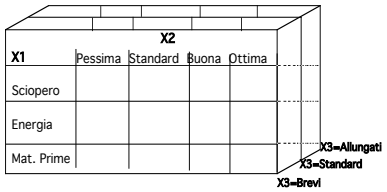
Distribuzioni discrete multidimensionali

Gli esperimenti possono avere una molteplicità di aspetti e generare più variabili casuali

ESEMPIO

Il ciclo di produzione può avere 3 tipi di interruzione: X1=(Sciopero, Energia,Materie prime). I prodotti sono di qualità X2=(Pessima, Standard, Buona, Ottima) ed i tempi di produzione: possono essere X3=(Standard, Ridotti, Allungati)



Per rappresentare l'esperimento usiamo la distribuzione congiunta trivariata

$$P(X_1, X_2, X_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \geq 0$$

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} P(_1 = x_1, X_2 = x_2, X_3 = x_3) = 1$$

La P(.) associa ad ogni possibile terna una probabilità non negativa con il vincolo di somma unitaria

Distribuzioni marginali

Se $X = (X_1, X_2, \ldots, X_n)$ **è una v.c. n-dimensionale la i-esima distribuzione marginale è**

$$P(X_i = x_i) = \sum_{x_1}\sum_{x_2}\cdots\sum_{\substack{x_n \\ Escluso\ x_i}} P(X_1, X_2, \ldots, X_n)$$

dove la somma è estesa a tutte le variabili casuali tranne la i-esima

ESEMPIO
la distribuzione congiunta è presentata a strati

$X1/X2$	X_3=1			X_3=2		
	-1	0	1	-1	0	1
-1	$4/24$	$1/24$	0	$1/24$	0	$2/24$
0	$2/24$	$1/24$	$2/24$	$1/24$	$1/24$	$2/24$
1	0	$1/24$	$2/24$	$3/24$	$1/24$	0

Calcoliamo P(X1)

Occorre sommare sia rispetto alla X2 che rispetto alla X3

$X1$		$P(X_1 = x_1)$
-1	$4/24 + 1/24 + 0 + 1/24 + 0 + 2/24 =$	$8/24$
0	$2/24 + 1/24 + 2/24 + 1/24 + 1/24 + 2/24$	$9/24$
1	$0 + 1/24 + 2/24 + 3/24 + 1/24 + 0$	$7/24$

Esempio

In questo esperimento definiamo
E1=Risultato del primo dado: E1=5
E2: Eisultato del secondo dado: E2=3

Gli aspetti che ci interessano sono:

X1= Somma dei punti: E1+E2
X2= Valore massimo: Max{E1,E2}
X3= Differenza: lE1-E2l

Ogni variabile casuale ha la sua distribuzione di probabilità marginale.

Ma ci sono anche le bivariate e la congiunta trivariata

X_1	$P(X_1 = x_1)$	X_2	$P(X_2 = x_2)$	X_3	$P(X_3 = x_3)$
2	1/36	1	1/36	0	6/36
3	2/36	2	3/36	1	10/36
4	3/36	3	5/36	2	8/36
5	4/36	4	7/36	3	6/36
6	5/36	5	9/36	4	4/36
7	6/36	6	11/36	5	2/36
8	5/36		1		1
9	4/36				
10	3/36				
11	2/36				
12	1/36				
	1				

Distribuzioni congiunte condizionali

Solo una estensione delle definizioni del caso bivariato

La distribuzione congiunta delle variabili $X_1, X_2, \ldots, X_m$ **condizionate dalle altre variabili** $X_1^*, X_2^*, \ldots, X_k^*$ **è data dal rapporto:**

$$P(X_1, X_2, \ldots, X_m | X_1^*, X_2^*, \ldots, X_k^*) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(X_1^* = x_1^*, X_2^* = x_2^*, \ldots, X_k^* = x_k^*)}$$

ESEMPIO

$$P(X_1, X_2 | X_3 = 1) = \frac{P(X_1, X_2, X_3)}{P(X_3 = 1)} =$$

X_1/X_2	-1	0	1
-1	$4/13$	$1/13$	0
0	$2/13$	$1/13$	$2/13$
1	0	$1/13$	$2/13$

$$P(X_1, X_2 | X_3 = 2) = \frac{P(X_1, X_2, X_3)}{P(X_3 = 2)} =$$

X_1/X_2	-1	0	1
-1	$1/11$	0	$2/11$
0	$1/11$	$1/11$	$2/11$
1	$3/11$	$1/11$	0

Distribuzioni congiunte marginali

Le variabili casuali multidimensionali consentono di definire le distribuzioni marginali di ogni sottoinsieme.

Dividiamo le "n" v.c. in due gruppi distinti:

VARIABILI CHE INTERESSANO: $\dot{X}_1,\ \dot{X}_2,\ldots,\dot{X}_m$

con k+m=n

VARIABILI CHE NON INTERESSANO: $X_1^*,\ X_2^*,\ldots,X_k^*$

Per ottenere la distribuzione (congiunta) delle variabili che interessano (marginali) Rispetto alle altre occorre sommare per le variabili che non interessano

$$P(\dot{X}_1,\dot{X}_2,\ldots,\dot{X}_m) = \sum_{x_1^*}\sum_{x_2^*}\ldots\sum_{x_k^*} P(X_1,X_2,\ldots,X_n)$$

Sommando si elimina l'influenza degli aspetti dell'esperimento che si vogliono tenere fuori.

Esempio

Per determinare $P(X_1,X_2)$ dobbiamo eliminare l'influenza di X_3 sommando - cella per cella- le due tabelle precedenti

$X2/X3$	-1	0	1
-1	$5/24$	$1/24$	$2/24$
0	$3/24$	$2/24$	$4/24$
1	$3/24$	$2/24$	$2/24$

Possiamo determinare le altre due distribuzioni congiunte-marginali eliminando di volta in volta l'influenza della terza variabile

$X1/X3$	1	2
-1	$5/24$	$3/24$
0	$5/24$	$4/24$
1	$3/24$	$4/24$

$X1/X3$	1	2
-1	$5/24$	$3/24$
0	$5/24$	$4/24$
1	$3/24$	$4/24$

E' possibile ottenere la marginale singola di ognuna delle altre v.c. usando una qualsiasi delle congiunte che la coinvolgono.

Esercizio

Si abbia la seguente distribuzione trivariata:

X_1	X_2	X_3	$P(X_1,X_2,X_3)$
1	1	1	0.05
1	1	2	0.10
1	2	1	0.15
1	2	2	0.20
2	1	1	0.20
2	1	2	0.15
2	2	1	0.10
2	2	2	0.05

Calcolare	*Alcune soluzioni*
$P(X_1=1)$	0.5
$P(X_1=1,X_2=2)$	0.35
$P(X_3<1.5)$	0.5
$P(X_1=1\ o\ X_3=2)$	0.7
$E(X_1)$	1.25
$P(X_1=1\|X_2=2)$	
$P(X_1=1,X_2=1\|X3=2)$	
$P(X_1=1,X_2=2\|X3=2)$	
$P(X_1\|X_2=2,X_3=2)$	$X_1=1,P(X_1=1)=0.4;$ $X_1=2,P(X_1=2)=0.6$

La distribuzione multinomiale

Un esperimento consiste di "n" prove indipendenti svolte in condizioni identiche. In ciascuna prova sono possibili "k" modalità distinte, anche qualitative:

$$(X_1,X_2,\ldots,X_k)$$

Le probabilità dei singoli risultati sono costanti di prova in prova

$$p_1,p_2,\ldots,p_k \geq 0;\quad \sum_{i=1}^{k} p_i = 1$$

Un modello adatto a tale esperimento è quello multinomiale con la seguente funzione di distribuzione:

$$P(X_1=x_1,X_2=x_2,\ldots,X_k=x_k) = \frac{n!}{x_1!*x_2!*\ldots*x_k!}p^{X_1}p^{X_2}*\ldots*p^{X_k}$$

$$x_1+x_2+\ldots+x_k = n$$

Esempio

Gli affidati di una banca sono:
X1=solvibili con p1=0.60,
X2=insolventi con p2=0.05,
X3=incerti, ma positivi con p3=0.30,
X4=incerti, ma negativi con p4=0.15.

Calcolare la probabilità che su n=10 ne risultino X1=4,X2=2,X3=1, X4=3

$$P(4,2,1,3)=\frac{10!}{4!2!1!3!}0.40^4*0.05^2*0.30*0.15^3=0.00081$$

Calcolare la probabilità che su n=20 ne risultino 5 di ciascun tipo:

$$P(5,5,5,5)=\frac{20!}{5!*5!*5!*5!}*0.6^5*0.05^5*0.30^5*0.15^5=0.000053$$

Ancora sulla multinomiale

Le distribuzioni marginali sono delle binomiali. Infatti, in ogni prova si verifica X=X oppure non si verifica e p_i rimane costante nelle prove indipendenti

$$P(X_i=x_i)=\binom{n}{x_i}p_i^{x_i}(1-p_i)^{n-x_i} \quad \textbf{con} \quad E(X_i)=np_i;\ \sigma^2(X_i)=np_i(1-p_i)$$

Note "k-1" variabili casuali componenti la variabile casuale multinomiale è nota anche la n-esima dato il vincolo di somma ad "n" dei risultati. Quindi la multinomiale ha (K-1) dimensioni

Le variabili componenti la multinomiale sono correlate (e quindi dipendenti)

$$Cov(X_i,X_j)=-np_ip_j \quad \forall \ i\neq j$$

L'aumento dei successi in X_i non può che avvenire a danno di X_j

V.C. continue multidimensionali

La capacità rappresentativa del modello rispetto all'esperimento aumenta se aumentano gli aspetti di cui riesce a tenere conto. Ciò vale anche per i fenomeni continui.

ESEMPIO
La valutazione del carico di lavoro di una unità di personale che svolge 4 diversi compiti tiene conto dei tempi di svolgimento di ciascuno. Se i compiti sono tra loro indipendenti un modello adatto è:

$$f(X_1,X_2,X_3,X_4)=\lambda_1*\lambda_2*\lambda_3*\lambda_4*e^{-\lambda_1x_1-\lambda_2x_2-\lambda_3x_3-\lambda_4x_4} \qquad Xi\geq 0$$

La definizione della funzione di densità ricalca quella bivariata

$$f(X_1,X_2,\ldots,X_n)\geq 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}f(X_1,X_2,\ldots,X_n)dx_1dx_2\ldots dx_n=1$$

$$P(X_1,X_2,\ldots X_n\in A)=\int_A\int\ldots\int f(X_1,X_2,\ldots,X_n)dx_1dx_2\ldots dx_n$$

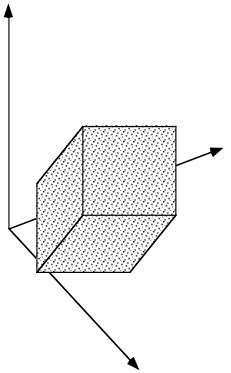
Gli eventi di cui si calcola la probabiiltà sono degli ipervolumi

Esempio: uniforme sul tetraedro

$$f(X_1,X_2,X_3)=\begin{cases}6 & per \quad x_1+x_2+x_3\leq 1; \qquad x_1,x_2,x_3\geq 0\\ 0\end{cases}$$

Ad ogni porzione del volume del tetraedro unitario la "f" assegna una densità di probabilità.

In questo modello la densità è costante



$$\int_0^1\int_0^{1-x_1}\int_0^{1-x_1-x_2}6dx_3dx_2dx_1=$$

$$6\int_0^1\int_0^{1-x_1}(1-x_1-x_2)dx_2dx_1=6\int_0^1 x_1-x_1^2dx_1$$

$$6\left(\frac{1}{2}-\frac{1}{3}\right)=1$$

Indipendenza di "n" variabili casuali

Siano $\{X_1, X_2, \ldots, X_n\}$ **"n" variabili casuali. Esse sono considerate indipendenti se**

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

Che è la naturale estensione della definizione data per il caso n=2

L'indipendenza può anche essere formulata in base alla funzione di ripartizione:

$$P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n} F(x_i)$$

Accomunando così le v.c. continue e discrete in una unica definizione

Riflessioni sull'indipendenza

L'indipendenza è una condizione molto forte a cui conseguono diversi risultati

- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo è qualsiasi loro sottoinsieme.**
- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo sono le rispettive trasformate**

$$\{g_1(x_1), g_2(x_2), \ldots, g_n(x_n)\}$$

- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo è qualsiasi combinazione di loro funzioni.**
- **Se "n" variabili casuali sono MUTUALMENTE INDIPENDENTI, cioè indipendenti due a due, questo non implica l'indipendenza n-dimensionale.**

Riflessione sull'indipendenza/2

Supponiamo che le tre variabili casuali discrete X,Y,Z abbiano distribuzione

$$P(X=x, Y=y, Z=z) = \frac{1}{4} \quad \text{se } (x,y,z) = \{(1,0,0);(0,1,0);(0,0,1);(1,1,1)\}$$

Come si vede, le coppie di v.c. sono indipendenti

$$P(X=0)=\frac{1}{2};\ P(X=1)=\frac{1}{2};\quad P(Y=0)=\frac{1}{2};\ P(Y=1)=\frac{1}{2};\quad P(Z=0)=\frac{1}{2};\ P(Z=1)=\frac{1}{2};$$

$$P(X=x, Y=y)=\frac{1}{4};\quad P(X=x, Z=z)=\frac{1}{4};\quad P(Y=x, Z=z)=\frac{1}{4}$$

La terna non è però indipendente

$$P(X=1, Y=0, Z=0)=\frac{1}{4} \neq P(X=1)*P(Y=0)*P(Z=0)=\frac{1}{8}$$

Variabili casuali I.D.

