Nota Statistica , Versione 1

Appunti di Analisi Statistica dei Dati

Fernando Palombo
Dipartimento di Fisica dell'Università and INFN, Milano

Chapter 1

Nozioni Introduttive

- Misure Sperimentali
- Estrarre Informazioni dai Dati Sperimentali
- Statistica Descrittiva e Statistica Inferenziale
- Probabilità
- Variabili Casuali

1.1 Misure Sperimentali

In fisica sperimentale noi facciamo esperimenti per misurare determinate grandezze fisiche (viscosità di un liquido, radioattività naturale in una determinata zona, misure di sezioni d'urto in fisica subnucleare, ecc). È noto che queste misure sono sempre affette da errori: il valore della misura non è prevedibile. Noi diciamo che la misura di una grandezza fisica è una variabile casuale.

Talvolta le misure sperimentali servono a verificare relazioni tra grandezze fisiche , relazioni ipotizzate o fornite dalla teoria fisica. Ad esempio la legge fondamentale del decadimento radioattivo è :

$$N = N_0 exp(-\lambda t)$$

dove N è il numero (grande) di nuclei radiattivi al tempo t e N_0 è il numero di nuclei radioattivi al generico istante iniziale t=0.

Noi possiamo esprimere la quantità misurata N come espressione matematica della proprietà da misurare λ (costante di decadimento dei nuclei considerati) . In questo modo possiamo confrontare le misure con un modello teorico.

Noi possiamo scegliere tra due ipotesi (per esempio segnale - rumore) (o eventualmente tra più ipotesi) e stimare l'errore che facciamo a scegliere una o l'altra ipotesi.

Le operazioni di misura devono essere chiaramente specificate (unità di misura, scale , approssimazioni, procedimenti di misura , ecc). Dalle misure fatte noi vogliamo estrarre informazioni sulla grandezza o relazione che stiamo studiando.

Mi sono riferito a misure in Fisica ma ovviamnete le misure possono essere le più svariate (in chimica, biologia, medicina, ecc).

La statistica è un ramo della matematica applicata . Tecniche statistiche per estrarre informazioni da dati sperimentali sono diventate di base praticamnete in ogni settore dell'attività umana. Il controllo di qualità di un determinato prodotto è fatto su basi statistiche analizzando campioni di questo prodotto, prima di fare un investimento una società controlla l'andamento delle vendite , la possibilità di espansione del mercato, i risultati preliminari in una elezione si fanno considerando campioni di persone che hanno votato, etc .

Le tecniche di analisi statistica sono numerose ed il loro utilizzo dipende dalle discipline e settori di applicazione. Noi introdurremo alcune tecniche statistiche di analisi dati comunemente usate in Fisica e che sono di generale applicazione in moltissimi altri campi.

1.2 Statistica Descrittiva e Statistica Inferenziale

La statistica si occupa dei procedimenti scientifici che permettono, fatta una serie di misure, di ordinare , presentare ed analizzare i dati, estraendone l'informazione.

La statistica descrittiva riguarda la classificazione e la sintesi delle informazioni relative ad un determinato campione di dati oggetto di studio. Dato un campione di misure, posso dare il grafico di una quantità caratteristica del campione, calcolare media, varianza , correlazioni tra le variabili, ecc. In modo conciso vengono sintetizzati i dati con pochi numeri o grafici. La statistica inferenziale utilizza il campione di dati per fare previsioni di tipo probabilistico sulla popolazione da cui il campione è stato estratto. Nella teoria e pratica delle misure , la statistica inferenziale è senza dubbio quella di maggiore interesse. Le aree prindipali dell'inferenza statistica sono due : la stima di parametri e la verifica di ipotesi. L'inferenza può essere induttiva o deduttiva

- **Inferenza Induttiva**

  Due esperimenti diversi che misurano la stessa quantità fisica avranno misure diverse (ed in generale un numero diverso di misure). Nel caso più elementare un esperimento misura una certa grandezza fisica 100 volte , un altro esperimento misura la stessa grandezza 1000 volte. Entrambi gli esperimenti misurano quella grandezza fisica a partire da un numero finito di misure. È chiaro che, solo considerando gli errori statistici, la misura più precisa è quella basata su un numero maggiore di misure. La misura più precisa sarebbe quella basata su un numero infinito di misure . La totalità degli elementi che sono oggetto della nostra indagine statistica è detta popolazione. I due esperimenti considerati nel nostro esempio rappresentano due campioni della stessa popolazione. Popolazione può essere ad esempio l'intera produzione di chip di una industria elettronica in un determinato periodo. Per controllare se il prodotto è stato realizzato con le richieste caratteristiche facciamo un controllo su un campione di chip prodotti. I risultati ottenuti dall'analisi di questo campione vengono attribuiti all'intera produzione (la popolazione). Quindi per esempio una certa produzione non viene immessa nel mercato se il controllo di qualità sul campione non mostra i requisiti minimi di qualità richiesti. Osserviamo che l'analisi statistica sull'intera popolazione è generalmente impossibile o poco pratica. Si pensi ad esempio al controllo di qualità di un prodotto che spesso è distruttivo o all'analisi di indagini su un campione per decidere cosa applicare alla popolazione o quali decisioni prendere o alla analisi di misure sperimentali che possono essere di numero infinito. Questo procedimento di passaggio dal particolare al generale è noto come inferenza induttiva. La generalizzazione non è mai assolutamente certa. Però l'analisi statistica permette di fare delle inferenze induttive e di misurare il loro grado di incertezza. Il campione deve essere casuale (vedremo dopo cosa significa) .

- **Inferenza Deduttiva**

Con l'inferenza deduttiva si deducono informazioni da altre accettate come vere. Ad esempo: 1) Ogni triangolo rettangolo ha un angolo interno di 90^0; 2) Il triangolo A è un triangolo rettangolo. Da queste due asserzioni per inferenza deduttiva concludo che il triangolo A ha un angolo interno di 90^0. Mentre le conclusioni dell'inferenza induttiva sono solo probabili, quelle dell'inferenza deduttiva sono definitive. L'inferenza deduttiva è molto usata in Matematica per dimostrare i teoremi.

1.3 Probabilità

L'impostazione assiomatica della teoria delle probabilità è dovuta a Kolmogorov (1933). Questa teoria si occupa di entità astratte che , nella fase di sviluppo della teoria, non necessitano di alcuna interpretazione.

Supponiamo di fare un esperimento. L'insieme di tutte le possibili misure in questo esperimento costituisce lo **spazio campione S**. Evento è un sottoinsieme di S. Un **evento semplice** è un evento che non può essere unione di altri eventi. Un evento **composto** è un evento che non è semplice. Ad ogni evento A di S associamo un numero reale P(A), detto probabilità di A , definito da questi assiomi:

1) Per ogni evento A in S, $P(A) \geq 0$

2) $P(S) = 1$

3) Se due generici eventi A e B sono mutuamente disgiunti ($A \cap B = \emptyset$, probabilità nulla che avvenga A e B) allora la probabilità che avvenga A oppure B è la somma delle corrispondenti probabilità: $P(A \cup B) = P(A) + P(B)$

Questi 3 assiomi definiscono in matematica la teoria assiomatica della probabilità. Si possono dedurre facilmente alcune proprietà della probabilità:

- Se due eventi A e $\overline{A}$ sono complementari ($A \cup \overline{A} = S$), allora $P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$. La proabilità $P(\overline{A})$ che non si verifichi A è uguale a 1 meno la probabilità P(A) che si verifichi A.

- $0 \leq P(A) \leq 1$

- $P(\emptyset) = 0$ (evento che non si può realizzare)

- Se $A \subset B$, allora $P(A) \leq P(B)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Esempio: lancio una moneta due volte.

Spazio degli eventi : TT, CT, TC,CC (l'ordine testa (T) , croce (C) conta!). Evento in cui la testa appare almeno una volta :

$$TT \cup CT \cup TC$$

1. **Combinazioni**

   In pratica spesso la probabilità di un evento composto si può ottenere come somma delle probabilità di tutti gli eventi semplici che lo costituiscono come si può dedurre dal postulato 3. Questo è particolarmente semplice quando si ha un numero finito di eventi semplici, tutti con eguale probabilità . Per esempio qual è la probabilità che si abbia un numero pari lanciando un dado ? Il numero di eventi semplici possibili è n(S)= 6. L'evento favorevole A si rializza con A = {2,4,6 }. Il numero di casi favorevoli è n(A)= 3. Quindi la probabilità di avere un numero pari è : P(A) = n(A)/n(S) = 0.5.

   In queste situazioni è centrale il concetto di combinazioni. Consideriamo n oggetti (tutti distinguibili , cioè diversi). Supponiamo che si trovino in una scatola. Estraiamo dalla scatola r di questi oggetti, uno alla volta (senza rimetterli nella scatola)(Combinazioni senza ripetizione di n oggetti di classe r). In quanti modi diversi riusciamo a fare questo ?

   $$n_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} = D_{n,r}$$

   Questo perchè la prima volta si sceglie su n oggetti, la seconda volta si sceglie su (n-1) oggetti e cosi via. Qui :

   $$k! \equiv k(k-1)(k-2)\cdots 1 \qquad 0! \equiv 1$$

   Si chiamano disposizioni senza ripetizione di n oggetti di classe r. Queste disposizioni differiscono tra di loro sia perchè contengono oggetti differenti sia perchè stesi oggetti si susseguono in ordine diverso.

   In quanti modi diversi è possibile estrarre r oggetti dagli n che si trovano nella scatola ? Vuol dire che prendiamo gli stessi oggetti indipendentemente dall'ordine in cui sono presi. Per avere questo numero devo dividere $D_{n,r}$ per il numero delle permutazioni degli r oggetti , cioè per r! Questo numero è detto coefficiente binomiale e si indica con $\binom{n}{k}$.

   Il numero di volte in cui è possibile pigliare gli stessi r oggetti è r!. Infatti alla prima presa prendi uno qualsiasi degli r oggetti; alla seconda presa, uno qualsiasi dei rimanenti r-1 ,ecc. Il coefficiente binomiale è il rapporto tra il numero n_r diviso il numero r!, quindi:

$$\binom{n}{r} = \frac{D_{n,r}}{r!} = \frac{n!}{(n-r)!r!}$$

Il nome deriva dal fatto che :

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r$$

Il numero di combinazioni di 3 oggetti di classe 2 è $\binom{3}{2}$: 3!/(3-2)! 2! = $\frac{3 \cdot 2}{2} = 3$

Con un mazzo di carte di bridge (52 carte) il numero di possibili mani è $\binom{52}{13}$. La probabilità di 5 quadri, 5 picche, 2 cuori e un fiori è :

$$\frac{\binom{13}{5}\binom{13}{5}\binom{13}{2}\binom{13}{1}}{\binom{52}{13}}$$

2. **Probabilità Condizionale** : Siano A e B due eventi di S e supponiamo che P(B) sia $\neq 0$. Definiamo probabilità condizionale P(A | B), la probabilità che si avveri A supponendo avvenuto B (diciamo probabilità P di A dato B) come :

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

I due eventi A e B si dicono statisticamente (o stocasticamente) indipendenti se P(A $\cap$ B) = P(A) $\cdot$ P(B). Se A e B sono eventi indipendenti, allora P(A | B) = P(A) e P(B | A) = P(B).

Se lancio una moneta ed ho testa, il lancio di un'altra moneta ha un risultato (testa o croce) che non dipende dal risultato del lancio della prima moneta . La nozione di indipendenza di due eventi si generalizza a quella di più eventi mutuamente indipendenti. La nozione di indipendenza degli eventi ha un ruolo importante nella teoria della probabilità e nelle sue applicazioni.

Esempio: Un dado viene lanciato due volte. Nell'ipotesi che si sappia che il punteggio totale dei due lanci è 6, qual è la probabilità che il punteggio del primo lancio sia 3 ?

Diciamo evento A "punteggio totale uguale a 6" ed evento B " punteggio primo lancio uguale a 3". Allora si hanno 36 possibili eventi, n(S)= 36 e

$$A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$$

$$B = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$$

Di conseguenza : $A \cap B$= (3,3) , P(A)= 5/36, P($A \cap B$)=1/36 e quindi P(B|A)=1/5

Esempio : Da un mazzo di 36 carte (con 4 assi) tiriamo due carte in successione.Trovare 1) la probabilità (non condizionale) di trovare come seconda carta un asso (non so qual è stata la prima carta) ; 2) la probabilità condizionale di trovare come seconda carta un asso sapendo che la prima era un asso.

evento A : avere come seconda carta un asso;

evento B : avere come prima carta un asso.

Allora la probabilità che si verifichi A è :

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

Nel tirare la seconda carta ho $36 \cdot 35$ possibili casi . Di questi $3 \cdot 4$ sono quelli favorevole all'evento $A \cap B$ e $4 \cdot 32$ sono quelli favorevoli all'evento $A \cap \bar{B}$. Quindi :

$$P(A) = \frac{4 \cdot 3}{36 \cdot 35} + \frac{32 \cdot 4}{36 \cdot 35} = \frac{1}{9}$$

Si noti che $P(A \cap B) = \frac{4 \cdot 3}{36 \cdot 35}$ e quindi

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)} = \frac{3}{35}$$

Il risultato può anche essere ottenuto cosi:

Se la prima carta è un asso , allora restano 3 assi e 35 possibilità per il secondo lancio e quindi :

$$P(A \mid B) = \frac{3}{35}$$

3. **Teorema di Bayes** : Poichè A $\cap$ B = B $\cap$ A e poichè

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

con P(A) $\neq$ 0, allora

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

cioe anche :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Questo importante risultato, che lega le due probabilità condizionali, è noto come teorema di Bayes. Questo teorema costituisce la base della " **Statistica Bayesiana** "

4. **Legge della Probabilità Totale** Sia S costituito da eventi disgiunti A_i ($S = \cup_i A_i$, $P(A_i \mid A_j) = 0$ for $i \neq j$) e sia $P(A_i) \neq 0$ per ogni i, allora un arbitrario evento B si scrive come $B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$. B e ogni A_i sono disgiunti, quindi:

$$P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i)P(A_i)$$

Questo risultato è noto come **legge della probabilità totale**

1.3.1 Una applicazione del teorema di Bayes

Consideriamo 3 contenitori B_1,B_2 e B_3, il primo contenente due monete d'oro, il secondo contenente una moneta d'oro ed una d'argento , il terzo contenente due monete d'argento. Scegliamo un contenitore a caso e prendiamo una moneta. Sia una moneta d'oro. Ci chiediamo qual è la probabilità che la seconda moneta dello stesso contenitore sia ancora d'oro.

Sia A l'evento (presa di una moneta d'oro dal contenitore), si tratta di calcolare la probabilità condizionale $P(B_1 \mid A)$, cioè che io scelga il contenitore B_1 con la condizione di trovare ancora una moneta d'oro. Le probabilità condizionali $P(A \mid B_i)$ di prendere una moneta d'oro dai contenitori B_i sono date da :

$$P(A \mid B_1) = 1, P(A \mid B_2) = 1/2, P(A \mid B_3) = 0$$

Avendo scelto a caso uno dei tre contenitori allora :

$$P(B_1) = P(B_2) = P(B_3) = 1/3$$

Applicando il teorema di Bayes abbiamo :

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{\sum_{j=1}^{3} P(A \mid B_j)P(B_j)} = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 1/3 + 0 \cdot 1/3} = 2/3$$

- Un'altra applicazione del teorema di Bayes: il contatore Cherenkov

  Si abbia un fascio di particelle costituita dal 90 % di pioni (π) e 10 % di kaoni (K). Il contatore dovrebbe dare segnale solo per i pioni. In pratica risponde ai pioni nel 95 % dei casi mentre da un segnale casuale per i kaoni in 6 % dei casi. Quindi se il contatore da un segnale la probabilità che sia un pione e :

$$P(\pi \mid segnale) = \frac{P(segnale \mid \pi) \cdot P(\pi)}{P(segnale \mid \pi)P(\pi) + P(segnale \mid K)P(K)}$$

$$= \frac{0.95 \cdot 0.90}{0.95 \cdot 0.90 + 0.06 \cdot 0.10} = 99.3\%$$

  In questo caso la probabilità che sia un K è 0.7 %

  Se il contatore non dà segnale, allora :

$$P(K \mid nessunsegnale) = \frac{P(nessunsegnale \mid K) \cdot P(K)}{P(nessunsegnale \mid \pi)P(\pi) + P(nessunsegnale \mid K)P(K)}$$

$$= \frac{0.94 \cdot 0.10}{0.05 \cdot 0.90 + 0.94 \cdot 0.10} = 67.6\%$$

Il teorema di Bayes , che è una semplice conseguenza delle regole del calcolo delle probabilità, ha interpretazioni che sono profondamente diverse tra loro e che vedremo tra un attimo.

1.3.2 Probabilità come Frequenza Relativa

Qualunque quantità che soddisfa i tre assiomi della teoria assiomatica della probabilità può essere interpretata come una probabilità. Infatti esistono due interpretazioni della probabilità che sono comunemente usate, che sono diverse tra di loro e che vanno tenute ben separate.

Una prima interpretazione della probabilità è data in termini di **frequenza relativa**. Faccio n volte un esperimento e supponiamo che un certo evento A si verifichi m volte. Allora quando n tende all'infinito il rapporto m/n tende ad un numero che definiamo P(A) di A. Questa interpretazione frequentista della probabilità è quella adottata più spesso(in particolare dalle scienze sperimentali). La statistica che fa uso di questa definizione di probabilità e detta **statistica frequentista (o classica)**. Qui si presuppone di poter effettuare più volte l'esperimento (cioè la misura di una certa grandezza fisica). Ad esempio se lancio un dado dico che la probabilità che venga un 3 sia 1/6. Voglio dire che se lancio il dado un numero elevato n di volte e conto il numero m di volte in cui appare il 3, il rapporto m/n =1/6. Il risultato è intuitivo in quanto tutte e sei le facce del dado sono equiprobabili (il dado è supposto " non manipolato").

1.3.3 Probabilità Soggettiva

È evidente che l'interpretazione frequentista della probabilità si basa sul presupposto che l'esperimento (lancio del dado) possa essere ripetuto. Ci sono situazioni in cui questo non è vero. Ad esempio lancio il dado e chiedo qual è la probabilità che nel lancio che sto per fare venga 3. Qui non sto parlando di un lancio qualsiasi ma di questo lancio particolare. Qui la frequenza è o 100 % o 0. Analogamente mi chiedo qual è la probabilità che domani 13 giugno a Milano piova. Posso aspettare e vedere se piove oppure no; intanto posso esprimere il mio " grado di fiducia " circa la possibilità che piova oppure no. Qui l'interpretazione della probabilità non può essere di tipo frequentista.

Elementi dello spazio campione sono ipotesi, cioe asserzioni che sono o false o vere. Probabilità P(A) che l'ipotesi A sia vera è il grado di fiducia che abbiamo che l'ipotesi A sia vera.

Consideriamo il teorema di Bayes dove l'evento A è l'ipotesi che una certa teoria sia vera mentre l'evento B è l'ipotesi che un esperimento misuri un particolare risultato (dati). Possiamo scrivere che :

$$P(teoria \mid dati) = \frac{P(dati \mid teoria) \cdot P(teoria)}{P(dati)}$$

P(teoria) è la **probabilità iniziale (prior)** che la teoria sia vera. P(dati | teoria) è la probabilità che si osservino effettivamente i dati misurati nell'ipotesi che la teoria sia vera. Questa probabilità è detta **verosimiglianza (likelihood)**. P(dati) è la probabilità che si ottengano i dati ottenuti sia che la teoria sia vera sia che la teoria sia falsa. P(teoria | dati) rappresenta la probabilità che la teoria sia vera , una volta viste le misure sperimentali. Questa probabilità è detta **probabilità finale (posterior)**. Parto dalla probabilità iniziale che una certa ipotesi sia vera. Faccio una misura che cambia la mia fiducia nella probabilità che l'ipotesi considerata sia vera.

La probabilità appena definita è detta soggettiva . La scelta della distribuzione iniziale è cosa abbastanza delicata. Per grandi statistiche la distribuzione finale è dominata dalla likelihood cosicchè la scelta della distribuzione iniziale è meno importante. La statistica che utilizza questa probabilità soggettiva è detta **Statistica Bayesiana**.

Noi abbiamo detto che la probabilità soggettiva viene utilizzata in eventi singoli (esperimenti non ripetibili). In pratica alcuni sostenitori della statistica bayesiana ritengono che non vi siano esperimenti ripetibili e che quindi questa interpretazione soggettiva della probabilità è l'unica valida.

La statistica bayesiana è la prima ad essersi sviluppata (Bernoulli, Laplace ecc). La statistica frequentista si è sviluppata nella prima parte del 1900 ad opera di Fisher, Neyman ed altri. Come detto in precedenza, la statistica frequentista è generalmente detta anche classica. Vi sono autori che chiamano classica la sta-

tistica bayesiana. E‘ innegabile che la statistica bayesiana in diverse circostanze è superiore alla statistica frequentista e vi sono studi volti a fondere le due statistiche in una sola. I risultati di questi sforzi però non sono ancora soffisfacenti. I contrasti tra le due correnti sono molto forti.

Noi riteniamo che le due statistiche rispondano ad esigenze diverse e siano da ritenersi perciò complementari : di volta in volta sceglieremo quella che risponde meglio alle nostre necessità. Le due statistiche comunque vanno tenute ben separate e deve essere chiaramente indicata la statistica usatanelle applicazioni. Come vedremo i risultati possono talvolta variare molto a seconda del tipo di statistica usata. Quando non specificato diversamente, la statistica utilizzata (da noi) sarà quella frequentista.

1.4 Variabili Casuali

Una variabile è detta casuale (o aleatoria) se assume un valore reale distinto per ogni elemento dello spazio campione . Una variabile casuale può essere discreta o continua. È detta discreta se può assumere un insieme numerabile di numeri reali con certe probabilità. Se l'insieme dei valori che la variabile casuale può assumere è continuo, la variabile casuale è detta continua. È possibile che la variabile casuale sia una combinazione di questi due tipi.

Quando parliamo di **variabile**, pensiamo a tutti i possibili valori che può assumere. Quando parliamo di **variabile casuale** noi pensiamo a tutti i possibili valori che la variabile assume ed inoltre alla distribuzione della probabilità in accordo alla quale la variabile assume tutti i valori.

- dati quantitativi

  Sono le misure di un esperimento.

- dati qualitativi

  Studio un campione di auto, analizzando per esempio il loro colore. In questi caso per una trattazione matematica associo ad ogni colore un numero. E quindi tratto statisticamente i numeri ottenuti.

Per vedere come sono distribuiti i numeri che ottengo uso un istogramma. Riporto sull asse x il valore della variabile per intervalli (bin) usualmente di pari larghezza. Per ogni bin riporto sull asse y il numero di volte che quel valore viene trovato nel campione considerato.

Se lancio una monetina mi aspetto che la probabilità che venga testa o croce debba essere la stessa (0.5). Invece che lanciare una monetina posso usare una sequenza di numeri casuali (per esempio da 0 a 100 con un programma sul computer) e poi dico croce se il numero ottenuto è compreso tra 0 e 49 mentre dico testa se il numero è compreso tra 50 e 99. Con questa tecnica posso generare un

campione casuale. Se il programma fosse sbagliato e non desse numeri casuali oppure se lancio veramente la monetina e questa per qualche motivo preferisce testa a croce , il campione non sarebbe casuale e l'inferenza statistica che farei sarebbe sbagliata. Considerazioni analoghe valgono in generale.

Chapter 2

Descrizione dei Dati

- Funzione Densità di Probabilità. Funzione Distribuzione Cumulativa. Funzione Caratteristica
- Funzioni di Variabili Casuali
- p.d.f della Somma di due Variabili Casuali Indipendenti
- p.d.f. del Prodotto di due Variabili Casuali Indipendenti
- Valori di Aspettazione
- Matrice di Covarianza
- Propagazione degli Errori

2.1 Funzione Densità di Probabilità. Funzione Distribuzione Cumulativa. Funzione Caratteristica

La nostra variabile casuale (misura di un qualche processo) assume valori diversi con diverse probabilità. Noi vogliamo vedere come caratterizzare la distribuzione delle probabilita'. Per semplicità noi consideriamo variabili casuali continue. La generalizzazione di quanto esposto al caso di variabili casuali disctrete è abbastanza semplice.

2.1.1 Funzione Densità di Probabilità

- **Misura caratterizzata da una sola variabile casuale**

  Consideriamo un esperimento per misurare una determinata grandezza. Supponiamo che questa misura sia una variabile continua compresa ad esempio tra - ∞ e + ∞. Indichiamo con $F(x_0)$ la probabilità che la variabile casuale (la misura) x sia minore di x_0 . Quindi $F(-\infty)=0$ e $F(+\infty)=1$. La probabilità che la variabile x sia conpresa tra x e x +dx è :

  $$P(x \in [x, x+dx]) = F(x+dx) - F(x)$$

  Definiamo **funzione densità di probabilità (p.d.f.)** la funzione :

  $$f(x) = \frac{dF}{dx}$$

  Quindi

  $$P(x \in [x, x+dx]) = F(x+dx) - F(x) = f(x)dx$$

  Ovviamente :

  $$\int_{-\infty}^{+\infty} f(x)dx = 1$$

  Se la variabile casuale assume valori discreti l'integrale si trasforma in una sommatoria.

- **Misura caratterizzata da due variabili casuali**

  Se la nostra misura è caratterizzata da due variabili casuali continue x e y, allora la probabilità che la nostra misura abbia x compreso tra x e x+dx e y compresa tra y e y+dy è data da :

$$P(x \in [x, x+dx], y \in [y, y+dy]) = f(x,y)dxdy$$

f(x,y) è detta **funzione densità di probabilità congiunta**. Integrando f(x,y) su tutti i possibili valori di x e y si ottiene 1 (condizione di normalizzazione).

Nota la p.d.f. congiunta f(x,y) possiamo essere interessati alla p.d.f. della variabile x indipendentemente dal valore assunto della variabile y. Questa p.d.f. si ottierne integrando (sommando) sulla variabile y :

$$f_x(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

$f_x(x)$ è detta **p.d.f. marginale** per x.

In modo analogo si definisce la p.d.f. marginale per y $f_y(y)$.

Esempio dello "scatter plot" con proiezioni sugli assi.

Calcolare la probabilità che y sia compreso tra y e y+dy per qualunque valore di x (evento B), con la condizione che x sia contenuto in x , x+dx con qualunque y (evento A) :

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x,y)dxdy}{f_x(x)dx}$$

la p.d.f. condizionale $h(y \mid x)$ che si verifichi y dato x è cosi definita :

$$h(y \mid x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y')dy'}$$

In modo analogo si definisce la p.d.f condizionale $g(x \mid y)$ che si verifichi y dato x. Si verifica facilmente che :

$$g(x \mid y) = \frac{h(y \mid x) f_x(x)}{f_y(y)}$$

che è il teorema di Bayes.

