LAB – LEZ. 2 – DISTRIBUZIONI DOPPIE E CONNESSIONE

LABORATORIO DI PROBABILITA' E STATISTICA

Docente: Bruno Gobbi

2 - DISTRIBUZIONE DOPPIE E CONNESSIONE

ESEMPIO DI TABELLA A DOPPIA ENTRATA

		CAPELLI	
		BIONDI	NERI
OCCHI	AZZURRI	25	10
	SCURI	15	60

 # SPESSO VENGONO CONDOTTI DEGLI STUDI CHE METTONO A CONFRONTO LE RELAZIONI FRA DUE CARATTERI RILEVATI CONGIUNTAMENTE IN UNA TABELLA DI FREQUENZE A DOPPIA ENTRATA.

CONSIDERIAMO UN CASO CLASSICO, QUELLO RIPORTATO QUI SOPRA DELLA TABELLA A DOPPIA ENTRATA NELLA QUALE VIENE RIPORTATA L'ASSOCIAZIONE FRA DUE FENOMENI, OSSIA IL COLORE DEGLI OCCHI E IL COLORE DEI CAPELLI. IN ALTRE PAROLE: C'E' UNA RELAZIONE FRA L'AVERE UN CERTO COLORE DEGLI OCCHI E UN CERTO COLORE DEI CAPELLI?

 # CREIAMO INNANZITUTTO LA MATRICE DEI DATI CON IL COMANDO matrix:

```
> eyehair=matrix(c(25, 10, 15, 60), nrow=2, byrow=TRUE)
```

 # nrow=2 E' IL NUMERO DI RIGHE, byrow=TRUE INDICA CHE I DATI VANNO LETTI PER RIGA

 # CREIAMO LE ETICHETTE PER LE RIGHE E LE COLONNE:

```
> eye=c("azzurri", "scuri")
> hair=c("biondi", "neri")
```

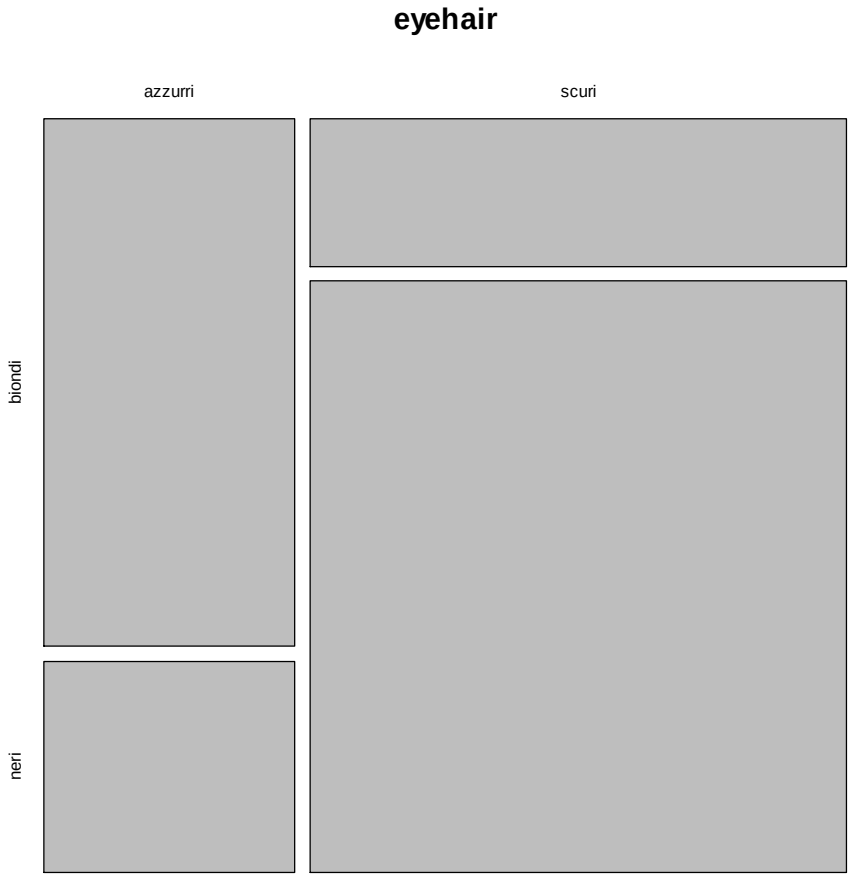
 # ASSEGNAMO LE ETICHETTE ALLA MATRICE CON IL COMANDO:

```
> dimnames(eyehair)=list(eye, hair)
> eyehair
```

```
        biondi neri
azzurri     25   10
scuri       15   60
```

