Quaderni di Statistica
Vol.2, 2000

Analisi statistica di un modello per le preferenze nel caso di tre alternative

Domenico Piccolo
Dipartimento di Scienze Statistiche, Universita' di Napoli Federico II
Centro di Specializzazione e Ricerche, Portici
E-mail: dopiccol@unina.it

Summary: We discuss a statistical model for preferences. The situation we examine is quite common in real applications: the subject is asked to give a score to a certain item on a three point scale or to rank a group of three objects. The paper discusses the probability structure underlying this kind of experiment. The proposed stochastic model is well parameterized by a single preference coefficient which has an interesting statistical interpretation. In particular, two estimators are derived for the preference parameter and their relationship is discussed. Finally, a simulation study of the proposals is performed in order to assess their relative performance.

Keywords: Preference Models, Rank Modelling, Maximum Likelihood Estimation

1. Introduzione

L'analisi statistica di informazioni qualitative richiede innovazioni sostanziali nella modellistica tradizionale (Agresti, 1984). In tale ambito, lo studio sui dati di preferenza coinvolge numerosi approcci (Taplin, 1997), anche perche' esso si connette a situazioni che si presentano in modo differenziato. L'utilizzo di tali informazioni conduce, generalmente, a schemi connessi alle possibili permutazioni degli oggetti a confronto (Marden, 1995). Tuttavia, l'approccio che qui viene privilegiato e' quello di una esplicita parametrizzazione di un

modello probabilistico che tiene conto della procedura di selezione da parte del soggetto (D'Elia, 2000a).

L'ordinamento di oggetti prefissati da parte di soggetti differenti puo' essere considerato come uno schema che semplifica dei comportamenti abbastanza diffusi, anche laddove non sia richiesto o presupposto un semplice ordinamento fisico da parte del soggetto. Tale constatazione riguarda ovviamente un numero qualsiasi di oggetti e di connesse valutazioni ma, come vedremo in questo lavoro, l'opzione fra tre alternative costituisce molto piu' di un caso particolare. A tal fine, per fornire una prima idea dell'ampiezza delle problematiche che cio' implica, proporremo alcuni ambiti reali nei quali le risultanze di questo lavoro possono essere applicate con maggiore immediatezza.

In primo luogo, la valutazione di qualsiasi entita' (un oggetto, un lavoro fisico o intellettuale, un'opera letteraria, un corso universitario, etc.) puo' essere spesso assimilata ad una collocazione (fisica o virtuale) dell'entita' medesima in una posizione ben definita scelta fra tre alternative disgiunte che, per semplicita', si possono definire come: "insufficiente", "sufficiente", "buono". Il fatto che, nelle situazioni contingenti, si utilizzino espressioni e locuzioni anche molto differenti tra loro non elimina il fatto che, comunque, la valutazione si esprima mediante tre modalita' che sintetizzano posizioni di avversione, neutralita' e preferenza, rispettivamente.

In secondo luogo, l'efficacia di una terapia, di un farmaco, di una concimazione (e, per analogia, la gravita' di una diagnosi o la profondita' di un intervento) possono essere giudicate in modo negativo, nullo o positivo (ovvero, profondo, intermedio, superficiale). Questa tipologia di classificazione puo' essere estesa, ovviamente, anche a misurazioni quantitative che siano opportunamente discretizzate mediante due soglie di efficacia: per esempio, individuando due valori che determinino rispettivamente il passaggio dal danno all'indifferenza e dall'indifferenza al miglioramento.

In terzo luogo, le reazioni psicologiche rispetto ad un evento specifico (per esempio, ad una notizia televisiva, ad un incidente, alla situazione internazionale, al grado di fiducia nell'economia, etc.) si

prestano ad una graduazione fra modalita' che, ancora una volta, per semplicita', potremmo definire: negative, indifferenti, positive.

Infine, le collocazioni ideologiche o, piu' specificamente, politiche o religiose registrano sovente una pluralita' di posizioni che possono essere caratterizzate (a volta in maniera anche riduttiva) come: conservatori-indifferenti-riformisti, destra-centro-sinistra, credenti-indifferenti-atei, e cosi' via. Ovviamente, a differenza dello schema precedente, tali graduazioni non implicano una corrispondenza tra modalita' collocate nel medesimo ordine (non vi sono motivi perche' necessariamente i conservatori corrispondano alla destra politica ne' che siano credenti) ma solo una convenzione diffusa che semplifica una posizione rispetto alle altre e che, molto spesso, include una sorprendente pluralita' di posizioni intermedie.

Senza voler essere esaustivi, ci sembra cosi' che in svariati campi della ricerca (dalla medicina al marketing, dalla psicologia alla pedagogia, dalla politica alla religione, etc.) sia possibile fare riferimento ad una schematizzazione semplice ed essenziale delle scelte e delle preferenze operate da un collettivo di soggetti in rapporto a determinate alternative.

In questo lavoro, dopo aver richiamato brevemente il modello statistico di riferimento per l'analisi delle preferenze, ci soffermeremo sul caso delle tre alternative investigandone, anzitutto, la struttura probabilistica ed il significato del parametro caratterizzante, anche mediante una conveniente rappresentazione grafica. Quindi, affronteremo il problema della stima del parametro mediante il metodo dei momenti e quello della massima verosimiglianza. Tali stimatori sono entrambi consistenti ma, in generale, non coincidono sullo spazio parametrico: il confronto tra essi procedera' valutandone le differenze numeriche e l'efficienza relativa. Quindi, saranno affrontati i problemi derivanti dalla valutazione approssimata della varianza dello stimatore del coefficiente di preferenza allo scopo di predisporre delle procedure inferenziali operativamente efficaci. Alcune considerazioni finali - concernenti la possibilita' di generalizzare i risultati qui ottenuti- concluderanno il lavoro.

2. Un modello probabilistico di riferimento

Una collezione di oggetti $\mathcal{O}_j$, $j = 1, 2, ..., m$, che rispondono a requisiti di similarita' in rapporto a contesti circoscritti e che sono distinguibili tra loro sulla base di qualche caratteristica, ben definita e con modalita' disgiunte ed esaustive, vengono proposti al confronto (reale o ipotetico) di n soggetti. Come gia' anticipato, ai fini della nostra discussione, e' irrilevante se trattasi delle m modalita' di un unico oggetto ovvero di m differenti oggetti ovvero di uno tra m distinti "stati" fisici o psicologici che possono essere assunti da una determinata entita'.

Si individui un prefissato "oggetto" tra quelli definiti che, per semplicita', sara' indicato come oggetto $\mathcal{O}$ e che viene proposto alla attenzione di n soggetti per essere collocato -in modo univoco- in una delle possibili posizioni tra 1 ed m. Il nostro obiettivo e' quello di studiare la distribuzione di probabilita' della variabile casuale (v.c.) discreta rango $\mathcal{R}$ dell'oggetto $\mathcal{O}$, che e' caratterizzata dai valori $\mathcal{R} = 1, 2, ...m$ e dalle corrispondenti probabilita'.

