

Statistica degli estremi

Richiami di probabilità e statistica

Il calcolo della probabilità di un evento è direttamente connesso con:

- la CONOSCENZA INCOMPLETA dell'evento stesso;
- l'assunzione di un RISCHIO, calcolato come la probabilità che un evento accada.

Esistono due modi fondamentali di utilizzare la teoria della probabilità applicata alla predizione di eventi:

- la via **PROBABILISTICA**, ossia il derivare le proprietà probabilistiche di un processo per via assiomatica;
- la via **STATISTICA**, ossia il derivare le proprietà probabilistiche di un processo dall'analisi del campione.

Diciamo CDF (FUNZIONE CUMULATIVA DI PROBABILITÀ) la funzione adimensionale, che prende valori compresi tra 0 e 1, definita come:

$$F_X(x) = P(X \leq x)$$

dove $P(X \leq x)$ sta per "probabilità che X sia minore o uguale a x "

X è detta variabile aleatoria

x_i per $i = 1, \dots, n$ è il campione di n misure

Diciamo poi PDF (DENSITÀ DI PROBABILITÀ) la funzione:

$$f_X(x) = \frac{dF_X}{dx} = \lim_{\Delta x \rightarrow 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x}$$

Supponendo di conoscere la probabilità del verificarsi di un evento A , la struttura di probabilità può cambiare se consideriamo un evento B che condiziona l'evento A : parleremo in tal caso di **PROBABILITÀ CONDIZIONATA** ed abbiamo:

$$P(A|B) = \frac{P(A \cdot B)}{P(B)}$$

Nel solo caso in cui i due eventi siano indipendenti abbiamo:

$$P(A|B) = P(A)$$

Dato uno spazio di probabilità $S = \bigcup_{i=1}^n B_i$ dove B_1, \dots, B_n è una partizione di S tale che

$B_i \cdot B_j = \emptyset$ per $B_i \neq B_j$, abbiamo il **TEOREMA DELLA PROBABILITÀ TOTALE**:

$$P(A) = \sum_{i=1}^n P(A \cdot B_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

Sempre nelle stesse ipotesi, abbiamo anche il TEOREMA DI BAYES:

$$P(B_i | A) = \frac{P(A | B_i) P(B_i)}{\sum_{j=1}^n P(A | B_j) P(B_j)}$$

importante per quanto riguarda il calcolo della probabilità A POSTERIORI: si è già ottenuto un certo risultato per A e si calcola la probabilità che esso derivi da una certa configurazione tra quelle possibili.

Un altro importante teorema è quello del LIMITE CENTRALE: supponiamo di avere una

collezione di variabili aleatorie X_i e cerchiamo una variabile $Y = \sum_{i=1}^n X_i$; possiamo allora

dire che, se $N \rightarrow \infty \forall f_{X_i}(x)$, allora la $f_Y(y)$ segue una distribuzione gaussiana. Allo stesso

modo, ipotizzando un modello moltiplicativo $Y = \prod_{i=1}^N X_i$, possiamo dire che se $N \rightarrow \infty$

$\forall f_{X_i}(x)$, allora la $f_Y(y)$ segue una distribuzione log-normale, poiché

$$\ln Y = \sum_{i=1}^N \ln X_i$$

Distribuzioni comunemente utilizzate

Le funzioni di probabilità discrete più utilizzate nello studio delle piene sono:

- la distribuzione di BERNOULLI, che calcola la probabilità delle alternative di successo (0) o insuccesso (1) ed è definita come:

$$P(1) = p$$

e quindi come:

$$P(0) = 1 - p$$

- la distribuzione BINOMIALE, che calcola la probabilità di avere m successi in un esperimento dato da una successione di n tentativi:

$$P(m) = \binom{n}{m} p^m (1-p)^{n-m}$$

con media $E(m) = np$ e varianza $\sigma^2(m) = np(1-p)$;

- la distribuzione GEOMETRICA, utile per calcolare il tempo di attesa per il primo successo e definita come:

$$P(W) = (1-p)^{W-1} p$$

dove W è il tempo di attesa per il primo successo

con media $E(W) = \frac{1}{p}$ e varianza $\sigma^2(W) = \frac{1-p}{p^2}$;

- la distribuzione di POISSON, definita a partire dalla distribuzione geometrica come:

$$P(m) = \lambda^m \frac{\exp(-\lambda)}{m!}$$

dove m è il numero dei successi

$\lambda = pn$ è l'intensità del successo

n è il numero dei tentativi

con media $E(W) = \lambda$ e varianza $\sigma^2(W) = \lambda$; tale distribuzione parte dalla considerazione che, quando il numero dei tentativi diventa molto grande ($n \rightarrow \infty$) la probabilità che si verifichi il singolo evento tende a 0;

Le funzioni di probabilità continue più utilizzate sono invece:

- la distribuzione NORMALE o di GAUSS, definita come:

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x-\bar{x}}{\sigma_x}\right)^2\right)$$

- la distribuzione LOG-NORMALE, che consiste in una distribuzione normale dove la variabile è $y = \ln x$;
- la distribuzione ESPONENZIALE, utile per il calcolo del tempo di attesa del primo successo, definita come:

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - \exp\left(-\frac{t}{w}\right)$$

con valore medio $E(t) = w$;

- la distribuzione GAMMA INCOMPLETA o TERZA DI PEARSON, utile per il calcolo del tempo di attesa del k -esimo successo, definita come:

$$f_{T_k}(t) = \frac{1}{\alpha(k-1)!} \left(\frac{t}{\alpha}\right)^{k-1} \exp\left(-\frac{t}{\alpha}\right)$$

$$F_{T_k}(t) = \int_0^t f_{T_k}(\tau) d\tau$$

con media $E(t) = \alpha k$ e varianza $\sigma^2(t) = \alpha^2 k$.

