Capitolo 3

Campionamento

In questo capitolo studieremo come ottenere un campione aleatorio estratto da una popolazione obiettivo finita di numerosità N. Con piano o disegno di campionamento si intende il procedimento utilizzato per costruire il campione partendo da una popolazione finita o infinita. Del disegno di campionamento possiamo evidenziare

- la struttura del campione (liste, sottoliste, attributi, strati);
- le regole seguite per identificare gli insiemi di unità da inserire nel campione;
- la probabilità di inclusione delle singole unità;
- il modo con cui si determina la numerosità ottima del campione e la relativa frazione di campionamento. "... la numerosità ottima di un campione è quella che permette di ottenere gli obiettivi dell'indagine al minimo costo, e sarà il più piccolo numero in base al quale le stime ragiungono il livello di attendibilità atteso dal ricercatore" (Fabbris, p. 26).

La conoscenza della tecnica di campionamento utilizzata è essenziale ai fini della determinazione delle proprietà probabilistiche (distribuzione congiunta) del campione nonché della validità e correttezza dei risultati. Un piano di campionamento può infatti essere

- probabilistico, in tal caso le unità vengono estratte secondo un meccanismo aleatorio;
- deterministico o non probabilistico, le unità sono scelte dalla popolazione tramite una regola *deterministica*.

Un esempio di piano di campionamento deterministico è il seguente: dall'elenco telefonico estraggo, per ciascuna lettera dell'alfabeto, i primi 10 individui. Se da un lato risulta evidente la sua semplicità di implementazione, il metodo deterministico di campionamento presenta alcuni svantaggi non indifferenti. Esso non permette di calcolare il grado di precisione con cui viene eseguita la stima. Inoltre, la validità dei risultati ottenuti è fortemente legata alle informazioni utilizzate a priori per la scelta del campione. Se tali informazioni sono sbagliate oppure obsolete e non corrispondono più alla realtà, i risultati della ricerca saranno distorti. Per tale motivo in questo corso tratteremo esclusivamente piani di campionamento basati sulla selezione casuale delle unità (individui) della popolazione obiettivo. In particolare vedremo le tecniche di campionamento casuale semplice, di campionamento stratificato, di campionamento sistematico. Tutti questi tipi di campionamento si basano su delle tecniche di selezione.

3.1 Campionamento tramite selezione con reinserimento

Supponiamo di avere quale popolazione obiettivo l'insieme delle automobili immatricolate in Svizzera ad una certa data. La numerosità della popolazione è indicata con la lettera N. Una casa automobilistica potrebbe essere interessata alla distribuzione del colore dell'automobile così da poi prevedere il consumo di vernice. Per semplicità supponiamo che i colori siano solo tre: nero, bianco e rosso e che la distribuzione sia tale per cui il 45% delle auto immatricolate è nero, il 25% è bianco ed il 30% è rosso. Definiamo ora il seguente esperimento aleatorio definito sull'insieme $\Omega = \{1, 2, ...N\}$ che consiste nel scegliere a caso in maniera *equiprobabile* un numero compreso tra 1 e N. Lo spazio di probabilità è dunque composto dalla solita tripla $(\Omega, \mathcal{E}, P)$ dove in questo caso P è la distribuzione uniforme su Ω

$$P(\{i\}) = \frac{1}{N} \text{ per ogni } i \in \Omega.$$

Supponiamo di avere una scheda per ogni automobile immatricolata. Nella scheda sono riportati il propritario ed i dati tecnici del veicolo: colore, cilindrata, anno d'immatricolazione, ecc.. Numeriamo le schede da 1 a N. Ad ogni automobile immatricolata in Svizzera sarà così assegnato un numero $i \in \Omega$. Definiamo ora la variabile aleatoria (funzione!) X da Ω verso l'insieme dei colori $\Omega' = \{$"nero","bianco","rosso"$\}$ in modo tale che per ogni $i \in \Omega$ la $V.A.$ X restituisca il colore della i-esima automobile

$$i \mapsto X(i) = \text{"il colore della } i\text{-esima automobile"} .$$

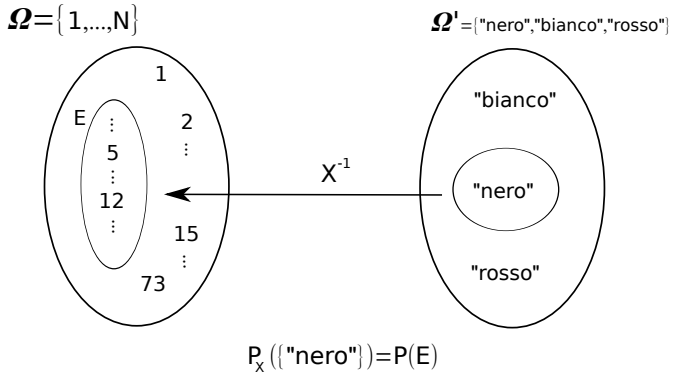

Figura 3.1: Estrazione colore automobile

Domanda 11. Qual è la legge di probabilità indotta dalla variabile aleatoria X, notata P_X, sull'insieme dei tre colori?

Risposta: la legge di probabilità P_X non è altro che la distribuzione statistica P_p della caratteristica in esame della popolazione obiettivo. Infatti, sia N_1 il numero di automobili di colore nero. Sappiamo che $\frac{N_1}{N} = 0.45$ è la frequenza relativa di automobili nere nella popolazione. Qual è la probabilità di selezionare una macchina nera? Questa probabilità corrisponde a

$$P_X(\{\text{“nero”}\}) = P(X = \text{“nero”}) = P(E)$$

dove E corrisponde agli indici delle schede delle automobili di colore nero. Il numero di esiti in E è evidentemente N_1. Come calcolato precedentemente la probabilità di ciascun esito è $\frac{1}{N}$ e quindi $P(E) = \frac{N_1}{N} = 0.45$. La situazione è rappresentata graficamente nella Figura 3.1.

La legge di probabilità P_X corrisponde esattamente alla distribuzione statistica P_p del colore della popolazione obiettivo.

Se la caratteristica in esame fosse stata la cilindrata in cm^3 dell'automobile anziché il suo colore, la variabile aleatoria X avrebbe una funzione di ripartizione F_X. Ancora una volta la situazione è riportata graficamente nella Figura 3.2.

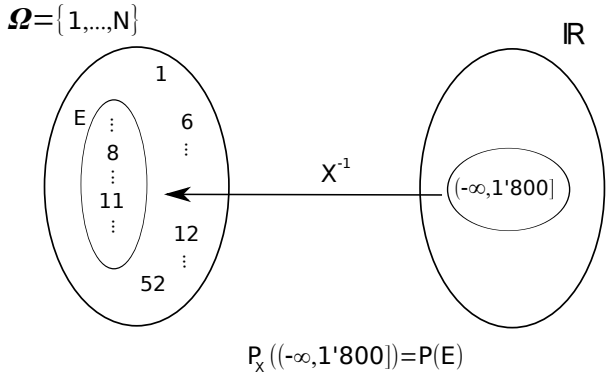

Figura 3.2: Estrazione cilindrata automobile

$$\begin{aligned} F_X(x) &= P_X\left((-\infty, x]\right) \\ &= P(X \in (-\infty, x])) \\ &= P(E) \\ &= (\#\text{ unità con caratteristica } \leq x)\frac{1}{N} \\ &= \frac{\#\text{ unità con caratteristica } \leq x}{\#\text{ unità della popolazione}} \end{aligned} \tag{3.1}$$

Come si evince dalla (3.1) la funzione di ripartizione della variabile aleatoria X è identica alla funzione di ripartizione della cilindrata della popolazione obiettivo.

È dunque possibile, partendo da un semplice spazio di probabilità discreto la cui legge di probabilità è uniforme, costruire una variabile aleatoria X la cui distribuzione P_X è uguale alla distribuzione dell'intera popolazione obiettivo.

$$(\Omega, \mathcal{E}, P) \underset{X}{\longrightarrow} (\mathbb{R}, \mathcal{E}, P_X)$$

La problematica della non conoscenza della distribuzione statistica P_p della popolazione obiettivo risulta quindi identica alla problematica della stima della legge di probabilità di una variabile aleatoria!

Ripetiamo l'esperimento n volte in maniera indipendente l'una dall'altra, reinserendo nell'urna il numero scelto dopo ogni estrazione e mescolando

nuovamente molto bene. Il nuovo spazio campionario è $\Omega^n = \{1, 2, ...N\}^n$, le liste di lunghezza n i cui elementi appartengono all'insieme $\Omega = \{1, 2, ...N\}$. Il numero di elementi di Ω^n è N^n. Poiché il numero selezionato ad ogni estrazione è reinserito nell'urna, è possibile osservare vettori in cui un numero si ripete più volte. Ad esempio, se la popolazione contasse $N = 5$ elementi e $n = 3$ (sorteggiamo 3 volte) un possibile esito potrebbe essere $\omega = (2, 5, 2)$. Qual è la probabilità di osservare un qualsiasi esito[1] $\omega \in \Omega^n$? Poiché le estrazioni sono indipendenti l'una dall'altra e ad ogni estrazione la probabilità di estrarre un qualsiasi numero è pari a $1/N$ vale

$$Q(\omega) = \frac{1}{N^n} \text{ per ogni } \omega \in \Omega^n \ . \tag{3.2}$$

Ecco così definito lo spazio di probabilità $\left(\Omega^n, \overset{n}{\underset{i=1}{\otimes}} \mathcal{E}_i, Q = \overset{n}{\underset{i=1}{\otimes}} P_i\right)$ relativo al nostro nuovo esperimento. Possiamo ora definire le n variabili aleatorie che notiamo con X_i, $i = 1, ..., n$ tramite la procedura

$$X_i(\omega) := \begin{cases} 1. \text{ Estraggo l' } i-\text{esimo elemento da } \omega, \text{ notato } \omega(i) \in \{1, 2, ...N\}. \\ 2. \text{ Associo la cilindrata della } \omega(i)-\text{esima automobile.} \end{cases}$$

Proseguendo con l'esempio in cui $N = 5$ e $n = 3$, data la realizzazione $\omega = (2, 5, 2)$ abbiamo $\omega(1) = 2$, $\omega(2) = 5$, $\omega(3) = 2$ (praticamente abbiamo selezionato due volte la seconda automobile ed una volta la quinta). Il valore che la *V.A.* X_i assume dipende *esclusivamente* dall'esito parziale della i-esima estrazione. Poiché le n estrazioni sono per costruzione indipendenti anche le variabili aleatorie X_i lo saranno e la legge di probabilità $Q_{X_1,...,X_n}$ indotta da $X_1, ..., X_n$ su $\mathbb{R}^n$ è semplicemente il prodotto delle singole leggi di probabilità P_{X_i}. L'esperimento aleatorio

$$(\Omega, \mathcal{E}, P) \underset{X}{\longrightarrow} (\mathbb{R}, \mathcal{E}, P_X)$$

viene così generalizzato in un esperimento statistico

$$(\Omega^n, \overset{n}{\underset{i=1}{\otimes}} \mathcal{E}_i, Q) \underset{X_1,...,X_n}{\longrightarrow} (\mathbb{R}^n, \overset{n}{\underset{i=1}{\otimes}} \mathcal{E}_i, Q_{X_1,...,X_n}) \ .$$

[1] Ora ω denota un qualsiasi elemento di Ω^n ed avrà quindi la forma

$$\omega = \underbrace{(5, 12, ..., 4)}_{n \text{ numeri tra 1 e } N}$$

di una lista a n componenti.

Quando la distribuzione statistica P_p della popolazione obiettivo è sconosciuta possiamo pensare di stimare P_X andando a selezionare da una particolare famiglia parametrica di distribuzioni $\mathcal{P}$ la distribuzione che riteniamo migliore rispetto all'evidenza empirica (le n osservazioni) in nostro possesso. È quindi possibile ricondurre il problema della stima della distribuzione P_p della popolazione obiettivo alla definizione di esperimento statistico e di n-campione ad esso associato. In altre parole, le n variabili aleatorie $i.i.d.$ $X_1, ..., X_n$ costruite estraendo a caso con reimmissione n unità dalla popolazione obiettivo costituiscono un n-campione di un esperimento aleatorio $(\mathbb{R}, \mathcal{E}, P_X)$. Poiché la legge di probabilità P_X è sconosciuta, scriveremo $(\mathbb{R}, \mathcal{E}, \mathcal{P})$ dove $\mathcal{P}$ rappresenta la famiglia parametrica a cui P_X appartiene.

La tecnica di selezione casuale con reinserimento utilizzata precedentemente per la costruzione di un n-campione a partire da una data popolazione obiettivo di numerosità N è riassunta nel modo seguente.

1. A ciascuna delle N unità della popolazione viene assegnato univocamente un numero da 1 a N.

2. Si estrae a caso una sola pallina da un'urna contenente N palline numerate da 1 a N. Il numero indicato sulla pallina corrisponde all'unità del campione da selezionare per l'osservazione della caratteristica o attributo in esame.

3. Si reintroduce la pallina estratta e si mescola nuovamente.

4. Si ripetono i punti 2. e 3. per un totale di n volte.

Le n osservazioni, raccolte in una lista $(x_1, ..., x_n)$, costituiscono una realizzazione delle n variabili aleatorie $i.i.d.$ $X_1, ..., X_n$ distribuite secondo la legge di probabilità definita dalla distribuzione della popolazione obiettivo.

