

CHAPTER 6

Test d'ipotesi

6.1 Test del χ^2

Consideriamo un campione di N misure $\{(x_i, y_i)\}_i^N$ IID legate tra loro da una funzione $y = \psi(x, \vec{\theta})$ avremo che le misure campionate rispetto alla variabile aleatoria y , possono essere riscritte come $y_i = \psi(x, \vec{\theta}) + \epsilon_i$ dove ϵ ipotizziamo essere una variabile aleatoria la cui $pdf(\epsilon)$ segue una distribuzione di probabilità Gaussiana. Nell'ipotesi in cui valga il TCL per le ϵ_i , si ha che il Q^2 associato alle misure e il modello segue la distribuzione di $\chi^2(N - k)$.

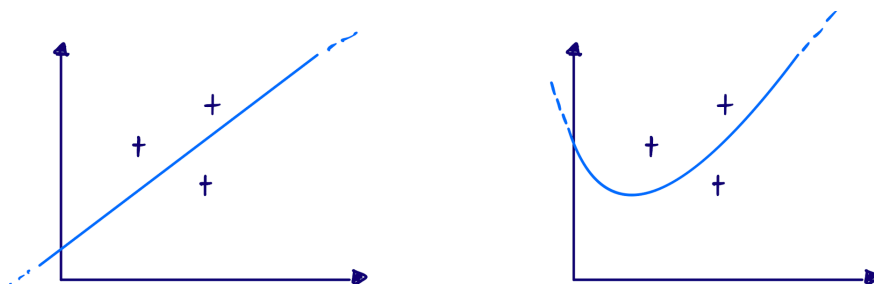
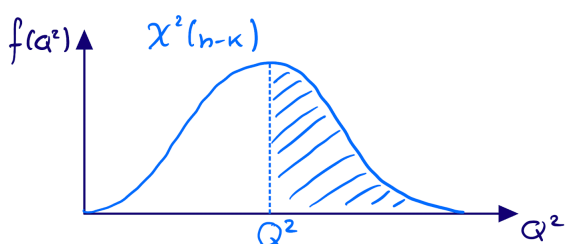


Figure 6.1: Due modelli differenti che interpolano lo stesso campione di dati.

Nel caso di destra in figura 6.1 gli scarti quadratici sono minori, mentre in quello di sinistra sono più grandi, di conseguenza possiamo aspettarci che il valore di aspettazione della distribuzione di χ^2 del modello di sinistra sia più grande di quello di destra. Definiamo il modello non corretto (quello

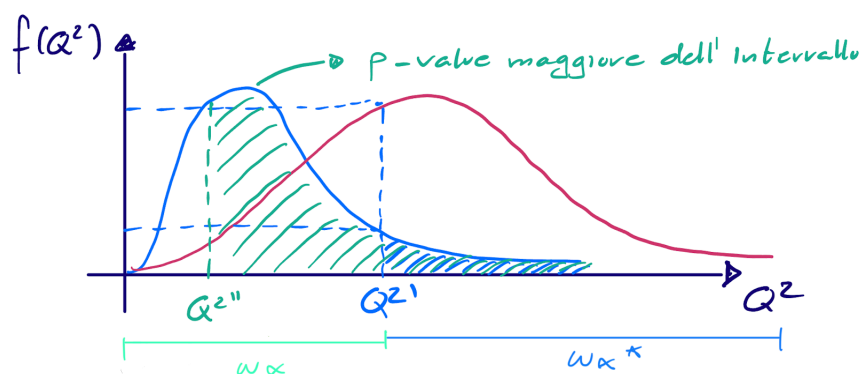
di sinistra) H_1 e il modello corretto (quello di destra) H_0 , la domanda che possiamo porci è: " Se partiamo da due ipotesi H_1 e H_0 e non sappiamo quale delle due sia corretta, come facciamo a determinare quella che descrive meglio la realtà sperimentale? ".