Le variabili casuali si dicono IDENTICAMENTE DISTRIBUITE se hanno la stessa funzione di ripartizione

$$P(X_1 \le x_1) = P(X_2 \le x_2)$$

ESEMPIO: **Supponiamo che X_1 ed X_2 abbiano distribuzione esponenziale con $\lambda=2$**

$$F(X_1) = 1 - e^{-2X_1};$$
$$F(X_2) = 1 - e^{-2X_2}$$

Le due variabili casuali descrivono lo stesso aspetto di un esperimento osservato in condizioni diverse ovvero REPLICHE DIVERSE della stesso esperimento

ID non significa che sono uguali cioè $X_1 \ \ I.D. \ \ X_2 \not\Rightarrow P(X_1 = X_2) = 1$

ESEMPIO: T e C si riferiscono al lancio di una moneta regolare con rispettivo successo se esce testa e se esce croce.

Sono quindi entrambe bernoulliane con p=0.5

	S	I
T	$\frac{1}{2}$	$\frac{1}{2}$
C	$\frac{1}{2}$	$\frac{1}{2}$

Tuttavia P(T=C)=0 dato che sono incompatibili

Variabili casuali I.I.D.

- **Molti esperimenti sono formati da "n" prove ripetute (repliche)**
- **Ogni prova è espressa da una variabile casuale (singola o multipla,discreta o continua).**
- **Le prove sono indipendenti.**
- **L'esito di ogni prova è modellabile con la stessa variabile casuale**

In tali casi si determinano "n" v.c. INDIPENDENTI ED IDENTICAMENTE DISTRIBUITE

Campione casuale

Le variabili casuali $\{X_1, X_2, \ldots, X_n\}$ ottenute da "n" repliche di una stessa prova sono solo una parte di di quelle che potevano essere effettuate

perciò la n-tupla è detta CAMPIONE CASUALE di ampiezza "n" estratto dalla "popolazione" delle possibili repliche.

I valori osservati ovvero le osservazioni campionarie: $(x_1, x_2, \ldots, x_n)$ sono la realizzazione del campione.

Spesso la funzione ri ripartizione "F" e/o la v.c. "X" sono indicate come la POPOLAZIONE da cui si "estrae" il campione

La distribuzione del campione

L'estrazione di un campione di ampiezza "n" dà luogo ad una v. c. n-dimensionale

$$\{X_1, X_2, \ldots, X_n\}$$

La funzione di distribuzione o densità congiunta del campione, data l'indipendenza, delle estrazioni sarà:

$$f(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i)$$

dove "f" è il modello di variabile casuale alla base dell'esperimento ovvero la distribuzione della popolazione da cui si estrae il campione casuale.

N.B. l'idea di indipendenza può essere estesa anche alle "estrazioni senza reimmissione" cioè variabili dipendenti,purché "n" sia piccolo rispetto alla popolazione delle possibili repliche.

Esempio

Supponiamo che il sesso alla nascita sia equiprobabile. Sia X il numero di figlie femmine in una famiglia con 5 figli.

La distribuzione di probabilità della popolazione delle famiglie è binomiale:

$$P(X = x) = \binom{5}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{5-x} \quad x = 0,1,2,3,4,5$$

Si sceglie a caso un campione casuale di n=3 famiglie di 5 figli. Costruiamo la funzione di distribuzione

$$P(X_1 = x_1) = \binom{5}{x_1}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{2}\right)^{5-x_1} \quad x_1 = 0,1,2,3,4,5$$

$$P(X_2 = x_2) = \binom{5}{x_2}\left(\frac{1}{2}\right)^{x_2}\left(\frac{1}{2}\right)^{5-x_2} \quad x_2 = 0,1,2,3,4,5$$

$$P(X_3 = x_3) = \binom{5}{x_3}\left(\frac{1}{2}\right)^{x_3}\left(\frac{1}{2}\right)^{5-x_3} \quad x_3 = 0,1,2,3,4,5$$

$$\Longrightarrow \begin{aligned} & P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ & = P(X_1 = x_1) * P(X_2 = x_2) * P(X_3 = x_3) \\ & = \binom{5}{x_1} * \binom{5}{x_2} * \binom{5}{x_3} * \left(\frac{1}{2}\right)^{15} \end{aligned}$$

Problema: quali ragioni giustificano p=0.5?

il rapporto dei sessi alla nascita è meglio approssimato da $\frac{105M}{100F}$

Esercizio

Nello studio sull'affidabilità di una componente elettronica si ipotizza che la sua durata sia una variabile casuale "Y" con funzione di densità esponenziale

$$f(y) = \lambda e^{-\lambda y} \quad \text{per} \quad y > 0$$

Si sottopone a controllo un campione casuale di n=5 componenti ovvero si replica 5 volte l'esperimento "durata della componente elettronica".

Quale sarà la loro densità congiunta?

$$f(y_1, y_2, \ldots, y_5) = \prod_1^5 f(y_i) = \prod_1^5 \lambda e^{-\lambda y_i} = \lambda^5 e^{-\lambda \sum_{i=1}^5 y_i}$$

Ad esempio, la probabilità che tutte le 5 componenti siano ancora in funzione dopo y=10 minuti è data da

$$P(Y_1 \geq 10, Y_2 \geq 10, \ldots, Y_5 \geq 10) = 1 - F(Y_1 < 10, Y_2 < 10, \ldots, Y_5 < 10) = e^{-\lambda \sum_{i=1}^5 y_i} = e^{-\lambda 50}$$

Se λ fosse noto, la valutazione di questa probabilità non sarebbe un problema

Combinazione lineare di variabili casuali

Le statistiche che più ci interessano sono delle funzioni del campione casuale $\{X_1, X_2, \ldots, X_n\}$ del tipo:

$$C_n = \sum_{i=1}^n w_i X_i$$

dove gli $\{w_i\}$ sono dei "pesi" che esprimono il contributo a "C_n" delle diverse v.c. facenti parti della combinazione

$$\{X_1, X_2, \ldots, X_n\}$$

In particolare w_i misura la variazione che interviene in "C_n" per una variazione unitaria in X_i

La "C_n" è detta combinazione lineare perché le v.c. vi entrano con potenza uno.

Valore atteso di C_n

Si era già dimostrato che E(X+Y)=E(X)+E(Y). Lo stesso risultato può essere estesc alla combinazione lineare di v.c.

$$E[C_n] = E\left[\sum_{i=1}^n w_i X_i\right] = \sum_{i=1}^n E[w_i X_i] = \sum_{i=1}^n w_i E[X_i]$$

ESEMPIO:
una libreria vende tre diversi dizionari in brossura al prezzo di 45, 55, 70 mila lire. Siano X_1, X_2, X_3 il numero di volumi venduti per ciascuno tipo in un dato periodo

La variabile casuale incasso per i dizionari sar $C_3 = 45X_1 + 55X_2 + 70X_3$

L'incasso atteso è quindi:

$$E(C_3) = E(45X_1 + 55X_2 + 70X_3) = 45 * E(X_1) + 55 * E(X_2) + 70 * E(X_3)$$

Esempio

Se le v.c. hanno lo stesso valore atteso "μ" (non necessariamente la stessa "F") allora il valore atteso della combinazione è proporzionale al comune valore atteso

$$\text{Se} \quad E(X_i) = \mu \quad \text{per} \quad i = 1, 2, \ldots, n \Rightarrow E(C_n) = \sum_{i=1}^n w_i E(X_i) = \mu \sum_{i=1}^n w_i$$

In particolare, se i pesi hanno somma unitaria, allora

$$E(C_n) = \sum_{i=1}^n w_i E(X_i) = \sum_{i=1}^n w_i \mu = \mu \sum_{i=1}^n w_i = \mu$$

Nel lancio di 6 dadi il valore atteso di ciascun esito è μ=3.5. Il valore atteso della somma sarà:

$$E\left(\sum_{i=1}^6 X_i\right) = \sum_{i=1}^6 E(X_i) = \sum_{i=1}^6 3.5 = 3.5 * 6 = 21$$

La varianza di C_n

Nel valore atteso di una combinazione lineare di variabili casuali non ha avuto alcun ruolo la dipendenza o meno delle stesse v.c.

La dipendenza, o meglio, l'incorrelazione, entra in gioco per la varianza di C_n.

$$\sigma^2(C_n) = \sum_{i=1}^{n} w_i^2 \sigma^2(X_i) + 2\sum_{i=1}^{n-1}\sum_{\substack{j=1 \\ i<j}}^{n} w_i w_j \mathrm{Cov}(X_i, X_j)$$

Se le covarianze sono nulle la varianza di C_n è una combinazione lineare delle varianze.

$$\sigma^2(C_n) = \sum_{i=1}^{n} w_i^2 \sigma^2(X_i)$$

Al livello di relazioni lineari, indipendenza e incorrelazione sono condizioni equivalenti

Applicazione:varianza di $\bar{x}$

Supponiamo di disporre di un campione casuale semplice estratto con reimmissione.

Sfruttando l'indipendenza e quindi l'incorrelazione otteniamo:

$$\sigma^2(\bar{x}) = \sigma^2\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Esempio

Sia D una v.c. discreta con distribuzione geometrica il cui parametro sia p.

Si vuole conoscere quante prove W siano necessarie per ottenere "r" successi.

W è la somma di "r" variabili geometriche: G_1 prove per il 1° successo, G_2 prove per il 2° e così via fino a G_r:

$$W = G_1 + G_2 + \ldots + G_r$$

Sfruttando l'indipendenza e quindi l'incorrelazione otteniamo:

$$E(W) = \sum_{i=1}^{r} E(G_i) = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}$$

$$\sigma^2(W) = \sum_{i=1}^{r} \sigma^2(G_i) = \sum_{i=1}^{r} \frac{1-p}{p^2} = \frac{r(1-p)}{p^2}$$

Come si vede non si è fatto altro che definire la v.c. Binomiale negativa

Ancora sulla varianza di C_n

Se la sequenza delle "n" repliche è assimilabile ad una estrazione senza reimmissione da una popolazione di "N" allora

- **Ogni X_i ha la stessa varianza σ^2**
- **Ogni coppia (X_i, X_j) ha la stessa covarianza.**

$$Cov(X_{i_1}, X_{i_2}) = d \quad per\ qualunque\ (i_1, i_j)$$

Ne consegue: $\sigma^2(C_n) = \sigma^2 \sum_{i=1}^{n} w_i^2(X_i) + 2d \sum_{i=1}^{n} \sum_{j<i} w_i w_j$

Se n=N allora C_n è costante perché riguarda l'intera popolazione e quindi: $\sigma^2(C_N)=0$. L'unico valore di "d" compatibile con le due formulazioni è il seguente:

$$d = \frac{-\sigma^2 \sum_{i=1}^{N} w_i^2}{2\sum_{i=1}^{N}\sum_{j<i} w_i w_j}$$

Ancora sulla varianza di $\bar{x}$

Nel caso del valore atteso la combinazione lineare ha pesi con somma uno, quindi

$$d = \frac{-\sigma^2}{N-1}$$

Ne consegue:

$$\sigma^2(\bar{x}) = \frac{1}{n^2}\left[n\sigma^2 - \frac{\sigma^2}{N-1}2\frac{n(n-1)}{2}\right] = \frac{\sigma^2}{n} * \left(\frac{N-n}{N-1}\right)$$

Dal confronto delle due formule si vede che la media campionaria è meno variabile se il campionamento avviene senza reimmissione

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

Questo non sorprende date le minori possibilità di sviluppo dell'universo dei campioni

Particolari combinazione di V.C.

Siamo interessati ad una particolare combinazione lineare: la somma ponderata di variabili casuali

$$C_n = \sum_{i=1}^{n} w_i X_i \qquad \text{con} \quad w_i \geq 0; \quad \sum_{i=1}^{n} w_i = 1$$

La distribuzione di "C" è difficile da determinare. Tuttavia, se le X sono Normali interviene una fondamentale proprietà di questa distribuzione: la RIPRODUCIBILITA'

$$C_n \sim N\left(\sum_{i=1}^{n} w_i \mu_i, \sum_{i=1}^{n} w_i^2 \sigma_i^2\right)$$

Cioè la somma ponderata di "n" variabili casuali gaussiane è pure gaussiana Con valore atteso e varianza date dalle somme di quelle delle variabili.

Se poi si tratta di variabili casuali I.I.D si ha: $C_n \sim N\left(\mu, \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i^2\right)$

Statistiche L

Gli stimatori che più ci interessano sono delle funzioni delle osservazioni campionarie $\{X_1, X_2, \ldots, X_n\}$ del tipo:

$$L_n = \sum_{i=1}^{n} w_i X_{(i)}$$

dove gli $\{w_i\}$ sono dei "pesi" che esprimono il contributo a "L_n" delle diverse osservazioni facenti parti dello stimatore.