3.2 Campionamento tramite selezione senza reinserimento

Questa tecnica è identica alla precedente con l'unica differenza che le palline estratte non vengono reinserite nell'urna. Cosa cambia rispetto al paragrafo precedente? Lo spazio campionario è lo stesso? Possiamo ancora parlare di esperimento statistico? Quali sono le proprietà della nuova legge di probabilità Q? Per rispondere a queste domande partiamo innanzi tutto dalla definizione del nuovo spazio campionario. Supponiamo per un momento di utilizzare ancora Ω^n. Poiché la tecnica di estrazione è senza reinserimento

lo spazio Ω^n conterrà esiti che non osserveremo mai. Ad esempio, prendiamo $N = 5$, $n = 2$ e l'esito $\omega = (2, 2)$. Ovviamente un simile ω non sarà mai osservabile in quanto la selezione ora è senza reinserimento. In teoria potremmo escluderlo dallo spazio campionario. Tuttavia per semplicità nella definizione del nostro spazio campionario lo lasciamo. L'esito $\omega = (2, 2)$ avrà semplicemente probabilità zero. In generale, dato un numero n qualsiasi di estrazioni, tutti gli esiti ω in cui un numero si ripete due o più volte avranno probabilità zero.

Dobbiamo ora calcolare la probabilità di un esito ω le cui componenti sono tutte diverse fra loro. Quanti esiti del genere ci sono? Per rispondere a questa domanda calcoliamo quante liste di lunghezza n si possono costruire partendo da un numero N di elementi diversi. Tenendo conto dell'ordine (disposizioni), ci sono $D_N^n = N!/(N-n)!$ modi diversi per estrarre n elementi da un insieme di N elementi. Se l'esperimento è condotto correttamente gli esiti con probabilità positiva saranno tutti equiprobabili. La probabilità di osservare un esito ω a componenti tutte diverse è dunque

$$Q(\omega) = \frac{1}{\#\text{ esiti a componenti diverse}} = \frac{(N-n)!}{N!} .$$

La nuova legge di probabilità Q è data da

$$Q(\omega) = \begin{cases} 0 & \text{se } \omega \text{ contiene almeno due valori uguali,} \\ \frac{(N-n)!}{N!} & \text{se i valori in } \omega \text{ sono tutti diversi fra loro.} \end{cases}$$

ed è dunque diversa dalla legge di probabilità vista nel caso di un'estrazione con reinserimento (confronta la 3.2).

Applichiamo la definizione delle n variabili aleatorie X_i, $i = 1, ..., n$ al nuovo spazio di probabilità:

$$X_i(\omega) := \begin{cases} 1.\ \text{Estraggo l' } i-\text{esimo numero da } \omega, \text{ notato } \omega(i) \in \{1, 2, ...N\}. \\ 2.\ \text{Associo la cilindrata dell'} \omega(i)-\text{esima automobile.} \end{cases}$$

Il nome delle variabili aleatorie è lo stesso ma la loro distribuzione è cambiata. In particolare, $X_1, ..., X_n$ non sono più indipendenti. Questo perché l'esito della variabile aleatoria X_i dipende dall'esito delle precedenti *V.A.* $X_1, ..., X_{i-1}$. Per mostrare la dipendenza facciamo un esempio prendendo quale popolazione obiettivo il colore delle automobili immatricolate. Supponiamo che nella popolazione obiettivo di numerosità N ci siano k automobili rosse. Calcoliamo

$$P(X_2 = \text{“}rosso\text{”} \mid X_1 = \text{“}rosso\text{”}) = \frac{k-1}{N-1}$$

mentre

$$P(X_2 = \text{“}rosso\text{”} \mid X_1 \neq \text{“}rosso\text{”}) = \frac{k}{N-1}.$$

Se X_1 e X_2 fossero indipendenti le due probabilità condizionate appena calcolate dovrebbero essere uguali fra loro e uguali a $P(X_2 = \text{“}rosso\text{”})$. Le due probabilità condizionate sono fra loro diverse e da ciò concludiamo che X_1 e X_2 non sono indipendenti. Questa conclusione è generalizzabile alle n variabili aleatorie $X_1, ..., X_n$ costruite tramite selezione senza reinserimento. La tripla $(\mathbb{R}^n, \overset{n}{\underset{i=1}{\otimes}} \mathcal{E}_i, Q_{X_1,...,X_n})$ è uno spazio di probabilità ma, sebbene utile ai fini dell'inferenza statistica, non costituisce un esperimento statistico. Le n *V.A.* X_i non sono *i.i.d.* e pertanto non costituiscono un n-campione. Per quanto riguarda questo corso utilizzeremo quindi il termine *n-campione di un esperimento statistico* solo in referimento a osservazioni di variabili aleatorie *i.i.d.*.

3.3 Campionamento tramite selezione sistematica

La selezione sistematica si effettua nella maniera seguente.

1. A ciascuna delle N unità della popolazione assegnamo univocamente un numero da 1 a N.

2. Calcoliamo il passo di campionamento, notato k, definito come il rapporto tra N (la numerosità della popolazione) e n (la numerosità del campione):
$$k = \frac{N}{n}.$$

3. Selezioniamo un numero a caso r compreso tra 1 e k:
$$1 \leq r \leq k. \tag{3.3}$$

4. Includiamo nel campione le n unità aventi posizioni
$$r,\ r+k,\ r+2k, \ldots, r+(n-1)k.$$

Osservazione 8. Il salto che si effettua tra due unità consecutive selezionate si chiama *passo di campionamento*. In pratica la selezione sistematica consiste

nel partizionare[2] la popolazione in k sottoinsiemi di numerosità n per poi selezionarne uno a caso. All'interno di ogni sottoinsieme la distanza tra le unità è costante ed uguale k.

Esercizio 4. Costruite lo spazio di probabilità che utilizzereste per eseguire un campionamento sistematico. Come sono definite le variabili aleatorie $X_1, ..., X_n$? Sono indipendenti? Quanto vale la probabilità di estrarre un qualsiasi elemento della popolazione?

Quando N/n non è un numero intero, utilizzeremo la tecnica della *lista circolare*. In pratica, la procedura sarà modificata nel modo seguente.

1. Calcoliamo k^*, il numero intero più vicino a $\frac{N}{n}$.
2. Estraiamo un numero casuale compreso tra 1 e N (si noti la differenza con (3.3)):
$$1 \leq r \leq N.$$
3. Includiamo nel campione le n unità aventi posizioni
$$r,\ r + k^*,\ r + 2k^*, \ldots,\ r + (n-1)k^*$$
dove, una volta esaurita la lista all'unità N-esima senza aver estratto tutte le n unità campionarie, si continuerà a contare dalla prima unità (da qui il termine lista circolare).

3.4 Varianza di - e covarianza fra - somme pesate di variabili aleatorie

Molti stimatori in statistica ed econometria si possono rappresentare come particolari *somme pesate* di variabili aleatorie. Al fine di studiare alcune delle loro caratteristiche quali ad esempio la correttezza (Definizione 17), la varianza o la covarianza con altre statistiche sarà necessario utilizzare le proprietà del valore atteso, della varianza e della covarianza di somme di variabili aleatorie.

Esempio 22. Rendimenti di scala costanti. Nell'Esempio 21 abbiamo visto come si potrebbe stimare una funzione di produzione di tipo Cobb-Douglas. I parametri β_2 e β_3 corrispondono rispettivamente alle elasticità della produzione rispetto al lavoro e al capitale. Potremmo chiederci se una simile

[2]Una partizione di un insieme Ω si può definire come una collezione di sottoinsiemi di Ω non vuoti, mutuamente disgiunti e tali che la loro unione è l'insieme Ω stesso. (Definizione tratta da Wikipedia, Partizione (teoria degli insiemi)).

tecnologia possiede rendimenti di scala costanti. In termine dei coefficienti β_2 e β_3 tale ipotesi implica

$$\beta_2 + \beta_3 = 1. \tag{3.4}$$

Tuttavia β_2 e β_3 sono dei parametri sconosciuti che andranno stimati tramite opportune statistiche $\widehat{\beta}_i = T_i(Q_1, L_1, K_1, ..., Q_n, L_n, K_n)$ $i = 1, 2, 3$. $\widehat{\beta}_2$ e $\widehat{\beta}_3$ sono quindi delle variabili aleatorie, generalmente correlate fra loro in quanto funzioni delle stesse variabili aleatorie. Per verificare tramite un opportuno test d'ipotesi la validità dell'ipotesi di rendimenti di scala costanti procederemo come segue:

1. sostituiremo i parametri sconosciuti β_2 e β_3 nella (3.4) con i rispettivi valori stimati, $\widehat{\beta}_2$ e $\widehat{\beta}_3$.

2. calcoleremo la differenza (o scarto) tra il valore osservato $\widehat{\beta}_2 + \widehat{\beta}_3$ ed il valore teorico che in questo caso è 1.

3. Se la differenza in valore assoluto risulterà essere "grande" rifiuteremo l'ipotesi di rendimenti di scala costanti altrimenti non la rifiuteremo.

Per stabilire se la differenza così calcolata è significativa ("grande") o non significativa ("piccola") sarà necessario calcolare la varianza di $\widehat{\beta}_2 + \widehat{\beta}_3$ che come già sapete è pari a

$$V\left(\widehat{\beta}_2 + \widehat{\beta}_3\right) = V\left(\widehat{\beta}_2\right) + V\left(\widehat{\beta}_3\right) + 2Cov(\widehat{\beta}_2, \widehat{\beta}_3).$$

Poiché $\widehat{\beta}_2$ e $\widehat{\beta}_3$ risulteranno essere delle particolari somme pesate[3] delle $V.A.$ $Q_1, ..., Q_n$ e cioè

$$\widehat{\beta}_2 = \sum_{i=1}^{n} a_i Q_i \quad \text{e} \quad \widehat{\beta}_3 = \sum_{j=1}^{n} b_j Q_j$$

dovremo essere in grado di calcolare espressioni del tipo

$$Cov(\widehat{\beta}_2, \widehat{\beta}_3) = Cov(\sum_{i=1}^{n} a_i Q_i, \sum_{j=1}^{n} b_j Q_j).$$

3.4.1 Definizione della varianza e della covarianza

Rivediamo brevemente alcune definizioni. Consideriamo le due variabili aleatorie reali X e Y la cui funzione di densità[4] congiunta è notata $f_{x,y}(x, y)$

[3] I termini a_i e b_i sono dei pesi deterministici (dei numeri).

[4] Consideriamo il caso di variabili aleatorie continue. Il caso discreto è del tutto identico.

mentre le funzioni di densità marginali sono notate rispettivamente $f_x(x)$ e $f_y(y)$. La varianza di X e definita come il valore atteso ... del quadrato ... degli scarti di X dal proprio valore atteso μ_X:

$$V(X) := E\left((X-\mu_X)^2\right) = \int_{\mathbb{R}} (x-\mu_X)^2 f_x(x)dx \ . \tag{3.5}$$

Quale formula alternativa vale anche

$$V(X) = E\left(X^2\right) - (\mu_X)^2 = \int_{\mathbb{R}} x^2 f_x(x)dx - (\mu_X)^2 \ . \tag{3.6}$$

La covarianza fra X e Y è definita come il valore atteso ... del prodotto ... degli scarti di X e Y dai rispettivi valori attesi:

$$Cov(X,Y) := E\left((X-\mu_X)(Y-\mu_Y)\right) = \int_{\mathbb{R}}\int_{\mathbb{R}} (x-\mu_X)(y-\mu_Y) f_{x,y}(x,y)dx\,dy \ . \tag{3.7}$$

Come per la varianza abbiamo una formula equivalente

$$Cov(X,Y) = E\left(XY\right) - \mu_X\mu_Y = \int_{\mathbb{R}}\int_{\mathbb{R}} xy\ f_{x,y}(x,y)dx\ dy - \mu_X\mu_Y \ .$$

Guardando alla definizione (3.7) di covarianza notiamo che abbiamo il prodotto di due termini $x-\mu_X$ il primo e $y-\mu_Y$ il secondo. Questo prodotto è ulteriormente moltiplicato per la rispettiva[5] "probabilità". Poiché la funzione di densità è sempre maggiore o uguale a zero ma mai negativa il prodotto

$$(x-\mu_X)(y-\mu_Y)f_{x,y}(x,y)$$

sarà negativo quando $(x-\mu_X)$ e $(y-\mu_Y)$ possiedono segno diverso ($+-$ oppure $-+$) e positivo quando $(x-\mu_X)$ e $(y-\mu_Y)$ hanno lo stesso segno ($++$ o $--$). La covarianza è la "somma" su tutti i possibili esiti pesati per la rispettiva "probabilità". Essa sarà dunque positiva se, mediamente, quando X è sopra o sotto il suo valore atteso lo è anche Y. La covarianza sarà invece negativa se, mediamente, quando X è sopra (sotto) il proprio valore atteso Y sarà sotto (sopra) μ_Y. Intuitivamente, sapendo ad esempio che X e Y sono correlate negativamente e che il valore osservato di X è maggiore al suo valore atteso mi aspetto che la realizzazione di Y sia inferiore a μ_Y.

[5]Ricordiamo che nel caso di variabili aleatorie continue la funzione di densità non indica la probabilità di un esito.

3.4.2 Tecnica di calcolo (difficoltà pari alla battaglia navale)

Iniziamo questo paragrafo prendendo due *V.A.* che chiameremo X e Y per le quali vale la seguente relazione

$$X = \sum_{i=1}^{n} a_i X_i \text{ e } Y = \sum_{j=1}^{m} b_j Y_j, \tag{3.8}$$

dove $X_1, ..., X_n$ e $Y_1, ..., Y_m$ sono delle *V.A.* tali per cui $Cov(X_i, Y_j) = c_{ij}$. c_{ij} rappresenta la covarianza (un numero quindi) fra le variabili aleatorie X_i e Y_j. Sottolineamo nuovamente che la variabile aleatoria X è semplicemente una somma pesata (si dice anche combinazione lineare) delle variabili aleatorie $X_1, ...X_n$. I pesi di tale somma sono i coefficienti a_i e sono dei numeri.