Introduciamo una nuova quantità definita **p-value** che ha la seguente espressione :

$$\text{p-value} = \int_{Q^2}^{\infty} \chi^2(N-k) d\chi^2 \quad (6.1)$$


la quantità così definita risulta essere una misura di probabilità. Per rispondere alla domanda precedente fissiamo una soglia di tolleranza del p-value oltre alla quale i valori ottenuti risultano essere dei **falsi negativi**. Riprendendo i modelli H_0 e H_1 che definiamo rispettivamente **null hypothesis** e **alternative hypothesis**, fissata una soglia del p-value, e definita una statistica x associata (come per esempio il Q^2) al $\chi^2(x|N-k)$ avremo che:

- H_0 è rigettata se x cade nella regione in azzurro ω_α^* in figura 6.2
- H_0 è accettata se x cade nella regione in verde ω_α nella figura sottostante.



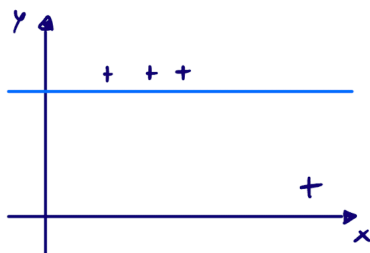
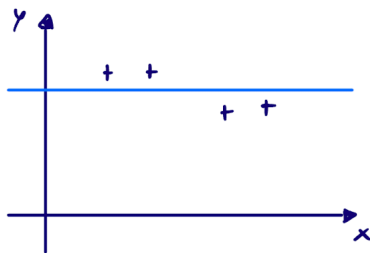
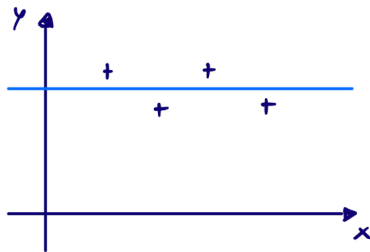
Dunque date due ipotesi H_0 e H_1 decidiamo di considerare affermativa quella

che restituisce il valore del p-value piú alto (ovvero quella con probabilità maggiore) rispetto al valore p_0 di soglia considerato, apparentemente sembrerebbe sufficiente per concludere quale delle due ipotesi sia verifica, ma in realtà come si vedrá nelle sezioni successive la realtà sperimentale é piú complicata e posso esserci casi di falso positivi.

Su quanto discusso fino ad ora possiamo fare le seguenti osservazioni:

- I minimi quadrati permettono di calcolare il Q^2 (e anche il metodo di ML calcola i parametri $\hat{\theta}$ da cui si pu calcolare il Q^2);
- Il test del χ^2 é applicabile se sappiamo calcolare il valore del $Q^2 \Rightarrow \sigma^2$ deve essere nota e ben stimata;
- Test del χ^2 un test integrale \Rightarrow Somma gli scarti su tutti gli eventi, e dunque puó presentare delle limitazioni (esempio sotto).

Esempio



Assumiamo che il p-value sia accettabile ovvero maggiore dell'intervallo di confidenza. Consideriamo gli stessi punti riorganizzati in un modo diverso, ma con stesso scarto quadratico. I due set di dati con y differenti hanno lo stesso p-value.

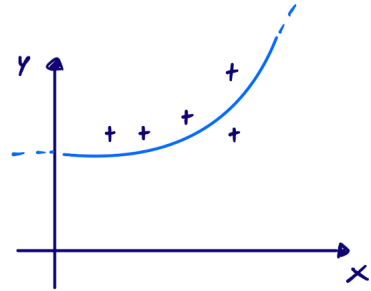
Essendo che il test del χ^2 ha una forma integrale, se graficamente si osserva una distribuzione differente delle misure, il test non lo tiene in considerazione.

