprietà .

Ora qual è l'altra proprietà che interessa? Si tratta di una proprietà che riguarda la larghezza della densità campionaria e che è fondamentalmente legata alla varianza della distribuzione campionaria. Se al variare dello stimatore invece di fare $\hat{t}_{HT} = \sum_{i \subset s} w_i y_i$ utilizziamo ad esempio un altro tipo di stimatore con una distribuzione di probabilità molto più stretta rispetto al valore vero della popolazione rendiamo più efficienti le stime poiché sbagliamo meno. In pratica se la proprietà di correttezza ci dice come si distribuisce il nostro stimatore, ora questa nuova proprietà ci dice che il nostro stimatore si allontana a probabilità più basse dal valore vero della popolazione, ossia ci dice di quanto si allontana in media dal valore vero.

Quindi avere minor variabilità o varianza, avere una funzione di densità di probabilità più schiacciata intorno al valore vero della popolazione sta a indicare che i valori vicino al valore vero della popolazione hanno più un probabilità di essere osservati, ossia abbiamo più probabilità di sbagliare meno.

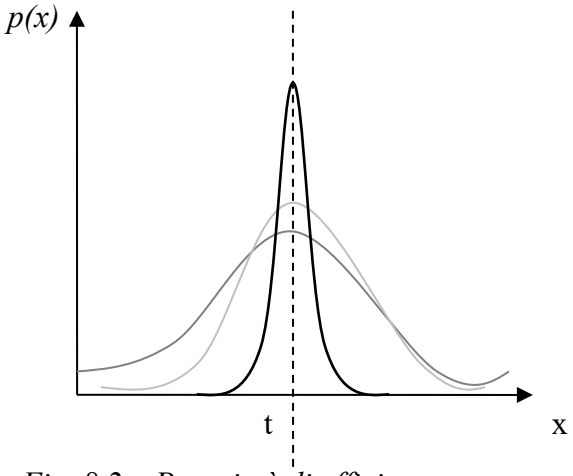



*Fig. 8.2 – Proprietà di efficienza*

Questa proprietà si chiama *efficienza* ed è un concetto relativo in quanto si può dire che uno stimatore è più efficiente di un altro - cioè la varianza di questo stimatore è minore rispetto a quella di un altro stimatore - solo se entrambi gli stimatori sono corretti. Infatti potrebbe accadere che un altro stimatore sia molto stretto, molto efficiente ma con un'alta probabilità di avere come valore medio qualcosa di totalmente diverso. Quindi gli stimatori devono essere tutti e due corretti, e in tal caso l'efficienza è definita in termini di varianza ovvero di larghezza intorno alla media. E se uno dei due estimatori non è corretto?

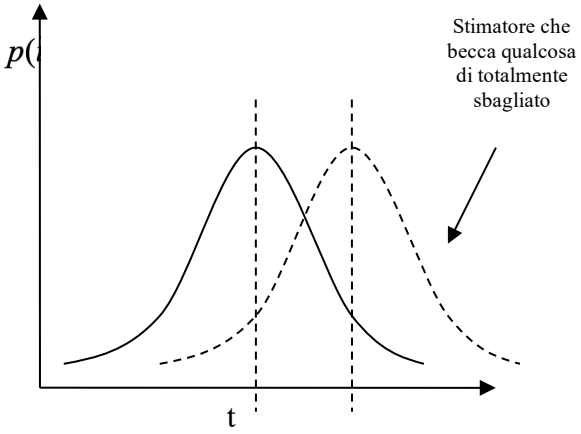



*Fig. 8.3 – Stimatore corretto e distorto*

Quando gli stimatori sono anche distorti al posto della varianza si usa il *MSE*, ad esempio dello stimatore di Horvitz-Thompson. Ma cos'è il *MSE? S*e la varianza è la distanza media di tutta la distribuzione dal valore centrale (ovvero dalla media), il *MSE* è la distanza quadratica media rispetto a qualsiasi altro valore, è la distanza quadratica media di tutta la distribuzione dal valore vero della popolazione, ovvero non dalla propria media ma dal valore vero della popolazione. Il *MSE* altro non è che la varianza dello stimatore più la distorsione al quadrato. E la *distorsione* è la differenza tra il valore medio dello stimatore ed il valore vero della popolazione elevato al quadrato, ovvero il *MSE* (errore quadratico medio) non è altro che la varianza, ossia quanto è larga la distribuzione più il quadrato della distanza fra la propria media ed il valore medio della popolazione.

$$MSE(\hat{t}_{HT}) = V(\hat{t}_{HT}) + B^2$$

dove la distorsione è

$$B^2 = \left[ E(\hat{t}_{HT}) - t \right]^2$$

Ora nel caso in cui gli stimatori sono corretti, la distorsione è uguale a zero e questa definizione assume una forma più generale. Se la distorsione è zero si ricade nel caso della varianza. In termini grafici non è altro che:

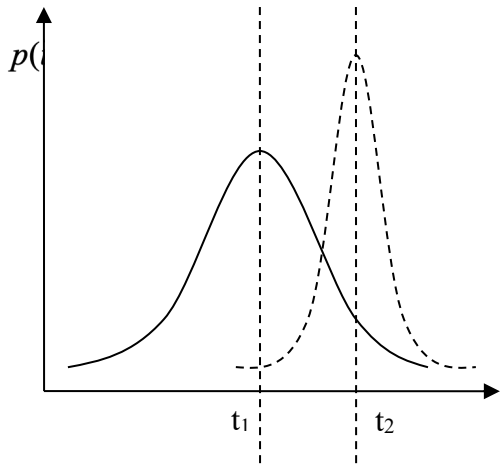



*Fig. 8.4 – L'errore quadratico medio e la distorsione*

Ora se $t_1$ è il valore vero della popolazione, ed abbiamo due stimatori uno corretto con varianza molto ampia ed un altro distorto ma con varianza molto stretta, il secondo sbaglia sempre per eccesso ma sbaglia di poco l'altro invece in media ha il valore giusto ma ha probabilità molto alte di discostarsi dal valore medio, quindi tra i due usando il nostro criterio dell'errore quadratico medio, tra i due potrebbe essere meglio prendere quello distorto; infatti la prima componente dello *MSE*, cioè la varianza, potrebbe essere talmente alta da compensare un piccolo valore di distorsione. Ciò vuol dire che entrano in gioco due fattori, è la somma di due fattori che conta: nel primo stimatore la distorsione è nulla mentre la varianza è altissima, nel secondo invece la varianza è piccolissima e c'è una distorsione, e la somma dei due potrebbe essere tale da essere minore della varianza del primo stimatore.

Facciamo un esempio pratico; tra due autobus che fanno lo stesso tragitto, uno dei due mezzi impiega $30\,\text{min.} \pm 20\,\text{min.}$ mentre l'altro ci mette $25\,\text{min.} \pm 20\,\text{min.}$
Quale prendiamo dei due? Dipende dalla somma dei due errori precedenti, varianza e distorsione, tempo medio di percorrenza più un errore dovuto alla varianza. La varianza rappresenta l'incertezza dello stimatore, come si muove al muoversi dei campioni mentre la distorsione rappresenta il tempo medio di percorrenza nel nostro esempio. Dunque quanto siamo disposti a sbagliare sistematicamente in positivo o in negativo in cambio di una forte o leggera riduzione dell'incertezza? La risposta è la somma tra varianza e distorsione, e lo stimatore che ha questa somma minore è preferibile ad un altro. Esistono svariate altre proprietà degli stimatori ma nella teoria dei campioni servono solo queste.

Cosa sono gli *intervalli di confidenza*? Un intervallo di confidenza, banalmente, è un metodo di stima che tiene conto dei problemi di cui abbiamo trattato prima, ovvero una volta che siamo sicuri di sbagliare, o per distorsione o per variabilità campionaria, per cercare di evitare il nostro dramma esistenziale dovuto al sicuro errore, per evitare questa frustrazione è più ragionevole dire, invece che un valore solo, ammettere che c'è una variabilità e che ci sono questi problemi, e quindi considerare un intervallo, un valore +/- un errore, e dire che il valore può stare in questo intervallo.

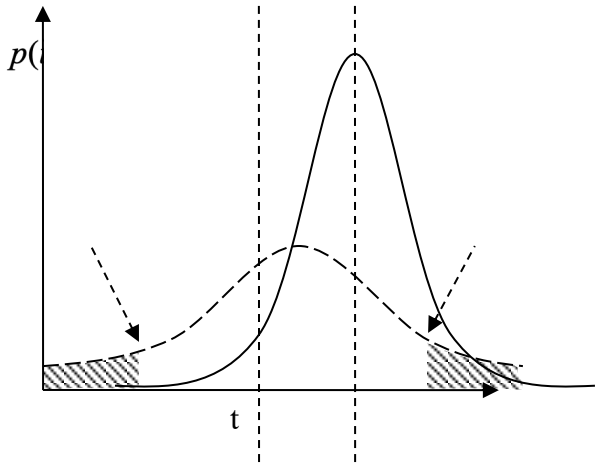



*Fig. 8.5 – Intervallo di confidenza*

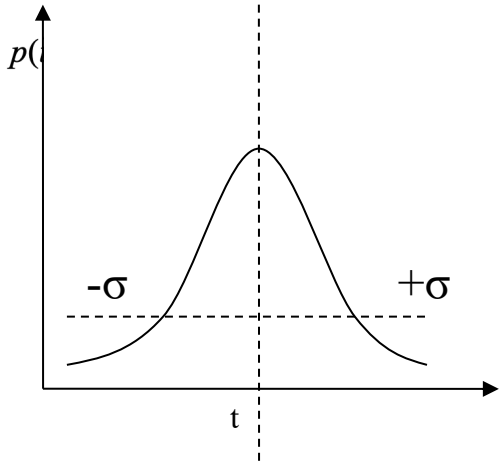



*Fig. 8.6 – Scarti dal valore vero della popolazione*

Data la distribuzione campionaria del grafico, escludiamo i casi estremi e quindi togliamo una percentuale. Una probabilità di campioni estremi rappresentano la "nostra sfortuna", ad esempio stimiamo che la sfortuna è misurata da uno scarto del *10%* (e quindi il 5% in più o in meno) e nel 90% dei campioni il valore vero della popolazione è compreso tra i valori evidenziati. Questa distribuzione campionaria la possiamo

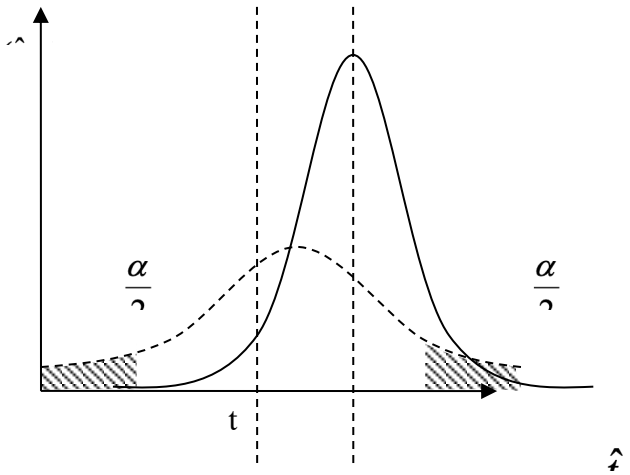
ricostruire sapendo che dipende dal valor medio che abbiamo stimato e dalla sua varianza. Una volta saputo dove la distribuzione è centrata e quanto è larga, si può sapere dove sono i campioni estremi. Per l'Horvitz-Thompson avevamo visto il valore medio, quindi lo conosciamo.

$$\hat{\mu}_{HT} = \frac{\sum y_i \cdot w_i}{\sum w_i}$$

oppure

$$\hat{\mu}_{HT} = \frac{\hat{t}_{HT}}{N}$$

a seconda dei casi, ossia scegliamo la formula in base al disegno campionario. Se assumiamo di conoscere la varianza, siamo in grado di calcolare i due valori tali che ciascuno esclude il 10% dei casi - e come si fa a calcolare le code? Allora, la *Gaussiana* è una distribuzione teorica che ha una tipica forma campanulare centrata su due parametri, su un valor medio $\mu$ e sulla varianza $\sigma^2$ , più è alta la varianza e più la campana si allarga, invece cambiando il valor medio la curva si sposta a destra ed a sinistra. Ora esiste un teorema che si chiama *teorema del limite centrale* che ci dice una cosa molto importante: se noi sommiamo o mediamo degli eventi casuali di qualsiasi forma o genere ed indipendenti, all'aumentare della dimensione di questa somma, più ne sommiamo o più ne mediamo e più ci avviciniamo alla Gaussiana. Di solito, per campioni $> 30$ elementi si può ritenere di essere abbastanza vicini alla Gaussiana, ed una volta che arriviamo a conoscere la media e la varianza, conosciamo in pratica la campana. Successivamente esistono delle tabelle che ci dicono quali sono i valori da prendere per far venir fuori la nostra percentuale 10%, 5%, 1% etc. Grazie alle tabelle siamo in grado di dire quali sono quei valori che ci fanno tenere fuori una percentuale di campioni. Quindi esprimendo il concetto in termini grafici ed analitici, un intervallo di confidenza è




$$\hat{t}_{HT} \pm z_{\frac{\alpha}{2}} \sqrt{V(\hat{t}_{HT})}$$

Quindi tutto questo per arrivare a dire cosa? Arrivati a questo punto, cosa ci manca in definitiva per fare delle stime complete? L'Horvitz-Thompson sappiamo che è corretto, e sappiamo anche che tra gli stimatori lineari, ovvero quelli che si trasformano in medie o in somme, è anche il più efficiente, ma non conosciamo ancora la varianza. Prima o poi dovremo conoscerla poiché è ragionevole pensare che se qualcuno ci viene a chiedere la stima del totale, ad esempio della ricerca e sviluppo in Italia, dovremo certamente dare un valore che però ammette un errore, e quindi per cautelarci diamo un valore ed un intervallo, una stima più o meno qualcosa. Ammettere un errore sulla nostra stima è ragionevole, quindi una misura quantitativa dell'incertezza è un ottimo parametro della qualità della nostra stima. Aggiungiamo che nelle indagini l'intervallo di confidenza non si usa mai ovvero non si fa la somma più o meno ma si dà la stima e la varianza, non si usa fare la doppia somma per il valore della Gaussiana, ci si limita a calcolare i due valori (stima e varianza); e questa è una cosa che vediamo spesso nella vita di ogni giorno, soprattutto nei telegiornali, anche se in molti casi non viene divulgato un intervallo ma solo un valore, come nel caso del dato sull'inflazione.

## 8.1 Exit Pools, proiezioni ed intenzioni di voto

L'unico caso in cui viene divulgato un intervallo è quello degli *Exit Pools*; in questo caso l'obiettivo vero è cautelarsi rispetto a possibili errori per non essere messi alla gogna in quanto società di statistica che non sa prevedere un esito elettorale. Soprattutto quando le situazioni di voto sono molto vicine tra i candidati, ad esempio *49%* e *51%*, si specifica sempre bene qual è il possibile errore poiché potrebbero nascere delle polemiche successivamente sulla qualità dell'indagine.

Rimaniamo sull'aspetto Exit Pools: siamo in clima elettorale e ci chiamano per avviare un'indagine. Innanzitutto facciamo una distinzione tra i risultati prima e dopo la votazione: quelli prima si chiamano *sondaggi.* Come ricordiamo infatti una cosa è chiedere un fatto ed una cosa è chiedere un'intenzione; da un punto di vista di marketing una cosa è chiedere l'acquisto passato mentre un'altra è chiedere l'intenzione, sono due cose molto diverse che quindi affrontiamo in modo separato. Noi rimaniamo sugli *Exit Pools*; ciò significa indagare prima dello spoglio ma dopo che il pubblico ha votato, equivale al censimento, durante la giornata quando il fatto è già accaduto possiamo chiedere. Inoltre gli *Exit Pools* sono qualcosa di diverso dalle *Proiezioni*. Le *Proiezioni* sono più affidabili degli *Exit Pools* statisticamente parlando. *Exit Pools* significa prendere una persona una volta uscita dal collegio, consegnarle una scheda, un fac-simile, farla accomodare in un'altra cabina simile e chiederle di votare. La *Proiezione* invece non chiede di ripetere un fatto e quindi non può essere diverso da quello reale, mentre con gli *Exit Pools* dentro la vera cabina i votanti possono aver dato un tipo di voto mentre nell'altra magari un voto diverso, si potrebbe mentire comunque. Mentre con la *Proiezione* non facciamo un campione di elettori ma un campione di seggi elettorali. Semplicemente ci si organizza con gli scrutatori per vie ufficiali, che danno il permesso o spingono questi seggi elettorali a velocizzare lo scrutinio. Quindi se con gli *Exit Pools* la nostra unità statistica è l'elettore con le *Proiezioni* la definizione di unità statistica è il seggio elettorale, avremo un campione di seggi elettorali. Ma qual è l'unità di analisi? L'unità di analisi naturalmente è il partito politico.

Gli *Exit Pools* però hanno un grandissimo vantaggio: si possono iniziare prima della chiusura dei seggi; gli unici dati però che ci permettono di collegare il partito votato con le caratteristiche strutturali della persona - ed è per questo che si fanno - sono le

*intenzioni di voto*, ovvero tra tutti questi sistemi di rilevazione ce n'è uno solo che ci permette di dire se le donne votano a destra e sinistra, se giovani votano a destra o a sinistra, etc. Ad esempio una società di statistica ci telefona e prima ci chiede le nostre caratteristiche, poi la nostra intenzione di voto, quindi i numeri che vengono divulgati prima del voto non servono affatto per dichiarare un vittoria o una sconfitta, perché l'intenzione può cambiare successivamente, ci serve solo per collegare le intenzioni di voto alle variabili esplicative in modo da indirizzare la campagna elettorale in un certo modo, le intenzioni non servono per stimare la percentuale di voto, ma servono per calibrare la politica di marketing politico, la possiamo vedere come la vendita di un prodotto, la stessa identica cosa infatti si fa anche per indagini sulle vendite al fine di calibrare la catena di montaggio, per poter discriminare tra chi lo comprerà e chi non lo comprerà, e successivamente, dopo la vendita, si può fare l'indagine sul consumo del nostro prodotto che è l'equivalente della proiezione.

Ma come facciamo a selezionare un campione di collegi elettorali? Qui la funzione del questionario è svolta dalla scheda elettorale, quindi tutto l'aspetto del questionario non è trattato.

Prima cosa stratifichiamo i seggi sulla base delle elezioni precedenti, magari facciamo tre strati soli, seggi di destra, seggi di centro e seggi di sinistra. Naturalmente utilizziamo i seggi con i risultati delle elezioni precedenti dello stesso tipo (amministrative, politiche, europee). Successivamente stratifichiamo per regioni se non addirittura per province; poi, a causa dei tempi di ritorno dei dati, viene fatto un campione di numerosità piccola e in seguito 1 campione di numerosità molto più grande fino a quando non si raggiunge una dimensione ottimale in base ai tempi di risposta e la varianza non è ridotta ad un livello accettabile.

Ma come faccio a fare un campione di questo tipo? Cioè come facciamo a fare tre campioni di numerosità diversa in modo che ognuno sia contenuto dentro l'altro? Semplicemente tenendo fermo il numero random, riordinandolo in modo tale che i primi *10* formeranno il primo campione, i primi *100* il secondo campione, gli ultimi *1000* il terzo campione.

Ma è possibile che nessuno tenga presente che i collegi elettorali possono essere di dimensioni diverse? L'estrazione deve essere fatta con probabilità proporzionale al

numero di votanti, quindi in base al numero di aventi diritti al voto e dunque alle dimensioni del collegio elettorale.
In questo caso il campionamento sistematico non viene utilizzato perché già abbiamo la lista dei nostri collegi; lo utilizza chi non ha nessuna competenza di estrazione di numeri random ovvero quando si richiede a qualcuno che sta per strada, o qualcuno al catasto o all'anagrafe, di estrarre a caso.
Il problema grosso di qualsiasi stima che viene fatta in una *proiezione*, un *Exit Pools* e così via, è proprio l'intervallo di confidenza perché la stima di una percentuale ha varianza pari a:

$$V = \frac{P(1-P)}{n}$$
$$Caso(1): 0,1(1-0,1) = 0,05$$
$$Caso(1): 0,5(1-0,5) = 0,25$$

ovvero se la probabilità di essere votato dei due contendenti è *0,10* e l'altra è *0,90* la varianza sarà molto bassa (*0,1 x 0,9)*, se invece i due sono molto vicini abbiamo 0,25 che come vediamo è cinque volte il caso uno. In quest'ultimo caso il fenomeno stesso ci indica che, più sono vicini i contendenti in termini di probabilità di vittoria e più si alza la varianza, quando invece sarebbe proprio questo il caso in cui avremmo bisogno di una varianza minore. E quindi nel caso degli *Exit Pools*, più ci avviciniamo a situazioni *41%-51%* più la numerosità campionaria che sta al denominatore deve essere molto grande per schiacciare la varianza. Proprio quando si necessita di maggiore certezza, gli intervalli di confidenza diventano più inefficienti e la stima campionaria più "ballerina": la varianza della stima, la varianza della percentuale che noi diamo al vincitore aumenta man mano che ci avviciniamo a valori sempre più stretti.

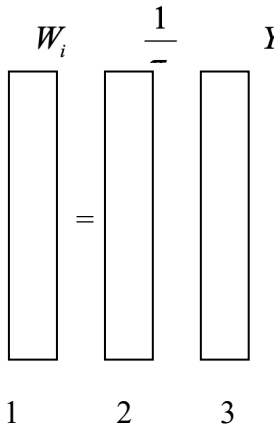
## 8.2 La varianza dell'Horvitz-Thompson tramite le probabilità d'inclusione

Come sappiamo una matrice non è altro che un rettangolo di numeri di *m*-righe ed *n*-colonne. Dagli argomenti precedenti, se ricordiamo, per calcolare gli intervalli di confidenza sul nostro stimatore ci serviva la varianza, sappiamo calcolare tutto tranne la

varianza. Per farlo, esprimeremo le formule in termini matriciali cercando di spiegare un concetto con un esempio. Iniziamo con lo stimatore di *Horvitz-Thompson*:

$$\hat{t}_{HT} = \sum_{i \subset s} W_i \cdot Y_i$$

Indichiamo $W_i \cdot Y_i$ con $\breve{Y}_i$ (dove il cappelletto si pronuncia "Tilde"), cioè il prodotto riportato all'universo della nostra variabile dove i valori sono già stati riportati al valore della popolazione. Supponiamo di spendere *€ 10.000* in ricerca e sviluppo e di rappresentare 10 aziende. Riportando al valore della popolazione avremo *€ 100.000*, ossia il valore è rilevato per quante unità esso rappresenta,




*1.Vettore dei pesi di riporto all'universo*
*2. Vettore dei pesi di riporto all'universo*
*3.Vettore della mia variabile osservata, rilevata*

**** E nota bene che tutti vettori sono da 1 a n****

Allora riscriviamo lo stimatore di Horvitz-Thompson come:

$$W^T \cdot Y = \hat{t}_{HT} = \sum_{i \subset s} W_i \cdot Y_i$$

Praticamente abbiamo fatto la matrice trasposta, abbiamo preso $W$ è l'abbiamo girato. Trasporre una matrice, significa semplicemente scambiare le righe con le colonne. In questo caso essendo una colonna sola avremo una riga sola e tante colonne quante sono

le righe che avevo prima. *(In Excel queste funzioni si chiamano MATR( ), in Excel esistono delle funzioni che non sono di singolo numero ma di un insieme di numeri, e queste ovviamente appartengono a questa categoria. Prima selezioniamo, anneriamo la zona di destinazione, poi scriviamo la funzione, e bisogna pigiare tre tasti contemporaneamente, il CONTROL, SHIFT, RETURN però da quel momento in poi la applica, è un blocco unico - cioè delle cellette che non possiamo modificare più di una alla volta, a meno che non togliamo la funzione. La stessa cosa se facciamo un prodotto di matrici o un'inversa )*

La varianza dello stimatore è pari alla doppia sommatoria per *K* ed *L* che sono due unità entrambe contenute nel campione e di $\Delta_{KL}$ e per $\breve{Y}_k \breve{Y}_L$ che è la coppia di unità, ovvero la *Y* già riportata all'universo, ovvero già moltiplicata per il *W* nella coppia di unità, ovvero tutti prodotti delle *Y* tra tutte le unità una contro l'altra, ovvero abbiamo *1000* unità ed ognuna la dobbiamo moltiplicare per le altre *999*.

$$V\left(\hat{t}_{HT}\right) = \sum_{k} \sum_{i \subset s} \Delta_{KL} \cdot \breve{Y}_K \cdot \breve{Y}_L$$

