Fit e test del Chi Quadro

In questa lezione conclusiva concludiamo l'argomento fit affrontato nella lezione precedente. Abbiamo visto con R come fare un fit di un insieme di misure di due variabili dipendenti (x,y) e di una distribuzione di n misure di una singola variabile x. L'algoritmo di fit ci permette di estrarre i parametri della funzione *f(x)* nel caso di due variabili e i parametri della funzione densità di probabilità nel caso della singola variabile. Determinati i parametri non sappiamo però quanto bene i dati misurati si adattano alla funzione fittata (quella che contiene i parametri risultati dal fit). Potrebbe infatti accadere che i dati si adattano bene, quindi le misure svolte sono significative oppure i dati non si adattano affatto alla funzione teorica attesa e questo ci può portare a concludere che i dati sono da scartare e l'esperimento deve essere ripetuto, oppure il modello teorico che stiamo usando non è quello che realmente descrive il comportamento delle variabili che stiamo misurando.

Per valutare quanto i dati misurati si adattano bene alla funzione teorica ipotizzata possiamo usare il test del chi quadrato. Il Chi Quadrato (indicato sempre con χ^2) è un numero che ci fornisce una indicazione di quanto i dati siano vicini alla funzione/distribuzione teorica attesa. Nell'uso del test del chi quadro dobbiamo però distinguere i casi in cui si stiano trattando con coppie di variabili correlate (x,y) dal caso in cui si tratti la distribuzione di una singola variabile. Affrontiamo quindi nel seguito i due casi separatamente.

Test del χ^2 per una coppia (x,y) di variabili correlate

Quando da un esperimento si cerca la correlazione fra due variabili (x,y) si effettuano le misure della variabile dipendente y in corrispondenza di diversi valori di y, ottenendo così un insieme di dati (x_i , y_i) i quali rappresentano nel piano x-y un numero di punto pari alle misure eseguite.

Dal modello teorico che descrive le variabili (x,y) ci aspettiamo una dipendenza di y da x secondo una certa funzione $y = f(x)$. La funzione $f(x)$ conterrà in generale un certo numero di parametri (la retta per esempio contiene 2 parametri, una esponenziale ne può contenere 3 etc etc) che saranno determinati dal fit. La procedura di fit infatti cerca quei valori dei parametri tali per cui la funzione si avvicina di più ai punti (x,y) misurati nell'esperimento.

Supponiamo che i nostri dati siano (x_i , y_i), che la variabile y_i sia stata misurata con un errore σ_i e che il modello teorico preveda una dipendenza da una qualche funzione $y = f(x)$.

Anzitutto determiniamo i parametri della funzione con il fit pesato, pesato significa che dobbiamo inserire gli errori σ_i. Più l'errore su una singola misura di y è piccolo più questa misura nella procedura di fit è importante, in quanto significa che è stata misurata con maggiore precisione. Nel comando per il fit in R si deve quindi specificare il vettore pesi. Capiamo che più piccolo è l'errore maggiore è il peso, quindi il vettore dei pesi sarà qualcosa di inverso all'errore. In generale il peso si definisce come l'inverso del quadrato dell'errore:

$$w_i = \frac{1}{\sigma_i^2}$$

Eseguito il fit pesato otteniamo i parametri fittati con i quali costruiamo la funzione fittata $f_{fit}(x)$, con la quale possiamo calcolare i valori di y fittati che ci aspettiamo dal modello teorico. Valutando le distanze dei valori misurati dai valori fittati e tenendo conto dei pesi possiamo costruire una variabile che indicherà quanto i valori misurati si adattano bene al fit. Costruiamo quindi il chi quadro con la seguente definizione:

$$\chi^2 = \sum_{i=1}^{N} w_i \left(y_i^{mis} - f(x_i) \right)^2$$

Dove N è il numero di misure analizzate. Se il valore di χ^2 risulta minore o circa uguale al numero di misure allora i punti si adattano bene al modello teorico usato. Teniamo presente che nel caso di accordo perfetto fra dati e modello teorico (situazione inverosimile in un esperimento) allora il chi quadro sarà uguale a 0.

Insieme al χ^2 si è anche soliti definire il chi quadro ridotto come segue:

$$\tilde{\chi}^2 = \frac{1}{d} \sum_{i=1}^{N} w_i \left(y_i^{mis} - f(x_i) \right)^2$$

Dove d è il *numero di gradi di libertà*, ossia il numero di misure analizzate meno il numero di parametri fittati:

$$d = N - N_{PAR}$$

Nel caso della retta per esempio il numero di parametri è 2 (intercetta e coefficiente angolare), nell'esempio di esponenziale che abbiamo visto nella lezione del 26/05/2009 i numero di parametri è 3.

