POLITECNICO DI MILANO

Facoltà di Ingegneria Civile Ambientale e Territoriale

Laurea specialistica in Ingegneria Civile

Orientamento Rilevamento e Controllo

APPLICAZIONE DEI METODI DI FOURIER NELL'AMBITO DEL TRATTAMENTO DELLE OSSERVAZIONI

Relatore: Prof. Luigi MUSSIO

Tesi di Laura

Roberto Nino Munevar

Matr Nr. 740543

Anno Accademico 2009-2010

Indice Generali

1. Introduzione	Pagina 3
2, Campionamento	Pagina 4
3. Test Sequenziali	Pagina 14
4. Metodo di Fourier	Pagina 30
5. Analisi ed Applicazioni	Pagina 48
6.Conclusioni e Futuri Studies	Pagina 59
7. Bibliografia	Pagina 68

1. INTRODUZIONE

Indice Capitolo 1

a. Riassunto....1
b. Statistica descrittiva....1
c. Stima di parametri di modelli....3
d. Classi di problemi....4
e. Inferenza statistica....7

a. Riassunto

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli da interpretare sono concepiti come popolazioni di dati ideali, costituenti un universo da cui estrarre dati reali (campionamento) costituiti, a loro volta, da campioni da interpretare. Le principali proprietà delle stime sono: la correttezza, la consistenza, l'efficienza, la sufficienza e la robustezza.

Nel campionamento, numerico e staticamente significativo, di basi di dati di notevoli dimensioni, esistono diverse problematiche collegate alla rappresentabilità dei campione1. è questo il caso per esempio di campionamenti poco significativi o di campionamenti con un numero di dati estremamente elevato. Tale argomento si pone in relazione con la stima delle frequenze, principalmente per quanto riguarda la ricerca dei limiti di confidenza di una frequenza osservata, per una popolazione distribuita binomialmente. Per stabilire la bontà dei modelli d'interpretazione delle osservazioni c'è bisogno di applicare test. I test statistici necessitano di un campionamento di dati significativo e rappresentativo della popolazione da cui è considerato estratto. Queste fatto può comportare l'esigenza di collezionare un numero elevato di estrazioni, prima di poter indicare ipotesi sensate e sopratutto svolgere il test.

Un'alternativa che permette di eseguire l'analisi di evidenze sperimentali quando sono presenti pochi dati è il test sequenziale dato che lo scopo di questo tipo di analisi è quello di arrivare a scegliere, tra ipotesi alternative, con il minimo numero di osservazioni. Parallelamente, la capacità di acquisire grosse moli di dati, durante un processo di misura, da un lato, permette una descrizione più puntuale del fenomeno, dall'altro, richiede maggior attenzione nella scelta del modello stocastico. Questo aspetto, a volte trascurato, può portare seri problemi sull'attendibilità delle stime, quando si ignorano eventuali correlazioni presenti tra le misure. Tuttavia in alcuni insiemi di dati, è possibile evidenziare e successivamente quantificare la correlazione, presente nelle osservazioni, mediante un approccio di tipo stocastico.

Un'altra alternativa alla soluzione dei problemi ed assumendo che i dati siano una sequenza estratta dalla realizzazione di un qualche processo stocastico (segnali stocastici), è fornita dai metodi di Fourier che, qui come altrove, trovano la loro naturale collocazione, in parallelo al metodo degli elementi finiti.

b. Statistica descrittiva

Un discorso preciso sulla statistica descrittiva si avvia con la definizione di variabile statistica e di variabile casuale, la postulazione di un'identità formale fra le stesse e la presentazione delle loro principali statistiche. Le variabili statistiche sono il risultato di esperimenti e, pertanto, sono concrete (ovvero costituite da dati reali od osservazioni, come la totalità dei dati a referenza spaziale, tempo varianti e non, quali, ad esempio, le misure geodetiche e geomatiche), finite (perché qualsiasi esperimento incontra evidenti limiti di spazio, tempo ed altre condizioni limitative) e discrete (perché qualsiasi esperimento è eseguito con una determinata accuratezza). Conseguentemente esse sono caratterizzate da un insieme di valori argomentali

(eventualmente raggruppati in classi), associati a frequenze elementari (assolute, come risultato di un conteggio, oppure relative, se la totalità è normalizzata ad uno) ed alle frequenze cumulate delle frequenze elementari.

Le variabili casuali sono modelli interpretativi e, pertanto, sono astratte (ovvero costituite da dati ideali od osservabili) ed, in generale, illimitate e continue (anche se, raramente, ad eccezione della teoria dei giochi, esse possono essere finite e discrete). Conseguentemente esse sono caratterizzate da un campo d'esistenza, associato ad una funzione densità di probabilità ed ad una funzione distribuzione di probabilità (comunemente detta: probabilità). L'identità formale fra variabili statistiche e variabili casuali discende dalla loro completa indistinguibilità, a valle della loro definizione. Allora la presentazione delle principali statistiche può essere eseguita congiuntamente per entrambe.

Per le variabili ad una dimensione, le principali statistiche rispondono alla quantizzazione delle idee di: centro, dispersione, simmetria e curtosi (comportamento delle code). Come noto, il centro può essere indicato tramite la moda, la mediana, le medie (aritmetica, geometrica, armonica, ponderata, potata, ecc.) od altro, la dispersione può essere valutata in base all'ampiezza, ai quantili, alla varianza, agli scarti assoluti medio o mediano, ecc., mentre gli indici di asimmetria e curtosi hanno, solitamente, poche varianti.

Per le variabili a due dimensioni, le principali statistiche (oltre a quelle monodimensionali marginali o condizionate) rispondono alla quantizzazione dell'idea di dipendenza. Come noto, dipendenza è un concetto molto generale che, fra totale e completa indipendenza e perfetta dipendenza (o dipendenza in legge), si articola in connessione (dipendenza vaga e generica), regressione (quasi-dipendenza funzionale) e correlazione (quasi-dipendenza lineare). Ancora numerosi sono gli indici ed i coefficienti che esprimono il grado della dipendenza o meno (si noti, a riguardo, come tutti siano normalizzati ad uno, assumendo anche valori negativi, fino a meno uno, se non intrinsecamente positivi).

Per le variabili a più di due dimensioni, a rigore, occorre continuare lo studio del loro raggrupparsi (come con gli indici di nuvolosità, ecc.). Tuttavia nel caso frequente in cui il modello interpretativo è fornito dalla variabile casuale normale, questo studio è del tutto superfluo. Si ricordi, inoltre, che detta variabile casuale è completamente caratterizzata dal vettore delle medie e dalla matrice di varianza-covarianza, cosa che rende superflue altre statistiche del centro, della dispersione e della dipendenza (covarianza comporta correlazione, ovvero dipendenza lineare e niente altro) e del tutto inutili le statistiche superiori (la variabile casuale normale è simmetrica e l'indice di curtosi vale, in ogni caso, tre).

Ulteriori vantaggi dell'adozione, quale modello interpretativo, della variabile casuale normale sono dati dall'invarianza della distribuzione di probabilità di detta variabile casuale, rispetto a trasformazioni lineari della variabile casuale stessa, e dell'ottimalità della stima dei parametri di modelli, supportati dalla variabile casuale normale, se le ipotesi di corrispondenza fra dati e modelli sono perfettamente soddisfatte (ovvero se i dati non sono affetti, in alcun modo, da dati anomali). A tutto ciò, si aggiunge la linearità dei sistemi da risolvere per la stima dei parametri di modelli, fatto di primaria importanza, in quanto solo i sistemi lineari ammettono, senza eccezioni e purché non-singolari, soluzioni esattamente determinabili, indipendentemente dal numero di equazioni ed incognite di cui si compongono.

La statistica descrittiva termina con alcuni teoremi limite. Fra questi il teorema di Bernoulli[1] (o legge dei grandi numeri) mostra la convergenza, in probabilità, delle frequenze di una variabile statistica alle probabilità di una corrispondente variabile casuale, mentre il teorema di Gauss[2] (o limite centrale della statistica) mostra la convergenza, in legge, della combinazione lineare (ovvero delle somme, come caso

---

1 $f(x)=P(X=x)=¿\binom{n}{x} p^x(1-p)^{n-x} per x=0,1,2,\ldots,n$

4

5

particolare) di variabili casuali qualsiasi, purché aventi ciascuna dispersione comparabile con le altre, alla variabile casuale normale. I due teoremi giustificano, rispettivamente, la comparazione fra variabili statistiche e variabili casuali, al di là della sopracitata identità formale, e le operazioni di media aritmetica, ponderata o potata fra osservazioni dirette per aumentare la normalità del comportamento dei dati.

c. Stima di parametri di modelli

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli, da interpretarsi, come già detto, quali variabili casuali, sono concepiti come popolazioni di dati ideali (od osservabili), costituenti un universo da cui estrarre dati reali (od osservazioni), costituenti, a loro volta, campioni da interpretare, pertanto, come le suddette variabili statistiche. L'operazione d'estrazione è detta campionamento[3] e, se i dati sono fra loro indipendenti, come avviene, auspicabilmente, nelle osservazioni dirette di fenomeni, il campionamento è detto bernoulliano; schemi di campionamento più complessi, attinenti alle problematiche della progettazione, simulazione ed ottimizzazione, sono considerati estranei agli scopi del presente lavoro.

Le principali proprietà delle stime sono la correttezza, la consistenza, l'efficienza e la sufficienza che, in base al significato letterale dei nomi, significano capacità di stimare parametri il cui centro coincide con il centro dei parametri dell'intera popolazione, capacità di stimare parametri con precisione ed accuratezza sempre maggiori, ed al limite infinite, al crescere della numerosità del campione, capacità di stimare parametri qualitativamente migliori delle informazioni presenti direttamente nelle osservazioni, capacità di stimare parametri conservando tutta la ricchezza di informazioni già presente nelle osservazioni dirette. Una proprietà aggiuntiva delle stime, estranea al corpus della statistica tradizionale, è la robustezza, intesa come capacità di stimare parametri indistorti, nonostante la presenza di eventuali dati anomali.

Le caratteristiche delle stime forniscono spesso, contemporaneamente, le modalità operative per effettuare le stime stesse. Infatti sono operativi tanto il metodo della minima varianza, ottimale per la statistica tradizionale, quanto diverse procedure robuste, certamente subottimali, ma capaci di evitare indesiderate distorsioni nelle stime. Altrettanto operativi sono il metodo della massima verosimiglianza[4] ed il metodo dei minimi quadrati, una particolarizzazione del primo nel caso in cui le stime avvengono in ambito lineare ed il modello interpretativo è fornito dalla variabile casuale normale. Si ricordi che, in questo caso, si ha l'ottimalità delle stime, in quanto tanto il metodo dei minimi quadrati[5], quanto quello della massima verosimiglianza, da cui discende, danno risultati perfettamente coincidenti con il metodo della minima varianza.

Tutto ciò conferma l'adozione del metodo dei minimi quadrati per il trattamento statistico delle osservazioni, agevola le sue generalizzazioni ed estensioni a tecniche complementari e giustifica un modo di procedere che prevede la centralità di detto metodo e riconduce ad esso, per quanto possibile, importanti tecniche complementari (cluster analysis, regressione multipla, analisi di varianza, delle componenti di varianza e della struttura di covarianza, procedure robuste). Si noti, in quest'ambito, il ruolo fondamentale ed indispensabile giocato dall'indissolubilità del legame fra un certo tipo di statistiche classiche elementari, la normalità e la linearità, per quanto riguarda tanto la definizione statistica delle metodologie, quanto la loro applicazione con elevate capacità risolutive in appropriati algoritmi numerici.

---

[2] la somma (normalizzata) di un grande numero di variabili casuali è distribuita approssimativamente come una variabile casuale normale standard

[3] Per campione se intende parte della popolazione che viene selezionata per l'analisi.

[4] Il metodo della massima verosimiglianza in statistica è un procedimento matematico per determinare uno stimatore.

[5] Il metodo dei minimi quadrati generalizzati di Aitken consente la stima di un modello lineare, sotto ipotesi più generali di quelle del modello classico di regressione lineare multivariata.

d. Classi di problemi

I problemi ai minimi quadrati si presentano, in generale e nell'ambito specifico delle discipline geodetiche e geomatiche, usualmente ripartiti in due classi fondamentali:

- problemi reticolari (o di compensazioni di reti);
- problemi d'interpolazione ed approssimazione di campi di punti.

Le stesse due classi si incontrano anche in problemi affini, quali ad esempio:

- i campionamenti delle osservazioni, l'ottimizzazione della configurazione di rilevamento e/o dello schema di misura, oppure dei pesi delle osservazioni;
- la "cluster analysis", l'analisi di varianza, la regressione multipla, l'analisi fattoriale (o studio delle componenti principali);
- lo studio dell'affidabilità delle osservazioni e le procedure di validazione dei dati e di stima dei parametri con procedure robuste.

In ogni caso, tutti i problemi minimi quadrati possono essere interpretati, topologicamente, come un grafo, dove:

- le osservazioni, i vincoli e le pseudo-osservazioni sovrappesate e non, come pure le informazioni a priori, le osservabili secondarie e le condizioni numeriche di regolarizzazione, costituiscono i lati del grafo;
- i parametri principali ed ausiliari (o di servizio) costituiscono i nodi dello stesso grafo.

Si noti, a riguardo, come l'interpretazione data della topologia sia indispensabile per una corretta comprensione dei casi e sottocasi in cui si articolano le suddette classi fondamentali. Le due classi fondamentali, già precedentemente enunciate, si articolano in vari e svariati casi e sottocasi, illustrati, dettagliatamente, nel prosieguo. I problemi reticolari (o di compensazione di reti) presentano come osservabili:

- differenze prime dei parametri;
- funzioni delle differenze prime dei parametri.

Il primo caso ha numerosi esempi, anche fuori dalle discipline geodetiche e geomatiche:

- discretizzazione di equazioni differenziali del primo ordine, tipiche della fisica, della chimica e delle scienze della terra;

- problemi di trasporto: schemi di circuitazione, traffico, circolazione e transazione, reti di comunicazione, distribuzione e telecomunicazione ed è costituito, per le suddette discipline, dalle reti di differenza di potenziale.

Il secondo caso è, invece, tipico delle discipline geodetiche e geomatiche, anche se non esclusivo (a questo caso, infatti, fanno riferimento ben particolari discretizzazioni di equazioni differenziali, sempre riferite ai sopracitati raggruppamenti di discipline fisiche e naturalistiche), e si articola nei seguenti sottocasi:

- l'informazione fluisce completa, bidirezionalmente, come nelle reti di differenza di potenziale, lungo ogni lato del grafo:
- l'informazione fluisce completa, unidirezionalmente, lungo ogni lato del grafo, costituendo nel suo fluire almeno un albero sul grafo stesso;
- l'informazione è irradiata, in modo completo, da alcuni nodi verso altri (senza ritorno), senza che né i primi, né i secondi si scambino alcuna informazione, costituendo nel suo fluire tanti alberi (costituiti da un solo livello, oltre la radice) sul grafo stesso, quanti sono i nodi d'emanazione:
- l'informazione è irradiata, in modo parziale, nelle stesse condizioni del sottocaso precedente, cosa che richiede l'individuazione di due o più co-alberi (sempre costituiti da un solo livello, oltre le radici) capaci di completare l'informazione trasmessa;
- l'informazione è irradiata, in modo parziale, senza restituzioni particolari.

Le osservabili differenze seconde dei parametri e loro funzioni richiedono la complessa sostituzione dei lati del grafo con triangoli fra i tre nodi interessati. Le osservabili differenze di ordine superiore e loro funzioni fanno riferimento, addirittura, a poligoni fra tutti i nodi coinvolti, cosa che rende la loro analisi ancora più complessa. Per queste ragioni, ad eccezione della discretizzazione di equazioni differenziali di secondo ordine o di ordine superiore e di loro trasformazioni funzionali, la loro adozione è estremamente rara. Per quanto riguarda la determinazione del numero di parametri principali, nel caso in cui i problemi ai minimi quadrati adottino lo schema principe delle equazioni d'osservazione, questo è sempre tale da determinare difetti di rango e singolarità del sistema da risolvere per cui sono necessari vincoli o pseudo-osservazioni sovrappesate.

I problemi d'interpolazione ed approssimazione di campi di punti presentano come osservabili funzioni dirette dei parametri principali, il cui numero, sempre nel caso in cui si voglia adottare il suddetto schema principe delle equazioni d'osservazione, non è mai tale da determinare difetti di rango e singolarità del sistema da risolvere. A tutto ciò, fanno eccezione eventuali problemi di sovra-parametrizzazione, rispetto al campionamento delle osservazioni effettuate, per cui sono indicate condizioni numeriche di regolarizzazione.

Esempi di problemi di interpolazione ed approssimazione di campi di punti sono dati da:

- ricostruzione (fitting) di linee, superfici, ipersuperfici aventi come dominio lo spazio 3D:
- descrittori di forma (form descriptors): contorni di figure (piane e/o gobbe), superfici (chiuse) di oggetti:
- centratura (matching) di segmenti, figure (immagini, mappe, disegni), oggetti (compresi modelli virtuali 3D) comunque conformati:

I problemi d'interpolazione ed approssimazione[6] di campi di punti, relativi alla ricostruzione di linee sono, ovviamente, assimilabili a quelli dello studio delle serie temporali storiche o di breve periodo, oppure frutto di simulazioni. Inoltre lo studio di serie temporali congiunto alla soluzione dei problemi reticolari (o di compensazione di reti) e/o d'interpolazione ed approssimazione di campi di punti, illustrati in precedenza, permette indagini accurate sugli aspetti dinamici delle osservabili a referenza spaziale di cui ai suddetti problemi, dando un'interpretazione unitaria a dati spazio-varianti, tempo-varianti.

Limitatamente alle discipline geodetiche e geomatiche, mentre le equazioni d'osservazione dei problemi reticolari (e di compensazioni di reti) fanno uso, in generale, di modelli grigi dedotti dalla geometria del problema in esame, le equazioni d'osservazione dei problemi d'interpolazione ed approssimazione di campi di punti fanno uso, in generale, di modelli neri.

Come noto, una vasta gamma di metodi deterministici e/o stocastici risponde positivamente alla bisogna. I primi annoverano fra i più comunemente impiegati:

- l'interpolazione polinomiale;
- il metodo degli elementi finiti e l'interpolazione con funzioni splines;
- l'analisi di Fourier, nel dominio delle frequenze;
- lo studio, sempre nel dominio delle frequenze, con ondine (wavelets).

I secondi prevedono l'interpretazione dei fenomeni in istudio come realizzazioni di un processo stocastico:

- stime di covarianza, filtraggio cross-validazione e predizione;
- studio della geometria frattale.

Uno studio dettagliato di esempi particolari e significativi di problemi reticolari (o di compensazioni di reti) può essere effettuato, nell'ambito delle discipline geodetiche e geomatiche, solo facendo riferimento a discipline specifiche, quali la geodesia, la navigazione, la topografia, la fotogrammetria ed il telerilevamento. Al contrario, uno studio dettagliato di esempi particolari e significativi di problemi d'interpolazione ed approssimazione di campi di punti richiede anche uno studio dei modelli neri. Per una migliore comprensione si tenga presente che l'insieme delle quantità osservate è sempre costituito da quattro parti distinte:

- le informazioni topologiche, ovvero i lati del grafo che indicano le connessioni esistenti fra i nodi del grafo stesso;
- le informazioni geometriche, ovvero la posizione ed altre caratteristiche degli stessi nodi;
- le informazioni metrologiche, ovvero le osservazioni (o quantità osservate) realmente effettuate;
- le informazioni stocastiche, ovvero la precisione delle osservazioni e le eventuali correlazioni fra queste.

[6] metodo per individuare nuovi punti del piano cartesiano a partire da un insieme finito di punti dati, nell'ipotesi che tutti i punti si possano riferire ad una funzione f(x) di una data famiglia di funzioni di una variabile reale.

Infatti questo insieme, altrimenti detto: base di dati provenienti da operazioni di misura, con riferimento a ciascuna delle sopraccitate quattro parti distinte, produce nei problemi ai minimi quadrati (come pure negli altri sopraccitati problemi affini), rispettivamente:

- la matrice disegno simbolica;
- la matrice disegno numerica;
- il vettore termine noto delle equazioni d'osservazione;
- la matrice di varianza-covarianza (a priori) delle quantità osservate o, più comunemente, se non esistono correlazioni fra le stesse quantità osservate, la matrice dei pesi.

Si noti che, con la sola eccezione delle osservazioni realmente effettuate, tutto quanto può essere noto già prima di compiere una sola osservazione. Da ciò derivano tutti i problemi di ottimizzazione della matrice di varianza-covarianza dei parametri:

- intervenendo nella matrice disegno per decidere sull'effettuazione o meno di ciascuna osservazione (1° ordine);
- sulla matrice di varianza-covarianza (a priori) delle quantità osservate per stabilire, note le osservazioni da effettuarsi, le precisioni delle stesse (2° ordine);
- su opportune parziali combinazioni dei due casi precedenti (3° ordine),

avendo cura di controllare, in ogni caso, l'affidabilità delle osservazioni, quale garanzia, sufficiente minimale, che il lavoro intrapreso, qualsiasi esso sia, risulti svolto a regola d'arte.

e. Inferenza statistica

La validazione dati e dei modelli si fonda sulle varie tecniche dell'analisi multivariata[7], di volta in volta, studiando la variabilità e l'interdipendenza fra gli attributi, entro una classe di oggetti. Essa prende in considerazione insiemi di dati, ciascuno dei quali, relativo ad un oggetto della classe, contiene i valori osservati di certe variabili statistiche. Questi insiemi possono, talvolta, essere completi; mentre, più A loro volta, le variabili osservate, sono in generale campioni estratti da variabili casuali, di tipo continuo o, raramente, discreto. Da una tale complessità di premesse, lo studio dell'analisi multivariata si può articolare, principalmente, nei seguenti punti.

- Semplificazione strutturale. Gli insiemi di dati devono essere ricondotti, se possibile, in forme più semplici con cambi di variabili, in particolare, con trasformazioni capaci di sciogliere variabili connesse in variabili indipendenti.

[7] Con statistica multivariata s'intende quella parte della statistica in cui l'oggetto dell'analisi è per sua natura formato da almeno due componenti.

- Classificazione degli oggetti. L'analisi dell'insieme di dati deve porre in evidenza la presenza di gruppi (clusters), ovvero di sottoinsiemi di oggetti, caratterizzati da valori preferenziali degli attributi o di parte di essi, cercando di ricondurre a poco le notevoli variabilità presenti.
- Raggruppamento degli attributi (clustering). L'analisi degli insiemi di dati deve far ricadere, per quanto possibile, differenti variabili in un unico gruppo.
- Analisi della connessione. Gli insiemi di dati devono essere studiati rispetto alla dipendenza vaga e generica o meno fra le variabili contenute (ovvero all'essere in connessione di queste ultime).
- Analisi della dipendenza funzionale. Gli insiemi di dati devono essere studiati rispetto alla dipendenza funzionale o meno fra le variabili contenute (ovvero all'essere in regressione di queste ultime), con particolare riferimento alla dipendenza lineare o correlazione (ovvero all'essere in regressione lineare o correlate).
- Costruzione e verifica d'ipotesi. Il confronto probabilistico, fra statistiche campionarie e valori teorici di riferimento, permette di formulare un giudizio critico sui risultati ottenuti nelle varie tappe dell'analisi multivariata.

I tests di validazione dei modelli permettono di sottoporre a verifica, mediante opportuni controlli e confronti d'ipotesi, le stime effettuate come, del resto, tutti i risultati ottenuti nell'ambito della statistica. Al solito, si possono avere errori nel modello deterministico: presenza di errori grossolani nelle osservazioni, ed errori nel modello stocastico: presenza di errori sistematici nelle osservazioni, ovvero cattiva conoscenza delle varianze delle osservazioni e/o delle eventuali covarianze fra le osservazioni stesse. Un'opportuna sequenza di tests permette di districarsi fra le varie cause d'errore.

