INFERENZA STATISTICA: Stima dei parametri e test delle ipotesi

Graziano Venanzoni

INFN/LNF

Incontri di Fisica

LNF 2 Ottobre 2007

E-mail: venanzoni@lnf.infn.it

Indice

- Introduzione al problema dell'inferenza statistica
- Esempio di impostazione bayesiana e frequentista del problema
- Definizione della funzione di verosimiglianza (Likelihood)
- Metodo della stima di massima verosimiglianza (ML)
- Soluzione di un problema con "stima" di ML
- Livello di confidenza
- Test delle ipotesi
- Esempio di un problema di test delle ipotesi: soluzione bayesiana
- Esempio di un problema di test delle ipotesi: soluzione frequentista

Nozioni richieste:

- Valore atteso e varianza di una variabile aleatoria
- Distribuzioni binomiale, poissoniana, gaussiana
- ….

INFERENZA STATISTICA:Stima dei parametri e test delle ipotesi

- Spesso ci troviamo in situazioni in cui da una o piu' osservazioni di un fenomeno fisico cerchiamo di ricavare delle informazioni sul modello (probabilistico) che utilizziamo per descriverlo
  - Es: dall'osservazione di n lanci di una stessa moneta che conclusioni possiamo trarre sulla moneta stessa?

- Questo tipo di informazioni possono riguardare sia: *(i)* la determinazione o stima dei parametri (del modello), che *(ii)* la bonta' delle ipotesi scientifiche su cui e' costruito il modello stesso. In questo secondo caso, ci si trova a dover scegliere tra due (o piu') ipotesi alternative

- Sia nel caso della stima dei parametri che nel test delle ipotesi e' necessario assegnare una quantita' (es. livello di confidenza, C.L.) che caratterizzi l'accuratezza della stima/test

- Il modello deve descrivere adeguatamente il fenomeno che stiamo trattando, in modo da passare dalle osservazioni ai parametri del modello (funzione di verosimiglianza)

Esempi (parametri Ipotesi):

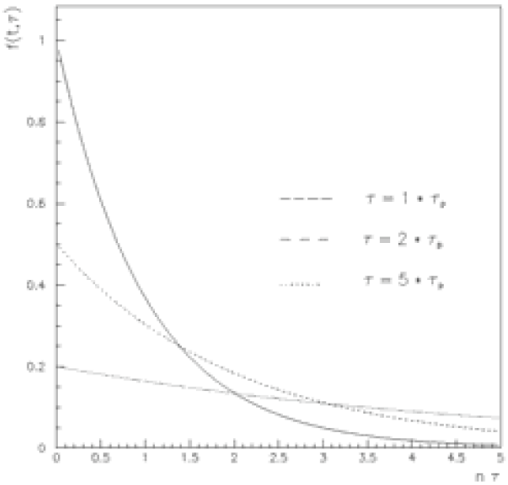
- In un apparato sperimentale si contano i tempi di decadimento del leptone μ. Qual e' la vita media del μ?
- In una misura di "calibrazione", qual e' l'efficienza di un rivelatore?
- Si osservano x globuli bianchi in un campione di sangue. Qual e' la densita' di globuli bianchi nel sangue?
- Si osservano l'evoluzione di alcune azioni finanziarie. Qual e' la probabilita' di guadagno delle azioni?
- Da una o piu' osservazioni di una specie animale, qual e' la popolazione della specie?
- Come si stima la percentuale di gradimento in un sondaggio?

- Dall'osservazione di ripetuti lanci di una moneta, quanto e' equilibrata la moneta?
- La particella che si e' rivelata e' un leptone o un mesone?
- Quanto i risultati di un esperimento sono in accordo con una certa predizione (per esempio il Modello Standard)? Al CERN si e' scoperto o no il bosone di Higgs?
- In un processo, la persona e' innocente o colpevole?
- In un esperimento ESP la persona ha poteri telepatici?

Esempio

- La vita media τ di una particella instabile determina la probabilita' dei diversi tempi di decadimento:

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

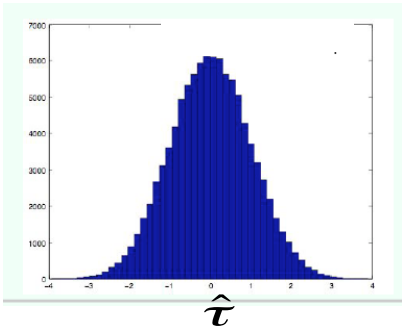


- $f(t,\tau)$ e' la funzione della densita' di probabilita' (o p.d.f), in quanto moltiplicata per l'intervallo di tempo infinitesimo dt da' la probabilita' che la particella decada tra t e t+dt
- L'osservazione di un certo numero di tempi di decadimento t_i, da' informazioni sul parametro τ

- Impostazione bayesiana del problema:
  - Prima dell'osservazione ho attribuito una certa probabilita' "a priori" ai diversi valori di τ. Come devo cambiare questa probabilita' per tenere conto dell'osservazione?:

$$f(\tau \mid t) \propto f(t \mid \tau) f_0(\tau)$$

- Impostazione frequentista del problema:
  - Posso costruire una funzione $\hat{\tau}$ delle osservazioni t_i che costitusce una stima della vita media. L'unico problema che mi posso porre e':qual e' la distribuzione di probabilita' di $\hat{\tau}$?
  - Esistono modi ingegnosi per stimare intervalli che hanno la stessa probabilita' di includere il valore vero, qualunque esso sia.