Esempio 23. Prendiamo $n = 3$, $m = 2$ ed i seguenti valori degli a_i e b_j:

$$\begin{array}{ll} \text{Indice } i\text{:} & a_1 = 2 \quad a_2 = -3 \quad a_3 = 1 \\ \text{Indice } j\text{:} & b_1 = 1 \quad b_2 = -2 \end{array}$$

Abbiamo dunque

$$\begin{aligned} X &= 2X_1 - 3X_2 + X_3 \ , \\ Y &= Y_1 - 2Y_2 \ . \end{aligned}$$

Supponiamo che le covarianze c_{ij} siano conosciute. Come possiamo calcolare $Cov(X, Y)$ in funzione delle covarianze c_{ij}? Per spiegare come fare (la dimostrazione è data in appendice) facciamo un passo indietro e rivediamo la proprietà di linearità del valore atteso. Dal corso di analisi matematica sapete che una funzione $f : \mathbb{R} \to \mathbb{R}$ è lineare se soddisfa le seguenti due condizioni:

1. Per qualsiasi $x_1, x_2 \in \mathbb{R}$ vale che

$$f(x_1 + x_2) = f(x_1) + f(x_2), \tag{3.9}$$

2. Per qualsiasi $\lambda, x \in \mathbb{R}$ vale che

$$f(\lambda x) = \lambda f(x). \tag{3.10}$$

L'operatore valore atteso, notato E, che associa ad una *V.A.* X il suo valore atteso

$$E(X) = \sum_{k=1}^{N} x_k p(x_k) \text{ oppure } E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

è anch'esso lineare. Infatti, come visto nel corso di Statistica I, le proprietà (3.9) (3.10) sono soddisfatte per qualsiasi *V.A.* X, X_1 e X_2 nonché scalare $\lambda \in \mathbb{R}$:

$$E(X_1 + X_2) = E(X_1) + E(X_2) \text{ e } E(\lambda X) = \lambda E(X) \ . \tag{3.11}$$

Dalla proprietà di linearità del valore atteso segue in maniera immediata che

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E\left(X_i\right). \tag{3.12}$$

A parole possiamo dire che "il valore atteso di una somma pesata è uguale alla somma pesata dei valori attesi".

Esempio 24. (continuato) Supponiamo che i valori attesi di X_1, X_2 e X_3 siano rispettivamente uguali a 1, -2, 3. Il valore atteso di X sarà pertanto uguale a

$$\begin{aligned} E\left(X\right) &= E(2X_1 - 3X_2 + X_3) \\ &= 2E(X_1) - 3E(X_2) + E(X_3) \\ &= 2 + 6 + 3 = 11. \end{aligned}$$

Torniamo ora al calcolo della covarianza fra (e alla varianza di) somme pesate di *V.A.*. Questo calcolo risulta estremamente semplice se si considera che la covarianza è una funzione (operatore) *bilineare*. Cosa significa bilineare? La spiegazione è molto semplice.

Innanzi tutto osserviamo che la funzione di covarianza ha due argomenti: quando calcoliamo la covarianza lo facciamo fra due *V.A.*

$$Cov\left(\underset{1^\circ \text{ argomento}}{X}, \underset{2^\circ \text{ argomento}}{Y}\right) \ .$$

Una funzione a due variabili $f(x, y)$ è detta bilineare se, fissato il secondo (primo) argomento risulta essere lineare nel primo (secondo). In pratica deve valere che

1. Per qualsiasi $x_1, x_2 \in \mathbb{R}$ e $y \in \mathbb{R}$ vale che

$$f(x_1 + x_2, y) = f(x_1, y) + f(x_2, y), \tag{3.13}$$

   Per qualsiasi $\lambda, x \in \mathbb{R}$ e $y \in \mathbb{R}$ vale che

$$f(\lambda x, y) = \lambda f(x, y). \tag{3.14}$$

2. Per qualsiasi $y_1, y_2 \in \mathbb{R}$ e $x \in \mathbb{R}$ vale che

$$f(x, y_1 + y_2) = f(x, y_1) + f(x, y_2), \tag{3.15}$$

Per qualsiasi $\lambda, y \in \mathbb{R}$ e $x \in \mathbb{R}$ vale che

$$f(x, \lambda y) = \lambda f(x, y). \tag{3.16}$$

Esempio 25. La funzione $f(x, y) = xy$ è bilineare. La dimostrazione è lasciata come esercizio.

La covarianza è un operatore bilineare. Questo significa che essa si comporta in maniera simile al valore atteso quando noi "fissiamo" uno dei due argomenti.

Ad esempio, supponiamo di voler calcolare $Cov\,(2X_1 - 3X_2 + X_3, Y_1 - 2Y_2)$. Partiamo considerando "fisso" il secondo argomento che in questo caso è dato da $Y_1 - 2Y_2$. Per semplicità chiamiamo $Z = Y_1 - 2Y_2$ l'argomento fisso.

$$\begin{aligned} Cov\,(2X_1 - 3X_2 + X_3, Z) &= Cov\,(2X_1, Z) + Cov\,(-3X_2, Z) + Cov\,(X_3, Z) \\ &= 2Cov\,(X_1, Z) - 3Cov\,(X_2, Z) + Cov\,(X_3, Z) \end{aligned}$$

Grazie alla bilinearità della covarianza possiamo ora sviluppare ciascuno dei tre termini rispetto al secondo argomento, Z, che riscriviamo nuovamente in funzione di Y_1 e Y_2:

$$\begin{aligned} Cov\,(X_1, Y_1 - 2Y_2) &= Cov\,(X_1, Y_1) + Cov\,(X_1, -2Y_2) \\ &= Cov\,(X_1, Y_1) - 2Cov\,(X_1, Y_2)\ , \end{aligned}$$

$$\begin{aligned} Cov\,(X_2, Y_1 - 2Y_2) &= Cov\,(X_2, Y_1) + Cov\,(X_2, -2Y_2) \\ &= Cov\,(X_2, Y_1) - 2Cov\,(X_2, Y_2)\ , \end{aligned}$$

$$\begin{aligned} Cov\,(X_3, Y_1 - 2Y_2) &= Cov\,(X_3, Y_1) + Cov\,(X_3, -2Y_2) \\ &= Cov\,(X_3, Y_1) - 2Cov\,(X_3, Y_2) \end{aligned}$$

Quale risultato finale otteniamo la seguente espressione

$$\begin{aligned} Cov\,(2X_1 - 3X_2 + X_3, Y_1 - 2Y_2) &= 2Cov\,(X_1, Y_1) - 4Cov\,(X_1, Y_2) \\ &\quad -3Cov\,(X_2, Y_1) + 6Cov\,(X_2, Y_2) \\ &\quad +Cov\,(X_3, Y_1) - 2Cov\,(X_3, Y_2) \\ &= 2c_{11} - 4c_{12} - 3c_{21} + 6c_{22} + c_{31} - 2c_{32}. \end{aligned}$$

Come potete vedere il calcolo della covarianza fra somme di variabili aleatorie è laborioso ma fondamentalmente semplice una volta capite le proprietà della covarianza.

È possibile aiutarsi utilizzando una tabella che ricorda la battaglia navale. La tabella consiste in n righe e m colonne, dove n è il numero di *V.A.* che figurano nella sommatoria della X (primo argomento) e m in numero di quelle della Y (secondo argomento). All'esempio precedente possiamo associare una tabella 3×2 come questa:

	Y_1	Y_2
X_1	c_{11}	c_{12}
X_2	c_{21}	c_{22}
X_3	c_{31}	c_{32}

Questa tabella è anche chiamata tabella (o matrice) delle covarianze fra X_1, X_2, X_3 e Y_1, Y_2. Per il calcolo della covarianza aggiungiamo dapprima i coefficienti della somma pesata (combinazione lineare) davanti ad ogni variabile

	$1Y_1$	$-2Y_2$
$2X_1$	c_{11}	c_{12}
$-3X_2$	c_{21}	c_{22}
$1X_3$	c_{31}	c_{32}

ed in seguito all'interno della tabella:

	$1Y_1$	$-2Y_2$
$2X_1$	$2 \cdot 1 \cdot c_{11}$	$2 \cdot (-2) \cdot c_{12}$
$-3X_2$	$(-3) \cdot 1 \cdot c_{21}$	$(-3) \cdot (-2) \cdot c_{22}$
$1X_3$	$1 \cdot 1 \cdot c_{31}$	$1 \cdot (-2) \cdot c_{32}$

Per terminare sommiamo tutti i termini della tabella ottenendo così la covarianza fra le due somme di *V.A.*:

$$Cov\,(2X_1 - 3X_2 + X_3, Y_1 - 2Y_2) = 2c_{11} - 4c_{12} - 3c_{21} + 6c_{22} + c_{31} - 2c_{32}.$$

Per completezza di informazione diamo la formula generale in termini di sommatoria per il calcolo della covarianza fra due combinazioni lineari qualsiasi di *V.A.*:

$$Cov(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)\ .$$

3.4.3 Varianza di una somma di variabili aleatorie

La varianza di una somma di variabili aleatorie, ad esempio

$$V\left(2X_1 - 3X_2 + X_3\right) \ ,$$

si calcola in maniera del tutto simile alla covarianza fra somme di *V.A.*. È infatti immediato verificare che per una qualsiasi *V.A.* X

$$V\left(X\right) = E\left[\left(X - E(X)\right)^2\right] = E\left[\left(X - E(X)\right)\left(X - E(X)\right)\right] = Cov(X, X).$$

La varianza di una somma pesata di *V.A.* altro non è che la somma debitamente pesata della varianze di e delle covarianze fra le *V.A.* che costituiscono la trasformazione lineare.

Esempio 26. Varianza di $2X_1 - 3X_2 + X_3$
Costruiamo la tabella delle covarianze fra X_1, X_2, X_3 e ... X_1, X_2, X_3:

	X_1	X_2	X_3
X_1	c_{11}	c_{12}	c_{13}
X_2	c_{21}	c_{22}	c_{23}
X_3	c_{31}	c_{32}	c_{33}

Notiamo dapprima che la tabella delle covarianze ha in questo caso lo stesso numero di righe e di colonne. Inoltre, sulla diagonale principale (cioè in posizione $(1,1)$, $(2,2)$ e $(3,3)$) abbiamo i termini c_{ii} che corrispondono a $Cov(X_i, X_i) = V(X_i)$. Queste covarianze non sono altro che le varianze degli X_i! Per tale motivo questa tabella è chiamata la tabella (o matrice) delle *varianze-covarianze* delle *V.A.* X_1, X_2, X_3.

Fuori dalla diagonale troviamo le covarianze fra X_i e X_j, con $i \neq j$. Poiché (verificatelo dalla definizione e dalla proprietà di commutatività della moltiplicazione) $Cov(X_i, X_j) = Cov(X_j, X_i)$ la tabella delle varianze-covarianze è *simmetrica*, cioè vale

$$c_{ij} = c_{ji}.$$

Per il calcolo della varianza procediamo come per il calcolo della covarianza fra somme di *V.A.*, andando ad aggiungere alla tabella (matrice) delle varianze-covarianze i coefficienti della trasformazione lineare (somma pesata)

	$2X_1$	$-3X_2$	$1X_3$
$2X_1$	$2^2 \cdot c_{11}$	$2 \cdot (-3) \cdot c_{12}$	$2 \cdot 1 \cdot c_{13}$
$-3X_2$	$(-3) \cdot 2 \cdot c_{21}$	$(-3)^2 \cdot c_{22}$	$(-3) \cdot 1 \cdot c_{23}$
$1X_3$	$1 \cdot 2 \cdot c_{31}$	$1 \cdot (-3) \cdot c_{32}$	$1^2 \cdot c_{33}$

Le varianze sono moltiplicate per il quadrato del coefficiente della trasformazione lineare. Fuori dalla diagonale troviamo le covarianze fra le rispettive *V.A.* moltiplicate per i pesi corrispondenti. La varianza di questa trasformazione lineare è uguale alla somma di tutti gli elementi della tabella, ovvero

$$\begin{aligned} V(2X_1 - 3X_2 + X_3) &= 4c_{11} + 9c_{22} + c_{33} + \\ &\quad -6c_{12} + 2c_{13} + \\ &\quad -6c_{21} - 3c_{23} + \\ &\quad +2c_{13} - 3c_{32} \end{aligned} \tag{3.17}$$

Infine, grazie alla simmetria della matrice delle varianze-covarianze ($c_{ij} = c_{ji}$) la (3.17) può essere riscritta come

$$\begin{aligned} V(2X_1 - 3X_2 + X_3) &= 4c_{11} + 9c_{22} + c_{33} + \\ &\quad -12c_{12} + 4c_{13} - 6c_{23} \end{aligned} \tag{3.18}$$

In generale, data una qualsiasi trasformazione lineare delle n *V.A.* $X_1, \ldots, X_n$ la varianza è calcolata tramite la seguente formula

$$\begin{aligned} V\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 V(X_i) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 V(X_i) + 2\sum_{\substack{i=1 \\ i>j}}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)\ . \end{aligned} \tag{3.19}$$

Caso particolare: variabili non correlate. Quando le *V.A.* $X_1, \ldots, X_n$ sono non correlate, cioè

$$Cov(X_i, X_j) = 0 \text{ per } i \neq j$$

la formula (3.19) per il calcolo della varianza di una combinazione lineare di *V.A.* si semplifica notevolmente in

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 V\left(X_i\right). \tag{3.20}$$

In pratica, nel caso di variabili non correlate, la varianza di una somma è la somma delle varianze pesate per il *quadrato* dei coefficienti della combinazione lineare. Ricordate che la varianza è il valore atteso "di un quadrato": per tale motivo i coefficienti sono al quadrato!

