# POLITECNICO DI MILANO

Facoltà di Ingegneria Civile, Ambientale e Territoriale

Laurea magistrale in Tecnologie di Risanamento Ambientale


# TECNICHE STATISTICHE PER LO STUDIO DI FENOMENI AMBIENTALI

Relatore:

Prof. Elio Piazza

Tesi di laurea Magistrale di:

Nicolò MANDELLI Matr. 767925

Anno Accademico 2013/2014

*Milano, 27 Novembre 2013*

*Rivolgo un grande ringraziamento al Professore Elio Piazza, per essersi fidato di me.*

*Ringrazio Francesca, complice impareggiabile, per avermi ascoltato e accompagnato fino alla fine, con un sorriso.*

*Grazie anche a Carlotta, Cristina e Massimiliano, che sempre assecondano le mie decisioni e che mi aiutano a renderle possibili, anzi, molto probabili.*

*A Elena*

*che ci ha cresciuto tutti*

# INDICE

# 1. DEGLI ERRORI E DELLE INCERTEZZE

*“Il dubbio non è piacevole, ma la certezza è ridicola”*
Voltaire

Col presente lavoro si intendono esporre ad un ingegnere ambientale gli strumenti statistici adeguati ad affrontare la vasta gamma di possibili problemi che potranno presentarsi nel corso di una fase progettuale e di monitoraggio, sia essa volta alla difesa, alla pianificazione o al risanamento del territorio, tenendo conto fin da subito che l'ingegneria ambientale si occupa di studiare fenomeni ambientali e situazioni di interazione tra uomo e natura, al fine di rendere agevole e possibile l'insediamento umano in un particolare tipo di ambiente e territorio.

Si guardi per esempio alla pianificazione e gestione dei parchi naturali, per i quali è necessaria un'approfondita conoscenza degli ecosistemi che li abitano, nonché sulle dinamiche *intra* e *inter*-specifiche delle specie presenti nel territorio, sia animali sia vegetali. Gli studi scientifici forniranno le basi conoscitive per determinare quale sia la dimensione idonea del parco per la sopravvivenza della flora e della fauna presenti, piuttosto che evidenziare la necessità di creare corridoi ecologici che lo connettano con altri territori e permettere la migrazione di determinate specie, o ancora determinare la quantità massima di individui prelevabili per raccolta o cacciagione che non comprometta la sopravvivenza delle coorti presenti.

I campi dell'ingegneria ambientale si estendono anche alla protezione civile. Infatti, per fare un esempio, attraverso lo studio delle piene di un determinato corso fluviale, è possibile ottenere gli strumenti per poter prevedere e scongiurare eventuali situazioni di rischio per la popolazione limitrofa, predisponendo opere di contenimento delle portate e sistemi efficienti di allarme. Anche qui, verranno fatti modelli di deflusso del bacino idrico in esame e, come si vedrà, ad ogni valore di portata si assegnerà una determinata probabilità di accadimento e ad ogni parcella di territorio limitrofo un livello di rischio e vulnerabilità.

Ancora, un ingegnere ambientale potrà trovarsi nella situazione di dover risanare, per esempio, una determinata area in cui il suolo e la falda sottostante sono caratterizzati dalla presenza di sostanze inquinanti. Anche in questo caso, tramite rilevazioni sul campo, carotaggi e successive simulazioni virtuali, verrà stimato il volume di suolo affetto da

contaminazione e gli eventuali futuri spostamenti dell'inquinante nella falda, per poi predisporre le azioni necessarie ad eliminare il rischio per l'uomo e l'ambiente.

Un fenomeno che si sta prendendo in considerazione sempre di più negli ultimi anni e che si sta manifestando con sempre maggiore impeto sono i cambiamenti climatici. La manifestazione di questo fenomeno si traduce in maniere differenti per luoghi differenti. Ad esempio in Europa è stato osservato un aumento delle piovosità in autunno e primavera mentre vengono periodicamente registrati record di ondate di calore nei periodi estivi. In altre zone, soprattutto nella fascia equatoriale, si sta invece registrando un aumento della desertificazione e della siccità del suolo, con conseguenze già evidenti in termini di calo della produzione agricola. Ancora, vi sono in tutto il mondo diverse isole e località costiere in serio pericolo di sommersione, dovuto all'innalzamento del livello dei mari e degli oceani, e questo ora sta dando luogo a fenomeni migratori consistenti. Ma come fare a quantificare l'entità del fenomeno? Come fare a prevedere quali potrebbero essere gli scenari futuri a cui stiamo andando incontro?

Tutti i casi sopra elencati prevedono uno studio scientifico dei fenomeni naturali coinvolti. Chiunque abbia avuto a che fare con problemi di questo tipo sa bene che la conoscenza delle dinamiche ambientali è affetta da elevata incertezza, dovuta all'impossibilità dell'uomo di conoscere e prevedere esattamente il comportamento della natura nel suo insieme. Questo si traduce nella difficoltà di formulare modelli basati esclusivamente su considerazioni teoriche e nella necessità di servirsi di più dati possibile per poter fare asserzioni sui fenomeni studiati. Inoltre, quando si affrontano problemi in questo campo, spesso i dati a disposizione non sono sufficientemente numerosi e non si dispone di serie storiche sufficientemente estese per poter generare modelli affidabili e previsioni a lungo termine. Da qui è nata la necessità di sviluppare metodi statistici adatti a trattare i dati disponibili, supportati da conoscenze scientifiche *a priori* riguardanti i processi studiati. L'uso della statistica nello studio della natura, in altre parole, deriva proprio dalla consapevolezza dell'uomo di non poter arrivare alla conoscenza assoluta, ma sempre ci sarà un margine di incertezza che indurrà a descrivere le manifestazioni di un fenomeno naturale non come quantità deterministiche, bensì come *variabili aleatorie*. In altre parole, gli errori, siano essi dovuti alla non completa conoscenza del fenomeno o errori di misura, nell'approccio statistico non vengono mascherati ma al contrario vengono ben evidenziati e su questi si concentra la maggior parte dello studio. Si ricordi che comunque lo scopo dell'ingegnere è quello di predisporre le tecnologie idonee a far fronte ad un certo problema o situazione, quindi spesso non vi è la necessità di conoscere

esattamente i processi naturali con cui si ha a che fare, ma sono sufficienti gli ordini di grandezza del fenomeno, per poi comunque applicare un margine di sicurezza ai valori di progetto.

Con il presente lavoro si vogliono illustrare le basi statistiche adeguate allo studio dei principali problemi ambientali.

Nel primo capitolo si illustreranno argomenti di statistica di base, fornendo la definizione e il trattamento delle variabili aleatorie e dei loro stimatori, a partire dall'enunciazione del teorema centrale del limite e passando attraverso la formulazione delle funzioni di ripartizione (fdr) e di densità di probabilità (fdp).

Nel secondo capitolo sì entrerà nei dettagli del metodo della formulazione di test d'ipotesi, a partire dalla formulazione di intervallo di confidenza e di statistica test.

Nel terzo capitolo si parlerà dello strumento fondamentale quale è la regressione lineare semplice e multivariata, risolta attraverso il metodo dei minimi quadrati.

Nel quarto capitolo si esporrà il grande problema dell'analisi di rischio, applicata a diversi settori dell'ingegneria ambientale e fornendo alcuni esempi reali. In particolare, verranno esplicitate le metodologie di valutazione del rischio geologico, idrogeologico e per siti contaminati.

Nel quinto capitolo si farà luce sui metodi di formulazione di modelli matematici descriventi fenomeni naturali. In particolare, verrà descritto il metodo classico in confronto con quello bayesiano, passando poi al metodo di simulazione Monte Carlo e all'algoritmo MCMC (*Markov Chain Monte Carlo*), dando un esempio di applicazione reale sulla stima dell'incertezza negli inventari di emissione per le combustioni domestiche. Segue un'illustrazione dei modelli ARMA e delle loro componenti. Il capitolo si conclude con una trattazione sui principali metodi di selezione dei modelli, fornendo anche qui un esempio di applicazione nello studio dello stambecco delle Alpi del parco nazionale del Gran Paradiso.

L'ultimo capitolo affronta il problema del trattamento dell'incertezza. In particolare, verrà presentato l'algoritmo HMC (*Hybrid Monte Carlo*), una sua applicazione alla stima dell'incertezza sugli inventari di emissione per la combustione domestica di legna, alcune tecniche di stima dell'incertezza sulle misure di portata e verrà infine descritto il processo di *rumore bianco*.

# 2. STATISTICA DI BASE

*"Seconno le statistiche di adesso*
*risurta che te tocca un pollo all'anno*
*e, se nun t'entra nelle spese tue,*
*t'entra ne la statistica lo stesso*
*perché c'è un antro che ne magna due"*
Trilussa

La necessità di ricorrere a metodi statistici, nell'ingegneria ambientale come in qualsiasi altro ambito scientifico, deriva dall'impossibilità di poter studiare il fenomeno in esame nel suo complesso. Si ricorre allora alla raccolta di più dati possibile riguardanti l'oggetto del nostro studio, come ad esempio gli eventi di piena di un fiume piuttosto che le concentrazioni di un dato inquinante atmosferico. I valori osservati vengono trattati come estrazioni di una variabile casuale. L'insieme che comprende tutti i valori che può assumere una variabile casuale, e che quindi nelle nostre misure ci aspettiamo di poter osservare, si chiama popolazione. Un campione è invece un qualsiasi sottoinsieme della popolazione ed è l'oggetto su cui si concentrerà lo studio della popolazione. L'intento, infatti, è quello di poter fare asserzioni sulla popolazione in esame basandosi sull'osservazione di alcuni suoi elementi. A tal proposito, è importante fin da subito sottolineare il fatto che se una popolazione $X$ è descritta da una funzione di densità $f_X(.)$ e $X_1,...X_n$ sono variabili aleatorie indipendenti identicamente distribuite secondo $f_X(.)$, allora la densità congiunta del vettore aleatorio $X = [X_1...X_n]'$ è $f_X(x_1...x_n) = f_X(x_1)f_X(x_2)....f_X(x_n)$ e questo vettore si chiama campione casuale. La distribuzione del campione casuale si chiama distribuzione campionaria.

## *2.1. Variabile aleatoria*

Prima di dare la definizione di variabile aleatoria è necessario capire cos'è lo spazio di probabilità. Lo spazio di probabilità è descritto da tre elementi: $(\Omega, A, F(.))$, dove $\Omega$ è lo spazio campionario, $F(.)$ è la funzione di probabilità assegnata ad $A$, che è la $\sigma$-algebra di $\Omega$. A prende il nome di spazio degli eventi ed è una collezione di sottoinsiemi di Ω che contiene almeno tutti gli eventi elementari tale che, presa un'infinità numerabile di

sottoinsiemi di $\Omega$ che stiano in $A$, la loro unione sta ancora in $A$. Tutte le collezioni di sottoinsiemi che godono di queste proprietà si chiamano $\sigma$-algebra. $\Omega$ è lo spazio campionario, cioè l'insieme di tutte le possibili realizzazioni di un esperimento. $F(.)$, come si è detto, prende il nome di funzione di probabilità ed associa ad ogni elemento di $A$ un valore compreso nell'intervallo $[0,1]$.

Una variabile aleatoria è una funzione che ha come dominio $\Omega$ e come codominio R (insieme dei numeri reali), cioè $X: \Omega \rightarrow R$, ed associa ad ogni elemento di $\Omega$ un numero reale $x \in R$ . Inoltre, una funzione $X: \Omega \rightarrow R$ è una variabile aleatoria se $\forall r \in R, A_r = [\omega : X(\omega) \leq r] \in A$ , cioè ogni semiretta $(-\infty, r]$ dell'asse reale ha come controimmagine un evento di A. Questo permette di trasferire la probabilità da $A$ ad $R$, cioè semplicemente associare ad ogni evento un numero reale. Si dice, infatti, che una variabile aleatoria $X$ codifica eventi $\omega \in \Omega$ con numeri $x \in R$ .

## *2.2. Misure di tendenza centrale e di dispersione*

Spesso i dati riguardanti un sistema naturale o qualche aspetto particolare di questo tendono a concentrarsi intorno ad un valore. Spesso si sceglie quindi questo valore, che chiameremo valore centrale, come rappresentativo del campione. Non esiste una particolare definizione di valore centrale. A seconda dei casi, sarà la discrezione dello statistico che porterà ad adottare la media, la moda o magari la mediana.

La media campionaria è una stima della media della popolazione, calcolata sulla base del campione osservato $[x_1 ... x_n]$:

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

La media campionaria sembra essere il valore più adatto se si vuole rappresentare il campione con un numero solo. Il difetto di questo stimatore è che più essere influenzato da misure che non sembrano appartenere alla popolazione in esame e che sono anomale rispetto alla distribuzione della variabile aleatoria osservata. Questi valori prendono il nome di *outliers*. La presenza di *outliers* può essere dovuta a tantissimi motivi: possono esservi ragioni fisiche, come per esempio uno spostamento di un punto di equilibro o un cambiamento di condizioni in cui si osserva il fenomeno; possono esservi altresì errori di misura, di registrazione o difetti nella strumentazione. Sarà dovere dell'attore dello studio individuare i vari *outliers* e, se possibile, le rispettive cause per poter intervenire modificando o eliminando il dato.

Un altro valore importante che può essere utilizzato in diverse circostanze è la mediana. La mediana è il valore centrale di un campione ordinato, se il numero dei valori è dispari, o la media dei due valori centrali, se il numero dei valori è pari. Uno degli utilizzi più diffusi della mediana avviene nello studio degli effetti nocivi di una certa sostanza sull'uomo. Nella ricerca della correlazione dose-effetto, la mediana rappresenta una buona misura e un riferimento per effettuare confronti tra inquinanti: la dose mediana è quella che uccide il 50% degli individui ed è chiamata $LD_{50}$ (*Lethal Dose for 50%*).

Infine, per moda si intende il valore che viene estratto più frequentemente. Anche se questo può essere distante dalla media, viene spesso usato in situazioni pratiche e può essere più utile dei valori della moda e della mediana. Si è osservato, infatti, che spesso il cervello umano tende ad assumere come valore medio quello che in realtà è solo il valore più frequente, e si assume quindi come valore medio di un campione quella che in realtà è la moda (con questo non si intende dire che sia sempre sbagliato fare un dimensionamento basandosi sul valore della moda piuttosto che su quello della media). Da notare che è possibile che vi siano più di un valore modale all'interno di un campione.

Le misure viste fin'ora ci danno un'indicazione sui valori centrali del campione e possono fornire un valore rappresentativo di quest'ultimo. Ciononostante, può risultare utile avere informazioni riguardo alla dispersione di un fenomeno attorno al valore centrale, oppure ottenere un indice di precisione dei dati. Il valore più semplice da calcolare è il *range*, definito come differenza fra valore massimo e valore minimo del campione. É una funzione non decrescente della numerosità del set di dati. Inoltre, è molto sensibile ai valori di massimi e minimi che possono essere considerati anomali anche se non classificabili come *outliers*. Per questo motivo si preferisce utilizzare il range interquartile, definito come la differenza da il 75-esimo quartile ed il 25-esimo.

La deviazione media assoluta *d* misura la distanza media assoluta dei dati del campione dalla media:

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

per un campione $x_1 ... x_n$.

La principale misura di dispersione rimane comunque la deviazione standard *S*, che è la radice quadrata della distanza media al quadrato dei valori dalla media del campione. Cioè, per un campione $x_1 ... x_n$:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

## *2.3. Convergenza in legge*

Prima di introdurre uno dei teoremi fondamentali della statistica, su cui si basano i metodi di analisi di campioni casuali, è necessario apprendere il concetto di convergenza in legge di una successione di variabili aleatorie $X_n$. Diremo che $X_n$ converge in legge a $X$, e lo indicheremo con $X_n \rightarrow X$, se e solo se $\{F_{Xn}(X_n)\} \rightarrow F_X(X)$, cioè se la successione delle finzioni di ripartizione delle variabili aleatorie Xn converge puntualmente a $F_X(.)$ per ogni punto di continuità di $F$, che prende il nome di funzione di ripartizione limite di $\{X_n\}$.

## *2.4. Teorema centrale del limite*

E' possibile ora dimostrare il seguente teorema (Lindeberg, 1922):

la distribuzione di probabilità della somma di $n$ variabili aleatorie indipendenti identicamente distribuite $(X_1....X_n)$, con media $\mu$ finita e varianza $\sigma^2$ finita e diversa da zero, converge in legge ad una distribuzione normale. In formule sarebbe:

$$\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$$

Definiamo una nuova variabile aleatoria $Z_n$:

$$Z_n = \frac{X_1 + ... + X_n - \mu n}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Dove $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è la *media campionaria* di $(X_1,...,X_n)$, mentre $Z_n$ è la *media campionaria standardizzata*. Si può dimostrare che la funzione di distribuzione di $Z_n$ converge ad una distribuzione normale con media zero e varianza unitaria, per $n$ che tende a infinito. Indicando con $F_n(x)$ la distribuzione di $Z_n$ e con $\Phi(x)$ la distribuzione della variabile aleatoria $Z \sim N(0,1)$, si avrà che:

$$Fn(x) \rightarrow \Phi(x), per\ n \rightarrow +\infty, \forall x \in \mathbb{R}$$

La distribuzione $N(0,1)$ è la funzione limite ed asintotica della media campionaria standardizzata. Questo significa che la $N(0,1)$ è il limite di $F_n(x)$ per $n$ che tende ad infinito (funzione limite) e, per $n$ grande, la $N(0,1)$ approssima la funzione di ripartizione di $F_n(x)$ (distribuzione asintotica).

Si può affermare inoltre che:

$$F_{\bar{X}}(x) \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Cioè che la distribuzione asintotica della media campionaria è una normale con media $\mu$ e varianza $\sigma^2/n$. Questo significa che la distribuzione limite di $\bar{X}$ è quella della variabile aleatoria degenere $\mu$ (costante). Il valore della varianza tende infatti a zero per $n$ che tende a infinito.

## *2.6. Stimatori e proprietà*

Una statistica è qualsiasi variabile aleatoria $T = T_n = T(X_1,..., X_n)$ che dipende dal campione $(X_1,..., X_n)$ e che non dipende da alcun parametro incognito. Inoltre, se $\tau(\theta)$ è una funzione di un parametro incognito $\theta$ della distribuzione della popolazione in esame (ad esempio la media $\mu$), considerato un campione di questa popolazione $(X_1,..., X_n)$ , uno stimatore di $\tau(\theta)$ è qualsiasi statistica $T = T_n = T(X_1,..., X_n)$ usata per stimare $\tau(\theta)$ (può darsi anche il caso in cui $\tau(\theta) = \theta$). Considerando una realizzazione $(x_1,..., x_n)$ del vettore aleatorio $(X_1,..., X_n)$, $t = T(x_1,..., x_n)$ prende il nome di stima di $\theta$.

Il più classico esempio di statistica usata come stimatore è la media campionaria:

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

Che può servire a stimare la media $\mu = E[X]$ di una popolazione.

Si vedranno ora tre proprietà fondamentali degli stimatori.

### *2.6.1 Correttezza*

Uno stimatore $T$ si dice corretto se $E[T] = \theta$, cioè se la sua media coincide col valore del parametro stimato. Se uno stimatore non è corretto si dice distorto.

Ad esempio, la media campionaria è uno stimatore corretto della media, mentre la quantità:

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

è uno stimatore corretto della varianza nel caso la media $\mu$ non fosse nota, e si dice varianza campionaria.

Se la media fosse invece nota, lo stimatore corretto della varianza sarebbe:

$$S_0^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Perché, essendo nota la media, $S_0^2$ può essere considerato una statistica e può essere usato per stimare $\sigma^2$.

Uno stimatore T si dice asintoticamente corretto se:

$$\lim_{n\to\infty} E[T]=\lim_{n\to\infty} E[T(X_{1,\dots},X_n)]=\theta$$

Se uno stimatore è corretto, allora è anche asintoticamente corretto ma non viceversa.

### *2.6.2 Consistenza*

Uno stimatore T è consistente se converge in legge al parametro incognito, cioè se è vero che:

$$\lim_{n\to\infty} Prob[|T-\theta|<\epsilon]=1$$

Con ε piccolo a piacere. Inoltre, se uno stimatore asintoticamente corretto ha varianza che tende a zero per n → ∞, cioè $\lim_{n\to\infty} var[T]=0$ , allora è consistente.

### 2.6.3 Efficienza

Si chiama errore quadratico medio di Q rispetto a θ il valore $E[(Q-\theta)^2]$ e si indica come $MSE_\theta(Q)$. Tra due stimatori di θ, Q e T, si dice che Q è più efficiente di T se vale la disuguaglianza:

$$MSE_\theta(Q)=E_\theta[(Q-\theta)^2]\le MSE_\theta(T)=E_\theta[(T-\theta)^2]$$

Q inoltre si dice ottimale se il suo errore quadratico medio rispetto al parametro è minore dell'errore quadratico di ogni altro stimatore di θ.

Ad esempio, in una popolazione normale del tipo $N(\mu,\sigma^2)$, con media e varianza ignote, le statistiche $\bar{X}$ e S2 sono gli stimatori ottimali di μ e $\sigma^2$.

## 2.7. Ricerca di uno stimatore

### 2.7.1 Metodo dei momenti

Si supponga di avere un campione casuale $(X_1,\dots,X_n)$, estratto da una popolazione con distribuzione $F$ che dipende da un parametro incognito $\theta$, che si vuole stimare. Il metodo dei momenti utilizza i momenti campionari per stimare il parametro incognito, se questo può essere espresso in funzione dei momenti di qualsiasi ordine, cioè se esiste una

relazione del tipo $\theta = g(E[X],E[X2],...,E[X^r])$. I momenti campionari $M_k = \frac{\sum_i X_i^k}{n}$, per

$k = 1,.. r$, vengono usati per stimare i momenti di ordine $k$ ed in seguito il parametro $\theta$.

### 2.7.2 Metodo della massima verosimiglianza

La funzione di verosimiglianza $L(\theta; x_1,...,x_n)$ è la densità congiunta $f_{x1,x2,...,xn}(x_1,...,x_n; \theta)$ di n variabili aleatorie $X_1,...X_n$, in funzione di $\theta$. Se le n variabili sono un campione casuale di una popolazione la cui densità congiunta è nota a meno di un parametro $\theta$, allora $L(\theta; x_1,...,x_n) = f_{x1}(x_i; \theta)*...*f_{xn}(x_n; \theta)$.

La stima di massima verosimiglianza di θ è il valore $\hat{\theta} = \theta(x_{1,...}, x_n)$ che rende massima $L(\theta; x_1,...,x_n)$, mentre lo stimatore massimo verosimile è la statistica $\hat{\Theta} = \theta(X_{1,...}, X_n)$ .

### 2.8. Vettori aleatori

Un vettore aleatorio $n$-dimensionale, detto anche variabile aleatoria n-dimensionale, è una funzione $\boldsymbol{X}: \Omega \rightarrow R^2$ tale per cui ogni sua componente sia una variabile aleatoria (sia essa continua, discreta o mista). Un vettore aleatorio è un vettore colonna che spesso viene indicato come vettore riga trasposto, per comodità:

$$\boldsymbol{X} = [X_1,...,X_n]'$$

Il vettore media di un vettore aleatorio $n$-dimensionale (o *vtan*) è definito come:

$$\boldsymbol{\mu_x} = (E[X_1],...,E[X_n])'$$

Se si ponesse un vettore $\boldsymbol{Y} = (Y_1,..., Y_n)$ tale che $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$, dove $A$ e $\boldsymbol{b}$ fossero rispettivamente una matrice $k \times n$ e un vettore colonna $k \times 1$ di numeri reali, la media di **Y** sarebbe così calcolata:

$$\boldsymbol{\mu_y} = A\boldsymbol{\mu_x} + \boldsymbol{b}$$

Si consideri ora il caso in cui due variabili aleatorie $X_1$ e $X_2$ siano le componenti di un vettore aleatorio bi-dimensionale, $\boldsymbol{X} = (X_1 , X_2)'$, e si supponga nota la densità congiunta delle due variabili, $f_X(.,.)$. La generica funzione $\boldsymbol{Y} = g(X_1 , X_2)'$ assume valori in corrispondenza degli stessi punti in cui $\boldsymbol{X}$ assume valori, cioè hanno lo stesso dominio. La media di $g(X_1 , X_2)'$ è data da:

- Caso discreto: $E[g(X_{1,}X_2)]=\sum_{(X_{1,}X_2)} g(X_{1,}X_2)f_X(x_{1,}x_2)$
- Caso continuo: $E[g(X_{1,}X_2)]=\int\int_{R^2} g(x_{1,}x_2)f_X(x_{1,}x_2)dx_1dx_2$

### 2.8.1 Funzioni generatrici di momenti

Sia $\boldsymbol{t} = (t_1,..., t_n)'$ un vettore di variabili reali. La funzione di generatrice di momenti (*fgm*) di un *vtan* $\boldsymbol{X} = (X_1,...,X_n)'$ è così definita:

$$m_{[X_1...X_n]}(t_1...t_n)=E\left[\exp\left(\sum_{i=1}^n t_i X_i\right)\right]=E[\exp(t'X)]$$

Inoltre, è possibile dimostrare che se le componenti di un vettore aleatorio sono indipendenti, allora la funzione generatrice di momenti del vettore è data dal prodotto delle singole funzioni generatrici delle componenti, cioè:

$$m_{[X_1...X_n]}(t_1...t_n)=\prod_{i=1}^n E\left[\exp(t_i X_i)\right]=m_{X_1}(t_1)...m_{X_n}(t_n)$$

Se invece si volesse considerare la *fgm* di una trasformazione lineare del tipo:

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$$

dove $A$ è una matrice $k \times n$ e $\boldsymbol{b}$ un vettore colonna $k \times 1$ di numeri reali, si può dimostrare che la funzione generatrice di momenti di $\boldsymbol{Y}$ data da:

$$m_Y(t)=\exp(t^T b)\,m_X(A^T t)$$

dove $t$ è un vettore colonna $k \times 1$.

### 2.8.2 Momenti centrali

Si consideri il caso in cui il vettore aleatorio $\boldsymbol{X}$ sia formato da due componenti, $\boldsymbol{X} = (X_1,X_2)'$.

Il momento centrale misto di ordine $|r| = r1 + r2$ di $\boldsymbol{X}$ è dato dal valore:

- Caso discreto:

$$\mu_{r_1,r_2}=E[(X_1-\mu_{X_1})^{r_1}(X_2-\mu_{X_2})^{r_2}]=\sum_{(x_1,x_2)}(x_1-\mu_{X_1})^{r_1}(x_2-\mu_{X_2})^{r_2}f_X(x_{1,}x_2)$$

- Caso continuo:

$$\mu_{r_1,r_2} = E[(X_1 - \mu_{X_1})^{r_1}(X_2 - \mu_{X_2})^{r_2}] = \int_{R^2} (x_1 - \mu_{X_1})^{r_1}(x_2 - \mu_{X_2})^{r_2} f_X(x_1, x_2)\, dx_1 dx_2$$

Il momento centrale misto di ordine due di $\boldsymbol{X}$ è detto covarianza di $X_1$ e $X_2$, se esiste, ed è dato dalla formula:

$$\mu_{1,1} = \sigma_{12} = \sigma_{X_1 X_2} = cov[X_1, X_2] = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$$

Si può dimostrare che $cov\ [X_1, X_2] = E[X_1X_2] - \mu_{X1}\mu_{X2}$. Inoltre, due variabili aleatorie $X_1$ e $X_2$ si dicono correlate positivamente o negativamente se la loro covarianza è maggiore o minore di zero rispettivamente; se la loro varianza è uguale a zero si dicono scorrelate, cosa che succede anche per variabili indipendenti.

La covarianza gode delle seguenti proprietà:

- $cov\ [X,X] = var[X]$;
- $cov\ [X_1, X_2] = cov\ [X_2, X_1]$;
- $cov\ [X_1,X_1 + X_2] = var[X_1] + cov\ [X_1, X_2]$;
- $cov\ [a_1X_1 + b_1, a_2X_2 + b_2] = a_1a_2cov\ [X_1, X_2]$;
- $cov\ [a, X] = 0$.

Sapere cos'è la covarianza è di fondamentale importanza per comprendere uno strumento fondamentale nello studio di dati ambientali e non solo: la matrice di covarianza.

Si consideri ancora una volta il caso bi-variato. La matrice di covarianza di un vettore aleatorio X = $(X_1 , X_2)'$ di media $(E[X_1],E[X_2])'$ è definita come la matrice *2 x 2* formata dai momenti centrali di ordine *2*, cioè:

$$C_X = \begin{bmatrix} var[X_1] & cov(X_1, X_2) \\ cov(X_2, X_1) & var[X_2] \end{bmatrix} = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1 X_2} \\ \sigma_{X_1 X_2} & \sigma^2_{X_2} \end{bmatrix}$$

Generalizzando al caso *n*-dimensionale, la matrice di covarianza di un vettore $\boldsymbol{X} = (X_1,..., X_n)'$ è una matrice *n x n* della forma:

$$\boldsymbol{C}_X = [cov[X_i, X_j]]$$

dove *i* e *j* variano tra *1* e *n*.

La matrice di covarianza può anche essere calcolata nel modo seguente:

$$\boldsymbol{C}_X = E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = E[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

La matrice di covarianza gode delle seguenti proprietà:

- è simmetrica;
- è semi-definita positiva, cioè $\forall a \in R^n, a^T C_X a \geq 0$ . Se il vettore **X** non

contiene variabili aleatorie con probabilità 1 (variabili degeneri), allora $\mathbf{C_X}$ è definita positiva.

- Se **a** è un vettore di costanti, allora $var[\boldsymbol{X} + \boldsymbol{a}] = \boldsymbol{C}_{\boldsymbol{X}+\boldsymbol{a}} = \boldsymbol{C}_{\boldsymbol{X}} = var[\boldsymbol{X}]$.

Infine, la legge di propagazione della covarianza dimostra che se X è un vettore colonna *n x 1*, A una matrice *k x n* e b un vettore di costanti reali *k x 1*, allora la matrice di covarianza della variabile trasformazione lineare $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ è data da:

$$\boldsymbol{C}_{\boldsymbol{Y}} = A\,\boldsymbol{C}_{\boldsymbol{X}}\,A^T = var[A\boldsymbol{X} + \boldsymbol{b}] = A\,var[\boldsymbol{X}]\,A^T$$

### 2.8.3 Coefficiente di correlazione lineare

Il coefficiente di correlazione lineare tra due variabili dà un'indicazione di quanto le due variabili, se poste rispettivamente sulle ascisse e sulle ordinate di un piano cartesiano, si avvicinino a una retta. Cioè, come dice il nome stesso, il coefficiente di correlazione lineare quantifica in qualche modo quanto la relazione tra le due variabili sia approssimabile con una retta.

Questo tipo di coefficiente viene generalmente indicato con la lettera greca $\rho$ con al pedice le due variabili in esame. Date due variabili $X_1$ e $X_2$, il coefficiente di correlazione lineare è così calcolato:

$$\rho_{X_1 X_2} = \frac{cov[X_1, X_2]}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma_{X_1 X_2}}{\sigma_1 \sigma_2}$$

se $cov[X_1,X_2]$, $\sigma_{X1}$ e $\sigma_{X2}$ esistono e queste due ultime sono maggiori di zero.

Il valore di $\rho$ varia tra *-1* e *+1*. Se le due variabili sono linearmente dipendenti, cioè se per esempio si avesse che $X_1 = aX_2 + b$, allora $\rho$ assume valore unitario in modulo e segno dipendente dal coefficiente angolare *a* della retta (*-1* se negativo e *+1* se positivo). Se due variabili $X_1$ e $X_2$ sono indipendenti, allora $\rho_{X1,X2} = 0$.

# 3. STIMA INTERVALLARE E TEST D'IPOTESI

*"Senza deviazione dalla media, il progresso non è possibile"*

Frank Zappa

Nell'applicare tecniche statistiche nello studio di una popolazione, si giunge a stimare uno o più valori di determinati parametri di tale popolazione, come ad esempio la media o la varianza. Si sa però che, a prescindere dalla tecnica di stima utilizzata, il valore dello stimatore puntuale può essere errato. Anzi, la probabilità che il valore di uno stimatore coincida col valore (sconosciuto) del parametro stimato è uguale a zero. Da qui nasce la necessità di ricercare i così detti intervalli di confidenza per gli stimatori, cioè degli intervalli di valori che contengano il valore del parametro ricercato, con una certa probabilità. In altre parole, al posto di ricercare una statistica che stimi il parametro $\theta$, ad esempio, si cerca un intervallo definito da due statistiche $a$ e $b$, che sono gli estremi dell'intervallo che contiene il valore del parametro con una data probabilità.

## *3.1. Quantità pivotale*

Prima di addentrarsi nell'argomento degli intervalli di confidenza e dei test d'ipotesi, si vuole fornire la definizione di quantità pivotale, di cui in seguito si comprenderà l'importanza:

si consideri una popolazione la cui distribuzione dipende dal parametro $\theta$. Si pensi di estrarre un campione casuale $(X_1,...X_n)$. Se la variabile aleatoria $Y = (X_1,...X_n;\theta)$ dipende da $\theta$ ma non la sua funzione di densità di probabilità, allora $Y$ si dice pivot o quantità pivotale.

Ad esempio, estraendo da una popolazione normale di media $\mu$ e varianza $\sigma^2$, cioè $N(\mu,\sigma^2)$, si ha che:

1. $t = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ è una *t di Student* con $n$ gradi di libertà;
2. $\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{(n-1)}{\sigma^2} S^2$ è una chi-quadrato con $n-1$ gradi di libertà;
3. $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ è una normale standard $N(0,1)$.

### *3.2. Intervallo di confidenza per la media nota la varianza*

Chiameremo *a* e *b* rispettivamente l'estremo inferiore e superiore dell'intervallo di confidenza per un certo parametro $\theta$. Per ora non vi è alcuna certezza che questo intervallo contenga il valore del parametro. Non cadremo in errore, quindi, affermando che vi sia una certa probabilità $\alpha$ che l'intervallo non contenga il parametro in questione. Potremo quindi scrivere:

$$Pr[a \leq \theta \leq b] = 1 - \alpha \quad (1)$$

con $0 < \alpha < 1$. Da notare che mentre il parametro $\theta$ è una costante, il suo stimatore $\hat{\theta}$, così come gli estremi dell'intervallo *a* e *b*, sono variabili casuali. La quantità $(1- \alpha)$ è detta livello di confidenza. I limiti di confidenza *a* e *b* dipendono dalla distribuzione di $\hat{\theta}$. La (1) definisce un intervallo di confidenza detto bilatero, perché caratterizzato da due estremi. In alcuni casi può risultare utile definire intervalli con limite superiore o inferiore, in tal caso scriveremo, rispettivamente:

$$Pr[a \leq \theta] = 1 - \alpha \quad \text{o} \quad Pr[\theta \leq b] = 1 - \alpha$$

Dove, come al solito, $0 < \alpha < 1$.

Il teorema centrale del limite ci dice che la deviazione standard $\sigma_{\bar{X}}$ della media campionaria $\bar{X}$ tende al valore:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (2)$$

per *n* che tende a infinito. Per *n* grandi, la distribuzione della media campionaria approssima una normale, anche se il campione non è estratto da una distribuzione normale. Se, invece, il campione in esame venga estratto da una distribuzione normale, la distribuzione della media campionaria è proprio una normale, qualsiasi sia la numerosità del campione. Possiamo ora usare la (2) per trovare gli estremi dell'intervallo di confidenza per la media $\mu$ della popolazione, infatti:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \quad (3)$$

Per esempio, utilizzando le tavole della distribuzione normale ed imponendo un livello di confidenza $(1- \alpha) = 0,95$, potremo affermare che:

$$Pr[-1,96 \leq Z \leq 1,96] = \int_{-1,96}^{1,96} F_Z(z)\,dz = 0,95 \quad (4)$$

Facendo un'analogia con quanto affermato prima, i limiti di confidenza *a* e *b* sono posti uguali rispettivamente a *-1,96* e *1,96* e, come già accennato, dipendono dalla distribuzione della variabile casuale *Z*. La (4) esprime il fatto che vi è una probabilità del

*95%* che la variabile *Z* sia inclusa nell'intervallo *[-1,96 ; 1,96]*.

Si può ora fornire la seguente definizione: dato un campione di numerosità *n*, con media campionaria $\bar{X}$ , estratto da una popolazione con deviazione standard $\sigma$ nota e media $\mu$ ignota, l'intervallo di confidenza bilatero con livello *100(1- α)%* per la media $\mu$ è dato da:

$$(\bar{X} - \Phi^{-1}(\alpha/2)\sigma/\sqrt{n}, \bar{X} + \Phi^{-1}(\alpha/2)\sigma/\sqrt{n})$$

Dove $\Phi^{-1}(.)$ indica il quantile della normale standard.

## *3.3. Intervallo di confidenza per la media con varianza ignota*

Nel caso stessimo cercando un intervallo di confidenza di un parametro appartenente a una popolazione di cui non si conoscono ne la media ne la varianza, l'approccio da applicare è differente rispetto al caso precedente. In particolare, verrà applicata la distribuzione *t di Student* per la variabile *T*, che rappresenta la media campionaria standardizzata come nella (3), solo che al posto della varianza si utilizza il valore della varianza campionaria $\hat{S}$ :

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim t_{n-1}$$

che ha una distribuzione *t di Student* con *(n-1)* gradi di libertà. Alla stregua del caso precedente, si da quindi la seguente definizione: dati i valori $\bar{X}$ e $\hat{S}$ , che rappresentano rispettivamente la media e la deviazione standard campionarie di un campione di numerosità *n* estratto da una certa popolazione, l'intervallo bilatero con livello *100(1- α)%* per la media $\mu$ della popolazione è dato da:

$$(\bar{X} - t_{n-1}(\alpha/2)\hat{S}/\sqrt{n}, \bar{X} + t_{n-1}(\alpha/2)\hat{S}/\sqrt{n})$$

dove $t_{n-1}(\alpha/2)$ è l' *α/2*-quantile della *t di Student* con *n-1* gradi di libertà.

## *3.4. Intervallo di confidenza per la varianza con media ignota*

Consideriamo un campione casuale di *n* variabili indipendenti e identicamente distribuite secondo una normale: $X_i$, con $i = 1,2...n$, con media $\mu_i$ e varianza $\sigma^2_i$. Ora supponiamo di standardizzare le $X_i$ sottraendo le medie e dividendo per la deviazione standard, ottenendo $Z_j$, con $j = 1,2,...n$, con media nulla e varianza unitaria. Si consideri ora la variabile casuale data dalla somma delle $Z_i$ al quadrato:

$$\chi^2 = Z_1^2 + Z_2^2 + ... + Z_n^2$$

La funzione generatrice di momenti di $\chi^2$ è:

$$M_{\chi^2}(t) = E[\exp(t\chi^2)] = \Pi_{i=1}^{n} E[\exp(tZ_i^2)] = \left(\frac{1/2}{1/2 - t}\right)^{n/2}$$

Si può affermare che la somma dei quadrati di variabili normali standard indipendenti ha distribuzione chi-quadro con *n* gradi di libertà.

Si consideri ora un campione casuale $X_1, X_2 ... X_n$ di variabili indipendenti normalmente distribuite con media comune $\mu$, varianza $\sigma^2$ e media campionaria $\bar{X}$ . Si può affermare che:

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad (5)$$

è uno stimatore corretto della varianza. Consideriamo ora la quantità:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

essendo $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ . Dividendo per la varianza $\sigma^2$ e sostituendo $(n-1)\hat{S}^2$ della (5), si ottiene:

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}$$

Il primo termine è quindi distribuito come una chi-quadro con *n* gradi di libertà e l'ultimo termine come una chi-quadro con un grado di libertà, quindi:

$$(n-1)\frac{\hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Questo risultato viene utilizzato per trovare finalmente l'intervallo di confidenza per la varianza di una popolazione normale. Infatti l'intervallo di confidenza bilatero con livello *(1-α)* è calcolato tramite la seguente:

$$Pr[\chi^2_{n-1}(1-\alpha/2) \le \frac{(n-1)\hat{S}^2}{\sigma^2} \le \chi^2_{n-1}(\alpha/2)] = 1-\alpha$$

Quindi, essendo $\hat{S}^2$ la varianza campionaria di un campione di numerosità *n* estratto da una popolazione normalmente distribuita con varianza ignota, l'intervallo bilatero al *100(1-α)%* per la varianza $\sigma^2$ è dato da:

$$\left(\frac{(n-1)\hat{S}^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)\hat{S}^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$$

## *3.5. Test d'ipotesi*

Strettamente correlati agli intervalli di confidenza sono i test d'ipotesi. Questi test possono riguardare i parametri di una popolazione (parametrici) o la forma della sua distribuzione di probabilità (non parametrici). I test prevedono di verificare se una dichiarazione (o ipotesi) fatta su una certa popolazione sia accettabile o no, con un certo margine di errore. Questa tecnica viene usata in numerosi campi dell'ingegneria e della scienza in generale: per esempio è possibile verificare se sia possibile assumere che due popolazioni in studio abbiano la stessa media, oppure che una certa distribuzione sia modificata nel tempo (ad esempio il regime delle piene di un fiume a seguito dei cambiamenti climatici). In ogni caso, come al solito, viene studiato un campione della popolazione in esame e vengono dichiarate le due ipotesi statistiche, che chiameremo ipotesi nulla e alternativa. Inoltre, viene scelto anche l'errore che si è disposti a commettere accettando o scartando l'ipotesi. Questo errore è detto livello di significatività $\alpha$, ed è il complementare a 1 del livello di confidenza del paragrafo precedente. L'ipotesi nulla è quella che dobbiamo testare, in base alle osservazioni disponibili. Può essere che queste ci portino a non scartare l'ipotesi nulla, che quindi verrà accettata con un certo margine di errore. In caso contrario, verrà assunta come veritiera l'ipotesi alternativa.

