# Introduzione all'inferenza statistica

Immagina di non avere tutti i dati della popolazione, ma un campione (validato da un estrazione sensata) di esso che però viene trattato come popolazione. Tali osservazioni hanno una famiglia di funzioni di probabilità.
Nel nostro studio supponiamo di conoscere il modello ma non i parametri di esso (si parla di inferenza parametrica), tale modello è identificabile se c'è corrispondenza biunivoca tra i vettori dei parametri e il modello. Le risposte dell'inferenza non sono risposte precise ma ben si sulla correttezza della stima.
Se non si conosce il modello dei dati allora possiamo stimarlo con l'inferenza non parametrica.
**Modello identificabile**: è cosi se c'è corrispondenza biunivoca tra le $\theta$ e la funzione di distribuzione.

# 1 Funzione di massima verosimiglianza

Nel corso tratteremo la **funzione di massima verosimiglianza** che rappresenta la funzione di probabilità del campione, si ipotizzi che essa dipenda dai parametri $\theta$ mentre le $x_i$ sono fisse e conosciute.
Può essere scritta anche come $L(\lambda, x) = p(x, \lambda) = p(X_1 = x_1 ..... X_n = x_n)$ e va da $\lambda \Rightarrow R$
Il nostro obbiettivo è quello di trovare $\lambda$ che massimizzi $L(\lambda, x)$ quindi troviamo lo 0 della derivata prima.
Se $L(\lambda') > L(\lambda'')$ preferiremo il primo in quanto le osservazione $x$ sono generate più verosimilmente da $\lambda'$ (è incorretto dire più probabilmente).
Se $f(x_i, \theta)$ è funzione di probabilità allora se $x_i$ è I.I.D allora la funzione di verosimiglianza è la produttoria delle marginali cioè $\prod_{i=1}^{n} f(x_i, \theta)$.

## 1.1 Proprietà di equivalenza

Se prendiamo un altro campione $Y$ allora si dirà che $L(\theta, x)$ è proporzionale equivalente a $L(\theta, y)$ se $L(\theta, x) = c * L(\theta, y)$ quindi se la loro funzione di massima verosimiglianza differisce di una costante moltiplicativa. I due campioni hanno la stessa informazione su $\theta$ quindi è indifferente quali usare. Lo si può mostrare se ne analizziamo i rapporti cioè $\frac{L(\theta', x)}{L(\theta'', x)} = \frac{c * L(\theta', y)}{c * L(\theta'', y)}$; quindi se scelgo $\theta'$ nella prima lo scelgo anche nella seconda.
**Principio di Verosimiglianza** Dato $E = (f(x_1 ... x_n, \theta); \theta = (\theta_1 ... \theta_n) \in \Theta \in R^k)$ se x e y campioni di osservazioni hanno Verosimiglianze equivalenti allora conducono a conclusioni inferenziali uguali.

## 1.2 LogVerosimiglianza

Per comodità noi però useremo la **LogVerosimiglianza** cioè $l(\theta, x) = log(L(\theta, x))$, inferenzialmente questa trasformazione non cambia nulla è solo più semplice da trattare perchè $l(\theta, x) = \sum_{i=1}^{n} log(f(x_i, \theta))$ quindi è più facile da trattare usando le sommatorie.
In questo caso l'equivalenza si ha se $l(\theta, x) = l(\theta, y) + log(c)$ quindi varia con una costante ed è facilmente identificabile.

## 1.3 Proprietà di Sufficienza

Dato l'insieme $E = (f(x_1 ... x_n, \theta); \theta = (\theta_1 ... \theta_n) \in \Theta \in R^k)$ se vogliamo fare inferenza su $\theta$ ed abbiamo n osservazioni ma vogliamo raccogliere queste osservazioni in una statistica con dimensione $K < n$ tale che l'inferenza mantenga tutte le informazioni del campione, se esiste tale statistica è detta **sufficiente**.

Quindi la statistica $S(x_1 ... x_n) : R^n \Rightarrow R^k$ è sufficiente se $f_x(x, \theta) = f_S(x, \theta) * f_{x|S=s}(x, \theta)$ dove l'ultimo termine non dipende da $\theta$ e quindi può essere considerata una costante e posso scrivere anche come $f_x(x, \theta) = f_S(x, \theta) * f_{x|S=s}(x)$ quindi indica che dopo aver osservato S la funzione marginale perde il collegamento con $\theta$. Quindi è sufficiente se tutta l'informazione è racchiusa in $f_S(x, \theta)$. Ovviamente la statistica sufficiente varia da problema a problema e dipende da cosa vogliamo analizzare.