Va fatta una importante osservazione per quanto riguarda quest'ultima distribuzione: essa ha la stessa forma dell'idrogramma unitario di Nash. L'IUH, infatti, che ha spiegazione fisica come efflusso, può essere interpretato come la probabilità che una particella, precipitata all'istante iniziale in un punto qualsiasi del bacino, attraversi la sezione di chiusura al tempo t .

Va inoltre ricordato che il calcolo dei fattoriali può essere effettuato mediante la funzione gamma completa:

$$\Gamma(k) = \begin{cases} (k-1)! & \text{per } k \in \mathbb{Z} \\ \int_0^\infty u^{k-1} \exp(-u) dt & \text{per } k \in \mathbb{R} \end{cases}$$

Distribuzioni EV

Una tipologia di distribuzioni statistiche di grande interesse per l'analisi di fenomeni rari come le piene sono le distribuzioni del VALORE ESTREMO o EV (extreme value), studiate prevalentemente da Gumbel.

Consideriamo un orizzonte temporale, ovvero un numero di tentativi, sufficientemente grande da permettere la ricerca di un valore massimo dei risultati sui tentativi. Ad esempio cerchiamo la massima tra le piogge in un'ora all'interno di un orizzonte temporale di un anno. Tale massimo raro Z_{\max} sarà una variabile aleatoria che può seguire 3 distribuzioni asintotiche:

- la distribuzione di Gumbel o EV1;

- la distribuzione di Fréchet o EV2;
- la distribuzione di Weibull o EV3.

La distribuzione di Gumbel è definita come:

$$F_{Z_{\max}}(z) = \exp(-\exp(-z))$$

con media $E(z) = 0.5772$ pari al numero di Eulero e varianza $\sigma^2(z) = \frac{\pi^2}{6}$.

Presentazione dei dati statistici

Dati i risultati x_i per $i = 1, \dots, n$ di una sequenza di esperimenti casuali tra loro indipendenti, costruiamo la CDF mediante le operazioni:

1. ordinamento degli x_i in ordine crescente ($X_i = \min(x_i)$, $X_n = \max(x_i)$, ecc);
2. assegnazione della frequenza assoluta di superamento ($F_{\text{ass}} = i$);
3. assegnazione della frequenza relativa di superamento
 - secondo la formula $F = \frac{i}{n}$;
 - secondo la formula di Hazen $F = \frac{i-1}{n}$;
 - secondo la formula di Weibull $F = \frac{i}{n+1}$;
4. adattamento (fitting) di una distribuzione teorica alla frequenza empirica.

Stima dei parametri

Il problema dell'applicazione di distribuzioni di probabilità a campioni statistici finiti è superato dallo stabilire una relazione tra le variabili standardizzate z e le variabili osservate x , generalmente nella forma:

$$x = az + b$$

Nel caso della distribuzione di Gumbel, ad esempio, otteniamo la distribuzione:

$$F_X(x) = \exp\left(-\exp\left(-\frac{x-b}{a}\right)\right)$$

$$f_X(x) = \frac{1}{b} \exp\left(-\frac{x-b}{a}\right) \exp\left(-\exp\left(-\frac{x-b}{a}\right)\right)$$

I parametri a e b non sono noti a priori, e bisogna quindi procedere mediante alcuni metodi di stima delle distribuzioni. I tre metodi proposti nel seguito danno luogo a tre risultati differenti, ma non si può affermare che uno sia più o meno preciso dell'altro.

Metodo dei momenti

Tale metodo di stima parte dalla considerazione che le distribuzioni di probabilità hanno dei momenti statistici ben definiti, ossia la media

$$\langle x \rangle = \int_{-\infty}^{+\infty} x f_X(x) dx,$$

che rappresenta il baricentro della funzione, e la varianza

$$\sigma_x^2 = \int_{-\infty}^{+\infty} [x - \langle x \rangle]^2 f_X(x) dx,$$

che ne rappresenta invece il momento di inerzia centrale.

Il metodo di stima consiste:

1. nel calcolare i momenti del campione x_i per $i = 1, \dots, N$:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

2. nell'uguagliarli ai momenti della distribuzione:

$$x = \hat{a}z + \hat{b} \Rightarrow \bar{x} = \hat{a}\bar{z} + \hat{b}$$

$$\hat{a}z = x - \hat{b} \Rightarrow \sigma_x^2 = \hat{a}^2 \sigma_z^2$$

In tal modo avremo le espressioni:

$$\hat{a} = \frac{\sqrt{6}}{\pi} \sigma_x$$

$$\hat{b} = \bar{x} - 0.5572\hat{a}$$

Metodo grafico

In tale caso si ipotizza che la relazione tra le variabili standardizzate ed osservate sia lineare. Innanzitutto si definisce, per le frequenze osservate ordinate in ordine crescente, la posizione grafica; ad esempio definiamo quella di WEIBULL come:

$$F_i = P(X \leq x_i) \cong \frac{i}{N+1}$$

In tal modo abbiamo un'espressione per le frequenze osservate pari a:

$$z_i = -\ln\left(\ln\frac{N+1}{i}\right).$$

Grazie alla velocità di elaborazione dei fogli elettronici commerciali, possiamo poi effettuare sulle frequenze una regressione lineare, utilizzando il METODO DEI MINIMI QUADRATI. Con questa procedura la retta interpolatrice risente molto dei valori estremi, che poi sono quelli d'interesse per l'analisi delle piene, ma si ha una grossa variabilità in funzione della variabilità del campione.