Funzione di Verosimiglianza (Likelihood)

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

- La funzione $f(t,\tau)$ e' una p.d.f. nella sua dipendenza dalla variabile (misurata) t, una volta fissato il parametro τ

- La stessa funzione pero' puo' essere vista nella sua dipendenza da τ, per fissati valori di t.Questo e' il caso che si utilizza nella statistica inferenziale; la funzione $f(t,\tau)$ viene chiamata funzione di verosimiglianza (likelihood) e si indica di solito con $f(t|\tau)$ o $L(\tau)$.

  - Oss: $L(\tau)$ non e' una p.d.f di τ

- Generalizzazione: se facciamo n misure di t $(t_1,t_2,\ldots,t_n)$ tra loro indipendenti e ripetute nelle stesse condizioni sperimentali:

$$L(\tau) = f(t_1,t_2,\ldots,t_n \mid \tau) = \prod_{i=1,n} f(t_i \mid \tau) = \prod_{i=1,n} \frac{1}{\tau} e^{-t_i/\tau}$$

Stima dei parametri

- Esistono vari modi per stimare un parametro θ: massima verosimiglianza (ML), minimi quadrati, ecc…

- La stima ottenuta da ciascun metodo si caratterizza per: (a) consistenza; (b) bias; (c) efficienza e (d) robustezza

- In questa lezione ci occuperemo del metodo di ML

Metodo di massima verosimiglianza (ML)

- La stima $\hat{\theta}$ e' quel valore che massimizza la funzione di verosimiglianza L(θ):

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

- Interpretazione intuitiva:
  - "Escludo di essere stato sfortunato": la miglior stima che posso fare e' quella per la quale la probabilita' dell'osservazione che ho fatto e' la piu' grande possibile.

- Ogni fattore costante nella likelihood e' irrilevante

- Buone proprieta' della ML:
  - Asintoticamente (ossia per un numero di osservazioni $\rightarrow\infty$) efficiente, priva di bias, ha una distribuzione gaussiana

  - E' invariante per trasformazione dei parametri:

    Se g_{ML} e' la stima di ML di g e θ_{ML} e' la stima di ML di θ:

    $g_{ML} = g(\theta_{ML})$

$$\frac{\partial L}{\partial \theta} = \left(\frac{\partial L}{\partial g}\right)_{g=g(\theta)} \frac{\partial g}{\partial \theta}; \quad \frac{\partial L}{\partial \theta} = 0 \Rightarrow \left(\frac{\partial L}{\partial g}\right)_{g=g(\theta)} = 0$$

Metodo di massima verosimiglianza

- In pratica e' piu' comodo usare il logaritmo della likelihood, specialmente in caso di *n* misure.
  - Es: stima della vita media di una particella elementare *instabile*:

$$\ln f(t_1, t_2, \ldots, t_n \mid \tau) = \sum_{i=1,n} \ln f(t_i \mid \tau) = -\sum_{i=1,n} \ln\left(\tau + \frac{t_i}{\tau}\right)$$

$$\frac{\partial}{\partial \tau} \ln f(\vec{t} \mid \tau) = \frac{n}{\tau} - \sum_{i=1,n} \frac{t_i}{\tau^2} = 0 \Rightarrow \hat{\tau} = \sum_{i=1,n} \frac{t_i}{n}$$

- Che errore associare alla stima $\hat{\tau}$?

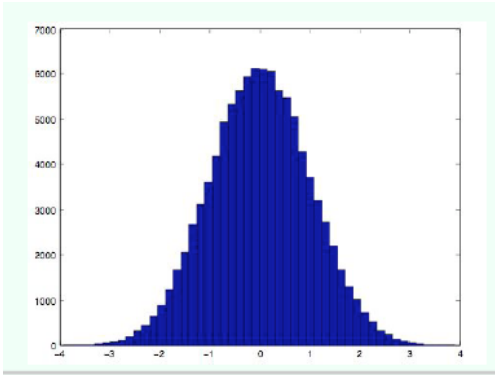
Metodo di massima verosimiglianza

- Che errore associare alla stima $\hat{\tau}$?

  - Possiamo riformulare la domanda nel seguente modo: se ripetiamo l'esperimento N volte, ognuno con n misure t_i, ogni esperimento darebbe una stima di τ differente. Quanto e' larga la distribuzione di $\hat{\tau}$?

$$\sigma_{\hat{\tau}}^2 = E(\hat{\tau}^2) - E(\hat{\tau})^2 = \frac{\hat{\tau}^2}{n}$$

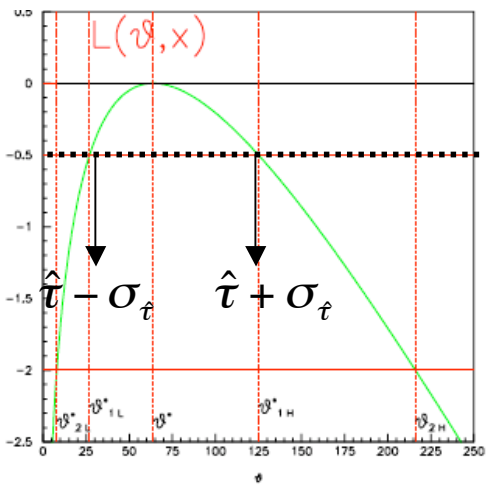
  - Nel limite di n grande, σ_τ rappresenta la probabilita' del 68% (approx. gaussiana)