Tra i possibili modelli, appare particolarmente convincente quello proposto da D'Elia (1999; 2000a) ed ampiamente discusso e sperimentato in molteplici situazioni reali. Infatti, assieme all'evidenza empirica, il modello e' derivato sulla base delle modalita' di scelta che il soggetto pone in essere in rapporto alle possibili collocazioni dell'oggetto, e si presta a numerose generalizzazioni e varianti. Peraltro, tale modello e' assimilabile a quello classico dell'estrazione di palline da un'urna ed appare, quindi, schematizzabile e generalizzabile in modo agevole.

In sostanza, si definisce la v.c. $\mathcal{R}$ come il numero di estrazioni senza ripetizione occorrenti perche' da un'urna, contenente palline "bianche" e "non-bianche" in proporzioni B e $m - 1$, rispettivamente, sia estratta *per la prima volta* una pallina "bianca". Tale struttura genera la v.c. Ipergeometrica Inversa -discussa con ampiezza ed in contesti differenti da Guenther (1975)- che nel seguito sigleremo con $\mathcal{R} \sim IGI(m, B)$, e la cui distribuzione di probabilita' e' definita da:

$$Pr(\mathcal{R}=r)=\frac{\binom{B+m-r-1}{m-r}}{\binom{B+m-1}{m-1}},\ r=1,2,...,m.$$

Si osservi esplicitamente che, essendo:

$$Pr(\mathcal{R}=1)=\frac{B}{B+m-1}=\theta;\quad B=(m-1)\frac{\theta}{(1-\theta)},$$

sia il parametro $B\in[0,\infty)$ che il parametro $\theta\in[0,1]$ costituiscono entrambi delle misure dirette della preferenza espressa verso l'oggetto. Per questo, definiamo il parametro θ come *coefficiente di preferenza* dell'oggetto $\mathcal{O}$.

Va sottolineato come il modello proposto non richieda necessariamente che l'urna predetta sia realizzabile in senso fisico. Infatti, essendo il parametro θ un numero reale, esso puo' essere approssimato con tutta l'accuratezza richiesta dal numero razionale $B/(B+m-1)$, ove B e' un intero appropriato, per un m prefissato. Sara', quindi, tale valore dell'intero B a determinare l'urna fisicamente realizzabile come quella piu' simile a quella presupposta dal modello.

Rinviando alla bibliografia citata per ulteriori sviluppi ed interpretazioni sul modello IGI, questo lavoro si sofferma con ampiezza sulla specificazione della v.c. $\mathcal{R}$ quando $m=3$, cioe' sulla v.c. $\mathcal{R}\sim IGI(3,\theta)$. In tale caso, quindi, la v.c. $\mathcal{R}$ esprime il numero di estrazioni senza ripetizione di palline da un'urna (che contiene 2 palline "non-bianche" e B palline "bianche") occorrenti affinche' esca per la prima volta una pallina "bianca".

Allora, la distribuzione di probabilita' della v.c. $\mathcal{R}$ diventa:

$$Pr(\mathcal{R}=r)=\frac{\binom{B+2-r}{3-r}}{\binom{B+2}{2}}=\frac{2}{(B+1)(B+2)}\binom{B+2-r}{3-r},\ r=1,2,3.$$

In particolare, esplicitando tale espressione (sia in termini di B che di θ) si ottiene lo schema seguente:

$Pr(\mathcal{R}=1)$	$\frac{B}{B+2}$	θ
$Pr(\mathcal{R}=2)$	$\frac{2\,B}{(B+1)\,(B+2)}$	$\frac{2\,\theta\,(1-\theta)}{(1+\theta)}$
$Pr(\mathcal{R}=3)$	$\frac{2}{(B+1)\,(B+2)}$	$\frac{(1-\theta)^2}{(1+\theta)}$
$Totali$	1	1

La seguente elementare verifica di calcolo delle probabilita' conferma tali risultati:

$i)$ L'evento $(\mathcal{R}=1)$ si verifica se e solo se la pallina "bianca" e' immediatamente estratta, il che ha probabilita' $B/(B+2)$.

$ii)$ L'evento $(\mathcal{R}=2)$ si verifica se e solo se la pallina "bianca" e' estratta dopo che alla prima estrazione si e' verificata pallina "non-bianca", il che ha probabilita' $(2/(B+2))*(B/(B+1),$ da cui l'espressione indicata.

$iii)$ L'evento $(\mathcal{R}=3)$ si verifica se e solo se la pallina "bianca" viene estratta dopo che si sono verificate due palline "non-bianche", per cui la probabilita' corrispondente e' $(2/(B+2))*(1/(B+1)*(1)$.

L'ultima colonna si ottiene parametrizzando la seconda mediante il parametro θ che esprime la probabilita' che l'oggetto $\mathcal{O}$ venga preferito fra i tre, cioe' venga collocato al primo posto nella graduatoria. Cio' conferma il ruolo del parametro θ come *coefficiente di preferenza* dell'oggetto $\mathcal{O}$.

Vale la pena di commentare le relazioni intercorrenti fra le tre probabilita' coinvolte nel modello e che sono esprimibili in funzione del

parametro θ. La Figura 1 evidenzia tale relazione e conferma l'ovvia corrispondenza inversa tra la prima e la terza probabilita'; meno immediato e' il comportamento della $Pr(\mathcal{R} = 2)$ che *non* e' compresa tra 0 e 1 ma varia nell'intervallo $[0, 1/3]$ raggiungendo il suo massimo in corrispondenza del valore di $\theta = 1/3$.

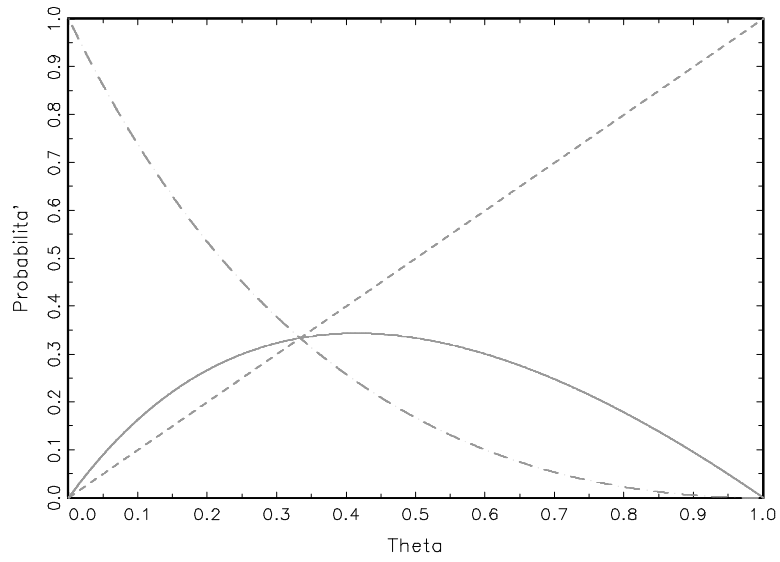

Figura 1. Probabilita' $Pr(\mathcal{R} = k)$, $k = 1, 2, 3$, *in funzione di* θ

Le tre probabilita' coincidono se e solo se $\theta = 1/3$ (ovvero, se e solo se $B = 1$), cioe' quando la v.c. IGI diventa una v.c. Uniforme discreta definita sui primi tre interi: in tale situazione, il soggetto colloca l'oggetto $\mathcal{O}$ indifferentemente in una delle tre collocazioni disponibili. Si puo' proporre, allora, il seguente schema interpretativo del modello in funzione del coefficiente di preferenza θ:

Campo di variazione di θ	*Campo di variazione di* B	*Atteggiamento verso l'oggetto* $\mathcal{O}$
$\theta \in [0, 1/3)$	$B \in [0, 1)$	repulsione
$\theta \equiv 1/3$	$B \equiv 1$	indifferenza
$\theta \in (1/3, 1]$	$B \in (1, \infty)$	attrazione

Una riflessione critica su tale modello appare comunque opportuna: la v.c. $\mathcal{R}$ cosi' definita presenta una distribuzione di probabilita' monotòna decrescente (se $B > 1$) e monotòna crescente (se $B < 1$), per cui tranne il caso di una distribuzione Uniforme discreta (quando $B = 1$), la moda della distribuzione (e quindi la presumibile frequenza piu' diffusa delle preferenza) puo' avvenire solo per $(\mathcal{R} = 1)$ oppure per $(\mathcal{R} = 3)$, rispettivamente. In altri termini, il modello prescelto richiede che sia maggioritaria una delle due scelte estreme (attrazione oppure repulsione, secondo lo schema precedente) per cui mal si adatta a quelle situazioni nelle quali la situazione intermedia riceve i maggiori consensi: d'altra parte, cio' deriva strettamente dalla natura probabilistica dello schema dell'estrazione dall'urna cui il modello medesimo si richiama[1].

Poiche' le tre probabilita' sono elementi di un simplesso definito sul piano (perche' non-negative e di somma unitaria), e' possibile fornire per esse una rappresentazione parametrica (in funzione di θ) di tipo triangolare, come nella seguente Figura 2.

[1] Tale possibilita', invece, sussiste per la v.c. Binomiale traslata (BIT), ampiamente discussa nel contesto dei modelli di preferenza da D'Elia (2000b, c).

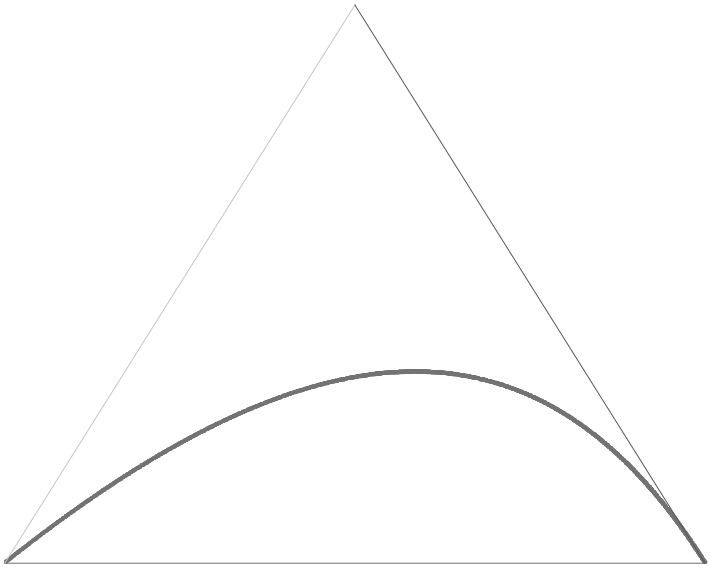

Figura 2. Rappresentazione triangolare di $Pr(\mathcal{R} = k|\theta)$, $k = 1, 2, 3$

In tale rappresentazione, ogni punto della curva e' caratterizzato dalla tripla: $\{Pr(\mathcal{R} = 1), Pr(\mathcal{R} = 2), Pr(\mathcal{R} = 3)\}$ per un fissato θ. Pertanto, la curva della Figura 2, al variare di θ, esprime il luogo geometrico di tali coordinate tali che $\sum_{k=1}^{3} Pr(\mathcal{R} = k) = 1$. In particolare, essa conferma che mentre le probabilita' degli eventi $(\mathcal{R} = 1)$ ed $(\mathcal{R} = 3)$ possono variare tra 0 e 1 (perche' le corrispondenti coordinate possono assumere tutti i valori del campo di definizione), la probabilita' dell'evento $(\mathcal{R} = 2)$ (che e' rappresentata dall'altezza della curva rispetto alla base del rettangolo) e' invece ristretta al sub-intervallo $[0, 1/3]$.

La funzione generatrice di tale v.c. risulta essere:

$$G(t) = \frac{e^t}{1+\theta}\Big\{\theta + \left[\theta + e^t(1-\theta)\right]^2\Big\}.$$

Da essa, oppure direttamente, e' agevole mostrare che i momenti di ordine k della v.c. $\mathcal{R}$ sono:

$$\mu_k = \left[3^k + \left(1 + 2^{k+1} - 2*3^k\right)\theta + \left(1 - 2^{k+1} + 3^k\right)\theta^2\right],\ k = 0, 1, ...$$

In particolare, si deduce che:

$$\mathbb{E}(\mathcal{R}) = \frac{3-\theta}{1+\theta};\ Var(\mathcal{R}) = \frac{2\,\theta\,(1-\theta)(3-\theta)}{(1+\theta)^2};$$

inoltre, il coefficiente di variazione e':

$$CV(\mathcal{R}) = \sqrt{\frac{3-\theta}{2\,\theta\,(1-\theta)}}.$$

Esaminando tali indicatori sullo spazio parametrico di $\theta \in [0, 1]$, risulta che al crescere di θ il valore medio decresce monotonicamente dal massimo valore $\mathbb{E}(\mathcal{R}) = 3$ (quando $\theta = 0$) al minimo valore $\mathbb{E}(\mathcal{R}) = 1$ (quando $\theta = 1$). Invece, la varianza raggiunge il suo minimo teorico ai due estremi dello spazio parametrico (in effetti, quando $\theta = 0$ oppure $\theta = 1$ la v.c. $\mathcal{R}$ e' degenere) ed un unico massimo (pari a $0.671006...$) quando $\theta = 0.299664...$

Inoltre, il coefficiente di asimmetria e' pari a:

$$Asym(\mathcal{R}) = \frac{(3\,\theta - 1)(5 - 3\,\theta)}{(3-\theta)\sqrt{2\,\theta\,(3-\theta)\,(1-\theta)}}.$$

Da tale risultato emerge che il segno dell'asimmetria della distribuzione e' concorde con il segno di $(\theta - 1/3)$, ovvero di $(B - 1)$. Se tale quantita' e' positiva, la probabilita' massima si verifica per $(\mathcal{R} = 1)$ -che quindi costituisce la moda della distribuzione- e l'asimmetria e' positiva;

se invece essa e' negativa, la moda si presenta in corrispondenza di $(\mathcal{R} = 3)$ e l'asimmetria e' negativa. L'asimmetria e' nulla se e solo se $\theta = 1/3$ (e, quindi, $B = 1$); in tale caso, $Pr(\mathcal{R} = k) = 1/3$, $k = 1, 2, 3$.