3.5 Il campionamento casuale semplice

Consideriamo una popolazione di numerosità N.

Definizione 18. Chiameremo campione casuale semplice un campione di numerosità n dove ciascuna unità possiede, *ad ogni passo*, una probabilità pari a $1/N$ di essere estratta.

Esercizio 5. Mostrate che entrambe le tecniche di campionamento con e senza reinserimento soddisfano la condizione imposta dalla definizione di campione casuale semplice. Suggerimento: per quanto riguarda la selezione senza reinserimento, aiutatevi utilizzando la proprietà condizionata.

È dunque possibile ottenere un campione casuale semplice applicando una delle due tecniche di selezione casuale con o senza reinserimento viste in precedenza.

> Il campionamento casuale semplice è raramente applicato nelle indagini statistiche, sia perché la selezione è completamente affidata al caso e non incorpora le informazione note a priori sulla popolazione o sulle caratteristiche distributive delle variabili, sia perché nelle indagini su vasta scala è pesante quanto a costi di relevazione dei dati e a organizzazione del lavoro "sul campo" ... Il campionamento casuale semplice è invece quello che si assume nella teoria dell'inferenza statistica quando non è precisato il disegno adottato. Per questo tipo di campionamento sono, infatti proposte *metodiche* di *stima* e di *verifica* della significatività statistica di ipotesi sui più disparati parametri delle distribuzioni, anche multivariate, di indici di relazione tra variabili, di dati organizzati in serie temporali ecc. (Fabbris, pp. 53-54).

Il campionamento casuale semplice è lo standard verso il quale confrontare gli altri tipi di campionamento. Per tale motivo sarà quello da noi maggiormente trattato.

3.5.1 Valore atteso e varianza della media campionaria

In questo e nei prossimi paragrafi desideriamo studiare le proprietà di correttezza e di precisione della media campionaria $\overline{X}_n$ quale stimatore del valore medio μ di una particolare caratteristica numerica X della popolazione (reddito da lavoro, patrimonio, spesa in beni alimentari, ecc.). In questo caso scalziamo dal trono la distribuzione statistica P_p della popolazione obiettivo

quale principale oggetto di studio. Con modestia ci accontentiamo di stimare il valore medio μ della caratteristica X della popolazione

$$\mu = \sum_{k=1}^{K} x_k p_k$$

dove K rappresenta il numero dei diversi valori osservabili della caratteristica in esame e p_k rappresenta la frazione di unità della popolazione (la frequenza relativa dunque) aventi un valore della caratteristica uguale a x_k. Sappiamo che la popolazione è costituita da N unità. Denotiamo con a_j il valore della caratteristica della j-esima unità. Ad esempio, se decidessimo di studiare la distribuzione del reddito della popolazione svizzera, a_j rappresenterebbe il reddito dello j-esimo individuo. Un possibile parametro sconosciuto al quale potremmo essere interessati potrebbe essere il valore medio della popolazione

$$\mu = \sum_{k=1}^{K} x_k p_k = \sum_{j=1}^{N} a_j \frac{1}{N}$$

oppure la varianza della distribuzione del reddito

$$\sigma^2 = \sum_{k=1}^{K} (x_k - \mu)^2 p_k = \sum_{j=1}^{N} (a_j - \mu)^2 \frac{1}{N} \ .$$

μ e σ^2 sono due quantità che potrebbero essere calcolate se avessimo a disposizione i dati sull'intera popolazione. Assumiamo di non essere interessati all'intera distribuzione della popolazione ma unicamente al primo momento ed alla varianza: per questa volta eviteremo di specificare una famiglia parametrica di distribuzioni.

Definizione 19. Frazione di campionamento. Chiameremo la quantità

$$f := \frac{n}{N} \tag{3.21}$$

frazione di campionamento. Essa indica semplicemente la % di unità estratte rispetto al totale della popolazione.

Ci interessiamo ora alla media del campione quale stimatore di μ.

3.5.1.1 Campionamento con reinserimento

Supponiamo di estrarre un campione di numerosità n utilizzando la tecnica di selezione con reinserimento. Indichiamo con $X_1, X_2, \ldots X_n$ le n variabili

aleatorie indipendenti ed identicamente distribuite che descrivono il risultato della prima, seconda, ... n-esima estrazione. È immediato verificare[6] che per un qualsiasi $i \in (1, \ldots, n)$

$$E(X_i) = \sum_{j=1}^{N} a_j \frac{1}{N} = \mu$$

e

$$V(X_i) = E\left((X_i - \mu)^2\right) = \sum_{j=1}^{N} (a_j - \mu)^2 \frac{1}{N} = \sigma^2.$$

Per quanto riguarda la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ avremo quindi

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \\ V(\overline{X}_n) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}. \end{aligned} \tag{3.22}$$

Conclusioni: la media campionaria è uno stimatore corretto del valore atteso della popolazione. La sua precisione, valutata utilizzando la varianza quale misura di dispersione, dipende da due fattori.

1. La numerosità del campione: la varianza della media $\overline{X}_n$ decresce in maniera inversamente proporzionale al numero di osservazioni nel campione.

2. La varianza della popolazione: il recercatore non ha alcun influsso su questo parametro.

[6] Potete verificare applicando le proprietà della sommatoria che $V(X_i)$ è anche uguale a

$$\sum_{j=1}^{N} a_j^2 \frac{1}{N} - \mu^2$$

che corrisponde alla formula alternativa (3.6) della varianza data da

$$V(X_i) = E(X_i^2) - (E(X_i))^2.$$

Da ultimo sottolineamo nuovamente che $\overline{X}_n$ è una variabile aleatoria la cui distribuzione dipende dalla distribuzione della popolazione campione. Purtroppo vale la seguente osservazione. Anche qualora la distribuzione della popolazione obiettivo fosse conosciuta, la distribuzione della media $\overline{X}_n$ non sarebbe, in generale, derivabile analiticamente. Ci sono però delle eccezioni. Una fra queste è la distribuzione Normale. Infatti, quando le variabili aleatorie $X_1, ..., X_n$ sono *i.i.d.* e distribuite secondo la legge Normale $N(\mu, \sigma^2)$, allora anche la loro media è una variabile aleatoria Normale.

Esercizio 6. Utilizzando il file di excel sulla distribuzione del numero di scarpe eseguite una simulazione ed una analisi dei risultati.

3.5.1.2 Campionamento senza reinserimento

In questo caso l'estrazione degli n individui dalla popolazione avviene utilizzando la tecnica di selezione senza reinserimento. Da un punto di vista teorico il campionamento senza reinserimento appare giustificato dal fatto che l'osservazione ripetuta dello stesso individuo (unità) non aumenta l'informazione disponibile sull'intera popolazione. Appare quindi sensato togliere dall'urna gli individui già osservati e procedere solo con la parte restante della popolazione. Vedremo che tale intuizione è corretta: la media di un campione estratto con la tecnica di selezione senza reinserimento possiede una varianza inferiore alla media di un campione estratto con reinserimento[7]. Come in precedenza abbiamo che

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{3.23}$$

Le variabili X_i sono ora definite nel seguente modo:

$$X_i = \sum_{j=1}^{N} a_j U_{ij} \tag{3.24}$$

dove la variabile aleatoria

$$U_{ij} = \begin{cases} 1 & \text{se lo } j\text{-esimo individuo è estratto alla } i\text{-esima estrazione,} \\ 0 & \text{altrimenti.} \end{cases}$$

Esercizio 7. Consideriamo una popolazione obiettivo di $N = 5$ unità ed un numero $n = 3$ di estrazioni senza reinserimento. Lo spazio di probabilità è

[7] Il prezzo da pagare consiste in una maggiore difficoltà nel calcolo di $V(\overline{X}_n)$.

quello studiato nella Sezione 3.2

$$\left(\Omega^3, \overset{3}{\underset{i=1}{\otimes}} \mathcal{E}_i, Q\right).$$

Prendiamo le variabili aleatorie U_{13}, U_{21} e U_{35}.

1. Definite a parole le variabili aleatorie X_1, X_2 e X_3.
2. Definite a parole le tre variabili aleatorie U_{13}, U_{21} e U_{35}.
3. Considerate gli esiti $\omega_a = (3, 2, 5)$, $\omega_b = (4, 2, 1)$ e calcolate per ciascuna delle tre variabili U il valore che esse assumono in ω_i, $i = a, b$.
4. Mostrate che in ω_a vale $X_1 = a_1U_{11} + a_2U_{12} + a_3U_{13} + a_5U_{15}$.
5. Mostrate che in ω_b non vale $X_1 = a_1U_{11} + a_2U_{12} + a_3U_{13} + a_5U_{15}$.

Possiamo riassumere il tutto nella seguente tabella

		Indiv. 1		Indiv. 2				Indiv. N
X_1	$=$	a_1U_{11}	$+$	a_2U_{12}	$+$	$\ldots$	$+$	a_NU_{1N}
X_2	$=$	a_1U_{21}	$+$	a_2U_{22}	$+$	$\ldots$	$+$	a_NU_{2N}
$\vdots$		$\vdots$		$\vdots$				$\vdots$
X_n	$=$	a_1U_{n1}	$+$	a_2U_{n2}	$+$	$\ldots$	$+$	a_NU_{nN}
$\sum_{i=1}^n X_i$	$=$	$a_1\underbrace{\sum_{i=1}^{n} U_{i1}}_{T_1}$	$+$	$a_2\underbrace{\sum_{i=1}^{n} U_{i2}}_{T_2}$	$+$	$\ldots$	$+$	$a_N\underbrace{\sum_{i=1}^{n} U_{iN}}_{T_N}$

Questa tabella possiede più colonne che righe: $n \leq N$ (non posso estrare più di N individui). Le N variabili aleatorie T_j sono la somma rispetto al numero di estrazioni (indice i) delle variabili U_{ij}. In pratica T_j assumerà il valore 1 o 0 a seconda che lo j-esimo individuo sia stato sorteggiato o meno. Inserendo la relazione (3.24) nella (3.23) possiamo riscrivere la media come una somma pesata delle *V.A.* T_j:

$$\begin{aligned} \overline{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N} a_jU_{ij} = \frac{1}{n}\sum_{j=1}^{N}\sum_{i=1}^{n} a_jU_{ij} \\ &= \frac{1}{n}\sum_{j=1}^{N} a_j \sum_{i=1}^{n} U_{ij} = \frac{1}{n}\sum_{j=1}^{N} a_jT_j. \end{aligned} \tag{3.25}$$

A questo punto è necessario calcolare valore atteso, varianza di - e covarianza fra - variabili aleatorie T_j.

Per quanto riguarda il valore atteso di T_j abbiamo

$$E(T_j) = 0\ P(T_j = 0) + 1\ P(T_j = 1) = P(T_j = 1)$$

$P(T_j = 1)$ è la probabilità che sull'arco delle n estrazioni lo j-esimo individuo venga sorteggiato. Poiché gli esiti con probabilità non nulla sono tutti equiprobabili, è possibile calcolare $P(T_j = 1)$ semplicemente come

$$P(T_j = 1) = \frac{\#\text{ esiti favorevoli}}{\#\text{ esiti possibili}}.$$

Il numero di esiti possibili è già stato calcolato nella Sezione 3.2 ed è

$$D_N^n = \frac{N!}{(N-n)!}.$$

Per l'individuo j, il numero degli esiti favorevoli è uguale al numero di esiti ω che hanno probabilità positiva[8] e che contengono il numero j. Tale numero è la risposta alla seguente domanda: in quanti modo posso disporre gli N individui $1, ..., N$ in n scatole, sapendo che una scatola sarà occupata dall'individuo j? Abbiamo n possibilità per disporre il numero j nelle n scatole e successivamente resteranno $n-1$ scatole libere che dovranno essere riempite scegliendo da $N-1$ elementi. Il risultato è dunque

$$n\, D_{N-1}^{n-1} = \frac{n\ (N-1)!}{(N-1-(n-1))!} = \frac{n\ (N-1)!}{(N-n)!}.$$

Otteniamo così

$$P(T_j = 1) = \frac{\frac{n\ (N-1)!}{(N-1-(n-1))!}}{\frac{N!}{(N-n)!}} = \frac{n}{N}. \tag{3.26}$$

Per quanto riguarda la varianza di T_j, essa è uguale a

$$V(T_j) = E(T_j^2) - (E(T_j))^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n(N-n)}{N^2}.$$

La covarianza fra T_j e T_k è invece uguale a

$$\begin{aligned} Cov(T_j, T_k) &= E(T_j T_k) - E(T_j)E(T_k) \\ &= E(T_j T_k) - \left(\frac{n}{N}\right)^2. \end{aligned} \tag{3.27}$$

[8]Ricordiamo che gli elementi ω di Ω^n in cui un numero appare due o più volte non vanno tenuti in considerazione in quanto non si realizzeranno mai (questa è un'estrazione senza reinserimento!).