Metodo della massima verosimiglianza

Con tale metodo si definisce la FUNZIONE DI VEROSIMIGLIANZA (likelihood) $L(x_i, a, b)$ in funzione della distribuzione standardizzata scelta $f_X(x, a, b)$ come:

$$L = \prod_{i=1}^N f_X(x_i, a, b) = f_X(x_1, a, b) \cdot \dots \cdot f_X(x_N, a, b)$$

I parametri a e b sono poi scelti in modo da massimizzare la funzione di verosimiglianza. Nel caso della distribuzione di Gumbel, potendo affermare che

$$\ln L = \sum_{i=1}^N \ln f_X(x)$$

cercheremo i valori massimi con le espressioni:

$$\frac{\partial \ln L(a, b)}{\partial a} = Na - \sum_{i=1}^N (x_i - b) + \sum_{i=1}^N (x - b) \exp\left(-\frac{x_i - b}{a}\right) = 0$$

$$\frac{\partial \ln L(a,b)}{\partial b} = Na - \sum_{i=1}^N \frac{1}{a} \exp\left(-\frac{x_i - b}{a}\right) = 0$$

Semplificando il sistema, possiamo iterare fino alla convergenza l'espressione

$$a = \bar{x} - \frac{\sum_{i=1}^N x_i \exp\left(-\frac{x_i}{a}\right)}{\sum_{i=1}^N \exp\left(-\frac{x_i}{a}\right)}$$

ed inserirla nella

$$b = a \left(-\ln \left(\frac{1}{N} \sum_{i=1}^N \exp\left(-\frac{x_i}{a}\right) \right) \right)$$

Outliers

Al fine di operare una scelta cosciente del metodo di stima, sul campione va effettuata anche una verifica dell'esistenza di valori molto rari, detti outliers, che si discostano in maniera significativa dall'andamento medio del campione e che stanno nelle code della distribuzione.

Il metodo più utilizzato per verificare la presenza di outliers nella distribuzione consiste nella verifica della disequazione:

$$\eta = \frac{\max(x_i)}{\text{mediana}(x_i)} > 3$$

Un metodo per tener conto della presenza, e quindi della dispersione, di questi valori nella distribuzione di Gumbel è quello di aumentare il numero dei parametri, e quindi il numero dei gradi di libertà. Con questo passaggio nascono dei problemi per la stima: per quanto riguarda il metodo dei momenti, basta calcolare anche i MOMENTI DI ORDINE MAGGIORE AL II della distribuzione ed eguagliarli a quelli del campione:

- il coefficiente di asimmetria (momento del III ordine);
- il momento centrale del IV ordine $m_4 = \int_{-\infty}^{+\infty} (x - \bar{x})^4 f_X(x) dx$, che dà origine al coefficiente di appiattimento di Kurtosis¹ pari a $\frac{m_4}{\sigma_x^4}$.

Statistiche regionali

Con il crescere dei parametri cala però la loro affidabilità, in quanto i parametri in aggiunta ai primi due momenti hanno una grande varianza di stima. In altre parole il significato del valore trovato è molto povero: cambiando il campione, otteniamo infatti valori molto diversi per i parametri. Un metodo per superare questo problema è quello di considerare le proprietà congiunte, ossia di passare dalle statistiche sulle stazioni indipendenti alle STATISTICHE REGIONALI, combinando le osservazioni di diverse stazioni di misura. In tal modo si ottengono campioni più numerosi e le stime dei parametri relativi ai momenti maggiori del II diventano più attendibili.

Mettendo insieme i dati delle stazioni regionali possiamo così ottenere anche una statistica degli outliers:

¹ Per la distribuzione di Gauss abbiamo asimmetria nulla e coefficiente di appiattimento pari a 3; per un coefficiente minore di 3 le distribuzioni sono meno disperse (più appuntite) al centro, per un coefficiente maggiore di 3 le distribuzioni sono più disperse (più appiattite) al centro.

$$\eta_{reg} = \frac{x_{locali}}{\bar{x}_{locale}}$$

e costruire così delle curve di crescita regionale su di un grafico di Gumbel: effettuiamo in questo modo il Flood Study Report.

TCEV

L'approccio italiano del Flood Study Report consiste nel metodo TCEV (two components extreme value). Ipotizziamo che la distribuzione di probabilità derivi dall'accoppiamento di 2 distribuzioni di Gumbel:

- una $F_1 = \exp\left(-\exp\left(-\frac{x-b_1}{a_1}\right)\right)$ relativa agli eventi ordinari;
- una $F_2 = \exp\left(-\exp\left(-\frac{x-b_2}{a_2}\right)\right)$ relativa ad eventi estremi ed eccezionali.

Ipotizziamo un rapporto di miscelazione λ che indichi quanto i fenomeni legati ad F_2 sono più rari rispetto a quelli legati ad F_1 :

$$F = (1-\lambda)F_1 + \lambda F_2 \quad \text{con } 0 \leq \lambda \leq 1$$

In tal modo abbiamo, per la distribuzione, i 5 parametri a_1 , b_1 , a_2 , b_2 e λ .

Questo tipo di distribuzione è stata testata dal GNDCI (gruppo nazionale per la difesa dalle catastrofi idrogeologiche) relativamente ai soli dati di pioggia. Ovviamente il peso del rapporto λ deve essere commisurato al rischio che si associa all'evento raro, ed in particolare al tempo di ritorno dell'evento eccezionale in relazione all'importanza dell'opera che si sta progettando.

Test statistici

Dopo aver stimato una funzione di probabilità che adatti la distribuzione di una variabile aleatoria campionaria, possiamo eseguire un test per verificare se l'adattamento è accettabile o meno.