La notazione $X_{(i)}$ indica che i valori delle osservazioni campionarie sono state ordinate in senso crescente.

Le statistiche L (cioè lineari) sono tali perché le osservazioni campionarie vi compaiono con potenza uno

Applicazione: media campionaria

Supponiamo che le: $\{X_1, X_2, \ldots, X_n\}$ **siano repliche indipendenti di un v.c. Normale con media "μ" e varianza** σ^2

Come si distribuisce la statistica media campionaria ?

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \sum_{i=1}^{n} w_i X_i \quad \text{con} \quad w_i = 1/n \quad \text{per ogni "i"}$$

Applicando il risultato precedente si ha $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \sqrt{n}\left(\frac{X-\mu}{\sigma}\right) \sim N(0,1)$

Quindi, per la media campionaria di un campione estratto da una v.c. Normale si ripropone lo stesso modello.

Esempio

Un particolare composto chimico ha un grado di impurità descritto da una v.c. con
media μ=4.0 g e deviazione standard σ=1.5 g.

Un lotto di questo prodotto è accettato se, scelte 50 confezioni secondo lo schema del campione casuale con reimmissione, la loro impurità MEDIA è compresa tra 3.5 e 3.8 g.Qual'è la probabilità che il lotto venga accettato?

Per rispondere dobbiamo conoscere la distribuzione della popolazione cioè il modello che descrive il comportamento dell'impurità nel composto.

Se è di tipo Normale allora:

$$P(3.5 \le \bar{X} \le 3.8) = P\left(\frac{3.5-4.0}{0.2121} \le Z \le \frac{3.8-4.0}{0.2121}\right) = \phi(-0.94) - \phi(-2.36) = 0.1645$$

Senza l'informazione sulla normalità nulla o quasi poteva essere detto sul valore atteso della media delle "n" variabili casuali.

Applicazione: totale campionario

il totale è una combinazione lineare con tutti i pesi pari ad uno. Quindi:

$$Q_n = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right)$$

ESEMPIO

La durata degli esami orali di statistica (sull'intero programma) è una v.c. Normale con μ=1.5 ore e σ=0.35 ore

Si presentano 5 frequentanti, scelti a caso, i cui tempi sono le $\{X_1, \ldots, X_5\}$

Qual'è la probabilità che il povero docente debba faticare per 6 - 8 ore?

$$Q = \sum_{i=1}^{5} X_i \sim N(5\mu, 5\sigma^2) = N(7.5, 0.6125)$$

Ne consegue:

$$P(6 \le Q \le 8) = P\left(\frac{6-7.5}{0.783} \le Z \le \frac{8-7.5}{0.783}\right) = P(-1.92 \le Z \le 0.64)$$
$$= \phi(0.64) - \phi(-1.92) = 0.7115$$

Applicazione: varianza campionaria

Se Z è una c.c. gaussiana standardizzata, è noto che:

$$E(Z) = 0,\ E\left(Z^2\right) = 1,\ E\left(Z^3\right) = 0,\ E\left(Z^4\right) = 3$$

Ci interessa la statistica $c^2 = \sum_{i=1}^{n} Z_i^2;\quad dove\ \ Z_i \sim N(0,1)$

Supponendo che il campionamento sia <u>con reimmissione</u>, avremo:

$$E\left(c^2\right) = E\left(\sum_{i=1}^{n} Z_i^2\right) = \sum_{i=1}^{n} E\left(Z_i^2\right) = \sum_{i=1}^{n} 1 = n$$

$$\sigma^2\left(c^2\right) = \sigma^2\left(\sum_{i=1}^{n} Z_i^2\right) = \sum_{i=1}^{n} \sigma^2\left(Z_i^2\right) = \sum_{i=1}^{n} E\left\{Z_i^4 - \left[E\left(Z_i^2\right)\right]^2\right\} = \sum_{i=1}^{n} (3-1) = 2n$$

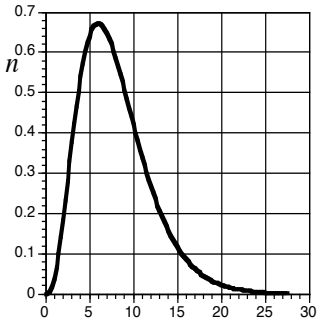
Valore atteso e varianza dipendono solo dal numero di variabili considerate

La distribuzione χ^2

La funzione di densità di c^2 , se Z è normale standardizzata, è:

$$f(x;n) = \left(\frac{x}{2}\right)^{\frac{n}{2}} \frac{e^{-\frac{x}{2}}}{d_n};\quad x > 0;\ d_n\ \ crescente\ con\ \ n$$

n=6



nota come distribuzione del chi quadro.

$$E\left(\chi^2\right) = g,\qquad Var\left(\chi^2\right) = 2g$$

Tale densità dipende da un solo parametro: "g" noto come "gradi di libertà" è fa riferimento al numero di informazioni da trasmettere per comunicare c^2 ,

Se x_1 è una v.c. chi quadro con g_1 gradi di liberà e x_2 una chi quadro con g_2 gradi di libertà allora $x_1 \pm x_2$ è una v.c. chi quadro con $g_1 \pm g_2$ gradi di libertà

Applicazione: varianza campionaria/2

Consideriamo la statistica: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{(n-1)}$ $dove$ $X_i \sim N(\mu, \sigma^2)$

Di questa non riusciamo a trovare la distribuzione. Ci riusciamo per:

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

Infatti,

$$c^2 = \frac{1}{\sigma^2}\left[\sum X_i^2 - n\bar{x}^2 \pm n\mu^2 \pm 2n\mu\bar{x}\right] = \frac{1}{\sigma^2}\left[\sum (X_i - \mu)^2 - n(\bar{x} - \mu)^2\right]$$

$$= \sum\left(\frac{X_i - \mu}{\sigma^2}\right)^2 - \frac{n(\bar{x} - \mu)^2}{\sigma^2}$$

il primo addendo è una chi quadro con "n" gradi di libertà la seconda è il quadrato di una normale standardizzata ed è perciò una chi quadro con 1 grado di libertà.

Applicazione: varianza campionaria/3

Ne consegue che $c^2 \sim \chi^2(n-1)$ **cioè una chi quadro con n-1 gradi di libertà.**

La perdita di un grado è dovuta al calcolo della media campionaria: noti (n-1) Valori e nota la media campionaria, l'n-esimo è determinato.

ESEMPIO

Un campione casuale di ampieza n=11 è stato estratto con reimmissione da unav.c. normale con varianza σ^2=5. Calcolare la probabilità che la varianza campionaria sia inferiore a 1.97015.

$$P\left(s^2 \le 1.97015\right) = P\left[\frac{(11-1)}{5}s^2 \le 3.9403\right] = P\left(c^2 \le 3.9403\right) = 5\%$$

Per ogni valore dei gradi di libertà c'è una distribuzione χ^2.

inferenza statistica

Se non si hanno dati attendibili su tutta la popolazione allora si deve trattare con un campione.

Il campione non interessa di per sé, ma in quanto consente di arrivare alla popolazione da cui è stato estratto.

Il processo induttivo dal NOTO (campione) all'iNCOGNITO (popolazione) prende il nome di INFERENZA STATISTICA.

Per proseguire dobbiamo però ipotizzare che il meccanismo di selezione delle unità sia soggetto alla sorte

Ad ogni campione deve essere possibile associare la probabilità di estrarlo (VEROSIMIGLIANZA)

Logica della Inferenza statistica

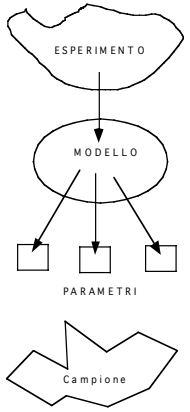
Le situazioni in cui la statistica si è più affermata sono gli esperimenti replicabili.

Le esigenze conoscitive si limitano spesso a poche caratteristiche dell'esperimento: valore atteso e varianza di una o più variabili.

Tali caratteristiche sono spesso i parametri del modello che descrive il comportamento delle variabili casuali.

Il modello di casualità è approssimabile dalla variabile casuale normale

Come sfruttare al meglio le informazioni del campione per determinare il valore dei parametri?



Esempio introduttivo

Una partita di N=5'000 articoli ne include alcuni difettosi. Vogliamo valutare la percentuale di pezzi difettosi.

I tempi ed i costi consentono di esaminarne n=40 che scegliamo casualmente.

Supponiamo che ricorrano le condizioni del modello bernoulliano cioè la v.c. che descrive la popolazione è:

$$f(X;\theta) = \theta^{x}(1-\theta)^{1-x};\ x = 0,1$$

La distribuzione di probabilità del campione è data da:

$$\prod_{i=1}^{40} f(X_i;\theta) = \prod_{i=1}^{40} \theta^{x_i}(1-\theta)^{40-x_i} = \theta^{\sum_{i=1}^{40} x_i}(1-\theta)^{40-\sum_{i=1}^{40} x_i}$$

Determiniamo la distribuzione della statistica:

$$T = \sum_{i=1}^{n} X_i \quad dove \quad X_i = \begin{cases} 1 \\ 0 \end{cases}$$

Esempio introduttivo/2

In questo caso la distribuzione di T è ottenibile in forma esplicita:

$$f(T;\theta) = \binom{40}{T}\theta^{T}(1-\theta)^{40-T};\ T = 0,1,\ldots,40$$

cioè una binomiale con n=40 e p=θ. Questo è lo stato informativo a priori.

Supponiamo ora di aver riscontrato T=5 articoli difettosi fra i 40 esaminati. Questa è l'informazione sperimentale.

La probabilità di osservare proprio T=5 è: che è funzione del solo parametro θ.

$$f(5;\theta) = \binom{40}{5}\theta^{5}(1-\theta)^{35}$$

Compito dell'inferenza statistica è di far confluire lo stato informativo a priori e le evidenze empiriche in una valutazione plausibile del parametro o della caratteristica di interesse

L'Esperimento di Rutherford-Geiger

X=numero di particelle alfa per numero di intervalli di 7.5 secondi

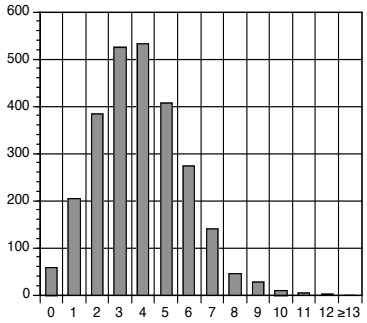
X	Intervalli
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	2
≥13	0
	2608

L'andamento ricorda la Poisson.

Supponiamo che le frequenze osservate coincidano con la Probabilità.

$$\frac{408}{2608} = 0.1564 \cong P(X=5)$$

Quanto varrebbe λ?



Poiché λ=E(X) si ha

$$\hat{\lambda} = \sum_{i=1}^{12} i * P(X=i) = 1*\frac{203}{2608} + 2*\frac{383}{2608} \ldots + 12*\frac{2}{2608} = 3.87$$

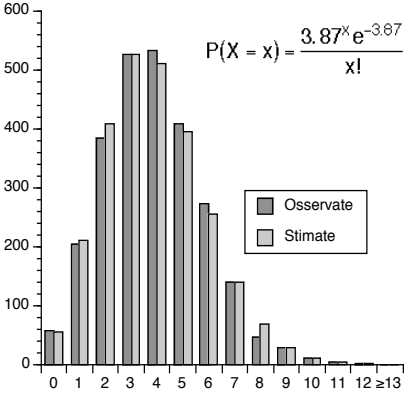
il cappello su λ ci ricorda che non è un valore noto, ma solo una congettura in base ai dati sperimentali.

L'esperimento di Rutherford-Geiger/2

La legge dei grandi numeri autorizza a sostituire alle probabilità (incognite) le lore frequenze. Supponiamo che n=2608 sia "grande"

X	Intervalli	Stime
0	57	54
1	203	210
2	383	407
3	525	526
4	532	510
5	408	394
6	273	255
7	139	140
8	45	68
9	27	28
10	10	11
11	4	4
12	2	1
≥13	0	0
	2608	2608

Valori stimati e osservati sono molto vicini.



il modello ha il grande vantaggio della sintesi. Se si deve trasmettere da un posto ad un altro basta il solo messaggio "Poisson con λ=3.87" invece dell'intera tabella

Le procedure inferenziali

Ciò che interessa il nostro corso è:

LA STIMA DEI PARAMETRI

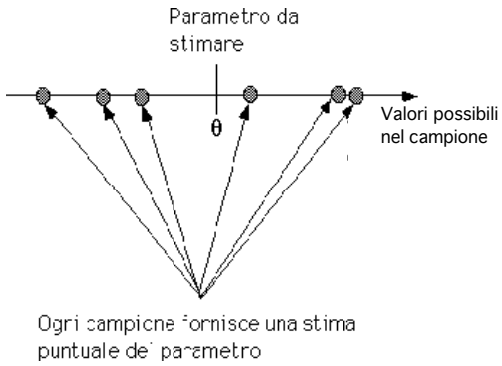
- **PUNTUALE:** quando si propone un singolo valore come stima di un parametro dell variabile casuale.
- **INTERVALLARE:** quando si propone un ventaglio di valori ragionevoli come stime.