Se consideriamo una dispersione delle misure come nella terza figura e assumiamo che abbia il medesimo p-value delle altre due, si ha che il test non considera che i dati diminuiscono di valore lungo l'asse delle ordinate e dunque si

ottiene un **falso positivo**, ovvero p-value è verificato, ma il modello non descrive adeguatamente il comportamento dei dati sperimentali. Il fatto che il test del χ^2 abbia forma integrale limita la generalità con cui possiamo decidere se il risultato ottenuto sia affidabile o meno.

Definizione di Overfitting

Ipotizziamo di avere un fit che ha $Q^2 = 0$ e p-value-1, ovvero i dati vengono interpolati perfettamente, questo non è un buon risultato. Si ha un caso di overfitting, dove si sono introdotti così tanti parametri che il risultato del fit si è completamente adattato alle misure, perdendo qualsiasi capacità di generalizzare il modello.



6.1.1 Applicazione del test di χ^2

Se il modello è corretto $y = \psi(x, \vec{\theta})$, allora il metodo dei MQ fornisce una stima dei parametri $\vec{\theta}$ che lo descrivono. Il valore stimato dei parametri rappresenta il punto di minimo della funzione di $Q^2_{min} = Q^2(x, \vec{\theta}_{MQ})$ rispetto al campione sperimentale, tale punto di minimo coincide anche con il massimo della distribuzione di χ^2 se pdf(ϵ) seguono una distribuzione gaussiana, che è dato da $E[Q^2] = N - k$ e quindi $Q^2_{min} = N - k$.

Il χ^2 ridotto è definito come $\chi^2_0 = \frac{\chi^2}{N-k}$ ciò implica che per $\chi^2 = Q^2_{min}$ il ridotto è $\chi^2_0 \sim 1$.

Se il valore del χ^2 è lontano dal suo valore di aspettazione $N-k$ (o 1 nel caso si usi quello ridotto), possiamo concludere che alcune delle ipotesi precedenti non sia corrette e dunque **i dati non confermano il modello**.

Le regioni a bassi valori di χ^2 corrispondono a scarti tra modello e dati molto piccoli, quindi a casi di overfitting. Tali risultati sono altamente improbabili. Tali risultati possono essere dovuti sovrastime degli errori, la pdf non è Gaussiana o se le incertezze considerate non hanno solo incertezze di tipo statistico.

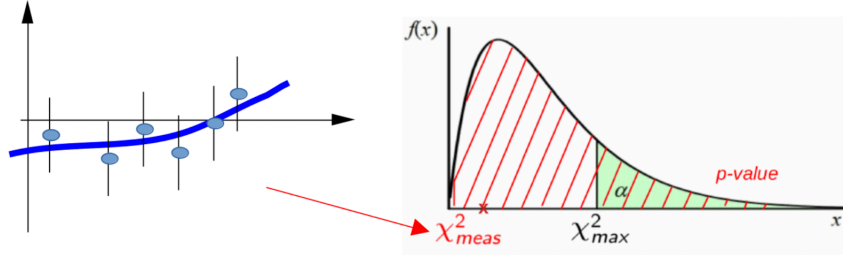


Figure 6.2: Overfitting

6.2 Errori di test statistici

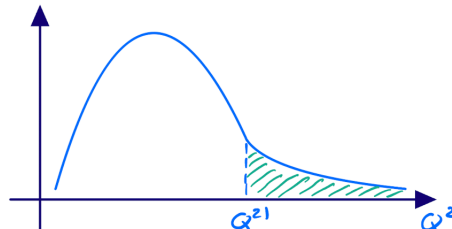
Quanto discusso in questa sezione fa riferimento al test del χ^2 , ma generalizzabile a qualsiasi statistica β per cui é possibile definire due distribuzioni di probabilità:

- $pdf(\beta|H_0) = \text{Null Hypothesis}$
- $pdf(\beta|H_1) = \text{Alternative Hypothesis}$

6.2.1 Errori del I° tipo

Un errore del primo tipo rappresenta il numero di casi veri per la null hypothesis H_0 che scartiamo fissata una soglia del p-value.

