Laboratorio di Probabilità e Statistica

lezione 3

Massimo Guerriero – Ettore Benedetti

Indice Lezione

- Requisiti dalla lezione scorsa
- Calcolo delle probabilità e spazio campionario
- Analisi di dipendenza: la connessione
  - Tabella di contingenza
  - Il caso del Titanic
  - Il paradosso di Simpson

Prerequisiti dalla lezione scorsa

- Confidenza con R ed RStudio
- Dataset dello scorso anno
  - capacità di trattarne le variabili
- Scelta del grafico più adatto per una certa variabile
- Trattare indici di posizione e dispersione in maniera opportuna

Calcolo delle probabilità e spazio campionario 1/4

Probabilità — Fiducia con cui ci aspettiamo che un evento si verifichi

Spazio campionario Ω — Insieme di tutti i possibili risultati di un esperimento casuale

$$P(E) = \frac{numero\ di\ casi\ favorevoli\ a\ E}{numero\ di\ casi\ possibili\ in\ \Omega}$$

Es. *Simulazione in R del lancio di un dado per 15 volte*

```
omega <- c(1, 2, 3, 4, 5, 6)
sample(omega, 15, replace=TRUE)
```

Calcolo delle probabilità e spazio campionario 2/4

Es. *Simulazione con R dell'estrazione di palline numerate da un'urna*

```
omega <- c(1, 2, 3, 4, 5, 6)
sample(omega, 6, replace=FALSE)
sample(omega, 7, replace=FALSE)
```

← Genera Errore

Probabilità Condizionata Probabilità di un evento A sapendo che si è già verificato un altro evento B.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Calcolo delle probabilità e spazio campionario 3/4

Es. *Abbiamo un'urna con 10 palline, 6 rosse e 4 nere.*

Estraggo senza reinserimento 2 palline: calcolare la probabilità che la seconda sia rossa dato che la prima era nera

P(B) = probabilità di pescare prima una pallina nera

P(A) = probabilità di pescare al secondo colpo una pallina rossa

$$P(A \mid B) = \frac{\frac{6}{9} \cdot \frac{4}{10}}{\frac{4}{10}}$$

P(B) = probabilità di pescare una rossa alla prima estrazione

P(A) = probabilità di pescare una rossa alla seconda estrazione

$$P(A \mid B) = \frac{\frac{5}{9} \cdot \frac{6}{10}}{\frac{6}{10}}$$

Calcolo delle probabilità e spazio campionario 4/4

- Il caso dei compleanni -

Es. *Quante persone ci devono essere in un'aula per avere una probabilità superiore al 50% che due di esse compiano gli anni nello stesso giorno?*

A = «Tutti compiono gli anni in giorni diversi»

$$P(A) = \frac{compleanni\ in\ date\ diverse\ per\ gli\ n\ studenti}{possibili\ date\ di\ compleanno}$$

$$= \frac{365 \cdot 364 \cdot 363 \ldots (365 - n+1)}{365 \cdot 365 \ldots 365}$$

```
compleanno <- function(n){
    1-prod((365:(365-n+1))/rep(365,n))
}
```

numeratori denominatori

Consegna

1. Utilizzare la funzione compleanno per rispondere al quesito riportato nell'esempio.
   (Funzione ed esempio si trovano nella slide precedente)

2. Verificare il risultato ottenuto nel punto 1 utilizzando il comando *sample* e il comando *table* per le frequenze assolute.

Indice Lezione

- Requisiti dalla lezione scorsa
- Calcolo delle probabilità e spazio campionario
- Analisi di dipendenza: la connessione
  - Tabella di contingenza
  - Il caso del Titanic
  - Il paradosso di Simpson

Analisi di dipendenza: la connessione

Il passo successivo allo studio univariato visto fino ad adesso, è verificare se esistono legami tra due o più fenomeni rilevati congiuntamente sugli stessi individui.

Es.

- L'utilizzo della rete dipende dal genere?
- La rinuncia della rete per il tempo libero dipende dalla situazione sentimentale?
- Il numero di persone decedute di tumore ai polmoni in Italia dipende dal loro consumo di sigarette?

Tabella di contingenza 1/5

E' utilizzata per leggere correttamente i dati relativi a due variabili.