Se il valore del $\tilde{\chi}^2$ è minore o circa uguale a 1 allora i dati si adattano bene al modello teorico utilizzato. Il chi quadro ridotto è grosso modo un chi quadro normalizzato, visto che come valore massimo indicativo ha 1.

Calcolato il χ^2_{fit} o il $\tilde{\chi}^2_{fit}$ (ovviamente i due sono equivalenti, basta sapere quale si sta usando) non si possiamo però valutare in modo soggettivo se il valore indica un buon adattamento oppure no. Supponiamo per esempio di ottenere $\tilde{\chi}^2_{fit} = 1.4$, con questo numero scartiamo l'ipotesi del nostro modello? Se fosse 10 sarebbe sicuramente peggio di 1.4, ma visto che questo è di poco superiore a 1 scartiamo tutto? Per concludere qualcosa in merito ai dati si deve procedere in modo oggettivo e per fare questo si deve valutare la probabilità di ottenere un valore di $\chi^2 > \chi^2_{fit}$. Se la probabilità è alta significa che è molto più probabile misurare dati che si adattano peggio di quelli in esame, quindi quelli in esame possono ritenersi buoni. Differentemente se la probabilità è bassa, allora vuol dire che il campione di dati in esame non è buono e si deve o ripetere l'esperimento oppure valutare se il modello teorico scelto è giusto o corretto.

Come valore limite standard per la probabilità di χ^2 tipicamente si scegli il 5%, quindi oltre il 5% i dati si tengono buoni, meno del 5% i dati si scartano.

A seconda dei casi si possono comunque scegliere limiti più stringenti per la probabilità.

Il valore della probabilità si può determinare da tabelle che riportano le probabilità in corrispondenza del valore di χ^2_{fit} o di $\tilde{\chi}^2_{fit}$ e del numero di gradi di libertà *d*. E' molto importante fare attenzione al numero di gradi di libertà. La probabilità si può calcolare integrando la funzione densità di probabilità $\chi^2(x,d)$, oppure con l'uso di software si cerca la funzione che restituisce la probabilità cercata per di χ^2_{fit} o di $\tilde{\chi}^2_{fit}$ e *d*.

Nota al caso di assenza di errori in *y*

Premesso che non è possibile da un esperimento avere misure senza incertezza, tutto ciò che riportiamo in questa breve nota potrebbero non essere corretto. Nel caso in cui gli errori delle misure y_i non siano conosciuti non è chiaro come si possa definire un chi quadro pesato. La procedura di fit ci permette di stimare un errore medio sulle misure di *y* facendo proprio riferimento alla distanza media dalla funzione fittata, ma questo è molto simile al chi quadro ridotto, quindi il risultato sarebbe proprio qualcosa di molto vicino a 1 per definizione. La differenza da 1 dipende dal numero di gradi di libertà in quanto la definizione di errore medio è:

$$\sigma_y^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i^{mis} - f(x_i)\right)^2$$

Quindi il chi quadro diventa:

$$\chi^2 = \sum_{i=1}^{N} w_i\left(y_i^{mis} - f(x_i)\right)^2 = \sum_{i=1}^{N}\frac{1}{\sigma_y^2}\left(y_i^{mis} - f(x_i)\right)^2 = \frac{N-2}{\sum_{i=1}^{N}\left(y_i^{mis} - f(x_i)\right)^2}\sum_{i=1}^{N}\left(y_i^{mis} - f(x_i)\right)^2 = N-2$$

Allora il chi quadro ridotto diventa: $\tilde{\chi}^2 = \frac{N-2}{N-N_{PAR}}$

Nel caso di un andamento lineare è proprio uguale a 1, quindi non stiamo affatto tenendo conto di quanto i dati distano dalla funzione fittata.

1. Un modo potrebbe essere quello di scegliere come peso il valore 1 e valutare eventualmente insieme a qualche coefficiente di correlazione e agli errori sui parametri la bontà del fit. Se in R si esegue un fit senza specificare i pesi, allora il RSE viene calcolato con peso = 1.