Un giudizio sui risultati può essere espresso in termini numerici e statistici. I controlli di tipo numerico rispondono a problemi di condizionamento ed affidabilità che comunemente accompagnano e seguono il metodo dei minimi quadrati, comprensivo delle sue estensioni e generalizzazioni; pertanto tutto quanto riguarda i controlli di tipo numerico è considerato estraneo agli scopi del presente lavoro. I secondi comprendono i tests statistici per la valutazione di osservazioni e parametri, della loro dispersione e, se del caso, della loro dipendenza. Per quanto riguarda le osservazioni, la validazione avviene in termini di entità

degli scarti - residui (oltreché numericamente in termini di affidabilità delle osservazioni all'interno dello schema di misura) e consiste essenzialmente nell'individuazione ed eliminazione degli errori grossolani.

L'inferenza statistica (multivariata) è quella parte dell'analisi multivariata dedicata alla costruzione e verifica d'ipotesi (per problemi di controllo di qualità, oppure controllo e confronto d'ipotesi di altri problemi). Infatti il confronto probabilistico permette di formulare un giudizio critico sui risultati ottenuti, nelle varie tappe dell'analisi multivariata. Così con le usuali strategie dell'inferenza statistica, vari tests multipli consentono di discriminare, tanto stime di parametri da campioni normali, quanto statisitche di modelli non-parametrici (ovvero modelli distribution free)[8].

Gli oggetti del giudizio critico sono, come detto, i risultati ottenuti nelle varie tappe dell'analisi multivariata, in particolare: frequenze relative, contingenze, medie campionarie o altri indicatori del centro (di una popolazione), varianze campionarie o altri indicatori della dispersione, coefficienti di correlazione campionari o altri indicatori della correlazione[9].

I test multipli si differenziano per i diversi oggetti in esame, come pure per la distribuzione di appartenenza della popolazione cui i campioni si riferiscono, se normale, oppure altra (spesso sconosciuta). In generale, comunque, tutti i campioni sono supposti fra loro indipendenti, mentre se quest'ipotesi non è soddisfatta, occorre procedere, come per i modelli non-parametrici, adottando strategie assolutamente generali, ma assai poco potenti, cioè meno capaci di discriminare fra ipotesi alternative vicine, a parità di numerosità dei campioni. Nelle seguenti tabelle, si presentano parecchi tests multipli, diversamente, rispondenti alla bisogna, facendo attenzione, in particolare, ai problemi pratici, legati alla loro applicazione ad esempi concreti. Per quanto riguarda, invece, i loro fondamenti teorici, questi si richiamano, in generale:

---

[8] L'aggettivo non parametrico (in letteratura inglese: distribution free) qualifica un particolare gruppo di tests statistici, sotto certe condizioni, sostitutivo dei tests statistici classici. Infatti i tests non-parametrici, rispetto ai test classici, presentano i seguenti vantaggi:

- la loro comprensione è immediata ed elementare;
- le condizioni di validità sono meno forti (più ampie);
- i calcoli necessari non presentano in generale difficoltà computazionali.

D'altra parte, i tests non-parametrici presentano alcuni svantaggi:

- molta informazione viene sprecata;
- la potenza del test è bassa.

Tests poco potenti tendono ad essere troppo conservativi, cioè l'ipotesi fondamentale (o nulla) è accettata anche quando dovrebbe valere l'ipotesi alternativa. Pertanto i tests statistici classici sono preferibili, quando le condizioni di validità sono soddisfatte.

In quest'ottica, ipotesi stringenti sulla normalità dei campioni e l'estrema specificità della grandezza da sottoporre al confronto d'ipotesi (coefficienti d'asimmetria e indici di curtosi, quantili nelle code) fanno, dei test di Pearson et al. e Hawkins, alcuni dei più potenti tests noti proprio per questo riportati nel seguito, pur convenendo sulla loro circoscrizione a classi di problemi particolari (ad esempio, individuazione ed eliminazione di dati anomali). Invece quando le condizioni di validità non sono soddisfatte, ad esempio se la distribuzione della popolazione non è quella normale, oppure se gli elementi della popolazione non sono statisticamente indipendenti, oppure se le varianze della popolazione sono significativamente diverse fra loro, allora i tests statistici non-parametrici devono essere utilizzati. Tutto ciò vale in particolare quando i campioni sono piccoli. Infatti una delle condizioni di validità dei tests classici è la dimensione grande dei campioni. L'analisi multivariata di campioni di fenomeni vari richiede spesso:

- il confronto fra medie di campioni aventi diversa varianza;
- il fra varianze di campioni non statisticamente indipendenti;
- il confronto fra contingenze di campioni con distribuzione diversa da quella normale.

In questi casi, i tests statistici non-parametrici sono indispensabili.

[9] Si chiama contingenza la differenza fra una probabilità doppia (o una frequenza relativa doppia) ed il prodotto delle corrispondneti probabilità marginali (o delle corrispondenti frequenze relative marginali). La contingenza fra due variabili casuali (o due variabili statistiche) indipendenti ha valore zero; in corrispondenza ad ogni altro valore (compreso fra -1 e 1), le due variabili si dicono connesse.

- alla definizione assiomatica di probabilità;
- alla legge dei grandi numeri ed al limite centrale della statistica;
- ai teoremi della normalità: conservazione per trasformazioni lineari, identità fra indipendenza ed incorrelazione;
- alla definizione delle variabili casuali χ^2 (chi quadrato), *t* (t di Student) e *F* (F di Fisher);
- al calcolo di probabilità estremali (ad es. di Kolmogorov–Smirnov e di Hawkins), ove necessario;
- al teorema di decomposizione ortogonale degli scarti;

e nello specifico, alla trasformazione della distribuzione di probabilità, secondo requisiti da definirsi, caso per caso, così come è costruito un determinato test multiplo. L'effettiva esecuzione di un qualsiasi test statistico si attua sempre compiendo i seguenti passi:

- formulazione di una determinata ipotesi fondamentale (o nulla);
- scelta del livello di significatività;
- costruzione di una statistica campionaria, a partire da dati osservati;
- partizione della distribuzione di probabilità della statistica campionaria;
- effettuazione del confronto d'ipotesi,

avendo cura di controllare la potenza del test, nel caso si voglia prendere in considerazione una o più ipotesi alternative (all'ipotesi fondamentale o nulla).

2. CAMPIONAMENTO

Indice Capitolo 2

a. Campionamento....1
b. Tipi di campionamento....3
c. Test di ipotesi....5

a. Campionamento

Nei lavori di controllo di qualità e, in generale, controllo e confronto d'ipotesi si pone spesso il problema di eseguire campionamenti su una base di dati che, per dimensione, eccede la possibilità di indagini esaustive. Le procedure di campionamento permettono l'estrazione di campioni, numericamente e statisticamente significativi, all'interno delle basi di dati da sottoporre ad indagini statistiche. Si ricorda, a riguardo, che la popolazione può essere finita (essa è costituita da un numero determinato di elementi ed il campionamento è eseguito in blocco) od infinita (quando si osservi, ad esempio, il valore di una variabile continua, oppure se si campiona con ripetizione) e che il piano di campionamento di una popolazione può essere: semplice[1] o sistematico[2], stratificato[3] o a cluster[4]. Dette procedure suggeriscono di effettuare la stratificazione o clusterizzazione delle basi di dati, al fine di ottenere una miglior spiegazione dei fenomeni e processi in studio riducendo, nel contempo, il rumore residuo, insito nei dati stessi e nelle metodologie del loro rilevamento.

Una **strategia logistica**, per il campionamento delle basi di dati oggetto di studio, prende in considerazione il numero di ripetizioni dello stesso tipo ed adotta, quale misura di campionamento:

- la totalità, quando la numerosità delle stesse è inferiore ad un valore di soglia prefissato;
- un certo valore percentile, quando la numerosità supera il suddetto valore di soglia, ma non eccede, nel contempo, un secondo valore di soglia prefissato;
- un dato valore (assoluto), quando è superiore al secondo valore di soglia.

Lo scopo di questa strategia è prevenire, insieme, un campionamento poco significativo di campioni poco numerosi ed un campionamento eccessivo di campioni enormemente grandi. La curva logistica[5] è una funzione algebrica, ben nota in demografia, per lo studio dell'accrescimento di popolazioni in ambienti limitati. Essa è caratterizzata da tre parametri, aventi il significato geometrico descritto per presentare la voluta approssimazione lineare. Si riportano, di seguito, alcune espressioni analitiche del suo studio:

$$y=a+(b-a)e^{\frac{-4(b-a)}{e^2 cx}} \qquad (2.1)$$

$$\lim_{x\to 0+} y=a \qquad \lim_{x\to\infty} y=b$$

[1] In cui ogni individuo ha la stessa probabilità di essere scelto.
[2] Dalla popolazione N e la dimensione del campione si calcola il quoziente interno R, che permette stabilire nel campione gli individui della lista che occupano i posti k, k+R, k+2R.
[3] Consiste in dividere gli N individui della popolazione in sottopopolazioni.
[4] Gli N individui della popolazione sono suddivisi in molti gruppi.
[5] Quando una popolazione è limitata da un valore limite L.

$$y^I=(b-a)e^{\frac{-4(b-a)}{e^2cx}}\frac{4(b-a)}{e^2cx^2} \qquad (2.2)$$

$$y^{II}=(b-a)c^{\frac{-۴(b-a)}{e^۲cx}}\frac{b(b-a)^۲}{e^۴c^۲x^۴}\left(۱-\frac{e^۲cx}{۲(b-a)}\right) \qquad (2.3)$$

$$y^{II}=0 \quad \text{Per} \qquad x=\frac{2(b-a)}{e^2c}=K$$

$$y(K)=a+(b-a)e^{-2} \qquad y^I(K)=c$$

E' chiaro che più il campione è grande, più le stime sono attendibili. D'altra parte, al crescere delle dimensioni del campione aumentano i costi e le difficoltà tecniche ed organizzative, oltretutto con rischi, non troppo remoti, di non riuscire più a garantire l'indipendenza fra le osservazioni e pertanto la beroullianità del campione ottenuto. L'esistenza di un limite superiore va nella direzione della ricerca di un equilibrio tra la diminuzione della varianza e la complessità dei problemi.

Per delineare i termini del problema, si fa riferimento alla **stima di frequenze**, in particolare alla ricerca dei limiti di confidenza di una frequenza osservata, per una popolazione distribuita binomialmente[6]. A riguardo, si ricordi che, essendo *p* e *q* le probabilità associate ai due valori argomantali 1 e 0, la media della variabile semplice è *p* e la varianza *pq*, la media della variabile somma è *np* e la varianza *npq*, mentre la media della variabile media è *p* e la varianza *pq/n*. Data la distribuzione binomiale[7], approssimata da una normale[8]: $z=(x-n\hat{p})/\sqrt{n\hat{p}q}$, essa viene corretta per continuità:

$$z_{\mathbf{inf}}=\frac{x_{\mathbf{inf}}+0.5-n\hat{p}}{\sqrt{n\hat{p}q}} \qquad (2.4)$$

$$z_{\mathbf{sup}}=\frac{x_{\mathbf{sup}}-0.5-n\hat{p}}{\sqrt{n\hat{p}q}} \qquad (2.5)$$

ovvero:

---

[6] O distribuzione di Bernoulli dove $f(x)=P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}\ per\ x=0,1,2,\ldots,n$

[7] Quando il numero di prove è grande, il calo con la distribuzione binomiale è molto lungo. In tal caso è possibile utilizzare la distribuzione normale per approssimare la distribuzione binomiale.

[8] Distribuzione di Gauss standardizzata, dove $f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}$ e $z=\frac{X-\mu}{\sigma} \quad -\infty<z<\infty$

$$z_{\inf}\sqrt{\frac{\hat{p}\,q}{n}}=\frac{x_{\inf}}{n}+\frac{1}{2\mathrm{n}}-\hat{p}=\Pi f_{inf}+\frac{1}{2\mathrm{n}}-\hat{p} \tag{2.6}$$

$$z_{\sup}\sqrt{\frac{\hat{p}\,q}{n}}=\frac{x_{\sup}}{n}-\frac{1}{2\mathrm{n}}-\hat{p}=\Pi f_{sup}-\frac{1}{2\mathrm{n}}-\hat{p} \tag{2.7}$$

ottenendo i seguenti risultati, in cui $z = z_{sup} = -z_{inf}$:

$$f_{\inf}=\left(\hat{p}-\frac{1}{2\mathrm{n}}\right)-z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.8}$$

$$f_{\sup}=\left(\hat{p}+\frac{1}{2\mathrm{n}}\right)+z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.9}$$

b. Tipi di campionamento

Tali equazioni riguardano il campionamento con riposizionamento; in caso di campionamento senza riposizionamento[9]; ovvero con correzione per la popolazione finita, si apportano correzioni adeguate. Come noto, la correzione per la popolazione finita ha espressione: $\sqrt{(N-n)/(\mathbf{N}-\square)}$, dove N è la numerosità della popolazione e n la numerosità del campione[10]; pertanto si ha:

$$f_{\inf}=\left(\hat{p}-\frac{\text{١}}{\text{٢n}}\right)-z_{\alpha}\sqrt{\frac{\hat{p}(\text{١}-\hat{p})}{n}\frac{N-n}{N-\text{١}}} \tag{2.10}$$

$$f_{\sup}=\left(\hat{p}+\frac{1}{2\mathrm{n}}\right)+z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1}} \tag{2.11}$$

A questo punto, espresso l'intervallo di confidenza:

$$f_{\sup}-f_{\inf}=\frac{1}{n}+2\,z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1}}=2\mathrm{a} \tag{2.12}$$

e sostituendo, al posto di $N-1$, il valore approssimato N:

$$\text{٢}\;z_{\alpha}\sqrt{\frac{\hat{p}(\text{١}-\hat{p})}{n}\frac{N-n}{N}}=\text{٢a}-\frac{\text{١}}{n} \tag{2.13}$$

[9] Campionamento Casuale.

[10] Per campioni estratti senza riposizionamento da una popolazione finita di ampiezza N la varianza della distribuzione della media campionaria è $\sigma_X^{\text{٢}}=\frac{\sigma^{\text{٢}}}{n}.\frac{N-n}{N-\text{١}}$

$$4z_{\alpha}^{2} n\left[\hat{p}(1-\hat{p})\frac{N-n}{N}\right]=4a^{2}n^{2}+1-4an \qquad (2.14)$$

$$۴\left[a^{۲}+z_{\alpha}^{۲}\hat{p}\frac{(۱-\hat{p})}{N}\right]n^{۲}-۴\left[a+z_{\alpha}^{۲}\hat{p}(۱-\hat{p})\right]n+۱=۰ \qquad (2.15)$$

si ottiene il valore atteso per la numerosità del campione e, nel caso in cui il livello di significatività (valga 5%, la variabile normale standardizzata 1.96: 2 e la probabilità elementare (nel caso più sfavorevole) p sia posta uguale a 0.5, trascurando $1/N$ e sviluppando in serie binomiale $(a+1)^{1/2}$, lo stesso valore diviene:

$$n=\frac{\left[a+z_{\alpha}^{۲}\hat{p}(۱-\hat{p})\right]\pm\sqrt{z_{\alpha}^{۴}\hat{p}^{۲}(۱-\hat{p})^{۲}+z_{\alpha\alpha}^{۲}\hat{p}(۱-\hat{p})\left(۲a-\frac{۱}{N}\right)}}{۲\left(a+z_{\alpha}^{۲}\hat{p}\frac{(۱-\hat{p})}{N}\right)} \qquad (2.16)$$

$$n=\frac{(a+۱)\pm\sqrt{۱+۱\left(۲a-\frac{۱}{N}\right)}}{۲\left(a+\frac{۱}{N}\right)}=\frac{N(a+۱)\pm\sqrt{۱+۲a}}{۲\left(a+\frac{۱}{N}\right)}=\frac{N\left[(a+۱)\pm(a+۱)\right]}{۲(۱+Na^{۲})} \qquad (2.17)$$

Trascurando la soluzione $n = 0$, si ha:

$$n=\frac{N(a+۱)}{۱+Na^{۲}}=\frac{N}{۱+Na^{۲}} \qquad (2.18)$$

in cui si è trascurato a, trattandosi di valore relativamente piccolo.[11]

[11] Allo stesso risultato, si poteva giungere, operando in ambito lineare, a partire dall'espressione dell'intervallo di confidenza nella quale trascurare il termine $1/2n$:

$$۲z_{\alpha\alpha}\sqrt{\frac{\hat{p}(۱-\hat{p})}{n}\frac{N-n}{N}}=۲a \qquad \text{al quadrato:} \qquad z_{\alpha\alpha}^{۲}\hat{p}(۱-\hat{p})\frac{N-n}{N}=a^{۲}n \qquad \text{ovvero:}$$

$$n=\frac{Nz_{\alpha\alpha}^{۲}\hat{p}(۱-\hat{p})}{z_{\alpha\alpha}^{۲}\hat{p}(۱-\hat{p})+a^{۲}N}$$

Adottando le stesse convenzioni numeriche, già assunte in precedenza, si ottiene: $n=N/۱+a^{۲}N,$, risultato, come già detto, identico a quello ottenuto, operando in ambito quadratico.

Test

c. Test di ipotesi

Tale risultato può venire trasformato nel caso in cui si eseguano altri tipi di test, ad esempio, **test per la distribuzione χ^2**. Infatti nel caso di campionamento con ripetizione, ricordando che la varianza della varianza (se le osservazioni sono normali) ha espressione: $\hat{\sigma}^2_{\hat{\sigma}^2} = \hat{\sigma}^4 2n$, si ottiene:

$$\hat{\sigma}^2 + \chi^2_{sup}\frac{\hat{\sigma}^2}{\sqrt{2n}} - \left(\hat{\sigma}^2 - \chi^2_{inf}\frac{\hat{\sigma}^2}{\sqrt{2n}}\right) = 2a \qquad (2.19)$$

$$\left(\chi^2_{sup} + \chi^2_{inf}\right)\frac{\hat{\sigma}^2}{\sqrt{2n}} = 2a \qquad (2.20)$$

$$n = \frac{\left(\chi^2_{sup} + \chi^2_{inf}\right)^2 \hat{\sigma}^4}{8a^2} \qquad (2.21)$$

Dopodiché utilizzando la correzione per la popolazione finita, ricordando che:

$$\hat{\sigma}^2_{\hat{\sigma}^2} = \frac{\hat{\sigma}^4}{2n}\frac{N-n}{N-1} \qquad (2.22)$$

e sostituendo N a N-1, si ottiene:

$$\hat{\sigma}^2 + \chi^2_{sup}\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}} - \left(\hat{\sigma}^2 - \chi^2_{inf}\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}}\right) = 2a \qquad (2.23)$$

$$\left(\chi^2_{sup} + \chi^2_{inf}\right)\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}} = 2a \qquad (2.24)$$

$$2a\sqrt{N}\sqrt{2n} = \hat{\sigma}^2\left(\chi^2_{sup} + \chi^2_{inf}\right)\sqrt{N-n} \qquad (2.25$$

$$n = \frac{\hat{\sigma}^4 \left(\chi^2_{sup} + \chi^2_{inf} \right)^2 N}{8a^2 N + \hat{\sigma}^4 \left(\chi^2_{sup} + \chi^2_{inf} \right)^2 \sigma} \qquad (2.26)$$

Introducendo i gradi di libertà ν e ricordando che: $E(\chi^2) = \nu$ per cui: $\chi^2_{\sup} + \chi^2_{inf} \simeq 2\nu$ si ottiene un'espressione, dalla struttura formale uguale alla precedente:

$$n = \frac{\nu^2 N}{\nu^2 + 2a^2 N} \qquad (2.27)$$

la quale fornisce il valore atteso per la numerosità del campione, nel caso di test per la distribuzione $\chi 2$. La tripartizione cui si è accennato nella strategia di campionamento proposta costituisce l'approssimazione lineare di una curva logistica, avente due asintoti orizzontali pari ai due valori di soglia ed un punto intermedio di flesso a tangente inclinata, dove il coefficiente angolare della retta tangente coincide con il suddetto valore percentile. Sarebbe desiderabile poter presentare valori numerici generali per i parametri rappresentativi della curva logistica: ciò presenta alcune difficoltà, per la variabilità, da caso a caso, di tali parametri. Tuttavia sono evidenti alcune costanti, all'interno di ragionevoli intervalli di precisione richiesti. Si può osservare l'esistenza di un limite inferiore per la numerosità del campione, corrispondente ad un valore al di sotto del quale è necessario sottoporre a test tutti gli elementi della popolazione, e di un limite superiore, corrispondente ad un valore al di sopra del quale il numero degli elementi della popolazione da sottoporre a test è approssimativamente lo stesso. Si osservi tuttavia come aumentino, all'aumentare della precisione richiesta, sia il limite inferiore, sia quello superiore. Per quanto riguarda il coefficiente angolare *c (equazione 2.1)*, si può procedere per tentativi, a partire da un valore iniziale ottenuto dal confronto tra il valore di *K*, ascissa della curva logistica nel punto di flesso, ed il valore di *N*, corrispondente a *n=y(K).*

Da ultimo, ricollegandosi al problema delle stime di covarianza e alle tecniche di ammassamento o clumping, appare chiaro, come e perché la prima e più importane proprietà richiesta ad un campionamento sia la sua Bernoullianità, ovvero l'indipendenza fra dati campionati. Infatti la loro eventuale non – indipendenza invalida il noto decremento della varianza di campionamento, in funzione della numerosità del campione, per effetto delle correlazioni fra i dati stessi (limitatamente all'ambito lineare).

D'altra parte, l'assoluta garanzia della Bernoullianità di un campionamento è cosa assai difficile da essere assicurata, specialmente per grandi campioni, quando i dati sono raccolti in modo denso e/o fitto. Tutto ciò porta inevitabilmente a dover accettare un limite inferiore alla varianza delle stime delle statistiche, oggetto di studio, rendendo praticamente inutili campionamenti troppo estesi. Infatti per quanto raffinate siano metodologie e procedure di campionamento, innumerevoli sono le cause di connessione / correlazione fra le osservazioni di un campione (ambiente, tarature, altre strutture latenti).

Nei limiti di queste precisazioni, ricollegandosi specificamente al problema dell'analisi di varianza, un'utilissima strategia per ridurre, il più possibile, la varianza di campionamento è data dal campionamento stratificato o da quello a cluster. Infatti la stratificazione o clusterizzazione delle osservazioni, grazie al teorema di decompisizione ortogonale della varianza, permette di ridurre la varianza residua, mettendo in

evidenza una o più varianze spiegate, frutto del campionamento a strati o a cluster. Nel prosieguo, verrà illustrata, a titolo d'esempio, la riduzione della varianza residua della media campionaria. Infatti data la numerosità n del campione, acquisito con un campionamento semplice (o sistematico) di Bernoulli, non è possibile ridurre la varianza della media campionaria: $\sigma_{\mu}^2=\sigma_x^2/n$, se non aumentando la numerosità stessa.

Un'alternativa vantaggiosa, capace di diminuire detta varianza senza accrescere n , è data dallo schema di campionamento stratificato (o clusterizzato) di Poisson, dove la suddivisione in m strati (o cluster) risponde a caratteristiche di omogeneità di determinati attributi, opportunamente prescelti per la classificazione della popolazione in esame,

Nel caso di campionamento in blocco di una popolazione finita, essendo N_i la numerosità di ciascuno strato (e pertanto: $N=\sum N_i$, la numerosità della popolazione), il peso di ciascun strato (o cluster) è dato dal rapporto: $p_i=N_i/N$. La scelta più elementare è effettuare, strato per strato (o cluster per cluster) , il campionamento con una numerosità n_i direttamente proporzionale alla numerosità N_i dello strato (o del cluster): $n_i=nN_i/N=np_i$, avendo preliminarmente scelto e fissato la numerosità del campione: $n=\sum n_i$.