Detta in breve deve essere pesata per $\Delta_{KL}$ che non è altro che la probabilità di inclusione doppia meno il prodotto delle probabilità di inclusione del primo ordine; di fatto questo $\Delta_{KL}$ è quanto il nostro sistema di selezione si allontana dall'indipendenza, perché la probabilità di inclusione doppia nel caso di indipendenza non è altro che il prodotto delle due singole, che nel caso di perfetta e completa indipendenza questo $\Delta_{KL}$ è uguale a zero. Però i campioni per come gli abbiamo estratti fino ad ora non sono indipendenti, perché sono senza reinserimento: ad esempio se l'impresa non la consideriamo più alla seconda estrazione, non si tratta più del prodotto delle probabilità singole, quindi abbiamo modificato la probabilità delle altre; c'è una leggera forma di dipendenza, quindi la differenza $\Delta_{KL}$ non è zero. Comunque è la probabilità di inclusione doppia che ci interessa perché anche se togliamo il prodotto delle singole, quello che pesa di più o di meno sulla distribuzione è la probabilità di inclusione doppia, ovvero ogni coppia di prodotti pesa di più quanto più pesa la probabilità che entrambe siano nel campione,

$$\Delta_{KL} = \pi_{KL} - \pi_K \cdot \pi_L$$

Ora riscriviamolo in termini matriciali. Se chiamiamo $\Delta = A - (B \cdot C)$ matrice quadrata di *n* righe e *m* colonne, in cui abbiamo le probabilità di inclusione doppie meno i prodotti delle probabilità singole, possiamo semplificare in

$$\Delta = \Delta_{KL} = \pi_{KL} - \pi_K \cdot \pi_L$$

Allora

$$V\left(\hat{t}_{HT}\right) = \breve{Y}^T \cdot \Delta \breve{Y}$$

Questa è una semplicissima forma quadratica che fornisce tutti i cross-prodotti del vettore *Y* pesati per una matrice che ci pesa i cross-prodotti. Facciamo un esempio: abbiamo un vettore $\breve{Y}$, riprendendo i valori precedenti che sono (1,2,3) ed i pesi (10,10,10) ossia $W_i = \frac{N}{n}$, le probabilità d'inclusione singole quindi sono (0,1-0,1-0,1) perché $W_i = \frac{1}{\pi_i}$.

| $\breve{Y}$ | $\pi_i$ |
|---|---|
| 20 | 0,1 |
| 10 | 0,1 |
| 30 | 0,1 |

La $\pi_{ij}$ è una matrice *3x3*, la probabilità d'inclusione del secondo ordine ci dà la probabilità che entrambe stiano nel campione quindi sulla diagonale abbiamo tutti *1*, mentre fuori dalla diagonale tutti i valori sono uguali a *0,11*. Per semplificare i calcoli,

$$\pi_{ij} = \begin{pmatrix} 1 & 0{,}11 & 0{,}11 \\ 0{,}11 & 1 & 0{,}11 \\ 0{,}11 & 0{,}11 & 1 \end{pmatrix}$$

Quant'è la matrice $\Delta$? Sono le probabilità doppie meno il prodotto delle probabilità di inclusione del primo ordine. Quindi, se le probabilità di inclusione del primo ordine sono

$$\pi_i = \pi_j = \begin{pmatrix} 1 & 0{,}1 & 0{,}1 \\ 0{,}1 & 1 & 0{,}1 \\ 0{,}1 & 0{,}1 & 1 \end{pmatrix}$$

ed il prodotto delle probabilità del primo ordine è

$$(\pi_i \cdot \pi_j) = \begin{pmatrix} 1 & 0{,}01 & 0{,}01 \\ 0{,}01 & 1 & 0{,}01 \\ 0{,}01 & 0{,}01 & 1 \end{pmatrix}$$

allora possiamo fare la differenza (il delta $\Delta$) tra le probabilità di inclusione del secondo ordine ed il prodotto delle probabilità d'inclusione del primo ordine,

$$\Delta = \pi_{ij} - (\pi_i \cdot \pi_j) = \begin{pmatrix} 1 & 0{,}1 & 0{,}1 \\ 0{,}1 & 1 & 0{,}1 \\ 0{,}1 & 0{,}1 & 1 \end{pmatrix}$$

Adesso dobbiamo applicare questa formula $V\left(\hat{t}_{HT}\right)=\breve{Y}^{T}\cdot\Delta\breve{Y}$, dove $\breve{Y}^{T}=(20,10,30)$; cioè abbiamo trasposto il nostro vettore iniziale delle $Y$ riportate all'universo: da una colonna abbiamo una riga, e

$$\Delta=\begin{pmatrix}1 & 0,1 & 0,1\\ 0,1 & 1 & 0,1\\ 0,1 & 0,1 & 1\end{pmatrix}$$

è quello che abbiamo ottenuto precedentemente. Facendo il prodotto tra la trasposta delle $\breve{Y}^{T}$ e $\Delta$ non trasposto otteniamo $(24,15,33)$ che è ancora un vettore. Successivamente dobbiamo moltiplicare per $\breve{Y}$ che è $(20,10,30)$. Fare un prodotto tra matrici significa prendere ogni riga, moltiplicarla per la prima colonna e sommare; poi moltiplicare per la seconda colonna e sommare etc. ed otteniamo tutti i singoli valori, poi prendiamo un'altra riga e facciamo lo stesso procedimento.

Ora ritornando all'esempio, il risultato, cioè la varianza dello stimatore (e non la stima della varianza che nessuno saprà mai), è 1620. La stima della varianza la si può solo stimare in base al campione che abbiamo osservato, infatti il valore vero non si può conoscere. La stima della varianza non si scrive più così

$$V\left(\hat{t}_{HT}\right)=\sum_{k}\sum_{i\subset s}\Delta_{KL}\cdot\breve{Y}_{K}\cdot\breve{Y}_{L}$$

ma metteremo la tilde sopra il delta e quindi non stimeremo più la varianza dello stimatore:

$$V\left(\hat{t}_{HT}\right)=\sum_{k}\sum_{i\subset s}\breve{\Delta}_{KL}\cdot\breve{Y}_{K}\cdot\breve{Y}_{L}$$

La differenza è qui

$$\breve{\Delta}_{KL}=\left(\frac{\Delta_{KL}}{\pi_{KL}}\right)$$

ovvero è uguale al delta precedente solo che è diviso per la probabilità d'inclusione doppia, ovvero per fare la stima di una varianza tutte le coppie di prodotti vanno riportate all'universo e per portare all'universo tutte le coppie si divide per la probabilità di inclusione della coppia (cioè la probabilità di inclusione doppia). Quindi quello che cambia tra la verità e la nostra stima è che dobbiamo riportare il campione all'universo. In questo caso per poter calcolare la varianza dobbiamo riportare delle coppie di valori e non dei singoli valori, e per riportare all'universo delle coppie di valori dobbiamo dividere per la probabilità di inclusione delle coppie. Riscritto in termini matriciali è identico a quello di prima; soltanto che la matrice che sta in mezzo, che dobbiamo pesare, diventa

$$V\left(\hat{t}_{HT}\right) = \breve{Y}^T \cdot \breve{\Delta} \breve{Y} .$$

Alla fine comunque è molto semplice. Nella sostanza è il prodotto di tre cose: pesi di riporto all'universo, probabilità di inclusione doppia trasformata ed il dato rilevato trasposto. La vera questione non è l'aspetto formale ma l'aspetto pratico di questa formula: è applicabile nella realtà? Ovviamente no. Supponiamo di avere il campione sui bilanci di famiglia di *3000* unità: la nostra matrice delta sarà uguale a *9000000x9000000*, ovvero la dimensione di calcoli di questa formula cresce secondo un ordine di grandezza di $n^2$; quindi questa formula per quanto bella e semplice è molto raro che venga applicata a meno che non si abbia tanta pazienza. Nella realtà invece come facciamo? Nella realtà quello che viene fatto è una cosa semplicissima. Poiché conosciamo già le probabilità d'inclusione in termini analitici come ad esempio nel campione casuale semplice

$$\pi_i = \frac{n}{N} \qquad i = 1, 2, ....., N$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \qquad i = 1, 2, ....., N \qquad j = 1, 2, ....., N$$

le inseriamo nella matrice ed alla fine otteniamo una formula, le probabilità d'inclusione saranno uguali su tutti i valori della matrice, quindi non abbiamo motivo di ripetere milioni di calcoli su una cosa che è costante. Quindi nella realtà, poiché per ogni piano di campionamento le probabilità di inclusione del primo ordine e del secondo ordine le sappiamo indicare per esteso, nella formula precedente sostituiremo i valori della probabilità del primo e del secondo ordine secondo il nostro specifico piano di campionamento - magari si allunga la formula però diminuiscono i calcoli. Quindi la formula precedente così come è stata espressa vale per tutti i piani di campionamento ma, sostituendo i valori, nel campione casuale semplice assume una forma, nel campione stratificato ne assume un'altra, nel campionamento a cluster un'altra, a più stadi un'altra, etc. Quindi se ci chiedono quant'è la varianza del campione stratificato dobbiamo costruire la formula per il campione stratificato - quella che abbiamo visto prima è quella generale. Vuol dire che dobbiamo specificare tutti gli strati, tutte le numerosità delle popolazioni $N_1, N_2, ....., N_h, ...., N_H$, tutte le numerosità $n_1, n_2, ....., n_h, ...., n_H$, le rispettive probabilità di inclusione del primo del secondo ordine per ogni strato, etc. Questo è un procedimento che si dipana in un "formulone" in cui è più semplice che eseguire tutti i calcoli. Successivamente vedremo come questa formula si specifica nei diversi piani di campionamento. Ad esempio, il piano di campionamento che complica di più di tutti questa formula è il campionamento a più stadi poiché abbiamo estrazioni di estrazioni e la formula mano a mano si allunga.

# 9. La numerosità campionaria

## 9.1 La numerosità nel campione casuale semplice

Ora, cosa ci facciamo con la varianza che abbiamo trovato? Sicuramente calcoliamo gli intervalli di confidenza. Bisogna tenere conto del fatto che gli errori si dividono in due parti, ovvero errori campionari ed errori non campionari. La varianza è un errore campionario. Consideriamo l'estrazione di un campione da 4 milioni di imprese italiane, perché ne prendiamo 30.000 invece di 50? La varianza certamente cambia al cambiare della numerosità campionaria e la varianza campionaria dell'*Horvitz-Thompson* è una funzione monotona decrescente della numerosità campionaria, ossia al crescere delle

osservazioni diminuisce la varianza; al limite massimo se guardiamo tutte le unità abbiamo varianza zero. Ora se questa funzione è monotona decrescente, per ogni *n* abbiamo la nostra varianza; se è monotona decrescente è anche invertibile, ossia - data una varianza prefissata, un errore campionario prefissato - possiamo stabilire quante unità osservare,

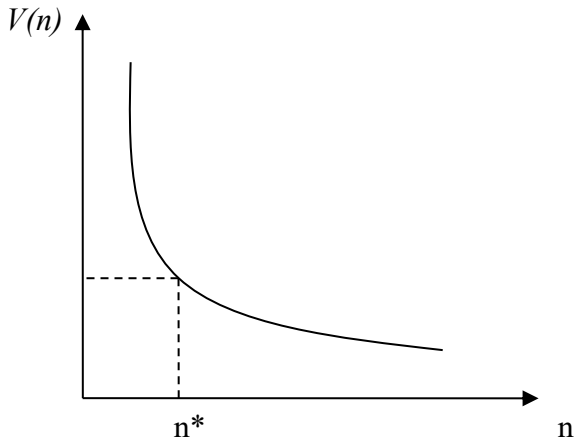



*Fig. 9.1 – l'inversione della varianza*

Oltre alla sua funzione principale che è dare la stima dell'errore campionario, dell'incertezza campionaria; la varianza campionaria ci permette anche di determinare una numerosità. Ora, quella curva è uguale per tutti i piani di campionamento? No, a livello logico il campione casuale semplice è il disegno campionario più banale che si possa fare perché si presuppone che funzioni meglio, nel senso che a parità di unità osservate qualsiasi altra cosa ha varianza più bassa. Quella curva nel caso del campione casuale semplice è semplice, invece per piani di campionamento più complessi diventa complicata e dà una formulazione che è difficile da invertire. Quindi noi calcoliamo quelle numerosità campionarie sempre come nel campione casuale semplice: al massimo può accadere che il piano di campionamento che utilizziamo avrà varianza più bassa. Questo lo facciamo per ogni piano di campionamento tranne che per lo stratificato. Infatti il campione stratificato è più particolare in quanto in ogni pezzo si ripete il problema di fare un piano di campionamento e, a prescindere dallo spezzettamento in strati, di solito si utilizza lo stesso procedimento del campione casuale semplice. Vediamo ora come si definisce questa numerosità campionaria,

$$n = \frac{N \cdot S_y^2}{c^2 \cdot t^2 + N \cdot S_y^2}$$

Nota bene che $N \cdot S_y^2$ ci dice che dobbiamo moltiplicare la nostra varianza per tutte le unità perché stiamo facendo una stima sul totale. Questa che abbiamo scritto didatticamente è molto imprecisa. In ogni caso questa formula è un'inversione della varianza del campione casuale semplice e indica la numerosità di un campione casuale semplice, dove $N$ è la numerosità della popolazione, $S_y^2$ deriva dagli anni precedenti oppure si calcola con le probabilità d'inclusione (anche se sarà una stima, bisogna dividere il delta della varianza dello stimatore per le probabilità d'inclusione doppie per ottenere la varianza della variabile *Y* di cui stiamo facendo una stima), $t^2$ è il totale al quadrato della popolazione della variabile oggetto di stima, $c^2$ è il livello di precisione espresso in termini di coefficiente di variazione,

$$CV = \frac{\sqrt{V\left(\hat{t}_{HT}\right)}}{\hat{t}_{HT}}$$

ovvero varianza dello stimatore di Horwitz-Thompson diviso lo stimatore di Horwitz-Thompson: indichiamo tutto in termini di coefficiente di variazione perché esprimere un vincolo in termini di varianza non è logico; ci risulta più facile dire, ad esempio, che vogliamo che la varianza sia intorno all'1%-3% del valore stimato ovvero preferiamo fissare dei limiti in termini di percentuale. Ad esempio supponiamo che vogliamo stimare la spesa in ricerca e sviluppo su una popolazione. Quante imprese dobbiamo prendere? Cosa c'è che non va? Praticamente dobbiamo calcolare una numerosità che ci dà una varianza e per calcolare la numerosità abbiamo bisogno della varianza.

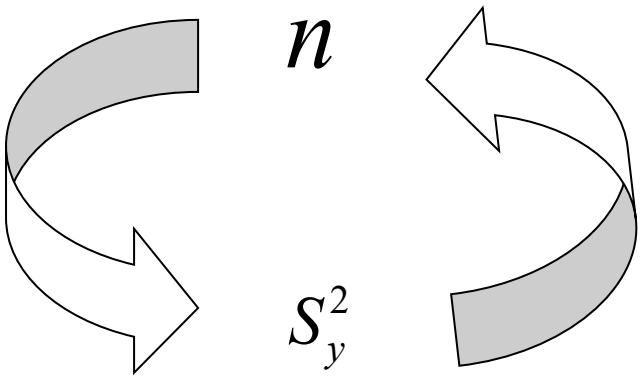



*Fig. 9.1 – Ciclicità sulla ricerca della numerosità*

Ma allora dove la prendiamo $S_y^2$? Semplicemente prendiamo quella dell'anno precedentemente. E se nessuno ha mai indagato quel fenomeno e non c'è nemmeno un fenomeno simile che ci può aiutare - come facciamo? Nel caso del questionario, parlando di campione pilota, abbiamo visto che potremmo utilizzare la stima della varianza su questo campione. Questa formula ha alcune caratteristiche interessanti: ad esempio, abbassare l'errore richiesto e dimezzarlo dal 10% al 5%, richiede un aumento di numerosità campionaria di gran lunga inferiore di quanto viene richiesto nel passare da 3% al 2%; in quest'ultimo caso la numerosità cresce in modo esponenziale fino al limite massimo. Questa formula non prevede blocchi su *N* ovvero, quando vogliamo un errore troppo piccolo, può dare degli *n* superiori a *N* sul campione casuale semplice e per altri tipi di campionamento. Se vogliamo un errore dell'*1%* ed abbiamo già messo la varianza dell'anno precedente, è possibile avere il coefficiente di variazione pari al *3%*? Sì, ci possono essere stati dei cambiamenti nel fenomeno studiato. Altro *worming* è che progettare il campione sulla base di un anno anomalo può dare dei dati; quando parlavamo degli *exit-pools* facevamo la stratificazione sulla base dei risultati precedenti - dov'è che tutto ci va a pallino? A cosa dobbiamo stare attenti perché nessuno ci rimproveri di aver sbagliato le previsioni? Ad esempio fare due indagini in due anni diversi in cui si è passati dal maggioritario al proporzionale utilizzando la varianza dell'anno precedente per fare una stima dell'anno in corso è assolutamente sbagliato. Con il sistema proporzionale le proiezioni funzionavano benissimo perché una sconfitta o un crollo di un partito significava perdere lo *0,05%*, da quando hanno messo il

maggioritario è stato creato un partito che non c'era prima ed è stato tutto stravolto, non abbiamo informazioni riguardo al nuovo fenomeno.

L'altro difetto è che quando facciamo un questionario non chiediamo una variabile sola, mentre questa formula è progettata per una sola variabile e quindi per ogni variabile avremo una numerosità diversa - quale prendiamo allora, quale numerosità? Prendiamo la numerosità più grande perché contiene tutte le altre. In ogni caso da un punto di vista applicativo dobbiamo stare attenti che non sia eccessivamente più grande delle altre. Rischiamo inoltre di porre un vincolo su variabili che non sono di grande interesse ai fini dell'indagine e magari costringiamo tutte le altre variabili ad avere una grande numerosità a causa di una variabile che non ci interessa. Ad esempio nelle indagini sulle imprese, tra le voci del bilancio - la *variazione delle scorte* - che è la classica voce residuale: quando al commercialista non tornano le altre voci, la differenza è inserita nelle scorte e quindi è una variabile non significativa.

Altro motivo per cui la varianza della voce sulla variazione delle scorte esplode è perché molto spesso non si tiene la contabilità di magazzino e quindi si fa tutto a fine anno: magari è zero per *364* giorni all'anno e viene scaricata magari per *100.000* a fine anno. Inoltre, alcune delle variabili delle imprese sono tipiche di alcuni settori e non di tutti i settori di attività economica: per esempio, nel settore dei servizi, la variazione della scorta è zero, quindi stiamo attenti dove prendiamo il *n-MAX*, ovvero la funzione di massimo è fortemente influenzata dai dati anomali.

## 9.2 La numerosità nel campione stratificato

Ora veniamo allo stratificato. Cosa succede quando andiamo a suddividere la popolazione? In questo caso, bisogna indicare anche dove andiamo a prendere le numerosità.

Ad esempio, fra *1000* imprese di quante in agricoltura, nell'industria, nei servizi? Oltre alla numerosità ci serve anche la posizione del campione. Un'idea potrebbe essere quella di fare lo stesso procedimento del campione casuale semplice per ogni strato? No, perché nella precedente formula della numerosità manca il vincolo $c^2$; non interessa fissarlo per i maschi e per le femmine ma interessa fissarlo per il totale. Quindi la cosa che cambia è che noi vogliamo sì sapere le numerosità per ogni pezzo ma stando ad un

vincolo che riguarda solo l'aggregato perché ci interessa solo che la varianza del campione sia di una certa entità, di quelle determinata entità. Si può ragionare anche al contrario: facciamo il calcolo come se non fosse stratificato e successivamente la numerosità totale la proporzioniamo all'interno degli strati sulla base di un qualcosa. Rovesciando il ragionamento, accade che la varianza risulta un po' più piccola per aver stratificato. Cambierà la varianza della formula ma il problema reale è come distribuire la numerosità del campione nei vari pezzi.

Allora, se abbiamo *10.000* imprese, come decidere quante prenderne in un settore o in un altro? Si prendono in modo proporzionale a quanto pesa ogni strato sulla popolazione ovvero proporzionale agli $N_h$ , ossia a quante unità della popolazione stanno nello strato. Questo si chiama *campione stratificato proporzionale* ed equivale a dire che il tasso di campionamento per una percentuale di unità è uguale per ogni strato, ovvero

$$\frac{n_h}{N_h} = \frac{n}{N} = W_h$$

e si usa solo ed esclusivamente quando non siamo in grado di calcolare le varianza all'interno degli strati. Se invece abbiamo le varianze per ogni strato, la numerosità sarà proporzionale alla varianza dello strato; se sappiamo quant'è la varianza calcoleremo la numerosità in proporzione della varianza dello strato; la formula è un po' più complicata e si chiama *allocazione ottima di Neyman*. Questa è molto simile alla precedente ma qui sottintendiamo che ci sia una sola variabile $J$ , quindi

$$a_h = \frac{N_h \cdot S_h^2}{c^2 \cdot t_h^{\,2} + \sum N_h \cdot S_h^2}$$

La calcoliamo per ogni strato, come se fosse una popolazione a se stante. Dobbiamo calcolare la quantità precedentemente scritta per ogni strato ed è simile alla formula del casuale semplice. Infatti si basa sull'idea che, facendo un campione stratificato e poi un campione casuale semplice per ogni strato, semplicemente si ha in aggiunta l'elemento

$\sum N_h \cdot S_h^2$ al posto della varianza della popolazione. Questo elemento è la varianza dello stratificato, cioè esso somma le varianze dei pezzi anziché fare la varianza su tutto; le numerosità finali sono proporzionali a queste quantità,

$$n_h = \sqrt{a_h} \sum_{1}^{H} {}_h a_h$$

perché in definitiva noi facciamo il calcolo per ogni singolo strato e poi riproporzioniamo sulla base delle numerosità ottenute per ciascuno strato. Ovvero per ottenere $c$ facciamo il calcolo a livello totale e poi lo spalmiamo sui diversi strati come se ciascuno fosse stato una popolazione a sé stante. Ed $a_h$ è solamente un calcolo intermedio, il primo calcolo serve solo per fare il secondo. Sulle varianze non c'è niente da dire se non le stesse cose dette sul campione casuale semplice, l'unica cosa da aggiungere è che finora abbiamo trattato una sola variabile. E se abbiamo più variabili cosa succede? Non abbiamo più *10* numerosità ma ogni variabile ci dice una numerosità diversa per ogni strato, quindi per ciascuno strato prendiamo la numerosità che ci viene data dalla variabile che ha la numerosità più grande. Prima prendevamo il più grande perché era quello che rispettava tutti i vincoli. Qui, per fare in modo di essere più grande in uno strato e rispettare un vincolo solo sul totale, potremmo prendere anche una media, e come una bilancia, abbassiamo un po' qua un po' là. La media comunque ci può far rispettare $c$ a livello totale: bisogna trovare questi piatti della bilancia e sistemarli in modo da far rispettare da una parte e da un'altra tutte le variabili e calcolare gli $n_h$ in modo che siano un po' meno del massimo. Qui però la soluzione a questo sistema delle bilance c'è ma non è analitica. La soluzione è numerica per approssimazione perché - tranne che in casi molto semplici come per lo stimatore di *Horvitz-Thompson* o il campione casuale semplice - in casi più complessi è difficile trovare una formula esplicita che ci dia il risultato: dobbiamo ricorrere a due sistemi che la calcolano numericamente. Si tratta dell'analisi numerica: una materia che si occupa di come calcolare numericamente delle cose tramite computer senza conoscerne la formula esplicita, come ad esempio il calcolo di un integrale tramite il metodo Montecarlo oppure il metodo Newton, il metodo della secante, etc. Questi sono metodi che partono dal valore precedentemente calcolato e continuano ad iterare fino a fermarsi ad esempio

ad un certo numero con una certa virgola. Questo ad esempio serve nell'allocazione dello stratificato a più variabili che si traduce in un problema di minimo: noi vogliamo il minimo numero di unità campionarie, vogliamo spendere di meno in modo che vengano rispettati tutti i $c$ su tutte le variabili.
Quindi non c'è una soluzione esplicita, dobbiamo cercare la soluzione in questo modo. Ma questo succedeva anche quando parlavamo di pesi di riporto alla popolazione: potremmo voler trovare dei pesi di riporto all'universo che rispettino tutti i vincoli della popolazione - totale addetti, totale fatturato, etc. - e che si discostino il meno possibile. Questo concetto (*il minimo condizionato a certi vincoli*) lo useremo spesso; non c'è una formula ma una funzione che ci dice che se partiamo da un certo punto e calcoliamo più volte arriviamo alla soluzione, ovvero ci verrà data una funzione che si mostrerà convergere ad un determinato valore.

### 9.2.1 L'allocazione ottima di Neyman

Cominciamo a risolvere il problema delle numerosità del campione quando abbiamo più vincoli, ovvero se adottiamo un campione stratificato. Per prima cosa ricostruiamo il coefficiente che abbiamo usato per il *campione di Neyman* per ciascuna variabile in ogni strato, ogni strato *h*, calcoliamo la varianza nello strato *h* della variabile *j,* l'errore sulla variabile *j* ovvero il vincolo che noi stessi abbiamo imposto e facciamo la sommatoria rispetto ad *h* sul numero degli strati della varianza. L'unica differenza che c'è rispetto al *campione di Neyman* è che qui abbiamo un indice *j* in più, ossia abbiamo più variabili, e quindi più vincoli, più numerosità, etc. Inoltre come abbiamo già detto c'è la possibilità di prendere la numerosità massima per ogni strato nel caso in cui abbiamo più variabili, ma è un procedimento che ci dà delle numerosità troppo elevate, per cui ora cerchiamo di individuare un metodo più preciso che tenga conto di tutto. Queste quantità $a_{ij}$ non entrano nel calcolo iterativo perché dipendono da numerosità e varianze della popolazione che sono ferme, sono sempre uguali.