Infine, si osservi che la generazione di numeri pseudo-casuali per tale famiglia di v.c. e' computazionalmente molto efficiente[2].

3. La stima del coefficiente di preferenza

Le informazioni contenute nel campione osservato $(r_1, r_2, ..., r_n)$ estratto casualmente dalla popolazione $\mathcal{R} \sim IGI(3, \theta)$ possono essere riassunte nella seguente distribuzione di frequenza osservata:

$Ranghi\ osservati$	1	2	3	$Totale$
$Frequenze\ assolute$	n_1	n_2	n_3	n
$Frequenze\ relative$	f_1	f_2	f_3	1

Sfruttando le relazioni tra frequenze assolute e relative, e' immediato dedurre che il rango medio campionario vale:

---

[2] La seguente procedura, scritta in linguaggio GAUSS, genera un vettore di n numeri pseudo-casuali da una v.c. $\mathcal{R} \sim IGI(m, B)$ sulla base di un risultato teorico che connette il coefficiente di preferenza alle v.c. Beta ed Uniforme, come discusso in D'Elia (2000a); Gelman et al. (1995).

@ Genera da una v.c. Beta(1,B) il parametro theta -mediante la tecnica dell'inversione della funzione di ripartizione inversa- e poi genera R da una v.c. Bin(m-1,theta)+1 mediante le v.c. Uniformi sull'intervallo (0,1). © D'Elia, 2000. @

```
PROC SIMULIGI(n,m,B);
LOCAL vettuno,vettore;
        vettuno= 1-(rndu(n,1))^(1/B);
        vettore=sumc(rndu(m-1,n) .<= vettuno')+1;
RETP(vettore);
ENDP;
```

Per rendersi conto della velocita' e dell'efficienza della procedura, si osservi che su un PC Pentium III 750, 128MbRam, essa genera 1 milione di numeri pseudo-casuali per la v.c. $\mathcal{R} \sim IGI(3, B)$ mediamente in 1.4 secondi.

$$\overline{r} = \frac{n_1 + 2\,n_2 + 3\,n_3}{n} = 2 + f_3 - f_1.$$

Pertanto, tale quantita' e' determinata -rispetto al valore centrale della distribuzione, pari a 2- dall'eccedenza tra le frequenze relative f_3 ed f_1. Una distribuzione campionaria quasi simmetrica produrra', quindi, un valore medio campionario dei ranghi vicino a 2.

Al variare del campione, la precedente stima determina lo stimatore media campionaria definito da:

$$\overline{R} = \frac{N_1 + 2\,N_2 + 3\,N_3}{n} = 2 + \mathcal{F}_3 - \mathcal{F}_1,$$

essendo (N_1, N_2, N_3) e $(\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3)$ le v.c. frequenze assolute e relative, rispettivamente. Tali v.c. triple possiedono, in effetti, una variabilita' bi-dimensionale poiche' valgono i rispettivi vincoli:

$$N_1 + N_2 + N_3 = n; \;\; \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3 = 1.$$

Il modello probabilistico prescelto e' specificato dal solo parametro θ, per cui e' importante individuare metodi di stima per la sua determinazione campionaria. In questo lavoro, esamineremo due soli metodi di stima consistenti[3].

Utilizzando il *metodo dei momenti*, dall'equazione: $\mathbb{E}(\mathcal{R}) = \overline{r}$ si deduce immediatamente lo stimatore:

[3] In effetti, il problema qui esaminato potrebbe essere affrontato anche mediante il metodo del minimo Chi-quadrato che, nello specifico, non conduce ad una formulazione esplicita dello stimatore. Peraltro, e' noto che tale metodo produce stimatori con proprieta' asintoticamente equivalenti a quelli della massima verosimiglianza (Rao, 1973, 352-353; Serfling, 1980, 163-165).

$$T_{mo} = \frac{3 - \overline{\mathcal{R}}}{1 + \overline{\mathcal{R}}} = \frac{1 - (\mathcal{F}_3 - \mathcal{F}_1)}{3 + (\mathcal{F}_3 - \mathcal{F}_1)}.$$

Per esempio, da un campione casuale per il quale si e' osservato: $f_1 = 1/2;\ f_2 = 1/3; f_3 = 1/6;$ si trae: $\overline{r} = 5/3$ e, quindi, la stima dei momenti diventa: $t_{mo} = 0.5$.

La formulazione stessa dello stimatore (che implica un rapporto tra stimatori del valore medio della v.c. $\mathcal{R}$) evidenzia come esso sia generalmente distorto per θ.

Utilizzando il *metodo della massima verosimiglianza* (ML), la funzione di verosimiglianza puo' essere espressa agevolmente se si tiene conto che le informazioni derivanti dal campione casuale $(r_1, r_2, ..., r_n)$ sono equivalenti a quelle derivabile dalle frequenze assolute (n_1, n_2, n_3) o relative (f_1, f_2, f_3) dei tre possibili ranghi.

Pertanto, si ha:

$$\mathcal{L}(\theta; n_1, n_2, n_3) = [\theta]^{n_1} \left[\frac{2\,\theta\,(1-\theta)}{(1+\theta)}\right]^{n_2} \left[\frac{(1-\theta)^2}{(1+\theta)}\right]^{n_3}.$$

Utilizzando i logaritmi (e ricordando i vincoli tra le frequenze), a meno di una costante additiva inessenziale, la funzione di log-verosimiglianza si puo' scrivere:

$$log\,\mathcal{L}(\theta) = (n - n_3) log\,\theta + (n_2 + 2n_3) log(1-\theta) - (n - n_1) log(1+\theta).$$

Uguagliando a zero la derivata prima di tale espressione, esplicitando poi il risultato tramite le frequenze relative f_1, f_3 (infatti, sono essenziali solo due frequenze su tre), si perviene alla seguente equazione di secondo grado in θ:

$$\theta^2 + [2(1 - f_1) + f_3]\,\theta - (1 - f_3) = 0.$$

la cui soluzione ammissibile costituisce la stima di massima verosimiglianza per θ.

Pertanto, lo stimatore ML per θ e':

$$T_{ML} = -1 + \left(\mathcal{F}_1 - \frac{\mathcal{F}_3}{2}\right) + \sqrt{\left(\mathcal{F}_1 - \frac{\mathcal{F}_3}{2}\right)^2 + 2(1 - \mathcal{F}_1)}.$$

Per esempio, da un campione casuale per il quale si e' osservato: $f_1 = 1/2;\ f_2 = 1/3; f_3 = 1/6;$ si trae l'equazione di secondo grado: $6\theta^2 + 7\theta - 5 = 0$, la cui soluzione ammissibile produce la stima: $t_{ML} = 0.5$. Si osservi che, per i particolari dati numerici prescelti, le stime ottenute con i due metodi hanno prodotto lo stesso risultato.

La successiva Figura 3 mostra i valori delle stime cosi' determinate al variare di f_1, f_3 (e, quindi, indirettamente di $f_2 = 1 - f_1 - f_3$), per valutare le differenze numeriche tra i due metodi qui esaminati sull'intero spazio campionario. I grafici mostrano che esiste una notevole similarita' tra le determinazioni numeriche dei due stimatori, quando applicati allo stesso campione osservato.