LA VERIFICA DI IPOTESI

Da esperienze precedenti o dalla logica delle indagini si può supporre che i parametri abbiano determinati valori. Sono compatibili con le risultanze campionarie?

La stima puntuale

E' la procedura più semplice: in base alle osservazioni campionarie si ottiene il valore da sostituire al parametro da stimare



C'è da aspettarsi un certo scarto tra la stima puntuale ed il parametro incognito, ma in genere non conosciamo né l'entità né il segno dell'errore

Stima puntuale basata su modelli

Si conosce il modello della casualità dell'esperimento, ovvero le condizioni sperimentali richiamano un particolare modello di cui però ignoriamo i parametri.

- **METODO DEI MOMENTI:** Eguagliando i momenti campionari e quelli della popolazione si ottengono delle equazioni da risolvere rispetto ai parametri.
- **METODO DELLA MASSIMA VEROSIMIGLIANZA:** A parità di condizioni si sceglie il valore che dà la massima probabilità ai fatti osservati

Il primo è più semplice dal punto di vista operativo, il secondo è più interpretabile

La scelta avverrà in base alle condizioni sperimentali ed alle diverse proprietà teoriche delle due procedure

Metodo dei momenti

Sia (X_1, X_2, … X_n) un campione casuale semplice da una v.c. $f(X;\theta_1,\theta_2,\ldots,\theta_k)$

il metodo dei momenti proposto da K. Pearson nel 1894 propone le relazioni

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} X_i}{n} = E(X)$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} X_i^2}{n} = E\left(X^2\right)$$

……

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} X_i^k}{n} = E\left(X^k\right)$$

Si usano tante relazioni quanti sono i parametri incogniti del modello.

Il sistema di equazioni è risolto rispetto ai parametri come funzioni dei valori campionari.

Esempio

Le aspettative inflazionistiche degli operatori di borsa sono descritte dal modello

$$f(X;\theta_1,\theta_2)=\frac{1}{\theta_2-\theta_1} \quad per \;\; \theta_1 \le X \le \theta_2$$

Dobbiamo stimare i parametri.

Disponiamo del campione casuale: (2.2, 1.8, 1.4, 1.6) (tassi previsti di inflazione)

$$\frac{\sum_{i=1}^{4} X_i}{4}=1.75; \quad \frac{\sum_{i=1}^{4} X_i^2}{4}=8.11712;$$

il metodo dei momenti suggerisce di risolvere le equazioni:

$$1.75=\frac{\theta_1+\theta_2}{2}; \quad 8.11712=\frac{4}{3}\left(\frac{\theta_1+\theta_2}{2}\right)^2-\frac{\theta_1\theta_2}{3}$$

$$\hat{\theta}_1=\bar{x}-\hat{\sigma}\sqrt{3}; \hat{\theta}_2=\bar{x}+\hat{\sigma}\sqrt{3} \Rightarrow \hat{\theta}_1=1.2377, \;\; \hat{\theta}_2=2.2623,$$

Difetti del metodo

- **Le relazioni potrebbero essere non lineari e quindi può succedere che le soluzioni non siano esplicite ovvero esserci più di una soluzione**
- **Non è univoca la scelta dei momenti: ad esempio si possono usare quelli centrati e quelli intorno alla media aritmetica.**

Disponiamo del campione casuale (3, 5, 7, 2, 8) da un modello di Poisson. Poiché c'è un solo parametro si usa una sola relazione, ma quale?

$$\frac{\sum_{i=1}^{n} X_i}{n}=\lambda \Rightarrow \hat{\lambda}=\frac{25}{5}=5$$

$$\frac{\sum_{i=1}^{n} X_i^2}{n}=\lambda^2+\lambda \Rightarrow \hat{\lambda}=5.018$$

In questo caso le due stime sono vicine, ma non sempre è così.

Metodo della massima verosimiglianza

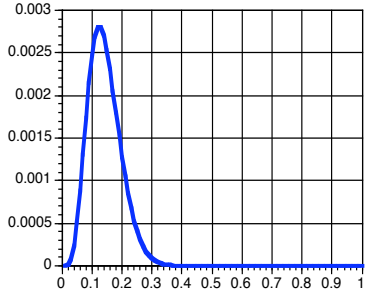
Una partita di N=5000 prodotti ne include certi con difetti. Che percentuale?

Usiamo un campione casuale semplice di n=40 in modo che valgano le condizioni della v.c. di Bernoulli.

La distribuzione del campione è

$$f(X_1,X_2,\ldots,X_n;\theta)=\theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i}$$

$$\theta^T(1-\theta)^{n-T}$$



La percentuale di successi richiede lo studio della distribuzione binomiale:

$$f(T;\theta)=\binom{n}{T}\theta^T(1-\theta)^{n-T}; \;\; T=\sum_{i=1}^{n} X_i$$

Supponiamo trovare T=5. Qual'è il valore di θ più compatibile con tale risultato?

Se scoprissimo che θ>0.3 ne saremmo molto sorpresi; ci sembrano plausibili valori tra 0.11 e 0.13

La funzione di verosimiglianza

Consideriamo la funzione di distribuzione o di densità del campione

$$f(X_1,X_2,\ldots,X_n;\theta)=\prod_{i=1}^{n} f(X_i;\theta)$$

Poiché la f(X;θ) è positiva nei punti che interessano conviene eliminare i prodotti e lavorare con le somme

$$Ln\left[f(X_1,X_2,\ldots,X_n;\theta)\right]=\sum_{i=1}^{n} Ln\left[f(X_i;\theta)\right]$$

Eliminiamo inoltre ogni addendo e/o fattore che non dipenda dai parametri da stimare.

La funzione L(θ) che ne risulta è la funzione di verosimiglianza del campione

La funzione di verosimiglianza dà valori proporzionali alla probabilità di osservare il campione per ciascun valore dei parametri

La stima di massima verosimiglianza

Tale metodo suggerisce di scegliere come valore presunto di θ il valore che dia la massima probabilità ai fatti osservati (campione)

$$\hat{\theta} \ \ tale\ che \quad L\left(\hat{\theta}\right) \geq \ L(\theta) \ \ per\ \ ogni \ \ \theta \in \Theta$$

dove Θ è lo spazio parametrico (insieme dei valori ammissibili per θ)

Molto spesso la stima di massima verosimiglianza è ottenuta dalla relazione:

$$L'(\theta) = 0; \quad dove \quad L'(\theta) = \frac{dL(\theta)}{d\theta}$$

N.B. Non sempre la funzione di verosimiglianza è dotata di derivate

Esempio

In una fabbrica che produce lampadine, in fase di controllo, si osserva la loro durata. La vita media X di ognuna è ritenuta un' esponenziale

$$f(X;\lambda) = \lambda e^{-\lambda X} \quad \text{per } X > 0$$

Ricordare che E(X)=1/λ

☐ **Scegliamo un campione di n=20 lampadine**

☐ **Ogni rilevazione genera una v.c.** $f(X_i;\lambda) = \lambda e^{-\lambda X_i} \quad \text{per } X_i > 0$

La funzione di verosimiglianza è

$$L(\lambda) = nLn(\lambda) - \lambda \sum_{i=1}^{n} X_i \Rightarrow \frac{n}{\lambda} - \sum_{i=1}^{n} X_i \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{x}}$$

Stima di "λ"

$x = \{13,15,18,10,16,12,13,19,11,13,12,11,12,14,10,14,12,18,11,12\} \quad \frac{1}{\hat{\lambda}} = 13.35 \Rightarrow \hat{\lambda} = 0.075$

Esercizio

Ho lanciato in aria cinque monete per n=20 volte ed ho riscontrato che un numero medio di teste pari a 2.47

$$f(X_1, X_2, \ldots, X_n; \theta) = \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$Ln(f) = \sum_{i=1}^{n} Ln\left[\binom{n}{x_i}\right] + x_i Ln(p) + (n - x_i) Ln(1-p)$$

$$L(\theta) = Ln(p) \sum_{i=1}^{n} x_i + Ln(1-p) \sum_{i=1}^{n} (n - x_i)$$

La stima di massima verosimiglianza si ottiene risolvendo: L'(θ)=0

$$\frac{\sum_{i=1}^{n} X_i}{p} - \frac{\sum_{i=1}^{n} (n - X_i)}{1-p} \Rightarrow \hat{p} = \frac{\bar{x}}{n} = \frac{2.47}{5} = 0.494$$

Difetti del metodo

La soluzione potrebbe non essere unica. ---- Modello di Laplace

$$f(X_i;\theta) = e^{-\frac{|X_i - \theta|}{2}} \quad \textbf{che implica} \quad L(\theta) = -\frac{\sum_{i=1}^{n} |X_i - \theta|}{2}$$

essendo negativa, basta cercare il minimo. Se "n" è dispari la soluzione è univoca: la mediana $x_{\left(\frac{n}{2}+1\right)}$ ma se "n" è pari, mancherà l'univocità

La soluzione potrebbe non esistere in una forma esplicita cioè L(θ) è tanto complessa che si può dare solo una soluzione approssimata con il computer

La soluzione potrebbe non esistere. --- Uniforme (0, θ)

$$f(X;\theta) = \frac{1}{\theta}; \ per \ \ 0 \leq x \leq \theta \Rightarrow L(\theta) = -nLn(\theta)$$

La scelta è ora molto ardua

In genere il metodo funzione male se il campo di variazione della X dipende da θ

Confronto Momenti/Verosimgilianza

Si è ottenuto il campione casuale (64, 32, 16, 8) dalla v.c. continua di Pareto

$$f(X_i;\theta) = \frac{\theta}{X_i^{\theta+1}};\ X_i > 1$$

La stima di massima verosimiglianza si ottiene risolvendo

$$L(\theta) = nLn(\theta) - (1+\theta)\sum_{i=1}^{n} Ln(x_i)$$

$$L'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} Ln(x_i) \Rightarrow \hat{\theta} = Ln(Mg) = 3.1192$$

La stima dei momenti si ottiene da $\bar{x} = \frac{\theta}{\theta - 1} \Rightarrow \hat{\theta} = \frac{\bar{x}}{\bar{x} - 1} = 1.0345$

Le due stime sono molto diverse e porterebbero a decisioni diverse

La stima senza modello

Non sempre si può individuare un modello di v.c. per un esperimento e quindi non è sempre nota la funzione di verosimiglianza.

In questi casi la nostra attenzione si rivolge ad alcuni aspetti del modello quali media e varianza che possono fare a meno del modello.

Le caratteristiche della popolazione sono stimate in base a statistiche campionarie scelte in modo opportuno.

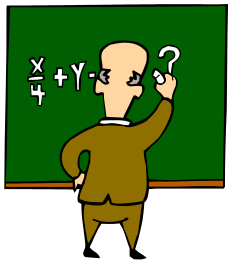
Una statistica è detta naturale se ciò che si calcola nel campione è scelto in stretta analogia con ciò che si deve stimare nella popolazione

il presupposto su cui si baseranno i nostri ragionamenti è che il campione sia tanto grande da richiamare il teorema del limite centrale

Gli stimatori

Lo stimatore è una funzione NOTA dei valori inclusi in un campione casuale. Il suo valore è la STIMA

E' caratteristica quantitativa della popolazione dalla quale il campione è stato estratto.



Esempi di stimatori:

$$\text{Totale: } Q = \sum_{i=1}^{n} X_i;\ \text{Campo di variazione: } R = X_{max} - X_{min}$$

$$\text{Media}: \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n};\ \text{Varianza}: s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1};\ \text{frazione di successi: } \pi = \frac{S_n}{n}$$

Uno stimatore è detto naturale se ciò che si calcola nel campione è in stretta analogia con ciò che si deve stimare nella popolazione

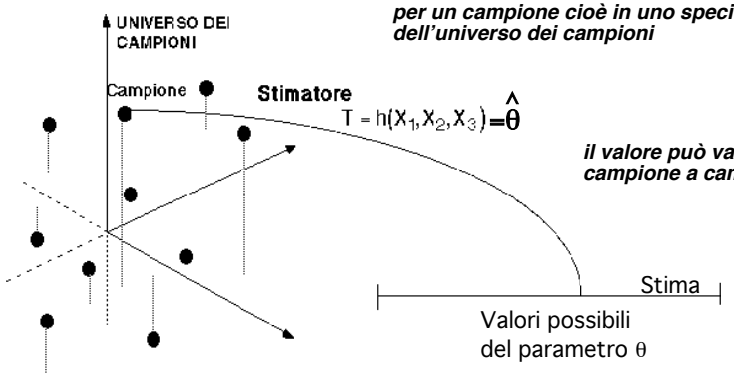
Stimatore e stima

ESEMPIO: Quale stipendio si può aspettare la manager di una USL?

Si sceglie un campione casuale diciamo di n=3 manager già in servizio e si calcola il valore atteso della loro retribuzione. Supponiamo che sia $\bar{x} = 65$

il valore "65 mila euro" è una STIMA del salario ipotetico, la media campionaria è uno STIMATORE del salario.

