Test d'ipotesi e tecniche multivariate

Laboratorio di Metodi Computazionali e Statistici (2022/2023)

Roberta Cardinale e Fabrizio Parodi

Dipartimento di Fisica - Università di Genova

Test statistici: introduzione

- I test d'ipotesi invece sono uno strumento per fare affermazioni sulla validità di un modello basandosi su una serie di dati raccolti (oppure tra due modelli in competizione, verificare quale è più consistente con i dati)
- strumento per trovare criteri per classificare eventi (selezionare eventi) provenienti da due diversi tipi

Test statistici

- Supponiamo di aver effettuato un esperimento/misura, assumiamo che i dati siano un set di N osservabili (caratteristiche dell'esperimento/misura)

$$\mathbf{x} = (x_1, ...x_N)$$

- **x** seguono una distribuzione di probabilità congiunta
- Ipotesi: specifica la distribuzione di probabilità dei dati (**x**)
- L'obiettivo è di effettuare una qualche affermazione basandosi sui dati osservati **x** sulla validità di una possibile ipotesi
- Il test può essere effettuato su una grandezza osservata nell'esperimento o su qualsiasi funzione di una o più grandezze osservate nell'esperimento
- Spesso invece di utilizzare tutte le variabili **x**, si costruisce una variabile $t(\mathbf{x})$ dall'insieme delle osservabili/misure detta "statistica di test" che riassume tutte le informazioni contenute nel campione misurato
- Statistica di test: è una variabile funzione dei dati che fornisce un metodo per fare un test delle ipotesi (che ha cioè un potere discriminante tra diverse ipotesi)

Ipotesi

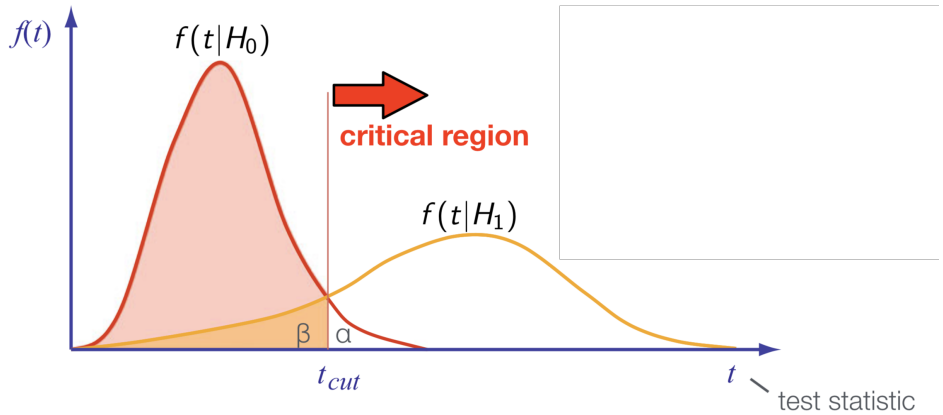
- Vogliamo determinare la validità di una ipotesi sulla base del $t(\mathbf{x})$ osservato (che dipende solo dai dati): possiamo rigettare l'ipotesi dai dati che abbiamo?
- La statistica $t(\mathbf{x})$ avrà una distribuzione di probabilità per l'ipotesi H_0 (ipotesi nulla): $f(t(\mathbf{x})|H_0)$
- Un test di ipotesi H_0 è definito specificando una regione critica W (detta anche regione di rigetto dell'ipotesi H_0) dello spazio dei dati S, tale che, assumendo H_0 sia corretta, la probabilità di osservare t in tale regione sia uguale o minore di α:

$$P(t \in W|H_0) \leq \alpha$$

- α è la significanza del test (ed è un valore scelto da noi).
- Se il valore ottenuto dai dati di t è in W, l'ipotesi H_0 è rigettata.
- La regione complementare a W è detta regione di accettanza (non possiamo rigettare l'ipotesi H_0).

Errori di Tipo I e Tipo II

- Consideriamo un'ipotesi alternativa rispetto a H_0: H_1



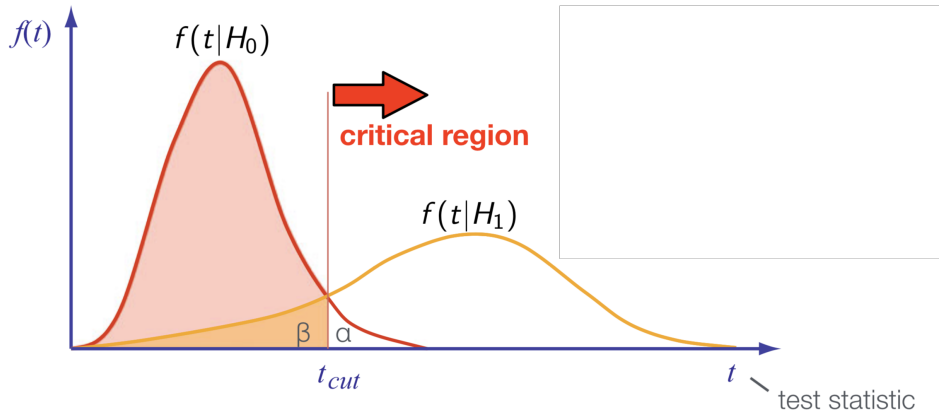
- Errore di tipo I: rigettare l'ipotesi H_0 se è vera. La massima probabilità per questo errore è

$$P(x \in W|H_0) \leq \alpha \quad \text{dove} \quad \alpha = \int_{t_{cut}}^{\infty} f(t|H_0)dt$$

α è il livello di significanza del test, ed è definito come la probabilità di t di essere nella regione W in cui rigetto H_0, quando H_0 è vera

Errori di Tipo I e Tipo II

- Consideriamo un'ipotesi alternativa rispetto a H_0: H_1



- Errore di tipo II: rigettare l'ipotesi alternativa H_1 se è vera

$$P(t \in S - W|H_1) \leq \beta \quad \text{dove} \quad \beta = \int_{-\infty}^{\mathrm{t_{cut}}} \mathrm{f(t|H_1)dt}$$

β è la probabilità che t cada nella regione di accettanza per H_0 se H_1 è vera

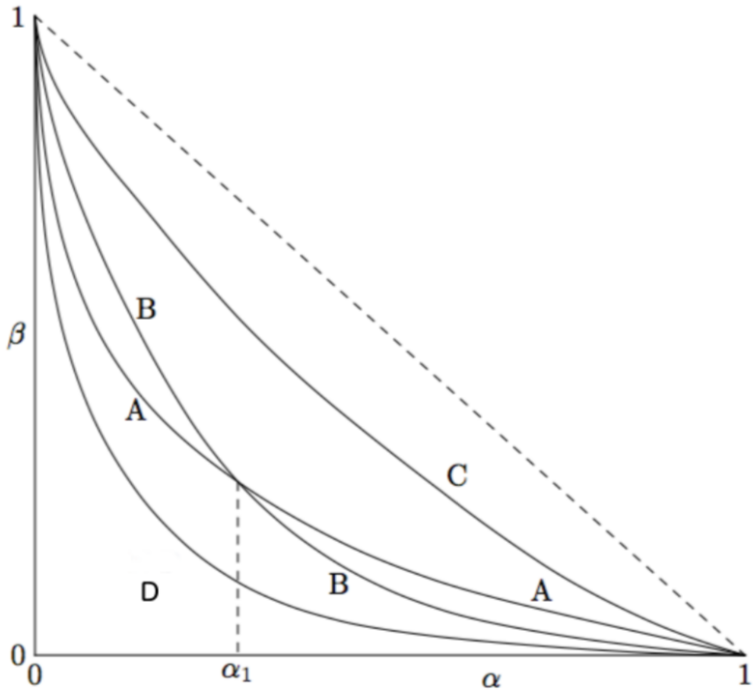
  - L'utilità di un test dipende dalla sua capacità di discriminare l'ipotesi alternativa H_1.
  - Per misurare l'utilità di un test si usa la potenza del test definita come la probabilità 1 - β cioè la probabilità che t cada nella regione critica se H_1 è vera $P(t \in W|H_1) \leq 1 - \beta$
  - cioè di rigettare H_0 se H_1 è vera

Errori di tipo I o II

		H_0 TRUE	H_1 TRUE
$X \notin w$	ACCEPT H_0	Acceptance good Prob $= 1 - \alpha$	Contamination Error of the second kind Prob $= \beta$
$X \in w$ (critical region)	REJECT H_0	Loss Error of the first kind Prob $= \alpha$	Rejection good Prob $= 1 - \beta$

Scelta del test statistico (I)

- Come si effettua la scelta di un test statistico?
- Dato un esperimento/una serie di dati, posso determinare più test statistici
- Ma quale è il migliore?