Rinviando al paragrafo successivo per la discussione sulla valutazione statistica comparata tra i due stimatori in termini di non-distorsione e di efficienza relativa, per campioni di dimensione finita, osserviamo qui che la differenza in valore assoluto tra le stime prodotte dai due stimatori, sullo spazio campionario di tutte le possibili realizzazioni, presenta il seguente campo di variazione:

$$0 \leq |t_{ML} - t_{mo}| \leq 0.080880...$$

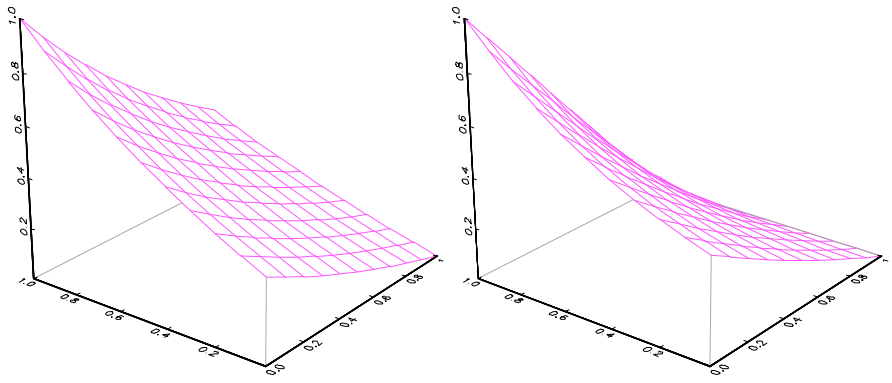

Figura 3. Possibili valori delle stime dei momenti e ML

Cio' conferma che, a tutti gli effetti pratici, anche lo stimatore ottenuto con il metodo dei momenti (di piu' semplice calcolo) puo' essere utilizzato in sostituzione di quello della massima verosimiglianza perche' -anche nel peggiore dei casi- il divario e' trascurabile.

4. Uno studio di simulazione

Discutiamo, ora, mediante un ampio studio di simulazione, l'efficienza relativa dei due stimatori individuati nel paragrafo precedente. La notevole efficacia computazionale della procedura di generazione dei numeri pseudo-casuali ci ha consentito di esaminare

con estremo dettaglio lo spazio parametrico di θ allo scopo di valutare le proprieta' statistiche degli stimatori nel caso di distribuzioni abbastanza differenti tra loro (rispetto all'asimmetria, per esempio). Per questo, invece, di presentare numerose tabelle che riportano gli indicatori essenziali dell'esperimento di simulazione, preferiamo visualizzare mediante alcuni grafici i comportamenti di tali misure sull'intero spazio parametrico.

L'analisi empirica ha mostrato che, in primo luogo, una simulazione di 1000 campioni casuali (per n e θ fissati) era piu' che sufficiente per stabilizzare il risultato. In secondo luogo, per numerosita' pari a quelle usuali per tale tipo di problematica (per esempio, con $n > 50, 100$), la distorsione e' sempre stata talmente trascurabile da poter essere ritenuta nulla; in ogni caso, e' stato valutato il MSE di ciascuno stimatore e il RMSE (perche' espresso nell'unita' di misura del parametro).

La Figura 4 evidenzia il MSE di entrambi gli stimatori (calcolato per ciascun valore prefissato di θ, sulla base di 1000 simulazioni di campioni osservati di $n = 150$ ranghi) e mostra come essi siano di fatto indistinguibili, soprattutto per valori di $\theta > 2/3$. Si nota, inoltre, in modo evidente, come il MSE diminuisca agli estremi dello spazio parametrico, anche se cio' avviene in modo asimmetrico.

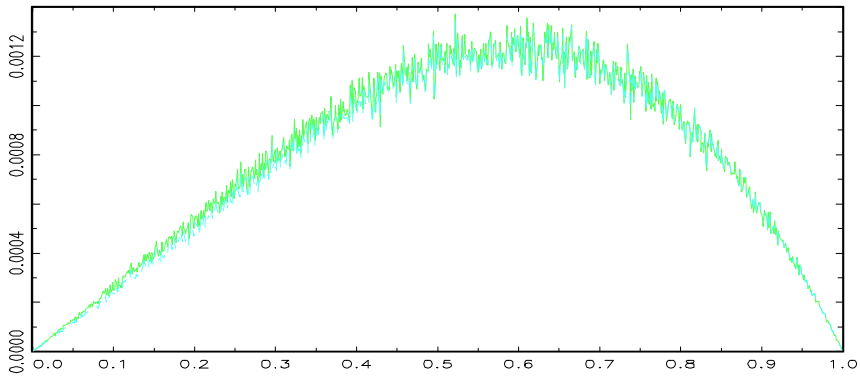

Figura 4. *MSE simulato per gli stimatori* T_{mo} *e* T_{ML}

Premesso, quindi, che l'ordine di grandezza dei MSE e' fra loro omogeneo, e' interessante una loro valutazione comparata mediante il rapporto $MSE(T_{mo})/MSE(T_{ML})$ che costituisce una misura dell'efficienza relativa tra i due stimatori. A tale riguardo, la Figura 5 conferma la (prevista) preferenza per lo stimatore T_{ML}. Inoltre, si vede come l'efficienza dello stimatore dei momenti tenda velocemente ad 1 quando $\theta > 2/3$ (circa) mentre per bassi valori del parametro θ la variabilita' dello stimatore dei momenti puo' superare quella dello stimatore di massima verosimiglianza anche del 10-18%.

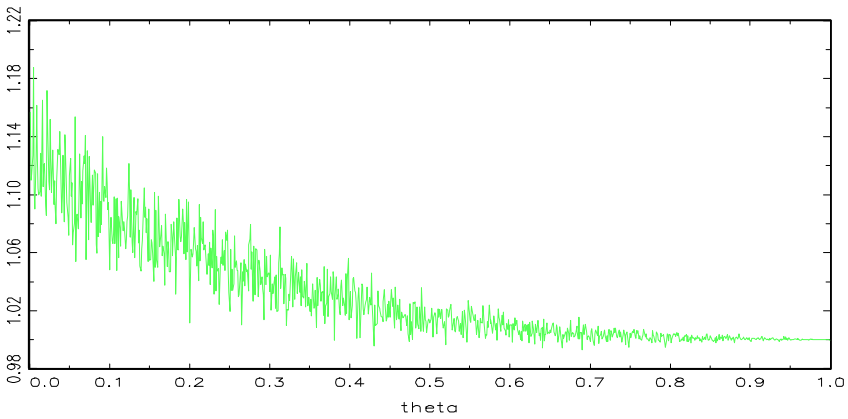

Figura 5. Efficienza relativa dello stimatore T_{mo} rispetto a T_{ML}

Poiche', nelle figure precedenti, puo' forse sorprendere la notevole variabilita' tra valori consecutivi del parametro θ, va chiarito che essa e' indotta dalla "finezza" della esplorazione adottata e dalla ovvia variabilita' campionaria di una simulazione non eccessivamente onerosa. Questo effetto, ovviamente, si riduce sensibilmente quando si passa da 1000 a 10000 oppure a 100000 simulazioni, per ogni prefissato valore del parametro.

