Dalla Statistica verso la Probabilità

Roberto Prisco

Indice

AVVERTENZA:

Questo testo è stato impaginato automaticamente usando il LaTex come implementato su WinShell. La versione usata non conosce la sillabazione della lingua italiana e quindi gli a capo appaiono quanto meno bizzarri. Il lettore è pregato di non tenerne conto. Sarà invece suo compito avvertire l'autore degli errori di contenuto e di richiamo di formule e tabelle che trovasse durante la lettura.

Entro qualche mese sarà prodotta una versione più completa, con l'aggiunta di esercizi e della trattazione della potenza dei test.

Capitolo 1

Teoria dei campioni

Dato un insieme Ω composto di N elementi (unità statistiche) che viene detto **popolazione** un elenco di n suoi elementi viene detto **campione**. La **Teoria dei Campioni** è la disciplina probabilistica che studia l'insieme di tutti i campioni che si possono estrarre seguendo una data regola dalla popolazione (*detta popolazione oggetto*) della quale è noto tutto quello che può interessare.

Conoscere l'insieme dei campioni e le relazioni che lo legano alla popolazione da cui sono estratti puó risultare utile in molti contesti. Ad esempio un candidato ad un esame, che sappia di avere una preparazione che gli consente di rispondere all'ottanta per cento della popolazione di tutte le domande possibili, può calcolare la probabilità di passare un esame nel quale la promozione richieda di dare almeno tre risposte corrette su un campione di quattro domande. Ricordiamo per inciso che, se la popolazione di domande possibili non è troppo numerosa, questa probabilità è calcolabile facendo ricorso alla distribuzione ipergeometrica. Un altro uso molto importante della Teoria dei Campioni riguarda l'inferenza statistica e ad essa sarà dedicato il prossimo capitolo, nel quale vedremo come sia possibile e con quali limiti, riportare sulla popolazione informazioni tratte da un campione.

Conduciamo lo studio dei campioni seguendo la scansione in fasi presentata come *operazioni della Statistica* (vedi Prisco, 2006 pag. 15).

1.0.1 Definire

Costruiamo per prima cosa l'insieme degli oggetti che verranno esaminati nelle successive parti dello studio. Stiamo studiando l'insieme di tutti i gruppi ordinati (*campioni ordinati*) di n elementi che possiamo formare partendo da un insieme più grande (*la popolazione*) formato di N elementi. Non poniamo alcuna restrizione ed ammettiamo che un elemento dell'insieme da cui siamo partiti possa comparire anche più volte nell'insieme più piccolo (*campione*) ed in qualsiasi posizione.

In generale quindi possiamo dire che da quella popolazione si possono formare N^n **campioni di numerosità n con ripetizione** degli elementi. Questo insieme di campioni ha come simbolo $\Omega_n^{r,o}$ ed è formato del prodotto cartesiano della popolazione con sé stessa per n volte, quindi (vedi Prisco, 2006 pag. 163) essendo $\Omega_n^{r,o} = \Omega \times \Omega \cdots \Omega$:

$$Card(\Omega_n^{r,o}) = Card(\Omega \times \Omega \cdots \Omega) = N * N \cdots N = N^n$$

Esempio 1.1 *Prendiamo come esempio una popolazione formata di 5 unità : $\Omega = \{A, B, C, D, E\}$ i campioni di 3 unità sono elencati nella prima parte della figura 1.1*

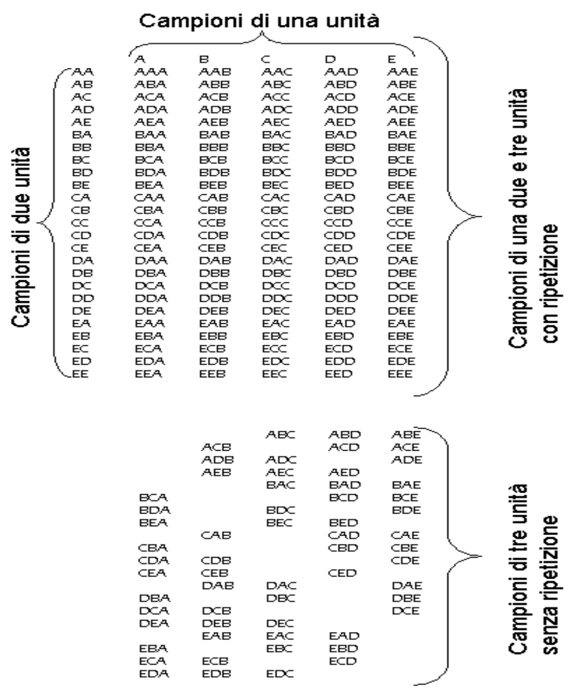


Figura 1.1: Campioni di tre unità

Una seconda regola di formazione del campione prevede che si escluda la possibilità di avere più volte la stessa unità nello stesso campione. In questo caso l'insieme dei **campioni di n elementi senza ripetizione** che ne risulta Ω_n^o è formato di ${}_nD_N$ elementi, che sono tutti compresi nell'insieme $\Omega_n^{r,o}$.

Si tratta proprio di quelle permutazioni-k (dette anche disposizioni; vedi Prisco, 2006 pag. 168) che si possono desumere dalle permutazioni e quindi:

$$Card(\Omega_n^o) =_n D_N = \frac{N!}{(N-n)!}$$

Esempio 1.2 *Si veda la seconda parte della figura 1.1.*

La partizione

Con la consistente eccezione, che noi non trattiamo, delle situazioni nelle quali è proprio la dipendenza temporale tra il prima ed il poi ad interessare, non si tiene conto dell'ordine con cui le unità statistiche entrano a far parte del (o sono scritte nel) campione; come conseguenza i campioni elencati negli insiemi $\Omega_n^{r;o}$ e Ω_n^o vengono riuniti in classi di equivalenza in modo da avere in un sottoinsieme dello spazio campionario tutti gli insiemi che differiscono soltanto per l'ordine di scrittura. Per quanto riguarda Ω_n^o chiamiamo $\mathcal{C}$ la partizione risultante che è formata di

$$Card(\mathcal{C}) = \left(\begin{array}{c} N \\ n \end{array} \right)$$

elementi. Ciascuno di questi sottoinsiemi comprende $n!$ elementi di Ω_n^o.

Viene applicata qui l'operazione logica che porta dalle permutazioni-k alle combinazioni (vedi Prisco, 2006 pag. 170).

Esempio 1.3 *La partizione $\mathcal{C}$ di Ω_n^o dell'esempio è mostrata nella figura 1.2 ed è formata di 10 elementi.*

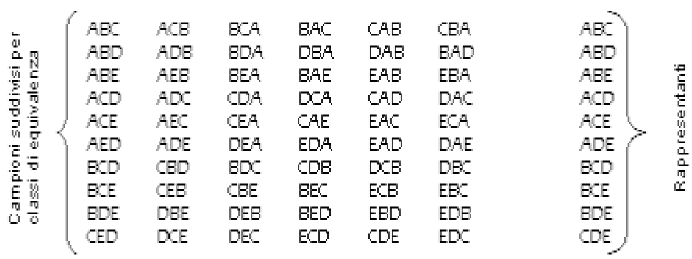


Figura 1.2: Campioni di tre unità per i quali non conta l'ordine

Più fastidiosamente complessa è la descrizione della analoga partizione fatta su $\Omega_n^{r;o}$. Si lascia alla sagacia del lettore verificare sull'esempio della figura 1.1 come i sottoinsiemi siano di numerosità diversa.

1.0.2 Misurare

I campioni mostrati negli esempi precedenti devono essere intesi come elenchi *di nomi propri* od *anagrafici* di elementi della popolazione. In una seconda fase, per mezzo della misurazione, si passa ad associare a questi oggetti gli elementi di un'insieme di contrassegni (ad esempio avendo misurato il colore dei capelli i contrassegni vengono scelti tra biondo o castano piuttosto che nero o rosso o bianco). Allora avremo un campione di misure di una certa variabile o caratteristica degli elementi del campione (colore dei capelli).

Esempio 1.4 *Le unità della popolazione degli esempi precedenti danno luogo alle seguenti misure.*

$Nome$	$Esami$	$Sesso$	$Voto\ medio$
A	5	F	25
B	10	F	27
C	10	M	24
D	5	M	23
E	5	M	29

e quindi l'elenco dei campioni di tre unità senza reinserimento e che differiscono solo per la composizione diventa, a seconda della misurazione che viene eseguita

$Unità$			$Misure$								
$nei\ campioni$			$Sesso$			$Esami$			$Voti$		
A	B	C	F	F	M	5	10	10	25	27	24
A	B	D	F	F	M	5	10	5	25	27	23
A	B	E	F	F	M	5	10	5	25	27	29
A	C	D	F	M	M	5	10	5	25	24	23
A	C	E	F	M	M	5	10	5	25	24	29
A	D	E	F	M	M	5	5	5	25	23	29
B	C	D	F	M	M	10	10	5	27	24	23
B	C	E	F	M	M	10	10	5	27	24	29
B	D	E	F	M	M	10	5	5	27	23	29
C	D	E	M	M	M	10	5	5	24	23	29

La scelta dell'unità

Finora abbiamo descritto l'insieme di tutti i campioni dal punto di vista soltanto combinatorio. Se consideriamo gli elementi di questo insieme come elementi di una partizione sulla quale eseguiamo una prova (estrazione) casuale costruiamo una partizione suscettibile di assegnazione di probabilità come previsto dagli assiomi di Kolmogoroff (vedi Prisco, 2006 pag. 25). A ciascuno degli elementi della partizione assegnamo una probabilità. L'insieme delle alternative con associate le probabilità viene detto **spazio campionario**.

Dato che questa trattazione riguarda il campionamento casuale semplice è evidente che, operando un'estrazione che non privilegia alcuno degli elementi dello spazio campionario $\mathcal{C}$, la probabilità di estrarre un certo campione (nell'esempio potrebbe essere ABE) è

$$\frac{1}{\left(\begin{array}{c} N \\ n \end{array}\right)}$$

e non varia da un campione all'altro. Questo è relativo all'estrazione del campione considerato dal punto di vista **anagrafico**, cioè della descrizione delle unità della popolazione che ne fanno parte.

Se al contrario ci rivolgiamo ai campioni di misure possiamo concludere che non tutti i campioni hanno la stessa probabilità in quanto questa dipende dalla diffusione delle singole modalità della grandezza (o attributo) nella popolazione.

Esempio 1.5 *I diversi campioni di misure, a seconda della misurazione fatta sui campioni dell'elenco dell'esempio precedente*

per quanto riguarda il sesso i tre campioni possibili sono

$Sesso$			$Prob$
F	F	M	$3/10$
F	M	M	$6/10$
M	M	M	$1/10$

per quanto riguarda il numero di esami superati, sono possibili questi altri tre campioni

$Esami$			$Prob$
5	5	5	1/10
10	5	5	6/10
10	10	5	3/10

e per quanto riguarda il voto medio conseguito, la struttura della popolazione consente dieci campioni diversi, dato che nessuna misura è comune a più elementi della popolazione

$Voto$			$Prob$
23	25	27	1/10
23	24	25	1/10
23	25	29	1/10
23	24	27	1/10
23	27	29	1/10
23	24	29	1/10
24	25	27	1/10
24	25	29	1/10
24	27	29	1/10
25	27	29	1/10

1.0.3 Descrivere

Il singolo campione estratto è formato di una molteplicità di misure, delle quali è necessario operare un riassunto descrittivo (detto *statistica di quel campione*) allo scopo di coglierne le caratteristiche che sono rilevanti per lo studio che si stà compiendo.

Esempio 1.6 *Le principali statistiche calcolate sul campione ABE sono: frequenza(F) = 2; p(F)=2/3; M(Esami) = 20/3; V(Esami) = 50/9; M(Voti)= 27; V(Voti) = 8/3. Poi, per ciascuna delle tre variabili misurabili, potremmo calcolare anche la mediana, la moda i percentili ecc.*

Le **statistiche campionarie** sono delle **funzioni aventi come dominio l'insieme delle misure possibili e, solitamente, come codominio un sottoinsieme di** $\Re$. Ad esempio la media campionaria ha come estremi il minimo ed il massimo dei valori di quella variabile nella popolazione; la varianza campionaria ha come estremi lo zero ed un estremo superiore che dipende dalla popolazione e dalla numerosità del campione, a parte casi particolari la frequenza relativa ha estremi zero ed uno.

Quando ci si riferisce al singolo campione di n unità, si parla di **media del campione** (o varianza o frequenza o percentile o .. ..) in generale di **statistica del campione** e ci si riferisce al particolare valore che quella statistica assume in *quel campione*.

La statistica campionaria, calcolata poi su tutti i campioni estraibili da quella popolazione con quella regola, assume solitamente valori non tutti uguali. Si potranno trovare infatti campioni con media diversa a seconda delle unità che compongono i diversi campioni su cui la si calcola. Riunire in un unico insieme questi valori tenendo conto delle diverse probabilità porta a costruire la **distribuzione della statistica campionaria**. Non sfugge al lettore attento che la procedura, che fa derivare la distribuzione della statistica campionaria dallo spazio campionario, non è altro che una applicazione della più generale procedura che ricava le distribuzioni delle variabili casuali dalle strutture probabilizzate secondo Kolmogoroff (vedi ad esempio Prisco, 2006 capitoli 2 e 3)

Questa distribuzione è importante in quanto descrive la previsione sul valore che la statistica assumerà come conseguenza del particolare campione di misure estratto dalla popolazione oggetto.

Esempio 1.7 *Dagli spazi campionari per estrazioni senza ripetizione dell'esempio che precede ricaviamo le seguenti distribuzioni di statistiche campionarie.*

Come prima ricaviamo la distribuzione della statistica campionaria frequenza di femmina *nel campione*

$Freq$	p	$P(p)$
0	0.00	1/10
1	0.33	6/10
2	0.67	3/10
3	1.00	0/10

come seconda quella di maschio

$Freq$	p	$P(p)$
0	0.00	0/10
1	0.33	3/10
2	0.67	6/10
3	1.00	1/10

come terza la statistica numero medio di esami *superati dai soggetti che formano il campione*

$Somma$	$media$	$P(m)$
15	5.00	1/10
20	6.67	6/10
25	8.33	3/10

come quarta la statistica voto medio degli esami *superati dai soggetti che formano il campione*

$Somma$	$media$	$P(m)$
72	24.00	1/10
74	24.67	1/10
75	25.00	1/10
76	25.33	2/10
77	25.67	1/10
78	26.00	1/10
79	26.33	1/10
80	26.67	1/10
81	27.00	1/10

Osserviamo che nei primi tre casi la distribuzione segue quella dei campioni, mentre nel quarto il passaggio alla media porta, tramite la somma, a ridurre la variabilità in quanto due campioni che hanno diversa composizione (ABC e CDE) hanno purtuttavia uguale media (25.33 *dato che la somma delle misure del campione vale* 76 *per entrambi i campioni).*

Esercizio 1.1 *Dalla popolazione descritta nell'esempio 1.7 si estrae un campione casuale semplice senza reinserimento di tre unità e ci si chiede auale sia la probabilità di ottenere un campione con*

- *A) frequenza relativa di femmina* > 0.5
- *B) frequenza relativa di femmina* $= 0.5$
- *C) frequenza relativa di femmina* < 0.5
- *D) voto medio* ≤ 25
- *E) voto medio* > 24.7

Soluzioni

- *A)* $\Pr(p > 0.5) = \Pr(p \geq 0.67) = \Pr(p = 0.67) + \Pr(p = 1) = \frac{3}{10} + \frac{0}{10} = \frac{3}{10}$
- *B)* $\Pr(p = 0.5) = 0$
- *C)* $\Pr(p < 0.5) = \Pr(p = 0) + \Pr(p = 0.33) = \frac{1}{10} + \frac{6}{10} = \frac{7}{10}$
- *D)* $\Pr(m \leq 25) = \Pr(m = 24) + \Pr(m = 24.67) + \Pr(m = 25) = \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{3}{10}$
- *E)* $\Pr(m > 25) = \Pr(m = 25) + \Pr(m = 25.33) + \Pr(m = 25.67) + \Pr(m = 26) + \Pr(m = 26.33) + \Pr(m = 26.67) + \Pr(m = 27)$

  $= \frac{1}{10} + \frac{2}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{8}{10}$

Esempio 1.8 *Situazioni analoghe si hanno per campioni formati ammettendo la ripetizione degli elementi, nel successivo esempio sono mostrate le distribuzioni corrispondenti a quelle appena viste, e che sono state ottenute dai campioni elencati nella prima parte della figura 1.1. Ciascun campione adesso ha probabilità*

$$\frac{1}{N^n}$$

Distribuzione della statistica campionaria frequenza di femmina *nel campione*

$Freq$	p	$P(p)$
0	0.00	27/125
1	0.33	54/125
2	0.67	36/125
3	1.00	8/125

la statistica numero medio di esami *superati ha distribuzione*

$Somma$	$media$	$P(m)$
15	5.00	27/125
20	6.67	54/125
25	8.33	36/125
30	10.00	8/125

rappresentiamo infine in una tabella il voto medio *riportato negli esami come risulta nell'insieme di tutti i campioni che possiamo formare da quella popolazione*

$Somma$	$media$	$P(m)$
69	23.00	1/125
70	23.33	3/125
71	23.67	6/125
72	24.00	7/125
73	24.33	9/125
74	24.67	9/125
75	25.00	13/125
76	25.33	12/125
77	25.67	15/125
78	26.00	9/125
79	26.33	12/125
80	26.67	6/125
81	27.00	10/125
82	27.33	3/125
83	27.67	6/125
85	28.33	3/125
87	29.00	1/125

Esercizio 1.2 *Calcolare le probabilità chieste nell'esercizio 1.1 sui campioni estratti con reinserimento ed elencati nell'esempio qui sopra*

Soluzioni:

$A)$ $\Pr(p > 0.5) = \frac{44}{125}$
$B)$ $\Pr(p = 0.5) = 0$
$C)$ $\Pr(p < 0.5) = \frac{81}{125}$
$D)$ $\Pr(m \leq 25) = \frac{48}{125}$
$E)$ $\Pr(m > 25) = \frac{90}{125}$

Queste distribuzioni sono descrittive dell'insieme di tutti i campioni che si possono formare con quella regola e sono previsive (in senso probabilistico) in relazione all'estrazione ancora da eseguire. Lo scopo principale della teoria dei campioni è infatti, ribadiamo, **lo studio delle distribuzioni probabilistiche delle statistiche campionarie poste in relazione con la distribuzione della popolazione**.

Dal grafico della figura 1.3 vediamo che i valori della frequenza relativa campionaria di Femmina più probabili sono vicini al valore della probabilità di Femmina nella popolazione che è 0.40

Similmente dal grafico 1.4 rileviamo che i valori più probabili della statistica campionaria sono vicini al valore di 25.60. Parametro riferito alla popolazione.

I grafici dell'esempio mostrano come la distribuzione delle statistiche campionarie frequenza relativa e media siano collegate rispettivamente con i parametri probabilità e media nella popolazione. In effetti ci si aspetta che la distribuzione di una statistica campionaria sia collegata con i parametri della popolazione (in questa prima fase dello studio non ci è ancora molto chiaro quali siano questi collegamenti). Lo studio della Teoria dei Campioni ha come primo scopo di investigare e di chiarire le relazioni di questo genere.

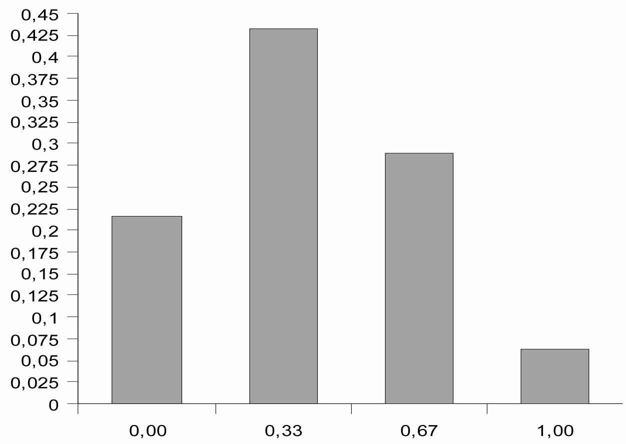


Figura 1.3: Frequenza campionaria

Passiamo a trattare alcuni teoremi che dimostrano come queste relazioni non siano dovute alla situazione particolare dell'esempio, ma siano legate alla procedura di formazione dei campioni. Vedremo in definitiva come sia possibile fissare in via generale alcune proprietà che legano i parametri della popolazione alle distribuzioni di alcune statistiche campionarie.

Nel caso dell'estrazione casuale semplice del campione i teoremi che vedremo valgono per qualsiasi distribuzione della popolazione.

distribuzione della frequenza relativa

Abbiamo una popolazione Ω composta di N elementi suddivisi secondo la partizione $\mathcal{A}$ nei due sottoinsiemi A e $\bar{A}$ di numerosità rispettivamente a e $\bar{a} = N - a$ e quindi nella popolazione abbiamo $\pi = a/N$.