### 3.5.1 Procedura

Come già anticipato in precedenza, il test si basa sul confronto tra due ipotesi: ipotesi nulla e alternativa. Per esempio, potremmo voler verificare se in un determinato sito le concentrazioni di un certo inquinante stiano diminuendo o no a seguito di un intervento di bonifica. In questo caso si potrebbero confrontare le medie di concentrazione dell'anno corrente $\mu_1$ e di quello passato $\mu_2$. Il confronto delle sole medie campionarie può fornire una prima idea della situazione però non deve trarre in inganno: non è vero, infatti, che la media campionaria coincide col valore reale della media della popolazione, e bisogna quindi tener conto anche di questo. Quindi il primo passo è proprio quello di formulare l'ipotesi nulla $H_0$, che sarà l'affermazione da verificare. Riguardo all'esempio citato, potremmo esprimere l'ipotesi:

$$H_0 : \ \mu_1 - \mu_2 = 0.$$

Il secondo passo è quello di formulare l'ipotesi alternativa, $H_a$, che sarà quella da

accettare se l'ipotesi nulla risultasse errata. Nel nostro caso potrebbe essere:

$$H_a : \ \mu_1 - \mu_2 < 0.$$

Dopodiché sarà necessario specificare una statistica test, che in questo caso potrebbero essere le medie campionarie $\bar{X}_1 e \bar{X}_2$ . Avremo poi bisogno di sapere la distribuzione della statistica test e della popolazione da cui è stato estratto il campione, anche in base a quanto detto nel capitolo precedente riguardo gli intervalli di confidenza. Il passo successivo sarà quello di determinare la regione critica, che è l'insieme dei valori della statistica test che ci porterebbero a rifiutare l'ipotesi nulla. Per questo sarà necessario determinare il livello di significatività $\alpha$, che è la probabilità di commettere l'errore di rifiutare l'ipotesi nulla essendo questa vera. Infine, in base ai dati disponibili verificheremo se il valore della statistica test ricade o meno nella regione critica.

Il test sopra citato è detto test a una coda (sinistra), perché la regione critica riguarda solo una delle due code della distribuzione. Se l'ipotesi alternativa fosse stata $H_a : \mu_1 - \mu_2 \neq 0$, sarebbe stato un test a due code.

Se indicassimo con $S$ la statistica test e con $R$ la regione critica, la probabilità di rifiutare l'ipotesi nulla $H_0$ sarebbe:

$$Pr[S \in R / H_0 = vera] = \alpha \quad .$$

Se l'ipotesi nulla viene scartata quando dovrebbe essere accettata, in quanto vera, si sta commettendo un errore del primo tipo:

$$Pr[Errore\ Tipo\ I] = Pr[S \in R / H_0 = vera] = \alpha$$

Se invece l'ipotesi nulla viene accettata quando dovrebbe essere scartata, in quanto falsa, si sta commettendo un errore del secondo tipo:

$$Pr[Errore\ Tipo\ II] = Pr[S \in A / H_0 = vera] = \beta$$

Dove con A è la regione di accettazione, complementare alla regione critica, tale che $A \cap R = \emptyset$ e $A \cup R = \Omega$ , dove Ω è lo spazio dei parametri e comprende tutti i valori che S può assumere.

Questi sono i casi in cui vengono commessi errori di giudizio nel test. Al contrario, la probabilità di accettare a ragione l'ipotesi nulla, in quanto vera, è complementare a quella di commettere un errore del primo tipo:

$$Pr[S \in A / H_0] = 1 - \alpha$$

Allo stesso modo, la probabilità di rifiutare l'ipotesi nulla essendo questa errata è complementare a quella di commettere un errore del secondo tipo. Quindi:

$$Pr[S \in R / H_a] = 1 - \beta$$

Il complementare a *1* di $\beta$ è detto anche potenza del test.

### 3.5.2 Valore P

Il valore p o *p-value* è definito come il più piccolo valore dell'ampiezza del livello di significatività α che permette di rifiutare l'ipotesi nulla, dato il valore calcolato della statistica test *S*.

Abbiamo visto come per determinare gli intervalli di confidenza sia necessario stabilire un livello di confidenza *(1-α)* e calcolare la statistica test: in questo modo la regione critica e di accettazione sono univocamente determinate. Si ricordi che α indica la probabilità che la statistica test cada nella regione critica, nonostante l'ipotesi nulla sia vera; se si scegliesse un α alto, la regione critica sarebbe più ampia e sarebbe più probabile rifiutare $H_0$. Ora si immagini di non scegliere un livello di significatività α a priori, ma che si voglia trovare il valore di α più piccolo possibile che permetta di rifiutare l'ipotesi nulla $H_0$, dato il valore della statistica test, cioè il valore di α che minimizzi l'ampiezza della regione critica. Questo equivale a trovare la densità di probabilità *F(S)* corrispondente alla statistica test *S* e porre (nel caso di distribuzioni normali):

- valore $p = 1 – F(S)$ se il test è a una coda destra;
- valore $p = 2 – 2F(S)$ se il test è a due code;
- valore $p = F(S)$ se il test è a una coda sinistra.

Se il valore del valore *p* è alto sarebbe come dire che l'errore del primo tipo ha una probabilità alta di verificarsi, quindi saremmo più portati ad accettare l'ipotesi nulla. Alternativamente, possiamo pensare di voler confrontare il valore *p* con un livello di significatività α imposto a priori. Sarà quindi sufficiente confrontare i due valori per decidere se accettare o rifiutare l'ipotesi nulla: se, infatti, il valore *p* assume valori più elevati di α, significa che molto probabilmente la statistica test cadrà nella regione di accettazione ed $H_0$ verrà quindi accettata, per valori bassi di *p* succede il contrario.

Si vedrà ora come impostare alcuni tra i test più utilizzati.

### 3.5.3 Test su una popolazione

#### 3.5.3.1 Test per la media di una popolazione con varianza nota

Impostiamo il seguente test:

$H_0 : \mu = \mu_0$

$H_a :$ 1) $\mu \neq \mu_0$

2) $\mu > \mu_0$

3) $\mu < \mu_0$

In questo caso consideriamo come stimatore della media la media campionaria $\bar{X}_n$ e come statistica test la media campionaria standardizzata: $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ .

A seconda dei tre casi si avranno tre differenti regioni critiche R, imponendo l'uguaglianza $\alpha = 1 - \gamma$:

1) $R = [\bar{X}_n \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\frac{1-\gamma}{2})] \cup [\bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\frac{1+\gamma}{2})]$

2) $R = [\bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\gamma)]$

3) $R = [\bar{X}_n \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\gamma)]$ .

#### 3.5.3.2 Test per la media di una popolazione normale con varianza ignota

Anche in questo caso viene impostato il test:

$H_0 : \mu = \mu_0$

$H_a :$ 1) $\mu \neq \mu_0$

2) $\mu > \mu_0$

3) $\mu < \mu_0$

Si supponga di avere a che fare con una variabile aleatoria X gaussiana, di cui non si conoscono ne la media ne la varianza, quindi $X \sim N(\mu, \sigma^2)$ . La statistica test $t_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ è una *t di Student* con *n-1* gradi di libertà. A seconda del tipo di ipotesi

alternativa, si avranno tre regioni critiche:

1. $R=[\bar{X}_n \leq \mu_0 - \frac{S}{\sqrt{n}} t_{n-1}(\frac{1-\gamma}{2})] \cup [\bar{X}_n \geq \mu_0 + \frac{S}{\sqrt{n}} t_{n-1}(\frac{1+\gamma}{2})]$
2. $R=[\bar{X}_n \geq \mu_0 + \frac{S}{\sqrt{n}} t_{n-1}(\gamma)]$
3. $R=[\bar{X}_n \leq \mu_0 - \frac{S}{\sqrt{n}} t_{n-1}(\gamma)]$

### 3.5.3.3 Test per la varianza di una popolazione con media nota

Si consideri ora il test nella forma:

$H_0 : \sigma = \sigma_0$

$H_a :$ 1) $\sigma \neq \sigma_0$

2) $\sigma > \sigma_0$

3) $\sigma < \sigma_0$

Il test si basa sull'osservazione che, dato un campione di *n* estrazioni di una variabile aleatoria, $\sum_{i=1}^{n} (\frac{x_i - \mu}{\sigma^2})^2 \sim \chi_n^2$ perché il teorema del limite centrale ci assicura che il termine tra parentesi è una normale di parametri *N (0,1)* e quindi la somma di *n* normali è una chi-quadro con *n* gradi di libertà.

Inoltre, si può dimostrare che per un campione di numerosità maggiore o uguale a *100* vale la relazione:

$$\frac{\sum_{i=1}^{n} (\frac{x_i - \mu}{\sigma^2})^2 - n}{\sqrt{2n}} \sim N(0,1)$$

Quindi le tre regioni critiche individuate in corrispondenza delle tre ipotesi alternative sarebbero:

1. $R=[S^2 \leq \sqrt{2/n}\,\Phi^{-1}(\frac{1-\gamma}{2}) - 1 ; S^2 \geq \sqrt{2/n}\,\Phi^{-1}(\frac{1+\gamma}{2}) + 1]$
2. $S^2 \geq \sqrt{2/n}\,\Phi^{-1}(\gamma) + 1$
3. $S^2 \leq \sqrt{2/n}\,\Phi^{-1}(\gamma) - 1$

Dove $S^2$ è la varianza campionaria $S^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$ , dove *n* è la numerosità del campione.

### 3.5.3.4 Test per la varianza di una popolazione con media ignota

Si consideri ancora il test:

$$H_0 : \sigma = \sigma_0$$

$$H_a : \text{1) } \sigma \neq \sigma_0$$

$$\text{2) } \sigma > \sigma_0$$

$$\text{3) } \sigma < \sigma_0$$

Si supponga che la variabile aleatoria in esame *X* sia una gaussiana del tipo $X \sim N(\mu, \sigma^2)$ . Denominando con $S^2$ la varianza campionaria su un campione di n realizzazioni, sappiamo che la quantità pivotale $\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$ è una variabile aleatoria con distribuzione chi-quadro con *n-1* gradi di libertà. Le regioni critiche saranno del tipo (ricordando che la chi-quadro non ammette valori negativi nel dominio e che non è simmetrica):

1. $R = [S^2 \leq \frac{\sigma_0^2 \chi^2_{n-1}(\frac{1-\gamma}{2})}{n-1}; S^2 \geq \frac{\sigma_0^2 \chi^2_{n-1}(\frac{1+\gamma}{2})}{n-1}]$
2. $R = [S^2 \geq \frac{\sigma_0^2 \chi^2_{n-1}(\gamma)}{n-1}]$
3. $R = [S^2 \leq \frac{\sigma_0^2 \chi^2_{n-1}(1-\gamma)}{n-1}]$

### 3.5.3.5 Test (non parametrico) di Wilcoxon

A differenza dei test visti fin'ora, si propone un approccio non parametrico per i test su una singola popolazione. I test non parametrici si chiamano così perché indipendenti dalla distribuzione della popolazione campionata.

Il test di Wilcoxon si applica alla media o alla mediana di una popolazione e considera un test del tipo:

$$H_0 : \mu \leq C$$

$$H_a : \mu > C$$

applicato a popolazioni di cui sia stata verificata la simmetria, dove *C* è un valore appartenente all'insieme dei numeri reali. Il test prevede due tipi differenti di

implementazione a seconda se il campione sia formato da un numero di dati $n \leq 20$ o da $n > 20$.

a) Campione poco numeroso ( n ≤ 20 )

Innanzi tutto va effettuata la correzione dei dati nel modo che segue:

$$d_i = C - Xi$$

Ottenendo tante deviazioni $d_i$ quante sono i dati. La numerosità del campione va poi ridotta eliminando i valori nulli di *d*.

Considerando i valori assoluti della deviazione $| d_i |$ e partendo dal più piccolo, si assegna il rango a ciascun dato. In seguito, viene assegnato segno al rango a seconda se il valore corrispondente di $d_i$ sia positivo o negativo. Infine, si calcola la somma R dei ranghi positivi e si determina il valore critico $w_\alpha$ tramite la Tavola 1 di seguito riportata.

| n | w.01 | w.05 | w.10 | w.20 |
|---|---|---|---|---|
| 4 | 0 | 0 | 1 | 3 |
| 5 | 0 | 1 | 3 | 4 |
| 6 | 0 | 3 | 4 | 6 |
| 7 | 1 | 4 | 6 | 9 |
| 8 | 2 | 6 | 9 | 12 |
| 9 | 4 | 9 | 11 | 15 |
| 10 | 6 | 11 | 15 | 19 |
| 11 | 8 | 14 | 18 | 23 |
| 12 | 10 | 18 | 22 | 28 |
| 13 | 13 | 22 | 27 | 33 |
| 14 | 16 | 26 | 32 | 39 |
| 15 | 20 | 31 | 37 | 45 |
| 16 | 24 | 36 | 43 | 51 |
| 17 | 28 | 42 | 49 | 58 |
| 18 | 33 | 48 | 56 | 66 |
| 19 | 38 | 54 | 63 | 74 |
| 20 | 44 | 61 | 70 | 82 |

*Tavola 1. Quantili del test del segno del rango di Wilcoxon*

In cui α rappresenta il livello di confidenza. Se:

$$R < w_\alpha$$

e

$$R > [n\,(n+1)/2 - w_\alpha]$$

L'ipotesi nulla può essere rigettata. In caso contrario, va calcolato il numero *m* di campioni necessario a raggiungere un errore di falsa accettazione:

$$m = \frac{s^2 \left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\left(\mu_1 - C\right)} + 0{,}5 \left(Z_{1-\alpha}\right)^2$$

Dove *s* è la deviazione standard campionaria, *Z* è il quantile della normale

standard *N (0,1)*, $\alpha$ è il livello di confidenza mentre $\beta$ è la probabilità di commettere errore del secondo tipo.

Se *1,16m < n* si accetta l'errore di falsa accettazione, cioè si respinge $H_0$ ma si accetta di stare commettendo un errore del primo tipo, quindi la media è probabilmente minore di *C*.

Altrimenti, se non viene soddisfatta nessuna condizione, si può affermare che la media è probabilmente inferiore a *C* ma il campione non è abbastanza numeroso per fare ulteriori considerazioni.

b) Campione numeroso (n > 20)

La procedura in questo caso è la stessa fino alla determinazione del valore di *w*, che viene invece calcolato nel modo seguente:

$$w = \frac{n(n+1)}{4} + Z_p \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

dove $p = 1 - \alpha$, *Z* ed *n* hanno lo stesso significato di cui sopra.

Se $R < w$ si rigetta l'ipotesi nulla. Altrimenti si procede come nel caso precedente.

### 3.5.4 Test di confronto fra due popolazioni

#### 3.5.4.1 Test di confronto tra medie di due popolazioni con varianza nota

Si ponga il caso che si abbiano due campioni appartenenti a due popolazioni differenti, A e B, descritte da due variabili aleatorie di cui si conosce la varianza, rispettivamente $\sigma^2_A$ e $\sigma^2_B$. Si può dimostrare che la statistica test

$$\frac{\bar{X}_A - \bar{X}_B - (\mu_a - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0,1)$$

può essere approssimata da una normale con media zero e varianza unitaria.

Consideriamo il test:

$$H_0 : \mu_A = \mu_B$$

$$H_a : 1) \ \mu_A \neq \mu_B$$

2) $\mu_A > \mu_B$

3) $\mu_A < \mu_B$

A seconda di quale delle tre ipotesi alternative si stiano considerando, le regioni critiche saranno della forma:

1. $$R=[\bar{X}_A-\bar{X}_B\leq(\mu_A-\mu_B)-\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}\Phi^{-1}(\frac{\gamma-1}{2});]$$
$$[\bar{X}_A-\bar{X}_B\geq(\mu_A-\mu_B)+\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}\Phi^{-1}(\frac{\gamma+1}{2})]$$

2. $$R=[\bar{X}_A-\bar{X}_B\geq(\mu_A-\mu_B)+\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}\Phi^{-1}(\gamma)]$$

3. $$R=[\bar{X}_A-\bar{X}_B\leq(\mu_A-\mu_B)-\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}\Phi^{-1}(\gamma)]$$

### 3.5.4.2 Test di confronto tra medie di due popolazioni con varianza ignota

Allo stesso modo, possono essere confrontate due medie appartenenti a due popolazioni differenti di cui non si conosca la varianza, considerando che la statistica test

$$\frac{\bar{X}_A-\bar{X}_B-(\mu_A-\mu_B)}{\sqrt{\frac{1}{n_a}-\frac{1}{n_B}}\frac{\sqrt{(n_A-1)S^2_{A(n-1)}+(n_B-1)S^2_{B(n-1)}}}{n_A+n_B-2}}\sim t_{n_A+n_B-2}$$

è una *t di Student* con ($n_A$+$n_B$ - *2*) gradi di libertà. Per snellire la formula, si indichi con β il denominatore della statistica test. Alla stregua dei risultati precedenti, sempre considerando il test nella formulazione:

$H_0 : \mu_A = \mu_B$

$H_a :$ 1) $\mu_A \neq \mu_B$

2) $\mu_A > \mu_B$

3) $\mu_A < \mu_B$

le regioni critiche assumeranno le seguenti forme:

1.

$$R=[\bar{X}_A-\bar{X}_B\leq(\mu_A-\mu_B)-\beta t_{(n_A+n_B-2)}(\frac{1-\gamma}{2});\bar{X}_A-\bar{X}_B\geq(\mu_A-\mu_B)+\beta t_{(n_A+n_B-2)}(\frac{1+\gamma}{2})]$$

2. $R=[\bar{X}_A-\bar{X}_B\geq(\mu_A-\mu_B)+\beta t_{(n_A+n_B-2)}(\gamma)]$

3. $R=[\bar{X}_A-\bar{X}_B\leq(\mu_A-\mu_B)-\beta t_{(n_A+n_B-2)}(\gamma)]$

### 3.5.4.3 Test t di Student per due campioni a varianza uguale

In questo test si mettono a confronto le medie di due popolazioni: $x_1,...x_n$ e $y_1,...,y_n$, di numerosità *m* e *n* rispettivamente. Per l'applicazione del test vanno prima verificate le seguenti ipotesi: le varianze dei campioni devono essere approssimativamente uguali, i campioni devono essere indipendenti e le loro medie campionarie devono avere distribuzione approssimativamente normale. (condizione soddisfatta dal teorema del limite centrale per campioni numerosi).

Il test viene così impostato:

$$H_0 : \mu_x - \mu_y \leq C$$

$$H_a : \mu_x - \mu_y > C$$

In primo luogo si calcolano le medie e le varianze campionarie dei due campioni: rispettivamente $\bar{X}, \bar{Y}, s_x^2 e\, s_y^2$ . Si calcola poi la deviazione standard congiunta $s_E$:

$$s_E=\sqrt{\frac{(m-1)s_x^2+(n-1)S_Y^2}{(m-1)+(n-1)}}$$

In seguito si calcola:

$$t=\frac{\bar{x}-\bar{y}-C}{s_E\sqrt{1/n+1/m}}$$

Dalle tavole della distribuzione *t di Student* si calcola il valore critico $t_{1-\alpha}$ con *(m+n-2)* gradi di libertà e $\alpha$ come livello di confidenza.

Se $t > t_{1-\alpha}$ l'ipotesi nulla viene rifiutata.

Altrimenti viene calcolata la dimensione del campione necessaria a ridurre le probabilità di commettere un errore del primo e del secondo tipo ($\alpha$ e $\beta$ rispettivamente), determinate a priori per una nuova differenza tra le medie $C_1$:

$$n^*=\frac{2s^2(Z_{1-\alpha}+Z_{1-\beta})^2}{(C_1-C)^2}+0{,}25\,Z_{1-\alpha}^2$$

Dove tutti i termini hanno significati noti. Se $n^* < n$ e $n^* < m$ allora l'ipotesi nulla non viene respinta e viene accettata la probabilità di commettere un errore del primo tipo, quindi probabilmente è vero che $\mu_A - \mu_B \leq C$. Se nessuna delle ipotesi precedenti viene soddisfatta, si tende ad accettare l'ipotesi nulla ma la numerosità del campione è troppo

piccola per poter fare ulteriori considerazioni.

### 3.5.4.3 Test t di Satterthwaite per due campioni a varianza diversa

Il test è impostato come quello precedente e le ipotesi sono le stesse, a parte quella di uguaglianza delle varianze. Nell'uso dei simboli si farà riferimento al test visto sopra.

Per il test di Satterthwaite si procede così: prima di tutto si calcola il valore di $s_{NE}$ dato da:

$$s_{NE} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

Si calcola poi il parametro t di Satterthwaite:

$$t = \frac{\bar{x} - \bar{y} - C}{s_{NE}}$$

Si ricava poi il valore critico $t_{1-\alpha}$ dalle tavole della t-Student con $v$ gradi di libertà, dove v si ottiene arrotondando al prossimo intero il seguente valore:

$$v' = \frac{\left[\frac{s_x^2}{m} + \frac{s_y^2}{n}\right]}{\frac{s_x^4}{m^2(m-1)} + \frac{s_y^4}{n^2(n-1)}}$$

Se $t > t_{1-\alpha}$ l'ipotesi nulla viene rifiutata. In caso contrario i calcoli per il valore di $n^*$ sono troppo complicati da affrontare in questa sede e comunque variano caso per caso: si ricorrerà quindi a formule suggerite dalla letteratura in questione o da esperti.

### 3.5.4.4 Approccio non parametrico: test di Kruskal-Wallis

Questo test sfrutta la forma e la posizione dei campioni per poter fare asserzioni su di essi, a prescindere cioè dalle distribuzioni di provenienza dei campioni, la cui conoscenza qui non è richiesta. Il test di Kruskal-Wallis verifica l'ipotesi nulla del tipo:

*$H_0$ : le distribuzioni delle popolazioni sono identiche;*

*$H_a$ : parte della distribuzione 1 è posta a destra o a sinistra di quella delle altre popolazioni.*

Questo tipo di test confronta la distribuzione di una o più popolazioni rispetto a

una di riferimento.

Utilizzando come parametro di riferimento il rango dei dati, questo test è adatto a paragonare i valori di un sito (ad es. di concentrazione di un inquinante) rispetto a quelli di fondo.

Prima di applicare il test, è necessario assegnare ai dati i rispettivi ranghi, ordinati in ordine crescente. I valori inferiori al limite di rilevabilità devono ricevere lo stesso rango: ad esempio, se vi sono quattro valori al di sotto di tale limite, il rango corrispondente sarà *(1 + 2 + 3 + 4) / 4 = 2,5*. Quindi, in ordine crescente, i primi quattro dati assumeranno valore uguale a *2,5*, il quinto sarà uguale a *5*, il sesto uguale a *6* e così via.

Il test viene applicato calcolando il valore:

$$H = \left[ \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

Dove $n$ è il numero complessivo dei campioni, $n_i$ il numero dei campioni dell'$i$-esimo gruppo, $k$ il numero dei gruppi e $R$ la somma dei ranghi dell'$i$-esimo gruppo.

Si esegue poi un aggiustamento per dati ripetuti due o più volte, nel modo seguente:

$$H' = \frac{H}{1 - \left( \frac{\sum_{i=1}^{g} t_i^3 - t_i}{n^3 - n} \right)}$$

Dove $g$ è il numero dei gruppi e $t_i$ il numero di osservazioni ripetute nell'$i$-esimo gruppo.

A questo punto si effettua una comparazione del valore di $H'$ con il valore tabellato della chi-quadro con k-1 gradi di libertà. Se $H' > \chi^2$ si rifiuta l'ipotesi nulla.

### 3.5.5 Trattamento dei valori estremi

#### 3.5.5.1 Metodo di Gilbert

All'interno di un territorio, i settori con elevate o anomale concentrazioni di inquinanti vengono detti *hot spot*. La loro individuazione è fondamentale nella fase di caratterizzazione di un sito. Il metodo di Gilbert è un metodo grafico supportato dall'uso della Tabella 1 e può essere applicato in due modi: o si stabilisce la probabilità accettabile

di non individuare un *hot spot* e si ricava la dimensione dell'*hot spot* per quella probabilità, o, nota la dimensione dell'*hot spot*, si può determinare la probabilità di individuarlo. Si suppone che vengano fatti sul territorio dei campionamenti a griglia regolare, che può essere quadrata o triangolare. La tabella fornisce i valori di probabilità di non trovare un hot spot ellittico; le colonne rappresentano il valore dell'asse minore confrontato con la spaziatura della griglia, le righe, invece, il valore dell'asse maggiore dell'ellisse confrontato col valore della spaziatura della griglia. Il valore in alto delle caselle si riferiscono ad una griglia quadrata mentre quelli in basso ad una griglia triangolare.

| | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| 10% | 0,97<br>0,95 | | | | | | | | | |
| 20% | 0,95<br>0,92 | 0,88<br>0,85 | | | | | | | | |
| 30% | 0,92<br>0,87 | 0,83<br>0,78 | 0,72<br>0,66 | | | | | | | |
| 40% | 0,88<br>0,85 | 0,75<br>0,71 | 0,65<br>0,55 | 0,50<br>0,41 | | | | | | |
| 50% | 0,85<br>0,82 | 0,69<br>0,63 | 0,54<br>0,44 | 0,38<br>0,27 | 0,21<br>0,08 | | | | | |
| 60% | 0,80<br>0,80 | 0,62<br>0,58 | 0,45<br>0,35 | 0,27<br>0,15 | 0,12<br>0,03 | 0,06<br>0,00 | | | | |
| 70% | 0,77<br>0,77 | 0,56<br>0,54 | 0,38<br>0,29 | 0,18<br>0,12 | 0,07<br>0,01 | 0,03<br>0,00 | 0,00<br>0,00 | | | |
| 80% | 0,75<br>0,75 | 0,54<br>0,50 | 0,32<br>0,23 | 0,12<br>0,08 | 0,05<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | | |
| 90% | 0,72<br>0,72 | 0,51<br>0,45 | 0,30<br>0,21 | 0,10<br>0,06 | 0,03<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | |
| 100<br>% | 0,70<br>0,66 | 0,45<br>0,37 | 0,24<br>0,18 | 0,08<br>0,04 | 0,01<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 | 0,00<br>0,00 |

*Tab.1 Probabilità di non individuare un hot spot ellittico*

### 3.5.5.2 Test statistici per l'individuazione di outlier

Si presenteranno ora due metodi da applicare ad uno o più campioni di dati per individuare la presenza di *outliers* statistici: il test di Dixon, per un numero n di dati inferiore o tuttalpiù uguale a *25*, e il test di Rosner per campioni di numerosità maggiore di *25*. In entrambi i test è necessario verificare che i campioni siano estratti da variabili

distribuite normalmente. Si ricorda che, all'interno di una serie di osservazioni, un outlier statistico è un valore anomalo o aberrante molto grande o molto piccolo, che si discosta dai valori del campione e che quindi è possibile eliminare dalla serie di dati, considerandolo non rappresentativo per la variabile osservata.

### 3.5.5.2.1 Test di Dixon

Siano $x_1$, $x_2$,..., $x_n$ i valori ordinati di una variabile $X$. Può darsi il caso in cui si pensi che $x_1$ sia un *outlier*: si calcola il valore $C$, a seconda del numero dei valori:

1. $C = \frac{x_2 - x_1}{x_n - x_1}$ per $3 \le n \le 7$;
2. $C = \frac{x_2 - x_1}{x_{n-1} - x_1}$ per $8 \le n \le 10$;
3. $C = \frac{x_3 - x_1}{x_{n-1} - x_1}$ per $11 \le n \le 13$;
4. $C = \frac{x_3 - x_1}{x_{n-2} - x_1}$ per $14 \le n \le 25$.

I valori di $C$ vanno poi confrontati con quelli in Tab.2 . Se $C$ supera il valore critico, si può ritenere con livello di significatività $\alpha$ che $x_1$ sia un *outlier*.

| n | α = 0,01 | α = 0,025 | α = 0,05 |
|---|---|---|---|
| 3 | 0,869 | 0,872 | 0,879 |
| 4 | 0,822 | 0,845 | 0,868 |
| 5 | 0,822 | 0,855 | 0,879 |
| 6 | 0,835 | 0,868 | 0,890 |
| 7 | 0,847 | 0,876 | 0,899 |
| 8 | 0,859 | 0,886 | 0,905 |
| 9 | 0,868 | 0,893 | 0,912 |
| 10 | 0,876 | 0,900 | 0,917 |
| 11 | 0,883 | 0,906 | 0,922 |
| 12 | 0,889 | 0,912 | 0,926 |
| 13 | 0,895 | 0.917 | 0,931 |
| 14 | 0,901 | 0,921 | 0,934 |
| 15 | 0,907 | 0,925 | 0,937 |
| 16 | 0,912 | 0,928 | 0,939 |
| 17 | 0,912 | 0,931 | 0,942 |

| | | | |
|---|---|---|---|
| 18 | 0,919 | 0,934 | 0,945 |
| 19 | 0,923 | 0,937 | 0,947 |
| 20 | 0,925 | 0,939 | 0,950 |
| 21 | 0,928 | 0,942 | 0,952 |
| 22 | 0,930 | 0,944 | 0,954 |
| 23 | 0,933 | 0,947 | 0,955 |
| 24 | 0,936 | 0,949 | 0,957 |
| 25 | 0,937 | 0,950 | 0,958 |

*Tab.2 Valori del test della carta di probabilità normale*

Nel caso invece si sospetti che il valore xn sia un outlier i valori di C sono cosi calcolati:

1. $C = \frac{x_n - x_{n-1}}{x_n - x_1}$ per $3 \le n \le 7$;
2. $C = \frac{x_n - x_{n-1}}{x_n - x_2}$ per $8 \le n \le 10$;
3. $C = \frac{x_n - x_{n-2}}{x_n - x_2}$ per $11 \le n \le 13$;
4. $C = \frac{x_n - x_{n-2}}{x_n - x_3}$ per $14 \le n \le 25$.

### 3.5.5.2.2 Test di Rosner

Il test di Rosner è un test parametrico con cui è possibile identificare fino a *10* outlier in campioni con numerosità maggiore o uguale a *25*, estratti da una variabile normalmente distribuita.

Si fornisce il numero massimo $r_0 \le 10$ di possibili *outliers*. Si ordinano quindi gli $r_0$ valori anomali dal più al meno estremo. I valori si confrontano con le tavole di Rosner per test su *outliers*. Se il valore del test è maggiore del valore critico, allora si hanno r0 *outliers*, altrimenti si esegue il test per ( $r_0 - 1$) possibili *outliers* e così via fino a superare il valore critico o fino a che $r_0 = 0$.

Si procede in maniera seguente: siano $x_1, x_2, ..., x_n$ $n$ valori ordinati di una variabile aleatoria $X$. Inoltre, siano $\bar{x}^0 e\, s^0$ la media campionaria e la deviazione standard di tutti i dati. Si elimina dalla serie il valore $y_0$ che più si discosta dalla media e si ricalcolano i nuovi $\bar{x}^1 e\, s^1$ . Si ripete l'operazione fino ad eliminare tutti i possibili *outliers*. Al termine di queste operazioni, si dovrebbero avere gli insiemi di valori:

$$[\bar{x}^0, s^0, y^0]; [\bar{x}^1, s^1, y^1]; \ldots; [\bar{x}^{r_0-1}, s^{r_0-1}, y^{r_0-1}]$$

dove:

$$\bar{x}^i = \frac{1}{n-i} \sum_{j=1}^{n-i} x_j$$

e

$$s^i = \left[ \frac{1}{n-j} \sum_{j=1}^{n-i} (x_j - \bar{x}^i)^2 \right]^{1/2}$$

Per controllare se nei dati ci non n outliers si calcola il parametro $R_r$:

$$R_r = \frac{|y^{r-1} - \bar{x}^{r-1}|}{s^{r-1}}$$

Per la verifica del test si procede poi come spiegato in precedenza.

# 4. REGRESSIONE LINEARE

*"Non fidatevi di ciò che le statistiche dicono*

*prima di aver attentamente considerato cosa non dicono"*

William Watt

Lo scopo della regressione è quello di mettere in relazione una variabile aleatoria con una o più variabili. Iniziando con l'ipotesi che queste variabili abbiano una relazione di tipo lineare, questa prende la forma:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad (1)$$

dove $Y$ è una variabile aleatoria detta variabile responso. $\beta_0$ e $\beta_1$ sono parametri fissi e sconosciuti, rispettivamente l'intercetta e la pendenza della retta di regressione. $X$ è la variabile osservata, detta predittore, mentre la variabile non osservabile $\varepsilon$ rappresenta la differenza fra il valore di $Y$ e la componente deterministica del modello, $\beta_0 + \beta_1 x$, comunemente detta errore. Se assumiamo che l'osservazione di $x$ sia sufficientemente precisa, l'errore racchiude in se tutti gli elementi sconosciuti o non misurabili che influenzano $Y$. Nella formulazione del modello si assume che $E[\varepsilon] = 0$. Inoltre si assume che la varianza dell'errore sia costante ed indipendente da $x$ e che si possa descrivere con una distribuzione normale, con varianza $\sigma^2$, quindi $\varepsilon \sim N(0 , \sigma^2)$. Da notare come queste assunzioni semplifichino molto l'implementazione del modello ma anche come a volte possano non essere più accettabili. Si può quindi osservare come il modello di predizione della variabile $Y$ sia composto da una combinazione lineare di una parte puramente deterministica e di una stocastica.

Alla luce di queste assunzioni, è ragionevole affermare la seguente relazione lineare condizionale:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x$$

## *4.1 Stima dei parametri*

Per prima cosa, quindi, è necessario procedere con la stima dei parametri della retta di regressione. Si supponga a tal scopo di possedere un campione di $n$ osservazioni di

$x_i$, a cui corrispondono altrettante osservazioni del responso $Y_i$, con $i = 1,2,...n$. É possibile quindi disporre i punti su un grafico. Il nostro obiettivo è quello di tracciare la retta che meglio interpola i dati e a questo punto risulta logico chiedersi quale sia il metodo migliore per farlo. Come si sa, l'occhio umano è il miglior interpolatore però non ha riscontri matematici. Si potrebbe pensare di voler minimizzare il valore assoluto delle differenze delle distanze dei punti dalla retta, ma in questo modo si rischia di mascherare punti la cui distanza dalla retta di interpolazione sia di molto superiore a quella degli altri punti. Si potrebbe pensare invece di minimizzare la somma delle differenze al quadrato dalla media per ottenere una stima dei parametri (Gauss, 1800, Legendre, Eulero), implementando il così detto metodo dei minimi quadrati.

Secondo il modello espresso dalla (1) la somma delle deviazioni al quadrato è data da:

$$S^2=\sum_{i=1}^{n}\epsilon^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

La soluzione con il metodo dei minimi quadrati per la stima dei parametri prevede la minimizzazione di tale somma. In particolare, la stima è ottenuta ponendo $\delta S^2/\delta\beta_0 = 0$ e $\delta S^2/\delta\beta_1 = 0$.

Da cui:

$$-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \quad \text{e} \quad -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i=0$$

Dove il simbolo ^ indica il valore stimato del parametro.

Indicando con:

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \quad \text{e con} \quad \bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$$

si ottiene:

$$\hat{\beta}_1=\frac{(\sum_{i=1}^{n}y_i x_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)/n)}{(\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2/n)}=\frac{[\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

e

$$\hat{\beta}_0=\frac{\sum_{i=1}^{n}y_i}{n}-\hat{\beta}_1\frac{\sum_{i=1}^{n}x_i}{n}=\bar{y}-\hat{\beta}_1\bar{x} \quad .$$

Per la computazione della regressione è inoltre utile introdurre i seguenti termini:

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n} \quad ,$$

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{(\sum_{i=1}^{n}y_i)^2}{n} \quad ,$$

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{nx_i}y_i-\frac{(\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i)}{n}$$

si può quindi scrivere $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$ .

Inoltre, per come è definito il modello, vale che $(y_i-\bar{y})=(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)$ . Si può dimostrare che vale la relazione:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad (2)$$

Infatti $(y_i-\bar{y})^2=(\hat{y}_i-\bar{y})^2+(y_i-\hat{y}_i)^2+2(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$ ,

quindi $\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2+2\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$ .

L'ultimo termine è uguale a zero, infatti, avendo già mostrato che $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ , si può scrivere $\hat{y}_i=\bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1x_i$ . Sostituendo questa relazione nell'ultimo termine della sommatoria si ottiene:

$$(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=\hat{\beta}_1(x_i-\bar{x})(y_i-\bar{y}+\hat{\beta}_1\bar{x}-\hat{\beta}_1x_i)=-\hat{\beta}_1^2(x_i-\bar{x})^2-\hat{\beta}_1(x_i-\bar{x})(y_i-\bar{y})$$

Facendo la sommatoria dei termini si ha che

$$\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=-\hat{\beta}_1^2\sum(x_i-\bar{x})^2-\hat{\beta}_1\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{S_{xy}^2}{S_{xx}}-\frac{S_{xy}^2}{S_{xx}}=0 \quad .$$

La (2) da un'informazione fondamentale: dice infatti che la varianza di *Y* è data dalla somma di due contributi: la varianza dovuta alla regressione, o varianza spiegata, e la varianza degli errori o varianza residua, che sono rispettivamente il primo ed il secondo termine alla destra della relazione e che chiameremo $SS_R$ e $SS_E$. Quindi: $S_{yy} = SS_R + SS_E$.

### *4.2 Proprietà di stimatori ed errori*

Come detto in precedenza, si assume che l'errore segua una distribuzione normale con media zero e varianza $\sigma^2$. Essendo inoltre gli errori indipendenti dai valori delle osservazioni $x_i$, anche le $Y_i$ sono variabili indipendenti. La varianza condizionata di Y vale

quindi:

$$Var\,[Y \mid X = x] = Var\,[\beta_0 + \beta_1 x + \varepsilon] = \sigma^2$$

Quindi si può affermare che $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$.

Si può inoltre dimostrare come $\beta_1$ sia uno stimatore corretto di $\beta_1$. Infatti, considerando i responsi $Y_i$ come variabili casuali e le osservazioni di $x_i$ come costanti, si ottiene che:

$$\beta_1 = \frac{[\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{[\sum_{i=1}^{n} Y_i(x_i-\bar{x}) - \bar{Y}\sum_{i=1}^{n}(x_i-\bar{x})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

e

$$E[\hat{\beta}_1] = \frac{1}{S_{xx}} E[Y_i(x_i-\bar{x})] = \frac{1}{S_{xx}} E[(\beta_0+\beta_1 x_i+\epsilon_i)(x_i-\bar{x})]$$

$$= \frac{1}{S_{xx}}[E[\beta_0\sum_{i=1}^{n}(x_i-\bar{x})] + E[\beta_1\sum_{i=1}^{n} x_i(x_i-\bar{x})] + E[\sum_{i=1}^{n}\epsilon_i(x_i-\bar{x})]] \cdot$$

$$= \frac{1}{S_{xx}}(0+\beta_1 S_{xx}+0) = \beta_1$$

E' possibile inoltre ricavare la varianza di $\beta_1$ :

$$Var[\hat{\beta}_1] = Var[\frac{(\sum_{i=1}^{n} Y_i(x_i-\bar{x}))}{S_{xx}}] = \frac{1}{S_{xx}^2}\sigma^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\sigma^2}{S_{xx}} \quad .$$

In maniera simile si può dimostrare che anche $\beta_0$ è uno stimatore corretto di $\beta_0$.