Se il valore di una variabile non dipende dal valore assunto dall'altra variabile, allora le due variabili si dicono **(statisticamente) indipendenti**. In questo caso f(x,y) = f(x) f(y).

In questo caso la p.d.f. marginale di x è semplicemente f(x).

Esempio: Supponiamo che $f(x,y) = \frac{3}{2}(x^2 + y^2)$ per $0 < x < 1$,$0 < y < 1$ e 0 altrove. Calcolare la probabilità che la variabile x assuma un valore compreso

nell'intervallo (0,0.5), supponendo che y sia compreso nell'intervallo (0,0.5). Calcolare inoltre la probabilità che x sia nell'intervallo (0,0.5) quando la variabile y assume il valore 0.5. Le due variabile x e y sono indipendenti ?

Sia A l'evento x compreso tra 0 e 0.5, B l'evento y compreso tra 0 e 0.5. Allora :

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

Il numeratore si ottiene per doppia integrazione della funzione $f(x, y)$ negli intervalli dati. Questa integrazione vale 1/16. Per calcolare il denominatore prima dobbiamo calcolate la pdf marginale per y:

$$f_y(y) = \frac{3}{2} \int_0^1 (x^2 + y^2)\, \mathrm{d}x = \frac{1}{2} + \frac{3}{2} y^2$$

La probabilità che si verifichi l'evento B è :

$$\int_0^{0.5} \left(\frac{1}{2} + \frac{3}{2} y^2 \right) \mathrm{d}y = \frac{5}{16}$$

Quindi la probabilità che si realizzi A supponendo realizzato B è uguale ad 1/5.

La p.d.f. condizionale che si verifichi x dato y è:

$$f(x|y) = \frac{f(x, y)}{f_y(y)} = \frac{\frac{3}{2}(x^2 + y^2)}{\frac{1}{2} + \frac{3}{2} y^2}$$

Ponendo $y = 0.5$, si ha :

$$f(x|y = 0.5) = \frac{3}{7} + \frac{12}{7} x^2$$

Quindi

$$P(A|B) = \int_0^{0.5} \left(\frac{3}{7} + \frac{12}{7} x^2 \right) \mathrm{d}x = \frac{2}{7}$$

Le due variabili x e y non possono essere indipendenti in quanto la p.d.f. congiunta non può esprimersi come prodotto di una funzione della variabile x e di una funzione della variabile y. Infatti si vede che :

$$\frac{3}{2}(x^2+y^2) \neq \left(\frac{1}{2} + \frac{3}{2}x^2\right)\left(\frac{1}{2} + \frac{3}{2}y^2\right)$$

Le cose dette si generalizzano facilmente al caso di misura caratterizzata da n variabili casuali.

2.1.2 Funzione Distribuzione Cumulativa

Un altro modo di caratterizzare la distribuzione delle probabilità di una variabile casuale x è dato dalla sua **funxione di distribuzione cumulativa** (c.d.f.) F(x) cosi definita :

$$F(x) = \int_{-\infty}^{x} f(u)du$$

dove f(x) è la p.d.f. F(x) rappresenta la probabilità che la variabile casuale assuma un valore minore od uguale ad x. la c.d.f. F(x) assume valori crescenti da o ad 1. Se la funzione F(x) è strettamente crescente, allora vi è una corrispondenza uno ad uno tra il valore della variabile casuale x e il valore assunto dalla c.d.f F(x). In questo caso la funzione F(x) può essere invertita. Si definisce punto α (o anche " quantile di ordine α ") x_α il valore della variabile x per il quale si ha :

$$F(x_\alpha) = \alpha$$

con $0 \leq \alpha \leq 1$. Il quantile non è altro che l'inverso della distribuzione cumulativa.

$$x_\alpha = F^{-1}(\alpha)$$

Il quantile più usato è la **mediana** che è il quantile corrispondente ad α=0.5 :

$$0.5 = \int_{-\infty}^{mediana} f(x)dx$$

2.1.3 Funzione Caratteristica

Un terzo modo di di rappresentare la distribuzione di una variabile casuale è dato dalla funzione caratteristica (ch.f.) ϕ. Questa è la trasformata di Fourier della p.d.f.. Supponiamo di avere una sola variabile casuale continua , allora

$$\phi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx$$

Qui vale una proprietà importantissima e cioè che vale una corrispondenza biunivoca tra p.d.f. e ch.f., cosicchè la conoscenza che si ha con una è del tutto equivalente alla conoscenza che si ha con l'altra. Naturalmente :

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ixt} \phi(t) dt$$

Molto spesso è più comodo utilizzare la ch.f.

Ovvia generalizzazione al caso di misure che comportano più variabili casuali.

2.2 Funzioni di Variabili Casuali

Una funzione di una variabile casuale è essa stessa una variabile casuale. Supponiamo di avere una variabile casuale X che assume valori continui x con p.d.f. f(x). Supponiamo che Y =ϕ(X) sia una funzione continua di x. Allora qual è la p.d.f. g(y) della variabile casuale Y ?

Noi abbiamo che :

$$P(Y < y) = P(\phi(X) < y)$$

Se la funzione ϕ è una funzione crescente, allora possiamo scrivere che :

$$P(Y < y) = P[X < \phi^{-1}(y)]$$

Notiamo che $\mathbf{x = \phi^{-1}(y)}$, essendo ϕ^{-1} la funzione inversa di ϕ. Indicando con G(y) ed F(x) le c.d.f. delle variabili casuali Y ed X, allora la relazione precedente possiamo riscriverla cosi:

$$G(y) = F[\phi^{-1}(y)]$$

Differenziando entrambi i membri rispetto ad y, otteniamo:

$$g(y) = f[\phi^{-1}(y)] \cdot \frac{d\phi^{-1}}{dy}$$

Se la funzione ϕ è una funzione decrescente, allora avremo che :

$$P(Y < y) = P[X > \phi^{-1}(y)]$$

che possiamo riscrivere cosi:

$$G(y) = 1 - F[\phi^{-1}(y)]$$

Differenziando entrambi i membri rispetto ad y, abbiamo:

$$g(y) = -f[\phi^{-1}(y)] \cdot \frac{d\phi^{-1}}{dy}$$

Quindi possiamo scrivere che :

$$g(y) = f[\phi^{-1}(y)] \cdot \left|\frac{d\phi^{-1}}{dy}\right|$$

Si noti che $\frac{d\phi^{-1}}{dy} = (\frac{d\phi}{dx})^{-1}$ e quindi possiamo anche scrivere che :

$$g(y) = f(x) \left|\frac{dx}{dy}\right|$$

Il termine in valore assoluto in questa espressione è detto Jacobiano della trasformazione.

- Sia f(x) =1 per x compreso tra 0 e 1 e 0 altrimenti per una certa variabile casuale X. Assumendo Y=X^2, determinare la p.d.f. g(y) della variabile casuale Y

  dy/dx =2x ; g(y) = $\frac{1}{2x}$ con ($0 < x < 1$).

  Poichè x =$\sqrt{y}$, allora g(y) = $\frac{1}{2\sqrt{y}}$ per y compreso tra 0 e 1.

  Si verifica facilmente che :

$$\int_0^1 \frac{1}{2\sqrt{y}} dy = 1$$

Questi risultati si generalizzano a funzioni di più variabili casuali.

2.3 p.d.f. della Somma di due Variabili Casuali Indipendenti

Sia Z = X + Y e siano g(x) e h(y) le p.d.f. delle variabili casuali ed indipendenti X e Y. La p.d.f. congiunta è f(x,y) = g(x)h(y). Sia P(z) la probabilità che Z= X + Y $\leq$ z. Sul piano cartesiano XY la regione corrispondente a X+Y $\leq$ z è la parte di piano posto a sinistra e sotto la retta di equazione X+Y=Z. Per ogni dato y la probabilità che X $\leq$ z-y è G(z-y) con :

$$G(z-y) = \int_{-\infty}^{z-y} g(u)du$$

Quindi P(z) si ottiene moltiplicando G(z-y) per la probabilità per y ed integrando su tutti gli y:

$$P(z) = \int_{-\infty}^{+\infty} G(z-y)h(y)dy$$

Differenziando rispetto a z otteniamo la p.d.f. f(z):

$$f(z) = \int_{-\infty}^{+\infty} g(z-y)h(y)dy$$

Questa è spesso scritta f = g $\otimes$ h ed è detta convoluzione (di Fourier) di g e h. f(z) può anche essere scritta come :

$$f(z) = \int_{-\infty}^{+\infty} g(x)h(z-x)dx$$

La ch.f. di Z è il prodotto delle ch.f. di X e Y :

$$\phi_z(t) = \phi_x(t)\phi_y(t)$$

f(z) si ottiene invertendo $\phi_z(t)$.

Esempio:

Supponiamo che X abbia una distribuzione uniforme nell'intervallo [0,1] e Y una distribuzione triangolare simmetrica nell'intervallo [0,2]. La variabile casuale Z = X + Y ha valori solo nell'intervallo [0,3]. Vogliamo determinare la p.d.f. della variabile Z.

La p.d.f. della variabile X è g(x) = 1 per $0 < x < 1$ mentre h(y) = y per $0 \leq y \leq 1$; h(y) = 2-y per $1 \leq y \leq 2$.

La funzione g(z-y) = 1 per i valori di y tra z-1 e z mentre è 0 per tutti gli altri valori. Di conseguenza:

$$f(z) = \int_{z-1}^{z} h(y)dy$$

Dalla definizione di h(y), f(z) ha una diversa espressione nei tre intervalli di z, cioè [0,1], [1,2] e [2,3]:

$$f(z) = \int_{0}^{z} ydy = \frac{z^2}{2}, \text{ per } 0 < \text{z} < 1$$

$$f(z) = \int_{z-1}^{1} ydy + \int_{1}^{z} (2-y)dy = 3(z-1/2) - z^2, \ 1< \text{z} < 2$$

$$f(z) = \int_{z-1}^{2} (2-y)dy = \frac{1}{2}(3-z)^2, 2 < \text{z} < 3$$

La p.d.f. f(z) è costituita da parti di 3 parabole. Essa è simmetrica rispetto a z=1/2.

2.4 p.d.f. del Prodotto di due Variabili Casuali Indipendenti

Con un procedimento simile al quello seguito nel paragrafo precedente si può far vedere che, se Z=XY con g(x) e h(y) le p.d.f. delle variabili casuali X e Y, allora la p.d.f. f(z) di Z è data da :

$$f(z) = \int_{-\infty}^{+\infty} g(x)h(\frac{z}{x})\frac{dx}{\mid x \mid}$$

o anche :

$$f(z) = \int_{-\infty}^{+\infty} g(\frac{z}{y})h(y)\frac{dy}{\mid y \mid}$$

Questa equazione è spesso scritta f = g $\otimes$ h . f è detta convoluzione (di Mellin) di g e h.

2.5 Valori di Aspettazione

- **Media (aritmetica)** di una variabile casuale.

  Se una variabile casuale x è distribuita secondo una p.d.f. f(x) , il suo valore di aspettazione E[x] è dato da :

  $$E[x] = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

  μ è detto anche media della popolazione (o semplicemente media). Questa è la media aritmetica (che è quella più usata).

- Media (aritmetica) di una funzione di variabile casuale

  Se a(x) è una funzione della variabile casuale x distribuita secondo una p.d.f. f(x), allora il valore di aspettazione della variabile a è :

  $$E[a] = \int_{-\infty}^{+\infty} a(x)f(x)dx$$

- **Momento n-esimo (momento di ordine n)**

  Si definisce momento n-esimo di x :

  $$E[x^n] = \int_{-\infty}^{+\infty} x^n f(x)dx = \mu_n^{'}$$

Ovviamente $\mu = \mu_1^{'}$

- **Momento centrale n-esimo**

  Si definisce momento centrale n-esimo la quantità :

$$E[(x - E[x])^n] = \int_{-\infty}^{+\infty} (x - \mu)^n f(x) dx$$

  Il momento centrale secondo è detto **varianza** di x ed è indicato con σ^2 oppure con V[x] :

$$E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \sigma^2 = V[x]$$

  Si noti che V[x] = $E[x^2]$ - μ^2. Infatti considerando per semplicità il caso di N misure si ha che :

$$V[x] = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i\mu + \frac{1}{N}\sum_{i=1}^{N}\mu^2$$

$$V[x] = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - 2\mu\frac{1}{N}\sum_{i=1}^{N} x_i + \mu^2\frac{1}{N}\sum_{i=1}^{N} 1$$

$$V[x] = \bar{x^2} - 2\mu^2 + \mu^2 = \bar{x^2} - \mu^2$$

  La varianza è data dalla media dei quadrati meno il quadrato della media.

  La varianza misura quanto larga è la distribuzione attorno al valore medio

- **Deviazione Standard**

  La radice quadrata della varianza è detta deviazione standard σ.

Esercizio 1 : Supponiamo di avere particelle distribuite in modo casuale in un quadrato di lato a con i lati paralleli agli assi. Calcolare $\overline{x}$ e la deviazione standard σ_x.

La probabilità lungo l'asse x è costante (k). Di conseguenza la condizione di normalizzazione per la pdf f(x) è : $\int_0^a k dx = 1$ e da questa si ha che $k = 1/a$. Allora il valore medio di x è:

$$\overline{x} = \int_0^a \frac{1}{a} x dx = \frac{a}{2}$$

$$\overline{x^2} = \int_0^a \frac{1}{a} x^2 dx = \frac{a^2}{3}$$

Di conseguenza $\sigma_x = \overline{x^2}$ - $\overline{x}^2$ e quindi

$$\sigma_x^2 = \frac{a^2}{3} - \left(\frac{a}{2}\right)^2 = \frac{a^2}{12}$$

La deviazione standard è $\sigma_x = a/\sqrt{12}$

Esercizio 2: Supponiamo di ruotare il duadrato dell'esercizio precedente portando le diagonali parllele agli assi coordinati. Calcolare $\overline{x}$ e la varianza σ_x. (Soluzione: $\overline{x} = 0$ e $\sigma_x = a/\sqrt{12}$).

2.6 Matrice di Covarianza

Le considerazioni viste prima per il caso di eventi con una sola variabile casuale vengono generalizzate al caso di eventi con più variabili casuali. Consideriamo per semplicità il caso di due variabili casuali x e y. Sia f(x,y) la p.d.f. congiunta delle due variabili. Oltre alle quantità viste in precedenza è possibile utilizzare l'eventuale correlazione di una variabile con l'altra. Questa correlazione è detta **covarianza** V_{xy} (indicata anche con cov(x,y)) ed è definita da :

$$V_{xy} = E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y$$

$$V_{xy} = \int_{-\infty}^{+\infty} xy f(x,y) dx dy - \mu_x \mu_y$$

con μ_x = E[x] e μ_y=E[y]

La matrice di covarianza (detta anche matrice degli errori) è una matrice simmetrica ($V_{xy} = V_{yx}$) con elementi diagonali positivi ($V_{xx} = \sigma_x^2$, $V_{yy} = \sigma_y^2$).

Coefficiente di correlazione

Il coefficiente di correlazione ρ è definito da :

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

ρ è compreso tra -1 e 1 ed è una grandezza senza dimension mentre gli elementi della matrice di covarianza hanno dimensioni. Il coefficiente di correlazione misura il grado di correlazione lineare tra le due variabili x e y. Se due variabili sono indipendenti hanno sempre grado di correlazione zero ma non vale il viceversa. La correlazione misura solo la dipendenza lineare. È possibile che X e X^2 abbiano correlazione nulla per quanto ovviamente non sono indipendenti.

Se consideriamo la somma di due variabili casuali , si ha:

$$V[x+y] = V[x] + V[y] + 2V_{xy}$$

Se le due variabili sono indipendenti, allora $V_{xy} = 0$ e la varianza della somma è la somma delle varianze.

Un esempio di correlazione positiva è per esempio quella tra altezza e peso. In genere le persone più alte pesano di più. Una correlazione negativa si ha ad esempio tra il peso di una persona e la sua agilità. Più una persona pesa e meno in generale è agile. Non c'è correlazione tra altezza e quoziente di intelligenza. Le persone più alte in genere non sono più intelligenti di quelle più basse.

Quanto detto si generalizza al caso di evento con n variabili casuali. Se l'evento dipende da n variabili casuali x_1, x_2, ... x_n , allora l'elemento V_{ij} della matrice degli errori è dato da :

$$V_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

dove E$[x_i] = \mu_i$ e E$[x_j] = \mu_j$. La matrice degli errori è una matrice simmetrica nxn. Il coefficiente di correlazione tra le due variabili x_i e x_j è :

$$\rho_{ij} = \frac{V_{ij}}{\sigma_i \sigma_j}$$

2.7 Propagazione degli Errori

Consideriamo una funzione f(x) lineare della variabile casuale x:

$$f(x) = ax + b$$

con a e b costanti. Le misure fatte della variabile casuale x permettono di conoscere il suo valore medio x_0 e la sua varianza V[x]. Dalla definizione di varianza si ha :

$$V[f] = E[(f - E[f])^2] = a^2 V[x]$$

e analogamente

$$\sigma_f = \mid a \mid \sigma_x$$

Consideriamo una generica funzione f(x) della variabile casuale x. Sviluppiamo la funxione f(x) in serie di Taylor attorno ad x_0:

$$f(x) = f(x_0) + (x - x_0) \left(\frac{df}{dx} \right) \Bigg|_{x=x_0}$$

Questa formula è valida per errori "piccoli" , cioé il differenziale deve variare poco all ' interno di alcune σ.

Se la funzione f dipende da due variabili casuali x_1 e x_2 ,

$$y = f(x_1, x_2)$$

Conoscendo i valori medi e le varianze di x_1 e x_2, possiamo sviluppare la funzione f attorno a questi valori ed ottenere :

$$V[y] = \left(\frac{\partial y}{\partial x_1}\right)^2 V[x_1] + \left(\frac{\partial y}{\partial x_2}\right)^2 V[x_2] + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) V_{x_1 x_2}$$

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \sigma_{x_2}^2 + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) \rho\sigma_{x_1}\sigma_{x_2}$$

dove i differenziali vanno calcolati nei punti medi delle variabili x_1 e x_2 (E[x_1] =μ_1 e E[x_2] =μ_2). La formula anche qui è valida per errori " piccoli".

Le relazione si generalizza al caso di funzione che dipende da n variabili casuali e si può scrivere cosi:

$$\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right] V_{x_i x_j}$$

Se si hanno m funzioni y_1, ... y_m delle n variabili casuali x_1 ,...., x_n allora la matrice degli errori diventa :

$$U_{kl} = V_{y_k y_l} = \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right] V_{x_i x_j}$$

Con notazione matriciale si ha :

$$U = AVA^T$$

essendo la matrice delle derivate A data da :

$$A_{ij} = \left[\frac{\partial y_i}{\partial x_j}\right]$$

A^T è la matrice trasposta di A.

Le relazioni viste permettono di propagare l'errore dalle variabili x_1 ,...., x_n alle variabili y_1 ,..., y_m (**propagazione degli errori**)

Nel caso di variabili x_i non correlate i termini covarianti nelle relazioni viste diventano nulli.

$$\sigma_y^2 = \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]^2 \sigma_i^2$$

e

$$U_{kl} = \sum_{i=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_i} \right] \sigma_i^2$$

Nel caso particolare di y = x_1 + x_2 si ha :

$$\sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2$$

cioè gli errori si sommano quadraticamente.

Se y = x_1 x_2 in generale si ha :

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2} + 2\frac{V_{x_1 x_2}}{x_1 x_2}$$

Se le variabili x_1 e x_2 non sono correlate, allora si ha :

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2}$$

Il rapporto tra σ e il valore della variabile è detto errore relativo di questa variabile. La relazione vista dice che per il prodotto si sommano quadraticamente gli errori relativi.

Questo risultato vale anche per il rapporto : y =$\frac{x_1}{x_2}$

Generalmente date n variabili casuali x_1, x_2,..., x_n, la matrice di covarianza non è diagonale. Spesso è utile riscrivere la matrice degli errori in termini di nuove variabili casuali y_1, y_2,..., y_n che non siano correlate, per le quali cioè la matrice di cavarianza è diagonale. Si può dimostrare che questo è sempre possibile. Le nuove variabili casuali si ottengono dalle iniziali variabili casuali mediante una trasformazione lineare:

$$y_i = \sum_{j=1}^{n} A_{ij} x_j$$

La nuova matrice di covarianza U è data da :

$$U_{ij} = \sum_{k,l=1}^{n} A_{ik} A_{jl} V_{x_k, x_l} = \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^T$$

Il problema quindi si riduce a trovare una matrice A tale che :

$$U = AVA^T$$

Questo equivale alla diagonalizzazione di una matrice simmetrica e reale.

Chapter 3

Distribuzioni di Probabilità Notevoli

- Distribuzione Binomiale
- Distribuzione di Poisson
- Distribuzione Gaussiana
- Distribuzione Uniforme
- Distribuzione Esponenziale
- Distribuzione χ^2
- Distribuzione t di Student
- Distribuzione di Cauchy
- Legge dei Grandi Numeri
- Teorema Limite Centrale

In questo capitolo introduciamo alcune famiglie parametriche di distribuzioni particolarmente importanti nelle applicazioni della statistica. Alcune di queste distribuzioni sono discrete, altre sono continue. Per le tabelle statistiche siveda ad esempio: http://calculators.stat.ucla.edu/cdf/ oppure cercare in rete surfstat.australia.

3.1 Distribuzione Binomiale

Si abbia un esperimento che possa avere come esito sono due possibilità A e B: per esempio lancio una moneta e posso avere come risultato o testa (A) o croce (B). La variabile è discreta. Sia p (costante) la probabilità che si verifichi A e q (costante) la probabilità che si verifichi B (p + q =1) . Ripetiamo N volte l'esperimento e in ogni esperimento la variabile casuale assume o il valore 1 (evento A) o il valore 0 (evento B). Qual è la probabilità di avere n volte l'evento A ?

- **La probabilità** che i primi n tentativi diano come risultato A ed i rimanenti N-n diano come risultato B è data da $p^n q^{N-n}$. In N misure quanti casi ho di avere n eventi A senza tener conto dell'ordine con il quale avvengono ? Il numero di questi casi è :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

  Quindi la probabilità di avere n volte A e N-n volte B è :

$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n q^{N-n}$$

  con n = 0,1, ... , N (n è la variabile casuale mentre N e p sono parametri).

- **Valore di aspettazione di n**

$$E[n] = \sum_{n=0}^{\infty} n \frac{N!}{n!(N-n)!} p^n q^{N-n} = Np$$

- **Varianza di n**

$$V[n] = E[n^2] - (E[n])^2 = Npq$$

In Fig. 3.1 sono mostrate le distribuzioni di probabilità binomiale per $N = 9$ = esperimenti e con diversa probabilità p . Fig. 3.2 sono invece mostrate le distribuzioni di probabilità binomiale per un diverso munero di esperimenti N e con fissata probabilità $p = 0.5$.

Figure 3.1: Distributioni di probabilità binomiali con numero di esperimenti costante e con diversi valori della probabilità p.

Figure 3.2: Distribuzioni di probabilità binomiali con numero di esperimenti variabile e costante probabilità p.

Esempio 1 : Sia n il numero di volte che ho testa in cinque lanci di una moneta. Dare la distribuzione di probabilità di n e calcolare il valore medio e la varianza.

In questo esempio la distribuzione di probabilià è binomiale con N=5 e p=0.5. Quindi si ha :

$$f(n;5,0.5) = \frac{5!}{n!(5-n)!}0.5^n 0.5^{5-n} = \frac{5!}{n!(5-n)!} \cdot \frac{1}{32}$$

Per ciascun n si ha:

$$f(0;5,0.5) = f(5;5,0.5) = 0.0312$$

$$f(1;5,0.5) = f(4;5,0.5) = 0.1562$$

$$f(2;5,0.5) = f(3;5,0.5) = 0.3125$$

Il valore medio è Np = 2.5 mentre la varianza è Npq = 1.25

Esempio 2 : Un certo strumento ha un tempo (misurato in ore) di durata la cui p.d.f. è data da :

$$f(t) = \frac{1}{2}e^{-\frac{1}{2x}} \text{per t} \geq 0$$

Qual è la probabilità che su cento strumenti (simili) 8 durino più di due ore ?

La probabilità che uno strumento duri più di 2 ore è data da :

$$p = \int_2^\infty \frac{1}{2}e^{-\frac{1}{2x}}dx = \frac{1}{e}$$

La probabilità P che 8 strumenti durino più di 2 ore è perciò uguale a :

$$P = \binom{100}{8} (e^{-1})^8 \cdot (1-e^{-1})^{100-8}$$

La **distribuzione multinomiale** è una generalizzazione della distribuzione binomiale al caso in cui il risultato può essere più di due tipi. Supponiamo che si possano avere m risultati ed indichiamo con p_i la probabilità che si verifiche il risultato i-imo. Poichè almeno uno dei risultati si deve realizzare allora:

$$\sum_{i=1}^{m} p_i = 1$$

La probabilità congiunta di avere n_1 risultati di tipo 1, n_2 risultati di tipo 2 ,$\cdots$ n_m di tipo m è data:

$$f(n_1, \cdots, n_m; N, p_1, \cdots, p_m) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

Questa distribuzione è detta multinomiale.

Il valore di aspettazione per il risultato i-mo è :

$$E[n_i] = Np_i$$

mentre la varianza per il numero di eventi con risultato i-imo è :

$$V[n_i] = Np_iq_i = Np_i(1 - p_i)$$

Esempio:

Compero 20 bulbi di giacinto, 10 aspettati di colore rosso e 10 di colore blu. Quando crescono, scopro che 16 sono blu. Faccio questo ragionamento:la σ sui fiori blu è 4 per cui ho una discrepanza di 1.56 σ trascurabile. Questo ragionamento è vero o falso ?

Il ragionamento è falso. La distribuzione è binomiale. La probabilità di avere 16 o più fiori blue è uguale alla probabilità di avere 4 o meno fiori rossi , cioè 0.0059 :

$$Pr(fiori \geq 16) = \sum_{r=16}^{20} \binom{20}{r} 0.5^r 0.5^{20-r} = 0.0059$$

3.2 Distribuzione di Poisson

Se in una distribuzione binomiale , N (dimensione del campione) è molto grande, p , la probabilità di ottenere un determinato valore della variabile, è molto piccolo (evento " raro") in modo che il valore di aspettazione del numero di successi sia un valore finito ν, allora la distribuzione binomiale diventa :

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

Questa distribuzione è detta di Poisson. Ha come parametro ν.

