Test esatto di Fisher in Matlab

Prima di iniziare:

- Inviare una mail all'indirizzo :

rmgmrc@unife.it

- Scaricare i files per la lezione di oggi dal sito:

http://www.unife.it/ing/lm.meccanica/insegnamenti/statistica-e-modelli-di-dati-sperimentali

Sezione materiale didattico->MATLAB

Creare la tabella di contingenza

- Individua la frequenza con cui valori di due variabili aleatorie vengono associati.
- La funzione «**crosstab**» permette di valutare una tabella di contingenza a partire da due vettori numerici. Individua automaticamente i valori osservabili per ogni vettore e ne conta la frequenza

	x	y
caso 1	x_1	y_2
caso 2	x_1	y_1
caso 3	x_2	y_2
caso 4	x_2	y_1
caso 5	x_1	y_1
....		
caso n	x_1	y_1

	y_1	y_2
x_1	# casi con x_1 & y_1	# casi con x_1 & y_2
x_2	# casi con x_1 & y_1	# casi con x_2 & y_2

COMANDI MATLAB (1)

- Aprire il dataset «dati2x2.mat»

- Eseguire il comando:

```
>> m=crosstab(x,y)

m =

    10     8
     7     5
```

Come calcolare la probabilità di soglia

Data una generica tabella di contingenza

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

la probabilità di ottenere un conteggio $\boldsymbol{a}$ di $\boldsymbol{y_1}$ tra i gli elementi x1 nella eventualità in cui si presentino indifferentemente nei casi $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}};$$

Es: y==salute (malato/sano) & x==genere(m/f), p indica la probabilità di osservare a malati maschi in un campione statistico di n=a+b+c+d soggetti dei quali a+c sono malati, nel caso in cui il morbo non faccia differenze di genere

COMANDI MATLAB (2)

- La formula per la probabilità di soglia può essere calcolata in MATLAB con l'utilizzo della funzione:

$$nchoosek(n,k) = \binom{n}{k};$$

- Eseguire i comandi:

```
>> a=m(1,1);
>> a=m(1,1);
>> b=m(1,2);
>> c=m(2,1);
>> d=m(2,2);
>> p_cut1=nchoosek(a+b,a)*nchoosek(c+d,c)/nchoosek(a+b+c+d,a+c)

p_cut1 =0.2894
```

COMANDI MATLAB (3)

- La probabilità di soglia può essere calcolata in MATLAB direttamente dalla funzione di densità di probabilità con l'utilizzo della funzione:

$$pdf('Hypergeometric', \#successi, \#elementi\ campione, \#elementi\ favorevoli, \#estrazioni)$$

- Eseguire i comandi:

```
>> p_cut2=pdf('Hypergeometric',a,a+b+c+d,a+c,a+b)

p_cut2 =    0.2894
```

Individuare le tabelle di contingenza ‘‘equivalenti’’

- Per proseguire occorre individuare le combinazioni per le quali le somme degli elementi in riga e in colonna siano invarianti (ovvero i possibili casi esistenti con lo stesso numero totale di casi $\boldsymbol{y}_{1\text{-}2}$ e lo stesso numero di estrazioni di $\boldsymbol{x}_{1\text{-}2}$)

- Es: continuando a considerare x==(maschio/femmina) & y==(malato/sano), individuo tutti i possibili casi (combinazioni) per i quali osservo lo stesso numero di malati e di sani, senza cambiare la percentuale di maschi e femmine nel campione.

COMANDI MATLAB (4)

- Utilizzando dei cicli è facile ottenere tutte le possibili combinazioni di tabella di contingenza con somma degli elementi in riga invariante, selezionando tra queste solo quelle che mantengono invariata la somma degli elementi in colonna (o viceversa).

- Eseguire il comando:

```
>> matrix_eq

R =(18 ; 12)
C = (17   13)
n_matrix =  12
```

Calcolare la probabilità dei casi

- Per ogni tabella individuata, va calcolata la probabilità come nel caso originale

- Le combinazioni con probabilità uguale o inferiore vanno sommati, e indicano la probabilità di errore nello scartare l'ipotesi che le variabili *x* e *y* siano tra loro indipendenti

COMANDI MATLAB (5)

- Eseguire il comando:

```
>> p_fisher

p1 =    1.0000

p2 =  1.0000
```

- Nb: lo script esegue i calcoli della probabilità in entrambe le modalità precedentemente spiegate. I due metodi sono equivalenti.

Matlab aiuta: funzione «fishertest»

- Matlab possiede una funzione che a partire da una tabella di contingenza è in grado di eseguire il test esatto di Fisher

$$[esito\ test(1-0), probabilità] = fishertest(tabella\ di\ contingenza)$$

- Nell'argomento della funzione è possibile specificare il parametro di rischio

$$fishertest(\ldots,'Alpha',\alpha)$$

- È inoltre possibile cambiare la ipotesi iniziale Ho da indipendenza tra variabili per verificare se la eventuale dipendenza tra v.a. aumenta (*"right"*) o diminuisce (*"left"*) la probabilità di osservare $\boldsymbol{y_1}$ in un campione $\boldsymbol{x_1}$

$$fishertest(\ldots,'right'\ or\ 'left',\ldots)$$

COMANDI MATLAB (6)

- Eseguire il comando:

```
[h,p]=fishertest(m)

h = 0
p =1.0000
```

- Nb: h=o quindi il test non rigetta la indipendenza delle due variabili
- Nb: h in nel workspace di MATLAB non è un numero intero ma una valore logico
- Nb: utile per controllare velocemente risultato esercizio (da formulazione della tabella di contingenza in poi)