Il vettore S contenente le statistiche sufficienti ha dimensione almeno pari a $\theta$ il vettore di parametri ma spesso ne ha anche di più.

### 1.3.1 Criterio di fattorizzazione di Newyman Fisher

E' un metodo generale per individuare la statistica sufficiente, se ho una famiglia
$F = (f(x_1...x_n, \theta), \theta = (\theta_1...\theta_n) \in \Theta \in R^k)$ allora la statistica $S(x, \theta)$ è sufficiente se vale il criterio di fattorizzazione quindi se $f_x(x) = k(x) * h(S(x, \theta))$, notiamo come prima che k dipende solo dal campione e h dal campione e dai parametri inferenziali, si può facilmente vedere che l'equazione del paragrafo precedente è una generalizzazione di questa.

### 1.3.2 Statistica sufficiente minimale

Di statistiche sufficienti ne esistono potenzialmente infinite, ma a noi interessa sopratutto quella minimale cioè quelle che rispetta le seguenti condizioni:

- S è sufficiente
- data una statistica S' allora $S(x) = f(S'(x))$, quindi S ha dimensione minore uguale a S'

Questa definizione è fortemente restrittiva perchè non è detto che esista una S di questa forma e inoltre non è operativa quindi ne limita di molto l'utilizzo.

Sappiamo anche che se S è sufficiente e S' e trasformazione biunivoca di S cioè $S'(x) = g(S(x))$ allora S' è sufficiente, invece se la tarsformazione non è biunivoca la statistica S' potrebbe non essere neanche statistica sufficiente.
E' facilmente dimostrabile infatti: $f_x(x) = k(x) * h(S(x, \theta)) = k(x) * h(g^{-1}(S'(x, \theta))) = k(x) * h'(S'(x, \theta))$ che è definizione di statistica sufficiente.

### 1.3.3 Partizionamento dello spazio campionario

La statistica minimale è un modo di partizionamento dello spazio nello specifico quello che accorpa un numero maggiore di osservazioni.
Uno spazio di campionario è l'insieme di disgiunti che compone tutti i valori osservati, il ruolo della statistica sufficiente è di ridurre al minimo questi insiemi disgiunti.
Sappiamo inoltre che le statistiche in corrispondenza biunivoca hanno lo stesso ripartizionamento, quindi esistono più statistiche minimali.

### 1.3.4 Partizione di Verosimiglianza

E' un modo per creare partizionamenti dello spazio campionario e capire se una statistica è minimale oppure no.

Sappiamo che se S è statistica sufficiente e $x = x'$ allora $S(x) = S(x')$ creeranno lo stesso partizionamento, sappiamo che $L(\theta, x) = k(x) * h(S(x), \theta)$, valutiamo il rapporto tra le funzione di verosimiglianza e otteniamo $\frac{L(\theta,x)}{L(\theta,x')} = \frac{k(x)}{k(x')}$ quindi notiamo che se due campioni hanno lo stesso partizionamento allora hanno anche verosimiglianza equivalente che non dipende da $\theta$, il vice versa è garantito esclusivamente nel nel caso cui la statistica sufficiente sia minimale.
Simbolicamente se $x = x'$ allora $S(x) = S(x') \Rightarrow L(\theta, x) = c * L(\theta, x')$
Se la S è minimale allora $S(x) = S(x') <==> L(\theta, x) = c * L(\theta, x')$

### 1.3.5 Statistica sufficiente per tutte le distribuzioni

Se $X = X'$ allora statistica sufficiente per tutte le distribuzione I.I.D. è $S(x) = S(x')$ dove S è ordinamento delle variabili del vettore.
Tale trasformazione anche se mantiene il numero dimensionale del campione, ha una partizione minore di esso, infatti S è una trasformazione non biunivoca del campione.
Quindi per valore la minimalità della statistica operativamente non basta la dimensione del vettore S.

Questa statistica viene usata quando non si riesce in nessuno caso a dividere x da $\theta$, come nel caso di una funzione di couchy.

# 2 Famiglie esponenziali e Tipi di statistiche

Sono particolari famiglie di distribuzioni con prorprietà interessanti per l'inferenza. Ed hanno forma: $F = (f(x,\theta), \theta \in \Theta)$ con $\theta = (\theta_1, ..., \theta_n)$ allora $f(x,\theta) = h(x) * \exp(\sum_{i=1}^{n}(\psi_i(\theta) * k_i(x)) - c(\theta))$.