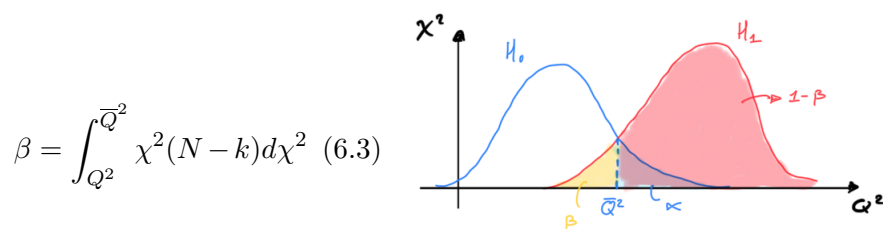
$$\alpha = \int_{Q^2}^{\infty} \chi^2(N-k) d\chi^2 \quad (6.2)$$



Il termine α prende il nome di **size del test**.

6.2.2 Errori del II° tipo

Un errore del secondo tipo rappresenta la probabilità di accettare H_0 quando è vera H_1 , in questo caso si parla di **falsi positivi**.



Si sta commettendo un errore poichè se H_1 è la forma funzionale sbagliata il fit dei dati supera ugualmente il test del χ^2 .

Il termine $1 - \beta$ prendere il nome di **power del test** e restituisce la probabilità di rifiutare H_0 quando H_1 è vera.

Fissate le due ipotesi alternative e definiti gli intervalli di confidenza, se si assume che l'errore di tipo uno sia quello più grave si procede scegliendo la percentuale di falsi negativi che si reputa accettabile e si cerca di definire gli intervalli ω_α e ω_α^* in modo tale che β sia il minore possibile (minor caso di falsi positivi). Il test così descritto viene definito il più potente per un determinato valore di soglia del p-value.

6.3 Test di Kolgomorov-Smirnov

Consideriamo un insieme di N misure della stessa grandezza fisica X vogliamo testare la null hypothesis H_0 che siano campionamenti di una determinata pdf che prendiamo come riferimento. Una possibilità è di usare il test del χ^2 applicandolo agli istogrammi costruiti con le misure raccolte e la pdf-modello. Tale procedura è corretta, ma richiedere di binnare i dati e dunque si ha una perdita d'informazione, inoltre l'esito del test può dipendere dal binning scelto per la costruzione degli istogrammi.

L'alternativa è data dal test di **Kolgomorv-Smirnof** che confronta le due distribuzioni cumulative (dati - pdf-modello) e in questo modo sfrutta

tuta l'informazione contenuta nei dati. Tale test è di tipo non parametrico ovvero non richiede la costruzione di stimatori rispetto ai dati raccolti sperimentalmente e viene utilizzato per variabili aleatorie continue.

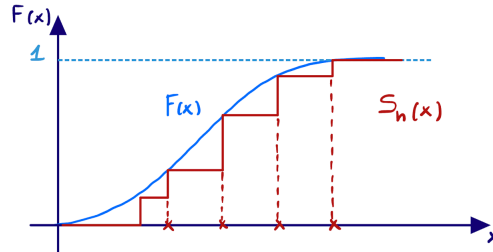
6.3.1 Costruzione del test

Date N misure ordinate in senso crescente, ricostruiamo la distribuzione cumulativa della pdf-modello, definendo una funzione a gradini $S_n(x)$ rispetto ai dati del campione, tale funzione prende il nome di **EDF (Empirical Distribution Function)**, la scelta ricade su una funzione di questo tipo poichè sono presenti dei buchi nell'informazione (campione) raccolta. La forma dell'EDF è data:

$$S_n(x) = \sum_{i=1}^n I(x_i \leq x) \quad (6.4)$$

dove la funzione $I(x_i \leq x)$ prende il nome di **indicator function** ed è espressa come:

$$I(x) = \begin{cases} 1 & x \leq x_{i+1} \\ 0 & \text{altrimenti} \end{cases}$$



Ci si domanda quanto bene S_n approssimi la cumulativa $F(x)$, la risposta è che dipende dal numero di misure contate prima del gradino successivo e quindi da come si scelgono gli intervalli. Per valutare l'approssimazione per ogni punto distinto che costituisce un estremo degli intervalli \tilde{x} valutiamo l'estremo superiore della differenza tra la EDF e la pdf con cui vogliamo confrontarla:

$$D_n = \sup_x |S_n(x) - F(x)|$$

la distanza $d_n \equiv \sqrt{n}D_n$ definisce il valore di riferimento per il test di KS che consiste nel confrontare tale numero con una grandezza di riferimento d_0 , che costituisce la quantità di soglia rispetto alla quale rigettare la null

hypothesis H_0 . Se $d > d_0$ l'ipotesi di compatibilità viene rigettata. Il valore di δ_0 è scelto in base alla probabilità che la variabile casuale δ sia maggiore di δ_0 quando il modello è corretto.

$$P(\delta > \delta_0 | H_0) = \alpha$$

Tale metodo non parametrico è anche utile per confrontare due campioni di dati al fine di determinare se provengono dalla stessa popolazione. Si noti anche che la **EDF**(**x**) costruita è anch'essa una distribuzione cumulativa di probabilità.

6.4 Confronto di una misura con il valore di riferimento

6.4.1 Distribuzione t-student

Si consideri un campione di N misure di cui si è calcolata la media campionaria $\bar{x} = \frac{1}{N} \sum x_i$ e si supponga di conoscere σ_i delle singole misure e $E[x] = \mu$, allora la media aritmetica \bar{x} è per il TCL è distribuita come una gaussian $G(\bar{x}, \mu, \frac{\sigma}{\sqrt{N}})$. Come facciamo a dire che \bar{x} e μ sono sufficientemente vicine tra loro rispetto alle incertezze ?

Per determinare la distanza tra le grandezze definiamo la distribuzione t-student data dalla variabile aleatoria:

$$t = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{N}}}$$

definita rispetto al caso descritto nelle righe precedenti. In generale la t-student per un parametro è data da:

$$t = \frac{|\hat{\theta}^* - \theta_t|}{\sigma_{\theta^*}} \quad (6.5)$$

Notare che nel caso in cui si conoscano a priori l'incertezza della misura che si sta confrontando, dunque non si è ottenuta mediante un processo statistico si ha che la pdf(t) é Gaussiana. Se \bar{x} segue una pdf Gaussiana e $Q^2 \sim \chi^2(N-1)$ la distribuzione di t-student ha la seguente forma funzionale.

6.4. CONFRONTO DI UNA MISURA CON IL VALORE DI RIFERIMENTO 9

$$f(t, \nu = N - 1) = \frac{1}{\sqrt{n\nu}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)} \quad (6.6)$$

Proprietà della distribuzione

$$\mu = E[t] = 0 \qquad \gamma_1 = 0$$

$$\sigma^2 = V[t] = \frac{\nu}{\nu-2} \quad \nu > 2 \qquad \gamma_2 = \frac{6}{\nu-4} \quad \nu > 4$$

Per $\nu \rightarrow \infty$ la distribuzione diventa Gaussiana. Si osserva che la pdf della t-student è un po' più larga della distribuzione Gaussiana (fig 6.2).