Test di Fisher per variabili con >2 valori possibili

- Il test esatto di Fisher è applicabile anche nei casi in cui le variabili aleatorie abbiano più di 2 possibili valori osservabili

- La probabilità di soglia in questo caso viene derivata dalla distribuzione ipergeometrica generalizzata

$$p = \frac{\prod_{i=1} \binom{\#elementi\ favorevoli_i}{\#\ successi_i}}{\binom{dimensione\ campione}{\#estrazioni}} \rightarrow \frac{\prod_{i=1} \binom{somma\ elementi\ riga\ i}{elemento\ colonna\ 1\ riga\ i}}{\binom{dimensione\ campione}{somma\ elementi\ prima\ colonna}}$$

COMANDI MATLAB (7)

- La implementazione in Matlab è analoga a quanto affrontato nel caso precedente, per la risoluzione è utilizzato uno script dal sito ufficiale di file exchange di Matlab (dettagli in appendice 3)

```
>> [H,P]=FisherExactTest(x,y)

H =    1

P = 4.1989e-04
```

- Nb: H= 1 quindi il test rigetta la indipendenza delle due variabili

APPENDICE 1: script «matrix_eq»

```
R=sum(m,2)  %calcolo la somma delle righe della tabella di contingenza
C=sum(m)    %calcolo la somma delle colonne della tabella di contingenza
n_matrix=1; %inizializzo indice matrice delle tabelle di contingenza equivalenti

%ciclo per calcolare tutte le possibili tabelle di contingenza equivalenti
for a=1:R(1)
   b=R(1)-a;   %individuo elementi prima riga in modo che la loro somma non cambi
   for c=1:R(2);
      d=R(2)-c;   %individuo elementi prima riga in modo che la loro somma non cambi
      if a+c==C(1) && b+d==C(2) && a~=m(1,1) && d~=m(2,2)    %verifico che la somma delle colonne non cambi\non ripetere la tabella iniziale
        M(:,:,n_matrix)=[a  b; c d];
        n_matrix=n_matrix+1;
      end
   end
end
n_matrix    %numero totale di combinazioni, tabella di contingenza originale compresa
```

APPENDICE 2: script «p_fisher»

```
p1=p_cut1;   %inizializzo la probabilità (1)
p2=p_cut2;  %inizializzo la probabilità (2)
%ciclo per tutte le tabelle di contingenza equivalenti
for n=1:size(M,3)
   a=M(1,1,n);    %calcolo elementi tab (solo per maggior chiarezza nei passaggi successivi)
   b=M(1,2,n);
   c=M(2,1,n);
   d=M(2,2,n);

   test1=nchoosek(a+b,a)*nchoosek(c+d,c)/nchoosek(a+b+c+d,a+c);   %calcolo la prob per la tabella selezionata
   test2=pdf('Hypergeometric',a,a+b+c+d,a+b,a+c);

   if test1<=p_cut1  %verifico se la prob della tabella è più estrema o ugualmente probabile
      p1=p1+test1;
   end

   if test2<=p_cut2
   p2=p2+test2;
   end
end
p1
p2
```

APPENDICE 3: script «FisherExactTest»

- Lo script è reperibile sul sito di Matlab al link:

  https://it.mathworks.com/matlabcentral/fileexchange/24379-fisher-s-exact-test-with-n-x-m-contingency-table

- Questo script è stato scritto da Lowell Guangdi il 2009/06/08, è stato riveduto il 2010/01/28

APPENDICE 4: funzioni utilizzate

Funzione	Utilizzo	input	Output
crosstab(x,y)	Calcolo tabella di contingenza	2 vettori numerici 1 d, ogni diverso valore viene associato a un tipo di osservabile	Matrice n x m con la frequenza di associazione tra
nchoosek(n,k)	Calcolo coefficiente binomiale $\binom{n}{k}$	2 scalari	1 scalare
pdf('tipo' v.a., x, p1,…,pn)	Calcolo valori funzione di densità di probabilità di una v.a. in uno o più punti	Tipo: nome v.a. , inserire come testo (es: 'Hypergeometric') x=valori nei quali viene calcolata la pdf. Può essere scalare o un vettore p1,…,pn: parametri per definire la distribuzione. Scalari. Importante l'ordine, sul sito matlab indicazioni per ogni tipo di v.a	In base a x, o scalare o vettore
[h,p]=fishertest(tbl, 'Alpha' , α, side)	Effettua test esatto di fisher	tbl: tabella di contingenza, matrice numerica 2 x 2, i valori devono essere tutti interi ≥0 'Alpha', α: coefficiente di rischio, indica l'errore statistico che si è disposti ad accettare nel rigettare H_0 , il parametro α è uno scalare compreso tra 0 e 1. nel caso non venisse specificato, il test viene comunque effettuato assumendo α=0.05 side: indica varianti del test esatto di fisher. Se non viene indicato o se si inserisce 'both' viene effettuato il test standard. Se viene inserito il 'right' si va ad testare se y1 e x1 interagiscono positivamente (aumento probabilità). Se viene inserito 'left' viene fatto il test opposto	Se nulla viene specificato, viene comunicato solo l'esito del test: 0==fallimento a rigettare H_0 , 1==rigetto H_0 h=esito test, p=probabilità che le due v.a. testate siano indipendenti