Essendo inoltre $\hat{\beta}_1 = \frac{[\sum_{i=1}^{n} Y_i(x_i-x)]}{S_{xx}}$ , combinazione lineare di n variabili aleatorie normalmente distribuite, è anch'esso una variabile aleatoria gaussiana:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \quad .$$

Si consideri il valore atteso della risposta ad un dato valore di $X$, $x_0$: $E[\,Y \mid X = x0\,] = \beta_0 + \beta_1 x_0$; uno stimatore corretto di tale valore è:

$$\mu_{(Y/\hat{X}=x_0)} = \hat{\beta}_0+\hat{\beta}_1 x_0 = \bar{Y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_0 = \bar{Y} + \hat{\beta}_1(x_0-\bar{x})$$

La varianza della media del responso è data da:

$$Var[\hat{Y}/X=x_0] = Var[\bar{Y}] + (x_0-\bar{x})^{2\mathrm{Var}}[\hat{\beta}_1] + 2(x_0-\bar{x})Cov[\bar{Y},\hat{\beta}_1]$$

Essendo gli $Y_i$ indipendenti con varianza costante pari a $\sigma^2$, quindi:

$$Var[\bar{Y}] = \frac{1}{n} Var[Y_i] = \frac{\sigma^2}{n}$$

Si può inoltre dimostrare che $Cov[\bar{Y},\hat{\beta_1}]=0$ , da cui:

$$Var[\hat{Y}/X=x_0]=\frac{\sigma^2}{n}+(x_0-\bar{x})^2\frac{\sigma^2}{S_{xx}} \quad .$$

Quindi la media stimata di un responso ad un dato valore di *X*, $x_0$, è distribuita come una normale:

$$\mu_{(Y/\hat{X}=x_0)}\sim N[\beta_0+\beta_1 x_0,\sigma^2(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}})] \quad .$$

Nel caso particolare in cui $x_0 = 0$, tale valore è uno stimatore di $\beta_0$, con distribuzione:

$$\hat{\beta}_0\sim N[\beta_0,\sigma^2(\frac{1}{n}+\frac{\bar{x^2}}{S_{xx}})] \quad .$$

Per quanto riguarda i residui, dalla (1) questi sono dati da:

$$\hat{\varepsilon}_i=y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i) \quad .$$

Questi errori possono essere individuati geometricamente come la distanza verticale dei dati dalla linea di regressione. Si può dimostrare che la loro media sia nulla, infatti:

$$\sum_{i=1}^{n}\hat{\varepsilon}_i=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)]=\sum_{i=1}^{n}[y_i-(\bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1 x_i)]=$$
$$=\sum_{i=1}^{n}(y_i-\bar{y})-\hat{\beta_1}\sum_{i=1}^{n}(x_i-\bar{x})=0-\hat{\beta}_1*0=0$$

Per trovare uno stimatore corretto della varianza degli errori sarà necessario dividere per (n – 2) perché due gradi di libertà sono stati persi per stimare $\beta_0$ e $\beta_1$. Quindi:

$\hat{\sigma^2}=\frac{1}{(n-2)}\sum_{i=1}^{n}\hat{\varepsilon_i^2}=\frac{1}{(n-2)}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2$ , oppure:

$$\hat{\sigma^2}=\frac{1}{(n-2)}\sum_{i=1}^{n}[(y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x})]^2=$$
$$=\frac{1}{(n-2)}\sum_{i=1}^{n}[(y_i-\bar{y})^2+\hat{\beta_1^2}(x_i-\bar{x})^2-2\hat{\beta}_1(y_i-\bar{y})(x_i-\bar{x})]$$
$$=\frac{1}{(n-2)}(S_{yy}-\frac{S_{xy}^2}{S_{xx}})$$

che può essere usata per stimare la varianza degli errori.

### *4.4 Il coefficiente $R^2$*

Il rapporto tra la varianza dovuta alla regressione, cioè la varianza che viene spiegata dal modello, e la varianza di y definisce il così detto coefficiente di variazione $R^2$:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$R^2$ è sempre compreso nell'intervallo *[0 , 1]* ed indica la percentuale di varianza dell'errore spiegata dal modello lineare. In altre parole, più il valore di $R^2$ si avvicina all'unità, più i nostri dati vengono spiegati bene dal modello lineare.

Se si volesse investigare il grado di correlazione tra *X* e *Y*, sappiamo che il coefficiente di correlazione lineare è definito come:

$$\rho = \frac{(Cov[X, Y])}{(\sigma_x \sigma_y)}$$

Per esempio, testare se *X* e *Y* sono indipendenti è equivalente a verificare che sia $\rho = 0$. Per un campione di *n* coppie $x_i$, $y_i$, lo stimatore massimo verosimile del coefficiente di correlazione lineare è dato da:

$$\hat{\rho} = \frac{[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]}{[\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}]} = \frac{S_{xy}}{\sqrt{(S_{xx} * S_{yy})}} = r \quad .$$

### *4.5 Test su $\beta_1$*

Fare un test sull'ipotesi che $\beta_1$ sia uguale a zero significa verificare che il responso *Y* del modello non dipenda dalla variabile predittore *x*. Si imposta quindi il test:

$H_0$: $\beta_1 = 0$

$H_a$: $\beta_1 \neq 0$

$SS_E$ è indipendente da $\beta_1$ e da $\beta_0$. Quindi la quantità $\frac{(\hat{\beta}_1 - \beta_1)}{(\sigma / \sqrt{S_{xx}})} \sim N(0,1)$ è

indipendente da $\frac{SS_E}{\sigma^2} \sim \chi^2_{(n-2)}$ . Si può allora verificare che

$$\left[\frac{(N(0,1))}{\sqrt{\left(\frac{\chi^2_{(n-2)}}{(n-2)}\right)}}\right] = \frac{\left(\sqrt{S_{xx}}\frac{(\hat{\beta}_1-\beta_1)}{\sigma}\right)}{\sqrt{\left(\frac{SS_E}{(\sigma^2(n-2))}\right)}} = \sqrt{\left((n-2)\frac{S_{xx}}{SS_E}\right)}(\hat{\beta}_1-\beta_1) \sim t_{(n-2)}$$

è una quantità pivotale.

Considerando l'ipotesi nulla secondo cui $\beta_0 = 0$ la statistica test diventa

$$\sqrt{\left((n-2)\frac{S_{xx}}{SS_E}\right)}\hat{\beta}_1 \sim t_{(n-2)} \quad .$$

L'intervallo bilatero di confidenza di livello $\gamma$ per $\beta_1$ è dato da:

$$\hat{\beta}_1 - t_{(n-2)}\frac{(1+\gamma)}{2}\sqrt{\frac{SS_E}{(n-2)S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{(n-2)}\frac{(1+\gamma)}{2}\sqrt{\frac{SS_E}{(n-2)S_{xx}}} \quad .$$

## *4.6 Regressione lineare multivariata*

Questo metodo di regressione lineare viene adottato nel caso in cui il fenomeno studiato venga descritto dalla relazione tra due o più variabili. Il modello di regressione lineare prende la seguente forma:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Dove $Y$ è il responso del modello, $x_1$, $x_2$... sono le $n$ variabili descrittive; $\beta_0$, $\beta_1$... sono gli $n+1 = k$ parametri del modello mentre $\varepsilon \sim N(0, \sigma^2)$ è l'errore casuale, che riveste lo stesso ruolo del caso univariato.

La prima cosa da fare è stimare i parametri del modello, che verranno rappresentati su un vettore colonna *(k x 1)* che verrà chiamato $\boldsymbol{\beta}$. Anche gli errori e i responsi verranno rappresentati con vettori colonna *(n x 1)*, che si chiameranno rispettivamente ε e $\boldsymbol{Y}$. Le osservazioni $x_{ij}$ sono contenute in una matrice *(n x k)*:

$$X = \begin{bmatrix} 1 & x_{11} & ... & x_{1(k-1)} \\ 1 & x_{21} & ... & x_{2(k-1)} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{n(k-1)} \end{bmatrix}$$

Quindi il modello di regressione multipla è scritto nella forma :

$$Y=\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}=\begin{bmatrix} 1 & x_{11} & \dots & x_{1(k-1)} & \beta_0 & \epsilon_1 \\ 1 & x_{21} & \dots & x_{2(k-1)} & \beta_1 & \epsilon_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{n(k-1)} & \beta_{k-1} & \epsilon_n \end{bmatrix} X + \epsilon$$

che nella forma compatta sarebbe:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Il vettore media di *Y*, *E[Y]*, è uguale a:

$$E[\boldsymbol{Y}] = \boldsymbol{X\beta}$$

Essendo l'errore caratterizzato da media nulla. Come nel caso univariato, si suppone che i dati osservati x non siano affetti da errori e che tutta l'incertezza del modello sia contenuta in $\varepsilon$. In altre parole, $\varepsilon$ è l'unico elemento aleatorio del modello che fa si che $\boldsymbol{Y}$ sia una variabile aleatoria.

Per quanto riguarda la stima dei parametri si può dimostrare che lo stimatore di $\boldsymbol{\beta}$ ottenuto col metodo dei minimi quadrati, cioè quella quantità che minimizza $|\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\beta}|^2$ è data da:

$$\hat{\beta}=(X^T X)^{-1} X^T Y$$

Quindi il vettore di stima della media di *Y* è:

$$\hat{Y}=X\hat{\beta}$$

Come nella regressione lineare semplice, i residui sono dati da:

$$\hat{\epsilon}=Y-X\hat{\beta}=Y-\hat{Y}$$

Cioè i residui rappresentano la differenza tra le media osservata e quella stimata dei valori di Y.

## *4.7 Proprietà del modello*

### 4.7.1 Proprietà dei parametri

Gli stimatori dei parametri ottenuti col metodo dei minimi quadrati sono stimatori corretti dei parametri. Si può dimostrare ipotizzando che gli errori siamo indipendenti dai valori di *X*. Come visto in precedenza, si può scrivere:

$$E[\hat{\beta}]=E[(X^T X)^{-1} X^T Y]$$

Si ricordi che solamente *Y* è una variabile aleatoria, mentre *X* rappresenta dei valori fissi e noti. Quindi:

$$E[\hat{\beta}]=(X^T X)^{-1}X^T E[Y]$$

Essendo $E[Y] = X\beta$, diventa $E[\hat{\beta}]=(X^T X)^{-1}X^T X\beta$ . Ora, si noti che $(X^T X)^{-1}(X^T X)=I$ dove I è la matrice identità. Concludendo, si può affermare che:

$$E[\hat{\beta}]=\beta$$

### 4.7.2 Matrice di covarianza dei parametri

La matrice di covarianza ***C*** degli stimatori dei parametri ottenuti col metodo dei minimi quadrati è data da:

$$C=\begin{bmatrix} Var[\hat{\beta}_0] & Cov[\hat{\beta}_0,\hat{\beta}_1] & ... & Cov[\hat{\beta}_1,\hat{\beta}_{(k-1)}] \\ Cov[\hat{\beta}_0,\hat{\beta}_1] & Var[\hat{\beta}_1] & ... & Cov[\hat{\beta}_1,\hat{\beta}_{(k-1)}] \\ ... & ... & ... & ... \\ Cov[\hat{\beta}_0,\hat{\beta}_{(p-1)}] & Cov[\hat{\beta}_1,\hat{\beta}_{(p-1)}] & ... & Var[\hat{\beta}_{(k-1)}] \end{bmatrix}$$

Per definizione sarebbe $C=E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T]$ . Dalle definizioni precedenti di $\beta$ ed essendo $(X^TX)^{-1}$ simmetrica:

$$C=E[(X^T X)^{-1}X^T(Y-E[Y])(Y-E[Y])^T X(X^T X)^{-1}] \quad .$$

Gli errori rappresentati da $\varepsilon = Y - E[Y]$ si assume abbiano media nulla e uguale varianza $\sigma^2$, come detto in precedenza. Essendo mutuamente indipendenti, si può scrivere:

$$E[(Y-E[Y])(Y-E[Y])^T]=\sigma^2 I$$

che sarebbe una matrice diagonale *(k x k)* dove gli elementi sulla diagonali sono pari a $\sigma^2$ e gli elementi fuori dalla diagonale sono pari a zero.

Ne segue che

$$C=\sigma^2(X^T X)^{-1}X^T X(X^T X)^{-1}=\sigma^2(X^T X)^{-1} \quad .$$

### 4.7.3 Varianza dell'errore

Essendo la varianza $\sigma^2$ dell'errore sconosciuta, vengono usati i residui per la sua stima nel modo seguente. La somma dei quadrati dei residui è così stimata:

$$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

In notazione matriciale sarebbe:

$$SS_E = (y - X\hat{\beta})^T (y - X\hat{\beta}) = y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta} = y^T y - \hat{\beta}^T X^T y$$

I *k* parametri sono da stimare, quindi uno stimatore corretto per $\sigma^2$ è:

$$\hat{\sigma}^2 = \frac{SS_E}{n-k} = \frac{y^T y - \hat{\beta}^T X^T y}{n-k} \quad .$$

L'intervallo di confidenza per la varianza dell'errore può essere ottenuto tenendo conto della quantità

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \approx \chi^2_{n-k} \quad .$$

## *4.8 Test sui singoli parametri*

Il metodo della regressione lineare è iterativo, nel senso che in principio vengono scelte delle variabili predittrici che vengono ordinate per importanza fisica decrescente. Dopodiché vengono fatti test sui valori dei coefficienti delle variabili e a seconda dei risultati il modello può venire modificato e la procedura reiterata. Quindi, una volta stimati i valori dei parametri, vengono fatti dei test d'ipotesi, come nel caso univariato, per verificare la relazione lineare tra il responso ed un particolare predittore. Le ipotesi da comprendere nel test saranno quindi:

*Ipotesi nulla:* $H_0 : \beta_i = 0$

*Ipotesi alternativa:* $H_a : \beta_i \neq 0$

Si consideri la seguente statistica:

$$T = \frac{\beta_i - \beta_i}{\sqrt{\hat{\sigma}^2 c_{ii}}} \approx t_{n-k}$$

dove $c_{ii}$, $i = 0, 1..., k-1$, sono gli elementi della diagonale della matrice $(X^T X)^{-1}$. Questa equazione ha una distribuzione di tipo *t di Student* con *n-k* gradi di libertà ed è usata per trovare gli intervalli di confidenza intorno al generico parametro $\beta_i$. Si noti che comunque questo è un test parziale sul parametro in quanto la stima del suo valore dipende da tutte le variabili predittrici usate nel modello.

L'intervallo al $100(1-\alpha)$% di confidenza per il parametro $\beta_i$ è calcolato con la seguente relazione:

$$Pr[\hat{\beta}_i - t_{n-p}(\frac{\alpha}{2})\sqrt{\hat{\sigma}^2 c_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{n-p}(\frac{\alpha}{2})\sqrt{\hat{\sigma}^2 c_{ii}}] = 1 - \alpha \quad .$$

### *4.9 Test su tutti i parametri*

Può risultare utile a questo punto confrontare il modello ottenuto con il modello ridotto $Y = \beta_0 + \varepsilon$. Si consideri il seguente test:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{k-1}$

$H_a : \beta_i \neq 0$, *per almeno un i tale che* $0 < i \leq k-1$

Indichiamo con $SS_R$ la somma dei quadrati dovuti alla regressione mentre con $SS_E$ la somma degli errori quadrati:

$$SS_R = \hat{\beta}^T X^T Y - n\bar{Y}^2$$

$$SS_E = Y^T Y - \hat{\beta}^T X^T Y$$

Se $H_0$ è vera, allora $SS_R / \sigma^2 \sim \chi^2_k$ mentre $SSE / \sigma^2 \sim \chi^2_{n-k}$. Quindi la statistica

$$F_0 = \frac{SS_R / k}{SS_E / (n-p)} = \frac{MS_R}{MS_E} \approx F_{k,n-p}$$

segue la distribuzione di *Fisher*. $H_0$ viene rifiutato con livello del test pari a $1 - \gamma$ se $F_0 > F_{\gamma,k,n-p}$.

# 5. ANALISI DI RISCHIO

*“L'uomo può credere all'impossibile, ma non crederà mai all'improbabile”*

Oscar Wilde

Generalmente un sistema soggetto ad analisi di rischio viene considerato caratterizzato da una capacità *x* di assolvere una richiesta *y*. Ad esempio *y* potrebbe essere il valore massimo annuale medio della piena di un corso fluviale ed *x* la sua capacità di smaltimento.

In principio si assume che *X* e *Y* siano due variabili aleatorie indipendenti.

Il *fattore di sicurezza* del sistema, definito come $Z = X/Y$, dà una prima quantificazione su quanta parte della richiesta il sistema è capace di assolvere, ed è quindi anch'esso una variabile aleatoria. Se il valore di *Z* è inferiore all'unità, significa che il sistema non è del tutto in grado di supportare la richiesta *Y*.

Si indica con $p_f$ la probabilità del sistema di fallire, ovvero:

$$p_f = Prob(z \leqslant 1) = F_z(1)$$

e con *r* la probabilità di successo:

$$r = 1\text{-}p_f .$$

## *5.1. Indici di rischio*

Un modo semplice e diffuso di affrontare il problema è quello di considerare i valori medi di *x* e *y*, ottenendo il così definito *fattore di sicurezza centrale*: $z_c = \frac{\mu_x}{\mu_y}$ .

Un approccio più cautelativo prevede di sottostimare *x* e sovrastimare *y*, attraverso opportune funzioni lineari delle deviazioni standard delle due variabili . In tal caso sarebbe $x^* = \mu_x - h_x\sigma_x$ e $y^* = \mu_y + h_y\sigma_y$, dove *h* è un numero arbitrario.

Nel caso in cui la funzione densità di probabilità della richiesta si sovrapponga del tutto o in parte alla capacità, ovvero vi siano dei valori di richiesta che la capacità non può supportare, la probabilità di fallimento del sistema sarebbe diversa da zero. Per studiare questa probabilità si introduce il margine di sicurezza $S = X\text{-}Y$, anch'esso descritto da una

variabile aleatoria, essendo differenza di variabili aleatorie. Risulta quindi che:

$$p_f = Prob[(x-y) \leqslant 0] = Prob(S \leqslant 0) \quad \text{, con} \quad \mu_s = \mu_x - \mu_y \quad \text{e}$$

$$\sigma_s^2 = \sigma_x^2 - 2\rho_{xy}\sigma_x\sigma_y + \sigma_y^2 \quad .$$

Nel caso non infrequente in cui si assumano per *x* e *y* distribuzioni normali, si avrà che:

$$p_f = \phi\left(-\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 - 2\rho_{xy}\sigma_x\sigma_y + \sigma_y^2}}\right) \quad .$$

Una misura di adeguatezza di un progetto ingegneristico è *l'indice di affidabilità* $\beta = \mu_s/\sigma_s$, che può essere interpretato come il numero di deviazioni standard tra la media del margine di sicurezza $\mu_s$ ed il suo valore critico $S = 0$.

Esprimendo $\mu_s$ in funzione di $\mu_x$ e $\mu_y$ e $\sigma_s$ in funzione di $\sigma_x$ e $\sigma_y$ si ottiene:

$$\beta = \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 - 2\rho_{xy}\sigma_x\sigma_y + \sigma_y^2}}$$

L'indice di affidabilità presenta un massimo per $\rho_{xy} = 1$ e un minimo per $\rho_{xy} = -1$. Per distribuzioni normali di *X* e *Y*, la probabilità di fallimento si può esprimere come:

$$p_f = 1 - \phi(\beta)$$

e

$$r = \phi(\beta).$$

Anche se spesso *X* e *Y* vengono viste come due variabili indipendenti, così facendo non si persegue l'obiettivo del progetto. Infatti i sistemi sono di solito progettati per poter soddisfare la domanda *Y*, cosicché al variare della domanda varierà la capacità del sistema.

Un approccio diffuso nei problemi ingegneristici è quello di calcolare l'indice di affidabilità e confrontarlo con valori suggeriti dall'esperienza. Per questo può risultare utile esprimere una relazione tra la domanda e la capacità in modo da individuare uno stato limite del sistema quando queste ultime si eguagliano.

Il fattore di sicurezza $Z = X/Y$ ne è un esempio: infatti il sistema è considerato in sicurezza se $Z > 1$, lo stato di fallimento è caratterizzato per valori di $Z < 1$ e lo stato limite si ha quando $Z = 1$.

## *5.2. Performance function*

Un'altra relazione che permette di individuare lo stato limite è $S = g(x,y) = X-Y$. In

questo caso lo stato limite si presenta quando $S = 0$. Utilizzare S invece che $Z$ conviene, in quanto $S$ è una funzione lineare mentre $Z$ è fortemente non lineare. Così definita, $g(x,y)$ prende il nome di *performance function* (Fig.2).

La *performance function* di un sistema è la funzione delle variabili $X$ e $Y$ che descrive la prestazione del sistema, relativamente al suo stato limite in cui $g(x,y) = 0$. Nel piano $XY$, la curva $g(x,y) = 0$ individua due regioni del piano (lo stato di sicurezza e quello di fallimento) separate dallo stato critico. La funzione $g(x,y) = X-Y$ ne è solo un esempio, ma possono esistere infinite *performance function*. Queste funzioni sono caratterizzate dal fatto che il luogo geometrico in cui assumono valore nullo, che prende il nome di *stato limite* o *critico*, separa lo spazio in due regioni, dette di sicurezza e di fallimento.

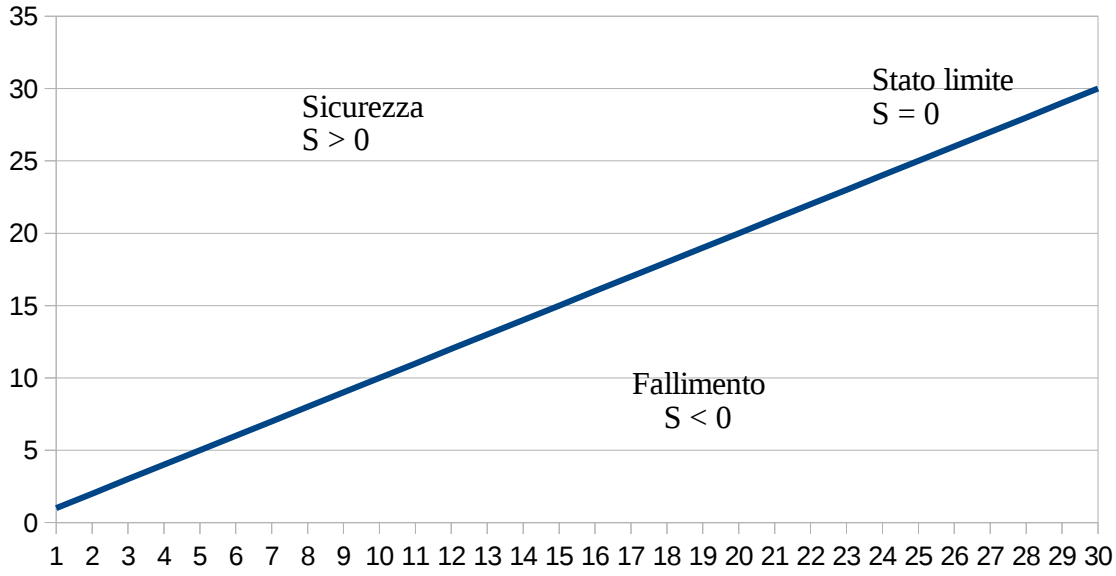


*Fig.1 Esempio di performance function*

Generalizzando, la *performance function* può essere funzione di due o più variabili di domanda e capacità, $X_1,X_2....X_n$. In questo caso lo stato critico è individuato da un'ipersuperficie, la cui minima distanza dall'origine è pari al valore dell'indice di affidabilità β.

### *5.3. FOSM – First Order Second Moment*

Un altro approccio al problema del definire l'affidabilità è il FOSM (First-Order Second-Moment). Si tratta di definire l'indice di affidabilità β come:

$$\beta \approx \frac{\ln(\mu_x/\mu_y)}{\sqrt{V_x^2+V_y^2}}$$

Con $V_x=\frac{\sigma_x}{\mu_x}$ e $V_y=\frac{\sigma_y}{\mu_y}$

Questo vale per qualsiasi distribuzione di *X* e *Y*. E' un metodo meno accurato e lo si usa quando si hanno pochi dati a disposizione. L'indice di affidabilità, come si è detto, viene poi utilizzato per determinare la probabilità di fallimento del sistema. Non vi sono metodi sistematici per determinare una probabilità di fallimento sufficientemente bassa per poter affermare che il sistema in esame sia "sicuro". A meno che non sia pari a zero, infatti, anche il più piccolo valore di $p_f$ indica che sussiste una probabilità di fallimento, per quanto remota. La scelta di intervenire o meno sul sistema per abbassare il valore di $p_f$ è responsabilità ultima del decisore politico.

### *5.4. Tempo di sopravvivenza*

Non è raro per un ingegnere dover stimare l'affidabilità di un sistema in un certo orizzonte temporale o, analogamente, prevedere in quanto tempo il sistema possa funzionare senza fallire. Questo lasso temporale è chiamato *tempo di sopravvivenza t*. Il tempo di sopravvivenza è descritto da una variabile aleatoria W, che rappresenta il tempo di attesa del primo fallimento del sistema, a partire da un certo istante iniziale. La probabilità $p_f(t)$ di fallire entro il tempo di sopravvivenza *t* è descritto dalla funzione cumulata di probabilità di *t*, $F_W(t)$, quindi:

$$p_f(t) = F_W(t)$$

La *funzione di affidabilità,* che chiameremo *R(t),* è la funzione che quantifica la probabilità che il sistema sopravviva almeno per un tempo *t* senza fallire. La funzione di affidabilità è definita come segue:

$$R(t)=\int_t^{\infty} f_W(t)\,dt = 1-F_W(t)$$

dove $f_W(t)$ è la funzione di densità di *t* e $F_W(t)$ la sua funzione di probabilità cumulata.

La funzione di probabilità del tempo di sopravvivenza dipende dal processo che

regola gli intervalli di occorrenza dei fallimenti. Se il processo che porta al fallimento fosse stazionario, cioè se ad ogni istante avesse la stessa probabilità di accadere a prescindere dalla storia passata, potrebbe essere rappresentato dall'intervallo di tempo $U$ che intercorre tra due fallimenti. Quindi il tempo di ritorno di un fallimento sarebbe il valore atteso di $U$, $\mu_U$, e il rateo di occorrenza medio di fallimento sarebbe $\lambda = 1/\mu_U$.

Ora, visualizzando la linea del tempo, si ponga l'origine in un punto casuale di questa. Per indagare la relazione tra il tempo di sopravvivenza $W$ ad un certo istante temporale ed il tempo $U$ che intercorre tra due fallimenti, si consideri la variabile aleatoria $U^*$, che descrive un intervallo di tempo che abbia l'origine come estremo destro.

Si può dimostrare che la pdf di $U^*$ sia:

$$f_{U^*}(u)=\left(\frac{u}{\mu_U}\right)f_U(u)=\lambda u f_U(u)$$

dove $f_U(u)$ è la funzione di densità di $U$. La funzione di probabilità condizionata di $W$ rispetto alla variabile $U$, che è il tempo che intercorre tra due fallimenti è:

$$f_W(w|u) = 1/u \quad \text{se } w < u,$$
$$= 0 \quad \text{altrimenti}$$

Inoltre la probabilità congiunta di $W$ e $U^*$ è:

$$f_{W,U^*}(w,u) = f_w(w|u)f_{U^*}(u) = (1/u)\,\lambda\, u\, f_U(u) = \lambda\, f_U(u), \quad \text{se } w < u,$$
$$= 0 \quad \text{altrimenti}$$

Integrando quest'ultima è possibile ottenere la funzione di probabilità del tempo di sopravvivenza $W$:

$$f_W(w)=\int_0^\infty f_{W,U^*}(w,u)\,du=\lambda\int_w^\infty f_U(u)\,du=\lambda[1-F_U(w)]$$

La probabilità che occorra un fallimento all'interno di un intervallo di tempo $t$ è data da:

$$p_f(t)=\int_0^t f_W(w)\,dw=\lambda\int_0^t [1-F_U(w)]\,dw=\lambda t-\lambda\int_0^t F_U(w)\,dw$$

Quindi, la funzione di affidabilità associata è:

$$R(t)=1-\lambda\int_0^t [1-F_U(w)]\,dw=1-\lambda t+\lambda\int_0^t F_U(w)\,dw$$

Si trova che:

$$p_f(t) \le \lambda t, \quad \text{per } t \le 1/\lambda,$$
$$p_f(t) \le 1, \quad \text{per } t > 1/\lambda.$$

I limiti di affidabilità associati risultano essere:

$$R(t) \geq 1 - \lambda t, \qquad per\ t \leq 1/\lambda,$$
$$R(t) \geq 0, \qquad per\ t > 1/\lambda.$$

Se, come solitamente accade, i fallimenti occorrono secondo una distribuzione di Poisson con intensità λ (ipotesi da verificare, ad esempio, con test di buon adattamento), il tempo che intercorre tra due eventi di fallimento è descritto da una distribuzione esponenziale:

$$p_f(t) = 1 - exp(-\lambda t)$$
$$R(t) = exp(-\lambda t).$$

## *5.5. Caso di studio: valutazione del rischio per una galleria*

## *(Modelli concettuali dinamici per l'analisi del rischio geologico a fini progettuali, Francani et al. 2005)*

Per effettuare un'analisi di rischio geologico ed idrogeologico è necessario innanzi tutto formulare il così detto *modello concettuale* del sito in esame. Questo tipo di modello dev'essere in grado di di rappresentare i vari elementi del sistema fisico oggetto di studio e le reciproche interazioni. Nella maggior parte dei casi, i fenomeni geologici non sono descrivibili da grandezze fisse o prevedibili da semplici relazioni causa-effetto. Questo tipo di grandezze vengono quindi considerate come variabili aleatorie, descritte da una funzione di densità di probabilità, a causa della variabilità dei processi naturali, dell'incapacità umana di comprendere un sistema fisico nella sua complessità o della mancanza di dati sufficienti per descriverle. Il modello concettuale che ne deriva dovrà anch'esso, per forza di cose, assumere connotazioni statistiche sia nella descrizione delle variabili in gioco sia nella descrizione dei loro rapporti causa-effetto.

### 5.5.1 Valutazione del rischio

Per la valutazione del rischio, quindi, una volta definito il modello concettuale del sito studiato, è necessario stabilire dei valori soglia di accettabilità del rischio, superate le quali risulta necessario intervenire per la salvaguardia dell'incolumità della popolazione, della sicurezza dell'opera e per la tutela dei comparti ambientali.

Generalmente, il procedimento seguito per effettuare un'analisi di rischio comprende le seguenti fasi: innanzi tutto si sceglie la *performance function* $g(x_1,..., x_n) = 0$ e i parametri da considerare come variabili aleatorie agenti sul sistema, con conseguente ricostruzione della loro distribuzione di probabilità. Da notare che le fonti di rischio ed il modello matematico che ne definisce le relazioni causa-effetto vengono stabiliti a priori. É necessario quindi stabilire quale sia l'effetto delle scelte a priori sui risultati, in modo tale da predisporre analisi di sensitività successive e poter modificare il modello in modo da non sottovalutare nessun possibile rischio. Anche attraverso la validazione, l'obiettivo è quello di formulare un modello che spieghi il più possibile i fenomeni osservati o osservabili.

Una volta eseguite queste operazioni bisogna determinare la distribuzione di probabilità di *g*, per esempio attraverso simulazioni numeriche (Monte Carlo). Il modello va successivamente validato e aggiornato attraverso opportune tecniche statistiche. Infine si stimano i valori di soglia e la $P_f$ (probabilità di fallimento del sistema) sulla base delle distribuzioni determinate.

Nel caso in esame di valutazione del rischio per una galleria, si è scelto descrivere il rischio come combinazione di due contributi principali: uno è quello relativo all'opera, che chiameremo $R_{gall}$, e uno è relativo invece all'ambiente circostante, detto $R_{amb}$. Questi due rischi sono a loro volta funzioni di diversi fenomeni idrogeologici e si possono esprimere nella forma:

$$R_{gall} = f_1(w_{H2O}, R_{H2O}; w_l,R_l; w_p,R_p; w_{gas},R_{gas})$$

$$R_{amb} = f_2(w_{abb}, R_{abb}; w_{sub},R_{sub}; w_{fr}, R_{fr}; w_{inq}, R_{inq})$$

$R_{gas}$ *e* $R_{H2O}$ rappresentano il rischio di infiltrazioni di acqua o gas in galleria, $R_I$ e $R_p$ sono rispettivamente il rischio di instabilità del fronte di scavo e di plasticizzazione del cavo. $R_{abb}$ e $R_{inq}$ costituiscono il rischio quantitativo e qualitativo connesso all'impoverimento delle risorse idriche sotterranee, $R_{sub}$ il rischio di subsidenza e $R_{fr}$ quello di frane indotto dallo scavo. Il rischio complessivo del sistema studiato può essere stimato utilizzando i pesi $w_i$ che danno maggiore o minore importanza ai vari fenomeni presi in considerazione a seconda del modello concettuale utilizzato per descrivere il sistema stesso.

Per esempio, il rischio idrogeologico, dovuto al rischio di venute d'acqua e di abbassamento del livello piezometrico, con potenziale estinzione delle sorgenti ($R_{H2O}$ e $R_{abb}$ rispettivamente), può essere valutato attraverso la *performance function* sotto forma di bilancio idrico condotto su un volume di controllo delimitato superiormente dal piano

campagna, inferiormente dalla galleria e lateralmente dal raggio di influenza della stessa galleria, cioè la zona in cui le linee di flusso vengono modificate a causa della presenza dell'opera.

La *performance function* viene quindi espressa come:

$$g(x_1 \dots x_n) = \frac{\Delta V}{\Delta t} = i + q_{monte} + q_{valle} - q_{gall}$$

dove $i$ è il tasso volumetrico di infiltrazione d'acqua nel sottosuolo, $\Delta V/\Delta t$ è la variazione nel tempo del contenuto d'acqua nella falda, $q_{monte}$ e $q_{valle}$ sono le portate defluenti a monte e a valle del volume di controllo considerato mentre $q_{gall}$ è la portata drenata dallo scavo.

Una volta definita la *performance function* è necessario esprimerne i contributi in termini statistici. In particolare, l'infiltrazione $i$ può essere stimata a partire dalla distribuzione statistica delle piogge col metodo SCS-CN, le portate $q_{monte}$ e $q_{valle}$ possono essere valutate con formule derivate dalla legge di Darcy:

$$Q = \frac{kA(P_b - P_a)}{\mu L}$$

Dove $Q$ è la portata nel mezzo poroso, $k$ è la permeabilità del mezzo poroso, $\mu$ è la viscosità del fluido, $P_b$ e $P_a$ sono le pressioni misurate in due sezioni del mezzo poroso, $A$ è la sezione attraverso cui passa il fluido ed $L$ è la distanza percorsa dal fluido in movimento. La permeabilità è considerata una variabile aleatoria mentre la portata drenata dalla galleria può essere calcolata con formule analitiche, come ad esempio la formula di Jacob-Lohman, in funzione della distribuzione statistica della permeabilità del mezzo. A tal scopo, se si dispone di un campione sufficientemente ampio di dati di permeabilità, possono essere eseguiti test non parametrici di adattamento statistico.

Con il metodo Monte Carlo è possibile ricostruire la distribuzione di probabilità della portata drenata dalla galleria, ottenendo sia il suo valor medio sia la probabilità che si verifichi una venuta d'acqua di una certa entità in galleria.

Utilizzando l'equazione di bilancio della *performance function* è possibile ottenere una stima dell'andamento del livello piezometrico nel tempo, descrivendone la distribuzione di probabilità ad ogni step temporale.

### 5.5.2 Determinazione del rischio

Come già accennato in precedenza, l'analisi di rischio va effettuata considerando il

punto di vista della galleria e quello ambientale. Per quanto riguarda la prima, si vuole determinare la probabilità che si verifichino infiltrazioni d'acqua superiori alla capacità di smaltimento della galleria. Il rischio ambientale è invece dovuto alla variazione di disponibilità idrica indotta dalla presenza dell'opera. Si valuterà quindi la probabilità che l'abbassamento del livello piezometrico dell'acquifero in cui è immersa la galleria non vada sotto una certa soglia di sicurezza. L'ultima operazione da eseguire in un'analisi di rischio, quindi, è la quantificazione vera e propria del rischio stesso.

Considerando il punto di vista della galleria, inteso come il rischio che si verifichino venute d'acqua all'interno dell'opera che possano comprometterne la funzionalità o la sicurezza, l'utilizzo della *performance function* fornisce al progettista indicazioni utili per la predisposizione degli interventi più adeguati per la mitigazione del rischio.

Per quanto riguarda il punto di vista ambientale, è necessario stabilire un valore di abbassamento accettabile Δh, che permetta poi di definire il rischio in funzione delle caratteristiche geologiche, morfologiche, climatiche e di utilizzo della risorsa idrica (variabili $x_1$, $x_2$...$x_n$) come:

$$p_f = Prob\ [g(x_1, x_2...x_n) < \Delta h]$$

Dopo aver valutato i vari termini della *performance function* è possibile procedere alla simulazione degli scenari di rischio, valutando la distribuzione di probabilità degli abbassamenti piezometrici prodotti dall'opera in regime stazionario, cioè una volta raggiunto lo stato di equilibrio.

Il valore di abbassamento accettabile viene stabilito sulla base di valutazioni di esperti ed espresso in termini di percentuale rispetto al livello precedente alla costruzione dell'opera; nel caso particolare di questo studio tale valore è pari al 20%, corrispondente ad un abbassamento di 10 m. Viene quindi valutata la probabilità di eccedere tale valore, che risulta essere dell'83%. Sempre attraverso valutazioni di esperti, si è poi proceduto con la valutazione della soglia di estinzione delle sorgenti, corrispondente ad un abbassamento di 25 m. Il rischio di superamento di tale soglia risulta pari al 18%.

L'accettabilità o meno dell'opera alla luce delle valutazioni eseguite viene ricondotta alla scelta politica del decisore (che può essere frutto di una valutazione di costi-benefici e/o di un processo decisionale partecipato dei portatori d'interesse). Nel caso in cui, per esempio, il rischio di superamento del valore accettabile di abbassamento venga ritenuto troppo elevato, verranno progettate opere di contenimento dell'acquifero, modificando il modello concettuale ed eseguendo successive simulazioni.

## *5.6. Rischio idrogeologico*

Per quanto riguarda la valutazione del rischio idrogeologico, che prevede il coinvolgimento della popolazione civile e di beni materiali mobili e immobili, viene adottata una metodologia che nasce dalla definizione dell'UNESCO di rischio di disastri naturali, risalente al 1972 e ancora accettata come la più adeguata: il rischio viene definito come la combinazione di tre componenti, che sono *pericolosità*, *vulnerabilità* e *valore esposto*.

La *pericolosità*, indicata con la lettera *H* (dall'inglese *hazard*) è definita come la probabilità che un fenomeno di una determinata intensità *I* si verifichi in un determinato periodo di tempo in una determinata area:

$$H = H(I)$$

La *pericolosità V* (dall'inglese *vulnerability*) è il grado di perdita prodotto su un certo elemento o su un gruppo di elementi dovuto al verificarsi di un evento di una data intensità *I*. Viene espressa da una scala che va da 0 a 1, in cui al valore nullo si associa l'evento “nessuna perdita” mentre all'unità corrisponde la “perdita totale” dell'elemento considerato. La vulnerabilità viene quindi espressa in funzione dell'intensità *I* del fenomeno e della tipologia dell'elemento *E* esposto al rischio:

$$V = V(I,E)$$

Il *valore esposto*, infine, che prende il nome di *W* (*worthiness*), rappresenta il valore economico o il numero di unità degli elementi a rischio, normalizzati in una scala da 0 a 1. Il valore di W è funzione solo del tipo di elementi a rischio:

$$W = W(E)$$

Il rischio totale *R*, associato ad un particolare elemento a rischio *E* e ad una data intensità *I* del fenomeno, è il risultato della convoluzione:

$$R(I,E) = H(I)*V(I,E)*W(E)$$

dove il simbolo “*” indica la convoluzione, che in questo caso è l'integrale del prodotto di distribuzioni di probabilità, sotto le ipotesi che queste ultime siano fissate e non condizionate alle altre. Quindi *H*, *V* e *W* risultano variabili casuali indipendenti ed il calcolo di *R* può essere effettuato in maniera analitica integrando il prodotto delle tre funzioni di probabilità espresse in maniera esplicita. In particolare, la pericolosità H risulta costante, V e W sono variabili nel tempo e nello spazio.

Nella maggior parte dei casi la convoluzione viene semplificata nel prodotto delle

componenti:

$$R(I,E) = H(I) \times V(I,E) \times W(E)$$

Questa operazione è possibile solo previa definizione di uno scenario di riferimento, che permetta di determinare i valori di *H*, *V* e *W* per fissati *I* ed *E*. Il fatto che l'intensità e gli elementi a rischio siano fissati implica tempi di ritorno e area fissati, quindi H, V e W risultano essere variabili indipendenti.