Il test parte dalle considerazioni:

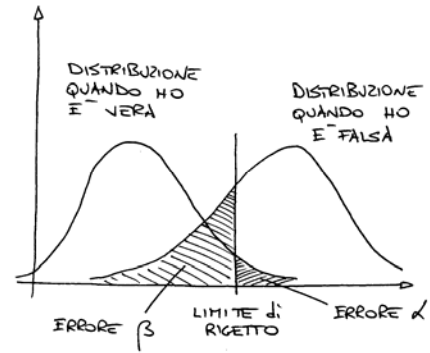
- i parametri a e b vengono stimati, quindi hanno una certa variabilità e possono così essere interpretati come variabili aleatorie;
- sotto particolari ipotesi, possiamo ritenere nota la distribuzione di probabilità di tali parametri;
- in funzione dei parametri è possibile valutare la STATISTICA DEL TEST, ossia una misura, determinata mediante opportune regole di calcolo, del grado di veridicità di una certa IPOTESI INIZIALE H_0 ; nel nostro caso la statistica è fondamentalmente la differenza tra la distribuzione del campione e la distribuzione teorica adottata;
- in base alle distribuzioni osservate e adottate, e alla misura della statistica calcolata, possiamo fissare un CRITERIO DI CONFRONTO, ossia un limite di rigetto per i valori della statistica, superato il quale l'ipotesi iniziale non è verificata;

Al di fuori delle fasce di confidenza, delimitate dai limiti di rigetto, è comunque presente una probabilità residua che l'ipotesi iniziale sia verificata: diciamo ERRORE DEL PRIMO TIPO la probabilità α di rigettare H_0 quando è vera. Ovviamente tale errore sarà sempre presente, ma va mantenuto il più basso possibile. Allo stesso modo avremo una certa probabilità β di accettare H_0 anche quando è falsa, che chiameremo ERRORE DEL SECONDO

TIPO. Come è facile osservare dalla figura, al crescere di α si assiste al calare di β , e viceversa.

La sequenza logica nella produzione di un test è quindi la seguente:

1. si identifica un'ipotesi H_0 ;
2. si definisce una statistica, tipica del test;
3. si definisce un criterio di rigetto C_α ;
4. si sceglie un livello α di errore del primo tipo;
5. si determina il valore di C_α in funzione di α ;
6. si calcola la statistica $P = P(x_1, \dots, x_N)$.
7. si controlla che l'ipotesi iniziale sia verificata.



Test del χ^2

L'ipotesi H_0 del test è che le frequenze osservate coincidono con quelle teoriche.

La prima operazione è quella di dividere l'asse delle variabili aleatorie x in M classi mediante $M - 1$ punti di estremo C_1, C_2, \dots, C_{M-1} , tali che per la prima classe avremo $x < C_1$, per la seconda classe avremo $C_1 < x < C_2$, eccetera. I valori X_1, \dots, X_N del campione si disporranno liberamente all'interno delle classi appena definite, ed otteniamo così una FREQUENZA OSSERVATA O_i per $i = 1, \dots, M$, ossia il numero di elementi del campione che ricadono in una certa classe.

Supponendo nota la PDF, possiamo dire che l'area sottostante tale curva e delimitata dalle linee di classe $x = C_{i-1}$ e $x = C_i$ rappresenta la probabilità che un generico valore x sia compreso tra C_{i-1} e C_i . Moltiplicando poi questa probabilità per N , otteniamo la FREQUENZA TEORICA E_i , ossia il numero atteso di uscite all'interno della singola classe. Va osservato che, affinché il test funzioni, il numero atteso in ciascuna classe deve essere maggiore o uguale a 5, e dobbiamo quindi disporre di un campione di almeno 20 elementi. La statistica da calcolare è poi:

$$C2 = \sum_{i=1}^M \frac{(O_i - E_i)^2}{E_i}$$

Quando l'ipotesi di base H_0 è verificata, ossia quando le frequenze osservate sono abbastanza simili alle frequenze teoriche, allora la statistica $C2$ segue la distribuzione $\chi^2(\nu)$, funzione dei gradi di libertà del sistema:

$$\nu = M - 1 - P$$

dove P è il numero di parametri della PDF²

Prima di calcolare la statistica va definito il livello di errore del primo tipo, che generalmente vale 0.01, 0.05 o 0.10: considerando ad esempio $\alpha = 0.05$ possiamo definire il valore critico

$$\chi_{crit}^2 = \chi^2(P(0.95), \nu)$$

Quando $C2 < \chi_{crit}^2$ accettiamo l'ipotesi di base, e quindi la distribuzione adottata è rappresentativa del campione; se invece $C2 > \chi_{crit}^2$ non possiamo accettare l'ipotesi di base. Essendo i valori tutti positivi, il test sarà ad una coda.

² Nel caso della distribuzione di Gumbel, $P=2$.

Test di Kolmogorov-Smirnov

L'ipotesi di questo test è identica a quella del test χ^2 , ossia che le frequenze osservate coincidono con quelle teoriche.

La prima operazione da compiere è la definizione di una frequenza di campionamento

$$F_E(i) = \frac{i}{N}$$

dove i è l'ordine (crescente) dell'elemento campionario

Definiamo poi la funzione

$$D(x) = \left| F_X(x, \hat{a}, \hat{b}) - F_E \right|$$

ossia la differenza, in ogni singolo punto x , tra il valore della frequenza di campionamento e la CDF adottata.

La statistica del test sarà:

$$D_N = \max_x [D(x)]$$

Osservando che il massimo deve necessariamente cadere in corrispondenza dei valori campionari, la statistica viene ad essere:

$$D_N = \max_i \left\{ \left| F_X(x_i, \hat{a}, \hat{b}) - \frac{i}{N} \right| - \left| F_X(x_{i-1}, \hat{a}, \hat{b}) - \frac{i-1}{N} \right| \right\}$$

Anche questo test è ad una coda, in quanto la statistica è il massimo di valori assoluti, e quindi il massimo di valori tutti positivi.

