Rivista di Statistica Applicata Vol. 20 n. 3 Settembre 1987

STATISTICA COMPUTA-ZIONALE

Rubrica a cura di Natale Lauro

Medie e covarianze esatte ed approssimate delle statistiche ordinate della distribuzione valori estremi (I tipo) standardizzata

Corrado Provasi, Dipartimento di Scienze Statistiche – Università di Padova

Vengono presentate alcune routine scritte in FORTRAN 77 che consentono di calcolare in forma esatta ed approssimata le medie e le covarianze delle statistiche ordinate della distribuzione valori estremi (I tipo) standardizzata. Nella discussione preliminare vengono descritte le procedure di calcolo impiegate e, in seguito, vengono illustrate la struttura delle routine, l'accuratezza dei calcoli e i tempi di esecuzione. Infine, viene fornito un esempio di utilizzo.

1. Descrizione e scopo

SOVEM1, SOVEC1 e SOVEM2, SOVEC2 sono routine scritte in FORTRAN 77 che consentono di calcolare, rispettivamente, in forma esatta ed approssimata le medie e le covarianze delle statistiche ordinate $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ corrispondenti ad un campione casuale di dimensione n proveniente dalla distribuzione valori estremi (I tipo) standardizzata, la cui funzione di ripartizione è data da

$$F(x) = \exp(-e^{-x}).$$

Tali medie e covarianze possono essere impiegate, tra l'altro, per costruire test funzionali (si veda, per esempio, Provasi, 1984) e per ottenere le migliori stime lineari non distorte dei parametri della distribuzione valori estremi (I tipo) generale (Lieblein e Zelen, 1956).

1.1. Calcolo esatto

Le espressioni dei momenti del primo e secondo ordine e dei momenti misti del primo ordine di $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ sono (Lieblein, 1953):

$$E(X_{i:n}) = n\binom{n-1}{i-1}\sum_{r=0}^{n-i}(-1)^r\binom{n-1}{r}g_1(i+r), \tag{1}$$

$$E(X_{i:n}^2) = n\binom{n-1}{i-1}\sum_{r=0}^{n-i}(-1)^r\binom{n-1}{r}g_2(i+r), \tag{2}$$

$i = 1, 2, \ldots, n$, e

$$E(X_{i:n}X_{j:n}) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\sum_{s=0}^{j-i-1}\sum_{s=0}^{n-j}(-1)^{r+s} \cdot \binom{j-i-1}{r}\binom{n-j}{s}\phi(i+r, j-i-r+s) \tag{3}$$

$i = 1, 2, \ldots, n-1, \quad j = i+1, i+2, \ldots, n$, dove

$$g_1(c) = (\gamma + \ln c)/c,$$

$$g_2(c) = [\pi^2/\sigma + (\gamma + \ln c)^2]/c,$$

$$\phi(t,u) = \{(u-t)g_2(t+u) + t^2[g_1(t)]^2 + 2L(1+t/u) -$$

$$- [\ln(t/u)]^2 - \pi^2/6\}/2tu, \qquad (t < u)$$

$$= \{(u-t)g_2(t+u) + t^2[g_1(t)]^2 - 2L(1+t/u) +$$

$$+ \pi^2/6\}/2tu, \qquad (t > u)$$

$$= (\gamma + \ln t)^2/2t^2, \qquad (t = u)$$

$c > 0$, $t, u > 0$, essendo γ = la costante di Eulero $0.5772156\ldots$, $\pi = 3,1415926\ldots$ e

$$L(1+v/z) = \int_1^{1+v/z} \ln w/(w-1)dw$$

l'integrale di Spence. E' facile notare che il calcolo di queste tre espressioni per tutte le statistiche ordinate di un campione di numerosità n si semplifica se vengono determinati in precedenza i valori di $g_1(c), c = 1, 2, \ldots, n$, di $g_2(c), c = 1, 2, \ldots, 2n$, e, poiché vale la relazione $\phi(t,u) + \phi(u,t) = g_1(t)g_1(u)$, di $\phi(t,u), t = 1, 2, \ldots, n-1, \; u = t, t+1, \ldots, n$.