Come noto, il teorema di decomposizione ortogonale della varianza fissa l'identità fra la varianza generale σ_x^2 e la somma della varianza spiegata (dalla stratificazione o clusterizzazione) $\sigma_{\bar{x}}^2$ e della varianza residua $\bar{\sigma}_x^2$. Pertanto essendo la varianza residua sempre minore della varianza generale, la varianza della media campionaria: $\sigma_{\mu}^2=\bar{\sigma}_x^2/n$, di un campionamento stratificato (o clusterizzato) risulta certamente minore a quella ottenuta dal campionamento semplice (o sistematico). Infatti essendo, nell'ordine, la varianza generale, la varianza spiegata e la varianza residua:

$$\sigma_x^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \tag{2.28}$$

$$\sigma_{\bar{x}}^{2}=\frac{m}{m-1}\sum_{i=1}^{m}p_i\left(\bar{x}_i-\bar{x}\right)^{2} \tag{2.29}$$

$$\bar{\sigma}_x^{2}=\frac{(n-1)\sigma_x^{2}-(m-1)\sigma_{\bar{x}}^{2}}{n-m} \tag{2.30}$$

la varianza della media campionaria (nell'ipotesi ragionevole che si abbia: $n>>m$) risulta:

$$\sigma_{\mu}^2=\frac{\bar{\sigma}_x^2}{n}=\frac{(n-1)\sigma_x^2}{n(n-m)}-\frac{(m-1)\sigma_{\bar{x}}^2}{n(n-m)}<\frac{\sigma_x^2}{n} \tag{2.31}$$

e, tenuto conto del sopraccitato fattore di correzione per il campionamento finito, diventa:

$$\sigma_{\mu}^2=\frac{\bar{\sigma}_x^2}{n}\frac{N-n}{N-1}=\left(\frac{(n-1)\sigma_x^2}{n(n-m)}-\frac{(m-1)\sigma_{\bar{x}}^2}{n(n-m)}\right)\frac{N-n}{N-1}<\frac{\sigma_x^2}{n}\frac{N-n}{N-1} \tag{2.32}$$

Nel caso di campionamento con ripetizione di una popolazione finita o infinita, il peso di ciascuno strato (o cluster) è dato dal rapporto: $p_i=\sigma_i/\sigma_0$, essendo σ_i la misura della dispersione del generico strato (o cluster), ovvero il suo sqm, $\sigma_0=\sum p_i\sigma_i$, la media ponderata degli sqm (e: $\bar{\sigma}_x^2=\sum p_i\sigma_i^2$, la media ponderata delle varianze). Questo campionamento è da preferirsi anche per il caso di campionamento in blocco di una popolazione finita. Infatti il campionamento ottimale fa aumentare o diminuire la numerosità dello strato (o del cluster), dove più ampia o ristretta è la dispersione dei dati, mentre il campionamento proporzionale prende in considerazioni solo le numerosità degli strati (o dei cluster), prescindendo dalla loro dispersione. Di consegfuenza, una scelta più attenta è effettuare, strato per strato (o cluster per cluster), il campionamento con una numerosità n_i direttamente proporzionale alla dispersione σ_i dello strato (o del cluster), oltreché alla sua numerosità: $n_i=nN_i\sigma_i/(N\sigma_\square)=np_i$, avendo preliminarmente scelto e fissato la numerosità del campione: $n=\sum n_i$ (nel caso di campionamento con ripetizione di una popolazione finita o infinita, il rapporto: N_i/N , vale: $1/m$). Introducendo la varianza degli sqm:

$$\sigma_\sigma^2=\sum_{i=1}^{m} p_i\left(\sigma_i-\sigma_0\right)^2=\sum_{i=1}^{m} p_i\sigma_i^2-\sigma_0^2=\bar{\sigma}_x^2-\sigma_0^2 \qquad (2.33)$$

e supponendo una popolazione sufficientemente grande, la varianza della media campionaria risulta:

$$\sigma_\mu^2=\frac{\sigma_\sigma^2}{n}=\frac{\bar{\sigma}_x^2}{n}-\frac{\sigma_0^2}{n}<\frac{\bar{\sigma}_x^2}{n}<\frac{\sigma_x^2}{n} \qquad (2.34)$$

e, tenuto conto del sopraccitato fattore di correzione per il campionamento finito, diventa:

$$\sigma_\mu^2=\frac{\sigma_\varsigma^2}{n}\frac{N-n}{N-1}=\left(\frac{\bar{\sigma}_x^2}{n}-\frac{\sigma_0^2}{n}\right)\frac{N-n}{N-1}<\frac{\bar{\sigma}_x^2}{n}\frac{N-n}{N-1}<\frac{\sigma_x^2}{n}\frac{N-n}{N-1} \qquad (2.35)$$

Tanto più grande la varianza degli sqm, quanto più elevato è l'ulteriore guadagno, ottenuto con il campionamento ottimale, rispetto al campionamento proporzionale, mentre i due campionamenti coincidono nel caso in cui ogni strato (o cluster) ha la stessa dispersione.

3. TEST SEQUENZIALI

Indice Capitolo 3

a. Introduzione....1
b. Test su campioni numerosi....3
c. Test non – parametrici....8
d. Verso i test multipli....10
e. Test multipli su campioni normali....11
f. Test multipli non – parametrici....12
g. Test per i minimi quadrati....13

a. Introduzione

Nonostante ad oggi solo pochi autori abbiano dedicato la loro attenzione ai test sequenziali, questo tipo di analisi dei dati sperimentali non è certo un tema nuovo per la statistica. Infatti in letteratura, si trovano alcuni esempi di test sequenziali, tuttavia le applicazioni sono elaborate solo per una piccola parte dei test più comunemente utilizzati.

Poiché l'inferenza statistica[1] nasce come strumento utile per la verifica di ipotesi espressionete su l'adattamento di dati a distribuzioni di probabilità[2], su parametri o, più in generale, sulla bontà dei modelli d'interpretazione delle osservazioni[3], l'applicazione di un test necessita un campione di dati significativo e rappresentativo della popolazione da cui è considerato estratto. Questo fatto può comportare l'esigenza di collezionare un numero elevato di estrazioni, prima di poter espressionere ipotesi[4] sensate e soprattutto svolgere il test[5]. Il vantaggio dei test sequenziali è quello di permettere la loro esecuzione già con pochissimi dati (in alcuni casi già con tre dati), arrivando a soluzione con un numero minore di estrazioni rispetto ai test tradizionali.

Infatti il numero di osservazioni su cui condurre il test non è fissato a priori, ma è determinato nel corso dell'esperimento. Il test è così effettuato dopo ogni osservazione (o gruppo di osservazioni) sull'insieme dei dati accumulati, fino a quel momento, e prosegue fino a quando non è possibile decidere quale ipotesi accettare.

Poiché lo scopo di questo tipo di analisi è quello di arrivare a scegliere, tra ipotesi alternative, con il minimo numero di osservazioni, i test sequenziali sono costruiti in modo da poter rappresentare graficamente, passo dopo passo, i risultati delle osservazioni, in funzione del numero di prove effettuate. Benché le funzioni di merito che si possono costruire siano numerose, si è scelto di utilizzare l'approccio conosciuto come test sequenziale del rapporto di verosimiglianza, poiché esso è applicabile a tutte le tipologie di test senza nessun tipo di adattamento.

L'idea è ottenere un grafico su cui siano riconoscibili due linee di confine che lo suddividano in tre aree, nel caso comune di una sola ipotesi alternativa:

---

1 Le caratteristiche della popolazione complessiva sono indotte da quelle osservate su un campione estratto dalla popolazione stess.

2 $f(x_i)=P(X=x)$ i=1,2,... che ad ogni valore assunto dalla variabile aleatoria discreta X associata la corrispondente probabilita e' detta distribuzione di probabilità della variabile aleatoria X. $f\left(xi\right)\geq 0$

3 Uno degli scopi più importanti di un'analisi statistica è quello di utilizzare dei dati provenienti da un campione per fare inferenza sulla popolazione da cui è stato tratto il campione.

4 Per prima cosa si stabilisce l'ipotesi da sottoporre a test, detta ipotesi nulla, indicata con H_o. Oltre all'ipotesi nulla occorre specificare anche un'adeguata ipotesi alternativa, indicata con H_1, ossia un'affermazione che contraddice l'ipotesi nulla.

5 La statistica test è una statistica che viene calcolata dai dati del campione e può assumere tanti valori quanti sono i possibili campioni dalla popolazione, quindi il particolare valore calcolato dipende dal campione estratto.

- la regione di accettazione dell'ipotesi fondamentale H_0
- la regione (intermedia) del dubbio

la regione di accettazione dell'ipotesi alternativa H_1

Le linee di confine λ_0 e λ_1 sono ricavate in funzione dell'entità dei rischi α ((livello di significatività od errore di prima specie) ▪ e β ▪ (potenza del test od errore di seconda specie):

$$\lambda_0 = (1-\alpha)/\beta \tag{3.1}$$

$$\lambda_1 = \alpha/(1-\beta) \tag{3.2}$$

La funzione λ che interpreta i risultati delle prove, è detta rapporto di verosimiglianza di Fisher e vale:

$$\lambda = \frac{P(\underline{x} \Rightarrow \text{ se } H_0 \text{ è vera})}{P(\underline{x} \Rightarrow \text{ se } H_1 \text{ è vera})} \tag{3.3}$$

Se la funzione di distribuzione di probabilità è continua, il rapporto di verosimiglianza si fa tra le densità di probabilità composte, ovvero fra i prodotti delle probabilità elementari per campioni Bernoulliani:

$$\lambda = \prod_{i=1}^{n} p_0(x_i) / \prod_{i=1}^{n} p_1(x_i) \tag{3.4}$$

$$\ln \lambda = \sum_{i=1}^{n} \ln\left(p_0(x_i)\right) - \sum_{i=1}^{n} \ln\left(p_1(x_i)\right) \tag{3.5}$$

dove la forma logaritmica è particolarmente utile, se la funzione di distribuzione di probabilità adottata è di classe esponenziale.

Di norma, la funzione si muove inizialmente nella regione del dubbio, per poi dirigersi in

una delle due regioni di accettazione: a questo punto, il test può essere interrotto, perché è arrivato ad una soluzione e, in probabilità, sarà accettata l'ipotesi relativa alla regione interessata (un esempio in figura 2.1)

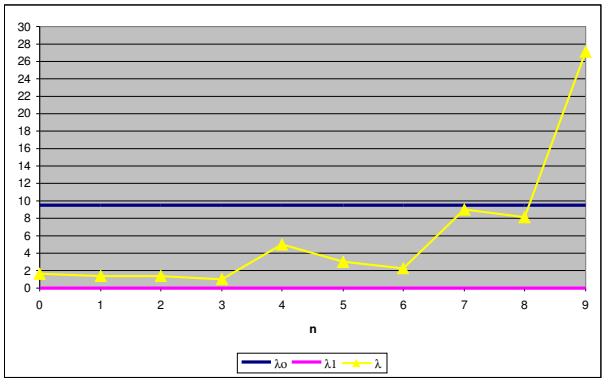


Fig. 2.1 – Un esempio di test sequenziale

Nel prosieguo, test per test (o per gruppi di test omogenei), è presentata l'applicazione dei test sequenziali, partendo dall'espressione del test tradizionale, fino ad arrivare ad ottenere il rapporto di verosimiglianza, avendo scelto e costruito due ipotesi alternative.

b. Test su campioni numerosi

I test parametrici semplici su campioni numerosi sottopongono a verifica d'ipotesi un solo parametro: la media o la differenza fra due medie, a varianza/e nota/e od incognita/e (considerata/e comunque stima/e corretta/e e consistente/i della/e varianza/e teorica/he, data la numerosità del/i campione/i). L'applicazione dei test sequenziali è molto semplice, in quanto utilizza l'espressione del test tradizionale, di volta in volta, inserendo la/e media/e ipotizzata/e nei due casi alternativi e calcolando, ad ogni passo di campionamento, la variabile casuale z nelle ipotesi H_0 e H_1. Dopodiché calcolate, passo dopo passo, le funzioni densità di probabilità normali nei due casi, il rapporto di verosimiglianza λ è rappresentato graficamente, in modo da poter essere messo a confronto direttamente con le rette di confine λ_0 e λ_1, finché non si arriva ad una soluzione:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \qquad (3.6)$$

$$z_1 = \frac{\bar{x} - \mu_1}{\sigma / \sqrt{n}} \tag{3.7}$$

da cui

$$\lambda = e^{-\frac{\square}{\square}\sum_{i=\square}^{n} z_{i,\square}^{\square}} / e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,1}^{2}} \tag{3.8}$$

$$z_0 = \frac{\Delta\bar{x} - \Delta\mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.9}$$

$$z_1 = \frac{\Delta\bar{x} - \Delta\mu_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.10}$$

da cui

$$\lambda = e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,0}^{2}} / e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,1}^{2}} \tag{3.11}$$

essendo solitamente nell'ipotesi fondamentale: $\Delta\mu_0 = 0$.

Test su campioni normali

Anche per questi test, effettuati su media, varianza e coefficiente di correlazione, si considerano due casi:

- test eseguito su un campione, verificando la rispondenza del parametro prescelto ad un valore di riferimento scelto e fissato;
- test eseguito su due campioni, verificando la corrispondenza (solitamente l'uguaglianza) dei rispettivi parametri.

Nel primo caso, è semplice costruire l'ipotesi fondamentale H_0 e quella alternativa H_1 , poiché si impone, in entrambi i casi, l'uguaglianza del parametro ad un valore di riferimento scelto e fissato. Pertanto il test sequenziale è costruito, passo dopo passo,

- andando a sostituire nella espressione del test il valore del parametro nelle due ipotesi;
- ricavando le variabili casuali t di Student per la media ($t=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$) e χ^2 per la varianza ($\chi^2=\frac{\nu\,\hat{\sigma}^2}{\sigma^2}$);
- calcolando le rispettive funzioni densità di probabilità;
- aggiornando il rapporto di verosimiglianza λ , come aggiornamento di quello calcolato al passo precedente;
- facendo il confronto rette di confine λ_0 e λ_1 .

Nel secondo caso, occorre modificare l'espressione dei test, evidenziando anche l'ipotesi alternativa, ovvero nei test sul confronto di medie, classico (di Gosset) con uguale varianza e di Welch con diversa varianza, si ha:

$$H_0 \;:\; \Delta\mu_0=0 \qquad t_0=\frac{\Delta\bar{x}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\Delta\bar{x}}{\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)\sigma_1^2+(n_1-1)\sigma_1^2}{n_1+n_0-2}}} \qquad t_0=\frac{\Delta\bar{x}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{3.12}$$

$$H_1 \;:\; \Delta\mu_1=\Delta \qquad t_1=\frac{\Delta\bar{x}-\Delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\Delta\bar{x}-\Delta}{\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}}} \qquad t_1=\frac{\Delta\bar{x}-\Delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{3.13}$$

dove nell'ipotesi fondamentale si è posto: $\Delta=0,$ eseguendo poi un test sequenziale parametrico per campioni indipendenti e normali, facendo uso della funzione densità di probabilità t di Student:

$$p(t)=f_0\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3.14}$$

Con:

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\left(\frac{\sigma_1^2}{n_1}\right)^2 \frac{1}{n_1+1} + \left(\frac{\sigma_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2 \qquad (3.15)$$

per il test di Welch

Come già detto in precedenza, non occorre aver accumulato un numero significativo di osservazioni, perché il test può essere eseguito già avendo due dati per campione. Infatti ad ogni passo n, si calcolano le medie dei due campioni e si confronta la loro differenza con il valore Δ scelto e fissato.

Nel test sul confronto di varianze, si ha invece:

$$H_0: \quad \sigma_0^2 = \sigma^2 \qquad \chi_{0,1}^2 = \frac{\nu_1 \hat{\sigma}_1^2}{\sigma^2} \qquad \chi_{0,2}^2 = \frac{\nu_2 \hat{\sigma}_2^2}{\sigma^2} \qquad F_0 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \qquad (3.16)$$

$$H_1: \quad \sigma_1^2 = k\sigma^2 \qquad \chi_{1,1}^2 = \frac{\nu_1 \hat{\sigma}_1^2}{\sigma^2} \qquad \chi_{1,2}^2 = \frac{\nu_2 \hat{\sigma}_2^2}{k\sigma^2} \qquad F_1 = \frac{k\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \qquad (3.17)$$

dove nell'ipotesi fondamentale si è posto: $k = 1$, eseguendo poi un test sequenziale parametrico per campioni indipendenti e normali, ricordando che le funzioni densità di probabilità delle variabili casuali χ^2 e F di Fisher sono rispettivamente:

$$p(\chi^2) = f_0 (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} \qquad (3.18)$$

$$p(F) = f_0 \left(\nu_2 F^{\frac{\nu_1}{2}-1} + \nu_1 F^{-\left(\frac{\nu_2}{2}+1\right)} \right) \qquad (3.19)$$

Il comportamento del rapporto di verosimiglianza λ, al crescere del numero dei dati n, è del tutto libero (infatti l'esempio illustrato in figura 2.2 mostra la conferma dell'ipotesi alternativa, contro quella fondamentale precedentemente ipotizzata), così come un test tradizionale è aperto a qualsiasi risultato. Inoltre anche tutte le ipotesi assunte sono proprio le stesse adottate per i corrispondenti test tradizionali.

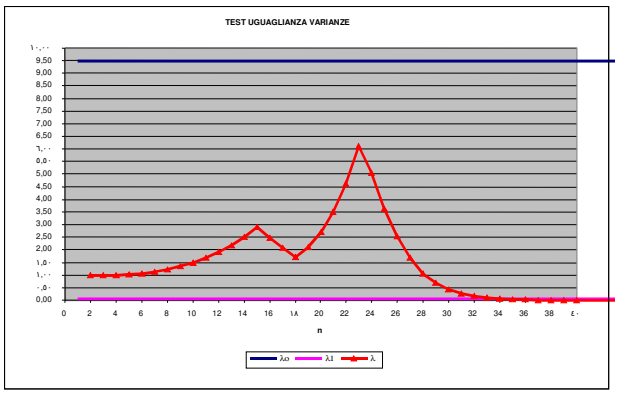


Fig. 2.2 – Un altro esempio di test sequenziale

Nel test sul confronto del coefficiente di correlazione, si ha:

$$H_0: \quad \rho_0 \qquad z_0 = \frac{Z - Z_0}{\sigma_Z} \tag{3.20}$$

$$H_1: \quad \rho_1 \qquad z_1 = \frac{Z - Z_1}{\sigma_Z} \tag{3.21}$$

essendo la trasformata Z di Fisher (per il coefficiente di correlazione):

$$Z = \frac{1}{2} * \ln \frac{1 + \hat{r}}{1 - \hat{r}} \tag{3.22}$$

$$Z_0 = \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0} \tag{3.23}$$

$$Z_1 = \frac{1}{2} \ln \frac{1 + \rho_1}{1 - \rho_1} \tag{3.24}$$

$$\sigma_Z^2 = \frac{1}{n-3} \tag{3.25}$$

dove nell'ipotesi fondamentale si pone spesso: $\rho_0 = 0$, eseguendo poi un test sequenziale parametrico per campioni normali[6], ricordando che anche la funzione densità di probabilità della variabile casuale z è normale.

c. Test non – parametrici

I test di rango di Mann – Whitney (per i valori centrali) e Siegel – Tuckey (per la dispersione) sono utilizzati rispettivamente per il confronto dei valori centrali o della dispersione di due campioni indipendenti, non necessariamente normali. Essi operano confrontando la somma dei ranghi di uno qualsiasi dei due campioni con la media teorica dei ranghi, standardizzata grazie alla varianza teorica dei ranghi stessi.

- Se i due campioni hanno valori centrali comparabili, il valore stimato $\hat{R}$ è vicino alla media teorica dei ranghi; in caso contrario, differisce per difetto od eccesso (e parimenti, in senso opposto, la somma dei ranghi dell'altro campione).
- Allo stesso modo, se hanno dispersione comparabile, il valore stimato $\hat{R}$ (a partire dalle differenze tra i valori argomentali di ogni campione con la rispettiva mediana) è prossimo alla media teorica dei ranghi.

Partendo da queste considerazioni e tenendo conto del limite centrale della statistica[7] (cui si può fare riferimento trattandosi di somme di valori – i ranghi – indipendenti ed equiponderati, benché di distribuzione incognita), si arriva a costruire l'espressione che permette di eseguire i test sequenziali con le due ipotesi alternative, modificando il denominatore della media teorica dei ranghi.

$$z_0 = \frac{\hat{R} - \bar{R}_0}{\sigma_R} \tag{3.26}$$

$$z_1 = \frac{\hat{R} - \bar{R}}{\sigma_R} \tag{3.27}$$

essendo

[6] Si definisce distribuzione di campionamento di una data statistica la distribuzione di tutti i possibili valori che possono essere assunti dalla statistica stessa, calcolati da campioni casuali della stessa dimensione estratti dalla stessa popolazione.

[7] Teorema del limite centrale Sia data una popolazione avente media μ e varianza σ, e da essa si estraggano campioni casuali di ampiezza n; indicando con X la media campionaria, la variabile $\frac{-\mu}{\frac{\sigma}{\sqrt{n}}}$ è una variabile aleatoria la cui distribuzione tende alla distribuzione normale standardizzata per $n \to \infty$.

$$R=\frac{n_1\left(n_1+n_2+1\right)}{A} \tag{3.28}$$

con $A=۲$ in H_0

Infatti nell'ipotesi alternativa, la media teorica è costruita come se si sapesse a priori che i ranghi del campione prescelto siano sistematicamente più piccoli o più grandi di quelli dell'altro.

I test di segno di Thompson (per i valori centrali e per la dispersione) sono utilizzati rispettivamente per il confronto dei valori centrali o della dispersione di due campioni qualsiasi. Con ragionamenti analoghi a quelli esposti per i test di rango, si arriva a costruire anche l'espressione che permette di eseguire i test sequenziali con le due ipotesi alternative:

$$z_{\cdot}=\frac{\hat{f}^{(+)}-\cdot.۵}{\frac{\cdot.۵}{\sqrt{\hat{f}^{(+)}+\hat{f}^{(-)}}}} \tag{3.29}$$

$$z_1=\frac{\hat{f}^{(+)}-A}{\frac{0.5}{\sqrt{\hat{f}^{(+)}+\hat{f}^{(-)}}}} \tag{3.30}$$

Infatti nell'ipotesi alternativa, la media teorica è ancora costruita come se si sapesse a priori che i ranghi del campione prescelto siano sistematicamente più piccoli o più grandi di quelli dell'altro.

Il test sul coefficiente di correlazione sui ranghi di Spearman si sviluppa invece proprio come il test classico sul coefficiente di correlazione, calcolato sui valori argomentali, in quanto identica è la distribuzione normale di comportamento della variabile casuale z, ottenuta dalla trasformata Z di Fisher dei suddetti coefficienti:

$$\hat{r}_S=1-\frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\Delta_i^2 \tag{3.31}$$

$$Z=\frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}} \tag{3.32}$$

$$z_0 = \frac{Z - Z_0}{\sigma_Z} \tag{3.33}$$

$$z_1 = \frac{Z - Z_1}{\sigma_Z} \tag{3.34}$$

dove tutti gli addendi Δ_i sono ottenuti come differenze puntuali fra i ranghi, assegnati separatamente alle due componenti del campione dato, ed i termini ($Z_0 Z_1 \sigma_Z^2$) gli stessi già definiti in precedenza.

d. Verso i test multipli

I test di buon adattamento sono eseguiti per verificare, se i dati di un campione possono essere interpretati come estratti o meno da una popolazione di distribuzione nota; i test d'indipendenza per verificare, se le componenti di un campione sono indipendenti fra loro o no.