- Il miglior test è quello per cui fissato α, β sia il più piccolo possibile

Scelta del test statistico (II)

Il lemma di Neyman-Pearson afferma che in un test d'ipotesi tra due ipotesi H_0 e H_1, la statistica di test ottimale, cioè quello che ha la potenza massima tra tutti i test (cioè quello con il più piccolo valore di β) con significatività α,la regione critica W deve essere tale che per tutti i valori $\mathbf{x}$ nella regione critica si ha:

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > c$$

all'interno di W e $\leq c$ fuori dalla regione, dove c è una costante scelta per avere un test per un determinato livello di significanza desiderato

Per cui dobbiamo richiedere che per tutti i punti $\mathbf{x}$ nella regione critica, il rapporto delle distribuzioni di probabilità sia maggiore di una costante c dove c corrisponde al valore di taglio dato il valore di significanza α scelto

Equivalente a dire che il test statistico ottimale è dato da:

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

p-value

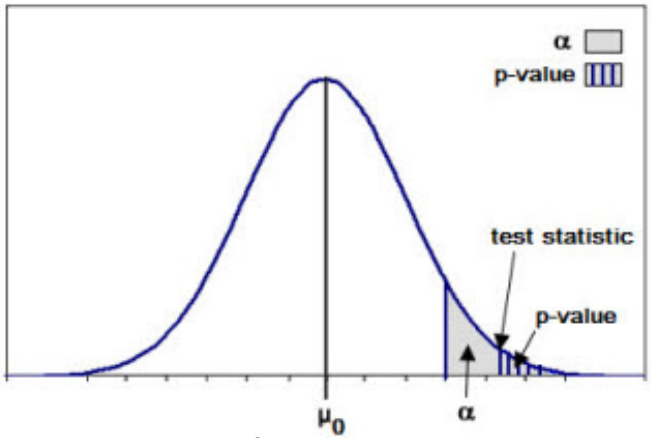
- Spesso si vuole esprime il livello di accordo dei dati con un'ipotesi H cioè vogliamo quantificare il livello di compatibilità tra i dati ottenuti e quelli attesi secondo l'ipotesi H
- Si definisce una statistica di test $t(\mathbf{x})$
- Si ha la distribuzione della statistica di test $f(t|H)$ data l'ipotesi considerata H
- Si utilizza il p-value definito come probabilità di osservare, assumendo H vera, dati $\mathbf{x}$ (o $t(\mathbf{x})$) che hanno minore o uguale compatibilità con H rispetto ai dati che abbiamo osservato ($\mathbf{x}_{\mathrm{oss}}$ o $t_{oss}(\mathbf{x})$)

$$p = P(\mathbf{x} \in \omega \leq \mathbf{x}_{oss}|H)$$

p-value

- Si calcola dalla distribuzione della statistica di test tenendo conto dell'ipotesi considerata $f(t|H)$

  $p = \int_{x:t>t_{obs})} f(x|H)dx$ per distribuzioni continue
  $p = \sum_{x:t>t_{obs}} f(x|H)dx$ per distribuzioni discrete



$$p = \int_{t_{obs}}^{+\infty} f(t|H)dt$$

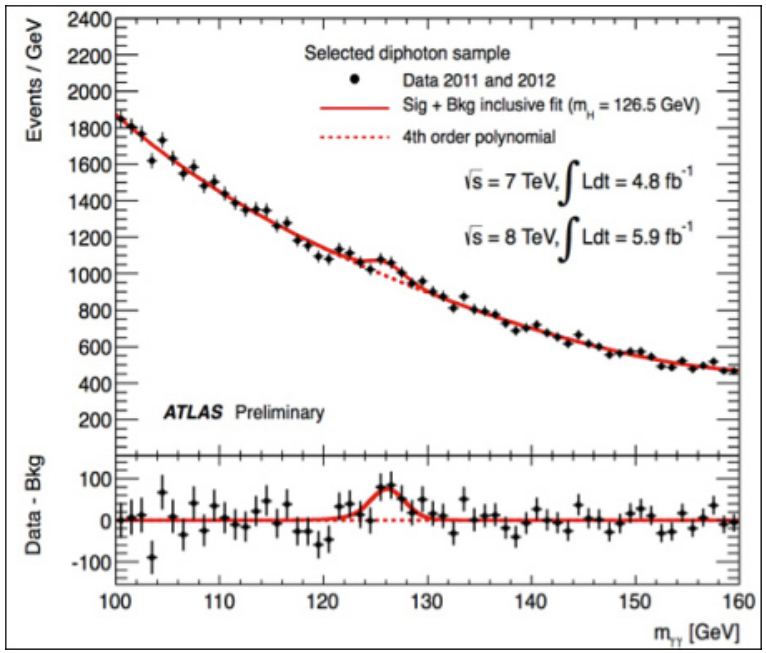
Decisa una significanza del test α, l'ipotesi H è rigettata se $p < \alpha$.

p-value

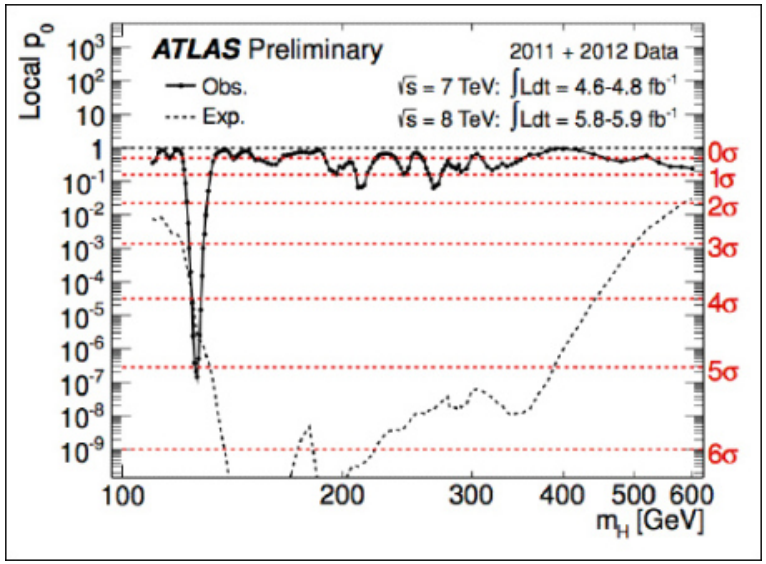
- Attenzione: p non è la probabilità che H sia vera dati i dati raccolti!
- Nell'approcio frequentista: non esiste la probabilità di H ma solo la probabilità di ottenere un altro set di valori $\mathbf{x}_{\text{oss}}$
- È la probabilità di ottenere dei dati incompatibili almeno tanto quanto i dati che abbiamo se l'ipotesi è vera.
- Noi ci concentreremo solo sull'approcio frequentista

Esempio: scoperta del bosone di Higgs (ATLAS e CMS)

Distribuzione di massa



p-value in funzione della massa



Alla scoperta del bosone di Higgs molti quotidiani hanno tradotta l'informazione (corretta) "la probabilità di una tale fluttuazione nei dati se non esiste il bosone di Higgs è 3×10^{-7}" nell'affermazione (falsa e fuorviante) "la probabilità che non esista il bosone di Higgs, osservata quella fluttuazione nei dati, è 3×10^{-7}"

p-value

- Il p-value non deve essere confuso con il livello di significanza α
  - il livello di significanza è un valore fisso scelto a priori
  - il p-value è una funzione dei dati e quindi è esso stesso una variabile aleatoria (con una data distribuzione)
- Un p-value piccolo è indice di una inconsistenza con l'ipotesi formulata
- L'ipotesi è rigettata se $p < \alpha$.

Esempio

- Supponiamo di effettuare un esperimento di conteggi e osserviamo n eventi
- Abbiamo due tipi di eventi possibili (evento interessante mai visto prima s, evento non interessante b): contiamo solo il numero totale di eventi