Per calcolare il prodotto delle tre componenti una delle scelte più semplici potrebbe essere quella di usare i valori medi:

$$\bar{R} = E[H] x E[V] x E[W]$$

Anche se la scelta del valore atteso non viene considerata adatta alle logiche di protezione civile, che si propone di evitare le perdite per lo scenario massimo affrontabile con le tecniche a disposizione. Altre possibili scelte possono essere quindi la moda o combinazioni di indici di variabilità come ad esempio *media + sqm*.

### 5.6.2 Rischio matematico e rischio percepito

Spesso, quando si deve affrontare un lungo viaggio, vi sono persone che non hanno fiducia nel prendere l'aereo per viaggiare perché considerato troppo pericoloso, preferendo di gran lunga affrontare il viaggio in macchina o in treno. Da notare che i concetti di *pericolo* e *rischio* vengono spesso interscambiati nel linguaggio corrente, nonostante matematicamente siano diversi: mentre il termine “pericoloso” fa riferimento alla probabilità di accadimento di un evento dannoso, il termine “rischioso” si riferisce alla combinazione fra danno e probabilità di accadimento. Se, infatti, si andassero a vedere le statistiche sulle morti annuali per incidenti automobilistici o ferroviari, si scoprirebbe che queste sono di molti ordini di grandezza superiori rispetto a quelle per incidenti aerei (in Europa muoiono sulle strade circa 120.000 persone all'anno mentre per gli incidenti aerei le vittime si aggirano sulle 1000 all'anno a livello globale).

Qual'è allora la motivazione, apparentemente irrazionale, che spinge a scegliere di viaggiare in macchina piuttosto che in aereo? Diversi studi di sociologia del rischio hanno messo in luce il fatto che l'essere umano considera molto rischioso un evento che causerebbe danni molto grandi, per quanto piccola sia la sua probabilità di accadimento. La percezione del rischio non è quindi lineare rispetto ai danni e alla probabilità di accadimento dell'evento, ma è piuttosto esponenziale rispetto ai primi. Inoltre, è stato dimostrato empiricamente che i decisori non si comportano con razionalità matematica di

fronte a formulazioni diverse dello stesso problema: ad esempio lo stesso problema presentato come una perdita o come un mancato guadagno può portare a decisioni differenti. Le motivazioni che portano ad evitare una perdita sono, per la maggior parte degli esseri umani, più forti di quelle di ottenere un guadagno: è più facile, ad esempio, rinunciare ad uno sconto piuttosto che accettare un aumento del prezzo, anche se magari il prezzo finale sarebbe identico.

Per supportare le decisioni, quindi, vengono introdotti altri indici che rendono l'analisi di rischio più facilmente interpretabile da tutti i portatori di interesse che partecipino al processo decisionale.

Uno di essi è il *danno atteso D*, o *gravità G*, espresso come la combinazione di vulnerabilità e valore esposto e rappresenta l'estensione e la gravità dei danni generati da un fenomeno di data intensità, per data tipologia di elementi:

$$D(E,I) = V(I,E) \times W(E)$$

Il fatto di esprimere il rischio come:

$$R^* = H \times D^k, \quad con \quad k > 1$$

pesa maggiormente il danno rispetto alla pericolosità, per spiegare la maggior percezioni di rischi più elevati, anche se molto rari.

Il *rischio specifico Rs* è invece la combinazione dei valori di pericolosità e vulnerabilità e rappresenta il livello di incidenza dei danni potenzialmente generati da un fenomeno di data intensità *I* per una tipologia di elementi *E*:

$$Rs(E,I) = H(I) \times V(I,E)$$

Il *rischio massimo* $R_{max}$, invece, per un evento di data intensità, si ottiene ponendo il valore di V=1, cioè supponendo che l'evento procuri tutti i danni massimi possibili:

$$R_{max} = H(I) \times W(E)$$

L'irreparabilità $I_{rr}$ può essere interpretato come un termine correttivo adimensionale per spiegare il grado di attenzione elevato di alcuni eventi per la popolazione civile:

$$R(I_{rr}) = H \times D \times I_{rr}$$

Un altro strumento di supporto alle decisioni è la curva di Farmer (Fig.2), che individua dei valori di rischio accettabili per date coppie di valori di danno e pericolosità. La curva è il rischio effettivo, o *soglia obiettivo*, calcolato con le tecniche sopra descritte. La regione sopra la curva rappresenta il rischio non accettabile mente la parte inferiore è il rischio accettabile.

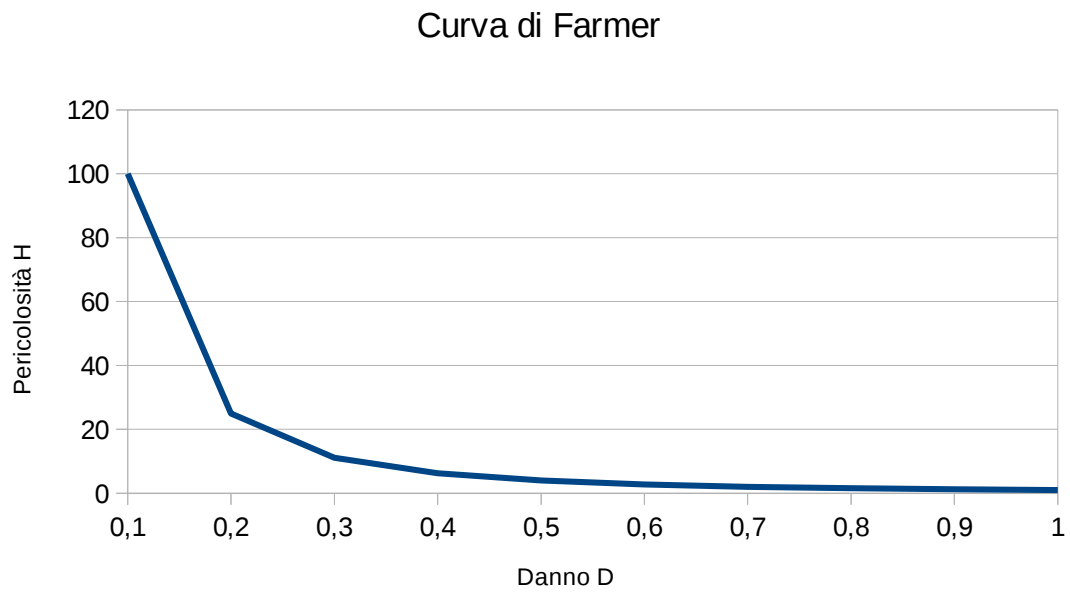


*Fig.2 Curva di Farmer generica*

### 5.6.3 Classi di pericolosità, danno e rischio

La normativa CE 2000/60 prevede che ciascuno Stato membro rediga entro il 2015 un elenco completo delle situazioni di rischio idrogeologico nel proprio territorio, elencando i potenziali rischi e danni e le misure di gestione e mitigazioni di questi.

A tal proposito vengono definite delle classi di pericolosità, danno e rischio per poter suddividere il territorio in zone omogenee:

#### 5.6.3.1 Classi di pericolosità

- Classe P1: pericolosità moderata, corrispondente ad un evento di piena con Tr = 500 anni;
- Classe P2: pericolosità media, riferita ad aree allagate da piene con Tr = 200 anni, con altezza d'acqua < 0,9 m se la velocità di corrente è > 2 m/s;
- Classe P3: pericolosità elevata, in corrispondenza di aree allagate da piene con Tr = 200 anni, con altezza d'acqua > 0,9 m per qualunque velocità di corrente o con altezza d'acqua < 0,9 m per velocità > 2 m/s;
- Classe P4: pericolosità molto elevata, riferita ad aree allagate da piene con Tr = 50 anni.

### 5.6.3.2 Classi di danno

- Classe D1: danno potenziale basso, riferito ad aree libere da insediamenti che consentano libero deflusso alle piene;
- Classe D2: danno potenziale medio, riferito ad aree in cui la presenza di persone è limitata ed il cui allagamento provocherebbe effetti limitati sul tessuto socio-economico;
- Classe D3: danno potenziale alto. Su tali aree fenomeni di esondazione possono provocare danni per la funzionalità del sistema economico e problemi all'incolumità delle persone;
- Classe D4: danno potenziale altissimo, riferito ad aree in cui fenomeni di esondazione

possono provocare ingenti danni ai beni e perdita di vite umane;

### 5.6.3.3 Classi di rischio

- R1: rischio moderato, per il quale sono possibili danni sociali ed economici ai beni ambientali e culturali marginali;
- R2: rischio medio, per il quale sono possibili danni minori agli edifici, alle infrastrutture e ai beni ambientali e culturali che non pregiudicano l'incolumità delle persone, l'agibilità degli edifici e la funzionalità delle attività socio-economiche;
- R3: rischio elevato, per il quale sono possibili problemi per l'incolumità delle persone, danni funzionali agli edifici, con conseguente inagibilità degli stessi, alle infrastrutture e ai beni ambientali e culturali, con l'interruzione delle funzionalità socio-economiche;
- R4: rischio molto elevato, per il quale sono possibili la perdita di vite umane e lesioni gravi alle persone, danni gravi agli edifici, alle infrastrutture e ai beni ambientali e culturali e la distruzione delle funzionalità delle attività socio-economiche.

### 5.6.4 Eventi estremi

Consideriamo un campione casuale di $n$ variabili casuali $X_1,X_2....X_n$ con distribuzione cumulata $F_x(x)$. Allora $X_{(1)} \leq X_{(2)} \leq ... X_{(n)}$ sono statistiche ordinate del campione in esame, dove le $X_{(i)}$ sono le $X_i$ ordinate per valori crescenti. Le sono $X_{(i)}$

statistiche perchè dipendono dal campione in esame e sono in un determinato ordine.

Per una data x, si definisce $L_i(x)=I_{(-\infty,x)}X_i$ , che significa che $L_i=1$ se $X_i$ è inferiore o tuttalpiù uguale al valore *x*, altrimenti $L_i=0$. Quindi la variabile casuale $Z=\sum L_i$ conta il numero di variabili $X_i$ che non superano *x*. *Z* ha una distribuzione binomiale di parametri *n* e $F_x(x)$.

Gli eventi $X_{(k)}\le x\, e\, Z\ge k$ sono equivalenti, cioè se la *k*-esima statistica è minore o uguale al valore *x*, allora il numero di $X_i$ minori o uguali a *x* è maggiore o uguale a *k* e viceversa.

Quindi la distribuzione cumulata marginale di una qualsiasi statistica $X_k$, con $k=1...n$, risulta essere:

$$F_{Xk}(x)=Pr(X_k\le x)=Pr(Z\ge k)=\sum_{(j=k)}^{n}\binom{n}{j}F_x(x)^j(1-F_x(x))^{(n-j)} \quad (1).$$

Questa formula fornisce la distribuzione marginale per molte applicazioni, come la ricerca della distribuzione del massimo $X_{max}=X_{(n)}$ o del minimo $X_{min}=X_{(1),}$ che sono, rispettivamente:

$$F_{Xn}(x)=\sum_{n}^{n}\binom{n}{j}F_X(x)^j(1-F_X(x))^{(n-j)}=F_X(x)^n \quad (2)$$

e

$$F_{X1}(x)=1-(1-F_X(x))^n \quad (3).$$

Se si assume che le variabili casuali $X_i$ siano continue e abbiano funzione di densità $f_x(.)$, allora le pdf di $X_{max}$ e $X_{min}$ risultano essere:

$$f_{Xmax}(x)=nf_X(x)F_X(x)^{(n-1)} \quad \text{e} \quad f_{Xmin}(x)=nf_X(x)(1-F_X(x))^{(n-1)} \quad .$$

### 5.6.4.1 Teoria dei valori estremi

Secondo la teoria dei valori estremi, il più grande o più piccolo valore di un set di variabili indipendenti identicamente distribuite tende ad una distribuzione asintotica che dipende solo dalla coda della variabile aleatoria di partenza.

Si consideri un campione $X_1,X_2...X_n$ di *n* variabili indipendenti con identica distribuzione di probabilità $F_X(.)$, dove *n* è il numero dei dati campionati ad intervalli di tempo regolari in un dato intervallo (es 1 anno).

$X_{(1)},X_{(2)},...X_{(n)}$ rappresentano le statistiche ordinate del campione, dove $X_{(1)}\le X_{(2)}\le ... X_{(n)}$ . La distribuzione di $X_{max}=X_{(n)}$ è data da:

$$F_{Xmax}=(F_X(x))^n$$

Al crescere di *n*, la distribuzione tende a zero per qualsiasi *x* nel dominio di *X*. Quindi la standardizzazione è necessaria per trovare la distribuzione limite.

Se $Y_n=\frac{(X_n-b_n)}{a_n}$ , dove $a_n>0$ è una costante di scala e $b_n$ è una costante di locazione, allora la distribuzione limite deve essere una di queste tre:

Tipo 1: $F_Y(y)=\exp(-\exp^{-y})$ , Gumbel;

Tipo 2: $F_Y(y)=\begin{matrix}\exp(-y^{-\gamma}) & se\ y>0\\ 0 & se\ y\le 0\end{matrix}$ , Fréchet;

Tipo 3: $F_Y(y)=\begin{matrix}\exp(-(-y)^{\gamma}) & se\ y<0\\ 1 & se\ y\ge 0\end{matrix}$ , Weibull.

Dove $\gamma>0$ è una costante e *Y* è l'asintoto di $Y_n$ per *n* tendente a infinito. Il fatto che la distribuzione asintotica sia una di queste tre è garantito dal postulato di stabilità, che afferma che se *X* ha una distribuzione di valore estremo, allora il massimo delle *n* osservazioni indipendenti di *X* ha la stessa distribuzione, ma con diversi parametri di scala e locazione.

Quindi:

$$F_{Xmax}=F_X(x)^n\sim F_X(\frac{x-b_n}{a_n})$$

dove *a* e *b* sono funzioni di *n*, fornisce tutte le possibili soluzioni asintotiche per *n* che tende a infinito.

Se $F_x(x)$ è strettamente monotona e continua, condizione sufficiente affinché $F_{xmax}(.)$ sia asintotica alla funzione di tipo 1 è che: $\lim_{(x\to\omega)}(\frac{d}{dx}(\frac{1-F_X(x)}{f_x(x)}))=0$ , con $\omega$ limite superiore di *x*.

L'espressione $\frac{f_x(x)}{1-F_x(x)}$ è chiamata funzione di rischio.

Condizione sufficiente di convergenza al tipo 2 è che: $\lim_{(x\to\infty)} x\frac{f_x(x)}{1-F_x(x)}=\gamma>0$ , mentre per il tipo 3: $\lim_{(x\to\omega)}(\omega-x)\frac{f_x(x)}{1-F_x(x)}=\gamma$ .

In molte applicazioni il modello di probabilità viene applicato nella forma inversa, fissando una probabilità *q* di non eccedere un dato valore e identificando il valore di progetto come in q-esimo quantile della variabile considerata. Per la distribuzione di tipo 1,

si ha che:

$$y = -\ln(-lnq)$$

con $F_y(y) = q$. $Y$ è il q-esimo quantile della distribuzione di tipo 1. Il q-esimo quantile della distribuzione di tipo 2 è $Q_{(q,2)} = \exp(y/\gamma)$ e per il tipo 3 è $Q_{(q,3)} = \exp(-y/\gamma)$ .

Quindi la formulazione generale della distribuzione dei valori estremi nella forma inversa è:

$$\xi_q = \frac{(1-\exp(-ky))}{k} \quad .$$

La distribuzione di probabilità data dalla forma inversa è chiamata GEV (General Extreme Value), che può essere scritta come:

$$F_y(y) = \exp(-(1-ky)^{(1/k)}) \quad .$$

Utilizzando tre parametri, la funzione cumulata GEV può essere riscritta come:

$$F_{Xmax}(x) = \exp\left(-\left(1-\frac{k(x-\varepsilon)}{\alpha}\right)^{(1/k)}\right)$$

dove α è un parametro di scala, ε è un parametro di locazione e *k* un parametro di forma. Per $k < 0$, GEV rappresenta una distribuzione di tipo 2, valida solo per $x > (\varepsilon+\alpha/k)$; per $k > 0$, GEV rappresenta una distribuzione di tipo 3, valida solo per $x < (\varepsilon+\alpha/k)$. Il caso di $k = 0$ corrisponde ad una distribuzione Gumbell (tipo1), con parametro di scala α e parametro di locazione *b*.

*Teorema di Fisher-Tippett-Gnedenko:* Data una successione $(X_n)$ di variabili aleatorie indipendenti identicamente distribuite e definita la variabile aleatoria $M_n = \max\{X_1...X_n\}$, se esistono costanti di normalizzazione $c_n$ e $d_n$ e una funzione di densità non degenere H tale che:

$$c_n^{-1}(M_n - d_n) \rightarrow H$$

allora *H* è la funzione di densità della GEV (Generalized Extreme Value), data da:

$$H(x) = \exp\left(-\left(1+\xi\frac{x}{\beta}\right)^{-\left(\frac{1}{\xi}\right)}\right) \quad \text{per} \quad \xi \neq 0$$

$$H(x) = \exp\left(-\exp\left(\frac{-x}{\beta}\right)\right) \quad \text{per} \quad \xi = 0$$

La media e la varianza della distribuzione GEV sono dati da:

$$E[X_{max}] = \varepsilon + \frac{\alpha}{k}[1-\Gamma(1+k)] \quad \text{per k>-1}$$

e $Var[X_{max}]=(\frac{\alpha}{k})^2[\Gamma(1+2k)-\Gamma^2(1+k)]$ per k>-0,5

Quindi la media e la varianza non sono definite per valori di *k* maggiori di *-1* e di *-0,5* rispettivamente.

La forma inversa della distribuzione GEV è data da:

$$\xi_q=\varepsilon+\frac{\alpha}{k}(1-(-lnq)^k)=\varepsilon+\frac{\alpha}{k}[1-\exp(-ky)]$$

e fornisce il *q*-esimo quantile, necessario a determinare il valore di progetto (es. il valore di una piena fluviale) per una data probabilità di non superamento o periodo di ritorno.

Il valore di progetto di una variabile casuale distribuita come una GEV è legato al periodo di ritorno di tale variabile:

$$X_{max}(T)=\varepsilon+\frac{\alpha}{k}[1-(-\ln\frac{T}{(T-1)})^k] \quad .$$

In altre parole, se la variabile considerata fosse il valore massimo di piena annuale, $X_{max}(T)$ sarebbe la piena associata al periodo di ritorno *T* o la portata che non verrebbe superata con probabilità *1-1/T*.

### 5.6.4.2 Eccessi sopra una soglia e metodo POT

Data una variabile aleatoria con funzione cumulativa *F*, la funzione di densità sopra una soglia *u* è data da:

$$F_u(x)=P(X-u|X>u)=\frac{[F(x+u)-F(u)]}{[1-F(u)]}$$

Il teorema di Pickands-Balkema-De Haan tratta della convergenza di questa distribuzione a una GPD (Generalized Pareto Distribution), ma prima forniamo la definizione di quest'ultima:

$$G(x)=1-(1+\xi\frac{x}{\beta})^{-(\frac{1}{\xi})} \quad \text{per} \quad \xi\neq 0$$

$$G(x)=1-\exp(\frac{-x}{\beta}) \quad \text{per} \quad \xi=0$$

Si chiama generalizzata perché quando $\xi > 0$ coincide con la distribuzione di Pareto, quando $\xi = 0$ coincide con l'esponenziale e quando $\xi < 0$ coincide con la distribuzione di Pareto di tipo II.

Si può ora enunciare il seguente:

*Teorema di Pickands-Balkema-DeHaan:* è possibile trovare una funzione misurabile positiva β(u) tale che:

$$\lim_{(u \to x_F)} \sup_{(0<x<x_F-u)} \left| \left( F_u(x) - G_{(\xi, \beta(u))} \right) \right| = 0$$

se e solo se *F* appartiene al dominio di attrazione dei massimi e $x_F = sup\{x \in \boldsymbol{R} : F(x)<1\}$. Con *u* si indica il valore della soglia stabilito.

Quindi informalmente, per *u* grande a piacere (cioè nella coda della distribuzione), la distribuzione degli eccessi sopra u è approssimativamente una GPD.

Inoltre le variabili indipendenti per cui la distribuzione asintotica dei massimi normalizzati è data dalla GEV sono esattamente le variabili aleatorie le cui distribuzioni degli eccessi convergono alla GPD.

Il metodo di stima della coda tramite GPD prende il nome di POT (Peacks Over a Threshold) e si basa sulla seguente affermazione: sia *F*(x) = 1-F(x)* e si supponga che, per una soglia u sufficientemente grande, $F_u(x) = G_{\xi,\beta}(x)$. Allora si può dimostrare che:

$$F^*(x) = F^*(u)\left(1 + \xi \frac{(x-u)}{\beta}\right)^{-\left(\frac{1}{\xi}\right)} = F^*(u)\, G^{-1}_{(\xi,\beta)}(x-u) \quad (4)$$

Il metodo POT prevede quindi i seguenti passi:

1. Stima di ξ e β (per esempio con il metodo della massima verosimiglianza) considerando gli eccessi $y_i = x_i - u$.

2. Stima di *F'(u)* a partire da $F(u) = N_u/N$, dove $N_u$ è il numero di superamenti della soglia u in un campione di numerosità *N*.

3. Nella (4) si sostituiscono le stime ottenute per ottenere una stima di *F'(x)*.

4. Dalla versione stimata della (4) si stima il quantile $x_p$ come:

$$\hat{x}_p = u - \frac{\hat{\beta}}{\hat{\xi}}\left[\left(N \frac{(1-p)}{N_u}\right)^{-\xi} - 1\right]$$

Per la scelta della soglia vi sono differenti criteri. Ad esempio, per la GPD la funzione di eccesso è lineare: $e(u) = E(X-u/X>u) = \frac{(\beta + \xi u)}{(1-\xi)}$ . Definendo la funzione di eccesso empirica $e_n(u) = \frac{1}{N_u} \sum_{x_i > u} (x_i - u)$ , si sceglie il più piccolo valore di u per cui la funzione è approssimativamente lineare.

### 5.6.4.3 Distribuzione dei minimi

La distribuzione dei massimi e dei minimi sono correlate dal principio di simmetria, introdotto da Gumbell nel 1958. Sia $X$ è una variabile con fdp $f_x(x)$ e $X^*$ una variabile con fdp simmetrica ad $f_x(x)$ rispetto all'asse dell'origine. Allora $1-F_x(x)=F_x^*(-x)$. Quindi $[1-F_x]^n=[F_x^*(x)]^n$ .

Ne consegue che, per l'equazione 3, si ha che:

$$[1-F_x(x)]^n=1-F_{x_1}(x)=1-F_{xmin}(x)$$

è la probabilità di non superare il più piccolo valore di X e $[F_x^*(-x)]^n$ è la probabilità di non superare il più grande valore di $X^*$.

Quindi:

$$1-F_{xmin}(x)=F_{(X^*max)}(-x) \quad .$$

Anche le fdp sono correlate:

$$f_{xmin}(x)=f_{(X^*max)}(-x) \quad .$$

Usando il principio di simmetria, la distribuzione asintotica dei valori minimi di una variabile aleatoria può essere dedotta a partire dalla distribuzione dei massimi di tale variabile invertendo il segno e considerando le probabilità complementari.

### 5.6.4.4 Esempio di applicazione: stima di una portata con determinato periodo di ritorno

La teoria dei valori estremi viene usata, nell'ambito dell'ingegneria civile ed ambientale, in una vasta gamma di applicazioni. Essa serve infatti da supporto alla gestione di eventi non ordinari a cui le opere devono far fronte, come le onde di piena in un invaso artificiale o a situazioni estreme come una forte alluvione o una frana. I dimensionamenti di opere come dighe o argini non possono prescindere dalla conoscenza delle possibili magnitudo dei processi naturali e dei carichi a cui possono essere sottoposti.

A tal proposito si apporta un esempio di applicazione all'idrologia della teoria dei valori estremi.

Per determinare il valore di portata associato ad un certo periodo di ritorno, in una certa sezione di fluviale, è necessario prima di tutto calcolare il valore della portata indice $q_{index}$, che ci da informazioni sui massimi di piena annuali. In generale, essa è definita come la media dei valori massimi annuali di piena: $q_{index}=E[Q_{max}]$.

Può darsi che si disponga di una serie abbastanza lunga di dati sulle portate. In tal caso la portata indice può essere calcolata come la media campionaria delle osservazioni:

$$q_{index} = \frac{1}{n}\sum_{i=1}^{n} q_i$$

Dove $n$ è il numero di anni per cui si hanno a disposizione dati e $q_i$ è il valore massimo annuale di portata dell'anno i-esimo. Essendo la stima della portata indice basata su dati osservati, questo metodo fa parte dei così detti metodi diretti e prende il nome di AFS (*estimation from Annual Flood Series*).

Nel caso gli anni di misurazioni a disposizione non siano un numero consistente, vi è un metodo diretto alternativo più adatto a questi casi. Si chiama metodo PDS (*estimation from Partial Duration Series*) e prevede la stima della portata indice come media di valori di portata $q'$ sopra una soglia (metodo POT, *peaks over a threshold*). In questo caso si ha che:

$$q_{PDS} = \frac{1}{n}\sum_{i=1}^{n} q'_i$$

dove si indica con $n$ il numero di anni di osservazioni a disposizione e con $q'$ il valore di portata che superi una data soglia.

Il corrispondente errore standard di stima è:

$$\sigma_{(q_{PDS})} = \sqrt{\frac{1}{[n(n-1)]}\sum_{i=1}^{n}(q'_i - q_{PDS})} \quad .$$

Il valore di portata indice è correlato al valore di $q_{PDS}$ attraverso il rateo di occorrenza λ, definito come il numero medio annuale di superamenti della soglia, e dai parametri della curva di crescita della $q_{PDS}$ . Nel caso italiano, le curve di crescita sono ben approssimate da distribuzioni GEV (De Michele e Rosso, 2001; Bocchiola et al. 2003). Si può dimostrare che i massimi sopra la soglia discendano da una distribuzione generalizzata di Pareto se occorrono come processi di Poisson (cioè in modo indipendente uno dall'altro e con stessa media e varianza). In questo caso il valore di portata indice è dato da:

$$\hat{q_{index}} = \frac{1}{\left[\varepsilon + \frac{\alpha}{k}\left(1 - \frac{\lambda^k}{(1+k)}\right)\right]} q_{PDS}$$

L'equazione fornisce il valore di portata indice dato il tasso di occorrenza λ, il valore di $q_{PDS}$ e i parametri regionali della GEV ε, α e $k$. L'errore standard di stima della portata indice è dato da:

$$\sigma_{(\hat{q_{index}})}=\sigma_{(q_{PDS})}\frac{\lambda^k}{[k(1-k)]}\sqrt{[(\Gamma(1+2k)-\Gamma^2(1+k))(1+2k)]}$$

dove Γ è la funzione Gamma.

Il metodo della portata indice prevede infine la stima di una portata $q_T$ caratterizzata da un periodo di ritorno T tramite la relazione: $q_T = \chi_T q_{index}$. $X_T$ prende il nome di fattore di crescita ed è una costante caratteristica della regione omogenea considerata. Esso viene determinato applicando la GEV della variabile aleatoria $\chi = Q/q_{index}$, dove Q è la portata massima annuale e $q_{index}$ la media dei massimi annuali. Introducendo

$y_T=-\ln(\ln\frac{T}{(T-1)})$ , variabile ridotta di Gumbell.

Il fattore di crescita è quindi della forma: $\chi_T=\varepsilon+\frac{\alpha}{k}(1-e^{-ky_T})$ .

Un concetto interessante introdotto da questa metodologia di stima delle portate è quello di regione omogenea. Si suppone infatti che i parametri della distribuzione GEV per i massimi di piena non varino, a meno di un fattore di scala, all'interno di una regione omogenea (in termini di fattori che determinano le variabili idrologiche come permeabilità, piovosità, ecc.). Ossia, stimando i parametri della GEV dei massimi di piena in una data sezione di un dato fiume, è possibile risalire alle distribuzioni dei massimi di piena di tutte le sezioni di tutti i fiumi compresi nella regione omogenea considerata.

### 5.6.5 Il metodo della portata indice

La valutazione del rischio idrogeologico di un area e la redazione della relativa carta di pericolosità prevede diverse fasi e studi di diversi aspetti del territorio. Innanzi tutto è necessario descrivere il regime piovoso (in termini statistici) del bacino fluviale, la geomorfologia e l'uso del suolo, nonché le caratteristiche idrauliche dell'alveo. In funzione di questi fattori viene poi valutata la probabilità di superamento di un certo valore di portata caratteristico del sito in esame, tale da provocare il superamento degli argini.

A questa prima fase segue poi quella di valutazione degli effetti delle piene: viene stimata la capacità di smaltimento del territorio inondato e, attraverso modelli mono e bi-dimensionali, si fanno simulazioni su possibili danni e conseguenze di vari eventi di piena con diversi periodi di ritorno assegnati.

Il primo passo da fare, quindi, è quello di studiare i valori di portata annuali del sito fluviale e la loro distribuzione di probabilità. Per fare questo il metodo più diffuso è

quello della portata indice. Questo metodo permette di stimare valori di portata con periodo di ritorno assegnato a partire dal valore della portata indice, $q_{index}$, definita come valore medio dei massimi annuali di piena registrati, e da un fattore di crescita $\chi_T$. Quest'ultimo assume valore costante per una data regione omogenea e per fissato periodo di ritorno $T$. Conoscendo il valore di questi due termini la portata corrispondente ad un determinato periodo di ritorno e per il particolare sito in esame, $q_T$, è:

$$q_T = \chi_T \text{ x } q_{index}$$

### 5.6.5.1 Il fattore di crescita

Il metodo della portata indice si basa sul concetto di frattalità statistica: se una data regione può essere considerata omogenea dal punto di vista delle variabili che influenzano i regimi fluviali (scabrezza di fondo, piovosità, tipologia di terreno, pendenze, uso del suolo, ecc..) allora le caratteristiche statistiche dei corsi fluviali che ne fanno parte (ad esempio la portata annuale media) sono anch'esse costanti e proporzionali alle loro dimensioni, ovvero bacini fluviali di pari dimensioni avranno comportamento analogo e le caratteristiche di quelli più grandi o più piccoli potranno essere stimate applicando dei fattori di scala.

Il fattore di crescita viene determinato applicando a scala regionale il modello probabilistico generalizzato del valore estremo, ossia la distribuzione GEV della variabile aleatoria $X = Q/q_{index}$, dove $Q$ indica il massimo annuale della portata al colmo e la portata indice per il generico sito fluviale corrisponde al valore atteso dei massimi annuali di portata al colmo nel sito stesso. $\chi_T$ viene poi calcolato come segue:

$$\chi_T = \epsilon + \frac{\alpha}{k}\left(1 - e^{-k y_T}\right)$$

Dove $y_T$ indica la variabile ridotta di Gumbell, funzione del periodo di ritorno $T$ in anni:

$$y_T = -\ln\left(\ln \frac{t}{T-1}\right) \quad .$$

ε, α e k sono rispettivamente il parametro di posizione, di scala e di forma e sono costanti all'interno della stessa regione omogenea. Il valore di $\chi_T$ rimane anch'esso costante all'interno della stessa regione omogenea.

Si vedranno ora alcuni tra i metodi più comuni di stima della portata indice.

### 5.6.5.2 Metodi diretti

#### 5.6.5.2.1 AFS - Estimation from maximum Annual Flood Series

Se in un particolare sito fluviale si dispongono n valori di massimi annuali di portata, la portata indice può essere stimata come la media dei dati campionati $q_1,q_2...,q_n$:

$$\hat{q_{index}} = \hat{q_{AFS}} = \frac{1}{n}\sum_{i=1}^{n} q_i$$

L'errore standard di stima è:

$$\sigma_{\hat{q_{index}}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(qi - \hat{q_{index}})^2}$$

Questo errore si riduce all'aumentare della numerosità del campione. Per un livello di significatività a, gli intervalli di confidenza del *100(1-a/2)%* sono $\hat{q_{index}} \pm F^{-1}(1-\frac{a}{2})\sigma_{\hat{q_{index}}}$ , dove $F^{-1}(.)$ è il quantile della distribuzione normale standard con probabilità di superamento *a/2*. I limiti $\hat{q_{index}} \pm \sigma_{\hat{q_{index}}}$ corrispondono ad un livello di confidenza del *84%*.

#### 5.6.5.2.2 PDS - Estimation from Partial Duration Series

Se sono disponibili n dati di colmo di piena, la portata indice può essere stimata a partire dalla media della serie di valori sopra una data soglia (metodo POT) $q'_1,q'_2...q'_n$

Si calcola:

$$\hat{q_{PDS}} = \frac{1}{n}\sum_{i=1}^{n} q'i$$

con un errore standard pari a

$$\sigma_{\hat{q_{index}}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(qi - \hat{q_{PDS}})^2}$$

La portata indice è correlata a $q_{PDS}$ dal rateo di occorrenza λ dei colmi sopra la soglia e dai parametri della curva di crescita PDS. Nel caso i colmi di piena sopra la soglia seguano un processo *poissoniano*, si può dimostrare che i colmi annuali di piena seguano

una distribuzione GEV.

La portata indice è data dalla relazione:

$$\hat{q_{index}} = \frac{1}{\varepsilon + \frac{\alpha}{k}\left(1 - \frac{\lambda^k}{(1+k)}\right)} \hat{q_{PDS}}$$

Dove λ è il numero medio di superamenti annuali, $\varepsilon$,$\alpha$ e $k$ sono parametri regionali della GEV.

L'errore standard della portata indice così stimata è dato da:

$$\sigma_{\hat{q_{index}}} = \sigma_{\hat{q_{PDS}}} \frac{\lambda^k}{k(1-k)} \sqrt{(\Gamma(1+2k) - \Gamma^2(1+k))(1+2k)}$$

dove Γ è la funzione Gamma.

Questo metodo è più accurato nel caso si disponga di pochi dati. Il valore di soglia è definito in modo tale che i superamenti seguano un processo di Poisson, ovvero che siano indipendenti. Questo si ottiene imponendo una soglia tale per cui il numero medio di superamenti annuali e la loro varianza siano uguali (proprietà della distribuzione di Poisson).

### 5.6.5.3 Metodi indiretti

#### 5.6.5.3.1 Invarianza di scala per la portata indice

Qualsiasi territorio può essere diviso in zone omogenee, in cui le proprietà statistiche di accadimento di eventi idrologici possono considerarsi costanti (De Michele e Rosso, 2002). Ad esempio, tutti i bacini fluviali della stessa zona omogenea sono caratterizzati da parametri costanti della distribuzione GEV per i colmi di piena. Allo stesso modo, le proprietà statistiche all'interno di queste zone rimangono costanti anche se riferite a bacini di diverse dimensioni, a meno di una costante di proporzionalità, introducendo il concetto di frattalità statistica. Da queste osservazioni nasce il metodo dell'invarianza di scala per determinare la portata indice di un corso fluviale a partire dalla conoscenza della portata indice di un altro corso fluviale appartenente alla stessa zona omogenea, attraverso la seguente formulazione:

$$q_{index2} = q_{index1} \left(\frac{A_2}{A_1}\right)^m$$

Dove $m$ è un esponente di scala, $q_{index2}$ è la portata indice ricercata, $q_{index1}$ è quella

conosciuta e $A_2$ e $A_1$ sono le rispettive aree sottese dal bacino 1 e 2. I due bacini possono far parte di due fiumi diversi o anche dello stesso fiume, considerando due sezioni di chiusura differenti; l'importante è che appartengano alla stessa zona omogenea. Una prima analisi di consistenza del metodo può essere la verifica di non superamento della capacità di invaso da parte della portata indice stimata. Se infatti $q_{index2}$ supera la soglia di sfioro, bisogna innanzi tutto controllare se non si è considerato un tempo di ritorno troppo elevato (di solito intorno ai due o tre anni), in seguito accertarsi che le proprietà statistiche dei due bacini si possano considerare le stesse e infine, se si reputa opportuno, scegliere un altro metodo di determinazione della portata.

#### 5.6.5.3.2 Formule empiriche

Le formule empiriche, grazie alla loro semplicità, vengono spesso utilizzate nella determinazione della portata indice. Il metodo consiste nel legare la portata indice alle caratteristiche del bacino in esame quali indici climatici (come il regime delle precipitazioni), parametri geolitologici e geopedologici, uso del suolo, parametri geomorfologici e forzanti antropiche. Queste formule sono sovente calibrate attraverso una regressione lineare multivariata del logaritmo della portata indice con i logaritmi dei valori dei parametri. Considerando $N$ parametri, le formule empiriche hanno la seguente struttura:

$$\hat{q_{index}} = const \prod_{i=1}^{N} X_i^{n_i}$$

Dove la costante e gli $n_i$ sono stimati da dati di portata osservati per bacini appartenenti alla stessa regione omogenea, mentre $X_i$ sono i parametri che caratterizzano il bacino di cui si è parlato in precedenza. Da notare come il metodo dell'invarianza di scala sia un caso particolare di formula empirica, dove l'unico parametro è l'area A drenata dal bacino. La robustezza del metodo andrà poi verificata attraverso una procedura *jaknife* o *bootstrap*. Per verificare la robustezza e l'affidabilità del metodo, comunque, dev'essere considerato un numero statisticamente significativo di parametri. E' stato osservato come la varianza di questo metodo possa variare in una range molto elevato (anche ± 100%). Nonostante questo, la procedura viene applicata nel caso la varianza sia comunque minore di altri metodi.

#### 5.6.5.3.3 Stima da dati storico-documentali

L'analisi degli eventi storici di piena viene eseguita a partire da dati osservati o

documentati durante la storia del bacino. Vengono considerati solo gli eventi di superamento di una data soglia, $q_t$, determinata a partire dalla capacità di invaso del corso fluviale in esame. Se in $n'$ anni vengono contati $n_t$ superamenti della soglia $q_t$, la frequenza attesa di superamento della soglia in un anno può essere così stimata:

$$Pr[q \geq q_t] = \frac{n_t + 1}{n' + 1}$$

Quindi, il periodo di ritorno di $q_t$ è:

$$\hat{T}_{q_t} = \frac{1}{Pr[q \geq q_t]} = \frac{n' + 1}{n_t + 1}$$

e la portata indice è data da:

$$\hat{q_{index}} = \frac{q_t}{\chi_{\hat{T}_{q_t}}}$$

dove $\chi_{\hat{T}_{q_t}}$ è il fattore di crescita riferito al tempo di ritorno stimato, calcolato a partire dalla distribuzione di probabilità GEV della regione omogenea.

L'individuazione degli effettivi superamenti della soglia $q_t$ non è banale. Innanzi tutto, detta soglia va stimata con il tracciamento dei profili di moto permanente. Per determinare i superamenti della soglia non vanno considerati, ad esempio, gli eventi di rottura degli argini, di ostruzione del corso fluviale o altre cause non naturali. Alla difficoltà di raccogliere dati storici si aggiunge quindi quella di determinare le cause delle inondazioni. Per questo il metodo si presta ad essere applicato in zone storicamente molto urbanizzate nelle quali questi eventi sono stati osservati e documentati.

## *5.7. Analisi di rischio assoluta per siti contaminati*

L'analisi di rischio per siti contaminati viene definita tecnicamente come “processo sistematico per la stima di tutti i fattori di rischio significativi che intervengono in uno scenario di esposizione, causato dalla presenza di pericoli”. In altre parole, la valutazione del rischio è la stima delle conseguenze sulla salute umana di un evento potenzialmente dannoso, in termini di probabilità che dette conseguenze si verifichino. Per come viene impostato, il processo di valutazione fornisce il grado di importanza dei rischi potenziali esaminati per il caso specifico, da confrontare con una base di riferimento univoca; tale base di giudizio è il livello di accettabilità/attenzione/necessità di bonifica, fissato in linee guida stabilite da enti ed organismi di gestione e salvaguardia ambientale nazionali e internazionali.

Tale valutazione di rischio si effettua, in genere, su siti che rappresentano un pericolo cronico per l'uomo e l'ambiente, stimando un livello di rischio e, conseguentemente, i valori limite di concentrazione, determinati in funzione delle caratteristiche della sorgente dell'inquinamento, dei meccanismi di trasporto e dei bersagli della contaminazione.

Il rischio *R* è inteso come il prodotto tra la probabilità di accadimento di un evento dannoso *P* e l'entità del danno provocato dall'evento stesso *D*:

$$R = P \times D$$

Il danno conseguente all'evento incidentale *D*, a sua volta, può essere dato dal prodotto tra un fattore di pericolosità $F_p$, dipendente dall'entità del possibile danno, e un fattore di contatto $F_e$, funzione della durata di esposizione:

$$D = F_p \times F_e$$

Nel caso di siti inquinati, la probabilità *P* di accadimento dell'evento è uguale a 1, il fattore di pericolosità è dato dalla tossicità dell'inquinante (*T* [mg/kg d]$^{-1}$ ) ed il fattore di contatto assume il valore della portata effettiva di esposizione (*E* [mg/kg d]); per cui, in generale, il rischio *R* derivante da un sito contaminato è dato dalla seguente espressione:

$$R = E \times T$$

Dove *E* ([mg/kg d]) rappresenta l'assunzione cronica giornaliera del contaminante e *T* ([mg/kg d]$^{-1}$) la tossicità dello stesso. Il risultato *R*, viene poi confrontato con i criteri di accettabilità individuali (cioè riferiti ad un solo inquinante) e cumulativi del rischio sanitario, per decidere se esistono o meno condizioni in grado di causare effetti sanitari nocivi.