Innanzitutto si presentano i test per una frequenza (p_0 o p_1), due frequenze ($\Delta p_0 = 0$ e $\Delta p_0 = \Delta$) ed una contingenza ($c_0 = f - pq$ e $c_1 = f - kpq$), precisando che le frequenze possono essere rilevate in un dominio di qualsiasi dimensione e la contingenza può essere calcolata utilizzando una frequenza doppia (come nella relazione presentata), oppure multipla:

$$z = \frac{\hat{f} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.35}$$

$$z = \frac{\Delta \hat{f} - \Delta p}{\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}} \tag{3.36}$$

$$z = \frac{\hat{c}}{\sqrt{\frac{pq(1-pq)}{n}}} \tag{3.37}$$

Dopodiché i test di buon adattamento ed indipendenza consistono nel confrontare le frequenze del

campione rispettivamente con la funzione densità di probabilità di una popolazione nota ed il prodotto delle frequenze marginali, dove l'ipotesi alternativa è formulata, classe per classe, come esposto appena sopra:

$$\chi^2_{m-h-1} = n \sum_{i=1}^{m} \frac{\left(\hat{f}_i - p_i\right)^2}{p_i} = \sum_{i=1}^{m} \frac{\left(\hat{F}_i - np_i\right)^2}{np_i} \qquad (3.38)$$

$$\chi^2_{(n-1)(m-1)} = n \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\left(\hat{f}_i - \hat{p}_i \hat{q}_j\right)^2}{p_i q_j} = \frac{1}{n} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\left(n\hat{F}_{ij} - \hat{P}_i Q_j\right)^2}{\hat{P}_i \hat{Q}_j} \qquad (3.39)$$

essendo h il numero dei parametri di disturbo ed i termini ($p_k \quad q_l$) gli stessi già definiti in precedenza.

Analogamente nei test più potenti di Kolmogorov – Smirnov[8] (di buon adattamento ed indipendenza) e di Pearson[9] et a. (per la normalità), occorre mettere in evidenza le ipotesi alternative rispettivamente sulla distribuzione di una popolazione (P_0 o P_1) dalla quale dati campionari possono essere interpretati come estratti o meno e sull'indipendenza o no delle componenti di un dato campione ($\left(\left[p\right]\left[q\right]\right)_0$ .o $\left(\left[p\right]\left[q\right]\right)_1$:

$$D_n = \max|\hat{t}_i - P_i| \qquad D_n = \max|\hat{t}_{ij} - \left[p_i\right]\left[q_j\right]|$$

$$\chi^2_2 = \frac{(\hat{\gamma} - A)^2}{6/N} + \frac{(\hat{\beta} - B)^2}{24/N} \qquad (3.40)$$

dove nell'ipotesi fondamentale, per il test di Pearson et al., si impone ($A=0$ e $B=3$), in conformità alla forma simmetrica e normocurtica della distribuzione normale.

e. Test multipli su campioni normali

I test multipli (su campioni normali) per l'analisi di varianza verificano l'ipotesi di uguaglianza delle medie parziali fra loro e con la media generale, con il test di Fisher nell'ipotesi di uguale varianza fra i campioni raccolti e con il test di Welch nell'ipotesi di diversa varianza:

[8] Il test di Kolmogorov-Smirnov è un test non parametrico che verifica la forma delle distribuzioni campionarie. È applicabile a dati per lo meno ordinali. Nella sua formulazione esatta prevede che le variabili siano continue. Non richiede di per sé alcuna ipotesi sulla distribuzione campionaria (salvo nel caso a un campione, in cui viene testata una distribuzione a propria scelta).

[9] Il test chi quadrato di Pearson (o della bontà dell'adattamento) è un test non parametrico applicato a grandi campioni quando si è in presenza di variabili nominali e si vuole verificare se il campione è stato estratto da una popolazione con una predeterminata distribuzione o che due o più campioni derivino dalla stessa popolazione.

$$F_{\nu_s \nu_r} = \frac{k\sigma_s^2}{\sigma_r^2} \qquad (3.41))$$

$$\upsilon_s = n-1 \qquad \nu_r = n(m-1) \qquad (1D)$$

$$\upsilon_s = n-l \qquad \nu_r = (m-l)(n-l) \qquad (2D)$$

$$\upsilon_s = n-1 \qquad \nu_r = \left(\sum_{j=1}^{n} \frac{\sigma_j^2}{m_j} \right)^2 / \sum_{j=1}^{n} \left(\frac{\sigma_j^2}{m_j} \right)^2 \frac{1}{m_j - 1} \qquad (3.42)$$

per il test di Welch

Inoltre il test multiplo di Bartlett per lo studio delle componenti della varianza verifica l'uguaglianza fra la varianza di una certa variabile casuale e le varianze di date variabili statistiche che si suppongono campioni estratti da una popolazione costituita dalla variabile casuale stessa:

$$\chi^2_{\nu=n-1} = -2\ln\Lambda \qquad \Lambda = \frac{k \prod_{j=1}^{n} \left(\sigma_j^2 \right)^{(m_j/2)}}{\left(\sum_{j=1}^{n} m_j \sigma_j^2 / \sum_{j=1}^{n} m_j \right)^{\sum_{j=1}^{n} m_j/2}} \qquad (3.43)$$

Analogamente il test multiplo di Hotelling per lo studio della struttura di covarianza verifica l'incorrelazione delle componenti di una certa variabile casuale multidimensionale, accertando l'annullarsi di opportuni indicatori della correlazione fra le componenti di una data variabile statistica multidimensionale che si suppone un campione estratto da una popolazione costituita dalla variabile casuale stessa:

$$\chi^2_{\nu=n(n-1)/2} = -2\ln\Lambda \qquad \sigma_{x_j}^2)^{m/2} \qquad (3.44)$$

$$\Lambda = \frac{k \left(detC_{xx} \right)^{m/2}}{\left(\prod_{j=1}^{n} \sigma_j^2 \right)^{m/2}} \qquad (3.45)$$

dove nell'ipotesi fondamentale si impone sempre: $k = 1$.

f. Test multipli non – parametrici

I test multipli non – parametrici eseguono l'analisi di varianza e lo studio delle componenti della

varianza, con il test di Kruskal - Wallis per campioni indipendenti e con il test di Friedman per campioni qualsiasi; mentre lo studio della struttura di covarianza è eseguito con il test di test di Wilcoxon – Wilcox modificato secondo Lawley:

$$\chi^2_{\nu=n-1}=\frac{A}{N(N+1)}\sum_{j=l}^{n}\frac{R_j^2}{m_j}-3(N+l) \qquad (3.46)$$

test di Kruskal – Wallis

$$\chi^2_{\nu=n-1}=\frac{A}{mn(n+1)}\sum_{j=1}^{n}R_j^2-3m(n+1)=\frac{B\sum_{j=1}^{n}(R_j-\bar{R})^2}{\sum_{j=1}^{n}R_j} \qquad (3.47)$$

test di Friedman

$$\chi^2_{\nu=n(n-1)/2}=\left(m-1-\frac{2n+5}{6}\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(r_{ij}-C\right)^2 \qquad (3.48)$$

test Wilcoxon – Wilcox / Lawley

dove nell'ipotesi fondamentale si impone ($A=12$, $B=6$ e $C=0$). Infine si noti, come il test di Lawley possa essere utilizzato anche per campioni normali, in sostituzione del test di Hotelling.

Un commento conclusivo di questa lunga collezione di test sequenziali, semplici e multipli, parametrici (o della normalità) e non – parametrici, rileva come la maggiore difficoltà, di trasformare un test tradizionale in un test sequenziale, stia spesso nella modalità adottata per inserire l'informazione legata all'ipotesi alternativa. Pertanto nei casi in cui non sia possibile costruire l'ipotesi alternativa con una semplice traslazione, una soluzione ragionevole è inserire un parametro, come una sorta di peso, con lo scopo di amplificare o smorzare il quoziente dello stimatore prescelto di cui calcolare la probabilità composta. Come già detto in precedenza, proprio questa probabilità composta, volutamente alterata rispetto all'ipotesi fondamentale, messa a quoziente con la probabilità composta, calcolata secondo la suddetta ipotesi fondamentale, dà il rapporto di verosimiglianza utilizzato per concludere il test sequenziale, mediante il confronto fra il valore atteso e le rette di confine, dipendenti dal livello di significatività e dalla potenza del test.

g. Test per i minimi quadrati

I test per i minimi quadrati sono strumenti importanti, utili alla validazione dei dati e dei modelli. Infatti alcuni di questi test accolgono o respingono insiemi di parametri e li confrontano con valori di riferimento o con valori precedenti. Altri test sono invece rivolti alla individuazione ed eliminazione dei dati anomali, in particolare errori accidentali, grossolani e sistematici che si possono commettere in relazione a varie cause, oppure possono dipendere da svariate sorgenti. Altri errori, ancora più sofisticati, sono detti errori di modello; fra questi: difetti di linearizzazione, malcondizionamento della configurazione dei

parametri, inaffidabilità dello schema di misura, cattiva conoscenza dei pesi delle osservazioni, presenza di correlazioni fra le stesse. Allora riuscire a valutare correttamente, sulla base dei dati, se e quanto un modello costruito sia accurato, preciso, affidabile e robusto, è uno strumento indispensabile e fondamentale per condurre a buon fine la suddetta validazione.

In particolare, i test globali di autoconsistenza e di crossvalidazione sono comunque test parametrici di verifica di ipotesi sulla varianza. Infatti come per questi ultimi, l'applicazione dei test sequenziali consiste, nel caso di confronto di ipotesi con un valore di riferimento, nell'esecuzione del test con le due ipotesi fondamentale ed alternativa, mentre nel caso di test sul rapporto di varianze, si introduce nel quoziente un fattore k che permette di esprimere l'ipotesi alternativa. Modalità analoghe sono altresì adottate nei test globali, parziali e locali di significatività dai parametri.

Infine nei test per l'identificazione e l'eliminazione dei dati anomali, ricordando che il test globale è dato dal test di Pearson et al. (cfr. verso i test multipli), tanto il test di Thompson per una selezione all'indietro, cosiddetto "data snooping", quanto il test estremale di Hawkins per una selezione in avanti ($H_e = \max\left(k\,\hat{v}_i^2 / \nu\sigma_{v_i}^2\right)$), dopo l'applicazione di procedure robuste, richiedono l'introduzione di un "peso" k , per formulare correttamente anche l'ipotesi alternativa, mentre nell'ipotesi fondamentale si impone sempre: $k=1$.

Alcune considerazioni conclusive sull'utilizzo dei test sequenziali rilevano:

- La rapidità con cui si arriva a determinare una soluzione è fortemente dipendente dal parametro stimato: in generale, si arriva a soluzione in un numero minore di passi, se si verifica un parametro di centro, rispetto ad un valore di dispersione, oppure e peggio ad un indice di dipendenza, in accordo con la diversa consistenza delle stime.
- Detta rapidità dipende, anche da come si scelgono le ipotesi fondamentale ed alternativa: quanto più vicine fra loro sono le due ipotesi, tanto meno velocemente si arriva ad una soluzione (del resto, è ben noto come sia fortemente sconsigliato verificare in alternativa ipotesi vicine anche con i test tradizionali).
- Inoltre la scelta dell'ipotesi alternativa deve avvenire dopo l'esame del comportamento delle osservazioni, in modo da fissare un'alternativa sensata che renda credibile il risultato. In altre parole, un'ipotesi alternativa che contraddica il comportamento delle osservazioni, se queste si discostano da quanto previsto con l'ipotesi fondamentale, porterebbe comunque ad accettare l'ipotesi fondamentale, anche se poco plausibile.

Infine è bene a sottolineare nuovamente, come un test sequenziale giunga a soluzione con un numero di osservazioni sempre inferiore rispetto a quelle necessarie per passare un test tradizionale, come mostrato in figura 2.3.

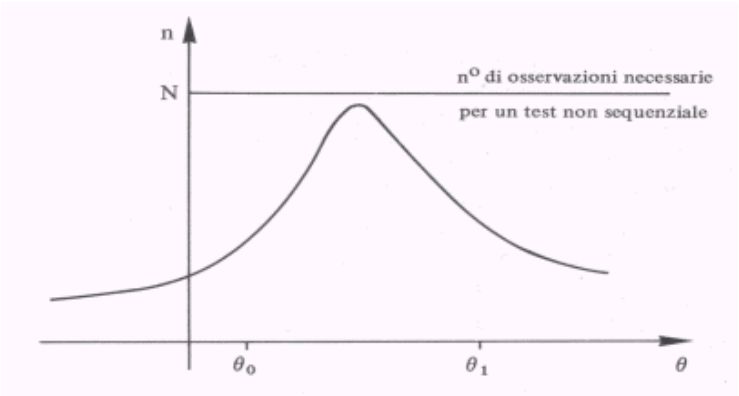


Fig. 2.3 – Andamento del numero medio di osservazioni richieste da un test sequenziale

4. METODI DI FOURIER

Indice Capitolo 4

a. Riassunto....1
b. Classificazione dei segnali....2
c. Elementi della serie di Fourier....3
c.1. Serie di Fourier trigonometriche....3
c.2. Serie di Fourier esponenziali....4
d. Trasformata di Fourier....5
d.1. Esistenza della Trasformata di Fourier....5
d.2. Proprietà della trasformata di Fourier....5
d.3. Teorema di Parseval....7
d.4. Trasformata di Fourier e funzione delta....7
d.5. Teorema del campionamento di Nyquist....10
d.5.1. Trasformata di Fourier Discreta (DFT)....10
d.6. Fast Fourier Transform (FFT)....11
d.6.1. Limitazioni della FFT....12
d.7. Dispersione spettrale....12
d.8. Aliasing....12
d.9. Risoluzione spettrale....13
d.10. Analisi di sequenze stocastiche....13
d.11. Analisi di segnali reali....14
d.12. Short Time Fourier Transform (STFT)....14
d.13. Wavelet Transform (WT)....15
d.13.1. Analisi Wavelet....15
d.14. La trasformata continua wavelet (CWT)....16

a. Riassunto

I metodi di Fourier [1] raggruppano una serie di tecniche matematiche che vanno dalla serie di Fourier alla sua trasformata (trasformata di Fourier) e possono estendersi alla trasformata Wavelet (altrimenti dette: ondine). Il loro campo d'elezione è quello dell'analisi matematica e questa osservazione spiega il perché di una certa loro relativa marginalità nell'ambito del trattamento delle osservazioni, dove la statistica è lo strumento principe. D'altra parte, in particolare oggigiorno con gli attuali mezzi digitali, un'importante branca della statistica rivolge la sua attenzione all'analisi numerica ed alla statistica computazionale, prendendo in considerazione metodi numerici che il calcolo numerico ha derivato non solo dall'algebra, ma anche proprio dall'analisi matematica. Allora questo fatto giustifica l'attenzione rivolta ai metodi di Fourier anche nell'ambito del trattamento delle osservazioni, arricchendo la collezione dei metodi offerti alla geostatistisca ed alla geomatica, nonché ad altre discipline (di carattere ingegneristico e non) attente ad un approccio simile ai problemi d'interesse.

Andando oltre gli aspetti matematici e la messa in evidenza dei legami tra le diverse parti della matematica e della matematica applicata, è interessante notare i contributi apportati dai metodi di Fourier alle discipline del rilevamento, storicamente in geodesia e cartografia, ma recentemente anche in geostatistica e geomatica. Infatti ad esempio, il geomonitoraggio, l'analisi d'immagine e la costruzione dei DTM (o modelli digitali del terreno), operazioni tipiche rispettivamente del controllo di movimenti e deformazioni, della fotogrammetria, della foto-interpretazione e del telerilevamento, e della cartografia numerica (ovvero della

[1] Pubblicato integralmente in Fava G., Mussio L. (2010): Metodi di Fourier. Rivista dell'agenzia del Territorio, anno X, n. 1, 2010.

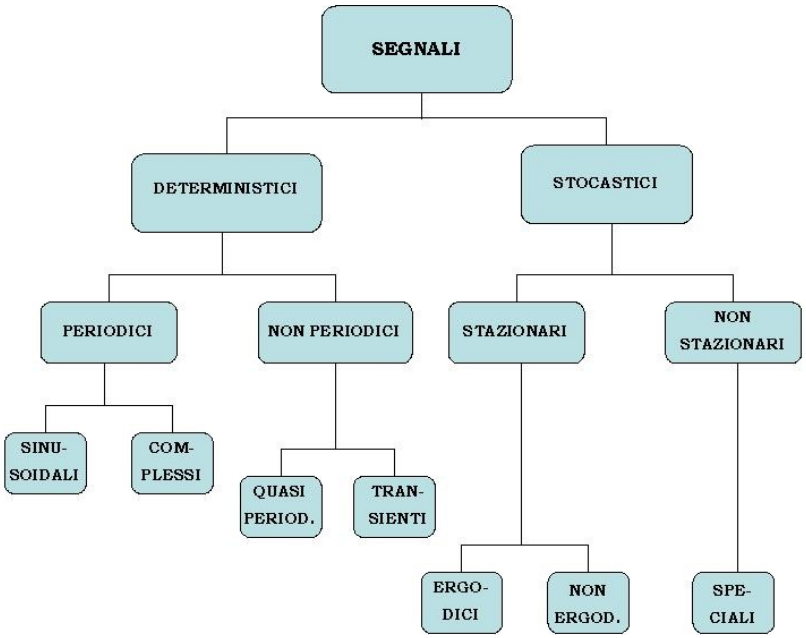
costruzione di DB topografici) possono vantaggiosamente avvalersi dei metodi di Fourier. Invece per quanto riguarda la geodesia e la cartografia, i metodi di Fourier sono impiegati rispettivamente per derivare le funzioni armoniche sferiche di Legendre (per lo studio del campo della gravità terrestre) e come approssimazione della soluzione di funzioni di equazioni differenziali alle derivate parziali delle rappresentazioni cartografiche (piane) dell'ellissoide di rotazione terrestre.

Approcci simili sono altresì adottati, ad esempio, nell'ambito della fisica terrestre, della geofisica applicata, dell'ingegneria strutturale, della meccanica applicata, di quella dei fluidi e dell'idraulica, dell'aerodinamica e della meccanica del volo. In particolare, la geofisica applicata e l'ingegneria strutturale sono solite rivolgere parte della loro attenzione alle condizioni dinamiche del sottosuolo e/o delle strutture. Questo studio avvicina queste discipline a quelle del rilevamento, completando lo studio di movimenti e deformazioni, proprio del geomonitoraggio. Infatti come questo, esso può vantaggiosamente avvalersi dell'apporto fornito dai metodi di Fourier che, qui come altrove, trovano la loro naturale collocazione, in parallelo al metodo degli elementi finiti, all'approccio fornito dai processi stocastici (assumendo che i dati siano una sequenza estratta da una realizzazione di un qualche processo stocastico), ecc. Da ultimo, è opportuno rilevare come la relativa vicinanza con tecniche tipiche dell'analisi statistica giustifichi l'inserimento dei metodi di Fourier, qui esposti, tra le metodologie e le procedure proposte dal trattamento delle osservazioni per l'analisi dei dati.

b. Classificazione dei segnali

In generale, due classi principali di segnali possono essere distinte: segnali continui e segnali discreti. I segnali continui sono descritti da una funzione continua che fornisce per ogni istante di tempo informazioni sul segnale stesso. I segnali discreti sono descritti da una sequenza che fornisce informazioni per punti discreti sull'asse del tempo.

A riguardo, si noti che il significato di tempo, benché usualmente scontato, non è esclusivo. Infatti a rigore, qualsiasi parametro uni-dimensionale può essere utilizzato in sua vece.



Classificazione dei segnali in base alle loro caratteristiche.

Dato un segnale $s(t)$, il campionamento di una sequenza avviene attraverso la seguente operazione:

$$s(m)=s(t)|_{t=mT_s} \tag{4.1}$$

dove $m\in Z$, T è l'intervallo di campionamento ed $f_s=1/T_s$ è la frequenza di campionamento.

Inoltre i segnali si distinguono in segnali deterministici e segnali stocastici. La trattazione completa dello studio di segnali stocastici è considerata estranea agli scopi del presente lavoro. I segnali deterministici possono essere integralmente descritti sia dal punto di vista matematico che grafico. Data la presenza di rumore aggiuntivo, i segnali in natura non sono quasi mai deterministici, ma può risultare è opportuno approssimare o modellare il segnale tramite funzioni deterministiche. Tra i segnali deterministici si distinguono quelli periodici che possono essere espressi come:

$$s(t)=s(t+nT) \tag{4.2}$$

con $n\in Z$, e T periodo del segnale. Il segnale periodico è costituito da una forma d'onda base avente durata pari a T che si ripete un infinito numero di volte. Altri segnali sono invece non-periodici e, tra questi, si distinguono quelli quasi – periodici e quelli transitori.

c. Elementi della serie di Fourier

Le espansioni in serie sono spesso utilizzate per semplificare i calcoli con valide approssimazioni o per esprimere una funzione localmente o globalmente. In particolare, in prossimità di un determinato punto, l'espansione in serie di Taylor permette di esprimere una funzione come un semplice polinomio in modo da fornire un'espressione accurata da applicare ad un dominio sempre più vasto al crescere dell'ordine del polinomio. Diversamente data una determinata funzione periodica, l'espansione in serie di Fourier separa la funzione in sinusoidi di diverse frequenze. Il miglior adattamento consiste nella minimizzazione dell'errore che, per una funzione h e la sua approssimazione h^a , è misurata nell'intero dominio come:

$$E=\int (h(t)-h^a(t))^2 dt \tag{4.3}$$

Le espansioni in serie di Fourier sono espresse sia in forma esponenziale che in forma trigonometrica:

c.1. Serie di Fourier trigonometriche

L'espansione in serie di Fourier, h^a è la combinazione lineare di funzioni {cos(*nt*)} e {sin(*nt*)}. In particolare,

$$h(t)=\frac{A_0}{2}+\sum_{n=1}^{\infty}(A_n\cos(nt)+B_n\sin(nt)) \tag{4.4}$$

Dove i coefficienti A_0 , A_n , B_n sono espressi come:

$$A_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} h(t)\,dt$$
$$A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(t)\cos(nt)\,dt \qquad (4.5)$$
$$B_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(t)\sin(nt)\,dt$$

I coefficienti A_n e B_n indicano l'influenza delle componenti armoniche di periodo $T = 2\pi/n$.

c.2. Serie di Fourier esponenziali

La serie di Fourier esponenziale è espressa mediante la formula di Eulero:

$$e^{\text{int}} = \cos nt + i\sin nt \qquad (4.6)$$

dove:

$$\cos nt = \frac{e^{\text{int}} + e^{-int}}{2} \qquad (4.7)$$

e

$$\sin nt = \frac{e^{\text{int}} - e^{-\text{int}}}{2\text{i}} \qquad (4.8)$$

Applicando queste ultime due espressioni alla forma trigonometrica delle serie di Fourier si ottiene:

$$h(t) = \sum_{n=1}^{\infty} C_n e^{in\frac{2\pi}{T}t} \qquad (4.9)$$

Con

$$C_n = \frac{1}{T}\int_{-T/2}^{T/2} h(t)\cdot e^{-in\frac{2\pi}{T}t}\,dt \qquad (4.10)$$

Condizioni sufficienti per la convergenza della serie di Fourier sono date dalle condizioni di Dirichlet:

1. $h(t)$ può avere solo un numero finito di discontinuità;
2. $h(t)$ può avere solo un numero finito di valori estremi.