Calcoliamo prima per ogni strato e per ogni variabile

$$a_{ij} = \frac{N_h^2 \cdot S_{hj}^2}{c_j^2 \cdot Y_j^{\,2} + \sum_{1}^{H}{}_h N_h \cdot S_{hj}^2}$$

ed otteniamo la matrice

$$\begin{pmatrix} a_{11} & \dots & a_{1J} & a_1 \\ \vdots & a_{hj} & \vdots & \vdots \\ a_{H1} & \dots & a_{HJ} & a_h \\ a_1 & \dots & a_J & a \end{pmatrix}$$

dove

$$A = \begin{bmatrix} a_{11} & \dots & a_{1J} \\ \vdots & a_{hj} & \vdots \\ a_{H1} & \dots & a_{HJ} \end{bmatrix}$$

e dove $\overline{a}_j$ è il vettore di $a_{ij}$ per ogni variabile

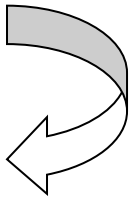
$$\overline{a}_j = \begin{pmatrix} a_1 \\ \vdots \\ a_H \end{pmatrix}$$

Adesso entriamo nel calcolo iterativo, ossia calcoliamo una quantità $x$ che all'iterazione *n* sarà funzione di un parametro calcolato al passo precedente che è uguale a :

*Condizioni iniziali*

$\alpha_j^0 = \frac{1}{j}$ Peso Iniziale (Media aritmetica )

$\left[ x_h\left(\alpha^{n-1}\right); x_h\left(\alpha^{n}\right)\right] < \varepsilon$ Quando ci fermiamo



$n_h = \dfrac{1}{x_h\left(\underline{\alpha}^{*}\right)}$ Numerosità Ottimale

Poiché abbiamo inserito una funzione della numerosità campionaria

$$x_h\left(\underline{\alpha}^{*}\right) = \frac{1}{n_h}$$

e potrebbe esserci anche una funzione che dipende da

$$x_h\left(\underline{\alpha}^{*}\right) = \frac{1}{n_h}$$

per determinare il costo dell'indagine ad esempio $g\left(x_h\right) = C_h$ a seconda della numerosità del campione dello strato. Naturalmente questo calcolo va effettuato per ogni strato,

$$x_h\left(\underline{\alpha}^{n-1}\right) = \frac{1}{\sqrt{\sum_{1}^{J}{}_j \alpha_j^{n-1} \cdot a_{hj}} \sum_{1}^{H}{}_h \sqrt{\sum_{1}^{J}{}_j \alpha_j^{n-1} \cdot a_{hj}}} \cdot \textit{ Nuovi Pesi}$$

dove

$$\sqrt{\sum_{1}^{J}{}_j \alpha_j^{n-1} \cdot a_{hj}}$$

rappresenta la nostra bilancia, ovvero moltiplichiamo i coefficienti $\alpha_j^{n-1}$ per le quantità $a_{hj}$ per ogni variabile nello strato di cui vogliamo calcolare la numerosità e ne

facciamo la sommatoria. Successivamente moltiplichiamo la bilancia che abbiamo calcolato per lo strato di nostro interesse per la sommatoria delle bilance

$$\sum_{1}^{H}{}_{h}\sqrt{\sum_{1}^{J}{}_{j}\alpha_j^{n-1}\cdot a_{hj}}$$

cioè per ogni strato. Ora, dopo aver fatto questo calcolo per ogni strato otteniamo un vettore,

$$\underline{x}\left(\underline{\alpha}^{n-1}\right)=\begin{pmatrix} x_1 \\ \vdots \\ x_h \\ \vdots \\ x_H \end{pmatrix}$$

dato dalle numerosità per ogni strato, vettore di *H* righe ed una colonna; lo inseriamo nella formula seguente e lo moltiplichiamo con il vettore trasposto della quantità $a_j^T$ della j-esima variabile ed otteniamo un nuovo peso,

$$\alpha_j^n=\frac{\alpha_j^{\alpha-1}\cdot\left[a_j^T\underline{x}\left(\underline{\alpha}^{n-1}\right)\right]^2}{\sum_{1}^{J}{}_{j}\alpha_j^{\alpha-1}\left[a_j^T\underline{x}\left(\underline{\alpha}^{n-1}\right)\right]^2}$$

sotto c'è la somma perché i pesi devono essere pari a *1* dove

$$\overline{a}_j=\begin{pmatrix} a_1 \\ \vdots \\ a_H \end{pmatrix}$$

è il vettore di $a_{ij}$ per ogni variabile. Alla fine se $J=1$ e quindi $\alpha_j^0=1$ ritorniamo a

$$n_h = \sqrt{a_h} \sum_{1}^{H} {}_h a_h$$

cioè *campione di Neyman*. Ora è da precisare che questo calcolo va fatto per ogni strato e se lo dobbiamo fare per 10 strati, sapendo che converge dopo sette iterazioni, più o meno dobbiamo fare 70 iterazioni. La formula di per se non è altro che il *campione di Neyman* perché in quella quantità, dove c'è alfa, $x_h\left(\alpha^{n-1}\right)$ è il reciproco della numerosità di *Neyman*, $\alpha$ è la nostra bilancia, ossia rappresenta come vengono pesate le variabili. *Il campione di Neyman* aveva una sola variabile quindi non c'era bisogno di ponderare, pesare le variabili, qui invece la numerosità campionaria la dobbiamo calcolare con la stessa formula, ma dobbiamo farne una media ponderata, con i pesi $\alpha$, ovvero non prendere il massimo ma prenderne una media, quindi il trucco del calcolo iterativo sta nel calcolare il vettore dei pesi $\alpha$, il resto è esattamente un *campione di Neyman*, né più né meno.

### 9.3 Disturbi sulla numerosità campionaria

a) Per praticità fissiamo un errore, ad esempio se ci viene una numerosità di *13,48* poiché le formule si trovano in un continuo e le numerosità no, il vincolo non può essere rispettato esattamente, se non con numerosità frazionarie - cosa si fa allora? Non si arrotonda mai per difetto poiché il 13 non può rispettare il nostro vincolo invece il *14* sì, quindi tutte le numerosità devono essere sempre approssimate per eccesso;

b) Una sola unità in uno strato ci sta bene? No, devono essere due o di più per poter calcolare la varianza.

c) Come altro accorgimento applicativo (quando ci occupiamo di un'indagine abbastanza conosciuta, di cui conosciamo i tassi di mancata risposta), se in uno strato sappiamo che il tasso di mancata risposta è del *50%,* naturalmente il nostro campione dovrà raddoppiare ovvero dividiamo per il tasso di mancata risposta, in modo da riproporzionare per la risposta attesa.

d) Esiste una pratica comune, un metodo che può essere considerato un aborto metodologico e che viene utilizzato in alternativa a quello che abbiamo appena descritto, ovvero dividere per il tasso di mancata risposta. Questa pratica comune

consiste nel sovra campionare, abbiamo in pratica quello che si definisce *campione suppletivo*, ovvero all'interno dello strato si fanno due campioni, uno è quello buono che ed un altro è suppletivo, si fanno in pratica due liste e se non risponde una unità su una lista se ne prende un'altra dalla seconda lista. Questo è un aborto metodologico perché dipende dal rilevatore, il quale scegliendo a sua discrezione dalla lista suppletiva rompe la casualità del campione. La numerosità del campione suppletivo dipende dal fenomeno, dal tasso di mancata risposta atteso. Se ci aspettiamo un tasso di mancata risposta dello *0,5* prenderemo un campione suppletivo con numerosità pari almeno alla metà del campione originale, e se non ci rispondono neanche le unità del campione suppletivo prendiamo le unità da un ulteriore campione suppletivo. Ma perché si usa tale metodologia anche se non è ragionevole? Perché in genere se noi facciamo un contratto con un'azienda per un'indagine che prevede una numerosità di *300.000* unità, 300.000 ne dobbiamo intervistare, per questo procediamo proprio come detto precedentemente, con il campione suppletivo. Questo è l'unico modo per rispettare la numerosità, si continua a pescare tra unità statistiche fin quando non si raggiunge la numerosità stabilita. L'unica soluzione sono è costituita da contratti particolari dove il compenso non dipende dalla numerosità campionaria. Ad esempio un'azienda dice di volere *300.000* unità, il rilevatore risponde: "te ne do *90* ma me ne paghi *150*, altrimenti non mi attivo perché non ci rientro nei costi".

e) Se in un'indagine su un campione casuale stratificato, dividendo in strati secondo una variabile ci accorgiamo che gli strati sono uniformi, e magari uno strato ha una sola unità, cosa facciamo – prendiamo quella sola unità? Attenzione: se vogliamo calcolare la varianza dobbiamo avere almeno due unità in uno strato, quindi per risolvere il problema collassiamo l'unità in un altro strato, ovvero uniamo quella singola unità ad un altro strato.

### 9.3.1 Campionamento per quote

Quello che è stato detto nel paragrafo precedente ci permette di aprire una parentesi su un tipo di campionamento che è una versione estremamente sporca del campione suppletivo, che non è un vero e proprio sistema di campionamento, ed è molto

utilizzato nelle società di analisi di mercato, che cercano di barattarlo per un campionamento stratificato, e si chiama *campionamento per quote.*

Funziona così, diamo delle classi come se fossero degli strati, poi per ciascuna classe diamo delle numerosità ma con la differenza che non diamo al rilevatore l'elenco estratto da noi, diciamo che servono *50* maschi e *50* femmine, ad esempio *10* individui dai *10* ai *20* anni, *10* individui dai *30* ai *40* anni, etc. Senza l'elenco di estrazione casuale, in realtà non è un campione casuale semplice. Tra i difetti abbiamo che:

1) Viene presentato come una rilevazione seria dicendo che comunque il rilevatore è soggetto a dei vincoli, a dei paletti, i limiti di età ad esempio. ma questo non è un vincolo niente perché potrebbero esserci altri possibili criteri di scelta ragionata da parte del rilevatore che lo statistico non conosce;
2) Questo metodo nasconde dei problemi: chi ha stabilito le numerosità non conosce le modalità di rilevazione, i tassi di mancata risposta e quindi la distorsione che ne deriva. Se avevamo selezionato *100* unità ed il rilevatore ne ha dovute intervistare *10.000*. Di solito questo metodo viene usato per le interviste faccia a faccia;
3) Chi utilizza questo metodo inoltre, ed è la cosa più bieca, applica le stesse formule che si usano per il campione casuale semplice;

## 10. Le varianze

### 10.1 Il campione casuale semplice

La varianza dello stimatore della variabile $Y$ è

$$\hat{\sigma}^2_{\hat{t}_{HT,SRS}} = N^2 \left[ \frac{(1-f)}{n} \right] S_Y^2$$

dove

$$f = \frac{n}{N}$$

Secondo l'inferenza che conosciamo da statistica di base la varianza sarebbe dovuta essere

$$\sigma^2_{\hat{t}_{HT,SRS}} = \frac{N^2}{n} S^2_Y$$

dove moltiplicare per $N$ un valore medio, ad esempio se volevamo il totale di una variabile dovevamo moltiplicare $N$ per il valore medio, per la varianza significa moltiplicare per $N^2$.

Qui $(1-f)$ viene definito *correttore da popolazioni finite*

$$(1-f) = \frac{(N-n)}{(N-1)}$$

ovvero ci dice di quanto va abbassata la varianza, in ogni caso molto simile perché i campioni non sono mai eccessivamente grandi, perché non stiamo campionando da popolazioni virtualmente infinite ma da popolazioni finite, quindi la varianza di tutti i possibili campioni è un po' più bassa. Se poi prendiamo tutte le unità la varianza va a zero. Per questo usiamo questa formula, perché è ad hoc per il campionamento da popolazioni finite, ossia $(1-f)$=0 ed $f=1$, questo succede nel caso di un censimento. Nel caso da popolazioni infinite invece, perché la varianza tendesse a zero, $N$ dovrebbe tendere all'infinito, quindi la varianza non è mai zero. Inoltre abbiamo che

$$s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}$$

rappresenta uno *stimatore corretto della varianza*, quella varianza che nel calcolare le numerosità non sapevamo dove prendere. Mentre l'espressione

$$f = \frac{n}{N}$$

sta ad indicare la *frazione di sondaggio*, ossia la percentuale di quante unità prendiamo nella popolazione.

## 10.2 Il campionamento stratificato

Lo stimatore di Horvitz-Thompson per il campione stratificato è

$$\hat{t}_{HT(STS)} = \sum_{1}^{H}{}_{h} \cdot \frac{N_h}{n_h} \cdot \sum_{1}^{n_h}{}_{i} y_i$$

Praticamente facciamo una somma di somme, ossia specificando il tipo di selezione campionaria non abbiamo più bisogno delle probabilità d'inclusione, praticamente sommiamo le risposte della variabile per ogni strato e poi sommiamo i totali di strato.

$$\hat{t}_{HT} = \sum_{i=1}^{n} W_i Y_i$$

La varianza viene invece calcolata all'interno di ciascun strato, che è molto simile a quella del campione casuale semplice,

$$\hat{\sigma}^2_{\hat{t}_{HT,STR}} = \sum_{1}^{H}{}_{h} N_h^2 \left[ \frac{(1-f_h)}{n_h} \right] S^2_{yh} .$$

Ma questo ce lo dovevamo aspettare perché, se stratificare significa dividere la popolazione in tanti piccoli pezzi, la varianza totale, l'incertezza totale, sarà la somma delle incertezze nei pezzi, ogni pezzo lo consideriamo come un campione casuale semplice, ci calcoliamo le varianze e poi le sommiamo. Per cui la varianza finale sarà una somma di varianze. E quindi sia la stima del totale che la sua varianza vengono fatte indipendentemente in ciascun pezzo. Sommando le varianze di ciascun pezzo otteniamo la stima finale.

Ma fare questo, cioè lo stratificato, invece di un campione casuale semplice su tutta la popolazione - è meglio o peggio in termini di variabilità ed efficienza della stima? Ossia se stratifichiamo otteniamo sempre una varianza più bassa? Da statistica di base sappiamo che la Variabilità o meglio la Devianza è uguale a

$$V_y = V_b + V_w$$

dove

$$V_b = \text{distanza tra le medie}$$

e da ciò sappiamo che se

$$\overline{y}_1 = \overline{y}_2 = ..... = \overline{y}_H = \overline{y}$$

ossia le medie condizionate alla variabile di stratificazione che chiamiamo $X$ $\left(x = 1, 2, ...., H\right)$ della nostra variabile $Y$ sono uguali alla media delle distribuzioni marginali della $Y$ totale, nel nostro caso

$$\overline{y} = \frac{\sum_{j=1}^{J} y_j \cdot N_j}{N}$$

allora c'è, in media, indipendenza della $Y$ dalla $X$ e quindi la devianza o varianza dipenderà esclusivamente dalla varianza dentro i gruppi

$$Dev\left(Y\right) = Dev\left(W\right) = V_w .$$

Ora la domanda che dobbiamo porci è - la variabile che stiamo analizzando ha più variabilità all'interno dei gruppi o tra un gruppo ed un altro? La risposta possiamo intuirla guardando la formula; quest'ultima per come l'abbiamo calcolata considera solo le varianze all'interno dei gruppi, lo stratificato della varianza totale utilizza solo quella all'interno dei gruppi, quindi solo una parte della varianza impatta sulla varianza della stima, a differenza del campione casuale semplice, quella dello stratificato è sistematicamente più bassa perché ne usa una sola di devianza, e quindi la varianza

totale sarà tanto più bassa quanto più bassa è l'eterogeneità all'interno del gruppo e quindi in gruppi omogenei. Il campione stratificato è tanto più efficiente quanto più c'è omogeneità all'interno dei gruppi o in altri termini eterogeneità tra i gruppi.
Tradotto in altri termini il coefficiente di correlazione

$$\eta_{y/x} = \frac{Dev(B)}{Dev(W)}$$

è normalizzato tra *0* ed *1*. Quindi se

$$\eta_{y/x} = 1$$

vuol dire che la mia stratificazione spiega pienamente la varianza della popolazione, se va a zero

$$\eta_{y/x} = 0$$

vuol dire che la divisione in gruppi è completamente correlata rispetto alla variabile e quindi la mia stratificazione non serve a nulla, ricordando che $Dev(B)$ è la devianza tra i gruppi.

### 10.3 Precisazione sulla varianza

Riassumendo: abbiamo visto tutte le fasi di un'indagine, dal disegno campionario fino alle numerosità ed ora alle varianze espresse in modo analitico. Ciò che potrebbe generare forse un po' di confusione è un incastro metodologico dovuto ad esigenze didattiche riguardo la ciclicità del calcolo tra le numerosità e la varianza. Abbiamo detto che le fasi sono

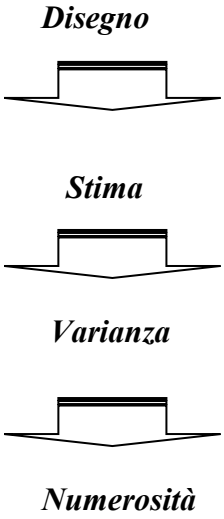



*Fig. 10.1 – Sequenza errata delle fasi di un'indagine*

Questa sequenza non è esatta, è assurda, è una sequenza esclusivamente didattica, perché in realtà il calcolo delle numerosità campionarie avviene nella fase di progettazione. Il punto è che didatticamente non possiamo rispettare quest'ordine, perché per poter calcolare le numerosità campionarie abbiamo prima bisogno delle varianze. Quindi il calcolo delle numerosità campionarie si trova didatticamente nel posto sbagliato rispetto a quanto succede nella realtà, chi progetta un'indagine per prima cosa calcola le numerosità campionarie, in pratica tra gli obiettivi che fissa c'è anche $c^2$, che se ricordiamo era il coefficiente di variazione, ovvero l'errore atteso in modo che posso stimare anche il budget dell'indagine, sulla base di quello calcoliamo le nostre numerosità campionarie. Didatticamente però non potevamo procedere in questo modo perché altrimenti non avremmo potuto calcolare le nostre numerosità o meglio far vedere come si calcolano.

Due sono le formule importanti per il calcolo delle numerosità: lo stimatore di Horvitz-Thompson e la relativa varianza, le altre possono sembrare più complesse ma lo sono solo da un punto di vista mnemonico, ricordiamo che

$$V\left(\hat{t}_{HT}\right) = \sum_{k} \sum_{i \subset s} \Delta_{KL} \cdot \breve{Y}_K \cdot \breve{Y}_L$$

quando togliamo i $\pi$ ed i $\Delta$ può essere più lunga ma certamente non richiederà di fare tutti i calcoli e si sviluppa con ordine di grandezza $n^2$, e quindi se al posto di questi ultimi ci mettiamo le probabilità d'inclusione costanti che conosciamo per i vari tipi di campionamento, la formula diventa più lunga ma non richiederà di fare tutti i calcoli. Quindi le formule che abbiamo enunciato riguardo al campione casuale semplice ed allo stratificato non richiederanno mai tutti i cross-prodotti tra gli elementi, non c'è mai una sommatoria doppia, ed anche se dobbiamo allungarle con gli $n$ ed $N$ non richiederanno mai $n^2$ calcoli; quindi in definitiva le nostro varianze in termini analitici diventano esasperatamente più semplici. Inoltre il trade-off tra velocità di calcolo e difficoltà di scrivere la formula non finisce qui. Calcolare le varianze ha tre livelli di complicazione, si possono presentare tre casi:

1) Sei in grado di scrivere la formula per esteso del tuo disegno campionario? Questo è il caso più semplice perché non richiede mai $n^2$ calcoli;

2) Non conosciamo la formula per esteso ma conosco le probabilità di inclusione, le conosciamo numericamente, in pratica abbiamo a disposizione una matrice *30000x 30000* ad esempio;

3) Non conosciamo la formula per esteso ne conosciamo le probabilità d'inclusione del primo e del secondo ordine, ovvero non sappiamo nulla, cosa facciamo? Lo vedremo a breve;

In pratica che differenza c'è in termini di calcolo tra la varianza dello stimatore $V\left(\hat{t}_{HT}\right)$ e la varianza che appare nelle nostre formule $S^2$? $S^2$ è la varianza che entra nella formula e la possiamo ottenere o dagli anni precedenti o tramite la stima con lo stimatore corretto $s^2$ per la varianza su un campione pilota, mentre $V\left(\hat{t}_{HT}\right)$ é la varianza che entra nel coefficiente di variazione e quindi rappresenta il vincolo per calcolare le numerosità, e può essere calcolata tramite i cross-prodotti di cui abbiamo parlato precedentemente; quindi potremmo dire che tutto il processo ha inizio da una

varianza, varianza calcolata sulla base o degli anni precedenti o tramite lo stimatore corretto su un campione pilota, quindi

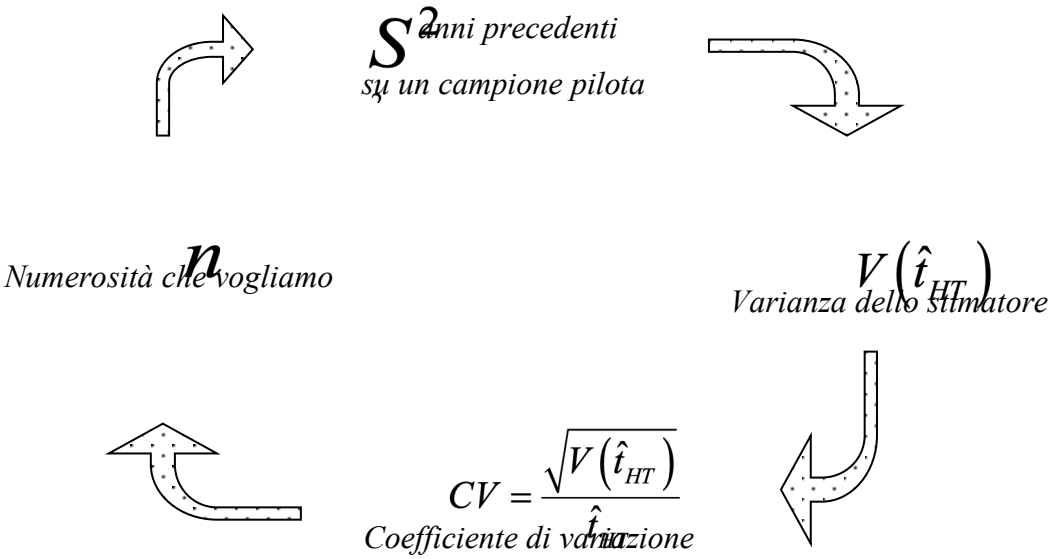
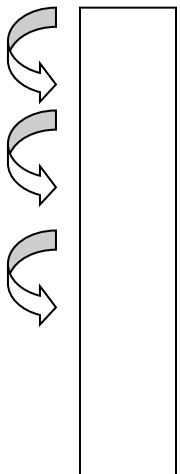



*Fig. 10.2 – Sequenza delle operazioni di calcolo*

Ora, perché esiste un secondo livello? Perché quelle formule di varianze anche per disegni campionari molto banali possono diventare molto complesse e bisogna essere molto bravi, invece calcolare le probabilità per via numerica è più semplice anche se più laborioso.

## 10.3 Il campionamento sistematico

Avevamo detto che nel campionamento sistematico riordiniamo la popolazione in base ad alcuni criteri, estraiamo un numero a caso e da qui in poi prendiamo le unità con passo fisso, sembra banale ma non lo è.



La formula della varianza per il nostro solito stimatore di Horvitz-Thompson nel campionamento sistematico è pari a:

$$\hat{\sigma}^2_{Sy}\left(\hat{t}_{HT}\right) = \frac{N^2 \cdot S_y}{n}\left[\left(1-f\right)+\left(n-1\right)\cdot\delta\right]$$

dove se togliamo

$$\left[\left(1-f\right)+\left(n-1\right)\cdot\delta\right]$$

ci rimane la varianza dello stimatore per il campione casuale semplice, inoltre

$$\left[\left(1-f\right)+\left(n-1\right)\cdot\delta\right]$$

dipende da $\delta$, e questo $\delta$ si chiama *coefficiente di correlazione intraclasse* e come tutti i coefficienti di correlazione che si rispettino è normalizzato tra -1 ed 1; ma cosa sta ad indicare? Siccome praticamente se ci pensiamo bene fare un sistematico equivale in qualche modo a dividere la popolazione in tante fette quante sono le unità campionarie e prenderne ognuna ad un passo fisso, è quasi una forma di stratificato se vogliamo, quel $\delta$ indica quanta dipendenza c'è tra i valori all'interno di ogni blocco, non negli strati perché non è una vera e propria stratificazione.

Spieghiamoci meglio: i tre valori tipici che può assumere questo coefficiente sono *1per* perfetta dipendenza, *0 per* indipendenza, *-1* per perfetta dipendenza negativa. Vuol dire che se abbiamo $\delta = 0$ i numeri all'interno di ogni blocco sono di tipo random, hanno un'alternanza casuale, se $\delta = 1$ significa che la varianza all'interno di ciascun blocco, gruppo, è superiore a quella che c'è tra un gruppo ed un altro, se $\delta = -1$ i dati all'interno di ciascun gruppo sono molto simili, ossia c'è omogeneità, ossia c'è poca varianza all'interno di ciascun gruppo; quindi se il coefficiente è $\delta = 0$ (i dati all'interno

dei gruppi sono casuali) vuol dire che mettendo zero abbiamo un campione casuale semplice nella formula, ovvero la nostra suddivisione in blocchi non ha cambiato niente rispetto ad una estrazione casuale su tutta la popolazione, tant'è che è il modo migliore per ordinare la popolazione in modo completamente random, ovvero ordiniamo la mia popolazione in modo random e poi facciamo un campionamento sistematico, quindi per definizione l'ordinamento è casuale e per $\delta = 0$ l'alternanza dei dati è casuale, uguale al campionamento casuale semplice.