Dove $h(x)$ deve essere sempre positivo perchè è parte di funzione di densità che per definizione è sempre positiva, ed inoltre possiamo capire il dominio della distribuzione vedendo dove $h(x) > 0$.
Notiamo quindi che il supporto non dipende da alcun parametro $\theta$ in questo tipo di famiglie, perciò se una distribuzione fa dipendere il suo supporta da $\theta$ non appartiene alla famiglia esponenziale come per esempio l'uniforme.

Se abbiamo un campione I.I.D per vedere se la sua distribuzione appartiene alla famiglia esponenziale basta verificare che solo una delle sue marginali ci appartiene e di conseguenze appartiene tutta la distribuzione.

Se invece non è un campione I.I.D bisogna vedere se la sua congiunta è esponenziale.

## 2.1 Statistica sufficiente minimale

Se il nostro modello appartiene alla famiglia esponenziale ed è scritto in forma ridotta allora $T(x) = (k_1(x), ..., k_r(x))$ è statistica sufficiente minimale.
Un modello è scritto in forma ridotta quando non esiste una combinazione lineare si $a * \psi_i(\theta)$ tale che dia una costante se non con a=0.
E' facilmente dimostrabile infatti se prendiamo due campioni x e y abbiamo:

$$\frac{f(x,\theta)}{f(y,\theta)} = \frac{h(x) * \exp(\sum_{i=1}^{n}(\psi_i(\theta) * k_i(x)) - c(\theta))}{h(y) * \exp(\sum_{i=1}^{n}(\psi_i(\theta) * k_i(y)) - c(\theta))} = \frac{h(x)}{h(y)} * \exp(\sum_{i=1}^{n} \psi_i(\theta) * (k_i(x) - k_i(y)) \qquad (1)$$

Qua notiamo subito che l'unico modo per rendere il rapporto non dipendente da $\theta$ è quando le due sommatorie sono uguali. Come detto prima essendo in forma ridotta l'unico modo che sia costante una loro combinazione lineare è se a=0 quindi se le sommatorie sono uguali. Ciò indica che la statistica è sia sufficiente che minimale.

Se abbiamo un campione I.I.D scritto in forma ridotta la sua statistica minimale è
$T = (\sum_{i=1}^{n} k_1(x_i), ..., \sum_{i=1}^{n} k_r(x_i))$.

## 2.2 Statistica completa

Una statistica $T(x)$ è detta completa se vale la seguente proprietà:
$E_\theta(g(T(x))) = 0$ solo se $g(T(x)) = 0$ e nessun'altra funzione.
Questo concetto è molto difficile da applicare.

**Teorema** Se $F \in$ famiglia parametrica esponenziale in forma ridotta allora la statistica $T(x) = (k_1(x), ..., k_r(x))$ è anche completa se $\psi_i(\theta) \in A$ dove A insieme aperto di R ovvero se $\psi_i(\theta)$ oltre che non essere combinazione lineare l'uno dell'altro non sono neanche legati da alcuna funzione. E' auto evidente che se $\psi_i(\theta)$ è di dimensione 1 è automaticamente completa.

# 3 Stimatori puntuali

Uno stimatore è una statistica del campione che abbiamo, scritta come $T(x)$, essa è una quantità aleatoria che stima un determinato parametro $\theta$.
L'ideale sarebbe che la stima coincidesse con $\theta$, ciò però non è realistico, infatti $P(T(x) = \theta) \sim 0$.
Però ha senso valutare un $T(x) - \theta$ più piccolo è meglio è, ciò però non è facile da valutare essendo $T(x)$ variabile aleatoria.
Allora si introduce **errore quadratico medio** ed è $E((T(x)-\theta)^2)$, con questa misura valutiamo quale stimatore è meglio, dati due stimatore $T'$ e $T''$ il migliore dei due sarà quello con errore quadratico medio inferiore.

Inoltre $E((T(x) - \theta)^2) = VAR(T) + (E(T) - \theta)^2$, notiamo quindi che lo stimatore ha due fonti di errore: la variabilità dello stimatore e la distorsione dello stimatore rispetto a $\theta$, questa stima però non basta per valutare la bontà dello stimatore.