 # DISEGNAMO IL GRAFICO AD AREE CHE RAPPRESENTA LA TABELLA:

```
> mosaicplot(eyehair)
```



 # IL GRAFICO RIPORTA LA RELAZIONE CHE ESISTE FRA I CARATTERI DEGLI OCCHI (AZZURRI O SCURI) E QUELLO DEI CAPELLI (BIONDI O NERI). L'AREA PIÙ GRANDE È QUELLA RELATIVA AGLI OCCHI SCURI E AI CAPELLI NERI, MENTRE SONO POCHI QUELLI CHE HANNO I CAPELLI NERI E GLI OCCHI AZZURRI. QUESTO GRAFICO PERMETTE DI AVERE SUBITO UN'IDEA DEI RAPPORTI DI "FORZA" CHE CI SONO FRA LE VARIABILI.

CALCOLO DEL CHI-QUADRATO

- Il test del chi-quadrato consiste in un test che mette a confronto le seguenti due ipotesi:
- **ipotesi nulla H0**: afferma che c'è indipendenza fra i due fenomeni;
- **ipotesi alternativa H1**: che invece dice che c'è una connessione fra i caratteri.

CALCOLO DEL CHI-QUADRATO

- In R il test del chi-quadrato viene condotto molto semplicemente con il comando: **chisq.test**

```
> testchiq=chisq.test(eyehair)
> testchiq

        Pearson's Chi-squared test with Yates' continuity correction

X-squared = 25.0983, df = 1, p-value = 5.448e-07
```

- **"X-squared"** è il chi-quadrato calcolato
- **"df"** sono i degrees of freedom, i gradi di libertà (g.d.l.), dati dal prodotto: df=(n. Righe-1)*(n. Colonne-1)
- **"p-value"** è il livello di significatività. Questo valore deve essere inferiore al 5% (ovvero 0,05) per considerare valido il risultato trovato con il test.

TAVOLA DEL CHI-QUADRATO

	alpha (significatività)	
g.d.l.	**1%**	**5%**
1	6,64	3,84
2	9,21	5,99
3	11,35	7,82
4	13,28	9,49
5	15,09	11,07
6	16,81	12,59
7	18,48	14,07
8	20,09	15,51
9	21,67	16,92
10	23,21	18,31

CONFRONTO DEL CHI-QUADRATO CALCOLATO CON LA SOGLIA TEORICA

- **3.84** per un livello di significatività del **5%** e 1 g.d.l.
- **6.64** per un livello di significatività dell'**1%** e 1 g.d.l.

- In questo caso abbiamo 25.0983, che è abbondantemente superiore non solo a 3.84, che è la soglia critica per sbagliare al massimo nel 5% dei casi, ma addirittura a 6.64, che è la soglia critica oltre la quale si rifiuta l'ipotesi nulla di indipendenza sbagliando solo nell'1% dei casi.

- Quindi **il test rifiuta l'ipotesi nulla H0 e conferma che al 99% c'è una connessione fra i fenomeni.**

CALCOLO DEL "V" DI CRAMER

- Una volta che abbiamo rilevato che c'è una connessione fra i 2 fenomeni, possiamo misurare quanto sono connessi fra di loro con un opportuno indice, il **V di Cramer**.

- Questo indicatore assume:
  - valore 0 nel caso di **perfetta indipendenza**;
  - valore 1 quando invece c'è la **massima connessione** fra i due fenomeni.