L'altro caso invece in cui c'è molta variabilità, dove $\delta = 1$ vuol dire che usando il sistematico abbiamo commesso un errore perché prendendo "*uno ogni tot"* ci siamo persi proprio la variabilità all'interno del "*tot"*, quando c'è molta variabilità all'interno di ogni gruppo; facciamo un esempio classico, l'andamento della popolazione è di questo tipo,

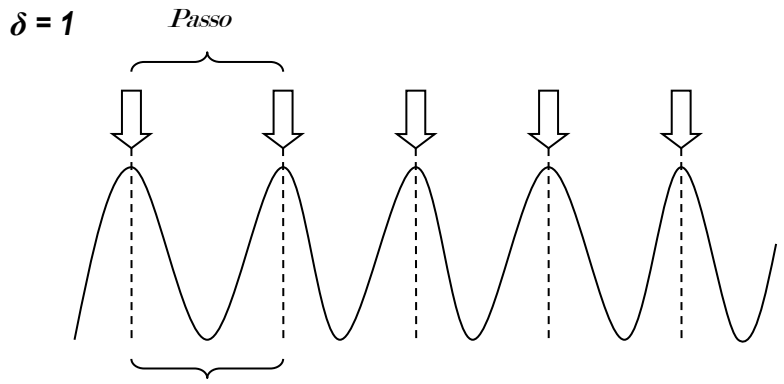
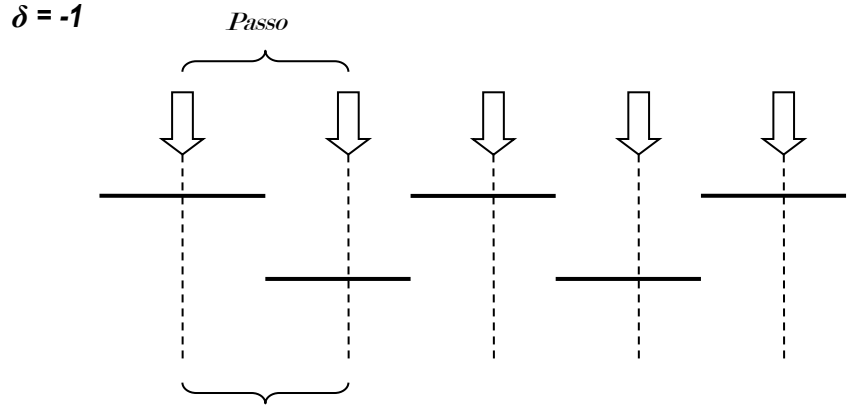



*Fig. 10.3 - Gruppo: valori tra un'unità e l'altra, prendendo le nostre unità con passo fisso ci siamo persi proprio la variabilità all'interno di questi gruppi*
$\delta = 1$

Se prendiamo una unità "*ogni tot"* perdiamo la variabilità all'interno del gruppo, ovvero abbiamo perso il fenomeno, poiché le unità che andiamo a prendere sono tutte molto simili tra loro per cui stanno falsando la variabilità che c'è nella popolazione, questo è il caso del delta $\delta = 1$, e come vediamo dalla formula la varianza aumenta tantissimo, di un ordine pari alla numerosità campionaria $(n-1)$; la situazione opposta si verifica nel caso in cui "*ogni tot"* le unità sono praticamente uguali, ossia tra quelle che prendo, ognuna di queste le rappresenta praticamente tutte ed il $\delta = -1$ riduce la

varianza moltissimo, se prima aumentava di un ordine $n$ adesso si riduce di un ordine $n$ nel caso contrario;

*Fig 10.4 - Gruppo: Valori tra un'unità e l'altra, prendendo le mie unità con passo fisso la varianza si riduce* $\delta = -1$




Ora poiché non abbiamo assolutamente controllo sul $\delta$, il campionamento sistematico viene utilizzato sempre in tre casi:

1) Caso in cui il campione non lo dobbiamo estrarre noi, caso in cui non abbiamo la lista, lo dobbiamo fare estrarre ad altri. Diciamo ad altri ad esempio di mischiare in modo casuale la popolazione, poi di prendere un'unità a caso e da lì in poi procedere con un passo fisso, generando così un $\delta = 0$ ed ottenere un campione casuale semplice;

2) Gli altri due casi in cui viene usato il sistematico sono casi di campionamento molto particolari, ovvero campionamenti che hanno come base di rilevazione due fattori che devono essere presi obbligatoriamente in modo sistematico, lo *spazio ed il tempo*; per quanto riguarda lo spazio se noi facciamo un campione di punti, supponiamo che abbiamo il territorio di una regione e ci dicono di rilevare qualcosa sul suolo, può essere qualsiasi cosa e ad esempio ci dicono – "fammi un campione di punti e dimmi cosa ci trovi" - (alberi, animali, etc.), per evitare che casualmente ci capitino degli agglomerati facciamo un campionamento sistematico, estraiamo un punto a caso fuori e poi seguiamo un passo fisso, tracciamo un grigliato regolare,

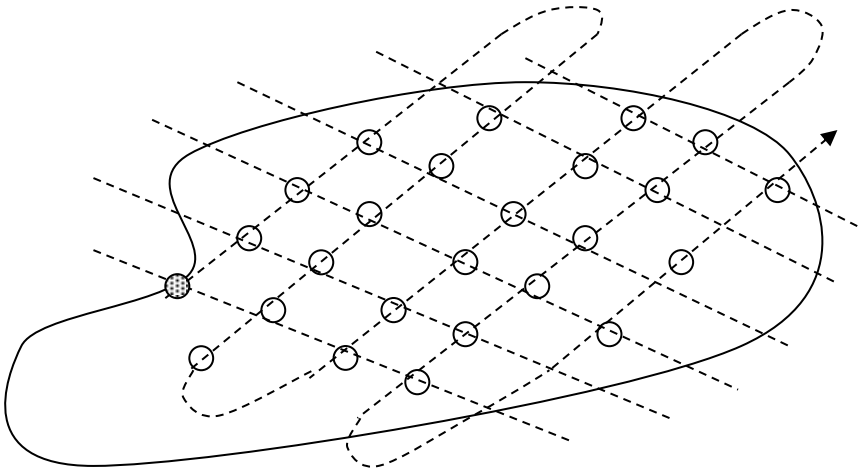


*Fig 10.5 - Campionamento sistematico*

e continuiamo, così distanziamo il più possibile i punti campione; questo vale per qualsiasi tipo di campionamento che riguarda lo spazio; la battaglia navale è qualcosa di simile; per quanto riguarda il tempo invece si tratta di campionamenti che vediamo fare praticamente tutti i giorni, è lo stesso principio che utilizza il lettore CD-ROM, detto anche, appunto, campionatore, legge le informazioni sui supporti "*ogni tot"*.

Quindi la varianza che ci è stata appena data è un falso totale, non la potremo mai stimare, per un motivo abbastanza semplice, quel $\delta$ non lo potremo mai stimare perché non abbiamo a disposizione tutti i dati sulla popolazione, a noi serve solo a livello concettuale e teorico, per capire quando il campionamento sistematico è meglio o peggio di un campione casuale semplice, da un punto di vista pratico quindi chi usa il

sistematico lo usa esclusivamente dopo aver fatto la seguente ipotesi $\delta = 0$. Si deve per forza ipotizzare di non poter fare confusione, altrimenti non occorre usarlo, ossia non è possibile stravolgere la struttura della popolazione, che tradotto in altri termini significa che i dati, come li abbiamo ordinati non hanno un ciclo che ha una base vicina a quella per cui andiamo a campionare i dati. Per ciclo intendiamo un andamento altalenante in cui per le varie fasi la frequenza non coincide con la frequenza del nostro campionamento, con la frequenza con cui andiamo a campionare; ovviamente, altra applicazione, il campionamento temporale esiste anche sulle borse mondiali, poiché il tasso di cambio varia continuamente e i database cui accediamo ci richiedono la frequenza con cui vogliamo ricevere i dati e non stiamo facendo altro che costruire un campione sistematico sui valori dei tassi di cambio.

## 10.4 Campionamento a grappoli

Questo tipo di campionamento ha varianza

$$\hat{\sigma}_G^2\left(\hat{t}_{HT}\right) = N_1^2 \frac{\left(1-f\right)}{n_1} S_{1y}^2$$

dove tutti gli 1 messi a pedice non riguardano le unità della popolazione né la numerosità del campione, ma il numero di grappoli in cui è suddivisa la popolazione. Abbiamo una popolazione, la suddividiamo in tanti gruppi ed estraiamo un campione di gruppi, ed all'interno di ogni gruppo osserviamo tutte le unità, quindi il l'*1* si riferisce al numero di gruppi mentre

$$f = \frac{n_1}{N_1}$$

é la frequenza di ogni gruppo campionato e $S_{1y}^2$ è la varianza delle medie all'interno di ciascun gruppo. Dobbiamo calcolare prima le medie all'interno di ciascun gruppo, quindi avremo due medie parziali, successivamente dobbiamo calcolare la media totale su tutti gli elementi estratti ed infine calcolare la varianza delle medie parziali dalla

media totale. Questo procedimento naturalmente ricorda il campione casuale semplice, l'unica differenza è che ora riguarda esclusivamente la parte gruppi, ma d'altra parte noi abbiamo fatto prima un campione casuale semplice di gruppi da $N_1$ gruppi, abbiamo preso $n_1$ gruppi e poi all'interno abbiamo fatto un censimento, quindi non c'è più variabilità dovuta alle unità perché all'interno di ciascun gruppo le abbiamo prese tutte, cioè non c'è variabilità nel campionamento delle unità; praticamente questo piano di campionamento non dipende dalla varianza all'interno dei gruppi ma da quella tra i gruppi, e quindi dalla varianza delle medie.

Ora il motivo per cui questo piano di campionamento è sempre totalmente inefficiente è che la varianza tra i gruppi di solito è infinitamente più alta di quella all'interno dei gruppi, questo piano di campionamento annulla la varianza all'interno dei gruppi anche se effettivamente non ce n'è bisogno perché è già abbastanza bassa; naturalmente assolutamente inefficiente a parità di numerosità campionaria, ovvero la somma di unità prese in ciascun gruppo, quindi per confrontarlo con un campione stratificato o con un campione casuale semplice bisognerebbe prendere la stessa numerosità. Tra le altre cose, e tra i tanti difetti di questo tipo di campionamento c'è il fatto che non abbiamo assolutamente controllo sulla numerosità campionaria a monte; ad esempio se dobbiamo condurre un'indagine sui comuni e dobbiamo estrarre dei comuni a caso, se siamo così sfortunati che ci capitano città come Milano, Roma, Napoli, tutte di grandi dimensioni il costo dell'indagine sarà elevatissimo, a quel punto potremmo fare un censimento, ma potrebbe capitare anche il contrario ed il campione non sarebbe molto rappresentativo; quindi la numerosità campionaria è a sua volta una variabile aleatoria tant'è che quando siamo costretti a fare piani di campionamento di questo genere di solito si ricorre ad un due stadi ovvero si cerca di selezionare il maggior numero di gruppi possibili ed invece di un censimento all'interno di ciascun gruppo facciamo dei campioni all'interno dei gruppi, in modo da tenere sotto controllo la numerosità campionaria, quindi si cerca di selezionare più unità di primo livello.

Quindi i vantaggi principali alla base del campionamento a più stadi sono prima di tutto il fatto che riduciamo un po' la varianza tra i gruppi perché ne prendiamo di più è poi che prendiamo un campione in ciascun grappolo e siamo noi a decidere la numerosità campionaria; naturalmente a meno di casi come quello degli aeroplani che abbiamo visto nella prima parte, il costo non cambia sia a due stadi sia ha grappoli.

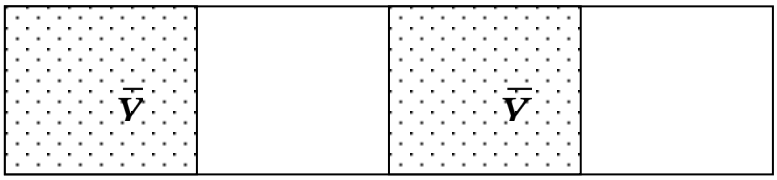



*Fig 10.6 - Abbiamo preso due gruppi su quattro ed in ogni gruppo facciamo un censimento*

## 10.5 Campionamento a più stadi

Prima ancora di arrivare a tutte le nostre sommatorie la cosa importante da tener presente è che quando lavoriamo a più stadi i nostri teoremi di composizione della varianza dicono che la varianza dello stimatore va spezzata nella la varianza dovuta alla prima estrazione ed in quella dovuta alla seconda estrazione; $\hat{V}_{PSU}$ viene detta *Primary Sampling Unit* mentre $\hat{V}_{SSU}$ viene detta S*econd Sampling Unit*, ovvero i gruppi e le unità, il gruppo gerarchico superiore ed il gruppo gerarchico inferiore. Il campione stratificato di fatto non è che un campione a due stadi che prende tutte le unità primarie e all'interno di ogni unità primaria fa un campione; e quindi nello stratificato $\hat{V}_{PSU} = 0$ perché prende tutti i gruppi e rimane solo la varianza dovuta al campionamento nei gruppi, mentre il campionamento a grappoli annulla $\hat{V}_{SSU} = 0$ perché prende tutte le unità a secondo stadio, quindi di fatto il campione stratificato ed il campione a grappoli sono due casi particolari di campione a due stadi, in cui uno censisce al primo stadio e campiona al secondo, l'altro invece campiona al primo stadio e censisce al secondo (quindi nota bene che quando viene inserita una casualità abbiamo bisogno di numeri random per poter ordinare le nostre unità, per lo stratificato al secondo stadio, per i grappoli al primo stadio, per il campionamento a più stadi in tutti gli stadi in cui dividiamo la selezione).

Ora il campionamento ad due stadi campiona al primo stadio e campiona al secondo, quindi il campione a due stadi include come casi particolari gli altri due

$$\hat{V}_{2S}\left(\hat{t}_{HT}\right)=\hat{V}_{PSU}+\hat{V}_{SSU}$$

quindi la varianza è

$$\hat{V}_{2S}\left(\hat{t}_{HT}\right)=N_1^2\frac{\left(1-f\right)}{n_1}\cdot S^2{}_{1y}+\frac{N_1}{n_1}\sum_{S_1}N_i^2\frac{\left(1-\breve{f}_i\right)}{n_i}\cdot S_i^2$$

dove

$$N_1^2\frac{\left(1-f\right)}{n_1}\cdot S_{1y}$$

è identica al campione a grappoli nel caso in cui si annulla la varianza al secondo termine, infatti avevamo visto che al secondo stadio del campione a grappoli veniva fatto un censimento,

$$\frac{N_1}{n_1}$$

rappresenta il numero di gruppi estratti dal primo stadio e

$$\sum_{S_1}N_i^2\frac{\left(1-\breve{f}_i\right)}{n_i}\cdot S_i^2$$

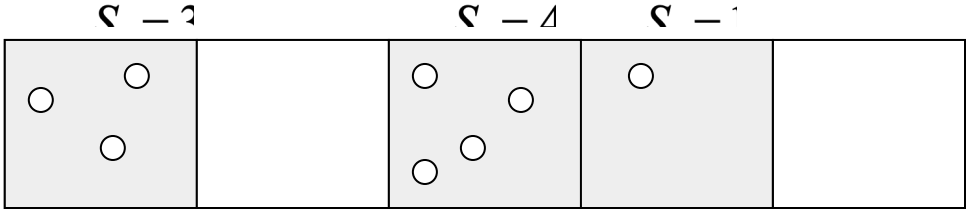
la varianza all'interno di ogni gruppo estratto , $N_i^2$ indica la numerosità dell'$i$-esimo gruppo che abbiamo estratto al primo stadio, quindi $i$ rappresenta l'indice di gruppo, e se abbiamo estratto *3* gruppi $i$ va da *1* a *3*. Quindi sommatoria sul numero di gruppi estratti sul campione di primo livello di

$$N_i^2\cdot\frac{\left(1-\breve{f}_i\right)}{n_i}\cdot S_i^2$$

quindi

$$\frac{N_1}{n_1}\sum_{S_1}N_i^2\cdot\frac{\left(1-\breve{f}_i\right)}{n_i}\cdot S_i^2$$

non è altro che la somma delle varianze all'interno dei tre gruppi.



$$S_1 = 3$$

*Fig. 10.7 – Il campionamento a più stadi*

*(Attenzione che in questo tipo di campionamento per calcolare la varianza dello stimatore devi calcolare prima medie parziali nei gruppi, media totale delle unità estratte che si trovano nei gruppi, e la varianza sulle unità estratte tramite lo stimatore corretto della varianza, poi svolgi il resto).*

Abbiamo quindi praticamente spiegato la formula di un campione casuale semplice, l'unica differenza è che era stata sommata per il numero di gruppi estratti, mentre sommare una varianza all'interno dei gruppi altro non è che la formula del campione stratificato, quindi la formula che abbiamo scritto altro non è che la formula del campione a grappoli più la formula del campione stratificato, nient'altro. Tutto questo in buona sostanza significa che aumentare il numero di stadi significa semplicemente calcolarsi più varianze diverse e poi sommarle e ciascuna di queste avrà un tipo di varianza a seconda di come è fatto il campione in ogni stadio, ad esempio se facciamo un censimento in uno stadio la varianza sarà naturalmente zero. Se noi avessimo utilizzato un campione *PPS* invece che il casuale semplice a livello di primo stadio la formula sarebbe rimasta uguale? Semplicemente al primo stadio la formula del campione casuale semplice sarebbe diventata quella del *PPS* .

Nel caso facessimo il *PPS* al primo stadio naturalmente, se ad esempio facciamo un campione sui comuni con probabilità proporzionali alla popolazione, in quel caso dobbiamo semplicemente sostituire al primo termine la formula del *PPS*, la varianza di primo livello, invece che con la formula del campione casuale semplice. Un aspetto molto importante è che *ogni stadio è indipendente da cosa facciamo nell'altro stadio*.

## 10.6 Il Sequential Poisson Sampling

La varianza del *PPS* come abbiamo detto non la facciamo, ma facciamo solo la versione semplificata del *PPS* che è il *Sequential Poisson Sampling*, poiché il *PPS* puro, vero e proprio, sia nel meccanismo di estrazione sia nella relativa varianza è molto più complesso, quello che abbiamo visto, come si estrae un *PPS* ne è una versione approssimata ed è molto più semplice da calcolare ed è semplice anche la sua varianza; otteniamo quindi

$$\hat{\sigma}^2_{SPS}\left(\hat{t}_{HT}\right)=\frac{1}{n}\cdot\left(\frac{N}{N-1}\right)\cdot\sum_{1}^{n}{}_{i}\left(1-n\cdot p_i\right)\left(\frac{y_i}{p_i}-\bar{y}\right)^2 p_i$$

Questa è la varianza del *Sequential Poisson Sampling* dove per $p_i$ intendiamo

$$p_i=\frac{x_i}{\sum_{n}^{1}x_i}$$

lo specifichiamo perché molte volte viene moltiplicata per $n$ o meno, qui intendiamo una cosa che somma uno su tutte le unità.

Ora nella formula la varianza del Sequential Poisson Sampling è

$$\hat{\sigma}^2_{SPS}\left(\hat{t}_{HT}\right)=\frac{1}{n}\cdot(.....)\cdot\sum_{1}^{n}{}_{i}\left(1-n\cdot p_i\right)\left(\frac{y_i}{p_i}-\hat{y}_{TOT}\right)^2 p_i$$

ma dobbiamo correggere un piccolo difetto, il piccolissimo difetto è che dall'espressione precedente ci aspettiamo che se le $p_i$ sono tutte uguali dovrebbe diventare identico alla varianza del campione casuale semplice, ovvero se estraiamo in modo proporzionale ad una variabile e le probabilità sono tutte uguali diventa un campione casuale semplice. Nel caso della varianza senza il fattore che abbiamo tolto di proposito non è più così; ovvero se le $p_i$ sono tutte uguali l'espressione senza il termine

$$\left(\frac{N}{N-1}\right)$$

diventa

$$\frac{N-1}{N} \text{ - n } volte$$

la varianza del campione casuale semplice, se le $p_i$ sono tutte uguali, ovvero

$$\frac{N-1}{N} \cdot \hat{\sigma}^2_{SRS}$$

dobbiamo aggiungere un *fattore correttivo* che fa sì che l'espressione senza il termine citato nel caso in cui le $p_i$ sono tutte uguali diventi la varianza del campione casuale semplice. Questa varianza si annulla in due casi - quali? Come vediamo all'interno della sommatoria c'è un prodotto di due differenze, quindi per annullarsi la varianza o si annulla una differenza o si annulla un'altra differenza, nel primo caso quando le $p_i = \frac{1}{n}$ vuol dire che la probabilità di selezione del primo ordine che come ricordiamo è pari ad $n \cdot p_i = \pi_i$ è uguale a $\pi_i = 1$, stiamo censendo la popolazione, prendiamo tutte le unità statistiche, quindi la varianza si annulla, in pratica non abbiamo preso le unità in modo casuale ma in modo deterministico quindi la varianza ovviamente sarà uguale a zero. Detta in breve questi di solito sono campioni che si fanno sulle imprese, grandi imprese come la Fiat, Pirelli, a probabilità uno, il contributo all'aleatorietà, all'incertezza, l'incertezza derivante da un'impresa che viene inserita per forza è zero, perché non c'è incertezza sul fatto che quella grande impresa sia inserita nel campione, quindi questa

varianza quando la probabilità di inclusione è uguale a uno, la componente, la somma relativa alle imprese con probabilità di inclusione uguale a uno è zero, quindi l'espressione precedente si annulla quando $\pi_i = 1$. Nel secondo caso invece, dove $y$ è un totale o una sua stima, è una costante, quest'espressione

$$\frac{y_i}{p_i} - \bar{y}$$

si annulla quindi quando anche quest'altra componente

$$\frac{y_i}{p_i}$$

è una costante, ovvero quando $y_i$ è proporzionale a $p_i$, ovvero $y_i \sim p_i$, la quale a sua volta è proporzionale ad $x_i$, in altri termini noi estraiamo ad esempio con probabilità proporzionale agli addetti se andiamo a fare una stima del fatturato, se questo è quasi proporzionale agli addetti è chiaro che la varianza verrà zero poiché il campione stesso sa già tutto degli addetti ed a maggior ragione sa tutto del fatturato, di conseguenza se una variabile è proporzionale all'altra non si aggiunge nessuna informazione, vuol dire che le costanti proporzionali non aggiungono nulla a livello informativo. Quindi questa varianza si annulla in questi due casi esaminati. Naturalmente tutto questo vuol dire che, comunque la proporzionalità esatta non avverrà mai, quando la *Y* sarà molto correlata con la *X*, sulla base della quale estraiamo, questa varianza si avvicinerà sempre di più a zero, ovvero quanto più il coefficiente di correlazione si avvicina ad uno, ossia quando il coefficiente di correlazione lineare tra le $y_i$ che stiamo stimando e le $x_i$ su cui abbiamo calcolato le

$$p_i = \frac{x_i}{\sum_n^1 x_i}$$

si avvicina ad uno, tanto più la varianza necessariamente si avvicinerà a zero, in altri termini,

$$\rho_{y_i x_i =} 1 \rightarrow \hat{\forall}_{SPS} = 0$$

ma chiaramente il coefficiente di correlazione non sarà mai uguale ad uno.

## 11. Altri metodi di stima

Finora abbiamo visto sia la varianza dell'Horwitz–Thompson generale espressa in forma matriciale e le formule analitiche ottenute conoscendo le probabilità d'inclusione costanti singole e doppie, lo abbiamo fatto per tutti i piani di campionamento. Chiaramente quando il piano di campionamento diventa sempre più complesso quindi, non siamo più in grado di usare le formule analitiche, ovvero le cose nel complicarsi vanno da piani di campionamento classici, semplici, noti, in cui sappiamo la formula a piani di campionamento più complessi in cui dovremmo calcolarci noi la formula e spesso non ci riusciamo, ma potremmo in questi casi calcolare numericamente le probabilità di inclusione del primo e del secondo ordine e quindi sappiamo calcolare la matrice dei $\Delta$ che abbiamo visto nelle pagine precedenti e successivamente applichiamo la formula dell'Horwitz–Thompson.

### 11.1 Metodi di ricampionamento

Ma esiste un terzo metodo. Una varianza ce la dobbiamo stimare per forza poiché dobbiamo conoscere il grado di precisione del nostro campione, e a tal proposito esistono dei metodi per stimare le varianze campionarie che valgono per qualsiasi piano di campionamento, non hanno bisogno di calcolare probabilità di inclusione né del primo né del secondo ordine, valgono per qualsiasi parametro si voglia stimare (finora abbiamo visto delle formule che si allungavano per la stima di una cosa fondamentalmente banale, figuriamoci cosa succede se noi su base campionaria volessimo stimare, studiare, un coefficiente di variazione o comunque un qualsiasi parametro che sia leggermente più complicato, come ad esempio il rischio di portafoglio delle nostre azioni). Quindi questi metodi servono non solo se il piano di campionamento diventa più complicato ma sono obbligatori, ovvero le varianze che

vediamo noi si conoscono per somme, se dobbiamo stimare su base campionaria anche il rapporto tra due variabili diventa molto ma molto più complicato, quindi non solo il piano di campionamento più complicato ci obbliga ad usare questi metodi, ma anche il parametro che andiamo a stimare.