Il criterio di rigetto è dato dalla condizione

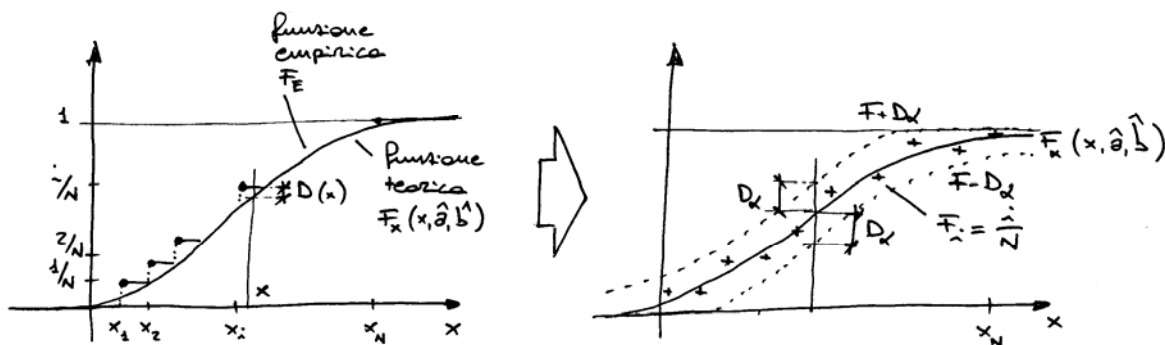
$$D_N > D_\alpha$$

dove $D_\alpha = D_\alpha(N)$ è una distribuzione molto complessa, rintracciabile unicamente su testi specialistici. I valori di D_α per i piccoli campioni sono tabellati; per campioni con

$N \geq 35 \div 40$ possiamo invece assumere i valori approssimati:

$$D_{0.10} = \frac{1.22}{\sqrt{N}}, \quad D_{0.05} = \frac{1.36}{\sqrt{N}}, \quad D_{0.01} = \frac{1.63}{\sqrt{N}}$$

Un metodo molto veloce di eseguire il test è quello per via grafica: definiamo il range $[F_X - D_\alpha; F_X + D_\alpha]$ attorno alla funzione di probabilità, ignorando i valori maggiori di 1 e minori di 0, e valutiamo se tutti i valori delle frequenze osservate cadono all'interno del range; se così non fosse, il test non sarebbe soddisfatto.



Applicazioni statistiche dell'idrologia

L'utilizzo di elaborazioni statistiche sui dati idrologici di piena è importante:

- in fase di VALUTAZIONE DEL RISCHIO IDROLOGICO;
- in fase di PROGETTAZIONE DEI SERBATOI ARTIFICIALI;
- in fase di PROGETTAZIONE DELLE OPERE DI PROTEZIONE IDRAULICA.

In riferimento a quanto detto sopra, i valori idrologici che ci interessa maggiormente ricavare sono:

- la PORTATA AL COLMO della piena Q_{colmo} , dalla quale possiamo ad esempio ricavare la quota di sicurezza dell'argine;
- il VOLUME DI PIENA V_{piena} in base al quale possiamo, ad esempio, dimensionare i serbatoi artificiali.

Quando parliamo di campione di eventi idrologici, ci riferiamo sostanzialmente al considerare le portate di piena o le misure di pioggia come variabili aleatorie. Non conoscendo deterministicamente il meccanismo di produzione delle piene, i valori Q_{colmo} e V_{piena} sono ottenibili a partire da un'inferenza statistica, ossia una stima della distribuzione di probabilità che meglio approssima l'evento osservato. L'inferenza, infatti, può essere

- diretta sulle osservazioni di portata di piena;
- indiretta, ossia con la distribuzione delle portate di piena ricavate dall'inferenza sui dati di pioggia mediante l'applicazione di un modello geomorfologico.

A seconda dei dati disponibili sulle portate di piena, ossia delle dimensioni del campione di portate, si adottano infatti approcci diversi:

- quando non sono presenti dati, oppure i dati sono relativi a pochi anni, è necessario costruire una trasformazione afflussi-deflussi;
- quando abbiamo dati relativi ad un periodo di osservazione di almeno 10 anni, possiamo invece effettuare un'analisi statistica delle portate di piena osservate.

Criteri di campionamento dei dati di portata

Esistono fondamentalmente tre metodi per effettuare il campionamento dei dati idrologici:

- possiamo effettuare un campionamento delle diverse portate di PICCO, ossia i valori massimi di portata tra i quali sia visibile la curva di decadimento delle sorgenti;
- possiamo fissare un valore di SOGLIA e considerare il massimo valore di portata per ogni intervallo che superi tale soglia, così da considerare gli eventi come indipendenti: otterremo in tal modo la serie POT (Peaks Over Threshold); tale serie è solitamente consigliata quando avessimo a disposizione dati relativi ad un decennio, in quanto solitamente si ha una media di 8-10 eventi di piena all'anno;
- possiamo considerare il PICCO MASSIMO Q_{ma} in un periodo annuale³, ottenendo in questo modo la serie MAF (Maximum Annual Flood); tale serie è solitamente adoperata quando si hanno dati relativi a più di 20 anni.

³ Utilizziamo il periodo annuale al fine di mantenere la stazionarietà del processo: in senso stretto la struttura di probabilità risulta essere sempre la stessa, in senso debole si ha una stazionarietà di media e varianza.