Si estraggono campioni di n elementi senza reinserimento e dei quali non si tiene conto dell'ordine di estrazione. Sia X [1] il numero di volte (la frequenza) che nel campione compare la modalità A. Inoltre sia $p = X/n$ la frequenza relativa campionaria, .

Dimostriamo i primi teoremi relativi alla distribuzione della frequenza relativa campionaria.

Teorema 1.1

$$M_{sc}(p) = \pi$$

Sotto le ipotesi fatte sulla estrazione casuale semplice senza reinserimento, X (numero di unità nel campione che danno la misura A) si distribusce come

[1] Allo scopo di evitare equivoci gli operatori media, varianza ecc. calcolati sulle statistiche campionarie portano in basso l'indice sc, iniziale di *sui campioni*, per intendere che vengono calcolati tutti i campioni.

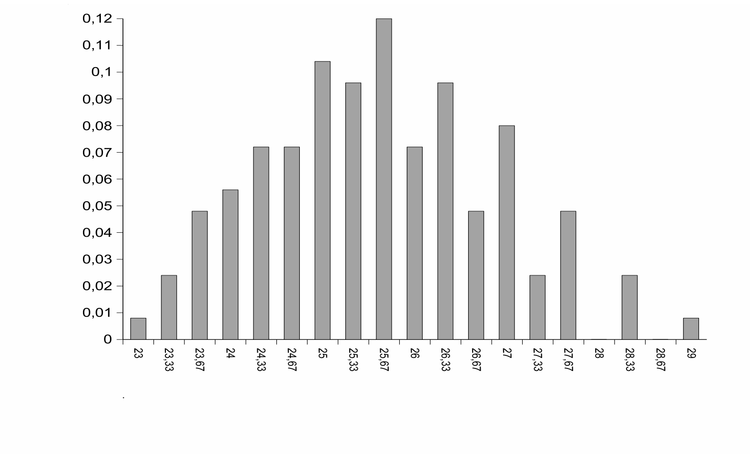


Figura 1.4: Media campionaria

una ipergeometrica con parametri N, n e a (vedi Prisco, 2006 pagg. 61 e 117) e quindi

$$M_{sc}(X) = n\pi$$

.

Essendo $p = X/n$, per il teorema della trasformata lineare (vedi Prisco, 2006 pag. 50)

$$M_{sc}(p) = \frac{1}{n} M_{sc}(X) = \frac{1}{n} n\pi = \pi$$

Teorema 1.2

$$V_{sc}(p) = \frac{\pi(1-\pi)}{n} \frac{N-n}{N-1}$$

Quanto considerato nella dimostrazione del teorema precedente ci porta (vedi Prisco, 2006 pag. 50) a

$$V_{sc}(p) = \frac{1}{n^2} V_{sc}(X) = \frac{1}{n^2} n\pi(1-\pi) \frac{N-n}{N-1} = \frac{\pi(1-\pi)}{n} \frac{N-n}{N-1}$$

Esempio 1.9 *Prendiamo in esame le due statistiche frequenza relativa campionaria di maschio e quella di femmina. Dopo aver eseguito i calcoli sull'insieme dei campioni verifichiamo che queste relazioni sono verificate.*

p	$P(p)$	$pP(p)$	$p^2P(p)$
0.00	0.10	0.00	0.0000
0.33	0.60	0.20	0.0667
0.67	0.30	0.20	0.1333
1.00	0.00	0.00	0.0000
	$Somma$	0.40	0.2000

p	$P(p)$	$pP(p)$	$p^2P(p)$
0.00	0	0.00	0.0000
0.33	0.3	0.10	0.0333
0.67	0.6	0.40	0.2667
1.00	0.1	0.10	0.1000
	$Somma$	0.60	0.4000

Infatti nella popolazione $\pi_M = 0.6$ *e nell'insieme dei campioni risulta* $M_{sc}(p) = 0.6$ *la varianza della statistica campionaria* $V_{sc}(p) = 0.4 - 0.6^2 = 0.04$ *che è uguale a* $\frac{\pi(1-\pi)}{n}\frac{N-n}{N-1} = \frac{0.6(0.4)}{3}\frac{5-3}{5-1} = 0.04$.

Per la frequenza di femmina risulta rispettivamente $M_{sc}(p) = 0.4$ *ed ancora* $V_{sc}(p) = 0.04$

Nel caso dell'estrazione con reinserimento risulta, con dimostrazioni condotte in modo simile e basate sulla distribuzione binomiale (vedi Prisco, 2006 pagg. 59 e 111) che:

$$M_{sc}(p) = \pi \tag{1.1}$$

$$V_{sc}(p) = \frac{\pi(1-\pi)}{n} \tag{1.2}$$

Sarebbe possibile trattare anche la statistica media campionaria secondo questa stessa impostazione. Questo procedimento risulterebbe però inutilmente complesso. Risulta invece molto più semplice dopo che con una nuova impostazione si passi, dal considerare il campione come un unico elemento tratto casualmente dall'elenco dei campioni, a trattarlo come il risultato di un insieme di estrazioni successive, una per ogni unità. In questa trattazione ci poniamo dal punto di vista probabilistico considerando le singole estrazioni come risultati di prove aleatorie che ancora non si sono eseguite.

frequenza relativa in singole prove

Esaminiamo l'estrazione senza reinserimento di due unità da un'urna di composizione prefissata:

$$\begin{aligned} card(\Omega) &= N \\ card(A) &= a \\ card(B) &= b \\ a+b &= N \end{aligned}$$

Per prima cosa verifichiamo che ciascuna unità riproduce la popolazione, sia quando l'estrazione è fatta con reinserimento (e questo parrebbe ovvio) sia quando viene fatta senza reinserimento. L'indice posto alla base a sinistra della lettera maiuscola individua la posizione dell'estrazione a cui si riferisce il risultato; ad esempio ${}_1A$ significa risultato A nella prima estrazione o prova, ${}_2B$ significa risultato B nella seconda estrazione o prova.

$\Pr(A) = \frac{a}{N}$
$\Pr(B) = \frac{b}{N}$

$\Pr({}_1A \cap {}_2A) = \Pr({}_1A) \Pr({}_2A|{}_1A) = \frac{a}{N} \frac{a-1}{N-1}$
$\Pr({}_1A \cap {}_2B) = \Pr({}_1A) \Pr({}_2B|{}_1A) = \frac{a}{N} \frac{b}{N-1}$
$\Pr({}_1B \cap {}_2A) = \Pr({}_1B) \Pr({}_2A|{}_1B) = \frac{b}{N} \frac{a}{N-1}$

$\Pr({}_1A) = \Pr({}_1A \cap {}_2A) + \Pr({}_1A \cap {}_2B) =$
$= \frac{a}{N} \frac{a-1}{N-1} + \frac{a}{N} \frac{b}{N-1} =$
$= \frac{a}{N} \left(\frac{a-1}{N-1} + \frac{b}{N-1} \right) = \frac{a}{N} \left(\frac{a+b-1}{N-1} \right) = \frac{a}{N}$

$\Pr({}_2A) = \Pr({}_1A \cap {}_2A) + \Pr({}_1B \cap {}_2A) =$
$= \frac{a}{N} \frac{a-1}{N-1} + \frac{b}{N} \frac{a}{N-1} =$
$= \frac{a}{N} \left(\frac{a-1}{N-1} + \frac{b}{N-1} \right) = \frac{a}{N} \left(\frac{a+b-1}{N-1} \right) = \frac{a}{N}$

Quindi la probabilità marginale di estrarre una delle A nella seconda estrazione è ancora uguale alla probabilità di estrarla nella prima.

Se la popolazione è ripartita in k sottoclassi in modo che

$$\bigcup_{i=1}^{k} A_i = \Omega$$

con

$$Card(A_i) = a_i$$

$$\sum_{i=1}^{k} a_i = N$$

per ciascuna delle quali la misurazione che interessa dà come risultato x_i allora, essendo $p_i = \frac{a_i}{N}$ la frequenza relativa di A_i e di x_i,

$$\mu = \sum_{i=1}^{k} x_i p_i$$

nella popolazione.

Nel campione dove le p_i sono le frequenze relative quando si considera l'elenco di tutti i campioni e le probabilità quando si considerino le estrazioni successive che formano un campione

$$ {}_i\mu = \sum_{i=1}^{k} x_i p_i = \mu $$

Analogamente poi per la varianza e per qualsiasi altro indice statistico. Abbiamo dimostrato con questo il **principio fondamentale del campionamento casuale semplice** con e senza reinserimento: *ciascun elemento del campione riproduce probabilisticamente la popolazione*.

Cioè la descrizione probabilistica della i-esima unità campionaria ha la stessa distribuzione della popolazione, sia per estrazioni con reinserimento sia per quelle senza reinserimento.

Quindi senza tenere conto dei risultati delle altre estrazioni

$$ M_{sc}(X_i) = M(X) = \mu \quad (1.3) $$

$$ V_{sc}(X_i) = V(X) = \sigma^2 \quad (1.4) $$

Esempio 1.10 *Verifichiamo sui campioni elencati nell'esempio 1.7 che*

$$ \Pr(A) = \Pr({}_1A) = \Pr({}_2A) = \Pr({}_3A) $$

.

Per prima cosa notiamo che, avendo posto l'indice a sinistra della modalità della popolazione che interessa, abbiamo individuato la posizione nel campione ordinato dell'unità che esaminiamo. Ci troviamo quindi a trattare con i campioni distinti non solo per la presenza delle diverse unità ma anche per l'ordine con cui vengono estratte.

Incominciamo dalla prima delle relazioni che devono essere dimostrate e cioè

$$ \Pr(A) = \Pr({}_1A) $$

Si tratta di verificare che nell'elenco dei campioni ordinati A *compare come primo estratto con la stessa probabilità* $(1/5)$ *con cui compare nella popolazione.*

Rielaboriamo l'elenco in modo da produrre da ciascuna delle terne che rappresentano i campioni non ordinati i $3! = 6$ *campioni ordinati. la terna* ACD *ad esempio fornisce i* 6 *campioni ordinati seguenti:* $ACD \ \ ADC \ \ CAD \ \ CDA \ \ DAC \ \ DCA$

Osserviamo come dalla terne ACD *si formino* 6 *campioni dei quali* 2 *presentano la* A *al primo posto. Avendo ben presente questo risultato possiamo concludere che l'elenco di* 10 *terne si sviluppa in un elenco di 60 campioni ordinati e in* 12 *di questi la* A *occupa il primo posto. Infatti in quell'elenco sono* 6 *le terne che contengono la* A *e dato che ciascuna terna la pone al primo posto in due campioni: sono* $6 \times 2 = 12$ *i campioni ordinati che iniziano per* A*.*

Concludendo i campioni con A *al primo posto sono* 12 *e quindi dato che nell'estrazione di campioni casuali semplici i campioni hanno tutti la stessa probabilità di essere estratti*

$$ \Pr({}_1A) = \frac{12}{60} = 0.2 $$

Che è la probabilità che dovevamo ottenere. Analogamente si opera per le altre posizioni.

Abbiamo dimostrato, e verificato con un esempio, che le diverse unità del campione hanno distribuzione uguale, che è anche la stessa della popolazione sia per estrazioni con reinserimento sia per estrazioni senza reinserimento. Verrebbe da ritenere allora che non valga la pena di distinguere tra le due tecniche di formazione del campione.

Prendiamo in esame un altro aspetto e cioè la dipendenza tra le estrazioni e vedremo che da questo punto di vista i due schemi differiscono.

Rileviamo facilmente dall'Appendice C che nel caso di estrazioni senza reinserimento la

$$cov(X_i X_j) = \frac{-\sigma^2}{N-1}$$

e quindi le unità del campione pur avendo uguale distribuzione sono correlate negativamente e quindi dipendenti. Al contrario le estrazioni fatte con reinserimento delle unità estratte sono indipendenti.

Siamo adesso in grado di dimostrare i seguenti teoremi che legano la distribuzione della media campionaria ai parametri della popolazione, essendo $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Teorema 1.3 *La media della media campionaria, estesa a tutti i campioni che è posssibile estrarre con e senza reinserimento è*

$$M_{sc}(\bar{X}) = \mu$$

Dimostrazione

$$M_{sc}(\bar{X}) = M_{sc}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n} M_{sc}\left(\sum_{i=1}^{n} x_i\right) =$$

dato che la media della somma è uguale alla somma delle medie (vedi Prisco, 2006 pag. 106), allora

$$= \frac{1}{n} \sum_{i=1}^{n} M_{sc}(x_i) =$$

per la formula 1.3

$$\frac{1}{n} (\underbrace{{}_1\mu + {}_2\mu + \cdots + {}_n\mu}_{sono\ n}) = \frac{1}{n} n\mu = \mu$$

Osserviamo che questo teorema vale sia per le estrazioni fatte senza reinserimento sia per quelle fatte con reinserimento. Per quanto riguarda l'analogo teorema relativo alla varianza ci sono invece delle differenze e quindi il teorema si sdoppia nei due sottocasi di estrazioni con e senza reinserimento.

Vediamo per primo il teorema per estrazioni con reinserimento

Teorema 1.4 *Sono estratti campioni casuali di n unità con il reinserimento delle unità estratte*

$$V_{sc}(\bar{X}) = \frac{\sigma^2}{n}.$$

Dimostrazione

$$V_{sc}(\bar{X}) = V_{sc}\left(\frac{\sum X_i}{n}\right) =$$

per il noto teorema sulla varianza della trasformata lineare di variabile (vedi Prisco, 2006 pag. 50) possiamo scrivere

$$= \frac{1}{n^2} V_{sc}\left(\sum X_i\right) =$$

per l'indipendenza delle estrazioni otteniamo

$$= \frac{1}{n^2} \sum V_{sc}(X_i) =$$

per la formula 1.3 in fine

$$= \frac{1}{n^2} \sum \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Teorema 1.5 *Sono estratti campioni casuali di n unità senza il reinserimento delle unità estratte.*

$$V_{sc}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

Dimostrazione

$$V_{sc}(\bar{X}) = V_{sc}\left(\frac{\sum X_i}{n}\right) =$$

per quanto appena visto (Prisco, 2006 pag. 50) possiamo scrivere

$$= \frac{1}{n^2} V_{sc}\left(\sum X_i\right) =$$

per la dipendenza tra le estrazioni

$$= \frac{1}{n^2}\left[\sum V_{sc}(X_i) + \sum_{i=1}^{n} \sum_{j \neq i} cov_{sc}(X_i; X_j)\right] =$$

come dimostrato nella appendice C le covarianze, che sono $n(n-1)$ sono tutte uguali e valgono $cov_{sc}(X_i; X_j) = \frac{-\sigma^2}{N-1}$; quindi si può scrivere

$$= \frac{1}{n^2}\left[\sum V_{sc}(X_i) + n(n-1)\frac{-\sigma^2}{N-1}\right] = \frac{1}{n^2} \frac{nN\sigma^2 - n\sigma^2 - n^2\sigma^2 + n\sigma^2}{N-1} =$$

$$= \frac{1}{n^2}\frac{n\sigma^2(N-n)}{N-1} = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Esempio 1.11

$\bar{X}$	$P(\bar{X})$	$\bar{X}P(\bar{X})$	$\bar{X}^2P(\bar{X})$
5.000	0.1	0.500	2.500
6.667	0.6	4.000	26.667
8.333	0.3	2.500	20.833
	Somma	7.000	50.000

24.000	0.1	2.400	57.600
24.667	0.1	2.467	60.846
25.000	0.1	2.500	62.500
25.333	0.2	5.0667	128.352
25.667	0.1	2.567	65.879
26.000	0.1	2.600	67.600
26.333	0.1	2.633	69.343
26.667	0.1	2.667	71.108
27.000	0.1	2.700	72.900
	Somma	25.600	656.133

Quindi per le due variabili la $M_{sc}(\bar{X})$ vale rispettivamente 7 e 25.6; mentre la $V_{sc}(\bar{X})$ vale rispettivamente 1 e 0.7733. Valori che si ottengono anche applicando le formule appena viste ai parametri della popolazione. Infatti per il numero di esami

$$V_{sc}(\bar{X}) = \frac{V(X)}{n}\frac{N-n}{N-1} = \frac{6}{3}\frac{2}{4} = 1$$

e per il voto medio

$$V_{sc}(\bar{X}) = \frac{V(X)}{n}\frac{N-n}{N-1} = \frac{4.64}{3}\frac{2}{4} = 0.7733$$

Osservazione

Possiamo affermare che per $N \to \infty$ i due schemi (con e senza reinserimento) coincidono, infatti il limite della formula dimostrata nel teorema 1.5 coincide con il risultato trovato nel teorema 1.6; inoltre se la popolazione è infinita la probabilità di pescare due volte la stessa unità è nulla. Dal punto di vista pratico si abbandonano le formule dello schema senza reinserimento quando la popolazione è abbastanza grande da rendere ininfluente (per le necessità di precisione dell'indagine) la differenza tra le formule stesse.

Teorema 1.6 *Sono estratti campioni casuali di n unità con il reinserimento delle unità estratte.*

$$M_{sc}(S^2) = \sigma^2$$

La dimostrazione è nell'appendice B.

1.1 Casi particolari

Ai fini di specifiche applicazioni sono state studiate le distribuzioni di statistiche campionarie interessanti per quei problemi. Presentiamo qui alcuni dei risultati che sono di maggior interesse dal punto di vista applicativo; si tenga sempre presente che tutto questo lavoro troverà senso pratico compiuto soltanto nell'ambito dello studio dell'Inferenza Statistica.

1.1.1 Secondo la distribuzione

Poniamo che la popolazione[2] da cui si estraggono i campioni sia distribuita come una normale con media μ e scarto quadratico medio σ, cioè $X \approx N\{\mu; \sigma\}$

Diamo senza dimostrazione che

essendo $\bar{x} = \sum_{i=1}^{n} x_i$

$$\bar{X} \approx N\left\{\mu; \frac{\sigma}{n}\right\} \tag{1.5}$$

essendo $s_1^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$

$$\frac{S_1^2 n}{\sigma^2} \approx \chi^2\left\{\nu = n\right\} \tag{1.6}$$

essendo $s_2^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$

$$\frac{S_2^2 n}{\sigma^2} \approx \chi^2\left\{\nu = n-1\right\} \tag{1.7}$$

essendo $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

$$\frac{S^2(n-1)}{\sigma^2} \approx \chi^2\left\{\nu = n-1\right\} \tag{1.8}$$

delle tre relazioni che trattano della varianza campionaria questa ultima è quella più utile per le applicazioni.

I due risultati qui esposti, relativi alle popolazioni distribuite normalmente, possono essere combinati per produrre l'utilissima distribuzione trovata da Gosset (Student) nel 1908

essendo

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

$$t \approx t\left\{\nu = n-1\right\} \tag{1.9}$$

[2]Se la popolazione è distribuita normalmente, e quindi è infinita, non serve distinguere l'estrazione con reinserimento da quella senza.

1.1.2 Secondo la numerosità

Per popolazioni distribuite in modo diverso dalla normale un importante teorema, detto **Teorema Centrale del Limite**, consente di conoscere la distribuzione, sia pure approssimata, della media campionaria. A questo teorema stato assegnato l'attributo di Centrale a causa della sua importanza.È purtroppo diffuso un errore di traduzione a causa del quale questo teorema viene chiamato in italiano Teorema del Limite Centrale, come se centrale fosse il limite e non il teorema stesso.

Nella versione che è più utile per noi il Teorema Centrale del Limite, che diamo senza dimostrazione, è:

Teorema 1.7 *Sia* $\{X_i\}$ *una successione di variabili casuali indipendenti ed identicamente distribuite. Poniamo poi che* $\mu = M(X_i)$ *e* $\sigma^2 = V(X_i)$ *esistano; chiamiamo* $S_n = \sum_i X_i$*. Allora*

$$\lim_{n\to\infty} \Pr\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \beta\right) = F_Z(\beta)$$

Dove $F_Z(\beta)$ *è la funzione di ripartizione della normale standardizzata calcolata per* $Z = \beta$*.*

La prima considerazione da fare riguarda la numerosità per lo più di difficile individuazione n_* a partire dalla quale la differenza in valore assoluto tra la vera funzione di ripartizione e quella limite della normale diventa irrilevante (vedi Prisco, 2006 pag. 76). Come già si è spiegato per altri casi simili, questo valore dipende dalla situazione (distribuzione probabilistica di X_i) e dal particolare problema che origina la necessità di quel calcolo. Per lo più si pone $n_* = 30$ e si conclude che per campioni di numerosità maggiore di 30 sia lecito usare la distribuzione normale per descrivere l'incertezza sulla media campionaria.

Di questo teorema esistono diverse versioni derivate dall'indebolimento delle condizioni iniziali e diversi collegamenti tra i quali richiamiamo quello sulla convergenza in distribuzione della v.c. Binomiale alla v.c. Normale.

Per qualche approfondimento vedi Feller, 1970 pag. 243 e seguenti del primo volume e 258 e seguenti del secondo.