## 3.1 Proprietà degli stimatori

**Correttezza**: Uno stimatore si definisce così se $E(T) = \theta$ quindi non ci si sposta sistematicamente dalla media. Questa caratteristica però ha due limiti: Misura la distorsione solo in media e le trasformate di uno stimatore corretto non è detto siano corrette. Inoltre non è condizione ne necessaira ne sufficiente per essere uno stimatore ottimo.
**Correttezza asintotica**: Uno stimatore si definisce così se $\lim_{n\to\infty} E(T_n) = \theta$, essa è più debole della correttezza in generale ma necessario per essere un buon stimatore.
**Consistenza debole**: E' così se $T_n$ per $n \to \infty$ converge in probabilità a $\theta$, quindi vale
$P(|T_n - \theta| > \epsilon) = 0$ per $n \to \infty$.
**Consistenza**: Se $T_n$ per $n \to \infty$ converge in media quadratica a $\theta$, questa convergenza implica quella in probabilità, quindi è più forte. Si può intendere come $E((T_n - \theta)^2) \to 0$ ed è uguale a $VAR(T) + (E(T) - \theta)^2$ perciò se va a 0 anche le componenti andranno a 0.
La consistenza è necessaria perchè uno stimatore sia buono ma non è sufficiente infatti bisogna studiare caso per caso.

Per trovare uno stimatore migliore possiamo valutare l'errore quadratico medio, però spesso esso dipende da $\theta$ e quindi è difficile valutarlo, per definire lo **stimatore ottimo (T)** richiediamo che sia corretto e valutiamo poi EQM(T)<EQM(T').

## 3.2 Score function e informazione attesa di Fisher

Se F famiglia di distribuzione, deve essere identificabile, $\Theta$ insieme aperto, e il supporto indipendente da $\theta$. Se l è la funzione di logVerosimiglianza ed è derivabile allora $l'(x)$ è la score function.
Ha due proprietà importanti:

- $\int \frac{d}{d\theta} f(x,\theta)dx = \frac{d}{d\theta} \int f(x,\theta)dx$ ed essendo $f(x,\theta)$ funzione di densità allora il suo integrale è 1, quindi la derivata di una costante è 0, perciò $\int \frac{d}{d\theta} f(x,\theta)dx = 0$ notiamo anche che questo è anche un valore atteso cioè $E(l'(x,\theta)) = 0$.
  Dimostrabile da $\int \frac{d}{d\theta} f dx = \int f' * \frac{f}{f} dx = \int (\frac{d}{d\theta} * log(f(x))) * f(x,\theta)dx = \int l'(x) * f(x,\theta)dx = E(l'(x,\theta))$.

- $\int \frac{d^2}{d\theta^2} f(x,\theta)dx = \frac{d^2}{d\theta^2} \int f(x,\theta)dx$ valendo questo e che il momento primo dello score function è 0 allora vale $Var(l'(x,\theta)) = E(l'(x,\theta)^2) = -E(l''(x,\theta)) = I_x(\theta)$.
  Dove $I_x(\theta)$ è l'informazione attesa di fisher.
  Dimostrabile so che: $0 = \frac{d}{d\theta} E(l'(x,\theta)) = \int \frac{d}{d\theta} l'(x,\theta) * f(x,\theta) = \int l''(x,\theta) * f(x,\theta) + \int l'(x,\theta) * f'(x,\theta) = E(l''(x,\theta)) + E(l'(x,\theta)^2)$ quindi vale l'affermazione precedente.

Ciò vale anche nel caso multi parametrico solamente che al posto della derivata si ha la derivata parziale. Il valore atteso è un vettore di zeri, e la varianza una matrice varianza-covarianza formata dal valore atteso della matrice hessiana di $l(\theta)$.

## 3.3 Disuguaglianza di Cramer-Rao

La disuguaglianza di Cramer-Rao dice che $Var(\tau(x)) \geq \frac{(\tau'(x))^2}{I_x(\theta)}$, uno stimatore ottimale per esserlo ha come statistica sufficiente che la disuguaglianza diventa uguaglianza, ovviamente dopo aver controllato la correttezza. Potrebbe essere ottimo anche se non raggiunge il limite inferiore però è più difficile da cercare. Se si stima direttamente $\theta$ allora $Var(\theta) \geq \frac{1}{I_x(\theta)}$. Notiamo quindi che più alta è l'informazione attesa di fisher migliore sarà la stima su $\theta$.

