REGIONE PUGLIA

ASSESSORATO AGRICOLTURA, ACQUACOLTURA, ALIMENTAZIONE, CACCIA E PESCA

PROGETTO

“SISTEMA INFORMATIVO CONGIUNTURALE - SIC”
P.O.R. PUGLIA 2000-2006 – SFOP
Asse IV – Misura 4.13 – D2

BENEFICIARIO

ASSOCIAZIONE ARMATORI DA PESCA MOLFETTA

ALLEGATO 2: METODOLOGIA DISEGNO DI CAMPIONAMENTO

IREPA ONLUS
ISTITUTO RICERCHE ECONOMICHE PER LA PESCA E L’ACQUACOLTURA

“Convenzione del 05.01.2007”

ALLEGATO 2: METODOLOGIA

La numerosità campionaria ottima per strato è stata definita in base alla procedura di Bethel (1989), l'estrazione dei battelli è basata sulla metodologia PPS (Probability Proportional to Size) e, più precisamente, mediante l'algoritmo di Hanurav-Vijayan; per la stima dei totali per strato, ossia per la fase di espansione, si è fatto ricorso allo stimatore di Horvitz – Thompson, mentre per la stima dei relativi errori campionari, alla formula di Sen-Yates-Grundy. Infine, questa fase di stima o riporto all'universo è stata preceduta da un'insieme di procedure di controllo e correzione dei dati campionari allo scopo di garantire risultati finali con livelli di qualità predeterminati.

Stima della numerosità campionaria: la procedura di Bethel

Il disegno di campionamento finalizzato all'individuazione della numerosità ottima del campione in una indagine campionaria multivariata, cioè con la rilevazione di più variabili obiettivo, nonché stratificata, è stato basato sull'applicazione della procedura di Bethel (1989).
Tale procedura è un algoritmo matematico il cui obiettivo è l'individuazione del campione di "costo minimo", dati i vincoli di precisione richiesta per ogni strato. Il costo C è definito come:

$$C = c_0 + \sum_{h=1}^{H} c_h n_h$$

dove c_0 rappresenta un costo fisso correlato con l'organizzazione della rilevazione, i c_h rappresentano i costi di campionamento di un'unità all'interno dello strato h-mo (h=1 … H), mentre n_h (n_h=1 … N_h) rappresenta il numero di unità estratte all'interno dell'h-mo strato.

Considerato che il campionamento è di tipo stratificato, i vincoli di precisione sulla stima sono esprimibili nel seguente modo[1]:

$$\mathrm{var}\left(\hat{Y}_j\right) = \sum_{h=1}^{H} N_h^{\,2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_{hj}^{\,2}}{n_h} \le \tilde{v}_j^{\,2} \qquad \text{j=1 … J} \qquad (1)$$

dove $\hat{Y}_j$ rappresenta la stima del totale per la j-ma variabile (j=1 … J), S_{hj}^2 rappresenta una stima (o un valore ipotetico) della varianza della j-ma variabile all'interno dell'h-mo strato e $\tilde{v}_j^{\,2}$ rappresenta il valore di soglia (il vincolo), in termini assoluti, per il valore della varianza della stima del totale per la j-ma variabile.
Questo insieme di J vincoli può essere equivalentemente espresso in una forma alternativa:

$$\sum_{h=1}^{H} N_h^{\,2} \frac{S_{hj}^{\,2}}{n_h} \le \tilde{v}_j^{\,2} + \sum_{h=1}^{H} N_h S_{hj}^{\,2} \Leftrightarrow \frac{\sum_{h=1}^{H} N_h^{\,2} \frac{S_{hj}^{\,2}}{n_h}}{\tilde{v}_j^{\,2} + \sum_{h=1}^{H} N_h S_{hj}^{\,2}} \le 1 \Leftrightarrow \sum_{h=1}^{H} \left(\frac{N_h^{\,2} S_{hj}^{\,2}}{\varepsilon_j^{\,2} Y_j^{\,2} + \sum_{h'=1}^{H} N_{h'} S_{h'j}^{\,2}} \cdot \frac{1}{n_h} \right) \le 1$$

dove Y_j rappresenta il totale stimato (o ipotizzato) per la variabile j-ma, e ε_j rappresenta l'errore relativo (errore assoluto della stima diviso per il valore della stima) ammesso per la j-ma variabile.