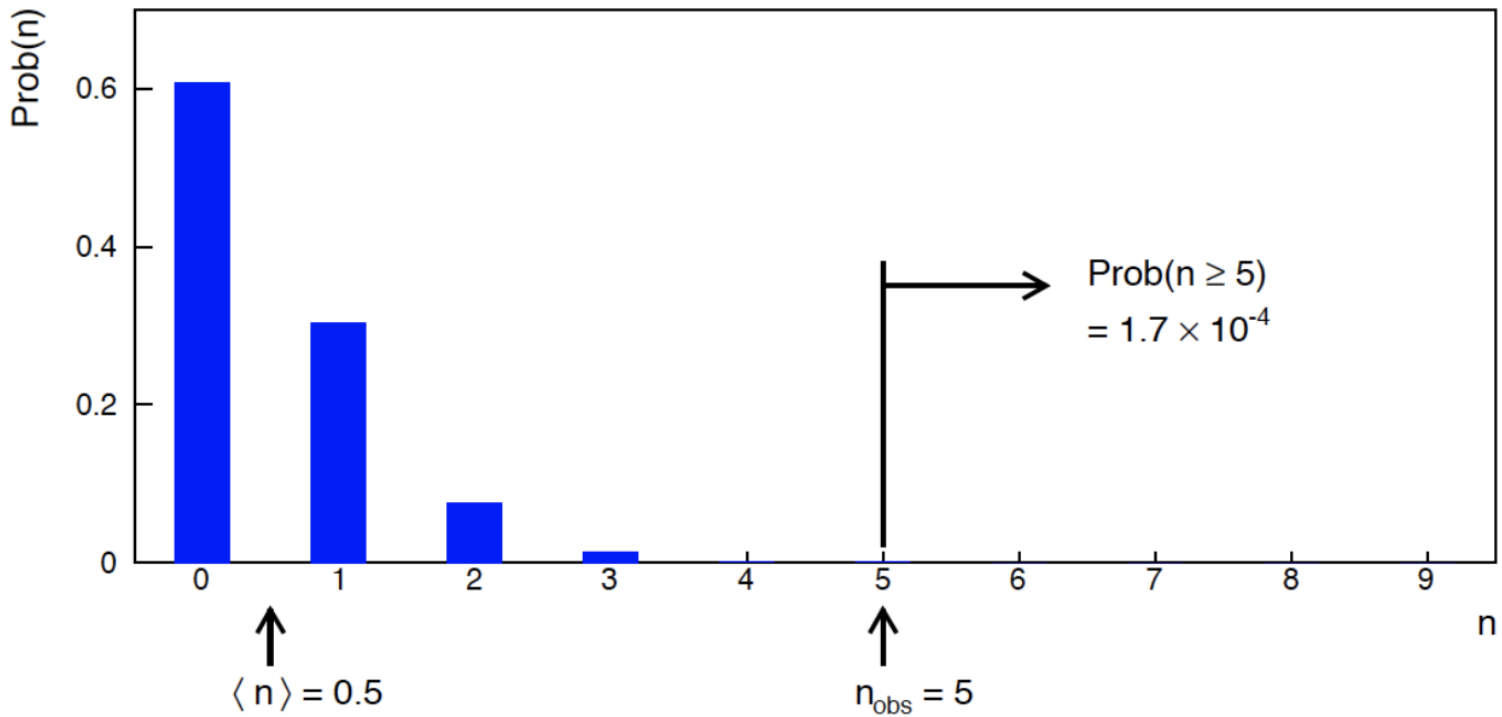
$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

- s = numero medio di eventi interessanti
- b = numero medio di eventi non interessanti
- Vogliamo testare l'ipotesi $H_0(s = 0)$ (rigettare l'ipotesi H_0: c'è un tipo di eventi mai visti prima)
- Testare $H_0(s \neq 0)$: intervalli di confidenza

Esempio

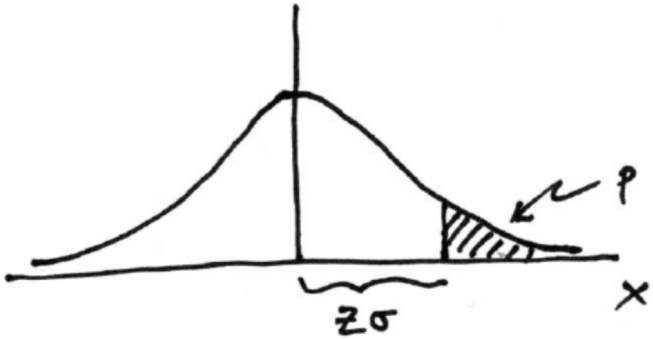
- Supponiamo noto $b = 0.5$ e osserviamo $n_{oss} = 5$
- Cosa possiamo dire di H_0?
- Calcoliamo il p-value per $H_0(s = 0)$

$$p = P(n \geq 5; b = 0.5, s = 0) = \sum_{n=5}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - \sum_{n=0}^{4} \frac{b^n}{n!} e^{-b} = 1.7 \times 10^{-4}$$



Quantile

- Spesso si converte il p-value in una probabilità equivalente (chiamata significanza o equivalente significanza Gaussiana o quantile)



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

In ROOT

```
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

In python (scipy.stats)

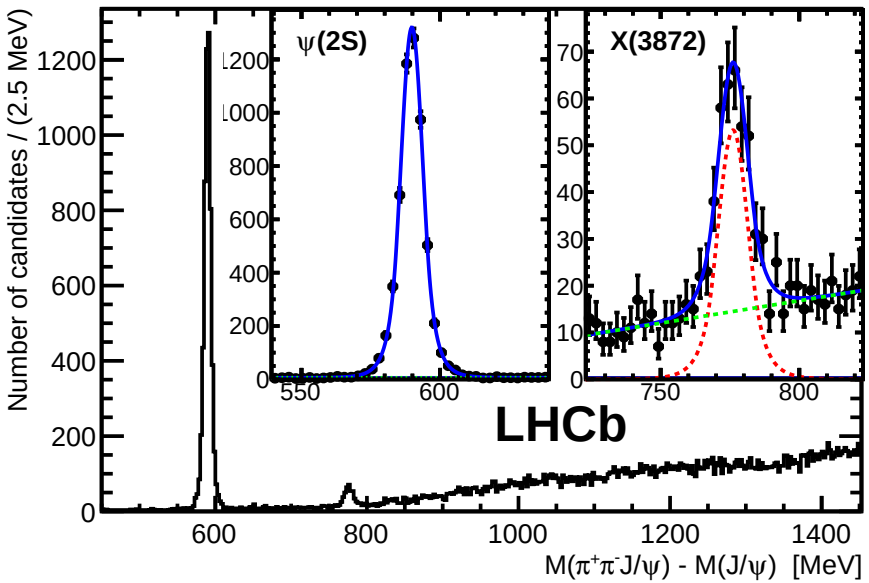
```
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

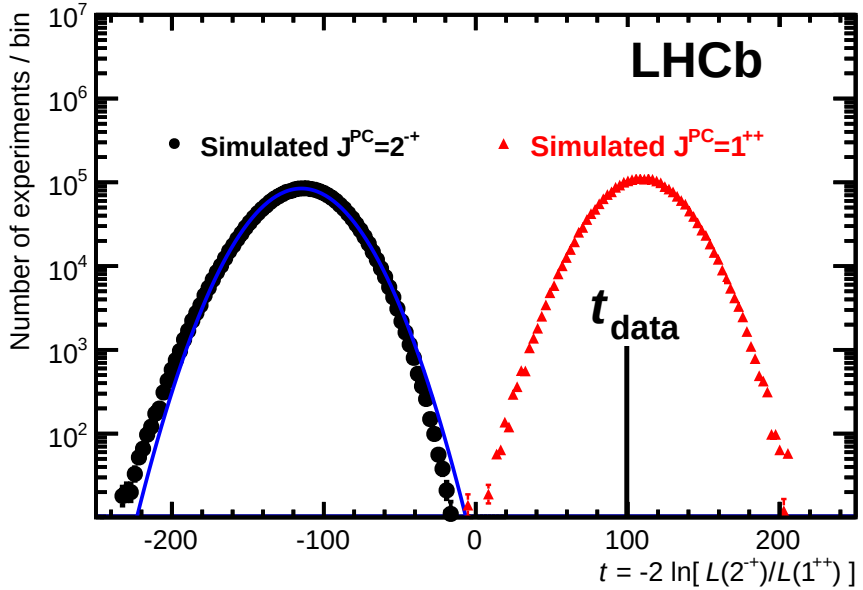
- La definizione di quantile permette anche di ricondurre la significanza ad un valore equivalente di σ
- Z è il numero di deviazioni standard che una variabile Gaussiana fluttuerebbe (in una direzione) per dare lo stesso p-value
- Per esempio Z= 2 (σ) equivale a 0.05, Z= 3 (σ) equivale a 0.003, Z= 5 (σ) equivale a 2.9×10^{-7}

Come procede il processo di test?

- Definiamo un'ipotesi nulla H_0 e possibili ipotesi alternative
- Selezioniamo un test statistico t
- Determiniamo la distribuzione attesa per t nel caso di ipotesi nulla $f(t|H_0)$
- Scegliamo il valore di α tenendo in conto degli errori di tipo I e di tipo II e definiamo la regione critica
- Determiniamo dai dati il valore di t
- Confrontiamo se il valore di t misurato si trova nella regione critica: se è nella regione critica, rigettiamo l'ipotesi nulla, altrimenti concludiamo che non c'è evidenza per rigettare l'ipotesi nulla

Determinazione spin particella esotica





Test parametrici e non-parametrici

I test di ipotesi si possono suddividere in:

- Test parametrici: riguardano ipotesi sul valore di un parametro della distribuzione (occorre assumere una distribuzione) come la media/la deviazione standard
- Non Parametrici: riguardano il tipo di distribuzione ipotizzabile. Non richiedono la conoscenza a priori della distribuzione dei dati
  - Test di bontà del fit o detti test sulla bontà dell'adattamento

Test parametrici: il problema dei due campioni

- Vogliamo confrontare due campioni di dati e vedere se sono compatibili (se provengono dalla stessa popolazione)
- Anche se provengono dalla stessa popolazione, avranno delle differenze dovute a fluttuazioni statistiche
  - Due campioni gaussiani con σ nota, la media dei due campioni è la stessa?
  - Due campioni gaussiani con stessa (non nota) σ, la media dei due campioni è la stessa?
  - Due campioni gaussiani, σ_1 è compatible con σ_2?