A seconda che si effettui il campionamento secondo una o l'altra serie, si adottano solitamente strutture di probabilità diverse: per la serie POT useremo una distribuzione esponenziale, mentre per la serie MAF utilizzeremo una distribuzione di Gumbel. Fissiamo l'attenzione su quest'ultimo caso: la probabilità di NON SUPERAMENTO, ossia la probabilità che la portata al colmo massima annuale superi un certo valore q , sarà data dalla funzione

$$F_x(x) = P(Q_{ma} \leq q) = \exp\left(-\exp\left(-\frac{Q_{ma} - b}{a}\right)\right),$$

mentre la probabilità di superamento, ovvero la probabilità che la portata al colmo massima annuale superi tale valore, sarà ovviamente data dalla funzione

$$F_1(x) = P(Q_{ma} \leq q) = 1 - F_x(x)$$

Definiamo il TEMPO DI RITORNO come il valore medio⁴ di attesa tra due superamenti successivi:

$$T_R = \frac{1}{F_1} = \frac{1}{1 - F_x}$$

Va notato che il concetto di tempo di ritorno è puramente statistico, in quanto valor medio; nella realtà non è detto che una piena eccezionale ritorni esattamente dopo T_R anni.

Così, sempre considerando un periodo di riferimento annuale, la probabilità che la portata massima annua superi il valore q è data da:

$$P(Q_{ma} > q) = \frac{1}{T_R}$$

mentre la probabilità del non superamento è

$$P(Q_{ma} > q) = 1 - \frac{1}{T_R}$$

Considerando invece un periodo di riferimento di m anni, invece, la probabilità di non superamento è data dalla

$$P(Q_{ma} > q) = \left(1 - \frac{1}{T_R}\right)^m$$

Quando poi avessimo $m = T_R$ la probabilità di non superamento viene ad essere

$$P(Q_{ma} > q) = \left(1 - \frac{1}{T_R}\right)^{T_R} \cong \frac{1}{e}$$

Poiché $e = 2.71 \cong 3$, su T_R anni avremo:

- circa un terzo di probabilità che la portata massima non sia mai superata;
- circa un terzo di probabilità che la portata massima sia superata una sola volta;
- circa un terzo di probabilità che la portata massima sia superata due o più volte.

Criteri di campionamento dei dati di pioggia

Per il campionamento delle misure di pioggia ci si riferisce ad altezze cumulate su intervalli di tempo prefissati, in quanto la misura puntuale è molto difficile da effettuarsi, come abbiamo già visto. Solitamente, il campione delle misure di pioggia sarà quindi

⁴ Il tempo di ritorno viene valutato in anni; dimensionalmente la definizione è corretta, in quanto l'1 al numeratore è l'unità di tempo di campionamento, ossia 1 anno.

costituito dalle massime altezze cumulate in 1,3,6,12 e 24 ore⁵, con un periodo di riferimento di un anno.

Ogni serie di misure $F_H^j(h^j)$, relativa ad uno degli intervalli di tempo $j = 1hr, \dots, 24hr$, sarà poi approssimata da una distribuzione EV1 di Gumbel:

$$F_Z^j(z) = \exp(-\exp(-z)) \quad \text{per } -\infty < z < +\infty$$

con l'assunzione $z = \frac{h^j - b^j}{a^j}$.

Al solito per la stima di a e b possiamo usare i metodi dei momenti, della massima verosimiglianza o dei minimi quadrati.

L'uso dei dati di pioggia così campionati ed approssimati pone però un'importante problema: non è detto che il parametro di tempo della trasformazione afflussi-deflussi sia uguale ad uno degli intervalli fissati dal Servizio Idrografico Nazionale.

Una rete di drenaggio urbana, ad esempio, è dimensionata in base alla portata nel collettore, stimata mediante l'uso del modello della corrivazione lineare sui dati di pioggia. Il tempo di corrivazione del bacino, in generale, non sarà esattamente 1,3,6,12 o 24 ore. Sarà quindi necessario ricavare una funzione di h e t che ci permetta di ottenere dei valori di ruscellamento per ogni durata di pioggia.

Assegnato quindi, anche per le piogge, un tempo di ritorno⁶, possiamo ricavare la :

$$h_j = \hat{b}_j - \hat{a}_j \ln \ln \frac{1}{F_H} \Rightarrow h_j = \hat{b}_j - \hat{a}_j \ln \ln \frac{T_R}{T_R - 1}$$

In base a queste considerazioni possiamo utilizzare la cosiddetta LSPP (Linea Segnalatrice di Possibilità Pluviometrica) che interpola gli h_j , una *linea di potenza per espressione di potenza, ossia una parabola ad asse orizzontale in senso lato*, definita come:

$$h(t) = at^n$$

dove a è pari a $0.2 \div 0.3$ per la Pianura Padana e $0.5 \div 0.6$ per le Alpi

n è compreso tra 0 e 1 e collegato alla collocazione geografica

entrambi i parametri variano in base al tempo di ritorno assegnato

A questo punto, sempre in riferimento al modello della corrivazione, è facile calcolare la portata

$$Q = A_B r \quad \text{con } r = \begin{cases} \phi j \frac{t}{T_C} & \text{per } t < T_C \\ \phi j & \text{per } t \geq T_C \end{cases}$$

Al crescere della durata t , abbiamo così per la portata:

- un'espressione crescente quando $t < T_C$:

$$Q = \frac{\phi A_b}{T_C} at^n$$

- un'espressione decrescente quando $t \geq T_C$:

$$Q = \phi A_b at^{n-1}$$

La massima portata di picco sarà così

$$Q = \phi A_b a T_C^{n-1}$$

⁵ Intervalli fissati dal Servizio Idrografico Nazionale.

⁶ Valori normalmente usati per il tempo di ritorno: 5 anni per le fognature, 100 anni per i fiumi, 1000 anni per le dighe.

Trasformazione afflussi-deflussi

Nel caso della trasformazione afflussi-deflussi, abbiamo i dati di partenza:

- le piogge $F_H(h)|t$ condizionate ad una certa durata t mediante la funzione LSPP

$$h(T_R, t) = at^n;$$

- una certa funzione $g(h)$, data dall' IUH, che trasforma le piogge in portate;

e vogliamo trovare la distribuzione $F_Q(q)$.