Per arrivare a questa legge di distribuzione possiamo seguire questo semplice ragionamento. Consideriamo un intervallo di tempo [0,T] e suddividiamo questo intervallo di tempo in N sottointervalli di lunghezza T/N. Sia p = $\lambda \cdot \frac{T}{N}$ la probabilità che l'eventi si verifichi in uno qualunque dei sottointervalli considerati (λ è un numero reale positivo) . La probabilità che in un sottointervallo l'evento non si verifichi è data da q = 1-p. Indicando con n il numero di volte in cui l'evento si realizza ed applicando la legge di distribuzione binomiale si ha:

$$f(n;N,p) = \frac{N!}{n!(N-n)!}p^n q^{N-n} = \frac{N!}{n!(N-n)!}\left(\lambda T/N\right)^n \left(1-\lambda T/N\right)^{N-n}$$

Questo può anche essere riscritto cosi:

$$f(n;N,p) = \frac{N\cdot(N-1)\cdots(N-n+1)}{n!}\cdot\left(\frac{\lambda T}{N}\right)^n\left(1-\frac{\lambda T}{N}\right)^{N-n}$$

$$f(n;N,p) = \frac{(\lambda T)^n}{n!}\cdot\frac{(N\cdot(N-1)\cdots(N-n+1))}{N^n}\cdot\left(1-\frac{\lambda T}{N}\right)^{N-n}$$

Facciamo aumentare il numero di sottointervalli N (cioè facciamo tendere N all'infinito) .

$$\frac{(N\cdot(N-1)\cdots(N-n+1))}{N^n} \to 1$$

$$\left(1-\frac{\lambda T}{N}\right)^{N-n} \to e^{-\lambda T}$$

Ponendo $\nu = \lambda T$ (= Np =costante), abbiamo infine:

$$f(n;\nu) = \frac{\nu^n}{n!}\cdot e^{-\nu}$$

- Valore di aspettazione.

  Il valore di aspettazione di una variabile poissoniana è :

$$E[n] = \sum_{n=0}^{\infty} n\frac{\nu^n}{n!}\exp\left[-\nu\right] = \nu$$

- Varianza

  La varianza è data da :

$$V[n] = \sum_{n=0}^{\infty}(n-\nu)^2\cdot\frac{\nu^n}{n!}\cdot\exp\left[-\nu\right] = \nu$$

In Fig. 3.3 sono mostrate le distribuzioni di una variabile poissoniana per diversi valori si aspettazione.

Se il valore di N diventa grande (N che tende all'infinito) e p diventa molto piccolo (p che tende a zero) in modo che il prodotto $N\cdot p$ resti costante, allora

la distribuzione binomiale diventa quella di Poisson. In Fig. 3.4 è mostrata la distribuzione binomiale con $N = 150$ e $p = 0.013$ che è da confrontare con la distribuzione poissoniana con valore di aspettazione 2 in Fig. 3.3.

La distribuzione poissoniana è tra le più adottate per descrivere fenomeni naturali.

Esempio : Il numero di particelle per pulse in un fascio ha una distribuzione di Poisson con valore medio 16 . Qual è la probabilità che un "pulse " abbia un numero di particelle compreso tra 12 e 20?

La distribuzione di Poisson in questo caso è :

$$f(n;16) = \frac{16^n}{n!}e^{-16} \qquad n = 0, 1, 2, ...$$

La probabilità richiesta è data da :

$$e^{-16}\sum_{n=12}^{20}\frac{16^n}{n!} =$$

3.3 Distribuzione Gaussiana

Questa è una distribuzione estremamente importante in statistica per ragioni che vedremo presto. Diciamo che una variabile casuale segue una distribuzione gaussiana (detta anche normale) se la p.d.f è :

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Questa p.d.f. ha due parametri, μ e σ.

- Valore di aspettazione di x

$$E[x] = \int_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx = \mu$$

- Varianza

$$V[x] = \int_{-\infty}^{+\infty} (x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx = \sigma^2$$

La distribuzione gaussiana è una curva a campana simmetrica attorno all'asse x = μ con due punti di flesso in x=μ -σ e in x = $\mu + \sigma$. Essa è spesso scritta come N(μ, σ^2).

La distribuzione gaussiana con parametri μ=0 e σ=1 è data da :

$$\phi(x) = f(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

La corrispondente c.d.f. $\Phi(x)$ è data da :

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$$

Questa p.d.f. è detta gaussiana standard (o anche funzione degli errori). Questa funzione non può essere calcolata esplicitamente ma è tabulata. Questa curva è particolarmente importante perchè se una variabile casuale y è distribuita gaussianamente con valore medio μ e deviazione standard σ , allora la variabile x = (y-μ)/σ segue la distribuzione gaussiana standard. Le corrispondenti c.d.f. sono legate dalla relazione F(y) = $\Phi(x)$. I valori di $\Phi(x)$ ed i quantili $x_\alpha = \Phi^{-1}(\alpha)$ sono tabulati. Ora è però più comodo calcolarseli al computer.

In Fig. 3.5.

Esempio 1 :

Una variabile casuale X ha una p.d.f. gaussiana con valore medio 5 e varianza 4. Calcolare la probabilità p che la variabile assuma un valore minore di 2.

La variabile $\frac{X-5}{2}$ ha una p.d.f. gaussiana standard, quindi :

$$p(X \leq 2) = \int_{-\infty}^{\frac{2-5}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \simeq 0.067$$

Si può verificare facilmente che il 68.27 % dell'area sotto la curva gaussiana si trova entro $\pm\sigma$ dal valore medio, il 95.45 % entro 2 σ ed il 99 % entro 3 σ. Spesso si calcola il 90 % dell'area e questa cade entro 1.645 σ, il 95 % entro 1.960 σ ed il 99 % entro 2.576 σ. Si tratta qui di intervalli centrali.

Tagli su una coda oppure su entrambe le code. ....

Esempio 2) :

Una variabile ha una distribuzione gaussiana con media uguale 10 e varianza uguale a 100. Calcolare la probabilità che $8 \leq x \leq 16$.

$$p(8 \leq x \leq 10) = \int_{\frac{8-10}{10}}^{\frac{16-10}{10}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \simeq 0.305$$

- **Distribuzione gaussiana come limite di quella poissoniana**

  Se il valore di aspettazione ν della distribuzione poissoniana è grande (ad esempio > 10) allora la distribuzione di Poisson è approssimata da una distribuzione gaussiana di valore medio $\mu = \nu$ e varianza $\sigma^2 = \nu$.

  In Fig. 3.6 si mostrano a confronto una distribuzione poissoniana con valore di aspettazione 25 ed una curva gaussiana con valore medio 25 e varianza 25.

Esempio :

In una zona del Canada ci sono in media 2 alci per lago. 1) Quale potrebbe essere la distribuzione del numero di alci per lago? 2) Se trovo 5 alci in un lago , qual è la probabilità che ciò sia accaduto per caso? 3) Se si approssima la distribuzione con una gaussiana, qual è la probabilità di trovare in un lago 5 o più alci? 4) Cosa direste se dichiarassi che ciò è avvenuto dopo aver visitato altri 19 laghi?

1) Poissoniana con media 2

2) f(alci = 5) = $\frac{e^{-2}2^5}{5!} = 0.0361$. La probabilità di trovare 5 o più alci in un lago è :

$$f(alci \geq 5) = 1 - Pr(alci \leq 4) = 0.0526$$

3) La distribuzione poissoniana potrebbe essere approssimata da una gaussiana N(2,2). In questo caso la probabilità di osservare 5 o più alci in un lago è :

$$p(alci \geq 5)) = \int_{\frac{5-2}{\sqrt{2}}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \simeq 0.017$$

Come si vede l'approssimazione gaussiana non è buona dato il valore troppo basso della valore medio.

4) Dopo 20 laghi la probabilità di non trovare 5 o più alci è data dalla distribuzionale binomiale come :

$$f = 1 - (1 - 0.0526)^{20} = 0.66$$

[(1-0.0526) rappresenta la probabilità di non trovare 5 o più alci in un lago. Quindi elevo a 20 se questo avviene dopo 20 laghi

Quindi non mi meraviglierebbe affatto l'averlo trovato.

- **Distribuzione gaussiana come limite di quella binomiale**

  Se nella distribuzione binomiale si prendono grandi valori di N (facciamo tendere N all'infinito) mantenendo p e q costanti, allora la distribuzione binomiale tende ad una gaussiana di valore medio $N \cdot p$ e varianza $N \cdot p \cdot q$. Si veda ad esempio la figure 3.2.

Distribuzione gaussiana multidimensionale Vi sono situazioni nelle quali siamo interessati a n variabili $x_1, x_2, \cdots, x_n$ che in forma compatta scriviamo $\vec{x}$, ognuna delle quali individualmente segue una distribuzione gaussiana. Sia $\vec{\mu} =$

$\mu_1, \mu_2, \cdots, \mu_n$ il vettore dei valori medi. I due vettori $\vec{x}$ e $\vec{\mu}$ son presi come vettori colonna. In generale queste variabili non sono indipendenti cosicchè la p.d.f. non è semplicemente il prodotto delle usuali distribuzioni gaussiane. Si ha invece che :

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{n/2}|V|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})\right)$$

dove $\vec{x}^T$ e $\vec{\mu}^T$ sono i vettori riga dei corrispondenti vettori colonna $\vec{x}$ e $\vec{\mu}$ mentre V è la matrice degli errori (matrice di covarianza).

Distribuzione binormale (o gaussiana a due dimensioni)

Vediamo ora il caso particolare di due variabili x e y , ognuna delle quali segue una distribuzione gaussiana. La matrice di covarianza è data da :

$$V = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Questa matrice si può invertire se e solo se $\rho \neq \pm 1$, cioè se e solo se le due vriabili non sono perfettamente correlate. In tal caso la matrice inversa è data da :

$$V^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

La p.d.f. binormale è data da :

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}$$

Per visualizzare questa funzione si possono considerare le linee contorno che sono quelle linee dove l'esponente dell'equazione vista è costante. Quindi :

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right] = costante$$

Questa è l'equazione di una ellisse. È possibile dimostrare che se la costante vale $-\frac{1}{2}$, allora l'ellisse è centrata sui valori μ_x e μ_y. Le tangenti all'ellisse sono nei punti $\mu_x \pm \sigma_x$ e $\mu_y \pm \sigma_y$ (Vedi figura ...).

La distribuzione in y , per un fissato valore di x, è una gaussiana la cui deviazione standard è $\sigma_y\sqrt{1-\rho^2}$ e media uguale a $\mu_y + \frac{\rho\sigma_y(x-\mu_x)}{\sigma_x}$

3.4 Distribuzione Uniforme

Questa distribuzione descrive una probabilità che sia costante in un certo intervallo e zero all'esterno :

$$P(x) = \begin{cases} \frac{1}{b-a} & \text{per } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases} \tag{3.1}$$

- Valore di aspettazione (valore medio)

$$E[x] = \int_a^b x \frac{dx}{b-a} = \frac{b-a}{2}$$

- Varianza

$$V[x] = \int_a^b \left[x - \frac{a+b}{2} \right]^2 \frac{dx}{b-a} = \frac{(b-a)^2}{12}$$

Notiamo che se a=0 e b=1, allora la cdf della nostra distribuzione uniforme è :

$$G(x) = \int_0^x dx = x$$

Esempio :

Risoluzione spaziale in una microstrip .....

3.5 Distribuzione Esponenziale

La distribuzione esponenziale di una variabile continua casuale x ($0 \leq x < \infty$) è definita da :

$$f(x;\xi) = \frac{1}{\xi} e^{-x/\xi}$$

- Valore di aspettazione di x

$$E[x] = \frac{1}{\xi} \int_0^\infty x e^{-x/\xi} dx = \xi$$

- Varianza di x

$$V[x] = \frac{1}{\xi}\int_0^\infty (x-\xi)^2 e^{-x/\xi} dx = \xi^2$$

Figure .....

Esempio :

Tempo di decadimento di una risonanza misurata nel suo sistema di riferimento a riposo. ξ in questo caso rappresenta il tempo di vita media della particella.

Si noti che :

$$f(t - t_0 | t \geq t_0) = f(t)$$

Non dipende dall'istante iniziale (vale solo per questa pdf).

3.6 Distribuzione χ^2

La distribuzione χ^2 (chi-quadrato) di una variabile casuale continua z ($0 \leq z < \infty$) è data da

$$f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \qquad n = 1, 2, ...$$

Il parametro n è detto **numero di gradi di libertà** . La funzione Γ è definita da :

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$$

ed ha le seguenti proprietà:

$$\Gamma(n) = (n-1)! \qquad \text{per n intero}$$

$$\Gamma(x+1) = x\Gamma(x)$$

e

$$\Gamma(1/2) = \sqrt{\pi}$$

- **Valore di aspettazione di z** :

$$E[z] = \int_0^\infty z \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} dz = n$$

- **Varianza di z** :

$$V[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} dz = n^2$$

Questa distribuzione ha particolare importanza in statistica. Si può dimostrare che se si hanno N variabili casuali x_i con distribuzioni di probabilità gaussiane di valor medio ν_i e varianza σ_i^2, la funzione :

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \nu_i)^2}{\sigma_i^2}$$

è distribuita secondo una distribuzione del χ^2 con N gradi di libertà. È legata a questa proprietà l'importanza che ha questa distribuzione nei test di bontà dei fit che vedremo in seguito.

Applicazione : Per sommare probabilità di Poisson , è comodo usare la relazione:

$$\sum_{n=0}^{n_{oss}} P(n;\nu) = \int_{2\nu}^{\infty} f_{\chi^2}(z; n_d = 2(n_{oss}+1))dz = 1 - F_{\chi^2}(2\nu; n_{dof} = 2(n_{oss}+1))$$

dove f_{χ^2} e' la p.d.f. del χ^2 per n_d gradi di liberta' mentre F_{χ^2} e' la corrispondente c.d.f.

3.7 Distribuzione di Cauchy

La p.d.f. della distribuzione di Cauchy (detta anche Breit-Wigner o anche Lorentziana) di una variabile continua x ($0 \leq x < \infty$) è definita da :

$$\frac{a}{\pi} \frac{1}{(a^2 + (x-b)^2)}$$

con $a > 0$.

In Fisica Subnucleare è usata (col nome di Breit-Wigner) per descrivere il decadimento di una particella in altre particelle piu‘ leggere.

Questa distribuzione non ha un valore di aspettazione ed una varianza (e neanche i momenti di ordine superiore) in quanto gli integrali che definiscono queste quantità sono divergenti. b può essere considerato come valore medio se l'integrale che lo definisce è interpretato come valore principle di Cauchy.

Dato l'integrale $\int_{-\infty}^{+\infty} f(x)dx$ si dice valore principale di Cauchy:

$$\lim_{c\to\infty} \int_{-c}^{c} f(x)dx$$

Questo limite può esistere anche se i due integrali separati (da $-\infty$ a zero e da zero a $+\infty$ sono divergenti.

a nella formula di Breit-Wigner è legato al tasso di decadimento della particella.

3.8 Distribuzione t di " Student"

Un'altra distribuzione di notevole rilevanza pratica è quella di "Student" [1] . Supponiamo di avere una variabile casuale Z che segua una distribuzione normale standardizzata ed un'altra variabile casuale U,indipendente dalla variabile Z, che segua una distribuzione χ^2 con n gradi di libertà, indipendente dalla prima. Si può dimostrare che la variabile casuale T, definita da :

$$T = \frac{Z}{\sqrt{\frac{U}{n}}}$$

segue la seguente distribuzione :

$$f(t;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}} \frac{1}{\Gamma(n/2)(1+t^2/n)^{(n+1)/2}}$$

detta distribuzione t di Student con n gradi di libertà.

Per n=1 la distribuzione t di Student diventa la distribuzione di Cauchy. Al crescere del numero di gradi di libertà la distribuzione t di Student si avvicina a quella gaussiana standardizata.

Figure .....

3.9 Legge dei Grandi Numeri

Consideriamo una variabile casuale che segua una determinata distribuzione di probabilità . Per conoscere i parametri di questa distribuzione faccio un certo numero di misure che mi permettono di stimare i parametri della distribuzione (in particolare il valore di aspettazione e la varianza). Come detto è in ciò che consiste l'inferenza statistica. Queste cose le vedremo in particolare in seguito.

Il valore di aspettazione E[x] $=\mu$ della nostra variabile casuale x presuppone una media su tutta la popolazione di dati (teoricamente infinita). Consideriamo un campione finito casuale costituito da n misure della variabile considerata e sia $\overline{x}_n$ la media aritmetica di queste misure (detta anche media campionaria). Il problema che mi pongo è : a partire da $\overline{x}_n$ posso fare delle inferenze attendibili su μ ?

[1] Pseudonimio di William S. Gosset, un pioniere della statistica che lavorara per la Guinness Brewery a Dublino.

Per dimostrare che questo è possibile si fa uso della legge (debole) dei grandi numeri che si può enunciare cosi :

È possibile determinare un intero positivo n tale che prendendo un campione casuale di dimensione maggiore od uguale ad n di una variabile casuale x distribuita con valore di aspettazione μ, la media campionaria $\overline{x}_n$ differisca da μ per una quantità piccola a piacere .

3.10 Teorema Limite Centrale

Questo teorema ha un ruolo fondamentale nella teoria della inferenza statistica. Supponiamo di avere n variabili continue x_i, indipendenti, con media μ_i e varianza σ_i^2. Il **teorema limite centrale** stabilisce che la variabile casuale $x = \sum_{i=1}^{n} x_i$ per grandi n (per n tendente ad ∞) tende ad essere distribuita secondo una gaussiana di valore medio $\mu = \sum_i \mu_i$ e varianza $\sigma^2 = \sum_i \sigma_i^2$.

La cosa veramente importante in questo teorema è che non ha importanza la natura delle distribuzioni di probabilità delle variabili casuali x_i. Qualunque quantità , che è originata dall'effetto cumulativo di molte altre quantità (comunque distribuite), almeno approssimativamente sarà distribuita secondo una gaussiana. L'errore casuale di una misura sperimentale, ad esempio, è distribuito secondo una gaussiana perchè è dovuto al contributo di molti effetti indipendenti.

È spiegabile con questo teorema l'importanza notevolissima che ha la distribuzione gaussiana in tutta la statistica e nel calcolo delle probabilità.

L'utilizzo nella pratica di questo teorema richiede attenzione in particolari casi. Il problema è questo: dato un campione finito di misure , che supponiamo distribuito secondo una gaussiana, in che misura questa nostra supposizione è vera ? Ci sono situazioni nelle quali particolari contributi alla somma danno contributi " non gaussiani " , cioè danno contributi alle code nella distribuzione , contributi che hanno effetti non gaussiani. Il problema di effetti non gaussiani è in certi tipi di analisi particolarmente delicato. Un esempio di ciò verrà visto quando parleremo di intervalli di confidenza.

Figure 3.3: Distribuzioni poissoniane con diversi valori di aspettazione.

Figure 3.4: Distribuzione binomiale con $N = 150$ e $p = 0.013$

Figure 3.5: Distribuzioni Gaussiane. A sinistra a) gaussiane standard ($\mu = 0$,$\sigma = 1$ linea continua), gaussiana con $\mu = 3$,$\sigma = 1.5$ linea tratteggiata, gaussiana con $\mu = 3$,$\sigma = 2$ punto-linea; a destra gaussiana con $\mu = 3$,$\sigma = 2$. Le linee verticali sono a distanza di una σ dal valore centrale. L'area compresa tra la gaussiana e queste due linee è pari al 68.27 % sotto tutta la curva.

Figure 3.6: Confronto tra una distribuzione poissoniana con valore di aspettazione 25 ed una gaussiana con valore medio 25 e deviazione standard 5.

Chapter 4

Il Metodo Monte Carlo

- Distribuzioni Uniformi di Numeri Casuali
- Metodo della Trasformazione
- Accettaziione-Reiezione
- Applicazioni del Metodo Monte Carlo

4.1 Distribuzione Uniforme di Numeri Casuali

Numeri casuali distribuiti in modo uniforme tra 0 e 1 vengono generati usando un **generatore di numeri casuali** in un computer. Tipicamente un generatore casuale di numeri uniformi tra 0 e 1 è chiamato RAND(), ecc. Il generatore viene inizializato con un seme e produce una sequenza di numeri casuali con un certo periodo, ecc Data una distribuzione uniforme tra 0 e 1 è possibile generare una distribuzione uniforme di numeri casuali in qualunque intervallo. Per esempio una distribuzione uniforme di un angolo ϕ tra 0 e 2π si può ottenere in questo modo :

$$\phi = 2\pi RAN()$$

essendo RAN() un generatore uniforme di numeri casuali tra 0 e 1.

4.2 Metodo della Trasformazione

Data una distribuzione uniforme di una variabile casuale r vogliamo determinare una funzione x(r) che sia distribuita secondo una specifia p.d.f. f(x). Sia g(r) la p.d.f. della variabile r. Quindi la probabilità di avere un valore di r compreso tra r e r +dr è dato da g(r)dr. Questa probabilità deve essere uguale alla probabilità di trovare un valore di x compreso tra x(r) e x(r) + d x(r) , cioè f(x)dx :

$$g(r)dr = f(x)dx$$

Per avere un x(r) che soddisfi a questa condizione chiediamo che la probabilità che r sia minore di un certo valore r' sia uguale la probabilità di avere x inferiore ad x(r'). Quindi bisogna trovare una funzione x(r) cosi che :

$$F(x(r)) = G(r)$$

dove F e G sono le c.d.f. corrispondenti alle p.d.f. f e g.

In una distribuzione uniforme si ha G(r) = r con $0 \leq r \leq 1$ e quindi :

$$F(x(r)) = \int_{-\infty}^{x(r)} f(x')dx' = \int_{-\infty}^{r} g(r')dr' = r$$

In taluni casi è possibile risolvere analiticamente la relazione precedente, ottenendo x(r).

Consideriamo ad esempio una distribuzione esponenziale. La relazione appena vista in questo caso diventa:

$$\int_{0}^{x(r)} \frac{1}{\xi} e^{-x'/\xi} dx' = r$$

Questa si puo‘ integrare, ottenendo:

$$x(r) = -\xi \log(1 - r)$$

Se r è una variabile casuale tra 0 e 1 , anche 1-r è una variabile casuale tra 0 e 1. Di conseguenza la variabile

$$x(r) = -\xi \log r$$

segue una distribuzione esponenziale con valore medio ξ.

4.3 Metodo Accettazione-Reiezione

In molti casi una soluzione analitica per avere la p.d.f. cercata è molto difficile per cui si ha un diverso approccio. Supponiamo che la p.d.f. f(x) cercata sia inscrivibile in un rettangolo dove la x è compresa tra un valore x_{max} e x_{min} ed f(x) è minore di un certo valore massimo f_{max}. Una distrribuzione f(x) può essere ottenuta in questo modo:

- Si genera un numero casuale compreso tra x_{max} e x_{min} :

  $$x = x_{min} + r_1(x_{max} - x_{min})$$

  essendo r_1 un numero casuale distribuito uniformemente tra 0 e 1.

- Si genera una seconda variabile casuale u distribuita uniformemente tra 0 ed f_{max} :

  $$u = r_2 f_{max}$$

  essendo r_2 un numero casuale distribuito uniformemente tra 0 e 1.

- Se u < f(x) allora x è accettato, altrimenti x viene rigettato e si ripete il procedimento

I valori cosi accettati sono distribuiti secondo una p.d.f. f(x). Infatti per costruzione la probabilità che x sia acettata è proporzionale ad f(x).

Talvolta il procedimento descritto è molto lento e quindi poco pratico. In questi casi la f(x) da generare viene inscritta in una qualunque curva $g(x)$ che rappresenta numeri casuali che possono essere generati in accordo a $g(x)/\int g(x')dx'$, usando il metodo della trasformazione. L'algoritmo nel caso generale è il seguente :

- Si genera un numero casuale che abbia una p.d.f. $g(x)/\int g(x')dx'$

- Si genera un secondo numero casuale u uniformemnte distribuito tra 0 e $g(x)$ Se u < f(x), x viene accettata altrimenti viene rigettata ed il procedimento viene ripetuto.

4.4 Applicazioni del Metodo Monte Carlo

Chapter 5

Stima Puntuale dei Parametri

- Problema della Stima Puntuale dei Parametri
- Proprietà degli Stimatori
- Stimatori di Media, Varianza e Covarianza
- Limite Inferiore della Varianza

5.1 Problema della Stima Puntuale dei Parametri

Si abbia una variabile casuale X e si voglia stimare la p.d.f. di questa variabile a partire da un campione di osservaxioni $x_1, x_2, ..., x_n$ della stessa variabile casuale. Per esempio misuriamo 10 volte la lunghezza di una sbarra. Noi assumiamo che le n osservazioni x_i siano n variabili casuali indipendenti ed identicamente distribuite (i.i.d.). Si noti che a rigore dovremmo scrivere il campione di dimensione n cosi : ($X_1, X_2, ..., X_n$) . Quando facciamo le n misure abbiamo la sequenza di numeri dati dalle misure. Noi vogliamo determinare la p.d.f. della variabile casuale X che chiamiamo la popolazione. L 'insieme delle n osservazioni costituisce un campione (di dimensione n) (Anche l' insieme delle n variabili X_i è detto campione (di dimensione n) . La distribuzione esatta delle osservazioni fatte è in generale non conosciuta. Spesso è nota a priori una qualche conoscenza o ipotesi di lavoro, per esempio che la p.d.f. della variabile X appartenga a qualche famiglia parametrica :

$$f(x) = f(x \mid \theta) \; \theta \subset \Theta$$

Noi vogliamo trovare un metodo per stimare il parametro (o i parametri) θ a partire dalle n osservazioni fatte (inferenza statistica). Per stimare questi parametri si costruiscono stimatori adatti allo scopo. Il processo che porta alla stima di un parametro si dice fit del parametro

5.2 Proprietà degli Stimatori

Supponiamo di voler misurare il valore medio della altezza di n studenti. Potrei farlo in molti modi:

1. Sommo le altezze degli n studenti e divido la somma per n (media aritmetica).
2. Sommo le altezze degli n studenti e divido per n-2
3. Sommo l'altezza dei primi 10 studenti e faccio la media.
4. Sommo l'altezza dei due studenti piu alti e dei due studenti più bassi e divido per quattro.
5. Tolgo i tre studenti più alti ed i tre più bassi e faccio la media (aritmetica) dei rimanenti.
6. Moltiplico le n altezze tra di loro e poi prendo la radice n-sima
7. Sommo l'altezza del secondo studente, del quarto, del sesto e cosi via e poi divido per n/2.

In ognuno di questi modi ottengo un valore medio (in generale i valori medi ottenuti sono diversi). Ognuno dei modi considerati è uno stimatore del valore medio.