Due gaussiane, σ nota

Supponiamo di avere dei campioni indipendenti $(x_1, \ldots x_n)$ e $(y_1, \ldots y_n)$ provenienti da due popolazioni normali di varianze note σ_x^2 e σ_y^2
Calcoliamo le medie dei due campioni $\bar{x}$ e $\bar{y}$, ovviamente questi due numeri non coincideranno

Ci chiediamo se i due campioni hanno la stessa media oppure no? Riformulazione: $\bar{x} - \bar{y}$ è compatibile con 0? cioè la differenza tra le medie è significativa? O dipende solo da fluttuazioni casuali dovute alla dimensione dei campioni di dati?

Esempio: due misure effettuate con apparati diversi con risoluzione σ_x e σ_y

Due gaussiane, σ nota

Test a una coda destro:

$\theta = \bar{x} - \bar{y}$

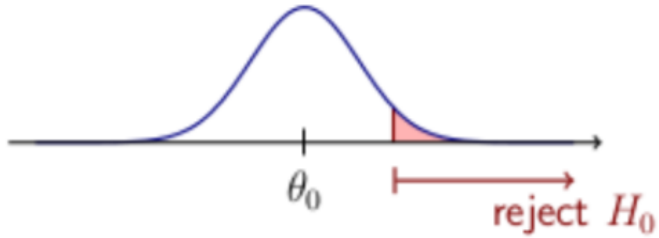
$\epsilon_{x-y} = \sqrt{\sigma_x^2/N_x + \sigma_y^2/N_y}$

$\theta_0 = 0$

H_0	H_1	test
$\theta = \theta_0$	$\theta > \theta_0$	$(\theta - \theta_0)/\epsilon$

$t = (\theta - \theta_0)/\epsilon = Z$

in numero di deviazioni standard



Due gaussiane, σ nota

Test a una coda sinistro:

$\theta = \bar{x} - \bar{y}$

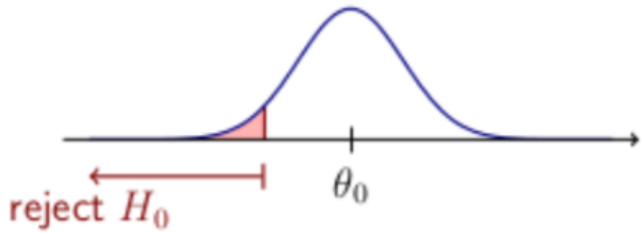
$\epsilon_{x-y} = \sqrt{\sigma_x^2/N_x + \sigma_y^2/N_y}$

$\theta_0 = 0$

H_0	H_1	test
$\theta = \theta_0$	$\theta < \theta_0$	$(\theta - \theta_0)/\epsilon$

$t = (\theta - \theta_0)/\epsilon = Z$

in numero di deviazioni standard

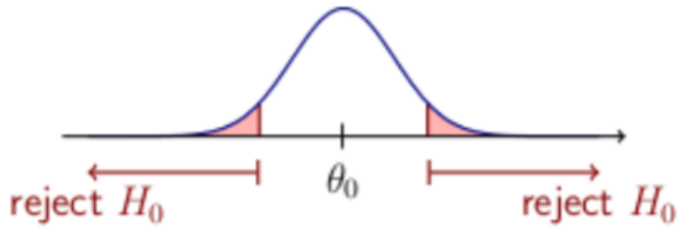


Due gaussiane, σ nota

Test a due code:
$\theta = \bar{x} - \bar{y}$, $\epsilon_{x-y} = \sqrt{\sigma_x^2/N_x + \sigma_y^2/N_y}$, $\theta_0 = 0$

H_0	H_1	test
$\theta = \theta_0$	$\theta \neq \theta_0$	$\lvert\theta - \theta_0\rvert/\epsilon$

$t = (\theta - \theta_0)/\epsilon = Z$
in numero di deviazioni standard



La questione si riduce al calcolo di quante σ la differenza è distante rispetto a zero. E si confronta il livello di significatività del test voluto con la tabella dell'integrale della Gaussiana.

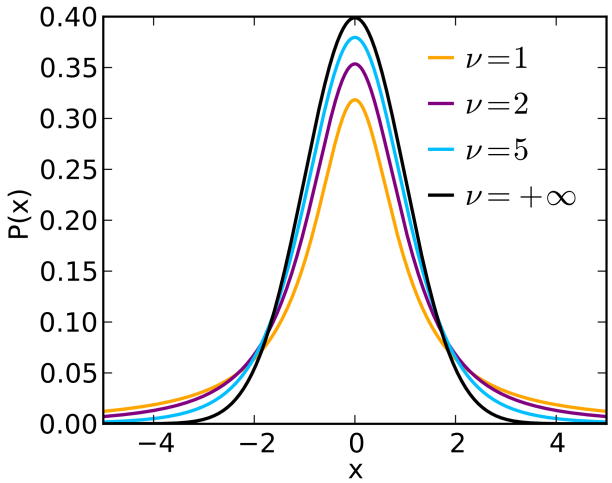
Due gaussiane, σ ignota

- Se non conosciamo la varianza della popolazione, possiamo stimarla con la varianza campionaria $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N_x - 1}}$ e $s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N_y - 1}}$
- La statistica di test

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{(1/N_x) + (1/N_y)}}$$

con $S^2 = \frac{(N_x - 1)s_x^2 + (N_y - 1)s_y^2}{N_x + N_y - 2}$

- è distribuita secondo la distribuzione di Student con numero di gradi di libertà $n = N_x + N_y - 2$



La significanza di una deviazione tra x e y è inferiore (la distribuzione ha code più lunghe) rispetto al caso precedente.

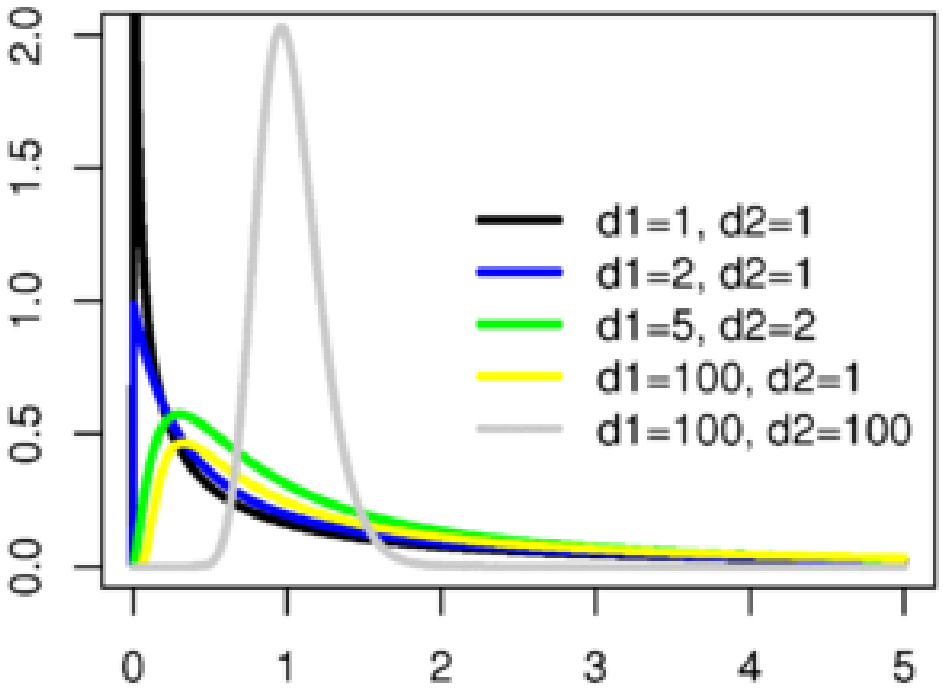
Test di Fisher F-test: test della varianza

Due gaussiane, σ_1^2 è compatibile con σ_2^2 ?

Prendiamo il rapporto tra gli stimatori delle varianze osservate:

$$F = \frac{V_1}{V_2} \qquad V = \sum_i \frac{(x_i - \bar{x})^2}{N-1}$$

segue la distribuzione di Fisher di parametri $F(n-1, m-1)$ e si definisce la regione di accettazione come l'intervallo tra i quantili di ordine $\alpha/2$ e $1-\alpha/2$ con α livello di significatività



Analisi della varianza

- La più importante applicazione del test di Fisher consiste nell'analisi della varianza (ANOVA, Analysis of Variance)
- Nelle scienze sociali l'analisi della varianza si usa per studiare caratteristiche simili in gruppi diversi
- Confronta G campioni confrontando la varianza interna ai gruppi con la varianza tra i gruppi
- Ipotesi nulla prevede che tutti i campioni provengano dalla stessa popolazione
- Test su G campioni sono compatibili con avere lo stesso valore atteso, nell'ipotesi di normalità con σ fissata ma non nota. Il test è utile in particolare per sapere se il valore atteso di una variabile x possa dipendere da una variabile categoriale in base alla quale è possibile suddividere il campione.