Come usuale, in SOVEM1 e SOVEC1 le medie e le covarianze sono date da

$$\mu_{i:n} = E(X_{i:n})$$

$i = 1, 2, \ldots, n$, e

$$\begin{aligned} \sigma_{ij:n} &= E(X_{i:n}^2) - \mu_{i:n}^2 && (i = j) \\ &= E(X_{i:n}X_{j:n}) - \mu_{i:n}\mu_{j:n}, && (i < j) \end{aligned}$$

$i = 1, 2, \ldots, n$, $j = i, i+1, \ldots, n$. Ovviamente, $\sigma_{ij:n} = \sigma_{ji:n}$ e $\sigma_{ii:n}$ indica la varianza di $X_{i:n}$.

I valori di $g_1(.), g_2(.)$ e $\phi(.,.)$ vengono calcolati, rispettivamente, mediante le routine ausiliarie VEG1, VEG2 e VEPHI, mentre i coefficienti binomiali e il rapporto tra fattoriali della (3) vengono calcolati a loro volta mediante le routine ausiliarie BINOM e BINOM2. Infine, l'integrale di Spence viene approssimato mediante il noto sviluppo in serie $(v/z)-(v/z)^2/4+\ldots+(-1)^{n+1}(v/z)^n/n^2+\ldots$ con la routine ausiliaria SPENCE.

1.2. Calcolo approssimato

Le medie e le covarianze, rispettivamente, in SOVEM2 e SOVEC2 vengono determinate mediante l'approssimazione di David e Johnson (1954).

In generale, l'approssimazione di David e Johnson ai momenti delle statistiche ordinate di una distribuzione continua X si ottiene mediante uno sviluppo in serie di Taylor di $X_{i:n} = Q_{(i)}[U_{i:n}]$ su $E(U_{i:n}) = i/(n+1), i = 1, 2, \ldots, n$, essendo $U_{i:n}$ la i–ma statistica ordinata in un campione di dimensione n proveniente dalla distribuzione rettangolore $R(0, 1)$ e $Q_{(i)}[.]$ l'inversa della funzione di ripartizione della statistica ordinata $X_{i:n}$ associata a X.

Le seguenti derivate di $Q_{(i)}[U_{i:n}] = -\ln(-\ln(u_{i:n}))$ sono necessarie per ottenere un'approssimazione di ordine $(n+2)^{-3}$ alle medie ed alle covarianze delle statistiche ordinate della distribuzione in discorso:

$$Q_{(i)}^{I}(u) = -1/(u \ln u),$$

$$Q_{(i)}^{II}(u) = (\ln u + 1)/(u \ln u)^2,$$

$$Q_{(i)}^{III}(u) = (-2\ln^2 u - 3\ln u - 2)/(u \ln u)^3,$$

$$Q_{(i)}^{IV}(u) = (6\ln^3 u + 11\ln^2 + 12\ln u + 6)/(u \ln u)^4,$$

$$Q_{(i)}^{V}(u) = (-24\ln^4 u + 50\ln^3 u - 70\ln^2 u - 60\ln u - 24)/(u \ln u)^5,$$

$$Q_{(i)}^{VI}(u) = (120\ln^5 u + 274\ln^4 u + 450\ln^3 u + 510\ln^2 u + 360\ln u + 120)/(u \ln u)^6,$$

$u = u_{i:n}$, determinate nel punto $i/(n+1)$, $i = 1, 2, \ldots, n$. Queste quantità vengono calcolate con la subroutine ausiliaria VEDER, mentre con le subroutine ausiliarie DJMED, DJVAR e DJCOV vengono calcolate rispettivamente, le medie, le varianze e le covarianze approssimate.

2. Struttura

Al fine di ovviare agli inevitabili errori di arrotondamento che si commettono utilizzando le formule (1), (2) e (3), in SOVEM1 e SOVEC1 tutti i calcoli vengono eseguiti in precisione multipla con l'ausilio del pacchetto *MP* scritto in FORTRAN IV da Brent (1978a).