Se ho due stimatori corretti $T_1$ e $T_2$ per decidere quale è il migliore se valuta la varianza di esso, se la varianza è un valore è facile prendere quello con la varianza minore. Se invece si ha una matrice di varianza-covarianza allora per valutare quale è lo stimatore migliore si fa la differenza matriciale tra $Var(T_1) - Var(T_2)$ devo vedere se tale matrice è semi definita positiva, se tale allora preferisco $T_2$ se no $T_1$.
Tale vale anche per le sue combinazione lineari $Var(c'T) = c'Var(T)c$.

Per valutare se è uno stimatore ottimale vedo se $Var(T) - I_x^{-1}(\theta)$.

## 3.4 Teorema di Rao Blackwell

Sia F una famiglia di distribuzione, su di essa non si fa nessuna restrizione.
Dobbiamo studiare $T(\theta) : \Theta \to R$.
Sia S(x) è una statistica sufficiente per $\theta$, e $E(T(x)) = T(\theta)$ per qualsiasi $\theta \in \Theta$ quindi è corretto.
Se costruiamo un nuovo stimatore $h(S(x)) = E(T|S)$ allora H è uno stimatore corretto e varrà: $Var(h(S(x))) \leq Var(T)$.
Per ottenere il miglior stimatore bisogna prendere la statistica sufficiente minimale, quindi uno stimatore ottimo è sempre funzione della SSM. Tale teorema però non afferma l'esistenza di esso.

Sappiamo che è corretta perchè $E(h(S(x))) = E(E(T|S)) = E(T) = T$.
Ora calcoliamo $Var(T) = E(Var(T|S)) + Var(E(T|S)) = E(Var(T|S)) + Var(h(S(x)))$ perciò $Var(h(S(x))) \leq Var(T)$.
Se $E(Var(T|S)) = 0$ allora $T = S$ quindi T era già funzione di statistica sufficiente.

## 3.5 Teorema di Lehman Schiefe

Se S è una statistica sufficiente minimale completa e T=f(S) è uno stimatore corretto funzione di S , allora T è unico, inoltre essendo funzione di statistica minimale per il teorema precedente T è anche stimatore ottimo.

Supponiamo che esista un $T_1 \neq T$ e che sia corretto quindi $E(T_1) = T$ se creiamo la funzione $g = T - T_1$ tale funzione è per forza una statistica completa, quindi $E(T - T_1) = 0$ solo se $T - T_1 = 0$ quindi $T = T_1$ sempre perciò non esiste un $T_1 \neq T$.

## 3.6 Stimatori di Massima Verosimiglianza

Possiamo trovare anche lo stimatore che massimizza la funzione di verosimiglianza.
Infatti trovo prima la funzione di verosimiglianza ($L(\theta)$) la trasformo in quella di Logverosimiglianza ($l(\theta)$), di questa ne trovo la derivata prima e la pongo a zero $l'(\theta) = 0$ trovo quindi $\theta^*$, vedo se $l''(\theta) > 0$ (o semi definita positiva se è una matrice), se si allora il $\theta$ trovato è quello che massimizza la funzione d verosimiglianza.
Operativamente posso dire che se:

- $\Theta$ è un insieme aperto.
- $l(\theta) \to -\infty$ nei punti di frontiera di $\Theta$, quindi il suo primo e ultimo valore.
- C'è un unica soluzione di $\theta$ che soddisfa l'equazione.

Allora il $\theta$ trovato è stimatore di massima verosimiglianza.

### 3.6.1 Proprietà esatte

**Funzione di statistica sufficiente** Se S è S.S. per $\theta$ allora lo stimatore di massima verosimiglianza è funzione di S quindi $\theta^* = f(s)$. Si dimostra infatti se S è sufficiente allora $f(x,\theta) = L(\theta,x) = k(x) * g(S((x),\theta)$, quindi per massimizzare L devo massimizzare g che a sua volta dipende da S.
**Equivarianza**: Se vogliamo stimatore $\theta$ e $\psi = g(\theta)$ con g biunivoca, ed abbiamo per $\theta$ lo stimatore di massima verosimiglianza allora la stima per $\psi$ è $g(\theta^*)$.

### 3.6.2 Proprietà asintotiche

Se X sono I.I.D, se il modello è identificabile e la verosimiglianza è regolare valgono.
**Consistenza** Se lo stimatore è SMV è automaticamente consistente.
**Distribuzione asintotica**: $\sqrt{n}*(\theta^*-\theta) \to N(0, \frac{1}{I_{x_i}}(\theta))$ in distribuzione. Quindi $\theta^* \to N(\theta, \frac{1}{n*I_{x_i}}(\theta))$. Da qua si vede che lo stimatore di massima verosimiglianza come sapevamo è corretto e ottimo asintoticamente, ed ha velocità di riduzione della varianza pari a $I_{x_i}$ quindi più alto è meglio è.