Quanto vale $E(T_jT_k)$? Il prodotto T_jT_k è sempre uguale a 0 salvo quando entrambi gli individui j e k figurano nel campione. In tal caso $T_jT_k = 1$. Quindi

$$\begin{aligned} E(T_jT_k) &= 0P(T_j = 0, T_k = 0) + 0P(T_j = 1, T_k = 0) \\ &+ 0P(T_j = 0, T_k = 1) + 1P(T_j = 1, T_k = 1) \\ &= P(T_j = 1, T_k = 1) \end{aligned}$$

Per calcolare $P(T_j = 1, T_k = 1)$ utilizziamo nuovamente la combinatoria:

1. Esiti favorevoli

   Degli n posti 2 sono occupati: uno dall' individuo j e l'altro dall'individuo k. Abbiamo $n(n-1)$ modi per disporre i nostri due individui nelle n scatole. Rimangono $n-2$ posti liberi che possono essere riempiti dagli altri $N-2$ individui. Abbiamo quindi

   $$\#\text{ esiti favorevoli} = n(n-1)D_{N-2}^{n-2}\ .$$

2. Esiti possibili (calcolati in precedenza) uguali a $\frac{N!}{(N-n)!}$.

Abbiamo quindi

$$P(T_j = 1, T_k = 1) = \frac{n(n-1)\frac{(N-2)!}{(N-2-(n-2))!}}{\frac{N!}{(N-n)!}} = \frac{n(n-1)}{N(N-1)}.$$

Inserendo quest'ultimo risultato nella (3.27) otteniamo (si noti il segno dell'espressione finale)

$$Cov(T_j, T_k) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n(N-n)}{N^2(N-1)}.$$

Siamo ora pronti a calcolare il valore atteso e la varianza di $\overline{X}_n$.

Il valore atteso di $\overline{X}_n$

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{j=1}^{N} a_jT_j\right) = \frac{1}{n}\sum_{j=1}^{N} a_jE\left(T_j\right) \\ &= \frac{1}{n}\sum_{j=1}^{N} a_j\frac{n}{N} = \frac{1}{N}\sum_{j=1}^{N} a_j = \mu. \end{aligned}$$

Quando il campionamento è effettuato senza reinserimento la media campionaria è uno stimatore corretto del valore atteso della popolazione. Questo

risultato è identico a quello ottenuto nel caso di campionamento con reinserimento. Sarà dunque interessante confrontare le varianze della media campionaria nei due casi per vedere quale delle due tecniche di campionamento fornisce i risultati migliori in termini di precisione.

La varianza di $\overline{X}_n$ Dall'equazione (3.25) otteniamo che

$$\begin{aligned} V(\overline{X}_n) &= V(\frac{1}{n}\sum_{j=1}^{N} a_j T_j) = \frac{1}{n^2} V(\sum_{j=1}^{N} a_j T_j) \\ &= \frac{1}{n^2}\sum_{j=1}^{N}\sum_{s=1}^{N} a_j a_s Cov(T_j, T_s) \\ &= \frac{1}{n^2}\left(\sum_{j=1}^{N} a_j^2 V(T_j) + \sum_{\substack{j=1 \\ j\neq s}}^{N}\sum_{s=1}^{N} a_j a_s Cov(T_j, T_s)\right). \end{aligned}$$

Utilizzando i risultati sulla varianza e covarianza delle *V.A.* T_j

$$\begin{aligned} V(\overline{X}_n) &= \frac{1}{n^2}\left(\frac{n(N-n)}{N^2}\sum_{j=1}^{N} a_j^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{\substack{j=1 \\ j\neq s}}^{N}\sum_{s=1}^{N} a_j a_s\right) \\ &= \frac{1}{n^2}\frac{n(N-n)}{N^2(N-1)}\left((N-1)\sum_{j=1}^{N} a_j^2 - \sum_{\substack{j=1 \\ j\neq s}}^{N}\sum_{s=1}^{N} a_j a_s\right) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{n}\frac{N-n}{N^2(N-1)}\left(N\sum_{j=1}^{N} a_j^2 - \sum_{j=1}^{N}\sum_{s=1}^{N} a_j a_s\right) \\ &= \frac{1}{n}\frac{N-n}{N^2(N-1)}\left(N\sum_{j=1}^{N} a_j^2 - \left(\sum_{j=1}^{N} a_i\right)\left(\sum_{s=1}^{N} a_s\right)\right) \\ &= \frac{1}{n}\frac{N-n}{N-1}\left(\frac{1}{N}\sum_{j=1}^{N} a_j^2 - \mu^2\right) \end{aligned} \tag{3.28}$$

Ma $\sum_{j=1}^{N} a_j^2 \frac{1}{N} - \mu^2$ è la formula alternativa della varianza della popolazione (confronta la Sezione 3.5.1.1). Per tale motivo otteniamo quale risultato

finale

$$V(\overline{X}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}. \tag{3.29}$$

Confrontando le due formule (3.22) e (3.29) ottenute per la varianza dello stimatore $\overline{X}_n$ notiamo che il campionamento senza reinserimento è più efficiente. Infatti la quantità $\frac{N-n}{N-1}$ è sempre minore di uno salvo nel caso particolare (in pratica non rilevante) in cui $n = 1$. Questo risultato conferma la nostra intuizione riguardo all'inutilità di reinserire l'unità osservata al fine di acquisire informazioni sulla popolazione. In questo caso la tecnica di reinserimento si rivela persino dannosa in termini della varianza di $\overline{X}_n$.

3.5.1.3 Campionamento sistematico

Supponiamo che il campione di numerosità n sia stato scelto utilizzando la tecnica di selezione sistematica (si veda il paragrafo 3.3) e per semplicità che $k = \frac{N}{n}$ sia un numero intero. Ricordiamo che k corrisponde al numero di gruppi in cui l'intera popolazione è partizionata. Ciascun gruppo conterrà dunque n unità della popolazione. La probabilità di includere un particolare individuo nel campione è uguale alla probabilità di selezionare il suo gruppo d'appartenenza: poiché in totale ci sono k gruppi ed ogni individuo appartiene ad esattamente un solo gruppo tale probabilità è $\frac{1}{k}$. Tutti gli individui hanno la medesima probabilità di essere estratti. Tale probabilità è però diversa da $\frac{1}{n}$ a meno che la numerosità N della popolazione non sia tale per cui $N = n^2$. La media campionaria $\overline{X}_n$ è come al solito uguale alla somma delle n variabili aleatorie X_i divisa per n. Tuttavia, mentre nei casi della tecnica di selezione con e senza reinserimento avevamo potuto definire le variabili aleatorie X_i come il reddito dell'individuo estratto all'i-esima estrazione, ora ci troviamo di fronte ad un'unica estrazione in cui tutti gli n individui sono estratti simultaneamente. In pratica è come se estraessimo in un unico colpo la somma $\sum_{i=1}^{n} X_i$ che, nel caso del campionamento sistematico, corrisponde alla somma della caratteristica degli n individui appartenenti al gruppo selezionato. Per tale motivo in questa Sezione utilizzeremo la seguente convenzione: l'indice j denoterà il gruppo mentre l'indice i verrà utilizzato per indicare l'individuo all'interno del gruppo. Definiamo per ciascun gruppo $j \in (1, ..., k)$ la quantità

$$g_j := \sum_{i=1}^{n} a_{ji}$$

dove a_{ji} corrisponde per definizione alla caratteristica della i-esima persona dello j-esimo gruppo. Le quantità g_j non sono aleatorie ma *sono dei numeri* in quanto il partizionamento della popolazione in k gruppi è avvenuto in

maniera deterministica. La casualità viene introdotta definendo le variabili aleatorie T_j nel modo seguente

$$T_j = \begin{cases} 1 & \text{lo } j\text{-esimo gruppo è selezionato} \\ 0 & \text{altrimenti} \end{cases}.$$

Come calcolato in precedenza è facile dimostrare che

$$\begin{aligned} E(T_j) &= \frac{1}{k}, \\ V(T_j) &= \left(\frac{1}{k} - \frac{1}{k^2}\right), \\ Cov(T_j, T_s) &= -\frac{1}{k^2}. \end{aligned}$$

A questo punto è semplice convincersi che la media campionaria di un campionamento sistematico altro non è che

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{k} g_j T_j.$$

Il valore atteso di $\overline{X}_n$ Utilizziamo le proprietà ormai note del valore atteso:

$$\begin{aligned} E\left(\overline{X}_n\right) &= E\left(\frac{1}{n}\sum_{j=1}^{k} g_j T_j\right) = \frac{1}{n}\sum_{j=1}^{k} g_j E\left(T_j\right) = \frac{1}{n}\frac{1}{k}\sum_{j=1}^{k} g_j \\ &= \frac{1}{N} \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n} a_{ji}}_{\text{somma sull'intera pop.}} = \mu. \end{aligned}$$

La media campionaria è dunque uno stimatore corretto del valore atteso della popolazione anche quando la tecnica di campionamento è quella del campionamento sistematico. Veniamo ora alla calcolo del suo grado di precisione.

La varianza di $\overline{X}_n$

$$\begin{aligned} V(\overline{X}_n) &= \frac{1}{n^2} V(\sum_{j=1}^{k} g_j T_j) = \frac{1}{n^2} \sum_{j=1}^{k} \sum_{s=1}^{k} g_j g_s Cov(T_j, T_s) \\ &= \frac{1}{n^2} \left(\sum_{j=1}^{k} g_j^2 V(T_j) + \sum_{\substack{j=1 \\ j \neq s}}^{k} \sum_{s=1}^{k} g_j g_s Cov(T_j, T_s) \right) \\ &= \frac{1}{n^2} \left(\left(\frac{1}{k} - \frac{1}{k^2} \right) \sum_{j=1}^{k} g_j^2 - \frac{1}{k^2} \sum_{\substack{j=1 \\ j \neq s}}^{k} \sum_{s=1}^{k} g_j g_s \right) \\ &= \frac{1}{n^2} \left(\frac{1}{k} \sum_{j=1}^{k} g_j^2 - \frac{1}{k^2} \sum_{j=1}^{k} \sum_{s=1}^{k} g_j g_s \right). \end{aligned}$$

Continuando esattamente come visto nella (3.28) si ottiene

$$V(\overline{X}_n) = \frac{1}{n^2} \frac{1}{k} \sum_{j=1}^{k} (g_j - \overline{g})^2 .$$

Interpretazione: la varianza di $\overline{X}_n$ dipenderà quindi dall'eterogeneità dei k gruppi rispetto alla caratteristica in esame. Se ad esempio g_1 contiene prevalentemente redditi bassi, g_2 prevalentemente redditi medi e g_3 prevalentemente redditi alti avremo una varianza elevata. Se invece g_1, g_2 e g_3 contengono tutti redditi bassi/medi/alti avremo che $g_1 \approx g_2 \approx g_3$ e la varianza dello stimatore sarà quindi più bassa.

3.6 Campionamento stratificato

Stratificare una popolazione consiste nel creare una partizione della stessa sulla base di determinati criteri. Ogni sottoinsieme è detto *strato*. I criteri utilizzati per definire gli strati dipendono dalle caratteristiche della popolazione nonché dall'obiettivo dello studio. Possibili criteri sono il sesso, l'età, il reddito, l'appartenenza ad una regione linguistica, la dimensione dell'azienda in termini di numero di dipendenti o cifra d'affari, il settore in cui opera l'azienda, ecc. Ovviamente gli strati non avranno quasi mai lo stesso numero di unità. Fabbris (p. 71) identifica i seguenti obiettivi che possono indurre ad effettuare una stratificazione della popolazione.

1. Evidenziare insiemi di unità significative per la ricerca.
2. Separare dalle altre le sottopopolazioni fisicamente isolate e con caratteristiche speciali.
3. Individuare certe unità che si vogliono osservare con tecniche particolari.
4. Introdurre sulla selezione il massimo controllo, pur mantenendola casuale.
5. Individuare sottopopolazioni al massimo omogenee rispetto alla variabile o alle variabili da rilevare e ricavare così stime più efficienti di quelle ottenibili con un campione casuale semplice.

Supponiamo per un istante di osservare l'intera popolazione. Potremmo suddividere gli uomini dalle donne e calcolare per ciascun strato i rispettivi valori attesi[9] μ_d e μ_u. Qual è la relazione tra il valore atteso μ dell'intera popolazione e quello dei due strati? Indichiamo con N_d e N_u il numero di donne e uomini contenuti nelle rispettive sottopopolazioni. Ovviamente varrà che $N = N_d + N_u$. Avremo dunque

$$\begin{aligned}
\mu &= \frac{1}{N}\left(\sum_{i=1}^{N} a_i\right) = \frac{1}{N}\left(\sum_{i=1}^{N_d} a_{d,i} + \sum_{j=1}^{N_u} a_{u,j}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N_d} a_{d,i} + \frac{1}{N}\sum_{j=1}^{N_u} a_{u,j} = \frac{N_d}{N}\frac{1}{N_d}\sum_{i=1}^{N_d} a_{d,i} + \frac{N_u}{N}\frac{1}{N_u}\sum_{j=1}^{N_u} a_{u,j} \\
&= \frac{N_d}{N}\mu_d + \frac{N_u}{N}\mu_u = \pi_d\mu_d + \pi_u\mu_u.
\end{aligned}$$

Il reddito medio della popolazione è dunque la *somma pesata* dei redditi medi dei due strati. Generalizzando al caso con H strati, varrà la seguente formula

$$\mu = \sum_{h=1}^{H} \pi_h \mu_h \tag{3.30}$$

dove

- π_h corrisponde al *peso* assegnato all'h-esimo strato e sarà uguale alla frazione di individui (unità) appartenenti allo strato h rispetto al numero totale di individui (unità) della popolazione

$$\pi_h = \frac{N_h}{N}. \tag{3.31}$$

[9]Per la precisione le due quantità μ_d e μ_u corrispondono ai valori attesi condizionati rispetto all'attributo uomo/donna.

- μ_h è il valore medio dell' h-esimo strato, ovvero

$$\mu_h = \frac{1}{N_h} \sum_{j=1}^{N_h} a_{h,j}.$$

L'idea fondamentale del campionamento stratificato è quella di sfruttare l'informazione disponibile e riguardante

- i pesi π_h sulla struttura della popolazione;
- l'omogeneità degli strati al loro interno

per migliorare la qualità della stima. Anziché stimare direttamente il valore atteso dell'intera popolazione si utilizzerà un approccio indiretto costituito da due fasi.