[1] Nell'articolo originario di Bethel (1989), non si considerava la correzione per popolazioni finite, pertanto (tenute in conto le differenze dovute al fatto che nell'articolo le quantità da stimare erano le medie e non i totali) la formula (*) in realtà si presentava come:

$$\mathrm{var}\left(\hat{Y}_j\right) = \sum_{h=1}^{H} N_h^{\,2} \frac{S_{hj}^{\,2}}{n_h} \le \tilde{v}_j^{\,2}$$

Indicando quindi con a_{hj} il termine a sinistra del prodotto tra parentesi dell'ultima disuguaglianza, e con x_h il valore $1/n_h$, tutta l'ultima disuguaglianza può essere espressa nella forma:
$a_j'x \leq 1 \quad j=1 \ldots J$
o, equivalentemente,
$A'x \leq 1$
dove $\underset{(H \times J)}{A} = [a_{hj}]$ e $\underset{(H \times 1)}{\underline{x}}$ è il vettore dei valori $1/n_h$.

Tutto il problema di minimo vincolato può, quindi, essere espresso nella seguente maniera:

$$\begin{cases} \min_{\underline{x}} g(\underline{x}) = \sum_{h=1}^{H} \frac{c_h}{x_h} \\ s.v.: A'\underline{x} \leq \underline{1} \end{cases}$$

Bethel dimostrò che questo problema ha sempre una soluzione, e che questa corrisponde alla seguente formula:

$$x_h^* = \frac{\sqrt{c_h}}{\sqrt{\sum_{j=1}^{J} \alpha_j^* a_{hj}} \cdot \sum_{k=1}^{H} \sqrt{c_k \sum_{j=1}^{J} \alpha_j^* \cdot a_{kj}}}$$

dove gli α_j^* sono opportune costanti ("moltiplicatori di Lagrange") normalizzate, cioè tali per cui $\sum_{j=1}^{J} \alpha_j^* = 1$.

Per la soluzione di tale problema di minimo vincolato, Bethel propose l'uso di un algoritmo non particolarmente efficiente e nemmeno di semplice implementazione. In realtà già all'epoca era disponibile un algoritmo, formulato da Chromy (1987) e proposto anche nella pubblicazione stessa di Bethel, che permetteva di trovare la soluzione in maniera più semplice, dal punto di vista dello sviluppo del codice, e più rapida, dal punto di vista dei tempi di elaborazione.

Una volta impostati i valori iniziali degli α_j pari a 1/J, tale algoritmo si sviluppa fondamentalmente su due passi, da ripetere iterativamente fino a raggiungere un opportuno criterio di convergenza:

calcolare:
$$x_h\left(\underline{\alpha}^{(r-1)}\right) = \frac{\sqrt{c_h}}{\sqrt{\sum_{j=1}^{J} \alpha_j^{(r-1)} a_{hj}} \cdot \sum_{k=1}^{H} \sqrt{c_k \sum_{j=1}^{J} \alpha_j^{(r-1)} \cdot a_{kj}}}$$

calcolare:
$$\alpha_j^{(r)} = \frac{\alpha_j^{(r-1)} \left[a_j' \underline{x}\left(\underline{\alpha}^{(r-1)}\right)\right]^2}{\sum_{k=1}^{J} \alpha_k^{(r-1)} \left[a_k' \underline{x}\left(\underline{\alpha}^{(r-1)}\right)\right]^2}$$

Nell'intento di applicare la procedura, seguendo lo sviluppo degli algoritmi sopra descritti, è stato necessario considerare, quali dati di input per l'avvio della procedura, le stime della varianza per singolo strato delle variabili oggetto di indagine e le stime dei totali; in genere, tali stime vengono ricavate dai dati disponibili, al momento dell'analisi, per l'anno più recente.

Per il campione dell'anno 2007 è stato opportuno considerare le catture e i ricavi delle 12 specie ittiche maggiormente rappresentative nell'ambito della produzione pugliese di cui costituiscono oltre il 75%: Alici, Sardine, Naselli, Triglie di Fango, Triglie di scoglio, Altri Pesci, Vongole, Seppie, Polpi, Moscardini, Gamberi bianchi, Scampi e Pannocchie.