Analisi della varianza

- Si calcola la varianza tra i gruppi $\sigma^2_{between\ groups}$ e la varianza interna ai gruppi $\sigma^2_{within\ groups}$

$$\sigma^2_{between\ groups} = \sum_{i=1}^{G} (\bar{x}_i - \bar{x})^2 \frac{n_i}{n-1}$$

$$\sigma^2_{within\ groups} = \frac{1}{n-1} \sum_{i=1}^{G} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_g)^2$$

Statistica del test:

$$t = \frac{(\sigma^2_{between\ groups}(n-1)/(G-1))}{(\sigma^2_{withingroups}(n-1)(n-G))}$$

da confrontare con la distribuzione di Fisher di parametri $F(G-1, n-G)$

Test di bontà del fit

- Sono un'importante tipo di test d'ipotesi
- Il nome "test di bontà del fit" è una classificazione fuorviante: non hanno a che fare direttamente con una procedura di fit
- Si ha un campione di dati e una distribuzione e la domanda che ci si pone è: quanto un campione di dati è descritto da una certa pdf?
- I test di bontà del fit confrontano i dati sperimentali con la pdf associata all'ipotesi nulla (oppure confrontano due set di dati sperimentali)
- Ipotesi H_0: i dati seguono la funzione ipotizzata e le differenze sono solo fluttuazioni
- Si costruisce anche in questo caso un test statistico
- Si calcola la probabilità di ottenere un valore del test statistico almeno grande quanto quello misurato, cioè di ottenere dati almeno compatibili con l'ipotesi come quelli misurati
- I test che vedremo hanno il vantaggio che sono indipendenti dal tipo di distribuzione (il test è lo stesso sia che i dati seguano per esempio una gaussiana o una distribuzione lineare)

Test binned e unbinned

- Combinando gli eventi in bin di istogrammi, abbiamo già visto che si perde informazione sui nostri dati
- La scelta di un binning implica infatti necessariamente perdita di informazioni e arbitrarietà nella scelta del binning
- La perdita di informazione sarà trascurabile se la larghezza del bin è piccola rispetto alla risoluzione sperimentale
- Ma in generale, i test binned hanno un potere inferiore rispetto ai test unbinned.
- Per alcuni test binnati è inoltre importante che il numero di eventi per bin sia sufficientemente grande da giustificare alcune assunzioni fatte

Test binnati

- Supponiamo che il set di dati misurati sia un set di numeri che possiamo rappresentare con un istogramma (conteggi)
- cioè abbiamo ottenuto un set di numeri $\mathbf{n} = (n_1, .., n_n)$ con cui posso riempire un istogramma
- $\mathbf{n} = (n_1, .., n_n)$ sono il numero di entries per ognuno dei bin dell'istogramma
- Voglio testare un'ipotesi che predice dei valori attesi in ogni bin dell'istogramma $\nu = (\nu_1, .., \nu_n)$
- χ^2 di Pearson è un test statistico che mi permette appunto di confrontare i valori osservati in ogni bin dell'istogramma con i valori predetti da un certo modello/teoria (confronto una distribuzione osservata con una attesa)

χ^2 di Pearson

Consideriamo tre casi:

- Il numero di entries in ogni bin dell'istogramma è sufficientemente grande e bin indipendenti, trattati come continui: $n_i \sim Gauss(\nu_i, \sigma_i)$

$$p(\mathbf{n}|\nu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(n_i-\nu_i)^2/2\sigma_i^2}$$

- $n_i \sim Poisson(\nu_i)$, bin indipendenti

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

- Se $n_{tot} = \sum n_i$ è fissato la probabilità in ogni bin è descritta dall multinomiale: $\mathbf{n} \sim Multinomiale(n_{tot}, \mathbf{n})$ con $n_{tot} = \sum_i n_i$ e $\mathbf{p} = \mathbf{nu}/n_{tot}$ $(p_i = \nu_i/n_{tot})$

$$P(\mathbf{n}|\nu) = \frac{n_{tot}!}{n_1!...n_N!} p_1^{n_1}...p_N^{n_N}$$

Statistica di χ^2 di Pearson

Statistica di test:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \qquad \text{dove} \qquad \sigma_i^2 = V(n_i)$$

χ^2 è la somma dei quadrati delle deviazioni del numero di entries misurate nel bin i-esimo dal valor medio atteso nel bin i-esimo divise per la varianza
Si può calcolare un p=value tenendo conto che $\chi^2 \geq \chi^2_{oss}$ definisce la regione a "uguale o minore compatibilità" (χ^2 elevato significa minor compatibilità):

$$p(\chi^2) = \int_{\chi^2_{oss}}^{+\infty} f(\chi^2) d\chi^2$$

dove $f(\chi^2)$= pdf della statistica di test sotto l'ipotesi che stiamo testando

Statistica di χ^2 di Pearson

- Se le misure n_i sono distribuite secondo una Gaussiana(): il χ^2 di Pearson segue la distribuzione di χ^2 per N gradi di libertà
- Se le misure n_i sono distribuite secondo Poisson allora

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

Il χ^2 di Pearson segue la distribuzione di χ^2 con N gradi di libertà se tutti i ν_i sono abbastanza grandi perché la distribuzione poissoniana sia approssimabile da una distribuzione gaussiana (in pratica $\nu_i \geq 5$ è sufficiente).

χ^2 per multinomiale

- Se $n_{tot} = \sum n_i$ è fissato (non distribuito Poissonianamente), il test controlla solo la forma della distribuzione e non la normalizzazione totale
- la probabilità in ogni bin è descritta dalla multinomiale: $n_i = p_i n_{tot}$ dove p_i è la probabilità per un evento di essere misurato nel bin i secondo l'ipotesi nulla
- In questo caso il test statistico del χ^2 di Pearson

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

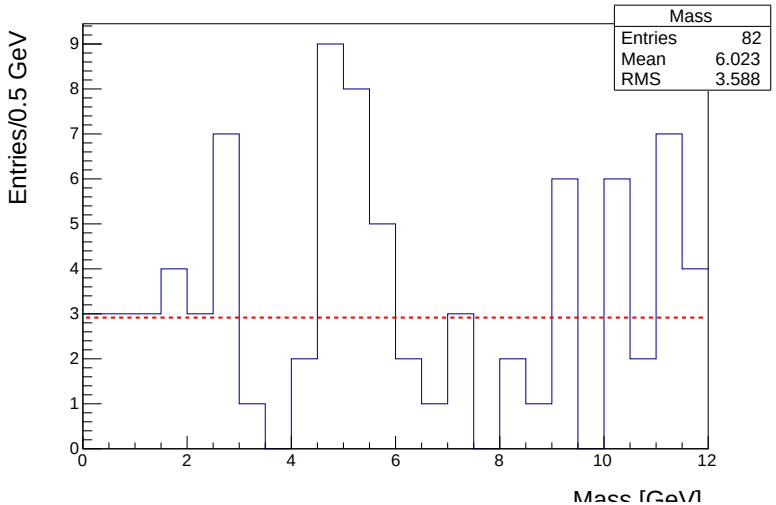
- Da notare che il denominatore non è la varianza $V[n_i] = n_{tot} p_i (1 - p_i)$
- Se tutti $p_i n_{tot} \gg 1$ allora la distribuzione di χ^2 segue la distribuzione di χ^2 per $N - 1$ gradi di libertà.

χ^2 di Pearson: esempio

- Supponiamo di avere un leggero eccesso di eventi (conteggi) su un fondo noto a priori
- Supponiamo di considerare $n_i \sim Poisson(\nu_i)$
  - Calcoliamo la statistica di test:

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

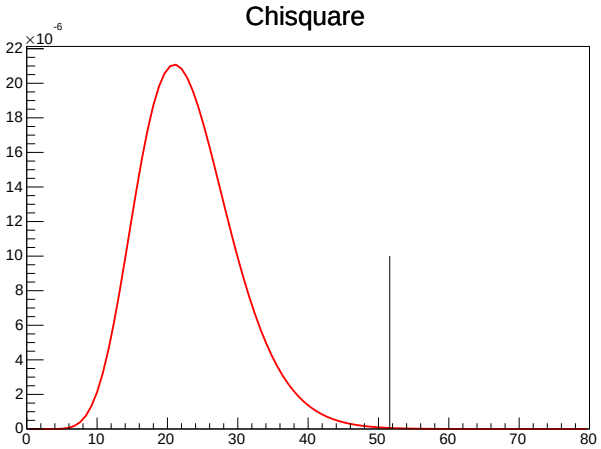
  - Ora vorremmo calcolare il p-value



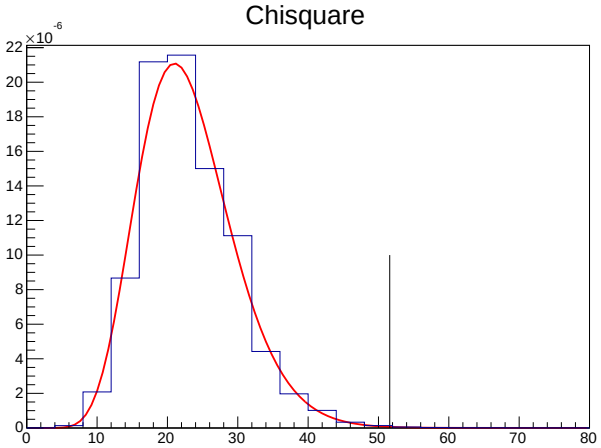
- Ma il numero di eventi in molti bins è piccolo (non è soddisfatta la condizione $\nu_i \geq 5$) per cui ci aspettiamo che il χ^2 non segua la distribuzione di χ^2
- Non si avrà più una distribuzione della variabile in esame che segue quella del χ^2 con N gradi di libertà
- Tuttavia il χ^2 di Pearson resta un test statistico valido, semplicemente non conosciamo la distribuzione