Per una visualizzazione piu' usuale, nella Figura 6, presentiamo un grafico analogo a quello della Figura 4 relativamente all'indice $RMSE(T_{ML})$ ove, pero', tale misura e' stata adeguatamente

perequata nei valori vicini per produrre una curva piu' regolare, al variare di θ.

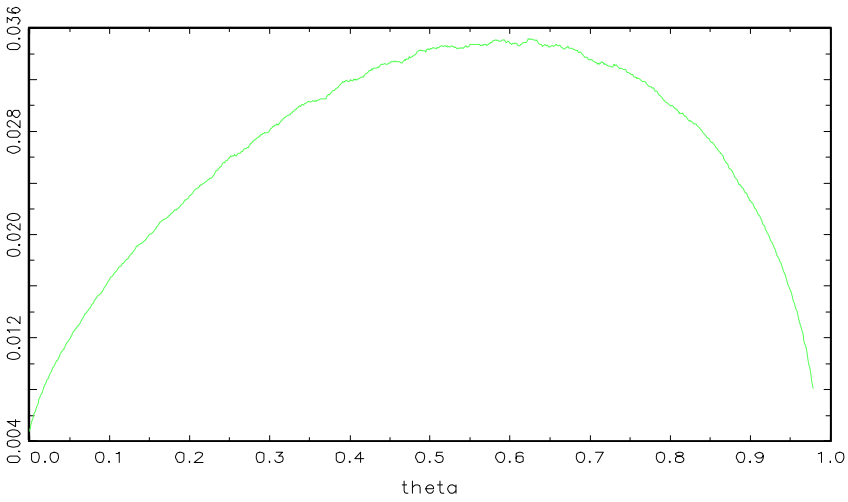

Figura 6. RMSE (T_{ML}) perequato per i dati della Figura 4

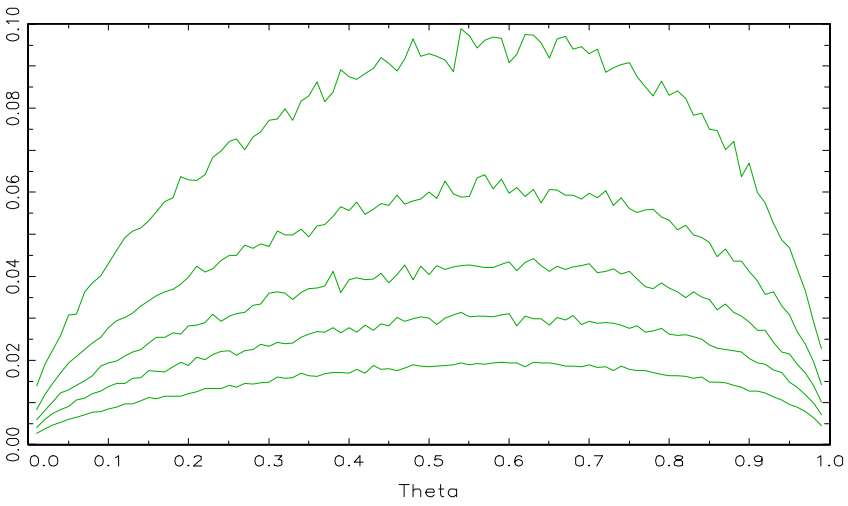

Figura 7. RMSE(T_{ML}) per n=20, 50, 100, 200, 500

Infine, per valutare la modifica della variabilita' dello stimatore rispetto alla numerosita' campionaria, la Figura 7 evidenzia la quantita' $RMSE(T_{ML})$ al variare di $n = 20, 50, 100, 200, 500$ (dall'alto in basso nel grafico). Emerge cosi' un comportamento coerente con le aspettative e simile a quello dei grafici precedenti.

La distribuzione dello stimatore di massima verosimiglianza e' asintoticamente Normale e lo studio di simulazione (i cui dettagli qui non riportiamo, per brevita') lo ha confermato per numerosita' finite sull'intero spazio parametrico: tale approssimazione e' piu' che accettabile anche per piccoli valori di n. Per esempio, la Figura 8 mostra la distribuzione perequata della stima di massima verosimiglianza quando $n = 50$ e per $\theta = 0.3$.

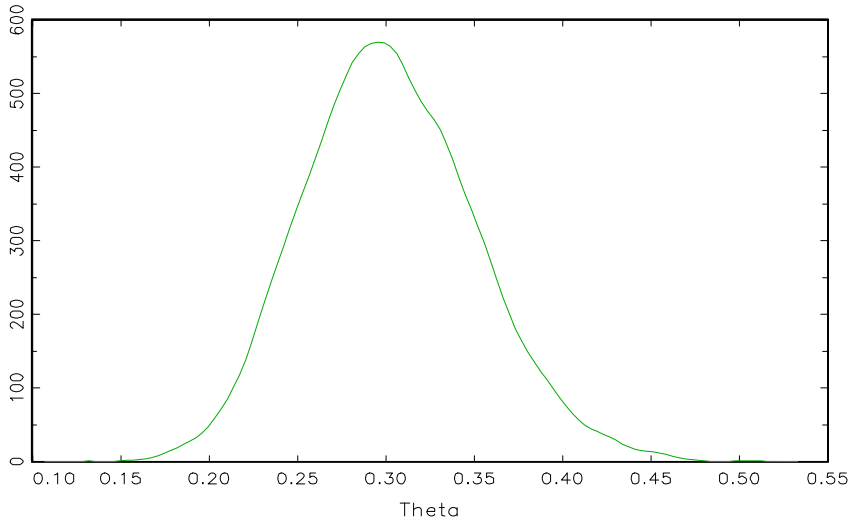

Figura 8. Un esempio di distribuzione simulata dello stimatore

5. Una valutazione approssimata della varianza dello stimatore

Le considerazioni precedenti hanno mostrato -assieme alla sostanziale non-distorsione di entrambi gli stimatori discussi- una notevole similarita' di comportamento. Inoltre, lo stimatore T_{ML} (che,

d'ora innanzi, indicheremo con T_n) ha confermato la validita' della distribuzione asintotica anche per campioni finiti di modeste dimensioni.

Tuttavia, perche' le procedure inferenziali della verifica di ipotesi e della costruzione di intervalli di confidenza per θ diventino operative, e' necessario procedere ad una determinazione della varianza dello stimatore, per ciascun prefissato n. Invece di procedere ad una determinazione analitica esatta (che e' connessa ai valori medi delle derivate della verosimiglianza), in questa sede, esploreremo una determinazione numerica approssimata di tale variabilita' espressa in funzione di n, la quale risultera' molto stabile in tale ambito.