La stima è il valore assunto dallo stimatore per un campione cioè in uno specifico punto dell'universo dei campioni



il valore può variare da campione a campione

Esempio

L'estrazione del campione produce la n-tupla $(x_1, x_2, \ldots, x_n)$ i cui elementi sono le osservazioni campionarie

Ogni n-tupla, a sua volta, produce un valore dello stimatore

Esempio:

Si esamina un campione casuale di 10 imprese e si rileva X il numero di dipendenti regolari.

Il valore della X è casuale perché non è certa quale azienda finirà nel campione

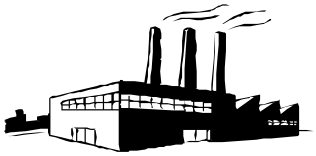
Osservazioni campionarie →

5	0
3	2
1	4
3	2
2	3

Calcoliamo alcuni stimatori

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{5+0+3+2+1+4+3+2+2+3}{10} = \frac{25}{10} = 2.5$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - 10)^2}{10-1} = \frac{2.5^2+2.5^2+0.5^2+0.5^2+1.5^2+ +1.5^2+0.5^2+0.5^2+0.5^2+0.5^2}{9} = 2.06$$



La distribuzione degli stimatori

Lo stimatore è una variabile casuale connessa all'esperimento: estrazione casuale di un campione.

Conoscere la sua distribuzione ci serve per descrivere l'andamento dei risultati che si possono osservare replicando il piano di campionamento.

Dobbiamo ricordare che…

- **Stimare qualcosa significa dare un valore a quel qualcosa**
- **La stima ottenuta da un campione può essere diversa da quella ottenuta con un altro campione**
- **La stima tende differire dal parametro da stimare, ma se conosciamo la distribuzione campionaria dello stimatore possiamo quantificare probabilisticamente l'errore**

Esempio

L'estrazione del campione produce la n-tupla $(x_1, x_2, \ldots, x_n)$ i cui elementi sono le osservazioni campionarie

Ogni n-tupla, a sua volta, produce un valore dello stimatore

Si esamina un campione casuale di 10 imprese e si rileva X il numero di dipendenti regolari.

Il valore della X è casuale perché non è certa quale azienda finirà nel campione

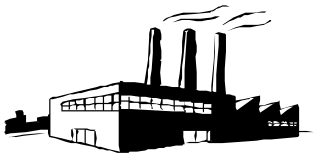
Osservazioni campionarie →

5	0
3	2
1	4
3	2
2	3

Calcoliamo alcuni stimatori

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{5+0+3+2+1+4+3+2+2+3}{10} = \frac{25}{10} = 2.5$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - 10)^2}{10-1} = \frac{2.5^2+2.5^2+0.5^2+0.5^2+1.5^2+ +1.5^2+0.5^2+0.5^2+0.5^2+0.5^2}{9} = 2.06$$



La distribuzione degli stimatori/2

Per costruire la distribuzione di uno stimatore si debbono considerare tutti i possibili campioni di ampiezza prefissata "n"

Esempio:

Una popolazione è composta dai valori {1, 3, 5}. Si estrae, con reimmissione, un campione di ampiezza n=2. Indichiamo con X_1 il valore osservato nella 1ª estrazione e con X_2 quello osservato nella 2ª.

Costruiamo la distribuzione dello stimatore $T = \frac{X_1^2 + X_2^2}{2}$

Basta elencare tutti i campioni di ampiezza n=2 ottenibili dalla popolazione e vedere che valori assume "T".

X_1	1	1	1	3	3	3	5	5	5
X_2	1	3	5	1	3	5	1	3	5
T	1	5	13	5	9	17	13	17	25

I valori sono poi accorpati per assegnare correttamente le probabilità:

t	1	5	9	13	17	25	
P(T = t)	1/9	2/9	1/9	2/9	1/9	1/9	1

Valore atteso

Degli stimatori ci interessa:

$$E(T) = \sum_{i=1}^{k} T_i \Pr(T = T_i)$$

il valore atteso è il valore della media aritmetica di "T" calcolata su tutti i possibili campioni di ampiezza "n".

Se la media E(T)=θ cioè il parametro da stimare, allora T è uno stimatore NON DISTORTO

Lo scarto E(T) - θ è detto Bias *(pron. baias)*

La varianza

Un altrfo aspetto essenziale è

$$\sigma^2(T) = \sum_{i=1}^{k} \left[T_i - E(T)\right]^2 \Pr(T = T_i)$$

$$= \sum_{i=1}^{k} \frac{T_i^2}{k} - \left[E(T)\right]^2$$

La varianza dello stimatore dà una indicazione delle fluttuazioni campionarie cioè quantifica le differenze tra i suoi valori potenziali nei diversi campioni.

Esempio

Riprendiamo la distribuzione della statistica $T = \frac{X_1^2 + X_2^2}{2}$ **nella popolazione {1,3,5} e per campioni di ampiezza n=2**

t	1	5	9	13	17	25	
P(T = t)	1/9	2/9	1/9	2/9	1/9	1/9	1

Calcoliamo il valore atteso e la varianza di "T"

$$E(T) = 1*\frac{1}{9} + 5*\frac{2}{9} + 9*\frac{1}{9} + 13*\frac{2}{9} + 17*\frac{1}{9} + 25*\frac{1}{9} =$$

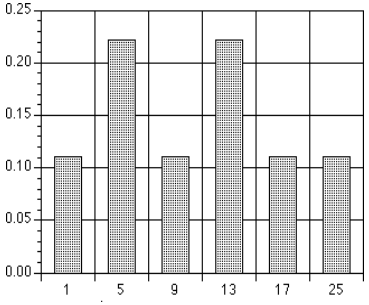
$$= \frac{68}{9} = 7.556$$

$$\sigma^2(T) = \frac{1}{9} + \frac{50}{9} + \frac{81}{9} + \frac{338}{9} + \frac{289}{9} + \frac{625}{9} - \left(\frac{68}{9}\right)^2$$

$$= \frac{1384}{9} - \left(\frac{68}{9}\right)^2 = 97.691$$

In media T assume valore 7.6 con una varianza vicino a 98.

Distribuzione dello stimatore



Esercizio

il tempo che Caterina Ruffolo impiega da casa all'università è una v.c. X_1 **con valore atteso 25 min. e deviazione standard 5 min.**

il ritorno, è una v.c. X_2 **, indipendente dalla prima, con valore atteso di 20 min. e deviazione standard di 4 min.**

Che comportamento ha la differenza $X_1 - X_2$ ***?***

Anche questa è una v.c. con valore atteso: $E(X_1 - X_2) = E(X_1) - E(X_2) = 25 - 20 = 4$

e con varianza: $\sigma^2(X_1 - X_2) = \sigma^2(X_1) + \sigma^2(X_2) = 25 + 16 = 41$

Che comportamento ha il tempo medio $\frac{X_1 + X_2}{2}$ ***?***

$$E\left(\frac{X_1 + X_2}{2}\right) = \left(\frac{1}{2}\right)E(X_1) + \left(\frac{1}{2}\right)E(X_2) = \left(\frac{25}{2}\right) + \left(\frac{20}{2}\right) = 22.5$$

$$\sigma^2\left(\frac{X_1 + X_2}{2}\right) = \left(\frac{1}{2}\right)^2 \sigma^2(X_1) + \left(\frac{1}{2}\right)^2 \sigma^2(X_2) = \frac{25}{4} + \frac{16}{4} = 10.25$$

Scelta tra stimatori diversi

Per ogni caratteristica della popolazione può proporsi più di uno stimatore. Lo stesso succede per i parametri di un modello.

Poiché lo stimatore indica come utilizzare le informazioni campionarie per stimare le caratteristiche o parametri noi sceglieremo quello che le impiega al meglio

LE INFORMAZIONI SI UTILIZZANO AL MEGLIO SE ...

Si utilizzano TUTTE

(cioè lo stimatore non deve disperdere alcuna informazione di quelle incluse nel campione).

Si utilizzano in modo EFFICIENTE

(cioè non deve essere possibile avere migliore conoscenza di ciò che è incognito cambiando stimatore)

Per stabilire tali condizioni dovremmo conoscere la distribuzione dello stimatore, ma può bastare la conoscenza del suo valore atteso e della varianza.

Sufficienza degli stimatori

Poiché lo stimatore è definito in R^n ed ha valori nello spazio parametrico (di solito l'asse reale) c'è il rischio che il processo di sintesi disperda qualche informazione.

E' <u>sufficiente</u> uno stimatore che non disperda alcuna informazione utile per la conoscenza del parametro da stimare.

Quando è noto il modello della popolazione è semplice stabilire la condizione di sufficienza:

La funzione di verosimiglianza deve dipendere dal campione solo attraverso lo stimatore

$$L(\theta) = g\left[T(X_1, X_2, \ldots, X_n); \theta\right]$$

Se si dovesse trasmettere come messaggio basterebbe comunicate "T" e non le singole X_i

Da notare che anche il "campione" è sufficiente di per sé, ma non svolge alcuna funzione di sintesi

Esempio

Supponiamo che il modello adatto ad una certa situazione sia la v.c. di Bernoulli

X assume valore zero oppure uno con $P(X = x) = p^x(1-p)^{1-x}$ in cui però si ignora la probabilità di successo. Esiste una statistica sufficiente per "p" ?

Si estrae un campione casuale . La distribuzione del campione è

$$= \prod_{i=1}^{n} p(X_i;\theta) = \prod_{i=1}^{n} p^{X_i} * (1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i} (1-p)^{n-\sum_{i=1}^{n} X_i} = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^{n} X_i} (1-p)^n$$

Ponendo

$$T(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} X_i; \quad g\left[T(x_1, x_2, \ldots, x_n;\theta)\right] = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^{n} X_i} (1-p)^n$$

T

Si vede subito che T è sufficiente per "p"

In generale la somma ed ogni altra funzione che coinvolga TUTTE le variabili casuali del campione è uno stimatore sufficiente.

Esempio

Si supponga che il campione $\{X_1, X_2, \ldots, X_n\}$ sia stato estratto da una v.c. di Poisson

La funzione di verosimiglianza può essere quindi espressa come

$$L(\lambda) = Ln(\lambda) \sum_{i=1}^{n} X_i - n\lambda$$

Poniamo. $T = \sum_{i=1}^{n} X_i$ che, come si vede, è sufficiente per λ.

Altri stimatori sufficienti sono però: $\sum_{i=1}^{n} 2 * X_i; \quad \sum_{i=1}^{n} Ln(X_i); \quad \sum_{i=1}^{n} a^{X_i}$

Invertibilità

Nell'esempio precedente si vede che $T = n\bar{x}$ e quindi sia il totale che la media campionaria sono statistiche sufficienti per il "λ" della Poisson.

Se $\hat{\theta}$ è sufficiente per "θ" e se g(.) è una funzione INVERTIBILE allora anche $g(\hat{\theta})$ è uno stimatore sufficiente per θ.

La sufficienza da sola non basta a selezionare lo stimatore

Da notare che se T è uno stimatore di massima verosimiglianza allora è anche uno stimatore sufficiente.

Gli stimatori ottenuti con il metodo dei momenti sono pure sufficienti purché lasoluzione di

$$\frac{\sum_{i=1}^{n} x_i^r}{n} = h(\theta) \Rightarrow \hat{\theta} = h^{-1}\left(\frac{\sum_{i=1}^{n} x_i^r}{n}\right)$$

non coinvolga altri parametri incogniti e/o singole osservazioni campionarie

il centramento (o non distorsione)

Lo stimatore è una funzione di v.c. e quindi esso stesso una v.c. il centramento fa riferimento al valore atteso di tale distribuzione.

Uno stimatore è <u>centrato</u> se, in media, è uguale al parametro da stimare

$$E[T(X_1, X_2, ..., X_n)] = \theta \quad \text{ovvero} \quad E[T(X_1, X_2, ..., X_n) - \theta] = 0 \quad \forall\ \theta \in \Theta$$

Questo è il BIAS dello stimatore

ESEMPIO

Sia un $\{X_1, X_2, ..., X_n\}$ campione casuale da una Bernoulli. Vediamo se: $H = \frac{\sum_{i=1}^{n} X_i}{n}$ è uno stimatore non distorto della probabilità di successo "p".

$$E(H) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Poichè l'aspettativa di ogni Bernoulli è "p", si ha: $\frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n} np = p$

Centramento - non distorsione/2

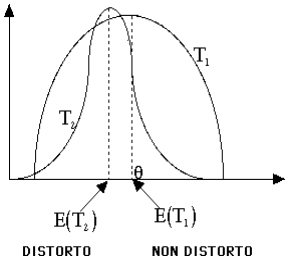
Tale caratteristica è spesso indicata come "correttezza" dello stimatore.

Ciò perché se $\tilde{T} \Rightarrow E(\tilde{T}) = a\theta + b$

allora $\hat{T} = \frac{\tilde{T} - b}{a} \Rightarrow E(\hat{T}) = \frac{1}{a}\left[E(\tilde{T}) - b\right] = \theta$

Quindi lo stimatore $\hat{T}$ è la versione "corretta" dello stimatore T nel senso che sono stale eliminate le tendenze alla deviazione sistematica.