X \ Y	y_1	y_2	…	y_j	…	y_k	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}	$n_{2.}$
⋮			**Frequenze congiunte**				⋮
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}	$n_{i.}$
⋮							⋮
x_h	n_{h1}	n_{h2}	…	n_{hj}	…	n_{hk}	$n_{h.}$
	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.k}$	n

Distribuzione di frequenza marginale di X

Distribuzione di frequenza marginale di Y

Tabella di contingenza 2/5

Es. Utilizzo della rete dal lunedì al venerdì **VS** luogo di domicilio

hinternet_lv domicilio	0-5	5-10	10-15	
Altro Comune	45	23	1	69
Verona e Provincia (VR)	23	14	2	39
	68	37	3	108

45 persone che alla domanda sul domicilio avevano risposto di abitare in un altro comune, hanno risposto che passano anche dalle 0 alle 5 ore (comprese) dal lunedì al venerdì in internet.

Tabelle di contingenza 3/5

Per costruire tabelle di contingenza con R possiamo utilizzare il comando "table" che abbiamo già visto per le frequenze assolute:

table(*variabile1, variabile2****)***

Es.

```
> HRange<-cut(dataset$hinternet_lv, breaks=c(0, 5, 10, 15), include.lowest=TRUE);
> tabella<-table(dataset$domicilio,HRange)
> tabella
   HRange
    [0,5] (5,10] (10,15]
  0    45     23       1
  1    23     14       2
```

Per ricavare poi le frequenze marginali

con: ***margin.table(****tabella, n°Variabile****)***

```
> margin.table(tabella,1)

 0  1
69 39
> margin.table(tabella,2)
HRange
  [0,5]  (5,10] (10,15]
     68      37       3
```

Tabella di contingenza 4/5

Applicando il comando summary sulla tabella di contingenza, si possono ottenere il **p-value** ed il **X^2**

```
> summary(tabella)
Number of cases in table: 108
Number of factors: 2
Test for independence of all factors:
 Chisq = 1.4161, df = 2, p-value = 0.4926
 Chi-squared approximation may be incorrect
```

Teoria: Se due variabili X e Y sono indipendenti tra loro, la frequenza congiunta n_{ij} deve essere pari al prodotto delle frequenze marginali ($n_{i.} \cdot n_{.j}$) diviso il totale delle osservazioni n.

$$X^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(nij - n^*_{ij})^2}{n^*_{ij}} \qquad \text{dove } n^*_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} =$$

Frequenze congiunte teoriche

Tabella di contingenza 5/5

Dal chi-quadro si possono ricavare diversi indici normalizzati.
Per esempio, per avere un numero da 0 a 1:

$$X^2_{max} = \max X^2 = n \cdot \min(h-1, k-1)$$

L'indice da impiegare sarà quindi:

$$0 <= \tilde{X}^2 = \frac{X^2}{X^2_{max}} <= 1$$

Consegna

1) Generare le tabelle di contingenza di queste variabili
   - *anni* ***vs*** *hinternet_we*
   - *studio* ***vs*** *hlav*
   - *single* ***vs*** *hlib_lv*
2) Trovare per ogni tabella le frequenze marginali
3) Calcolare per ogni tabella il X^2
   - Anticipazione:

     Valutare se le due variabili di ogni tabella sono statisticamente dipendenti

     (p-value < 0.05)

Indice Lezione

- Requisiti dalla lezione scorsa
- Calcolo delle probabilità e spazio campionario
- Analisi di dipendenza: la connessione
  - Tabella di contingenza
  - Il caso del Titanic
  - Il paradosso di Simpson

Il caso del Titanic 1/4

Dall'inchiesta ufficiale di Lord Mersey

«…Mi ritengo soddisfatto della spiegazione che l'elevata proporzione di perdite non deve essere ricercata nella discriminazione dei passeggeri di terza classe. Essi non sono stati discriminati».