Le serie di Fourier sono la base della trasformata di Fourier che si ottiene trasformando C_n da una variabile discreta ad una variabile continua ponendo $T \to \infty$ e $n/L \to f$.

d. Trasformata di Fourier

Un processo fisico può essere descritto sia nel dominio del tempo, come una funzione $h(t)$, che nel dominio delle frequenze, specificando il processo tramite la definizione della sua ampiezza $H(f)$, con $-\infty < f < +\infty$. Poiché $h(t)$ e $H(f)$ sono due diverse rappresentazioni della stessa funzione, è possibile alternare le due rappresentazioni mediante le equazioni della trasformata di Fourier [2]:

$$\int_{-\infty}^{\infty} |h(t)|\, dt < \infty \qquad (4.13)$$

ed abbia, al più, un numero finito di discontinuità finite. Questo requisito è generalmente soddisfatto per i segnali reali, ma può essere violato se:

1. la funzione ha energia infinita (come ad esempio, le funzioni di Eulero o quelle periodiche);
2. la funzione contiene impulsi od altre funzioni limite.

Entrambe le eccezioni sono utili per applicazioni ingegneristiche: considerando la trasformata di Fourier come limite di una sequenza per la quale esiste una trasformazione che permette l'applicazione dell'analisi di Fourier a molti di questi casi.

d.2. Proprietà della trasformata di Fourier

La trasformata di Fourier è un'operazione lineare che segue le proprietà di omogeneità e di additività. La proprietà di omogeneità consiste nel fatto che la trasformata di una costante posta a moltiplicare una determinata funzione è la stessa costante moltiplicata per la trasformata della funzione. La proprietà di additività è determinata dal fatto che la trasformata della somma di diverse funzioni è uguale alla somma delle trasformate delle singole funzioni.

[2] L'operazione di trasformata è involutoria.

Nel dominio del tempo, la funzione $h(t)$ può presentare particolari simmetrie, come essere puramente reale o puramente immaginaria, altrimenti pari: $h(t)=h(-t)$, o dispari: $h(t)=-h(-t)$. Nel dominio della frequenza, queste simmetrie portano a relazioni tra $H(f)$ e $H(-f)$. La seguente tabella fornisce la corrispondenza tra le simmetrie nei due domini:

Se:		allora :
$h(t)$	è reale	$H(-f)=[H(f)]$
$h(t)$	è immaginario	$H(-f)=-[H(f)]$
$h(t)$	è dispari	$H(f)$ è dispari
$h(t)$	è pari	$H(f)$ è pari
$h(t)$	è reale e dispari	$H(f)$ è reale e dispari
$h(t)$	è reale e pari	$H(f)$ è reale e pari
$h(t)$	è immaginario e dispari	$H(f)$ è immaginario e dispari
$h(t)$	è immaginario e pari	$H(f)$ è immaginario e pari

dove il simbolo (*) indica la funzione complessa coniugata. Queste simmetrie sono utilizzate al fine di aumentare l'efficienza computazionale. Inoltre le proprietà della trasformata di Fourier permettono di definire il passaggio matematico da un dominio ad un altro. Di seguito sono indicate alcune proprietà elementari della trasformata di Fourier (dove il simbolo $\Leftrightarrow$ indica la coppia trasformata).

Se $h(t)\Leftrightarrow H(f)$ è una coppia, allora le altre coppie trasformate sono:

$h(at)\Leftrightarrow\frac{1}{\lvert a\rvert}H\left(\frac{f}{a}\right)$	scalatura temporale
$\frac{1}{\lvert b\rvert}h\left(\frac{t}{b}\right)\Leftrightarrow H(bf)$	scalatura di frequenza
$h(t-t_0)\Leftrightarrow H(f)\exp(2\pi ift_0)$	traslazione nei tempi
$h(t)\exp(2\pi itf_0)\Leftrightarrow H(f-f_0)$	traslazione nelle frequenze
$h(t)*g(t)\Leftrightarrow H(f)G(f)$	convoluzione (operatore)

Per quanto riguarda la proprietà di scalatura temporale, se si riduce la larghezza della funzione mantenendo costante la sua altezza si ricava uno spettro di Fourier più ampio e basso. Al contrario, un

aumento della larghezza della funzione rende lo spettro più stretto e più alto. Simili proprietà sono date dalla proprietà di scalatura di frequenza. Inoltre se $a=-1$ o $b=-1$:

$$h(-t)\Leftrightarrow H(-f) \tag{4.14}$$

Per quanto riguarda la proprietà di traslazione nei tempi, la trasformata di Fourier di una funzione traslata consiste nella trasformata della funzione non traslata moltiplicata per un fattore esponenziale la cui fase è lineare. Proprietà analoghe si osservano nella proprietà di traslazione nelle frequenze.

d.3. Teorema di Parseval

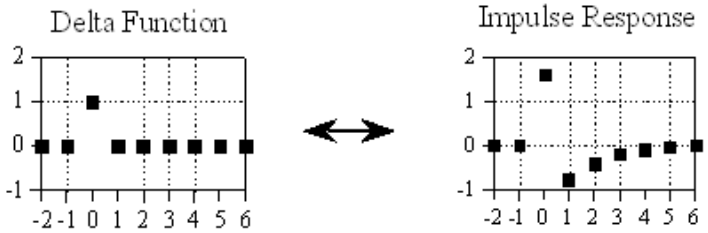
La potenza di un segnale rappresentato da una funzione $h(t)$ è la stessa, se calcolata a partire dal segnale nel dominio del tempo, oppure a partire dalla trasformata nel dominio di frequenza

$$h^2(t)\cdot dt=|H(f)|^2\cdot df \tag{4.15}$$

Dove la potenza dello spettro $P(f)$ è data da $P(f)=|H(f)|^2$.

d.4. Trasformata di Fourier e funzione delta

A causa della linearità, si può applicare il principio di sovrapposizione degli effetti: Il segnale su cui deve essere valutata la trasformata è scisso in semplici componenti additive ognuna delle quali è trattata individualmente, per riunire i risultati solo alla fine. Il segnale è generalmente decomposto in molteplici impulsi normalizzati: segnali i cui valori sono tutti nulli ad ccezione di un unico punto il cui valore risulta unitario. Questo processo è descritto matematicamente da una convoluzione. Nella figura sottostante sono definiti l'impulso unitario o funzione delta: $\delta(n)$, e l'impulso di risposta: $h(t)$.



Funzione delta: $\delta(n)$ **,** e impulso di risposta: $h(t)$, [Modificato da Smith].

Quando si calcola la trasformata di Fourier della funzione δ si ricava:

$$\Delta(f)=\int_{-\infty}^{\infty}\delta(t)\exp(-2\pi ift)\,dt=\exp(-2\pi ift)\big|_{t=0}=1 \tag{4.16}$$

Dai segnali continui ai segnali discreti

La maggior parte dei segnali direttamente incontrati nelle scienze ed in ingegneria sono continui e, al fine di consentire ai computer digitali di interagire con essi, è necessaria una conversione.

L'informazione digitale è diversa dal quella continua (o analogica), dal momento che è campionata e quantizzata. Il campionamento converte la variabile indipendente (ad esempio, il tempo) da continua a discreta. La quantizzazione converte la variabile dipendente (il segnale) da continua a discreta. Sia il campionamento e che la quantizzazione limitano le informazioni del segnale digitale.

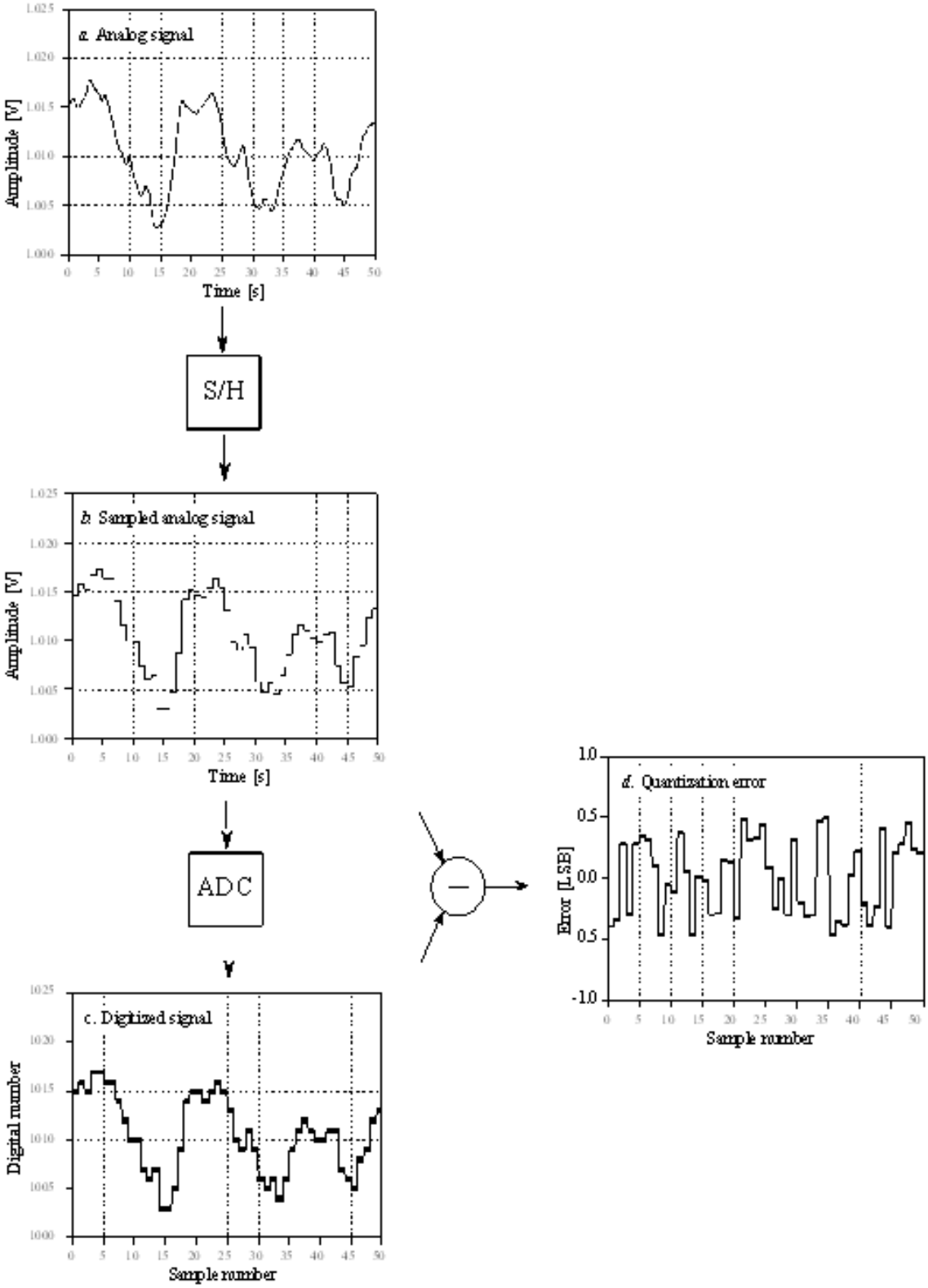
Le forme d'onda, rappresentate nella figura sottostante, illustrano il processo di digitalizzazione.

- La prima fase, dal segnale analogico originale (a) al segnale analogico campionato (b) è denominata *sample-and-hold* (S/H) e permette di convertire il tempo da continuo a discreto.
- Nella seconda fase, dal campione del segnale analogico (b) al segnale digitalizzato (c), l'ADC converte la variabile dipendente nel numero intero più prossimo.

Pertanto solo l'informazione relativa al valore istantaneo del segnale, quando si effettua il campionamento periodico, è conservata. Invece i cambiamenti nel segnale di ingresso che si verificano tra due tempi di campionamento sono trascurati. Anche i dati in uscita possono cambiare solo ad intervalli periodici.

Ciascun campione nel segnale digitalizzato ha un errore massimo di ± ½ LSB (*Least Significant Bit*, ossia distanza tra livelli di quantizzazione adiacenti), come mostrato dall'errore di quantizzazione (d), ricavato sottraendo (b) a partire da (c), con opportune conversioni.

L'errore di quantizzazione appare come un rumore casuale e, di solito, può essere modellato come semplice aggiunta di rumore al segnale.



Processo di digitalizzazione di una forma d'onda [modificato da Smith].

d.5. Teorema del campionamento di Nyquist

Al fine di ricostruire esattamente il segnale analogico dai campioni che descrivono le informazioni chiave del processo il segnale continuo deve essere correttamente campionato. Secondo il teorema del campionamento di Nyquist, un segnale continuo non può contenere componenti di frequenza superiore alla metà della frequenza di campionamento.

Si consideri, ad esempio, un segnale rilevato con dati campionati ad intervalli Δt , caratterizzato da frequenza di campionamento: $f_s = 1/\Delta t$ (campioni al secondo). La componente di frequenza più elevata prende nome di frequenza di Nyquist: $f_N = f_s/2 = 1/2\Delta t$.

Se esistono frequenze più elevate della frequenza limite di Nyquist, esse assumeranno valori compresi tra 0 e f_N , e risulteranno combinate con le frequenze dell'intervallo $\left[0, f_N\right]$. Il teorema del campionamento di Nyquist è un'applicazione del principio fisico di indeterminazione di Heisenberg, dato che l'area del rettangolo definito da Δf e Δt ha un minimo pari a $1/2$.

d.5.1. Trasformata di Fourier Discreta (DFT)

Si consideri un segnale discreto composto da una serie complessa $x(k\Delta t)$ per $k = 0, \ldots, N-1$, costituita da N campioni nell'intervallo di tempo $T = N\Delta t$; la trasformata di Fourier Discreta (DFT), definita come $X(k\Delta f)$ e composta da N campioni è definita come segue:

$$X(k\Delta f) = \Delta t \sum_{k=0}^{N-1} x(k\Delta t) \exp[-i2\pi (k\Delta t)(n\Delta f)] \qquad (4.17)$$

dove:

Δt è l'intervallo di campionamento di x

$\Delta f = 1/T$ è l'intervallo di campionamento di x

L'operatore inverso, detto Anti – Trasformata di Fourier (IDTF), trasforma la sequenza $X(k\Delta f)$ nella sequenza di partenza $x(k\Delta t)$:

$$x(k\Delta t) = \Delta f \sum_{k=0}^{N-1} X(k\Delta f) \exp[i2\pi (k\Delta t)(n\Delta f)] \qquad (4.18)$$

Lo spettro di Fourier risulta:

$$S(k\Delta f)=\frac{1}{N\Delta t}|X(k\Delta f)|^2 \qquad (4.19)$$

$$S(k\Delta t)=\frac{1}{N\Delta t}|x(k\Delta t)|^2 \qquad (4.20)$$

d.6. Fast Fourier Transform (FFT)

Tra i modi per calcolare la trasformata Discreta di Fourier (DFT), l'algoritmo Fast (Discrete) Fourier Transform (FFT) è uno dei più popolari ed efficaci per la riduzione dei tempi di calcolo.

La tecnica è inizialmente scoperta da K.F. Gauss (1777-1855); tuttavia la tecnica, risultata poco pratica (mancando allora mezzi digitali), è in gran parte dimenticata. Nel 1965, all'inizio della rivoluzione tecnologica portata dai computer [3], J.W. Cooley e J.W. Tukey ripropongono l'algoritmo FFT.

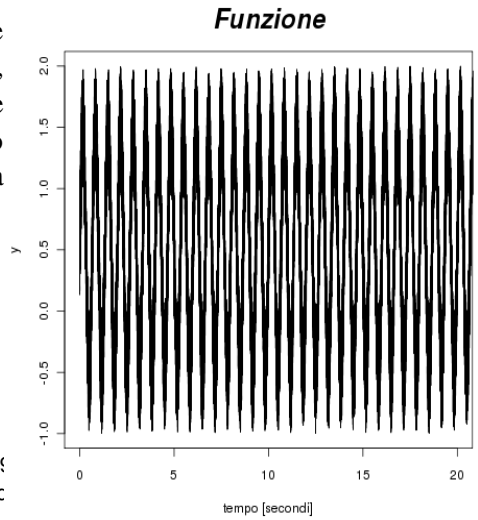
Il successo dell'algoritmo FFT consiste nel fatto che nella trasformata Discreta di Fourier sono presenti ridondanze che aumentano il carico computazionale. Inoltre scegliendo la frequenza di analisi del campione, per tale frequenza sono generate sia un'onda sinusoidale che un'onda cosinusoidale. Pertanto l'ampiezza della trasformata di Fourier è data da

$$X(k\Delta f)=\Delta t\left[\left(\sum_{k=0}^{N-1} x(k\Delta t)\sin[2\pi(k\Delta t)(n\Delta f)]\right)^2+\left(\sum_{k=0}^{N-1} x(k\Delta t)\cos[2\pi(k\Delta t)(n\Delta f)]\right)^2\right]^{12} \qquad (4.21)$$

Questa è una operazione di ordine N^2 , dove sono coinvolte due variabili (numero di frequenze da analizzare e numero di campioni) ed il tempo di calcolo è proporzionale al prodotto di queste due variabili.

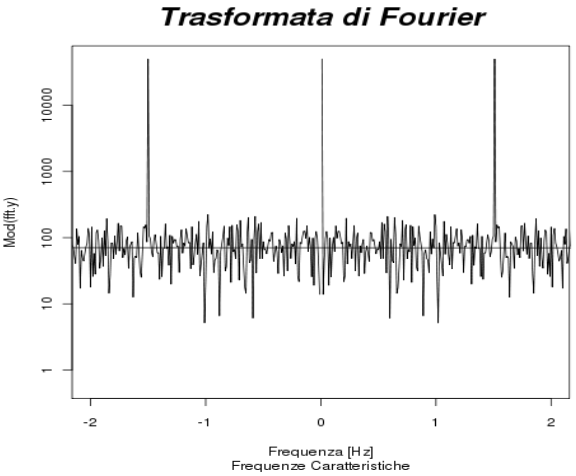
In primo luogo, le onde seno e coseno sono le stesse onde in diverse posizioni. Considerando solamente l'onda sinusoidale, la cosinusoide si ottiene spostando la fase di $\pi/2$. Inoltre utilizzando le proprietà di simmetria delle onde si riducono ulteriormente i tempi di calcolo. Infatti nell'equazione in forma complessa, è presente un termine noto:

$$W(k,n)=\exp[-i2\pi(k\Delta t)(n\Delta f)]$$



[3] L'avvento e la diffusione, come ben noto, datano a partire solo dagli anni '50 del '9
addirittura a B. Pascal, nel secolo d'oro del '600 francese, e via, via perfezionate ec

Se $N=2^{\gamma}$, $W(k,n)$ può essere espresso come il prodotto di γ matrici di dimensioni $N \times N$ al fine di ridurre l'onere computazionale da un'operazione di ordine di N^2 ad una di ordine di $N \log_2 N$. e consegue che uno dei requisiti dell'algoritmo è che il numero di campioni del segnale sia pari ad una potenza di 2 . Se tale requisito non è soddisfatto, ai campioni del segnale deve essere aggiunto un supplemento di campioni, di valore fissato pari a zero, per ottenere il numero di campioni richiesto pari ad una potenza di 2 . In generale, l'aggiunta di campioni di valore nullo non modifica le stime di densità spettrale.



d.6.1. Limitazioni della FFT

La FFT è dal punto di vita computazionale un approccio efficace, ma rivela alcune limitazioni quando è applicata a serie temporali. Di conseguenza, devono essere presi in considerazione metodi alternativi che superino tali limitazioni. I principali punti deboli delle tecniche di Fourier sono la dispersione spettrale, l'*aliasing* e la risoluzione,

d.7. Dispersione spettrale

La dispersione spettrale (*spectral leakage*) si manifesta come frammentazione dell'energia (ampiezza) da distinte caratteristiche spettrali che si distribuisce nei canali di frequenza adiacenti, dando luogo a componenti spurie nello spettro delle frequenze del segnale. La componente di frequenza del segnale non appare una singola linea ma come una diffusione di frequenze e la risposta di un segnale più debole può essere coperta da quella di un segnale più forte.

La dispersione spettrale si verifica quando la componente di frequenza di un segnale non finisce esattamente in uno dei canali di frequenza dello spettro calcolato. Se il periodo del segnale campionato è $N\Delta t$, dove N è il numero di campioni e Δt l'intervallo di campionamento, i canali di frequenza sono le armoniche della frequenza fondamentale $1/N\Delta t$.

Per evitare completamente tale rischio, occorrerebbe far sì che tutte le componenti di frequenza del segnale coincidano esattamente con i canali di frequenza dello spettro calcolato. Tuttavia questo è poco pratico, soprattutto per segnali arbitrari contenenti molte (e di solito sconosciute) componenti di frequenza.

Gli effetti della dispersione spettrale possono inoltre essere ridotti mediante tecniche di finestratura. I valori del segnale campionato sono moltiplicati per una funzione che si riduce a zero alle estremità della finestra, in modo che il segnale possa avere origine ed interrompersi in maniera graduale e non bruscamente. In questo modo si riduce l'effetto delle discontinuità e, di conseguenza, la dispersione spettrale. Esiste tuttavia uno svantaggio, perché le linee dello spettro di un segnale diventano più ampie, rendendo difficile distinguere componenti di frequenza separate e riducendo la risoluzione spettrale.

d.8. Aliasing

Un alias rappresenta un identità assunta, spesso falsa. Se una componente di frequenza del segnale è sottocampionata (vale a dire con un tasso di campionamento inferiore al minimo prescritto dal teorema di campionamento di Nyquist), si ritrovano frequenze di alias più basse. Pertanto le componenti di frequenza del segnale più elevate potrebbero apparire come componenti a bassa frequenza ed è impossibile ricostruire correttamente le informazioni contenute nel segnale originale.

Dato che la frequenza di Nyquist è la più alta componente di frequenza che può essere correttamente identificata senza aliasing, il tasso di campionamento deve soddisfare i teorema di campionamento di Nyquist.

d.9. Risoluzione spettrale

Sulla base del teorema del campionamento di Nyquist, la banda di frequenza è fissata tra 0 Hz e la frequenza di Nyquist e non tra le frequenze presenti nel segnale. Se N è il numero di campioni e Δt l'intervallo di campionamento, la risoluzione spettrale è data da $1/N\Delta t$.

La risoluzione spettrale indica il livello di dettaglio dello spettro di frequenza, ovvero la capacità di distinguere la risposta spettrale di due o più segnali. Se la risoluzione spettrale è troppo rozza, si possono perdere dettagli dello spettro. Pertanto $N\Delta t$ deve essere amplificato, aumentando N od amplificando Δt , oppure entrambi. Tuttavia se Δt è troppo grande può verificarsi l'aliasing. Al contrario, l'aumento di N può richiedere un'elevata quantità di memoria necessaria per immagazzinare i dati da processare e tempi di calcolo troppo lunghi. Pertanto N e Δt devono essere scelti in modo da fornire un'adeguata risoluzione di frequenza ed evitare inconvenienti.

d.10. Analisi di sequenze stocastiche

L'analisi di Fourier di segnali stocastici richiede particolare attenzione, perché devono essere considerate quantità medie.

Tutte le medie dipendono dalla deviazione standard e pertanto, noti i valori degli scarti quadratici medi, si conoscono i valori di tutte le funzioni di correlazione del processo stocastico. Per processi ergodici gaussiani, la densità spettrale è la trasformata di Fourier della funzione di correlazione cosicché, una volta determinata la densità spettrale, si ricavano tutte le proprietà statistiche del processo stocastico.