Capitolo 2

Teoria dell'inferenza statistica

> ... la credenza che possiamo partire da delle pure osservazioni, senza niente di simile a una teoria, è davvero assurda (*Congetture e Confutazioni* di Karl R. Popper)

L'inferenza logica è (vedi Lalande, 1975) *Ogni operazione per mezzo della quale si ammette una proposizione la cui verità non è conosciuta direttamente, in virtù del suo legame con altre proposizioni già ritenute per vere. Questo legame può essere tale che la proposizione inferita sia giudicata necessaria, o solamente verosimile.*

Possiamo quindi ritenere che inferire sia sinonimo di far derivare, cioè l'inferenza è una forma di legame tra proposizioni. A seconda di come opera questo legame distinguiamo tra inferenze deduttive ed inferenze induttive. Le prime sono ad esempio gli strumenti di dimostrazioni matematiche e procedono di solito partendo da proposizioni generali e sono certe, le seconde procedono dal singolare e terminano al generale e sono per loro natura incerte (o verosimili).

Per quanto riguarda la deduzione lasciamo alla curiosità del lettore il compito di indurlo ad affrontare gli studi di logica; per quanto riguarda l'induzione ricordiamo i quattro metodi elencati da J. S. Mill: della concordanza, della differenza, dei residui e delle variazioni concomitanti. Per una comprensione di questi termini si rimanda ai manuali di filosofia.

Sarebbe certamente interessante esplicitare i paralleli esistenti tra questi metodi e le tecniche statistiche di investigazione dei dati, ma questo esula dalle necessità di questi appunti. A noi è sufficiente comprendere come le inferenze statistiche siano un tipo formalizzato di inferenza induttiva e quindi anche se vengono condotte rigorosamente sono soggette alla possibilità dell'errore. Ebbene studiare la entità e la probabilità degli errori e cercare di ridurle è il compito della **Teoria dell'Inferenza Statistica**. Come si vedrà le probabilità di compiere inferenze statistiche errate sarà condizionata ad alcune conoscenze di contesto e non sarà mai incondizionata. Le inferenze statistiche hanno lo scopo di trasferire **induttivamente** informazioni ottenute attraverso le misure di un campione alla popolazione da cui si è estratto quel campione.

Le due componenti essenziali per le inferenze statistiche sono quindi **la popolazione oggetto** e la **tecnica di formazione del campione**. Quest'ultima deve essere conosciuta con precisione e deve essere stata applicata rigorosamente alla popolazione oggetto che è *quell'insieme di unità statistiche delle quali si vogliono conoscere alcune caratteristiche*. Quali siano le informazioni e con quali gradi di errore le si voglia ottenere influisce sulla scelta sia della tecnica di formazione sia sulla dimensione del campione. Le eventuali conoscenze sulla popolazione delle quali eventualmente si dispone prima di formare il campione influenzano tutto il procedimento di studio e di inferenza.

La tecnica di formazione del campione deve essere applicata alla popolazione oggetto con rigore, altrimenti la procedura statistica oltre ad essere soggetta alla variabilità che le è propria (derivata dalla aleatorietà dell'estrazione campionaria) è soggetta anche ad un errore extrateoretico di difficile quantificazione dovuto al fatto che l'estrazione compiuta in modo non corretto ha avuto come conseguenza che la popolazione da cui si è estratto il campione non è quella che interessa ma un'altra.

Lo strumento principale dell'inferenza statistica è quindi la Teoria dei Campioni, che descrive le attese probabilistiche relative ai valori delle misure campionarie e delle statistiche calcolate sui loro valori. I parametri della popolazione hanno relazioni con le distribuzioni probabilistiche delle statistiche campionarie; questo fatto permette, ripetiamo non senza errore, di compiere un'inferenza che basata sul risultato di un certo campione fa esprimere conclusioni sui valori dei parametri della popolazione.

Quanto visto nel capitolo precedente riguardava la descrizione di entità ben definite: popolazioni da cui si estraggono campioni, attese di misure provenienti da prove aleatorie ben definite, rendiconti di misure ottenute da oggetti conosciuti ecc. La Teoria dei Campioni riguarda, come visto, *prove casuali ripetute secondo regole note*.

L'argomento di questo capitolo riguarda al contrario il trasferimento di informazioni parziali ad un ambito più vasto.

Pensiamo per esempio alla quantità di sondaggi politici e sociali dai cui risultati siamo sommersi tramite i giornali e gli altri mezzi di comunicazione.

Esempio 2.1 *Il sito repubblica.it riportava il 17 ottobre 2008 i risultati dei sondaggi preelettorali condotti negli USA da due diversi istituti demoscopici* Zogby *e* Gallup *relativamente alle percentuali di votanti dei due candidati alla presidenza degli USA.*

Istituto	*Obama*	*McCain*
Zogby	49	44
Gallup	49	47

Torneremo ancora ad esaminare questi dati.

Le inferenze statistiche si raggruppano in due grandi ambiti: la **stima** e la **prova di ipotesi**.

Scopo della stima è fornire informazioni sul valore di un parametro della popolazione, mentre la prova di ipotesi ha lo scopo di rispondere alla domanda se il parametro può avere un certo valore. Quindi la prova di ipotesi (o *test statistico*) presuppone che si abbia già una informazione sul parametro (che viene detta **ipotesi base** solitamente fondata su conoscenze attendibili) e ci si chiede se regge al confronto con il campione.

2.1 Stima e valore

Dobbiamo avere presente per prima la distinzione tra **valutazione** e **stima**. La prima è l'assegnazione di valore ad un parametro fatta disponendo di tutte le informazioni necessarie; quando queste non sono tutte disponibili abbiamo una **stima** che è sempre incerta a causa della assenza di informazioni rilevanti. La contrapposizione tra i due significati si coglie appieno considerandoli nel linguaggio comune. L'esempio posto qui sotto è utile a chiarire le differenze di significato.

Esempio 2.2 *Una certa persona intende vendere un appartamento e si rivolge ad un mediatore per sapere quanto potrà ricavarne. Il mediatore tiene conto di tutti i dati oggettivi dell'immobile la superficie, l'età, ecc. poi l'andamento del mercato, i prezzi a cui si sono concluse precedenti trattative ecc. e come conclusione di una serie di elaborazioni e di considerazioni fornisce una stima del valore. Soltanto però in un secondo tempo, quando si presenteranno i possibili acquirenti, la trattativa giungerà a definire il valore di scambio al quale verrà stipulato il contratto. Il valore al quale avviene lo scambio appartamento contro denaro è la valutazione.*

Se ad esempio interessa conoscere la media di una certa grandezza misurata su tutte le unità di una certa *popolazione oggetto* (**indagine totale**) è necessario disporre delle misure di tutti i suoi elementi per farne la media. Per i più vari motivi, solitamente legati ai costi della misurazione od al tempo che richiederebbe una tale operazione, talvolta si può disporre soltanto di alcune di tutte le misure di cui si avrebbe bisogno (**indagine campionaria**). Dalla rilevazione parziale si otterrà perciò una stima, che potrà essere più o meno buona a seconda della bontà del processo che ha portato alla selezione (scelta) delle unità da misurare.

Per poter eseguire una buona stima è necessario, quindi, sapere come sono state scelte le unità che costituiscono la rilevazione parziale, se le misurazioni così ottenute sono in relazione con il parametro che si intende stimare e come da queste relazioni si ottengano informazioni sul parametro.

Nel capitolo precedente abbiamo mostrato come alcune statistiche campionarie siano in relazione con i parametri della popolazione; ad esempio abbiamo visto che ciascuna unità riproduce probabilisticamente la popolazione (vedi il paragrafo 1.0.3), e questo avviene sia per le estrazioni semplici con reinserimento sia per quelle senza, poi abbiamo studiato le relazioni che queste statistiche campionarie hanno con i parametri della popolazione.

Da questo si ricava che un campione di misurazioni contiene delle valide informazioni relativamente ai parametri della popolazione. Il problema teorico della statistica inferente consiste nel trovare e valutare le procedure che *estraggono* questa informazione dai dati del campione. Il problema pratico poi consiste nell'operare la formazione del *campione casuale semplice* in modo che le unità della popolazione abbiano tutte la stessa probabilità di essere estratte. In modo cioè che la popolazione che interessa sia realmente quella investigata.

Esempio 2.3 *Per continuare l'esempio precedente possiamo immaginare che i mediatori abbiano la tendenza a sovrastimare gli immobili parlando con i venditori ed a sottostimarli parlando con i possibili compratori.*

Esempio 2.4 *Un errore compiuto frequentemente in indagini di tipo sociologico (non professionale ovviamente) consiste nel rilevare i dati, avendo per unità statistica la famiglia, e nell'interpretarli poi come se le unità statistiche fossero le persone. Prendiamo come esempio i risultati della seguente rilevazione condotta su un comune di un migliaio di abitanti*

$Componenti$	$Numero\, di\, televisori$				
$della\, famiglia$	0	1	2	3	Tot
1	30	20	3	0	53
2	14	40	60	2	116
3	0	20	70	20	110
4	0	5	50	30	85
Tot	44	85	183	52	364

Da elaborazioni condotte su questa tabella si ricava che, essendo 364 *le famiglie e* 607 *i televisori, vi sono in media* 1.67 *televisori per famiglia. Fin qui nulla di strano, ma se si passa a calcolare la percentuale di copertura televisiva di questo comune si trova che il* 12.09% *delle famiglie, ma soltanto il* 6.07% *delle persone non dispone di televisore. Talvolta si vede commettere l'errore di interpretare questi dati come se circa il* 12% *della popolazione non disponesse di televisore.*

Teniamo presente quindi nell'interpretare i risultati qual è stata la popolazione reale da cui si sono estratte le unità del campione. In questo caso si è trattato di una indagine totale, ma gli errori si ingigantiscono quando si passi all'interpretazione delle inferenze condotte sui risultati di una indagine campionaria condotta su una popolazione diversa da quella che si riteneva di esaminare.

La cautela principale consiste nel tenere sempre ben presente quali sono state le unità statistiche esaminate ed oggetto della elaborazione.

Un altro errore classico ormai per gli studiosi di sondaggi di opinione fu commesso in occasione delle elezioni presidenziali americane del 1936 quando la rivista Literary Digest considerò i propri lettori (rinforzati da elenchi esterni come quello dei possessori di telefono ed altri) come un campione rappresentativo degli elettori e sbagliò le previsioni sull'esito elettorale, avendo anche congiurato il fatto che rispose soltanto un terzo dei repubblicani ed un quarto dei democratici.

Ma non basta aver affinato le tecniche di raccolta dei dati per evitare gli errori. Durante la campagna elettorale per le elezioni americane del 2008 vediamo, infatti, che contemporaneamente vengono pubblicati sondaggi con percentuali diverse per lo stesso candidato. Questo significa che le diverse agenzie di ricerche hanno di fatto sondato popolazioni diverse.

Un modo spontaneo (*naive* come dicono gli inglesi) di risolvere il problema della stima consiste nel calcolare sulle unità del campione la statistica che si vorrebbe calcolare sulla popolazione.

Quindi per stimare la media di una popolazione dovrebbe essere sufficiente estrarne un campione preferibilmente formato di molti elementi, dato che già il buon senso ritiene poco credibili i campioni di poche unità, e calcolarne la media.

La **teoria dell'inferenza statistica** è invece più attenta, in quanto la sua funzione è appunto di studiare sotto quali condizioni questo particolare tipo di inferenza è possibile e quali ne sono i punti di forza e di debolezza.

Per quanto riguarda il procedimento di stima lo strumento operativo è lo **stimatore campionario**: *una funzione delle misure del campione la cui distribuzione probabilistica dipende dal parametro della popolazione che deve essere stimato.*

Stima viene detto sia il risultato dell'operazione sia la procedura statistica.

Per lo studio degli stimatori campionari si utilizzano i risultati ottenuti nello studio della **Teoria dei Campioni**.

Le procedure di stima a loro volta a seconda dello scopo si dividono in: **stime del valore** (dette anche stime puntuali) e **stime per intervallo** (dette anche intervalli fiduciari o intervalli di confidenza). Le prime terminano con l'assegnazione di un valore che sulla base delle proprietà dello stimatore usato viene ritenuto credibile; le seconde terminano con un intervallo di valori che, con una probabilità prefissata, contiene il parametro delle popolazione.

I **test di ipotesi**, che sono l'altra branca della statistica inferente, hanno lo scopo di saggiare se la descrizione probabilistica della popolazione possiede certe particolari caratteristiche. Queste procedure, come d'altronde anche i procedimenti di stima, sono soggette intrinsecamente alla possibilità di sbagliare. Gli errori che ne conseguono non dipendono dal fatto che *siamo uomini e quindi ci capita di sbagliare* ma dalla intrinseca debolezza del metodo campionario dovuta alla aleatorietà dell'estrazione. La **teoria dell'inferenza statistica** studia questa debolezza per permettere di limitarne gli aspetti negativi.

2.1.1 Stima del valore

Esaminiamo per prima la stima del valore. Dato che alla base della stima si trova un campione casuale, le proprietà dello stimatore vengono studiate sull'insieme di tutti i campioni estraibili dalla popolazione oggetto con quella tecnica di estrazione. Il campione che viene misurato è infatti solo uno dei possibili campioni e non essendo noi in grado di valutare se il campione è più o meno somigliante alla popolazione secondo il criterio adottato[1], il giudizio di bontà che si può dare allo stimatore deve essere dato non alla singola stima ma a tutta la procedura di stima, cioè allo stimatore. L'esame allora riguarderà tutte le diverse applicazioni possibili della procedura e potrà essere condotto prima di conoscere gli esiti della misurazione.

Trovandoci nel dominio della probabilità, a priori si potrà giudicare la bontà della procedura, ma a posteriori non si potrà giudicare la bontà del risultato, che consiste nella sua verità o aderenza alla realtà.

Affinché uno stimatore possa essere adottato è necessario quindi valutarlo sotto diversi punti di vista e cioè bisogna sapere se, almeno in media, riesce a fornire il valore del parametro (*correttezza*), se questo valore medio risulta da stime tra loro vicine (*efficienza*), con quale *distribuzione di probabilità*, ed anche se caratteristiche interessanti possono trovarsi almeno con campioni di numerosità infinita (vedi *comportamento asintotico*). In questi ultimi casi si intende dire che quanto più grandi sono i campioni tanto *migliori* sono almeno tendenzialmente i risultati campionari che si ottengono.

In definitiva la griglia di esame degli stimatori del valore è formata, al primo ed elementare livello al quale è fatta questa trattazione, degli argomenti:

- media
- variabilità
- distribuzione probabilistica
- comportamento asintotico

Esaminiamo ad uno ad uno questi aspetti degli stimatori tenendo presente che l'esame deve essere compiuto in relazione alla tecnica adottata per la formazione del campione e per la numerosità campionaria che interessa. Cominciamo dalla media dello stimatore.

correttezza

Poniamo che sia l uno stimatore campionario di un parametro λ. Questo stimatore l viene detto **corretto** (in inglese *unbiased*) se

$$M_{sc}(l) = \lambda$$

allora dello stimatore l si dice che gode della proprietà della **correttezza**. Questa proprietà è la più importante in quanto assicura che lo stimatore, il quale può anche assumere un valore distante da quello vero, purtuttavia almeno in media è in grado di cogliere il valore vero del parametro della popolazione.

[1]Parametro sotto studio

Quando questa proprietà non è verificata e quindi

$$M_{sc}(l) = \lambda + B$$

si parla di stimatori non corretti o tendenziosi (in inglese *biased*) ed il valore B è detto distorsione (*bias*).

Esempio 2.5 *La varianza campionaria, nella formula S_2^2 (vedi formula 1.7) può essere uno stimatore di σ^2 in quanto (vedi appendice B) la sua media è $\frac{n-1}{n}$ volte la varianza della popolazione, e quindi è collegata con il parametro da stimare; non è però uno stimatore corretto. Al contrario S^2 (vedi formula 1.8) è uno stimatore corretto.*

efficienza

Dato che lo stimatore è descrivibile come una variabile casuale, cioè assume valori almeno in parte imprevedibili, oltre alla media dispone anche di una varianza che indichiamo come

$$V_{sc}(l)$$

Poniamo che l_1, l_2 ed l_3 siano stimatori corretti dello stesso parametro λ, ciascuno di essi possiede una sua varianza che è rispettivamente $V_{sc}(l_1)$, $V_{sc}(l_2)$ e $V_{sc}(l_3)$. La varianza è un indice inversamente proporzionale all'efficienza dello stimatore, infatti quanto minore è la varianza tanto maggiore è la capacità di quello stimatore di fornire stime *mediamente vicine* al valore del parametro. Quindi uno stimatore viene definito più efficiente di un altro se ha varianza minore. In questo caso si parla di **efficienza relativa**

Efficienza assoluta Viene definito **il più efficiente** quello stimatore corretto l_e tale che

$$V_{sc}(l_e) \leq V_{sc}(l)$$

essendo l un qualsiasi altro stimatore corretto dello stesso parametro.

Tra diversi stimatori corretti della stessa numerosità deve essere preferito quello a varianza minima.

Esempio 2.6 *Per $n = 3$ lo stimatore media campionaria è corretto ed ha varianza $\sigma^2/3$.*

Anche lo stimatore $m_2 = \frac{X_1+2X_2+X_3}{4}$ è corretto, come si può dimostrare seguendo la stessa tecnica di dimostrazione del teorema 1.3.

Riprendendo la tecnica del teorema 1.4 si può dimostrare che $V(m_2) = \sigma^2 6/16 > \sigma^2 6/18 = \sigma^2/3$.

Sorge allora il problema di sapere se si possa trovare un altro stimatore migliore di uno corretto già noto e con una certa varianza. Per certi casi è possibile rispondere negativamente alla domanda. Infatti, con la disuguaglianza di Cramer-Rao è possibile individuare il valore minimo sotto il quale la $V_{sc}(l)$ non può scendere; per una esposizione e dimostrazione vedi Azzalini, 2001 pag. 79 od anche Piccolo, 2000 pag. 558 e pag. 524 per le condizioni, per altro assai ampie, di validità del teorema. Vediamo ora come è fatta questa disuguaglianza

Teorema 2.1 *Sia X una popolazione descritta dalla densità di probabilità $p(x;\lambda)$ e λ uno dei parametri che la caratterizzano. Dato lo stimatore corretto l di λ allora*

$$V_{sc}(l) \geq \frac{[\tau'(\lambda)]^2}{nM\left(\left[\frac{\partial}{\partial\lambda}\ln p(x;\lambda)\right]^2\right)}$$

dove $\tau(\lambda)$ è una funzione biunivoca e derivabile di λ.

Teniamo presente che se, come è usuale in questo livello elementare di trattazione non si stimano delle funzioni particolari di λ,

$$\tau(\lambda) = \lambda$$
$$\tau'(\lambda) = 1$$

Il teorema della unicità dello stimatore più efficiente ci consente di affermare poi che se si trova uno stimatore corretto la cui varianza è uguale a quella minima del teorema di Cramer Rao allora è unico.

Esempio 2.7 *Consideriamo adesso una popolazione distribuita normalmente con media μ e varianza σ^2. La media campionaria di campioni di n elementi è stimatore corretto di μ ed ha varianza pari a $\frac{\sigma^2}{n}$. Si può verificare, tramite il teorema di Cramer Rao, che il valore minimo della varianza per stimatori corretti di μ è proprio $\frac{\sigma^2}{n}$. Quindi non solo nessun altro stimatore corretto può essere più efficiente, ma nemmeno avere la stessa efficienza.*

Infatti se confrontiamo la mediana campionaria con quello stimatore (vedi Piccolo, 2000 pag. 506) si può dimostrare che, per popolazioni normali, asintoticamente $M_{sc}(Med) = \mu$ e $V_{sc}(Med) = \frac{\sigma^2}{n}\frac{\pi}{2}$.

Quindi anche la mediana campionaria è uno stimatore (sia pure asintoticamente) corretto ma più variabile (e quindi meno efficiente) della media campionaria, dato che $\frac{\sigma^2}{n}\frac{\pi}{2} > \frac{\sigma^2}{n}$ per qualsiasi n finito.

distribuzione probabilistica

La distribuzione probabilistica degli stimatori possiede informazioni che se da una parte sono essenziali per le stime intervallari, dall'altra permettono di avere una conoscenza più dettagliata delle relazioni che legano la popolazione e gli stimatori.

Per le necessità elementari sono sufficienti, come vedremo tra poco, le tre distribuzioni: normale standardizzata, chi quadrato e t di Student. Le tavole disponibili permettono, infatti, di calcolare le probabilità per gli stimatori ed in generale per le statistiche campionarie che si distribuiscono secondo quelle variabili.