Si può ricavare sapendo che la score function è una variabile con media 0 e varianza $I_x$ quindi $\frac{-l'(\theta)}{\sqrt{n}} \to N(0, I_x(\theta))$. Se sviluppo arrestandomi al primo ordine con Taylor ottengo che $l'(\theta^*) = l'(\theta) + (\theta^* - \theta) * l''(\theta) + R$ sapendo però che $l'(\theta^*) = 0$ ottengo $-l'(\theta) = (\theta^* - \theta) * l''(\theta) + R$, se moltiplichiamo da entrambe le parte $\sqrt{n}$ otteniamo che $\frac{-l'(\theta)}{\sqrt{n}} = (\theta^* - \theta) * \frac{l''(\theta)}{n} * \sqrt{n}$ sappiamo che la parte destra tende in probabilità a $N(0, I_x(\theta))$ quindi anche quella sinistra, ora dividiamo per $\frac{l''(\theta)}{n}$ ed otteniamo che $(\theta^* - \theta) * \sqrt{n} \to \frac{N(0,I_x(\theta))}{I_x(\theta))} = N(0, I_x(\theta)^{-1})$

Sappiamo che l'informazione osservata di fisher ($Ioss_x(\theta)$) è $= -l''(x,\theta)$ e quindi $E(Ioss_x(\theta)) = I_x(\theta)$. Per la legge dei grandi numeri se n tende ad infinito allora $\frac{\sum Ioss_{x_i}(\theta)}{n} \to I_x(\theta)$ in probabilità, quindi se $n \to \infty$ allora $\theta^* \to N(\theta, \frac{1}{Ioss_x(\theta)})$ in probabilità.
Notiamo quindi un importante differenza che l'informazione osservata è una quantità che dipende dal campione mentre quella attesa è deterministica. Se è una famiglia esponenziale questi due valori coincidono.

Notiamo inoltre che se $\theta^*$ stimatore di massima verosimiglianza e proviamo a sviluppare taylor per $l(\theta) = l(\theta^*) + (\theta - \theta^*)l'(\theta^*) + \frac{(\theta-\theta^*)^2}{2}l''(\theta^*)$ noi sappiamo che $l'(\theta^*) = 0$ e -$l''(\theta^*) = Ioss_x(\theta)$ allora vediamo che $l(\theta) = l(\theta^*) + \frac{(\theta-\theta^*)^2}{2} * (-Ioss_x(\theta))$ perciò vediamo che ha forma parabolica concava e la sua ampiezza dipenderà da quanto è grande l'informazione osservata, infatti maggiore è essa meglio è, visto che è più stretta è la parabola.

Ovviamente tutte queste proprietà sono estendibili anche al caso multi parametrico.

### 3.6.3 Verosimiglianza profilo

Se ho un insieme di parametri del tipo $\theta = (\psi, \lambda)$ ma a noi interessa fare inferenza solo su $\psi$ posso costruire la verosimiglianza profilo cioè fissato $\psi$ trovo $l_p(\psi) = sup_\lambda l(\psi,\lambda) = l(\psi, \lambda_{\hat{\psi}})$. In questo modo trovo il valore massimo per il parametro di non interesse, ciò non porta al cambiamento di alcun risultato solo di semplicità dei calcoli in particolare nel test d'ipotesi.

## 3.7 Metodo delta

Se definisco un nuovo parametro $\psi = f(\theta)$ con f biunivoca allora vale anche $\hat{\psi} = f(\hat{\theta})$. Se $r_n \to 0$ allora $\frac{1}{r_n}(x_n - c) \to N(0, \sigma^2)$ in distribuzione, se conosco la distribuzione di $x_n$ conosco anche $f(x_n)$. Posso dire che $\frac{1}{r_n}(f(x_n) - f(c)) \to z * f'(c)$ in distribuzione, quindi basta che f'(c) esista e sia diverso da 0 perchè questo metodo valga.
Ovviamente questo ragionamento vale anche per $R^n$ quindi si estende multi dimensionalmente.
Se invece $f'(c) = 0$ uso il secondo metodo delta che è ancora più efficiente del primo infatti so che $\frac{1}{r_n^2}(f(x_n) - f(c)) \to \chi^2 * \sigma^2/2 * f''(c)$ questo però vale solo se z è una normale.