La formula che viene usata per la determinazione della dose cronica giornaliera *E* per un inquinante *j* e una generica via di esposizione *i* (ingestione, contatto dermico e inalazione) è:

$$E_{i,j} = \frac{C_{i,j} \, x \, CR_i \, x \, EF \, x \, ED}{BW \, x \, AT}$$

Dove $C_{j,i}$ è la concentrazione dell'inquinante *j* al punto di esposizione nello specifico comparto ambientale associato alla via di esposizione *i;* $CR_i$ è il fattore di contatto, cioè il rateo di comparto ambientale associato alla via di esposizione *i* assunto dal soggetto [$m^3 d^{-1}$ o kg $d^{-1}$], *EF* è la frequenza di esposizione [d/anno], *ED* è la durata di esposizione [anni], *BW* il peso corporeo in kg e *AT* il tempo di mediazione dell'esposizione [giorni].

Il calcolo del rischio si differenzia, inoltre, a seconda se l'inquinante sia

cancerogeno oppure non cancerogeno. Per le sostanze cancerogene si ha che:

$$R = E \times SF$$

Dove *R* (Rischio [%]) rappresenta la probabilità di casi incrementali di tumore nel corso della vita, causati dall'esposizione alla sostanza, rispetto alle condizioni di vita usuali, *SF* (Slope Factor [mg/kg d]$^{-1}$) indica la probabilità di casi incrementali di tumore nella vita per unità di dose.

Per le sostanze tossiche non cancerogene la valutazione del rischio si basa sulla formula:

$$HQ = E / RfD$$

Dove *HQ* (Hazard Quotient [adim]) è un indice di pericolo che esprime di quanto l'esposizione alla sostanza supera la dose tollerabile o di riferimento; *RfD* (Reference Dose [mg kg$^{-1}$ d$^{-1}$]) è la stima dell'esposizione media giornaliera che non produce effetti avversi apprezzabili sull'organismo umano durante il corso della vita.

Il rischio per la salute umana viene differenziato tra individuale e cumulativo. Si definisce:

- rischio e indice di pericolo individuale (*R* e *HQ*): rischio dovuto ad un singolo contaminante per una o più vie d'esposizione;
- rischio e indice di pericolo cumulativo ($R_{TOT}$ e $HQ_{TOT}$): rischio dovuto alla cumulazione

degli effetti di più sostanze per una o più vie d'esposizione.

Il calcolo del rischio per la salute umana associato a più specie chimiche inquinanti e a una o più

modalità di esposizione (rischio cumulativo) è il seguente:

$$R_{TOT} = \sum_{i=1}^{n} R_i \quad \text{e} \quad HQ_{TOT} = \sum_{i=1}^{n} HQ_i$$

dove $R_{TOT}$ e $HQ_{TOT}$ rappresentano il rischio cumulativo e l'indice di pericolo cumulativo causati dall'esposizione contemporanea alle *n* sostanze inquinanti. Un modo rigoroso e giustificato scientificamente di calcolo del rischio cumulato è tutt'oggi fonte di incertezza: non è infatti ancora chiaro quale sia l'azione congiunta di più inquinanti tossici o cancerogeni agenti in contemporanea su un individuo. Il fatto di sommare i singoli rischi è da considerarsi un approccio conservativo, perché nel caso gli effetti delle varie sostanze non siano sinergici l'analisi porterebbe a sovrastimare il rischio reale.

Per quanto riguarda il criterio di cumulazione delle concentrazioni individuali dovute a più vie d'esposizione, il calcolo del rischio per la salute umana viene svolto in

funzione delle sorgenti di contaminazione considerate, che sono: suolo superficiale, suolo profondo, falda e prodotto libero.

Per il suolo superficiale il rischio viene stimato scegliendo il valore più conservativo tra il rischio derivante dalle modalità di esposizione che hanno luogo in ambienti confinati (*indoor*) e aperti (*outdoor*). Le modalità di esposizione considerate sono: inalazione di vapori o polveri, ingestione e contatto dermico.

In maniera analoga, per il suolo profondo e la falda vengono presi in considerazione i valori più conservativi di rischio, assumendo che la modalità di esposizione sia solo l'inalazione di vapori o polveri, sempre però distinguendo la situazione *indoor* da quella *outdoor*.

La procedura di analisi di rischio può essere condotta in due modalità:

- la modalità diretta (*forward mode*) permette il calcolo del rischio associato al recettore esposto, derivante da una sorgente di contaminazione di concentrazione nota. In particolare, nota la concentrazione rappresentativa della sorgente, si stima l'esposizione da parte del recettore, tenendo conto, sulla base della modalità di esposizione, dell'attenuazione dovuta ai fattori di trasporto; si valuta la tossicità delle sostanze mediante i parametri RfD e SF ed infine si calcola il rischio.
- La modalità inversa (*backward mode*) permette il calcolo della massima concentrazione ammissibile in sorgente compatibile con il livello di rischio ritenuto accettabile per il recettore esposto. Tale concentrazione rappresenta l'obiettivo di bonifica specifico per il sito in esame. In particolare, stabilita la soglia di rischio tollerabile, si ottiene una concentrazione accettabile nel punto di esposizione ed infine, per mezzo dei fattori di trasporto, si arriva a stimare la concentrazione accettabile in sorgente.

### 5.7.1 Criteri di accettabilità del rischio

Per quanto riguarda gli effetti cancerogeni sulla salute umana, si reputa necessario stabilire una soglia di rischio al di sotto della quale si ritiene tollerabile la probabilità incrementale di effetti cancerogeni sull'uomo. Nel caso della procedura diretta, il valore di soglia viene utilizzato a valle della procedura come termine di confronto con il valore del rischio *R* calcolato; nel caso di applicazione della procedura inversa, viene invece utilizzato a monte dell'analisi, per risalire ai valori obiettivi di bonifica sito-specifici, detti Concentrazioni Soglia di Rischio (o *CSR*).

In Italia, solitamente, viene considerato accettabile un valore di rischio per la

salute umana pari a $10^{-6}$ per il rischio di una singola sostanza, il che significa che la probabilità incrementale per un individuo di contrarre un tumore è pari a 1/1000000. Il valore soglia di rischio cumulato, sommatoria dei rischi di tutte le sostanze cancerogene presenti nel sito, è invece di $10^{-5}$.

Il valore di riferimento per il rischio di sostanze tossiche non cancerogene è 1, il che significa che la dose giornaliera assunta dev'essere inferiore alla dose quotidiana accettabile di riferimento (*RfD*).

Sia per le sostanze cancerogene sia per quelle tossiche, quindi, l'analisi di rischio prevede di calcolare i valori assunti individuali e cumulativi e il confronto con i rispettivi valori soglia.

### 5.7.2 Determinazione di Reference Dose e Slope Factor

Come accennato in precedenza, la *Reference Dose* o *RfD* è la dose di sostanza tossica per la quale in letteratura non vengono riportati casi effetti avversi per l'uomo, espressa in mg $Kg^{-1}$ $d^{-1}$. Come per le sostanze cancerogene, questo valore viene estrapolato da curve dose-risposta determinate con dati sperimentali, individuando il valore *NOAEL* (“*no observed adverse effect concentration”*), in corrispondenza del quale non si registrano effetti negativi, e dividendolo per fattori di incertezza (*UF*) e/o di correzione (*MF*).

Per quanto riguarda le sostanze cancerogene, si assume che non esista un valore soglia di dose sotto il quale non vi siano effetti, cioè non esiste un livello di esposizione che non abbia probabilità, per quanto piccola, di generare effetti tumorali. In questo caso viene quindi stimata la probabilità di un individuo di contrarre neoplasia nel corso della vita. A tale scopo, viene calcolato lo *Slope Factor*, equivalente al rischio per dose unitaria, corrispondente al coefficiente angolare della retta interpolante i punti osservati di dose/risposta. Essendo comunque questo valore affetto da incertezza, come approccio cautelativo viene considerato il valore corrispondente al limite superiore dell'intervallo di confidenza al 95%.

In figura 1 si riporta, a titolo d'esempio, l'andamento tipico delle curve dose-risposta per inquinanti cancerogeni e tossici.

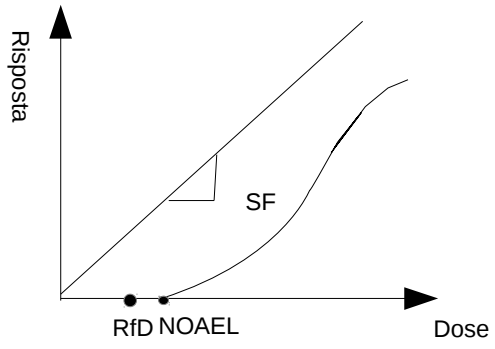

*Fig.1 Curve dose-risposta*

In figura 1 sono riportati due andamenti tipici rispettivamente per la curva dose-risposta di sostanze con effetti cancerogeni (la linea retta) e tossici. Lo slope factor coincide col valore del coefficiente angolare della retta. L'indice NOAEL, spiegato in precedenza, è il punto in cui la curva assume valore nullo mentre il valore RfD è pari al valore NOAEL ridotto di un certo fattore di sicurezza.

# 6. FORMULAZIONE DI MODELLI

*"Se torturi i numeri abbastanza a lungo, confesseranno qualsiasi cosa"*

Gregg Easterbrook

Lo studio scientifico dei fenomeni naturali, come per esempio le migrazioni di specie animali o il loro andamento demografico, piuttosto che il modello di deflusso di un bacino idrografico o la previsione di determinati eventi di piena, avviene attraverso uno studio statistico dei dati a disposizione. Qualsiasi sia l'approccio statistico di studio delle osservazioni, l'idea di base è quella di formulare un modello matematico stocastico in cui i dati osservati sono estrazioni di una variabile aleatoria, che chiameremo $X$, funzione di uno o più parametri $\theta$. Questa funzione verrà indicata come $Pr(X \mid \theta)$ o $f(X \mid \theta)$. L'intento è quello di poter fare asserzioni su tali parametri, per esempio determinarne la media o la varianza, basandosi sui valori osservati della variabile aleatoria $X$.

L'approccio classico si differenzia da quello bayesiano per quanto riguarda lo studio dei parametri. Nel primo caso, infatti, questi sono visti come costanti da stimare (ad esempio il tasso di mortalità annuale degli individui di una popolazione). La stima avviene, nella maggior parte dei casi, attraverso il metodo della massima verosimiglianza, come si vedrà in seguito, che porterà ad assegnare ai vari parametri un valore fisso. L'approccio classico prevede poi la formulazione di diversi modelli che descrivono il fenomeno studiato ed in seguito verrà selezionato il più idoneo rispetto a determinati criteri, che possono essere differenti a seconda delle necessità. Da notare che, qualsiasi sia il criterio di selezione, il modello migliore sarà comunque tra quelli che meglio descrivono il set di dati disponibili e non il fenomeno di per sé. Questo metodo, quindi, è tanto più affidabile quanto più il set di dati disponibili è numeroso.

Nell'approccio bayesiano, invece, i parametri, alla stregua dei dati, sono considerati come variabili aleatorie, quindi descritti da una certa distribuzione di probabilità. Come meglio verrà descritto successivamente, il metodo prevede una descrizione *a priori* di questa distribuzione, basata sulle conoscenze dello statistico o su esperienze di casi simili simili (si vedrà anche il caso in cui non vi sia alcun tipo di informazione a priori). In seguito alla raccolta dei dati, conoscendo sia la distribuzione di questi ultimi in funzione dei parametri sia la distribuzione a priori dei parametri, si ricava

la distribuzione dei parametri in funzione dei dati attraverso la semplice applicazione del teorema di Bayes. Questa verrà chiamata distribuzione a posteriori ed è l'obiettivo dell'analisi. Ciò che si ottiene, in particolare, è la distribuzione congiunta dei parametri, dalla quale, se necessario, è possibile risalire alle singole distribuzioni marginali. Come si capirà meglio in seguito, l'approccio bayesiano presenta numerosi vantaggi rispetto a quello classico, ma soprattutto fornisce un'interpretazione tutta nuova della realtà: abbandona infatti la visione deterministica secondo cui i parametri sono costanti il cui valore può essere esattamente calcolato e lascia spazio ad una descrizione più flessibile del mondo reale, le cui innumerabili variabili continue variano in maniera quasi sempre a noi sconosciuta. Spesso la scelta tra approccio classico o bayesiano è più filosofica che pragmatica, e dipende, tra le altre cose, dal tipo di problema esaminato e dall'accuratezza dei risultati che viene ricercata. Uno dei punti a favore del metodo bayesiano è che permette di affrontare problemi complessi con un piccolo sforzo computazionale aggiuntivo rispetto al metodo classico, evitando assunzioni e semplificazioni poco realistiche; inoltre la scelta e la combinazione di più modelli può essere effettuata in maniera relativamente semplice anche all'interno di un set numeroso.

### *6.1. Nota storica*

Dalla sua comparsa nella seconda metà del diciottesimo secolo, il metodo bayesiano è stato quello dominante per più di centocinquant'anni. Anche se affonda le sue basi nel celebre teorema formulato dal reverendo Thomas Bayes (1702 – 1761), non è da imputare a lui la messa a punto del metodo di analisi statistica. Con l'avvento del ventesimo secolo l'approccio bayesiano fu abbandonato in favore dell'approccio classico per due principali motivi: innanzi tutto presentava rilevanti difficoltà computazionali rispetto agli strumenti disponibili all'epoca, difficoltà che diventavano via via più ingenti a causa dei sempre più complessi problemi statistici da affrontare; inoltre, divenne sempre più diffusa tra i ricercatori dell'epoca la perplessità riguardo la soggettività del metodo, che prevede la scelta arbitraria di una distribuzione a priori per i parametri nell'analisi. Venne così adottato e messo a punto il metodo classico di analisi statistica dei dati, che dominò la scena per tutto il ventesimo secolo.

Il ritorno all'approccio bayesiano che vediamo in questi anni è dovuto ad innovazioni sia nel campo della statistica sia nell'avvento di computer di ultima generazione, che rendono trattabili con con questo metodo anche i problemi più complessi.

In particolare, un nuovo algoritmo chiave per l'analisi bayesiana è il Markov Chain Monte Carlo (MCMC), che consente di effettuare simulazioni a partire dalla distribuzione a priori; allo stesso tempo, la sua implementazione con calcolatori abbastanza potenti permette di risolvere un numero adeguato di iterazioni per raggiungere la convergenza (concetto che verrà spiegato meglio in seguito).

## *6.2. Approccio Classico*

La funzione di verosimiglianza, come già accennato, è un concetto statistico basilare nell'approccio classico di formulazione di modelli matematici. Come descritto sopra, lo scopo dell'inferenza statistica è di poter fare asserzioni sui parametri della popolazione ($\theta$) basandosi sull'osservazione di valori di una o più variabili aleatorie ($X$). Il nostro interesse si sposta quindi sui valori di $\theta$ , condizionati ai dati osservati $X$, attraverso una funzione che indicheremo come $L\ (\theta \mid X)$. Una volta determinata la distribuzione di probabilità di una variabile casuale $f\ (X|\theta)$ , calcolare la funzione di verosimiglianza è solo una questione di notazione e non è richiesto alcun metodo matematico addizionale, infatti $L(\theta\ |X) \equiv f\ (X|\theta)$. In altre parole, determinare la funzione di verosimiglianza equivale a determinare la probabilità di osservare un particolare set di dati. Lo scopo di questo metodo è quello di produrre una stima dei parametri di interesse. Gli stimatori massimo-verosimili (*MLEs*) sono i valori dei parametri che massimizzano la funzione di verosimiglianza, data l'osservazione del campione. Questo equivale a chiedersi: dato il modello implicito, quali sono i valori dei parametri che meglio descrivono i dati osservati?

Per ottenere questi valori sono diversi gli approcci utilizzabili. A volte può essere sufficiente visualizzare i dati su un grafico e ricavare geometricamente il massimo della funzione. Un'alterativa più spesso utilizzata è quella di calcolare le derivate parziali della funzione di interesse, porle uguali a zero e risolvere le equazioni risultanti.

### 6.2.1. Esempio

In una data area geografica vi sono cinque siti circoscritti di interesse in cui ci si aspetta di poter osservare una data specie animale. Dopo aver eseguito delle campagne di osservazione, la specie in esame è stata avvistata in solo $x$ siti su cinque. Senza ulteriori informazioni teoriche a disposizione, si può considerare che la variabile $x$ sia descritta da una binomiale, data la probabilità $\theta$ di osservare la specie in un sito qualsiasi:

$$Pr(X|\theta)=\binom{5}{X}\theta^{x}(1-\theta)^{5-X} \qquad (1)$$

Immaginiamo ora di voler calcolare la probabilità di osservare due siti occupati su cinque. La (1) diventa:

$$Pr(2|\theta)=\binom{5}{2}\theta^{2}(1-\theta)^{3} \qquad (2)$$

Si ponga il caso, ora, di aver effettivamente osservato due siti su cinque occupati dalla specie sotto studio. Quindi *x = 2* sono i dati disponibili. L'approccio classico prevede che la funzione di verosimiglianza dei parametri *L (θ|X)*, come detto, sia la stessa funzione di probabilità dei dati in funzione dei parametri, *Pr (X|θ)*. Quindi:

$$L(\theta|x=2)=\binom{5}{2}\theta^{2}(1-\theta)^{3}$$

Dove il parametro $\theta$ indica la probabilità che un qualsiasi sito dei cinque sia occupato. Si può dimostrare che il valore che rende massima la verosimiglianza è $\hat{\theta}=\frac{x}{n}=\frac{2}{5}=0{,}4$ . In questo caso è anche possibile individuare graficamente il valore di $\theta$ massimo verosimile, facendo variare i valori di $\theta$ sulle ascisse e sulle ordinate i corrispondenti valori assunti dalla (2).

### 6.2.2. Proprietà degli stimatori massimo-verosimili

Lo stimatore MLE $\hat{\theta}$ ha approssimativamente una distribuzione normale per campioni con elevata numerosità. Inoltre, la sua distribuzione converge asintoticamente a una gaussiana al crescere della numerosità del campione.

Anche se lo stimatore non è corretto, il suo valore atteso converge al parametro $\theta$ al crescere del campione osservato ed è quindi asintoticamente corretto.

Uno stimatore MLE è consistente, cioè all'aumentare dell'informazione, ossia della numerosità del campione, la sua distribuzione di probabilità si concentra in corrispondenza del valore del parametro da stimare.

La varianza di $\hat{\theta}$ è asintoticamente minima, cioè la sua varianza è la più piccola tra tutti gli stimatori corretti di $\theta$ quando il campione è grande.

### 6.2.3. Intervalli di confidenza

Gli intervalli di confidenza sono uno tra i modi di esprimere l'incertezza di uno stimatore. A volte è possibile ottenere intervalli di confidenza esatti, in altri casi si ottengono intervalli asintotici o molto ampi, basati o sulla distribuzione normale asintotica degli *MLE* o su un appropriato rapporto log-verosimile con distribuzione asintotica chi-quadro.

L'intervallo asintotico con livello di confidenza del 100(1-α)% basato sulla distribuzione normale è della forma:

$$\hat{\theta} \pm z_{(\frac{\alpha}{2})} \sqrt{Var(\hat{\theta})}$$

dove $z_{\alpha/2}$ è il quantile della distribuzione normale.

Un intervallo di confidenza basato sul rapporto *log*-verosimile può essere determinato a partire dalla seguente quantità:

$$\varphi(\theta_0) = 2 \ln \frac{\boldsymbol{L}(\hat{\theta_0}, \hat{\theta})}{\boldsymbol{L}(\hat{\theta_0}, \hat{\theta}*)}$$

dove al numeratore vi è la funzione di verosimiglianza valutata nei valori degli *MLE* di tutti i parametri e il denominatore è la funzione di verosimiglianza valutata in corrispondenza di tutti gli *MLE* tranne quello del parametro in esame, cioè non è presente la stima di $\theta_0$, che può assumere qualsiasi valore. $\hat{\theta} e \hat{\theta}^*$ Sono da intendersi come vettori. Si trovano i due valori di $\theta_0$ che soddisfano l'equazione $\varphi(\theta_0) = \chi_1^2(\alpha)$ dove $\chi_1^2(\alpha)$ è il 100α-esimo percentile della distribuzione chi-quadro con un grado di libertà.

## *6.3. Approccio Bayesiano*

L'approccio bayesiano è piuttosto differente da quello classico, nonostante entrambi si basino sulla verosimiglianza. Come si è detto, in questo tipo di procedura i dati sono visti come variabili connotate da una certa distribuzione di probabilità, esattamente come i dati osservati. Prima dell'estrazione del campione da analizzare, i parametri sono descritti da una distribuzione a priori o *prior*. Dopo l'acquisizione dei dati, questa distribuzione è sostituita da quella a posteriori o *posterior*. Al contrario dell'approccio classico, non si prevede alcuna ottimizzazione. Il passaggio dalla distribuzione a priori a quella a posteriori si basa sulla semplice applicazione del teorema di Bayes. Il problema

dell'applicazione del teorema è che fornisce la distribuzione a posteriori congiunta di tutti i parametri; se si fosse interessati alla funzione di probabilità di un solo parametro, sarà quindi necessario ricavare la sua distribuzione marginale. Per fare questo, come consuetudine, sarà necessario integrare la funzione congiunta ed è qui che nascono le complicazioni di questo metodo. Per la complessità che spesso comporta questo calcolo, la soluzione non è quasi mai ricavata per via analitica, bensì implementando tecniche di simulazione che forniscono campioni di dati estratti dalla distribuzione a posteriori. Se si ottengono campioni della distribuzione a posteriori congiunta dei parametri, questi deriveranno automaticamente dalle distribuzioni marginali dei singoli parametri. Su questa importante osservazione si basa l'algoritmo Catena di Markov Monte Carlo o MCMC *(Markov Chain Monte Carlo).*

### 6.3.1. Teorema di Bayes

Nell'applicare il metodo bayesiano si parte da alcune assunzioni a priori riguardo al sistema in esame, che poi verranno aggiornate sulla base dei dati osservati, *x*. La procedura di aggiornamento è basata sul teorema di Bayes:

$$\pi(\theta/x)=\frac{f(x/\theta)\,p(\theta)}{f(x)}$$

Dove x sono i dati osservati e θ i parametri del modello. La funzione *f* è la distribuzione di probabilità dei dati condizionata ai parametri del modello. Con *p* si indica la distribuzione a priori dei parametri e con *π* quella a posteriori. Al contrario del metodo classico, si ottiene una distribuzione di probabilità dei parametri e non una stima puntuale. La funzione *f(x)* è funzione solo dei dati e non dei parametri ed è solitamente omessa dal teorema, che viene espresso nella forma

$$\pi(\theta/x)\propto f(x/\theta)\,p(\theta)$$

Il termine *f(x)* al denominatore è semplicemente la costante di normalizzazione, che fa si che la distribuzione a posteriori sia propriamente una distribuzione, cioè che il suo integrale sia pari all'unità. Prima di applicare il teorema, quindi, è necessario definire un adeguato modello probabilistico dei dati in funzione dei parametri e un'appropriata distribuzione a priori dei parametri. Si ottiene una nuova distribuzione dei parametri, che formalmente combina le assunzioni a priori descritte da *p* con la nuova informazione derivante dall'osservazione del campione.

Da notare come, col metodo classico, non si fa nessuna asserzione sulla

distribuzione a priori ma semplicemente si impone $f(\theta \mid x) \equiv f(x \mid \theta)$, dove $f(\theta \mid x)$ si utilizza come funzione di verosimiglianza.

### 6.3.2. Distribuzione a priori

La scelta di una distribuzione a priori dei parametri è evidentemente ricca di soggettività e rimane oggi oggetto di dibattito scientifico, specialmente quando si hanno poche o nessuna informazione a disposizione. Quando è possibile, la distribuzione a priori è ricavata da opinioni di esperti e/o da studi precedenti sull'argomento in esame. In conclusione, i casi possono essere due: o non si ha nessun tipo di informazione a priori o si ha un'informazione che va espressa sotto forma di un'adeguata distribuzione di probabilità. A seconda della situazione, quindi, verranno formulate due tipi di distribuzioni a priori, chiamate rispettivamente non-informative ed informative.

#### 6.3.2.1. Distribuzioni a priori uniformi

In assenza di informazioni a priori su uno o più parametri, ci si deve assicurare che questa incertezza si rifletta nella distribuzione a priori. Tipicamente questo si traduce nel scegliere una distribuzione con una grande varianza. Considerare una distribuzione a priori uniforme per i parametri può sembrare un approccio rischioso, ma raramente questo tipo di distribuzioni sono invarianti nella riparametrizzazione del modello. In altre parole non hanno una così grande influenza nella distribuzione a posteriori, come invece si potrebbe immaginare. Un altro problema della distribuzione uniforme come distribuzione a priori è che per variabili continue queste sono distribuzioni improprie, a meno che non siano assegnati dei limiti allo spazio dei parametri. Distribuzioni a priori improprie portano a distribuzioni a posteriori improprie; d'altro canto, l'imposizione di limiti troppo restrittivi si può tradurre in distribuzioni a posteriori restrittive.

#### 6.3.2.2. Distribuzioni a priori coniugate

Si dice che una distribuzione a priori è coniugata quando ha la stessa forma della distribuzione a posteriori dei parametri. Per esempio, se si ha un campione di n variabili indipendenti identicamente distribuite $x_1,....x_n$ tale che $X_i \sim N(\mu, \sigma^2)$ con μ ignota e $\sigma^2$ nota. Specifichiamo una distribuzione a priori normale per μ, tale che $\mu \sim N(0, \tau)$. Allora

la distribuzione a posteriori è data da:

$$f(\mu \mid X=x) \sim N\left(\tau n \frac{\bar{x}}{(\tau^2 n+\sigma^2)}, \sigma^2 \frac{\tau^2}{(\tau^2 n+\sigma^2)}\right)$$

La distribuzione a priori e a posteriori di $\mu$ sono entrambe normali. In altre parole, si può dire che una distribuzione a priori normale per la media di una distribuzione normale è una distribuzione coniugata. Vi sono molti altri casi di distribuzioni a priori coniugate per il cui approfondimento si rimanda a Chapman et al. (2010). Il fatto di scegliere una distribuzione coniugata semplifica molto l'implementazione del metodo ma questa motivazione non giustifica una sua scelta incondizionata, date le grandi potenzialità dei moderni calcolatori. Nel senso che, nel caso si posseggano informazioni a priori sulla distribuzione dei parametri, sarà sempre preferibile farne uso piuttosto che assumere una distribuzione a priori che faciliti i calcoli.

### 6.3.2.3. Distribuzione di Jeffrey

Questo tipo di distribuzione intende minimizzare l'influenza della prior sulla distribuzione a posteriori. É basata sulla stima della matrice informazione di Fisher. L'informazione di Fisher è la varianza della derivata logaritmica associata a una data funzione di verosimiglianza. Può essere interpretata come l'ammontare di informazione contenuta in una variabile casuale osservabile *X*, concernente un parametro non osservabile *θ*, da cui dipende la distribuzione di *X*. Utilizzando le stesse notazioni di cui sopra, la distribuzione a priori di Jeffrey è data da:

$$p(\theta) \propto \sqrt{(I(\theta / x))}$$

Dove *I( θ | x)* è l'informazione di Fisher:

$$I(\theta / x) = -E\left[\frac{d}{d\theta} \ln L(\theta \mid x)\right]^2$$

La distribuzione a priori di Jeffrey è una distribuzione impropria.

### 6.3.2.4. Distribuzioni informative

L'intento delle distribuzioni informative è quello di riflettere nelle distribuzioni a posteriori l'informazione sui parametri disponibile a priori. Questo significa scegliere una famiglia di distribuzioni per ciascun parametro del modello e poi tentare di cercare i parametri della prior che riflettano l'informazione disponibile in maniera accurata. Spesso

risulta essere un processo iterativo dove in principio si ottiene un range di valori entro cui possono variare i parametri. Si può anche pensare di usare dati indipendenti per costruire una distribuzione informativa. Si supponga, per esempio, di avere due serie di dati indipendenti, $x_1$ e $x_2$. Inizialmente si assume una prior per $\theta$, $p(\theta)$, e si calcola la distribuzione a posteriori dei parametri usando solo i dati $x_1$:

$$\pi(\theta/x_1) \propto f(x_1/\theta)\, p(\theta)$$

Dopodiché si usa questa distribuzione a posteriori come *prior* corrispondente all'analisi dei dati $x_2$:

$$\pi(\theta/x_1, x_2) \propto f(x_2/\theta)\pi(\theta/x_1) \propto f(x_2/\theta) f(x_1/\theta)\, p(\theta) = f(x_1, x_2/\theta)\, p(\theta)$$

se i dati $x_1$ e $x_2$ sono indipendenti. Si può dire che costruire una distribuzione a priori informativa da dati indipendenti è equivalente ad analizzare in maniera congiunta tutti i dati, utilizzando la stessa distribuzione a priori dei parametri.

### 6.3.3. Analisi di sensitività

Risulta necessario verificare quanto la scelta della *prior* influisca sui risultati della distribuzione a posteriori. Nel caso si ritenga che i dati contengano più informazioni sulla *posterior* rispetto alla *prior*, quest'ultima dovrebbe avere una bassa sensitività, in modo che i risultati siano dominati dall'informazione derivante dai dati. Una semplice tecnica di analisi di sensitività prevede di far variare la varianza della *prior* per vedere quanto questa influisca nei risultati. Questo si può fare, ad esempio, imponendo tre livelli di varianza per ogni parametro e condurre analisi separate per ogni combinazione.

## *6.4. Metodo Monte Carlo*

L'applicazione del metodo Bayesiano ha visto un grande sviluppo agli inizi degli anni '90 grazie all'avvento dei computer ed all'introduzione nella letteratura statistica dell'algoritmo noto come *Markov Chain Monte Carlo* (MCMC, Smith and Gelfand, 1992). Si intende ora descrivere separatamente il metodo di integrazione Monte Carlo e le catene di Markov, per poi vederli combinati nel suddetto algoritmo.

### 6.4.1. Introduzione

Come è noto, quando si ha a che fare con lo studio di un processo naturale, vengono eseguiti in laboratorio esperimenti fisici che riproducono le dinamiche del fenomeno. Vengono quindi effettuate delle simulazioni artificiali del sistema studiato e i dati ottenuti vengono poi elaborati statisticamente per poi poter fare asserzioni sul sistema reale, per esempio stimando i parametri di un modello o valutando la risposta del sistema soggetto a determinate perturbazioni.

Il metodo Monte Carlo fa parte dei così detti metodi di simulazione teorici, in cui vengono fatte simulazioni numeriche, e non fisiche, del processo in esame. In particolare, viene usato nel caso in cui si abbia a che fare con variabili aleatorie. Il metodo permette di generare un set di valori della variabile aleatoria considerata estraendoli dalla sua distribuzione di probabilità. Ripetendo la procedura si possono ottenere diverse serie di soluzioni che hanno la stessa probabilità di verificarsi e che sono statisticamente equivalenti alle osservazioni delle simulazioni sperimentali. Per applicare il metodo Monte Carlo è necessario conoscere o assumere a priori una certa distribuzione di probabilità per la variabile in esame. Può risultare interessante notare come questa procedura permetta di usare un metodo statistico per risolvere problemi deterministici.

### 6.4.2. L'ago di Buffon

Uno dei primi esempi di applicazione del metodo Monte Carlo risale al XVIII secolo quando Georges Louis Leclerc conte di Buffon pose il seguente problema matematico:

*si consideri un pavimento in parquet, costituito da strisce di legno parallele tutte della stessa larghezza. Se faccio cadere un ago sul pavimento in maniera del tutto casuale, qual'è la probabilità che questo si posizioni a cavallo di una linea fra le due strisce?*

Di seguito si vedrà come il problema possa essere risolto con un procedimento del metodo Monte Carlo e come si possa ottenere un'approssimazione del valore di $\pi$.

Riformuliamo il problema in termini matematici: si consideri un piano orizzontale su cui giacciono rette parallele a distanza $a$. In modo casuale viene tracciato sul piano un

segmento di lunghezza $b<a$. Il problema è determinare quale sia la probabilità che il segmento intersechi una delle due rette.

Per la risoluzione del problema torna utile definire due variabili aleatorie. Sia $X$ la variabile che rappresenta la distanza del centro del segmento rispetto alla linea più vicina e $Y$ il valore dell'angolo acuto tra il segmento e le linee.

La funzione densità di probabilità di $X$ tra $a/2$ e 0 (cioè dal punto medio tra due linee e una linea) può essere considerata distribuzione uniforme, perché l'ago cade in maniera del tutto casuale ed il suo centro può cadere in qualsiasi punto tra le due rette con la stessa probabilità:

$$f_X(x)=\frac{2}{a}I_{[0;\frac{a}{2}]}$$

Analogamente, tra 0 e $\pi/2$ la variabile $Y$ avrà distribuzione uniforme:

$$f_Y(y)=\frac{2}{\pi}I_{[0;\frac{\pi}{2}]}$$

Se è vero che $X$ e $Y$ sono indipendenti, allora la loro distribuzione di probabilità congiunta è uguale al prodotto delle distribuzioni marginali:

$$f_{(x,y)}(x,y)=\frac{4}{a\pi}I_{[0,\frac{a}{2}]}I_{[0;\frac{\pi}{2}]}$$

Se dal punto medio del segmento si traccia una linea parallela alle rette del piano, $(b/2)sinY$ è la distanza dell'estremità del segmento da questa linea. Quindi il segmento tracciato a caso sul piano interseca una delle rette se $X\leq\frac{b}{2}sinY$ .

E' possibile dunque calcolare la probabilità che il segmento intersechi una retta:

$$p=\frac{4}{a\pi}\int_0^{(\frac{\pi}{2})}\int_0^{(\frac{b}{2}\sin y)}dx\,dy=\frac{2\mathrm{b}}{a\pi} \quad (1)$$

Il procedimento di risoluzione del problema con il metodo Monte Carlo può essere riassunto come segue. Innanzi tutto si esplicitano i valori di $a$ e di $b$. Dopo di che è necessario generare una serie di coppie indipendenti delle variabili $X$ e $Y$, i cui valori sono estratti da una distribuzione uniforme.

Successivamente la probabilità $p$ viene stimata statisticamente come il rapporto tra il numero di coppie che soddisfano la condizione $X\leq\frac{b}{2}sinY$ ed il numero totale di coppie generate.

Il valore di $\pi$ viene infine stimato semplicemente invertendo la (1), esplicitando il

valore π.

E' intuitivo fin da subito immaginare come all'aumentare della numerosità della serie di coppie estratte aumenti l'accuratezza del metodo. Su questo aspetto specifico ci soffermeremo nei paragrafi successivi.

### 6.4.3. Trasformazione integrale di probabilità

L'aspetto critico del metodo Monte Carlo è quindi quello di generare una serie sintetica di valori della variabile aleatoria in esame estratti dalla sua distribuzione di probabilità. Di seguito si espone un metodo che permette di generare serie numeriche estratte da qualsivoglia distribuzione a partire dalla distribuzione uniforme nell'intervallo (0,1), riducendo il problema alla generazione di numeri casuali nel suddetto intervallo.

Si consideri una variabile aleatoria $X$ continua e la sua funzione cumulata di probabilità $F_x(.)$, monotona crescente in senso stretto. Si definisce una seconda variabile aleatoria $U = g_x(x) = F_x(X)$, trasformazione della variabile $X$. Ne consegue che $F^{-1}{}_X(u) = \xi(u) = x$ (l'invertibilità di $F_x(.)$ è garantita dal fatto che sia definita continua monotona crescente). $F^{-1}{}_X(u)$ è quindi definita per valori di u compresi tra 0 e 1.

Considerando quindi la variabile aleatoria $U = F_x(x)$, si possono scrivere le seguenti relazioni:

$$F_U(u) = P[U \le u] = P[F_X(x) \le \xi(u)] = P[X \le \xi(u)] = F_X(\xi(u)) = u$$

Ne consegue che la funzione densità di probabilità di $U$ è data da

$$f_U(u) = \frac{dF_U(u)}{du} = 1 \quad per\ 0 < u < 1.$$

Si è dimostrato quindi che $U$ è una variabile uniforme nell'intervallo $(0,1)$:

$$U \sim Uniforme(0,1).$$

La trasformazione $U = F_X(x)$ si chiama trasformazione integrale di probabilità.

Generalizzando si può affermare che: se $X$ è una variabile aleatoria continua con funzione di ripartizione $F_X(.)$ continua strettamente crescente, allora la variabile aleatoria $U = F_X(x)$ ha distribuzione uniforme nell'intervallo $(0,1)$. Viceversa, se $U \sim Unif(0,1)$ allora la distribuzione di $X = \xi(u)$ è $F_X(.)$.

Se si vuole ottenere un set di valori $x_i$ della variabile $X$ con funzione di ripartizione $F_X(x)$ è quindi sufficiente (ma non banale) generare un set di valori casuali $u_i$ estratti da una distribuzione uniforme $U(0,1)$.

### 6.4.4. Generazione di numeri casuali

#### 6.4.4.1. Numeri casuali estratti da una variabile uniforme U(0,1)

I metodi di estrazione di valori casuali da una variabile uniforme nell'intervallo *(0,1)* sono deterministici, cioè vi sono procedure sistematiche in grado di fornire valori pseudo-casuali di una variabile uniforme a partire da un valore iniziale arbitrario, chiamato *seed*. Per esempio, i generatori di numeri casuali dei computer utilizzano quasi sempre il così detto generatore lineare congruenziale. Questo è un algoritmo basato sul calcolo ricorsivo di una sequenza di numeri interi $n_1,n_2,n_3....$, ognuno compreso tra 0 e $m$-1, ottenuti tramite una trasformazione lineare:

$$n_{i+1}=(an_i+c)(modulo\ m)$$

dove $a$ e $c$ sono numeri interi chiamati moltiplicatore ed incremento rispettivamente. "Modulo m" significa che $n_{i+1}$ è il resto della divisone $(an_i+c)/m$. Indicando con $\eta_i=Int[(an_i+c)/m]$ allora $n_{i+1}=an_i+c-m\eta_i$,

Definendo $u_{(i+1)}=\frac{n_{(i+1)}}{m}=\frac{an_i}{m}+\frac{c}{m}-\mathrm{Int}\left[\frac{(an_i+c)}{m}\right]$ , gli $u_i$ sono estratti da una distribuzione uniforme *U(0,1)*. In fase di implementazione dell'algoritmo si nota come i numeri generati si ripetano con un certo periodo e per questo vengono chiamati pseudo-casuali. La qualità dei risultati e l'ampiezza del periodo di un ciclo dipendono dalla scelta delle costanti *a*, *c* e *m*. In particolare, il periodo di un ciclo non è maggiore di *m* e cresce al crescere di *m*, per questo si usa scegliere *m* il più grande possibile. Inoltre *c* e *m* non devono avere nessun fattore comune ed *a* dev'essere sufficientemente grande. Infine, essendo che tutti i numeri compresi nell'intervallo *(0 , m-1)* occorrono dopo un certo periodo di tempo, la scelta di $n_0$ non compromette l'esito della simulazione.

Nell'implementazione di questa tecnica al computer il vantaggio sta nella rapidità del processo, essendo richiesti pochi calcoli. Lo svantaggio è che una volta selezionato il *seed* ed i parametri tutta la serie è prevedibile.

#### 6.4.4.2. Numeri casuali estratti da variabili continue

Attraverso la trasformazione integrale di probabilità è possibile ottenere una realizzazione $x$ di una variabile aleatoria $X$ con distribuzione continua $F_X(x)$ a partire da un

valore u della variabile uniforme standard, dove *x* è il valore dell'*u*-quantile di *X*. Questo è possibile però solo nel caso in cui la distribuzione di *X* sia invertibile analiticamente, come nel caso di una distribuzione esponenziale. Vi sono altresì diverse distribuzioni non invertibili. Per ottenere dei valori appartenenti a queste distribuzioni esistono metodi alternativi alla trasformazione integrale di probabilità. Ad esempio, se una variabile aleatoria *X* può essere espressa come funzione di altre variabili aleatorie appartenenti a distribuzioni invertibili, cioè

$X = f(Y_1,Y_2...Y_n)$, e sono disponibili metodi per generare $y_1,y_2...y_n$, allora *X* può essere determinata come $f(y_1,y_2...y_n)$ .