MP lavora con numeri normalizzati in virgola mobile di t cifre e con base b, essendo $t \geq 2$, $b \geq 2$ e $8b^2 - 1$ rappresentabile in una parola di memoria. Questi parametri devono essere passati alle routine che compongono MP mediante il seguente BLANK COMMON:

```
      COMMON B, T, LUN, MXR, R
*     integer B,T,M,LUN,MXR,R(MXR)
```

dove

B base dei numeri in precisione multipla;
T numero di cifre dei numeri in precisione multipla;
M esponente massimo allocabile;
LUN unità logica per i messaggi di errore;
MXR dimensione di R in COMMON;
R vettore di lavoro.

Sia SOVEM1 che SOVEC1 richiamano la routine ausiliaria PARMP nella quale devono essere state inizializzate, al momento della installazione, LUN, MXR e la variabile intera BIT con il numero di bit per parola di memoria; tale routine provvede, quindi, a determinare in esecuzione B e M sulla base del valore di BIT. T, invece, viene reso equivalente ad IC cifre di un numero con base decimale mediante la istruzione FORTRAN

```
T=INT(2E0+FLOAT(IC)+ALOG(10E0)/ALOG(FLOAT(B)))
```

(per maggiori dettagli sulla inizializzazione dei parametri di MP, si veda BRENT, 1978b).

Dal pacchetto MP vengono richiamate le routine MPEUL, MPPI, MPLN, MPADD, MPADDI, MPSUB, MPMUL, MPMULI, MPMULQ, MPDIV, MPDIVI, MPCMPA, MPSTR, MPCIM, MPCMI e MPCMP per approssimare la costante di Eulero, π e i logaritmi naturali, per eseguire le quattro operazioni

fondamentali e le operazioni logiche fra variabili in precisione multipla e fra variabili miste (in precisione singola e multipla) e per convertire le variabili dalla singola alla multipla precisione e viceversa.

Allora, le specificazioni delle routine SOVEM1 e SOVEC1 sono:

```
      CALL SOVEM1(N,VM,IC,W,NW,IER)
*     integer N, IC, W, NW, IER
*     real VM

      CALL SOVEC1(N,V,VM,IC1,IC2,W,NW,IER)
*     integer N, IC1, IC2, W, NW, IER
*     real VM, V
```

con il seguente significato degli argomenti:

N in input, numerosità campionaria , N$\geq$1;

VM vettore di dimensione N contenente le medie, in output per SOVEM1 e in input per SOVEC1; la media della i–ma statistica ordinata viene memorizzata nella posizione $I, I = 1, 2, \ldots, N$;

V in output, vettore di dimensione $N*(N+1)/2$ contenente le covarianze; la covarianza fra le statistiche ordinate i–ma e j–ma viene memorizzata nella posizione $(J*(J-1)/2+I)$, $I = 1, 2, \ldots, N$, $J = I, I+1, \ldots, N$;

IC, IC1 in input, numero di cifre decimali equivalenti a T cifre in precisione multipla, IC, IC1$\geq$0; in output, numero di cifre corrette ottenute nel controllo della formula (1) in SOVEM1 e della formula (2) in SOVEC1 rispetto al numero di cifre decimali date in input (si veda il prossimo paragrafo);

IC2 in output, numero di cifre corrette ottenute nel controllo della formula (3) in SOVEC1 rispetto al numero di cifre decimali IC1 date in input (si veda il prossimo paragrafo);

W vettore di lavoro di dimensione almeno pari a $(T+2)*(N+6)$ per SOVEM1 e $(T+2)*(N*(N+5)/2+7)$ per SOVEC1 utilizzato per memorizzare le variabili in precisione multipla e i valori delle funzioni $g_1(.), g_2(.), \phi(.,.)$ (si veda la sezione 2.1.);

NW in input, dimensione di W nell'unità chiamante;

IER in output, condizione di errore:

1. N minore di uno;
2. IC o IC1 minore di zero;
3. dimensione di W insufficiente;
4. risultati finali non attendibili (si veda il prossimo paragafo);
0. altrimenti;

una segnalazione di errore si può avere da MP se il dimensionamento del vettore di lavoro R è insufficiente per eseguire i calcoli in precisione multipla con il numero di cifre decimali richiesto).