Esempio 2.8 *Si possono trovare svariati esempi di distribuzioni di stimatori campionari nel paragrafo 1.1.*

Talvolta però non si conosce la formula probabilistica della distribuzione ma se ne conoscono media e varianza. In tal caso, qualora, data l'esiguità del campione, non si possa ricorrere al Teorema Centrale del Limite (vedi teorema 1.7), si può utilizzare il teorema di Cebiceff per avere un valore limite della probabilità :

Teorema 2.2 *Per variabili casuali X con densità $p(x;\mu,\sigma)$ e media e varianza finiti*

$$\Pr\left(|x-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

o anche nella formulazione equivalente

$$\Pr\left(|x-\mu| < k\sigma\right) > 1-\frac{1}{k^2}$$

Da queste se l è uno stimatore corretto di λ e σ_l il suo scarto quadratico medio le disuguaglianze divengono

$$\Pr\left(|l-\lambda| \geq k\sigma_l\right) \leq \frac{1}{k^2} \tag{2.1}$$

$$\Pr\left(|l-\lambda| < k\sigma\right) > 1-\frac{1}{k^2} \tag{2.2}$$

Esempio 2.9 *Confrontiamo le probabilità che si ottengono per un intervallo simmetrico rispetto alla media ricorrendo al teorema di Cebiceff ed alla distribuzione probabilistica della variabile.*

Sia X una variabile casuale normale con $\mu = 500$ e $\sigma = 50$. Si estraggono campioni casuali semplici di 10 unità e si vuole conoscere $\Pr(470 < \bar{X} \leq 530)$. Con la distribuzione normale si ottiene:

$$\Pr(470 < \bar{X} \leq 530) = \Pr\left(\frac{470-500}{\frac{50}{\sqrt{10}}} < \bar{X} \leq \frac{530-500}{\frac{50}{\sqrt{10}}}\right)$$

$$\Pr(-1.90 < Z \leq 1.90) = F_Z(1.90) - F_Z(-1.90)$$

Per usare le tavole che danno la $F_Z(z)$ solo per $Z \geq 0$ $F_Z(-z_0) = 1 - F_Z(-z_0)$ e quindi

$$F_Z(1.90) - F_Z(-1.90) = F_Z(1.90) - 1 + F_Z(1.90) =$$

$$= 2 * F_Z(1.90) - 1 = 2 * 0.9713 - 1 = 0.9426$$

Ricorrendo al teorema di Cebiceff si ottiene invece

$$\Pr(470 < \bar{X} \leq 530) = \Pr(|\bar{X} - 500| < 30)$$

dato che $k = 30/\frac{50}{\sqrt{10}} = 1.90$

$$\Pr(|\bar{X} - 500| < 30) = \Pr(|\bar{X} - 500| < 1.90\frac{50}{\sqrt{10}}) =$$

$$= \Pr(|\bar{X} - 500| < 1.90\sigma_{\bar{X}}) \geq 1 - \frac{1}{1.90^2} = 0.7230$$

Verifichiamo quindi che, ignorando la forma della distribuzione dello stimatore, riusciamo con il teorema di Cebiceff ad avere soltanto un limite inferiore ($\Pr \geq 0.7230$) *della probabilità relativa all'intervallo assegnato, mentre con l'informazione della distribuzione possiamo dare la probabilità esatta* ($\Pr = 0.9426$)*.*

comportamento asintotico

Una prima proprietà asintotica interessante è la **consistenza**. Si dice allora che l **converge in probabilità** a λ. Una statistica campionaria l è uno **stimatore consistente** di λ se $\forall \epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(|l - \lambda| \leq \epsilon\right) = 1$$

Questa relazione, che non è sempre agevole da dimostrare, per gli stimatori corretti può essere sostituita da quest'altra che le è equivalente:

$$\lim_{n\to\infty} V_{sc}(l) = 0$$

Esempio 2.10 *La media campionaria è stimatore corretto di* μ *e dato che*

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

è anche consistente.

Tra le molteplici proprietà asintotiche possiamo evidenziarne un'altra che ci risulta particolarmente utile: la **convergenza in distribuzione** [2].

Per certe statistiche campionarie la distribuzione probabilistica per n finito non è disponibile con precisione oppure è molto complessa da trattare, se ne dispone al contrario della distribuzione limite per $n \to \infty$; il **Teorema Centrale del Limite** è un esempio di come possa essere trattata questa convergenza. A questo Teorema è stato dato l'attributo di **centrale** dato che (grazie soprattutto a versioni più generali di quella mostrata nel Teorema 1.7), permette di risolvere casi di distribuzioni di statistiche campionarie altrimenti di difficile soluzione.

Possiamo infatti dire che per popolazioni non normali la distribuzione della media campionaria con n finito non è facilmente trattabile matematicamente, ma che al divergere di n si avvicina sempre più alla gaussiana.

[2] La problematica della convergenza in distribuzione è trattata per esempio nei paragrafi 3.4.1, 3.4.2 e 3.5.2 di Prisco, 2006.

Altra proprietà molto importante è la **correttezza asintotica** della quale godono quegli stimatori la cui media è uguale al parametro da stimare soltanto per $n \rightarrow \infty$.

Esempio 2.11 *La varianza campionaria, nella formula 1.7 è uno stimatore asintoticamente corretto di σ^2 in quanto al divergere di n la sua media è uguale alla varianza della popolazione.*

Passiamo ora a trattare gli stimatori campionari, che nell'ambito della teoria dei campioni abbiamo studiato come statistiche campionarie. È importante qui ricordare come nell'ambito della teoria dei campioni la statistica campionaria sia considerata come una variabile casuale e la popolazione come nota; al contrario nella teoria dell'inferenza il campione, e quindi le statistiche campionarie sono considerati come noti ed è invece la popolazione (con i valori dei suoi parametri) ad essere trattata come non nota.

2.1.2 I principali stimatori

Frequenza relativa campionaria

Data una popolazione nella quale la probabilità di estrarre una unità che possiede la caratteristica A è uguale a π ed è costante in ognuna delle n estrazioni, la statistica campionaria X numero di elementi del campione che possiedono quella caratteristica si distribuisce come una binomiale con parametri n e π. Come conseguenza immediata la statistica $p = \frac{X}{n}$ gode di varie proprietà tra le quali sono importanti ai fini della stima puntuale le seguenti (vedi formule 1.1):

$$M_{sc}(p) = \pi \tag{2.3}$$

$$V_{sc}(p) = \frac{\pi(1-\pi)}{n} \tag{2.4}$$

considerando la statistica campionaria p come uno stimatore di π ne evidenziamo le proprietà seguendo lo schema appena visto:

- è corretto
- è efficiente
- per campioni di piccole dimensioni la sua distribuzione è una trasformata lineare della binomiale, per campioni di grandi dimensioni la distribuzione è normale
- è consistente

Queste proprietà valgono anche per estrazioni senza reinserimento avendo l'avvertenza di tenere conto della diversa formula della varianza dello stimatore che è $V_{sc}(p) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}$ e la distribuzione per campioni finiti è una trasformata lineare dell'Ipergeometrica.

Media campionaria

Data una popolazione, nella quale una certa grandezza X ha media μ e varianza σ^2, la statistica media aritmetica campionaria

$$\bar{X} = \frac{\sum X_i}{n}$$

è uno stimatore di μ, ed essendo per estrazioni con reinserimento

$$\begin{array}{rcl} M_{sc}(\bar{X}) & = & \mu \\ V_{sc}(\bar{X}) & = & \frac{\sigma^2}{n} \end{array}$$

quest'ultima per le estrazioni senza reinserimento è invece

$$V_{sc}(\bar{X}) \quad = \quad \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

possiamo concludere che lo stimatore

- è corretto
- ha efficienza nota
- per campioni di piccole dimensioni la sua distribuzione è difficile da trattare a meno che la popolazione d'origine sia normale, nel qual caso anche la $\bar{X}$ è distribuita normalmente; per campioni di grandi dimensioni la distribuzione è tendenzialmente normale
- è consistente

Quindi la media campionaria è uno stimatore della media della popolazione che non è stravagante, dato che la sua media coincide con il valore incognito da stimare ed essendo efficiente è da preferire ad altri stimatori corretti. Inoltre il valore μ della media della popolazione è quello che con maggiore probabilità viene o scelto o per lo meno approssimato da campioni di grandi dimensioni.

Varianza campionaria

Data una popolazione con varianza σ^2 la statistica varianza campionaria corretta

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

ottenuta da n unità campionarie estratte con reinserimento è tale che

$$M_{sc}(S^2) = \sigma^2$$

$$V_{sc}(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

nella quale $\mu_4 = M(X - \mu)^4$
ed inoltre se la popolazione è normale ed allora $\mu_4 = 3\sigma^4$

$$V_{sc}(S^2) = \frac{2\sigma^4}{n-1}$$

.

Da questi risultati si ricava che questo stimatore della varianza della popolazione

- è corretto
- ha efficienza nota
- se la popolazione d'origine è normale, la quantità $\frac{S^2(n-1)}{\sigma^2}$ si distribuisce come un χ^2 con $\nu = n - 1$ gradi di libertà
- è consistente

2.1.3 Stima per intervallo

Gli stimatori puntuali hanno il pregio di fornire un valore che potrebbe essere quello del parametro e, se godono di buone proprietà, possiamo credere che sia una buona stima. Non possiamo tuttavia escludere che il valore vero possa essere anche distante. In particolare l'efficienza dello stimatore, dopo che ha svolto la funzione di criterio per la scelta dello stimatore, esce dalla procedura di stima; nelle stime per intervallo al contrario la variabilità campionaria dello stimatore fa parte integrante della procedura di stima attraverso la distribuzione dello stimatore.

Sulla base della distribuzione probabilistica dello stimatore si può argomentare che valori molto distanti siano poco probabili, ma basandoci sulla sola stima del valore non si è in grado di fissare dei limiti che definiscano cosa è troppo distante per essere creduto. Per fissare questi limiti è necessario passare ad una tecnica diversa, quella della **stima per intervallo**. Mentre alcune delle proprietà che abbiamo visto per gli stimatori puntuali possono essere dimostrate anche nell'ignoranza della distribuzione probabilistica della statistica campionaria (ad esempio la correttezza) la stima per intervallo richiede come essenziale che questa sia conosciuta.

I limiti dell'intervallo di stima (detto anche *intervallo di confidenza* o meglio *intervallo fiduciario*) sono probabilistici dato che il valore della stima dipende dal particolare campione (di n elementi) che è stato estratto ed il campione è stato formato con criteri casuali. La tecnica della stima per intervallo ha lo scopo di fissare un intervallo entro il quale si trovi con la probabilità assegnata il valore vero del parametro. Se questa probabilità, che di solito ha come simbolo $1-\alpha$, è vicina ad uno si ottiene un intervallo che contiene quasi certamente il parametro. La probabilità di estrarre un campione il cui intervallo non contenga il parametro è α e quindi piccola, ma quell'intervallo è grande e quindi poco utile. Se al contrario α è grande si produce un intervallo stretto e quindi utile, ma con una maggiore probabilità di sbagliare.

Si abbia una popolazione e si voglia stimare un parametro incognito λ che la caratterizza. Una statistica campionaria l è un suo stimatore. Esiste una sua funzione $u = f(l, \lambda, n, ..)$ della quale si è in grado di calcolare la funzione di ripartizione

$$F(u) = P(U \leq u)$$

dopo aver fissato α si possono individuare molti, se non infiniti intervalli per i quali

$$\Pr(u_1 < U \leq u_2) = 1 - \alpha$$

di solito però si formano degli intervalli probabilisticamente simmetrici e quindi si trovano univocamente i valori u_1 e u_2 tali che $F(u_1) = \alpha/2$ e $F(u_2) = 1 - \alpha/2$; da questa doppia disequazione si ricavano con passaggi algebrici le due

$$\Pr\left[f_1(u_1, u_2, l, n, ...) < \lambda \leq f_2(u_1, u_2, l, n, ...)\right] = 1 - \alpha$$

nella quale le funzioni $f_i(\Box)$ non dipendono da λ

In questo modo si possono calcolare degli intervalli che con probabilità prefissata contengono il valore del parametro della popolazione. Vediamo ora come si formino in particolare gli intervalli di stima per i più usati parametri delle popolazioni.

Nell'ipotesi di validità delle distribuzioni delle singole statistiche campionarie abbiamo:

media

$$\Pr\left(z_1 \leq Z \leq z_2\right) = 1 - \alpha$$

$$\Pr\left(z_1 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_2\right) = 1 - \alpha \tag{2.5}$$

$$\Pr\left(z_1 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_2 \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Pr\left(-\bar{X} + z_1 \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_2 \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Pr\left(\bar{X} - z_2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_1 \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

poiché si usano intervalli di confidenza probabilisticamente simmetrici e dato che la Z è normale si può vedere facilmente che $z_1 = -z_2$ e quindi quest'ultima espressione diventa

$$\Pr\left(\bar{X} - z_2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_2\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \qquad (2.6)$$

il valore più comune di α è 0.05 e per ciò

$$\Pr\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Esempio 2.12 *Un campione di 10 unità estratto da una popolazione distribuita normalmente secondo il carattere X, ha dato un campione con $\bar{x} = 200$. Essendo la varianza della popolazione $\sigma^2 = 900$ si dia una stima per intervallo della media della popolazione con $\alpha = 0.05$.*

La media campionaria si distribuisce normalmente e quindi:

$$\Pr\left(200 - 1.96\frac{30}{\sqrt{10}} < \mu \leq 200 + 1.96\frac{30}{\sqrt{10}}\right) = 1 - 0.05$$

$$\Pr(181,41 < \mu \leq 218.59) = 0.95$$

Dato che il campione è di piccole dimensioni, se ignoriamo che la popolazione (o sappiamo che non) si distribuisce normalmente dobbiamo ricorrere alla disuguaglianza di Cebiceff. Elemento essenziale di questa è k; vediamo per prima cosa di calcolare quanto vale.

$$0.95 = 1 - \frac{1}{k^2} \to k^2 = \frac{1}{0.05} \to k = 4.47$$

Possiamo quindi scrivere

$$\Pr(|\bar{X} - \mu| \leq 4.47\sigma_{\bar{X}}) \geq 1 - \frac{1}{k^2}$$

$$\Pr(|\bar{X} - \mu| \leq 4.47\sigma_{\bar{X}}) \geq 0.95$$

e quindi

$$\Pr\left(200 - 4.47\frac{30}{\sqrt{10}} < \mu \leq 200 + 4.47\frac{30}{\sqrt{10}}\right) \geq 0.95$$

$$\Pr(157.59 < \mu \leq 242.41) \geq 0.95$$

Dal confronto tra i due intervalli si rileva come quello ottenuto con la conoscenza della distribuzione del carattere nella popolazione è più piccolo e quindi più utile dell'altro che risulta garantito probabilisticamente in modo forse eccessivo.

Questo nella situazione un po' artificiale nella quale si pone di conoscere la varianza della popolazione e di ignorarne la media. In una trattazione più realistica si pone di ignorare anche la varianza ed allora si fa riferimento alla distribuzione T dato che non sussitono le condizioni per l'utilizzo della Z. Sotto queste altre condizioni è necessario ricorrere a

$$\Pr\left(t_1 \leq T \leq t_2\right) = 1 - \alpha$$

che per il problema specifico diventa

$$\Pr\left(t_1 \leq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_2\right) = 1 - \alpha \tag{2.7}$$

da cui

$$\Pr\left(\bar{X} - t_2 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} - t_1 \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

e poi, valendo ancora le proprietà di simmetria per cui $t_1 = -t_2$, si ottiene

$$\Pr\left(\bar{X} - t_2 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_2 \frac{s}{\sqrt{n}}\right) = 1 - \alpha \tag{2.8}$$

I valori di t_2 (con $n - 1$ gradi di libertà) dipendono dalla numerosità del campione e si ricavano dalle tavole della distribuzione di t di Student e sono tali che:

$$\Pr\left(t_2 \leq T\right) = 1 - \frac{\alpha}{2}.$$

.

Esempio 2.13 *Un campione casuale di 12 elementi ha dato come media 320 e varianza non corretta 5200; si dia un intervallo di stima ($\alpha = 0.05$) per la media della popolazione sapendo che la popolazione è distribuita normalmente riguardo a quel carattere.*

Per prima cosa si deve trasformare la varianza di quel campione nella varianza corretta

$$s^2 = 5200 \frac{12}{11} = 5672,73$$

poi si trovano i valori dei percentili della distribuzione della t con 11 gradi di libertà. Dalle tavole si trova che $t_{0.975} = 2.201$, quindi

$$\Pr\left(320 - 2.201\sqrt{\frac{5672.73}{12}} < \mu \leq 320 + 2.201\sqrt{\frac{5672.73}{12}}\right) \geq 0.95$$

$$\Pr(272.15 < \mu \leq 367.85) \geq 0.95$$

frequenza relativa

$$\Pr\left(z_1 \leq Z \leq z_2\right) = 1-\alpha$$
$$\Pr\left(z_1 \leq \frac{p-\pi}{\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}} \leq z_2\right) = 1-\alpha$$

da cui ripetendo i soliti passaggi

$$\Pr\left(p - z_2\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \leq \pi \leq p + z_2\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}\right) = 1-\alpha \qquad (2.9)$$

Non è quindi possibile calcolare gli estremi di questi intervalli in quanto contengono il parametro che deve essere stimato. Per superare questo inconveniente si può scegliere tra queste due possibii soluzioni.

- si sostituisce negli estremi dell'intervallo la π con il valore 0.5 che produce l'intervallo più ampio

$$\Pr\left(p - z_2\frac{0.5}{\sqrt{n}} \leq \pi \leq p + z_2\frac{0.5}{\sqrt{n}}\right) \geq 1-\alpha \qquad (2.10)$$

  che quindi fornisce una protezione maggiore di $1-\alpha$

- si ricorre alla formula approssimata (per una dimostrazone vedi ad esempio Cicchitelli, 1984 pag. 176)

$$\Pr\left(p - z_2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \pi \leq p + z_2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha \qquad (2.11)$$

Esempio 2.14 *Sono state estratte casualmente senza reinserimento 100 unità che sottoposte a misurazione dicotomica hanno dato 45 esiti favorevoli alla modalità cercata. Dare un intervallo al 95% per la probabilità nella popolazione.*

$$\Pr\left(0.45 - 1.96\sqrt{\frac{0.45*0.55}{100}} < \pi \leq 0.45 + 1.96\sqrt{\frac{0.45*0.55}{100}}\right) = 1-0.05$$

$$\Pr(0.35 < \pi \leq 0.55) = 0.95$$

Esempio 2.15 *Un campione di 85 unità estratto senza reinserimento da una popolazione di 2800 unità ha dato 34 esiti favorevoli; dare un intervallo di stima per la proporzione (o probabilità) nella popolazione di quella modalità.*

$$p = \frac{34}{85} = 0.40$$

Dato che le estrazioni sono state fatte senza reinserimento da una popolazione di numerosità non elevata la varianza della frequenza relativa campionaria è

$$\sigma_P^2 = \frac{\pi(1-\pi)}{n} * \frac{N-n}{N-1}$$

$$\Pr\left(0.4 - 1.96\sqrt{\frac{0.4 * 0.6}{85}\frac{2715}{2799}} < \pi \leq 0.4 + 1.96\sqrt{\frac{0.4 * 0.6}{85}\frac{2715}{2799}}\right) = 1 - 0.05$$

$$\Pr(0.30 < \pi \leq 0.50) = 0.95$$

varianza

Lo stimatore varianza campionaria opera attraverso la variabile casuale χ^2 secondo la relazione

$$\chi^2 = \frac{s^2\nu}{\sigma^2}$$

nella quale $\nu = n - 1$ sono i gradi di libertà. Quindi si pone l'intervallo relativo alla variabile χ^2

$$\Pr\left(\chi_1^2 < \chi^2 \leq \chi_2^2\right) = 1 - \alpha$$

con

$$\Pr\left(\chi^2 < \chi_1^2\right) = \frac{\alpha}{2}$$

e

$$\Pr\left(\chi^2 < \chi_2^2\right) = 1 - \frac{\alpha}{2}$$

da cui

$$\Pr\left(\chi_1^2 < \frac{s^2\nu}{\sigma^2} \leq \chi_2^2\right) = 1 - \alpha$$

da cui eseguendo semplici passaggi

$$\Pr\left(\frac{s^2\nu}{\chi_2^2} \leq \sigma^2 < \frac{s^2\nu}{\chi_1^2}\right) = 1 - \alpha \qquad (2.12)$$

Esempio 2.16 *Da un campione di 30 unità estratte da una popolazionenormalmente distribuita secondo il carattere X, si è ottenuta la varianza corretta $s^2 = 3900$; dare un intervallo ($\alpha 0.05$) per la varianza della popolazione.*

$$\Pr\left(\frac{3900 * 29}{45.722} \leq \sigma^2 < \frac{3900 * 29}{16.047}\right) = 1 - 0.05$$

$$\Pr\left(2473.645 \leq \sigma^2 < 7048.046\right) = 1 - 0.05$$

2.2 Metodi per la costruzione degli stimatori

La forma probabilistica della popolazione è l'ingrediente principale per l'uso dei metodi (ma sarebbe più corretto chiamarli criteri) utili per la costruzione degli stimatori. Con questi criteri si selezionano gli stimatori dei parametri di interesse che soddisfano certe condizioni. I criteri sono tre:

- Minimi quadrati
- Momenti
- Massima verosimiglianza

Del primo di questi criteri sono state viste le proprietà principali nello studio della stima dei modelli di regressione che solitamente viene fatta nel corso dei fondamenti della Statistica e quindi qui non ci dilunghiamo su questo criterio dato che i modelli a più variabili non vengono affrontati in questa rassegna elementare dell'inferenza statistica.