Una generica funzione t delle n osservazioni a disposizione è detta statistica :

$$t = t(x_1, ..., x_n)$$

Stimatore è una statistica concepita per stimare il valore del parametro (o dei parametri) θ di una p.d.f. . Noi lo stimatore del parametro θ lo indichiamo con $\hat{\theta}$. Nella lista sopra sono indicate diverse possibili statistiche allo scopo di stimare il parametro valore medio della distribuzione delle altezze degli studenti presi in considerazione. Naturalmente non ci possiamo aspettare :

$$\hat{\theta} = \theta$$

in modo esatto ma possiamo cercare che $\hat{\theta}$ sia uguale a θ in media o che sia il valore vicino a θ con più alta probabilita ecc.

L' insieme delle n misure costituisce esso stesso una variabile casuale $\vec{X}$ ad n-dimensioni. Se ripetessimo l' esperimento , cioè rifacessimo le n misure, avremmo un nuovo valore $\vec{x}$ della variabile casuale $\vec{X}$. Per ognuno di questi valori nello spazio ad n-dimensioni lo stimatore $\hat{\theta}(\vec{x})$ assumerà valori diversi distribuiti secondo una p.d.f. g($\hat{\theta}$; θ), dipendente in generale dal valore "vero " θ. La p.d.f. di una statistica è detta p.d.f. di campionamento.

Poichè si suppone che tutte le misure siano i.i.d., allora ognuna delle n quantità x_i è descritta dalla stessa p.d.f. f(x_i) è la p.d.f. dell'insieme delle n-osservazioni :

$$f_{campione}(x_1, x_2, ...x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

Il valore di aspettazione di uno stimatore $\hat{\theta}$ con una p.d.f. di campionamento g($\hat{\theta}$; θ) è dato da :

$$E[\hat{\theta}(\vec{x})] = \int \hat{\theta} g(\hat{\theta};\theta) d\hat{\theta} = \iint \theta(\vec{x}) f(x_1;\theta)\cdot\cdot f(x_n;\theta) dx_1 \cdot\cdot dx_n$$

Questo è il valore di aspettazione dello stimatore pensando di ripetere infinite volte l'esperimento (ognuno costituito da una n-tupla di valori casuali).

La differenza tra il valore di aspettazione dello stimatore ed il valore vero della variabile stimata è detta distorsione (bias) :

$$b = E[\hat{\theta}] - \theta$$

Non esistono stimatori giusti o sbagliati ma buoni o cattivi. Uno stimatore lo diciamo buono se è consistente, non distorto (unbiased) ed efficiente.

- **Stimatore Consistente**

  Uno stimatore che tende al valore vero quando il numero di dati tende all' infinito è detto stimatore consistente. Sia $\hat{a}$ uno stimatore del parametro a, allora questo stimatore sarà **consistente** se :

  $$\lim_{n\to\infty} \hat{a} = a$$

  Questo significa che considerata una sequenza di stimatori $\hat{a}_n$ di dimensione n, allora al crescere della dimensione n del campione usato la distribuzione di $\hat{a}_n$ è concentrata in un intervallo sempre più piccolo attorno al valore (ignoto) del parametro a che si sta stimando. Nell' esempio dell' altezza media degli studenti lo stimatore media (aritmetica) è uno stimatore consistente della altezza media. Il secondo estimatore è anch'esso consistente ;il terzo stimatore invece non è consistente.

- **Stimatore non Distorto (unbiased)**

  Se il valore di aspettazione di uno stimatore E[$\hat{a}$] è uguale al valore vero del parametro a, lo stimatore è detto **non distorto** .

  Lo stimatore media aritmetica è uno stimatore non distorto:

  $$E[\hat{a}] = E[\frac{a_1 + a_2 + ... + a_n}{n}] = \frac{E[a] + E[a] + ... + E[a]}{n} = \frac{nE[a]}{n} = a$$

  Nel caso dello stimatore 2 si ha :

  $$E[\hat{a}] = \frac{n}{n-2}a$$

  Quindi lo stimatore 2 è uno stimatore distorto.

  Se al crescere della dimensione n del campione il valore di aspettazione del campione tende al valore vero, allora lo stimatore è detto **asintoticamente non distorto**.

  $$\lim_{n\to\infty} E[\hat{a}_n] = a$$

  Lo stimatore 2 dell'esempio precedente è asintoticamente non distorto.

- **Stimatore Efficiente** Nella scelta di uno stimatore , certamente preferibile è quello che dà il parametro con la minima varianza. Tra due o più stimatori di un parametro a quello più efficiente è quello che ha varianza più piccola. Lo stimatore 7 e‘ meno efficiente dello stimatore 1 perchè ha una varianza $\sqrt{2}$ maggiore.

Un'altra misura della qualità di uno stimatore è l' **errore quadratico medio (MSE)** , cosi definito:

$$MSE = E[(\hat{\theta} - \theta)^2]$$

Si consideri l'identità :

$$E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}) + (E[(\hat{\theta}] - \theta)]^2 = E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta} - \theta])^2$$

La seconda equaglianza sussiste in quanto

$$E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] = (E[\hat{\theta}] - \theta)E[\hat{\theta} - E[\hat{\theta}]] = 0$$

Dall'identità sopra vista si ha che :

$$MSE = V[\hat{\theta}] + b^2$$

MSE è la somma della varianza e del quadrato del bias. Si interpreta come la somma degli errori statistici e sistematici

Un'altra caratteristica dello stimatore è la **sufficienza**. Uno stimatore del parametro θ a partire da n misure $x_1, x_2, \cdots, x_n$ è detto sufficiente se l'informazione che dà sul parametro stimato e' esaustiva di tutta l'informazione sul parametro presente nelle n misure fatte. Ad esempio la media del campione $\bar{x}$ presa come stimatore della media μ della popolazione è esaustiva di tutta l'informazione su μ presente nelle n misure. Non posso trovare alcuna altra funzione delle n osservazioni che aggiunga conoscenza su μ (ulteriore a quella che ho ottenuto con lo stimatore media del campione (media aritmetica). Noi diciamo che questo stimatore è sufficiente.

Nelle situazioni pratiche spesso si deve cercare un compromesso. Per esempio lo stimatore più efficiente può essere distorto per cui bisogna scegliere se é preferibile, nel caso pratico considerato, uno stimatore non distorto e meno efficiente di uno più efficiente ma un pò distorto.

5.3 Stimatori di Media, Varianza e Covarianza

5.3.1 Stimatore della Media

Come abbiamo visto nel paragrafo precedente, la media aritmetica $\bar{x}$ è uno stimatore non distorto del valore di aspettazione del campione μ (o E[x]) . Infatti:

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

dal momento che

$$E[x_i] = \iint x_i f(x_1) f(x_2) ... f(x_n) dx_1 dx_2 ... dx_n = \mu$$

La varianza dello stimatore della media è data da :

$$V[\hat{\mu}] = \frac{\sigma^2}{n}$$

Attenzione : questa è la varianza sul valore di aspettazione di X (valore medio) ed è data dalla varianza della variabile x diviso il numero di misure. La radice quadrata della varianza sul valore di aspettazione è nota **errore standard della media ,** σ_μ :

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}}$$

5.3.2 Stimatore della Varianza

Supponiamo per ora di essere nel caso fortunato ma raro di conoscere il valore vero μ di una variabile casuale. In questo caso possiamo stimare la varianza col seguente stimatore:

$$\hat{V}[x] = \frac{1}{n} \sum (x_i - \mu)^2$$

Questo stimatore è consistente e non distorto. Infatti :

$$E[\hat{V}[x]] = \frac{nE[(x-\mu)^2]}{n} = E[(x-\mu)^2] = V[x]$$

Come detto , in genere il valore vero μ non è noto . Noi possiamo usare al suo posto il valore stimato $\bar{x}$:

$$\hat{V}[x] = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum (x_i^2 - \bar{x}^2)$$

Si verifica facilmente che questo stimatore è distorto. Prendendo infatti il valore di aspettazione si ha :

$$E[\hat{V}[x]] = n \frac{E[x^2 - \bar{x}^2]}{n} = E[x^2] - E[\bar{x}^2]$$

Poichè E[x] = E[$\bar{x}$] , allora :

$$E[\hat{V}[x]] = E[x^2] - E[x]^2 - (E[\bar{x}^2] - E[\bar{x}]^2) = V[x] - V[\bar{x}]$$

Da questa si deduce che :

$$E[\hat{V}[x]] = \left(1 - \frac{1}{n}\right) V[x] = \frac{n-1}{n} V[x] \neq V[x]$$

Quindi questo stimatore della varianza è distorto. Nota la distorsione (il bias) possiamo però correggerla, introducendo un altro stimatore della varianza cosi definito :

$$\hat{V}[x] = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

Questo stimatore, oltre che consistente, è anche non distorto. Infatti il valore di aspettazione di questo stimatore è dato da :

$$E[\hat{V}[x]] = \frac{n}{n-1} E[x^2 - \bar{x}^2] = \frac{n}{n-1} \left(E[x^2] - E[\bar{x}^2]\right) = \frac{n}{n-1} V[x] \left(1 - \frac{1}{n}\right) = V[x]$$

Questo stimatore consistente e non distorto è chiamato varianza del campione.

5.3.3 Stimatore della Covarianza

In modo analogo si può introdurre uno stimatore non distorto della matrice di covarianza V_{xy} nel modo seguente:

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1} (\overline{xy} - \bar{x}\bar{y})$$

Da questa otteniamo facilmente uno stimatore r_{xy} per il coefficiente di correlazione ρ

$$r_{xy} = \frac{\hat{V_{xy}}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{j=1}^{n} (x_j - \bar{x})^2 \cdot \sum_{k=1}^{n} (y_k - \bar{y})^2\right)^{1/2}} =$$

$$\frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

Questo risultato si generalizza facilmente al caso di n variabili $x_1, x_2, \cdots x_n$

5.3.4 Limite Inferiore della Varianza

Se abbiamo diversi stimatori consistenti e non distorti di un parametro incognito, noi scegliamo ovviamente quello più efficiente, cioè quello con varianza più piccola. Come facciamo a sapere che abbiamo trovato il migliore possibile stimatore? Noi vedremo che la varianza di uno stimatore consistente e non distorto non può essere inferiore ad un certo limite inferiore che dipende dal tipo di distribuzione e dalla dimensione del campione.

Se abbiamo un campione di osservazioni $x_1, x_2, ..., x_n$ i.i.d., la p.d.f. congiunta calcolata nei punti misurati è data da :

$$L(\theta) = L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

L è detta **funzione di verosimiglianza (Likelihood Function)**(LF) del campione.

Consideriamo ora un generico stimatore $\hat{\theta}(x_1, x_2, ...x_n)$, non distorto ($E[\hat{\theta}] = \theta$) , allora se LF è sufficientemente regolare $V[\hat{\theta}]$ ha un limite inferiore dato da :

$$V[\hat{\theta}] \geq \frac{1}{E\left(\left[\frac{\partial \log L}{\partial \theta}\right]^2\right)_\theta}$$

dove θ è il valore vero del parametro incognito.

Sotto la condizione di sufficiente regolarità della LF si può dimostrare che :

$$E\left(\left[\frac{\partial \log L}{\partial \theta}\right]^2\right)_\theta = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]_\theta$$

Si ha quindi che :

$$V[\hat{\theta}] \geq -\frac{1}{E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]_\theta}$$

Questo valore minimo della varianza è detto anche **limite inferiore di Cramer-Rao**. L' inverso del limite inferiore è detto **informazione di Fisher** del campione :

$$I(\theta) = -E\left[\frac{\partial^2 logL}{\partial \theta^2}\right]_\theta$$

Noi siamo interessati ai casi in cui questo limite inferiore della varianza viene raggiunto. Questi stimatori si chiamano stimatori con limite di minima varianza (MVB).

Si può dimostrare che uno stimatore $\hat{\theta}$ di θ raggiunge il MVB se e solo se la derivata parziale di log L rispetto a θ fattorizza nel seguente modo :

$$l' = \frac{\partial logL}{\partial \theta} = A(\theta)(\hat{\theta} - \theta)$$

dove A(θ) non dipende dalle variabili osservate. A(θ) è uguale all' informazione di Fisher ed inoltre:

$$V[\hat{\theta}] = \frac{1}{A(\theta)}$$

Stimatori MVB esistono solo per particolari classi di distribuzioni (alcune delle quali però molto importanti dal punto di vista delle applicazioni pratiche).

La maggior parte degli stimatore ha una varianza maggiore di quella degli stimatori MVB. Noi definiamo efficienza ϵ di uno stimatore il rapporto tra MVB e valore V della varianza dello stimatore :

$$\epsilon = \frac{MVB}{V}$$

- Stima della vita media di una particella.

  Segue una distribuzione esponenziale (si trascurano effetti di risoluzione, errori sperimentali):

  $$f(t) = \frac{1}{\tau} e^{-t/\tau}$$

  Per un campione di n misure si ha :

  $$l' = \frac{\partial \log L}{\partial \tau} = -n/\tau + \sum t_i/\tau^2 = (n/\tau^2)\left(\sum t_i/n - \tau\right)$$

  Si ha quindi che $\bar{t}_n$ è uno stimatore MVB di τ, A(τ) = n/τ^2 e V[$\bar{t}_n$] = τ^2/n

- Nel caso di una distribuzione di Cauchy si ha :

  $$l' = \frac{\partial logL}{\partial \theta} = 2\sum \frac{x_i - \theta}{1 + (x_i - \theta)^2}$$

  e qui come si vede non esiste uno stimatore MVB.

  Se il MVB non esiste , possiamo cercare uno stimatore non distorto che tra tutti quelli non distorti ha varianza minima. Questi stimatori si dicono di **Minima Varianza (MV)** . Si può dimostrare che gli stimatori MV , se esistono, sono unici.

Chapter 6

Metodo della Massima Verosimiglianza

- Stimatori di Massima Verosimiglianza
- Varianza negli Stimatori di Massima Verosimiglianza
- Massima Verosimiglianza Estesa
- Massima Verosimiglianza con Dati Istogrammatiq
- Bontà del Fit col ML
- Combinazione di più Esperimenti col ML
- Stimatori Bayesiani
- Postulato di Bayes e Scelta della Distribuzione Iniziale

6.1 Stimatori di Massima Verosimiglianza

Quando non esiste alcun ovvio stimatore MVB, come facciamo a costruire un buon stimatore? La più importante prescrizione per cercare uno stimatore è data dal **Principio della Massima Verosimiglianza** (**Maximum Likelihood (ML)**) :

In presenza di un campione di dati (x_1,x_2,x_3, $\cdots$,x_n) , la LF è una funzione del parametro ignoto θ , L = L(θ). Lo **stimatore di maximum likelihood (MLE)** $\tilde{\theta}$ è definito come quel valore del parametro θ che massimizza la LF :

$$L(x_1, \cdots, x_n; \tilde{\theta}) \geq L(x_1, \cdots, x_n; \theta)$$

per tutti i possibili valori di θ. Se L è due volte differenziabile rispetto a θ, allora MLE può essere trovato tra gli zeri della derivata prima della LF :

$$\frac{\partial L}{\partial \theta} = 0$$

con la condizione che la derivata seconda della LF calcolata nello zero della derivata prima sia negativa.

Se ci sono più massimi locali, bisogna prendere quello maggiore.

Spesso è più comodo lavorare con il logaritmo della LF (log likelihood function). Il MLE si ottiene dagli zeri della derivata prima rispetto a θ del logaritmo della LF. Questa equazione talvolta può essere risolta analiticamente , negli altri casi bisogna risolverla numericamente.

Esempio 1. Sia $(x_1, \cdots, x_n)$ un campione che segue una distribuzione gaussiana con media non nota μ e varianza non nota σ^2. Calcoliamo la LF :

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

Si ha perciò

$$\log L(x_1, \cdots, x_n; \mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Differenziando rispetto a μ e σ^2 si ottiene:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu) = 0$$

Quindi il MLE di μ è :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Come si vede il MLE per la media è non distorto mentre la stima ML della varianza è distorta. Il valore di aspettazione dello stimatore della varianza è infatti:

$$E[\widehat{\sigma^2}] = \frac{n-1}{n}\sigma^2$$

Questa stima comunque è asintoticamente non distorta. Il bias in questo caso può essere corretto, considerando la varianza del campione s^2 :

$$\hat{s^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Questo è uno stimatore non distorto della varianza della distribuzione gaussiana.

Si noti che che:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

Lo stimatore di ML raggiunge il limite inferiore di Cramer-Rao.

Esempio2. Supponiamo che una variabile casuale sia distribuita secondo una poissoniana:

$$f(x;\theta) = \frac{\theta^x}{x!}e^{-\theta}$$

con $\theta > 0$ parametro ignoto da stimare e x = 0,1, ...

e si abbia un campione di n misure di questa variabile $(x_1, x_2, .., x_n)$. Allora la LF e' data da:

$$L(\theta) = L(x_1, x_2, .., x_n; \theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$$logL(\theta) = -n\theta + \left(\sum_{i=1}^{n} x_i\right) log\theta - log\prod_{i=1}^{n} x_i!$$

Derivando rispetto al parametro θ, otteniamo:

$$\frac{\partial logL(\theta)}{\partial \theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0$$

che ha la soluzione

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Poiche' la derivata seconda rispetto a θ calcolata nello zero della derivata prima e' negativa, lo zero trovato e' un punto di massimo. Anche in questo caso lo stimatore di ML raggiunge il limite di Cramer-Rao. Il valore trovato (media aritmetica) rappresenta la stima di massima verosimiglianza per il parametro θ.

Vi possono essere diversi stimatori consistenti e non distorti di un certo parametro. Ad esempio la mediana del campione è il valore del parametro che divide in campione in due sottocampioni con stesso numero di misure. Se la distribuzione delle misure è simmetrica, la mediana del campione è anche uno stimatore consistente non distorto della media del campione.

Se la distribuzione del campione è gaussiana (media μ, varianza σ^2) , allora la varianza sulla media aritmetica $V[\bar{x}_n]$ e quella sulla mediana $V[\tilde{x}_n]$ sono date da :

$$V[\bar{x}_n] = \frac{\sigma^2}{n}$$

$$V[\tilde{x}_n] = \frac{\sigma^2}{n} \cdot \frac{\pi}{2}$$

La mediana del campione ha una varianza maggiore della media del campione. La mediana però è uno stimatore più robusto (cioè è influenzato meno fortemente da misure lontane o poco precise).

Lo stimatore ML è invariante sotto trasformazioni funzionali . Questo significa che se T è MLE di θ e se u(θ) è una funzione di θ, allora u(T) è il MLE per u(θ). Questa proprietà di invarianza non è vera per tutti gli stimatori. Ad esempio se T è uno stimatore non distorto di θ , non segue che T^2 sia uno stimatore non distorto di θ^2.

Perchè dovremmo adottare il principio di maximum likelihood ? Gli stimatori ML hanno alcune proprietà ottimali che giustificano il loro uso.

1. La prima ragione importante consiste nel fatto che se esiste uno stimatore MVB, esso è invariabilmente trovato applicando il principio di maximum likelihood.

2. Sotto condizioni abbastanza generali essi sono consistenti.

3. Al crescere della dimensione del campione il MLE non è distorto e la sua varianza è uguale al MVB : MLE è (asintoticamente) efficiente.

4. Al crescere della dimensione del campione, MLE segue una distribuzione gaussiana.

6.2 Varianza negli Stimatori di ML

Abbiamo visto come col ML sia possibile stimare i parametri di una p.d.f. (o anche PDF) in un campione di n osservazioni. Ora vogliamo vedere che errori assegnare ai valori stimati per i parametri. Ripetendo l' esperimento avremmo un altro insieme di n misure e cosi stimerei valori diversi dei parametri. Se ripetessi un numero elevato di volte l'esperimento i valori dei parametri ottenuti avrebbero una determinata distribuzione. Noi vogliamo determinare la varianza (deviazione standard) di queste distribuzioni. Teniamo presente che la distribuzione dei parametri stimati per n sempre più grande diventa gaussiana.

6.2.1 Metodo Analitico

In taluni casi è possibile calcolare la varianza per via analitica. Come esempio possiamo calcolare la varianza dello stimatore della vita media in un decadimento esponenziale.

Stimatore di ML :

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Si ha :

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2 = \frac{\tau^2}{n}$$

τ è il valore incognito della vita media. Nella formula precedente si sostituisce il valore vero con quello fornito dallo stimatore . Questo è possibile grazie alla invarianza sotto trasformazione degli stimatori di maximum likelihood. Quindi :

$$V[\hat{\tau}] = \frac{\hat{\tau}^2}{n}$$

La deviazione standard sul valore del parametro stimato e‘ quindi:

$$\hat{\sigma}_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

6.2.2 Limite Inferiore di Cramer-Rao

La varianza può essere calcolata a partire dal limite inferiore di Cramer-Rao (assumendo zero bias ed efficienza). L'informazione di Fisher è determinata mediante le derivate seconde calcolate con i dati misurati e nel valore del parametro θ=θ_0 stimato dal ML :

$$V[\hat{\theta}] = -\left(\frac{1}{\frac{\partial^2 logL}{\partial\theta^2}}\right)\Bigg|_{\theta=\theta_0}$$

Questo è il metodo generalmente usato per determinare la varianza , quando la LF è massimizzata numericamente (per esempio usando MINUIT). Questo risultato si generalizza al caso di fit di più parametri.

6.2.3 Metodo Monte Carlo

Nei casi in cui non sia possibile calcolare analiticamente la varianza dei parametri stimati col ML , questo può essere realizzato tramite Monte Carlo. Dai valori dell'esperimento si deducono i valori dei parametri. Si assumono questi come valori veri e si simulano tanti esperimenti e per ognuno si stimano i parametri. Dalle distribuzioni dei parametri cosi ottenute si determinano le deviazioni standard dei parametri fittati. .....

6.2.4 Metodo Grafico

Consideriamo due casi, il primo con un singolo parametro da stimare ed il secondo con due parametri da stimare:

Metodo Grafico in casi con un parametro

Sviluppo il $logL(\theta)$ in serie di Taylor attorno al valore $\hat{\theta}$ stimato dal ML :

$$\log L(\theta) = \log L(\hat{\theta}) + \left[\frac{\partial \log L}{\partial\theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \log L}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \cdots$$

Trascurando termini di ordine superiore , notando che il secondo termine è nullo e che il primo è uguale a $\log L_{max}$ e riscrivendo il terzo termine mediante il limite inferiore di Cramer-Rao, si ha :

$$\log L(\theta) = \log L_{max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma^2}_{\hat{\theta}}}$$

o anche ponendo $\theta - \hat{\theta} = \pm\hat{\sigma}_{\hat{\theta}}$:

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{max} - \frac{1}{2}$$

Cioè variando di una deviazione standard il valore del parametro stimato il valore del $\log L$ diminuisce di $1/2$ rispetto al suo valore massimo. Questa relazione

permette di determinare graficamente il valore della deviazione standard. La funzione $\log L$ al limite di grandi n diventa una parabola (perchè la funzione di verosimiglianza diventa una curva gaussiana).

Metodo Grafico in casi con due parametri

Nel caso di due parametri da stimare la funzione di likelihood $\mathrm{L}(\theta_1, \theta_2)$ è una superficie tridimenzionale. La forma della LF può essere visualizzata dando curve di livello corrispondenti a valori costanti di LF in un piano (θ_1, θ_2). Con più di un parametro da fittare , spesso si hanno più massimi. Usualmente è relativamente facile identificare quello giusto. Con un massimo solo e con LF regolari, gli errori sui valori stimati dei parametri possono essere ottenuti a partire dal particolare profilo del log-likelihood per il quale $\log L = \log L_{max} - 0.5$. Nel caso con due parametri da fittare e nell'ipotesi di grandi campionamenti, il logaritmo della LF assume la forma :

$$\log L(\theta_1, \theta_2) = logL_{max} - \frac{1}{2(1-\rho^2)} \left[\left(\frac{\theta_1 - \hat{\theta}_1}{\sigma_{\hat{\theta}_1}} \right)^2 + \left(\frac{\theta_2 - \hat{\theta}_2}{\sigma_{\hat{\theta}_2}} \right)^2 - 2\rho \left(\frac{\theta_1 - \hat{\theta}_1}{\sigma_{\hat{\theta}_1}} \right) \left(\frac{\theta_2 - \hat{\theta}_2}{\sigma_{\hat{\theta}_2}} \right) \right]$$

dove ρ è il coefficiente di correlazione tra $\hat{\theta}_1$ e $\hat{\theta}_2$. Il profilo corrispondente a $\log L = \log L_{max} - 0.5$ è in questo caso dato da :

$$\frac{1}{1-\rho^2} \left[\left(\frac{\theta_1 - \hat{\theta}_1}{\sigma_{\hat{\theta}_1}} \right)^2 + \left(\frac{\theta_2 - \hat{\theta}_2}{\sigma_{\hat{\theta}_2}} \right)^2 - 2\rho \left(\frac{\theta_1 - \hat{\theta}_1}{\sigma_{\hat{\theta}_1}} \right) \left(\frac{\theta_2 - \hat{\theta}_2}{\sigma_{\hat{\theta}_2}} \right) \right] = 1$$

Questa è l'equazione di una ellisse centrata attorno ai valori stimati $(\hat{\theta_1}, \hat{\theta_2})$. È detta **ellisse di covarianza**. L'asse principale d'ellisse di covarianza forma con l'asse θ_1 un angolo α dato da :

$$\tan 2\alpha = \frac{2\rho\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}}{\sigma^2_{\hat{\theta}_1} - \sigma^2_{\hat{\theta}_2}}$$

Si può far vedere che (sempre al limite di grandi campioni) le tangenti all'ellisse parallele agli assi distano sempre $\pm\hat{\sigma}_{\hat{\theta_1}}$, $\pm\hat{\sigma}_{\hat{\theta_2}}$ dal punto (θ_1, θ_2), indipendentemente dal valore assunto dal coefficiente di correlazione ρ.

6.3 Massima Verosimiglianza Estesa

Nel calcolo della Massima verosimiglianza basata su un campione di n misure non si tenuto conto che il numero delle misure fatte è esso stesso una variabile casuale. Per tener conto di questo fatto si può moltiplicare la musuale ML per n misure per la probabilità che si osservino n misure :

$$L(\nu,\theta) = \frac{\nu^n}{n!}e^{-\nu}\prod_{i=1}^{n} f(x_i;\theta) = \frac{e^{-\nu}}{n!}\prod_{i=1}^{n}\nu f(x_i;\theta)$$

dove $f(x;\theta)$ è la pdf della variabile casuale x, θ è il parametro (o l'insieme dei parametri) da stimare. Si suppone che nell'esperimento la variabile casuale sia stata misurata n volte ($x_1, x_2, \cdots x_n$). La funzione di verosimiglianza cosi introdotta è detta **funzione di verosimiglianza estesa** . Possiamo considerare i due casi interessanti che la variabile ν dipenda dal parametro θ o che invece venga trattato come un parametro indipendente.