- Fase 1: si stimano i valori attesi μ_h delle sottopopolazioni (strati) selezionando da ciascuna sottopopolazione un campione casuale di numerosità n_h.
- Fase 2: utilizzando le stime dei singoli μ_h e grazie alla conoscenza dei pesi π_h si stima il valore atteso μ della popolazione con la formula (3.30), sostituendo ai valori attesi sconosciuti μ_h le rispettive medie campionarie, notate $\overline{X}_{n_h}$:

$$\overline{X}_{str,n} = \sum_{h=1}^{H} \pi_h \overline{X}_{n_h}.$$

Esempio 27. In Svizzera il numero di persone occupate (dati del secondo trimestre 2007, in milioni) sono 4,369 di cui 2,415 uomini e 1,954 donne. Avremo quindi $\pi_u = \frac{2,415}{4,369}$ e $\pi_d = \frac{1,954}{4,369}$. Supponiamo che $n_u = 200$ e $n_d = 100$. Sulla base dei due campioni abbiamo calcolato $\overline{x}_{n_u} = 60,000$ e rispettivamente, $\overline{x}_{n_d} = 50,000$. La stima del reddito medio dei lavoratori svizzeri sarebbe quindi

$$\begin{aligned} \overline{x}_{str,300} &= \frac{2,415}{4,369}\, 60,000 + \frac{1,954}{4,369}\, 50,000 \\ &= 33,165 + 22,362 = 55,527. \end{aligned}$$

La numerosità del campione dell' h-esimo strato è indicato con n_h. I sottocampioni potranno essere di numerosità diversa. La numerosità del campione è semplicemente la somma delle H numerosità dei sottocampioni

$$n = n_1 + n_2 + ... + n_H = \sum_{h=1}^{H} n_h.$$

Il campione è l'unione degli H sottocampioni.

Definizione 20. Frazione di campionamento. La quantità

$$f_h = \frac{n_h}{N_h} \tag{3.32}$$

è chiamata frazione di campionamento dell'h-esimo strato.

Quando si estrae la stessa frazione di unità da ogni strato, o in altre parole quando la frazione di campionamento è uguale per tutti gli strati

$$f_h = c \quad \forall h \tag{3.33}$$

allora il campione si dice *stratificato proporzionale*. Quando invece f_h non è costante si parlerà di campione *stratificato non proporzionale* o *a probabilità variabili*. Si noti che nel caso di campione stratificato proporzionale, c è uguale alla frazione di campionamento f definita dalla (3.21) e pari a $\frac{n}{N}$. Infatti, utilizzando le (3.32) e (3.33) otteniamo

$$\begin{aligned} n_h &= cN_h \quad \forall h \\ \sum_{h=1}^{H} n_h &= c\sum_{h=1}^{H} N_h \\ n &= cN \\ c &= \frac{n}{N}. \end{aligned}$$

Per un campione stratificato proporzionale vale dunque la relazione

$$n_h = \frac{N_h}{N} n \text{ o equivalentemente } \frac{n_h}{n} = \frac{N_h}{N} = \pi_h \ \forall h. \tag{3.34}$$

Fissata la numerosità del campione sulla base di vincoli economici dettati dalla limitatezza delle risorse a disposizione, il numero di unità da estrarre dall'h-esimo strato è dato dalla formula (3.34). Il campione selezionato avrà in questo caso le stesse caratteristiche della popolazione in termini di rappresentatività di ogni strato al suo interno (formula (3.33)).

Osservazione 9. Un campione stratificato proporzionale non è condizione né necessaria né sufficiente per garantire la correttezza dello stimatore $\overline{X}_{str,n}$.

Osservazione 10. Un campione di numerosità n verrà generalmente costruito eseguendo H campionamenti casuali semplici di numerosità n_h ciascuno, $h = 1, ..., H$. Se il campione è stratificato proporzionale, avremo che individui appartenenti a strati diversi avranno la medesima probabilità di figurare nel campione. Infatti, la probabilità di figurare nel campione di un qualsiasi

individuo (prendiamo ad esempio il primo) dello strato h è uguale (cf. formula (3.26))

$$P(\text{individuo 1 dello strato } h \text{ è estratto}) = \frac{n_h}{N_h} = c = \frac{n}{N}.$$

Se, al contrario, il campione non è stratificato proporzionale, il rapporto $\frac{n_h}{N_h}$ varierà con h e quindi le probabilità di inclusione dei singoli individui varieranno anch'esse da strato a strato.

Osservazione 11. Se il campione è stratificato proporzionale, la formula per il calcolo di $\overline{X}_{str,n}$ è semplificabile. Infatti

$$\begin{aligned}\overline{X}_{str,n} &= \sum_{h=1}^{H} \pi_h \overline{X}_{n_h} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X}_{n_h} = \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{n_h} \sum_{j=1}^{n_h} X_{h,j} \\ &\underset{n_h = \frac{N_h}{N} n}{=} \sum_{h=1}^{H} \frac{N_h}{N} \frac{N}{N_h} \frac{1}{n} \sum_{j=1}^{n_h} X_{h,j} = \frac{1}{n} \underbrace{\sum_{h=1}^{H} \sum_{j=1}^{n_h} X_{h,j}}_{\text{somma sull'intero campione}}.\end{aligned}$$

Tabella 3.1: Tabella riassuntiva

Strato	h	1,	...,	h,	...,	H	Popolazione
# Unità componenti	(N_h)	N_1,	...,	N_h,	...,	N_H	N
Peso	$\left(\pi_h = \frac{N_h}{N}\right)$	π_1,	...,	π_h,	...,	π_H	1
Varianza	(σ_h^2)	σ_1^2,	...,	σ_h^2,	...,	σ_H^2	σ^2
Unità campionarie	(n_h)	n_1,	...,	n_h,	...,	n_H	n
Frazione di campionamento	(f_h)	f_1,	...,	f_h,	...,	f_H	f

3.6.1 La correttezza di $\overline{X}_{str,n}$

Se gli stimatori $\overline{X}_{n_h}$ di μ_h sono tutti degli stimatori corretti, $\overline{X}_{str,n}$ sarà automaticamente uno stimatore corretto di μ:

$$E\left(\overline{X}_{str,n}\right) = E\left(\sum_{h=1}^{H} \pi_h \overline{X}_{n_h}\right) = \sum_{h=1}^{H} \pi_h E\left(\overline{X}_{n_h}\right) = \sum_{h=1}^{H} \pi_h \mu_h = \mu.$$

La stima di μ richiede dunque la stima di H valori attesi, uno per strato. Per ciascun strato siamo liberi di scegliere la tecnica di selezione nonché il tipo di stimatore più appropriato alle sue caratteristiche.

3.6.2 La varianza di $\overline{X}_{str,n}$

Poiché i campionamenti eseguiti sui vari strati sono fra loro indipendenti, la varianza di $\overline{X}_{str,n}$ è semplicemente uguale a

$$V\left(\overline{X}_{str,n}\right) = V\left(\sum_{h=1}^{H} \pi_h \overline{X}_{n_h}\right) = \sum_{h=1}^{H} \pi_h^2 V\left(\overline{X}_{n_h}\right)$$

Quest'ultima formula è valida indipendentemente dalle tecniche di campionamento applicate ai vari strati (purché sia mantenuta l'indipendenza fra un campionamento e l'altro). Ammettiamo ora che ad ogni strato venga applicata la tecnica di campionamento senza reimmissione. Abbiamo visto nei paragrafi precedenti che in tal caso la varianza[10] di $\overline{X}_{n_h}$ (cf. formula (3.29)) è uguale a

$$V(\overline{X}_{n_h}) = \frac{N_h - n_h}{N_h - 1}\frac{\sigma_h^2}{n_h} = \frac{1 - f_h}{1 - 1/N_h}\frac{\sigma_h^2}{n_h}$$

dove ora σ_h^2 rappresenta la varianza dello strato h, ovvero

$$\sigma_h^2 = \frac{1}{N_h}\sum_{j=1}^{N_h}\left(a_{h,j} - \mu_h\right)^2.$$

Nel caso di un campionamento senza reinserimento avremo quindi

$$V\left(\overline{X}_{str,n}\right) = \sum_{h=1}^{H} \pi_h^2 \frac{1 - f_h}{1 - 1/N_h}\frac{\sigma_h^2}{n_h}. \tag{3.35}$$

Osservazione 12. La varianza σ^2 della popolazione non compare più nella formula (3.35) della varianza dello stimatore $\overline{X}_{str,n}$.

Osservazione 13. $V\left(\overline{X}_{str,n}\right)$ risulterà essere tanto più piccola quanto più omogenei saranno i diversi strati al loro interno. Ricordiamo che se uno strato h è perfettamente omogeneo (tutte le unità dello strato presentano lo stesso valore della caratteristica in esame) allora la sua varianza è evidentemente nulla. Nasce da qui il vantaggio di stratificare la popolazione quando gli strati possiedono una certa uniformità rispetto alla caratteristica studiata.

Esempio 28. Caso estremo di stratificazione. Supponiamo che

- la popolazione conti 100 individui;
- i possibili valori dell'attributo in esame siano solo 4;

[10] Rispetto alla formula (3.29) occorre evidentemente sostituire N con N_h e n con n_h.

- i 4 valori siano sconosciuti;
- si sappia quali e quanti siano gli individui col medesimo attributo.

Si potrebbe dunque stratificare la popolazione in quattro strati perfettamente omogenei al loro interno. La varianza delle sottopopolazioni è nulla! Sarebbe sufficiente campionare $n = 4$ individui - un individuo da ogni strato - per riuscire in questo caso estremo a stimare con precisione assoluta il valore atteso della popolazione!

Osservazione 14. La formula (3.35) esprime la varianza di $\overline{X}_{str,n}$ in funzione delle numerosità n_h dei sottocampioni. Sapendo che per un campione stratificato proporzionale vale (cf. la (3.34))

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

è possibile riscrivere la varianza di $\overline{X}_{str,n}$ come

$$V(\overline{X}_{str,n}) = \frac{1-f}{n} \sum_{h=1}^{H} \pi_h \frac{N_h}{N_h - 1} \sigma_h^2. \tag{3.36}$$

3.6.3 Effetto della stratificazione sulla precisione di stima

Abbiamo visto in precedenza che quando il campione è ottenuto tramite campionamento casuale semplice senza reinserimento la varianza di $\overline{X}_n$ è uguale a (cf. (3.29))

$$V(\overline{X}_n) = \frac{N-n}{N-1} \frac{\sigma^2}{n} = \frac{1-f}{n} \frac{N}{N-1} \sigma^2. \tag{3.37}$$

Possiamo confrontare questo resultato con la formula (3.36) della varianza di $\overline{X}_{str,n}$ di un campione stratificato proporzionale. Tuttavia, prima di procedere al confronto vero e proprio, occorre chiarire che relazione sussiste tra la varianza dell'intera popolazione e la varianza di ogni singolo strato.

Teorema 1. Per qualsiasi popolazione stratificata vale la seguente relazione

$$\sigma^2 = \sum_{h=1}^{H} \pi_h \sigma_h^2 + \sum_{h=1}^{H} \pi_h \left(\mu_h - \mu\right)^2 \tag{3.38}$$

Interpretazione: La varianza della popolazione può essere decomposta come la somma pesata delle varianze di ogni singolo strato (misura della eterogeneità interna agli strati) più la varianza dei valori medi degli strati (misura di eterogeneità fra strati).

Utilizzando l'enunciato del Teorema 1 per il confronto fra le varianze $V(\overline{X}_n)$ e $V(\overline{X}_{str,n})$ dei due stimatori notiamo che

$$\begin{aligned} V(\overline{X}_n) - V(\overline{X}_{str,n}) &= \frac{1-f}{n}\frac{N}{N-1}\sigma^2 - \frac{1-f}{n}\sum_{h=1}^{H}\pi_h\frac{N_h}{N_h-1}\sigma_h^2 \\ &= \frac{1-f}{n}\left(\frac{N}{N-1}\sigma^2 - \sum_{h=1}^{H}\pi_h\frac{N_h}{N_h-1}\sigma_h^2\right) \end{aligned}$$

Per N e N_h sufficientemente grandi $\frac{N}{N-1} \simeq 1$ e $\frac{N_h}{N_h-1} \simeq 1$ dimodoché

$$\begin{aligned} V(\overline{X}_n) - V(\overline{X}_{str,n}) &\simeq \frac{1-f}{n}\left(\sigma^2 - \sum_{h=1}^{H}\pi_h\sigma_h^2\right) \\ &\underset{(3.38)}{=} \frac{1-f}{n}\sum_{h=1}^{H}\pi_h\left(\mu_h-\mu\right)^2 \geq 0. \end{aligned}$$

Da quest'ultima disuguaglianza deduciamo che la varianza della media calcolata utilizzando un campione stratificato proporzionale è in genere inferiore a quella di un campione casuale semplice di uguale numerosità. Il guadagno risultante dal processo di stratificazione della popolazione, rappresentato dalla quantità

$$\frac{1-f}{n}\sum_{h=1}^{H}\pi_h\left(\mu_h-\mu\right)^2 \ ,$$

è proporzionale alla varianza dei valori medi di strato. Esso sarà nullo se tutti i valori medi sono uguali fra loro.