CALCOLO DEL "V" DI CRAMER

- Per calcolare il V di Cramer bisogna usare la seguente formula:

$$V = \sqrt{\frac{\chi^2}{\mathrm{N} * (\min(righe, colonne) - 1)}}$$

- χ^2= valore della variabile chi-quadrato ricavato dal test chi quadrato (**$statistic**)
- N = numero totale di casi (**N=sum(eyehair)**)
- min(*righe*, *colonne*) - 1 = si sceglie il minore fra il numero delle righe e delle colonne; quindi si sottrae 1 (ES. tab. 2 righe e 3 colonne: si sceglie 2, quindi si toglie 1: 2-1=1)

ESEMPI DI DETERMINAZIONE DEL MINORE FRA IL NUMERO DI RIGHE E DI COLONNE PER IL V DI CRAMER:

- ES. tabella 2 x 2:

	BIONDI	NERI
AZZURRI	25	10
SCURI	15	60

- n. righe = 2
- n. colonne = 2
- In questo caso il numero di righe e di colonne è lo stesso, quindi scelgo 2.
- Da 2 sottraggo 1: 2-1 = 1

- ES. tabella 4 x 3:

	ALPHA	BETA	GAMMA
TIPO A	25	10	12
TIPO B	15	60	48
TIPO AB	22	10	36
TIPO 0	21	6	12

- n. righe = 4
- n. colonne = 3
- In questo caso il minore fra il numero di righe e di colonne è 3.
- Da 3 sottraggo 1: 3-1 = 2

PER CALCOLARE IL COEFFICIENTE V DI CRAMER, DEVO QUINDI PRIMA RICAVARMI LE SINGOLE COMPONENTI: IL CHI QUADRATO E LA NUMEROSITA' TOTALE "N"

 # IL VALORE DEL CHI-QUADRATO SI RICAVA DA:

```
> chiquadrato= testchiq$statistic
> chiquadrato
X-squared
 25.09833
```

 # IL TOTALE DI ELEMENTI PRESENTI SI OTTIENE COSÌ:

```
> N = sum(eyehair)
> N
[1] 110
```

 # IL VALORE DEL CHI QUADRATO:

```
> V=sqrt( chiquadrato / (N*(2-1)) )
> V
X-squared
0.4776679
```

 # IL VALORE DEL V DI CRAMER (0.4776679) PORTA A RITENERE CHE C'E' UNA BUONA CONNESSIONE FRA I DUE CARATTERI "COLORE DEI CAPELLI" E "COLORE DEGLI OCCHI", NEL SENSO CHE E' CORRETTO IPOTIZZARE CHE DI SOLITO AD UN CERTO COLORE DEI CAPELLI CORRISPONDE UN CERTO COLORE DEGLI OCCHI.

ES. EFFICACIA FARMACO

Ipotizziamo di avere i risultati di un test sull'efficacia di un nuovo farmaco su N=400 pazienti.

		EFFETTO	
		MIGLIORAMENTO	PEGGIORAMENTO
TRATTAMENTO	FARMACO	250	50
	PLACEBO	50	50

	alpha (significatività)	
g.d.l.	1%	5%
1	6,64	3,84
2	9,21	5,99
3	11,35	7,82
4	13,28	9,49
5	15,09	11,07
6	16,81	12,59
7	18,48	14,07
8	20,09	15,51
9	21,67	16,92
10	23,21	18,31

```
> farmaco=matrix(c(250, 50, 50, 50), nrow=2, byrow=TRUE)
> trattamento=c("farmaco", "placebo")
> effetto=c("miglioramento", "peggioramento")
> dimnames(farmaco)=list(trattamento, effetto)
> farmaco
        miglioramento peggioramento
farmaco           250            50
placebo            50            50
```