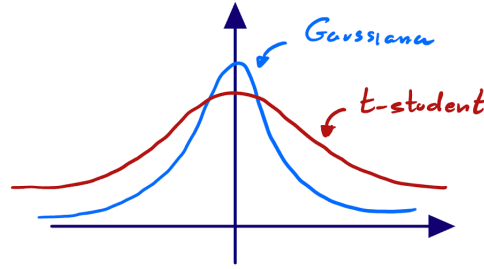


Figure 6.3: Confronto distribuzioe di Gauss e t-student

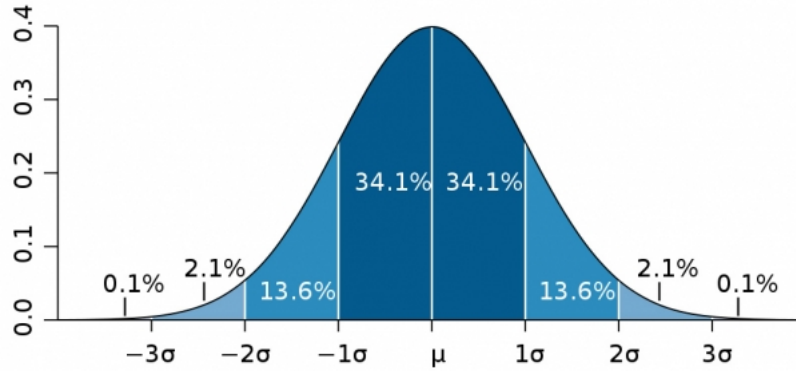
In generale possiamo vedere la distribuzione della t-student come il rapporto tra una Gaussiana normalizzata e la radice di $\frac{\chi^2}{N-1}$.

$$t = \frac{|\bar{x} - \mu|}{\left[\frac{\hat{\sigma}^2}{N}\right]^{\frac{1}{2}}} = \frac{\left[\frac{\sigma^2}{N}\right]^{\frac{1}{2}}}{\left[\frac{\hat{\sigma}^2}{N}\right]^{\frac{1}{2}}} \cdot \frac{\left[\frac{\sigma^2}{N}\right]^{\frac{1}{2}}}{\left[\frac{\hat{\sigma}^2}{N}\right]^{\frac{1}{2}}} = \frac{|\bar{x} - \mu|}{\left[\frac{\sigma^2}{N}\right]^{\frac{1}{2}}} \cdot \left[\frac{\chi^2}{N-1}\right]^{\frac{1}{2}}$$

la parte in azzurro segue la pdf di una Gaussiana normalizzata $N(0,1)$, mentre la parte in rosso segue $\frac{\chi^2}{N-1}$. Dove:

$$\frac{1}{\left[\frac{\sigma^2}{\hat{\sigma}^2}\right]^{\frac{1}{2}}} = \frac{1}{\left[\frac{N-1}{\chi^2}\right]^{\frac{1}{2}}} = \left[\frac{\chi^2}{N-1}\right]^{\frac{1}{2}}$$

Per la t-student fissato un valore di soglia t_0 , ci permette di determinare la compatibilità tra il valore stimato e quello atteso, se $t > t_0$ allora le due misure risultano essere non compatibili tra loro. Dove gli intervalli di compatibilità risultano essere in multipli di deviazioni standard.



Confronto tra la stima di due parametri

La distribuzione di t-student può essere utilizzata non solo per confrontare una misura con un valore atteso, ma anche due stime del medesimo parametro tra loro.

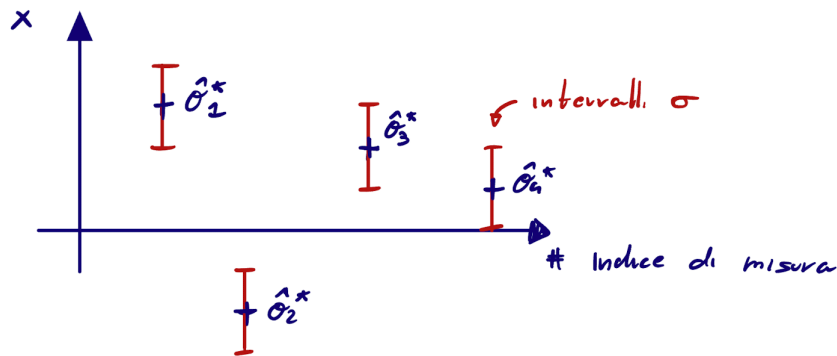
$$z = \frac{|\theta_1^* - \theta_2^*|}{\left[\frac{\hat{\sigma}_1^2}{N} + \frac{\hat{\sigma}_2^2}{N}\right]^{\frac{1}{2}}} \quad (6.7)$$

dove la pdf(z) è Gaussiana se σ_1 e σ_2 sono senza errore, mentre altrimenti segue una distribuzione di t-student con $N-M-2$ gradi di libertà.