Del secondo possiamo dire che non ha più molta importanza in quanto è stato soppiantato nella pratica statistica dal terzo, ci limitiamo qui a dire che si basa sull'uguaglianza tra momenti del campione e momenti della popolazione.

Esempio 2.17 *Una popolazione di tempi di attesa si distribuisce come una esponenziale negativa. La densità di probabilità è*

$$p(x) = \alpha e^{-\alpha x}$$

con

$$\mu = \frac{1}{\alpha}; \, \sigma^2 = \frac{1}{\alpha^2}$$

essendo la media una funzione di α allora lo stimatore di α è una funzione della media campionaria, più esattamente lo stimatore è $\frac{1}{\bar{x}}$.

2.2.1 stimatori di massima verosimiglianza

Questa procedura (proposta da R. A. Fisher negli anni venti) parte dalla probabilità del campione di n elementi $(x_1, x_2, \cdots x_i, \cdots x_n)$, calcolata sulla base del modello probabilistico della popolazione considerata come funzione dei parametri $(\lambda_1, \lambda_2 \cdots)$. Se il campione è estratto con reinserimento:

$$\Pr(x_1, x_2, \cdots x_i, \cdots x_n | \lambda_1, \lambda_2 \cdots) = \prod_{i=1}^{n} p(x_i; \lambda_1, \lambda_2 \cdots)$$

Esempio 2.18 *Da una popolazione distribuita come una Poisson si è ottenuto un campione di 5 unità formato delle unità* 0, 1, 0, 2, 1; *la probabilità di ottenere quel campione è*

$$\prod_{i=1}^{n} p(x_i;\lambda) = \frac{e^{-\lambda}\lambda^0}{0!}\frac{e^{-\lambda}\lambda^1}{1!}\frac{e^{-\lambda}\lambda^0}{0!}\frac{e^{-\lambda}\lambda^2}{2!}\frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-5\lambda}\lambda^4}{2!}$$

In generale per popolazioni distribuite come la Poisson

$$\prod_{i=1}^{n} p(x_i;\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\cdots\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\cdots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod x_i!}$$

in generale la probabilità di ottenere un campione generico è

$$\Pr(x_1, x_2, \cdots x_i, \cdots x_n|\lambda) = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod x_i!}$$

La probabilità del campione è costruita nell'ambito della teoria dei campioni e considera le unità del campione come variabili casuali ed i parametri della popolazione come dati. L'uso di questa quantità nella teoria della stima prevede che si scambi il significato di queste quantità e si considerino certi i valori assunti dalle unità del campione ed incerti i valori dei parametri. [3] Dopo questo rovesciamento di ruoli questa funzione non è più una probabilità e nemmeno una densità di probabilità, ma una **verosimiglianza** e si rappresenta come una L (dall'iniziale della parola inglese *Likelihood*).

$$L(\lambda_1, \lambda_2 \cdots |x_1, x_2, \cdots x_i, \cdots x_n) =\propto \prod_{i=1}^{n} p(x_i; \lambda_1, \lambda_2 \cdots) \tag{2.13}$$

Il **principio di verosimiglianza** dice che la credibilità di un certo valore del parametro λ è proporzionale ($\propto$) alla quantità scritta a secondo membro della formula 2.13. Da questo si ricava il precetto pratico: *tra i possibili valori del parametro λ, si preferisce quello che corrisponde alla massima probabilità di generare i dati osservati* (vedi Piccolo, 2000 pag. 587 ma soprattutto Azzalini, 2001). Poi si trascura essendo non rilevante il coefficiente non sempre ben definito di proporzionalità.

Non solo per motivi pratici di semplificazione della funzione, ma anche per convenienze teoriche per le quali rinviamo al testo di Azzalini, la funzione di verosimiglianza viene sostituita dal suo logaritmo naturale

$$\ell(\lambda_1, \lambda_2 \cdots |x_1, x_2, \cdots x_i, \cdots x_n) = \ln[L(\lambda_1, \lambda_2 \cdots |x_1, x_2, \cdots x_i, \cdots x_n)] \tag{2.14}$$

diciamo intanto che per lo meno viene semplificata il trattamento matematico della funzione.

[3] Esula dalle nostre necessità di esposizione elementare vedere i riferimenti al teorema di Bayes che per la applicazione alla teoria classica dell'inferenza viene *castrato* delle probabilità iniziali.

Esempio 2.19 *Riprendiamo l'esempio del campione tratto da una popolazione poissoniana*

$$L(\lambda|x_1, x_2, \cdots x_i, \cdots x_n) = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\Pi x_i!}$$

da cui

$$\ell = \ln(L) = -n\lambda + \Sigma x_i \ln(\lambda) - \ln(\Pi x_i!)$$

ed in relaziona al campione

$$L(\lambda|0,\, 1,\, 0,\, 2,\, 1) = \frac{e^{-5\lambda}\lambda^4}{2!}$$

passando ai logaritmi

$$\ell(\lambda|0,\, 1,\, 0,\, 2,\, 1) = -5\lambda + 4\ln(\lambda) - \ln(2)$$

L'applicazione del criterio della massima verosimiglianza richiede quindi sia da preferire come stimatore la statistica campionaria che rende massima la funzione ℓ, infatti essendo la logaritmica una trasformata monotona crescente in senso stretto non altera gli ordinamenti, e quindi il valore delle variabili indipendenti che rendono massima la ℓ rendono massima anche la L.

Sotto ampie condizioni, soddisfatte da tutti i modelli probabilistici più diffusi, il massimo si può trovare cercando il valore del parametro che annulla la derivata prima di ℓ.

Esempio 2.20 *Restiamo sul campione da popolazione poissoniana*

$$\ell(\lambda|0,\, 1,\, 0,\, 2,\, 1) = -5\lambda + 4\ln(\lambda) - \ln(2)$$

$$\frac{d\ell}{d\lambda} = -5 + \frac{4}{\lambda} = 0$$

da cui

$$\lambda = \frac{5}{4}$$

ed in generale

$$\frac{d\ell}{d\lambda} = -n + \frac{\Sigma x_i}{\lambda} = 0$$

$$\lambda = \frac{\Sigma x_i}{n} = \bar{X}.$$

La media campionaria è lo stimatore di massima verosimiglianza del parametro λ di una popolazione poissoniana.

In altri casi nei quali la densità del modello probabilistico della popolazione dipende da due o più parametri è necessario ricorrere al sistema delle derivate parziali per individuare il massimo della funzione di log-verosimiglianza. I casi più semplici sono dati dalla distribuzione normale che dipende da due parametri (μ e σ^2), dal modello di regressione che dipende da tre (α, β e σ^2) e da quello della normale doppia che dipende da cinque (μ_X, σ_X^2, μ_Y, σ_Y^2 e ρ_{XY}).

Esempio 2.21 *Troviamo gli stimatori di massima verosimiglianza dei parametri di una popolazione normale, usando i valori di un campione di n unità estratto con campionamento casuale semplice. Si tenga presente che la varianza è un parametro in quanto tale e che quindi il simbolo σ^2 non deve essere considerato come un quadrato.*

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$L = \left(\frac{1}{\sqrt{\sigma^2 2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

$$\ell = n \ln \left((\sigma^2 2\pi)^{-1/2} \right) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ell}{\partial \mu} = 0 - \frac{1}{2\sigma^2} 2 \sum (x_i - \mu)(-1) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = -n \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} \sqrt{2\pi} \frac{1}{2} \frac{1}{\sqrt{\sigma^2}} - \sum (x_i - \mu)^2 \frac{1}{2} \frac{-1}{(\sigma^2)^2} = 0$$

Bisognerebbe risolvere il sistema formato di entrambe queste derivate, ma è sufficiente risolverle ad una ad una.

Dalla prima

$$\frac{\partial \ell}{\partial \mu} = \frac{\sum (x_i - \mu)}{\sigma^2} = 0$$

ed in fine per ben nota proprietà della media aritmetica

$$\mu = \frac{\sum x_i}{n} = \bar{X}$$

Per quanto riguarda la varianza

$$\frac{\partial \ell}{\partial \sigma^2} = -n \frac{1}{\sqrt{\sigma^2}} \frac{1}{2} + \frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

$$\frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2} = \frac{n}{2\sigma^2}$$

$$\frac{\sum(x_i - \mu)^2}{\sigma^2} = n$$

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n} = S_2^2$$

Quindi lo stimatore di massima verosimiglianza della varianza di una distribuzione normale non è corretto.

proprietà degli stimatori di massima verosimiglianza

Gli stimatori di massima verosimiglianza sono molto importanti per le proprietà di cui godono. Infatti anche se non sono necessariamente corretti sono asintoticamente corretti in quanto stimatori consistenti. Se esiste uno stimatore corretto, che raggiunge il limite inferiore della varianza come indicato dalla disuguaglianza di Cramer-Rao, è anche di massima verosimiglianza e quindi gli stimatori che soddisfano questo criterio sono asintoticamente efficienti. Ultimo aspetto positivo, sia pure valido anche questo al divergere di n, la distribuzione dello stimatore è normale.

Per modelli complicati talvolta caratterizzati da un numero elevato di parametri gli stimatori vengono trattati implicitamente. In questi casi le stime vengono ottenute con procedimenti di ottimizzazione.

Gli stimatori di massima verosimiglianza godono della proprietà di **invarianza**, cioè se l è uno stimatore di λ allora $\tau(l)$ è uno stimatore di $\tau(\lambda)$

Capitolo 3

Test di ipotesi

3.1 Osservazioni preliminari

Nelle procedure di stima, come abbiamo visto, lo statistico si pone nella situazione di avere poche o anche nulle informazioni sul valore del parametro che è oggetto dello studio e, attraverso gli stimatori dei quali conosce le proprietà, ottiene informazioni o sul valore più credibile (stima del valore) o su un intervallo di valori che ha una elevata probabilità di contenere quel parametro (stime per intervallo). Ben diversa è la situazione in cui si trova lo statistico che usa un test di ipotesi [1].

In quest'altra situazione possiede informazioni dalle quali si aspetta che il parametro dovrebbe assumere un certo valore. Allora non opera più in quella quasi completa mancanza di informazioni, ma avendo lo scopo ridotto di dover semplicemente accertare se quel valore ipotizzato è vero o falso.

In questa trattazione elementare prendiamo in esame soltanto le statistiche campionarie la cui media è uguale al parametro da conoscere. Se il campione estratto per verificare l'ipotesi fornisce una statistica campionaria il cui valore è vicino a quello dell'ipotesi, allora tutti siamo indotti ad affermare che quell'ipotesi è accettabile. Ma se la statistica campionaria assume un valore molto distante allora sarà giocoforza concludere che quell'ipotesi è falsa. L'inferenza statistica cerca di fissare i limiti di questo confronto in modo da rendere metodologicamente corretta la definizione di quali differenze siano *piccole* e di quali siano *troppo grandi*.

Esempio 3.1 *Alcuni dei residenti nel Comune di Xxxxx leggono abitualmente il quotidiano L'Araldo della Vittoria.*

[1] La parola ipotesi viene usata in matematica e nei test statistici con significati diversi. In matematica è una affermazione valida, solo ai fini del teorema ma comunque vera ed indubitabile, nei test statistici al contrario è si una affermazione valida soltanto per il problema ma della quale si cerca di sapere se può reggere il confronto con l'esito del campione.

Questo fatto, tradotto nel linguaggio statistico viene descritto in questo modo: in quel comune ed in relazione a quel giornale è presente, con una frequenza relativa incognita, la caratteristica di essere lettore abituale*. La popolazione di quel comune, tenendo presente soltanto questa caratteristica, è descrivibile quindi con uno schema estremamente semplice, infatti Ω è ripartibile nei due sottoinsiemi A_1 dei lettori ed A_2 dei non lettori; essendo $card(\Omega) = N$; $card(A_1) = a$; $card(A_2) = N - a$, la frequenza relativa dei lettori in questa popolazione vale $\pi = a/N$. Sulla base di informazioni relative alla diffusione ed a conoscenze relative agli anni precedenti viene avanzata l'ipotesi che questo parametro valga* 0.10.

Dalla teoria dei campioni sappiamo che la frequenza relativa campionaria p, per campioni casuali semplici, si distribuisce, per n abbastanza elevato come una normale con $M(p) = \pi$ e $V(p) = \frac{\pi(1-\pi)}{n}$. Se da un campione di 200 *residenti si ricava una frequenza relativa in quel campione pari a* 0.09 *saremo indotti a ritenere che l'ipotesi possa essere vera, se invece la statistica di quel campione valesse* 0.20 *ci sentiremmo di dire che forse l'ipotesi non era corretta; se invece si ottenesse una frequenza relativa pari a* 0.15*? o a* $0,14$*? Scopo operativo dei test di ipotesi è risolvere questi casi sfumati per i quali l'approccio non formalizzato non sarebbe sufficiente a dare una risposta.*

Con i test di ipotesi non si cerca, quindi, di stimare il valore di un parametro della popolazione ma di rispondere alla domanda: *il parametro può assumere questo valore?* Se la misurazione fosse eseguita su tutte le unità della popolazione avremmo una valutazione e sapremmo in modo certo se il valore ipotizzato corrisponde a quello ottenuto dal calcolo fatto sulle misure. Ma, operando attraverso le statistiche campionarie, ci confrontiamo con una variabilità di valori dovuta al fatto che da una popolazione possono essere estratti campioni di composizione anche molto varia. Per alcuni campioni la statistica campionaria assume un valore vicino a quello della popolazione per altri è distante, e talvolta anche di molto. La **teoria dei test** tiene conto proprio della variabilità campionaria.

Si tratta in buona sostanza di rispondere a quest'altra domanda: *la differenza riscontrata tra il valore che si ipotizza per la popolazione e quello che si ottiene nel campione è attribuibile alla sola estrazione campionaria?*

Essendo il risultato dell'estrazione campionaria incerto è quindi tuttalpiù soggetto alle elaborazioni probabilistiche e la risposta non potrà che essere incerta, eventualmente corredata della probabilità di aver preso una decisione giusta. Come vedremo, una caratteristica fondamentale dei test di ipotesi è la loro fallibilità, non soltanto perché, come insegna la saggezza popolare, tutti possono sbagliare, ma perché è la procedura stessa a comportare un margine ineliminabile di errore, presente anche se si sono seguite tutte le regole, le attenzioni e le cautele, che la Teoria dell'Inferenza Statistica e l'esperienza suggeriscono.

3.2 Aspetti terminologici

La procedura dei test statistici parte da una descrizione probabilistica della popolazione dalla quale si ricava con le tecniche della Teoria dei Campioni la distribuzione campionaria di una statistica campionaria stimatore del parametro incognito. Quest'ultimo definisce un aspetto della popolazione stessa e sul suo valore si vuole avere una maggiore informazione, che nella fattispecie dei test consiste nel sapere se il campione è per così dire compatibile probabilisticamente col valore del parametro che è stato fissato nell'ipotesi.

La **Teoria dei Test** ha lo scopo di definire quali valori della statistica campionaria sostengono l'ipotesi fatta e quali al contrario la smentiscono. L'impiego della Teoria nel caso pratico consiste poi nel verificare se in quel caso la statistica campionaria ha assunto un valore che sostiene o al contrario porta a rigettare l'ipotesi principale.

Sulla popolazione vengono, quindi, poste due ipotesi complementari e disgiunte, delle quali cioè una deve essere vera e l'altra falsa. La tecnica del test statistico ha lo scopo di far scegliere una delle due. La prima che solitamente comprende la relazione di uguaglianza e descrive ciò che si sa (o si ritiene di sapere) già prima dell'estrazione del campione viene detta **ipotesi base** oppure **ipotesi di nullità** o anche più diffusamente ma impropriamente **ipotesi nulla** ed ha come simbolo H_0. L'altra ipotesi viene detta **ipotesi alternativa** ed ha come simbolo H_1.

In generale le ipotesi assumono la forma

$$\begin{aligned} H_0 &: \lambda = \lambda_0 \\ H_1 &: \lambda \neq \lambda_0 \end{aligned}$$

che per i parametri che abbiamo visto sono:

$$\begin{aligned} H_0 &: \pi = \pi_0 \\ H_1 &: \pi \neq \pi_0 \end{aligned}$$

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0^2 \\ H_1 &: \sigma^2 \neq \sigma_0^2 \end{aligned}$$

In queste espressioni $\pi = \pi_0,\ \mu = \mu_0,\ \sigma^2 = \sigma_0^2$ sono le proposizioni la cui verità si cerca di saggiare mediante la procedura.

Esempio 3.2 *A proposito del problema della diffusione dell'Araldo della Vittoria si tratta di verificare se l'ipotesi fatta sulla proporzione di lettori nella popolazione* (10%) *regge se confrontata con gli esiti di una indagine campionaria.*

La procedura del test di ipotesi, allo scopo di conseguire una maggiore snellezza e precisione delle operazioni, opera solitamente per mezzo di una statistica campionaria, funzione di uno stimatore, che sia una variabile casuale tabulata. Questa funzine viene detta **funzione test**. I moderni software statistici ricorrono ad altre procedure equivalenti basate sul *p-value* che qui non sviluppiamo.

Lo spazio dei valori possibili della funzione test viene diviso in due insiemi disgiunti e complementari uno dei quali (detto *intervallo di accettazione*) comprende i valori con densità di probabilità maggiore ed è costruito in modo da avere una probabilità totale elevata di verificarsi dato H_0. L'altro (detto *intervallo di rigetto* o anche *di rifiuto*) riunisce i valori con densità minore avente un totale di probabilità piuttosto piccola, sempre nella stessa condizione. La probabilità associata al primo insieme è $1-\alpha$ e quella associata al secondo è α. I valori più usati sono per la prima probabilità 0.95 e per la seconda 0.05.

Sul campione estratto si calcola la statistica campionaria e da questa la funzione test. Si verifica se il valore della funzione test appartiene al primo insieme ed allora si conclude che l'ipotesi accettata è la H_0, in caso contrario si decide per l'ipotesi H_1.

3.3 Le diverse procedure su un campione

In questo paragrafo esaminiamo le procedure per la verifica delle ipotesi su alcuni parametri.

3.3.1 frequenza relativa

Ciascuna unità della popolazione sottoposta ad una misurazione dicotomica [2] fornisce il risultato positivo con probabilità π incognita. L'estrazione di un campione casuale semplice di n unità dà luogo a n misure che valgono 1 per i risultati positivi e 0 per quelli negativi. Poniamo la variabile X uguale alla somma di questi n valori; in altri termini X rappresenta il numero di esiti favorevoli sulle n prove. Sappiamo che $p = X/n$ è uno stimatore corretto di π, possiamo quindi costruire un test di ipotesi su π stessa.Sappiamo dalla Teoria dei Campioni che per n sufficientemente grande

$$p \approx N\left\{\mu_p = \pi; \sigma_p^2 = \frac{\pi(1-\pi)}{n}\right\}$$

allora, avendo poste le ipotesi

[2] Una misurazione di questo genere consiste di una domanda alla quale si può rispondere SI oppure NO a seconda della situazione reale. L'esempio più diffuso e dato dalle domande poste per i sondaggi di opinione come: *Dai giudizio favorevole alla politica del Governo in materia di .. .. ?*, ma anche altre domande sono di questo genere come *Sei lettore abituale del giornale .. .. ?* o anche *Hai compiuto meno di .. .. anni.*

$$H_0 : \pi = \pi_0$$
$$H_1 : \pi \neq \pi_0$$

ricorriamo alla funzione test

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}},$$

che riferisce il valore della statistica campionaria p alla variabile casuale Z, cioè alla normale standardizzata. Usando le tavole di questa variabile casuale possiamo risolvere la disequazione probabilistica

$$\Pr\left(-z_{\alpha/2} < Z \leq z_{\alpha/2}\right) = 1 - \alpha \tag{3.1}$$

nella quale $z_{\alpha/2}$ è tale che

$$\Pr\left(Z \geq z_{\alpha/2}\right) = \frac{\alpha}{2}$$

e poi

$$\Pr\left(Z \leq z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

gli intervalli simmetrici costruiti con questa regola contengono i valori con la densità di probabilità più elevata ed escludono quelli con densità minore e sono quindi i più piccoli tra tutti quelli che contengono $1 - \alpha$ di probabilità. L'intervallo individuato dalla formula 3.1 è la zona (o intervallo) di accettazione dell'ipotesi H_0.

In particolare per $\alpha = 0.05$ abbiamo

$$\Pr\left(-1.96 < Z \leq 1.96\right) = 1 - 0.05 = 0.95$$

Quindi l'intervallo di accettazione dell'ipotesi H_0 è dato dai valori della funzione test compresi tra -1.96 e 1.96 e l'unione dei due intervalli esterni costituisce la zona di rigetto di H_0.