- Esiste un altro modo (grafico) per ricavare l'errore statistico associato alla stima: la deviazione standard di $\hat{\tau}$ si trova facendo decrescere LnL di 0.5 (approx. limite gaussiano):

$$\ln L(\hat{\tau} \pm \sigma_{\hat{\tau}}) = \ln L(\hat{\tau}) - 0.5$$

  - Nel caso di n grande, $\sigma_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$



Stima di ML e stima bayesiana

- Nell'approccio bayesiano si puo' dare un significato piu' precisio alla espressione "escludo di essere stato sfortunato"

- La stima bayesiana piu' naturale e' quella che massimizza la probabilita' a posteriori di τ:

$$\frac{\partial f(\tau \mid t)}{\partial \tau} = 0 \qquad f(\tau \mid t) = \frac{f(t \mid \tau) f_0(\tau)}{\int d\tau f(t \mid \tau) f_0(\tau)}$$

- Nell'ipotesi di ignoranza iniziale espressa da una densita' di probabilita' a priori costante ($f_0(\tau)$=const) la stima bayesiana coincide con la stima di ML

- Ma attenzione:
  - La stima bayesiana dipende dall'ipotesi di probabilita' a priori
  - Non e' invariante (in generale) per trasformazione dei parametri (diversamente dalla ML)

Esempio di applicazione della ML

- Si vuole misurare la densita' d di leucociti (L) nel sangue e la frazione r di una loro varieta', i linfociti (l) ,osservando uno "striscio" con un microscopio ottico.

- Inizialmente si contano 25 L, di cui 9 l, in $V=5x10^{-3}$ mm^3:
  1) determinare la stime di d e r.
  2) determinare il livello di confidenza (C.L.) al 68% per r
- Per migliorare la precisione, si osserva un altro campione contando l'=13
  3) Determinare come cambiano le stime di d e r

$$P(l,L) = P(l \mid L) \cdot P(L) = \underbrace{\binom{L}{l} r^l (1-r)^{L-l}}_{P(l|L)} \cdot \underbrace{\frac{e^{-\mu}\mu^L}{L!}}_{P(L)}$$

p.d.f.

P(l,L) = probabilita' di osservare L leucociti e l linfociti nel campione di sangue. Non sono eventi indipendenti in quanto i linfociti sono leucociti.
P(L) = probabilita' di osservare L leucociti nel campione di sangue (poissoniana)
P(l|L) = probabilita' di osservare l Infociti condizionata ad avere osservato L leucociti (binomiale). Ciascun leucocita ha una probabilita' r di essere linfocita.

Esercizio di ML

- Risolviamo il problema. La p.d.f. scritta da' la likelihood:

$$\frac{\partial}{\partial \mu} \ln L(l,L \mid r,\mu) = 0 \Rightarrow \hat{\mu} = L \Rightarrow \hat{d} = \frac{L}{V}$$

$$\frac{\partial}{\partial r} \ln L(l,L \mid r,\mu) = 0 \Rightarrow \hat{r} = \frac{l}{L}$$

- Rimandiamo a dopo il punto 2. Facciamo ora il punto 3:

$$P(l') = \sum_L P(l' \mid L) P(L) = \sum_L \binom{L}{l'} r^{l'} (1-r)^{L-l'} \frac{e^{-\mu}\mu^L}{L!} = \frac{e^{-(r\mu)}(r\mu)^{l'}}{l'!}$$

$$P(l,L,l') = P(l,L) \cdot P(l') = \underbrace{\binom{L}{l} r^{l} (1-r)^{L-l} \frac{e^{-\mu}\mu^L}{L!}}_{P(l,L)} \cdot \underbrace{\frac{e^{-(r\mu)}(r\mu)^{l'}}{l'!}}_{P(l')}$$

In questo caso la likelihood e' il prodotto delle probabilita' di ciascuna osservazione

Esercizio di ML

- Otteniamo quindi:

$$\frac{\partial}{\partial \mu}\ln L(l,L,l' \mid r,\mu) = 0 \Rightarrow \hat{\mu} = L - \frac{2}{3}l + \frac{l'}{3} \Rightarrow \hat{d} = \frac{\hat{\mu}}{V}$$

$$\frac{\partial}{\partial r}\ln L(l,L \mid r,\mu) = 0 \Rightarrow \hat{r} = \frac{l + l'}{3L - 2l + l'}$$

- Ricordiamo come si puo' risolvere il quesito con un approccio bayesiano:

$$P(\mu,r \mid D_2,D_1)\cdot P(D_2 \mid D_1) = P(D_2 \mid \mu,r,D_1)\cdot P(\mu,r \mid D_1)$$

$$P(\mu,r \mid D_2,D_1) \propto \underbrace{P(D_2 \mid \mu,r)}_{Likelihood}\cdot \underbrace{P(\mu,r \mid D_1)}_{prob.a\ posteriori}$$

- Abbiamo assunto D2,D1 indipendenti. Assumiamo "priors" uniformi:

$$P(\mu,r \mid D_2,D_1) = \frac{e^{-(r\mu)}(r\mu)^{l'}}{l'!}\cdot\binom{L}{l}r^{l}(1-r)^{L-l}\frac{e^{-\mu}\mu^{L}}{L!}$$

Che (come aspettato) coincide con la stima di ML (ma sarebbe stato diverso per una differente scelta di P_0)

Intervalli di confidenza I

- Nella stima di un parametro θ, occorre dare un intervallo che rifletta la precisione statistica della misura.
  - Un modo per esprimere il risultato, e' quello di dare l'intervallo $\hat{\theta} \pm \sigma_{\hat{\theta}}$. Se $f(\hat{\theta})$ e' gaussiana, l'intervallo $[\hat{\theta} - \sigma_{\hat{\theta}}; \hat{\theta} + \sigma_{\hat{\theta}}]$ contiene il 68% di probabilita'
- Se la *p.d.f.* non e' gaussiana non e' ovvio che probabilita' rappresenti $\hat{\theta} \pm \sigma_{\hat{\theta}}$
- In questo caso si usa "quotare" un intervallo (di confidenza) a cui sia associata una certa probabilita' in cui trovare il valore vero θ.