Questi metodi comportano formule semplicissime ma non sono adatte per i calcoli da fare al computer; questi metodi si basano su un'idea fondamentalmente abbastanza banale.

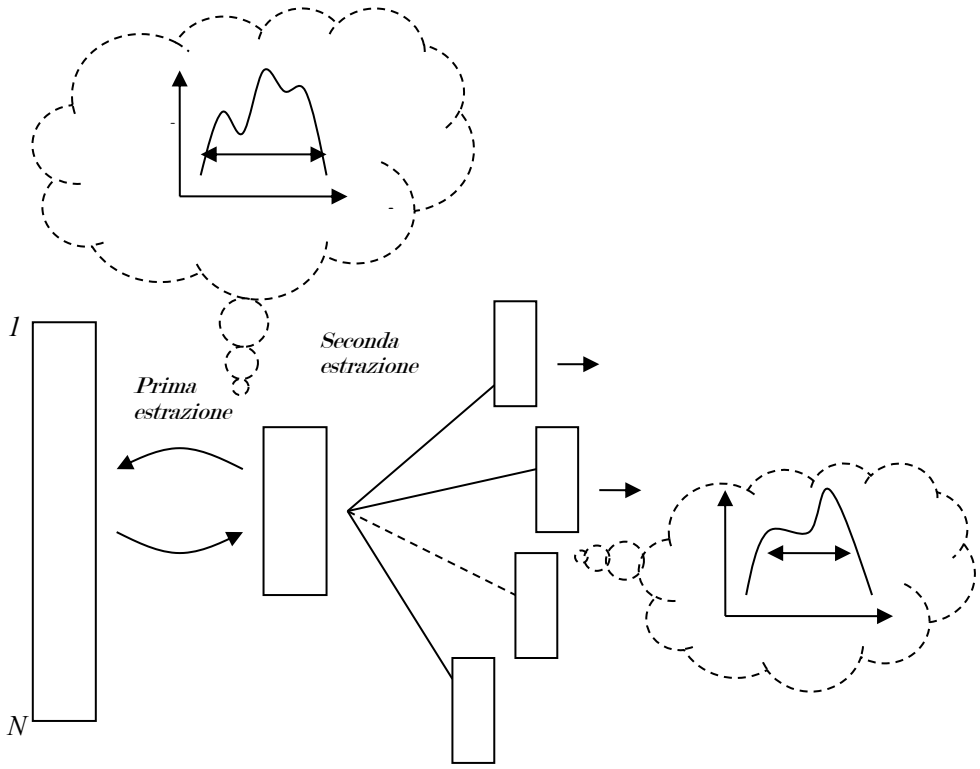



*Fig. 11.1 – Come operano questi metodi a livello logico*

L'idea è questa, noi partiamo da una popolazione di $N$ elementi, estraiamo delle unità per la prima volta, $n$ elementi. Tutto quello che abbiamo fatto finora è tentare di rovesciare l'estrazione per tornare indietro e fare stime sulla popolazione quindi lo stimatore ci permette di tornare sul valore della popolazione. Il calcolo della varianza campionaria di fatto è utile perché di tutti i campioni estraibili, in cui abbiamo ogni volta una stima diversa ed una distribuzione campionaria dello stimatore, quello che vogliamo fare noi è calcolare la variabilità della distribuzione, ovvero l'ampiezza.

L’idea di questi metodi è la seguente - perché non estraiamo,  cioè non ripetiamo l’operazione ad un secondo livello? Prendiamo il nostro campione, ne estraiamo un pezzo non una volta ma tante volte e per ciascuno di questi facciamo la nostra stima, e sulla base di tutte le varie stime ottenute ci costruiamo la distribuzione campionaria delle stime secondo la seconda estrazione. Praticamente stiamo estraendo un campione dal nostro campione, ad esempio, abbiamo un milione di imprese, ne abbiamo prese *1000* e le abbiamo rilevate - come si distribuisce lo stimatore di tutti i possibili campioni di *1000*?

Se non lo sappiamo fare con i metodi canonici facciamo così, dalle nostre *1000* ne prendiamo *100,* infinite volte, infiniti campioni, e guardiamo come si distribuisce lo stimatore nella seconda estrazione; ma la domanda è - questa distribuzione campionaria mi permette di calcolare indicatori sulla prima distribuzione campionaria? Ovvero nel passare dalla popolazione al campione abbiamo una distribuzione campionaria e nel passare dal campione al sotto-campione abbiamo un'altra distribuzione campionaria, e questa seconda distribuzione campionaria la calcoliamo senza fatica semplicemente estraendo e guardando, infinite volte – quindi la domanda è possiamo tornare indietro?

Sì, questa seconda estrazione, questa seconda distribuzione campionaria è un'ottima stima della prima distribuzione campionaria; ma perché questo modo di procedere è generale? Perché si può fare per qualsiasi piano di campionamento? Come detto prima la seconda estrazione è completamente indipendente dalla prima, possiamo farla a due strati, a due stadi, a grappoli, etc., e la seconda la facciamo in un altro modo, sono completamente indipendenti, non c'entrano niente l'una con l'altra. Quindi per quanto sia complicata la prima estrazione la seconda la facciamo come vogliamo e ci ricostruiamo la distribuzione campionaria.

Inoltre sul campione nella prima estrazione ci possiamo stimare tutto quello che vogliamo, coefficiente di correlazione, rischio di portafoglio, cose complicate come vogliamo, qualsiasi funzione complicata dei dati che estraiamo, e comunque delle stime ne abbiamo tante alla seconda estrazione e calcoliamo la distribuzione della prima estrazione, e non c'è nessuna ipotesi di *gaussiana* o di altro, questo metodo è indipendente da qualsiasi forma di ipotesi, né dalla selezione campionaria, né dalla funzione che è il nostro stimatore, né dalla distribuzione dei dati.

L’unica cosa è che bisogna stare attenti a questa seconda estrazione perché non vanno tutte bene, ce ne sono alcune che fanno in modo che poi la seconda distribuzione campionaria sia la stima della prima distribuzione campionaria, e noi ne vedremo solo due, le più utilizzate ma ce ne sono tantissime.

## 11.2 Il metodo BOOTSTRAP

Il termine “bootstrap” nasce dalla storia di Alice nel paese delle meraviglie che cercava di prendersi per gli stivali e tirarsi su, e questa operazione veniva definita con il termine inglese “bootstrap”, in effetti quello che stiamo facendo è più o meno questo.

Il primo criterio di selezione funziona più o meno in questo modo, dal nostro campione $n$ ne estraiamo un altro con la stessa numerosità ma *con reinserimento*, ovvero se abbiamo come al solito un'urna con le solite palline, invece di estrarre e buttare via la pallina la rimettiamo dentro, ne possiamo estrarre *1000*, *2000*, quante ne vogliamo, l’importante è rimetterle dentro.

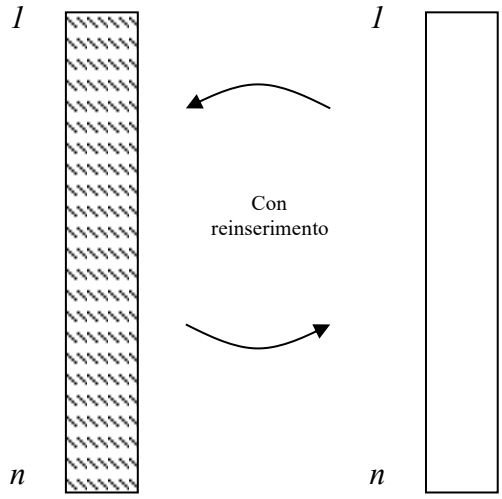



*Fig. 11.2 – Il metodo Bootstrap*

Questo significa che è un po' remota la possibilità che estrarre per la seconda volta lo stesso campione in cui le unità sono tutte diverse, è un po' raro, anzi la maggior parte delle volte avremo un campione in cui una unità riappare più di una volta, *1,2,3* volte,

quindi di campioni così, con la stessa dimensione di quello originale possiamo estrarne infiniti, diciamo $k$ per semplicità, e per ciascuno di questi campioni stimiamo un qualsiasi parametro, non necessariamente deve essere lo stimatore del totale ma qualsiasi altro stimatore,

$$\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, ....., \hat{\theta}_k .$$

E poi i cosa facciamo, chiamando con $\bar{\theta}$ la media aritmetica dei $k$ valori,

$$\bar{\theta} = \frac{\sum_{1}^{k} {}_i \hat{\theta}_i}{k}$$

potremo calcolarci la stima della varianza campionaria di questo parametro nel modo più ovvio:

$$\hat{V}_B\left(\hat{\theta}\right) = \frac{\sum\left(\hat{\theta} - \bar{\theta}\right)^2}{k}$$

dove la $B$ sta per " bootstrap ", la varianza di tutte le stime ottenute, e per quanto possa essere complicato il piano di campionamento alla prima estrazione, questa operazione la possiamo ripetere ogni volta.

### 11.3 Il metodo JACKNIFE

Letteralmente sta a significare "coltellaccio di jack", il tipico coltello che vediamo nei film di orror. Si chiama così perché il metodo di selezione funziona in questo modo: dal nostro campione ordiniamo in modo casuale supponiamo $n$ unità, inseriamo il numero random e ordiniamo, le mischiamo, e dividiamo l'ordinamento casuale in $k$ fette, proprio per questo il nome deriva dal coltellaccio di jack, perché dobbiamo tagliare il campione in $k$ fette, possibilmente di pari dimensione, quindi supponiamo che $n/k$ sia

un numero intero, ad esempio abbiamo *1000* imprese, *100* fette da *10*, questo è sempre possibile perché se è un numero primo ci limitiamo ad una fetta.

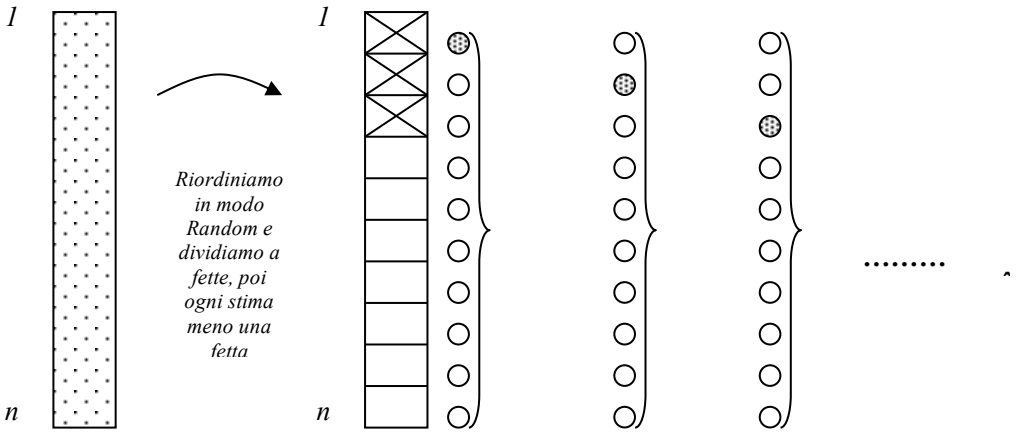



*Fig. 11.3 – Il metodo Jacknife*

Ed a quel punto i $k$ sotto-campioni come li otteniamo? Togliendo una fetta per volta, ovvero la prima stima viene fatta togliendo la prima fetta, la seconda stima viene fatta rimettendo dentro la prima fetta e togliendo la seconda, vuol dire che ad ogni campione manca una sola fetta, e quindi la stima della media di questi $k$ sotto-campioni

$$\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, ....., \hat{\theta}_k$$

è uguale a

$$\bar{\theta} = \frac{\sum_{1}^{k} {}_i \hat{\theta}_i}{k}$$

mentre la varianza è leggermente diversa ma non molto,

$$\hat{V}_j\left(\hat{\theta}\right) = \frac{\sum\left(\hat{\theta}_j - \bar{\theta}\right)^2}{k(k-1)}$$

Questi naturalmente sono metodi applicabili e fattibili solo se si ha un computer a disposizione, in pratica dobbiamo realizzare un programma che automaticamente estrae *1000*, *2000* campioni, e ne fa la stima, mette da parte le stime, e poi ne calcola la varianza, e naturalmente il $k$ deve essere molto grande perché, chiaramente per poter ottenere una stima di varianza dignitosa si tratta di migliaia di stime, e questo è il motivo per cui questi metodi per quanto esasperatamente facili e generali ci potrebbero permettere di calcolare parametri per qualsiasi piano di campionamento. Invece quando possiamo utilizzare le formule il calcolo è molto più veloce, per usare questi metodi ci possiamo mettere anche delle ore, per campioni seri si parla di solito di un $n$ pari a *30.000*, poi andiamo a estrarre da questo *3000*, *4000* sotto-campioni, ci mettiamo una vita, quindi per quanto semplici e generali sono molto lunghi.

I software di statistica più avanzati funzionano più o meno in questo modo. Ci viene chiesto un possibile piano di campionamento noto oppure la matrice delle probabilità di inclusione doppia, e come ultima opzione ci viene chiesto uno di questi metodi, quindi o sappiamo le formule, o la matrice, o uno di questi metodi.

Per curiosità generale aggiungiamo che questi metodi nascono all'interno delle tecniche di indagine per il calcolo di stime di varianza campionaria, ma è vero anche che non valgono solo per queste cose, la generalità di questi due metodi è tale che anche se non lo usiamo per calcolare varianze di popolazioni finite ma per calcolare variabilità di parametri di qualsiasi genere, l'importante è che nel modello che genera i dati ci sia una componente di casualità. In pratica se abbiamo un modello teorico, ad esempio sul rischio di portafoglio, e pari ad una serie di componenti più una componente casuale, simuliamo la componente casuale, estraiamo dei campioni, lo ripetiamo $k$ volte e ci calcoliamo la varianza, quella sarà la varianza, l'incertezza del nostro modello.

# 12. Stime non dirette

## 12.1 La forzatura di una matrice - Iterative Proportionaly Fitting

Riprendiamo un esempio di calcolo iterativo, che probabilmente risulta ancora poco chiaro, con un esempio più semplice; noi abbiamo un elenco di numeri che hanno un

loro totale che conosciamo e non coincide con la somma, abbiamo *100* numeri ad esempio che devono sommare *10.000* ed invece ci danno *9.998*, è tipico nella redazione dei bilanci

$$\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \\ \hline \sum x_i \neq \bar{x} \end{array}$$

*con* $\bar{x}$ *totale conosciuto*

Ritrasformiamo ogni singolo valore, lo dividiamo per la somma calcolata e lo moltiplichiamo per la somma desiderata. Questa è la soluzione più banale, spalmiamo l'errore in quote uguali su tutto, quindi quando non torna il bilancio spalmiamo, fino a qui nulla di iterativo,

$$\frac{x_i}{\sum x_i} \cdot \bar{x}$$

Ma se invece di avere una sola colonna, i conti che dobbiamo far tornare sono due, per esempio abbiamo una tabella a doppia entrata, o una matrice di dati, avremo delle somme di riga e di colonna che calcoliamo ed abbiamo dei totali noti me che non coincidono con le somme calcolate, ad esempio se lavoriamo in un'azienda e ci viene chiesto di compilare una tabellina sulle risorse per i progetti, per esempio: ci sono *100* impiegati e *10* progetti, ciascun impiegato deve lavorare in uno o più progetti, quindi deve tornare sia la somma di riga, che deve essere uno poiché non è che l'impiegato può fare lo straordinario, sia il totale del progetto perché quel progetto può costare molto. Ci chiedono di compilare una tabella, a ciascuna persona chiediamo - tu dove lavori? E quando andiamo a fare le somme vediamo che non ci tornano, ovviamente non è possibile che un progetto che costa *10* lo facciamo pagare *100* - quindi come facciamo a far tornare i totali?

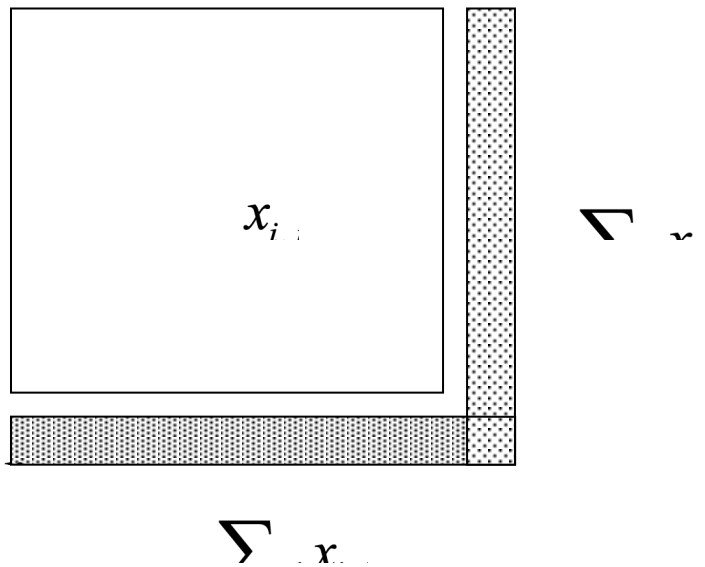


*Fig. 12.1 – La forzatura di una matrice*

Come si fa? Si rispalma prima e su tutte le righe e viene fuori un altro numero, poi risparmiamo questo numero sulle colonne, poi facciamo tornare prima le righe e poi le colonne, un colpo al cerchio ed uno alla botte fino a quando non finisce il problema, ovvero non coincidono i totali, questo è un calcolo iterativo. Fino a quando i numeri non cambiano poco questo procedimento si mostra convergere ad un'altra tabella che è quella di distanza minima da quella originale che avevamo, è la tabella più vicina possibile a quella da cui siamo partiti e che rispetta i totali di riga e di colonna. Questo metodo si chiama *IPF (Iterative Proportionaly Fitting)* ovvero stima proporzionale iterativa.

Uno dei motivi per cui questo metodo è il più usato in applicazioni economiche, è perché esistono delle tabelle la cui costruzione è molto costosa. Ad esempio una tabella origine-destinazione dove abbiamo le stesse intestazioni di riga e di colonna. Supponiamo che in un'indagine sui comuni vogliamo sapere ogni persona in quale comune lavora e conoscere i flussi di pendolarismo. Costruire questa tabella è molto costoso perché significa distribuire tanti questionari e di solito viene condotta ogni *10* anni.

Questa matrice ha delle implicazioni a livello territoriale per l'economia importantissime perché è inutile costruire fabbriche dove non ci abita nessuno oppure dove non ci va nessuno a lavorare; la matrice è un parametro fondamentale per studiare i flussi di traffico, ma anche per aprire uno sportello bancario. E' più ragionevole aprire uno sportello bancario dove lavora la gente dato che gli spostamenti durante il giorno sono fondamentali a livello economico.

Ma come si aggiorna questa matrice? I totali di questa matrice se vogliamo li possiamo conoscere anche mensilmente, ogni mese sappiamo quanta gente risiede in un comune e quanta gente lavora in un comune, il fatto è che non sappiamo da dove vengono, da dove viene una persona, sappiamo i due totali ma non sappiamo gli incroci, per sapere gli incroci prendiamo l'ultima matrice e la riadattiamo ai nuovi totali, è una modalità di *aggiornamento di tabelle a doppia entrata.*

L'origine-destinazione è molto importante dal punto di vista economico ad esempio per i flussi di import-export, è importante sapere da dove vengono le merci e dove vanno.

Un'altra tipica matrice origine-destinazione, riguarda le tavole input-output per i flussi intersettoriali, dove abbiamo i settori di attività economica sia sulle righe sia sulle colonne ed all'interno le caselle indicano quanti prodotti ogni settore vende ad un altro. Ad esempio facciamo una matrice *3x3* (anche se in realtà sono molto più grandi) in cui abbiamo tre settori, Agricoltura, Industria, Servizi, ed ogni singolo settore per produrre una quantità di *100* miliardi di prodotto acquista tot. dal settore servizi, tot. dal settore industria, tot. dal settore agricoltura. Abbiamo praticamente una funzione di produzione che ci dice che per ogni lira prodotta da ogni settore questa è composta da tot. lire provenienti da altri settori, ad esempio nel settore calzaturiero per ogni lira prodotta dovremo prendere dei fattori di produzione come cuoio, gomma, lacci, che hanno un costo, da altri settori.

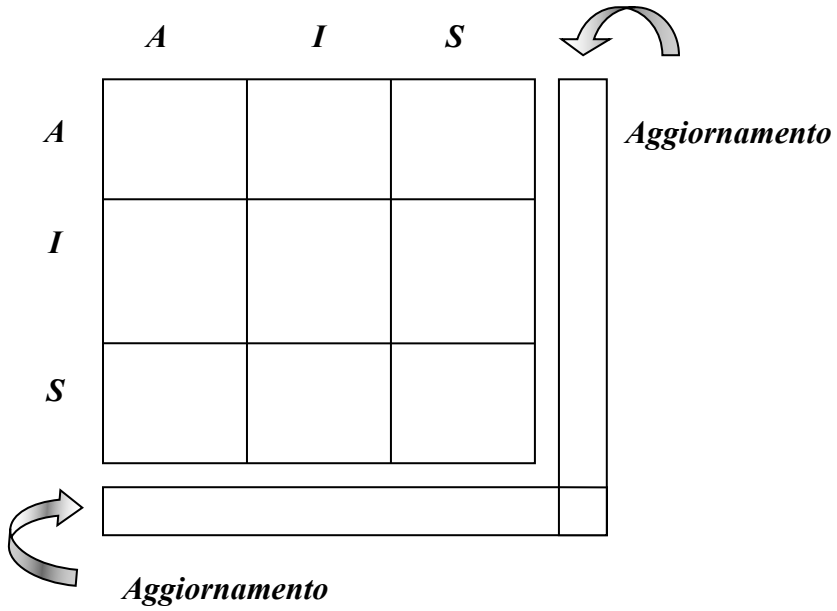



*Fig. 12.2 – La forzatura della matrice con tre settori di attività economica*

Questa tabella è fondamentale per capire su quali settori investire o no, poiché se un settore rimane sempre lo stesso, ovvero non compra nulla da nessun'altro, investire in quel settore è da pazzi (per investimento intendiamo le sovvenzioni del governo); dare sovvenzioni ad un settore invece che rimettere in circolazione ogni euro che è stato sovvenzionato perché compra da altri è molto più produttivo, per la logica del moltiplicatore che rimette in circolazione il denaro secondo una serie geometrica, quindi da questa matrice si riesce a fare programmazione economica, e questo è il motivo per cui alcuni settori sono più favoriti. In questo modo abbiamo una misura quantitativa del flusso degli scambi e di come il denaro viene rimesso in circolazione. Queste tabelle si costruiscono facendo indagini sulle imprese e per ciascuna impresa in ogni settore chiediamo gli acquisti in tutti gli altri settori, è abbastanza facile ma il problema è che bisogna chiederlo a troppe imprese, l'indagine molto pesante e costosa.

Questo è il motivo per cui questa indagine viene effettuata solo ogni tanto, ma come si realizza l'aggiornamento? Viene fatto sulla base di totali noti che sono quelli provenienti dalla contabilità nazionale, ovvero sono il totale dei costi intermedi che ha ogni settore, e questo totale lo otteniamo come differenza tra produzione e valore aggiunto, la produzione ed il valore aggiunto per ogni settore si conoscono, ne facciamo

la differenza ed otteniamo il totale dalla nuova matrice. Cosi come il totale di riga, che rappresenta i consumi intermedi, viene ottenuto come differenza tra risorse disponibili e consumi finali nel senso che ogni settore produce per vendere agli altri settori e per vendere alle famiglie o ai consumatori finali; ad esempio nel caso dell'agricoltura è vero che alcuni prodotti sono sia intermedi di trasformazione sia finali per la famiglia, ovvero il latte può essere utilizzato sia dalla azienda di trasformazione per fare il formaggio sia consumato dalla famiglia, semplicemente noi sottraiamo alla quantità di latte totale il latte consumato dalla famiglia ed otteniamo il totale dei consumi intermedi ovvero quanto latte viene venduto agli altri settori, e questi totali sono noti ogni anno. In pratica ogni anno abbiamo nuovi totali, li mettiamo nella tavola input-output, procediamo in modo iterativo e otteniamo l'aggiornamento della matrice. Ma quali sono i settori in cui ogni euro si moltiplica di più? Qual è il settore più trainato e trainante? Quello delle assicurazioni e quello manifatturiero, sono i settori che richiedono più fattori produttivi e quindi distribuiscono a pioggia praticamente su tutta l'Economia.