2. Un'altra alternativa è quella di fare riferimento alla definizione di chi quadro per le distribuzioni, che vedremo nel seguito, data da questa relazione:

$$\chi^2 = \sum_{i=1}^{N}\frac{\left(y_i^{mis} - f(x_i)\right)^2}{\left|f(x_i)\right|}$$

Con questa definizione stimiamo una somma di variazioni dal valore teorico atteso relative a questo stesso valore, quindi sembrerebbe essere la definizione più sensata.

3. Un'altra idea potrebbe essere quella di stimare gli errori sui valori teorici stimati facendo la propagazione degli errori sui parametri ottenuti dal fit. In realtà per applicare questo metodo bisognerebbe verificare di non riottenere un risultato assurdo come quello visto nella parte introduttiva di questa nota. Comunque se la funzione dipende da N_{PAR} parametri P_i, allora si può stimare l'errore sul valore atteso di y come segue:

$$\sigma_{y_i} = \sum_{j=1}^{N_{PAR}}\left|\frac{\partial f\left(x_i, P_{1,\ldots}, P_N\right)}{\partial P_j}\sigma_{P_j}\right|$$

Si dovrebbero quindi poter usare questi errori come pesi per il calcolo del chi quadro.

È importante ribadire che senza errori le misure non hanno molto senso, quindi poco senso ha anche eseguire il test del chi quadro. In questo caso comunque il metodo più sensato sembrerebbe essere il secondo, anche se nel terzo il chi quadro dovrebbe tendere ad essere maggiore di quello calcolato al punto 2. Per essere tranquilli forse conviene calcolare i tre valori di chi quadro corrispondenti ai tre metodi e alla fine usare quello maggiore, al fine di essere conservativi, per stimare la bontà del fit.

Uso di R per il Test del χ^2 per una coppia (x,y) di variabili correlate

Definiti x, ymis, symis (errori su y) e la funzione con cui fittare, calcoliamo il vettore pesi come

w=1/symis^2

quindi eseguiamo il fit pesato:

lm(ymis ~ x, weights=w)

oppure con

*fitnls <- nls(ymis ~ Amis * x + Bmis, start = list(Amis = 1, Bmis = 1), weights=w)*

richiamiamo le informazioni sui con il commando summary:

summary(fitnls)

notiamo che a un certo punto compare l'imformazione:

Residual standard error: 1.03 on 8 degrees of freedom

Il Residual Standard Error calcolato su 8 gradi di libertà (10 misure – 2 parametri nel caso della retta) che chiameremo RSE_d altro non è che la radice quadrata del chi quadro ridotto:

$$RSE_d = \sqrt{\frac{1}{d}\sum_{i=1}^{N} w_i\left(y_i^{mis} - f(x_i)\right)^2} = \sqrt{\tilde{\chi}^2}$$

Verifichiamo in R questa cosa calcolando chi quadro ridotto come riportato nell'esercizio allegato.

Ora chiediamo a R quale è la probabilità di avere un chi quadro maggiore di quello calcolato con il numero di gradi di libertà del fit. Per fare questo usiamo la funzione *pchisq*, la quale restituisce la probabilità di avere un $\chi^2 < \chi^2_{fit}$ o $\chi^2 > \chi^2_{fit}$ per un certo numero di gradi di libertà. Bisogna fare molta attenzione in quanto questa funzione vuole come parametri d'ingresso il chi quadro e non il chi quadro ridotto e per richiedere la probabilità $P_d\left(\chi^2 > \chi^2_{fit}\right)$ bisogna aggiungere una opzione come segue:

pchisq(χ^2_{fit} , d, lower.tail = FALSE)

Useremo quindi

pchisq($RSE^2 \times d$, d, lower.tail = FALSE)

visto che il fit restituisce RSE.

In base alla probabilità che otteniamo possiamo decidere se i dati si accordano bene con il modello teorico oppure no.

Test del χ^2 per una distribuzione di una variabile misurata x

Quando da un esperimento si misura ripetutamente una singola variabile, di questa se ne vuole studiare la distribuzione e la si vuole confrontare con la funzione densità di probabilità (*pdf – probability density function*) prevista del modello teorico. Per esempio *x* può distribuirsi secondo una gaussiana, una poissoniana, una binomiale etc etc.