Esempio 3.3 *Riprendiamo i dati dell'esempio 3.1 e scriviamoli nei termini di un test di ipotesi. Le unità statistiche sono i residenti in quel Comune e la loro totalità costituisce la popolazione oggetto, da cui si estrae un campione casuale semplice di* 200 *elementi. Le ipotesi sono:*

$$H_0 : \pi = 0.10$$
$$H_1 : \pi \neq 0.10$$

la funzione test diventa

$$z = \frac{p - 0.10}{\sqrt{\frac{0.10(0.90)}{200}}};$$

per $\alpha = 0.05$ *la zona di accettazione è* $-1.96 < Z \leq 1.96$. *Il campione ha dato* $p = 0.15$ *e qundi* $z = 2.36$. *La decisione è di dichiarare respinta* H_0. *Fin qui il test.*

Vediamo cosa sarebbe successo se invece il campione avesse dato come frequenza relativa gli altri valori, elencati nell'esempio 3.1. Una tabellina ci aiuta a confrontare i risultati:

p	z
0.20	4.71
0.15	2.36
0.14	1.88
0.09	−0.47

Vediamo dalla tabellina che valori di $p \leq 0.14$ *verificano* H_0 *mentre per* $p \geq 0.15$ *la stessa ipotesi viene rigettata. Più precisamente i valori soglia sono* $p = 0.058$ *e* $p = 0.142$ *ottenuti invertendo la formula della funzione test per* -1.96 *e* 1.96.

3.3.2 media

Ciascuna unità della popolazione, sottoposta a misurazione fornisce un valore x preso da un sottoinsieme di numeri reali [3]. La descrizione probabilistica della popolazione, fatta secondo questa grandezza è una variabile casuale con media μ e varianza σ^2. Un campione casuale semplice con reinserimento di n elementi trattati secondo quella misurazione è descrivibile come un'ennupla di valori: $x_1; x_2; x_3; \cdots x_i; \cdots x_n$. Su questo insieme di valori si possono calcolare le statistiche campionarie corrette:

$$\bar{x} = \frac{\sum x_i}{n}; s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

sappiamo poi, per n sufficientemente grande (e per popolazioni distribuite normalmente anche con $n \geq 1$) che

$$\bar{x} \approx N\left\{\mu_{\bar{x}} = \mu; \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\right\}$$

.

Questo combinato con le ipotesi

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

permette di scrivere la funzione test

[3] Le misurazioni che terminano con un numero sono le misurazioni quantitative e sono formulate con domande simili ad esempio a: *Quanto anni hai compiuto? Quanti centimetri sei alto?"Quanti milligrammi di colesterolo per decilitro di sangue hai?*

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}},$$

con le altre conseguenze relativamente all'intervallo di accettazione viste nel paragrafo relativo alla frequenza relativa.

Esempio 3.4 *Un campione di 17 unità estratto da una popolazione normale con* $\sigma^2 = 4000$ *ha dato* $\bar{x} = 2750$ *ci si chiede con* $\alpha = 0.05$ *se la media della popolazione può essere* $\mu = 2800$.

$$H_0 : \mu = 2800$$
$$H_1 : \mu \neq 2800$$

la funzione test è la z *e quindi*

$$z = \frac{2750 - 2800}{\sqrt{\frac{4000}{17}}} = -3.26$$

essendo i valori critici di $z = \pm 1.96$ *la funzione test assume valore nella zona di rigetto e quindi si respinge* H_0

Nelle situazioni più comuni però si ignora il valore della varianza della popolazione ed allora la funzione test, a patto che la popolazione sia distribuita normalmente, diventa:

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{S^2}{n}}},$$

Le altre considerazioni sono ancora valide. Bisogna comunque tenere conto che la distribuzione di t varia a seconda dei gradi di libertà (gdl, in simbolo ν) che sono $n - 1$.

$$\Pr\left(-t_{\alpha/2} < t \leq t_{\alpha/2}\right) = 1 - \alpha \tag{3.2}$$

nella quale $t_{\alpha/2}$ è tale che

$$\Pr\left(t \geq t_{\alpha/2}\right) = \frac{\alpha}{2}$$

e poi

$$\Pr\left(t \leq t_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

Poiché la distribuzione di t ha un andamento campanulare simile a quello della Z l'intervallo individuato dalla formula 3.2 è la zona di accettazione dell'ipotesi H_0 ed è il più piccolo di tutti quelli che contengono $1-\alpha$ di probabilità.

Non è ovviamente possibile definire gli estremi dell'intervallo di accettazione (detti anche valori critici) validi per qualsiasi numerosità, ma devono essere trovati di volta in volta sulle tavole a seconda dei gradi di libertà.

Esempio 3.5 *Una popolazione distribuita in modo normale ha fornito un campione di $n = 25$ unità ceh hanno dato $\bar{x} = 17400$ e $s^2 = 1500000$ ci si chiede se la popolazione può avere $\mu = 17000$.*

$$H_0 : \mu = 2800$$
$$H_1 : \mu \neq 2800$$

La varianza della popolazione non è nota e quindi la funzione test è t

$$t = \frac{17400 - 17000}{\sqrt{\frac{1500000}{25}}} = 1.63$$

essendo $t_{tab} = \pm 2.064$ viene accettata H_0.

3.3.3 varianza

Nella situazione reale appena descritta l'ipotesi da saggiare è relativa alla varianza; cioè

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

La statistica campionaria S^2 è uno stimatore corretto di σ^2 ed inoltre, se la popolazione si distribuisce normalmente, disponiamo della funzione test

$$\chi^2 = \frac{S^2(n-1)}{\sigma^2}$$

Teniamo presente la numerosità del campione, e dato che la variabile χ^2 dipende dai gradi di libertà, che in questa situazione valgono $\nu = n - 1$, troviamo tramite le tavole i valori della variabile che isolano le due code a destra ed a sinistra di 0.025

Essendo χ_1^2 tale che

$$\Pr(\chi^2 \leq \chi_1^2) = 0.025$$

e χ_2^2 tale che

$$\Pr(\chi^2 \leq \chi_2^2) = 0.975$$

la zona di accettazione di H_0 è

$$\chi_1^2 < \chi^2 \leq \chi_2^2$$

La variabile χ^2 assume soltanto valori positivi e per ν finito non è simmetrica. Quindi si deve scegliere se l'intervallo di accettazione deve essere simmetrico in senso probabilistico come quelli fissati sopra oppure essere il più piccolo di quelli possibili. Questa seconda possibilità, anche se caldeggiata da Dennis Lindley che aveva prodotto le tavole necessarie, non ha avuto seguito e nella pratica dei test statistici si segue la strada illustrata più sopra di creare intervalli di accettazione simmetrici in probabilità.

Esempio 3.6 *Ci si chiede se la varianza della popolazione da cui si è estratto il campione dell'esempio precedente può essere* $\sigma^2 = 1100000$

$$H_0 : \sigma^2 = 1100000$$
$$H_1 : \sigma^2 \neq 1100000$$

La funzione test è χ^2 *ed allora*

$$\chi^2 = \frac{1500000(25-1)}{1100000} = 34.09$$

Dalle tavole troviamo $\chi_1^2 = 12.401$ *e* $\chi_2^2 = 39.364$ *il valore assunto dalla funzione test è compreso nell'intervallo e quindi si accetta* H_0

3.4 Le procedure su due campioni

Affrontiamo adesso la problematica del confronto tra i parametri di due popolazioni diverse fatto per mezzo di test di ipotesi, che impiegano campioni provenienti dalle due popolazioni, identificate dall'indice posto al piede del parametro e poi della statistica campionaria.

Il vero problema richiederebbe che si risolvesse il dilemma tra queste due ipotesi:

$$H_0 : \lambda_1 = \lambda_2$$
$$H_1 : \lambda_1 \neq \lambda_2$$

Invece, data la disponibilità delle distribuzioni delle statistiche campionarie, il problema viene trasferito sulla differenza tra i parametri per la media e la probabilità, e sul rapporto tra le varianze.

I parametri vengono trasformati in questo modo:

$$\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = \delta = 0$$
$$\pi_1 = \pi_2 \rightarrow \pi_1 - \pi_2 = \delta = 0$$
$$\sigma_1^2 = \sigma_2^2 \rightarrow \frac{\sigma_1^2}{\sigma_2^2} = 1$$

le ipotesi vengono poste in funzione di queste composizioni di parametri, e le statistiche campionarie impiegate dalle funzioni test sono gli stimatori delle differenze tra medie e probabilità, e del rapporto tra le varianze.

3.4.1 frequenza relativa

La situazione è questa: si hanno due popolazioni dalle quali si estraggono due campioni casuali semplici ed indipendenti di rispettivamente n_1 e n_2 elementi, i due campioni hanno dato le frequenze relative p_1 e p_2. Per saggiare l'ipotesi

$$H_0 : \pi_1 = \pi_2$$

contro l'altra

$$H_1 : \pi_1 \neq \pi_2$$

si usa la funzione test:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

nella quale

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Esempio 3.7 *Un certo articolo viene commercializzato attraverso due diversi canali di vendita i Supermercati ed i Negozi Specializzati. Due campioni casuali di acquirenti sono stati interpellati in relazione al fascicolo di istruzioni allegato chiedendo loro:*Trovi esurienti le istruzioni? *i risultati sono stati*

Freq assolute	*SI*	*NO*	*Totale*
Supermercato	30	5	35
Negozio	10	10	20

Ci si chiede se il grado di soddisfazione dei clienti sia diverso a seconda del canale di approvvigionamento preferito. Le statistiche campionarie sono $p_1 = 30/35 = 0.86$ *e* $p_2 = 10/20 = 0.50$ *con rispettivamente* $n_1 = 35$ *e* $n_2 = 20$. *Le ipotesi sono*

$$H_0 : \pi_1 = \pi_2 H_1 : \pi_1 \neq \pi_2$$

Per prima cosa calcoliamo il valore combinato p

$$p = \frac{30 + 10}{35 + 20} = 0.73$$

da cui

$$z = \frac{0.86 - 0.50}{\sqrt{0.73(0.27)}\sqrt{\frac{1}{35} + \frac{1}{20}}} = 2.86$$

Poiché i valori critici sono $z = \pm 1.96$ *si respinge l'ipotesi che il gradimento non dipenda dal canale di approvvigionamento.*

3.4.2 varianza

Dalle due popolazioni si sono estratti due campioni casuali semplici di numerosità n_1 e n_2 che hanno fornito le statistiche campionarie: $\bar{x}_1$ e $\bar{x}_2$, s_1^2 e s_2^2. Essendo le due ipotesi in gioco:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

la funzione test è

$$F = \frac{s_1^2}{s_2^2}$$

i cui valori critici vanno trovati sulle tavole della distribuzione F. Si tenga presente che tra gli statistici prevale l'uso di mettere al numeratore la s^2 maggiore e quindi il test viene ad essere ad una coda, dato che l'ipotesi di nullità può essere contraddetta soltanto da valori grandi. Le tavole forniscono il valore critico in relazione a $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$ gradi di libertà.

L'ipotesi di nullità per le varianze viene detta di *omoschedasticità*.

Esempio 3.8 *Da due campioni di rispettivamente 21 e 18 unità si sono ottenute le varianze corrette* $s_1^2 = 2000$ *e* $s_2^2 = 1500$ *è credibile che le popolazioni abbano varianza uguale? Le ipotesi sono*

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

La funzione test è

$$F = \frac{2000}{1500} = 1.33$$

Con $\nu_1 = 20$ *e* $\nu_2 = 17$ *risulta* $F_{tab} = 2.23$
La conclusione porta ad accettare l'omoschedasticità delle due popolazioni.

3.4.3 media

Dalle due popolazioni si hanno le stesse statistiche campionarie del punto precedente, le ipotesi per il confronto tra le medie sono:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

e, se è verificata l'ipotesi di nullità a proposito delle varianze, la funzione test è

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

dove

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}$$

la distribuzione di t di riferimento ha $\nu = n_1 + n_2 - 2$ gradi di libertà.

Si tralascia la trattazione sia della situazione in cui l'ipotesi di omoschedasticità non è verificata sia di quella con le varianze della popolazione note.

Esempio 3.9 *Due campioni hanno dato i seguenti valori:* $n_1 = 15; \bar{x}_1 = 735; s_1^2 = 2500$ *e poi* $n_2 = 23; \bar{x}_2 = 815; s_2^2 = 1400$. *Si deve decidere sull'uguaglianza delle medie delle due popolazioni.*

Prima di sotto porre a test le medie si deve verificare che le varianze delle due popolazioni siano uguali. Quindi con il test F

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{2500}{1400} = 1.79$$

Con $\nu_1 = 14$ *e* $\nu_2 = 22$ *risulta* $F_{tab} = 2.15$*; viene quindi confermata* H_0 *e si può procedere con il test* t*.*

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

$$s^2 = \frac{14(2500) + 221400}{14 + 22} = 1827.77$$

$$s = \sqrt{1827.77} = 42.75$$

quindi

$$t = \frac{735 - 815}{42.75\sqrt{\frac{1}{15} + \frac{1}{23}}} = -5.64$$

con $\nu = 36$ *si trova dalle tavole* $t_{tab} = 2.042$ *che ci porta a ritenere respinta l'ipotesi di uguagliuanza delle medie.*

3.4.4 Breve sintesi

Prima di procedere ripensiamo le procedure che abbiamo appena fissato.

Nella popolazione, che è descritta in modo probabilistico, si prende in esame un carattere, appartenente ad una certa scala di misurazione. La distribuzione della popolazione secondo quel carattere dipende da un parametro sul cui valore si esegue il test di ipotesi.

Dati l'ipotesi di base ed il valore del parametro che interessa si costruisce la distribuzione di probabilità di una statistica campionaria che sia stimatore di quel parametro e condizionata al valore del parametro (λ_0) posto nell'ipotesi. La distribuzione cambia a seconda della tecnica di formazione del campione e della sua numerosità.

Se ad esempio interessa la media di una popolazione infinita (o comunque molto grande), lo stimatore che interessa è la media campionaria, che, per n abbastanza grande, si distribuisce normalmente con media uguale a μ_0 e varianza σ^2/n.

Sfruttando la distribuzione probabilistica della statistica campionaria condizionata a H_0, si divide l'insieme dei suoi valori possibili in due intervalli (incompatibili e disgiunti) tali che uno sia (data l'ipotesi H_0) molto probabile e l'altro (sotto la stessa ipotesi) improbabile. La probabilità che la statistica cada nel primo intervallo è $1 - \alpha$, che cada nel secondo è α. I due intervalli vengono detti rispettivamente **di accettazione** e **di rigetto.**

Fissati questi intervalli si procede all'estrazione di un campione, fatta secondo le regole previste, e se il valore della statistica cade nell'intervallo di accettazione, si conclude dicendo che si accetta l'ipotesi di base, altrimenti si accetta l'ipotesi alternativa.

L'accettazione di una delle due ipotesi comporta che si agisca come se l'ipotesi accettata (sia H_0 oppure H_1) fosse una descrizione vera della realtà.

3.5 Gli errori nei test

I test che abbiamo visto finora sono stati costruiti cercando di difendere il più possibile l'ipotesi base. Una volta che l'ipotesi base risulta non più sostenibile allora ci si trova costretti a respingerla. Questo assomiglia al percorso che si segue non per sviluppare nuove ipotesi scientifiche, ma per verificare se quelle che si posseggono reggono alla prova dei fatti. Un procedimento più interessante di test può essere costruito mettendo in competizione due teorie diverse confrontando i due modelli probabilistici che ne derivano. Non seguiremo questo percorso di studio, ma passeremo ad esaminare più in dettaglio gli errori dei test soprattutto in vista della costruzione di test per la verifica di ipotesi puntuali sul valore dei parametri.

Un aspetto ineliminabile dei test statistici, ricordiamo infatti, è la possibilità di sbagliare, la possibilità cioè di commettere un errore [4]. I modi di sbagliare sono due e cioè

- a) ritenere l'ipotesi H_0 falsa quando invece è vera

- b) ritenere che l'ipotesi H_0 sia vera quando al contrario è falsa.

Il primo viene detto errore del primo tipo (o di *rigetto*), il secondo errore del secondo tipo (o di *accettazione*). Si noti, che l'ipotesi di nullità è privilegiata rispetto a quella alternativa. Le scelte linguistiche sottolineano questa sua posizione di maggior rilievo.

La teoria statistica ha come scopo anche di calcolare le probabilità di questi tipi di errore e le designa rispettivamente con le lettere α e β.

Ricapitolando

[4] L'errore è una descrizione non adeguata della realtà.

$$\alpha = \Pr(errore\ di\ primo\ tipo) = \Pr(H_1|H_0)$$
$$\beta = \Pr(errore\ di\ secondo\ tipo) = \Pr(H_0|H_1)$$

con le ultime espressioni si intende rispettivamente la probabilità di accettare H_1 al posto della vera H_0 e la probabilità di accettare H_0 al posto della vera H_1.

Le probabilità complementari sono

$$1 - \alpha = \Pr(scelta\ corretta\ dato\ H_0) = \Pr(H_0|H_0)$$
$$1 - \beta = \Pr(scelta\ corretta\ dato\ H_1) = \Pr(H_1|H_1)$$

Si noti di passaggio che queste non sono probabilità marginali di commettere errori o di compiere scelte corrette ma probabilità condizionate.

3.5.1 Valutazione della procedura

Errore di prima specie

La procedura, per il modo stesso con cui è stata definita la zona di accettazione, porta con sé il rischio di commettere errori. Poniamo che sia vera H_0 ed applichiamo correttamente la procedura; se la statistica campionaria cade al difuori dell'intervallo di accettazione la decisione è errata, e questo evento ha probabilità α. Quindi la possibilità di sbagliare non dipende dal fatto che, come si dice fatalisticamente *ci si sbaglia*, ma è intrinseca alla struttura stessa della teoria dei test.

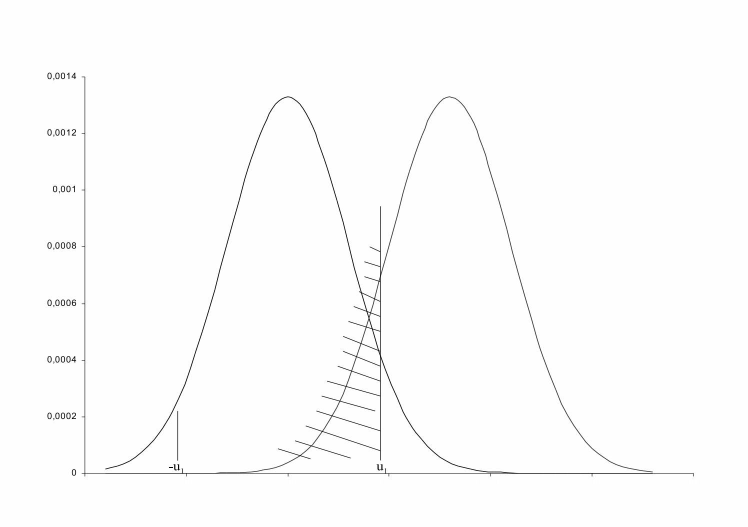
Errore di seconda specie

Arrivati a questo punto della trattazione verrebbe spontaneamente da dire: *riduciamo α in modo da ridurre la possibilità di sbagliare*. Ma ridurre α porta come conseguenza che si allarghi la zona di accettazione e quindi si venga ad aumentare la possibilità di trovare nella zona di accettazione statistiche campionarie provenienti da popolazioni il cui parametro λ non vale λ_0. Infatti dai disegni delle figure 3.1 si vede che passando da α_1 ad α_2 la zona di accettazione passa da $-u_1; u_1$ a $-u_2; u_2$ ed al diminuire di α si accompagna l'aumento di β, rappresentato dall'area con tratteggio orizzontale.

3.6 Caratteri a più modalità

Con il campione investighiamo una caratteristica della popolazione descrivibile con una scala nominale di misurazione a più classi. Detto in altro modo l'esito della misurazione non è suddiviso in due sole modalità (come avviene per i caratteri dicotomici), ma può essere attribuito ad una classe presa da un elenco di k alternative. In questo paragrafo, per la necessità di semplificare la trattazione, restringiamo l'esame alle popolazioni infinite oppure alle estrazioni fatte con reinserimnto.