A tale proposito, l'esperimento di simulazione condotto nel paragrafo precedente ha mostrato che la variabilita' dello stimatore dipende sia dalla numerosita' n (in modo monotonicamente decrescente) che dal parametro θ (in modo convesso sullo spazio parametrico). Per questo, al fine di sviluppare un'approssimazione accettabile per $n\,Var(T_n)$, cercheremo di stimare con il metodo dei minimi quadrati una relazione analitica per la determinazione campionaria simulata di tale varianza in funzione di θ.

Operativamente, tra le infinite scelte possibili per adattare l'andamento osservato di $n\,Var(T_n)$, abbiamo preferito una funzione cubica nel parametro θ:

$$f(\theta) = \beta_0 + \beta_1\,\theta + \beta_2\,\theta + \beta_3\,\theta^3, \;\; \theta \in [0,1],$$

determinando i coefficienti mediante il metodo dei minimi quadrati, al variare di $n = 20, 50, 100, 200, 500$.

In questa analisi, generalmente, il numero di simulazioni effettuato e' stato fissato pari a 10000, eccetto per i casi di $n = 20, 50$, quando si e' preferito effettuare 100000 simulazioni allo scopo di accentuare la regolarita' dei risultati. La successiva Tabella 1 presenta tali valori numerici.

Tabella 1. Coefficienti stimati della funzione $n\,Var(T_n)$

n	β_0	β_1	β_2	β_3
20	-0.00328	0.40409	0.20794	-0.61240
50	-0.00185	0.36688	0.30034	-0.66836
100	-0.00207	0.36302	0.31500	-0.67876
200	-0.00106	0.34532	0.35368	-0.70090
500	-0.00096	0.34473	0.35249	-0.69858

La possibilita' di una efficace utilizzazione dell'espressione $f(t)$ calcolata tramite i coefficienti esposti nella Tabella 1 si basa su molteplici elementi:

i) l'adattamento tra $n\,Var(T_n)$ ottenuta dalle simulazioni e la sua determinazione teorica tramite la detta funzione stimata -sull'intero spazio parametrico- e' notevole (l'indice R^2 supera sempre, anche per n basso, il valore di 0.998), come confermano i grafici della Figura 9, ove e' esemplificato il caso di $n = 100$;

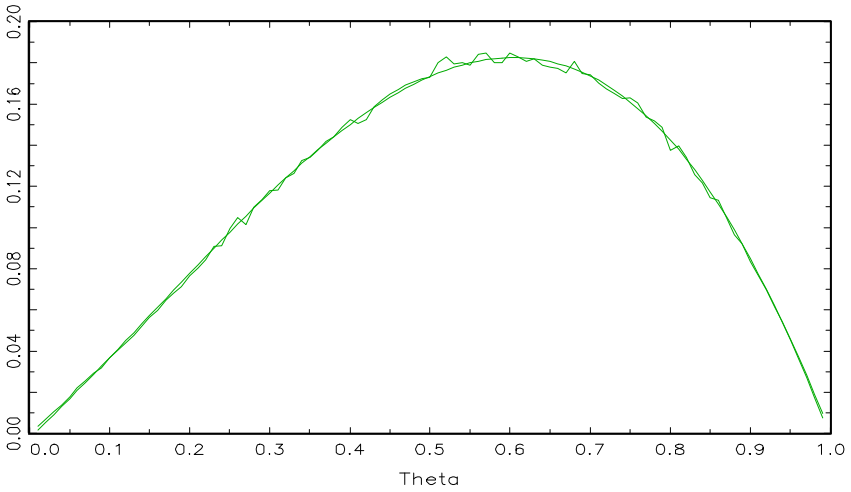

Figura 9. Determinazione simulata e adattata di $n\,Var(T_n)$*, n=100*

ii) la correttezza dell'approssimazione prescelta e' confermata dall'andamento praticamente costante di $n\,Var(T_n)$, sullo spazio parametrico di θ, per qualsiasi valore di n, come mostra la Figura 10;

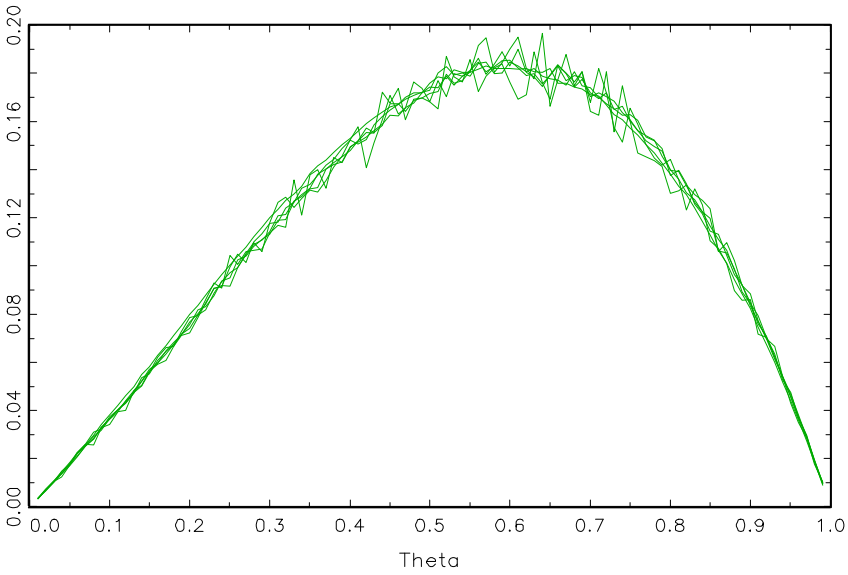

Figura 10. Determinazioni di $n\,Var(T_n)$,n=20,50,100,200,500,1000

iii) le variazioni dei coefficienti stimati nella Tabella 1 sono coerenti e concordi in segno, muovendosi nella stessa direzione al crescere di n: cio' facilita' la loro utilizzazione per ampie classi di numerosita' campionaria.

Ovviamente, sono possibili numerosi ulteriori miglioramenti[4] nella specificazione di una formulazione analitica per $n\,Var(T_n)$, ma l'entita'

[4] Per esempio, si potrebbe osservare che al crescere di n il peso e la significativita' della costante β_0 tendono a diminuire, oppure che una funzione razionale potrebbe fornire adattamenti migliori, e cosi' via. Inoltre, si potrebbero perequare i coefficienti presentati nella Tabella 1 effettuando un numero molto piu' elevato di simulazioni e per valori piu' ravvicinati della numerosita' n.

dei vantaggi addizionali che si raggiungono non giustifica quasi mai la opportunita' di procedure a specificazioni complesse e non traducibili in procedure operative semplici ed immediate come quella proposta.

6. Procedure inferenziali per il coefficiente di preferenza

E' possibile ora suggerire delle procedure inferenziali di tipo approssimato per il coefficiente di preferenza θ. Quanto diremo vale in modo rigoroso per lo stimatore di massima verosimiglianza (T_n) ma l'evidenza dei paragrafi precedenti conferma che e' possibile, quasi sempre, sostituirvi anche lo stimatore dei momenti.