6.5 Intervalli di confidenza

Ipotizziamo che lo stimatore $\hat{\theta}$ segue una distribuzione gaussiana $G(\hat{\theta}, \theta_t, \sigma)$ determinato il suo valore θ^* ci domandiamo quale sia la probabilità che tale valore disti σ dal valore vero θ_t , ovvero $P[\theta^* \in (\theta_t - \sigma, \theta_t + \sigma)] =$

0,68. Essendo θ^* una variabile aleatoria dipendente dal campione mentre θ_t no, l'affermazione precedente non è corretta, in quanto θ_t non è una variabile aleatoria, per questo motivo riscriviamo l'intervallo di confidenza in $\theta_t \in (\theta^* - \sigma, \theta^* + \sigma)$ e la probabilità come $P[\theta_t \in (\theta^* - \sigma, \theta^* + \sigma)] = 0,68$, determinando un 68 % di confidenza nell'intercettare θ_t ripetendo l'esperimento.



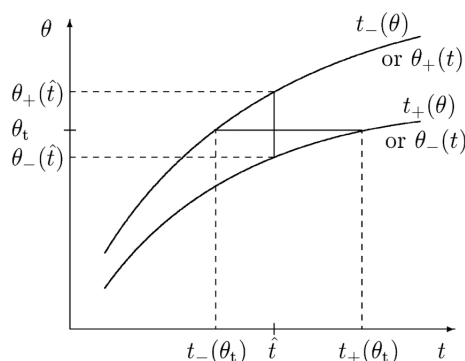
6.5.1 Metodo della cintura di confidenza

Definiamo uno stimatore $\hat{\theta} \equiv \hat{\theta}(x)$, di un parametro θ rispetto ad un campione di N variabili aleatorie IID di una grandezza x , poichè lo stimatore dipende da variabili aleatorie anch'esso è una r.v. di conseguenza seguirà una distribuzione di probabilità $f(\hat{\theta} | \theta)$. Per ciascun campione raccolto $\hat{\theta}$ definirà una stima θ^* del parametro, costruendo la pdf associata. Ripetendo l'esperimento con diversi campionamenti ciascuna distribuzione relativa avrà un valore atteso $E[\hat{\theta}] = \theta_t^i$ e una varianza $V[\hat{\theta}] = \sigma_{\theta_t^i}^2$.

Per ciascun esperimento non conosciamo la forma analitica della pdf oppure non siamo in grado di definirla di conseguenza per stimare l'intervallo di confidenza di un certo valore di θ^* stimato usiamo il seguente metodo:

- Per ogni valore di θ_t^i definiamo $pdf_i(\hat{\theta} | \theta_t)$;

- Si determinano i punti della



$\text{pdf}(\hat{\theta}|\theta_t)$ che definiscono un intervallo di confidenza per il θ_t corrispettivo;

- Si tracciano due rette parallele che attraversano ciascuna i punti estremanti di ciascun intervallo di confidenza, definendo una banda nel piano (θ_t, θ^*) che prende il nome di **Confidence Band**.