Nel primo caso considerando il logaritmo della funzione di likelihood estesa si ha :

$$logL(\theta) = nlog\nu(\theta) - \nu(\theta) + \sum_{i=1}^{n} logf(x_i;\theta) = -\nu(\theta) + \sum_{i=1}^{n} log(\nu(\theta)f(x_i;\theta))$$

Qui i termini che non dipendono dai parametri sono stati eliminati. Includendo la variabile poissoniana ν , la varianza dello stimatore in generale decresce.

Nel secondo caso nel quale le ν e θ sono trattati come indipendenti, prendendo il logaritmo della ML estesa si trova che lo stimatore di ν è n e che lo stimatore $\hat{\theta}$ (o gli stimatori $\hat{\theta}$) è lo stesso che nella usuale ML. La differenza dei due casi ora sta nel fatto che una variabile che dipende sia da n che da $\hat{\theta}$ ha una sorgente in più di fluttuazione statistica (n è trattata come variabile casuale!). Anche in questo secondo caso comunque ci sono situazioni in cui si dimostra utile l'utilizzo della ML estesa.

Supponiamo infatti che la pdf di una variabile x sia somma di più contributi :

$$f(x;\theta) = \sum_{i=1}^{m}\theta_i f_i(x)$$

Si vuole determinare i parametri θ_i che rappresentano i relativi contributi delle varie componenti. I vari parametri θ_i non sono indipendenti perchè la loro somma deve dare 1. Questo significa che uno dei parametri può essere espresso in funzione degli altri m-1 parametri e la funzione di likelihood verrà scritta in termini dei rimanenti m-1 parametri. Un parametro come si vede è trattato diversamente che gli altri.

Utilizzando la ML estesa e prendendo il logaritmo (e come al solito eliminando i termini che non contengono i parametri) si ha :

$$logL(\nu,\theta) = -\nu + \sum_{i=1}^{n} log\left(\sum_{j=1}^{m}\nu\theta_j f_j(x_i)\right)$$

Possiamo porre $\mu_i = \theta_i\nu$ e riscrivere il logaritmo della LF nel modo seguente :

$$logL(\mu) = -\sum_{j=1}^{m}\mu_j + \sum_{i=1}^{n} log\left(\sum_{j=1}^{m}\mu_j f_j(x_i)\right)$$

I parametri $\mu = \mu_1, \mu_2, ..\mu_m$ sono tutti trattati allo stesso modo, simmetricamente.

6.4 Massima Verosimiglianza con Dati Istogrammati

Come abbiamo visto un modo comodo di visualizzare i dati è quello di presentarli sotto forma di istogramma. L'istogrammazione (a causa della larghezza finita dell'intervallo (bin)) comporta una perdita di informazione. Comunque l'istogrammazione è molto comune non solo perchè permette di visualizzare bene i dati ma anche, come vedremo, di applicare il test del χ^2 per la bontà del fit.

Sia n_{tot} il numero di misure di una variabile casuale X distribuita secondo una p.d.f. $f(x;\theta)$. Noi dobbiamo stimare i parametri $\theta = (\theta_1, \theta_2, \cdots \theta_n)$. Indiciamo con n_i il numero di eventi nel bin i e con :

$$\nu_i(\theta) = n_{tot} \int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx$$

i valori aspettati del numero di dati nel bin i. x_i^{min} e x_i^{max} sono i limiti del bin i. Sia N il numero di bin. L'istogramma può essere visto come una singola misura di un vettore casuale di N dimensioni per il quale la p.d.f. congiunta è una distribuzione multinomiale :

$$f_{congiunta}(\mathbf{n};\nu) = \frac{n_{tot}!}{n_1! \cdots n_N!} \left(\frac{\nu_1}{n_{tot}}\right)^{n_1} \cdots \left(\frac{\nu_N}{n_{tot}}\right)^{n_N}$$

La probabilità di essere nel bin i è data come rapporto tra il valore di aspettazione nel bin i ν_i diviso il numero totale n_{tot}.

Prendendo il logaritmo si ottiene:

$$logL(\theta) = \sum_{i=1}^{N} n_i log\nu_i(\theta)$$

Gli stimatori $\hat{\theta}$ si ottengono massimizzando $\log L$. Quando la larghezza del bin tende a zero (grande numero di eventi) la funzione ML diventa la stessa di quella trovata senza binning. Quindi il " binned ML" tende al "unbinned ML" per valori grandi del numero di misure.

Anche in questo caso il numero totale di eventi n_{tot} puó essere considerato come una variabile casuale con una distribuzione di Poisson con media ν_{tot}. Quindi prima bisogna determinare il numero n_{tot} e quindi fare la distribuzione degli eventi secondo la formula multinomiale vista sopra. Quindi la p.d.f. congiunta per n_{tot} e per $n_1, n_2, \cdots, n_N$ è data da :

$$f_{congiunta}(\mathbf{n};\nu) = \frac{\nu_{tot}^{n_{tot}} e^{-\nu_{tot}}}{n_{tot}!} \frac{n_{tot}!}{n_1!, \cdots, n_N!} \left(\frac{\nu_1}{\nu_{tot}}\right)^{n_1}, \cdots, \left(\frac{\nu_N}{\nu_{tot}}\right)^{n_N}$$

dove $\nu_{tot} = \sum_1^N \nu_i$ e $n_{tot} = \sum_1^N n_i$. Utilizzando queste si ha che :

$$f_{congiunta}(\mathbf{n};\nu) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

con

$$\nu_i(\nu_{tot}, \theta) = \nu_{tot} \int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx$$

Si noti che la pdf congiunta è quella che si ottiene trattando il numero di eventi in ogni bin come una variabile Poissoniana indipendente n_i con valolre medio ν_i. Passando dalla pdf congiunta alla log-lilelihood ed eliminando i termini che non dipendono dai parametri si ha:

$$\log L(\nu_{tot}, \theta) = -\nu_{tot} + \sum_{i=1}^{N} n_i \log \nu_i(\nu_{tot}, \theta)$$

Questa è la funzione log-likelihood estesa per i dati istogrammati. Come nel caso di dati non istogrammati, se non c'è alcuna relazione funzionale tra ν_{tot} e θ, i risultati coincidono con quelli in cui n_{tot} è preso costante. Se invece c'è una relazione funzionale, allora le varianze che si ottengono sono più piccole quando si introduce l'informazione su n_{tot}.

6.5 Bontà del Fit con ML

Stimato il parametro mediante un fit col ML , il problema che ci poniamo è : qual è la bontà del fit? Bisogna innanzi tutto stabilire come si fa a giudicare la bontà di un fit. Noi introduciamo una statistica il cui valore rifletta il grado di qualità del fit. Questa statistica è diversa dalla statistica scelta per stimare un parametro. Si determina la p.d.f. di questa statistica. Noto il valore della statistica di test per il fit in questione , si definisce **valore P** (**P-value**) la probabilità di avere un valore della statistica di test che corrisponde ad un risultato egualmente compatibile o meno compatibile di quanto effettivamente osservato nell'esperimento. Il P-value è detto anche **livello di significanza** o **livello di confidenza** del test del fit. Questo vale in generale. Vediamo ora come operare con fit usando il ML.

- **Uso di** $\log L_{max}$ **come statistica di test**

  Come statistica del test del fit possiamo usare il valore $\log L_{max}$. Noi però non conosciamo come è fatta la p.d.f. di questa statistica. Possiamo però determinarla mediante Monte Carlo. Supponiamo che dal fit ML dei nostri dati per stimare un parametro otteniamo $\log L_{max}$ = -15246. Noi possiamo simulare un certo numero (ad esempio 500) di esperimenti uguali al nostro. Ognuno di questi esperimenti avrà lo stesso numero di eventi del nostro esperimento (reale). In ognuno di questi esperimenti determino col ML il parametro da stimare e da questo fit ottengo il valore $\log L_{max}$. In Fig. 6.1 è riportata la distribuzione dei valori di $\log L_{max}$. Se le PDFs sono parametrizzate bene, il valore $\log L_{max}$ per l'esperimento reale (indicato con la freccia rossa in Fig. 6.1 deve essere entro 1-2 σ dal massimo della distribuzione. Normalizzando la curva ad 1 ed integrando questa crva dal valore effettivamente trovato a $+\infty$ si ottiene il valore P o livello di significanza osservato.

- **Uso di una statistica basata sulla LF**

  Si fa l'istrogramma del nunero di eventi $\mathbf{n} = (n_1, n_2, \cdots, n_N)$ con n_{tot} valori misurati. Quindi usando i valori dei parametri stimati dal ML si calcolano i valori medi $\nu = (\nu_1, \nu_2, \cdots, \nu_N)$. Tutto ciò si può fare anche se il fit Ml è stato fatto unbinned. I dati $\mathbf{n}$ ed i valori stimati ν possono essere usati per costruire una statistica dei bontà del fit.

  Consideriamo come statistica della bontà fit il rapporto λ cosi definito:

$$\lambda = \frac{f_{congiunta}(\mathbf{n};\nu)}{f_{congiunta}(\mathbf{n};\mathbf{n})} = \frac{L(\mathbf{n}|\nu)}{L(\mathbf{n}|\mathbf{n})}$$

  Queste funzioni le abbiamo gia' viste per dati distribuiti sia multinomialmente che poissonianamente. In questo secondo caso per esempio si ha che:

$$\lambda_P = e^{n_{tot}-\nu_{tot}} \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

  Supponiamo che i parametri fittati siano stati m. Allora al limite di grandi campioni la statistica $(-2 \cdot logL = \chi^2)$:

$$\chi^2_P = -2\log\lambda_P = 2\sum_{i=1}^{N}\left(n_i \log\frac{n_i}{\hat{\nu_i}} + \hat{\nu_i} - n_i\right)$$

  segue una distribuzione del χ^2 con N-m gradi di libertà. $\hat{\nu_i}$ sono i valori aspettati una volta stimati gli m parametri dai dati.

  Se i dati sono distribuiti in modo multinomiale, allora si ha :

$$\lambda_M = \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

  e

$$\chi^2_M = -2\log\lambda_M = 2\sum_{i=1}^{N}\left(n_i \log\frac{n_i}{\hat{\nu_i}}\right)$$

  Al limite di grandi campioni questa è una distribuzione χ^2 con N-m-1 gradi di libertà.

  Il livello di significanza osservato (valore P) del fit si ottiene integrando la p.d.f. della distribuzione del χ^2 per N-m o N-m-1 gradi di libertà rispettivamente dal valore di χ^2 osservato all'infinito:

$$P = \int_{\chi^2}^{\infty} f(z; n_d)dz$$

dove $f(z; n_d)$ è la p.d.f. del χ^2 con n_d gradi di libertà.

Si noti che la statistica λ usata per il test di bontà del fit differisce dalla funzione di likelihood per il fattore $L(\mathbf{n}|\mathbf{n})$ che non dipende dai parametri. Di conseguenza questa statistica può essere usata sia per la stima dei parametri che per la bontà del fit.

- **Uso della statistica χ^2 (di Pearson)**

  Consideriamo l'istogramma di una variabile x. Per esempio possiamo considerare la distribuzione dei valori osservati di x (n_{tot} eventi con N bin). Siano n_i in numero di eventi nel bin i e ν_i il numero degli eventi stimato nel bin i. Supponiamo di trattare n_{tot} come variabile poissoniana. Per il test di bontà del fit possiamo usare la statistica del χ^2 , detta di Pearson, :

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

  Se in ogni bin c'è un numero minimo di eventi (tipicamente maggiore od uguale a 5) e se in ogni bin il numero di eventi è distribuito secondo una poissoniana (ν_i è il valore medio della distribuzione poissoniana di n_i nel bin i), allora la statistica di Pearson segue una distribuzione del χ^2 per N-m gradi di libertà. Si noti che questo è sempre vero indipendentemente dalla forma della distribuzione della variabile x. Per questo motivo il test del χ^2) si dice indipendente dal tipo di distribuzione ("distribution free").

  Se n_{tot} è preso come costante allora si ha :

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i \cdot n_{tot})^2}{p_i \cdot n_{tot}}$$

  dove $p_i = \frac{\nu_i}{n_{tot}}$ è la probabilità stimata per una misura di trovarsi nel bin i. Quest'ultima distribuzione segue la distribuzione del χ^2 con N-m-1 gradi di libertà. Il valore P si calcola come sopra.

  Se si ha nella distribuzione ci sono bin con meno di 5 entrate , allora le distribuzioni appena viste non seguono più la distribuzione del χ^2. In questi casi la vera p.d.f. va calcolata con uno studio Monte Carlo e poi dalla p.d.f. ottenuta di calcola il livello di significanza (valore P) della bontà del fit.

6.6 Più Esperimenti Combinati Mediante ML

Supponiamo che due esperimenti cerchino di determinare uno stesso parametro incognito θ: il primo esperimento misura n volte una quantità x mentre il secondo esperimento misura m volte una quantità y. Siano $f_1(x; \theta)$ e $f_2(y; \theta)$ le due p.d.f. delle variabili x e

y nei due esperimenti , funzioni dello stesso parametro incognito θ. La LF di tutte le osservazioni dei due esperimenti è :

$$L(\theta) = \prod_{i=1}^{n} f_1(x_i;\theta) \prod_{i=1}^{m} f_2(y_i;\theta) = L_1(\theta) \cdot L_2(\theta)$$

Se dei due esperimenti sono note le due funzioni di verosimiglianza $L_1(\theta)$ e $L_2(\theta)$, allora possiamo calcolare la LF totale e quindi costruire uno stimatore di ML del parametro θ tenendo presente le misure di tutte e due gli esperimenti.

Più frequentemente vengono riportate le stime dei parametri e non le LF. Per esempio il primo esperimento con le misure di x dà lo stimatore $\hat{\theta}_x$ del parametro θ mentre dello stesso parametro il secondo esperimento dà lo stimatore $\hat{\theta}_y$. Questi due stimatori sono variabili casuali distribuiti secondo le p.d.f. $g_1(\hat{\theta}_x;\theta)$ e $g_2(\hat{\theta}_y;\theta)$. Se i due stimatori $\hat{\theta}_x$ e $\hat{\theta}_y$ sono indipendenti, allora la funzione log-likelihood totale è data dalle somma:

$$\log L(\theta) = \log g_1(\hat{\theta}_x;\theta) + \log g_2(\hat{\theta}_y;\theta)$$

Per campioni molto grandi le due p.d.f. g_1 e g_2 diventano gaussiane e gli errori sui valori stimati dei parametri sono le due deviazioni standard : $\hat{\sigma}_{\hat{\theta}_x}$ e $\hat{\sigma}_{\hat{\theta}_y}$ In questo caso la stima combinata di θ è data dalla media pesata:

$$\hat{\theta} = \frac{\frac{\hat{\theta}_x}{\hat{\sigma}^2_{\hat{\theta}_x}} + \frac{\hat{\theta}_y}{\hat{\sigma}^2_{\hat{\theta}_y}}}{\frac{1}{\hat{\sigma}^2_{\hat{\theta}_x}} + \frac{1}{\hat{\sigma}^2_{\hat{\theta}_y}}}$$

Lo stimatore della varianza è :

$$\hat{V}[\hat{\theta}] = \frac{1}{\frac{1}{\hat{\sigma}^2_{\hat{\theta}_x}} + \frac{1}{\hat{\sigma}^2_{\hat{\theta}_y}}}$$

Caso particolare di quello studiato qui è quello in cui i due esperimenti misurano la stessa quantità (x=y). Naturalmente i due esperimenti possono essere anche lo stesso esperimento : le prime n misure e le seconde m misure sarebbero due sottoinsiemi di dati.

Misure del branching fraction dello steddo decadimento fatte con sottodecadimenti diversi vengono combinate utilizzando -ln $\mathcal{L}$. Nella Fig. 6.2 si mostra - 2 ln $\mathcal{L}$ per i due sottodecadimenti e per la loro combinazione.

6.7 Stimatori Bayesiani

Nel trattare il problema diella stima puntuale, noi siamo partiti da un campione di dati per il quale abbiamo supposto nota la p.d.f.. Questa p.d.f. è stata considerata funzione di un parametro θ che , benchè ignoto, è considerato costante. Nella statistica frequentista non si tiene conto di ogni altra informazione sia nota riguardo il parametro θ. Ad esempio cerco con uno stimatore di ML di stimare il branching ratio (BR) di un

certo decadimento. Se questo è vicino a zero (decadimento raro) allora dal fit posso avere un risultato negativo. Io mostro come valore sperimentale il valore ottenuto dal fit più o meno una sigma. Frequentisticamente significa che se facessi molti esperimenti (ognuno con un numero di eventi uguale a quello dell'esperimento effettivamente fatto) troverei nel 68 % dei casi un valore del parametro fittato compreso nell'intervallo definito prima (valore centrale più o meno 1 σ). Naturalmente qui non sto sfruttando il fatto che a priori io so che sto misurando una quantità fisica che deve essere positiva. Se imponessi questa condizione, costringendo il parametro da fittare ad assumere solo valori positivi, avrei un diverso valore del parametro fittato che però non posso più interpretare frequentisticamente. Sono passato cioè ad un atteggiamento di tipo bayesiano.

Nella statistica bayesiana il parametro θ è trattato come una variabile casuale e qundi per esso si definisce una distribuzione di probabilità nello spazio del parametro. È ovvia la generalizzazione al caso in cui i parametri da determinare siano, m $\theta = (\theta_1, \cdots, \theta_m)$.

Qui però la probabilità non è da intendersi frequentisticamente bensi come " grado di fiducia " che noi riponiamo sulla veridicità di quello che stiamo affermamdo. L'atteggiamento bayesiano è di ritenere che gli eventi non siano mai ripetibili , essi esistono come fatti singoli. Di conseguenza la frequenza con cui avviene un evento non ha nulla a che vedere col concetto di probabilità. È sul fatto singolo che la statistica bayesiana esprime la plausibilità che una data situazione si verifichi. Esempi pratici : qual è la probabilità che domani piova? Qual è la probabilità che Paola superi l'esame di Analisi Statistica dei dati ? Qual è la probabilità che l'Inter vinca questo campionato di calcio? In casi del genere sarebbe ardua una interpretazione della probabilità dal punto di vista frequentista.

Noi abbiamo una distribuzione bayesiana (o del grado di fiducia) iniziale $\pi(\theta)$ che riassume tutto ciò che conosciamo (o il grado di ignoranza) del parametro θ prima che facciamo l'esperimento. L'esperimento comporta nuova conoscenza che arrichisce quella che avevo in precedenza. La nuova conoscenza porta ad una distribuzione bayesiana finale $\pi(\theta \mid x_1, \cdots, x_n)$ che si puo ottenere tramite il teorema di Bayes:

$$\pi(\theta \mid x_1, \cdots, x_n) = \frac{\pi(\theta) \cdot L(x_1, \cdots, x_n \mid \theta)}{\int_\Theta \pi(\theta) \cdot L(x_1, \cdots, x_n \mid \theta) d\theta}$$

Per un determinato campione di misure il denominatore è costante e quindi:

$$\pi(\theta \mid x_1, \cdots, x_n) \sim \pi(\theta) \cdot L(x_1, \cdots, x_n \mid \theta)$$

La LF può essere interpretata come una distribuzione (bayesiana) finale con una distribuzione (bayesiana) iniziale costante.

Dalla distribuzione finale è possibile stimare il parametro θ prendendo il valore di aspettazione:

$$\hat{\theta} = \int_\Theta \theta \cdot \pi(\theta \mid x_1, \cdots, x_n) d\theta$$

$\hat{\theta}$ è detto **stimatore bayesiano**. La sua varianza è uguale alla varianza della distribuzione finale.

Con una distribuzione iniziale costante, la differenza tra lo stimatore ML e quello bayesiano consiste nel fatto che il primo sceglie il massimo della distribuzione finale

mentre il secondo sceglie il valore medio. In questo senso lo stimatore bayesiano rispetto a quello di ML tiene in conto anche della forma della distribuzione finale.

È possibile anche stimare θ cercando il massimo della distribuzione finale :

$$\pi(\hat{\theta} \mid x_1, \cdots, x_n) \geq \pi(\theta \mid x_1, \cdots, x_n)$$

per tutti i valori possibili di θ. Se la distribuzione iniziale è presa costante, questo stimatore si riduce all'ordinario stimatore di ML (però l'interpretazione del risultato è diversa nei due casi).

Esempio:

Sia una variabile casuale X distribuita uniformemente sull'intervallo (θ, 10.5). Assumendo che la distribuzione iniziale di θ sia data da :

$$\pi(\theta) = \begin{cases} 5 & \text{per } 9.5 < \theta < 9.7 \\ 0 & \text{altrimenti} \end{cases} \tag{6.1}$$

calcolare la distribuzione finale di θ, noto che una misura di X è 10.

Noi abbiamo che :

$$\pi(\theta|x = 10) = \frac{\pi(x = 10|\theta) \cdot \pi(\theta)}{\int_{9.5}^{9.7} \pi(x = 10|\theta)\pi(\theta)d\theta}$$

$$= \frac{\frac{5}{10.5-\theta}}{\int_{9.5}^{9.7} \frac{5}{10.5-\theta} d\theta}$$

$$= -\frac{1}{(\log 0.8)(10.5 - \theta)} \cong \frac{4.48}{10.5 - \theta} \qquad \text{per } 9.5 < \theta < 9.7.$$

6.8 Postulato di Bayes e Scelta della Distribuzione Iniziale

L'inferenza statistica bayesiana si basa sulla distribuzione finale e richiede che venga assegnata la distribuzione iniziale. Questa distribuzione iniziale riflette generalmente il giudizio dello sperimentatore e quindi non ha quel carattere oggettivo che invece la statistica frequentista vuol assegnare alla probabilità. Per questo motivo la probabilità bayesiana è detta anche **probabilità soggettiva**. Essa, come detto, rappresenta il grado di fiducia che un individuo assegna al verificarsi di un dato evento (incerto).

Il punto debole della statistica bayesiana sta nella scelta della distribuzione iniziale. Secondo i fautori della statistica frequentista il carattere soggettivo della probabilità rende l'inferenza statistica bayesiana non adatta per risultati che devono avere invece un carattere oggettivo. Effettivamente la scelta della distribuzione iniziale è in taluni casi ovvia, in molti casi però e' difficile da definire e non univoca ,in altri casi non si sa come assegnarla. Teniamo presente comunque che per grandi campioni la distribuzione finale è largamente dominata dalla funzione di verosimiglianza, cosicchè la scelta della distribuzione iniziale è meno importante.

Una prima prescrizione per la distribuzione iniziale venne data dallo stesso Bayes col seguente postulato (**detto postulato di Bayes o principio di indifferenza)** :

In assenza di ogni tipo di conoscenza tutte le distribuzioni iniziali devono essere uguali (uniformi).

Questo postulato sembra apparentemente innoquo ma si verifica subito che se la distribuzione di un parametro θ è uniforme, non lo è in generale quella di una funzione f(θ).

Questo postulato all'inizio veniva applicato in modo acritico e generalizzato col risultato che l'intera impostazione bayesiana finì col cadere in discredito (nonostante il prestigio di Bernoulli, Laplace , ecc)

Nella prima parte del XX secolo si è sviluppata la statistica frequentista ad opera di Fisher, Pearson, Neyman ed altri.

In tempi recenti si è avuta una rinascita della teoria bayesiana (**statistica neo-bayesiana)** . La definizione ed i criteri per una comune scelta della distribuzione iniziale ha portato allo sviluppo della **teoria bayesiana oggettiva** .

Comunque la separazione tra statistica frequentista e quella bayesiana resta netta. Sono stati inutili sinora i tentativi di arrivare ad una sintesi delle due. Noi a seconda dei casi utilizzeremo o l'una o l'altra.

Noi in generale applichiamo la statistica frequentista. In taluni casi utilizzeremo quella bayesiana , specificando chiaro che la significanza statistica, l'intervallo di confidenza, ecc sono calcolati in modo bayesiano. Quando non specificato , si sottintende l'uso della statistica frequentista.

chapterMetodo dei Minimi Quadrati

- Stimatori di Minimi Quadrati
- Massima Verosimiglianza e Minimi Quadrati
- LS Fit di Dati Istogrammati
- Test di Bontà del Fit col χ^2
- Combinazione di più Esperimenti con LS

Figure 6.1: Goodness-of-fit con 500 "toy experiments" : La freccia rossa indica il valore di -log L_{max} ottenuto nel ML fit dei dati sperimentali.

Figure 6.2: Combinazione di due misure di rapporto di decadimento del decadimento $B \to \eta K^0$: linea punteggiata (rosa) $\eta_{\gamma\gamma} K^0_S$, linea a tratti blu : $\eta_{3\pi} K^0_S$

6.9 Stimatori di Minimi Quadrati

Consideriamo due variabili casuali x e y . Immaginiamo di misurare y dopo aver misurato x. Per esempio a determinati istanti di tempo misuro la posizione di un corpo che si muove di moto rettilineo uniforme. Quindi negli istanti $x_1, \cdots, x_n$ misuro le posizioni $y_1, \cdots, y_n$. Queste misure hanno deviazione standard σ_i , in generale diverse per ogni misura. Supponiamo di conoscere la forma della funzione $\lambda(x;\theta)$ che per ogni x_i mi permette di determinare il corrispondente valore di y_i. La funzione λ comunque contiene un parametro (o piu' parametri) che io devo determinare. Nell'esempio fatto la relazione che lega le due variabili e'di tipo lineare per cui il problema è di determinare i parametri di questa retta.

Indichiamo con $\lambda(x_i;\theta)$ e y_i il valore stimato e quello misurato per la coordinata x_i rispettivamente. Le differenze tra valore misurato e quello predetto si chiama residuo. Ogni residuo lo pesiamo con l'inverso della deviazione standard corrispondente alla misura considerata e sommiamo i quadrati di questi residui pesati per tutte le n misure. Questa somma è detta χ^2 :

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - \lambda(x_i;\theta)}{\sigma_i} \right]^2$$

Il parametro incognito cercato è quello che minimizza il χ^2 (da cui il nome del metodo).