Nella procedura di fit vogliamo quindi estrarre i parametri della *pdf* e per determinare quanto i dati si adattano bene a questa distribuzione possiamo eseguire il test del chi quadro. In questo caso però il chi quadro è definito in modo diverso dal precedente, in quanto non dobbiamo più valutare la differenza fra y misurato e y fittato, ma dobbiamo confrontare le frequenze di *x* negli intervalli Δx_i (che chiamiamo frequenze

osservate O_i) con le frequenze attese E_i, determinate dalla *pdf* fittata. Il chi quadro risulta essere quindi definito come segue:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \text{e} \qquad \tilde{\chi}^2 = \frac{1}{d} \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \begin{array}{l} n = \text{numero di classi} \\ d = n - N_{PAR} \end{array}$$

Dove n è il numero di classi, o intervalli (suddivisione dell'istogramma), nelle quali sono state raggruppate le misure di x e d è il numero di gradi di libertà definito come la differenza fra il numero di classi e il numero di parametri determinati dal fit.

Distinzione fra variabili continue e variabili discrete

Nel momento in dobbiamo valutare le frequenze attese con la pdf fittata dobbiamo fare attenzione se stiamo lavorando con variabili discrete (cioè che possono assumere solo numeri interi), come potrebbe essere una variabile distribuita secondo una binomiale, oppure con variabili continue (cioè che possono assumere tutti i valori reali), come potrebbe essere una variabile distribuita secondo una gaussiana.

Nel caso di variabili continue la frequenza attesa, o probabilità se normalizzata, per quel dato valore di x_i sarà dato dal valore della pdf in corrispondenza di x_i:

$$E_i = P(x_i) = pdf(x_i)$$

Nel caso di variabili continue invece la probabilità relativa all'intervallo Δx_i (di estremi $x_{i,\min}$ e $x_{i,\max}$) è dato dall'integrale della pdf nell'intervallo considerato:

$$E_i = P(x_{i,\min} < x \le x_{i,\max}) = \int_{x_{i,\min}}^{x_{i,\max}} pdf(x)dx$$

Determinate le frequenze attese e calcolato il chi quadro si procede come visto in precedenza con la valutazione della probabilità di chi quadro per valutare la qualità del fit.

Uso di R per il Test del χ^2 per la distribuzione di una variabile x

Supponiamo di aver eseguito N=1000 misure di una variabile x che in R chiamiamo dati. Possiamo visualizzarne la distribuzione con un istogramma:

hist(dati)

Se non specificato R procede con una suddivisione in classi che ritiene più adatta.
Per fare il fit di una distribuzione R mette a disposizione uno strumento chiamato fitdistr, come abbiamo già visto nella lezione del 26/05/2009:

fitdistr(dati, "normal") #nell'ipotesi in cui i nostri dati debbano distribuirsi secondo una gaussiana (normale)

dal summary di questo fit non viene visualizzato nessun RSE o chi quadro, in quanto per le distribuzioni R mette a disposizione uno strumento per il test del chi quadro con il comando chisq.test, nel quale si devono specificare le frequenze osservate e le probabilità attese.

Calcoliamo quindi nell'esempio le frequenze attese o usando le funzioni specifiche oppure integrando la funzione pdf.

Usando chisq.test ci si accorge anzitutto che non si possono specificare il numero di gradi di libertà, quindi i risultati che restituisce devono essere valutati attentamente. Notiamo inoltre che in questo test dobbiamo fornire noi le frequenze osservate e quelle misurate, quindi il comando si limita a calcolare il chi quadro e a determinarne la probabilità, tra l'altro senza considerare il numero di gradi libertà corretto. A questo punto appare chiaro che il chi quadro possiamo calcolarlo senza problemi visto che abbiamo già tutti gli ingredienti necessari. Determinare poi la probabilità di chi quadro abbiamo visto essere semplice con il comando pchisq come nell'esempio precedente.

Concludiamo quindi che il comando chisq.test di R non è molto utile per gli scopi qui visti e probabilmente nemmeno esatto, in quanto con il calcolo a mano otteniamo risultati migliori.

La comodità di fittare una distribuzione con fitdistr, anziché con nls, sta nel fatto che se volessimo fittare una distribuzione con nls dovremmo costruire l'istogramma estrarre una variabile dipendente y che corrisponderebbe alle frequenze, o densità, quindi costruire due variabili (x,y) e fare il fit. In fitdistr c'è anche la possibilità di fittare con una funzione personalizzata, si veda l'esempio riportato nel manuale:

```
mydt <- function(x, m, s, df) dt((x-m)/s, df)/s
fitdistr(x2, mydt, list(m = 0, s = 1), df = 9, lower = c(-Inf, 0))
```