χ^2

- Si può calcolare la distribuzione di χ^2 con metodi Montecarlo
  - Si generano le n_i poissoniane con valor medio ν_i
  - Si calcola il χ^2
  - SI ripete il processo N volte fino ad ottenere una distribuzione del χ^2 dal MonteCarlo
  - Si integra la distribuzione per ottenere il p-value



p = 0.00056

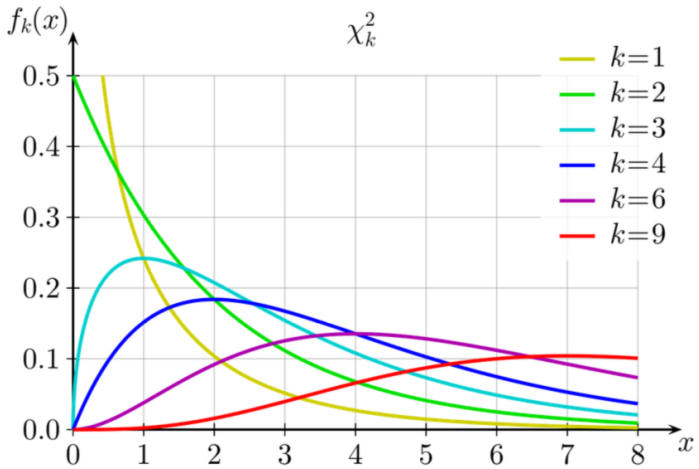


p = 0.0007

Ipotesi H0 (distribuzione uniforme) esclusa al ‰

Distribuzione di χ^2

$$f_{\chi^2}(\chi^2, \mathrm{ndf}) = \frac{1}{2^{\mathrm{ndf}/2}\Gamma(\mathrm{ndf}/2)}((\chi^2)^{\mathrm{ndf}/2-1})\mathrm{e}^{-\chi^2/2} \quad \text{dove} \quad \Gamma(\mathrm{x}) = \int_0^{+\infty} \mathrm{e}^{-\mathrm{t}}\mathrm{t}^{\mathrm{x}-1}\mathrm{dt}$$



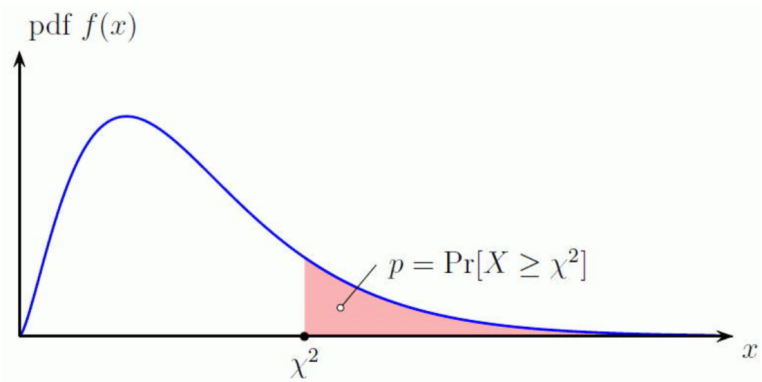
- La distribuzione ha valor medio $E[\chi^2] = \mathrm{ndf}$ e varianza $V[\chi^2] = 2\mathrm{ndf}$
- Spesso si riporta il χ^2 diviso i gradi di libertà come stima della bontà dell'accordo dati/valori aspettati

χ^2

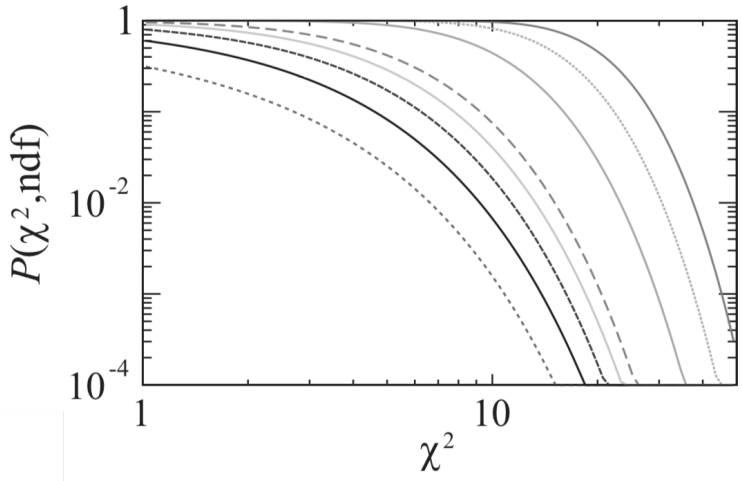
- In realtà è preferibile utilizzare il p-value
- La probabilità di χ^2 misura la probabilità che, data l'ipotesi nulla, un set di misure fornisce un χ^2 quanto, o più elevato di quello ottenuto

$$P(\chi^2) = \int_{\chi^2}^{+\infty} f_{\chi^2}(z)dz$$

- Un valore piccolo del p-value: cattivo accordo dei dati con l'ipotesi testata
- χ^2=TF1::GetChisquare *ndf*=TF1::GetNDF *p*=TF1::GetProb



$\chi^2 = 15$ $N = 10$ $p = 0.13$
$\chi^2 = 150$ $N = 100$ $p = 9 \times 10^{-4}$



ndf = 1, 2, 3, 4, 5, 10, 15, 20

Lo stesso valore di χ^2/N corrisponde a *p*-value molto diversi al variare di N !

Binned Likelihood (Likelihood χ^2)

- Per test di bontà di adattamento per istogrammi possiamo anche usare la distribuzione nota di eventi in un bin (nel caso in cui non sia Gaussiana)
  - Poisson: se i contenuti del bin sono indipendenti (non c'e' vincolo sul numero totale di eventi, il numero totale di eventi non è fissato)
  - Multinomiale: se il numero totale di eventi dell'istogramma è fissato
- La statistica di test sarà la likelihood binnata che si comporta come una distribuzione di χ^2 ($\chi^2 = -2lnL$) con numero di gradi di libertà uguale a $N - r$ dove N è il numero di bin dell'istogramma, r numero di parametri da stimare sui dati nel caso di dati distribuiti secondo una Poissoniana e $N - r - 1$ nel caso di dati distribuiti secondo una Multinomiale

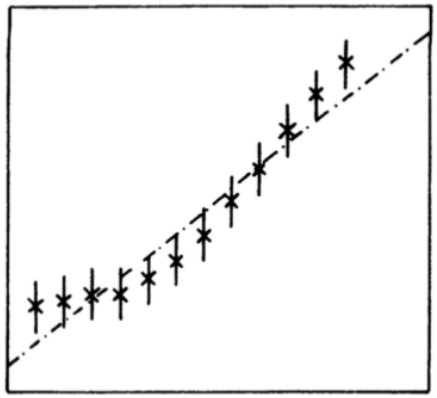
Run test

- Il test di χ^2 è insensibile al segno delle deviazioni
- Un test complementare al test del χ^2 che si basa sul segno delle deviazioni: Run test
- H_0=tutti i pattern di segno hanno uguale probabilità
- Chiamiamo N_+ il numero di deviazioni positive e N_- il numero di deviazioni negative e R il numero di run, cioè una sequenza di bin consecutivi dove i dati mostrano deviazioni dello stesso segno (positivo/negativo)
- Il valore di aspettazione e la varianza di R sono dati:
  $E[R] = 1 + \frac{2N_+N_-}{N_++N_-}$
  $V[R] = \frac{2N_+N_-(2N_+N_- - N_+ - N_-)}{(N_++N_-)^2(N_++N_- -1)} = \frac{(E[R]-1)(E[R]-2)}{N-1}$
- Per una distribuzione con più di 20 bins, il numero di run può essere approssimato da una distribuzione gaussiana per cui la significanza della deviazione di un numero di runs r osservato dal valore atteso è

$$Z = \frac{r - E[r]}{\sqrt{V[r]}}$$

Run test (esempio)

- Dal test del χ^2 fit ok...
- Proviamo ad applicare il Run test che fornisce un'informazione aggiuntiva rispetto al test di χ^2



+++ − − − − − − − +++
$N = N_+ + N_- = 12$ e $N_+ = 6$ e $N_- = 6$
2 "+" runs e 1 "-" run
Per cui se calcolo:

$$E[R] = 1 + \frac{2N_+N_-}{N}$$

$$V(r) = 2N_+N_- \frac{(2N_+N_- - N)}{N^2(N-1)}$$

Ottengo:

$$E[r] = 7 \text{ e } V[r] = 2.73$$

$Z = \frac{r - E[r]}{\sqrt{V[r]}} = \frac{7-3}{\sqrt{2.73}} = 2.4\sigma$ che equivale ad una significatività dell'1%
Ipotesi nulla rigettata: fit non buono!