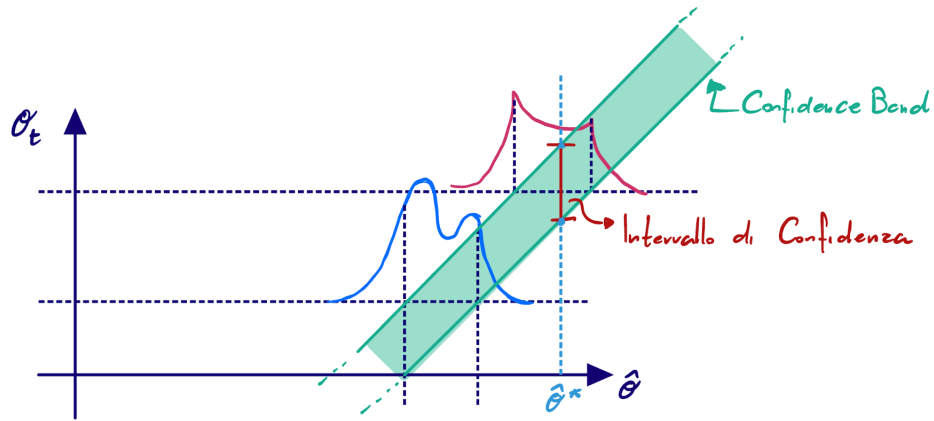


Figure 6.4: Banda di confidenza ed intervallo di confidenza

Quando si effettua una misura si trova θ^* e dove tale valore interseca la confidence band, punti d'intersezione determinano l'intervallo di confidenza di θ_t . Si dimostra che l'intervallo trovato ha copertura uguale a quello scelto per le singole p.d.f.

6.6 Discovery Significance

Consideriamo di rilevare un numero di eventi n_0 in un intervallo di tempo, e che tale numero di eventi può essere distinto in due sotto categorie date da n_s che è il numero di eventi dovuto a un processo fisico e n_b numero di eventi che definiscono il "rumore di fondo" (e rappresenta fenomeni non

legati al processo fisico osservato) dove $n_0 = n_s + n_b$. A priori non abbiamo modo di sapere nel conteggio quanti fenomeni fisici e non compongano n_0 . In compenso conosciamo il numero medio di eventi di entrambi i conteggi $E[n_s] = \nu_s$ e $E[n_b] = \nu_b$.

Costruiamo la nostra null hypothesis H_0 assumendo che del numero complessivo di fenomeni buona parte siano dati dal rumore di fondo; per confutare tale ipotesi utilizzando il p-value è necessario che questo sia più piccolo del valore di soglia p_0 , in fisica delle particelle si sceglie $p_0 = 3 \times 10^{-7}$ per il segnale che una particella sia stata rilevata. La Poissoniana che descrive la probabilità di n_b è data da:

$$Poiss(n_b, \nu_b) = \frac{\nu_b^{n_b}}{n_b!} e^{-\nu_b}$$

la probabilità di misurare un valore n_b di quello misurato è data da:

$$\beta = P(n > n_b^0) = \sum_{k=n_b^0+1}^{\infty} \frac{\nu_b^k}{k!} e^{-\nu_b} = 1 - \sum_{k=0}^{n_b^0} \frac{\nu_b^k}{k!} e^{-\nu_b} \quad (6.8)$$

Possiamo riscrivere l'equazione (6.7) come:

$$1 - \beta = \sum_{k=0}^{n_b^0} \frac{\nu_b^k}{k!} e^{-\nu_b} \approx \int_{2n_b^0}^{\infty} \chi^2(2n_b^0 + 2) d\chi^2 \quad (6.9)$$

approssimabile alla distribuzione del $\chi^2(2n_b^0 + 2)$ con $2(n_b^0 + 1)$ g.d.l., ovvero l'espressione di sinistra della 6.8 coincide con la sua c.d.f. Di conseguenza avremo che il:

$$\text{p-value} = 1 - \beta$$

Per $\beta \approx 1$ il p-value è molto piccolo e quindi possiamo rigettare la null hypothesis H_0 formulata all'inizio e quindi il segnale osservato è effettivamente un processo fisico. Un valore grande di β ci dice che si ha un alta probabilità che per valori più grandi di n_b^0 i fenomeni osservati siano composti per la maggior parte da rumore di fondo. Mentre per $1 - \beta$ molto piccolo si ha una bassa probabilità che quanto osservato sia dato da rumore di fondo.