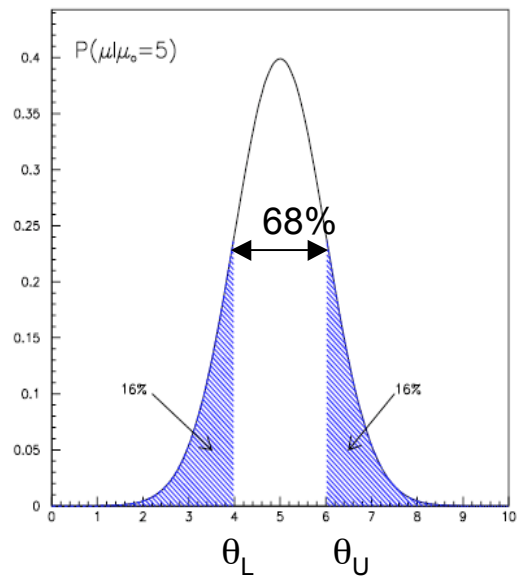
Approccio bayesiano

- *Tutta* la conoscenza su θ e' data dalla probabilita' a posteriori f(θ|x)

$$f(\theta \mid x) = \frac{f(x \mid \theta) f_0(\theta)}{\int d\theta f(x \mid \theta) f_0(\theta)}$$

- L'intervallo al 68% di C.L. e' appresentato da $[\theta_L < \theta < \theta_U]$ tale che:

$$\int_{\vartheta_L}^{\vartheta_U} f(\vartheta \mid x) d\vartheta = 0.68$$



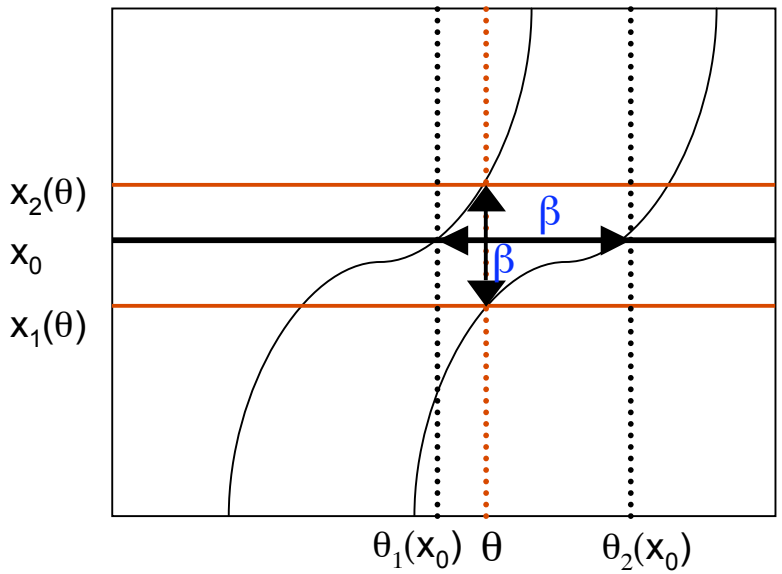
- Questa definizione non definisce univocamente l'intervallo. Es: si puo' richiedere che (1-P) =16% su entrambi i lati (intervallo centrale)

Approccio frequentista

- Nell'approccio frequentista il CL rappresenta un intervallo che include il valore vero del parametro con una probabilita' = livello specificato (per es. 68%)

- Costruzione di Neyman (non andiamo in dettaglio):

Intervallo chiuso



- Per ogni valore di θ, si definisce un intervallo di valori $x_1(\theta)$, $x_2(\theta)$ t.c.: $P(x_1(\theta) \leq x \leq x_2(\theta)) = \beta$ (x e' uno stimatore di θ)
- Il valore "osservato" x_0 definisce un intervallo t.c. $P(\theta_1(x_0) \leq \theta \leq \theta_2(x_0)) = P(x_1(\theta) \leq x \leq x_2(\theta)) = \beta$,ossia l'intervallo di confidenza si ottiene leggendo il grafico orizzontalmente.
- IL valore di β non definisce univocamente l'intervallo (ossia l'andamento funzionale delle curve $x(\theta)$). Per definirlo si impongono altre condizioni, tra cui:
  - Larghezza minima;
  - intervallo centrale ($P(X>X_2(\theta)) = P(X<X_1(\theta)) = \alpha = (1-\beta)/2$;
  - intervalli aperti