Tabella 3.1: Relazione tra errori di prima e seconda specie

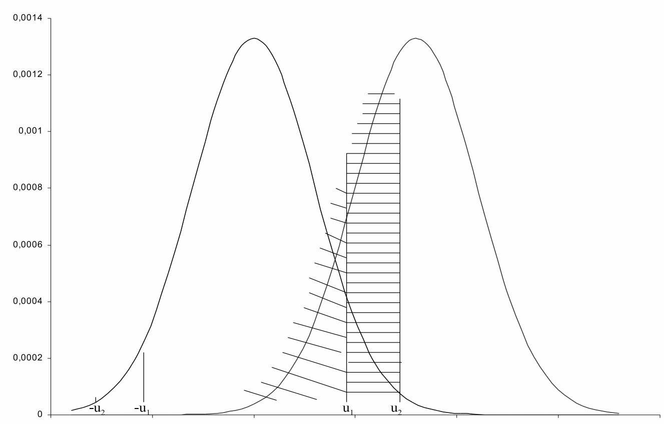


Esempio 3.10 *Gli acquirenti di un trapano elettrico della marca AA vengono intervistati nell'ambito di una ricerca sulla soddisfazione del cliente (customer satisfaction). È chiesto di compilare una cartolina inserita nella confezione del prodotto. Le risposte vengono classificate a seconda che il prodotto sia stato acquistato presso un supermercato, un negozio specializzato o direttamente dallo spaccio aziendale. Il singolo cliente viene assegnato ad una delle tre classi rappresentate nella tabellina. Le frequenze sono il rendiconto di 67 cartoline spedite dai clienti*

$Canale$ $di\,vendita$	$Frequenza$	$Freq\,relat$
$Supermercato$	35	0.52
$Negozio$	20	0.30
$Spaccio$	12	0.18

Dal punto di vista probabilistico possiamo dire che viene ripetuta per n volte una prova (esperimento od osservazione) il cui risultato appartiene ad uno ed uno solo dei k sottoinsiemi A_i, aventi probabilità π_i, in cui è stato diviso l'insieme dei risultati possibili elementari Ω. Registriamo come n_i il numero di volte con cui si è verificata la modalità A_i. Ne segue che $\sum_{i=1}^{k} n_i = n$ e $\sum_{i=1}^{k} \pi_i = 1$ ed inoltre $p_i = n_i/n$ con quindi $\sum_{i=1}^{k} p_i = 1$.

La probabilità di una certa $k - pla$ di valori di n_i è esprimibile per mezzo della **distribuzione multinomiale di probabilità** e cioè, essendo N_i le variabili casuali marginali che descrivono il numero di prove (o misurazioni) nelle quali appare il risultato A_i:



$$\Pr\left[(N_1 = n_1) \cap (N_2 = n_2) \cap \cdots (N_i = n_i) \cap \cdots (N_k = n_k)\right] =$$

$$= \frac{n!}{n_1! n_2! \cdots n_i! \cdots n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_i^{n_i} \cdots \pi_k^{n_k}$$

Esempio 3.11 *Per il caso in esame, lasciando inespresse le π_i abbiamo*

$$\Pr\left[(N_1 = 35) \cap (N_2 = 20) \cap (N_3 = 12)\right] =$$

$$= \frac{67!}{35!20!12!} \pi_1^{35} \pi_2^{20} \pi_3^{12}$$

Ricordiamo che

$$M(N_i) = n\pi_i$$
$$V(N_i) = n\pi_i(1 - \pi_i)$$
$$COV(N_i, N_h) = -n\pi_i\pi_h$$

Queste relazioni valgono prima di eseguire le misurazioni e dal loro esame ricaviamo interessanti informazioni sul fatto che le p_i sono buoni stimatori delle π_i.

Possiamo concludere, dall'esame di una sola di queste mutabili (A_i) considerata contro tutte le altre, che si distribuisce come una binomiale con media e varianza che dipendono da π_i. Quindi per quanto riguarda la stima ed i test di ipotesi, che coinvolgano una sola delle modalità di Ω, vale quanto detto a proposito della frequenza relativa (vedi paragrafo 2.1.2).

Esempio 3.12 *La direzione ritiene che tra i rispondenti il 70% provenga dai clienti di supermercati. Verifichiamo con il solito test sulla frequenza relativa se questa ipotesi sia sostenuta dai dati o sia invece contraddetta.*

$$H_0 : \pi = 0.70$$
$$H_1 : \pi \neq 0.70$$

$$z = \frac{0.52 - 0.70}{\sqrt{\frac{0.70 \cdot 0.30}{67}}} = 3.22$$

Questo valore è maggiore di 1.96 e quindi si conclude che l'ipotesi di base è respinta.

La procedura cambia quando si intenda verificare un'ipotesi relativa alla distribuzione della probabilità su tutte le modalità considerate contemporaneamente. In questo caso l'ipotesi di nullità assume la forma:

$$H_0 : \pi_1 = \pi_{1\ 0} \cap \pi_2 = \pi_{2\ 0} \cdots \cap \pi_k = \pi_{k\ 0}$$
$$H_1 : H_0\, false$$

La funzione test che saggia la bontà di questa ipotesi è

$$X^2 = \sum_{i=1}^{k} \frac{n(p_i - \pi_{i\ 0})^2}{\pi_{i\ 0}}$$

Di questa statistica campionaria è stato dimostrato che si distribuisce approssimativamente (per $n \to \infty$) come un χ^2 con $k-1$ gradi di libertà; è quindi da questa variabile casuale che si traggono i valori critici.

A parte casi molto particolari [5] la zona di accettazione comprende anche il valore zero della variabile χ^2. Questo valore infatti si ottiene quando i dati osservati rispecchiano esattamente le frequenze ipotizzate. Quindi il valore critico è uno solo χ^2_1 ed è tale che

$$\Pr(\chi^2 > \chi^2_\alpha) = \alpha$$

La zona di accettazione di H_0 è

$$\chi^2 \leq \chi^2_\alpha$$

e quella di rigetto

$$\chi^2 > \chi^2_\alpha.$$

[5] Tra i quali va ricordata la rivisitazione critica degli esperimenti di G. Mendel fatta da R. A. Fisher negli anni trenta.

Esempio 3.13 *Sulla base delle statistiche di vendita si sa che quel prodotto viene venduto dai tre canali nelle proporzioni*

$Canale$ $divendita$	$Probab.$
$Supermercato$	0.70
$Negozio$	0.25
$Spaccio$	0.05

Ci si aspetta quindi che anche le risposte seguano le stesse percentuali. Ci si chiede se le risposte siano in linea con le vendite per canale. Una differenza che non sia attribuibile al caso viene considerata come indizio di una attenzione verso il prodotto da parte dei clienti positiva o negativa a seconda del canale di acquisto.

Il test viene costruito così

$$H_0 : \pi_1 = 0.70 \cap \pi_2 = 0.25 \cap \pi_3 = 0.05$$
$$H_1 : H_0\, falsa$$

$$X^2 = 67\left[\frac{(0.52-0.70)^2}{0.70}+\frac{(0.30-0.25)^2}{0.25}+\frac{(0.18-0.05)^2}{0.05}\right] = 26.42$$

con due gradi di libertà. Dalle tavole si trova

$$\chi^2_\alpha = 5.991$$

L'ipotesi H_0 viene respinta. Si può quindi ritenere che la scelta del canale influisca sulla maggiore (spaccio e negozio) o minore (supermercato) attenzione verso il prodotto.

Le scale di misurazione quantitative e quelle ordinali possono essere trattate come scale nominali e quindi questo test può essere impiegato per verificare se un certo carattere della popolazione si distribuisce secondo una legge probabilistica che si pone come ipotesi. La procedura prevede che l'insieme di tutti i valori possibili sia classificato secondo un numero finito di sottoinsiemi in modo che l'elenco di queste classi di valori costituisca l'insieme Ω su cui si applica questo test. A ciascuna classe A_i viene attribuita la probabilità $\pi_{i\,0}$ calcolata secondo quel modello e poi con la funzione test applicata alle frequenze relative ottenute dal campione e confrontate con le probabilità ipotizzate si giunge ad una conclusione.

Esempio 3.14 *Il settore tecnico dell'industria afferma che quel prodotto è stato progettato per avere una durata descrivibile con una variabile esponenziale negativa avente media di* 500 *ore. I 67 clienti hanno dichiarato diverse durate che raccolte in classi hanno dato la seguente tabella di frequenze:*

$Estremo$ $inferiore$	$Estremo$ $superiore$	$Frequenze$ $osservate$	$Freq\,rel$ $osservate$	$Prob$
0	200	18	0.27	0.33
200	400	16	0.24	0.22
400	800	19	0.28	0.25
800	$in\,poi$	14	0.21	0.20

Le ipotesi sono:

$$H_0 : \pi_1 = 0.33 \cap \pi_2 = 0.22 \cap \pi_3 = 0.25 \cap \pi_4 = 0.20$$
$$H_1 : H_0\, falsa$$

Le probabilità dell'ipotesi sono state calcolate con la formula ben nota

$$\Pr(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

nella quale, essendo $\lambda = 1/\mu = 1/500 = 0.002$

$$F(x) = 1 - e^{-\lambda x}$$

Il calcolo dà:
$X^2 = 1.2234.$
Essendo il valore critico $\chi^2_\alpha = 7.815$
L'ipotesi H_0 *viene accettata.*

Talvolta l'ipotesi non comprende tutti i parametri necessari per definire il modello probabilistico, ma alcuni vengono stimati attraverso il campione; in questo caso per ogni parametro stimato si deve ridurre di uno il numero dei gradi di libertà.

Esempio 3.15 *Sono state misurate le altezze di 200 studenti universitari, allo scopo di verificare l'ipotesi secondo la quale l'altezza degli adulti si distribuirebbe secondo una distribuzione normale, avente media e scarto quadratico medio diversi a seconda delle circostanze di tempo e di luogo delle misurazioni. Le misure sono state riunite in classi di valori e sono riassunte nella tabella seguente.*

$Estremo$ $inferiore$	$Estremo$ $superiore$	$Valore$ $centrale$	$Frequenze$ $osservate$	$Freq\,rel$ $osservate$	$Prob$
$fino\,a$	150	145	5	0.025	0.041
150	160	155	30	0.150	0.157
160	170	165	80	0.400	0.320
170	180	175	50	0.250	0.308
180	190	185	25	0.125	0.141
190	$e\,oltre$	195	10	0.050	0.034

I calcoli eseguiti con i valori centrali delle classi danno $\bar{x} = 169.5$ *e* $s^2 = 125.38$. *Queste stime vengono imposte al modello normale per determinare i valori della standardizzata (ad esempio* $(150 - 169.5)/\sqrt{125.38} = -1.75$ *e così via). Le probabilità dell'ultima colonna sono state calcolate appunto ricorrendo a questi valori standardizzati. Risulta*
$X^2 = 9.47.$
Le ipotesi concorrenti sono:

$$H_0 : \pi_1 = 0.041 \cap \pi_2 = 0.157 \cap \pi_3 = 0.320 \cap \pi_4 = 0.308 \cap \pi_5 = 0.141$$
$$\cap \pi_6 = 0.034$$
$$H_1 : H_0\, falsa$$

essendo $\nu = 6 - 2 - 1 = 3$
$\chi^2_\alpha = 7.815$
che porta a respingere l'ipotesi di normalità, dato che le probabilità di H_0 *sono state calcolate con quel modello probabilistico.*

Un caso particolare è dato dall'ipotesi di indipendenza in una tabella a doppia entrata. Non ci è possibile (dato il carattere introduttivo di questo manuale) affrontare il problema di ricavare dai principi di cui sopra, integrati da altri strumenti matematici come lo studio dei sistemi di equazioni lineari, le regole operative per questo test, accontentiamoci di vederne il funzionamento.

Abbiamo eseguito n misurazioni sugli elementi di un campione casuale semplice e le abbiamo classificate secondo le n_r modalità del carattere Ω_r e le n_c modalità del carattere Ω_c. La distribuzione congiunta è formata di $n_r n_c$ caselle di una tabella a doppia entrata nella quale il carattere Ω_r descrive le righe ed il carattere Ω_c descrive le colonne. La frequenza relativa p_{ij} dice quant'è la frequenza congiunta della modalità $i - esima$ del carattere riga e della $j - esima$ del carattere colonna. Solitamente le frequenze marginali di riga $p_{i.} = \sum_{j=1}^{n_c} p_{ij}$ sono trattate come parametri certi e non come stime, analogamente viene fatto per le marginali di colonna $p_{.j} = \sum_{i=1}^{n_r} p_{ij}$.

I valori teorici da confrontare con le frequenze osservate p_{ij} si calcolano seguendo il teorema delle partizioni indipendenti e cioè, per ogni coppia ij

$$p^*_{ij} = p_{i.} p_{.j}.$$

La funzione test allora diventa:

$$X^2 = \sum_{i=1}^{n_c} \sum_{j=1}^{n_r} \frac{n(p_{ij} - p^*_{ij})^2}{p^*_{ij}}$$

che si distribuisce come un χ^2 con $(n_c - 1)(n_r - 1)$ gradi di libertà.

Esempio 3.16 *Le cartoline di cui nell'esempio 3.10 sono classificate secondo il canale di acquisto* (Ω_r) *e secondo la risposta data alla domanda:* Ha trovato sufficienti le istruzioni allegate?(Ω_c). *Le frequenze sono le seguenti:*

Freq assolute	*SI*	*NO*	*Totale*
Supermercato	30	5	35
Negozio	10	10	20
Spaccio	8	4	12
Totale	48	19	67

Da queste si ottiene la tabella delle frequenze relative osservate (p_{il})*:*

Freq relative	*SI*	*NO*	*Totale*
Supermercato	0.4478	0.0746	0.5224
Negozio	0.1493	0.1493	0.2985
Spaccio	0.1194	0.0597	0.1791
Totale	0.7164	0.2836	1.0000

e da questa quella delle probabilità ipotizzate nella condizione di indipendenza tra i due caratteri:

Freq relative	*SI*	*NO*	*Totale*
Supermercato	0.3742	0.1481	0.5224
Negozio	0.2139	0.0847	0.2985
Spaccio	0.1283	0.0508	0.1791
Totale	0.7164	0.2836	1.0000

infine
$X^2 = 8.1686$

H_0*: i due caratteri sono indipendenti*
H_1*:* H_0 *falsa*

essendo $\nu = (3-1)(2-1) = 2$

$\chi^2_\alpha = 5.991$
Si conclude quindi riconoscendo la presenza di dipendenza nella relazione che lega i due caratteri.

Appendice A

Cenni di storia del concetto di probabilità

A.1 Storia della parola *probabilità*

La parola *probabilità* ha avuto una storia alquanto articolata, i cui inizi possono risalire al filosofo greco Carneade.

Secondo la corrente dello scetticismo radicale è impossibile dare un giudizio morale data l'impossibilità di conoscere ciò che è bene e distinguerlo da ciò che è male. Carneade propone, per uscire da questo stallo di scegliere secondo il grado di *probabilità* basato sull'esperienza. Il termine in questo contesto significa, *possibilità di essere ritenuto buono.*

Da questa posizione metafisica deriva il sistema morale detto *probabilismo* secondo il quale l'incertezza relativa a ciò che è più adeguato al bene sarebbe dissipata facendo riferimento ad una autorità in materia che dispone delle conoscenze e della esperienza su ciò che costituisce oggetto di decisione [1].

Cicerone, seguendo appunto Carneade, nel trattato *de Bono et de Malo* ha per primo usato il termine *probabilia*[2] per intendere quei comportamenti che possono essere ritenuti buoni da una autorità in materia. Da qui, applicando un principio della metafisica classica (*ens bonum et verum convertuntur*[3]) questa parola aggiunge al significato *ciò che può essere buono* quell'altro: *ciò che può verificarsi*. Questo secondo significato è stato attribuito al termine per la prima volta (in forma scritta) da Gerolamo Fracastoro (1478-1553) quando, riferendosi ai segni premonitori delle pestilenze, diceva che erano segni probabili e non certi del loro verificarsi. Dopo 15 secoli questo nuovo significato viene

[1] Qui si è data una versione semplificata, forse all'eccesso, di queste dottrine. Con queste poche righe si è voluto infatti soltanto richiamare il substrato concettuale che ha prodotto questa parola.

[2] In latino: *probabilia* = cose che possono essere approvate.

[3] *Ente, bene e vero si convertono.* In altre parole i termini ente, vero e buono possono essere sostituiti nelle proposizioni l'uno all'altro trasformando una proposizione vera in un'altra vera.

aggiunto a quello precedentemente fissato da Cicerone quando aveva creato il termine.

Il termine *probabilismo* rimane comunque nella filosofia morale dove definisce uno dei sistemi morali accanto al *lassismo,* al *tuziorismo*, ecc.[4] .

A.2 Storia del concetto

Da questo punto in avanti la parola *probabile* significa per noi soltanto che un certo (cioè ben definito) risultato di una misurazione *può verificarsi*. Sia la possibilità che ha di verificarsi sia la valutazione numerica di quella possibilità vengono chiamate *probabilità*; avendo l'intesa che i valori di probabilità più alti vengono assegnati ai risultati che si presentano più facilmente, e che i valori più bassi vengono assegnati ai risultati che si presentano più difficilmente. Come conseguenza i risultati rari delle misurazioni hanno probabilità piccole.

Lo studio scientifico della probabilità richiede che sia possibile dire qualcosa di razionale a proposito di fatti che avvengono in modo imprevedibile e quindi, almeno per lo stadio di conoscenze di quel momento, irrazionali.

Come conseguenza di questa contrapposizione tra razionalità ed irrazionalità lo studio delle probabilità non ha avuto sviluppo scritto fino al sedicesimo secolo. Purtuttavia è possibile rinvenire una antica traccia del fatto che una scienza dell'alea sia possibile.

Lo statistico indiano Godambe, infatti, fece notare come nell'antica storia del principe Nala, inserita nel Mahabharata nel sesto secolo dell'era cristiana, si spiega che può esistere una scienza del gioco dei dadi accessibile a chi sappia trattare i numeri. Non viene data la descrizione di questa scienza, ma viene mostrato come sia in relazione con quella che noi chiameremmo una stima campionaria. Null'altro, ma è perlomeno curioso trovare traccia dell'intuizione che l'irrazionalità del risultato del lancio del dado sia interpretabile nell'ambito della matematica anche se ne sembrerebbe irrimediabilmente distante.

Torniamo alla cultura europea dove la teoria delle probabilità si è sviluppata, come abbiamo ricordato, a partire dal sedicesimo secolo. In questo periodo assistiamo alle prime relazioni scritte riguardanti applicazioni al gioco d'azzardo, che risultano essere precise nello sviluppo, ma intuitive nel metodo probabilistico. Nelle elaborazioni possiamo trovare l'impiego, per lo più implicito, di concetti che poi diverranno familiari di casi favorevoli, frequenze ecc. I massimi matematici dell'epoca si cimentarono con successo in queste episodiche applicazioni; tra tutti basta citare il Tartaglia, Cardano ed infine Pascal.

Appoggiandosi sull'opera di quest'ultimo Huygens pubblicò nel 1657 un primo testo dedicato per intero alla probabilità. Huygens conservò l'impostazione frammentaria della trattazione costituita di una serie di problemi ai quali viene data soluzione seguendo comunque un metodo di lavoro omogeneo. Grazie comunque ad Huygens, e poi a de Witt ed all'inglese Graunt, abbiamo

[4]Si lascia al lettore, desideroso di sapere, l'impegno di decifrare questi termini così *esotici*.

in quella fine del secolo diciassettesimo l'estensione dell'applicazione della probabilità dai giochi d'azzardo alla demografia ed al calcolo attuario.

Giacomo I Bernoulli fu il primo a dare veste completa alla teoria matematica delle probabilità ed a darne una elaborazione teorica disgiunta dalle applicazioni direttamente interessate. Questa novità è chiaramente percepibile dal confronto dei titoli di queste due ultime opere: Huygens chiamò la sua *De Ratiociniis in Ludo Alee*, mentre Bernoulli intitolò la sua *Ars Conjectandi*.

A.3 Filosofia della probabilità

A.3.1 Pierre Simone de Laplace

Il primo testo che abbia affrontato il problema di definire la natura della probabilità è il fondamentale *Saggio Filosofico sulle Probabilità* di Pierre Simon de Laplace (1749-1827) (vedi Laplace, 1967a).

La sua concezione parte da una visione della realtà del mondo basata su un determinismo fisico assoluto:

> *Dobbiamo dunque considerare lo stato presente dell'universo come effetto del suo stato anteriore e come la causa del suo stato futuro. Un'Intelligenza che, per un dato istante, conoscesse tutte le forze da cui è animata la natura e la situazione rispettiva degli esseri che la compongono, se per di più fosse abbastanza profonda per sottomettere questi dati all'analisi, abbraccerebbe nella stessa formula i movimenti dei più grandi corpi dell'universo e dell'atomo più leggero: nulla sarebbe incerto per essa e l'avvenire, come il passato, sarebbe presente ai suoi occhi.*

Non serve e non è scopo di questa esposizione criticare questa visione del mondo. È invece essenziale rilevare come in questa visione determinista vi sia posto per la probabilità. Infatti Laplace, parlando degli avvenimenti in genere, afferma:

> *Ignorando i legami che li uniscono al sistema intero dell'universo, li si è fatti dipendere dalle cause finali o dal caso, a seconda che si manifestassero e si succedessero con regolarità o senza ordine apparente; ma queste cause immaginarie sono state successivamente arretrate sino ai limiti delle nostre conoscenze e spariscono del tutto davanti alla sana filosofia la quale non vede in esse che l'espressione dell'ignoranza in cui ci troviamo circa le vere cause.*

È quindi l'ignoranza umana il fondamento della probabilità. Orietta Pesenti Cambussano, curatrice del volume delle *Opere* di Laplace, commenta giustamente nella nota di pagina 245di Laplace, 1967a:

> *Il concetto di probabilità appare invece ambiguo, per il suo partecipare contemporaneamente all'ambito logico della possibilità e all'ambito empirico della misurazione dei dati che possediamo circa un determinato*

oggetto. Da qui la difficoltà di determinare esattamente tale concetto nelle applicazioni pratiche, ma da qui anche la sua funzione di ponte di passaggio tra esperienza e matematica.