Le specificazioni di SOVEM2 e SOVEC2, invece sono:

```
      CALL SOVEM2(N,VM,IER)
*     integer N, IER
*     real VM

      CALL SOVEC2(N,V,IER)
*     integer N, IER
*     real V
```

con lo stesso significato degli argomenti visto per SOVEM1 e SOVEC1. Una condizione di errore, IER=1, si verifica solamente quando N è minore di uno.

3. Accuratezza

L'accuratezza dei calcoli nelle quattro routine viene controllata mediante le note relazioni

$$\sum_{i=1}^{n} E(X_{i:n}) = n\gamma,$$

$$\sum_{i=1}^{n} E(X_{i:n}^2) = n(\pi^2/6 + \gamma),$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} E(X_{i:n}X_{j:n}) = n(n-1)\gamma^2/2,$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{ij:n} = n\pi^2/6,$$

con il significato dei simboli visto in precedenza. In particolare, le prime tre relazioni vengono utilizzate in SOVEM1 e SOVEC1 per controllare l'accuratezza dei risultati finali in precisione multipla delle formule (1), (2) e (3) rispetto al numero di cifre decimali date in input, mentre in SOVEM2 e SOVEC2, mediante un fattore di correzione, viene imposto alle medie ed alle covarianze approssimate di soddisfare alla prima ed alla quarta relazione. Da notare che in queste due ultime routine la media e la varianza della n–ma statistica ordinata, rispettivamente, pari a $\gamma + \ln n$ e $\pi^2/6$, vengono calcolate esattamente.

Come si è detto in precedenza, in SOVEM1 e SOVEC1 i calcoli sono eseguiti in precisione multipla, mentre l'output è in precisione singola; allora, al fine di ottenere dei risultati finali accurati è sufficiente che, rispetto al numero di cifre decimali dato in input alle due routine per un prefissata numerosità campionaria, le relazioni riportate sopra siano soddisfatte con un numero di cifre tale da garantire che il numero di cifre significative dei valori delle medie

e delle covarianze ottenuto nella conversione dalla multipla alla singola precisione, sia il massimo rappresentabile in una parola di memoria dell'elaboratore sul quale le due routine vengono eseguite (per esempio, 7 cifre sugli elaboratori a 32 bit e 15 su quelli a 64 bit).

Qui di seguito si riporta, per alcune numerosità campionarie , il numero di cifre decimali da dare in input a SOVEM1 e SOVEC1 per ottenere dei risultati finali con 7 cifre significative su un elaboratore a 32 bit, unitamente all'errore massimo relativo in valore assoluto che si commette calcolando le medie e le covarianze con l'approssimazione di David e Johnson in luogo delle formule esatte.

N	Numero di cifre decimali		Errore massimo relativo	
	SOVEM1	SOVEC1	SOVEM2	SOVEC2
10	9	9	0,0023	0,0065
20	13	13	0,0035	0,0032
30	17	22	0,0032	0,0018
40	22	26	0,0028	0,0013
50	26	30	0,0024	0,0012
75	38	47	0,0018	0,0011
100	51	68	0,0014	0,0010

E' da osservare che l'utilizzo in SOVEM1 e SOVEC1 di un numero di cifre troppo basso rispetto alla numerosità campionaria, può produrre condizioni di overflow o valori abnormi nei risultati finali; questi ultimi vengono segnalati con la condizione di errore IER=3, in SOVEM1 se le medie non sono ordinate o non risultano comprese nell'intervallo (David, 1981; pag. 59)

$$\gamma - \frac{(n-1)\pi}{[3(2n-1)]^{\frac{1}{2}}} \leq \mu_{i:n} \leq \gamma + \frac{(n-1)\pi}{[3(2n-1)]^{\frac{1}{2}}},$$

$i = 1, 2, \ldots, n$, e in SOVEC1 se le covarianze risultano negative.