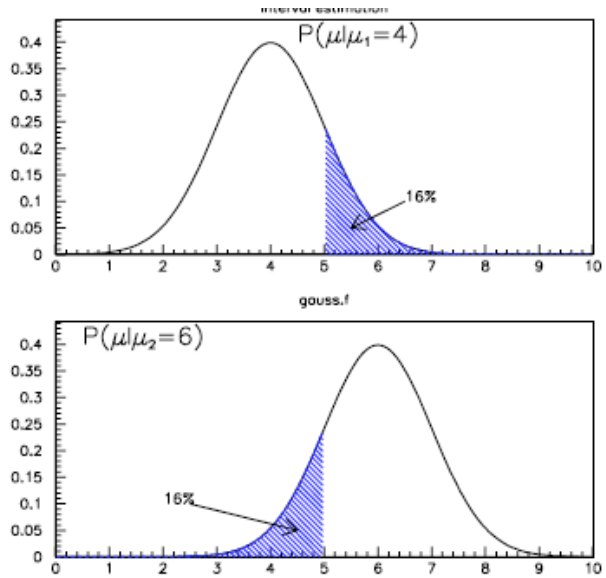
Intervallo centrale: caso gaussiano

- Intervallo centrale di confidenza al 68% (=β)
  - $P(X>X_2(\theta))= P(X<X_1(\theta)) = \alpha =(1-\beta)/2= 0.158$. Supponiamo che la p.d.f f(x) gaussiana con valor medio incognito μ:

$$\int_{x_0}^{\infty} f(x\,|\,\mu_1)dx = 0.158 \Rightarrow \mu_1 = x_0 - \sigma_x$$

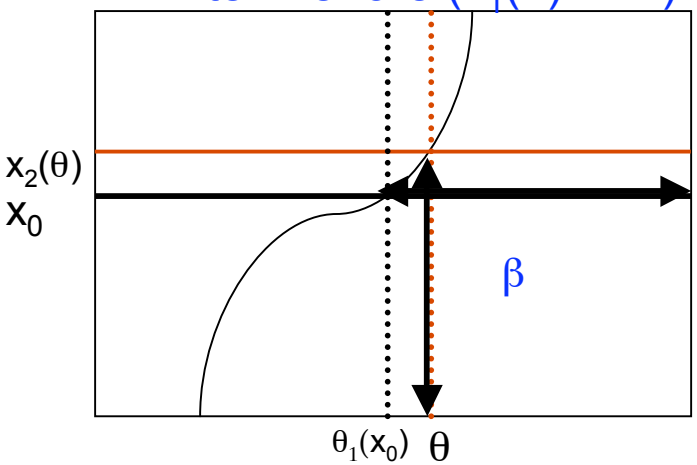
$$\int_{-\infty}^{x_0} f(x\,|\,\mu_2)dx = 0.158 \Rightarrow \mu_2 = x_0 + \sigma_x$$

- Quindi l'intervallo (centrale) al 68% (per μ) e' $(x_0-\sigma_x;\ x_0+\sigma_x)$



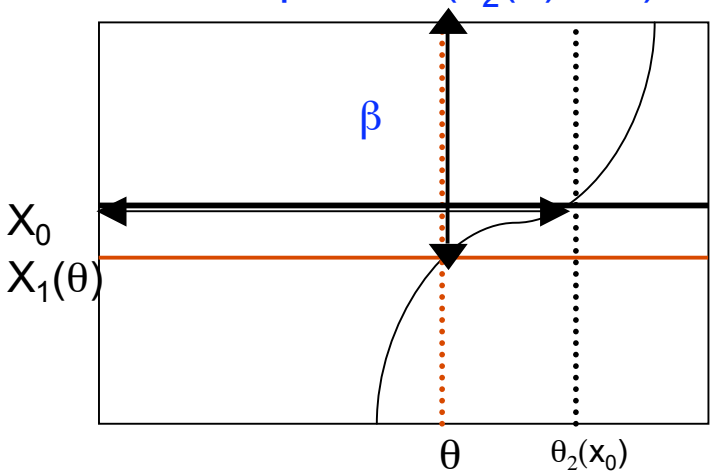
Intervalli aperti

Limite inferiore ($x_1(\theta)\to-\infty$)



$P(x<x_2(\theta))=\beta \Rightarrow P(\theta_1(x_0)\leq \theta)= \beta$;
$\theta_1(x_0)$ e' detto l.i.

Limite superiore ($x_2(\theta)\to\infty$)



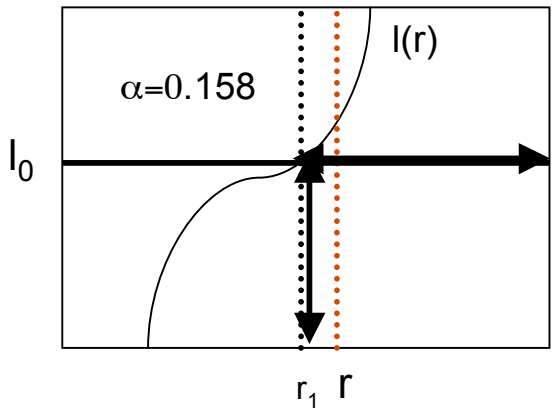
$P(x>x_1(\theta)) =\beta \Rightarrow P(\theta_2(x_0)\geq \theta)= \beta$;
$\theta_2(x_0)$ e' detto l.s.