Queste ambiguità appartengono in genere alle misurazioni che mettono in relazione fatti reali e segni astratti. Poiché assegnamo alla probabilità questa funzione di essere ponte tra l'esperienza attesa e quella misurata a posteriori, non possiamo far altro che porla in questa posizione intermedia. La definizione di probabilità parte infatti dal concetto di indecisione e su questo esercita la procedura di misurazione (vedi pagina 246 di Laplace, 1967a):

La teoria dei casi consiste nel ridurre tutti gli eventi dello stesso genere ad un numero di casi ugualmente possibili, cioè tali da renderci indecisi circa la loro esistenza, e nel determinare il numero dei casi favorevoli di cui si ricerca la probabilità.

Quindi l'indecisione rispetto all'accadere di quel particolare risultato è, come detto, il componente essenziale delle probabilità. Essendo uno stato psicologico è sottratta alla procedura matematica e costituisce il componente essenziale della probabilità che è (vedi Laplace, 1967a pagina 245)

Relativa in parte [alla nostra] ignoranza, in parte alle nostre conoscenze.

A.3.2 Ludwig von Mises

Una visione complementare può essere trovata nell'opera di Ludwig von Mises (1883-1953) che fu generata quando, all'inizio del '900 cominciarono ad affermarsi gli studi sull'atomo e sullo scandalo della natura della luce che, a seconda degli esperimenti che venivano eseguiti, appariva in certe circostanze un corpuscolo ed in altre un'onda. Come conseguenza divenne difficile, se non impossibile praticare la suddivisione in classi di indifferenza dei risultati possibili di una misurazione.

Il sistema che produce gli oggetti da misurare è infatti sconosciuto e forse inconoscibile del tutto. Se lo scopo della probabilità secondo Laplace era in definitiva di fornire una descrizione del meccanismo che produce certi risultati piuttosto di altri, adesso con von Mises, vista la difficoltà di definire di cosa si stia parlando, questa finalità viene abbandonata.

Il punto essenziale dell'analisi della probabilità viene spostato, infatti, dall'apparato che produce i risultati, in qualche misura imprevedibili, ai risultati stessi.

Nel suo trattato sulla probabilità von Mises pone una serie di condizioni perché si possa trattare scientificamente questo concetto. I requisiti posti definiscono dapprima il collettivo e poi su questo giungono a definire cosa si debba intendere per probabilità[5]:

[5]Vedi von Mises, 1964 pag. 12

> *Poniamo che $S = \{a_i\}$ sia un insieme discreto di etichette e $K = \{x_j\}$ una successione di elementi di S.*
>
> *Poniamo che G sia un sistema di scelte di posizioni con la potenza dell'infinito numerabile. Noi assumiamo che:*
>
> 1. *per ogni etichetta a_i il limite della frequenza relativa p_i esiste in K;*
> 2. *$\sum_{i=1}^{\infty} p_i = 1$, essendo la somma estesa su tutti gli elementi di S;*
> 3. *Qualsiasi scelta di posizioni Γ appartenente a G applicata a K produce una sottosuccessione infinita di K nella quale di nuovo, per ogni a_i, il limite della frequenza relativa esiste ed è uguale a p_i.*
>
> *Allora, K, o più esplicitamente $K(G, S)$, è definito un collettivo (rispetto a G ed S) e $p(a_i) = p_i$ è la probabilità di incontrare l'etichetta a_i in K.*

E poi nei commenti della stessa pagina aggiunge:

> *Il concetto di collettivo è alla base della nostra teoria della probabilità. Il termine probabilità è significativo per noi soltanto in relazione ad un chiaramente definito collettivo (o popolazione). Questo è il significato della nostra originale asserzione, del paragrafo precedente, che dice, nella attuale terminologia, che ogni proposizione probabilistica si rivolge ad un collettivo, un fenomeno aggregato, piuttosto che ad un evento isolato individuale. La probabilità di morire entro il prossimo anno può essere p_1 per un certo uomo se è considerato come un maschio di 42 anni, e può essere $p_2 \neq p_1$ se è considerato come un maschio con età tra 40 e 45 anni, e p_3 se è considerato come appartenente alla classe di tutti gli uomini (maschi e femmine) degli Stati Uniti. (Applichiamo qui la nostra teoria ad una popolazione grande ma finita).*

Da queste citazioni possiamo ricavare che in questa impostazione la casualità è una proprietà ineliminabile dei risultati e fa quindi parte in modo irriducibile del processo da cui sono prodotti. La seconda visione caratteristica è che non si possa attribuire la probabilità ad un caso singolo, ma che esista soltanto per prove indefinitamente ripetibili.

A.3.3 Confronto tra le due impostazioni

Ora come l'altra impostazione anche questa non va intesa in senso esclusivo, cioè come se a tutto ciò che non rispetta le clausole dovesse essere vietata l'attribuzione di probabilità. Si veda l'esempio riportato qui sopra.

D'altronde nessuna di queste due definizioni può essere verificata nella realtà. Pensiamo alla situazione più semplice: un dado.

Per quanto riguarda la definizione laplaciana riusciamo ben facilmente a negare che le facce abbiano tutte la stessa indecidibilità (ma sarebbe meglio parlare subito di probabilità). È sufficiente immergere il dado in una soluzione zuccherina sufficientemente concentrata per vedere come il dado, lasciato libero, mostri sempre la stessa faccia. Quindi ci viene spontaneo concludere che, per quanto piccola, questa faccia avrà una preferenza[6].

L'utilità della visione laplaciana però si manifesta proprio in situazioni come queste nelle quali stimola a spingere la ricerca per così dire *dentro il dado*, allo scopo di individuare quali sono le cause che favoriscono una delle facce piuttosto delle altre. Un modello probabilistico completo e corretto del dado dovrebbe infatti coinvolgere le proprietà geometriche del dado, quelle strettamente legate alla sua stabilità, come la posizione del baricentro, la sua elasticità ed in genere la sua capacità a rotolare, la risposta della tavola alla caduta, poi chissà quanti altri fattori, e soprattutto il modo con cui lo si lancia.

Quest'ultima componente, l'unica variabile da lancio a lancio, dovrebbe essere descrivibile in termini di equidistribuzione probabilistica.

Ma poi affinando lo studio potremmo trovare che nemmeno questo è vero e sorgerebbe un altro problema sul movimento del braccio, sull'atteggiamento di chi esegue i lanci e così via.

Da questo vediamo che l'impostazione di Laplace è una sorgente di problemi, dalla soluzione di ciascuno dei quali sorgono conoscenze più approfondite sul *sistema che genera* i risultati.

Per la seconda definizione possiamo intanto osservare che il dado non potrà essere lanciato per un numero infinito di volte. Nessuna struttura umana riesce a durare così tanto.

Potremmo pensare, per scherzo, ad una società segreta cinese che da 5000 anni abbia lanciato un dado per 24 ore al giorno, al ritmo di 30 lanci al minuto e ne abbia registrato i risultati; ebbene, ad oggi avrebbe totalizzato meno di 80 miliardi di lanci, che sono molti, ma molti meno dell'infinito richiesto alla lettera dalla teoria frequentista esaminata. Ancora di più, dopo così tanti lanci il dado come si sarà consumato?

Anche questa definizione va quindi presa per quanto ha di buono e cioè come una guida a cercare le condizioni che fanno variare la probabilità del risultato. Si veda a proposito l'esempio che von Mises stesso pone a chiarimento della sua definizione: è un collettivo formato di un numero finito di elementi e che considera come un sottocollettivo di altri collettivi aventi probabilità diversa.

Si tratterà allora di verificare se, nell'esempio scherzoso a cui si è appena accennato, si ottengono sottocollettivi diversi cambiando il lanciatore dei dadi, oppure il supporto su cui si lancia, ecc. Si vede come questa impostazione non sia in opposizione a quella laplaciana, ma come, ai fini della ricerca scientifica, la integri e la potenzi.

Se da un punto di vista operativo le due impostazioni possono convivere, non dobbiamo nasconderci che da quello più propriamente filosofico sono al-

[6]Laplace ha studiato un problema analogo relativo alla moneta e lo troviamo alla pagina 229 delle *Opere* nella sua *Teoria Analitica delle Probabilità* (vedi Laplace, 1967b). Da questo si vede come non ritenesse così tassativa e restrittiva la sua definizione di probabilità da calcolare come rapporto tra casi favorevoli e possibili.

l'antitesi una dell'altra. Infatti la concezione di Laplace si pone in un contesto metafisico di assoluto determinismo fisico con il quale si confronta la limitatezza delle conoscenze umane. Quella di von Mises al contrario si pone (con il ricorso a tutte le scelte $\Gamma \in G$) in un contesto di indetermisimo irriducibile, nel quale è ammissibile una sorgente di variabilità non dominabile neppure con maggiori dosi di conoscenza, che verrebbero inglobate nei Γ.

A.3.4 Bruno de Finetti

Su una posizione opposta si pone la visione soggettivistica delle probabilità. Secondo questa filosofia della probabilità l'incertezza, qualsiasi ne sia l'origine è una proprietà del pensiero ed è questo l'ambito nel quale deve essere trattata. De Finetti diceva l'unica cosa di cui sono certo è che talvolta sono incerto.

Da questa impostazione si ottiene come guadagno l'ammissione del caso singolo ad argomento dello studio probabilistico. Ad esempio il quarantaduenne dell'esempio di von Mises si trova all'intersezione di infiniti o quanto meno moltissimi sottocollettivi, a seconda degli studi fatti, dello stato anagrafico, della professione, delle malattie sofferte in passato, ecc. Quindi se teniamo conto di tutte le circostanze relative a quel soggetto troviamo alla fine che si tratta di un caso singolo. Ma non solo: in definitiva a lui che interessa *di tutte* le altre persone che eventualmente condividessero l'intersezione *di tutte* le sue caratteristiche? Il suo sarebbe pur sempre un caso singolo. Se alla scala opposta badiamo al lancio di un dado; che interessa il fatto che nelle migliaia di lanci precedenti abbia fornito una certa successione di risultati se è sul prosssimo lancio che facciamo la scommessa?

Una risposta può essere che la probabilità relativa a quel lancio futuro non è trattabile scientificamente se non sono rispettate quelle certe condizioni di scientificità, ma che comunque si agisce soltanto si ricorre ad una definizioni indebolita di probabilità. La risposta di de Finetti è stata che sempre di probabilità si tratta, e di probabilità intesa nello stesso senso.

Lo scopo della teoria delle probabilità diventa così studiare le condizioni di coerenza per le previsioni da fare nelle situazioni di incertezza. La probabilità quindi può essere valutata in una qualsiasi situazione in cui un soggetto si trovi a dover agire senza poter conoscere pienamente il risultato delle sue azioni. Questa valutazione avviene per via indiretta attraverso due diverse procedure la più semplice delle quali simula una scommessa e l'altra tiene conto di certe penalizzazioni che seguirebbero dal non aver indovinato il risultato che si verificherà. Nella pagina 92 del suo importante trattato il de Finetti (vedi de Finetti, 1970) dà questa definizione della scommessa attraverso la quale si è sempre in grado di fissare la probabilità di un evento (o risultato E):

> *La probabilità $P(E)$ che Tu attribuisci ad un evento E è quindi il guadagno certo p che giudichi equivalente al guadagno unitario subordinato al verificarsi di E.*

A questa definizione aggiunge poi nel seguito della trattazione diverse cautele come la sostituzione del guadagno unitario con una somma più elevata, ma non enorme, e la disponibilità che si deve dare ad essere costretti a scommettere su $\bar{E}$.

Ma l'aspetto più interessante di questa impostazione, rispetto alle due che precedentemente abbiamo tratteggiato, è che espunge il problema metafisico dalla definizione di probabilità. Le altre due impostazioni ponevano forti limitazioni alla descrizione della realtà (per una era completamente deterministica, per l'altra era presente una componente di indeterminatezza irriducibile) riguardo la quale questa rimane al contrario perfettamente indifferente.

Dal punto di vista metafisico, quindi, questa impostazione può convivere con qualsiasi visione del mondo sia deterministica (vedi Laplace) sia indeterministica (vedi von Mises) ed anche con una visione creazionista del mondo, secondo la quale esiste un creatore che conosce e fa istante per istante un mondo del quale noi uomini abbiamo una conoscenza limitata, ma sempre perfettibile.

Appendice B

Media della varianza campionaria

Per estrazioni con reinserimento, qualsiasi distribuzione della popolazione

$$M_c\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right] = \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\right] =$$

$$= \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + (\bar{x}-\mu)2\sum_{i=1}^{n}(x_i-\bar{x}) + \sum_{i=1}^{n}(\bar{x}-\mu)^2\right] =$$

dato che $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$

$$= \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + \frac{1}{n}M_c\left[\sum_{i=1}^{n}(\bar{x}-\mu)^2\right] =$$

$$= \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + \frac{1}{n}\sum_{i=1}^{n}\left[M_c(\bar{x}-\mu)^2\right] =$$

$$= \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + \frac{1}{n}n\frac{\sigma^2}{n} =$$

$$= \frac{1}{n}M_c\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + \frac{\sigma^2}{n} =$$

nel primo membro

$$M_c\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{M_c(x_i-\mu)^2}{n}\right] = \frac{1}{n}n\sigma^2 = \sigma^2$$

e quindi

$$\sigma^2 = \frac{1}{n} M_c \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right] + \frac{\sigma^2}{n}$$

da cui

$$\sigma^2 - \frac{\sigma^2}{n} = \frac{1}{n} M_c \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]$$

$$\frac{n-1}{n} \sigma^2 = M_c \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \right]$$

Appendice C

Covarianza di due unità

La popolazione è formata di N elementi $\{x_1, x_2, \cdots x_N\}$ con $T = \sum x_s$ per intendere un generico elemento della popolazione usiamo l'indice s oppure t; per indicare un generico alemento del campione usiamo le variabili X_i e X_j. Calcoliamo la $cov(X_i; X_j)$ per i due schemi, quello con la restituzione e quello senza.

Prima verifichiamo che nell'estrazione con reinserimento $cov(X_i; X_j) = 0$. Potremmo risolvere facilmente questo compito facendo riferimento al fatto che in questo schema le estrazioni (e quidni le grandezze X_i e X_j) sono indipendenti e perciò (vedi CORRELAZIONE) la covarianza è nulla. Vediamo invece come si giunge in modo diretto ed esplicito a questo risultato.

$$cov(X_i; X_j) = M(X_iX_j) - \mu_{X_i}\mu_{X_j} = \frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N} x_s x_t - \mu\mu$$

per la definizione di cov e per la formula 1.3

Esempio C.1 *Riprendendo l'esempio l'elenco dei possibili campioni di due elementi è formato di*

AA	AB	AC	AD	AE
AB	BB	BC	BD	BE
AC	CB	CC	CD	CE
AD	DB	DC	DD	DE
AE	EB	EC	ED	EE

e gli esiti della misurazione voto medio sono rispettivamente

$25; 25$	$25; 27$	$25; 24$	$25; 23$	$25; 29$
$25; 27$	$27; 27$	$27; 24$	$27; 23$	$27; 29$
$25; 24$	$24; 27$	$24; 24$	$24; 23$	$24; 29$
$25; 23$	$23; 27$	$23; 24$	$23; 23$	$23; 29$
$25; 29$	$29; 27$	$29; 24$	$29; 23$	$29; 29$

dai quali si ottiene la corrispondente tabella dei prodotti

625	675	600	575	725
675	729	648	621	783
600	648	576	552	696
575	621	552	529	667
725	783	696	667	841

Somma dei prodotti 16384 *da cui* $M_c(X_iX_j) = \frac{16384}{25} = 655.36$ *valore che si ottiene anche da* $\mu\mu = 25.6^2 = 655.36$

In generale, dato che siamo in condizioni di indipendenza logica x_s può essere portato fuori dalla seconda sommatoria

$$\frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N} x_s x_t = \frac{1}{N^2}\sum_{s=1}^{N} x_s \sum_{t=1}^{N} x_t =$$

e dopo aver eseguito le due sommatorie, prima quella relativa alla t e poi quella della s

$$\frac{1}{N^2}TT = \mu\mu$$

e la covarianza

$$cov(X_i; X_j) = \mu\mu - \mu\mu$$

Esempio C.2 *Segue immediatamente*

$$cov(X_i; X_j) = 655.36 - 655.36 = 0$$

Nel caso invece di estrazione senza reinserimento

$$cov(X_i; X_j) = M(X_iX_j) - \mu_{X_i}\mu_{X_j} = \frac{1}{N(N-1)}\sum_{s \neq t}\sum_{t=1}^{N} x_s x_t - \mu\mu$$

dato che la non restituzione impedisce che la stessa unità compaia in entrambe le estrazioni; siamo in una condizione di dipendenza logica e la tabella dei prodotti contiene una casella vuota per ogni riga

Esempio C.3 *Riprendendo l'esempio l'elenco dei possibili campioni di due elementi è formato di*

$----$	AB	AC	AD	AE
AB	$----$	BC	BD	BE
AC	CB	$----$	CD	CE
AD	DB	DC	$----$	DE
AE	EB	EC	ED	$----$

le cui misure sono

$----$	$25;27$	$25;24$	$25;23$	$25;29$
$25;27$	$----$	$27;24$	$27;23$	$27;29$
$25;24$	$24;27$	$----$	$24;23$	$24;29$
$25;23$	$23;27$	$23;24$	$----$	$23;29$
$25;29$	$27;29$	$29;24$	$29;23$	$----$

ed i prodotti

$----$	675	600	575	725
675	$----$	648	621	783
600	648	$----$	552	696
575	621	552	$----$	667
725	783	696	667	$----$

la cui somma vale13084 *da cui* $M_c(X_iX_j) = \frac{13084}{20} = 654.20$ *e la*

$$cov(X_i; X_j) = 654.20 - 655.36 = -1.16$$

possiamo scomporre la somma di tutti i prodotti nel seguente modo

$$\sum_{s=1}^{N}\sum_{t=1}^{N} x_s x_t = \sum_{t=1}^{N}\sum_{s\neq t} x_s x_t + \sum_{t=1}^{N} x_t^2$$

$$\sum_{s=1}^{N}\sum_{t=1}^{N} x_s x_t - \sum_{t=1}^{N} x_t^2 = \sum_{t=1}^{N}\sum_{s\neq t} x_s x_t \tag{C.1}$$

inoltre

$$\sum_{s=1}^{N}\sum_{t=1}^{N} x_s x_t = N^2\mu^2$$

e dalla formula per il calcolo indiretto della varianza ($\sigma^2 = M(X^2) - \mu^2$)

$$\sum_{t=1}^{N} x_t^2 = N(\sigma^2 + \mu^2)$$

sostituendo nella C.1 si ottiene

$$N^2\mu^2 - N(\sigma^2 + \mu^2) = \sum_{t=1}^{N}\sum_{s\neq t} x_s x_t$$

ed infine

$$
\begin{aligned}
cov(X_i; X_j) &= \frac{1}{N(N-1)} N(N\mu^2 - \sigma^2 - \mu^2) - \mu^2 \\
&= \frac{N\mu^2 - \sigma^2 - \mu^2 - N\mu^2 + \mu^2}{N-1} \\
&= \frac{-\sigma^2}{N-1}
\end{aligned}
$$

Esempio C.4 *Nel solito esempio* $\frac{-\sigma^2}{N-1} = \frac{-4.64}{4} = -1.16$

Bibliografia

Azzalini, A. (2001). *Inferenza Statistica*. Springer, Milano.

Cicchitelli, G. (1984). *Probabilità e statistica*. Maggioli, Rimini.

de Finetti, B. (1970). *Teoria delle probabilità*. Einaudi, Torino.

Feller, W. (1970). *An introduction to probability theory and its applications*. John Wiley and Sons, New York.

Lalande, A. (1975). *Dizionario Critico di Filosofia*. ISEDI, Milano.

Laplace, P. S. (1967a). Saggio filosofico sulle probabilità. In Cambursano, O. P., curatore, *Opere*, pagine 241–404. UTET.

Laplace, P. S. (1967b). Teoria analitica delle probabilità. In Cambursano, O. P., curatore, *Opere*, pagine 221–231. UTET.

Piccolo, D. (2000). *Statistica*. Mulino, Bologna.

Prisco, R. (2006). *Dalla Probabilità Verso La Statistica*. Libreria Universitaria Editrice, Verona.

von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic Press, New York.