# 4 Verifica di Ipotesi

Data una famiglia $F = (f(x,\theta), \theta \in \Theta)$, noi vogliamo capire le spazio parametrico è stato creato da $\Theta_1$ o $\Theta_0$ per farlo utilizziamo un test.
Identifichiamo un ipotesi nulla $H_0 : \theta \in \Theta_0$ e ipotesi alternativa $H_1 : \theta \in \Theta_1$, in questo modo creeremo una regione di accettazione di $H_0$ e una di rifiuto, per collocare i nostri campioni in una di queste due regioni usiamo la statistica test T(x). (Esistono anche test casuali che creano una regione di incertezza)
Nella creazione di questi test si incontrano due tipi di errore:

- Errore di 1° tipo, cioè che $H_0$ è vera ma si rifiuta il test. Ed ha probabilità $P(x \in R; \theta) = \alpha(\theta)$ ma $\theta \in \Theta_0$ dove R regione di accettazione.
- Errore del 2° tipo, cioè $H_1$ è vera ma si accetta il test. Ed ha probabilità $P(x \in A; \theta) = \beta(\theta)$ ma $\theta \in \Theta_1$ dove A regione di rifiuto.

Per trovare un test ideale si dovrebbe trovare uno tale minimizzi entrambi gli errori.

Nella costruzione reali di Test impostiamo un valore $\alpha$ che siamo disposti ad accettare, e a parità di errore del 1° tipo scegliamo il test con il valore di $\beta$ uniformemente più basso.
Tale valutazione può essere fatta sfruttando la funzione di potenza.

## 4.1 Funzione di potenza

E' una funzione di $\theta$ costruita come $\pi(\theta) = P(x \in R)$:

$$\begin{cases} \alpha(\theta) \textbf{ se } \theta \in \Theta_0 \\ 1 - \beta(\theta) \textbf{ se } \theta \in \Theta_1 \end{cases} \tag{2}$$

In questo caso preferirò la funzione di potenza maggiore sistematicamente nell'intervallo $\theta \in \Theta_1$ per decidere quale test scegliere.

## 4.2 Test di verosimiglianza

Si può creare un test con la verosimiglianza, infatti fissato un livello $\alpha$ di errore del primo tipo, posso creare un test $\lambda(x) = \frac{L(\theta_0,x)}{L(\theta_1,x)}$, più basso sarà più la differenza di spiegazione a favore del modello $\Theta_1$ sarà maggiore. Perciò creo una regione di rifiuto R=$(x : \lambda(x) < k)$ dove K viene deciso in base all'alpha scelto, infatti $P(\lambda(x) < k, \Theta_0) = \alpha$. Il test che troviamo con P$(\lambda(x) < k; \theta_0)$=$\alpha$ è detto test ottimo, questo deriva dal **Lemma di Neyman Pearson**.

Operativamente noi usiamo il test di Wald. E' definito come $w(x) = -2log(\lambda(x)) = 2(log(\hat{\theta}) - log(\theta_0)) > k$ ha come regione di rifiuto $R : (w(x) > k)$.
Questo nuovo test è molto utile infatti sotto $H_0$ il $\lim_{n\to\infty} w(x) \to \chi^2_m$ dove m i gradi di libertà sono uguali al numero dei parametri vincolati in $H_0$, sapendo la distribuzione possiamo calcolare facilmente K sapendo che $P(w(x) > k) = \alpha$ quindi $k = \chi^2_{m_{1-\alpha}}$.

## 4.3 P-value

Il P-value è una misura che ci dice quanto è compatibile il valore t(x) con la distribuzione sotto l'ipotesi $H_0$, quindi misura quanto è anomala l'osservazione, infatti esso misura la probabilità di trovare dei valori più grandi rispetto a quello osservato. Esso è una funzione casuale del campione x definita come $P(W > w_{oss}, \theta_0)$.
Accetto $H_0$ se P-value è maggiore di $\alpha$ e rifiuterò $H_0$ quando è più basso.
Un p-value molto grande o piccolo non ci dice di quanto divergono due misura ma semplicemente che divergono. Inoltre al crescere di n-osservazioni anche la minima differenza sui valori presi in esame da un p-value bassissimo.
Ecco perchè il modo operativo corretto per analizzare è usando $H_0 = |\mu_0 - \mu_1| < \epsilon$. Purtroppo quest'ultimo è di difficile implementazione visto che è difficile creare il test. Ecco perchè spesso si fa ricorso agli intervalli di confidenza.