- Carichiamo il dataset del titanic, già presente in R in un formato speciale
  - ***data****(Titanic)*
  - ***ftable****(Titanic)*

			Survived	No	Yes
Class	Sex	Age			
1st	Male	Child		0	5
		Adult		118	57
	Female	Child		0	1
		Adult		4	140
2nd	Male	Child		0	11
		Adult		154	14
	Female	Child		0	13
		Adult		13	80
3rd	Male	Child		35	13
		Adult		387	75
	Female	Child		17	14
		Adult		89	76
Crew	Male	Child		0	0
		Adult		670	192
	Female	Child		0	0
		Adult		3	20

Il caso del Titanic 2/4

Creiamo le nostre tabelle di contingenza con il comando *as.table* e *apply*:

Es.

tabsex <- ***(as.table(apply(****Titanic, c(2,4),sum****)))***

Sex	Survived No	Survived Yes
Male	1364	367
Female	126	344

Valutiamo se il numero di sopravvissuti è legato al sesso

```
test <- chisq.test(tabsex);
chi <- test[1];
chi <- round(as.numeric(chi),3);
pvalue <- test[3];
pvalue <- round(as.numeric(pvalue),3);
chiN <- round(as.numeric(chi/2201),3);
```

pvalue = 0
chi = 454.5
chiN= 0.206

Il caso del Titanic 3/4

Valutiamo se il numero di sopravvissuti è legato all'età:

	Survived	
Age	No	Yes
Child	52	57
Adult	1438	654

pvalue = 0
chi = 20.005
chiN = 0.009

…E se è legato alla classe (escludendo l'equipaggio):

```
tabclass <- apply(Titanic, c(1,4),sum)
tabclass <- as.table(tabclass[1:3,])
test<-chisq.test(tabage);
…
```

	Survived	
Class	No	Yes
1st	122	203
2nd	167	118
3rd	528	178

pvalue = 0
chi = 133.052
chiN = 0.101

Il caso del Titanic 4/4

Il sesso sembra essere la variabile più legata alla sopravvivenza, se guardiamo solo il p-value possiamo quindi compiere degli errori.

Class	Sex: Male	Sex: Female
1st	180	145
2nd	179	106
3rd	510	196
Crew	862	23

Le donne sono in percentuale maggiore in prima classe (45%) rispetto alla terza (28%).

Riguardiamo infatti i $\tilde{X}^2$ ottenuti riportati in tabella:

Sesso	Età	Classe
0.206	0.009	0.101

Si ricorda poi che a bordo erano imbarcate 2201 persone, ma i mezzi di salvataggio a disposizione potevano salvare solo 1184 persone.

Il paradosso di Simpson 1/3

L'andamento generale del legame fra due fenomeni statistici può apparire alterato se ci limitiamo ad analizzarlo in sottogruppi

Comandi creazione dataset di esempio:

http://benedettiettore.altervista.org/Statistica/dati/simpson.R

```
> ftable(simpson)
                        ospedale   1   2
trattamento decesso
A           FALSE                 40  85
            TRUE                 160  15
B           FALSE                 30 300
            TRUE                 170 100
```

Il paradosso di Simpson 2/3

```
tabS <- table(simpson)
osp1 <- tabS[,,1]
osp2 <- tabS[,,2]
osp1[1,] <- osp1[1,]/sum(osp1[1,])
osp1[2,] <- osp1[2,]/sum(osp1[2,])

osp2[1,] <- osp2[1,]/sum(osp2[1,])
osp2[2,] <- osp2[2,]/sum(osp2[2,])
```

```
            decesso
trattamento FALSE TRUE
          A  0.20 0.80
          B  0.15 0.85
```

```
            decesso
trattamento FALSE TRUE
          A  0.85 0.15
          B  0.75 0.25
```

Il trattamento A vince sul trattamento B in tutti e due gli ospedali

Il paradosso di Simpson 3/3

Ma unendo i risultati degli ospedali…

```
ospedali <- apply(tabS,c(1,2),sum)
ospedali[1,] <- ospedali[1,]/sum(ospedali[1,])
ospedali[2,] <- ospedali[2,]/sum(ospedali[2,])
```

```
                decesso
trattamento      FALSE      TRUE
          A 0.4166667 0.5833333
          B 0.5500000 0.4500000
```

Il trattamento B vince sul trattamento A!

Consegna

Presso un grande ateneo americano si ebbe la seguente contestazione:
Di 1000 posti disponibili per la facoltà di Economia e Lettere, 819 studenti maschi ottennero l'ammissione a fronte di solo 181 studentesse.
Le domande furono 1000 per i maschi e 1000 per le femmine,
ci si chiede se vi sia stata discriminazione.

L'amministrazione dell'ateneo dimostrò che la contestazione era sbagliata.

Senza dati alla mano, come si potrebbe spiegare questa conclusione?