Torniamo al problema sui leucociti:

- Dobbiamo stimare il C.L. (centrale) al 68% per r. Analogamente al caso gaussiano i due valori (inferiore e superiore) r_1 ed r_2 sono t.c. $P(r \le r_1) = 0.158$, $P(r \ge r_2) = 0.158$, ossia $P(r_1 \le r \le r_2) = 0.68$, e si trovano come soluzione delle rispettive equazioni:

$$P(l > l_0 \mid r_1) = 0.158 \Leftrightarrow \sum_{l=l_0,L} \binom{L}{l} r_1^l (1-r_1)^{L-l} = 0.158 \Rightarrow r_1 = 0.29$$

$$P(l < l_0 \mid r_2) = 0.158 \Leftrightarrow \sum_{l=0,l_0} \binom{L}{l} r_2^l (1-r_2)^{L-l} = 0.158 \Rightarrow r_2 = 0.48$$



- Possiamo confrontare questi limiti con quanto atteso da:

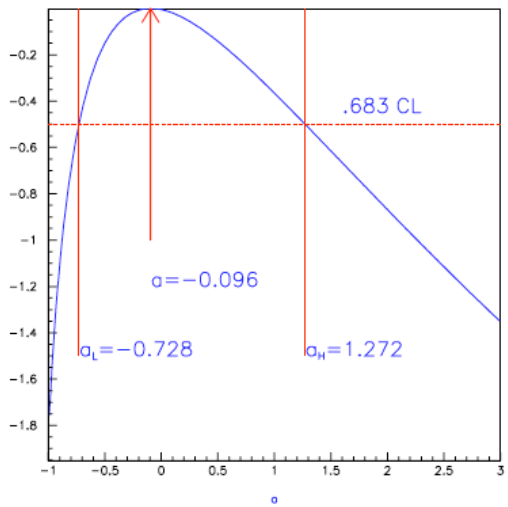
- Varianza della Binomiale: $\sigma_{\hat{r}}^2 = \dfrac{\hat{r}(1-\hat{r})}{L}$

$$\hat{r} - \sigma_{\hat{r}} < r < \hat{r} + \sigma_{\hat{r}} \Rightarrow 0.26 < r < 0.46$$

  - Tale procedure equivale ad assumere che r tenda ad una distribuzione normale e che la sua varianza possa essere calcolata sostituendo al valore vero r la sua stima $\hat{r}$.

- C.L. della M.L: $\ln L(\hat{r} \pm \sigma_{\hat{r}}) = \ln L(\hat{r}) - 0.5$

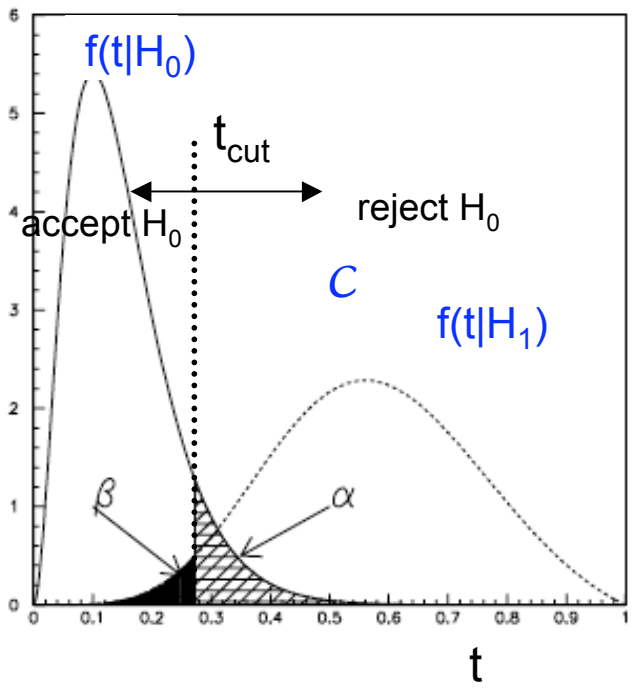
  - Si risolve numericamente ottenendo 0.27 < r < 0.46. Anche in questo caso si e' assunta l'approx. gaussiana per la likelihood



Test delle ipotesi

- Consideriamo un esperimento la cui uscita sia caratterizzata da un vettore di dati $\mathbf{x}=(x_1,\ldots,x_n)$. Un ipotesi e' una affermazione che riguarda la distribuzione di **x** (che puo' o no dipendere da parametri liberi).

- Un *test statistico* t(**x**) e' un metodo che permette di scegliere per quali valori di **x** una data ipotesi H_0 (tipicamente l'ipotesi nulla) va rigettata.
  - Si definisce una regione "critica" C nello spazio **t**-dimensionale (o x-dimensionale); se $t\in C$ H_0 va rigettata. Analogamente se $t\in(\Omega-C)$ H_0 va accettata.
  - In figura la regione critica e' $t>t_{cut}$



- Possibilta' di errori:

  - Rigettare H_0 se e' vera $\Rightarrow$ Errori del primo tipo. La probabilita' α per cui cio' accada e' detta livello di significativita' del test.

  - Accettare H_0 se e' falsa $\Rightarrow$ Errori del secondo tipo. La probabilita' β per cui cio' accada dipende da un ipotesi alternativa H_1 (vera) e $1-\beta$ e' detto potere del test per rigettare H_1.

- In Fisica delle alte energie si definisce l'efficienza di segnale ($\equiv H_0$) $\varepsilon_S=1-\alpha$ e l'efficienza per il fondo ($\equiv H_1$) $\varepsilon_B=\beta$

Neyman-Parson lemma

- La regione di accettanza con piu' alta purezza del segnale (ossia che minimizza l'efficienza di fondo) e' tale per cui

$$r = \frac{f(t \mid H_0)}{f(t \mid H_1)} > c$$

- *c* e' una costante determinata dall'efficienza del segnale desiderata.
- *r* e' anche detto rapporto delle likelihood.