4. Tempi di esecuzione

I tempi di esecuzione di SOVEM2 e SOVEC2 sono molto bassi e le due routine possono essere utilizzate anche su microelaboratori. Non altrettanto si può dire di SOVEM1 e SOVEC1 i cui tempi di esecuzione aumentano considerevolmente al crescere della numerosità campionaria. A questo proposito, il tempo di esecuzione in secondi di SOVEM1 e SOVEC1 per ottenere le medie e le covarianze con una accuratezza di 7 cifre sul VAX 8600 del Centro di Calcolo d'Ateneo dell'Università di Padova sono stati i seguenti:

N	SOVEM1	SOVEC1
10	0,26	3,36
20	0,70	23,07
30	1,37	95,39
40	2,54	240,12
50	4,06	532,94
75	11,63	2726,44
100	21,51	9843,15

5. Esempio

In questo esempio vengono calcolate le medie e le covarianze esatte ed approssimate della distribuzione valori estremi (I tipo) per N=6. I risultati finali possono essere confrontati con i valori tabulati da Lieblein e Zelen (1956).

Riferimenti bibliografici

Brent R.P., 1978a, A Fortran multiple–precisione arithmetic package, *ACM Transactions on Mathematical Software*, **4**, 1, 57–70
Brent R.P., 1978b, MP, a Fortran multiple–precisione arithmetic package, Algorithm 524, *ACM Transactions on Mathematical Software*, **4**, 1, 71–81
David F.N. e Johnson N.L., 1954, Statistical treatment of censored data. I. Fundamental formulae, *Biometrika*, **41**, 228–240
David H.A., 1981, *Order statistics, 2nd edition*, Wiley, New York
Lieblein J., 1953, On the exact evolution of the variances and covariances of order statistics in samples from the extreme–value distribution, *Annals of Mathematical Statistics*, **24**, 282–287
Lieblein J. e Zelen M., 1956, Statistical investigation of the fatigue life of deep–groove ball bearings, *Journal of Research of the National Bureau of Standards*, **57**, 273–316
Provasi C., 1984, Un test funzionale per variabili casuali standardizzabili, *Statistica*, **4**, 687–698

Summary

Exact and approximate means and covariances of order statistics of the standardized extreme value distribution (I type)

Some FORTRAN 77 routines are presented for computing exact and approximate means and covariances of the standardized extreme value distribution (I type). Firstly, the algorithm features are described. Secondly, the routine structure, accuracy and timing are quoted. A numerical example illustrates the usage of the routines.

Input:

```
      REAL VM1(6),V1(21),VM2(6),V2(21)
      INTEGER N,IC,IC1,IC2,NW,W(240)

      N   = 6
      IC  = 9
      IC1 = 9
      NW  = 240
      CALL SOVEM1(N,VM1,IC,W,NW,IER)
      CALL SOVEC1(N,V1,VM1,IC1,IC2,W,NW,IER)
      CALL SOVEM2(N,VM2,IER)
      CALL SOVEC2(N,V2,IER)
      .
      .
      .
      END
```

Output:

```
IC  =    9
IC1 =    9
IC2 =    9
IER =    0
```

$$VM1 = \begin{bmatrix} -0.777294 \\ -0.254534 \\ 0.188385 \\ 0.662716 \\ 1.275046 \\ 2.368975 \end{bmatrix} \qquad VM2 = \begin{bmatrix} -0.780409 \\ -0.256409 \\ 0.189155 \\ 0.665532 \\ 1.276451 \\ 2.368975 \end{bmatrix}$$

$$V1 = \begin{bmatrix} 0.246582 & 0.154967 & 0.121216 & 0.102915 & 0.091162 & 0.082854 \\ & 0.248546 & 0.196706 & 0.168065 & 0.149453 & 0.136191 \\ & & 0.297616 & 0.256165 & 0.228879 & 0.209255 \\ & & & 0.401855 & 0.361456 & 0.332045 \\ & & & & 0.647700 & 0.599857 \\ & & & & & 1.644934 \end{bmatrix}$$

$$V1 = \begin{bmatrix} 0.250587 & 0.156086 & 0.121964 & 0.103505 & 0.091657 & 0.083285 \\ & 0.250035 & 0.197609 & 0.168748 & 0.150000 & 0.136640 \\ & & 0.298433 & 0.256642 & 0.229178 & 0.209424 \\ & & & 0.401428 & 0.360730 & 0.331146 \\ & & & & 0.642450 & 0.594256 \\ & & & & & 1.644934 \end{bmatrix}$$