Supponiamo che la relazione funzionale tra x e y sia lineare : y=mx . Allora **lo stimatore di minimi quadrati (LS)** di m (denotato $\hat{m}$) si trova dalla minimizzazione del χ^2:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - mx_i}{\sigma_i} \right]^2$$

Per determinare il minimo di questa funzione deriviamo rispetto ad m ed uguagliamo a zero questa derivata parziale:

$$\frac{\partial \chi^2}{\partial m} = -2 \sum_{i=1}^{n} x_i \frac{y_i - mx_i}{\sigma_i^2} = 0$$

Se le varianze delle misure sono tutte uguali , possiamo scrivere:

$$-\frac{2}{\sigma^2} \sum_{i=1}^{n} (x_i y_i - m x_i^2) = 0$$

Trasformiamo le sommatorie in valori medi, si ottiene per lo stimatore $\hat{m}$:

$$\hat{m} = \frac{\overline{xy}}{\overline{x^2}}$$

Questo risultato può essere scritto anche cosi:

$$\hat{m} = \sum_{i=1}^{n} \frac{x_i}{n\overline{x^2}} y_i$$

e da questa propagando gli errori da ogni y_i ad m otteniamo:

$$V[\hat{m}] = \sum_{i=1}^{n} \left(\frac{x_i}{n\overline{x^2}} \right)^2 \sigma^2 = \frac{\sigma^2}{n\overline{x^2}}$$

Questo risultato si generalizza agli stimatori di pendenza e intercetta all' origine quando la retta da fittare e' del tipo:

$$y = ax + b$$

6.10 Massima Verosimiglianza e Minimi Quadrati

Supponiamo che le distribuzioni dei valori y misurati abbiano distribuzioni gaussiane e che le n misure y_i fatte siano indipendenti. Sia λ il vettore dei valori stimati $(\lambda_1, \lambda_2, \cdots, \lambda_N)$ con $\lambda_i = \lambda(x_i; \theta)$.

Allora fissato x_i, la p.d.f. di y_i è :

$$g(y_i; \lambda_i, \sigma_i^2) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{[y_i - \lambda(x_i;\theta)]^2}{2\sigma_i^2}}$$

Per le N misure il logaritmo della LF è :

$$logL = -\frac{1}{2} \sum_{i=1}^{n} \frac{[y_i - \lambda(x_i;\theta)]^2}{\sigma_i^2} - \sum_{i=1}^{n} log\sigma_i \sqrt{2\pi}$$

Per massimizzare la LF , bisogna minimizzare la quantità :

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{[y_i - \lambda(x_i;\theta)]^2}{\sigma_i^2}$$

Questa come si vede è il procedimento degli stimatori di minimi quadrati.

Come si vede da queste relazioni, a meno di termini che non contengono i parametri, si ha:

$$logL = -\chi^2/2$$

Se le n misure non sono indipendenti ma descritte da una p.d.f. di tipo gaussiamo multidimensionale con matrice di covarianza V nota e valori medi non noti, allora la funzione di log-likelihood é ottenuta usando la p.d.f. congiunta :

$$logL = -\frac{1}{2} \sum_{i,j=1}^{n} (y_i - \lambda(x_i;\theta)) V_{ij}^{-1} (y_j - \lambda(x_j;\theta))$$

dove i termini non contenenti i parametri θ sono stati eliminati.

Questa funzione si massimizza , minimizzando la quantitá:

$$\chi^2(\theta) = \sum_{i,j=1}^{n} (y_i - \lambda(x_i;\theta))(V^{-1})_{ij}(y_j - \lambda(x_j;\theta))$$

I parametri θ che mimimizzano il χ^2 sono detti stimatori di minimi quadrati (LS).

In contrasto con gli stimatori ML , gli stimatori LS non hanno proprietà generali ottimali ad eccezione di un caso particolare e ciò quando la la relazione funzionale e' di tipo lineare. In questo caso lo stimatore LS è non distorto ed ha la minima varianza tra tutti gli stimatori che sono funzioni lineari delle variabili.

Questo procedimento di fit viene usato anche quando le singole misure y_i non sono gaussiane. La quantità da minimizzare è detta χ^2 perchè sotto determinate condizioni ha una p.d.f χ^2. Mantiene questo nome anche nei casi più generali in cui questo non è vero. Quanto detto si generalizza al caso di piú parametri da stimare. Gli stimatori LS sono probabilmente quelli più comunemente usati per la loro semplicità.

6.11 Fit Lineari

Supponiamo che $\lambda(x;\theta)$ sia una funzione lineare dei parametri $\theta = (\theta_1, \theta_2, \cdots, \theta_m)$

$$\lambda(x;\theta) = \sum_{j=1}^{m} a_j(x)\theta_j$$

dove le funzioni $a_j(x)$ sono funzioni qualsiasi in x linearmente indipendenti tra di loro. In questo caso stimatori e loro varianza si possono trovare analiticamente. (Spesso è più semplice e comodo trovare la soluzione per via numerica, minimizzando il χ^2).

Scriviamo la funzione $\lambda(x_i;\theta)$ in x_i cosi :

$$\lambda(x_i;\theta) = \sum_{j=1}^{m} a_j(x_i)\theta_j = \sum_{j=1}^{m} A_{ij}\theta_j$$

Di conseguenza in notazione matriciale il χ^2 puó essere scritto cosi:

$$\chi^2 = (\mathbf{y} - \lambda)^T V^{-1}(\mathbf{y} - \lambda)$$

$$= (\mathbf{y} - A\theta)^T V^{-1}(\mathbf{y} - A\theta)$$

I vettori delle misure ed i vettori dei valori predetti sono vettori colonna.

Per minimizzare il χ^2 si mettono uguali a zero le sue derivate rispetto ai parametri:

$$-2(A^T V^{-1}\mathbf{y} - A^T V^{-1} A\theta) = 0$$

Se la matrice $A^T V^{-1} A$ non é singolare, allora si ha :

$$\hat{\theta} = (A^T V^{-1} A)^{-1} A^T V^{-1}\mathbf{y}$$

Questi sono i valori dei parametri stimati.

Propagando gli errori si puó deteminare la matrice di covarianza degli stimatori U_{ij}

$$U = (A^T V^{-1} A)^{-1}$$

L'inverso della matrice di covarianza é :

$$(U_{ij})^{-1} = \frac{1}{2}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]_{\theta=\hat{\theta}}$$

Se le misure y_i hanno distribuzione gaussiana allora log L = -$\chi^2/2$ e la formula vista prima coincide con il limite di Cramer-Rao.

É possibile far vedere che (con λ lineare nei parametri) il χ^2 é quadratico in θ:

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + \frac{1}{2}\sum_{i,j}^{m}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]_{\theta=\hat{\theta}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

La linea di livello del $\chi 2$ corrispondente al χ^2_{min} +1 ha tangenti nei punti $\hat{\theta}_i \pm \hat{\sigma}_i$, corrispondenti ad una deviazione standard dalle stime LS.

$$\chi^2(\theta) = \chi^{\mathbf{2}}(\hat{\theta}) + \mathbf{1} = \chi^{\mathbf{2}}_{\mathbf{min}} + \mathbf{1}$$

Confrontare quetso risultato con quello ottenuto con ML. In particolare quando i parametri sono due, la linea di contorno è una ellisse e dalle tangenti si ottengono le deviazioni standard.

Se la funzione λ non é lineare nei parametri , allora la linea di livello non é piú ellittica.

6.12 LS Fit di Dati Istogrammati

Supponiamo che i nostri n eventi siano riportati in un istogramma. Sia N il numero di bin (numerati da 1 a N). Il generico bin i-simo ha centro in x_i e contiene y_i eventi. Generalmente la larghezza è la stessa per tutti i bin . Sia $\lambda(x;\theta)$ la p.d.f. ipotizzata con θ il parametro (i parametri) da stimare. Il numero di eventi (entries) previsti nel bin i-simo è dato dal valore di aspettazione di y_i, quindi :

$$\lambda_i(\theta) = n\int_{x_i{}^{min}}^{x_i{}^{max}} \lambda(x;\theta)dx = np_i(\theta)$$

dove $p_i(\theta)$ è la probabilità che l' evento appartenga al bin i-simo. La funzione $\lambda_i(\theta)$ viene fittata sui dati sperimentali per stimare il valore di θ. Quindi si minimizza il χ^2 :

$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda_i(\theta))^2}{\sigma_i^2}$$

Se y_i è piccolo in confronto ad n , allora y_i può essere considerata una variabile poissoniana. La varianza di y_i è uguale al valore medio e quindi si può scrivere:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{\lambda_i(\theta)} = \sum_{i=1}^{N} \frac{(y_i - np_i(\theta))^2}{np_i(\theta)}$$

È possibile usare come varianza il numero di entrate nel bin i-simo (**metodo dei minimi quadrati modificato (MLS)**). In questo caso la funzione che si minimizza è la seguente :

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{y_i} = \sum_{i=1}^{N} \frac{(y_i \ - np_i(\theta))^2}{y_i}$$

6.13 Test di Bontà del Fit col χ^2

Con valori distribuiti secondo una gaussiana, lo stimatore LS coincide con lo stimatore di ML. Inoltre il valore del χ^2 può essere utilizzato per stimare la qualità del fit. Infatti se le misure sono distribuite in modo gaussiano, le ipotesi $\lambda(x;\theta)$ sono lineari nei parametri θ_i e la dipendenza funzionale dell'ipotesi λ è corretta, allora il valore minimo del χ^2 segue la distribuzione del χ^2 con $N--m$ gradi di libertà (N numero bin, m numero parametri fittati).

Come valore P si prende la probabilità che l'ipotesi fatta abbia un χ^2 peggiore (più grande) di quello χ^2_0 effettivamente osservato:

$$P = \int_{\chi^2_0}^{\infty} f(z; n_d) dz$$

con $n_d = N - m$ numero di gradi di libertà.

Ricordiamo che il valore di aspettazione di una variabile casuale distribuita secondo la distribuzione del χ^2 è uguale al numero di gradi di libertà. Quindi il valore del χ^2 diviso per n_d viene considerato spesso come strima di qualità del fit. Se il rapporto è circa 1 , allora le cose vanno nel giusto verso. Se il rapporto è diverso da uno allora qualcosa non va (tipicamente gli errori sono sottostimati o sovrastimati).

6.14 Combinazione di più Esperimenti con LS

Supponiamo che la variabile casuale Y sia stata misurata da N esperimenti, ognuno dei quali ottiene un valore y_i con deviazione standard σ_i. Sia λ il valore vero aspettato (uguale per tutti gli esperimenti). Allora si ha :

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2}$$

qui λ ha il ruolo di θ.

Derivando rispetto a λ, ponendo uguale a zero e risolvendo per l'λ, si ottiene:

$$\lambda = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2}$$

che è la ben nota formula della media pesata. Dalla derivata seconda del χ^2 si ottiene la varianza di λ :

$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

Questa procedura può facilmente essere generalizzata al caso di misure non indipendenti, tenendo conto della matrice di covarianza.

Chapter 7

Stima di Parametri per Intervalli

- Intervalli di Confidenza
- Intervalli di Confidenza per Stimatori a Distribuzione Gaussiana
- Intervalli di Confidenza per Stimatori a Distribuzione Poissoniana
- Intervalli di Confidenza con funzioni di ML o χ^2
- Limiti sulla Media di una Variabile Poissoniana
- Dati Gaussiani vicini ad un Limite Fisico
- Dati Poissoniani con Piccoli Campioni di dati

7.1 Introduzione agli Intervalli di Confidenza

La stima puntuale di un parametro (con la stima della varianza ecc) in talune situazioni non è adeguata. Talvolta bisogna tener conto ad esempio della natura e caratteristica delle code della distribuzione. Cè quindi la possibilità di utilizzare una descrizione statistica di tipo diverso basata sul concetto di stima per intervalli. In una serie di esperimenti ripetuti, si dà la frazione di volte che in un certo intervallo cada il valore vero del parametro. Come per la stima puntuale anche in questo caso bisogna trovare degli stimatori per intervallo, stimatori per intervalli buoni o ottimi , ecc.

Cominciamo con un semplice esempio . Si abbiamo n misure estratte estratte da una popolazione di misure distribuite in modo normale con una media μ non conosciuta e varianza nota uguale a σ^2. La stima di ML per μ è data da :

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

La quantità

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

segue una distribuzione normale standardizzata (cioè con media nulla e varianza 1). Quindi la p.d.f. della variabile z è data da :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Ci possiamo chiedere quanto vale la probabilità che z si trovi tra due valori scelti arbitrariamente. Per esempio

$$P[-1.96 < z < 1.96] = \int_{-1.96}^{+1.96} \phi(z)dz = 0.95$$

Cioè la probabilità che z sia compresa tra -1.96 e 1.96 è 0.95 (95 %). La diseguaglianza -1.96 < z implica che :

$$\mu < \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Analogamente la diseguaglianza z< 1.96 implica che :

$$\mu > \overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Possiamo perciò scrivere che :

$$P[\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}] = 0.95$$

Interpreto questa relazione , dicendo che se estraessi campioni di n misure dalla popolazione distribuita in modo normale con media non nota e varianza σ^2 nota, allora la probabilità che l'intervallo $(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$ contenga la vera media incognita μ è

0.95. La probabilità scelta 0.95 è detta livello confidenza. Quindi l'intervallo considerato $(\overline{x}-1.96\cdot\frac{\sigma}{\sqrt{n}}, \overline{x}+1.96\cdot\frac{\sigma}{\sqrt{n}}$) è chiamato intervallo di confidenza ad un livello di confidenza di 0.95 (del 95 %)

Si possono ottenete analogamente intervalli di confidenza con ogni livello di confidenza tra 0 e 1 (tipici valori sono 0.90, 0.95 e 0.99). Si possono ottenere intervalli di confidenza per la media, per la varianza e per la media e per la varianza insieme. Noi ci occuperemo esclusivamente degli intervalli di confidenza per la media.

7.2 Intervalli di Confidenza

Supponiamo di avere n misure di una variabile casuale x $(x_1, x_2, \cdots, x_n)$ e di utilizzare queste misure per stimare un parametro θ. Sia $\hat{\theta}_{oss}$ il valore ottenuto dallo stimatore $\hat{\theta}$. Supponiamo di conoscere la p.d.f. dello stimatore $\hat{\theta}$, $g(\hat{\theta};\theta)$,dove il valore vero di θ è preso come parametro. La p.d.f. può essere calcolabile o in forma analitica o mediante Monte Carlo. Si tenga presente che il valore vero di θ non è noto ma per ogni fissato θ so calcolarmi la p.d.f. $g(\hat{\theta};\theta)$ dello stimatore $\hat{\theta}$.

Nota la p.d.f. $g(\hat{\theta};\theta)$ so calcolare il valore u_α in modo che sia α la probabilità di osservare un valore di $\hat{\theta} \geq u_\alpha$

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta) \tag{7.1}$$

G è la distribuzione cumulativa della p.d.f. $g(\hat{\theta};\theta)$.

In modo analogo possiamo determinare il valore ν_β tale che sia β la probabilità di osservare $\hat{\theta} \leq \nu_\beta$:

$$\beta = P(\hat{\theta} \leq \nu_\beta(\theta)) = \int_{-\infty}^{\nu_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta} = G(\nu_\beta(\theta);\theta) \tag{7.2}$$

Al variare di θ le funzioni u_α e ν_β rappresentano due curve. La regione tra queste due curve è detta fascia di confidenza (" confidence belt"). Per qualunque valore di θ (e quindi anche per il valore $\hat{\theta}$ dato dallo stimatore) per costruzione si ha che :

$$P(\nu_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta) = 1 - \alpha - \beta$$

Se $\hat{\theta}$ è uno stimatore buono di θ , allora le due funzioni u_α e ν_α sono monotone crescenti . Invertendole si ha :

$$a(\hat{\theta}) \equiv u_\alpha^{-1}(\hat{\theta})$$

$$b(\hat{\theta}) \equiv \nu_\beta^{-1}(\hat{\theta})$$

La disequaglianza $\hat{\theta} \geq u_\alpha(\theta)$ implica $a(\hat{\theta}) \geq \theta$. Analogamente la disequaglianza $\hat{\theta} \leq \nu_\beta(\theta)$ implica $b(\hat{\theta}) \leq \theta$. Si ha di conseguenza che :

$$P(a(\hat{\theta}) \geq \theta) = \alpha$$

$$P(b(\hat{\theta}) \leq \theta) = \beta$$

Mettendole assieme si ha che :

$$P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta$$

Determiniamo i valori di $a(\hat{\theta})$ e $b(\hat{\theta})$ in corrispondenza al valore effettivamente ottenuto dallo stimatore $\hat{\theta}_{oss}$ e siano a e b questi valori (vedi figura). L'intervallo [a,b] è detto **intervallo di confidenza** ad un **livello di confidenza** di $1 - \alpha - \beta$. Ripetendo più volte l'esperimento posso affermare che in una frazione $1 - \alpha - \beta$ di esperimenti il valore vero del parametro θ sarà contenuto nell'intervallo [a,b].

L'intervallo di confidenza non è univocamentre determinato dal livello di confidenza. Talvolta si prende $\alpha = \beta = \gamma/2$. In questo caso si parla di **intervallo di confidenza centrale** con livello di confidenza $1 - \gamma$. Si noti che nell'intervallo di confidenza centrale α e β sono uguali ma non necessariamente a e b sono equidistanti dal valore stimato.

Talvolta si è interessati ad **intervalli di confidenza da un lato solo**. Per esempio si vuole determinare un limite inferiore sul parametro θ cosicchè si abbia $a < \theta$ con una probabilità $1 - \alpha$. Analogamente si può introdurre un limite superiore b del parametro θ cosicchè la probabilità che $b > \theta$ sia $1 - \beta$.

Per costruzione si ha che :

$$\alpha = \int_{\hat{\theta}_{oss}}^{\infty} g(\hat{\theta}; a) d\hat{\theta} = 1 - G(\hat{\theta}_{oss}; a) \tag{7.3}$$

$$\beta = \int_{-\infty}^{\hat{\theta}_{oss}} g(\hat{\theta}; b) d\hat{\theta} = G(\hat{\theta}_{oss}; b) \tag{7.4}$$

In definitiva bisogna risolvere numericamente queste equazioni per determinare a e b.

L'intervallo di confidenza [a,b] viene espresso talvolta riportando una misura sperimentale come $\hat{\theta}^{+d}_{-c}$ [c,d] , dove c= $\hat{\theta}$-a e d= b- $\hat{\theta}$. Quando si utilizzano le barre di errore per indicare un intervallo di confidenza, convenzionalmente ci si riferisce ad un intervallo centrale di confidenza con un livello di confidenza $1 - \gamma = 0.683$.

Uno può mostrare il valore stimato del parametro e l'intervallo di confidenza oppure solo l'intervallo di confidenza, a seconda che del grado di significanza statistica che ha il valore stimato del parametro.

7.3 Intervalli di Confidenza per Stimatori a Distribuzione Gaussiana

Un caso molto importante è quello in cui la p.d.f. dello stimatore $\hat{\theta}$ è una curva gaussiana con valore medio θ e deviazione standard $\sigma_{\hat{\theta}}$. Bisogna trattare separatamente i due casi

in cui la varianza sia nota oppure no. Supponiamo di considerare il primo caso e cioè che la deviazione standard $\sigma_{\hat{\theta}}$ sia nota. Allora se il valore del parametro stimato è $\hat{\theta}_{oss}$, l'intervallo di confidenza si ottiene risolvendo le due equazioni:

$$\alpha = 1 - G(\hat{\theta}_{oss}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{oss} - a}{\sigma_{\hat{\theta}}}\right)$$

$$\beta = G(\hat{\theta}_{oss}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{oss} - b}{\sigma_{\hat{\theta}}}\right)$$

dove G è espressa mediante la distribuzione cumulativa della gaussiana standardizzata Φ. Da queste equazione si ha anche che :

$$a = \hat{\theta}_{oss} - \sigma_{\hat{\theta}}\Phi^{-1}(1-\alpha)$$

$$b = \hat{\theta}_{oss} + \sigma_{\hat{\theta}}\Phi^{-1}(1-\beta)$$

Φ^{-1} è la funzione inversa di Φ , cioè il quantile della gaussiana standardizzata. I quantili $\Phi^{-1}(1-\alpha)$ e $\Phi^{-1}(1-\beta)$ ci dicono quanto a e b distano dal valore stimato $\hat{\theta}$ in unità di deviazione standard $\sigma_{\hat{\theta}}$. Generalmente questi quantili sono presi pari a piccoli numeri interi (1, 2 ,3 ...). Cioè si cercano limiti dell'intervallo di confidenza che distano dal valore stimato di una o due o tre deviazioni standard. Nella tavola Table 7.1 sono riportati i livelli di confidenza corrispondenti a valori interi di alcuni quantili sia per intervalli centrali (a sinistra) che per intervalli da una sola parte.

$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

Table 7.1: Livello di confidenza per valori diversi del quantile della gaussiana standardizzata per intervalli centrali e da una sola parte

Supponiamo di considerare un intervallo di confidenza centrale con $\alpha = \beta = \gamma/2$. Il quantile uguale ad 1 corrisponde ad un livello di confidenza $1-\gamma$ di 0.6827 (68.3 %) . In questo caso la barra di errore è di una σ :

$$[a, b] = \left[\hat{\theta}_{oss} - \sigma_{\hat{\theta}}, \hat{\theta}_{oss} + \sigma_{\hat{\theta}}\right]$$

e la misura finale è riportata come :

$$\hat{\theta}_{oss} \pm \sigma_{\hat{\theta}}$$

Per un intervallo di confidenza da un solo lato (a destra) per fare un taglio ad una sigma si prende uguale ad uno il quantile $\Phi^{-1}(1-\alpha)$. In questo caso il livello di confidenza $1-\alpha$ è pari a 0.8413 (84.1 %).

Se si vogliono livelli di confidenza di 0.90, 0.95 e 0.99, i quantili dell'intervallo di confidenza centrale devono essere presi uguali a 1.645,1.960 e 2.576 rispettivamente. Per gli stessi di livelli di confidenza in intervalli da un solo lato (a destra), bisogna prendere quantili pari a 1.282, 1.645 e 2.326 rispettivamente. Nella tavola Table 7.2 sono riportati i valori di alcuni quantili corrispondenti per determinati valori del livello di confidenza sia per intervalli centrali (a sinistra) che per intervalli da una sola parte.

$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

Table 7.2: Quantile della gaussiana standardizzata per determinati valori del livello di confidenza per intervalli centrali e da una sola parte

Se la deviazione standard della distribuzione non è nota , allora la situazione si complica ; infatti in questo caso $\hat{\sigma}_{\hat{\theta}}$ dipende in generale da $\hat{\theta}$ e la funzione cumulativa $G(\hat{\theta};\theta,\sigma\theta)$ (dove al posto della deviazione standard non nota si mette il valore stimato) non ha un legame semplice con la distribuzione cumulativa della gaussiana standardizzata. Nella pratica se si tratta di un campione molto grande in modo che la deviazione standard stimata sia una buona approssimazione di quella vera, allora quanto abbiamo detto sinora si applica sostituendo alla deviazione standard vera e non nota quella stimata (varianza del campione).

Se il campione delle misure è piccolo , consideriamo le seguenti due variabili Z e U :

$$Z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$$

ha una distribuzione gaussiana $N(\mu,\sigma^2)$, anche se noi non conosciamo il valore di σ

$$U = \frac{(n-1)s^2}{\sigma^2}$$

dove s^2 la varianza del campione, cioè :

$$s^2 = \frac{1}{(n-1)}\sum_i (x_i-\overline{x})^2$$

La variabile U ha una distribuzione del χ^2 con n-1 gradi di libertà. Le due variabili Z ed U son tra di loro indipendenti. Di conseguenza la variabile:

$$t = \frac{z}{\sqrt{u/(n-1)}} = \frac{\overline{x}-\mu)/\sigma/\sqrt{n}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} = \frac{\overline{x}-\mu}{s/\sqrt{n}}$$

segue una distribuzione t di Student con n-1 gradi di libertà . La distribuziont t è simmetrica rispetto al valore t=0. Per questo motivo di solito l'intervallo centrale è preso simmetrico :

$$P(-b \leq \frac{\overline{x}-\mu}{s/\sqrt{n}} \leq +b) = \int_{-b}^{+b} f(t; n-1)dt = 1-\gamma$$

essendo $1-\gamma$ il livello di confidenza scelto. Questa relazione può essere riscritta così :

$$P(\overline{x} - b\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + b\frac{s}{\sqrt{n}}) = 1-\gamma$$

Fissato il livello di confidenza 1- γ, b viene determinato dalle tavole della distribuzione di Student per n-1 gradi di libertà.

Esempio: Un test di 25 professori mostra un valore per il quoziente di intelligenza QI di 128 con una deviazione standard di 15. Quali sono i limiti dell'intervallo di confidenza con il livello di confidenza di 0.95 (95 %) sul valore vero del valore medio QI di tutti i professori?

Con professori scelti a caso , il valore della stima dell'errore sulla media è $15/\sqrt{25} = 3$. Se usassimo una distribuzione gaussiana, noi avremmo $\pm 1.96\sigma$, ottenendo i limiti [122.1 , 133.9]. Se invece usiamo la distribuzione t di Student, allora il valore (critico) t per 24 gradi di libertà ad un livello di confidenza di 95 % è 2.06. In questo caso i limiti dell'intervallo al 95 % di livello di confidenza sono [121.8, 134.2].

7.4 Intervalli di Confidenza per Stimatori a Distribuzione Poissoniana

Nel caso di distribuzione poissoniana (e in generale discreta) gli integrali nelle equazioni 7.3 e 7.4 vanno sostituiti con sommatorie. Inoltre se ν il valore di aspettazione di n, la probabilità di osservare n è data da :

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu}$$

Poichè la variabile da stimare assume solo valori interi, allora non è sempre possibile trovare i valori a e b dalle equazioni 7.3 e 7.4 . Noi possiamo ancora utilizzare queste equazioni richiedendo che la α sia uguale alla probabilitá che lo stimatore dia un valore uguale o maggiore a quello effettivamente trovato. Analogamente per la probabilitá β che deve essere uguale o minore a quello effettivamente trovato. In questo modo sovrastimiamo l'intervallo di confidenza.

Supponiamo di aver misurato di una variabile poissoniana n (n=0,1,2...) un valore n_{oss}. Vogliamo definire un intervallo di confidenza per il valore medio ν. Dalle equazioni 7.3 e 7.4 e tenendo conto di quanto appena detto, si ha che :

$$\alpha = P(\hat{\nu} \geq \hat{\nu}_{oss}; a)$$

$$\beta = P(\hat{\nu} \leq \hat{\nu}_{oss}; b)$$

che nel nostro caso poissoniano diventano:

$$\alpha = \sum_{n=n_{oss}}^{\infty} f(n;a) = 1 - \sum_{0}^{n_{oss}-1} f(n;a) = 1 - \sum_{0}^{n_{oss}-1} \frac{a^n}{n!} e^{-a}$$

e

$$\beta = \sum_{0}^{n_{oss}} f(n;b) = \sum_{0}^{n_{oss}} \frac{b^n}{n!} e^{-b}$$

Per un valore osservato n_{oss} e per date probabilità α e β, le equazioni precedenti possono essere risolte numericamnete ottenendo a e b.