Test delle ipotesi - Bayesiana

$$r = \frac{P(H_0 \mid t)}{P(H_1 \mid t)} = \frac{f(t \mid H_0)P(H_0)}{f(t \mid H_1)P(H_1)} > c$$

Test delle ipotesi: approccio Bayesiano

- Supponiamo di mettere alla prova una persona che dice di avere dei poteri telepatici (ESP). Estraiamo una carta e chiediamogli di indovinarla. Sia p la probabilita' di successo dovuta al "caso" (es. Su 5 carte p=0.2), individuata dall'ipotesi "nulla" H_0 (no poteri telepatici). Il numero di successi r in n prove e' una binomiale, con media:

$$\bar{r} = np \pm \sqrt{np(1-p)}$$

- Per n=37100 $\bar{r}$ e' 7420 ± 77. Assumiamo che il numero di successi osservato sia 9410, (percentuale di successi *f* =0.254) ossia ~26(!)σ dal valore aspettato nel caso di successi "casuali" (dato difficilmente compatibile con H_0). Possiamo costruirci la Likelihood:

$$L_0 = P(D \mid H_0) = 3.15 \times 10^{-139}$$

- Consideriamo ora l'ipotesi alternativa in cui assegno una probabilita'di successo q=0.254. Posso valutare la likelihood in questa ipotesi

- Da cui

$$\frac{L_0}{L_f} = \frac{P(D \mid H_0)}{P(D \mid H_f)} = 7 \times 10^{-137} (!!!)$$

- Ossia non ci sarebbero dubbi che i dati supportano l'ipotesi di ESP. Ma…

Test delle ipotesi: appr. bayesiano

- Consideriamo la probabilita' a posteriori dell'ipotesi ESP:

$$P(H_f \mid D) = \frac{P(D \mid H_f)P(H_f)}{P(D)} = \frac{L_f P(H_f)}{L_f P(H_f) + L_0 P(H_0)}$$

- Assumiamo ora che ci siano altre ipotesi (che prima abbiamo trascurato), con likelihood e probabilita' a priori L_i, P_i

$$P(H_f \mid D) = \frac{L_f P_f}{L_f P_f + L_0 P_0 + \sum L_i P_i}$$

- Ora per essere $P(H_f|D)\sim 1$ occorre che $L_0P_0 + \Sigma P_iL_i << L_fP_f$. Se assumiamo che $L_i \sim L_f$, ossia che le ipotesi alternative possano produrre gli stessi risultati dell'ipotesi ESP, ed essendo L_0P_0 trascurabile ($\leq 10^{-139}$):

$$\sum P_i << P_f$$

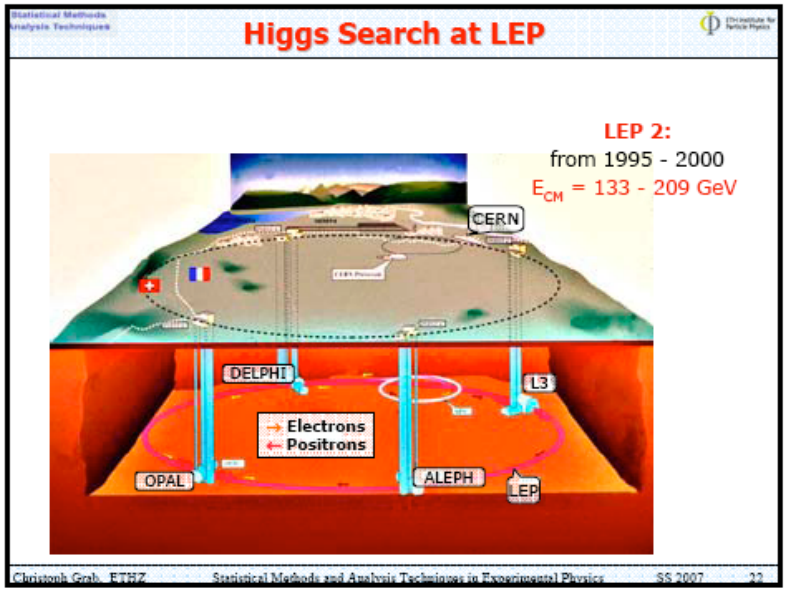
- **Dal momento che ogni ipotesi alternativa non puo' essere esclusa a priori ⇒ la disuguagliana sopra, difficilmente e' verificabile !**

- Quindi l'esperimento ESP difficilmente convince, non perche' dogmaticamente $P_f=0$, ma perche' ci possono essere moltissime ipotesi alternative che non possono essere scartate a priori.

Test delle ipotesi: approccio frequentista

- In un esperimento, quando si osserva un discrepanza tra i dati e quanto atteso, occorre decidere se quanto osservato sia realmente un segnale di nuova fisica o solo una fluttuazione (statistica) del fondo (=eventi di fisica nota).

  - Un esempio di questo problema e' quanto e' accaduto al CERN di Ginevra, dove alcuni esperimenti del collisore LEP hanno osservato una discrepanza nei dati compatibile con un segnale del bosone di Higgs con massa 115 GeV/c^2.