## 12.2 I salti di strato

Ora cerchiamo di rispondere ad una domanda, ad un probabile problema che possiamo riscontrare nelle nostre indagini. Abbiamo una popolazione di *N* unità, facciamo il caso tipico, le imprese, e stratifichiamo per classe di addetti, da *0* a *50* addetti, da *50* a *500* addetti, *Strato > 500* addetti.

| *Addetti* | |
|---|---|
| *0 - 50* | |
| *50 - 100* | |
| *X > 500* | |

Quando facciamo l'indagine però, quando andiamo a fare il campione succede una cosa strana, e che tra l'altro non è affatto strana, è frequentissima, succede sempre. Noi in teoria avremmo potuto prendere *100* unità del primo strato, *100* del secondo, e *100* del terzo.

| | *0 - 50* | *50 - 100* | *X > 500* | |
|---|---|---|---|---|
| *0 - 50* | *80* | *20* | *0* | *100* |
| *50 - 100* | *20* | *60* | *20* | *100* |
| *X > 500* | *0* | *10* | *90* | *100* |
| | *100* | *90* | *110* | |

*Fig. 12.3 – Esempio di totali di colonna non tutti uguali*

Durante la rilevazione però ci accorgiamo che il numero di addetti che avevamo chiesto in fase di costruzione dell'archivio non corrisponde a quello registrato sui questionari, ci accorgiamo che appartengono ad un altro strato, ovvero succede che quando andiamo a fare la rilevazione ci accorgiamo che l'unità che avevamo inserito nell'archivio per un determinato numero di addetti non corrisponde, non è inserita nello strato che avevamo stabilito, magari perché l'impresa dalla fase di costruzione dell'archivio fino alla somministrazione dei questionari è cresciuta, probabilmente quando abbiamo costruito l'archivio era in una fase iniziale di sviluppo, quindi la domanda è - quando andiamo a fare le nostre stime quale codice di strato utilizziamo per la nostra unità ballerina, quello con cui le abbiamo estratte o quello che effettivamente verifichiamo dai questionari essere il codice attuale di strato?

Questo succede nelle indagini reali quasi sempre, a meno che non si utilizzino codici di strato immutabili. Per esempio anche la regione di residenza che dovrebbe essere la più stabile può cambiare. Teniamo presente che a meno che non stratifichiamo rispetto al sesso, ed anche qui ogni tanto c'è qualche sorpresa, la maggior parte delle variabili di stratificazione come gli addetti cambiano spesso. Pensiamo di andare a stratificare una famiglia per numero di componenti, succede di tutto.

*Cerchiamo ora di rispondere alle seguenti domande:*

*1) In definitiva quale codice di strato utilizziamo, quello prima della rilevazione e quindi dell'archivio, o quello risultante dalla rilevazione?*

*2) Questa aspetto delle unità che saltano da uno strato all'altro e quindi il passaggio di strato, è un problema piccolo, medio, grande, gigantesco o insormontabile?Perché?*

*3) Nel caso in cui questo problema sia insormontabile come lo risolviamo?*

*Risposte:*

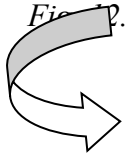
L'Horwitz-Thompson si basa esclusivamente sulle probabilità d'inclusione e quindi dei pesi di riporto all'universo basati sulle unità presenti nello strato nel momento in cui le abbiamo estratte, i pesi di riporto all'universo quindi sono funzione esclusivamente della probabilità di inclusione iniziale e la probabilità di inclusione è funzione solo dell'urna da cui abbiamo estratte le unità, quindi le imprese che abbiamo visto prima le abbiamo prese a prescindere dalla loro dimensione, per cui logicamente dovremmo prendere le unità in entrata, i codici di strato in entrata per rispettare l'Horwitz- Thompson.

Allora supponiamo che lo stratificato non sia accettabile perché non ha alcuna differenza rispetto al campionamento casuale semplice e quindi gli strati sono tutti uguali in termini di varianza cioè c'è poca varianza tra gli strati e molta dentro, che un'unità passi da uno strato all'altro non comporta grossi problemi, quindi più c'è differenza tra le varianze di ogni strato e quindi strati omogenei al loro interno e più questo problema diventa rilevante, mentre se c'è molta varianza negli strati e gli strati sono omogenei tra di loro questo problema non è molto rilevante. Il peso di questo problema è una *funzione inversa della qualità dello stratificato*, ovvero una funzione diretta della varianza all'interno di ogni strato e inversa alla varianza tra gli strati.

Naturalmente questo *dipenderà dal tipo di variabile*, ad esempio se facciamo un campione d'imprese in grandi medie e piccole, e rileviamo una cosa che non centra nulla con la dimensione (di solito più è grande l'impresa e più il fatturato per addetto cala per le economie di scala quindi non va bene il fatturato per addetto), ad esempio quanto paga l'impresa di *ICI*, sulle variabili che sono *scorrelate con la stratificazione* e quindi su cui la stratificazione non aiuta questo problema, non crea molti disastri, invece paradossalmente più la variabile è correlata con lo strato, con la stratificazione, e quindi ad esempio il fatturato, e più potremmo incorrere in problemi, ad esempio,

| | *W (ordine di grandezza, a titolo esemplificativo, del peso di riporto all'universo*) | *X (spesa in ricerca e sviluppo)* |
|---|---|---|
| *G* | *1* | *1000.000* |
| *M* | *10* | *10.000* |
| *P* | *1000* | *0* |

*Figura 2.2 – Esempio di un salto di strato*



come si può ben vedere, le imprese grandi le prendiamo sicuramente, se abbiamo assegnato

una probabilità, dopo aver stratificato, in maniera proporzionale al numero di addetti. Ora se una unità di quelle grandi, e quindi con probabilità uguale ad uno e rispettivo peso di riporto, passa in un altro strato, si capisce subito il danno che si crea dopo aver effettuato o nell'effettuare la stima, noi applichiamo in pratica l'Horwitz-Thompson sui pesi di riporto che abbiamo stabilito all'inizio, quindi se un'unità che ha *1 milione di euro* in ricerca e sviluppo e dopo la rilevazione ci casca tra le imprese piccole dovremo calcolare l'Horwitz-Thompson con un peso di riporto di *1000* ed un valore di *1 milione*, (*1000x1000.000*), otterremo delle cifre abnormi, invece dall'esempio che abbiamo fatto una unità grande rappresenta solo se stessa e non *1000* imprese, il danno è tale che una singola unità riportata all'universo potrebbe essere più grande di tutto il campione messo insieme, e più c'è differenza tra gli strati secondo la variabile di stratificazione e più creiamo danni, perché in pratica abbiamo valori estremi, e più sono estremi i valori, più sono distanti, e più sarà estremo il danno.

Supponiamo di avere un campione d'imprese e un'unità del nostro campione che si è appena iscritta alla camera di commercio, e che già sapeva che durante l'anno avrebbe assunto *300* persone - come facciamo ad impedire a questa impresa di crearci dei problemi? Cosa facciamo? La togliamo? Sì, ma in un modo molto particolare, supponiamo di avere un archivio con *N* unità ed un campione di *n*, e supponiamo che

l'unità che ha saltato lo strato sia proprio l'ultima - allora possiamo utilizzare le prime *(n -1)* unità per fare una stima su *(N – 1)* unità? Sì, quindi prima facciamo una stima sulle unità che non hanno saltato lo strato, quelle giuste insomma, come abbiamo sempre fatto, e poi aggiungiamo il valore di quella anomala, è come se la censissimo, come quando nel *PPS* otteniamo delle probabilità d'inclusione uguali ad 1, insomma facciamo la nostra stima su (*n -1)* unità e poi censiamo l'unità anomala, e quindi,

$$\hat{t}_{HT} = \hat{t}_{n-1} + \hat{t}_{n-esima}$$

se poi le unità anomale sono più di una, ovvero un numero limitato, queste situazioni vengono chiamate *Outlyer* (dati anomali), se abbiamo un numero di unità limitato rispetto alla numerosità del campione, ad esempio *10* o *15*, ne togliamo *10* o *15.* Quindi le *(n – 10)* le utilizziamo per fare una stima sulle *(N – 10).*

In ogni caso è il fenomeno stesso che ci da i dati anomali e la statistica serve proprio a fare in modo che torni il conto, proprio con i dati anomali, e proprio perché il dato è anomalo deve rappresentare solo se stesso.

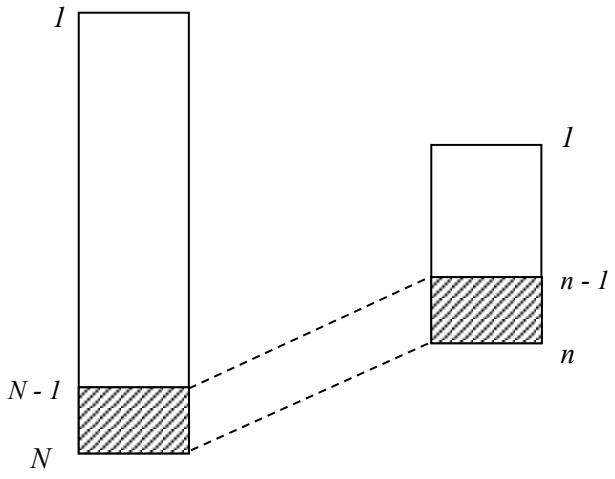



*Fig. 12.3 – Correzione dei salti di stato*

E se abbiamo tutti dati anomali? Ad esempio nell'anno *2005* si è verificato che in un'indagine sulle aziende agricole dopo aver fatto le stime è venuto fuori che la

produzione agricola a frumento è salita del 30%, quindi tutti dati anomali - li metto tutti da parte?

No, dopo aver telefonato ai venditori di sementi è venuto fuori che le sementi vendute erano aumentate del 30%; quindi come vediamo si potrebbe anche per mancanza di fede non tenere conto della teoria dei campioni ma è praticamente impossibile che tutte le unità che abbiamo estratto siano, migliaia e migliaia di unità, tutti dati anomali, invece è più verosimile che sia successo qualcosa al fenomeno, come effettivamente è stato, perché la comunità europea aveva messo a disposizione dei fondi per estendere le superfici a frumento.

In questo caso bisogna sempre mettersi in discussione ma non esagerare e quindi verificare effettivamente se ci sono dati anomali o se il fenomeno è anomalo solo per quell'anno.

### 12.3 La post-stratificazione

Ora precisiamo che per quanto riguarda tutto l'apparato del salto strato, anche se sistemiamo la questione dei dati anomali, tutto l'apparato comunque crea qualche problema, ad esempio come nell'esempio precedente *1000.000x1000*.

Quindi come soluzione generale dobbiamo si usare i pesi di riporto all'universo in entrata per l'Horvitz-Thompson ma poi dobbiamo fare in modo che le stime rispettino dall'altra parte i totali degli strati di destinazione. Quello che è importante capire è che non stiamo usando più l'Horvitz-Thompson come stimatore, ma stimatori che non si basano sui pesi di riporto all'universo calcolati sulle probabilità di inclusione ma che utilizzano *variabili ausiliarie* contenute nell'archivio.

Supponiamo di avere il nostro archivio ed avere dei dati ausiliari, supponiamo il numero addetti su cui possiamo fare un *PPS* ma invece di fare un *PPS* subito facciamo un campione casuale semplice e alla fine utilizziamo l'informazione ausiliaria che ci fornisce la variabile addetti dopo aver fatto la rilevazione in fase di stima. Supponiamo di avere il nostro campione e che ogni unità del nostro campione ha un peso di riporto uguale a *15*; e supponiamo di sapere quanto pesa la somma degli addetti su tutta la popolazione, ad esempio il *10%*; oppure potremmo anche pensare ad un *1.000.000* di imprese, ne prendiamo *1000* ed avremo un peso di riporto all'universo di *1000* e le *1000*

che abbiamo preso ad esempio rappresentano l'1% degli addetti della popolazione, cosa che possiamo calcolare perché abbiamo il totale degli addetti. Quindi il nostro campione ha preso un'unità ogni *15* della popolazione, la variabile addetti per la popolazione la conosciamo perché c'era nell'archivio e nel nostro campione, sappiamo fare la somma degli addetti nel nostro campione, sappiamo fare la somma degli addetti per tutta la popolazione, se il nostro campione ha un peso del *10%* sugli addetti di tutta la popolazione potrebbe essere ragionevole invece di usare il peso *15* usare il peso *10* che è basato sugli addetti; questo però ci porta ad allontanarci dall'Horvitz-Thompson che conta solo la probabilità di inclusione, *che si basa solo sul disegno campionario* e che aveva peso di riporto all'universo uguale a 15; quindi lo stimatore che useremo in questo caso non è lo stimatore di Horvitz-Thompson, e si basa non sulla estrazione, cosa di cui tiene conto l'Horvitz-Thompson sempre, ma tiene conto del fatto che possono essere introdotte delle variabili ausiliarie, in questo caso gli addetti non li abbiamo usati per l'estrazione ma per fare stime.

Nel nostro problema dei salti di strato i due stimatori sono diversi, c'è una differenza, uno è lo stimatore di Horvitz-Thompson che si basa sulle probabilità ed i pesi riguardanti l'estrazione del campionamento stratificato, invece nell'altro caso abbiamo uno stimatore che si basa su variabili ausiliarie nella *post-stratificazione*, cosa completamente diversa dalla *pre-stratificazione* fatta dall'Horvitz-Thompson, perché facciamo tornare i totali della popolazione ottenuti sulla base della stratificazione ottenuta a valle e non a monte, ovvero sulla base degli strati ottenuti e non sulla base di quelli estratti; ora, anche se la stima è più o meno uguale le *varianze sono completamente diverse* e tutte le formule che abbiamo visto non valgono più. Quindi usare delle variabili ausiliarie per post-stratificare alla fine è il modo generale per trattare questi dati, vuol dire che per ogni strato in entrata post-stratifichiamo sugli strati in uscita. Il problema è che con questo metodo viene fuori uno stimatore che non centra niente con l'Horvitz-Thompson.

| | *ADDETTI* | | | |
|---|---|---|---|---|
| | | | | |
| | *10%* | | *1* | *15-HT* |
| | *100%* | | | |

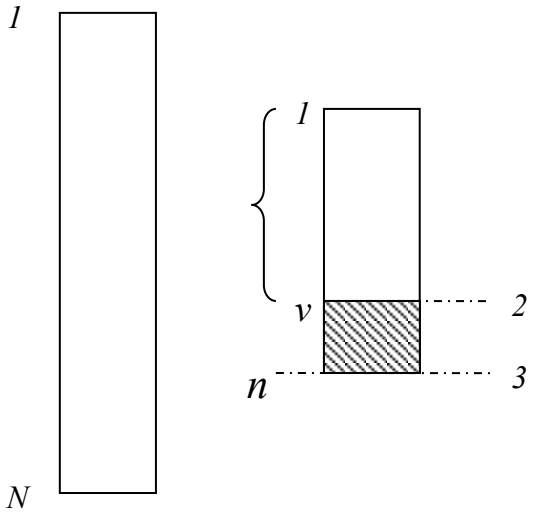
*n*

## 12.4 Stima della sovra-copertura dell'archivio

Sappiamo che nell'archivio abbiamo *N* unità, estraiamo il nostro campione di *n* unità e di queste sappiamo che solo *v* risultano vive, nel senso che sono uscite dal campo di osservazione, per morte s' intende che stanno fuori dalla rilevazione. All'inizio del corso avevamo detto che quando facciamo gli archivi tendiamo ad eccedere poiché l'errore di sovra-copertura lo possiamo stimare, riusciamo a correggere facilmente la sovra-copertura, ed una volta appresa la teoria dei campioni possiamo vedere effettivamente come si corregge. In questo caso in pratica quando andiamo ad effettuare la rilevazione ci accorgiamo che l'unità non esiste più. Come si corregge allora? Proposte di alcuni studenti:

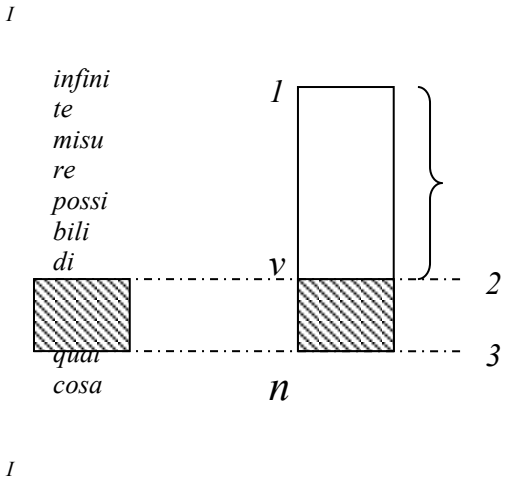
1) Ad esempio per un compagno si corregge in questo modo, calcolando le probabilità di inclusione

$$w = \frac{N}{v}.$$




2) Per un altro compagno si corregge in questo modo, togliendo dall'archivio anche le unità $(n-v)$ e poi ricalcolando le probabilità di inclusione come

$$\left[\frac{N-(n-v)}{v}\right];$$




3) Per un altro compagno invece bisogna fare $\hat{N}=\frac{v}{n}\cdot N$ e $\hat{w}=\frac{\hat{N}}{v}$;

4) Ancora per un'altra compagna invece bisogna considerarle mancate risposte il che equivale quasi alla seconda proposta;

Cerchiamo si spiegare ora le lacune di alcune risposte o la ragionevolezza di altre. Nel primo caso non consideriamo che tipo di unità sono, in pratica fingiamo che le unità siano tutte vive e riportiamo le nostre unità ad *N* barando sulla situazione reale, ci tappiamo gli occhi e diciamo che non è morto nessuno, ed è esattamente come considerarle mancate risposte, quando invece potremmo toglierle dall'archivio; la seconda ipotesi invece è totalmente folle; nel caso della terza

$$\hat{N} = \frac{v}{n} \cdot N$$

non è così ma

$$\hat{N} = \frac{\sum_{i=1}^{v} w_i}{\sum_{i=1}^{n} w_i} \cdot n$$

che alla fine è la stessa cosa dato che siamo in un casuale semplice. Ma

$$\hat{w} = \frac{\hat{N}}{v}$$

è inutile, non ce n'è bisogno;
Ora sappiamo che lo stimatore dell'Horvitz-Thompson

$$\hat{t}_{HT} = \sum x_i \cdot w_i$$

determina il totale della popolazione. Partendo da questo inseriamo all'interno della formula, per $x_i$, una bernoulliana, dove gli 0 ed 1 stanno ad indicare se l'unità è presente o no, e otteniamo il numero di vivi ed il numero di morti. *Dal punto di vista formale significa introdurre una colonna in più dove stimiamo il nuovo totale della popolazione come se stessimo facendo l'Horwitz–Thompson, solo per precisare che abbiamo un certo numero di imprese cessate, per il resto operiamo con l'Horvitz-Thompson come sempre senza ricalcolare i nuovi pesi, altrimenti avrei dei pesi basati*

*sui soli vivi.* Ragionevolmente potremmo anche fare uno stratificato perché magari la stragrande maggioranza delle mortalità si ha nello strato delle piccole imprese, ed anche se non abbiamo stratificato a monte il campione, quando stimiamo la numerosità *n* lo facciamo stratificando per omogeneità "dei morti", dei tassi di mortalità, quindi calcoliamo la Y normalmente con gli stessi pesi

$$\hat{t}_{HT} = \sum y_i \cdot w_i$$

e ricalcoliamo la numerosità della popolazione

$$\hat{N} = \hat{t}_{HT} = \sum x_i \cdot w_i$$

con

$$x_i = 0,1(MORTE, VIVE)$$

Ora visto che spesso utilizziamo degli esempi sulle imprese queste cose che abbiamo detto finora spesso sono più importanti dell'indagine stessa perché la *demografia d'impresa*, cioè le imprese che sono cessate o ancora vive, imprese che si attivano e imprese che cessano, le stime di demografia d'impresa sono a volte più importanti di altre indagini sulle imprese, a volte fare la stima di $\hat{N}$ è molto più importante di qualsiasi altra stima, esiste addirittura una materia a sé stante che riguarda questo argomento.

## 13. Esercizi

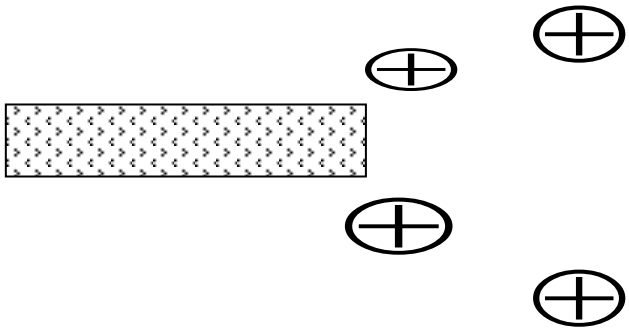
### 13.1 Le domande randomizzate

Cominciamo con un'indagine in cui ci sono domande che violano fortemente la privacy delle persone, e conduciamo l'indagine tramite domande randomizzate e nota bene che le domande randomizzate valgono soltanto per domande cui è possibile rispondere solo

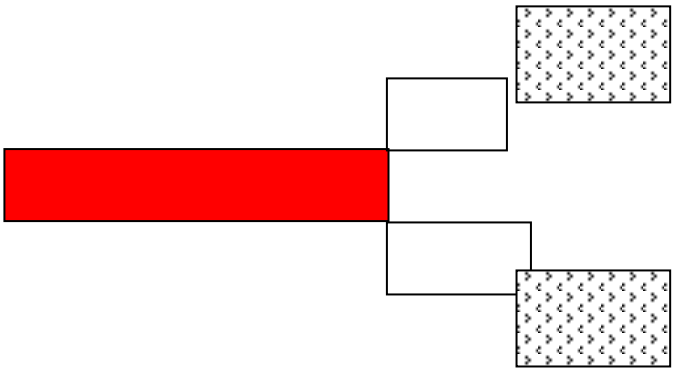
con sì e no; esiste una forma con più modalità ma inapplicabile per variabili quantitative.

| | | |
|---|---|---|
| Domanda vera (NRS): | *hai copiato all'esame?* | *(si-no)* |
| Domanda falsa, di controllo(NRC): | *sei pugliese?* | *(si-no)* |

Allora, per quanto riguarda la domanda falsa ne conosciamo la frequenza, cioè vediamo quanti sono pugliesi e quanti non lo sono,



ed intervistiamo *100* persone di cui *60* persone rispondono (si) e *40* persone (no);



Ora utilizzando il nostro meccanismo di estrazione cerchiamo la frequenza delle persone che hanno copiato all'esame:

$$NR\% = \pi \cdot NRC\% + (1-\pi)NRS\%$$

$$NRS\% = \frac{NR\% - \pi \cdot NRC\%}{1-\pi}$$

$$60 = 0,5 \cdot (0,2 \cdot 100) + (0,5) \cdot NRS$$

$$NRS = \frac{0,6 - 0,5 \cdot (0,2 \cdot 100)}{0,5} = 100$$

*Attenzione, perché qui abbiamo già moltiplicato all'interno della formula le frequenze x gli intervistati, se invece avessimo lavorato solo con le frequenze avremmo avuto come risultato (1) che poi avremmo dovuto moltiplicare per il numero di intervistati;*

$$0,6 = 0,5 \cdot (0,2 \cdot 100) + (0,5) \cdot NRS\%$$

$$NRS\% = \frac{0,6 - 0,5 \cdot (0,2)}{0,5} = 1 \quad \Rightarrow 1 \cdot 100 = 100\%$$

Vuol dire che tecnicamente chi ha copiato all'esame è sicuramente pugliese, e chi ha risposto *no* non è pugliese.

## 13.2 Teoria dei campioni

*(Attenzione, qualche numero potrebbe essere sbagliato ma i procedimenti sono giusti ed anche i campioni estratti, per il resto basta applicare le formule, magari confrontate i risultati).*

***Esercizio .1***

Siamo un comune che vuole effettuare dei controlli sul pagamento dell'Ici, abbiamo l'archivio delle abitazioni del nostro comune, abbiamo il codice identificativo da *1* a *10* ma di solito c'è il codice catastale e quello che ci serve fondamentalmente è sapere il numero di vani.

| *Codice* | *N° Vani* | *Mq* | *1-2 casa* | *Y (ICI Pagata) €* | *r* |
|---|---|---|---|---|---|
| 1 | 3 | 90 | 0 | 200 | 0,7 |
| 2 | 4 | 110 | 1 | 600 | 0,6 |
| 3 | 2 | 70 | 0 | 150 | 0,4 |
| 4 | 2 | 60 | 0 | 140 | 0,5 |
| 5 | 4 | 130 | 1 | 700 | 0,5 |
| 6 | 6 | 180 | 1 | 1000 | 0,4 |
| 7 | 8 | 250 | 0 | 500 | 0,9 |
| 8 | 1 | 30 | 1 | 50 | 0,1 |
| 9 | 2 | 70 | 1 | 150 | 0,2 |
| 10 | 3 | 80 | 0 | 180 | 0,3 |
| | | | | | |

*1-2 casa :* *0 – prima casa*
*1 – seconda casa*

Dobbiamo fare un campione perché non possiamo intervistare tutte le case, il rilevatore costituito da un ufficiale comunale rileva i dati ovvero l'ici pagata. Qui il valore dell'ici lo diamo per tutte e *10* le unità poiché qualsiasi campione estraiamo abbiamo sempre una risposta, ma naturalmente nella realtà non è così in quanto la risposta viene data dopo l'estrazione e quindi risponderanno solo le unità del campione,

in pratica abbiamo messo le risposte che si suppone vengano ottenute dopo l'estrazione dalle unità, tutte.