Corretto non significa "giusto" o "onesto"



Lo stimatore è centrato se non ha tendenza sistematica a deviare (sempre per difetto o sempre per eccesso) rispetto al valore vero del parametro o caratteristica da stimare.

Esempio

il tempo di reazione ad un certo stimolo ha distribuzione uniforme tra 0 e θ con θ>0.

$$f(x, \theta) = \frac{1}{\theta} \quad \text{per} \quad 0 \le X \le \theta$$

Si vuole stimare θ in base ad un campione casuale e poiché θ è il tempo di reazione più grande riscontrabile sembra naturale stimarlo con la statistica:

$$x_{max} = \max\{X_1, X_2, ..., X_n\}$$ *E' uno stimatore sufficiente?*

Con n=5 ed i valori campionari sono: $\{4.2, 1.7, 2.4, 3.9, 1.3\} \Rightarrow x_{max} = 4.2$

E' intuitivo che questo stimatore sia distorto: se non lo fosse sarebbero possibili sia stime errate per difetto che per eccesso. Queste però sono escluse per costruzione:

E' infatti possibile dimostrare che: $E(x_{max}) = \frac{n}{n+1}\theta \quad \text{Bias} = E(x_{max}) - \theta = \left(\frac{n}{n+1} - 1\right)\theta = -\frac{\theta}{n+1}$

Tuttavia, X_{max} è <u>asintoticamente non distorto</u>, ovvero il bias tende a zero man mano che l'ampiezza del campione diventa grande

Esempio

Verifichiamo che la varianza campionaria $V = \dfrac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n}$ **sia uno stimatore distorto**

$$E(V) = E\left(\frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n}\right) = E\left(\frac{\sum_{i=1}^{n}X_i^2}{n} - \hat{\mu}^2\right) = E\left(\frac{\sum_{i=1}^{n}X_i^2}{n}\right) - E(\hat{\mu}^2) = \frac{\sum_{i=1}^{n}E(X_i^2)}{n} - -E(\hat{\mu}^2)$$

Poiché $\sigma^2 = E(X_i^2) - \mu^2 \Rightarrow E(X_i^2) = \sigma^2 + \mu^2; \qquad E(\hat{\mu}^2) - \mu^2 = \dfrac{\sigma^2}{n} \Rightarrow E(\hat{\mu}^2) = \dfrac{\sigma^2}{n} + \mu^2$

Si ha $E(V) = \dfrac{\sum_{i=1}^{n}(\sigma^2 + \mu^2)}{n} - \left[\dfrac{\sigma^2}{n} + \mu^2\right] = \sigma^2 + \mu^2 - \dfrac{\sigma^2}{n} - \mu^2 = \left(\dfrac{n-1}{n}\right)\sigma^2$

Che è solo asintoticamente non distorto. Ecco perchè per stimare σ^2 **si usa** $s^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{x})}{n-1} = \left[\dfrac{n}{n-1}\right] * V$ **che è centrato**

Esempio

a) Supponiamo che il campione casuale provenga dalla v.c. geometrica. La media campionaria è uno stimatore corretto?

In questo caso: $E(X_i) = \dfrac{1}{p}; \quad i = 1,2,\ldots,n$ **Quindi** $E(\bar{x}) = E\left(\dfrac{\sum_{i=1}^{n}X_i}{n}\right) = \dfrac{1}{n} * \dfrac{n}{p} = \dfrac{1}{p}$

In generale $E[g(x)] \neq g[E(x)]$ **per cui la media campionaria non è sempre uno stimatore corretto**

b) Supponiamo che il campione casuale provenga dalla v.c. uniforme (θ-1,θ+1). La media campionaria è uno stimatore corretto?

In questo caso si ha $E(X_i) = \dfrac{(\theta-1)+(\theta+1)}{2} = \theta \quad \text{per } i = 1,2,\ldots,n$

e quindi $E(\bar{x}) = \dfrac{n\theta}{n} = \theta$ **In questo caso è corretto**

Variabilità dello stimatore

Ogni somma ponderata delle v.c. del campione è uno stimatore centrato. Quale scegliere?

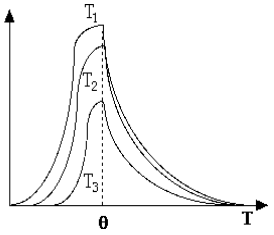
$$\tilde{\mu} = \sum_{i=1}^{n} w_i X_i \Rightarrow E\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i E(X_i) = \sum_{i=1}^{n} w_i \mu = \mu \sum_{i=1}^{n} w_i = \mu$$

Dobbiamo fare entrare in gioco la variabilità (errore standard o scarto tipo) dello stimatore: σ(T)

PRINCIPIO DELLA VARIANZA MINIMA:

Minore è la variabilità intorno a "θ", più attendibili è lo stimatore

T2 è preferito a T1 e T3 è preferito a T2



Secondo il teorema di Cramér -Rao esiste un limite inferiore alla varianza di uno stimatore corretto. Se qualcuno la raggiunge allora ci sarà unico stimatore che ha la varianza pari al limite minimo.

Esempio

Ragioniamo ancora nell'ambito della uniforme U(0,θ) e definiamo vari stimatori non distorti per θ a partire da un campione casuale di ampiezza "n":

Poiché $E(X_i) = \dfrac{\theta}{2} \Rightarrow E(2X_i) = \theta$ **e quindi** $T_1 = 2\bar{x}$ **è uno stimatore corretto**

La varianza di T1 è $\sigma^2(2\bar{x}) = 4\sigma^2(\bar{x}) = 4\sigma^2\left(\dfrac{\sum_{i=1}^{n}X_i}{n}\right) = 4\dfrac{\sum_{i=1}^{n}\frac{(\theta-0)^2}{12}}{n^2} = \dfrac{\theta^2}{3n} \qquad \sigma(T_1) = \dfrac{\theta}{\sqrt{3n}}$

Un altro stimatore corretto $T_2 = \dfrac{n+1}{n}x_{max}$ **con errore standard** $\sigma(T_2) = \dfrac{\theta}{\sqrt{n(n+2)}}$

Se si ha una sola osservazione campionaria si usa T1 e se ne hanno due o più si usa T2

Esempio

Consideriamo i seguenti stimatori della media aritmetica basati su di un campione casuale semplice conreimmissione da una distribuzione qualsiasi

$$\hat{\mu}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3; \quad \hat{\mu}_2 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3; \quad \hat{\mu}_3 = \frac{3}{6}X_1 + \frac{2}{6}X_2 + \frac{1}{6}X_3;$$

Gli stimatori sono tutti sufficienti e non distorti (somma ponderata dei valori).

Qual'è lo stimatore efficiente?

Sarà quello con varianza minima.

$$\sigma_1^2 = \sigma^2 * \left[\frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^2}\right] = \sigma^2 * \left(\frac{1}{3}\right) = \sigma^2 * 0.3333$$

il primo stimatore è preferibile al secondo che, a sua volta, è preferibile al terzo

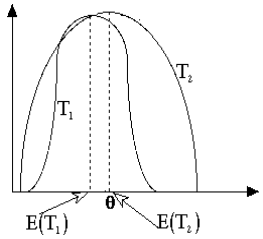
$$\sigma_2^2 = \sigma^2 * \left[\frac{1}{4^2} + \frac{1}{2^2} + \frac{1}{4^2}\right] = \sigma^2 * \left(\frac{3}{8}\right) = \sigma^2 * 0.3750$$

$$\sigma_3^2 = \sigma^2 * \left[\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{6^2}\right] = \sigma^2 * \left(\frac{7}{18}\right) = \sigma^2 * 0.3889$$

L'errore quadratico medio

E' chiaro che tra stimatori non distorti si sceglie quello con varianza minima, ma è possibile compensare la maggior distorsione con una minor dispersione?

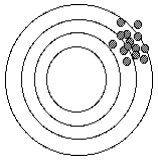
Lo stimatore T1 è distorto, ma molto meno disperso di T2 che è sì non distorto, ma poco preciso



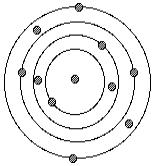
Distorsione e Dispersione si fondono nell'Errore quadratico medio. Tale misura ha due componenti additive:

$$E(T-\theta)^2 = E\{[T - E(T)] + [E(T) - \theta]\}^2 =$$
$$= E\{[T - E(T)]^2\} + E\{[E(T) - \theta]^2\} + 2[E(T) - \theta]E\{[T - E(T)]\}$$
$$= \sigma^2(T) + [E(T) - \theta]^2$$
$$= \textit{Varianza dello stimatore} + \textit{Bias}$$

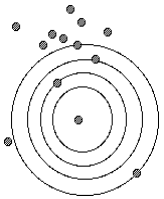
illustrazione



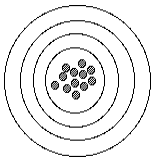
Bias elevato, moderata variabilità



Bias moderato, elevata variabilità



Bias elevato, elevata variabilità



Bias moderato, moderata variabilità

Errore quadratico medio/2

L'E.Q.M. è un criterio importante per la scelta tra stimatori diversi

In generale, T_1 è preferibile a T_2 se $EQM(T_1) \le EQM(T_2) \Rightarrow \frac{EQM(T_2)}{EQM(T_1)} \ge 1$ ← **indice di efficienza**

ESEMPIO: dato il campione casuale $\{X_1, X_2, \ldots, X_n\}$ **da una v.c. con F(X;μ;σ²) si considerano due stimatori**

$$T_1 = X_1 \Rightarrow \quad E(T_1) = \mu \quad \sigma^2(T_1) = \sigma^2$$
$$T_2 = \hat{\mu} \qquad E(T_2) = \mu \quad \sigma^2(T_2) = \frac{\sigma^2}{n}$$

L'indice di efficienza diventa:

$$\frac{EQM(T_2)}{EQM(T_1)} = \frac{\sigma^2/n}{\sigma^2} = \frac{1}{n} \le 1 \quad per \quad n > 1$$

Quindi, T2 è preferibile a T1 a meno che n=1. In tal caso hanno uguale efficienza

Gli stimatori di M.V. tendono ad essere più efficienti di quelli dei momenti

Esempio

Sia $\{X_1, X_2, \ldots, X_n\}$ **un campione casuale dalla Poisson:** $P(X;\theta) = \frac{\theta^x e^{-\theta}}{x!} \quad x = 0,1,2,\ldots,$

In tale modello si ha $E(X_i) = \sigma^2(X_i) = \theta$ **per cui sia la media che la varianza campionaria possono servire a stimare θ.**

Nella Poisson abbiamo inoltre: $\bar{\mu}_4 = 3\theta^2 + \theta$

Stimatore	Valolere atteso	Varianza
Media	θ	$\frac{\theta}{n}$
Varianza	θ	$\frac{\theta}{n}\left[\frac{2n}{(n-1)}\theta + 1\right]$

$$\frac{Var(s^2)}{Var(\bar{x})} = \frac{2n}{(n-1)}\theta + 1 > 1$$

Ne consegue che la media campionaria, nell'ambito del modello di Poisson, è uno stimatore più efficiente della varianza campionaria.

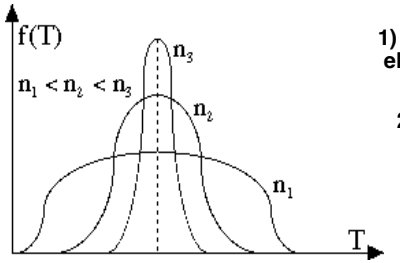
La consistenza

Limitarsi agli stimatori non distorti non basta: per un dato modello potrebbe non esistere lo stimatore non distorto dei parametri.

Se ne esiste più di uno si sceglie quello a varianza minima. Ciò suggerisce una nuova proprietà

Uno stimatore valido dovrebbe avere una funzione di distribuzione che se "n" aumenta si riduce ad un solo punto: il parametro da stimare.

Questo perchè, maggiori sono le informazioni, più precisa risulta la nostra stima e ciò deve essere recepito dallo stimatore



1) La probabilità di scarti dalla media molto elevati deve tendere a zero

2) La varianza dello stimatore deve tendere a zero

Se ciò si verifica lo stimatore è detto CONSISTENTE o COERENTE

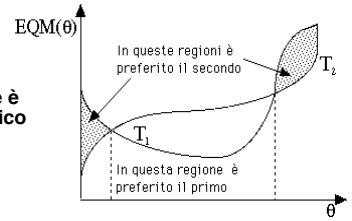
Errore quadratico medio/3

L'EQM non dà sempre indicazioni univoche. Per certi valori è preferibile uno stimatore, in altre regioni l'altro.

Ad esempio T=5 è uno stimatore migliore o non peggiore di nessun altro se θ=5

L'EQM dipende da θ. E poiché θ è incognito non si potrà mai decidere con certezza.

Nei rari casi in cui si riesce a dire che uno stimatore è preferibile a tutti gli altri su tutto lo spazio parametrico è detto OTTIMALE.



Uno di questi è il raggiungimento del limite di Cramér-Rao nel qual caso l'efficienza è massima.