Test unbinned

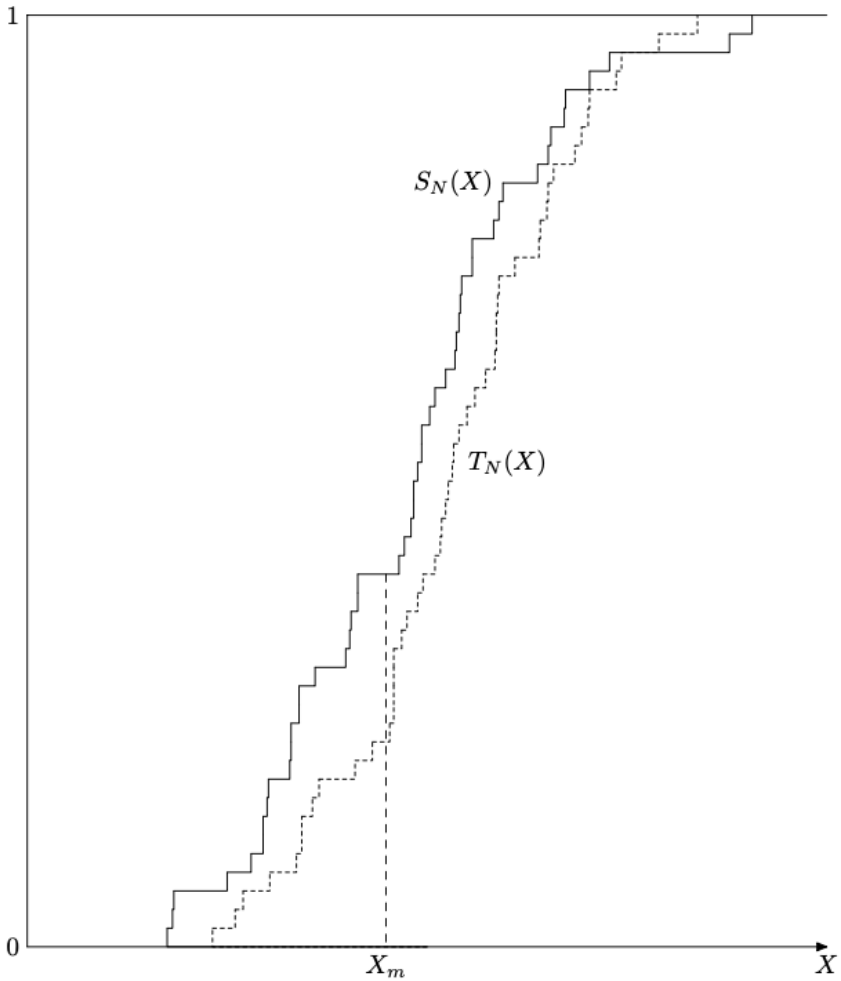
- Abbiamo N osservazioni indipendenti: x_1, ... x_N della variabile aleatoria x
- Riordiniamo le osservazioni in ordine crescente per cui $x_{(1)} \leqslant x_{(2)} \leqslant ... \leqslant x_{(N)}$
- Le osservazioni ordinate $x_{(i)}$ sono chiamate statistica d'ordine
- Si scrive la distribuzione cumulativa empirica $S_N(x)$ per N misure come

$$S_N(x) = \begin{cases} 0 & x < x_{(1)} \\ 1/N & x_{(i)} \leq x < x_{i+1}, i = 1, ..., N-1 \\ 1 & x_n \leq x \end{cases}$$

Test unbinned

- I test unbinned utilizzando si basano sul confronto tra la distribuzione cumulativa della distribuzione di probabilità $F(x) = \int_{-\infty}^{x} f(x')dx'$ sotto l'ipotesi H_0 con l'equivalente distribuzione dei dati $S_N(x)$
- La statistica di test utilizza una qualche differenza tra la distribuzione cumulativa sperimentale e l'ipotetica distribuzione cumulativa: $S_N(x) - F(x)$ cioè misura la "distanza" tra le due funzioni cumulative
- Analogamente nel caso di confronto tra due distribuzioni sperimentali si può utilizzare $S_N(x) - T_N(x)$ dove $T_N(x)$ è la distribuzione cumulativa empirica di un campione di dati

Test unbinned: statistica ordinata



- In questo esempio, la distanza massima $S_N(x) - T_N(x)$ avviene a $x = x_m$
- A seconda di come si misura la "distanza" tra le due funzioni cumulative si hanno diversi tipi di test

Kolmogorov-Smirnov test

- Utilizza la deviazione massima tra la distribuzione osservata $S_N(x)$ e la distribuzione $F(x)$ attesa sotto l'ipotesi H_0
- È definita come:

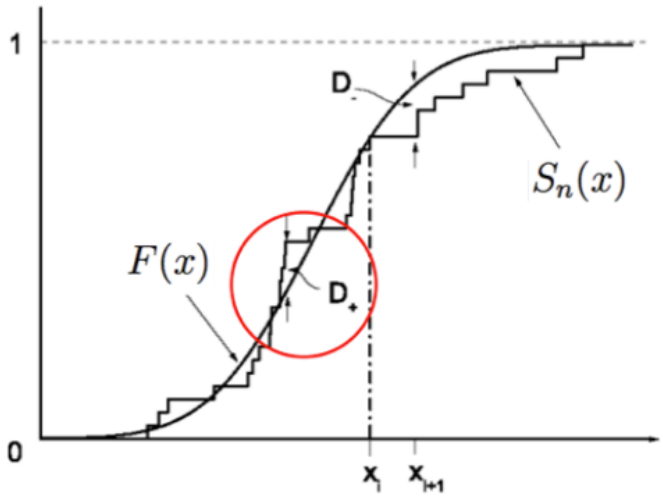
$$D_n = max|S_n(x) - F(x)|$$

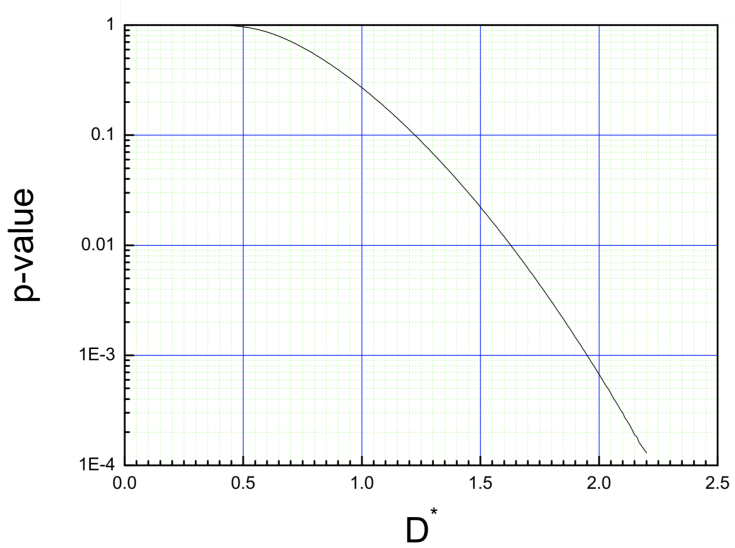
$$D_n^{\pm} = max\{\pm|S_n(x) - F(x)|\}$$

quando si considera test one-side

Test size α	Critical value of $\sqrt{N}D_N$
0.01	1.63
0.05	1.36
0.10	1.22
0.20	1.07

- Al 5%, il valore critico è $D = \frac{1.36}{\sqrt{N}}$ dove N è il numero di misure
- l'ipotesi nulla non è rigettata se $D < D_{\text{crit}}$





$$D^* = \sqrt{N}D$$

Kolmogorov-Smirnov test: 2 campioni

- Il test di KS può essere anche utilizzato per confrontare due campioni di dati $(x_1,...x_n)$ e $(y_1,...y_n)$ per verificare l'ipotesi che entrambi i campioni provengano dalla stessa distribuzione
- L'equivalente statistica per confrontare due distribuzioni cumulative empiriche $S_N(x)$ e $S_M(x)$ è:

$$D_{MN} = max|S_N(x) - S_M(x)|$$

$$D^{\pm}_{MN} = max\{\pm[S_N(x) - S_M(x)]\}$$

per il test one-side

Kolmogorov-Smirnov test: osservazioni

- I test unbinned basandosi sulla statistica d'ordine (per cui una distribuzione ha dei valori ordinabili secondo un qualche criterio), non sono facilmente estendibili a N dimensioni per cui sono applicabili a dati che dipendono da un'unica variabile aleatoria
- Se alcuni parametri della distribuzione $f(x)$ sono determinati dal campione di dati x_1,...x_n (i.e. procedura di fit), il test non può essere applicato (non c'è un qualcosa di equivalente al numero di gradi di libertà come nel caso del test di χ^2)
- Ha d'altra parte tutti i vantaggi di un test unbinned: più potente rispetto ai test binned, possibile effettuarlo anche con un numero piccolo di dati