L'estrazione di battelli campionari: l'algoritmo di Hanurav-Vijayan

Il disegno di campionamento adottato ha previsto l'estrazione, senza ripetizione, delle unità campionarie in base alla metodica PPS (probability proportional to size); più semplicemente, in tale piano di campionamento si è provveduto ad estrarre le varie unità con probabilità di inclusione del primo ordine non costante, ma proporzionali ad una variabile ausiliaria opportunamente scelta. L'utilizzo di un tale piano di campionamento e, quindi, il suo utilizzo in luogo di un campionamento casuale semplice, è giustificato dall'intenzione di volere sfruttare l'informazione fornita dalla variabile ausiliaria. Tale variabile ausiliaria, ovviamente, dovrà essere nota per tutte le unità dell'universo di riferimento e dovrà essere "legata" alla variabile ignota che si cerca di stimare. Tale legame, in termini statistici, si traduce in "relazione di proporzionalità" tra variabile da stimare e variabile ausiliaria nota.
L'utilizzo dell'informazione fornita dalla variabile ausiliaria è stata finalizzata al miglioramento della stima; detto diversamente, quanto più "forte" sarà questa relazione di proporzionalità, tanto più piccola sarà la variabilità dello stimatore (ossia la varianza), e tanto più precisa sarà la stima. Nella situazione limite teorica di esatta proporzionalità, lo stimatore avrebbe varianza nulla e ad assumerebbe, in qualsiasi campione, l'esatto totale da stimare.
Nel caso in questione, la variabile ausiliaria nota è la LFT, il cui utilizzo come variabile accessoria è stato preceduto da un ulteriore analisi esplorativa per validare l'ipotesi di proporzionalità tra LFT da un lato e quantitativo pescato e ricavi dall'altro (ovviamente, non si parla di relazione "esatta" tra le variabili).
L'algoritmo di Hanurav-Vijayan definisce una serie di passaggi per effettuare l'estrazione di un campione di numerosità prefissata (n), senza reinserimento, e con probabilità di inclusione nel campione per le singole unità non uniformi.
Seguendo il suddetto algoritmo si ottiene un campione che gode di una serie di proprietà, alcune delle quali apprezzabili:

- $\pi_i = n X_i / X$, dove π_i rappresenta la probabilità di inclusione nel campione (detta anche probabilità di inclusione del primo ordine) dell'i-ma unità, n indica la dimensione prefissata del campione, X_i rappresenta la dimensione della variabile (o misura "accessoria") nota, da cui si ricava il valore della probabilità di inclusione, e X è la somma dei valori X_i per i=1… N; dove con N viene indicata la dimensione dell'universo da cui si campiona. Questa identità è "richiesta per costruzione" e richiede alcuni trattamenti particolari in determinate circostanze (si veda più avanti).
- $\pi_{ij}>0$, dove π_{ij} rappresenta la probabilità (detta "di secondo ordine") di contemporanea presenza delle unità i e j nel campione. Il fatto stesso di poter determinare in maniera esatta e relativamente semplice tali probabilità, conseguenza della procedura di campionamento, è già un risultato notevole che assicura l'esistenza di uno stimatore non distorto per la varianza.
- $\pi_{ij} \leq \pi_i \pi_j$. Questa caratteristica è apprezzabile perché garantisce la positività dello stimatore della varianza del totale di Sean-Yates-Grundy.
- $\pi_{ij} - \pi_i \pi_j > \beta$, per β non troppo prossimo a 0. Questa proprietà garantisce la stabilità dello stimatore di Sean-Yates-Grundy.

I valori π_i e π_{ij} (per i,j=1 … N) soddisfano le seguenti due identità:

$$\sum_{i=1}^{N} \pi_i = n$$

$$\sum_{i=1}^{N} \sum_{j>i}^{N} \pi_{ij} = \frac{n(n-1)}{2}$$

È interessante osservare come la somma delle probabilità di primo ordine non sia mai pari a 1 (a meno che il campione sia composto di una sola unità). Stessa cosa dicasi per le probabilità del secondo ordine (a meno che il campione sia composto di 2 unità).
C'è inoltre da osservare come l'applicazione della formula (i) possa portare talvolta a probabilità di inclusione del primo ordine maggiori di 1. In tal caso sono state apportare delle correzioni alla procedura di estrazione e alle probabilità di inclusione. Nello specifico, sono state assegnate probabilità di inclusione del primo ordine pari a 1 alle k unità la cui probabilità di inclusione risultasse superiore a 1 e sono state estratte n-k unità all'interno della popolazione complessiva, una volta escluse le unità con probabilità 1. È chiaro che, laddove "accantonate" (o meglio, estratte con probabilità 1) le unità con probabilità maggiore di 1, se sono comparirse altre con probabilità di inclusione maggiore di 1 all'interno delle N-k rimanenti, si è provveduto al graduale "accantonamento" anche di queste ultime e di tutte le altre unità fino ad ottenere una popolazione di unità da estrarre casualmente aventi tutte probabilità del primo ordine inferiori a 1.