Dimostrazione del Teorema 1. Notiamo con SS (SS_h) la somma del quadrato degli scarti dal valore atteso, ovvero

$$SS = \sum_{i=1}^{N}\left(a_i-\mu\right)^2 \text{ e } SS_h = \sum_{j=1}^{N_h}\left(a_{h,i}-\mu_h\right)^2.$$

$$
\begin{aligned}
SS &= \sum_{i=1}^{N} (a_i - \mu)^2 = \sum_{i=1}^{N} a_i^2 - N\mu^2 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} a_{h,i}^2 - N\mu^2 \\
&= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(a_{h,i}^2 - \mu_h^2 + \mu_h^2\right) - N\mu^2 \\
&= \sum_{h=1}^{H} \left(\left(\sum_{i=1}^{N_h} \left(a_{h,i}^2 - \mu_h^2\right) \right) + N_h \mu_h^2 \right) - N\mu^2 \\
&= \sum_{h=1}^{H} \left(\underbrace{\left(\sum_{i=1}^{N_h} a_{h,i}^2 - N_h \mu_h^2 \right)}_{SS_h} + N_h \mu_h^2 \right) - \underbrace{N}_{\sum_{h=1}^{H} N_h} \mu^2 \\
&= \sum_{h=1}^{H} SS_h + \sum_{h=1}^{H} N_h \mu_h^2 - \sum_{h=1}^{H} N_h \mu^2 \\
&= \sum_{h=1}^{H} SS_h + \sum_{h=1}^{H} N_h \left(\mu_h^2 - \mu^2\right)
\end{aligned}
$$

Il secondo termine, $\sum_{h=1}^{H} N_h \left(\mu_h^2 - \mu^2\right)$ è uguale a $\sum_{h=1}^{H} N_h \left(\mu_h - \mu\right)^2$. Infatti

$$
\begin{aligned}
\sum_{h=1}^{H} N_h \left(\mu_h - \mu\right)^2 &= \sum_{h=1}^{H} N_h \left(\mu_h^2 - 2\mu_h\mu + \mu^2\right) = \\
&= \sum_{h=1}^{H} N_h \mu_h^2 - 2\mu \sum_{h=1}^{H} N_h \mu_h + \mu^2 \sum_{h=1}^{H} N_h \\
&= \sum_{h=1}^{H} N_h \mu_h^2 - 2N\mu^2 + N\mu^2 = \sum_{h=1}^{H} N_h \left(\mu_h^2 - \mu^2\right)
\end{aligned}
$$

Per il termine SS vale dunque

$$
SS = \sum_{h=1}^{H} SS_h + \sum_{h=1}^{H} N_h \left(\mu_h - \mu\right)^2 .
$$

Poiché $\sigma^2 = \frac{SS}{N}$ abbiamo che

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{h=1}^{H} SS_h + \frac{1}{N}\sum_{h=1}^{H} N_h \left(\mu_h - \mu\right)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H} \frac{N_h}{N_h} SS_h + 1\sum_{h=1}^{H} \frac{N_h}{N} \left(\mu_h - \mu\right)^2 \\
&= \sum_{h=1}^{H} \frac{N_h}{N}\frac{SS_h}{N_h} + 1\sum_{h=1}^{H} \frac{N_h}{N} \left(\mu_h - \mu\right)^2
\end{aligned}$$

e quindi

$$\sigma^2 = \sum_{h=1}^{H} \pi_h \sigma_h^2 + \sum_{h=1}^{H} \pi_h \left(\mu_h - \mu\right)^2 .$$

3.7 La stima di σ^2

Nei paragrafi precedenti è stata analizzata la proprietà di correttezza della media campionaria sotto diverse tipologie di campionamento: con e senza reinserimento, sistematico e stratificato. Oltre alla proprietà di correttezza dei vari stimatori ci siamo interessati anche alla loro varianza, ovvero alla precisione con la quale il valore atteso della popolazione è approssimato. Nelle diverse formule (3.22), (3.29), (3.35) e (3.36) inerenti alla varianza della media campionaria, figura sempre σ^2 (σ_h^2), la varianza della popolazione (dell'h-esimo strato). Ad esempio, per un campione casuale semplice estratto senza reimmissione la varianza della media campionaria è pari a

$$V(\overline{X}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

N e n sono noti mentre σ^2 (σ_h^2) può o non può essere conosciuta.

1. Quando σ^2 è noto non sussiste alcun problema: $V(\overline{X}_n)$ è calcolabile.

2. Quando σ^2 (σ_h^2) non è noto.

   Quando la varianza σ^2 (σ_h^2) non è conosciuta la varianza degli stimatori del valore atteso μ studiati in precedenza non è calcolabile. Affinché le suddette formule siano utilizzabili, occorre stimare σ^2 (σ_h^2). Ai fini di questo corso - a meno che non venga specificato diversamente - utilizzeremo sempre S^2 (S_h^2) quale stimatore di σ^2 (σ_h^2)

   $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

o rispettivamente

$$S_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(X_{h,i} - \overline{X}_h \right)^2,$$

dove

- n_h rappresenta la numerosità del sottocampione estratto dall'h-esimo strato,
- $X_{h,i}$ la variabile aleatoria relativa all'i-esima estrazione dallo strato h
- $\overline{X}_h$ la media calcolata sull'h-esimo sottocampione.

$V(\overline{X}_n)$ rappresenta la vera varianza della *media campionaria* $\overline{X}_n$, mentre S^2 è unicamento lo stimatore della varianza σ^2 di un singolo X_i. Inoltre, quando σ^2 (σ_h^2) non è nota ed è stimata tramite lo stimatore S^2 (S_h^2) aggiungeremo l'accento circonflesso alla V della varianza di $\overline{X}_n$. Ad esempio, scriveremo

$$\widehat{V}(\overline{X}_n) = \frac{N-n}{N-1} \frac{S^2}{n}$$

per indicare che si tratta della stima della varianza di $\overline{X}_n$ e non della sua vera varianza $V(\overline{X}_n)$.

3.8 L'intervallo di confidenza per μ

Quando si presenta il risultato di una stima si è soliti fornire tale risultato sotto forma di un valore (stima puntuale) più o meno una certa quantità. Ad esempio, se foste interessati ad organizzare il ballo dell'università potreste stimare il numero di partecipanti a $1'000 \pm 100$ persone, intendendo in tal modo che il numero di persone sarà compreso con "molta probabilità" fra le 900 e le $1'100$ persone. Un secondo esempio potrebbe essere quello di una manifestazione sportiva dove il numero esatto di partecipanti non è noto. Gli organizzatori dovranno stimare un intervallo plausibile di partecipanti. Sulla base di tale informazione verranno poi dimensionare le infrastrutture necessarie allo svolgimento della manifestazione. Capita dunque spesso di vedere delle stime in forma di intervallo. Vedremo nei successivi paragrafi come costruire un intervallo di confidenza per μ, ovvero un intervallo dentro il quale, con molta probabilità, è contenuto il valore μ. La varianza della popolazione o della variabile aleatoria X di cui si desidera stimare il valore

atteso μ è, se non indicato diversamente, da considerarsi conosciuta. Prima di affrontare tale discussione è necessario presentare due importanti argomenti della teoria asintotica: la legge dei grandi numeri ed il teorema del limite centrale.

3.8.1 La legge (debole) dei grandi numeri

La legge dei grandi numeri è utile al fine di comprendere il comportamento *asintotico* di medie di variabili aleatorie. Studiare il comportamento asintotico di uno stimatore significa descrivere il suo comportamento quando la numerosità del campione diventa molto grande tendendo all'infinito. Supponiamo infatti di avere a disposizione una successione infinita $X_1, X_2, X_3, ...$ di variabili aleatorie

- *indipendenti*
- di valore atteso μ
- di varianza σ^2.

In questo paragrafo non imponiamo la condizione che le variabili aleatorie X_i abbiano la stessa distribuzione. Per mezzo di questa successione infinita costruiamo una seconda successione, la successione delle medie dei primi n elementi

$$\begin{array}{rcc} \overline{X}_1 & = & X_1 \\ \overline{X}_2 & = & (X_1 + X_2)/2 \\ \overline{X}_3 & = & (X_1 + X_2 + X_3)/3 \\ & \vdots & \\ \overline{X}_n & = & \frac{1}{n}\sum_{i=1}^{n} X_i \\ & \vdots & \end{array}$$

La successione delle medie $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$ è ancora una successione di variabili aleatorie. È immediato calcolare, per fisso n, il valore atteso e la varianza di $\overline{X}_n$:

$$\begin{aligned} E(\overline{X}_n) &= \mu\ , \\ V(\overline{X}_n) &= \frac{\sigma^2}{n}\ . \end{aligned}$$

Come per la successione $X_1, X_2, X_3, ...$, le *V.A.* $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$ possiedono tutte il medesimo valore atteso. Tuttavia questo non è vero per la loro

varianza, che tende a diminuire al crescere di n. Asintoticamente abbiamo che

$$\lim_{n\to\infty} V(\overline{X}_n) = 0 \ .$$

Ma questo significa che quando n è grande le realizzazioni di $\overline{X}_n$ saranno molto concentrate attorno al valore atteso μ.

Esempio 29. Per meglio comprendere quanto accade, aggiungiamo alle tre ipotesi precedenti l'ipotesi di normalità, vale a dire

- $X_1, X_2, X_3, ...$ sono distribuite secondo la legge Normale.

Questa ulteriore ipotesi non è necessaria per la conclusione a cui arriveremo, ma ci permette di eseguire il grafico della situazione. Infatti, poiché la somma pesata di *V.A.* normali indipendenti è una *V.A.* Normale, possiamo eseguire il grafico della funzione di densità degli $\overline{X}_n$, ad esempio per $n = 1, 2$ e 16. Scegliamo, senza perdita di generalità, $\mu = 1$ e $\sigma^2 = 4$.

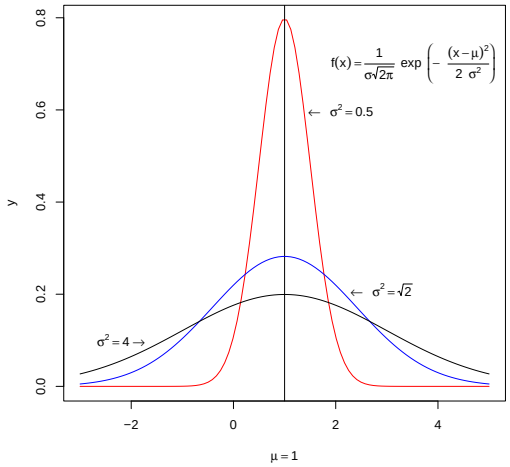

Figura 3.3: Funzione di densità di $\overline{X}_1$, $\overline{X}_2$ (blu) e $\overline{X}_{16}$(rossa)

Al crescere di n, la funzione di densità (e quindi la probabilità) tende a concentrarsi attorno al valore atteso μ (in questo esempio $\mu = 1$). Il prossimo teorema formalizza quanto visibile nella figura precedente al caso non Normale.

Teorema 2. Legge (debole) dei grandi numeri. Sia $X_1, X_2, X_3, ...$ una successione di *V.A.* indipendenti, di valore atteso μ e di varianza $\sigma^2 < \infty$. Allora per qualsiasi $\varepsilon > 0$ piccolo a piacere si avrà che

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \varepsilon) = 0. \tag{3.39}$$

Per fisso $\varepsilon > 0$, la probabilità $P(|\overline{X}_n - \mu| \leq \varepsilon)$ corrisponde alla probabilità che il valore realizzato di $\overline{X}_n$ cada in un intervallo di centro μ e raggio ε. La (3.39) ci permette di affermare che questa probabilità tende a 1 al crescere di n all'infinito.

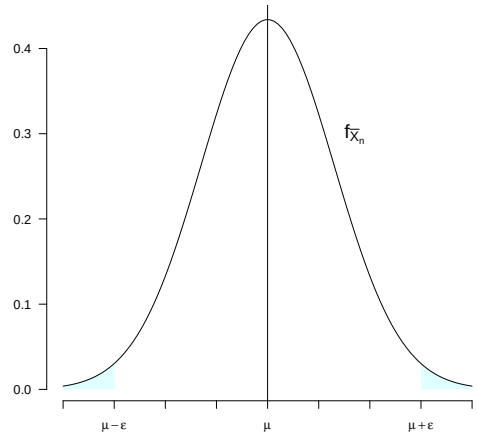

Figura 3.4: Legge dei grandi numeri

3.8.2 Uguaglianza e convergenza in distribuzione

Nel corso di Statistica I avete studiato che ad ogni variabile aleatoria a volori reali X è associata una funzione di ripartizione $F_X : \mathbb{R} \to [0, 1]$ definita per qualsiasi numero $c \in \mathbb{R}$ da

$$F_X(c) := P(X \leq c).$$

La funzione di ripartizione F_X caratterizza la variabile aleatoria X in termini della sua *probabilità*. Ora, date due variabili aleatorie X e Y è lecito chiedersi se esse abbiano la medesima distribuzione. Diremo che le due variabili aleatorie X e Y sono uguali in distribuzione, e in tal caso scriveremo $X \sim Y$, se

$$F_X(c) = F_Y(c) \quad \forall c \in \mathbb{R}. \tag{3.40}$$

Se le due variabili aleatorie X e Y sono uguali in distribuzione avremo dunque che

$$P(X \leq c) = P(Y \leq c) \quad \forall c \in \mathbb{R}. \tag{3.41}$$

A scanso di equivoci è bene precisare quanto segue. Affermare che $X \sim Y$ non significa che la realizzazione di X sarà uguale a quella di Y. Ad esempio,

supponiamo di lanciare due volte un dado. Siano X e Y le $V.A.$ che descrivono il risultato del primo e, rispettivamente, del secondo lancio: entrambe le $V.A.$ hanno una distribuzione discreta uniforme $U(1,6)$. X e Y sono uguali in distribuzione: $X \sim Y$. Ciò non toglie che nel primo lancio potrò osservare un cinque mentre nel secondo un tre.