Poiche' e' lecito assumere qui che lo stimatore T_n sia asintoticamente Normale e asintoticamente non-distorto per θ, e con una varianza asintotica pari a quella determinata numericamente nel paragrafo precedente:

$$T_n \rightarrow \mathcal{N}(\theta,\ f(\theta)/n),$$

standardizzando, si ha:

$$\sqrt{n}\ \frac{T_n - \theta}{\sqrt{f(\theta)}} \rightarrow \mathcal{N}(0,1).$$

Tale risultato consente la determinazione di regioni critiche asintotiche nonche' la costruzione di intervalli di confidenza, sui quali ci soffermeremo ora con maggiore dettaglio.

Per la costruzione di un intervallo di confidenza per il parametro θ, con coefficiente di confidenza $(1-\alpha)$, occorre "invertire" rispetto a θ l'espressione $f(\theta)$; cio' e' analiticamente complesso, in generale, a meno di non ricorrere a tecniche numeriche (determinando pero' il risultato volta per volta).

Alternativamente, grazie alla consistenza dello stimatore T_n e alla continuita' della funzione $f(\theta)$, e' possibile sostituire $f(T_n)$ al posto di

$f(\theta)$. In tal modo, con agevoli passaggi, si giunge al seguente intervallo asintotico per θ con coefficiente di confidenza $1-\alpha$:

$$T_n - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{f(T_n)} < \theta < T_n + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{f(T_n)}\,,$$

dove abbiamo indicato con $z_{\alpha/2}$ il quantile $(1-\alpha/2)$ della v.c. $\mathcal{Z} \sim \mathcal{N}(0,1)$.

Per esempio, supponiamo che -nella graduatoria effettuata da parte di un campione casuale di $n = 100$ soggetti- un determinato oggetto $\mathcal{O}$ sia stato collocato al primo posto da 70 soggetti e al secondo posto da 20 soggetti. Quindi, le frequenze relative osservate sono: $f_1 = 0.7$; $f_2 = 0.2$; $f_3 = 0.1$, e la stima di massima verosimiglianza[5] risulta:

$$t_n = -1 + (0.7 - 0.1/2) + \sqrt{(0.7 - 0.1/2)^2 + 2(1 - 0.7)} = 0.66119.$$

Ora, utilizzando la funzione approssimante per $n\,Var(T_n)$ quando $n = 100$, cioe':

$$f(t_n) = -0.00207 + 0.36302\,t_n + 0.31500\,t_n^2 - 0.67876\,t_n^3\,,$$

e, sostituendovi la precedente stima, si ha: $f(0.66119) = 0.17947$. Poiche', per $\alpha = 0.05$, risulta: $z_{\alpha/2} = 1.96$, alla fine l'intervallo di confidenza asintotico per il coefficiente di preferenza θ, al livello di confidenza 95%, e' determinato da:

[5] Si noti che, per questo esempio, la stima ottenuta con il metodo dei momenti risulterebbe pari a $t_n = 0.666...$, con una differenza relativa in piu', rispetto a quella di massima verosimiglianza, di appena lo 0.8 %.

$$0.66119 - \frac{1.96}{\sqrt{100}}\sqrt{0.17947} < \theta < 0.66119 + \frac{1.96}{\sqrt{100}}\sqrt{0.17947}$$

ovvero da:

$$0.57816 < \theta < 0.74422.$$

Con gli opportuni accorgimenti e tenuto conto delle proprieta' di invarianza degli stimatori di massima verosimiglianza, e' immediato tradurre i risultati ottenuti sin qui in funzione del parametro B che, alternativamente, puo' essere utilizzato per caratterizzare il modello probabilistico prescelto.

7. Considerazioni finali

In questo lavoro abbiamo discusso un modello probabilistico efficace per descrivere ed interpretare le situazioni di classificazione e graduazione di un oggetto da parte di una pluralita' di soggetti quando le alternative possibili sono tre. Quindi, ci siamo soffermati sul confronto tra lo stimatore dei momenti e quello di massima verosimiglianza i quali -nella fattispecie- consentono entrambi di pervenire ad una formulazione analitica esplicita per la determinazione del coefficiente di preferenza θ che caratterizza il modello proposto.

E' emerso che i due stimatori sono praticamente coincidenti sull'intero spazio parametrico, entrambi asintoticamente non-distorti e con una prevedibile maggiore efficienza dello stimatore di massima verosimiglianza, che tende a ridursi al crescere del parametro. In particolare, e' stato possibile pervenire ad una formulazione analitica approssimata ma molto accurata della varianza asintotica dello stimatore. In tal modo, le tradizionali procedure inferenziali possono essere applicate con notevole semplicita'.

Ulteriori sviluppi di questo approccio possono riguardare i seguenti punti:

i) confrontare questo modello con alcune proposte alternative, come per esempio quelle derivante dalla v.c. Binomiale traslata;

ii) generalizzare i risultati qui discussi mediante l'analisi di un modello a $m > 3$ alternative.

In linea con tali sviluppi, per esempio, e' lecito congetturare che la soluzione esatta per lo stimatore di massima verosimiglianza del coefficiente di preferenza nel caso di m alternative sia la soluzione ammissibile di un polinomio di grado $m - 1$ in θ.

Ringraziamenti: Il presente lavoro ha beneficiato di contributi per la ricerca provenienti da progetti di Ateneo e MURST, afferenti al Dipartimento di Scienze Statistiche della Universita' di Napoli Federico II. Si ringraziano i referees per i suggerimenti critici e le richieste di modifiche che hanno consentito di apportare miglioramenti importanti alla versione finale di questo articolo.

Riferimenti bibliografici

Agresti A. (1984) *Analysis of Ordinal Categorical Data*, John Wiley & Sons, New York

D'Elia A. (1999) A Proposal for Ranks Statistical Modelling, *Proceedings of the 14th International Workshop on Statistical Modelling*, (Friedl H., Berghold A., Kauermann G. editors), Graz, Austria, 468-471.

D'Elia A. (2000a) Un modello lineare generalizzato per i ranghi: aspetti statistici, problemi computazionali e verifiche empiriche, *Italian Journal of Applied Statistics*, 12, in corso di stampa.

D'Elia A. (2000b) A Shifted Binomial Model for Rankings, *Proceedings of the 15th International Workshop on Statistical Modelling* (V. Nuñez-Anton & E. Ferreira, editors), Servicio Editorial de la Universidad de Pais Vasco, Bilbao, Spagna, 412-416.

D'Elia A. (2000c) Il meccanismo dei confronti appaiati nella modellistica per graduatorie: sviluppi statistici ed aspetti critici, *Quaderni di Statistica*, 2, 173-204.

Gelman A., Carlin J. B., Stern H. S., Rubin D. B. (1995) *Bayesian Data Analysis*, Chapman & Hall, London.

Guenther W. C. (1975) The Inverse Hypergeometric - A Useful Model, *Statistica Neerlandica*, 29, 129-144.

Marden J. I. (1995) *Analyzing and Modeling Rank Data*, Chapman & Hall, London.

Rao C. R. (1973) *Linear Statistical Inference and Its Applications*, 2^{nd} edition, John Wiley & Sons, New York.

Serfling R. J. (1980) *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Taplin R. H. (1997) The Statistical Analysis of Preferences Data, *Applied Statistics*, 46, 49-512.