*a) Come selezioniamo il campione?*

In primis facciamo un campione di cinque unità, stratificato per prima e seconda casa e *PPS* sui metri quadri, del numero vani non ce ne importa nulla in quanto prima di tutto è quasi proporzionale ai metri quadri e poi perché si paga sui metri quadri e non sui vani.
Generiamo il nostro numero Random, stratificato rispetto alla prima e alla seconda casa, delle prime case ne prendiamo due e delle seconde case ne prendiamo tre e successivamente eseguiamo un campionamento *PPS* in entrambi gli strati rispetto ai metri quadri; sul campione estratto poi, sia per lo stratificato sia per il *PPS* dobbiamo fare la stima dell'ici totale e calcolare la relativa varianza.

*b) Come calcoliamo le numerosità?*

Successivamente ipotizziamo che il nostro campione di cinque case sia un campione pilota, sulla base del quale, sempre stratificato rispetto a prima e seconda casa, estraiamo un campione casuale semplice stratificato, con l'unico fine di calcolarci le varianze per strato, e calcoliamo le numerosità di un eventuale campione definitivo per avere una stima dell'ici con coefficiente di variazione $C(\hat{t}_{HT}) \leq 10\%$ , ovvero ci calcoliamo le numerosità in modo da avere tale coefficiente di variazione, sarebbe Neyman con il $C(\hat{t}_{HT})$ uguale a 0,1.

*a) Svolgimento*

La prima cosa da tenere in considerazione è che ora abbiamo definito le numerosità e vogliamo conoscere la varianza dello stimatore $V(\hat{t}_{HT})$. Per estrarre il campione secondo il piano di campionamento *PPS* dobbiamo legare la nostra variabile $Y(ICI)$ ad un'altra variabile di tipo quantitativo $X(Mq)$,

*Totale 1° strato(Mq)= 550*
*Totale 2° strato(Mq)= 520*

| *Codice* | *r* | *Riordiniamo rispetto al Codice di Strato* | $\pi_{i,h} = \frac{x_{i,h}}{\sum_{i=1}^{N_h} x_i} \cdot n_h$ *Attenzione al denominatore, vanno i totali di strato* | $R = \frac{\pi_i}{r}$ | *(Ora possiamo riordinare in senso decrescente)* | $W_i = \frac{1}{\pi_i}$ | *Y (ICI Pagata) €* |
|---|---|---|---|---|---|---|---|
| 1 | 0,7 | 1 (0) | 0,327 | 0,467 | 1,01 (7) | 1,100 | 500 |
| 2 | 0,6 | 3 (0) | 0,255 | 0,638 | 0,97 (10) | 3,436 | 180 |
| 3 | 0,4 | 4 (0) | 0,218 | 0,436 | 0,638 (3) | .... | |
| 4 | 0,5 | 7 (0) | 0,909 | 1,01 | 0,467 (1) | …. | |
| 5 | 0,5 | 10 (0) | 0,291 | 0,97 | 0,436 (4) | …. | |
| 6 | 0,4 | 2 (1) | 0,635 | 1,058 | * * * (6) | *1(censita)* | 1000 |
| 7 | 0,9 | 5 (1) | 0,750 | 1,5 | 2,02 (9) | 2,475 | 150 |
| 8 | 0,1 | 6 (1) | 1,038 | 2,595 | 1,73 (8) | 5,780 | 50 |
| 9 | 0,2 | 8 (1) | 0,173 | 1,73 | 1,5 (5) | …. | |
| 10 | 0,3 | 9 (1) | 0,404 | 2,02 | 1,058 (2) | | |

*** abbiamo censito l'unità con codice identificativo 6 poiché ha probabilità *1* di essere estratta, inoltre non calcoliamo il peso di riporto in quanto rappresenta solo se stessa.

*Attenzione!! Dopo aver effettuato la stratificazione al primo stadio e quindi dopo aver teoricamente censito, estraiamo dei campioni al secondo stadio e naturalmente qualsiasi parametro verrà stimato tenendo conto delle nuove probabilità d'inclusione basate sulle numerosità dei rispettivi strati, qualsiasi sia il piano di campionamento che facciamo, le probabilità si staccano dalla numerosità totale del nostro archivio, ad esempio tra piano PPS ad ogni strato SRS e avremo probabilità diverse e quindi una stima diversa ma ambedue i piani si basano sulle numerosità degli strati.*

Quindi prendiamo il seguente campione *(7,10,6,9,8)*

*I. Stima del totale*

$$\hat{t}_{HT,1STR} = (1{,}100 \cdot 500) + (3{,}436 \cdot 180) = 1168{,}48$$

$$\hat{t}_{HT,2°STR} = (1 \cdot 1000) + (2{,}475 \cdot 150) + (5{,}780 \cdot 50) = 1660{,}25$$

$$\hat{t}_{HT} = (1{,}100 \cdot 500) + (3{,}436 \cdot 180) + (1 \cdot 1000) + (2{,}475 \cdot 150) + (5{,}780 \cdot 50) =$$
$$= 550 + 618{,}48 + 1000 + 371{,}25 + 289 = 2828{,}73$$

*II. Stima della varianza*

Dobbiamo tenere presente che:

$$\hat{V}_{SPS}\left(\hat{t}_{HT}\right) = \frac{1}{n} \cdot \left(\frac{N}{N-1}\right) \cdot \sum_{1}^{n}{}_{i} \left(1 - n \cdot p_i\right) \left(\frac{y_i}{p_i} - \hat{y}_{TOT}\right)^2 p_i$$

con

$$p_i = \frac{x_i}{\sum_{n}^{1} x_i}$$

la dobbiamo calcolare per ciascun strato, trattare i totali di strato come indipendenti dalla numerosità totale, quindi nel nostro caso,

$$\hat{V}_{SPS}\left(\hat{t}_{HT}\right) = \frac{1}{n_h} \cdot \left(\frac{N}{N-1}\right) \cdot \sum_{1}^{n_h}{}_{i} \left(1 - n_h \cdot p_i\right) \left(\frac{y_i}{p_i} - \hat{y}_{TOT}\right)^2 p_i$$

con

$$\pi_{i,h} = \frac{x_{i,h}}{\sum_{i=1}^{N_h} x_i} \cdot n_h = n_h \cdot p_i$$

quindi fare attenzione a non fare confusione. Per il *1° strato* avremo

$$\hat{V}_1\left(\hat{t}_{HT}\right) = 0,5\cdot(0,625)\cdot\left[\left(1-0,909\right)\cdot\left(\frac{500}{0,455}-1168,48\right)^2\cdot 0,455\right]+$$

$$+0,5\cdot(0,625)\cdot\left[\left(1-0,291\right)\cdot\left(\frac{180}{0,145}-1168,48\right)^2\cdot 0,145\right]=$$

$$=0,3125\cdot\left[\left(0,091\right)\cdot 2202,757\right]+0,3125\cdot\left[\left(0,709\right)\cdot 770,575\right]=$$

$$=466,743$$

Per il *2° strato* avremo:

$$\hat{V}_2\left(\hat{t}_{HT}\right)=\frac{1}{3}\cdot(0,625)\cdot\left[\left(1-1\right)\cdot(.....)^2\cdot 0,346\right]+\frac{1}{3}\cdot(0,625)\cdot\left[\left(1-0,404\right)\cdot\left(\frac{150}{0,135}-1660,25\right)^2\cdot 0,135\right]+$$

$$+\frac{1}{3}\cdot(0,625)\cdot\left[\left(1-0,173\right)\cdot\left(\frac{50}{0,058}-1660,25\right)^2\cdot 0,058\right]=$$

$$=0+0,248\cdot(68266,67)+0,345\cdot 12383,448=22860,689$$

e poi

$$\hat{V}\left(\hat{t}_{HT}\right)=\hat{V}_1(\hat{t}_{HT})+\hat{V}_2(\hat{t}_{HT})=23327,432$$

*b) Svolgimento*

| *Codice* | *r* | *Riordiniamo rispetto al codice di strato* | *r* | *Ordiniamo in senso crescente il numero random)* | $\pi_i=\frac{n_h}{N_h}$ | $W_i=\frac{1}{\pi_i}$ | *Y (ICI Pagata) €* |
|---|---|---|---|---|---|---|---|
| 1 | 0,7 | 1 (0) | 0,7 | 0,3 (10) | 0,4 | 2,5 | 180 |
| 2 | 0,6 | 3 (0) | 0,4 | 0,4 (3) | 0,4 | 2,5 | 150 |
| 3 | 0,4 | 4 (0) | 0,5 | …. | 0,4 | …. | …. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4 | 0,5 | 7 (0) | 0,9 | …. | 0,4 | …. | …. |
| 5 | 0,5 | 10 (0) | 0,3 | …. | 0,4 | …. | …. |
| 6 | 0,4 | 2 (1) | 0,6 | 0,1 (8) | 0,6 | 1,67 | 50 |
| 7 | 0,9 | 5 (1) | 0,5 | 0,2 (9) | 0,6 | 1,67 | 150 |
| 8 | 0,1 | 6 (1) | 0,4 | 0,4 (6) | 0,6 | 1,67 | 1000 |
| 9 | 0,2 | 8 (1) | 0,1 | …. | 0,6 | …. | …. |
| 10 | 0,3 | 9 (1) | 0,2 | …. | 0,6 | …. | …. |
| | | | | | | | |
| | | | | | | | |

Quindi il nostro campione casuale stratificato sarà *(10,3,8,9,6)*

Tieni a mente che la numerosità viene trovata in modo proporzionale alla varianza per prendere un maggior numero di unità negli strati più variabili ed un minor numero di unità negli strati più omogenei, è la logica dei nostri calcoli. La numerosità per ogni strato la dobbiamo trovare tramite,

$$n_h = \sqrt{a_h} \sum_{1}^{H} {}_h a_h$$

con

$$a_h = \frac{N_h \cdot S_h^2}{c^2 \cdot t_h{}^2 + \sum N_h \cdot S_h^2}$$

- $t_h$ ce lo stimiamo per ogni strato tramite

$$\hat{t}_{HT(STS)} = \frac{N_h}{n_h} \cdot \sum_{1}^{n_h} {}_i y_i$$

diversamente dal casuale semplice in cui avremmo usato

$$\hat{t}_{HT(STS)} = \sum_{1}^{H}{}_{h} \cdot \frac{N_h}{n_h} \cdot \sum_{1}^{n_h}{}_{i} y_i$$

quindi il totale per strato e poi sommatoria.

- $S_h^2$ la stimiamo invece tramite

$$s^2 = \frac{\sum_{1}^{n} \left( y_i - \overline{y} \right)^2}{n-1}$$

con

$$\overline{y} = \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)$$

ma avremmo potuto usare anche la stima della varianza di indagini precedenti.

- $C = 0,1$ quindi $C^2 = 0,01$

Ora naturalmente avremo quindi due totali di strato

$$\hat{t}_1 = \left(2,5 \cdot 180\right) + \left(2,5 \cdot 150\right) = 825$$

$$\hat{t}_2 = \left(1670\right) + \left(83,5\right) + (250,5) = 2004$$

il totale di tutti e due gli strati è

$$\hat{t}_{HT(STS)} = \sum_{1}^{H}{}_{h} \cdot \frac{N_h}{n_h} \cdot \sum_{1}^{n_h}{}_{i} y_i = 825 + 2004 = 2829$$

ma in ogni caso ora non c'interessa. Avremo due medie di strato che ci servono per calcolare le due varianze di strato

$$\overline{y}_1 = \left( \frac{180 + 150}{2} \right) = 165$$

$$\overline{y}_2 = \left( \frac{150 + 50 + 1000}{3} \right) = 400$$

$$s_1^2 = \frac{(180 - 165)^2 + (150 - 165)^2}{1} = 225 + 225 = 500$$

$$s_2^2 = \frac{(150 - 400)^2 + (50 - 400)^2 + (1000 - 400)^2}{2} = 62500 + 122500 + 360000 = 272500$$

quindi

$$\sum_{h=1}^{H} N_h \cdot S_h^2 = (500 \cdot 5) + (272500 \cdot 5) = 2500 + 1365000 = 1367500$$

Ora non ci rimane altro che calcolare la numerosità, dopo aver definito tutte le nostre variabili

$$a_1 = \frac{5 \cdot 500}{0,01 \cdot 680625 + 1367500} = 0,00182$$

$$a_2 = \frac{5 \cdot 272500}{0,01 \cdot 4016016 + 1367500} = 0,968$$

e quindi dopo il calcolo intermedio poiché

$$\sum_{1}^{2}{}_{h} a_h = 0,970$$

Avremo le nostre agognate numerosità

$$n_1 = \sqrt{0,00182} \cdot 0,970 = 0,041$$

$$n_2 = \sqrt{0,968} \cdot 0,970 = 0,954$$

Ma quante unità prendiamo? Una parte di unità? Attenzione perché vengono numerosità con la virgola dato che il nostro archivio è troppo piccolo, di solito gli archivi contengono molte più unità. Come si può notare il campione di Neyman ci obbliga praticamente a censire la popolazione nello strato meno omogeneo, quello con varianza altissima e a prendere una sola unità nello strato più omogeneo.

*Approfondimento*

Se avessimo provato a calcolare la varianza dello stimatore per il nostro campione pilota avremmo visto che:

$$\hat{V}\left(\hat{t}_{HT,STR}\right) = \sum_{1}^{H}{}_{h} N_h^2 \left[\frac{\left(1-f_h\right)}{n_h}\right] S_{yh}^2$$

$$\hat{V}\left(\hat{t}_{HT,STR}\right) = \left[5^2 \cdot \frac{\left(1-\frac{2}{5}\right)}{2} \cdot 500\right] + \left[5^2 \cdot \frac{\left(1-\frac{3}{5}\right)}{3} \cdot 272500\right] = 3750 + 908333,333 = 912083,333$$

con coefficiente di variazione:

$$CV = \frac{\sqrt{V\left(\hat{t}_{HT}\right)}}{\hat{t}_{HT}} = \frac{\sqrt{912083,333}}{2829} = 0,34$$

Ma questa varianza non ci serve per calcolare le numerosità, in ogni caso se ora ricalcolassimo sulle numerosità ottenute la varianza dello stimatore o il coefficiente di variazione questi ultimi rispetterebbero il vincolo che abbiamo imposto.

***Esercizio .2***

| *Codice* | *r* | *X* | *Y* | *Cod. Strato* |
|---|---|---|---|---|
| A | 0,13 | 200 | 90 | 0 |
| B | 0,54 | 600 | 110 | 0 |
| C | 0,35 | 150 | 70 | 1 |
| D | 0,10 | 140 | 60 | 0 |
| E | 0,25 | 700 | 130 | 1 |
| F | 0,63 | 1000 | 180 | 1 |
| G | 0,71 | 500 | 250 | 0 |
| H | 0,40 | 50 | 30 | 1 |
| I | 0,31 | 150 | 70 | 1 |
| L | 0,90 | 180 | 80 | 0 |

*a) Come selezioniamo il campione?*

Prima di tutto facciamo un campione di cinque unità, stratificato sui codici di strato e p.p.s. sulla x. Generiamo il nostro numero random, prendiamo le unità in base alle nuove probabilità d'inclusione basate sui totali di strato e poi dobbiamo fare stima dell' $\hat{t}_{HT}$ e relativa varianza $\hat{V}_{SPS}\left(\hat{t}_{HT}\right)$.

b) *Come selezioniamo il campione se al secondo stadio facciamo un campionamento casuale semplice? S*tima dell' $\hat{t}_{HT}$ e relativa varianza $\hat{V}_{SPS}\left(\hat{t}_{HT}\right)$ se al secondo stadio facciamo un casuale semplice.

c) Successivamente ipotizziamo che il nostro campione sia un campione pilota stratificato sulla base del quale estraiamo un campione casuale semplice con l'unico fine di calcolarci le varianze per strato e calcoliamo le numerosità di un eventuale campione definitivo per avere una stima della $Y$ con coefficiente di variazione $C(\hat{t}_{HT}) \leq 10\%$ , ovvero ci calcoliamo le numerosità in modo da avere tale coefficiente di variazione, ovvero Neyman con il $C(\hat{t}_{HT})$ a 0,1.

a) *Svolgimento*

La prima cosa da tenere in considerazione è che ora abbiamo definito le numerosità e vogliamo conoscere la varianza dello stimatore $V(\hat{t}_{HT})$. Per estrarre il campione secondo il piano di campionamento *PPS* dobbiamo legare la nostra variabile $Y$ ad un'altra variabile di tipo quantitativo $X$ ,

*totale 1° strato(0)= 1620*

*totale 2° strato(1)= 2050*

| *codice* | *r* | *x* | *y* | *str.* | $\pi_{i,h} = \frac{x_{i,h}}{\sum_{i=1}^{N_h} x_i} \cdot n_h$ *attenzione al denominatore vanno i totali di strato* | $R = \frac{\pi_i}{r}$ | *(ora possiamo riordinare in senso decrescente)* | $W_i = \frac{1}{\pi_i}$ |
|---|---|---|---|---|---|---|---|---|
| A | 0,13 | 200 | 90 | 0 | 0,617 | 4,746 | B(0) | 1 |
| B | 0,54 | 600 | 110 | 0 | 1,852 | *SI* | G(0) | 1 |
| C | 0,35 | 150 | 70 | 1 | 0,366 | 0,68 | A(0) | 1,621 |
| D | 0,10 | 140 | 60 | 0 | 0,432 | 4,32 | …. | …. |
| E | 0,25 | 700 | 130 | 1 | 1,707 | *SI* | …. | …. |
| F | 0,63 | 1000 | 180 | 1 | 2,439 | *SI* | F(1) | 1 |
| G | 0,71 | 500 | 250 | 0 | 1,543 | *SI* | E(1) | 1 |
| H | 0,40 | 50 | 30 | 1 | 0,122 | 0,305 | …. | …. |
| I | 0,31 | 150 | 70 | 1 | 0,073 | 0,235 | …. | …. |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| L | 0,90 | 180 | 80 | 0 | 0,556 | 0,62 | | |

Quindi prendiamo il seguente campione *(B,G,A,F,E).* Questo esercizio è identico a quello precedente, eseguiamo i passaggi senza specificarli.

$$\hat{t}_{HT,1°STR} = (1 \cdot 110) + (1 \cdot 250) + (1,621 \cdot 90) = 505,89$$

$$\hat{t}_{HT,2°STR} = (1 \cdot 180) + (1 \cdot 130) = 310$$

$$\hat{t}_{HT} = (1 \cdot 110) + (1 \cdot 250) + (1,621 \cdot 90) + (1 \cdot 180) + (1 \cdot 130) =$$
$$= 110 + 250 + 145,89 + 180 + 130 = 815,89$$

Poiché sappiamo che la formula della varianza del *PPS* è

$$\hat{V}_{SPS}\left(\hat{t}_{HT}\right) = \frac{1}{n} \cdot \left(\frac{N}{N-1}\right) \cdot \sum_{1}^{n}{}_{i} \left(1 - n \cdot p_i\right) \left(\frac{y_i}{p_i} - y\right)^2 p_i$$

con

$$p_i = \frac{x_i}{\sum_{n}^{1} x_i}$$

che dobbiamo calcolare per ogni strato, quindi nel nostro caso,

$$\hat{V}_{SPS}\left(\hat{t}_{HT}\right) = \frac{1}{n_h} \cdot \left(\frac{N}{N-1}\right) \cdot \sum_{1}^{n_h}{}_{i} \left(1 - n_h \cdot p_i\right) \left(\frac{y_i}{p_i} - \overline{y}\right)^2 p_i$$

con

$$\pi_{i,h} = \frac{x_{i,h}}{\sum_{i=1}^{N_h} x_i} \cdot n_h = n_h \cdot p_i$$

per il *1° strato* avremo

$$\hat{V}_1\left(\hat{t}_{HT}\right) = 0,33\overline{3}\cdot(0,625)\cdot\left[(1-1)\cdot\left(\frac{110}{0,370}-505,89\right)^2\cdot 0,370\right]+$$

$$+0,33\overline{3}\cdot(0,625)\cdot\left[(1-1)\cdot\left(\frac{250}{0,309}-505,89\right)^2\cdot 0,309\right]+$$

$$+0,33\overline{3}\cdot(0,625)\cdot\left[(1-0,617)\cdot\left(\frac{90}{0,123}-505,89\right)^2\cdot 0,123\right]=0+0+3317,701=3317,701$$

per il *2° strato* avremo

$$\hat{V}_2\left(\hat{t}_{HT}\right)=\frac{1}{2}\cdot(0,625)\cdot\left[(1-1)\cdot\left(\frac{\cdots}{\cdots}-310\right)^2\cdot(\ldots)\right]+\frac{1}{2}\cdot(0,625)\cdot\left[(1-1)\cdot\left(\frac{\cdots}{\cdots}-310\right)^2\cdot(\ldots)\right]=0$$

quindi la varianza totale sarà solo quella del primo strato

$$\hat{V}\left(\hat{t}_{HT}\right)=\hat{V}_1(\hat{t}_{HT})+\hat{V}_2(\hat{t}_{HT})=3317,701$$

Non vengono svolti gli altri passaggi poiché sono identici a quelli dell'esercizio n°1, cambiano solo i calcoli, abbiamo praticamente la stessa logica.

*Precisazioni*

Per quanto riguarda poi altri tipi di campionamento come il campionamento a grappoli ed il campionamento a più stadi bisogna tener presente che ogni volta che siamo costretti ad estrarre abbiamo per forza bisogno di numeri random, non possiamo estrarre altrimenti, dunque bisogna solo stare attenti a dove, in un piano di campionamento, incontriamo la casualità legata alla nostra estrazione. Ad esempio nel campionamento a grappoli la incontreremo al primo stadio di estrazione quindi lì si troveranno i numeri random, nel campionamento stratificato come abbiamo visto abbiamo dovuto estrarre e dobbiamo estrarre al secondo stadio, mentre nel campionamento a più stadi o a due stadi

la ritroviamo rispettivamente in tutti gli stadi o al primo o al secondo stadio, quindi avremmo più colonne di numeri random, ognuna in ogni stadio; poi non rimane altro che calcolare i pesi di riporto all'universo.

Riportiamo ora due semplicissimi esercizi, sul campionamento a più stadi ed a grappoli, solo a scopo esemplificativo per fare capire ancor di più la logica che sottintende alle estrazioni e quindi l'applicazione successiva di una stima.

***Esercizio .3***

*(Campionamento a grappoli)*

Con $n_1 = 2$ $N_1 = 4$ $f_1 = 0,5$

| *Cod.Unità* | *Cod.Strato* | *Y* | *r* | *Riordiniamo* | $W_i = \frac{N_1}{n_1}$ |
|---|---|---|---|---|---|
| 1 | 1 | 10 | 0,6 | 0,01 | 2 |
| 2 | 1 | 20 | | | 2 |
| 3 | 1 | 30 | | | 2 |
| 4 | 2 | 40 | 0,01 | 0,3 | 2 |
| 5 | 2 | 50 | | | 2 |
| 6 | 2 | 60 | | | 2 |
| 7 | 2 | 70 | | | 2 |
| 8 | 3 | 80 | 0,3 | 0,5 | 2 |
| 9 | 3 | 90 | | | 2 |
| 10 | 4 | 100 | 0,5 | 0,6 | 2 |
| 11 | 4 | 110 | | | 2 |
| 12 | 4 | 120 | | | 2 |

Quindi la stima del totale sarà $\hat{t}_{HT,GR} = 2 \cdot (40 + 50 + 60 + 70 + 80 + 90) = 780$

e dobbiamo calcolare prima di applicare la formula della varianza.

$$\bar{y}_{GRUPPO(2)} = \left( \frac{40+50+60+70}{4} \right) = 55$$

$$\bar{y}_{GRUPPO(3)} = \left( \frac{40+50+60+70}{4} \right) = 85$$

$$\bar{y}_{TOTALE} = \left( \frac{40+50+60+70+80+90}{6} \right) = 65$$

$$S^2_{1,y} = \left( \frac{\left(\bar{y}_{GRUPPO(2)} - \bar{y}_{TOTALE}\right)^2 + \left(\bar{y}_{GRUPPO(3)} - \bar{y}_{TOTALE}\right)^2}{1} \right)$$

***Esercizio .4***

*(Campionamento a più stadi)*

In questo esempio mostriamo solo come ci verrebbe proposto un esercizio su questo tipo di campionamento. Dovremo come già specificato nella parte relativa a questo piano di campionamento estrarre prima un campione di gruppi e poi dai gruppi dei campioni di unità, non osserviamo tutte le unità come nel campionamento a grappoli. (In questo esempio abbiamo impostato uno stesso numero di unità per tutti i gruppi e per ogni gruppo abbiamo preso sempre lo stesso numero di unità, ma il numero di unità campionate nei gruppi può cambiare, lo si vede bene dalla formula della varianza.

| *Cod.Unità* | *Cod.Gruppo* | $r_1$ | *Y* | $r_i$ | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 00,6 | 10 | 0,37 | | |
| 2 | | | 20 | 0,40 | | |
| 3 | | | 30 | 0,21 | | |
| 4 | 2 | 0,01 | 40 | 0,08 | | |
| 5 | | | 50 | 0,56 | | |
| 6 | | | 60 | 0,32 | | |
| 7 | 3 | 0,2 | 70 | 0,50 | | |
| 8 | | | 80 | 0,01 | | |
| 9 | | | 90 | 0,03 | | |
| 10 | 4 | 0,4 | 100 | 0,4 | | |
| 11 | | | 110 | 0,18 | | |
| 12 | | | 120 | 0,20 | | |

Con

$n_1 = 2 \qquad N_1 = 4 \qquad N_i = 3 \qquad n_i = 2$

E dallo stimatore del totale possiamo vedere come otteniamo i pesi di riporto dal primo stadio e dal secondo stadio, li moltiplichiamo tra di loro;

$$\hat{t}_{HT,2S} = \frac{N_1}{n_1} \cdot \sum_{S_1} \frac{N_i}{n_i} \cdot y_i$$

Naturalmente in questo esercizio dobbiamo prima riordinare i numeri random per poi estrarre i gruppi e le unità, quindi formalmente non va bene come abbiamo fatto nella tabella, ovvero annerire le caselle dei gruppi scelti, e delle unità estratte, prima di aver riordinato.