# 5 Intervalli di confidenza

Attraverso questo metodo voglio capire quali siano i valori reali plausibili per i nostri parametri incogniti, ed in base al campione che abbiamo associamo ad essi un intervallo. So che $P(\theta \in c(x)) = 1 - \alpha$ con c(x) intervallo di confidenza.

Esso può essere interpretato come un inversione del test di ipotesi. Sia $H_0 = \theta = \theta_0$ ed esso è associato una regione di rifiuto, con relativo $\alpha$ errore. Allora c(x) è costruito come l'intervallo dei $\theta$ che accetti sapendo che $\theta_0$ è il valore reale, sappiamo che ad ogni $\theta$ è associto un certo campione x, quindi se $\theta \in c(x)$ è come dire che $x \in A_{\theta_0}$.

## 5.1 Metodo Verosimiglianza

Per costruire tale intervallo prenderemo i $\theta$ con verosimiglianza più alta, la soglia di alto lo decido con un test di ipotesi. L'inversione si può fare partendo dal test W, sia $c(x) = (w(x,\theta) < K_\alpha) = (2(l_p(\hat{\theta}) - l_p(\theta)) < K_\alpha) = (l_p(\theta) > l_p(\hat{\theta}) - \frac{K_\alpha}{2})$.

**Approssimazione di massima verosimiglianza** so che se $\hat{\theta}$ è SMV allora si dispone come $N(\theta, I_{nx}^{-1})$ da qua è facile costruire il suo intervallo di confidenza come $\hat{\theta} \pm Z_{1-\alpha/2} * \sqrt{I_{nx}^{-1}}$, questa è un'approssimazione molto forte infatti rende gli intervalli non simmetrici in simmetrici.

# 6 Model selection and fitting

Le inferenze fatte fino ad ora sono corrette se si utilizza una particolare famiglia parametrica. Ma per verificare ciò bisogna:

- Fare un test $\chi^2$ per vedere l'adattamento dei dati campionari a quelli teorici.
- Aggiungere eventuali parametri da stimare così da migliorare l'accuratezza del modello, e confrontare i modelli creati avendo come ipotesi nulla che il modello base è corretto. Oppure se ho due modelli uno più semplice e uno più complesso posso pensare ad una media pesata delle stime dei modelli.
- Verifica dei residui.

## 6.1 Criterio di scelta dei modelli

Date due famiglie parametriche g(x) (vera famiglia di appartenenza) e $f_\theta(x)$ (famiglia empirica) vogliamo valutare la differenza tra le due ed useremo la divergenza di
Kullback-Leibler KL=$\int log(\frac{g(x)}{f_\theta(x)}) * g(x) = E_g(log(\frac{g(x)}{f_\theta(x)}))$.
Sappiamo che se $g = f_\theta$ lo stimatore è consistente. Io cercherò quindi quando $g = f_\theta$ perciò il valore minimo di KL.
Se vogliamo aggiungere modello il parametro stimato prenderò $E_g(log(\frac{g(x)}{f(x,\hat{\theta})})$, dove $\hat{\theta}$ stima di $\theta$, dovrò minimizzare questo valore atteso. So che è equivalente a scrivere $E_g(log(g(x)) - E_g(f(x,\hat{\theta}))$, dove la prima parte è una costante e la seconda una parte da massimizzare.

Da qua si crea l'AIC cioè un statistica che ci indica che modello scegliere infatti più alto è meglio è, ed è definito come $2(l_m(\hat{\theta}) - P_m)$, dove $l_m$ log-verosimiglianza del modello e $P_m$ parametri del modello. Notiamo quindi che penalizza i modello con molti parametri.
L'AIC è consistente, cioè se abbiamo diverse famiglie parametriche, ed $F_1$ è quella reale, se $n \to \infty$ allora $P(AIC$ **scelga** $F_1 | F_1$ **vera** $) = 1$.

Se invece le famiglie sono annidati per esempio $F_1 \subseteq F_2$, allora l'aic non è asintoticamente consistente infatti $P(l_p(\hat{\theta}_p) - P_m > 2l_{p^*}(\hat{\theta}_{p^*}) - P_m^*) = P(l_p - l_{p^*} > p - p^*) > 0$ e si dispone come una chi quadro con $p - p^*$ gradi di libertà perciò ha una probabilità positiva di sbagliare.
Useremmo il BIC= $2l_m(\hat{\theta}) - P_m * log(n)$, dove n è il numero delle osservazioni, quindi penalizza ancora di più i modelli complessi.