### 6.4.4.2.1 Metodo della decomposizione

Secondo il teorema delle probabilità totali, la funzione distribuzione di probabilità di una variabile *X* può essere espressa come la somma di altre funzioni di distribuzione:

$$f_X(x) = \sum_{i=1}^{m} f_{Xi}(x)\, p_i$$

dove $f_{Xi}(x) = f_X(x|B_i)$, $i = 1,...m$ è un set di funzioni di probabilità e $p_i = P[B_i]$ è la probabilità o peso relativo associato alla $f_{Xi}(x)$ dell'*i*-esimo componente. Una funzione di distribuzione complessa può quindi essere scomposta in una combinazione di funzioni più semplici, le cui funzioni di densità possono essere invertite analiticamente.

### 6.4.4.2.2 Metodo dello scarto

Il metodo dello scarto permette di generare valori da una variabile X con pdf nota e calcolabile $f_X(x)$. Non è necessario che sia calcolabile anche la cdf. Si disegna la funzione $f_X(x)$ su un piano cartesiano. In seguito si disegna un'altra curva $y = g(x)$ che sottende un'area finita e tale per cui $g(x) \geq f_X(x)$ per tutti i possibili valori di *X*. Questa funzione è chiamata funzione di comparazione e giace ovunque sopra $f_X(x)$. Il metodo dello scarto standard prevede la generazione di un valore *x* della variabile aleatoria *X* di densità $g(x)/\alpha$, dove α è l'area sottesa da $g(x)$. In seguito si estrae un valore *u* da una variabile uniforme standard *U*. Quindi si verifica la condizione

$u < f_X(x)/g(x)$

se è soddisfatta *x* viene accettato, altrimenti viene scartato.

Riassumendo, viene estratto casualmente un valore di *x* appartenente a $g(x)$, se

questo è incluso nell'area sottesa dalla $f_X(x)$ allora si accetta, altrimenti si rifiuta e si ripete l'esperimento.

Questo metodo viene utilizzato nel caso sia difficile estrarre casualmente dalla funzione $f_X(x)$, quindi si sceglie una funzione $g(x)$, che non abbia uno scarto eccessivo rispetto a $f_X(x)$, dalla quale sia più semplice estrarre.

### 6.4.4.3. Numeri casuali estratti da variabili discrete

Il metodo della trasformazione dell'integrale di probabilità può essere utilizzato anche per generare numeri casuali estratti da distribuzioni di variabili discrete a partire dall'estrazione di numeri casuali u estratti da una distribuzione uniforme standard. La relazione $F_X(x_{(i-1)}) < u \leq F_X(x_i)$ fornisce il corrispondente valore casuale $x_i$.

## 6.4.5. Integrazione Monte Carlo

Il metodo di integrazione Monte Carlo può essere utilizzato nei casi in cui si debba affrontare la valutazione di integrali troppo complessi da calcolare esplicitamente, peraltro non infrequenti nel campo della ricerca scientifica. Riprendendo il caso dell'inferenza bayesiana, si è visto come questa sia basata sulla stima di valori della distribuzione a posteriori, come per esempio la media o la valutazione delle distribuzioni a posteriori marginali dei parametri. Quindi si prevede un'integrazione della densità di probabilità a posteriori, che può prendere forme anche molto complesse. Per esempio, si può pensare di dover valutare il valore atteso di una funzione di un parametro $\theta$, $\varphi(\theta)$, date le osservazioni $x$:

$$E_\pi[\varphi(\theta)] = \int \varphi(\theta)\pi(\theta/x)\,d\theta$$

La tecnica di simulazione Monte Carlo può fornirci una stima di questo integrale.

Come già descritto, il metodo genera un campione di osservazioni estratte dalla distribuzione della variabile di interesse ed in seguito la media può essere calcolata semplicemente come media campionaria. Per esempio, dato un campione di osservazioni $\theta_1 ... \theta_n \sim \pi(\theta / x)$, possiamo stimare l'integrale di cui sopra con la media campionaria:

$$\hat{\varphi}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\theta_i)$$

### 6.4.6 Accuratezza del metodo e numerosità dei campioni

Nell'impostare una simulazione numerica con il metodo Monte Carlo è necessario innanzi tutto determinare la numerosità del campione da generare tale da avere un'accuratezza soddisfacente. In particolare, se si hanno *n* punti distribuiti in maniera casuale e uniforme si ottiene un errore che decresce con $n^{-(\frac{1}{2})}$, perché ogni nuovo punto si aggiungerà alla somma dei quadrati che sarà la varianza e l'errore deriva dalla radice quadrata della varianza.

Consideriamo l'esempio iniziale dell'ago di Buffon. Per avere una stima accurata della probabilità *p* è possibile a priori determinare quante coppie *XY* è necessario generare per avere una certa accuratezza: se *N* è il numero di successi e *n* la numerosità del campione, *p* viene stimata dal rapporto *N*/*n*. Se gli eventi sono indipendenti allora *N* è una binomiale di parametri *n* e *p*, e l'errore standard della stima del rapporto *N*/*n* è:

$$\sigma_{\hat{p}} = \sqrt{p\frac{(1-p)}{n}}$$

Si può dimostrare che per $np > 5$ e $n(1-p) \geq 5$ la distribuzione approssima una normale di media *np* e varianza *np(1-p)*. Sostituendo a *p* la stima di *p* vengono determinati il limiti dell'intervallo bilatero di confidenza al 100(1-α)% del vero valore di *p*:

$$\hat{p} \pm \Phi^{-1}_{(1-\alpha/2)}\sqrt{\hat{p}\frac{(1-\hat{p})}{n}}$$

indicando con $\Phi$ la funzione di distribuzione cumulata gaussiana. La numerosità *n* del campione necessaria ad assicurare che questi limiti siano minori o uguali al $100\varepsilon\%$ vero valore di *p* (con ε compreso tra 0 e 1) deve soddisfare la relazione:

$$\Phi^{-1}_{(1-\alpha/2)}\sqrt{p\frac{(1-p)}{n}} \leq \varepsilon p$$

quindi

$$n \geq \left(\Phi^{-1}_{(1-\alpha/2)}\right)^2 \frac{(1-p)}{p\varepsilon^2}$$

### 6.4.7. Controllo della varianza

#### 6.4.7.1. Variabili antitetiche

Nonostante si sia visto come l'accuratezza del metodo Monte Carlo dipenda dalla

numerosità del campione *n*, la varianza dei risultati della simulazione può essere ridotta senza incrementare *n*. Questo risultato può essere ottenuto con tecniche di riduzione della varianza che sfruttano le proprietà di campioni correlati.

Per esempio, si considerino due stimatori corretti della variabile *X*, $X_1$ *e* $X_2$. Questi due stimatori possono esse combinati per ottenere un nuovo stimatore $X' = \frac{(X_1 + X_2)}{2}$ .

La media di *X'* è ancora $\mu_x$, infatti:

$$E[X'] = E[\frac{(X_1 + X_2)}{2}] = \frac{(E[X_1] + E[X_2])}{2} = \frac{(\mu_X + \mu_X)}{2} = \mu_X$$

Ne consegue che *X'* è uno stimatore corretto di $\mu_x$.

La varianza di *X'* è:

$$Var[X'] = Var[\frac{(X_1 + X_2)}{2}] = \frac{(Var[X_1] + Var[X_2] + 2\mathrm{Cov}[X_1, X_2])}{4}$$

La varianza di *X'* è minore di $(Var[X_1]+Var[X_2])/4$ (caso in cui gli stimatori siano indipendenti) se i due stimatori sono correlati negativamente.

Il metodo delle variabili antitetiche (Hammersley and Morton, 1956) è una procedura di simulazione che assicura la correlazione negativa tra $X_1$ *e* $X_2$. Questo risultato è ottenuto generando una serie di variabili uniformi indipendenti $u_1, u_2, ...u_n$ per ottenere un campione di numerosità *n* per lo stimatore $X_1$. Per ottenere un altro campione di numerosità *n* per lo stimatore $X_2$ si usa la sequenza $1-u_1, 1-u_2, ...1-u_n$. In questo modo $X_1$ e $X_2$ sono automaticamente correlati negativamente.

### 6.4.7.2. Variabile di controllo

L'accuratezza del metodo Monte Carlo può essere incrementata in alcuni casi introducendo uno stimatore indiretto *Y'* di *X* a partire dallo stimatore *X'*. Questo si ottiene introducendo la variabile di controllo *Z* come segue:

$$Y' = X' - \eta(Z - \mu_z)$$

dove $\eta$ è un coefficiente e $\mu_z$ è la media di *Z*, essendo *Z* una variabile aleatoria correlata a *X*. Se *X'* è uno stimatore corretto di *X*, allora:

$$E[X'] = X$$

$$E[Y'] = E[X'] - \eta(E[Z] - \mu_z) = E[X'] = X$$

Quindi anche *Y'* è uno stimatore corretto di *X*. La varianza di *Y'* è calcolata con la

seguente:

$$Var[Y'] = Var[X'] + \eta^2 Var[Z] - 2\eta \mathrm{Cov}[X', Z] \quad (2)$$

Se $2\eta \mathrm{Cov}[X', Z] > \eta^2 Var[Z]$ allora *Var[Y']* < *Var[X']* e *Y'* sarebbe uno stimatore più accurato di *X'*.

Si può scegliere $\eta$ in modo tale che la varianza di *Y'* sia minima, cioè:

$$\frac{(d\, Var[Y'])}{d\eta} = 2\eta \mathrm{Var}[Z] - 2\mathrm{Cov}[X', Z] = ¿$$

Da cui:

$$\eta = \frac{(Cov[X', Z])}{(Var[Z])} \quad (3)$$

Sostituendo la (3) nella (2) si ottiene il valore della minima varianza di *Y'*.

### 6.4.8. Possibili applicazioni

Il metodo di simulazione Monte Carlo, ed in generale la tecnica della simulazione numerica, viene usato nel caso in cui una soluzione analitica del problema in esame sia difficilmente perseguibile o non lo sia affatto.

Consideriamo per esempio il caso in cui una variabile di progetto non sia direttamente osservabile, per esempio il carico di nutrienti un lago (fluoro e potassio, responsabili del fenomeno dell'eutrofizzazione) è dato dalla somma degli apporti umani delle acque di rifiuto e da quelli naturali. Per determinare la frazione di nutrienti derivata da attività antropiche è necessario esplicitare una certa relazione tra la variabile osservata (la concentrazione del lago) e la variabile in esame (la frazione di concentrazione causata dall'apporto umano). Non sempre questo tipo di relazioni sono deterministiche e prive di incertezza. Quindi sono stocastiche. La simulazione numerica diventa necessaria nel momento in cui metodi analitici non siano perseguibili per la complessità del sistema, a meno di fare semplificazioni che rischiano di risultare eccessive.

Un altro caso molto comune riguarda la ricerca del valore di un determinato quantile della variabile di progetto, per esempio nella previsione dei periodi di ritorno di una piena di una certa entità. Quindi si necessitano, teoricamente, dei campionamenti diretti per ottenere delle statistiche da studiare. Spesso, però, non sono disponibili serie di dati o si dispone di campioni troppo piccoli per poter fare studi affidabili. In questi casi il metodo Monte Carlo può essere implementato per ottenere questo tipo di statistiche, basandosi sui modelli probabilistici usati, sul metodo usato per stimare i parametri e dalla

numerosità del campione. Inoltre, la simulazione numerica può essere utilizzata per valutare la sensitività dei modelli probabilistici ad eventuali errori sistematici o casuali.

Un altro esempio di applicazione riguarda i sistemi varianti nel tempo o nello spazio. Per esempio nello studio dell'andamento demografico di una popolazione si può assumere che le nascite e le morti siano sequenza di eventi che occorrono a intervalli di tempo regolari (modelli annuali o stagionali) oppure che siano punti casuali sull'asse del tempo. Quindi si fissa un determinato periodo di tempo $\Delta t$ ed il modello esamina lo stato del sistema al termine di ciascun intervallo. Così facendo però alcuni eventi possono non essere osservati, perché si sviluppano e si esauriscono all'interno dell'intervallo di tempo e quindi, misurando lo stato del sistema all'inizio e alla fine di $\Delta t$, non si osservano perturbazioni. Una simulazione di situazioni di questo genere con metodo Monte Carlo considera le serie di eventi osservabili ed il tempo di occorrenza tra un evento e l'altro è modellato come una variabile casuale.

Infine, le simulazioni numeriche via metodo Monte Carlo sono largamente diffuse nel designare alternative di progetto ed ottimizzarne il dimensionamento. Il motivo del suo largo impiego si deve alla semplicità e versatilità matematica, che rende possibile stimare risposte sia di tipo fisico sia di tipo economico. Per raggiungere la soluzione ottima di progetto, si comincia con una situazione di prova e si ottiene la risposta del sistema con differenti simulazioni variando la configurazione del sistema fino a raggiungere quella ottima, magari supportati da un apposito software di simulazione.

## *6.6. Catene di Markov*

Una catena di Markov è un processo stocastico in cui, noto lo stato attuale, passato e futuro sono indipendenti. In altre parole, è una sequenza stocastica di numeri dove ogni valore dipende solamente dall'ultimo. Questa è nota come proprietà di Markov:

$$P[\theta_{(k+1)} \in A / \theta_k = x, \theta_{(k-1)} \in A_{(n-1)} \ldots \theta_0 \in A_0] = P[\theta_{(k+1)} \in A / \theta_k = x]$$

Dove $\theta$ sono gli elementi della catena, $x$ indica un valore noto ed $A$ è lo spazio dei parametri.

Questa probabilità dipende quindi da $x$, $A$ ed $n$. Nel caso non dipenda da $n$ si dice che la catena è omogenea e si può definire il nucleo di transizione, che descrive univocamente la dinamica della catena. Se volessimo costruire una sequenza $\theta_0$, $\theta_1$, $\theta_2$,...ecc, con una catena di Markov, dove $\theta_0$ è un valore arbitrario scelto dalla distribuzione iniziale, generiamo il nuovo elemento della catena $\theta_{k+1}$ estraendolo da una densità

dipendente solo da $\theta_k$:

$$\theta_{(k+1)} \approx P(\theta_k, \theta) = P(\theta / \theta_k)$$

*P* è il nucleo di transizione, che gode delle seguenti proprietà:

1) *P(x,.)* è una distribuzione di probabilità, per ogni *x* appartenente allo spazio delle osservazioni;

2) la funzione $x \rightarrow P(x,A)$ può essere valutata, per ogni *A* contenuto nello spazio delle osservazioni.

Sotto certe condizioni, cioè che la catena sia aperiodica ed irriducibile, la distribuzione degli stati della catena di Markov convergerà ad una distribuzione stazionaria. Dobbiamo sempre assumere che queste condizioni siano soddisfatte.

Forniamo ora alcune definizioni utili:

Una distribuzione $\pi$ è stazionaria per una catena di Markov con probabilità di transizione *P(x,y)* se

$$\sum_{x \in S} \pi(x) P(x, y) = \pi(y)$$

dove *S* è lo spazio delle osservazioni.

Un insieme *C* si dice irriducibile se vi è una probabilità non nulla che a qualsiasi step temporale un qualsiasi stato x appartenente a *C* abbia probabilità non nulla di raggiungere un altro stato y sempre appartenente a *C*. Una catena è irriducibile se l'insieme *spazio delle osservazioni* è irriducibile.

Il periodo di uno stato di una catena di Markov con un numero di stati finito o infinito numerabile è definito come il minimo numero di step temporali affinché vi sia una probabilità diversa da zero di tornare sullo stesso stato, partendo dallo stato $\theta_k$ al tempo *k*. Il periodo *T* è matematicamente definito come il massimo comun divisore dell'insieme:

$$T = [n \geqslant 1 : P(\theta_{(k+n)} = \theta_k / \theta_k) > 0]$$

Lo stato $\theta_k$ è detto aperiodico se il suo periodo è uguale a 1. Una catena di Markov è detta aperiodica se tutti i suoi stati sono aperiodici, altrimenti è detta periodica.

## *6.7. Markov Chain Monte Carlo*

Il metodo MCMC implementa l'integrazione Monte Carlo usando una catena di Markov per generare osservazioni dalla generica distribuzione $\pi$. L'aggiornamento avviene in modo tale che la distribuzione di probabilità associata al k-esimo elemento si avvicini

sempre di più alla distribuzione target, $\pi ( \theta / x )$, al crescere di k. Si dice che la catena converge a π. Una volta avvenuta la convergenza a π, possiamo estrarre i valori della catena e ritenere che essi appartengano alla distribuzione di interesse, per poi utilizzarli in qualsiasi tipo di stima empirica (Monte Carlo) di quantità come la media. Questo significa che possiamo considerare solo le realizzazioni della catena che avvengono dopo la convergenza, e scartare i valori iniziali. Questo periodo iniziale della catena viene chiamato *burn-in*. La strategia di campionamento MCMC consente la costruzione di catene di Markov aperiodiche e irriducibili, per le quali la distribuzione stazionaria sia esattamente la distribuzione target π. Vi sono diversi algoritmi che ci permettono di generare una catena di Markov, di cui i principali sono noti come *Metropolis Hasings* e *Gibbs Sampling*.

### 6.7.1. Metropolis Hastings

Consiste in una forma generalizzata di campionamento col metodo dello scarto, come quello visto in precedenza nella generazione di osservazioni di una variabile casuale per il metodo Monte Carlo.

I valori della catena vengono estratti da distribuzioni che approssimano la distribuzione obiettivo, π, e sono "corretti" in maniera tale che si comportino asintoticamente come estrazioni casuali di π. L'aspetto saliente di questo algoritmo è proprio il fatto che ad ogni step di simulazione vengono migliorate le approssimazioni della distribuzione target.

Viene scelto un nucleo di transizione $\boldsymbol{K ( \theta , \varphi )}$ tale che π sia la distribuzione stazionaria della catena. In particolare sarà:

$$\pi(\theta)K(\theta,\varphi)=\pi(\varphi)K(\varphi,\theta)$$

Per ogni coppia $( \theta , \varphi )$ la catena è reversibile, quindi π è la distribuzione stazionaria. La distribuzione generatrice del valore al generico step $k+1$, $\theta_{k+1}$ è scelta arbitrariamente e dipende solo dallo stato attuale $\theta_k$ e si indica con $q ( \varphi \mid \theta_k)$. Generalmente, così costruita, la catena non soddisferà le condizioni di reversibilità necessarie ad assicurare la convergenza. Si introduce quindi una funzione di accettazione, $\alpha ( \theta_k , \varphi )$. L'osservazione φ viene accettata con probabilità $\alpha ( \theta_k , \varphi )$, in tal caso si pone $\theta_{k+1} = \varphi$. In caso contrario si rimane sul valore attuale ponendo $\theta_{k+1} = \theta_k$. La forma ottima della funzione di accettazione è fornita da Peskun (1973), in modo tale che i valori generati siano scartati meno frequentemente possibile e l'efficienza computazionale sia ottimizzata:

$$\alpha(\theta_k,\varphi)=min(1,\frac{\pi(\varphi/x)q(\theta_k/\varphi)}{\pi(\theta_k/x)q(\varphi/\theta_k)})$$

Il nucleo di transizione è così definito, essendo *A* un generico sottoinsieme dello spazio dei parametri *Θ*:

$$K(\varphi,A)=\int_A q(\varphi/\theta)\alpha(\theta,\varphi)d\varphi+I_A(\theta)[1-\int_\Theta q(\theta,\varphi)\alpha(\theta,\varphi)d\varphi]$$

La quantità $1-\int_\Theta q(\theta,\varphi)\alpha(\theta,\varphi)d\varphi$ quantifica la probabilità che un valore φ proposto venga scartato. Il termine $I_A(\theta)$ è la funzione identità tale che:

$$I_A(\theta)=1\ se\ \theta\in A$$
$$0\ se\ \theta\notin A$$

Descriviamo ora l'algoritmo che ci permette di passare dallo stato $\theta_k$ allo stato $\theta_{k+1}$:

1) Si inizializza il contatore *i = 1* e si seleziona arbitrariamente un valore iniziale $\theta_0$;

2) Si muove la catena ad un nuovo valore $\varphi \sim q(\theta_k, .)$;

3) Si valuta la probabilità di accettazione dello spostamento da $\theta_k$ a $\varphi$, pari a $\alpha(\theta_k, \varphi)$ e si accetta φ con probabilità α in questo modo:

- si genera un valore indipendente $u \sim U[0,1]$;
- se $u \leq \alpha$ si accetta lo spostamento e si impone $\theta_{k+1} = \varphi$;
- se $u \geq \alpha$ si rifiuta lo spostamento e sarà $\theta_{k+1} = \theta_k$.

4) Si incrementa il contatore *i = i+1* e si ritorna al passo 2.

In generale, q offre spostamenti simmetrici rispetto alle precedenti posizioni, quindi $q(\theta,\varphi) = q(\varphi,\theta)$ e diventa:

$$\alpha(\theta,\varphi)=min(1,\frac{\pi(\varphi)}{\pi(\theta)})$$

Il che può risultare intuitivo: l'accettazione del nuovo stato dipende solo dallo spostamento rispetto allo stato precedente.

Con questo metodo si sostituisce il campionamento da π, che può risultare molto complesso, con tanti campionamenti da q più semplici. Essendo che i valori generati da q non saranno necessariamente accettati, ma vengono "proposti", si fa riferimento a *q* anche come distribuzione *proposta*.

### 6.7.2. Campionamento di Gibbs

Questo metodo prende in considerazione le distribuzioni dei parametri valutate singolarmente e non in maniera congiunta come nel caso precedente (si ricorda che nel metodo bayesiano la distribuzione a posteriori risultante è quella congiunta di tutti i parametri). In particolare, si setta la distribuzione *proposta* di ogni parametro come la distribuzione a posteriori del parametro condizionata al valore noto di tutti gli altri parametri. Questo tipo di distribuzione prende anche il nome di full-conditional e si indica con $\pi\ (\theta_i / \theta^k_{-i})$ o semplicemente $\pi_i\ (\theta_i)$, dove $k$ è il numero dei parametri. In questo caso la probabilità di accettazione è sempre pari a 1. Condizione necessaria è che le $\pi\ (\theta_i / \theta^k_i)$ siano note e campionabili.

Si supponga di avere un vettore di $p$ parametri da stimare $\boldsymbol{\theta} = (\theta_1,..., \theta_p)$, con distribuzione congiunta $\pi\ (\boldsymbol{\theta})$. Il campionamento di Gibbs usa le singole full-conditional di $\pi$ per campionare indirettamente dalle distribuzioni marginali dei parametri.

Vediamo ora come si presenta l'implementazione dell'algoritmo di questo metodo:

1) Si inizializza il contatore $i = 1$ u un set arbitrario di valori iniziali $\boldsymbol{\theta}_0 = (\theta^1_0 ..... \theta^p_0)$;

2) Ottengo $\boldsymbol{\theta}_i$ a partire da $\boldsymbol{\theta}_{i-1}$ con la generazione successiva dei seguenti valori:

$$\theta_1^i \approx \pi\left(\theta_1 / \theta_2^{(i-1)} \ldots \theta_p^{(i-1)}\right)$$
$$\theta_2^i \approx \pi\left(\theta_2 / \theta_1^i, \theta_3^{(i-1)} \ldots \theta_p^{(j-1)}\right)$$
$$\vdots$$
$$\theta_p^i \approx \pi\left(\theta_p / \theta_1^i \ldots \theta_{(p-1)}^i\right)$$

3) si incrementa il contatore $i = i+1$ e si torna al punto 2.

Una volta raggiunta la convergenza, i $\theta_i$ possono essere considerati campionamenti di $\pi\ (\boldsymbol{\theta})$.

### 6.7.3. Convergenza

Vi sono due elementi principali da considerare nel determinare la lunghezza della catena, che sono il numero di iterazioni necessarie a raggiungere la convergenza e la numerosità del campione necessaria per avere errori accettabili nella simulazione Monte

Carlo. Come già accennato in precedenza, è necessario un periodo iniziale della simulazione necessario a raggiungere la convergenza, chiamato burn-in. I valori della catena utili al campionamento sono solo quelli generati dopo questo periodo.

Il metodo più semplice per determinare il periodo di burn-in è quello di visualizzare i valori dei parametri in un grafico dove l'asse delle ascisse rappresenta il numero di iterazioni e quello delle ordinate il valore del parametro. Sempre è possibile osservare come i parametri, partendo dal loro valore iniziale, si stabilizzino intorno ad un valore medio. Questo metodo può essere migliorato facendo diverse simulazioni, inizializzando i parametri ogni volta con valori anche molto diversi, in modo tale da poter verificare la robustezza del metodo. Così facendo, se ad ogni iterazione i parametri si assestano sempre intorno allo stesso valor medio, si può dire che la catena ha converso. Questo tipo di approccio viene formalizzato nell'analisi BGR (*Brooks-Gelman-Rubin*, Brooks and Gelman, 1998).

## *6.8 Analisi e previsione di serie storiche – modelli ARMA*

Con il termine serie storica si intende una successione finita di numeri del tipo

$$x_1,\ldots, x_n$$

che è ottenuta tramite osservazioni di un fenomeno reale durato nel tempo. Si immagina quindi che questi valori siano legati da un certo tipo di relazione e seguano una certa logica, oltre ad elementi di casualità. La serie storica rappresenta il passato ed è l'unica base che abbiamo per poter prevedere come questa serie potrà continuare nel futuro, Modellare un fenomeno con una serie storica rappresenta il punto di partenza di uno studio del fenomeno stesso mediante un modello matematico. Il ruolo dell'analista è quello di utilizzare gli strumenti matematici per analizzare gli eventi passati e riprodurli nel futuro. Per esempio, può non essere conveniente prendere in considerazione l'intera serie storica di dati ma solo una parte più recente, e qui l'analista dev'essere in grado di capire dove sia più conveniente tagliare la serie e considerarne solo una parte $x_k$, .., $x_n$; oppure deve saper individuare trend e periodicità nella serie, per capire quali sono i dati più rappresentativi e più idonei a prevedere i valori futuri. A questo scopo, la prima operazione da fare è quella di rappresentare la serie storica in maniera adeguata e saperla interpretare, soffermandosi su determinati periodi o valori e saperli contestualizzare. Una serie storica può dire molte cose se la si osserva su diverse scale (mensile, giornaliera, annuale, decennale, ecc.). Va sempre tenuto a mente che la mente umana è il processore più potente e l'occhio umano è il

miglior interpolatore. Non sempre, però, la nostra mente è allenata ad effettuare operazioni di un certo tipo e, per, mancanza di esperienza su un determinato problema, potrebbe non tenere in conto alcuni aspetti, che invece non sfuggirebbero ad un'analisi matematica. Per questo la matematica dev'essere vista come uno strumento di supporto ai nostri ragionamenti ed intuizioni, ma è inevitabile e, anzi, auspicabile che vada utilizzata con un certo grado di soggettività. Uno dei pregi della nostra mente è quello di poter considerare un'infinità di variabili in una volta sola, variabili che spesso è difficile tradurre in termini matematici. Si pensi, ad esempio, all'atto di attraversare una strada: la nostra mente deve calcolare in quanto tempo la macchina raggiungerà l'incrocio e quanto tempo ci metteremo ad attraversare la strada, faremo una stima del tempo di frenata, magari considerando le condizioni dell'asfalto, assicurandoci comunque che il conducente ci abbia visto incrociando il nostro sguardo. Tradurre questo in termini matematici è sicuramente più complicato e dispendioso in termini di tempo, mentre nella nostra mente queste operazioni sono avvenute in una frazione di secondo. Il pregio di un modello matematico, però, è quello di esplicitare un valore che noi non sapremmo fornire. Non sapremmo mai dire a che velocità sta andando la macchina o a quanti metri si fermerà da noi. Un modello matematico sì, quindi può fornire informazioni che non abbiamo e che completerebbero la nostra capacità di previsione.

Da tener presente, però, che il valore di previsione fornito da un modello matematico non è mai assoluto, cioè non corrisponde alla verità, ma se le ipotesi di partenza sono ragionevoli si può pensare che il risultato sia vicino a quello reale. Ma quanto vicino? La nostra intuizione spesso ci porta a formulare un ventaglio di previsioni, individuando quelle più o meno probabili; allo stesso modo, un modello previsionale dovrebbe fornire un gamma di soluzioni, individuando gli intervalli più probabili. Questo avviene, ad esempio, dichiarando un intervallo di confidenza.

Vi sono modelli matematici all'interno dei quali sono stati elaborati algoritmi atti ad analizzare le serie storiche, individuando e quantificando trend e periodi e fornendo una proiezione della serie nel futuro. Per esempio i modelli ARIMA. Con questo tipo di algoritmi, si vuole infatti cercare un modello ricorsivo che sia aderente il più possibile ai dati, che ne catturi la struttura e che ne fornisca una previsione.

Si andrà ora ad analizzare le varie parti da cui sono composti questo tipo di modelli

### 6.8.1. Modelli AR

Si chiama modello autoregressivo di ordine *p*, o modello *AR(p)*, l'equazione lineare:

$$X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p} + \varepsilon_t$$

e la sua soluzione si chiama processo *AR(p)*, dove *p* è l'ordine, gli $a_i$ sono parametri (numeri reali), $\varepsilon_t$ è il termine di errore e gli $X_i$ sono le realizzazioni della serie storica all'istante di tempo *i*.

Il metodo più semplice in assoluto per fornire una previsione è quello di ripetere l'ultimo valore:

$$x_{n+1} = x_n$$

In questo caso si ipotizza che la serie sia stazionaria e che il valori passati non influiscano su quello presente. Da notare che questa è una particolare realizzazione di un modello *AR(1)*.

Una versione più elaborata è il metodo della media mobile, che a due passi è formulato come:

$$x_{n+1} = \frac{x_n + x_{n-1}}{2}$$

Il metodo può considerare un qualsivoglia numero di passi, fino a considerare la media complessiva:

$$x_{n+1} = \frac{x_n + x_{n-1} + \ldots + x_1}{n}$$

Questi sono tutti esempi di modelli *AR* in cui l'errore è considerato nullo e i coefficienti $a_i$ uguali per qualsiasi istante temporale. Questo tipo di approccio vede la serie storica come una realizzazione di un campione sperimentale, ignorando la struttura temporale dei dati. Non è detto, comunque, che questo sia un approccio sbagliato; anzi, se i dati hanno un'elevata componente aleatoria e non presentano trend o periodicità può essere l'unica strada ragionevolmente percorribile. Vi sono comunque software che decidono i valori migliori per i coefficienti $a_i$, e che quindi piuttosto che $x_{n+1} = (x_n + x_{n-1})/2$ potrebbero decidere che sia meglio usare $x_{n+1} = 0{,}3\, x_n + 0{,}7\, x_{n-1}$, per meglio adattarsi ai dati.

Un modello *AR* del tipo:

$$AR(1): \quad X_t = a_1 X_{t-1} + \varepsilon_t$$

è adatto a descrivere situazioni a media nulla e non contiene intercetta. Inoltre, se $|a| < 1$ il processo è stazionario.

Se $X_t$ è il processo che ci interessa e $\mu$ è la sua media, il processo $(X_t - \mu)$ è a media

nulla e per esso si può considerare il processo *AR(p)*:

$$(X_t - \mu) = a_1(X_{t-1} - \mu) + \ldots + a_p(X_{t-p} - \mu) + \varepsilon_t$$

Quindi si può scrivere:

$$X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p} + \varepsilon_t + (\mu - a_1 \mu - \ldots - a_p \mu)$$
$$= a_1 X_{t-1} + \ldots + a_p X_{t-p} + \varepsilon_t + b$$

che è un modello *AR(p)* con intercetta

$$b = (\mu - a_1 \mu - \ldots - a_p \mu).$$

Si pensi ora al caso particolare di un *AR(1)* con intercetta e coefficiente $a = 1$:

$$X_n = X_{n-1} + b + \varepsilon_n$$

Si può scrivere anche, iterativamente, che:

$$X_n = X_{n-2} + b + \varepsilon_{n-1} + b + \varepsilon_n = X_{n-2} + 2b + \varepsilon_{n-1} + \varepsilon_n$$

Fino a risalire al valore iniziale:

$$X_n = X_0 + nb + \sum_{i=1}^{n} \epsilon_i$$

Questo esempio mostra che $X_n$ ha un trend lineare di coefficiente *b*.

Un altro esempio che può aiutare a capire il comportamento dei modelli *AR* è il seguente:

Si consideri il modello *AR(1)* :

$$X_n = a\, X_{n-1} + b + \varepsilon_n$$

con $|a| < 1$.

Risulta essere:

$$X_n = a\,(a\, X_{n-2} + b + \varepsilon_{n-1}) + b + \varepsilon_n = a^2\, X_{n-2} + (a+1)\, b + a\, \varepsilon_{n-1} + \varepsilon_n$$

Infine:

$$X_n = a^n X_0 + (a^{n-1} + \ldots + a + 1)\, b + a^{n-1}\, \varepsilon_1 + \ldots + a\, \varepsilon_{n-1} + \varepsilon_n$$

Essendo $|a| < 1$, vale che $(a^{n-1} + \ldots + a + 1) \rightarrow 1 / (1-a)$, quindi non vi sono trend ma $(a^{n-1} + \ldots + a + 1)$ b tende ad un valore costante pari a $b /(1-a)$. Il processo $X_n$ ha però una media non nulla.

Nel caso sia $|a| > 1$, i modelli *AR* hanno comportamento esponenziale.

### 6.8.2. Operatore L

Sia *S* l'insieme di tutte le successioni $x = (x_t)$ di numeri reali, con t appartenente a *Z*.

L'operatore di traslazione temporale è l'applicazione $L : S \rightarrow S$ definita da:

$$L_{xt} = x_{t-1}$$

Data una successione $x$, $L$ calcola una nuova successione, il cui valore al tempo $t$, $L_{xt}$, è dato da $x_{t-1}$. L'operatore $L$ prende quindi una sequenza e la trasla all'indietro. Le potenze positive o negative di $L$ verranno indicate con $L^k$. Esse sono la composizione di $L$ fatta $k$ volte, cioè:

$$L^k_{xt} = x_{t-k}$$

Il modello *AR(p)* può essere quindi scritto nella forma:

$$\left(1-\sum_{k=1}^{p} a_k L^k\right) X_t = \epsilon_t$$

Questo può essere utile ad evidenziare la relazione tra $X_t$ ed $\varepsilon$, dove quest'ultimo non sia un errore ma piuttosto un input del modello come può essere un controllo, ad esempio.

### 6.8.3. Modelli MA

Un modello a media mobile di ordine $q$, o modello *MA(q)*, è definito dall'equazione lineare:

$$X_t = \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q}$$

Dove i $\beta_i$ sono coefficienti numerici appartenenti all'insieme dei numeri reali. A differenza del caso *AR(p)*, qui il processo è definito esplicitamente dal rumore, attraverso una sua media pesata.

### 6.8.4. Modelli ARMA

Si chiama modello *ARMA(p,q)* (Auto-Regressive Moving Average di ordini $p$ e $q$) l'equazione lineare:

$$X_t = \alpha_1 X_t + \ldots + \alpha_p X_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q}$$

Alternativamente, utilizzando la forma implicita con l'operatore $L$, sarebbe:

$$\left(1-\sum_{k=1}^{p} \alpha_k L^k\right) X_t = \left(1+\sum_{k=1}^{q} \beta_k L^k\right) \epsilon_t$$

Il modello ora descritto si adatta a situazioni a media nulla. Volendo esaminare processi a media non nulla μ, introduciamo $Z_t = X_t - \mu$ nel modello *ARMA (p,q)*:

$$Z_t = \alpha_1 Z_{t-1} + ... + \alpha_p Z_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + ... + \beta_q \epsilon_{t-q}$$

Esplicitando $X_t$ diventa:

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + b + \epsilon_t + \beta_1 \epsilon_{t-1} + ... + \beta_q \epsilon_{t-q}$$

dove *b*, l'intercetta, è data da:

$$b = \mu - \alpha_1 \mu - ... - \alpha_p \mu \quad .$$

## *6.9. Strategie di campionamento*

Si intende ora fare cenno alle principali strategie di campionamento di dati sperimentali. In particolare ci si concentrerà sul campionamento di matrici ambientali, quasi sempre al fine di stimare concentrazioni di sostanze inquinanti. La scelta dello schema di campionamento è funzionale ad ottenere risultati statisticamente e scientificamente significativi. Inoltre, per valutare la qualità dei dati raccolti, non solo è fondamentale un buono schema di campionamento, ma è inoltre necessario essere a conoscenza degli errori associati alle misure, all'accuratezza delle analisi di laboratorio e alla rappresentatività dei dati in relazione all'obiettivo finale dell'analisi.

Una campagna di campionamento non può prescindere dall'indicare luogo e data del campionamento, e deve consentire lo svolgimento di diverse operazioni, come il supporto alle decisioni in situazioni di superamento delle soglie di rischio, poter confrontare medie e varianze di due o più popolazioni, localizzare punti critici e fonti di inquinamento, determinare la natura e l'estensione di un certo fenomeno di contaminazione, individuare trend e periodicità. Vi sono diverse strategie di campionamento per le quali è stata studiata una determinata procedura. Tra di esse si vedranno il campionamento ragionato, casuale semplice, stratificato, sistematico, pesato, cluster adattato e composito.

- Il *campionamento ragionato* si basa su conoscenze pregresse sulla conformazione e sulla storia del sito in esame. Viene applicato su piccola scala e per campioni poco numerosi. L'obiettivo è spesso quello di individuare aree contaminate. Ha il vantaggio di essere rapido ed economico ma resta comunque un'indagine puntuale ed è impossibile conoscere l'incertezza associata al metodo. In sostanza consiste nel campionare zone in maniera mirata, stimando poi le concentrazione col supporto di modelli fisici della matrice ambientale e delle sue interazioni con l'inquinante.
- La procedura del *campionamento casuale semplice* non è standardizzata. Si

basa sulla selezione casuale dei punti di campionamento, spesso col supporto di generatori di numeri casuali che possano determinare le coordinate dei punti di campionamento o gli istanti di tempo in cui campionare. Questa strategia viene spesso utilizzata congiuntamente ad altre. Per come viene impostato, si è sicuri che il campione ottenuto sia rappresentativo della zona in esame, se il numero di osservazioni è adeguato. Si fa presente, però, che vi è il rischio che i campioni possano non essere distribuiti in maniera uniforme nello spazio e nel tempo, a causa della scelta casuale; inoltre, viene ignorata ogni tipo di informazione disponibile a priori sul sito in esame, ed è per questo che la tecnica viene affiancata da altre metodologie che tengano in conto anche questi ultimi aspetti.

- Nel *campionamento stratificato* le informazioni disponibili sulla popolazione vengono utilizzate per identificare gruppi considerati omogenei sotto determinati aspetti per poi campionare singolarmente all'interno dei vari gruppi. Ad esempio se il terreno è formato da strati con diverse percentuali di argilla, l'acqua avrà comportamenti diversi nei singoli strati e quindi ne risentiranno anche le concentrazioni degli inquinanti. Inoltre, gli strati possono non essere solo geografici ma anche temporali, come nel caso di periodicità di determinati fenomeni inquinanti. Ulteriori considerazioni statistiche, che variano da caso a caso, potranno far luce sulle metodologie di campionamento da adottare in ogni singolo strato. Il metodo è efficace, però, solo se si dispone di una buona conoscenza pregressa del sito in analisi.

- Nel *campionamento sistematico* prevede una copertura uniforme (a maglia) del sito in esame. La spaziatura regolare (spaziale e/o temporale) permette l'individuazione di correlazioni e tendenze. É utile nel caso di ricerca di hot spot e per stima di parametri di una popolazione, anche grazie alla buona rappresentatività dei campioni. La procedura non richiede conoscenze pregresse del sito; per ottimizzare la campagna di campionamento possono essere usate congiuntamente anche altre tecniche che tengano in conto le conoscenze a priori disponibili sul luogo analizzato.

- Una tecnica che combina il campionamento casuale semplice con le conoscenze iniziali del sito è il *campionamento pesato*. Questo metodo permette di ridurre il numero di campioni necessari e l'individuazione di caratteristiche o strutture che indichino un aumento delle concentrazioni dell'inquinante ricercato (come macchie, odori), anche grazie a misure preliminari (come fluorescenza a raggi x, foto aeree...). Quindi la tecnica prevede un campionamento mirato nelle zone in cui si suppone vi sia la presenza dell'inquinante.

- Il *campionamento cluster adattato* si esegue a valle di un campionamento iniziale al fine di indirizzare ed ampliare le misurazioni fatte in un primo momento. Nella prima fase, quindi, si esegue un campionamento casuale; in quelle successive i campionamenti vengono mirati in zone di interesse individuate nella prima. É una procedura di tipo iterativo che può richiedere diverso tempo e l'impiego di diverse risorse economiche, anche se comunque queste ultime vengono concentrate nelle aree di maggior interesse.
- Il *campionamento cluster composito*, infine, prevede l'omogeneizzazione di vari campioni per ottenerne uno nuovo (composito). Il metodo prevede una riduzione significativa del numero di analisi da effettuare, a costo di una considerevole perdita di informazione. Si ottiene una buona stima della media della variabile in esame, ma bisogna valutare se sia il metodo adatto da applicare alla particolare situazione. Il vantaggio è che possono essere considerate aree molto più grandi a parità di costi. Inoltre non è detto che l'omogeneizzazione dei campioni sia completa e questo è fonte di incertezza.