Definizione 21. Convergenza in distribuzione. Diremo che una *qualsiasi* successione $X_1, X_2, X_3, ...$ di $V.A.$ converge in distribuzione verso Y se per ogni $c \in \mathbb{R}$ con F_Y continua in c

$$\lim_{n\to\infty} F_{X_n}(c) = F_Y(c) \text{ o equivalentemente } \lim_{n\to\infty} P(X_n \leq c) = P(Y \leq c).$$

3.8.3 Il Teorema del Limite Centrale

Il Teorema del Limite Centrale (TLC) è di estrema importanza in probabilità e statistica in quanto è il fondamento sul quale sono costruiti moltissimi test d'ipotesi (rimandiamo al capitolo sull'inferenza statistica per la definizione formale di un test d'ipotesi statistica). Nelle pagine precedenti abbiamo visto che, sotto certe condizioni, una media di $V.A.$ converge verso il suo valore atteso. Il valore limite non è più una variabile aleatoria ma un numero ben preciso.

L'idea del teorema del limite centrale è quella di evitare, tramite un'opportuna trasformazione, che la successione di $V.A.$ $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$ converga verso un numero (in questo caso il valore atteso μ). In altre parole, si desidera mantenere l'aleatorietà del limite della successione e, se possibile, ottenere come valore limite una $V.A.$ *la cui distribuzione sia nota*. Come fare dunque? L'idea è quella di standardizzare le $V.A.$ $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$ così da ottenere una nuova successione di $V.A.$ che chiameremo $Y_1, Y_2, Y_3, ...$. In pratica dovremo eseguire i due passi necessari alla standardizzazione, ovvero

1. sottrarre a $\overline{X}_n$ il suo valore atteso che in questo caso è μ ;
2. dividere per la sua deviazione standard $\sqrt{\frac{\sigma^2}{n}}$,

così da ottere una nuova successione $Y_1, Y_2, Y_3, ...$ di $V.A. \sim (0,1)$

$$Y_n = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \ . \tag{3.42}$$

Esempio 30. Riprendiamo l'Esempio (29) in cui $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. Ancora una volta, grazie alla proprietà della distribuzione Normale, dopo la standardizzazione le variabili aleatorie Y_n saranno ancora tutte normali $N(0,1)$ per qualsiasi valore di $n = 1, 2, ...$.

La proprietà di normalità delle *V.A.* Y_n nell'Esempio 30 non è vera in generale. Se la successione di *V.A.* $X_1, X_2, ...$ non è costituita da variabili normali, la media $\overline{X}_n$ non sarà più distribuita secondo la legge Normale e, di riflesso, pure Y_n. La conseguenza della non normalità delle *V.A.* $X_1, ..., X_n$ è la non normalità di Y_n, sebbene che $Y_n \sim (0, 1)$ per ogni $n = 1, 2, ...$. Cosa si potrà affermare sulla distribuzione di Y_n in questo caso? Il TLC è la soluzione a questa domanda.

Prima di presentare la versione classica del TLC facciamo notare una particolarità relativa alla standardizzazione di $\overline{X}_n$ (formula (3.42)). Tramite una serie di semplici trasformazioni algebriche è possibile riscrivere Y_n come una somma pesata di *V.A.* $\sim (0, 1)$.

Esercizio 8. Dimostrate dapprima che

$$\overline{X}_n - \mu = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)$$

e successivamente che

$$Y_n = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \ , \tag{3.43}$$

dove $\varepsilon_i = \frac{X_i - \mu}{\sigma}$.

In pratica Y_n, che corrisponde alla media $\overline{X}_n$ standardizzata, puo' essere interpretato come la somma debitamente pesata delle *V.A.* $\varepsilon_1, ..., \varepsilon_n$ dove ogni ε_i altro non è che $\frac{X_i - \mu}{\sigma} \sim (0, 1)$. La somma $\sum_{i=1}^{n} \varepsilon_i$ delle n variabili aleatorie *i.i.d.* $\sim (0, 1)$ ha valore atteso zero e varianza n. La standardizzazione richiede che il peso applicato alla somma $\sum_{i=1}^{n} \varepsilon_i$ sia $\frac{1}{\sqrt{n}}$, come appare nell'uguaglianza (3.43).

Teorema 3. Teorema del Limite Centrale (TLC). Sia $\varepsilon_1, \varepsilon_2, ...$ una successione di *V.A.* indipendenti, di valore atteso nullo e varianza unitaria. Vale allora

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \underset{n \to \infty}{\sim} N(0, 1). \tag{3.44}$$

Osservazione 15. Dal Teorema del Limite Centrale e la relazione (3.43) si deriva che, sotto le condizioni sopraelencate (quali sono?) sulla sequenza di *V.A.* $X_1, X_2, ...$, per valori di n sufficientemente grandi la media standardizzata è distribuita come una variabile aleatoria Normale standard. È importante notare e *ricordarsi* che il denominatore nella parte sinistra della (3.42) non è altro che la deviazione standard di $\overline{X}_n$.

Una variabile aleatoria Normale standard è generalmente indicata con la lettera Z:

$$Z \sim N(0,1).$$

Il vantaggio di lavorare con una variabile aleatoria Normale standard risiede nel fatto che per essa è possibile calcolare probabilità del tipo

$$P(Z \leq c)$$

utilizzando le tavole che già conoscete. Ora se $\overline{X}_n$ soddisfa il TLC avremo che per n sufficientemente grande[11]

$$\frac{\overline{X}_n - \mu}{\sqrt{V(\overline{X}_n)}} \underset{\text{circa}}{\sim} Z.$$

Poiché Z e $\frac{\overline{X}_n - \mu}{\sqrt{V(\overline{X}_n)}}$ sono uguali in distribuzione avremo (confronta (3.40) e (3.41)) che

$$P\left(\frac{\overline{X}_n - \mu}{\sqrt{V(\overline{X}_n)}} \leq c\right) = P\left(\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq c\right) \underset{TLC}{\simeq} P\left(Z \leq c\right) \quad \forall c \in \mathbb{R}.$$

Sarà questo il punto di partenza per la costruzione dell'intervallo di confidenza di μ.

3.8.4 L'intervallo di confidenza

Iniziamo la costruzione dell'intervallo di confidenza ragionando sulla variabile aleatoria Z. Siamo interessati a calcolare la probabilità

$$P(|Z| \leq 1.96) = P(-1.96 \leq Z \leq 1.96).$$

Detto a parole stiamo semplicemente calcolando la probabilità che la realizzazione di Z sia compresa nell'intervallo $[-1.96, 1.96]$. Poiché Z è una variabile aleatoria Normale standard è immediato verificare che

$$P(|Z| \leq 1.96) = 0.95\ . \tag{3.45}$$

[11] L'uguaglianza non sarà esatta ma al crescere di n all'infinito eventuali differenze saranno in pratica trascurabili.

Ora, poiché per n sufficientemente grande Z e $\frac{\overline{X}_n-\mu}{\sqrt{V(\overline{X}_n)}}$ sono approssimativamente uguali in distribuzione[12], possiamo scrivere

$$P(|Z| \leq 1.96) = P(|\frac{\overline{X}_n - \mu}{\sqrt{V(\overline{X}_n)}}| \leq 1.96)$$

da cui ricaviamo

$$P(\mu - 1.96\sqrt{V(\overline{X}_n)} \leq \overline{X}_n \leq \mu + 1.96\sqrt{V(\overline{X}_n)}) = 0.95. \tag{3.46}$$

All'uguaglianza (3.46) si dà la seguente interpretazione: 95 volte su 100 lo stimatore $\overline{X}_n$ assumerà dei valori compresi nell'intervallo di centro μ e raggio $1.96\sqrt{V(\overline{X}_n)}$, ovvero

$$P(\overline{X}_n \in \left[\mu - 1.96\sqrt{V(\overline{X}_n)}\ , \mu + 1.96\sqrt{V(\overline{X}_n)}\right]) = 0.95. \tag{3.47}$$

Tuttavia, essendo μ sconosciuto, questa informazione è poco utile. Infatti dove sarà posizionato sulla retta dei numeri reali l'intervallo

$$\left[\mu - 1.96\sqrt{V(\overline{X}_n)}\ , \mu + 1.96\sqrt{V(\overline{X}_n)}\right] \ ?$$

Per mezzo di semplici trasformazioni algebriche possiamo però riscrivere la (3.46) nel seguente modo

$$P(\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)} \leq \mu \leq \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}) = 0.95. \tag{3.48}$$

Quale interpretazione possiamo ora assegnare alla nuova rappresentazione (3.48) della precedente uguaglianza (3.46)? Per dare un significato a questa espressione osserviamo i due estremi dell'espressione

$$\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)} \leq \mu \leq \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}$$

che definiscono l'intervallo

$$\left[\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)}\ , \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}\right]. \tag{3.49}$$

L'estremo sinistro di questo intervallo è dato da $\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)}$. Il termine $1.96\sqrt{V(\overline{X}_n)}$ non è aleatorio ma è un numero. $\overline{X}_n$ per contro è

[12]Ricordiamo che l'uguaglianza è vera solo al limite.

una variabile aleatoria. Lo stesso ragionamento si applica all'estremo destro dell'intervallo. Per tale motivo l'intervallo definito dalla (3.49) è chiamato *intervallo aleatorio* o più comunemente *intervallo di confidenza al 95%*. La (3.48) può essere riscritta in maniera simile alla (3.47) utilizzanzo la notazione insiemistica "$\in$" ovvero

$$P(\mu \in \left[\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)}\ , \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}\right]) = 0.95. \tag{3.50}$$

Siamo ora pronti a dare un'interpretazione alla (3.48): con una probabilità del 95% l'intervallo aleatorio (3.49) conterrà il valore μ. Oppure: 95 volte su 100 l'intervallo $\left[\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)}\ , \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}\right]$ conterrà μ.
Esiste una sottile ma importante differenza fra la (3.47) e la (3.50). Nella (3.47) l'intervallo è deterministico ma sconosciuto e quindi inutilizzabile. Nella (3.50) per contro l'intervallo è osservabile ma aleatorio ed il punto μ è deterministico ma sconosciuto. È importante sottolineare l'equivalenza delle due espressioni: è possibile passare dall'una all'altra tramite semplici operazioni algebriche. È tuttavia la seconda espressione in termini dell'intervallo aleatorio

$$\left[\overline{X}_n - 1.96\sqrt{V(\overline{X}_n)}\ , \overline{X}_n + 1.96\sqrt{V(\overline{X}_n)}\right]$$

quella che ci fornisce uno strumento pratico per la stima del valore atteso μ. Infatti sulla base dei valori realizzati di $X_1, ..., X_n$, notati $x_1, ..., x_n$, calcoleremo dapprima la realizzazione di $\overline{X}_n$, notata $\overline{x}_n$, ed in seguito la realizzazione dell'intervallo aleatorio ovvero l'intervallo[13]

$$\left[\overline{x}_n - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]. \tag{3.51}$$

Se μ fosse conosciuto, potremmo verificare se l'intervallo così calcolato lo contiene oppure no. Tuttavia questa verifica è impossibile. Quello che sappiamo è che nel 95% dei casi questa procedura genera un intervallo contenente μ.

3.8.5 L'ampiezza dell'intervallo di confidenza

Vogliamo studiare l'ampiezza dell'intervallo di confidenza e ci chiediamo quali siano i fattori che la determinano. Osservando la formula (3.49) di intervallo aleatorio notiamo che l'ampiezza dell'intervallo è data dal termine

[13]Sappiamo che quando le osservazioni sono i.i.d. $V(\overline{X}_n) = \frac{\sigma 2}{n}$. Ricordiamo che in questo paragrafo la varianza σ^2 è da considerarsi come nota.

$1.96\sqrt{V(\overline{X}_n)}$, ovvero dalla combinazione dei due termini 1.96 e $\sqrt{V(\overline{X}_n)}$. Come già sappiamo la varianza di $\overline{X}_n$ dipende dalla numerosità n del campione, dalla varianza della popolazione e dalla tecnica di campionamento utilizzata. La costante 1.96 dipende invece dalla livello di probabilità che era stato scelto all'inizio di questo paragrafo e che ammonta al 95%. Ricordiamo infatti che la costante 1.96 è stata calcolata come la soluzione in c della seguente equazione (confronta la (3.45))

$$P(|Z| \leq c) = 0.95$$

e corrisponde al 0.975-quantile della distribuzione Normale.

La probabilità $\alpha = 0.95$ è chiamata il *livello di confidenza*. Essa rappresenta la confidenza che riponiamo nel fatto che la realizzazione dell'intervallo aleatorio (3.51) contenga μ. È il ricercatore a determinare il livello di confidenza. $\alpha = 90\%$, 95% e 99% sono i valori comunemente utilizzati. Fissato dunque il livello di confidenza α desiderato, occorre risolvere l'equazione

$$P(|Z| \leq c) = \alpha$$

in funzione di c. Le soluzioni per i tre livelli α di confidenza indicati in precedenza corrispondono agli $(\frac{1+\alpha}{2})$-quantili della distribuzione Normale e sono nell'ordine uguali a 1.64, 1.96 e 2.58. Essi rimpiazzerano il valore della costante 1.96 nella costruzione dell'intervallo di confidenza (3.51). Riassumendo: il valore che la costante c assume in funzione del livello α di confidenza corrisponde all'$(\frac{1+\alpha}{2})$-quantile della distribuzione Normale. Ad esempio, per $\alpha = 90\%$, lo 0.95-quantile è uguale a 1.64.

Esercizio 9. Descrivete come varia l'ampiezza dell'intervallo di confidenza al variare

- del livello di significatività α;
- della numerosità n del campione.