- L'ipotesi "nulla" H_0 e' l'ipotesi di solo fondo, ossia che i dati possano essere capiti con le spiegazioni fisiche esistenti. L'ipotesi alternativa H_1 (che e' favorita quando l'ipotesi nulla e' rigettata con un alta *significativita'*) e' quella di scoperta, ossia che ci sia bisogno di nuova fisica per spiegare i dati

Test delle ipotesi: approccio frequentista

- Si usa costruire un test statistico X in cui si confronta l'ipotesi che i dati siano di "segnale+fondo" (ipotesi H_1), con l'ipotesi di solo "fondo" H_0:

$$X(m_H) = -2\ln Q(m_H) = -2\ln \frac{L(x \mid s+b)}{L(x \mid b)}$$

- Quanto piu' l'ipotesi di "s+b" e' vera tanto piu' X(x) e' negativa (e viceversa).

- $\mathcal{L}$(x|b) e' la funzione di likelihood nell'ipotesi di solo fondo. x e' in generale un set di variabili rappresentative dei dati. Facciamo l'ipotesi che sia x_i sia il numero di eventi osservati, nel bin i-esimo, e assumiamo che il fondo sia distribuito secondo una poissoniana:

$$L(x \mid b) = \prod_i^N \frac{e^{-b_i} b_i^{x_i}}{x_i!}$$

- Calcolare X per i seguenti valori di eventi s= 16.0, b=71.7, dati=70.0

Test delle ipotesi

- Consideriamo x come il numero di eventi totali. Se il fondo fluttua in modo poissoniano (cosi' come come il segnale), anche s+b e' una variabile poissoniana:

$$L(x \mid b) = \frac{e^{-b} b^{x}}{x!}$$

$$L(x \mid s+b) = \frac{e^{-(s+b)} (s+b)^{x}}{x!}$$

- Si ottiene:

$$X = -2\left(x \ln(s+b) - x \ln b - s\right)$$

- Da cui X=3.8, che favorisce l'ipotesi di solo fondo.

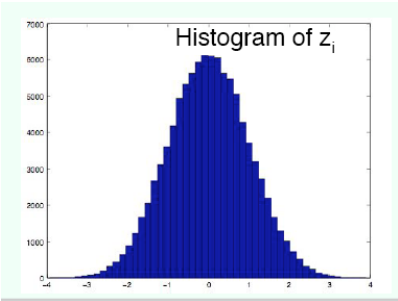
- Vogliamo ora quantificare se questo valore e' significativo (e a che livello).

- Facciamo quello che si chiama un Monte Carlo. Ripetiamo idealmente il nostro esperimento 50000 volte e per ogni volta calcoliamo il valor di X nell'ipotesi che il dato sia una fluttazione del fondo (estrazione poissoniana di media b) o del segnale+fondo (estrazione poissoniana di media s+b)

$$X_B(i) = -2\left(x_i^b \ln(s+b) - x_i^b \ln b - s\right)$$

$$X_{S+B}(i) = -2\left(x_i^{s+b} \ln(s+b) - x_i^{s+b} \ln b - s\right)$$

$x_i^{b,s+b}$ e' estratta poissonianamente



Test delle ipotesi

- Calcoliamo la probabilita' che ciascuna delle due ipotesi ci dia un valore di X ≥ di quello osservato:

$$CL_B = P(X \geq X_0 \mid B) = \frac{N(X_B(i) \geq X_0)}{50000} = 0.4048$$

$$CL_{S+B} = P(X \geq X_0 \mid S+B) = \frac{N(X_{S+B}(i) \geq X_0)}{50000} = 0.02264$$

- Da cui si ricava:

$$CL_S = \frac{CL_{S+B}}{CL_B} = 0.0559$$

- Siccome $1\text{-}CL_S < 0.95$, il segnale **non** puo' essere **escluso** al 95% di CL.

- Esercizio: generalizzare al caso di piu' intervalli

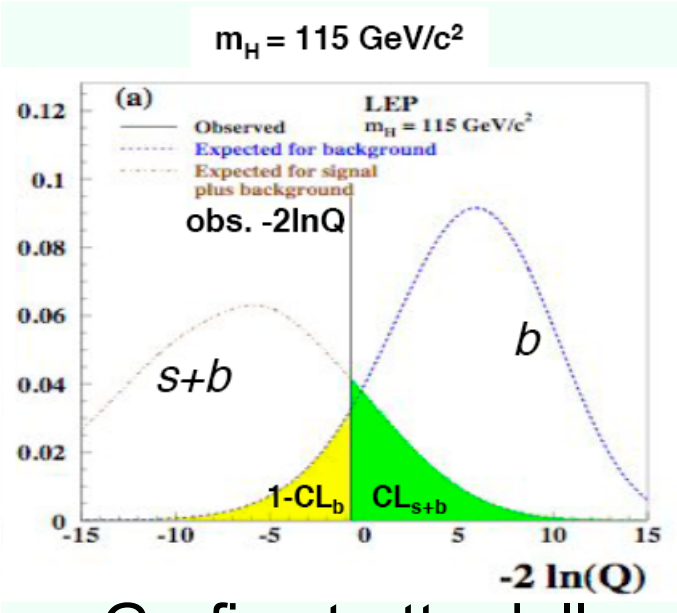


Grafico tratto dalla pubblicazione scientifica del LEP

Bibliografia

- G. Cowan "Statistical Data Analysis" , Oxford Science Publications, 1998
- E.T. Jaynes "Probability Theory", Cambridge University Press, 2003
- The LEP Working Group for Higgs Boson Searches "Search for the Standard Model Higgs Boson at LEP", Physics Letter B, **565** 2003, 61-75
- D.S. Sivia "Data Analysis, a Bayesan tutorial", Oxford Science Publications, 2006