Si parla talvolta di efficienza relativa di uno stimatore come rapporto tra il suo errore standard e il limite Cramér-Rao.

Consistenza in probabilità

Lo stimatore $T_n(X_1,\ldots,X_n)$ **del parametro "θ" è CONSISTENTE IN PROBABILITA' se**

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1 \quad \forall\theta\in\Theta \text{ e } \forall\varepsilon > 0$$

Questo richiede in pratica che la varianza dello stimatore sia una funzione inversa dell'ampiezza campionaria. Infatti, se nella Chebychev

$$P[|T_n - \theta| < k\sigma(T_n)] \geq 1 - \frac{1}{k^2}$$

si pone $\sigma(T_n) = \frac{f(\sigma)}{\sqrt{n}};\ \varepsilon = k * \frac{f(\sigma)}{\sqrt{n}} \Rightarrow k = \sqrt{n}\frac{\varepsilon}{f(\sigma)}$ **si avrà**

$$P[|T_n - \theta| < \varepsilon] \geq 1 - \frac{1}{n} * \left(\frac{f(\sigma)}{\varepsilon}\right)^2$$ **che appunto tende ad uno all'aumentare di "n"**

il centramento o on distorsione non è direttamente presente

Esempio

Sia X la durata, in ore, di un fusibile ed X una v.c. con funzione di densità

$$f(x,\beta) = \frac{e^{\frac{-x}{\beta}}}{\beta} \quad x>0 \quad \text{(esponenziale con } \lambda = 1/\beta)$$

consideriamo un campione casuale di ampiezza " $\{X_1, X_2, \ldots, X_n\}$ $\quad E(X_i) = \beta$

Lo stimatore naturale di β è la media aritmetica campionaria che, come si può dimostrare è corretto. E' anche consistente?

La disuguaglianza di Chebyschev ci dice: $P(|\bar{x} - \beta| \geq \varepsilon) \leq \frac{\sigma^2(\bar{x})}{\varepsilon^2} = \frac{\sigma^2(x)}{n\varepsilon^2} = \frac{\beta}{n\varepsilon^2}$ **dato che** $\sigma^2(x) = \beta$

Quindi, all'aumentare di "n", la probabilità di scarti -anche molto piccoli- tra stimatore e parametro incognito tende a zero.

Gli stimatori la cui varianza tenda a zero all'aumentare di "n" sono degli stimatori consistenti

Esercizio

Una organizzazione di consumatori sottopone a controllo la durata in ore di scrittura di una penna a biro di tipo economico.

A questo fine acquista n=10 penne e le inserisce in un apposito dispositivo che lemantiene in scrittura fino all'esaurimento.

Sia $\{X_1, X_2, \ldots, X_{10}\}$ **un campione casuale di ampiezza n=10 da una distribuzione che rappresenti il consumo delle penne.**

I valori campionari effettivi sono stati

$x_1 = 26.3$	$x_2 = 35.1$	$x_3 = 23.0$	$x_4 = 28.4$	$x_5 = 31.6$
$x_6 = 30.9$	$x_7 = 25.2$	$x_8 = 28.0$	$x_9 = 27.3$	$x_{10} = 29.3$

Calcoliamo media e varianza campionaria

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 28.50; \qquad s^2 = \frac{\sum_{i=1}^{10} (x_i - 28.5)^2}{9} = 11.90$$

La consistenza in media quadratica

Lo stimatore $T_n(X_1, \ldots, X_n)$ **di "θ" è CONSISTENTE IN MEDIA QUADRATICA se**

$$\lim_{n\to\infty} E\left[(T_n - \theta)^2\right] = 0$$

Questo implica che l'EQM tende a zero man mano che aumenta l'ampiezza campionaria

$$EQM(T_n) = Var(T_n) + d_n^2(\theta)$$

Poichè la consistenza si otterrà se n-=> ∞ tendono a zero sia la varianza che la distorsione per cui lo stimatore consistente in MQ è anche ASINTOTICAMENTE NON DISTORTO

SE, nella disuguaglianza di Chebyschev, si pone $f(\theta) = \sqrt{E\left[(X-\theta)^2\right]}$

Si ottiene

$$P(|X - \theta| \geq \varepsilon) \leq \frac{E\left[(X-\theta)^2\right]}{\varepsilon^2}$$

quindi, la consistenza in MQ implica quella in probabilità cioè la prima è una condizione più restrittiva per la scelta tra stimatori diversi.

Esempio

Lo stimatore x_{max} **del parametro che della uniforme U(0,θ) è consistente in media quadratica**

In generale, se T è uno stimatore di θ, abbiamo l'identità

$$\text{Errore quadratico medio} = \text{Varianza} + \text{bias} \Rightarrow E\left[(T-\theta)^2\right] = \sigma^2(T) + d^2(\theta)$$

Come visto nei risultati precedenti, per lo stimatore in questione, abbiamo:

$$\sigma^2(x_{max}) = \frac{n}{(n+1)^2}\theta^2; \quad d^2(\theta) = \frac{1}{(n+1)^2}\theta^2 \Rightarrow E\left[(x_{max} - \theta)^2\right] = \frac{n}{(n+1)^2}\theta^2 + \frac{1}{(n+1)^2}\theta^2 = \frac{\theta^2}{(n+1)}$$

Ne consegue che

$$\lim_{n\to\infty} E\left[(x_{max} - \theta)^2\right] = \lim_{n\to\infty} \frac{\theta^2}{(n+1)} = 0$$

che attesta la consistenza in media quadratica come del resto c'era da aspettarsi, visto che, all'aumentare di "n" diminuisce sia la varianza che il bias.

Consistenza della media campionaria

Lo stimatore media campionaria è consistente in probabilità. Infatti

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

nell'ambito di un campione casuale (fatto da v.c. i.i.d.) ha varianza

$$\text{caso particolare con } w_i = \frac{1}{n} \Rightarrow \text{Var}(\hat{\mu}) = \sigma^2 \left[\sum_{i=1}^{n} \left(\frac{1}{n} \right)^2 \right] = \sigma^2 n * \frac{1}{n^2} = \frac{\sigma^2}{n}$$

Secondo la disuguaglianza di Tchebycheff (in cui ora f(σ)=σ) abbiamo

$$P\left(|\hat{\mu} - \mu| < \varepsilon \right) \geq 1 - \frac{1}{n} \left(\frac{\sigma}{\varepsilon} \right)^2$$

Esisterà un "n" che "con probabilità uno" afferma che lo stimatore ha una distanza prestabilita dal parametro da stimare quantunque piccola essa sia

Consistenza della varianza campionaria

Per stimare la varianza di una popolazione è stata proposta la statistica

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

che si è già dimostrato privo di errore sistematico: $E(s^2) = \sigma^2$

Si dimostra che $\sigma^2(s^2) = \frac{E(X-\mu)^4}{n} + \frac{3-n}{n-1} * \frac{\sigma^4}{n}$

All'aumentare di "n" si ha

$$\frac{E(X-\mu)^4}{n} \to 0 \quad se \quad E(X-\mu)^4 < \infty$$

$$\frac{3-n}{n-1} * \frac{\sigma^4}{n} \to 0 \quad se \quad \sigma < \infty$$

La consistenza si ha se il momento 4° intorno alla media non diverge.

Ciò implica che siano anche finiti tutti i momenti inferiori al 4°

Stima puntuale delle proporzioni

Spesso si è interessati a sapere quale porzione "π" di una popolazione possiede una certa caratteristica (unità speciali)

Se si estrae un campione casuale con riposizione si è di fronte ad un modello di tipo binomiale in cui il "successo" è la estrazione di una unità speciale con "π" come probabilità di successo.

Ne consegue che stimare "π" significa stimare il parametro della binomiale con un numero di prove pari ad "n" cioè all'ampiezza del campione.

Lo stimatore di "π" è naturalmente:

$$H = \frac{\text{numero di unità speciali nel campione}}{\text{ampiezza del campione}}$$

Che caratteristiche ha tale stimatore?

Stima delle propozioni/2

Come è noto "H" ha pure una distribuzione binomiale con

$$E(H) = \pi; \qquad \sigma(H) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Quindi "H" è centrato e consistente per la frazione di unità speciali nella popolazione

Da notare che se il campionamento è fatto senza reimmissione da una popolazione finita e piccola interviene il modello ipergeometrico con

$$E(H) = \pi; \qquad \sigma(H) = \left[\sqrt{\frac{\pi(1-\pi)}{n}} \right] * \left[\sqrt{\frac{N-n}{N-1}} \right]$$

La proporzione campionaria è ancora una stima non distorta. E' ancora consistente?

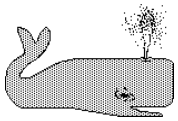
Anche in questo caso la performance dello stimatore dipende dal parametro da stimare

Distribuzione delle statistiche/3

Poiché la disamina completa dell'universo dei campioni non è in genere possibile, la distribuzione delle statistiche è determinata secondo due impostazioni

APPROCCIO BASATO SU MODELLI (piccoli campioni)

il modello della popolazione è noto nella forma, ma non nei parametri. Si usa il campione per ricostruire la distribuzione delle statistiche per poi fare inferenza sui parametri.
Il metodo MLE e dei momenti rientra in questa categoria



APPROCCIO DEI GRANDI CAMPIONI

Sul modello si fanno ipotesi minimali e grazie all'elevato numero di informazioni si determina la distribuzione delle statistiche per fare inferenza sulle caratteristiche più rilevanti (media e varianza).

Algoritmo di Tchebycheff

Studiamo le caratteristiche per la statistica "media campionaria" nel caso di un campione casuale, cioè di "n" v.c. I.I.D con valore atteso μ e varianza σ^2

$$\bar{x}=\frac{\sum_{i=1}^{n}X_i}{n};\quad E(\bar{x})=\frac{\sum_{i=1}^{n}E(X_i)}{n}=\frac{\sum_{i=1}^{n}\mu}{n}=\frac{n\mu}{n}=\mu;\quad \sigma^2(\bar{x})=\frac{\sum_{i=1}^{n}\sigma^2(X_i)}{n^2}=\frac{\sum_{i=1}^{n}\sigma^2}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

il valore atteso della media campionaria è pari al valore atteso della popolazione

La varianza è σ^2/. Applichiamo la disuguaglianza di Tchebycheff

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

Algoritmo di Tchebycheff/2

Se poniamo $k=\varepsilon\frac{\sqrt{n}}{\sigma}$

la disuguaglianza diventa: $P(|\bar{x}-\mu|\geq\varepsilon)\leq\frac{\sigma^2}{\varepsilon^2}*\frac{1}{n}$

e dunque

$$0\leq\lim_{n\to\infty}P(|\bar{x}_n-\mu|\geq\varepsilon)\leq\lim_{n\to\infty}\left(\frac{\sigma^2}{\varepsilon^2}\right)*\frac{1}{n}=0$$

All'aumentare dell'ampiezza del campione la distribuzione della media campionaria tende sempre più a concentrarsi sulla media della popolazione

Algoritmo di Tchebycheff/3

Riporendiamo la FREQUENZA di successi su "n" prove indipendenti con probabilità di successo costante "p".

$$E(H)=p;\qquad \sigma^2(H)=\frac{p^*(1-p)}{n}$$

Posto h_n pari al numero di successi di un campione di "n" prove, si ottiene

$$P\left(|h_n-p|\geq k\frac{\sqrt{p^*(1-p)}}{\sqrt{n}}\right)\leq\frac{1}{k^2}\Rightarrow P(|h_n-p|\geq\varepsilon)\leq\frac{p^2(1-p)^2}{\varepsilon^2}*\frac{1}{n}$$

$$\lim_{n\to\infty}P(|h_n-p|\geq\varepsilon)=0$$

che conferma il postulato empirico del caso sul fatto che la frequenza di successi su di un numero elevato di prove fornisce una buona stima della probabilità .

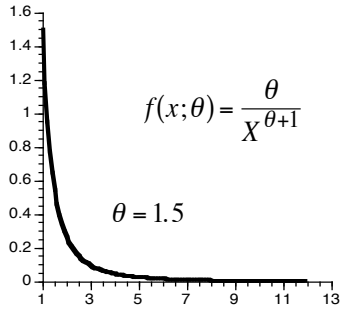
Inapplicabilità della Tchebycheff

Richiede che $E(X^2)< \infty$ e non è sempre applicabile

Nel modello di Pareto abbiamo

$$F(x) = 1 - x^{-\theta}, x > 1;\ E(X) = \frac{\theta}{\theta - 1}; E\left(X^2\right) = \frac{\theta}{\theta - 2}$$

il momento secondo non esiste (la formula lo dà negativo per θ=1.5).

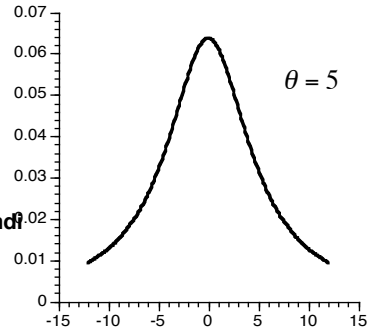


Inapplicabilità della Tchebycheff/2

Non abbiamo garanzie che la casualità del nostro esperimento sia modellabile con una v.c. che abbia $E(|X|)< \infty$.