Risulta evidente che ciascun metodo sia connotato da vantaggi e svantaggi. Una sua scelta ragionata può comunque sempre portare ad ottenere ottimi risultati. Nell'impostazione della strategia di campionamento vanno considerate tutte le possibili metodologie, prevedendo i possibili sviluppi e conseguenze delle scelte, senza sottovalutare le disponibilità economiche e di tempo, oltre agli obiettivi dell'analisi.

## *6.10. Selezione dei modelli*

Si prenderanno ora in considerazione metodi di selezione di modelli generati sia con approccio classico (in cui i parametri hanno un valore fisso) sia con approccio bayesiano (in cui i parametri sono visti come variabili aleatorie, alla stregua dei dati).

Come già visto, le osservazioni su cui si basano i modelli possono essere raccolte sul campo, in laboratorio, a seguito di determinati esperimenti, oppure generate tramite simulazioni numeriche. In ogni caso, di frequente si ha a che fare con un set di modelli che sotto diverse ipotesi concorrono a descrivere il fenomeno in esame. Può risultare quindi di particolare interesse ordinare i modelli ottenuti secondo determinati criteri di bontà, per poi poter selezionare il modello migliore da utilizzare. Come si vedrà in seguito, vi sono metodi idonei per effettuare una discriminazione tra i modelli di un set. Di questi sarà

possibile scegliere quello che meglio soddisfa determinate condizioni rispetto agli altri, ma ciò non significa che sia quello che in generale descrive meglio il fenomeno. Per questo è necessario che tutti i modelli candidati siano scientificamente fondati e che le ipotesi con cui sono stati generati siano ragionevoli e documentate.

### 6.10.1. SRM – Structural Risk Minimisation

SRM è un criterio di selezione di modelli basato sulla teoria SLT (*Statistical learning theory*). Questa è basata sull'indice *VC-dimension*, che da informazioni sulla complessità di funzioni parametriche, lineari e non (Vapnik, 1999).

L'approccio al problema di selezione del miglior modello è il seguente: si dispone di una variabile casuale input *x*, estratta da una fissata distribuzione incognita. L'output *y* è restituito dal sistema secondo una probabilità condizionata a x, incognita anch'essa. Chiameremo *p(x,y)* la probabilità congiunta delle variabili input e output. L'obiettivo è scegliere la funzione *f(x,θ)*, chiamando con $\theta$ i parametri della funzione, tale che minimizzi la media della perdita:

$$R(\theta)=\int_x \int_y (y-f(x,\theta))^2 p(x,y)\,dxdy$$

Questo valore viene indicato con la lettera R in quanto può essere interpretato come il rischio di ottenere output diversi da quelli campionati stimandoli con la funzione *f*, a parità di input.

Nell'espressione del rischio, *p(x,y)* è ignota e l'unica informazione di cui si dispone è contenuta nei *q* campioni di input/output. La sola cosa che è possibile calcolare è l'errore di adattamento durante la calibrazione, chiamato rischio empirico ($R_{emp}$):

$$R(\theta)_{emp}=\frac{1}{q}\sum_{i=1}^{q}(y_i-f_j(x_i,\theta))^2$$

Il limite superiore del rischio che si vuole calcolare (*R( θ))* è una funzione del rischio empirico. In particolare, secondo Cherkassky et al. (1999), la probabilità che:

$$R(\theta)\leqslant R(\theta)_{emp}\left(1-\sqrt{p-p\ln(p)+\frac{\ln(q)}{2q}}\right)$$

è pari a $\left(1-\frac{1}{\sqrt{q}}\right)$ , dove $p = d/q$ e d è il numero di parametri della funzione. In questo caso viene calcolato il limite superiore del rischio, cioè quello garantito, non quello medio.

### 6.10.2. AIC – Akaike's Information Criterion

Il metodo di Akaike per la selezione dei modelli si basa sull'informazione di Kullback-Leibler, detta *K-L information*. L'informazione di *K-L* può essere vista come una quantificazione del concetto di statistica sufficiente introdotto da Fisher, cioè una funzione del campione di osservazioni che sia in grado di descrivere in maniera sintetica l'informazione contenuta nel campione stesso.

Se si indica con f la funzione della realtà (priva di parametri), con *g* si indica il modello di approssimazione della realtà e con *I* l'informazione persa nell'usare *g* per approssimare *f* (la *K-L information*). L'informazione *I* è definita come:

$$I(f,g)=\int f(x)\ln\frac{f(x)}{g(x/\theta)}\,dx$$

Si dovrebbe quindi minimizzare *I* in funzione di *g*. $\boldsymbol{\theta}$ è il vettore dei parametri del modello *g*.

La funzione della realtà *f* è considerata costante ed unica. Per valutare la bontà di un modello l'informazione *K-L* non può essere direttamente utilizzata, in quanto richiederebbe la conoscenza della realtà *f*. Il metodo Akaike fornisce quindi una graduatoria dei modelli $g_i$ presi in considerazione per descrivere la realtà, considerando migliore quello che minimizza l'informazione *K-L*. Quest'ultima può essere scritta come:

$$I(f,g)=\int f(x)\ln(f(x))\,dx-\int f(x)\ln(g(x/\theta))\,dx=E_f(\ln f(x))-E_f(\ln g(x/\theta))$$

Essendo *f(x)* una costante:

$$E_f(\ln f(x))=C$$

Allora:

$$I(f,g)=C-E_f(\ln(g(x/\theta)))$$

Quindi per ogni modello del set di modelli considerato è necessario stimare solo $E_f(\ln(g(x/\theta)))$ .

Akaike mostra come il problema da risolvere per avere un criterio rigoroso di selezione basato sull'informazione K-L sia la stima di $E_x[E_y(\ln(g(x/\hat{\theta}(y))))]$ , dove *θ(y)* è uno stimatore MLE di *θ*.

*X* indica le risposte predette dal modello, dato un un campione di dati *y*. *X* e *y* sono variabili casuali dipendenti appartenenti ad una certa distribuzione di probabilità.

Akaike mostra come il massimo valore log-verosimile di *g* (che indicheremo come *L(θ/dati)*) sia uno stimatore distorto di $E_xE_y(\ln(g(x/\hat{\theta}(y))))$ , e come la loro differenza sia approssimativamente uguale al numero di parametri stimati dal modello. Questo

numero viene indicato come *K*.

Quindi uno stimatore corretto di $E_x E_y(\ln(g(x/\hat{\theta}(y))))$ è

*ln L(θ/dati)-K*.

Conseguentemente si può scrivere che $\ln(L(\hat{\theta}/dati))-K=C-\hat{E}_{\hat{\theta}}(I(f,\hat{g}))$ , con $\hat{g}=g(./\hat{\theta})$

Akaike trova quindi uno stimatore della media dell'informazione *K-L* basato sulla massimizzazione della funzione di log-verosimiglianza corretta con *K*. Moltiplicando per -2 si trova il criterio di informazione di Akaike:

$$AIC=-2\ln(L(\hat{\theta}/dati))+2\mathrm{K}$$

Nel caso particolare in cui sia stata usata una stima con il metodo dei minimi quadrati e distribuzione normale degli errori, AIC può essere scritto come:

$$AIC=n\ln(\sigma^2)+2\mathrm{K}$$

dove $\sigma^2=\frac{\sum \varepsilon_i^2}{n}$ , indicando con $\varepsilon_i$ gli errori stimati dal modello e con n il numero dei dati.

Assumendo un set di modelli selezionato a priori, ben supportato da considerazioni scientifiche, l'indice AIC viene calcolato per ciascun modello. Infine, viene fatta semplicemente una graduatoria dei modelli con indice AIC crescente. Il modello migliore nel set sarà quello con indice AIC minore.

Da notare come, all'aumentare della numerosità n del campione, lo stimatore MLE di $\theta$ converga al valore $\theta_0$ che minimizza l'informazione *K-L* per un dato modello *g*. La media di AIC per grandi campioni converge a

$$E(AIC)=-2\mathrm{C}+2\mathrm{I}(f,g(./\theta_0))+K \quad .$$

Nel caso in cui il numero di parametri stimati da un modello sia grande rispetto alla numerosità dei dati, quindi per piccoli campionamenti, si usa l'indice AIC corretto (correzione di secondo ordine):

$$AIC_c=AIC+\frac{2\mathrm{K}(K+1)}{n-K-1}$$

e si usa per *n/K < 40*.

### 6.10.3. BIC – Bayesian Information Criterion

Shwarz definisce Il criterio di informazione Bayesiano come:

$$BIC=-2\ln(L)+K\ln(n)$$

Dove $L$ è la funzione massima log-verosimile del modello, $K$ il numero di parametri stimati dal modello e $n$ la numerosità del campione considerato. Come nel caso AIC, l'indice BIC viene calcolato per ciascun modello e viene fatta una graduatoria per valori di BIC crescenti, individuando come miglior modello quello caratterizzato da BIC minore.

Si definisce $\Delta BIC_i$ la differenza tra l'indice BIC di un modello e il BIC minimo del set di modelli. In questo modo si individua il modello migliore come quello caratterizzato da $\Delta BIC = 0$.

Un uso più rigoroso del criterio di informazione bayesiano prevede il calcolo della probabilità a posteriori di ciascun modello, $p_i$, definita come:

$$p_i = Pr(g_i / dati) = \frac{\exp(\frac{-1}{2}\Delta BIC_i)}{\sum \exp(\frac{-1}{2}\Delta BIC_i)}$$

La probabilità $p_i$ indica la probabilità di un modello, rispetto agli altri modelli presenti nel set, di essere quello che minimizza l'informazione $K$-$L$, rispetto al quale un aumento del numero di parametri stimati non fornirebbe alcun miglioramento aggiuntivo nel descrivere la realtà.

Matematicamente, in un set di modelli, ve ne sarà uno, che chiameremo $g_t$, tale per cui $p_t \rightarrow 1$

per $n \rightarrow \infty$ e le altre $p_i \rightarrow 0$. Questo non significa che il modello $g_t$ descriva pienamente la realtà. Infatti $g_t$ sarà il modello più parsimonioso, in termini di parametri da stimare, che all'interno del set minimizza l'informazione $K$-$L$.

### 6.10.4. DIC – Deviance Information Criterion

Nei metodi sopra descritti i parametri vengono visti come quantità fissi e stimati con il metodo della massima verosimiglianza. Possono essere utilizzati anche nell'ambito dell'analisi bayesiana calcolando i valori attesi dei parametri rispetto alla loro distribuzione a posteriori. Un differente criterio di informazione, DIC (Spiegelhalter et al. 2002), è stato sviluppato apposta per confrontare modelli generati con approccio bayesiano, tenendo cioè in conto che i parametri sono variabili aleatorie descritte da una funzione di probabilità, invece che da un valore fisso. Questo criterio è simile ai precedenti (AIC e BIC), nel senso che considera sia la complessità del modello sia il buon adattamento alle osservazioni. Il

buon adattamento del modello *m* è valutato in termini di devianza, mentre la complessità del modello è definita in termini di *numero effettivo di parametri*, indicato con $p_D(m)$ ed espresso nella forma:

$$p_D(m) = -2\,E_\pi(\log f_m(x/\theta)) + 2\log f_m(x/\hat{\theta})$$

Dove la media è calcolata con rispetto alla distribuzione a posteriori e $\theta$ è la stima a posteriori di $\theta$. Una scelta tipica è quella di considerare $\hat{\theta} = E_\pi(\theta/x)$ . Quindi il numero effettivo di parametri è la differenza tra la media a posteriori della devianza e la devianza valutata sulla media a posteriori dei parametri.

L'indice DIC del modello m è espressa nella forma:

$$DIC_m = -2\,E_\pi(\log f_m(x/\theta)) + p_D(m)$$

### 6.10.5. Probabilità del modello a posteriori

Questo metodo di discriminazione viene usato sempre nell'ambito dell'analisi bayesiana. Le probabilità del modello a posteriori sono ottenute semplicemente tramite un'estensione del teorema di Bayes tenendo in conto l'incertezza del modello. Ciò è possibile considerando il modello stesso come un parametro discreto da stimare. Il set di possibili valori che può assumere il parametro è il set stesso di modelli considerati nella comparazione. Applicando il teorema di Bayes possiamo valutare la distribuzione a posteriori congiunta del modello e dei parametri. Formalmente sarebbe:

$$\pi(\theta_m, m/x) \approx f_m(x/\theta_m)\,p(\theta_m/m)\,p(m)$$

dove $\theta_m$ denota il set di parametri del modello *m*, $f_m(x|\theta_m)$ è la funzione di verosimiglianza del modello *m*, $p(\theta_m|\, m)$ è la distribuzione a priori dei parametri nel modello *m* e $p(m)$ è la probabilità a priori di *m*.

Dato che stiamo trattando il modello come un parametro da stimare, sarà necessario specificare la funzione di probabilità a priori associata a ciascun modello. Tipicamente uno stesso parametro è presente in più modelli del set, come per esempio un coefficiente di regressione lineare. In generale, si specificano differenti distribuzioni a priori per i parametri all'interno del set di modelli. Se invece si intende intraprendere la via della selezione dei modelli tramite il metodo delle probabilità a posteriori, è usanza comune specificare le stesse distribuzioni a posteriori per gli stessi parametri dei modelli.

La distribuzione marginale a posteriori è l'elemento che permette di discriminare i vari modelli del set. Si consideri un set di modelli $\boldsymbol{m} = (m_1,....,m_k)$. La probabilità marginale

a posteriori corrispondente al modello $m_i$ è:

$$\pi(m_i/x) = \frac{[f(x/m_i)\,p(m_i)]}{\sum_{i=1}^{K}(f(x/m_i)\,p(m_i))}$$

dove

$$f(x/m_i) = \int f_{(m_i)}(x/\theta_{(m_i)})\,p(\theta_m/m_i)\,d_{(m_i)}$$

dove $f_{(mi)}(x|\theta_{mi})$ indica la funzione di verosimiglianza dei dati *x*, dato il modello $m_i$, corrispondente ai parametri $\theta_m$. Le proprietà marginali a posteriori dei modelli così ottenuti è l'elemento discriminante tra i vari modelli.

### *6.10.6. Esempio: modelli di dinamica demografica strutturati per età e sesso dello stambecco delle Alpi Capra ibex ibex (Mignatti et al., 2012)*

In questo paragrafo si vuole dare un esempio dell'utilizzo dei criteri di selezione dei modelli visti fin'ora, applicati ad un caso reale. Lo studio in questione riguarda lo stambecco delle Alpi (*Capra ibex ibex*), presente nel parco nazionale del Gran Paradiso, di cui si dispone di una serie particolarmente lunga di dati (disponibili dal 1956), catalogati inoltre per sesso e per età. In questa sede ci si focalizzerà sui risultati dei metodi di selezione dei modelli e non su considerazioni teoriche riguardanti la loro formulazione, per le quali si rimanda all'articolo "*Sex- and age-structured models for Alpine ibex Capra ibex ibexpopulation dynamics*" (A. Mignatti et al., 2012).

I dati raccolti fin'ora mostrano un andamento medio pressoché costante fino agli anni '80. Dal 1980 in poi la popolazione è cresciuta fino a raggiungere un picco negli anni '90, dopodiché il numero di individui è diminuito costantemente fino ad oggi. Questo andamento unimodale è in contrasto con il comportamento della profondità del manto nevoso, che nella seconda metà degli anni '70 ha visto un costante trend negativo. Quest'ultima osservazione ha portato gli autori dello studio a chiedersi quali altri fattori, oltre a quello climatico, potessero incidere sulla dinamica della popolazione.

In particolare, si è voluto indagare l'influenza del sesso e dell'età sull'andamento demografico di questi ungulati, tenendo in considerazione studi precedenti che ne evidenziavano la dipendenza dalla copertura nivale.

Uno dei modelli più affidabili fin'ora è quello messo a punto da Jacobson et al. (2004), che hanno indagato diverse relazioni possibili tra andamento demografico, densità totale della popolazione e altezza del manto nevoso. L'indagine ha portato gli autori a

formulare dei così detti “modelli soglia”, che, a differenza dei “modelli continui”, prevedono differenti valori dei parametri della popolazione per anni con elevata o scarsa copertura nivale (valutata appunto attraverso una data soglia). Dopo aver applicato le tecniche di selezione dei modelli, gli autori sono stati in grado di formulare due modelli ugualmente affidabili, che descrivevano bene il trend positivo degli anni '80 ed il picco degli anni '90, ma che non erano in accordo con i dati più recenti della popolazione, la cui numerosità veniva largamente sovrastimata.

Nello studio di Mignatti et al. si è deciso di contrastare il modello soglia di Jacobson et al. con un modello continuo, consistente in un polinomio di secondo grado:

$$\log\left(\frac{N_{t+1}}{N_t}\right)=\beta_0+\beta_1 N_t+\beta_2 S_t+\beta_3 S_t N_t+\beta_4 S_t^2+\beta_5 N_t^2+\rho_t$$

dove $N_t$ rappresenta la densità totale della popolazione, $S_t$ è la profondità media del manto nevoso, $\rho_t$ è un fattore stocastico che rappresenta il rumore ambientale e i processi non modellizzati mentre i $\beta_i$ sono coefficienti numerici. Nello studio di Jacobson, $\beta_4$ e $\beta_5$ erano posti uguali a zero.

Si noti che in questo modello non si fanno ancora distinzioni di sesso ed età. A questo proposito, diversi studi etologici hanno messo in mostra come lo stambecco delle Alpi viva in gruppi separati, a seconda del sesso e dell'età, che occupano diversi tipi di habitat. In particolare, le femmine e i maschi si incontrano solo nel periodo dell'accoppiamento tra novembre e gennaio, i cuccioli stanno con le madri nel primo anno di età, dopodiché i maschi cominciano a separarsi per raggiungere i gruppi di maschi adulti. Inoltre, maschi, femmine e cuccioli hanno diversi ratei di sopravvivenza, che dipendono in modi differenti dalla densità e da fattori ambientali.

Alla luce di queste osservazioni, si è deciso di formulare modelli strutturati per sesso ed età. La popolazione è stata quindi divisa in classi:: cuccioli $K_t$, individui nel secondo-terzo anno di età (*yearlings*) $Y_t$, adulti maschi $M_t$ e adulti femmine $F_t$. Le prime due classi includono sia maschi sia femmine. In un generico anno $t$, il rateo di variazione demografica dei cuccioli ($\sigma_{K,t}$), dei maschi ($\sigma_{M,t}$) e delle femmine ($\sigma_{F,t}$) e il successo dello svezzamento $w_t$ sono stati definiti come segue:

$$\sigma_{K,t}=\frac{Y_{t+1}}{K_t} \quad ,$$

$$\sigma_{M,t}=\frac{M_{t+1}}{M_t+Y_t/2} \quad ,$$

$$\sigma_{F,t} = \frac{F_{t+1}}{F_t + Y_t/2} \quad ,$$

e

$$w_t = \frac{K_{t+1}}{F_{t+1}} \quad .$$

Sono stati poi formulati dei modelli strutturati non lineari, che descrivessero le relazioni tra i logaritmi dei ratei appena calcolati e due variabili: la profondità del manto nevoso *S* e la densità $D_C$, dove con *C* si indica il generico comparto (*K*, *F* o *M*):

$$\log(\sigma_{C,t}) = \beta_{0,C} + \beta_{1,C} D_{C,t} + \beta_{2,C} S_t + \beta_{3,C} S_t D_{C,t} + \beta_{4,C} S_t^2 + \beta_{5,C} D_{C,t}^2 + \rho_{C,t}$$

e

$$\log(w_t) = \beta_{0,w} + \beta_{1,w} D_{w,t} + \beta_{2,w} S_t + \beta_{3,w} S_t D_{w,t} + \beta_{4,w} S_t^2 + \beta_{5,w} D_{w,t}^2 + \rho_{w,t}$$

Questi modelli possono essere poi modificati rendendoli discontinui e facendo variare i parametri seconda se il valore di $S_t$ sia sopra o sotto una data soglia $\underline{S}$. Inoltre può anche darsi il caso in cui si vogliano considerare i parametri $\beta_4 = \beta_5 = 0$.

Per la selezione dei modelli vera e propria, sono state definite due famiglie di modelli: strutturati e non strutturati. Considerando varie combinazioni, in cui uno o più termini $\beta_i$ vengono posti uguali a zero e considerando i modelli come “continui” o “soglia”, sono stati individuati 32 possibili modelli non strutturati e 72 strutturati. Per orientarsi in un così elevato numero di candidati, sono stati applicati tre criteri di selezione: AIC, BIC e SRM. Dato che i criteri di selezione non individuano un modello ottimo ma un set gerarchico di modelli ottimi, sono state individuate delle soglie da rispettare per i tre metodi, in modo da accettare solo i modelli che rispettassero le condizioni in tutti e tre i criteri di selezione. In particolare, sarebbero stati selezionati i modelli che avrebbero ottenuto un $\Delta AIC < 4$, $\Delta BIC < 2$ e $SRM < 1{,}06 SRM_{best}$, dove $\Delta$ indica la differenza rispetto al modello migliore e $SRM_{bes}$ il punteggio SRM ottenuto dal modello migliore. Per i modelli selezionati a valle della procedura sono stati poi calcolati i valori di $R^2$ e si è proceduto con la validazione: in particolare, sono stati calibrati con i dati dei primi vent'anni (1960 – 1980) e poi validati utilizzando i dati di copertura nivale dal 1981 fino ad oggi.

I modelli che hanno rispettato i criteri di selezione sono dieci, di cui due non strutturati. Questi ultimi sono entrambi modelli di tipo soglia, con dipendenza da densità e da copertura nivale, e di primo grado, il che indica che l'aggiunta di termini di secondo grado non ha portato ad un miglioramento dei risultati. Tutti i modelli strutturati selezionati sono caratterizzati da segregazione spaziale, cioè la dinamica dei maschi è influenzata solo

dalla densità dei maschi stessi, piuttosto che dell'intera popolazione, e lo stesso vale per le femmine. Inoltre, è emerso come gli effetti soglia siano più incidenti sugli adulti che sui cuccioli. Si è potuto inoltre rilevare come annate molto nevose siano dannose alla riproduzione alla stregua di quelle in cui la profondità del manto nevoso è molto scarsa. Infine, è stato constatato che nelle simulazioni a lungo termine, sia i modelli strutturati sia quelli non strutturati tendono a sovrastimare la numerosità della popolazione: questa osservazione spinge ad indagare più a fondo sull'argomento, introducendo ulteriori fattori influenti sulla dinamica demografica, come l'incidenza di malattie o competizioni interspecifiche.

Si può notare come la tecnica di selezione dei modelli sia un elemento chiave dello studio, in quanto ha portato a scartare o accettare numerose ipotesi sulla dinamica della popolazione in esame e formulare modelli affidabili, considerando, tra quelli più parsimoniosi, quelli che meglio descrivessero i dati osservati. Non solo, ma dalla coniugazione fra considerazione teoriche e matematiche sono emersi nuovi elementi che potranno migliorare ancora di più la comprensione l'andamento demografico dello stambecco delle Alpi, come per esempio l'introduzione dell'incidenza di malattie o parassiti.

## *6.11 Filtro di Kalman*

Il filtro di Kalman fornisce la stima dello stato $x \in \mathbf{R}^n$ di un sistema lineare (discreto o continuo). Nel caso discreto, indicando con $k$ l'istante di tempo considerato, il sistema è descritto dalle equazioni:

$$x_k = Ax_{(k-1)} + Bu_{(k-1)} + w_k \quad (1)$$

$$y_k = Cx_k + v_k \qquad (2)$$

### 6.11.1. Definizione del problema

Il problema di base è la stima dello stato $x$ del sistema, avendo a disposizione un modello lineare che lo descrive (prima equazione) e la misura di un sensore che misura direttamente lo stato del sistema o un suo indicatore, detta uscita (seconda equazione). Per esempio se lo stato del sistema in esame fosse il valore di portata di un fiume e avessimo a disposizione un misuratore di portata su una determinata sezione, $x$ rappresenterebbe il

valore stimato tramite il modello di deflusso e *y* sarebbe la misura del sensore.

Entrambe le relazioni sono affette da errore: $w_k$ è il vettore aleatorio delle incertezze del processo, detto anche rumore di processo, mentre $v_k$ rappresenta il vettore aleatorio degli errori di misura. Entrambi sono assunti come indipendenti, bianchi e con distribuzione di probabilità normale, di media zero e varianza nota:

$$p(w) \sim N(0,Q)$$

$$p(v) \sim N(0,R)$$

I due vettori contengono tutte le incertezze del sistema. Le rispettive matrici di covarianza possono essere determinate con un modello statistico del processo e del sensore.

Lo stato *x* e la misura *y* sono quindi vettori gaussiani perché combinazione lineare di vettori gaussiani.

La matrice quadrata A lega lo stato al tempo *k* con lo stato al tempo *k-1* in assenza di input *u* e di rumore di processo. La matrice B lega l'input *u* con lo stato *x* (non è detto che ci sia sempre un input, ossia che B può essere una matrice nulla). La matrice C lega lo stato alla misurazione *y,* cioè si assume che il valore di misurazione si possa prevedere conoscendo la stima dello stato del sistema e l'errore commesso nel misurarlo.

### 6.11.2. L'idea di Kalman

L'idea di Kalman è quella di sfruttare la conoscenza del modello e del metodo di misura con i relativi errori, trovando un modo di combinare le due informazioni ed avere così una stima più accurata dello stato del sistema. Questo, come si vedrà, avviene in maniera ricorsiva: per stimare lo stato attuale si utilizza l'informazione della stima precedente, una stima attuale di primo tentativo (cioè il valore dello stato fornito dalla prima equazione) e la misura attuale con la quale poi verrà aggiornata la stima attuale. Questo avviene per mezzo di un algoritmo ricorsivo formato da due “blocchi” comunicanti. Prima di fornire l'algoritmo, però, si danno alcune definizioni.

Si indica con $x^-_k \in \mathbf{R}^n$ la stima dello stato eseguita a priori, cioè effettuata al passo *k* data la conoscenza del processo all'istante *k-1,* e $x^+_k$ è la stima dello stato eseguita a posteriori, cioè effettuata al passo *k* una volta noto il valore della misurazione $y_k$.

Gli errori di stima a priori e posteriori possono essere così definiti:

$$e^-_k = x_k - x^-_k \quad (3)$$

$$e^+_k = x_k - x^+_k \quad (4)$$

Le rispettive matrici di covarianza sono date da:

$$P^{-}_{k}=E[e^{-}_{k}e^{-}_{k}{}^{T}] \qquad (5)$$

$$P^{+}_{k}=E[e^{+}_{k}e^{+}_{k}{}^{T}] \qquad (6)$$

Si definisce inoltre la così detta innovazione di misurazione o residuo come la differenza: $(y_k - C_k x^{-}_{k})$. Essa riflette la discrepanza tra il valore della misurazione *y* e la previsione di misurazione $C_k x^{-}_{k}$.

Per derivare le equazioni del filtro di Kalman si cerca un'equazione che effettui una stima a posteriori $x^{+}_{k}$ come combinazione lineare della stima a priori e del residuo pesato:

$$x^{+}_{k} = x^{-}_{k} + K_k(y_k - C_k x^{-}_{k}) \qquad (7)$$

La matrice K è detta guadagno di Kalman o blending factor e si determina in modo da minimizzare la covarianza dell'errore a posteriori (6). La minimizzazione viene solitamente effettuata sostituendo la (7) nella definizione dell'errore (4), introducendo poi la nuova espressione dell'errore nella (6), derivando rispetto a K, eguagliando a zero ed infine risolvendo rispetto a K.

Una delle possibili forme di guadagno è data dall'espressione:

$$K_k=P^{-}_{k}C^{T}_{k}(C_k P^{-}_{k}C^{T}_{k} + R_k)^{-1}$$

Il termine tra parentesi è l'espressione della matrice di covarianza dell'innovazione. Si può notare come al tendere a zero della matrice di covarianza dell'errore di misurazione $R_k$, il guadagno di Kalman pesa il residuo in modo sempre più rilevante. Cioè più è precisa la misurazione più il suo valore viene preso in considerazione nel fornire una stima dello stato attuale. In particolare:

$$\lim_{(R_k \to 0)} K_k = C_k^{-1}$$

Mentre al tendere a zero della covarianza dell'errore di stima a priori $P^{-}_{k}$, K pesa il residuo in modo via via meno rilevante:

$$\lim_{(P_K^{-} \to 0)} K_k = 0$$

In altre parole, nella stima dello stato come combinazione del residuo e della stima a priori viene data più importanza al valore meno affetto da incertezza. Così, quando la covarianza dell'errore di misura *R* tende a zero, la misurazione si avvicina al valore dello stato reale, mentre succede il contrario per la previsione della misurazione. In modo analogo, al tendere a zero della covarianza dell'errore di stima a priori data dalla (5), la previsione di misurazione si avvicina sempre più al valore reale mentre la misurazione *y* va via via discostandosi da tale valore.

Il guadagno di Kalman può anche essere espresso nella forma:

$$K_k = P_k C_k R^{-1}$$

con

$$P_k = P^-_k (I + C^T_k R^{-1} C_k P^-_k)^{-1}$$

Se $C_k = I$, allora:

$$K_k = P_k R^{-1}$$

Quindi il guadagno è direttamente proporzionale alla covarianza dell'errore ed inversamente proporzionale alla covarianza dell'errore di misurazione, cioè che all'aumentare dell'errore commesso nella stima precedente aumenta l'affinamento della stima attuale (il guadagno). Inoltre le stime ottenute sono sempre più precise all'aumentare della precisione delle misurazioni.

### 6.11.3. L'algoritmo del filtro di Kalman discreto

Le equazioni del filtro di Kalman si dividono in due gruppi: da una parte vi sono quelle di predizione, che forniscono il valore della stima a priori dello stato e la sua covarianza, dall'altra vi sono le equazioni di aggiornamento, che attraverso il calcolo dell'innovazione, della sua covarianza e del guadagno, forniscono una stima a posteriori accurata dello stato attuale combinando, come visto, le informazioni della stima a priori e della misurazione. Si può pensare quindi al primo gruppo come le equazioni di previsione, mentre il secondo gruppo è formato dalle equazioni di correzione della previsione.

Le specifiche equazioni dell'algoritmo sono le seguenti.

Predizione:

$$x^-_k = A_{k-1} x^+_{k-1} + B_{k-1} u_{k-1}$$

$$P^-_k = A_{k-1} P^-_{k-1} A^T_{k-1} + Q_{k-1}$$

Aggiornamento:

$$K_k = P^-_k C^T_k (C_k P^-_k C^T_k + R_k)^{-1}$$

$$x^+_k = x^-_k + K_k (y_k - C_k x^-_k)$$

$$P_k = (I - K_k C_k) P^-_k$$

al concludersi dell'esecuzione dei due gruppi di equazioni, la stima a priori viene aggiornata con quella posteriori ed usata nella stima a priori del passo successivo.

### 6.11.4. Conclusioni

Il filtro di Kalman condiziona quindi in maniera ricorsiva la stima dello stato corrente del sistema a tutte quelle passate, nel caso in cui le incertezze siano modellabili come gaussiane. Ovviamente le matrici Q ed R devono essere determinate prima dell'implementazione del ciclo. Per scegliere i valori dei parametri si possono avere delle basi razionali oppure no. In ogni caso possono essere regolati in modo da ottimizzare la perfomance del filtro, operazione generalmente eseguita fuori linea con l'aiuto di altri filtri di Kalman. Inoltre l'algoritmo va inizializzato con il valore dello stato all'istante iniziale, rappresentato da un vettore gaussiano. Anche in questo caso, se non si hanno basi razionali per fornire valori dello stato iniziale si utilizza un valore casuale. Più vicino al valore reale sarà il valore dello stato iniziale, più velocemente si otterrà la convergenza del filtro allo stato del sistema.

Uno dei vantaggi del filtro di Kalman è quello di fornire stime di stati passati, presenti e futuri conseguite senza conoscere con precisione la natura del sistema modellato.

# 7. TRATTAMENTO DELLE INCERTEZZE

*“Non v'è causa d'errore più frequente che la ricerca della verità assoluta”*

Samuel Butler

## *7.1. Monte Carlo ibrido probabilistico-possibilistico*

Questo tipo di algoritmo viene spesso usato nella determinazione della propagazione dell'incertezza nella modellizzazione di processi naturali. Uno degli ambiti in cui sta prendendo piede la stima dell'incertezza tramite metodo Monte Carlo è la stima di emissioni di gas inquinanti e climalteranti. Essendo questi studi volti al rispetto di limiti legislativi o di vincoli dettati da accordi internazionali, una stima dell'incertezza associata agli inventari delle emissioni diventa necessaria (Galante et al., 2013).

Un altro ambito dell'ingegneria ambientale in cui viene applicato questo metodo è la stima del rischio, come ad esempio la determinazione del rischio di esondazione di un corso fluviale al fine di dimensionare eventuali argini (Baraldi et al., 2011).

Ancora, vi sono esempi di valutazione di adeguatezza e affidabilità di sistemi integrati di generazione di energia da fonti rinnovabili tramite modelli la cui incertezza viene stimata sempre col metodo Monte Carlo Ibrido (Li & Zio, 2011).

La necessità di questo tipo di algoritmo deriva dal fatto che non tutte le variabili di un modello possono essere descritte da variabili aleatorie; è il caso in cui, per esempio, non si abbiano dati sufficienti o conoscenze nulle sul comportamento di queste. Le variabili che non vengono descritte da funzioni di distribuzione di probabilità verranno chiamate epistemiche, mentre alle altre si darà il nome di probabilistiche. Come fare allora ad inserire le variabili epistemiche nel modello?

Le variabili epistemiche possono essere descritte da funzioni di possibilità, indicate con $\pi(.)$, che associano ad ogni valore della variabile, o ad un suo intervallo di valori, un certo grado di possibilità, nell'intervallo di valori *[0 , 1]*. Ovvero, se un certo valore è ritenuto possibile gli viene assegnato il valore 1, se impossibile il valore 0, altrimenti qualsiasi valore contenuto nell'intervallo a seconda di vari gradi di possibilità di accadimento. Come si vedrà in seguito, vi sono comunque varie tecniche per determinare

la funzione π(.).

Il termine “possibilità” assume un significato più debole rispetto a “probabilità”. Dire che un certo evento è possibile, anche nel linguaggio corrente, non equivale a dire che sia probabile: per esempio il fatto che, durante un temporale, un fulmine colpisca in pieno la propria bicicletta parcheggiata in strada in mezzo a tante altre è sicuramente possibile, ma poco probabile. Quindi nella funzione di possibilità “*dove cade il prossimo fulmine”* all'evento “*cade sulla mia bicicletta*” verrà assegnato il valore 1, mentre magari, potendone calcolare la probabilità, questa sarebbe di gran lunga inferiore all'unità.

Un sistema può essere quindi descritto da variabili probabilistiche (nel caso si abbiano informazioni sufficienti sulle pdf) o possibilistiche. Un sistema che contiene entrambi i tipi di variabili viene detto *ibrido* e l'applicazione del metodo Monte Carlo a sistemi ibridi viene detto Monte Carlo Ibrido o HMC (*Hybrid Monte Carlo*) probabilistico – possibilistico.

### 7.1.1. Modellizzare l'incertezza probabilistica

Come si è visto in precedenza, un sistema probabilistico è descritto da una variabile output, funzione di variabili input aleatorie, descritte da una certa funzione di probabilità: $Y = f(X_1, X_2, ..., X_n)$, dove le $X_i$ sono le variabili aleatorie del modello descritte da una funzione di probabilità $p_{xi}(x)$. La distribuzione della variabile *Y*, output del nostro modello, può essere trovata analiticamente in casi semplici o tramite sistema di campionamento Monte Carlo (MCS – *Monte Carlo Sampling*). Il metodo *MCS* prevede un certo numero *m* di iterazioni in cui vengono campionate le variabili aleatorie del modello, generando ogni volta un vettore di valori: per la *i*-esima iterazione questo sarà della forma $(x_1^i, x_2^i, ...x_n^i)$. Ad ogni iterazione viene calcolato il valore di *Y*. Al termine delle *m* iterazioni verrà effettuata una stima empirica della pdf di *Y*.

### 7.1.2. Teoria della possibilità

Secondo la teoria della possibilità, una variabile *X* viene descritta dalla sua funzione di possibilità $\pi_X(x)$, con *x* appartenete all'insieme dei valori di *X*. Come accennato in precedenza, se un certo evento $X = x_i$ è considerato impossibile, la sua funzione di possibilità assumerà un valore nullo: $\pi_X(x_i) = 0$, se invece l'evento $x_i$ viene considerato

possibile la funzione assumerà il valore *1*: $\pi_X(x_i) = 1$. La funzione di possibilità può assumere tutti gli infiniti valori all'interno dell'intervallo *[0,1]* (Dubois, 2006). Come già accennato, dire che un evento $x_i$ abbia possibilità *1* è un'affermazione più debole rispetto a dire che la sua probabilità sia *1*, che renderebbe $x_i$ un evento certo.

A partire dalla funzione di possibilità, è possibile definire dei limiti di possibilità. La possibilità (o plausibilità) di un evento *A* è definita dalla funzione *Π*:

$$\Pi(A) = \sup_{x \in A}[\pi_X(x)]$$

La misura così detta di necessità di un evento *A* è data dalla funzione *N*:

$$N(A) = 1 - \Pi(not\ A) = 1 - \sup_{x \notin A}[\pi_X(x)]$$

La funzione *Π* soddisfa la seguente relazione:

$$\forall\, A, B \subseteq R; \Pi(A \cup B) = max[\Pi(A); \Pi(B)]$$

La funzione *N*, invece, soddisfa la seguente:

$$\forall\, A, B \subseteq R; N(A \cap B) = min[N(A); N(B)]$$

Indicando con *R* l'insieme dei numeri reali.

La possibilità può essere legata alla probabilità (Baudrit et al., 2006). Consideriamo una famiglia di pdf $\underline{P}(\pi)$, tale per cui per ogni insieme *A*, si ha che:

$$N(A) \leq \underline{P}(A) \leq \Pi(A).$$

La funzione di necessità e di possibilità rappresentano quindi il limite inferiore e superiore, rispettivamente, della funzione di probabilità dell'evento *A*, qualsiasi essa sia, appartenente alla famiglia $\underline{P}(\pi)$.

Non si è quindi in grado di definire la funzione di probabilità della variabile, ma solo di definire una famiglia $\underline{P}$ che la contiene.

### 7.1.3. Metodo α-cut

Se si è in presenza di un modello possibilistico, descritto cioè da sole variabili epistemiche $X_i$, l'output *Y* è una funzione multivariata di queste ultime: $Y = f(X_1,....,X_n)$. Se si è a conoscenza delle distribuzioni di possibilità di ciascuna $X_i$, attraverso il metodo α-cut è possibile definire la funzione di possibilità dell'ouput del sistema, *Y*. La funzione α-cut è così definita:

$$F_\alpha = [x \in U | \pi_X(x) \geq \alpha, 0 \leq \alpha \leq 1]$$

$$F_\alpha = [\underline{F}_\alpha, \bar{F}_\alpha]$$

Dove $F_\alpha \, e \, \bar{F}_\alpha$ sono rispettivamente l'estremo inferiore e superiore individuati col metodo.

Con *U* insieme delle realizzazioni di *X*. Ovvero, si sceglie un valore di α compreso tra 0 e 1, ad esempio 0,3. In corrispondenza di questo valore, si "taglia" la funzione $\pi_X(x)$. La funzione α-cut è definita come l'intervallo dei valori di x che soddisfano la condizione di avere possibilità maggiore o uguale ad α, nell'esempio in fig.1 sarebbe [3,5 ; 6] .