Dopodichè si considera la funzione di autocorrelazione della sequenza stocastica. L'autocorrelazione è una funzione deterministica ed applicando la trasformata di Fourier si ricava una funzione deterministica nel dominio della frequenza. La trasformata di Fourier della funzione di autocorrelazione è definita densità spettrale di potenza (*Power Spectral Density* o *PSD*) e descrive la densità di potenza del segnale stocastico sull'asse delle frequenze:

$$PSD[s(t)] = S_{ss}(\omega) = F\{ r_{ss}(\tau)\} = \int_{-\infty}^{\infty} r_{ss}(\tau) e^{-j\omega\tau} d\tau \qquad (4.23)$$

Se la funzione di autocorrelazione è pari, la *PSD* è reale. Spesso interessa analizzare la relazione tra due processi che, nel dominio del tempo, è data dalla funzione di cross-correlazione:

$r_{xy}(t,\tau)=E\left[x(t)\,y(t-\tau)\right]$, e, nel dominio della frequenza, la sua trasformata di Fourier prende il nome di densità spettrale di *cross-power* o *cross-spectrum*:

$$S_{xy}(\omega)=F\left[r_{sy}(\tau)\right]=|S_{sy}(\omega)|e^{j\theta_{sy}(\omega)} \qquad (4.24)$$

Si noti che, in questo caso, si abbia implicitamente fatto l'assunzione di stazionarietà dei segnali $s(t)$ ed $y(t)$ per cui la funzione di cross-correlazione non è funzione del tempo, ma della differenza τ . Inoltre $r_{sy}(\tau)$ non è pari, pertanto la sua trasformata di Fourier non è reale e necessita della conoscenza di modulo e fase. Il valore assoluto della densità spettrale di cross-power è limitato:

$$S_{sy}^2(\omega)\le S_{ss}(\omega)S_{yy}(\omega) \qquad (4.25)$$

Il valore assoluto di questa ultima espressione può essere normalizzato, in tal caso, si parla di funzione di coerenza:

$$\gamma_{sy}^2=\frac{S_{sy}^2(\omega)}{S_{ss}(\omega)S_{yy}(\omega)}\le 1 \qquad (4.26)$$

d.11. Analisi di segnali reali

L'analisi in frequenza del segnale, mediante le tecniche tradizionali finora presentate, è effettuata per segnali assimilabili a processi stocastici a banda limitata con distribuzione gaussiana e media nulla. Tali condizioni sono in realtà difficilmente realizzabili. Tuttavia il segnale per intervalli di tempo relativamente brevi, può essere assunto come stazionario. Per caratterizzare il segnale sia nel dominio del tempo che in quello della frequenza sono introdotte trasformate tempo-frequenza (*Time-Frequency Representation* o *TFR*), quali:

- *Short-Time Fourier Transform (STFT)*, altrimenti detti: spettri evolutivi;
- *Wavelet* (*WT)*, altrimenti dette: ondine.

In entrambi i casi, per analizzare un segnale variabile rispetto al tempo, si calcola lo spettro del segnale in un intorno dello stesso, troncando il segnale con una finestra di durata limitata ed effettuando l'analisi spettrale. Forma e durata della finestra influiscono sui risultati dell'analisi. Nella STFT ampiezza e forma della finestra rimangono costanti durante tutta l'analisi, mentre nella trasformata Wavelet la durata della finestra aumenta al diminuire della frequenza.

Uno svantaggio comune alle due tecniche consiste invece nella contrapposizione tra la risoluzione nel dominio del tempo e quello della frequenza, dato che il prodotto è costante e pari ad un valore che dipende dalla finestra utilizzata. Si perde inoltre accuratezza nell'analisi di segnali il cui contenuto armonico varia talmente velocemente da rendere impossibile la scelta di una finestra appropriata.

d.12. Short Time Fourier Transform (STFT)

Nella trasformata di Fourier, il segnale è analizzato su tutti i tempi per una durata infinita. Questo deriva dal fatto che non esiste il concetto di tempo nel dominio della frequenza e, di conseguenza, il concetto di cambiamento di frequenza rispetto al tempo. Pertanto la trasformata di Fourier è adatta a segnali stazionari perché permette di evidenziare nettamente fenomeni che avvengono nel dominio della frequenza, ma non discrimina eventi che avvengono nel dominio del tempo. Infatti da un punto di vista matematico, la frequenza ed il tempo sono ortogonali, mentre la maggior parte dei segnali reali fisici è di tipo non stazionario, cioè con caratteristiche variabili nel tempo e le componenti di frequenza dei segnali che, a loro volta, cambiano con il tempo.

La Short Time Fourier Transform (STFT) è un metodo per valutare il modo in cui le frequenze cambiano nel tempo. Il segnale è suddiviso in piccole parti e, per ogni parte, si calcola la Trasformata di Fourier. Pertanto ogni spettro di frequenza fa riferimento ad un breve periodo di tempo ed il confronto, tra spetti consecutivi, mostra l'evolversi delle frequenze contenute nel segnale al passare del tempo.

La principale contraddizione della Short Time Fourier Transform consiste nel considerare che le frequenze siano valide solo se si fa uso di onde sinusoidali infinitamente lunghe, mentre allo stesso tempo accetta di applicare trasformate di Fourier a brevi tratti del segnale.

d.13. Wavelet Transform (WT)

Nella trasformata Wavelet la durata della finestra aumenta al diminuire della frequenza. In questo modo, è possibile migliorare la risoluzione dell'analisi spettrale nel campo delle alte frequenze, cosa molto utile nel caso di analisi di transitori.

d.13.1. Analisi Wavelet

La trasformata Wavelet è particolarmente appropriata per l'analisi di segnali transitori e di sistemi tempo – varianti, poiché è localizzata sia nel tempo che in frequenza (Todorovska, 2001).

Le Wavelet sono funzioni matematiche che separano i dati in diverse componenti di frequenza, in modo da studiare ogni componente con una risoluzione abbinata alla scala. Paragonate ai tradizionali metodi dell'analisi di Fourier, le Wavelet sono adatte per approssimare dati con cuspidi e discontinuità. Pertanto i dati sono processati a differenti scale e risoluzioni:

- larghe finestre determinano gli aspetti più generali;
- piccole finestre permettono di analizzare i dettagli.

Matematicamente le Wavelet sono funzioni a durata limitata localizzate sia nel tempo che nella frequenza e con valore medio nullo. Affinché la condizione di media nulla sia soddisfatta, la funzione dovrà essere oscillatoria da cui il nome Wavelet (Vetterly e Kovacenic, 1995).

La procedura dell'analisi Wavelet consiste nell'adottare una funzione prototipo $\psi(t)$, chiamata Wavelet madre. La famiglia di Wavelet $\psi_{(a,b)}(t)$ può essere costruita tramite operazioni elementari che consistono di traslazioni e ridimensionamenti (dilatazioni e contrazioni) della Wavelet madre:

$$\psi_{(a,b)}(t)=\frac{1}{a}\psi\left(\frac{t-b}{a}\right) \qquad b\in R,\quad a>0 \tag{4.27}$$

dove a è il fattore di scala che indica la larghezza della Wavelet e b lo sfasamento temporale che ne definisce la posizione. La Wavelet prototipo corrisponde a $b=0$ e $a=1$

- se il fattore di scala a è maggiore dell'unità, la trasformazione corrisponde ad una dilatazione della Wavelet madre;
- se il fattore di scala a è miniore dell'unità, la trasformazione corrisponde ad una contrazione della Wavelet madre.

Il coefficiente $1/a$ è un fattore di normalizzazione per l'ampiezza della Wavelet, selezionato in modo tale che tutte le Wavelet nella famiglia abbiano la medesima norma [Carmona, 1998; Deplart, 1992].

$$\int_{-\infty}^{+\infty} |\psi_{(a,b)}(t)|dt = \int_{-\infty}^{+\infty} |\psi(t)|dt \qquad (4.28)$$

In funzione di a e b, è possibile osservare un auto-similarità tra le basi Wavelet e, data la funzione di madre, tutto ciò che riguarda la base è nota. L'accuratezza è controllata selezionando la larghezza della finestra della Wavelet: grandi finestre temporali forniscono una migliore localizzazione in frequenza, ma più povera nel tempo e viceversa. Uno dei principali vantaggi dell'analisi Wavelet è l'abilità di effettuare analisi locali, ovvero di analizzare una parte localizzata di un segnale più grande.

- L'analisi temporale è effettuata utilizzando versioni contratte, ad alta frequenza della Wavelet prototipo.
- L'analisi di frequenza è effettuata con versioni dilatate, a bassa frequenza della medesima Wavelet.

Se il segnale originale è rappresentato in termini di espansione Wavelet (tramite i coefficienti di una combinazione lineare delle funzioni Wavelet), le operazioni sui dati possono essere effettuate utilizzando soltanto i corrispondenti coefficienti Wavelet.

d.14. La trasformata continua wavelet (CWT)

Come l'analisi di Fourier consiste nella scomposizione di un segnale in onde sinusoidali di differenti frequenze, in maniera simile, nell'analisi Wavelet il segnale è scomposto in versioni traslate e ridimensionate della Wavelet madre. La trasformata continua Wavelet (CWT) di una funzione è definita come la somma sul tempo di un segnale moltiplicato per versioni scalate e traslate della funzione Wavelet. I risultati della CWT sono differenti coefficienti Wavelet T_f, una funzione di scala e posizione.

$$T_f(a,b) = \int_{-\infty}^{+\infty} f(t)\,\bar{\psi}_{(a,b)}(t)\,dt \qquad (4.29)$$

Moltiplicando ciascun coefficiente T_f per la wavelet opportunamente ridimensionata e spostata è possibile ricostruire le funzioni wavelet che costituiscono il segnale originale. Uno dei principali vantaggi offerta dall'analisi wavelet è la capacità di effettuare analisi locali - cioè, di analizzare una zona localizzata di un segnale più grande.

Come riportato da Todorovska [2001], la Trasformata Wavelet è particolarmente adatta per l'analisi di segnali transitori e di sistemi variabili nel tempo, perché è localizzata sia nel tempo che nella frequenza. Con l'analisi wavelet, è possibile utilizzare funzioni approssimate perfettamente contenute in domini finiti. Per impulsi con variazione continua della loro frequenza, la precisione della localizzazione è povera se le ampiezze dell'impulso sono piccole.

A partire dai coefficienti Wavelet, il metodo più diretto per la determinazione della frequenza istantanea di un segnale consiste nel trovare, per ogni istante b, la scala a per la quale l'ampiezza della trasformata risulta massima. In questo modo, non si ha bisogno di usare nessuna informazione aprioristica circa i massimi, poiché il metodo si basa sull'osservazione euristica che, in un certo intervallo ristretto, l'energia del segnale è concentrata vicino alla sua frequenza istantanea.

Una volta determinato il massimo, le frequenze istantanee sono calcolate come:

$$\omega_S(b)=\frac{\omega_0}{a_r(b)} \qquad (4.30)$$

in cui ω_0 è la frequenza centrale della Wavelet madre, implicando che la Wavelet dilatata ($a>1$) ha una frequenza minore della Wavelet prototipo di a volte.

5. ANALISI ED APPLICAZIONI

Indice Capitolo 5

a. Riassunto....1
b. Descrizione dei Dati....1
c. Osservazione dei Dati Grezzi....3
d. Classi di problemi....3
e. Applicazione della Trasformata di Fourier....7
f. Analisi statistico....9
g. Risultati....9

a. Riassunto

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli da interpretare sono concepiti come popolazioni di dati ideali, costituenti un universo da cui estrarre dati reali (campionamento) costituiti, a loro volta, da campioni da interpretare. Le principali proprietà delle stime sono: la correttezza, la consistenza, l'efficienza, la sufficienza e la robustezza.

Nel campionamento, numerico e staticamente significativo, di basi di dati di notevoli dimensioni, esistono diverse problematiche collegate alla rappresentabilità dei campione1. è questo il caso per esempio di campionamenti poco significativi o di campionamenti con un numero di dati estremamente elevato. Tale argomento si pone in relazione con la stima delle frequenze, principalmente per quanto riguarda la ricerca dei limiti di confidenza di una frequenza osservata, per una popolazione distribuita binomialmente. Per stabilire la bontà dei modelli d'interpretazione delle osservazioni c'è bisogno di applicare test. I test statistici necessitano di un campionamento di dati significativo e rappresentativo della popolazione da cui è considerato estratto. Queste fatto può comportare l'esigenza di collezionare un numero elevato di estrazioni, prima di poter indicare ipotesi sensate e sopratutto svolgere il test.

Un'alternativa che permette di eseguire l'analisi di evidenze sperimentali quando sono presenti pochi dati è il test sequenziale dato che lo scopo di questo tipo di analisi è quello di arrivare a scegliere, tra ipotesi alternative, con il minimo numero di osservazioni. Parallelamente, la capacità di acquisire grosse moli di dati, durante un processo di misura, da un lato, permette una descrizione più puntuale del fenomeno, dall'altro, richiede maggior attenzione nella scelta del modello stocastico. Questo aspetto, a volte trascurato, può portare seri problemi sull'attendibilità delle stime, quando si ignorano eventuali correlazioni presenti tra le misure. Tuttavia in alcuni insiemi di dati, è possibile evidenziare e successivamente quantificare la correlazione, presente nelle osservazioni, mediante un approccio di tipo stocastico.

Un'altra alternativa alla soluzione dei problemi ed assumendo che i dati siano una sequenza estratta dalla realizzazione di un qualche processo stocastico (segnali stocastici), è fornita dai metodi di Fourier che, qui come altrove, trovano la loro naturale collocazione, in parallelo al metodo degli elementi finiti.

b. Descrizione dei Dati

Un discorso preciso sulla statistica descrittiva si avvia con la definizione di variabile statistica e di variabile casuale, la postulazione di un'identità formale fra le stesse e la presentazione delle loro principali statistiche. Le variabili statistiche sono il risultato di esperimenti e, pertanto, sono concrete (ovvero costituite da dati reali od osservazioni, come la totalità dei dati a referenza spaziale, tempo varianti e non, quali, ad esempio, le misure geodetiche e geomatiche), finite (perché qualsiasi esperimento incontra evidenti limiti di

spazio, tempo ed altre condizioni limitative) e discrete (perché qualsiasi esperimento è eseguito con una determinata accuratezza). Conseguentemente esse sono caratterizzate da un insieme di valori argomentali (eventualmente raggruppati in classi), associati a frequenze elementari (assolute, come risultato di un conteggio, oppure relative, se la totalità è normalizzata ad uno) ed alle frequenze cumulate delle frequenze elementari.

Le variabili casuali sono modelli interpretativi e, pertanto, sono astratte (ovvero costituite da dati ideali od osservabili) ed, in generale, illimitate e continue (anche se, raramente, ad eccezione della teoria dei giochi, esse possono essere finite e discrete). Conseguentemente esse sono caratterizzate da un campo d'esistenza, associato ad una funzione densità di probabilità ed ad una funzione distribuzione di probabilità (comunemente detta: probabilità). L'identità formale fra variabili statistiche e variabili casuali discende dalla loro completa indistinguibilità, a valle della loro definizione. Allora la presentazione delle principali statistiche può essere eseguita congiuntamente per entrambe.

Per le variabili ad una dimensione, le principali statistiche rispondono alla quantizzazione delle idee di: centro, dispersione, simmetria e curtosi (comportamento delle code). Come noto, il centro può essere indicato tramite la moda, la mediana, le medie (aritmetica, geometrica, armonica, ponderata, potata, ecc.) od altro, la dispersione può essere valutata in base all'ampiezza, ai quantili, alla varianza, agli scarti assoluti medio o mediano, ecc., mentre gli indici di asimmetria e curtosi hanno, solitamente, poche varianti.

Per le variabili a due dimensioni, le principali statistiche (oltre a quelle monodimensionali marginali o condizionate) rispondono alla quantizzazione dell'idea di dipendenza. Come noto, dipendenza è un concetto molto generale che, fra totale e completa indipendenza e perfetta dipendenza (o dipendenza in legge), si articola in connessione (dipendenza vaga e generica), regressione (quasi-dipendenza funzionale) e correlazione (quasi-dipendenza lineare). Ancora numerosi sono gli indici ed i coefficienti che esprimono il grado della dipendenza o meno (si noti, a riguardo, come tutti siano normalizzati ad uno, assumendo anche valori negativi, fino a meno uno, se non intrinsecamente positivi).

Per le variabili a più di due dimensioni, a rigore, occorre continuare lo studio del loro raggrupparsi (come con gli indici di nuvolosità, ecc.). Tuttavia nel caso frequente in cui il modello interpretativo è fornito dalla variabile casuale normale, questo studio è del tutto superfluo. Si ricordi, inoltre, che detta variabile casuale è completamente caratterizzata dal vettore delle medie e dalla matrice di varianza-covarianza, cosa che rende superflue altre statistiche del centro, della dispersione e della dipendenza (covarianza comporta correlazione, ovvero dipendenza lineare e niente altro) e del tutto inutili le statistiche superiori (la variabile casuale normale è simmetrica e l'indice di curtosi vale, in ogni caso, tre).

Ulteriori vantaggi dell'adozione, quale modello interpretativo, della variabile casuale normale sono dati dall'invarianza della distribuzione di probabilità di detta variabile casuale, rispetto a trasformazioni lineari della variabile casuale stessa, e dell'ottimalità della stima dei parametri di modelli, supportati dalla variabile casuale normale, se le ipotesi di corrispondenza fra dati e modelli sono perfettamente soddisfatte (ovvero se i dati non sono affetti, in alcun modo, da dati anomali). A tutto ciò, si aggiunge la linearità dei sistemi da risolvere per la stima dei parametri di modelli, fatto di primaria importanza, in quanto solo i sistemi lineari ammettono, senza eccezioni e purché non-singolari, soluzioni esattamente determinabili, indipendentemente dal numero di equazioni ed incognite di cui si compongono.

La statistica descrittiva termina con alcuni teoremi limite. Fra questi il teorema di Bernoulli[1] (o legge dei grandi numeri) mostra la convergenza, in probabilità, delle frequenze di una variabile statistica alle probabilità di una corrispondente variabile casuale, mentre il teorema di Gauss[2] (o limite centrale della statistica) mostra la convergenza, in legge, della combinazione lineare (ovvero delle somme, come caso

[1]

[2] la somma (normalizzata) di un grande numero di variabili casuali è distribuita approssimativamente come una variabile casuale normale standard

particolare) di variabili casuali qualsiasi, purché aventi ciascuna dispersione comparabile con le altre, alla variabile casuale normale. I due teoremi giustificano, rispettivamente, la comparazione fra variabili statistiche e variabili casuali, al di là della sopracitata identità formale, e le operazioni di media aritmetica, ponderata o potata fra osservazioni dirette per aumentare la normalità del comportamento dei dati.

c. Osservazione dei Dati Grezzi

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli, da interpretarsi, come già detto, quali variabili casuali, sono concepiti come popolazioni di dati ideali (od osservabili), costituenti un universo da cui estrarre dati reali (od osservazioni), costituenti, a loro volta, campioni da interpretare, pertanto, come le suddette variabili statistiche. L'operazione d'estrazione è detta campionamento[3] e, se i dati sono fra loro indipendenti, come avviene, auspicabilmente, nelle osservazioni dirette di fenomeni, il campionamento è detto bernoulliano; schemi di campionamento più complessi, attinenti alle problematiche della progettazione, simulazione ed ottimizzazione, sono considerati estranei agli scopi del presente lavoro.

Le principali proprietà delle stime sono la correttezza, la consistenza, l'efficienza e la sufficienza che, in base al significato letterale dei nomi, significano capacità di stimare parametri il cui centro coincide con il centro dei parametri dell'intera popolazione, capacità di stimare parametri con precisione ed accuratezza sempre maggiori, ed al limite infinite, al crescere della numerosità del campione, capacità di stimare parametri qualitativamente migliori delle informazioni presenti direttamente nelle osservazioni, capacità di stimare parametri conservando tutta la ricchezza di informazioni già presente nelle osservazioni dirette. Una proprietà aggiuntiva delle stime, estranea al corpus della statistica tradizionale, è la robustezza, intesa come capacità di stimare parametri indistorti, nonostante la presenza di eventuali dati anomali.

Le caratteristiche delle stime forniscono spesso, contemporaneamente, le modalità operative per effettuare le stime stesse. Infatti sono operativi tanto il metodo della minima varianza, ottimale per la statistica tradizionale, quanto diverse procedure robuste, certamente subottimali, ma capaci di evitare indesiderate distorsioni nelle stime. Altrettanto operativi sono il metodo della massima verosimiglianza[4] ed il metodo dei minimi quadrati, una particolarizzazione del primo nel caso in cui le stime avvengono in ambito lineare ed il modello interpretativo è fornito dalla variabile casuale normale. Si ricordi che, in questo caso, si ha l'ottimalità delle stime, in quanto tanto il metodo dei minimi quadrati[5], quanto quello della massima verosimiglianza, da cui discende, danno risultati perfettamente coincidenti con il metodo della minima varianza.

Tutto ciò conferma l'adozione del metodo dei minimi quadrati per il trattamento statistico delle osservazioni, agevola le sue generalizzazioni ed estensioni a tecniche complementari e giustifica un modo di procedere che prevede la centralità di detto metodo e riconduce ad esso, per quanto possibile, importanti tecniche complementari (cluster analysis, regressione multipla, analisi di varianza, delle componenti di varianza e della struttura di covarianza, procedure robuste). Si noti, in quest'ambito, il ruolo fondamentale ed indispensabile giocato dall'indissolubilità del legame fra un certo tipo di statistiche classiche elementari, la normalità e la linearità, per quanto riguarda tanto la definizione statistica delle metodologie, quanto la loro applicazione con elevate capacità risolutive in appropriati algoritmi numerici.

d. Classi di problemi

[3] Per campione se intende parte della popolazione che viene selezionata per l'analisi.
[4] Il metodo della massima verosimiglianza in statistica è un procedimento matematico per determinare uno stimatore.
[5] Il metodo dei minimi quadrati generalizzati di Aitken consente la stima di un modello lineare, sotto ipotesi più generali di quelle del modello classico di regressione lineare multivariata.

I problemi ai minimi quadrati si presentano, in generale e nell'ambito specifico delle discipline geodetiche e geomatiche, usualmente ripartiti in due classi fondamentali:

- problemi reticolari (o di compensazioni di reti);
- problemi d'interpolazione ed approssimazione di campi di punti.

Le stesse due classi si incontrano anche in problemi affini, quali ad esempio:

- i campionamenti delle osservazioni, l'ottimizzazione della configurazione di rilevamento e/o dello schema di misura, oppure dei pesi delle osservazioni;
- la "cluster analysis", l'analisi di varianza, la regressione multipla, l'analisi fattoriale (o studio delle componenti principali);
- lo studio dell'affidabilità delle osservazioni e le procedure di validazione dei dati e di stima dei parametri con procedure robuste.

In ogni caso, tutti i problemi minimi quadrati possono essere interpretati, topologicamente, come un grafo, dove:

- le osservazioni, i vincoli e le pseudo-osservazioni sovrappesate e non, come pure le informazioni a priori, le osservabili secondarie e le condizioni numeriche di regolarizzazione, costituiscono i lati del grafo;
- i parametri principali ed ausiliari (o di servizio) costituiscono i nodi dello stesso grafo.

Si noti, a riguardo, come l'interpretazione data della topologia sia indispensabile per una corretta comprensione dei casi e sottocasi in cui si articolano le suddette classi fondamentali. Le due classi fondamentali, già precedentemente enunciate, si articolano in vari e svariati casi e sottocasi, illustrati, dettagliatamente, nel prosieguo. I problemi reticolari (o di compensazione di reti) presentano come osservabili:

- differenze prime dei parametri;
- funzioni delle differenze prime dei parametri.