Le sommatorie appena viste si calcolano facilmente utilizzando la relazione esistente tra la distribuzione poissoniana e quella χ^2

$$\sum_{0}^{n_{oss}} \frac{\nu^n}{n!} e^{-\nu} = \int_{2\nu}^{\infty} f_{\chi^2}(z; n_d = 2(n_{oss}+1)) dz$$

$$= 1 - F_{\chi^2}(2\nu; n_d = 2(n_{oss}+1))$$

dove f_{χ^2} è la p.d.f. del χ^2 per n_d gradi di libertà. F_{χ^2} è la c.d.f. della f_{χ^2}. Utilizzando questa relazione si ha che :

$$a = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; nd = 2n_{oss})$$

e

$$b = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; nd = 2(n_{oss}+1))$$

I quantili $F^{-1}_{\chi^2}$ possono essere ottenuti consultando le Tavole statistiche.

Si noti che se si osservano 0 eventi, il limite inferiore non può essere determinato. In questo caso è particolarmente interessante il calcolo del limite superiore b .

$$\beta = \sum_{n=0}^{0} \frac{b^n e^{-b}}{n!}$$

Da questa si ottiene $b = -\log\beta$. Al 95 % di livello di confidenza , si ha $b = -\log(0.05) = 2.996 \sim 3$ Nella tabella Table 7.3 sono riportati i limiti superiore ed inferiore poissoniani per determinati valori n di eventi osservati.

7.5 Intervalli di Confidenza con Funzioni di ML o χ^2

La funzione L di Ml (o i l χ^2 con $L = e^{-\chi^2/2}$) permette un calcolo semplice ed approssimato dell'intervallo di confidenza.

	Inferiore			Superiore		
	90%	95%	99%	90%	95%	99%
$n = 0$	–	–	–	2.30	3.00	4.61
$n = 1$	0.11	0.05	0.01	3.89	4.74	6.64
$n = 2$	0.53	0.36	0.15	5.32	6.30	8.41
$n = 3$	1.10	0.82	0.44	6.68	7.75	10.05
$n = 4$	1.74	1.37	0.82	7.99	9.15	11.60
$n = 5$	2.43	1.97	1.28	9.27	10.51	13.11
$n = 6$	3.15	2.61	1.79	10.53	11.84	14.57
$n = 7$	3.89	3.29	2.33	11.77	13.15	16.00
$n = 8$	4.66	3.98	2.91	12.99	14.43	17.40
$n = 9$	5.43	4.70	3.51	14.21	15.71	18.78
$n = 10$	6.22	5.43	4.13	15.41	16.96	20.14

Table 7.3: Alcuni limiti poissoniani.

Consideriamo per ora lo stimatore di ML $\hat{\theta}$ per un parametro θ in un campione molto grande di misure. Si può dimostrare che in questo limite la p.d.f. $g(\hat{\theta};\theta)$ diviene gaussiana :

$$g(\hat{\theta};\theta) = \frac{1}{\sqrt{2\pi\sigma^2_{\hat{\theta}}}} e^{-(\frac{(\hat{\theta}-\theta)^2}{2\sigma^2_{\hat{\theta}}})}$$

Si può dimostrare che sempre nelle stesse condizioni limite anche la funzione di ML L diviene essa stessa gaussiana centrata attorno al valore $\hat{\theta}$ stimato dal ML :

$$L(\theta) = L_{max} e^{-(\frac{(\hat{\theta}-\theta)^2}{2\sigma^2_{\hat{\theta}}})}$$

Si noti che $\sigma_{\hat{\theta}}$ è la stessa sia nella funzione di ML che nella p.d.f.. Questa $\sigma_{\hat{\theta}}$ è stata già incontrata quando ci siamo occupati del calcolo (per via grafica) della varianza degli stimatori di ML. Abbiamo visto che cambiando il parametro θ di N deviazioni standard, la funzione di log-likelihood decresce di $N^2/2$ dal suo valore massimo :

$$\log L(\hat{\theta} \pm N\sigma_{\hat{\theta}}) = \log L_{max} - \frac{N^2}{2}$$

Di conseguenza l'intervallo centrale di confidenza al 68.3 % di livello di confidenza lo si può costruire dal valore stimato del parametro θ e dal valore stimato della deviazione standard $\hat{\sigma}_{\hat{\theta}}$ come [a,b] = [$\hat{\theta} - \hat{\sigma}_{\hat{\theta}}$, $\hat{\theta} + \hat{\sigma}_{\hat{\theta}}$] .

Se la funzione di massima verosimiglianza non è gaussiana l'intervallo di confidenza centrale può ancora essere ottenuto come [a,b]= $[\hat{\theta} - c, \hat{\theta} + d]$ usando la relazione :

$$\log L(\hat{\theta}^{+d}_{-c}) = \log L_{max} - \frac{N^2}{2}$$

dove $N = \Phi^{-1}(1 - \gamma/2)$ è il quantile della gaussiana standardizzata con $1 - \gamma$ livello di confidenza voluto.

Nel caso di di minimi quadrati con errori gaussiani , cioè con $\log L = -\chi^2/2$, allora si ha :

$$\chi^2(\hat{\theta}^{+d}_{-c}) = \chi^2_{min} + N^2$$

7.6 Limiti sulla Media di una Variabile Poissoniana in Presenza di Fondo

Abbiamo già visto come si possono calcolare limiti sulla media di una variabile poissoniana. In quell'esempio però non tenevamo presente il fatto che generalmente sono anche presenti eventi di fondo (riconosciuti come segnale ma che segnale non sono). Supponiamo perciò di avere un campione di n eventi che è la somma di n_s eventi di segnale e n_b eventi di fondo.

$$n = n_s + n_b$$

n_s e n_b sono variabili poissoniane con valori medi ν_s e ν_b. Mediante studi con eventi Monte Carlo è possibile conoscere il numero medio di eventi di fondo ν_b . Supponiamo che questo numero sia noto con errore zero. Vogliamo cosi determinare il limite superiore di ν_s, avendo osservato in totale n eventi di cui in media ν_b sono da considerare di fondo.

La variabile n è anch'essa una variabile poissoniana con una funzione di probabilità data da :

$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}e^{-(\nu_s+\nu_b)}$$

Lo stimatore di ML per ν_s è data da :

$$\hat{\nu}_s = n - \nu_b$$

Per determinare i limiti dell'intervallo centrale di confidenza , bisogna risolvere le equazioni:

$$\alpha = P(\hat{\nu}_s \geq \hat{\nu}_s^{oss};\nu_s^{lo}) = \sum_{n\geq n_{oss}} \frac{(\nu_s^{lo}+\nu_b)^n e^{-(\nu_s^{lo}+\nu_b)}}{n!}$$

$$\beta = P(\hat{\nu}_s \leq \hat{\nu}_s^{oss};\nu_s^{up}) = \sum_{n\leq n_{oss}} \frac{(\nu_s^{up}+\nu_b)^n e^{-(\nu_s^{up}+\nu_b)}}{n!}$$

La soluzione numerica di queste equazioni permette di determinare i limiti inferiori e superiori ν_s^{lo} e ν_s^{up}

Tenendo presente le soluzioni trovate nell'ipotesi di $\nu_b = 0$, si ha che :

$$\nu_s^{lo} = \nu_s^{lo}(senzafondo) - \nu_b$$

$$\nu_s^{up} = \nu_s^{up}(senzafondo) - \nu_b$$

Se il campione di dati n è piccolo, a causa di fluttuazioni di n_s e n_b , è possibile che ν_s^{up} sia negativo.

7.7 Stimatore Bayesiano per Intervalli

Parlando in termini di statistica frequentista la relazione :

$$P[\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}] = 0.95$$

significa che in n ripetuti esperimenti, nel 95 % dei casi il valore vero e non noto μ si trova nell'intervallo considerato. In termini bayesiani non si possono avere esperimenti ripetuti. Nel solo esperimento fatto il valore medio della misura è una costante mentre il valore vero μ ha una distribuzione di probabilità (grado di fiducia) finale $\pi(\theta|x_1, x_2, \cdots, x_n)$ data da :

$$\pi(\theta|x_1, x_2, \cdots, x_n) = \frac{\pi(\theta)\mathcal{L}(x_1, x_2, \cdots, x_n)|\theta)}{\int_\theta \pi(\theta)\mathcal{L}(x_1, x_2, \cdots, x_n|\theta)d\theta} \tag{7.5}$$

La probabilità che il valore da stimare sia compreso tra a e b ad un livello di confidenza γ è data da :

$$P(a \leq \theta \leq b) = \int_a^b \pi(\theta|x_1, x_2, \cdots, x_n)d\theta = \gamma \tag{7.6}$$

L'intervallo [a,b] è una stima bayesiana per intervalli al livello di confidenza γ. La relazione vista non definisce in modo univoco l'intervallo di confidenza. Spesso a questa relazione viene aggiunta la condizione che la distanza b-a sia minima.

Il denominatore nella relazione 7.5 è un numero per cui può essere inserito nella normalizzazione della funzione di likelihood $\mathcal{L}(x_1, x_2, \cdots, x_n|\theta)$. Inoltre la distribuzione iniziale può essere scelta uniforme nella zone fisica e zero altrove . Ad esempio se so che θ deve essere un numero positivo o nullo, allora definisco la distribuzione iniziale così :

$$\pi(\theta) = \begin{cases} 1 & \theta \geq 0 \\ 0 & \theta < 0 \end{cases} \tag{7.7}$$

La distribuzione finale in questo caso coincide con la funzione di likelihood normalizzata ad 1 e l'intervallo di confidenza al livello di confidenza γ si ottiene direttamente dall'integrazione della funzione di likelihood:

$$P(a \leq \theta \leq b) = \int_a^b \mathcal{L}(x_1, x_2, \cdots, x_n|\theta)d\theta = \gamma \tag{7.8}$$

Integriamo la funzione di likelihood da zero sino al punto A tale che l'area sottesa sia una frazione γ del totale :

$$\int_0^A \mathcal{L}(x_1, x_2, \cdots, x_n|\theta)d\theta = \gamma \tag{7.9}$$

In questo caso A rappresenta il limite superiore bayesiano al livello di confidenza γ . Se ad esempio γ fosse 0.90, allora potrei dire che è del 90 % la probabilità che il valore della variabile che sto stimando sia inferiore od uguale ad A.

Si noti che in statistica frequentista prima bisogna definire uno stimatore e quindi usando lo stimatore si costruisce un intervallo di confidenza. Nella statistica bayesiana la distribuzione finale è ottenuta direttamente dal campione di misure effettuate senza necessità di definire uno stimatore. Si noti anche che come detto altre volte nella statistica frequentista il parametro da stimare è una costante e non aiuta in alcun modo sapere che si trova in qualche intervallo . Per questo motivo la statistica frequentista è incapace ad usare l'informazione racchiusa nella distribuzione iniziale (come per esempio la 7.7).

Chapter 8

Verifica di Ipotesi

- Ipotesi
- Errori di tipo I e di tipo II
- Lemma di Neyman-Pearson
- Identificazione di Particelle
- Statistica di Test : Discriminante di Fisher e Reti Neurali
- Bontà del Fit
- Significanza di un Segnale Osservato
- Test del χ^2 di Pearson
- Test di Kolmogorov-Smirnov

8.1 Ipotesi

La verifica di ipotesi rappresenta una delle due principali aree della inferenza statistica, l'altra essendo la stima dei parametri. Nella indagine sperimentale capita molto spesso di dover distinguere all'interno di un insieme di misure tra un tipo che che per esempio chiamiamo segnale ed un altro che chiamiamo fondo. In ogni settore dell'attività umana si hanno problemi di stima di ipotesi. Si si produce un nuovo farmaco bisogna confrontalo col farmaco precedente e stabilire tra i due quale è il migliore: si considerano due campioni del vecchio medicinale e del nuovo medicinale e si verifica l'ipotesi che per esempio il secondo è da preferire. Un test statistico serve a decidere se una certa assunzione sulla distribuzione di una variabile casuale è accettabile sulla base delle osservazioni fatte oppure no. Questa assunzione è detta **ipotesi**. Generalmente viene anche specificata una **ipotesi alternativa** ed il test statistico in ultima analisi deve scegliere tra queste due ipotesi. L'ipotesi presa in considerazione è per tradizione detta **ipotesi nulla** H_0. Se questa ipotesi determina univocamente la p.d.f. f(x) di una variabile casuale x, allora l'ipotesi si dice **semplice**. Se la forma della p.d.f. è definita ma non cosi almeno uno dei suoi parametri θ, f(x, θ) allora l'ipotesi è detta **composta**. Noi consideriamo solo il caso di ipotesi semplici.

Sia $\vec{x} = (x_1, \cdots, x_n)$ un insieme di n misure della nostra variabile casuale x. Supponiamo di dover scegliere tra due ipotesi, quella nulla H_0 ed una alternativa H_1. L'ipotesi nulla specifica una p.d.f. congiunta, $f(\vec{x} \mid H_0)$, mentre quella alternativa specifica la p.d.f. congiunta $f(\vec{x} \mid H_1)$. Per vedere quale delle due ipotesi si accorda meglio con le misure sperimentali, ci serviamo di una **statistica di test** $t(\vec{x})$. Ad ognuna di queste due ipotesi corrisponderà una determinata p.d.f. per la statistica $t(\vec{x})$, cioè ad esempio $g(t \mid H_0)$ e $g(t \mid H_1)$. La statistica di test può essere un vettore a più dimensioni :

$$\vec{t} = (t_1, \cdots, t_m)$$

La dimensione m viene presa minore di n in modo da diminuire la quantità di dati senza comunque perdere la capacità di discriminare tra le due ipotesi. Noi per semplicità assumiamo che la statistica di test sia una funzione scalare $t(\vec{x})$. Si veda la figura : definiamo un taglio t_{cut} in base al quale decidiamo se l'ipotesi nulla debba essere accettata o meno. L'ipotesi nulla H_0 è respinta se t ha un valore maggiore del taglio considerato. Questa regione è detta **regione critica**. La regione complementare è detta regione d'accettanza. Se t e' misurato in questa regione l'ipotesi nulla è accettata (e naturalmente quella alternativa è respinta). Consideriamo ora l'integrale della p.d.f. dell'ipotesi nulla tra il valore di taglio accettato e l'infinito :

$$\alpha = \int_{t_{cut}}^{\infty} g(t \mid H_0) dt$$

α è detto **livello di significanza del test**. α è anche detto **misura del test**. Lo sperimentatore fissa la misura del test. Naturalmente deve essere la più piccola possibile per minimizzare la perdita di buoni dati. D'altro canto deve essere grande abbastanza da rigettare il maggior numero possibile di eventi di fondo.

Se decidiamo di non rigettare (e quindi di accettare) l'ipotesi nulla per i valori di t minori di t_{cut}, allora abbiamo una probabilità α di rigettare H_0 quando H_0 è vera. Questo errore si dice **errore di prima specie** oppure **errore di tipo I**.

È possibile che quando t è minore del t_{cut} l'ipotesi vera non sia quella nulla H_0 che abbiamo accettato ma sia quella alternativa (che abbiamo rigettato). Questo tipo di errore si dice **errore di seconda specie** oppure **errore di tipo II**. La probabilià β di commettere un errore di tipo II è data da :

$$\beta = \int_{-\infty}^{t_{cut}} g(t \mid H_1)dt$$

La quantità 1-β è la probabilità di rigettare H_0 quando questa ipotesi nulla è falsa. La quantità 1-β è detta potere del test. Si noti che il concetto di potere del test è legato alla presenza dell'ipotesi alternativa e che o è vera H_0 o è vera H_1. L'insieme (α, β) rappresentano la caratteristica del test

Fissato un determinato valore di taglio t_{cut}, noi possiamo stimare quanto vale l'errore di prima specie e quindi l'efficienza della mia selezione e quanto vale l'errore di seconda specie e quindi la purezza del campione selezionato. In questo caso di statistica di test monodimensionale, il valore di taglio fissa automaticamente l'efficienza della selezione e la purezza del campione selezionato. Se aumento il valore di t_{cut} aumento l'efficienza e diminuisco la purezza. Questo lo posso trovare utile nella ricerca di eventi rari nei quali devo avere alta efficienza. In altri casi ho bisogno di maggiore purezza e percio' diminuisco il valore del t_{cut}. Questo e' il caso ad esempio della selezione di campioni di un particolare tipo da utilizzare nella calibrazione di un rivelatore.

Nel caso di statistiche di test multidimensionali $\vec{t}$ la scelta delle regioni critiche e di accettanza non e' altrettanto ovvia e semplice. Vi possono essere diverse regioni critiche ω_α con la stessa misura α del test. Tra le possibili regioni critiche noi scegliamo la regione critica ω_α in modo tale che il potere del test sia massimizzato per la data ipotesi alternativa. Regioni critiche di questo tipo si dicono **regioni critiche migliori (BCR)**. Un test basato su una BCR è detto **più potente (MP)**. Il test piu' potente assicura per un fissato α il valore massimo per la probabilita' $(1 - \beta)$. Se il test MP di una data misura α è ritenuto insufficiente a rigettare una determinata quantità di fondo in un esperimento, allora bisogna aumentare la misura α del test, sacrificando dati buoni.

L'esistenza e la individuazione del test piu' potente per la verifica di due ipotesi semplici tra loro in alternativa sono assicurate dal Lemma di Neyman-Pearson.

8.2 Lemma di Neyman-Pearson

Supponiamo di dover scegliere tra due ipotesi semplici H_0 e H_1 tra loro in alternativa. Il problema che ci poniamo è il seguente : data una statistica di test t multidimensionale [$\vec{t} = (t_1, t_2, \cdots, t_n)$] come facciamo a costruire la regione migliore critica che per una determinata efficienza (misura del test α) dia il massimo di purezza (cioe' il massimo potere del test $(1 - \beta)$? La risposta ci viene dal **Lemma di Neyman-Pearson** : la regione di accettanza con la più elevata purezza (più elevata potenza) per una data

misura del test (cioe' quindi per una data efficienza) è data dalla regione dello spazio $\vec{t}$ nella quale:

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c$$

dove c è una costante che dipende dalla efficienza richiesta. Questo rapporto è detto rapporto di massima verosimiglianza (likelihood ratio) per le ipotesi semplici H_0 e H_1.

8.3 Identificazione di Particelle

Un caso molto importante di verifica di ipotesi è la identificazione di particelle. Ad esempio in esperimenti di fisica subnucleare è necessario saper distinguere se una particella è un pione o un kaone. Ad esempio se vogliamo individuare un decadimento raro del mesone B in η' K è fondamentale saper individuare il kaone (e non confonderlo con un pione che im media viene prodotto in quantità sette volte maggiore). Per fare questo si introduce nell'apparato sperimentale un sotto-rivelatore dedicato alla identificazione delle particelle. Informazione sulla identità delle particelle viene anche da altri rivelatori dell'apparato sperimentale (come la parte del tracciamento delle tracce, il calorimetro e.m., ecc).

8.3.1 PDF

La risposta di un rivelatore ad un tipo di particelle è data dalla p.d.f. $P(x; p, H)$: essa descrive la probabilià che una particella di quantità di moto p e di tipo H (dove H può essere un e,π, K, p, ecc) lasci un segnale x (dE/dx, angolo cherenkov,ecc).

Quindi $P(x; p, H)$ dx è la probabilià che il rivelatore al passaggio di una determinata particella di tipo H e quantità di moto p dia una misura nell'intervallo (x , x + dx)

La p.d.f. viene determinata con campioni di controllo (cioè eventi veri caratterizzati dal fatto che con elevata purezza conosco la natura delle tracce figlie del B ricostruito con semplici tagli cinematici (B in $D^*\pi$ con D^* in $D\pi$).

8.3.2 Likelihood

La funzione di likelihood per una determinata particella di quantità di moto p e di tipo H in una misura x è definita da :

$$L(H; p, x) \equiv P(x; p, H)$$

La funzione di likelihood L e la p.d.f. sono cose differenti. La p.d.f. è una funzione della misura x per una fissata quantità di moto p e per un fissato tipo di particella H).

La funzione di likelihood *L(H; p,x)* è una funzione del tipo di particella H per una fissata misura della quantità di moto p e per una fissata misura x. Ipotesi alternative su diversi tipi di particelle possono essere confrontate, utilizzando il rapporto delle rispettive

funzioni di likelihood. Per esempio per discriminare tra pioni positivi (π^+) e K^+ si può utilizzare il rapporto :

$$L(K^+; p_{oss}, x_{oss})/L(\pi^+; p_{oss}, x_{oss})$$

8.3.3 Consistenza e Livello di Significanza

Un test statistico di **consistenza** non cerca di distinguere tra due ipotesi competitive bensì si pone come problema di vedere quanto bene le quantità misurate si accordino con quelle aspettate da una particella di tipo H. La domanda viene posta generalmente cosi: Qual è la frazione di tracce vere di tipo H che sembrerebbero meno di tipo H di questa traccia?

Sia P(x | H) la p.d.f. della quantità x che stiamo misurando data l'ipotesi H. Il **livello di significanza** di una osservazione x_{oss} data l'ipotesi H è definita da :

$$SL(x_{oss} \mid H) = 1 - \int_{P(x|H)>P(x_{oss}|H)} P(x \mid H)dx$$

Si noti che la consistenza è definita sulla p.d.f. della quantità x mentre il range è specificato in termini di p.d.f. e non di x.

Essendo la p.d.f. normalizzata ad 1 , allora $SL((x_{oss} \mid H)$ può essere definito equivalentemente da :

$$SL(x_{oss} \mid H) = \int_{P(x|H)<P(x_{oss}|H)} P(x \mid H)dx$$

Facciamo l'esempio di una quantità che nel rivelatore viene misurata con una p.d.f. gaussiana:

$$P(x \mid H) = \frac{1}{\sqrt{2\pi}\sigma(H)} \exp\left[-\frac{1}{2}\left(\frac{(x-\mu(H))}{\sigma(H)}\right)^2\right]$$

allora $SL(x_{oss} \mid H)$ per una ipotesi H di una quantità misurata x_{oss} è definita da :

$$SL(x_{oss} \mid H) = 1 - \int_{\mu(H)-x_{oss}}^{\mu(H)+x_{oss}} P(x; H)dx$$

Qui si sono presi limiti simmetrici attorno al valore centrale. Si parla perciò di un test a due lati. Sono possibili anche test da un lato (da x_{oss} a $+\infty$ o da $-\infty$ a x_{oss}) nel caso di p.d.f. non simmetriche.

SL può essere usata per eliminare tracce inconsistenti con l'ipotesi fatta. SL può essere usato anche per discriminare tra due ipotesi. SL $> \alpha$ implica che l'errore di tipo I è α e di conseguenza l'efficienza è uguale a 1- α.

8.3.4 Probabilità

In alcuni casi (come nella identificazione delle particelle) le probabilità a priori $P_A(H)$ delle ipotesi competitive sono conosciute . Per esempio si sa che sette pioni sono prodotti per ogni kaone. In questi casi le funzioni di likelihood possono essere utilizzate per calcolare le purezze aspettate delle date selezioni. Consideriamo il caso della separazione π/K. Allora la frazione di K in un campione con un vettore $\vec{x}$ misurato, è data da:

$$F(K;x) = \frac{L(K;x) \cdot P_A(K)}{L(\pi;x) \cdot P_A(\pi) + L(K;x) \cdot P_A(K)} = \frac{L(K;x)}{L(\pi;x) \cdot 7 + L(K;x) \cdot 1}$$

La quantità $F(K;x)$ è detta probabilità a posteriori (o anche probabilità relativa o probabilità condizionale).

Supponiamo di voler selezionare un campione di dati con $F(K;x) > 0.9$; la purezza di questo campione si ottiene calcolando il numero di K osservati nel rilevante intervallo di valori di $F(K;x)$ e normalizzando al numero totale di tracce osservate, cioè :

$$frazione(F_H > 0.9) = \frac{\int_{0.9}^{1} \frac{dN}{dF(H;x)} F(H;x) dF(H;x)}{\int_{0.9}^{1} \frac{dN}{dF(H;x)} dF(H;x)}$$

dove la variabile di integrazione è il valore di $F(H;x)$.

8.4 Statistica di test: Discriminante di Fisher e Reti Neurali

Supponiamo di dover distinguere tra due ipotesi H_0(per esempio evento segnale) e H_1 (evento di fondo). Sia $\mathbf{x}$ un vettore di n dati sperimentali [$\mathbf{x} = (\ x_1, x_2, \cdots, x_n\ )$]. Una statistica di test possiamo scriverla grazie al Lemma di Neyman-Pearson come rapporto della massima verosimiglianza:

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}$$

Questo calcolo presuppone la conoscenza di $f(\mathbf{x}|H_0)$ e $f(\mathbf{x}|H_1)$. Questo potrebbe essere fatto con una simulazione Monte Carlo dei due tipi di eventi ma spesso ciò non è affatto pratico (le p.d.f sono istogrammi a molte dimensioni!). Per questo motivo si fanno delle approssimazioni, scegliendo particolari funzioni della statistica di test dalle misure sperimentali $\mathbf{x}$. Noi consideriamo funzioni lineari e funzioni non linerari. Noi useremo dati MC di segnale per caratterizzare le p.d.f. del segnale e dati off-resonance oppure on-resonance side band per caratterizzare le p.d.f. del fondo.

8.4.1 Discriminante di Fisher

Nel discriminante di Fisher la statistica di test è una funzione lineare delle misure :

$$t(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T \mathbf{x}$$

avendo indicato con $\mathbf{a}^T$ il vettore trasposto dei coefficienti. Bisognarà determinare i coefficienti a_i in modo da massimizzare la distanza tra la pdf per una ipotesi e la pdf per l'altra ipotesi. La definizione di distanza può essere fatta in modi diversi. Noi ora consideriamo l'approccio di Fisher.

Consideriamo i valori medi e la matrice di covarianza dei dati sperimentali per le due ipotesi(k=0 e k=1):

$$(\mu_k)_i = \int x_i f(\mathbf{x}|H_k) dx_1 \cdots dx_n$$

$$(V_k)_{ij} = \int (x - \mu_k)_i (x - \mu_k)_j f(\mathbf{x}|H_k) dx_1 \cdots dx_n$$

Analogamente le due ipotesi hanno un certo valore di aspettazione ed una certa varianza per la statistica t :

$$\tau_k = \int t \cdot g(t|H_k) dt = \mathbf{a}^T \mu_k$$

$$\Sigma_k^2 = \int (t - \tau_k)^2 g(t|H_k) dt = \mathbf{a}^T V_k \mathbf{a}$$

Per aumentare la separazione, si può massimizzare la differenza $|\tau_0$ - $\tau_1|$. Si vuole anche che gli eventi di un certo tipo siano il più possibile concentrati tra di loro attorno a τ_0 e τ_1 e questo dipende dalle varianze Σ_0^2 e Σ_1^2. Volendo tenere in conto le due cose si prende come misura della separazione la quantità :

$$J(\mathbf{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}$$

Scriviamo il numeratore e denominatore di questa espressione in termini delle misure a_i :

$$(\tau_o - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^{n} a_i a_j B_{ij} = \mathbf{a}^T B \mathbf{a}$$

La matrice B è definita da :

$$B_{ij} = (\mu_o - \mu_1)_i (\mu_0 - \mu_1)_j$$

con i,j = 1,2,...,n.