In realtà la distribuzione di probabilità delle piogge sarebbe una funzione complessa del tipo $F_{H,T}(h, t)$, a dominio bidimensionale (h, t) e condominio monodimensionale.

Adottiamo quindi una convenzione lavorando a t costante, ossia condizionando le piogge

alle durate assegnando una frequenza $T_R : F = 1 - \frac{1}{T_R}$ che determina la funzione LSPP.

Al variare di t la funzione $g(h)$ fornisce una portata di piena $Q(h, t)$ della quale cerco il valore massimo. La particolare durata t^* che produce la PORTATA CRITICA $Q_{\max} = g(h^*, t^*)$

viene detta DURATA CRITICA, mentre la coppia $(h^* = a(t^*)^n, t^*)$ viene detta EVENTO CRITICO.

Secondo consuetudine assumiamo lo stesso tempo di ritorno T_R sia per la portata critica che per le piogge, nonostante nella realtà il tempo di ritorno delle portate sia minore di quello delle piogge, e quindi il rischio aumenti con tale assunzione.

Valutiamo ora le espressioni per la durata critica a seconda che si utilizzi uno o l'altro modello di IUH.

Per quanto riguarda il modello della corrivazione, abbiamo già visto che la durata critica coincide con il tempo di corrivazione del bacino.

Invaso

Per quanto riguarda invece il modello dell'invaso, partiamo da un IUH del tipo

$$IUH = \frac{1}{k} \exp\left(-\frac{t}{k}\right)$$

La portata di picco sarà così determinata dall'espressione

$$Q_p = Q(t) = \phi A_b \frac{h}{t} \left[1 - \exp\left(-\frac{t}{k}\right) \right] = \phi A_b a t^{n-1} \varepsilon(t)$$

dove $\varepsilon(t) = 1 - \exp\left(-\frac{t}{k}\right)$ è detto fattore di riduzione della piena

Visto che $\phi A_b a$ è un prodotto tra costanti, possiamo cercare il massimo come:

$$\max_t (Q_p) = \max_t [t^{n-1} \varepsilon(t)] = \max_t \left\{ t^{n-1} \left[1 - \exp\left(-\frac{t}{k}\right) \right] \right\}$$

Moltiplicando poi per la costante $\frac{1}{k^{n-1}}$ e ponendo $c = \frac{t_{crit}}{k}$ il massimo diventa:

$$\max_t \left\{ \left(\frac{t}{k}\right)^{n-1} \left[1 - \exp\left(-\frac{t}{k}\right) \right] \right\} = \max_t \left\{ c^{n-1} [1 - \exp(-c)] \right\} = \max_t F(c)$$

Poniamo quindi uguale a zero la derivata prima della funzione:

$$\frac{dF(c)}{dc} = (n-1)c^{n-2} (1 - e^{-c}) + c^{n-1} e^{-c} = 0$$

$$(n-1)(1-e^{-c}) + ce^{-c} = 0$$

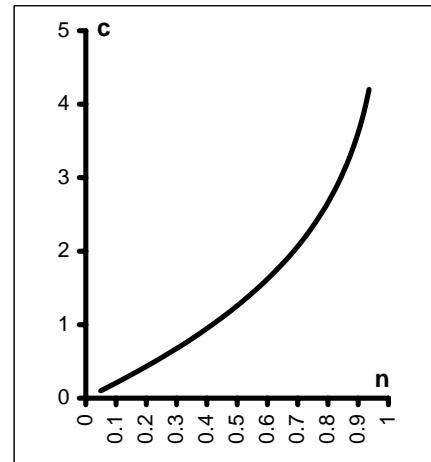
$$ce^{-c} = (n-1)(1-e^{-c})$$

Vista la difficoltà di risolvere quest'ultima equazione, e visto che l'espressione

$$n = 1 - \frac{ce^{-c}}{1-e^{-c}}$$

dà luogo al grafico univocamente determinato di figura, possiamo ricavare c a partire da n e calcolare poi la durata critica mediante l'espressione:

$$t_{crit} = ck$$



Nash

Per quanto riguarda il modello di Nash possiamo condurre lo stesso ragionamento svolto per il modello dell'invaso.

Partiamo da un'espressione per l'idrogramma unitario istantaneo

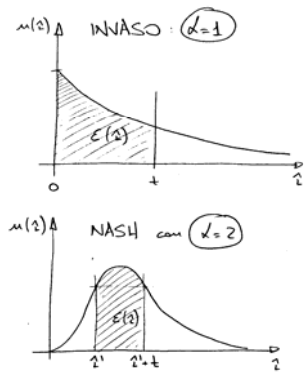
$$u(\tau) = \frac{1}{k\Gamma(\alpha)} \left(\frac{\tau}{k}\right)^{\alpha-1} \exp\left(-\frac{\tau}{k}\right)$$

e cerchiamo una durata di pioggia t tale che

$$\frac{d\varepsilon}{dt} = (1-n) \frac{\varepsilon}{t}$$

dove l'area ε è pari a

$$\varepsilon = \max_{\tau'} \int_{\tau'}^{\tau'+t} u(\tau) d\tau$$



Il problema si riduce quindi sostanzialmente al calcolo dell'area $\varepsilon(\tau)$, difficile poiché dobbiamo calcolarne il massimo sugli estremi d'integrazione.