Il primo caso ha numerosi esempi, anche fuori dalle discipline geodetiche e geomatiche:

- discretizzazione di equazioni differenziali del primo ordine, tipiche della fisica, della chimica e delle scienze della terra;
- problemi di trasporto: schemi di circuitazione, traffico, circolazione e transazione, reti di comunicazione, distribuzione e telecomunicazione ed è costituito, per le suddette discipline, dalle reti di differenza di potenziale.

Il secondo caso è, invece, tipico delle discipline geodetiche e geomatiche, anche se non esclusivo (a questo caso, infatti, fanno riferimento ben particolari discretizzazioni di equazioni differenziali, sempre riferite ai sopracitati raggruppamenti di discipline fisiche e naturalistiche), e si articola nei seguenti sottocasi:

- l'informazione fluisce completa, bidirezionalmente, come nelle reti di differenza di potenziale, lungo ogni lato del grafo:
- l'informazione fluisce completa, unidirezionalmente, lungo ogni lato del grafo, costituendo nel suo fluire almeno un albero sul grafo stesso;
- l'informazione è irradiata, in modo completo, da alcuni nodi verso altri (senza ritorno), senza che né i primi, né i secondi si scambino alcuna informazione, costituendo nel suo fluire tanti alberi (costituiti da un solo livello, oltre la radice) sul grafo stesso, quanti sono i nodi d'emanazione:
- l'informazione è irradiata, in modo parziale, nelle stesse condizioni del sottocaso precedente, cosa che richiede l'individuazione di due o più co-alberi (sempre costituiti da un solo livello, oltre le radici) capaci di completare l'informazione trasmessa;
- l'informazione è irradiata, in modo parziale, senza restituzioni particolari.

Le osservabili differenze seconde dei parametri e loro funzioni richiedono la complessa sostituzione dei lati del grafo con triangoli fra i tre nodi interessati. Le osservabili differenze di ordine superiore e loro funzioni fanno riferimento, addirittura, a poligoni fra tutti i nodi coinvolti, cosa che rende la loro analisi ancora più complessa. Per queste ragioni, ad eccezione della discretizzazione di equazioni differenziali di secondo ordine o di ordine superiore e di loro trasformazioni funzionali, la loro adozione è estremamente rara. Per quanto riguarda la determinazione del numero di parametri principali, nel caso in cui i problemi ai minimi quadrati adottino lo schema principe delle equazioni d'osservazione, questo è sempre tale da determinare difetti di rango e singolarità del sistema da risolvere per cui sono necessari vincoli o pseudo-osservazioni sovrappesate.

I problemi d'interpolazione ed approssimazione di campi di punti presentano come osservabili funzioni dirette dei parametri principali, il cui numero, sempre nel caso in cui si voglia adottare il suddetto schema principe delle equazioni d'osservazione, non è mai tale da determinare difetti di rango e singolarità del sistema da risolvere. A tutto ciò, fanno eccezione eventuali problemi di sovra-parametrizzazione, rispetto al campionamento delle osservazioni effettuate, per cui sono indicate condizioni numeriche di regolarizzazione.

Esempi di problemi di interpolazione ed approssimazione di campi di punti sono dati da:

- ricostruzione (fitting) di linee, superfici, ipersuperfici aventi come dominio lo spazio 3D:
- descrittori di forma (form descriptors): contorni di figure (piane e/o gobbe), superfici (chiuse) di oggetti:
- centratura (matching) di segmenti, figure (immagini, mappe, disegni), oggetti (compresi modelli virtuali 3D) comunque conformati:

I problemi d'interpolazione ed approssimazione[6] di campi di punti, relativi alla ricostruzione di linee sono, ovviamente, assimilabili a quelli dello studio delle serie temporali storiche o di breve periodo, oppure frutto di simulazioni. Inoltre lo studio di serie temporali congiunto alla soluzione dei problemi reticolari (o di compensazione di reti) e/o d'interpolazione ed approssimazione di campi di punti, illustrati in precedenza, permette indagini accurate sugli aspetti dinamici delle osservabili a referenza spaziale di cui ai suddetti problemi, dando un'interpretazione unitaria a dati spazio-varianti, tempo-varianti.

Limitatamente alle discipline geodetiche e geomatiche, mentre le equazioni d'osservazione dei problemi reticolari (e di compensazioni di reti) fanno uso, in generale, di modelli grigi dedotti dalla geometria del problema in esame, le equazioni d'osservazione dei problemi d'interpolazione ed approssimazione di campi di punti fanno uso, in generale, di modelli neri.

Come noto, una vasta gamma di metodi deterministici e/o stocastici risponde positivamente alla bisogna. I primi annoverano fra i più comunemente impiegati:

- l'interpolazione polinomiale;
- il metodo degli elementi finiti e l'interpolazione con funzioni splines;
- l'analisi di Fourier, nel dominio delle frequenze;
- lo studio, sempre nel dominio delle frequenze, con ondine (wavelets).

I secondi prevedono l'interpretazione dei fenomeni in istudio come realizzazioni di un processo stocastico:

- stime di covarianza, filtraggio cross-validazione e predizione;
- studio della geometria frattale.

Uno studio dettagliato di esempi particolari e significativi di problemi reticolari (o di compensazioni di reti) può essere effettuato, nell'ambito delle discipline geodetiche e geomatiche, solo facendo riferimento a discipline specifiche, quali la geodesia, la navigazione, la topografia, la fotogrammetria ed il telerilevamento. Al contrario, uno studio dettagliato di esempi particolari e significativi di problemi d'interpolazione ed approssimazione di campi di punti richiede anche uno studio dei modelli neri. Per una migliore comprensione si tenga presente che l'insieme delle quantità osservate è sempre costituito da quattro parti distinte:

- le informazioni topologiche, ovvero i lati del grafo che indicano le connessioni esistenti fra i nodi del grafo stesso;
- le informazioni geometriche, ovvero la posizione ed altre caratteristiche degli stessi nodi;
- le informazioni metrologiche, ovvero le osservazioni (o quantità osservate) realmente effettuate;
- le informazioni stocastiche, ovvero la precisione delle osservazioni e le eventuali correlazioni fra queste.

[6] metodo per individuare nuovi punti del piano cartesiano a partire da un insieme finito di punti dati, nell'ipotesi che tutti i punti si possano riferire ad una funzione f(x) di una data famiglia di funzioni di una variabile reale.

Infatti questo insieme, altrimenti detto: base di dati provenienti da operazioni di misura, con riferimento a ciascuna delle sopraccitate quattro parti distinte, produce nei problemi ai minimi quadrati (come pure negli altri sopraccitati problemi affini), rispettivamente:

- la matrice disegno simbolica;
- la matrice disegno numerica;
- il vettore termine noto delle equazioni d'osservazione;
- la matrice di varianza-covarianza (a priori) delle quantità osservate o, più comunemente, se non esistono correlazioni fra le stesse quantità osservate, la matrice dei pesi.

Si noti che, con la sola eccezione delle osservazioni realmente effettuate, tutto quanto può essere noto già prima di compiere una sola osservazione. Da ciò derivano tutti i problemi di ottimizzazione della matrice di varianza-covarianza dei parametri:

- intervenendo nella matrice disegno per decidere sull'effettuazione o meno di ciascuna osservazione (1° ordine);
- sulla matrice di varianza-covarianza (a priori) delle quantità osservate per stabilire, note le osservazioni da effettuarsi, le precisioni delle stesse (2° ordine);
- su opportune parziali combinazioni dei due casi precedenti (3° ordine),

avendo cura di controllare, in ogni caso, l'affidabilità delle osservazioni, quale garanzia, sufficiente minimale, che il lavoro intrapreso, qualsiasi esso sia, risulti svolto a regola d'arte.

e. Applicazione della Trasformata di Fourier

La validazione dati e dei modelli si fonda sulle varie tecniche dell'analisi multivariata[7], di volta in volta, studiando la variabilità e l'interdipendenza fra gli attributi, entro una classe di oggetti. Essa prende in considerazione insiemi di dati, ciascuno dei quali, relativo ad un oggetto della classe, contiene i valori osservati di certe variabili statistiche. Questi insiemi possono, talvolta, essere completi; mentre, più A loro volta, le variabili osservate, sono in generale campioni estratti da variabili casuali, di tipo continuo o, raramente, discreto. Da una tale complessità di premesse, lo studio dell'analisi multivariata si può articolare, principalmente, nei seguenti punti.

- Semplificazione strutturale. Gli insiemi di dati devono essere ricondotti, se possibile, in forme più semplici con cambi di variabili, in particolare, con trasformazioni capaci di sciogliere variabili connesse in variabili indipendenti.
- Classificazione degli oggetti. L'analisi dell'insieme di dati deve porre in evidenza la presenza di gruppi (clusters), ovvero di sottoinsiemi di oggetti, caratterizzati da valori preferenziali degli attributi o di parte di essi, cercando di ricondurre a poco le notevoli variabilità presenti.

[7] Con statistica multivariata s'intende quella parte della statistica in cui l'oggetto dell'analisi è per sua natura formato da almeno due componenti.

- Raggruppamento degli attributi (clustering). L'analisi degli insiemi di dati deve far ricadere, per quanto possibile, differenti variabili in un unico gruppo.
- Analisi della connessione. Gli insiemi di dati devono essere studiati rispetto alla dipendenza vaga e generica o meno fra le variabili contenute (ovvero all'essere in connessione di queste ultime).
- Analisi della dipendenza funzionale. Gli insiemi di dati devono essere studiati rispetto alla dipendenza funzionale o meno fra le variabili contenute (ovvero all'essere in regressione di queste ultime), con particolare riferimento alla dipendenza lineare o correlazione (ovvero all'essere in regressione lineare o correlate).
- Costruzione e verifica d'ipotesi. Il confronto probabilistico, fra statistiche campionarie e valori teorici di riferimento, permette di formulare un giudizio critico sui risultati ottenuti nelle varie tappe dell'analisi multivariata.

I tests di validazione dei modelli permettono di sottoporre a verifica, mediante opportuni controlli e confronti d'ipotesi, le stime effettuate come, del resto, tutti i risultati ottenuti nell'ambito della statistica. Al solito, si possono avere errori nel modello deterministico: presenza di errori grossolani nelle osservazioni, ed errori nel modello stocastico: presenza di errori sistematici nelle osservazioni, ovvero cattiva conoscenza delle varianze delle osservazioni e/o delle eventuali covarianze fra le osservazioni stesse. Un'opportuna sequenza di tests permette di districarsi fra le varie cause d'errore.

Un giudizio sui risultati può essere espresso in termini numerici e statistici. I controlli di tipo numerico rispondono a problemi di condizionamento ed affidabilità che comunemente accompagnano e seguono il metodo dei minimi quadrati, comprensivo delle sue estensioni e generalizzazioni; pertanto tutto quanto riguarda i controlli di tipo numerico è considerato estraneo agli scopi del presente lavoro. I secondi comprendono i tests statistici per la valutazione di osservazioni e parametri, della loro dispersione e, se del caso, della loro dipendenza. Per quanto riguarda le osservazioni, la validazione avviene in termini di entità degli scarti - residui (oltreché numericamente in termini di affidabilità delle osservazioni all'interno dello schema di misura) e consiste essenzialmente nell'individuazione ed eliminazione degli errori grossolani.

f. Analisi statistico

L'inferenza statistica (multivariata) è quella parte dell'analisi multivariata dedicata alla costruzione e verifica d'ipotesi (per problemi di controllo di qualità, oppure controllo e confronto d'ipotesi di altri problemi). Infatti il confronto probabilistico permette di formulare un giudizio critico sui risultati ottenuti, nelle varie tappe dell'analisi multivariata. Così con le usuali strategie dell'inferenza statistica, vari tests multipli consentono di discriminare, tanto stime di parametri da campioni normali, quanto statisitche di modelli non-parametrici (ovvero modelli distribution free)[8].

Gli oggetti del giudizio critico sono, come detto, i risultati ottenuti nelle varie tappe dell'analisi multivariata, in particolare: frequenze relative, contingenze, medie campionarie o altri indicatori del centro (di una popolazione), varianze campionarie o altri indicatori della dispersione, coefficienti di correlazione campionari o altri indicatori della correlazione[9].

I test multipli si differenziano per i diversi oggetti in esame, come pure per la distribuzione di appartenenza della popolazione cui i campioni si riferiscono, se normale, oppure altra (spesso sconosciuta). In generale, comunque, tutti i campioni sono supposti fra loro indipendenti, mentre se quest'ipotesi non è soddisfatta, occorre procedere, come per i modelli non-parametrici, adottando strategie assolutamente generali, ma assai poco potenti, cioè meno capaci di discriminare fra ipotesi alternative vicine, a parità di numerosità dei campioni. Nelle seguenti tabelle, si presentano parecchi tests multipli, diversamente, rispondenti alla bisogna, facendo attenzione, in particolare, ai problemi pratici, legati alla loro applicazione ad esempi concreti. Per quanto riguarda, invece, i loro fondamenti teorici, questi si richiamano, in generale:

g. Risultati

- alla definizione assiomatica di probabilità;

[8] L'aggettivo non parametrico (in letteratura inglese: distribution free) qualifica un particolare gruppo di tests statistici, sotto certe condizioni, sostitutivo dei tests statistici classici. Infatti i tests non-parametrici, rispetto ai test classici, presentano i seguenti vantaggi:

- la loro comprensione è immediata ed elementare;
- le condizioni di validità sono meno forti (più ampie);
- i calcoli necessari non presentano in generale difficoltà computazionali.

D'altra parte, i tests non-parametrici presentano alcuni svantaggi:

- molta informazione viene sprecata;
- la potenza del test è bassa.

Tests poco potenti tendono ad essere troppo conservativi, cioè l'ipotesi fondamentale (o nulla) è accettata anche quando dovrebbe valere l'ipotesi alternativa. Pertanto i tests statistici classici sono preferibili, quando le condizioni di validità sono soddisfatte.

In quest'ottica, ipotesi stringenti sulla normalità dei campioni e l'estrema specificità della grandezza da sottoporre al confronto d'ipotesi (coefficienti d'asimmetria e indici di curtosi, quantili nelle code) fanno, dei test di Pearson et al. e Hawkins, alcuni dei più potenti tests noti proprio per questo riportati nel seguito, pur convenendo sulla loro circoscrizione a classi di problemi particolari (ad esempio, individuazione ed eliminazione di dati anomali). Invece quando le condizioni di validità non sono soddisfatte, ad esempio se la distribuzione della popolazione non è quella normale, oppure se gli elementi della popolazione non sono statisticamente indipendenti, oppure se le varianze della popolazione sono significativamente diverse fra loro, allora i tests statistici non-parametrici devono essere utilizzati. Tutto ciò vale in particolare quando i campioni sono piccoli. Infatti una delle condizioni di validità dei tests classici è la dimensione grande dei campioni. L'analisi multivariata di campioni di fenomeni vari richiede spesso:

- il confronto fra medie di campioni aventi diversa varianza;
- il fra varianze di campioni non statisticamente indipendenti;
- il confronto fra contingenze di campioni con distribuzione diversa da quella normale.

In questi casi, i tests statistici non-parametrici sono indispensabili.

[9] Si chiama contingenza la differenza fra una probabilità doppia (o una frequenza relativa doppia) ed il prodotto delle corrispondneti probabilità marginali (o delle corrispondenti frequenze relative marginali). La contingenza fra due variabili casuali (o due variabili statistiche) indipendenti ha valore zero; in corrispondenza ad ogni altro valore (compreso fra -1 e 1), le due variabili si dicono connesse.

- alla legge dei grandi numeri ed al limite centrale della statistica;
- ai teoremi della normalità: conservazione per trasformazioni lineari, identità fra indipendenza ed incorrelazione;
- alla definizione delle variabili casuali χ^2 (chi quadrato), *t* (t di Student) e *F* (F di Fisher);
- al calcolo di probabilità estremali (ad es. di Kolmogorov–Smirnov e di Hawkins), ove necessario;
- al teorema di decomposizione ortogonale degli scarti;

e nello specifico, alla trasformazione della distribuzione di probabilità, secondo requisiti da definirsi, caso per caso, così come è costruito un determinato test multiplo. L'effettiva esecuzione di un qualsiasi test statistico si attua sempre compiendo i seguenti passi:

formulazione di una determinata ipotesi fondamentale (o nulla);

- scelta del livello di significatività;
- costruzione di una statistica campionaria, a partire da dati osservati;
- partizione della distribuzione di probabilità della statistica campionaria;
- effettuazione del confronto d'ipotesi,

avendo cura di controllare la potenza del test, nel caso si voglia prendere in considerazione una o più ipotesi alternative (all'ipotesi fondamentale o nulla).

CONCLUSIONI E FUTURI STUDI

Indice Capitolo 6

a. Metodo di Fourier....1
b. Probabilità....1
c. Conclusioni....3
d. Classi di problemi....3
e. Futuri Studi....7

a. Metodo di Fourier

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli da interpretare sono concepiti come popolazioni di dati ideali, costituenti un universo da cui estrarre dati reali (campionamento) costituiti, a loro volta, da campioni da interpretare. Le principali proprietà delle stime sono: la correttezza, la consistenza, l'efficienza, la sufficienza e la robustezza.

Nel campionamento, numerico e staticamente significativo, di basi di dati di notevoli dimensioni, esistono diverse problematiche collegate alla rappresentabilità dei campione1. è questo il caso per esempio di campionamenti poco significativi o di campionamenti con un numero di dati estremamente elevato. Tale argomento si pone in relazione con la stima delle frequenze, principalmente per quanto riguarda la ricerca dei limiti di confidenza di una frequenza osservata, per una popolazione distribuita binomialmente. Per stabilire la bontà dei modelli d'interpretazione delle osservazioni c'è bisogno di applicare test. I test statistici necessitano di un campionamento di dati significativo e rappresentativo della popolazione da cui è considerato estratto. Queste fatto può comportare l'esigenza di collezionare un numero elevato di estrazioni, prima di poter indicare ipotesi sensate e sopratutto svolgere il test.

Un'alternativa che permette di eseguire l'analisi di evidenze sperimentali quando sono presenti pochi dati è il test sequenziale dato che lo scopo di questo tipo di analisi è quello di arrivare a scegliere, tra ipotesi alternative, con il minimo numero di osservazioni. Parallelamente, la capacità di acquisire grosse moli di dati, durante un processo di misura, da un lato, permette una descrizione più puntuale del fenomeno, dall'altro, richiede maggior attenzione nella scelta del modello stocastico. Questo aspetto, a volte trascurato, può portare seri problemi sull'attendibilità delle stime, quando si ignorano eventuali correlazioni presenti tra le misure. Tuttavia in alcuni insiemi di dati, è possibile evidenziare e successivamente quantificare la correlazione, presente nelle osservazioni, mediante un approccio di tipo stocastico.

Un'altra alternativa alla soluzione dei problemi ed assumendo che i dati siano una sequenza estratta dalla realizzazione di un qualche processo stocastico (segnali stocastici), è fornita dai metodi di Fourier che, qui come altrove, trovano la loro naturale collocazione, in parallelo al metodo degli elementi finiti.

b. Probabilità

Un discorso preciso sulla statistica descrittiva si avvia con la definizione di variabile statistica e di variabile casuale, la postulazione di un'identità formale fra le stesse e la presentazione delle loro principali statistiche. Le variabili statistiche sono il risultato di esperimenti e, pertanto, sono concrete (ovvero costituite da dati reali od osservazioni, come la totalità dei dati a referenza spaziale, tempo varianti e non, quali, ad esempio, le misure geodetiche e geomatiche), finite (perché qualsiasi esperimento incontra evidenti limiti di spazio, tempo ed altre condizioni limitative) e discrete (perché qualsiasi esperimento è eseguito con una determinata accuratezza). Conseguentemente esse sono caratterizzate da un insieme di valori argomentali

(eventualmente raggruppati in classi), associati a frequenze elementari (assolute, come risultato di un conteggio, oppure relative, se la totalità è normalizzata ad uno) ed alle frequenze cumulate delle frequenze elementari.

Le variabili casuali sono modelli interpretativi e, pertanto, sono astratte (ovvero costituite da dati ideali od osservabili) ed, in generale, illimitate e continue (anche se, raramente, ad eccezione della teoria dei giochi, esse possono essere finite e discrete). Conseguentemente esse sono caratterizzate da un campo d'esistenza, associato ad una funzione densità di probabilità ed ad una funzione distribuzione di probabilità (comunemente detta: probabilità). L'identità formale fra variabili statistiche e variabili casuali discende dalla loro completa indistinguibilità, a valle della loro definizione. Allora la presentazione delle principali statistiche può essere eseguita congiuntamente per entrambe.

Per le variabili ad una dimensione, le principali statistiche rispondono alla quantizzazione delle idee di: centro, dispersione, simmetria e curtosi (comportamento delle code). Come noto, il centro può essere indicato tramite la moda, la mediana, le medie (aritmetica, geometrica, armonica, ponderata, potata, ecc.) od altro, la dispersione può essere valutata in base all'ampiezza, ai quantili, alla varianza, agli scarti assoluti medio o mediano, ecc., mentre gli indici di asimmetria e curtosi hanno, solitamente, poche varianti.

Per le variabili a due dimensioni, le principali statistiche (oltre a quelle monodimensionali marginali o condizionate) rispondono alla quantizzazione dell'idea di dipendenza. Come noto, dipendenza è un concetto molto generale che, fra totale e completa indipendenza e perfetta dipendenza (o dipendenza in legge), si articola in connessione (dipendenza vaga e generica), regressione (quasi-dipendenza funzionale) e correlazione (quasi-dipendenza lineare). Ancora numerosi sono gli indici ed i coefficienti che esprimono il grado della dipendenza o meno (si noti, a riguardo, come tutti siano normalizzati ad uno, assumendo anche valori negativi, fino a meno uno, se non intrinsecamente positivi).

Per le variabili a più di due dimensioni, a rigore, occorre continuare lo studio del loro raggrupparsi (come con gli indici di nuvolosità, ecc.). Tuttavia nel caso frequente in cui il modello interpretativo è fornito dalla variabile casuale normale, questo studio è del tutto superfluo. Si ricordi, inoltre, che detta variabile casuale è completamente caratterizzata dal vettore delle medie e dalla matrice di varianza-covarianza, cosa che rende superflue altre statistiche del centro, della dispersione e della dipendenza (covarianza comporta correlazione, ovvero dipendenza lineare e niente altro) e del tutto inutili le statistiche superiori (la variabile casuale normale è simmetrica e l'indice di curtosi vale, in ogni caso, tre).

Ulteriori vantaggi dell'adozione, quale modello interpretativo, della variabile casuale normale sono dati dall'invarianza della distribuzione di probabilità di detta variabile casuale, rispetto a trasformazioni lineari della variabile casuale stessa, e dell'ottimalità della stima dei parametri di modelli, supportati dalla variabile casuale normale, se le ipotesi di corrispondenza fra dati e modelli sono perfettamente soddisfatte (ovvero se i dati non sono affetti, in alcun modo, da dati anomali). A tutto ciò, si aggiunge la linearità dei sistemi da risolvere per la stima dei parametri di modelli, fatto di primaria importanza, in quanto solo i sistemi lineari ammettono, senza eccezioni e purché non-singolari, soluzioni esattamente determinabili, indipendentemente dal numero di equazioni ed incognite di cui si compongono.