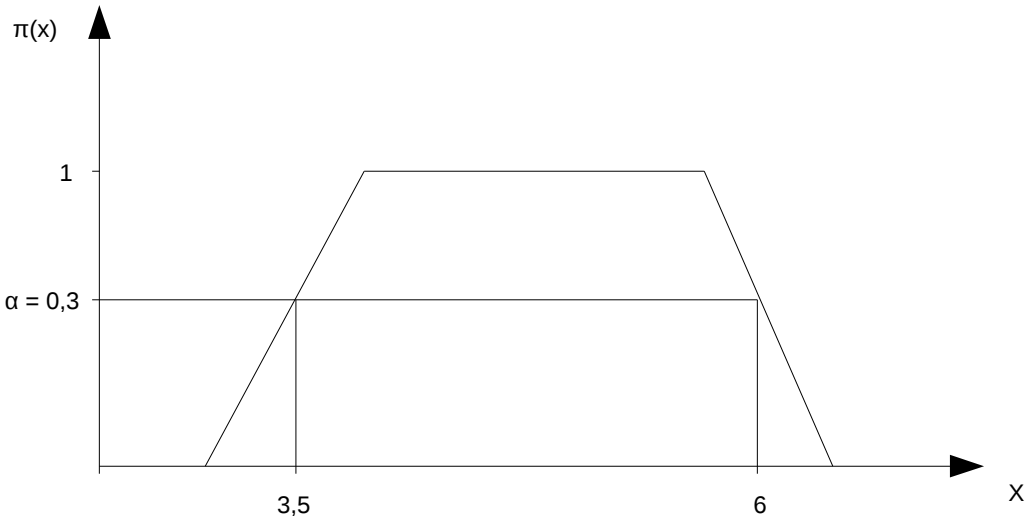


*Fig.1 Esempio di funzione di possibilità e metodo α-cut*

Ora, date le funzioni α-cut di ciascuna variabile possibilistica del modello, possiamo risalire alla distribuzione di possibilità dell'output *Y* facendo variare opportunamente α e calcolandone la funzione α-cut per ogni valore di α. Generalizzando, dato un certo α, la funzione α-cut di Y sarebbe:

$$F_Y^\alpha = [\underline{F_Y^\alpha}, \bar{F}_Y^\alpha]$$

con

$$\underline{F_Y^\alpha} = f(\underline{F_{X1}^\alpha}, ...., \underline{F_{Xn}^\alpha})$$

e

$$\bar{F}_Y^\alpha = f(\bar{F}_{X1}^\alpha, ..., \bar{F}_{Xn}^\alpha)$$

Per ogni α-cut, quindi, si ottengono così un valore massimo ed un valore minimo di Y.

### 7.1.4. Algoritmo di propagazione dell'incertezza possibilistica-probabilistica

Si consideri ora un modello il cui output *Z* sia la funzione $Z = f(Y_{1,\dots}, Y_k, \tilde{Y_{k+1}}, \dots, \tilde{Y_n})$ , in cui le prime *k* variabili, indicate come $Y_i$, sono aleatorie con distribuzione di probabilità $p_{Yi}(y)$, mentre le ultime *n-k* sono possibilistiche, rappresentate dalle rispettive distribuzioni di possibilità: $(\pi(y_{k+1}), \dots, \pi(y_n))$ .

L'algoritmo per la determinazione della propagazione dell'incertezza è una combinazione del metodo Monte Carlo classico applicato alle variabili aleatorie e del metodo α-cut applicato alle variabili epistemiche. Il metodo consiste nell'eseguire due algoritmi annidati e può essere riassunto come segue (Baraldi et al., 2006):

1a. Inizializzare il contatore i = 0;

1b. porre arbitrariamente m = 10000 (ad esempio);

2. porre i = i+1;

3. campionare la i-esima realizzazione del vettore aleatorio ($Y_1^i$,...,$Y_k^i$) con il metodo Monte Carlo tradizionale: ad esempio si può generare un numero casuale compreso tra 0 e 1 ed ottenere il valore della variabile aleatoria attraverso la trasformazione integrale di probabilità; questo per ciascuna variabile del vettore;

4. si pone il valore di possibilità α = 0;

5. si sceglie un Δα (ad esempio Δα = 0,02) ed si applica la funzione α-cut alle variabili possibilistiche, per ciascun valore di α, che sarà α = 0, Δα, 2 Δα,...,1.

6. Per ciascun valore di α e per ciascuna variabile possibilistica si ottengono quindi due valori, infatti $F_Y^\alpha = [\underline{F_Y^\alpha}, \bar{F_Y^\alpha}]$ . Per ogni α si calcola quindi il massimo ed il minimo della funzione output $f(Y_{1,\dots}, Y_k, \tilde{Y_{k+1}}, \dots, \tilde{Y_n})$ , dati rispettivamente dai valori massimi e minimi della funzione α-cut delle variabili possibilistiche.

7. si registrano i due valori di F ottenuti $\underline{F_\alpha}\, e\, \bar{F}_\alpha$ per ciascun α;

8. se $i < m$, si torna al passo 2, se no si chiudono le iterazioni.

Riassumendo, per ciascuna iterazione *i* ottengo un vettore aleatorio campionato col metodo MC e *n* valori massimi e minimi delle variabili possibilistiche, a seconda del valore di $\Delta\alpha$ selezionato. Ad ogni *i*-esima iterazione si calcola quindi il massimo ed il minimo della funzione *f* per ciascun $\alpha$-cut, dato il vettore aleatorio campionato. Eseguendo

*m* iterazioni ottengo quindi *m* funzioni di possibilità di *Z*, definite dai valori massimi e minimi di ciascun α-cut. Per ogni realizzazione $\pi_i(z)$ e per ogni sottoinsieme di *Z*, *A*, possiamo ottenere la misura di possibilità e di necessità. Con $A = (-\infty,z]$, si avrà:

$$\Pi_i^f(A) = \sup_{z \in A}[\pi_i^f(z)]$$

e

$$N_i^f(A) = \inf_{z \notin A}[1 - \pi_i^f(z)] = 1 - \Pi_i^f(\bar{A})$$

Le m realizzazioni di possibilità e necessità si possono combinare per ottenere le funzioni di confidenza (*belief*) *Bel(A)* e plausibilità *Pl(A)*:

$$Bel(A) = \sum_{i=1}^{m} p_i N_i^f(A) = \sum \frac{N_i^f(A)}{m}$$

e

$$Pl(A) = \sum_{i=1}^{m} p_i \Pi_i^f(A) = \sum \frac{\Pi_i^f(A)}{m}$$

Dove $p_i$ indica la probabilità di campionare la i-esima realizzazione del vettore aleatorio $(y^i_1,..., y^i_k)$.

Se $A = (-\infty,z]$, *Bel(A) e Pl(A)* possono considerarsi come il limite rispettivamente inferiore e superiore della distribuzione cumulata di probabilità di z:

$$\bar{F}(z) = Pl(A) \quad \text{e} \quad \underline{F}(z) = Bel(A)$$

Questa tecnica, quindi, prevede il calcolo, per ogni insieme A, della possibilità media associata all'evento A pesata con la probabilità di ogni intervallo di valori.

Ciò che si ottiene sono quindi due funzioni, *Bel(A)* e *Pl(A)*, all'interno delle quali si è sicuri che sia contenuta la funzione cumulata di probabilità dell'output del modello.

### 7.1.5. Propagazione probabilistica

Nel caso si voglia indagare la propagazione dell'incertezza puramente probabilistica, in presenza di variabili possibilistiche, è necessario convertire la distribuzione di possibilità delle variabili epistemiche in funzioni di distribuzione di probabilità. Vi sono diverse tecniche per farlo, la più semplice prevede una semplice normalizzazione nel modo seguente:

$$p_{X_i}(x) = \frac{\pi_{X_i}(x)}{\int_0^{\infty} \pi_{X_i}(x)\,dx}$$

Dopo aver ottenuto le probabilità di ciascuna variabile possibilistica, per

determinare la distribuzione di probabilità della variabile output sarà sufficiente applicare il metodo Monte Carlo probabilistico visto nei capitoli precedenti.

### 7.1.6. Costruire le distribuzioni di possibilità

Si esporranno ora due tra i metodi più diffusi per ottenere una funzione distribuzione di possibilità di una variabile epistemica (Baraldi et al., 2006).

#### 7.1.6.1. Funzione triangolare

Si supponga di conoscere il range di valori entro il quale può variare la variabile incerta in esame, [*a*,*b*]. Questi possono essere, per esempio, i limiti fisici della variabile studiata. Supponiamo inoltre che vi sia un valore preferito della variabile, *c*. Questo può essere per esempio un valore assunto dalla variabile nella maggior parte delle osservazioni (moda). La funzione di possibilità triangolare associata a questa situazione è costruita come un triangolo di base [*a*,*b*] e vertice in *c*. La possibilità risulta quindi nulla agli estremi dell'intervallo [a,b] e per tutti i valori esterni a questo; assume valore uguale a 1 in corrispondenza del valore *c*. Si può dimostrare che una funzione di possibilità triangolare così costruita contiene tutte le famiglie di distribuzioni di probabilità di dominio [*a*,*b*] e moda *c*. Può anche darsi il caso in cui invece di un solo valore preferito *c*, si osservi un range di valori preferiti [*c*,*d*]. In questo caso la funzione di possibilità assumerà forma trapezoidale di base maggiore *ab* e base mino *cd*.

#### 7.1.6.2. Disuguaglianza di Chebyshev

La disuguaglianza di Chebyshev può essere utilizzata nella costruzione della funzione di possibilità di una variabile nel caso si conoscano media μ e varianza $\sigma^2$ di questa. La disuguaglianza di Chebyshev fornisce un'approssimazione “raggruppata” degli intervalli di confidenza di una variabile Y attorno alla sua media μ, nota la sua deviazione standard σ. La disuguaglianza di Chebishev può essere scritta come segue:

$$Pr\left(|Y-\mu|<k\,\sigma\right)\geq 1-\frac{1}{k^2}$$

Questa disuguaglianza permette di definire una funzione di possibilità della variabile che contenga tutte le famiglie di distribuzioni di probabilità di questa variabile

con media μ e deviazione standard σ, a prescindere se la funzione di probabilità sconosciuta sia simmetrica o no, unimodale o no. Si considerino a tal proposito gli intervalli *[μ – kσ, μ + kσ]* come α-cuts della distribuzione di possibilità *π(y)* e siano *π(μ – kσ) = π(μ + k) = 1/k²*. Dalla disuguaglianza di Chebyshev si può affermare che $Pr(Y \in [\mu - k\sigma, \mu + k\sigma]) \geq 1 - \frac{1}{k^2}$ per *k > 1*. Inoltre, per definizione degli α-cuts, si ha che $Pr(Y \in [\mu - k\sigma, \mu + k\sigma]) \geq 1 - \alpha$ . Quindi per costruzione $\alpha = 1/k^2$. Facendo quindi variare k, la funzione di possibilità di *Y* può essere costruita punto per punto.

### 7.1.7. Esempio: stima dell'incertezza sulle emissione di benzo(a)pirene da combustione domestica di legna (Galante et al, 2013)

Le diverse applicazioni del metodo Monte Carlo (in versione probabilistica e ibrida possibilistico- probabilistica) sono state utilizzate per valutare l'incertezza nella stima delle emissioni di benzo(a)pirene (B(a)P) dalla combustione domestica della legna.

Lo studio intende comparare i risultati ottenibili mediante i diversi approcci suggeriti dalla metodologia AEIG.

L'AEIG utilizza due approcci per il calcolo delle emissioni: il metodo *Tier 1*, che definisce un fattore di emissione medio per l'intero settore, ed un metodo più dettagliato (*Tier 2*) che definisce fattori di emissione specifici per i diversi apparecchi di combustione (camino aperto, camino chiuso, stufa a legna, stufa a pellet, ecc.). L'utilizzo del metodo *Tier 2* richiede ovviamente un livello più dettagliato di informazione, cioè la suddivisione dei quantitativi di legna complessivamente consumati per tipologia di apparecchio.

Le equazioni utilizzate sono le seguenti:

*Tier 1*:

$$E = F \cdot A \cdot 10^{-6} \quad (1)$$

Dove E sono le emissioni di B(a)P (kg a-1) , F il fattore di emissione medio di B(a)P per la combustione residenziale della biomassa (mg $GJ^{-1}$) e A la biomassa consumata nel settore domestico (GJ $a^{-1}$) .

*Tier 2*:

$$E = \sum_{i=1}^{n} F_i A_i 10^{-6} \quad (2)$$

Dove, come sopra, E sono le emissioni di B(a)P [kg a-1] , $F_i$ il fattore di emissione di B(a)P per la combustione della biomassa nell'apparecchio i [mg GJ-1] e $A_i$ la biomassa consumata nell'apparecchio i [GJ a-1].

Le variabili A e $A_i$ sono state descritte da una pdf, e quindi rappresentano la parte puramente probabilistica del modello. Per il fattore di emissione invece è noto solamente un valore 'suggerito' ed un intervallo di confidenza: è pertanto possibile considerare $F_i$ come una variabile di tipo possibilistico. I fattori di emissione utilizzati, ed i relativi intervalli di confidenza, sono quelli suggeriti dall'AEIG.

Nell'applicazione del metodo HMC, si è assunta per $F_i$ una distribuzione di possibilità triangolare (Fig.2).

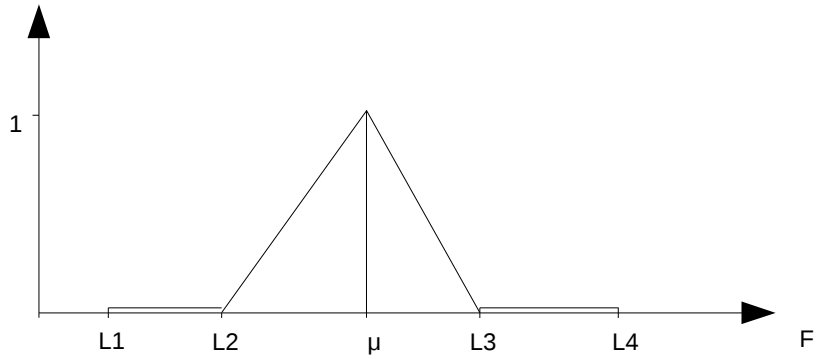


*Fig.2 Funzione di possibilità di F*

L'informazione contenuta nella funzione di possibilità è più coerente con i dati effettivamente disponibili sul fattore di emissione, evitando gli elementi arbitrari introdotti con la scelta della pdf.

Nel caso Tier 1, il modello HMC procede come segue:

1. Nel ciclo esterno si campiona la variabile A utilizzando il PMC;

2. Per ciascun valore di A ottenuto viene eseguito un ciclo interno. Nell'intervallo di valori tra 0 e 1, sono stati generati 41 intervalli della variabile α (α –cuts), e per ciascun valore, in base alla distribuzione di possibilità di F, si calcola il massimo ed il minimo di E mediante l'equazione (1): viene generata quindi una distribuzione di possibilità per ciascun valore di α generato nel ciclo più esterno;

3. Concluse le iterazioni del ciclo più esterno, viene fatta una media delle 10.000 distribuzioni cumulative generate (misure di possibilità e necessità) utilizzando la formula 10.4: si ottengono così le curve di probabilità cumulate massima e minima per E.

Nell'approccio *Tier 2* si segue la medesima struttura, mai il calcolo è modificato

nel modo seguente:

1. Nel ciclo esterno vengono campionate separatamente le 7 variabili *Ai* (una per ciascun tipo di apparecchio);

2. Per ogni insieme di valori *Ai*, nel ciclo interno vengono calcolate le distribuzioni di possibilità di *Ei* (emissioni dell'apparecchio i) mediante la distribuzione di possibilità di *Fi*;

3. Si ottiene la distribuzione di possibilità di E sommando le 7 distribuzioni relative ai singoli settori;

4. Alla fine del ciclo esterno viene seguita la procedura di aggregazione già spiegata ai punti 3 e 4. del caso *Tier 1*.

I due casi *Tier 1* e *Tier 2* portano a risultati simili, con una lieve riduzione nell'ampiezza dell'intervallo di confidenza per il caso Tier 2, che utilizzando informazioni più dettagliate dovrebbe fornire risultati più attendibili. L'approccio possibilistico fornisce come risultato due curve di probabilità cumulate che costituiscono un limite inferiore e superiore, all'interno dei quali è compresa la curva di probabilità della variabile. Considerato lo scarso livello di conoscenza associato al fattore di emissione, l'approccio fornisce una rappresentazione della probabilità più congruente alla realtà. I risultati sono riportati in Fig.3. I risultati del metodo HMC sono stati inoltre confrontati con quelli ottenuti utilizzando il metodo puramente probabilistico, sempre utilizzando sia la metodologia *Tier 1* sia *Tier 2*, e rappresentati in Fig.3 rispettivamente dalla curva verde e dalla curva rossa.

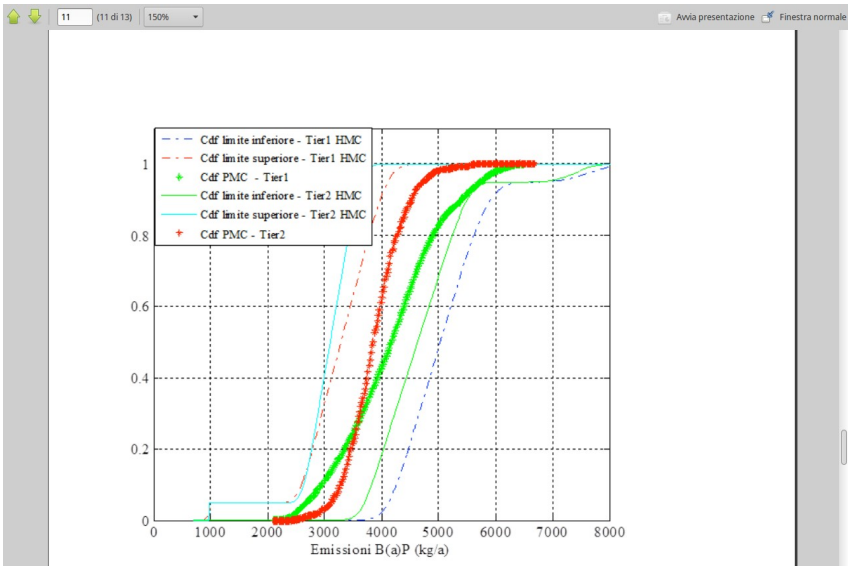


*Fig.3 Risultati dell'analisi sui fattori di emissione*

La scelta dei due parametri L1 ed L4 (valori estremi dell'intervallo all'interno dei quali è contenuta la variabile con il 100% di probabilità)è stata fatta senza una reale base, essendo l'informazione disponibile limitata all'intervallo nel quale la variabile è contenuta con un 95% di probabilità. Si è perciò effettuata un'analisi di sensitività sui due parametri. Si è proceduto a diverse simulazioni variando in ugual modo l'ampiezza dei due intervalli L1-L2 ed L3-L4, che sono stati ridotti del 25%, 50% e 90%. L'analisi ha mostrato come sia trascurabile l'ampiezza dell'intervallo L1-L4.

## 7.2. Incertezza nella misura di portata

Si intende ora presentare un metodo semplice e molto diffuso di valutazione dell'incertezza sulla misura della portata di un determinato corso fluviale, a partire dalla scala delle portate (G.Baldassarre, A.Montanari, 2009).

La scala delle portate è un diagramma che correla in maniera biunivoca il livello idrico di un corso fluviale con la corrispondente portata osservata. La portata di una data sezione viene generalmente determinata col metodo area/velocità ed in seguito applicata la formula di Gaukler-Strikler per il moto assolutamente turbolento:

$$Q = K_S \cdot A \cdot R_H^{\frac{2}{3}} \cdot i_f^{\frac{1}{2}}$$

Dove Q è la portata in esame, $K_s$ il coefficiente di scabrezza di Strikler, A è l'area della sezione, $R_h$ è il raggio idraulico e $i_f$ è la pendenza di fondo.

Nel metodo che si andrà ad esporre non verranno considerati gli errori di misura del livello idrico, considerati trascurabili rispetto ad altri (Shmidt, 2002; Pappenberger et al., 2006). Inoltre la geometria fluviale viene considerata stazionaria, ovvero le uniche variazioni presenti nel corso fluviale sono dovute ai cambi stagionali che affettano la scabrezza di fondo. In altre parole, non vengono considerati i cambiamenti di geometria del fiume nel corso dell'anno. Questo perché detti cambiamenti variano a seconda della situazione specifica analizzata e non è possibile formulare un modello generale. Da tenere in conto, però, che il corso di un fiume può subire rilevanti modifiche anche repentine in caso, per esempio, di trasporto solido ingente.

Alla luce di queste premesse, l'errore da cui è affetta la misura della portata *Q(x,t)* si considera dovuto ai seguenti fattori: errore di misura dovuto al metodo area-velocità $\varepsilon_1(Q(x,t))$, errori relativi alla curva di scala delle portate, $\varepsilon_2(Q(x,t))$, da suddividersi in: errore di interpolazione ed estrapolazione dalla curva di interpolazione dei punti osservati di altezza/portata, $\varepsilon_{2.1}(Q(x,t))$, errore dovuto all'ipotesi di moto permanente anziché moto

vario, $\varepsilon_{2.2}(Q(x,t))$e cambiamenti stagionali della scabrezza di fondo, $\varepsilon_{2.3}(Q(x,t))$.

Per come sono definiti, $\varepsilon_1$ ed $\varepsilon_2$ si possono considerare indipendenti. L'errore totale può quindi essere calcolato come la combinazione lineare dei due:

$$\epsilon(Q(x,t)) = \epsilon_1(Q(x,t)) + \epsilon_2(Q(x,t)) \quad (1)$$

### 7.2.1. Metodo area/velocità

Il metodo area/velocità consiste nello stimare la portata Q integrando il valore di velocità v del flusso sull'area A della sezione in esame: $Q = \int\int_A v(x,y)\,dxdy$ . Essendo oggi impossibile ottenere una misura della velocità in ogni punto di una sezione, questa viene divisa in segmenti verticali su cui poi si misura la velocità media (misurando la velocità in due o più punti a diverse profondità). Questo metodo, ampiamente utilizzato per la sua semplicità e relativa accuratezza, presenta diverse fonti di incertezza: innanzi tutto, per come viene impostato, prevede che nel corso fluviale vi siano condizioni di flusso permanente, mentre nella realtà questo non accade mai; inoltre la variabilità spaziale della velocità in una sezione porta ad errori di stima della velocità media su una sezione verticale; da non sottovalutare, infine, l'errore indotto dalla misura dell'area della sezione considerata, essendo nella maggioranza dei casi affetta da irregolarità geometriche.

### 7.2.2. Incertezza sulla misura con metodo area/velocità

Per quantificare l'errore dovuto alla misura della portata col metodo area/velocità si dovrebbe quantificare ciascuna delle sopra elencate fonti di errore. All'interno delle norme ISO EN 748 viene fatto questo studio, sotto le seguenti assunzioni:

- Il misuratore di flusso opera in condizioni ideali di moto permanente, senza errori sistematici e in assenza di vento;
- gli errori sono indipendenti e distribuiti normalmente;
- il numero di segmenti verticali è almeno pari a 20.

Sotto queste ipotesi, l'incertezza da cui è affetta la misura di portata, con livello di confidenza del 95%, è pari al 5,3%.

Si può concludere che qualsiasi misura di portata effettuata col metodo area/velocità è affetta da circa il 5% di incertezza, con livello di confidenza del 95%.

Essendo tutti i contributi dell'errore gaussiani, $\varepsilon_1(Q(x,t))$ è anch'esso gaussiano con media zero e varianza uguale a $0{,}027Q(x,t)$.

### 7.2.3. Incertezza nella scala delle portate

Secondo la norma ISO EN 748, $\varepsilon_2(Q(x,t))$ è da considerarsi una variabile binaria, senza informazioni disponibili per poterne determinare il segno positivo o negativo. La situazione peggiore si ha ovviamente quanto tutte e tre le componenti di $\varepsilon_2$ sono concordi. Seguendo un approccio cautelativo, si assume che questi errori siano additivi in valore assoluto. Quindi l'errore assoluto che affetta $Q(x,t)$ indotto dall'incertezza sulla scala delle portate $|\varepsilon_2(Q(x,t))|$ si ottiene come segue:

$$|\varepsilon_2(Q(x,t))| = |\varepsilon_{2.1}(Q(x,t))| + |\varepsilon_{2.2}(Q(x,t))| + |\varepsilon_{2.3}(Q(x,t))|$$

$\varepsilon_2(Q(x,t))$ viene quindi assunto come variabile binaria, con la stessa probabilità di assumere valore positivo $+|\varepsilon_2(Q(x,t))|$ o negativo $-|\varepsilon_2(Q(x,t))|$.

Per determinare il valore di $|\varepsilon_2(Q(x,t))|$ sono state effettuate simulazioni numeriche con il modello monodimensionale HEC-RAS, considerando il tratto del fiume Po da Isola S.Antonio a Pontelagoscuro.

### 7.2.4. Incertezza indotta dall'interpolazione ed estrapolazione della scala delle portate

Per stimare $\varepsilon_{2.1}(Q(x,t))$ attraverso simulazioni numeriche sono state ottenute 11 coppie di valori (Q(x,t) ; h(x,t)) per ciascuna sezione del corso fluviale, in un range di portate che va da 1000 a 6000$m^3/s$, con passo 500$m^3/s$. Per interpolare questi punti ed ottenere la scala delle portate sono state usate le seguenti equazioni:

- $Q(x, t) = c_1 \times (h(x, t) - c_2)^{C3}$ (2)
- $Q(x, t) = c_1 \times h(x,t) + c_2 \times h(x, t)^2 + c_3 \times h(x, t)^3$ (3)

dove $c_1$, $c_2$ e $c_3$ sono parametri di calibrazione. Di solito per l'interpolazione si considerano valori di portata ordinari, che in questo caso sono compresi nel range 1000-6000$m^3/s$. Per effettuare l'estrapolazione dei punti della scala delle portate si considerano invece valori di portata straordinari, contenuti nel range 6500-12000$m^3/s$.

L'analisi dell'errore ha messo in luce il fatto che l'equazione (3) è più idonea ad interpolare i punti rispetto alla (2). Dall'analisi è emerso che gli errori di interpolazione ed

estrapolazione sono gaussiani e che la media di $|\varepsilon_{2.1}(Q(x,t))|$ è pari al 1,2% e al 11,5% di *Q(x,t)*, con livello di confidenza del 95%, per l'errore di interpolazione ed estrapolazione rispettivamente.

### 7.2.5. Incertezza indotta dalle condizioni di moto vario

Come è noto, in condizioni di moto vario non vi è una corrispondenza biunivoca tra altezza del pelo libero e portata. In particolare, la fase ascendente di un'onda di piena provoca portate superiori alla fase discendente: quindi se nella prima fase si misura una coppia di punti livello idrico-portata, questa sarà diversa in corrispondenza dello stesso livello idrico in fase di onda discendente, a cui corrisponderà una portata inferiore.

Per valutare l'errore dovuto alla presenza di moto vario, sono state fatte 2000 simulazioni di eventi di piena e per ciascuna è stata stimata la scala delle portate per moto vario.

Per ciascun valore di *Q(x,t)* nel range 1000-120000$m^3$/s, con step 500$m^3$/s, e per ogni sezione trasversale sono stati considerati gli errori assoluti più elevati, in modo da ottenere una corrispondenza biunivoca tra $|\varepsilon_{2.2}(Q(x,t))|$ e *Q(x,t)*. Assumendo che l'errore sia gaussiano, la media di $|\varepsilon_{2.2}(Q(x,t))|$ risulta essere pari al 9,8% di *Q(x,t)*, con livello di confidenza del 95%.

### 7.2.6. Incertezza indotta dalla variazione stagionale della scabrezza di fondo

La scabrezza di fondo, rappresentata dal valore del coefficiente di Manning, dipende dallo stato della vegetazione, ed è quindi affetta da variazioni stagionali. Questo provoca cambiamenti nella scala delle portate ed influisce sulla stima della portata. L'errore $\varepsilon_{2.3}(Q(x,t))$ è stato calcolato per ciascun valore di portata nel range 1000-12000$m^3$/s, con step di 500$m^3$/s, per le condizioni autunnali e primaverili. Assumendo un errore ad andamento gaussiano, il valore medio di $|\varepsilon_{2.3}(Q(x,t))|$ è pari al 4,9% di *Q(x,t)*, con livello di confidenza del 95%.

### 7.2.7. Incertezza totale sulla scala delle portate

L'incertezza totale indotta dalla scala delle portate è stata ottenuta sommando i tre contenuti di cui sopra. In termini percentuali, il valore di $|\varepsilon_2(Q(x,t))|$ varia tra il 1,8% e il 38,4% del valore di $Q(x,t)$, con valor medio di 21,2% e deviazione standard pari al 10,8%.

### 7.2.8. Calcolo dell'errore totale

Attraverso la (1) è stato infine possibile stimare l'errore totale su $Q(x,t)$ al 95% di confidenza, sotto l'assunzione di indipendenza di $\varepsilon_1(Q(x,t))$ e $\varepsilon_2(Q(x,t))$. Nell'analisi si è tenuto conto che $\varepsilon_1(Q(x,t))$ è una variabile aleatoria gaussiana con media zero e deviazione standard pari a $0{,}027Q(x,t)$, mentre $\varepsilon_2(Q(x,t))$ è una variabile binaria che assume i valori $+|\varepsilon_2(Q(x,t))|$ e $-|\varepsilon_2(Q(x,t))|$ con pari probabilità.

I limiti di confidenza al 95% di un assegnato valore di $Q(x,t)$ sono stati valutati calcolando la relazione:

$$Q(x,t) \pm \{\alpha \times 0.027Q(x,t) + |\varepsilon 2(Q(x,t))|\} = Q(x,t) \pm \varepsilon * (Q(x,t))$$

dove $\alpha$ è lo 0,95 quantile della distribuzione normale standard e $\varepsilon * (Q(x,t))$ è la semi-ampiezza dell'intervallo di confidenza del 95%. L'errore indotto dal metodo della scala delle portate risulta essere compreso nel range 6,2% - 42,8%, con valor medio pari al 25,6%.

## 7.3. *Rumore Bianco*

Il rumore bianco viene spesso utilizzato nel descrivere le incertezze ambientali all'interno di un modello. Ad esempio, nei modelli ARMA descriventi processi naturali, la parte a media mobile fa riferimento a questo tipo di segnale. Oppure, nel modellizare le dinamiche di popolazioni ecologiche, i termini di incertezza derivanti da fattori ambientali o biologici vengono sempre descritti da un rumore bianco.

Il rumore bianco è un processo statistico e, come tale, può essere descritto dalla sua *funzione di autocorrelazione*:

$$R(t_1, t_2) = E[x_{t1} x_{t2}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_{t1} x_{t2} F_{x_{t1} x_{t2}}(x_{t1}, x_{t2}) dx_{t1} dx_{t2}$$

Dove t indica un determinato istante temporale, $x_{t1}$ e $x_{t2}$ i valori aleatori output del processo negli istanti $t_1$ e $t_2$ rispettivamente.

Un'alternativa equivalente alla funzione di autocorrelazione è la *funzione di*

*autocovarianza*:

$$\gamma(t_{1,}t_2)=E[(x_{t1}-E[x_{t1}])(x_{t2}-E[x_{t2}])]=R(t_{1,}t_2)-E[x_{t1}]E[x_{t2}]$$

Se i valori attesi delle due variabili sono nulli, autocovarianza e autocorrelazione coincidono. L'autocovarianza è la covarianza fra le due variabili $x_{t1}$ e $x_{t2}$.

Il rumore bianco è un processo stocastico con funzione di autocovarianza nulla per ogni coppia ($t_1$, $t_2$), con $t_1 \neq t_2$, cioè:

$$\gamma(t1, t2) = 0 \qquad \text{se } t1 \neq t2$$
$$= VAR\,[xt1] \qquad \text{se } t1 = t2$$

Cioè, ad ogni istante *t* la variabile aleatoria $x_t$ è incorrelata a tutte la variabili che la precedono e che la seguono. Inoltre in molte applicazioni si richiede che il rumore bianco abbia media nulla in ogni istante temporale, cioè $E[x_t] = 0$ per ogni t. Inoltre, l'errore bianco gaussiano prevede che ogni realizzazione del processo sia una variabile aleatoria stocastica distribuita come una normale di media nulla e varianza pari all'ampiezza dell'errore N (0, $\sigma^2$).

# 8. CONCLUSIONI

*“Ho imparato tanto dai miei errori, che sto pensando di continuare a farne”*

Anonimo

Nei primi tre capitoli del testo si sono visti i principali strumenti statistici di base, partendo dalla definizione di variabile aleatoria e passando poi, attraverso il teorema centrale del limite, alla ricerca e definizione dei principali stimatori. Si è proseguito poi con la presentazione del metodo della regressione lineare col metodo dei minimi quadrati, soffermandosi sulle proprietà dei parametri e sui principali metodi di verifica della bontà del metodo. In seguito sono stati affrontati gli intervalli di confidenza e i test d'ipotesi, soffermandosi sui principali test in uso nello studio di dati ambientali.

Nel quinto capitolo sono state affrontate le principali metodologie di analisi di rischio. Sono stati forniti dapprima gli strumenti necessari ad affrontare un analisi di rischio generica, le cui origini storiche risalgono alla valutazione dei rischi di fallimento di catene di montaggio industriali e che consistono nel calcolo di determinati indici di rischio. Si è poi entrati nello specifico dell'analisi di rischio geologico, idrogeologico e per siti contaminati, descrivendo le attuali metodologie con cui vengono svolte.

Nel capitolo sesto sono stati illustrati i principali metodi di formulazione di modelli matematici riferiti a fenomeni ambientali. In particolare, ci si è concentrati in prima istanza sul metodo classico paragonato a quello bayesiano, il che ha presupposto la trattazione del metodo Monte Carlo e dell'algoritmo MCMC. Si è poi passati alla trattazione dei modelli ARMA, il cui utilizzo rimane oggi il più diffuso in molti campi dell'ingegneria ambientale. Il capitolo si conclude con l'esplicitazione delle principali tecniche di selezione dei modelli, fornendo un esempio reale di applicazione nello studio di una popolazione ecologica.

Nel settimo ed ultimo capitolo è stato affrontato il tema del trattamento delle incertezze. Nello specifico, è stato inizialmente esposto il metodo Monte Carlo ibrido probabilistico-possibilistico, con un'applicazione reale volto alla determinazione dell'incertezza relativa all'inventario di emissioni dovute alla combustione domestica di legna. Segue la trattazione dei principali metodi di determinazione dell'incertezza nelle

misure di portata fluviale in Italia. Il capitolo si conclude con una breve descrizione del processo di rumore bianco, che viene oggi utilizzato nel descrivere le principali fonti di incertezza dei fenomeni naturali ed ambientali.

L'obiettivo del presente lavoro, come anticipato nell'introduzione, è quello di fornire ad uno studente di ingegneria ambientale un quadro generale dei metodi statistici necessari ad affrontare i principali problemi derivanti dallo studio di fenomeni naturali, qualsiasi si la specializzazione che si intende conseguire. Per una piena padronanza dei suddetti metodi è necessaria una conoscenza approfondita di molti strumenti matematici di cui non si è fornita, per scelta, una trattazione capillare. Si suppone, infatti, che il presente lavoro, che prende la forma di un manuale, vada utilizzato congiuntamente a testi che affrontino nello specifico gli argomenti trattati.

# 9. BIBLIOGRAFIA


- D.R. Anderson, K. P. Burnham, *Multimodel Inference: Understanding AIC and BIC in Model Selection,* Sociological Methods Research; 33; 261; 2004*;*
- D. R. Anderson, K.P. Burnham, K. P. Huyvaert, *AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons,* Behav Ecol Sociobiol 65:23–35, 2011;
- L. L. Bailey, J. E. Hines, D. I. MacKenzie, J. D. Nichols, K. H. Pollock, J.A. Royle, *Occupancy Estimation and Modeling, Inferring Patterns and Dynamics of Species Occurrence, ELSEVIER, 2006;*
- G. Di Baldassarre, A. Montanari, *Uncertainty in river discharge observations: a quantitative analysis,* Hydrol. Earth Syst. Sci., 13, 913–921, 2009;
- A. Ballabene, tesi di laurea, *Applicazione della teoria dei valori estremi al valore di rischio,* Univ. Degli studi Roma Tre, 2003*;*
- P. Baraldi, F. Cadini, G. Lonati, G. Ripamonti, E. Zio, *Uncertainty propagation methods in dioxin/furans emission estimation models, ESREL, 2011;*
- P. Baraldi,M. Couplet, E. Ferrario, A. Pasanisi, N. Pedroni, E. Zio , *Monte Carlo and fuzzy interval propagation of hybrid uncertainties on a risk model for the design of a flood protection dike, ESREL, 2011;*
- *G. P. Beretta, Il trattamento e l'interpretazione dei dati ambientali, qualità dei suoli e delle acque sotterranee nella bonifica dei siti contaminati, Pitagora Ed., Bologna, 2004;*
- D. Bocchiola, *Programma Tecnico di Sperimentazione sul Deflusso Minimo Vitale nell'alto corso del fiume Serio , ALLEGATO 3 : Valutazione della scala delle portate e misura dei deflussi per il fiume Serio a Ponte Nossa, 2005;*
- D. Bocchiola, C. De Michele, R. Rosso, *Review of recent advances in index flood estimation,* Hydrology & Heart System Sciences, 7(3), 283-296, 2003*;*
- R. Casagrandi, M. Gatto, A. Hardenberg, A. Mignatti, A. Provenzale, *Sex-and age-structured models for Alpine ibex Capra ibex ibex population dynamics,* Wildlife Biology, 18(3):318-332, 2012*;*
- S. Caserini, M. Giugliano, C. Pastorello, *Analisi dell'incertezza delle emissioni da traffico,* 2004;

- *G. Corani, M. Gatto, Structural risk minimization: a robust method for density-dependence detection and model selection,* Ecography 30: 400 Á 416, 2007*;*
- S. Cucco, tesi di laurea, Inventario delle emissioni di carbonio elementare e o rganico in Lombardia, Politecnico di Milano, 2011;
- C. De Michele, H. Pavlopoulos, R.J. Scholes, R. Vezzoli, *A minimal model of soil water–vegetation interactions forced by stochastic rainfall in water-limited ecosystems*, Ecological Modelling, 212, 397-407, 2008;
- D. Dubois, *Possibility Theory and Statistical Reasoning*, Institut de Recherche en Informatique de Toulouse, 2006;
- V. Francani, D. Fumagalli, P. Gattinoni, S. Mottini, *Modelli concettuali dinamici per l'analisi del rischio geologico a fini progettuali, Quaderni di geologia applicata, Pitagora Ed., 2005;*
- V. Francani, P. Gattinoni, *Applicazione di tecniche per l'analisi del rischiodi inquinamento delle acque sotterranee*, Aquifer Vulnerability and Risk, 2nd international Workshop, 4th congress on the Protection and Management of Groundwater, Parma, 2005;
- V. Francani, P. Gattinoni, *Depletion risk assessment of the Nossana Spring (Bergamo, Italy)*
- *based on the stochastic modeling of recharge, Hydrogeology Journal 18: 325–337, 2010;*
- V. Francani, P. Gattinoni, *Previsione e prevenzione degli inquinamenti delle acque sotterranee: l'approccio sistemico, Quaderni di geologia, Pitagora Ed., 2003;*
- P. Gattinoni, E. Pizzarotti, A. Rizzella, *L. Scesi, Geological Risk in Tunneling: the Example of Teheran Underground, 2008;*
- R. King, B. J. T Morgan, O. Gimenez, S. P. Brooks, *Bayesian Analysis for population ecology,* Chapman & Hall/CRC, Interdisciplinary Statistics Series, *2010;*
- N. T. Kottegoda, R. Rosso, *Applied statistics for civil and environmental engeneers*, second ed., Blackwell Publishing, 2008*;*
- Y. Li, E. Zio, *Uncertainty Analysis of the Adequacy Assessment Model of a Distributed Generation System, 2011;*
- *E. Piazza, Probabilità e statistica, appunti di teoria ed esrcizi risolti, Progetto Leonardo, Bologna, 2008;*
- F. Ruggeri , F. Faltin, R. Kenett, *Bayesian Networks*, Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons, 2007*;*

# 10. SITOGRAFIA


- *digilander.libero.it/statistici/inferenza/cap8.pdf;*
- *home.deib.polimi.it/gatto/ecologia2/03-Stocasticita.pdf*
- *http://mox.polimi.it/~nobile/Teaching/StatComp/Lezione06/Presentazione_MCMC.pdf;*
- *http://www-3.unipv.it/webidra/materialeDidattico/fugazza/la%20difesa%20dalle%20piene.pdf;*
- *http://www.diet.unina.it/giacinto.gelli/corsoTDSinfo_9CFU/03_Segnali_aleatori_gaussiani.pdf;*
- *http://www.diiar.polimi.it/cimi/corso.asp?id=110;*
- *http://www.diiar.polimi.it/cimi/corso.asp?id=6*
- http://www.dismi.unimo.it/Members/csecchi_
- http://www.micro.dibe.unige.it/maurizio_valle/Elettronica_Industriale_2/Attachment_Kalman_filter.pdf
- http://www.sburover.it/psice/statistica/13_Test_non_parametrici.pdf
- *http://www.science.unitn.it/~sala/events2012/MatFinTN2012_04_Bee_Extreme%20Value%20Theory.pdf;*
- http://www.unipa.it/~laura.giarre/kalman_app.pdf
- *users.dma.unipi.it/~flandoli/dispense_Flandoli_2011_versione3.pdf;*