Likelihood unbinned

- La likelihood unbinned non fornisce strumenti per la bontà del fit (al contrario del χ^2)
- Potremmo pensare che il valore di $-ln\mathcal{L}$ possa essere una buona statistica di test per il test di GOF: sappiamo che il massimo della likelihood fornisce la miglior stima dei parametri
- Ma nella stima di maximum likelihood, stiamo usando la likelihood per un set di dati fissi in funzione dei parametri, mentre i test di GOF, usiamo la likelihood per un'ipotesi fissa in funzione dei dati
- Vedremo un esempio in cui può essere visto facilmente che la likelihood non ha nessun poter come statistica di test per i test di GOF poichè tutti i campioni di dati presi in considerazione avrebbero lo stesso valore di likelihood, indipendentemente da quanto seguono la distribuzione sotto l'ipotesi H_0

Likelihood unbinned

Esempio: fit esponenziale (vita media)

$$\mathcal{L}(\lambda, x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda} x_i = \lambda^n e^{-\lambda \sum x_i}$$

$$\frac{d\mathcal{L}(\lambda, x_1, x_2, ..., x_n)}{d\lambda} = n\lambda^{n-1} e^{-\lambda \sum x_i} - \sum x_i \lambda^n e^{-\lambda \sum x_i} = 0$$

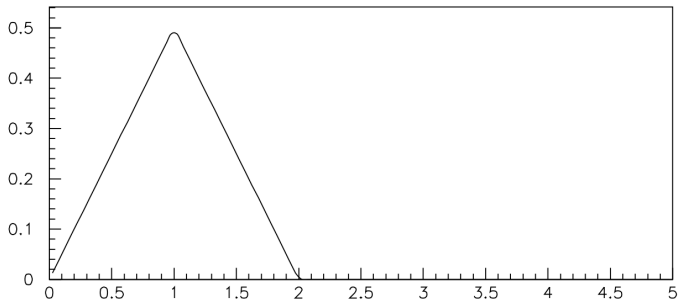
$$n - (\sum x_i)\lambda = 0 \quad \rightarrow \lambda = \frac{n}{\sum x_i}$$

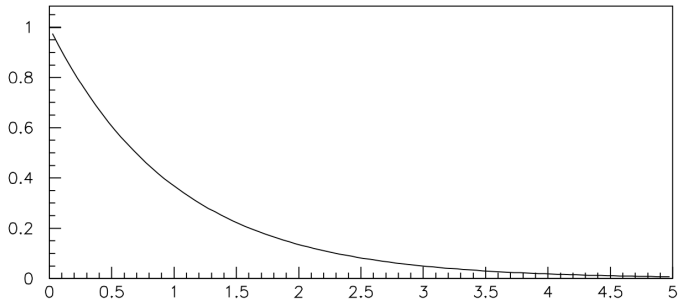
Sostituendo ottengo

$$\begin{aligned} -2\ln(\mathcal{L})_{min} &= -2n\ln(\lambda) + 2\lambda \sum x_i \\ &= -2n\ln(\lambda) + 2n = 2n(1 - \ln(\lambda)) = f(n, \lambda) \end{aligned}$$

Siccome $-2\ln(\mathcal{L})$ è funzione unicamente di n e λ non può essere utilizzata come test di bontà del fit.
Quindi qualunque distribuzione con valor medio λ e stesso numero di eventi n danno lo stesso valore di likelihood

Likelihood unbinned





Poichè entrambe le distribuzioni hanno la stessa media forniscono lo stesso valore di $-2\ln(\mathcal{L})_{min}$ quando fittate con un'esponenziale.
Tutti i set di dati con stessa media $1/\lambda$ forniscono lo stesso valore di likelihood indipendentemente dalla distribuzione. Per il test di likelihood, significherebbe che entrambe le distribuzioni sono un buon fit con un'esponenziale!
La likelihood unbinned non fornisce strumenti per la bontà del fit!

```
Double_t TH1::Chi2Test ( const TH1 *  h2,
                         Option_t *   option = "UU",
                         Double_t *   res = 0
                       )              const                    virtual
```

χ^2 test for comparing weighted and unweighted histograms

Function: Returns p-value. Other return values are specified by the 3rd parameter

Parameters

[in] **h2** the second histogram

[in] **option**

- "UU" = experiment experiment comparison (unweighted-unweighted)
- "UW" = experiment MC comparison (unweighted-weighted). Note that the first histogram should be unweighted
- "WW" = MC MC comparison (weighted-weighted)
- "NORM" = to be used when one or both of the histograms is scaled but the histogram originally was unweighted
- by default underflows and overflows are not included:
  - "OF" = overflows included
  - "UF" = underflows included
- "P" = print chi2, ndf, p_value, igood
- "CHI2" = returns chi2 instead of p-value
- "CHI2/NDF" = returns χ^2/ndf

[in] **res** not empty - computes normalized residuals and returns them in this array

scipy.stats.ks_2samp

scipy.stats.ks_2samp(*data1*, *data2***)** **[source]**
(http://github.com/scipy/scipy/blob/v0.15.1/scipy/stats/stats.py#L3966)

Computes the Kolmogorov-Smirnov statistic on 2 samples.

This is a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.

Parameters: **a, b** *: sequence of 1-D ndarrays*

two arrays of sample observations assumed to be drawn from a continuous distribution, sample sizes can be different

Returns: **D** *: float*

KS statistic

p-value *: float*

two-tailed p-value

scipy.stats.kstest

scipy.stats.kstest(*rvs*, *cdf*, *args=()*, *N=20*, *alternative='two-sided'*, *mode='approx'*) **[source]**
(http://github.com/scipy/scipy/blob/v0.14.0/scipy/stats/stats.py#L3307)

Perform the Kolmogorov-Smirnov test for goodness of fit.

This performs a test of the distribution G(x) of an observed random variable against a given distribution F(x). Under the null hypothesis the two distributions are identical, G(x)=F(x). The alternative hypothesis can be either 'two-sided' (default), 'less' or 'greater'. The KS test is only valid for continuous distributions.

Parameters:

rvs *: str, array or callable*

If a string, it should be the name of a distribution in scipy.stats (../stats.html#module-scipy.stats). If an array, it should be a 1-D array of observations of random variables. If a callable, it should be a function to generate random variables; it is required to have a keyword argument *size*.

cdf *: str or callable*

If a string, it should be the name of a distribution in scipy.stats (../stats.html#module-scipy.stats). If *rvs* is a string then *cdf* can be False or the same as *rvs*. If a callable, that callable is used to calculate the cdf.

args *: tuple, sequence, optional*

Distribution parameters, used if *rvs* or *cdf* are strings.

N *: int, optional*

Sample size if *rvs* is string or callable. Default is 20.

alternative *: {'two-sided', 'less','greater'}, optional*

Defines the alternative hypothesis (see explanation above). Default is 'two-sided'.

mode *: 'approx' (default) or 'asymp', optional*

Defines the distribution used for calculating the p-value.

- 'approx' : use approximation to exact distribution of test statistic
- 'asymp' : use asymptotic distribution of test statistic

Returns:

D *: float*

KS test statistic, either D, D+ or D-.

p-value *: float*

One-tailed or two-tailed p-value.

Hands on

Riassunto GOF test: Binned

Dipendono dall'arbitrarietà del binning

- χ^2:
  - I dati devono essere distribuiti gaussianamente in ciascun bin (applicabile se il numero di eventi per ogni bin è maggiore di 5)
  - Segue la distribuzione di χ^2 con ndf = N gradi di libertà per la Poissoniana e ndf = N-1 per la multinomiale (N = numero di bin dell'istogramma)
  - Se la distribuzione dipende da *r* parametri da stimare sui dati: $\sim \chi^2(\mathrm{ndf} - \mathrm{r})$
- Likelihood binnata ($\chi^2 = -2\ln\lambda$)
  - I dati non devono essere distribuiti gaussianamente in ciascun bin (minori di 5)
    - distribuiti Poissonianamente se il contenuto dei bin è indipendente (no vincoli sul numero totale di eventi)
    - distribuiti secondo una Multinomiale (se il numero totale degli eventi nell'istogramma è fissato)
- Per numero di eventi piccoli è sempre comunque consigliato utilizzare un metodo MC per ottenere la distribuzione di χ^2

Riassunto GOF test: Unbinned

- Kolmogorov-Smirnov test
  - Unbinned
  - Funziona anche con numeri piccoli di eventi
  - Implementazioni disponibili standard solo per distribuzioni continue (non discrete) [anche se il test è estendibile a distribuzioni discrete]
  - Non applicabile quando i parametri della distribuzione sono stimati dal campione di dati
- Likelihood non binnata: non fornisce GoF intrinseca

Riassunto sui metodi di fit (GOF)

- χ^2:
  - GOF ok
  - Dipende dal binning
  - I dati devono essere distribuiti gaussianamente in ciascun bin (applicabile se i dati per ogni bin sono maggiori di 5)
- Likelihood binnata:
  - GOF ok (χ^2 modificato)
  - Dipende dal binning
  - I dati non devono essere per forza distribuiti gaussianamente in ciascun bin (minori di 5)
- Likelihood non binnata
  - Molto più performante, non richiede che i dati siano binnati
  - No GOF intrinseca
  - I dati devono essere binnati per confrontarli con la funzione

La Likelihood in generale non fornisce strumenti per la bontà del fit (al contrario del χ^2)