Nel modello di Cauchy non è applicabile

$$f(x;\theta) = \frac{1}{\pi} * \frac{\theta}{\theta^2 + x^2};\ -\infty < x < \infty$$



il momento primo non esiste: E(|X|)-->∞ (e quindi nemmeno il secondo).

In questo modello il calcolo della media non ha alcun effetto ed i campioni casuali estratti da tale v.c. hanno una media che divegbe all'aumentare di n.

il fallimento è dovuto alla presenza di valori grandi con una probabilità elevata anche per posizioni molto estreme (elevato spessore delle code).

I grandi numeri

Di solito si ignora la variabile casuale che può descrivere in modo soddisfacente un dato aspetto della popolazione.

Di conseguenza non è possibile costruire la distribuzione di uno stimatore.

Inoltre, uno stesso stimatore ha una distribuzione campionaria diversa in dipendenza del tipo di variabile casuale che descrive la popolazione.

C'è una via d'uscita?

Se la distribuzione non è nota, ma il campione casuale è abbastanza numeroso e le estrazioni sono indipendenti (o virtualmente tali) è possibile approssimare la distribuzione delle statistiche con il MODELLO NORMALE

Teorema del limite centrale

Supponiamo che :

$$\{X_1, X_2, \ldots, X_n\}$$

siano estrazioni casuali da una popolazione descritta da una v.c. con varianza finita

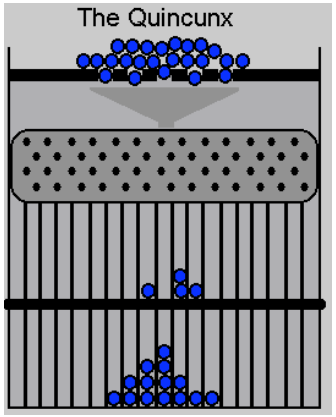
Allora, all'aumentare di "n", il poligono delle frequenze della statistica espressa in unità standard

$$\frac{L_n - E(L_n)}{\sigma(L_n)} \qquad \text{dove} \quad L_n = \sum_{i=1}^{n} w_i X_{(i)}$$

1) Tende ad essere ben approssimato dalla curva Normale standardizzata

2) Il bias e la varianza tendono a zero

Esempio di T.L:C. con il Quincux



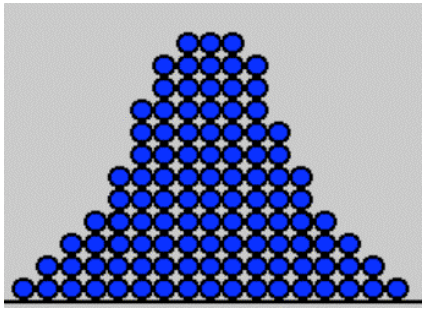
1) Le biglie entrano nell'imbuto dai vari fori

2) Le biglie escono dall'imbuto una alla volta

3) Le biglie rimbalzano a caso tra i vari pioli

4) Ogni biglia imbocca unasola scanalatura

Risultato finale



Esempio

il T.L.C. afferma, in sostanza che, per campioni casuali di ampiezza abbastanza grande, la media campionaria ha distribuzione normale quale che sia la "F" di base.

Qual'è la probabilità che un fenomeno soggetto a variazioni casuali esprima una media campionaria compresa in $\pm\sigma/\sqrt{n}$ dalla media della popolazione di partenza?

$$P\left(|\bar{x}-\mu| < \sigma/\sqrt{n}\right) = P\left(\mu-\sigma/\sqrt{n} < \bar{x} < \mu+\sigma/\sqrt{n}\right)$$

sfruttando il T.L.C. si ha:

$$P\left(\frac{\mu-\sigma/\sqrt{n}-\mu}{\sigma/\sqrt{n}} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{\mu+\sigma/\sqrt{n}-\mu}{\sigma/\sqrt{n}}\right) = P(-1 \le Z \le 1) = 2\phi(1)-1 = 0.6826$$

In questo caso la Chebyshev darebbe come limite inferiore lo zero.

Teorema del limite centrale/2

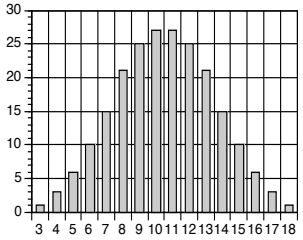
La tendenza alla normale non richiede che le v.c. siano identicamente distribuite r che siano continue.

E' necessario che nessuna delle varianze delle variabili componenti predomini su altre (sorte benigna)

Somma del lancio di "n" dadi

$$S = X_1 + X_2 + \ldots + X_n$$

$$E(S) = n*3.5;\ \sigma^2(S) = n\frac{35}{12};\ Z = \frac{S - n*3.5}{\sqrt{n}*2.9167}$$



oscillazione di un titolo di borsa (sorte selvaggia)

Alcune di queste possono essere catastrofiche o superpositive e non c'è convergenza alla normale perché la loro varianza è infinita

Esercizio

La produzione di una linea di biscotti utilizza macchinari tali che il contenuto in grammi delle confezioni sia una v.c. X con μ=450 e σ=30.

Calcolare la probabilità che il contenuto medio di una scatola di 25 confezioni sia almeno di 460g.

il peso medio è una v.c. data dalla media dei singoli pesi $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

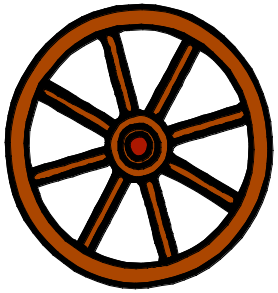
Se la produzione è realizzata in modo da non generare dipendenza nelle varie confezioni sembrano ricorrere le condizioni del T.L.C. e quindi

$$P(\bar{x} \ge 460) = P\left(\frac{\bar{x}-450}{\frac{30}{5}} \ge \frac{460-450}{\frac{30}{5}}\right) = 1 - P(Z \le 1.67) = 0.0475$$

Questa asserzione di probabilità è valida qualunque sia la distribuzione di partenza (purché siano valide le condizioni del T.L.C.)

Teorema del limite centrale/3

Il teorema del limite centrale è un risultato fondamentale della scienza



Con esso è possibile stabilire la distribuzione di vari stimatori senza conoscere quale sia il modello che descrive la casualità dell'esperimento

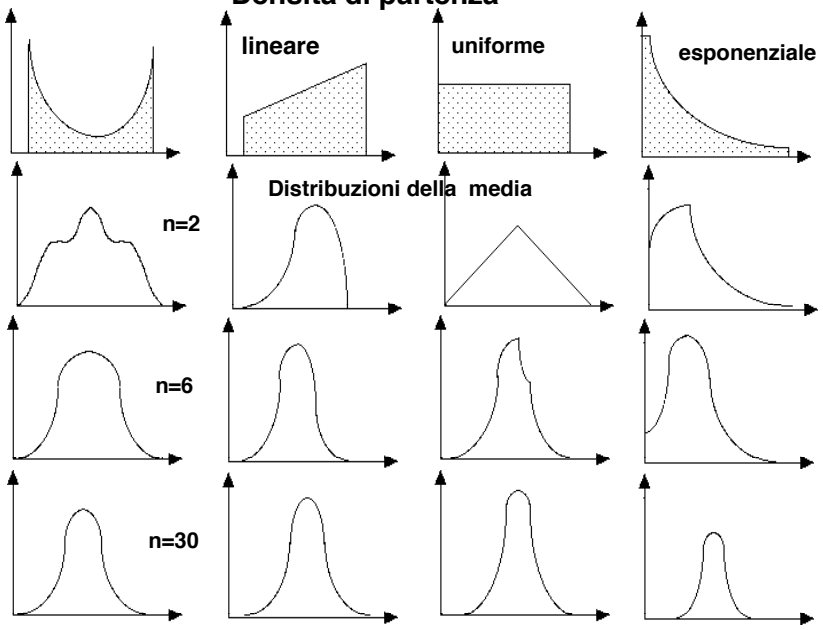
Quando può essere applicato tale risultato?

1) Le estrazioni campionarie debbono essere indipendenti.

2) L'ampiezza del campione deve essere grande.

3) L'aspetto considerato deve essere il risultato di molte concause

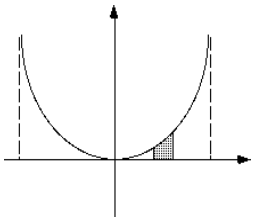
4) Non ci deve essere una causa predominate rispetto alle altre



Esempio

Sia $\bar{X}$ la media campionaria di un campione casuale di ampiezza n=15 dalla v.c. con funzione di densità:

$$f(x) = \frac{3x^2}{2} \quad \text{per} \quad -1 < x < 1$$



Si verifica subito che: $E(x) = 0; \quad \sigma^2(x) = 0.6$

Per calcolare la probabilità che la media sia compresa in un certo intervallo usiamo il T.L.C.

$$P(0.03 \le \bar{x} \le 0.15) = P\left(\frac{0.03-0}{\sqrt{\frac{0.6}{15}}} \le \frac{\bar{x}-0}{\sqrt{\frac{0.6}{15}}} \le \frac{0.15-0}{\sqrt{\frac{0.6}{15}}}\right) = P(0.15 \le Z \le 0.75) = 0.2138$$

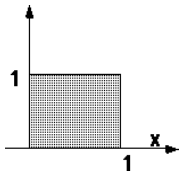
Anche ignorando la reale distribuzione della media campionaria possiamo fare delleaffermazioni. Certo, approssimate, ma almeno ragionevoli.

Esercizio

Sia $\{X_1, \ldots, X_{20}\}$ un campione casuale di ampiezza n=20 dalla uniforme U(0,1)

In questo caso abbiamo:

$$E(X_i) = \frac{0+1}{2} = 0.5; \quad \sigma^2(X_i) = \frac{(1-0)^2}{12} = 0.0833$$



Indichiamo con Q la statistica campionaria: $Q = \sum_{i=1}^{20} X_i$

Calcoliamo la probabilità che Q sia inferiore a 9.1. A questo fine recuperiamo il legame tra media campionaria e totale campionario

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{n(\bar{x}-\mu)}{n\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{n\bar{x}-n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

Quindi: $P(Q \le 9.1) = P\left(\frac{Q-10}{0.2887\sqrt{20}} \le \frac{9.1-10}{0.2887\sqrt{20}}\right) = P(Z \le -0.70) = 0.2423$

Applicazione alla frazione campionaria

La distribuzione della frazione campionaria di successi "H" è Binomiale.

Poichè la Binomiale è la somma di "n" bernouliiane indipendenti e con la stessa probabilità di successoricorrono le condizioni del T.L.C.

Pertanto, la distribuzione di "H", all'aumentare di "n", diventa:

$$Z = \frac{H - \pi}{\sqrt{\frac{\pi * (1 - \pi)}{n}}} \sim N(0,1)$$

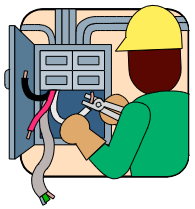
Siccome "n" è grande non c'è bisogno della correzione per la continuità

(usata se si approssima una discreta con una continua)

Esempio

Una partita di circuiti elettrici include il 30% di pezzi difettosi. Si estrae un campione casuale di 500 pezzi.

Qual'è la probabilità che la frazione di pezzi difettosi sia inferiore a 0.31?



$$P(H \le 0.31) = P\left(Z \le \frac{0.31 - 0.3}{\sqrt{\frac{0.3 * 0.7}{500}}}\right) = P\left(Z \le \frac{0.01}{0.0205}\right) = P(Z \le 0.49) = 0.6879$$

Esercizio

In un campione casuale di n=200 studenti si sono trovate 36 (H=0.18) d'accordo sull'università "a distanza " come equivalente a quella da loro frequentata.

Determinare la probabilità che, replicando il campione, la proporzione campionaria rimanga compresa tra 12.5 e 24.5

$$0.0272 = \sqrt{\frac{0.18(1 - 0.18)}{200}}$$

$$P(12.5 \le H \le 24.5) = P\left(\frac{0.125 - 0.18}{0.0272} \le \frac{H - 0.18}{0.0272} \le \frac{0.245 - 0.18}{0.0272}\right) = P(-2.0221 \le Z \le 2.3897)$$
$$= 0.99157 - 0.02158 \cong 0.97$$

il 97% dI TUTTI I POSSIBILI campioni costruiti ipotizzando che sia π=0.18 avrà una proporzione compresa tra il 12.5% ed il 24.5%,

Distribuzione della varianza campionaria

L'asimmetria della distribuzione del χ^2 non è tale da pregiudicare l'approssimazione con la normale.

$$c^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{\sigma^2} \Rightarrow \frac{c^2 - n}{\sqrt{2n}} \to N(0,1)$$

ESEMPIO

$$P(s^2 \le 1.97015) = P\left[\frac{(11-1)}{5}s^2 \le 3.9403\right] = P(c^2 \le 3.9403)$$
$$\cong P\left(\frac{c^2 - 11}{\sqrt{22}} \le -1.51\right) = 1 - \phi(1.51) = 6.55\%$$

Non è eccellente, ma per molti scopi può bastare.