 # DISEGNARE IL GRAFICO AD AREE

```
> mosaicplot(farmaco)
```

 # CALCOLIAMO IL TEST DEL CHI QUADRATO

```
> testchiq=chisq.test(farmaco)
> testchiq

        Pearson's Chi-squared test with Yates' continuity correction

data:  farmaco
X-squared = 42.6844, df = 1, p-value = 6.432e-11
```

 # POICHE' IL VALORE CALCOLATO DEL CHI-QUADRATO E' 42.6844, BEN SUPERIORE ALLA SOGLIA CRITICA DI 6.64 VALIDO ALL'1% CON 1 GRADO DI LIBERTA' (G.D.L.)(O df=degrees of freedom NELL'OUTPUT), SI RIFIUTA L'IPOTESI NULLA DI INDIPENDENZA E SI CONFERMA LA CONNESSIONE FRA I FENOMENI

 # CALCOLIAMO IL VALORE DELLA STATISTICA V DI CRAMER

```
> chiquadrato= testchiq$statistic
> chiquadrato
X-squared
 42.68444
```

 # IL TOTALE DI ELEMENTI PRESENTI SI OTTIENE IN QUESTO MODO:

```
> N = sum(farmaco)
> N
[1] 400

> V=sqrt( chiquadrato / (N*(2-1)) )
> V
X-squared
0.3266667
```

 # IL RISULTATO PORTA AD AFFERMARE CHE C'È UNA DISCRETA CONNESSIONE FRA I DUE FENOMENI

ES. SOPRAVVISSUTI DEL TITANIC

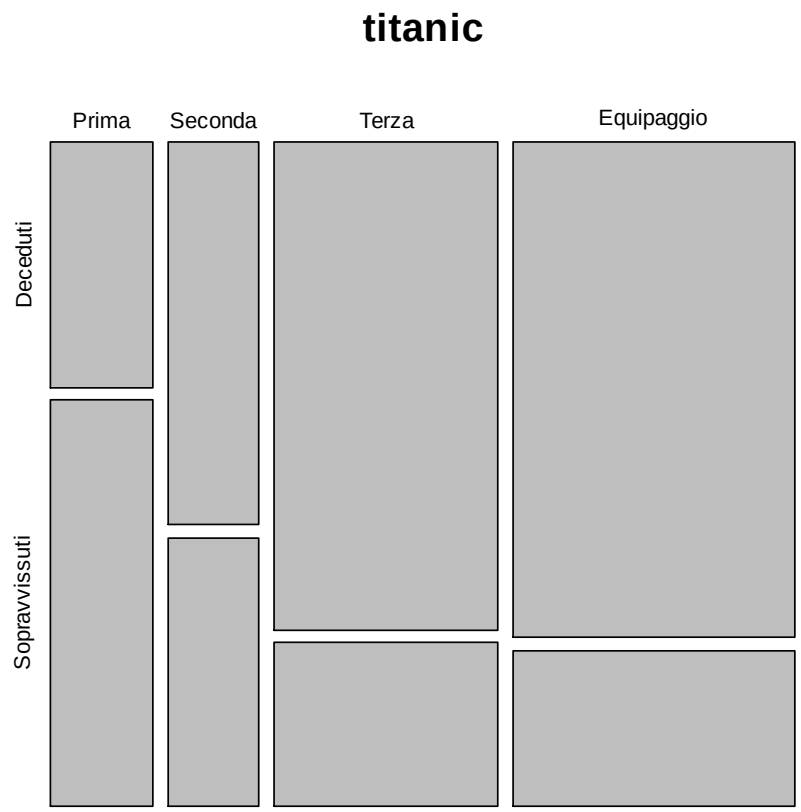
La tabella riporta i sopravvissuti e i deceduti fra i passeggeri del Titanic a seconda della classe di appartenenza.