La statistica descrittiva termina con alcuni teoremi limite. Fra questi il teorema di Bernoulli[1] (o legge dei grandi numeri) mostra la convergenza, in probabilità, delle frequenze di una variabile statistica alle probabilità di una corrispondente variabile casuale, mentre il teorema di Gauss[2] (o limite centrale della statistica) mostra la convergenza, in legge, della combinazione lineare (ovvero delle somme, come caso particolare) di variabili casuali qualsiasi, purché aventi ciascuna dispersione comparabile con le altre, alla variabile casuale normale. I due teoremi giustificano, rispettivamente, la comparazione fra variabili

[1]

[2] la somma (normalizzata) di un grande numero di variabili casuali è distribuita approssimativamente come una variabile casuale normale standard

statistiche e variabili casuali, al di là della sopracitata identità formale, e le operazioni di media aritmetica, ponderata o potata fra osservazioni dirette per aumentare la normalità del comportamento dei dati.

c. Conclusioni

La teoria della stima ha lo scopo di definire proprietà, caratteristiche e modalità della stima dei parametri di modelli, dove questi modelli, da interpretarsi, come già detto, quali variabili casuali, sono concepiti come popolazioni di dati ideali (od osservabili), costituenti un universo da cui estrarre dati reali (od osservazioni), costituenti, a loro volta, campioni da interpretare, pertanto, come le suddette variabili statistiche. L'operazione d'estrazione è detta campionamento[3] e, se i dati sono fra loro indipendenti, come avviene, auspicabilmente, nelle osservazioni dirette di fenomeni, il campionamento è detto bernoulliano; schemi di campionamento più complessi, attinenti alle problematiche della progettazione, simulazione ed ottimizzazione, sono considerati estranei agli scopi del presente lavoro.

Le principali proprietà delle stime sono la correttezza, la consistenza, l'efficienza e la sufficienza che, in base al significato letterale dei nomi, significano capacità di stimare parametri il cui centro coincide con il centro dei parametri dell'intera popolazione, capacità di stimare parametri con precisione ed accuratezza sempre maggiori, ed al limite infinite, al crescere della numerosità del campione, capacità di stimare parametri qualitativamente migliori delle informazioni presenti direttamente nelle osservazioni, capacità di stimare parametri conservando tutta la ricchezza di informazioni già presente nelle osservazioni dirette. Una proprietà aggiuntiva delle stime, estranea al corpus della statistica tradizionale, è la robustezza, intesa come capacità di stimare parametri indistorti, nonostante la presenza di eventuali dati anomali.

Le caratteristiche delle stime forniscono spesso, contemporaneamente, le modalità operative per effettuare le stime stesse. Infatti sono operativi tanto il metodo della minima varianza, ottimale per la statistica tradizionale, quanto diverse procedure robuste, certamente subottimali, ma capaci di evitare indesiderate distorsioni nelle stime. Altrettanto operativi sono il metodo della massima verosimiglianza[4] ed il metodo dei minimi quadrati, una particolarizzazione del primo nel caso in cui le stime avvengono in ambito lineare ed il modello interpretativo è fornito dalla variabile casuale normale. Si ricordi che, in questo caso, si ha l'ottimalità delle stime, in quanto tanto il metodo dei minimi quadrati[5], quanto quello della massima verosimiglianza, da cui discende, danno risultati perfettamente coincidenti con il metodo della minima varianza.

Tutto ciò conferma l'adozione del metodo dei minimi quadrati per il trattamento statistico delle osservazioni, agevola le sue generalizzazioni ed estensioni a tecniche complementari e giustifica un modo di procedere che prevede la centralità di detto metodo e riconduce ad esso, per quanto possibile, importanti tecniche complementari (cluster analysis, regressione multipla, analisi di varianza, delle componenti di varianza e della struttura di covarianza, procedure robuste). Si noti, in quest'ambito, il ruolo fondamentale ed indispensabile giocato dall'indissolubilità del legame fra un certo tipo di statistiche classiche elementari, la normalità e la linearità, per quanto riguarda tanto la definizione statistica delle metodologie, quanto la loro applicazione con elevate capacità risolutive in appropriati algoritmi numerici.

d. Classi di problemi

I problemi ai minimi quadrati si presentano, in generale e nell'ambito specifico delle discipline geodetiche e geomatiche, usualmente ripartiti in due classi fondamentali:

[3] Per campione se intende parte della popolazione che viene selezionata per l'analisi.
[4] Il metodo della massima verosimiglianza in statistica è un procedimento matematico per determinare uno stimatore.
[5] Il metodo dei minimi quadrati generalizzati di Aitken consente la stima di un modello lineare, sotto ipotesi più generali di quelle del modello classico di regressione lineare multivariata.

- problemi reticolari (o di compensazioni di reti);
- problemi d'interpolazione ed approssimazione di campi di punti.

Le stesse due classi si incontrano anche in problemi affini, quali ad esempio:

- i campionamenti delle osservazioni, l'ottimizzazione della configurazione di rilevamento e/o dello schema di misura, oppure dei pesi delle osservazioni;
- la "cluster analysis", l'analisi di varianza, la regressione multipla, l'analisi fattoriale (o studio delle componenti principali);
- lo studio dell'affidabilità delle osservazioni e le procedure di validazione dei dati e di stima dei parametri con procedure robuste.

In ogni caso, tutti i problemi minimi quadrati possono essere interpretati, topologicamente, come un grafo, dove:

- le osservazioni, i vincoli e le pseudo-osservazioni sovrappesate e non, come pure le informazioni a priori, le osservabili secondarie e le condizioni numeriche di regolarizzazione, costituiscono i lati del grafo;
- i parametri principali ed ausiliari (o di servizio) costituiscono i nodi dello stesso grafo.

Si noti, a riguardo, come l'interpretazione data della topologia sia indispensabile per una corretta comprensione dei casi e sottocasi in cui si articolano le suddette classi fondamentali. Le due classi fondamentali, già precedentemente enunciate, si articolano in vari e svariati casi e sottocasi, illustrati, dettagliatamente, nel prosieguo. I problemi reticolari (o di compensazione di reti) presentano come osservabili:

- differenze prime dei parametri;
- funzioni delle differenze prime dei parametri.

Il primo caso ha numerosi esempi, anche fuori dalle discipline geodetiche e geomatiche:

- discretizzazione di equazioni differenziali del primo ordine, tipiche della fisica, della chimica e delle scienze della terra;
- problemi di trasporto: schemi di circuitazione, traffico, circolazione e transazione, reti di comunicazione, distribuzione e telecomunicazione ed è costituito, per le suddette discipline, dalle reti di differenza di potenziale.

Il secondo caso è, invece, tipico delle discipline geodetiche e geomatiche, anche se non esclusivo (a questo caso, infatti, fanno riferimento ben particolari discretizzazioni di equazioni differenziali, sempre riferite ai sopracitati raggruppamenti di discipline fisiche e naturalistiche), e si articola nei seguenti sottocasi:

- l'informazione fluisce completa, bidirezionalmente, come nelle reti di differenza di potenziale, lungo ogni lato del grafo:
- l'informazione fluisce completa, unidirezionalmente, lungo ogni lato del grafo, costituendo nel suo fluire almeno un albero sul grafo stesso;
- l'informazione è irradiata, in modo completo, da alcuni nodi verso altri (senza ritorno), senza che né i primi, né i secondi si scambino alcuna informazione, costituendo nel suo fluire tanti alberi (costituiti da un solo livello, oltre la radice) sul grafo stesso, quanti sono i nodi d'emanazione:
- l'informazione è irradiata, in modo parziale, nelle stesse condizioni del sottocaso precedente, cosa che richiede l'individuazione di due o più co-alberi (sempre costituiti da un solo livello, oltre le radici) capaci di completare l'informazione trasmessa;
- l'informazione è irradiata, in modo parziale, senza restituzioni particolari.

Le osservabili differenze seconde dei parametri e loro funzioni richiedono la complessa sostituzione dei lati del grafo con triangoli fra i tre nodi interessati. Le osservabili differenze di ordine superiore e loro funzioni fanno riferimento, addirittura, a poligoni fra tutti i nodi coinvolti, cosa che rende la loro analisi ancora più complessa. Per queste ragioni, ad eccezione della discretizzazione di equazioni differenziali di secondo ordine o di ordine superiore e di loro trasformazioni funzionali, la loro adozione è estremamente rara. Per quanto riguarda la determinazione del numero di parametri principali, nel caso in cui i problemi ai minimi quadrati adottino lo schema principe delle equazioni d'osservazione, questo è sempre tale da determinare difetti di rango e singolarità del sistema da risolvere per cui sono necessari vincoli o pseudo-osservazioni sovrappesate.

I problemi d'interpolazione ed approssimazione di campi di punti presentano come osservabili funzioni dirette dei parametri principali, il cui numero, sempre nel caso in cui si voglia adottare il suddetto schema principe delle equazioni d'osservazione, non è mai tale da determinare difetti di rango e singolarità del sistema da risolvere. A tutto ciò, fanno eccezione eventuali problemi di sovra-parametrizzazione, rispetto al campionamento delle osservazioni effettuate, per cui sono indicate condizioni numeriche di regolarizzazione.

Esempi di problemi di interpolazione ed approssimazione di campi di punti sono dati da:

- ricostruzione (fitting) di linee, superfici, ipersuperfici aventi come dominio lo spazio 3D:
- descrittori di forma (form descriptors): contorni di figure (piane e/o gobbe), superfici (chiuse) di oggetti:
- centratura (matching) di segmenti, figure (immagini, mappe, disegni), oggetti (compresi modelli virtuali 3D) comunque conformati:

I problemi d'interpolazione ed approssimazione[6] di campi di punti, relativi alla ricostruzione di linee sono, ovviamente, assimilabili a quelli dello studio delle serie temporali storiche o di breve periodo, oppure frutto di simulazioni. Inoltre lo studio di serie temporali congiunto alla soluzione dei problemi reticolari (o di

[6] metodo per individuare nuovi punti del piano cartesiano a partire da un insieme finito di punti dati, nell'ipotesi che tutti i punti si possano riferire ad una funzione f(x) di una data famiglia di funzioni di una variabile reale.

compensazione di reti) e/o d'interpolazione ed approssimazione di campi di punti, illustrati in precedenza, permette indagini accurate sugli aspetti dinamici delle osservabili a referenza spaziale di cui ai suddetti problemi, dando un'interpretazione unitaria a dati spazio-varianti, tempo-varianti.

Limitatamente alle discipline geodetiche e geomatiche, mentre le equazioni d'osservazione dei problemi reticolari (e di compensazioni di reti) fanno uso, in generale, di modelli grigi dedotti dalla geometria del problema in esame, le equazioni d'osservazione dei problemi d'interpolazione ed approssimazione di campi di punti fanno uso, in generale, di modelli neri.

Come noto, una vasta gamma di metodi deterministici e/o stocastici risponde positivamente alla bisogna. I primi annoverano fra i più comunemente impiegati:

- l'interpolazione polinomiale;
- il metodo degli elementi finiti e l'interpolazione con funzioni splines;
- l'analisi di Fourier, nel dominio delle frequenze;
- lo studio, sempre nel dominio delle frequenze, con ondine (wavelets).

I secondi prevedono l'interpretazione dei fenomeni in istudio come realizzazioni di un processo stocastico:

- stime di covarianza, filtraggio cross-validazione e predizione;
- studio della geometria frattale.

Uno studio dettagliato di esempi particolari e significativi di problemi reticolari (o di compensazioni di reti) può essere effettuato, nell'ambito delle discipline geodetiche e geomatiche, solo facendo riferimento a discipline specifiche, quali la geodesia, la navigazione, la topografia, la fotogrammetria ed il telerilevamento. Al contrario, uno studio dettagliato di esempi particolari e significativi di problemi d'interpolazione ed approssimazione di campi di punti richiede anche uno studio dei modelli neri. Per una migliore comprensione si tenga presente che l'insieme delle quantità osservate è sempre costituito da quattro parti distinte:

- le informazioni topologiche, ovvero i lati del grafo che indicano le connessioni esistenti fra i nodi del grafo stesso;
- le informazioni geometriche, ovvero la posizione ed altre caratteristiche degli stessi nodi;
- le informazioni metrologiche, ovvero le osservazioni (o quantità osservate) realmente effettuate;
- le informazioni stocastiche, ovvero la precisione delle osservazioni e le eventuali correlazioni fra queste.

Infatti questo insieme, altrimenti detto: base di dati provenienti da operazioni di misura, con riferimento a ciascuna delle sopraccitate quattro parti distinte, produce nei problemi ai minimi quadrati (come pure negli altri sopraccitati problemi affini), rispettivamente:

- la matrice disegno simbolica;
- la matrice disegno numerica;
- il vettore termine noto delle equazioni d'osservazione;
- la matrice di varianza-covarianza (a priori) delle quantità osservate o, più comunemente, se non esistono correlazioni fra le stesse quantità osservate, la matrice dei pesi.

Si noti che, con la sola eccezione delle osservazioni realmente effettuate, tutto quanto può essere noto già prima di compiere una sola osservazione. Da ciò derivano tutti i problemi di ottimizzazione della matrice di varianza-covarianza dei parametri:

- intervenendo nella matrice disegno per decidere sull'effettuazione o meno di ciascuna osservazione (1° ordine);
- sulla matrice di varianza-covarianza (a priori) delle quantità osservate per stabilire, note le osservazioni da effettuarsi, le precisioni delle stesse (2° ordine);
- su opportune parziali combinazioni dei due casi precedenti (3° ordine),

avendo cura di controllare, in ogni caso, l'affidabilità delle osservazioni, quale garanzia, sufficiente minimale, che il lavoro intrapreso, qualsiasi esso sia, risulti svolto a regola d'arte.

e. Futuri Studi

La validazione dati e dei modelli si fonda sulle varie tecniche dell'analisi multivariata[7], di volta in volta, studiando la variabilità e l'interdipendenza fra gli attributi, entro una classe di oggetti. Essa prende in considerazione insiemi di dati, ciascuno dei quali, relativo ad un oggetto della classe, contiene i valori osservati di certe variabili statistiche. Questi insiemi possono, talvolta, essere completi; mentre, più A loro volta, le variabili osservate, sono in generale campioni estratti da variabili casuali, di tipo continuo o, raramente, discreto. Da una tale complessità di premesse, lo studio dell'analisi multivariata si può articolare, principalmente, nei seguenti punti.

- Semplificazione strutturale. Gli insiemi di dati devono essere ricondotti, se possibile, in forme più semplici con cambi di variabili, in particolare, con trasformazioni capaci di sciogliere variabili connesse in variabili indipendenti.
- Classificazione degli oggetti. L'analisi dell'insieme di dati deve porre in evidenza la presenza di gruppi (clusters), ovvero di sottoinsiemi di oggetti, caratterizzati da valori preferenziali degli attributi o di parte di essi, cercando di ricondurre a poco le notevoli variabilità presenti.
- Raggruppamento degli attributi (clustering). L'analisi degli insiemi di dati deve far ricadere, per quanto possibile, differenti variabili in un unico gruppo.
- Analisi della connessione. Gli insiemi di dati devono essere studiati rispetto alla dipendenza vaga e generica o meno fra le variabili contenute (ovvero all'essere in connessione di queste ultime).

[7] Con statistica multivariata s'intende quella parte della statistica in cui l'oggetto dell'analisi è per sua natura formato da almeno due componenti.

- Analisi della dipendenza funzionale. Gli insiemi di dati devono essere studiati rispetto alla dipendenza funzionale o meno fra le variabili contenute (ovvero all'essere in regressione di queste ultime), con particolare riferimento alla dipendenza lineare o correlazione (ovvero all'essere in regressione lineare o correlate).
- Costruzione e verifica d'ipotesi. Il confronto probabilistico, fra statistiche campionarie e valori teorici di riferimento, permette di formulare un giudizio critico sui risultati ottenuti nelle varie tappe dell'analisi multivariata.

I tests di validazione dei modelli permettono di sottoporre a verifica, mediante opportuni controlli e confronti d'ipotesi, le stime effettuate come, del resto, tutti i risultati ottenuti nell'ambito della statistica. Al solito, si possono avere errori nel modello deterministico: presenza di errori grossolani nelle osservazioni, ed errori nel modello stocastico: presenza di errori sistematici nelle osservazioni, ovvero cattiva conoscenza delle varianze delle osservazioni e/o delle eventuali covarianze fra le osservazioni stesse. Un'opportuna sequenza di tests permette di districarsi fra le varie cause d'errore.

Un giudizio sui risultati può essere espresso in termini numerici e statistici. I controlli di tipo numerico rispondono a problemi di condizionamento ed affidabilità che comunemente accompagnano e seguono il metodo dei minimi quadrati, comprensivo delle sue estensioni e generalizzazioni; pertanto tutto quanto riguarda i controlli di tipo numerico è considerato estraneo agli scopi del presente lavoro. I secondi comprendono i tests statistici per la valutazione di osservazioni e parametri, della loro dispersione e, se del caso, della loro dipendenza. Per quanto riguarda le osservazioni, la validazione avviene in termini di entità degli scarti - residui (oltreché numericamente in termini di affidabilità delle osservazioni all'interno dello schema di misura) e consiste essenzialmente nell'individuazione ed eliminazione degli errori grossolani.

L'inferenza statistica (multivariata) è quella parte dell'analisi multivariata dedicata alla costruzione e verifica d'ipotesi (per problemi di controllo di qualità, oppure controllo e confronto d'ipotesi di altri problemi). Infatti il confronto probabilistico permette di formulare un giudizio critico sui risultati ottenuti, nelle varie tappe dell'analisi multivariata. Così con le usuali strategie dell'inferenza statistica, vari tests multipli consentono di discriminare, tanto stime di parametri da campioni normali, quanto statisitche di modelli non-parametrici (ovvero modelli distribution free)[8].

[8] L'aggettivo non parametrico (in letteratura inglese: distribution free) qualifica un particolare gruppo di tests statistici, sotto certe condizioni, sostitutivo dei tests statistici classici. Infatti i tests non-parametrici, rispetto ai test classici, presentano i seguenti vantaggi:

- la loro comprensione è immediata ed elementare;
- le condizioni di validità sono meno forti (più ampie);
- i calcoli necessari non presentano in generale difficoltà computazionali.

D'altra parte, i tests non-parametrici presentano alcuni svantaggi:

- molta informazione viene sprecata;
- la potenza del test è bassa.

Tests poco potenti tendono ad essere troppo conservativi, cioè l'ipotesi fondamentale (o nulla) è accettata anche quando dovrebbe valere l'ipotesi alternativa. Pertanto i tests statistici classici sono preferibili, quando le condizioni di validità sono soddisfatte.

In quest'ottica, ipotesi stringenti sulla normalità dei campioni e l'estrema specificità della grandezza da sottoporre al confronto d'ipotesi (coefficienti d'asimmetria e indici di curtosi, quantili nelle code) fanno, dei test di Pearson et al. e Hawkins, alcuni dei più potenti tests noti proprio per questo riportati nel seguito, pur convenendo sulla loro circoscrizione a classi di problemi particolari (ad esempio, individuazione ed eliminazione di dati anomali). Invece quando le condizioni di validità non sono soddisfatte, ad esempio se la distribuzione della popolazione non è quella normale, oppure se gli elementi della popolazione non sono statisticamente indipendenti, oppure se le varianze della popolazione sono significativamente diverse fra loro, allora i tests statistici non-parametrici devono essere utilizzati. Tutto ciò vale in particolare quando i campioni sono piccoli. Infatti una delle condizioni di validità dei tests classici è la dimensione grande dei campioni. L'analisi multivariata di campioni di fenomeni vari richiede spesso:

Gli oggetti del giudizio critico sono, come detto, i risultati ottenuti nelle varie tappe dell'analisi multivariata, in particolare: frequenze relative, contingenze, medie campionarie o altri indicatori del centro (di una popolazione), varianze campionarie o altri indicatori della dispersione, coefficienti di correlazione campionari o altri indicatori della correlazione[9].

I test multipli si differenziano per i diversi oggetti in esame, come pure per la distribuzione di appartenenza della popolazione cui i campioni si riferiscono, se normale, oppure altra (spesso sconosciuta). In generale, comunque, tutti i campioni sono supposti fra loro indipendenti, mentre se quest'ipotesi non è soddisfatta, occorre procedere, come per i modelli non-parametrici, adottando strategie assolutamente generali, ma assai poco potenti, cioè meno capaci di discriminare fra ipotesi alternative vicine, a parità di numerosità dei campioni. Nelle seguenti tabelle, si presentano parecchi tests multipli, diversamente, rispondenti alla bisogna, facendo attenzione, in particolare, ai problemi pratici, legati alla loro applicazione ad esempi concreti. Per quanto riguarda, invece, i loro fondamenti teorici, questi si richiamano, in generale:

- alla definizione assiomatica di probabilità;
- alla legge dei grandi numeri ed al limite centrale della statistica;
- ai teoremi della normalità: conservazione per trasformazioni lineari, identità fra indipendenza ed incorrelazione;
- alla definizione delle variabili casuali χ^2 (chi quadrato), t (t di Student) e F (F di Fisher);
- al calcolo di probabilità estremali (ad es. di Kolmogorov–Smirnov e di Hawkins), ove necessario;
- al teorema di decomposizione ortogonale degli scarti;

e nello specifico, alla trasformazione della distribuzione di probabilità, secondo requisiti da definirsi, caso per caso, così come è costruito un determinato test multiplo. L'effettiva esecuzione di un qualsiasi test statistico si attua sempre compiendo i seguenti passi:

- formulazione di una determinata ipotesi fondamentale (o nulla);
- scelta del livello di significatività;
- costruzione di una statistica campionaria, a partire da dati osservati;
- partizione della distribuzione di probabilità della statistica campionaria;
- effettuazione del confronto d'ipotesi,

avendo cura di controllare la potenza del test, nel caso si voglia prendere in considerazione una o più ipotesi alternative (all'ipotesi fondamentale o nulla).

- il confronto fra medie di campioni aventi diversa varianza;
- il fra varianze di campioni non statisticamente indipendenti;
- il confronto fra contingenze di campioni con distribuzione diversa da quella normale.

In questi casi, i tests statistici non-parametrici sono indispensabili.

[9] Si chiama contingenza la differenza fra una probabilità doppia (o una frequenza relativa doppia) ed il prodotto delle corrispondneti probabilità marginali (o delle corrispondenti frequenze relative marginali). La contingenza fra due variabili casuali (o due variabili statistiche) indipendenti ha valore zero; in corrispondenza ad ogni altro valore (compreso fra -1 e 1), le due variabili si dicono connesse.

BIBLIOGRAFIA

Barrile V., Bellone T., Mussio L. (1999): Trattamento di osservazioni a referenza spaziale e/o di serie temporali nei problemi di controllo. Geoingegneria ambientale e mineraria, Rivista dell'Associazione Georisorse e Ambiente, n. 1, 1999.

Bochner S., Chandrasekharan K. (1949), Fourier Transforms, Princeton University Press.

Bracewell, R. N. (2000), The Fourier Transform and Its Applications (3rd ed.), Boston: McGraw-Hill.

Campbell, George; Foster, Ronald (1948), Fourier Integrals for Practical Applications, New York: D. Van Nostrand Company, Inc. .

Fava G., Mussio L., Paolucci R. (2004): Applicazione di tecniche spettrali per la valutazione della sicurezza strutturale. Atti della Conferenza Standardizzazine, Interoperabilità e Nuove Tecnologie – 8° Conferenza Nazionale dell'ASITA, vol. 2. Roma, p. 1047-1052

Fava G., Paolicci R., Higashihara I., Autoregressive time series analysis and its application to ambient vibration records - Tesi datt., Milano : Politecnico, 2002/2003. - VII, 162 , [18] c. ; 30 cm. 1. Fac. di ingegneria, Laurea in ingegneria civile. - Matr. 635852.

Rotondi A., Pedroni P., Pievatolo A. Probabilita Statistca e Simulazione. Programmi applicativi scriti con Scilab, Milano, Springer-Verlag Italia, 2005.

Torabi H. and Mirhosseini S. M. , Sequential Probability Ratio Tests for Fuzzy Hypotheses Testing , Applied Mathematical Sciences, Vol. 3, 2009, no. 33, 1609 – 1618 .