Data una generica funzione $f(x)$, dovendo calcolare

l'integrale $\varepsilon = \int_x^{x+t} f(\xi) d\xi$, possiamo cercare la ε che

annulla la derivata prima rispetto alla x . Sappiamo che per un incremento $x+dx$ abbiamo un incremento $d\varepsilon = f(x+dx)dx - f(x)dx$, e quindi

$$\frac{d\varepsilon}{dx} = f(x+dx) - f(x) = 0.$$

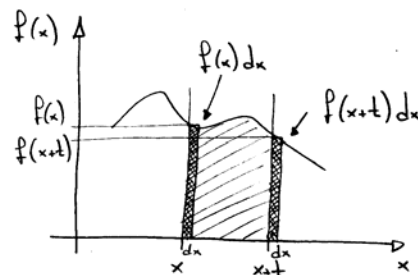
Nel caso di $\alpha > 1^7$, quindi, cerchiamo una τ' tale che

$$u(\tau'+t) = u(\tau')$$

In tal modo possiamo ricavare:

$$(\tau'+t)^{\alpha-1} \exp\left(-\frac{\tau'+t}{k}\right) = \tau'^{\alpha-1} \exp\left(-\frac{\tau'}{k}\right)$$

$$(\tau'+t)^{\alpha-1} \exp\left(-\frac{\tau'}{k}\right) \exp\left(-\frac{t}{k}\right) = \tau'^{\alpha-1} \exp\left(-\frac{\tau'}{k}\right)$$



⁷ Per $\alpha = 1$ (IUH dell'invaso) la relazione non vale in quanto la funzione è vincolata ed il massimo si trova sul confine del dominio di definizione della funzione.

$$(\tau' + t)^{\alpha-1} \exp\left(-\frac{t}{k}\right) = \tau'^{\alpha-1}$$

dove quest'ultima espressione va risolta per tentativi, almeno quando $\alpha \notin \mathbb{N}$.

In generale, quindi, il metodo per trovare la durata critica è:

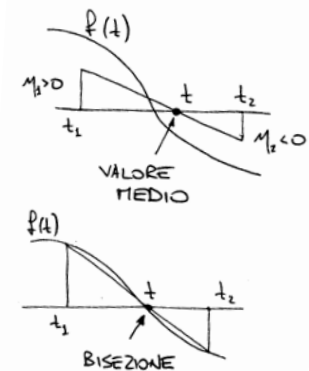
1. fissare una durata t di primo tentativo;
2. calcolare numericamente o analiticamente τ' ;
3. calcolare $\varepsilon(t) = \Gamma'(\tau' + t; \alpha, k) - \Gamma'(\tau'; \alpha, k)$, dove $\Gamma'(x; \alpha, k) = \int_0^x u(\tau) d\tau$ ⁸;
4. calcolare la derivata $\frac{d\varepsilon}{dt}$ in modo analitico, poiché $\frac{d\varepsilon}{dt} = u(\tau') = u(\tau' + t)$;
5. verificare se l'equazione $\frac{d\varepsilon}{dt} = (1-n)\frac{\varepsilon}{t}$ è soddisfatta; in caso contrario ripetere il procedimento scegliendo una durata t di secondo tentativo.

Oggi giorno è ovviamente possibile far eseguire l'iterazione ad un software, ad esempio ad un foglio di calcolo Excel. Tuttavia è utile conoscere il procedimento di scelta delle nuove t , in quanto i metodi di iterazione usati dal software si basano sullo stesso procedimento.

Innanzitutto definiamo lo scarto

$$\eta = \frac{d\varepsilon}{dt} - (1-n)\frac{\varepsilon}{t}$$

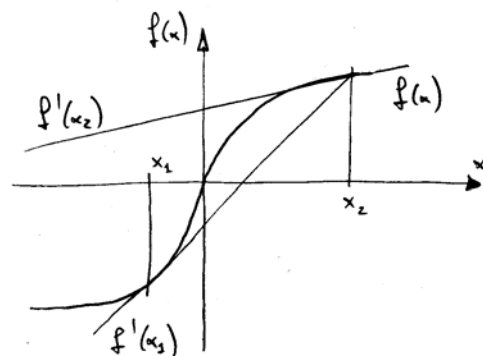
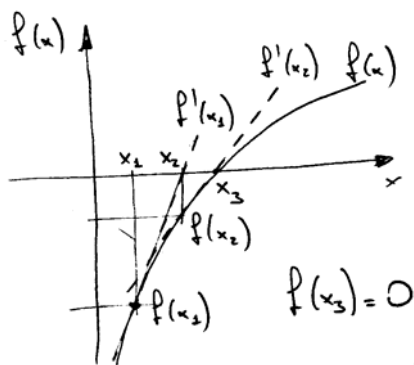
Quando $\eta > 0$ la t va aumentata, mentre quando $\eta < 0$ la t va diminuita. Ad un certo punto dell'iterazione, avremo una durata t_1 di poco minore del valore cercato, ed una durata t_2 di poco maggiore del valore cercato. A questo punto possiamo usare il metodo del valore medio o della bisezione per trovare la durata cercata.



Un altro metodo piuttosto usato è invece quello di Newton-Raphson, che consiste nel calcolare, per un generico valore x_1 , i

valori $f(x_1)$ e $f'(x_1)$. A partire da questi valori troviamo il punto $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$, per il

quale calcoliamo ancora il valore della funzione e della derivata prima: l'iterazione si ferma quando $f(x_i) = 0$. Il metodo va molto bene in quanto si ha una convergenza molto rapida, ma può avere grossi problemi con alcune funzioni.



⁸ La funzione gamma incompleta è disponibile, nella forma proposta e con i parametri definiti, nel software di calcolo Excel.

Nash secondo Matteo Mainetti

Una strada alternativa per calcolare la durata critica è quella di ricavare, anche per Nash, un grafico di n in funzione di c .

Nel caso di un modello di Nash con 2 serbatoi abbiamo l'espressione:

$$n = 1 - \frac{c^2}{e^c - c - 1}$$

Nel caso invece di m serbatoi abbiamo:

$$n = 1 - \frac{c^m}{(m-1)!e^c - \left[\sum_{i=1}^{m-1} (m-i)!c^i \right] - (m-1)!}$$

La durata critica sarà sempre:

$$t_{crit} = ck$$