***Esercizio .5***

*(Campionamento a 2 Stadi)*

| *Cod.Unità* | *Cod.Gruppo* | $r_1$ | *Y* | $r_i$ | *Riordiniamo i gruppi* | *Riordiniamo le unità* | $W_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,01 | 20 | 0,01 | 0,01 | 0,01 | 1,5 |
| 2 | | | 31 | 0,6 | | 0,32 | 1,5 |
| 3 | | | 72 | 0,32 | | 0,6 | 1,5 |
| 4 | 2 | 00,6 | 58 | 0,7 | 0,06 | 0,2 | 1,5 |
| 5 | | | 15 | 0,3 | | 0,3 | 1,5 |
| 6 | | | 16 | 0,2 | | 0,7 | 1,5 |
| 7 | 3 | 0,7 | 26 | 0,5 | 0,3 | | |
| 8 | | | 17 | 0,25 | | | |
| 9 | | | 18 | 0,27 | | | |
| 10 | 4 | 0,3 | 11 | 0,31 | 0,7 | | |
| 11 | | | 48 | 0,05 | | | |
| 12 | | | 25 | 0,3 | | | |

Con $n_1 = 2$ $N_1 = 4$ $N_i = 3$ $n_i = 2$

$$\hat{t}_{HT,2S} = \frac{N_1}{n_1} \cdot \sum_{S_1} \frac{N_i}{n_i} \cdot y_i = 0,5 \cdot 1,5 \cdot (20 + 72 + 16 + 15) = 92,5$$

con medie tra le unità

$$\overline{y}_{i,1} = \frac{(20 + 72)}{2} = 46$$

$$\overline{y}_{i,2} = \frac{(16 + 15)}{2} = 15,5$$

con varianze

$$S^2_{i,1} = \frac{(20 - 46)^2 + (72 - 46)^2}{1} = 1352$$

$$S^2_{i,1} = \frac{(0,25) + (0,25)}{1} = 0,5$$

con media tra tutte le unità

$$\overline{y} = \left( \frac{20 + 72 + 16 + 15}{4} \right) = 30,75$$

con varianza tra i gruppi

$$S^2_{1,y} = (30,75 - 46)^2 + (30,75 - 46)^2 = 465,25$$

e quindi la varianza dello stimatore è

$$\hat{V}_{2S}(\hat{t}_{HT}) = 4^2 \cdot \frac{(1 - 0,5)}{2} \cdot 465,25 + 2 \cdot$$
$$\cdot \left[ 3^2 \cdot \frac{(1 - 0,667)}{2} \cdot 1352 + 3^2 \cdot \frac{(1 - 0,667)}{2} \cdot 0,5 \right] = 3887,221$$

***Esercizio .6***

*(Campionamento sistematico)*

| *COD.* | *Y* |
|---|---|

| | |
|---|---|
| 1 | 10 |
| 2 | 50 |
| 3 | 25 |
| 4 | 70 |
| 5 | 32 |
| 6 | 83 |
| 7 | 58 |
| 8 | 70 |
| 9 | 25 |
| 10 | 30 |

$N = 10$

$n = 2$

$Passo \quad \frac{N}{n} = k = 5$

$\pi_i = \frac{1}{k} = \frac{n}{N} = \frac{1}{5}$

$\delta = 0,35$

Prima effettuiamo l'estrazione di un numero casuale compreso tra 1 e k e supponiamo che ci venga 3, quindi partiamo dalla terza unità, delle nostre unità allora prenderemo la terza e l'ottava. Quindi stimiamo il totale

$$\hat{t}_{HT,SIST} = k \cdot \sum_{1}^{n} {}_i\, y_i = (5 \cdot 25) + (5 \cdot 70) = 475$$

la varianza

$$V(\hat{t}_{HT,SIST}) = \frac{N^2 \cdot S_y^2}{n} \cdot \left[ (1 - f) + (n - 1) \cdot \delta \right]$$

prima

$$\overline{y}_y = \frac{(25 + 70)}{2} = 47,5$$

$$S_y^2 = \frac{(25 - 47,5)^2 + (70 - 47,5)^2}{1} = 1012,5$$

$$V(\hat{t}_{HT,SIST}) = \frac{10^2 \cdot 1012,5}{2} \cdot \left[ (1 - 0,2) + (1) \cdot 0,35 \right] = 58218,75$$

### 13.3 Indagini di mercato

Un'impresa ha realizzato un nuovo prodotto da lanciare sul mercato (omogeneizzati) e vuole conoscere l'ampiezza del mercato a cui è diretto il prodotto, in pratica il consumo

di prodotti della stessa categoria; dal momento che i costi per il lancio di un nuovo prodotto sono elevati vuole conoscere il mercato di riferimento e si rivolge alla sua agenzia che si occupa di indagine di mercato e le commissiona quest'indagine;

*Descrivi:*

1. *La popolazione di riferimento poiché l'indagine è a carattere nazionale;*
2. *L'unità statistica e/o l'unità di rilevazione;*
3. *Periodo di Riferimento;*
4. *Costruzione dell'Archivio;*
5. *La tipologia di indagine, nel senso se è campionaria o censuaria, Panel etc.;*
6. *Modalità di somministrazione del questionario;*
7. *Il piano di campionamento se è un'indagine campionaria;*
8. *Alcune domande di esempio del questionario;*

Naturalmente motivare brevemente ciascuna delle risposte; brevemente poiché per ciascuna risposta si potrebbe scrivere un tema, a noi interessa la sintesi, 3/4 righe per ciascuna argomentazione bastano ed avanzano.

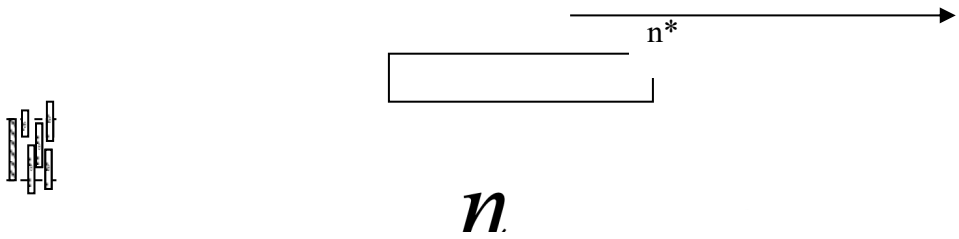
*Svolgimento*

Ovviamente il nostro obiettivo è quello di tradurre i bisogni, le domande del direttore d'azienda in obiettivi di tipo operativo, fare in modo insomma che alla fine la nostra indagine conduca a delle risposte utili all'imprenditore.

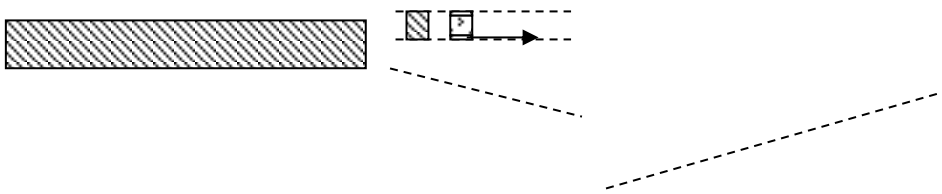
- ✓ *Definizioni ed obiettivi*

Prima di tutto dobbiamo capire cosa vuol dire *misurare il mercato di riferimento*. Il direttore d'azienda volle sapere il numero di potenziali clienti, e ciascun cliente possibilmente pesato per consumi del suo possibile prodotto, in questo caso parliamo di omogeneizzati, quindi per mercato di riferimento intendiamo i *consumi delle famiglie per omogeneizzati;* ovviamente il direttore dell'azienda vorrà conoscere i consumi *per zona territoriale* in modo da sapere dove il prodotto sarà più o meno consumato e magari progettare la propria campagna pubblicitaria in modo più appropriato;

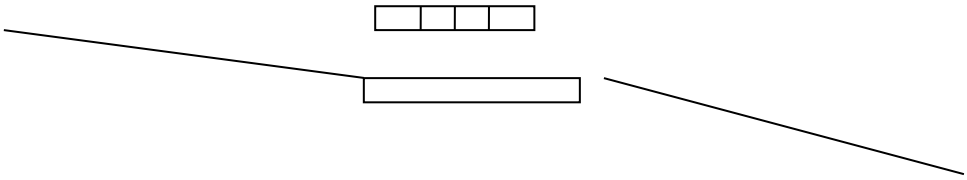
probabilmente l'imprenditore li vorrà conoscere anche *per tipologia familiare*, magari nella pubblicità rivolta a famiglie ricche o povere si comporterà in un modo, con famiglie grandi o famiglie piccole si comporterà in un modo diverso; quindi per tipologia familiare intendiamo una classificazione sia *dimensionale* sia dal punto di vista della *classe sociale,* in termini di condizione professionale del capofamiglia o altri aspetti del genere.




La *popolazione delle unità statistiche* è costituita naturalmente da coloro che consumano omogeneizzati quindi i neonati, che costituiscono la nostra unità di analisi in quanto sono i consumatori del nostro prodotto; d'altro canto non possiamo somministrare il questionario ad un neonato, quindi la nostra unità di rilevazione è la famiglia, e quindi la popolazione di riferimento è costituita dalle *famiglie che hanno avuto almeno una nascita negli ultimi 5/6 mesi*, e questo collegamento possiamo effettuarlo facilmente ricercando nelle anagrafi. Insomma abbiamo costruito la nostra popolazione di riferimento basandoci su un'unità complessa, la nostra popolazione di riferimento è costituita da tutte le unità complesse (famiglie che hanno avuto una nascita negli ultimi 5/6 mesi). Inoltre prendiamo le nascite degli ultimi 5/6 mesi in quanto gli omogeneizzati cominciano ad essere consumati a quest'età. Quindi come prima soluzione la nostra unità statistica sarà un'unità complessa.

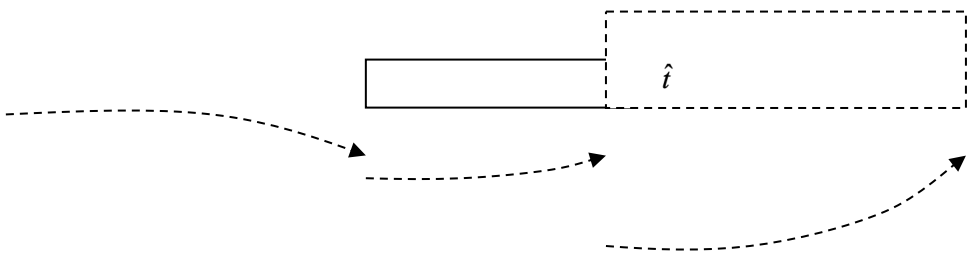


Oppure come seconda alternativa potremmo prendere i rivenditori.



In questo secondo caso dobbiamo analizzare le vendite degli esercizi commerciali, dunque i consumi aggregati, quindi non potremo mai sapere il consumo per neonato; il difetto fondamentale di questo metodo è non poter conoscere il consumo per tipologia famigliare ma si velocizza la rilevazione in quanto i rivenditori sono in numero inferiore rispetto alle famiglie che hanno avuto un bambino negli ultimi *5/6* mesi e la qualità del dato è maggiore, ma non possiamo conoscere il consumo per neonato; per risolvere il problema per cui non siamo in grado di individuare la tipologia famigliare, possiamo chiedere alla rivendita di somministrare il questionario agli acquirenti dell'omogeneizzato che poi sarà rispedito a noi.

Abbiamo, ancora, una terza alternativa: considerare il caso in cui la popolazione delle unità statistiche è costituita naturalmente dalle *famiglie che hanno avuto almeno una nascita negli ultimi 5/6 mesi sotto la supervisione di un pediatra,* insomma abbiamo costruito la nostra popolazione di riferimento basandoci nuovamente su un'unità complessa, la nostra popolazione di riferimento è costituita da tutte le unità complesse, ovvero le famiglie che hanno avuto una nascita negli ultimi 5/6 mesi sotto la supervisione di un pediatra.



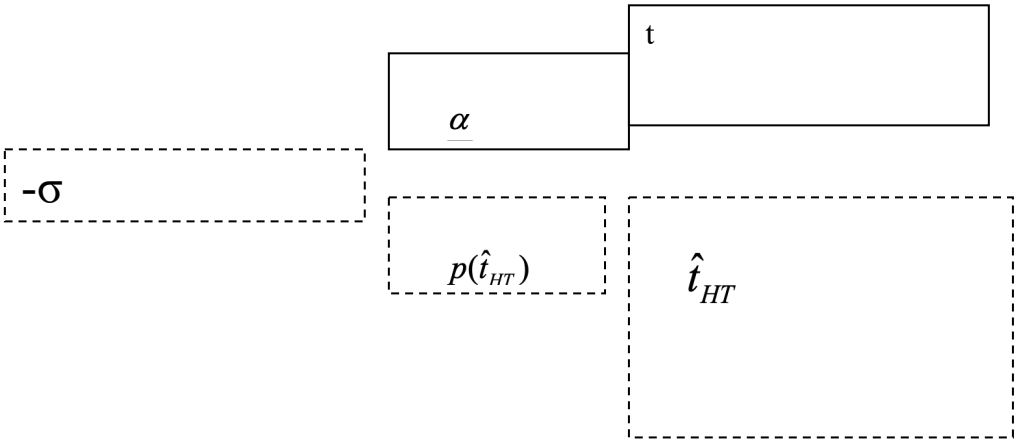

In questo terzo caso il pediatra può dirci anche qualcosa sulla tipologia familiare a differenza della seconda alternativa, inoltre è lui a prescrivere gli omogeneizzati; per

quanto riguarda i difetti invece potremmo avere sottocopertura, in quanto non tutti vanno dal pediatra. Inoltre avremo un alto tasso di mancata risposta e la qualità del dato sarà bassissima. Il problema della mancata risposta si può risolvere utilizzando i pediatri come unità di rilevazione di primo stadio, dare i questionari e dire di somministrarli alle famiglie in modo da non importunare il pediatra per la rilevazione, quindi avremo due unità di rilevazione, il pediatra di primo stadio e la famiglia di secondo stadio.

Probabilmente la soluzione migliore è quella che riguarda le rivendite in quanto abbiamo sia un dato sui consumi aggregati delle rivendite e sia i dati riguardanti le famiglie nel caso vogliamo informazioni sulla tipologia famigliare dicendo ai rivenditori di somministrare il questionario e poi rispedircelo, insomma dobbiamo prendere i pregi ed evitare i difetti delle varie alternative. Tra l'altro se ci pensiamo bene questa è la metodologia che spesso si usa nelle indagini di mercato di questo genere, in cui per sollecitare la risposta e quindi spedire il questionario spesso si usa distribuire anche gadget, viaggi, premi, etc.

✓ *Il periodo di riferimento*

Per quanto riguarda il periodo di riferimento dobbiamo prendere due decisioni di carattere temporale, per quanto tempo portare avanti l'indagine ed il periodo relativo alla definizione della popolazione, ovvero sarebbe opportuno consultare un esperto è chiedere per quanto tempo i bambini mangiano omogeneizzati;




quindi supponendo che i bambini mangiano omogeneizzati per un anno dobbiamo prendere tutte le famiglie che hanno avuto bambini negli ultimi *12* mesi mentre per quanto riguarda la durata dell'indagine sappiamo che più o meno tutte le variabili e tutti

fenomeni con cui abbiamo a che fare quotidianamente sono su base annuale e quindi per correlare la stima derivante dalla nostra indagine con altri fenomeni sembra opportuno farla durare per *12* mesi, quindi una cosa è la durata dell'indagine ed un'altra la definizione del periodo di riferimento per la definizione della popolazione.

✓ *Costruzione dell'archivio*

Naturalmente se abbiamo scelto come definizione di unità statistica la prima alternativa per costruire l'archivio basta andare all'anagrafe e indagare tutte le famiglie che hanno avuto un bambino negli ultimi *12* mesi; l'unico problema che potrebbe sorgere è quello dei doppioni, e comunque non costituisce un grave problema dal un punto di vista statistico, rischieremo solo di mandare lo stesso questionario alla stessa famiglia, ma cerchiamo di evitarlo  per non fare una figura barbina.

✓ *Tipo d'indagine*

Se abbiamo utilizzato la prima alternativa non possiamo utilizzare un'indagine panel in quanto l'unità di rilevazione, la famiglia esce dalla popolazione. Poiché nascono circa *200/250* mila bambini l'anno in Italia l'indagine si potrebbe anche fare *censuaria* quindi i costi sono affrontabili; l'unico problema che si crea in questo caso è che ci affidiamo totalmente alle mancate risposte, è già tanto se ci risponde un 40%, faremmo meglio a fare l'indagine *campionaria* e dedicarci maggiormente all'insistenza nel ricevere una risposta, in modo da avere una migliore qualità del dato. Il campione potrebbe risultare un po' distorto poiché la *mancata risposta è auto selettiva* in quanto di solito chi non risponde appartiene spesso alla stessa categoria di soggetti. Inoltre sarebbe opportuno selezionare un campione di famiglie e poi ruotarle mese per mese in modo da evitare *molestia statistica*, in pratica dividiamo il nostro campione per *12* mesi in modo casuale e poi ogni parte la intervistiamo in un mese; se ci accontentiamo di una qualità del dato minore non c'è bisogno di fare ciò, ad esempio nell'indagine sui bilanci di famiglia chiedevamo in modo dettagliato gli acquisti, i consumi, e quindi in quel caso sarebbe stato opportuno ruotare il campione, nel caso invece ci fossimo accontentati ad esempio della spesa complessiva, "*signora quanto spende in una settimana?*", e non il consumo per scontrino non sarebbe stato necessario ruotarle. In ogni caso per un solo prodotto di solito ruotare il campione non è necessario, la molestia non è alta. la rotazione del

campione dopo averlo divisa in *12* parti in modo casuale è necessaria poiché dobbiamo evitare stagionalità, in quando di solito i bambini nascono maggiormente in determinati periodi dell'anno. Ecco perché si usa spesso l'esempio delle cicogne che portano i bambini, non perché i bambini sono effettivamente portati dalle cicogne ma perché c'è una stagionalità delle nascite correlata al passaggio delle cicogne.

✓ *Modalità di somministrazione del questionario*

Naturalmente la modalità di somministrazione del questionario dipenderà dalle scelte fatte precedentemente; se passiamo per i rivenditori naturalmente questa andrà fatta come un'indagine postale, mentre se passiamo direttamente per le famiglie potrebbe essere un'indagine telefonica. Il fatto è che poiché abbiamo utilizzato come fonte dell'archivio "anagrafe" la somministrazione telefonica è complessa dato che sull'anagrafe il numero di telefono non esiste. Oppure potremmo utilizzare un'indagine faccia a faccia visto che sull'anagrafe abbiamo l'indirizzo e mandiamo un rilevatore anche se ci costa di più. Oppure per far costare meno l'indagine faccia a faccia potremmo direttamente mettere un rilevatore nei punti vendita di omogeneizzati in modo da coprire una maggiore utenza con costi minori. Di quest'ultima possibilità uno dei difetti possibili è che la persona che sta facendo la spesa non sa nulla riguardo al consumo familiare di omogeneizzati ma è stata demandata a comprarli; ma il problema più grosso di questo metodo che tra l'altro è il più utilizzato dalle ditte che fanno indagini di mercato è che *l'effetto rilevatore è rilevantissimo* (donne affascinanti oppure si prendono solo le prime unità disponibili), quindi se il rilevatore non è fidato, l'indagine potrebbe essere seriamente compromessa. Tornando all'indagine telefonica l'aggancio tra l'archivio Telecom con l'anagrafe quasi mai coincide, se si riesce a trovare il *10% - 20%* di aggancio è già molto. Quindi, per concludere o utilizziamo un'indagine postale con un basso grado di risposta, oppure un'indagine faccia a faccia stando attenti alla preparazione dei rilevatori quindi con un costo maggiore.

✓ *Piano di campionamento*

Dobbiamo essere in questo caso coerenti con gli obiettivi e le definizioni dell'indagine dopo aver definito i consumi per regione e per tipologia famigliare, poiché di solito definire i consumi per regione o per tipologia famigliare significa stabilire i domini di

stima, cioè gli aggregati all'interno dei quali effettuare una stima che sia coerente con essi, e di solito esiste una corrispondenza biunivoca tra domini di stima e variabile di stratificazione, poiché stratificare è l'unico metodo che ci garantisce di avere una stima per ogni dominio di stima; ad esempio se vogliamo fare le stime di addetti per regione, dove il dominio di stima è la regione, l'unico modo che abbiamo di essere sicuri di avere una stima attendibile per ogni regione e quindi di avere un tot. d'imprese per ogni regione è stratificare in base alle regioni. Quindi dobbiamo analizzare due casi, quello in cui mandiamo i questionari per posta alle famiglie e quello faccia a faccia presso i rivenditori; *nel primo caso* quindi poiché abbiamo due obiettivi, tipologia famigliare e regioni, stratifichiamo in base a queste variabili; certamente non utilizzeremo un *PPS* per reddito famigliare in quanto il consumo di omogeneizzati è più o meno fisso, è possibile che una famiglia acquista omogeneizzati con un prezzo più basso ma comunque il quantitativo rimane più o meno lo stesso; se proprio dobbiamo utilizzare un *PPS* è meglio farlo con probabilità proporzionale in base al numero dei bambini ma in ogni caso non ha senso in quanto le nascite sono in genere di un bambino per la famiglia nell'arco dell'anno di riferimento, a meno che non si tratti di parti gemellari.
Per quanto riguarda *la numerosità* sarebbe opportuno prendere una numerosità per ogni strato proporzionale alla numerosità totale dello strato ossia $n_h \sim N_h$ che indica la proporzionalità. *Nel secondo caso* quello in cui utilizziamo i rivenditori dopo essere andati alla camera di commercio ed aver selezionato tutti i rivenditori, al primo stadio *stratifichiamo per regione* e poi facciamo un *PPS* sul fatturato oppure sugli addetti, in modo da distinguere i rivenditori grandi dai piccoli o un'ulteriore stratificazione per classi di fatturato o di addetti. Al secondo stadio, dove ci sono tutte le famiglie che vanno ad acquistare e che costituiscono una popolazione esageratamente numerosa in capo ad un anno, abbiamo il forte effetto rilevatore, dato che le unità non siamo in grado di estrarle noi, utilizziamo un sistematico e demandiamo al rilevatore il compito di effettuare l'estrazione.

- ✓ *Il questionario*
- ➢ *Prima sezione:*

Riguarderà le caratteristiche generali della famiglia, posizione professionale etc., le caratteristiche del neonato, maschio o femmina, età, etc., questo per discriminare l'uso di omogeneizzati a seconda dell'età.

- *Seconda sezione:*

Per evitare la molestia statistica non registriamo gli scontriamo, però dobbiamo tenere conto del fatto che ci potrebbe essere l'effetto memoria e quindi dobbiamo decidere il lasso di tempo del quale vogliamo chiedere il consumo, ad esempio quanti omogeneizzati si comprano al mese. Poi chiedere che tipi di omogeneizzati si comprano, perché magari di omogeneizzati si differenziano per tipologia di composizione chimica, ingredienti, etc.

- *Terza sezione:*

Dopo aver *esaurito l'obiettivo* dell'indagine nella seconda sezione, possiamo passare a chiedere qualche informazione aggiuntiva; ad esempio se il nostro prodotto è diretto ai bambini allergici, sapere il tipo di omogeneizzato acquistato sarà un'informazione molto rilevante, oppure sapendo che la presenza dei musulmani incide sempre di più sulla popolazione, il dato sulla fede religiosa sarà molto importante in quanto alcuni omogeneizzati nella loro composizione prevedono la presenza di carne di suino di cui la fede musulmana vieta il consumo. Solitamente, quando si eseguono indagini sulla soddisfazione del consumatore, che qui non c'interessa, per sapere se si consumano anche altri prodotti,una cosa fondamentale è non dire per quale azienda si lavora altrimenti la risposta sarà di bassa qualità, magari sarebbe più opportuno chiedere "quali altre marche di omogeneizzati conosce?"

*Domanda di una studentessa (caso limite)*

Potrebbe succedere che una famiglia che ha avuto un bambino quando abbiamo iniziato l'indagine, destinata ad essere intervistata in uno dei gruppi che avevamo estratto a caso per ogni mese, dopo 12 mesi ovvero al dodicesimo mese sia uscita dalla nostra popolazione? Si è possibile, in ogni caso noi, è vero che effettuiamo l'estrazione dei gruppi da allocare in ogni mese in modo casuale ma comunque possiamo tenere sotto controllo la nostra popolazione, siamo perfettamente in grado di controllare la selezione