Analogamente il denominatore si può scrivere :

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \mathbf{a}^T W \mathbf{a}$$

con $W_{ij} = (V_0 + V_1)_{ij}$. La misura della separazione si scrive perciò :

$$J(\mathbf{a}) = \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T W \mathbf{a}}$$

Possiamo derivare rispetto ad a_i e uguagliare a zero, ottenendo :

$$\mathbf{a} \propto W^{-1}(\mu_0 - \mu_1)$$

I coefficienti sono determinati a meno di un arbitrario fattore di scala. La martrice W^{-1} ed i valori di aspettazione μ_0 e μ_1 sono determinati mediante dati simulati .

La scala della variabile t può essere cambiata moltiplicando i coefficienti a_i per una costante. La definizione di t può essere generalizzata nel modo seguente:

$$t(\mathbf{a}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

In questo modo si può utilizzare la scala arbitraria e l'offset a_0 per fissate τ_0 e τ_1 a qualunque valore desiderato. Si può far vedere che se le pdf f($\mathbf{x}$;H_0) e f($\mathbf{x}$;H_1) sono multigaussiane e con la stessa matrice di covarianza V = V_0 = V_1, allora il discriminante di Fisher è buono quanto il test del rapporto delle likelihood. Attenzione bisogna porre particolare attenzione alla scelta delle variabili discriminanti con cui costruire il discriminante di Fisher.

8.4.2 Reti Neurali Artificiali

Se il discriminante di Fisher non ha le caratteristiche viste (cioè p.d.f. non gaussiane ouppre sono gaussiane ma con diverse matrici di covarianza), allora si possono provare altri forme più generali per la statistica di test, utilizzando le reti neurali artificiali (spesso abbreviate in reti neurali). Queste reti neurali, appartengono al settore dell'intelligenza artificiale che ha lo scopo di imitare meccanismi che avvengono in naturale. Le reti neurali infatti imitano le reti neurali bilogiche come ad esempio il cervello umano. Queste reti in anni recenti hanno avuto un forte sviluppo e notevoli applicazioni.

Un neurone e' una cellula in grado di ricevere impulsi e trasmettere impulsi Queste cellule hanno un corpo centrale (detto anche soma) che contiene il patrimonio genetico dell'individuo (DNA) e le funzioni cellulare. Questo corso centrale e' coperto da centinaia di ramificazioni (dette dendriti) e da una lunga estnsione (detta assone). I dendriti ricevono informazioni da altri neuroni trasferendole al corpo centrale mentre l'assone riceve le informazioni elaborate dal corpo centrale e le trasmette ad un altro neurone (neurone post-sinaptico) o verso altre cellule di destinazione. Il neurone quindi ha porte di ingresso da cui ricevono informazioni. In base alla intensita' delle informazioni ricevute si attiva (si eccita) oppure no. Il neurone ha una porta di uscita da cui trasmette. La trasmissione avviene solo quando il neurone si e' attivato.

Il percettrone e' la rete neurale piu' semplice. E' costituito da un solo neurone (detto anche nodo) avente un certo numero n di ingressi (variabili discriminanti x_1, x_2, ... , x_n).

Nel nodo le informazioni entranti, pesate con i pesi a_1, a_2, ..., a_n, vengono opportunamente sommate in modo da calcolare il potenziale di attivazione del nodo. La funzione di attivazione puo' avere forme diverse. Puo' dare il segno della funzione (-1, +1), oppure essere una funzione a scalino (0,1) oppure puo' dare in uscita una distribuzione continua mediante la funzione sigmoidea definita da:

$$\sigma(x) = \frac{1}{1 - e^{-x}}$$

La curva logistica di questa funzione e' rappresentata in Fig. 8.3.

Il valore calcolato viene trasmesso al nodo di uscita. La statistica di test t(x) e' data valore dato dal nodo di uscita (Fig. 8.2). La formula di uscita del percettrone e' data da:

$$t(x) = \sigma(a_0 + \sum_{i=1}^{n} a_i x_i)$$

con a_0 termine di offset (denominato bias). Questo bias puo' essere considerato il peso di un nodo fittizio. In questo modo la formula precedente puo' essere riscritta cosi:

$$t(x) = \sigma(\sum_{i=0}^{n} a_i x_i)$$

L'architettura della rete neurale e' varia. Siano n i nodi di ingresso , corrispondenti alle variabili discriminanti scelte. In uscita si possono avere uno o più nodi. Inoltre tra lo strato dei nodi di ingresso e lo strato dei nodi di uscita ci possono essere uno o piu' strati di nodi detti nascosti o intermedi (percettroni multistrato) . Fig. 8.4 e' mostrata l'architettura di una rete neurale con uno strato di ingresso di 4 nodi, uno strato di uscita con tre nodi ed uno strato nascosto con due nodi.

In queste reti possiamo decidere che i valori in input ad un certo strato dipendono esclusivamente dai nodi dello strato precedente. Questa è detta rete "feed-forward".

Noi ora vogliamo struire la rete. Il processo di addestramento viene realizzato con un campione di eventi di addestramento (training set) di tipo H_0 e di tipo H_1. Vengono dati in pasto alla rete eventi dei due tipi e la rete conosce il tipo di evento in ingresso . Il numero di cicli dell'apprendimento e' il numero di eventi usati per l'apprendimento. Ad ogni ciclo la rete riaggiusta i pesi in modo da ridurre l'errore tra il valore di uscita nel nodo e il valore vero (che la rete conosce a priori). La rete gia' addestrata viene poi usata su un campione di convalida (validation set), indipendente dal training set, in modo da valutare la qualita' dell'addestramento (in termini di risposte corrette fornite dalla rete).

Come faccio a convincermi che non ho introdotto bias addestrando la rete col campione di addestramento? Una verifica puo' essere fatta suddividendo il training set in K sottocampioni. Addestrro la rete in un sottocampione e la verifica sui K-1 sottoinsiemi aggregati. Itero K volte e prendo la media dei risultati (K-fold cross-validation).

Un ulteriore problema nell'addestramento e' quello dell' overtraining. Il training e' fatto su un campione di eventi. Se il numero di eventi di questo campione e' elevato il sistema si adatta sempre piu' alle caratteristiche di questo particolare campione.

Aumentando il numero di eventi usati nel training, l'errore che la rete commette nella separazione tra i due tipi di eventi tende a zero! In realta' quello che vogliamo e' un set di parametri ottimizzati che valga per tutta la popolazione e non solo per il campione considerato. Per risolvere questo problema di overtraining oltre al campione di training considero un altro campione di dati (validation set), indipendente dal training set. Durante la dase di addestramento della rete verific la qualita' dell'addestramento sul validation set. Quando noto che l'errore di identificazione comincia ad aumentare, arresto il training (Fig. 8.5).

Una volta che la rete sia istruita e validata, utilizzo ul altro campione di dati (test set) per valutare l'accuratezza finale della rete. Una volta istruita la rete (con i pesi a_0, ..., a_n gia' ottimizzati) , essa viene utilizzata per la discriminazione tra eventi (diversi da quelli utilizzati per l'addestramento). Questo tipo di apprendimento e' detto supervisionato.

Le fasi di learning, validation e test si applicano in generale a tutti i classificatori multivariati.

8.5 Test χ^2 di Bontà del Fit

Abbiamo già visto il concetto di consistenza e livello di significanza nel caso dell'identificazione delle particelle. Questo vale in generale. Un test di bontà del fit viene fatto dando il cosiddetto valore P , cioè la probabilità sotto l'ipotesi fatta di ottenere un risultato compatibile o meno compatibile di quello effettivamente osservato. P è detto anche livello di confidenza del test o livello di significanza osservata.

8.6 Significanza di un Segnale Osservato

Un test di bontà del fit viene usualmente fatto per verificare l'attendibilità o meno di un segnale.

Supponiamo che in una certa regione del segnale siano presenti n candidatti. Supponiamo di sapere che n_b sia il numero di eventi di fondo aspettati (errore zero per ora!). Quindi il numero di eventi di segnale aspettato n_s sarebbe dato da n_s = n - n_b. Il numero di candidati osservati n ed i numeri di eventi di fondo aspettati n_b e di eventi di segnale n_s sono tutte variabile poissoniane. I valori medi sono legati dalla relazione $\nu = \nu_s + \nu_b$. La probabilità di osservare n candidati è :

$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}\exp\left[-(\nu_s+\nu_b)\right]$$

Per dire quanto io sono confidente di aver visto un segnale di n_s eventi devo controllare se il numero di eventi di fondo osservati può fluttuare sino a dare il numero di eventi uguale ad n o superiore. Quindi pongo uguale a zero il numero di eventi di segnale osservato ($\nu_s = 0$) e calcolo la probabilità che la fluttuazione del solo fondo possa dare il numero di eventi osservati n_{oss} o anche di più:

$$P(n \geq n_{oss}) = \sum_{n=n_{oss}}^{\infty} f(n; \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{oss}-1} f(n; \nu_s = 0, \nu_b)$$

$$= 1 - \sum_{n=0}^{n_{oss}-1} \frac{\nu_b^n}{n!} \exp\left[-\nu_b\right]$$

Per esempio supponiamo di aver osservato 5 eventi e di aspettare ν_b=0.5 eventi di fondo. In questo caso il valore P è $1.7 \cdot 10^{-4}$. Consideriamo come ipotesi nulla H_0 l'ipotesi nulla che cio' che osservo e' una fluttuazione del fondo in alto dal valore medio. In questo caso in termini frequentisti accettando l'ipotesi nulla farei una cosa giusta 1 su 5882 volte. Quindi rigetto l'ipotesi nulla con un p-value di $1.7 \cdot 10^{-4}$. Poiche' stiamo cercando una fluttuazione in alto dal valore medio, e' comodo determinare in una distribuzione gaussiana standard il punto P tale che l'area sotto la curva da questo punto all'infinito sia pari al p-value. Il numero di σ con cui P dista dal valore medio rappresenta la significanza (statistica con cui rigetto l'ipotesi nulla. Nell'esempio che abbiamo fatto l'ipotesi di fluttuazione del fondo e' rigettata con una significanza statistica di 3.6 σ.

Bisogna tener conto della precisione con cui è noto il numero di eventi di fondo aspettato. Bisogna determinare un intervallo di possibili valori di ν_b e conseguentemente calcolare i possibili valori P.

8.7 Test del χ^2 di Pearson

È possibile dare la significanza di un segnale attraverso il calcolo del χ^2. Consideriamo la distribuzione della variabile considerata X in N intervalli. L'istogramma viene fittato con una curva che descrive il fondo piu' un'altra curva che descrive un eventuale eccesso di eventi in una regione dove e' aspettato un segnale. Per ogni bin posso determinare dal fit il numero di eventi aspettato come segnale ed il numero di eventi aspettato come fondo. Come faccio ad essere sicuro che n_s eventi trovati nella regione del segnale sono veramente eventi di segnale e non il frutto di una fluttuazione statistica del fondo in alto dal valore medio aspettato? Faccio l'ipotesi che ci sia solo fondo e vedo quanto i dati sperimentali sono inconsistenti con questa ipotesi. Rifaccio il fit ponento nel fit $n_s = 0$ e calcolo il χ^2 a partire dan numero di eventi n_i trovati nel bin i-esimo e dal numero di eventi ν_i aspettati dal fit:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

Come visto altre volte questa variabile segue la distribuzione del χ^2 indipendentemente dalla distribuzione seguita dalla variabile X. (È necessario che in ogni bin ci siano almeno 5 o più eventi, altrimenti si passa ad esperimenti simulati).

Il livello di significanza osservato (valore P) si ottiene integrando la distribuzione del χ^2 dal valore osservato all'infinito :

$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz$$

done n_d = N è il numero di gradi di libertà. Questa quantita' (o la corrispondente significanza statistica) mi dice quanto eventualmente i miei dati sperimentali sono inconsistenti con l'ipotesi di fluttuazione del fondo.

Se non sappiamo a priori dove e' la regione del segnale cercato allora dobbiamo tener conto del fatto che la fluttuazione del fondo sia avvenuta in quel punto casualmente. Trovato un certo disaccordo in un bin, dovrei dare la probabilita' di osservare lo stesso disaccordo in qualunque altro bin (look elsewhere effect). Questo abbassa la significanza dell'eventuale disaccordo tra ipotesi di solo fondo e dati sperimentali. Notiamo che a rigore trovare un disaccordo tra dati e ipotesi di fondo non significa che abbiamo dimostrato l'esistenza di un segnale!

8.8 Test di Kolmogorov-Smirnov

Il test del χ^2 si basa sull'uso di dati istogrammati. L'uso di dati istogrammati comporta sempre una perdita di informazione e quindi di potere nel test. Con campioni di dati piccoli inoltre può essere impossibile trovare un bin di dimensioni sensibili. Il test di bontà del fit di Kolmogorov-Smirnov utilizza dati non istigrammati. Questo test comunque non e' propriamente un test per verificare la bonta' del fit di parametri ma quanto l'attendibilita' che un campione di dati segua una prefissata p.d.f. a parametri costante. Il test e' molto utile per verificare se due campioni di dati provengono dalla stessa popolazione.

Supponiamo di avere un campione di n misure $(x_1, x_2, \cdots, x_n)$ e vogliamo verificare se questa distribuzione segue una certa p.d.f. (continua a parametri noti). Possiamo calcolare la c.d.f. F della p.d.f. f . La funzione F è una curva non decrescente che va da zero ad 1. Possiamo quindi disporre i dati del campione in ordine crescente ed ottenere la c.d.f. del campione. Questa è una curva a scalino che va da zero ad 1. Queste due curve dovrebbero coincidere (se il campione effettivamente segue la p.d.f. f) o almeno dovrebbero coincidere i loro valori di aspettazione. È chiaro che io posso misurare quanto le due curve differiscono tra di loro per stimare se il campione è estratto dalla popolazione data. Per vedere quanto le due curve differiscono tra di loro , calcolo la funzione $S_n(x)$ cosi definita:

$$S_n(x) = \begin{cases} 0 & \text{per } x < x_{(1)} \\ \frac{r}{n} & x_{(r)} \leq x \leq x_{(r+1)} \\ 1 & x_{(n)} \leq x \end{cases} \tag{8.1}$$

dove $x_{(r)}$ è la statistica di ordine r del campione x ($x_{(n/2)}$ è la mediana del campione di dati).

Il grafico di $S_n(x)$ è una funzione a scalino : ad ogni misura $S_n(x)$ aumenta di un passo pari ad 1/n. Possiamo definire la statistica :

$$D_n \equiv Max|S_n(x) - F(x)|$$

D_n è usualmente chiamata statistica di Kolmogorov-Smirnov. Una proprietà di questa statistica è quella di essere " distribution free". Infatti noi possiamo applicare qualunque trasformazione continua ad x senza che D_n ne risenta in alcun modo: la distorsione dell'asse x non può cambiare i valori sull'asse y.

Moltiplicando D_n per $\sqrt{n}$ si ha :

$$d_n = D_n\sqrt{n}$$

Se l'accordo è buono d_n deve assumere valori piccoli. Queste funzioni sono tabulate e i quantili della distribuzione di d_n sono riportate nelle tavole statistiche e si possono calcolare meglio calcolare utilizzando in rete un calcolatore statistico.

Il test di Kolmogorov-Smirnov risulta particolarmente utile quando si vuole controllare se due campioni provengono dalla stessa popolazione. In questo caso dette n_1 e n_2 le dimensioni dei due campioni di dati e S_{n_1} e S_{n_2} le corrispondneti funzioni cumulative empiriche, si pone:

$$D_{n_1-n_2} \equiv Max|S_{n_1} - S_{n_2}|$$

e si usano queste funzioni per misurare l'attendibilitá che i due campioni provengano dalla stessa popolazione. Anche i quantili della distribuzione sono riportate nelle tavole statistiche.

Il test di Kolmogorov-Smirnov e' molto piu' sensibile del test del χ^2. Si veda la Fig. 8.6 in cui si mostra una situazione in cui il test del χ^2 (a sinistra) e' meno efficace del test di Kolmogorov-Smirnov (a destra). I valori del χ^2 dei fit dei due istogrammi con una retta potrebbero essere egualmente buoni anche se e' evidente che la retta si dovrebbe accordare meglio col secondo istogramma. Questo fatto e' chiaramente visibile nel test di Kolmogorov-Smirnov (a destra).

Figure 8.1: Struttura di un neurone.

Figure 8.2: Schema di un percettrone.

Figure 8.3: Andamento della curva logistica della funzione sigmoidea.

Figure 8.4: Struttura di una rete neurale.

Figure 8.5: Analisi dell'overtraining. In ordinata l'errore usato per modificare i pesi ad ogni ciclo

Figure 8.6: Due diversi istogrammi fittati con una retta. A sinistra test del χ^2 e a destra quello di Kolmogorov-Smirnov.

Chapter 9

Confronto Teoria - Esperimento

- Introduzione
- Funzione di Risoluzione
- Accettanza ed efficienza di Rivelazione

9.1 Introduzione

Se noi simuliano un certo numero di eventi e poi li ricostruiamo, notiamo subito che le quantità ricostruite dal rivelatore sono diverse da quelle simulate. per esempio il vertice del decadimento di una particella , simulato con un certo valore V(x,y,z), verrà trovato in V(x_1, y_1, z_1). Questo è dovuto al fatto che il rivelatore ricostruisce le varie quantità con un certo errore. Se questo errore fosse nullo, allora il valore ricostruito sarebbe uguale a quello simulato ed il rivelatore avrebbe una risoluzione infinita. Nella realtà tutti i rivelatori hanno una risoluzione finita. Questo effetto che abbiamo visto per gli eventi simulati è presente anche per gli eventi reali. Quindi una quantità che in produzione era x dopo la ricostruzione dell'evento nel rivelatore diventa y. È evidente che se vogliamo confrontare dati sperimentali a modelli teorici dobbiamo tener conto di questo effetto. Poichè la risoluzione sperimentale dipende dal rivelatore , bisogna correggere per questo effetto prima di confrontare risultati di due esperimenti diversi.

9.2 Funzione di risoluzione

Consideriamo la p.d.f. vera di una variabile casuale x, f(x;θ). Quando io misuro la variabile x nel mio rivelatore ottengo la quantità x'. La funzione r(x';x) che da la distribuzione del valore misurato x' per un dato valore vero x è detta **funzione di risoluzione** (del nostro rivelatore) per la variabile x.

La p.d.f. della osservabile x' è data da :

$$f'(x';\theta) = \int f(x;\theta)\, r(x';x,a)dx = f(x;\theta) \otimes r(x';x,a)$$

dove con a abbiamo determinato uno (o più parametri) da cui eventualmente dipende la funzione di risoluzione. Noi diciamo che la funzione f(x;θ) è stata " smeared out" in f'(x';x). Si dice anche che la p.d.f. vera è "folded" con la funzione di risoluzione e il compito di ricavare la p.d.f. vera è detto " **unfolding (o anche unsmearing)** ".

La funzione di risoluzione r(x';x) può essere determinata utilizzando eventi simulati. Per la variabile X distribuisco la differenza tra il valore misurato (distorto dall'apparato) e il valore vero (usato nella simulazione). Questa distribuzione generalmente segue un andamento di tipo gaussiano (o somma di due o più gaussiane).

$$r(x';x) = \frac{1}{\sqrt{2\pi}R} \exp\left[-\frac{1}{2}\left(\frac{x'-x}{R}\right)^2\right]$$

Dal fit mi ricavo i valori dei parametri (R nel nostro esempio) della funzione di risoluzione. La deviazione standard R e' anche detta larghezza della risoluzione. In alcuni casi l'integrale di convoluzione visto può essere calcolato analiticamente mentre altre volte il calcolo va eseguito in modo numerico.

9.2.1 p.d.f. Esponenziale e Funzione di Risoluzione Gaussiana

Supponiamo di avere una variabile casuale x con una p.d.f. di tipo esponenziale:

$$f(x;\lambda) = \lambda \exp(-\lambda x)$$

con λ vita media di una particella che decade in volo.

Supponiamo anche che la funzione di risoluzione abbia forma gaussiana su tutto il range in cui è definita la variabile casuale X. Si ha in questo caso :

$$f'(x') \simeq \int_0^\infty \lambda \exp[-\lambda x] \exp\left[-\frac{1}{2}\left(\frac{x'-x}{R}\right)^2\right] dx$$

Questo integrale può essere calcolato analiticamente :

$$f'(x') \simeq \exp\left[(\lambda r)^2/2\right] \cdot G\left(\frac{x'}{R} - \lambda R\right) \cdot \lambda \exp(-\lambda x')$$

dove G è la distribuzione cumulativa della normale standard.

In Fig. 9.1 si può vedere come era la distribuzione esponenziale vera e come si deforma per effetto della risoluzione finita dell'apparato sperimentale per diversi valori della vita media λ e larghezza della risoluzione R.

9.2.2 p.d.f. Gaussiana e Funzione di Risoluzione Gaussiana

Un altro caso particolare in cui l'integrale di convoluzione puo' essere calcolato analiticamente e' quando sia la p.d.f. sia la funzione di risoluzione sono di tipo gussiano. La p.d.f. abbia ad esempio una distribuzione normale con media x_0 e e deviazione standard σ:

$$f(x;x_0,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]$$

in tutto l'intervallo di variabilita' della variabile x ($-\infty \leq x \leq +\infty$). La funzione di risoluzione abbia anch'essa una forma gaussiana con larghezza della risoluzione R:

$$r(x';x) = \frac{1}{\sqrt{2\pi}R} \exp\left[-\frac{1}{2}\left(\frac{x'-x}{R}\right)^2\right]$$

In questo caso la soluzione dell'integrale di convoluzione e' data da:

$$f(x') = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+R^2}} \exp\left[-\frac{1}{2}\left(\frac{x-x_0}{\sqrt{\sigma^2+R^2}}\right)^2\right]$$

la p.d.f. della osservabile x' e' anch'essa di tipo normale con lo stesso valore media della osservabile x ma con una varianza data dalla somma delle varianze della originaria osservabile x e della funzione di risoluzione.

9.3 Accettanza ed Efficienza di Rivelazione

Spesso gli apparati sperimentali non hanno la possbilita' di rivelare l'evento in tutto l'angolo solido. Questo perche' ci sono difficolta' sperimentali che lo impediscono o perche' costerebbe troppo (in rapporto ai vantaggi che se ne trarrebbero). Quindi un rivelatore puo' avere limitata accettanza. Sia f(x;θ) la p.d.f. teorica (gia' convoluta con la funzione di risoluzione) della variabile X. Supponiamo che nel range di variabilitá della osserbabile X, $[-\infty, +\infty]$, la regione di accettanza del rivelatore sia conpresa tra A e B. Allora la p.d.f. f'(x; θ) della osservabile X troncata all'intervallo [a, b] e' data da:

$$f'(x;\theta) = \frac{f(x;\theta)}{\int_A^B f(x;\theta)dx} = \frac{f(x;\theta)}{F(B) - F(A)}$$

dove F denota la distribuzione cumulativa della p.d.f.. Questa p.d.f. e' definita nulla all'esterno della regione di accettanza. Si noti si tratta di limitare il range di variabilitá della osservabile X e di rinormalizzare in questo intervallo la p.d.f. $f(x;\theta)$.

Spesso nell'intervallo di accettanza il rivelatore ha diversa efficienza di rivelazione a seconda dei particolari valori della variabile X. Il rivelatore rivela con efficienza diversa la variabile X nei diversi punti dello spazio delle fasi accessibile. Nel confronto tra modello teorico ed esperimento bisogna correggere per questa inefficienza dell'apparato sperimentale. Questo compito puo' essere svolto in due modi diversi: o alterando la p.d.f. f(x;θ)del modello teorico o agendo sui dati sperimentali. Nel primo modo possiamo determinare l'efficienza di rivelazione D dell'osservabile X in corrispondenza al valore x. Questa efficienza di rivelazione in generale dipende non sola dalla variabile X ma anche da altre variabili che globablmente indichiamo con Y che a loro volta possono essere correlate con la variabile X. In questi casi bisognerebbe determinare la probabilita' condizionale $P(y|x)$ di y dato x e calcolare la nuova p.d.f. ideale corretta come:

$$f'(x;\theta) = \frac{\int f(x;\theta)D(x,y)P(y|x)dy}{\int\int f(x;\theta)D(x,y)P(y|x)dydx}$$

dove D(x,y) e' la efficienza di rivelazione. L'integrazione in y e' fatta perche' non abbiamo alcun interesse in questa variabile.

La efficienza di rivelazione, essendo una proprieta' dell'apparato sperimentale, puo' essere determinata prima di fare l'esperimento. Questo puo' essere fatto ad esempio simulando molti eventi e ricostruendoli nell'apparato sperimentale. La efficienza di rivelazione nell'intorno di un valore x della variabile X si ottiene dal rapporto tra numero di eventi ricostruiti e numero di eventi simulati nell'intorno del valore x considerato. La determinazione di $P(y|x)$ puo' essere un compito estremamente difficile e talvolta irrealizzabile prima di aver eseguito l'esperimento. In questi casi la probabilita' condizionale dovrebbe essere ottenuta dai dati a scapito della precisione con cui si determina il parametro θ. Vi sono casi particolari nei quali l'efficienza di rivelazione dipende solo dalla osservabile X e in questi casi la p.d.f. ideale corretta f'(x;θ) sarebbe data da :

$$f'(x;\theta) = \frac{f(x;\theta)D(x)}{\int f(x;\theta)D(x)dx}$$

dove abbiamo indicato con D(x) l'efficienza di rivelazione della variabile X e l'integrale e' esteso alla regione di accettanza del rivelatore.

Un diverso modo di risolvere il problema del confronto teoria-dati in presenza di efficienza di rivelazione variabile da punto a punto e' di agire sui dati sperimentali e correggerli prima di fare il confronto col modello teorico. Se abbiamo osservato un valore x_i della osservabile X, possiamo correggere la misura, pesamdo questo evento con un peso w_i, dove w_1 e' l'inverso della probabilita' di rivelazione di questo evento:

$$w_i = \frac{1}{D(x_i, y_i)}$$

Questa tecnica di pesare gli eventi sperimentali prima del confronto col modello teorico e' quella piu' spesso usata nella pratica (vedi esempi di analisi in ampiezza e Dalitz plot).

Figure 9.1: Distorsioni di una p.d.f. esponenziale con vita media λ dovute a funzioni di risoluzioni gaussiane con diversi valori di λ e della larghezza di risoluzione R