		ESITO	
		SOPRAVVISSUTI	DECEDUTI
CLASSE	PRIMA	122	203
	SECONDA	167	118
	TERZA	528	178
	EQUIPAGGIO	673	212

	alpha (significatività)	
g.d.l.	1%	5%
1	6,64	3,84
2	9,21	5,99
3	11,35	7,82
4	13,28	9,49
5	15,09	11,07
6	16,81	12,59
7	18,48	14,07
8	20,09	15,51
9	21,67	16,92
10	23,21	18,31

```
> titanic=matrix(c(122, 203, 167, 118, 528, 178, 673, 212), nrow=4, byrow=TRUE)
> classe=c("prima", "seconda", "terza", "equipaggio")
> esito=c("deceduti", "sopravvissuti")
> dimnames(titanic)=list(classe, esito)
> titanic
           Deceduti Sopravvissuti
Prima           122           203
Seconda         167           118
Terza           528           178
Equipaggio      673           212
```

```
> mosaicplot(titanic)
```




```
> testchiq=chisq.test(titanic)
> testchiq

        Pearson's Chi-squared test

data:  titanic
X-squared = 190.4011, df = 3, p-value < 2.2e-16
```

 # POICHE' IL VALORE CALCOLATO DEL CHI-QUADRATO E' 190.4011, BEN SUPERIORE ALLA SOGLIA CRITICA DI 11.35 VALIDO ALL'1% CON 3 GRADI DI LIBERTA', SI RIFIUTA L'IPOTESI NULLA DI INDIPENDENZA E SI CONFERMA LA CONNESSIONE FRA I FENOMENI, OVVERO FAR PARTE DELLA PRIMA, SECONDA, TERZA CLASSE O DELL'EQUIPAGGIO FACEVA DIFFERENZA FRA LA VITA E LA MORTE. I GRADI DI LIBERTA' SONO 3 PERCHE' DATI DA (r-1)*(c*1)=(4-1)*(2-1)

 # CALCOLIAMO IL VALORE DELLA STATISTICA V DI CRAMER

```
> chiquadrato= testchiq$statistic
> chiquadrato
X-squared
 190.4011
```

 # IL TOTALE DI ELEMENTI PRESENTI SI OTTIENE IN QUESTO MODO:

```
> N = sum(titanic)
> N
[1] 2201
```

 # SI SCEGLIE IL MINORE FRA IL NUMERO DI RIGHE E DI COLONNE E SI SOTTRAE 1

```
> V=sqrt( chiquadrato / (N*(2-1)) )
> V
X-squared
0.2941201
```

 # IL RISULTATO PORTA AD AFFERMARE CHE C'È UNA DISCRETA CONNESSIONE FRA I DUE FENOMENI

ES. STAGE E ASSUNZIONE (CASO NORMALE)

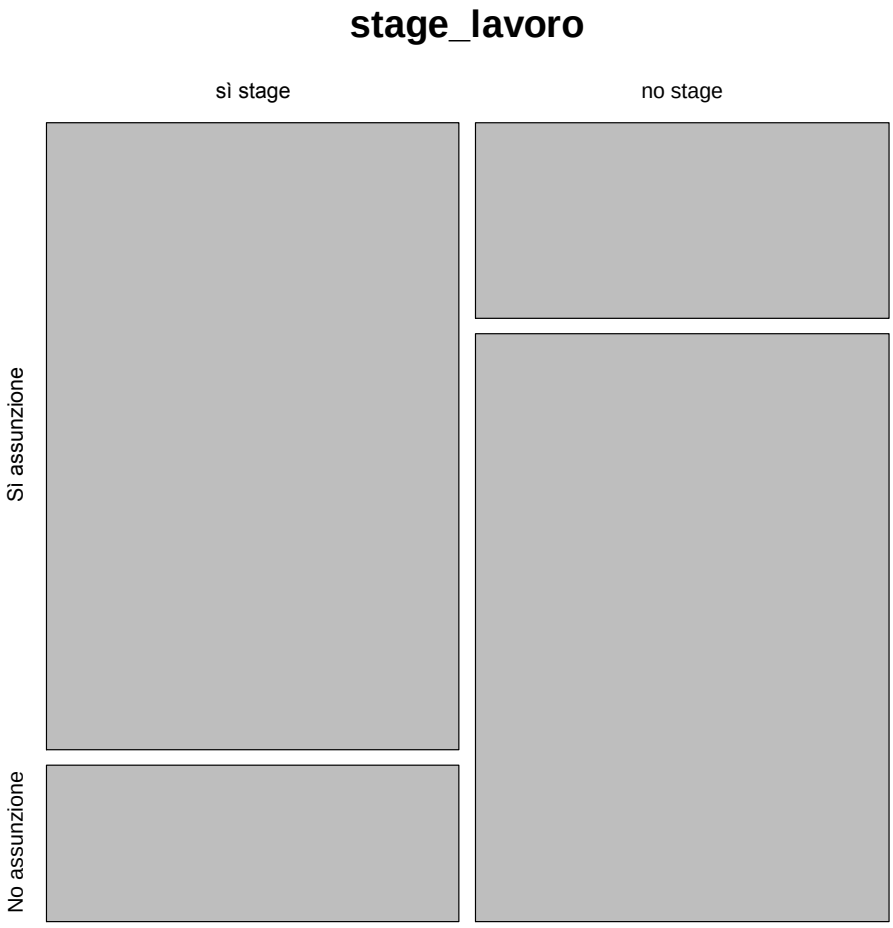
Si vuole verificare se esiste una relazione fra il fatto di svolgere uno stage presso un importante istituto di credito e la successiva eventuale assunzione. Sono stati così presi in considerazione 200 ragazzi così distribuiti:

		ASSUNZIONE?		
		SI'	NO	Totale
STAGE?	SI'	80	20	100
	NO	25	75	100
	Totale	105	95	200

	alpha (significatività)	
g.d.l.	1%	5%
1	6,64	3,84
2	9,21	5,99
3	11,35	7,82
4	13,28	9,49
5	15,09	11,07
6	16,81	12,59
7	18,48	14,07
8	20,09	15,51
9	21,67	16,92
10	23,21	18,31

```
> stage_lavoro=matrix(c(80, 20, 25, 75), nrow=2, byrow=TRUE)
> stage=c("sì stage", "no stage")
> lavoro=c("Sì assunzione", "No assunzione")
> dimnames(stage_lavoro)=list(stage, lavoro)
> stage_lavoro
          Sì assunzione No assunzione
sì stage             80             20
no stage             25             75
```

```
> mosaicplot(stage_lavoro)
```




```
> testchiq=chisq.test(stage_lavoro)
> testchiq

        Pearson's Chi-squared test with Yates' continuity correction

data:  stage_lavoro
X-squared = 58.4662, df = 1, p-value = 2.068e-14
```

 # POICHE' IL VALORE CALCOLATO DEL CHI-QUADRATO E' 58.4662, BEN SUPERIORE ALLA SOGLIA CRITICA DI 6.64 VALIDO ALL'1%, SI RIFIUTA L'IPOTESI NULLA DI INDIPENDENZA E SI CONFERMA LA CONNESSIONE FRA I FENOMENI, OVVERO FARE UNO STAGE COMPORTA MAGGIORI PROBABILITA' DI ESSERE ASSUNTI. I GRADI DI LIBERTA' SONO 1 PERCHE' DATI DA (r-1)*(c*1)=(2-1)*(2-1)

 # CALCOLIAMO IL VALORE DELLA STATISTICA V DI CRAMER

```
> chiquadrato=testchiq$statistic
> chiquadrato
X-squared
 58.46617
```

 # IL TOTALE DI ELEMENTI PRESENTI SI OTTIENE IN QUESTO MODO:

```
> N = sum(stage_lavoro)
> N
[1] 200
```

 # SI SCEGLIE IL MINORE FRA IL NUMERO DI RIGHE E DI COLONNE E SI SOTTRAE 1

```
> V=sqrt( chiquadrato / (N*(2-1)) )
> V
X-squared
0.5406763
```

 # IL RISULTATO PORTA AD AFFERMARE CHE C'È UNA BUONA CONNESSIONE FRA I DUE FENOMENI