6 – ASSOCIAZIONE TRA DUE VARIABILI CATEGORIALI

TECNICHE DI ANALISI DI DATI I

In precedenza abbiamo descritto e verificato ipotesi relative a una sola distribuzione.

D'ora in avanti, **descriveremo e verificheremo ipotesi relative a più distribuzioni congiuntamente considerate**: prima **bivariate** (**due** variabili : categoriali, ordinali, metriche), poi **multivariate**.

Cominciamo con **l'associazione tra due variabili categoriali.**

Useremo il dataset **fumo**: scaricatelo da Elly, apritelo in R e leggetene la descrizione, prima di proseguire con la lettura.

DESCRIVERE L'ASSOCIAZIONE TRA DUE VARIABILI CATEGORIALI

In una distribuzione bivariata categoriale, con $k \geq 2$ categorie in X_1 e $k \geq 2$ categorie in X_2, si considerano **due modalità appaiate dello stesso caso**, ovvero la sua appartenenza alla categoria $\boldsymbol{k_a}$ della variabile $\boldsymbol{X_1}$ **e** alla categoria k_a della variabile $\boldsymbol{X_2}$: si contano quanti casi del campione condividono le modalità $X_{1a}X_{2a}$, $X_{1a}X_{2b}$, $X_{1b}X_{2a}$, $X_{1b}X_{2b}$,.... $X_{1k}X_{2k}$.

Tabella di contingenza a due vie: le categorie di X_1 rappresentano le righe e quelle di X_2 le colonne. Il caso più semplice è una tabella di contingenza $\mathbf{2 \times 2}$

Usiamo `table(x1, x2)`: `x1` dà le frequenze assolute delle righe, `x2` quelle delle colonne.

*Esploriamo **l'associazione tra il genere** (X_1: a = femmina, b = maschio) e **l'esito del trattamento dopo tre mesi di terapia** (X_2: a = astinente, b = fumatore):*

```
table(fumo$genere,fumo$outcome_3_mesi)
   astinente fumatore
F         32       21
M         54       19
```

Con `margin.table(table)` ricaviamo: il **totale delle osservazioni** (N, di default), i marginali di **riga** (`margin=1`) e i marginali di **colonna** (`margin=2`).

```
margin.table(table(fumo$genere, fumo$outcome_3_mesi))
[1] 126
```

```
margin.table(table(fumo$genere,fumo$outcome_3_mesi),1)
 F  M
53 73
```

```
margin.table(table(fumo$genere,fumo$outcome_3_mesi),2)
astinente  fumatore
       86        40
```

Con **prop.table(table)** si ottengono **proporzioni** (*100: **%**) per **riga (margin=1)**, per **colonna (margin=2)** o sul **totale** (**default**: nessuna specificazione)

Per riga

32 F astinenti e 21 F fumatrici su 53 F

54 M astinenti e 19 M fumatori su 73 M

```
  astinente fumatore
F        32       21
M        54       19
```

```
round(prop.table(table(fumo$genere,
 fumo$outcome_3_mesi),margin=1)*100,1)
  astinente fumatore
F      60.4     39.6
M      74.0     26.0
```

Per colonna

32 F e 54 M astinenti su 86 astinenti.

21 F e 19 M fumatori su 40 fumatori

```
round(prop.table(table(fumo$genere,
 fumo$outcome_3_mesi),margin = 2)*100,1)
  astinente fumatore
F      37.2     52.5
M      62.8     47.5
```

Sul totale

32 F astinenti su 126 sogg., 21 F fumatrici su 126 sogg.,

54 M astinenti su 126 sogg. e 19 M fumatori su 126 sogg.

```
round(prop.table(table(fumo$genere,
 fumo$outcome_3_mesi))*100,1)
  astinente fumatore
F      25.4     16.7
M      42.9     15.1
```

Quanto è **forte la differenza tra le proporzioni** di donne astinenti e fumatrici (.604 *versus* .396)?

	astinente	fumatore
F	60.4	39.6
M	74.0	26.0

Un coefficiente di effect size della differenza tra proporzioni è il **coefficiente *h* di Cohen**.

Si può ricavare come **differenza tra le proporzioni trasformate in $2 \times arcoseno$** (phi, Φ)

$$\phi_i = 2 \times \arcsin(proporzione_i)$$

$$h = \Phi_1 - \Phi_2$$

L'arcoseno è dato dal **seno inverso** (`asin`) della **radice quadrata di X**: `asin(sqrt(x))`

```
phi1<-2*(asin(sqrt(.604)))
phi2<-2*(asin(sqrt(.396)))
(h<-phi1-phi2)
[1] 0.4190596
```

… ma possiamo servirci di `ES.h(prop1-prop2)` di `pwr`, usato nella power analysis:

```
ES.h(p1 = .604, p2=.396)
[1] 0.4190596
```

Poiché: ≤.2 effetto trascurabile, .2-.5 debole, .5-.8 discreto e ≥.8 forte, la differenza tra le proporzioni di donne astinenti e fumatrici è debole.

Più rilevante è quella tra i **maschi**:

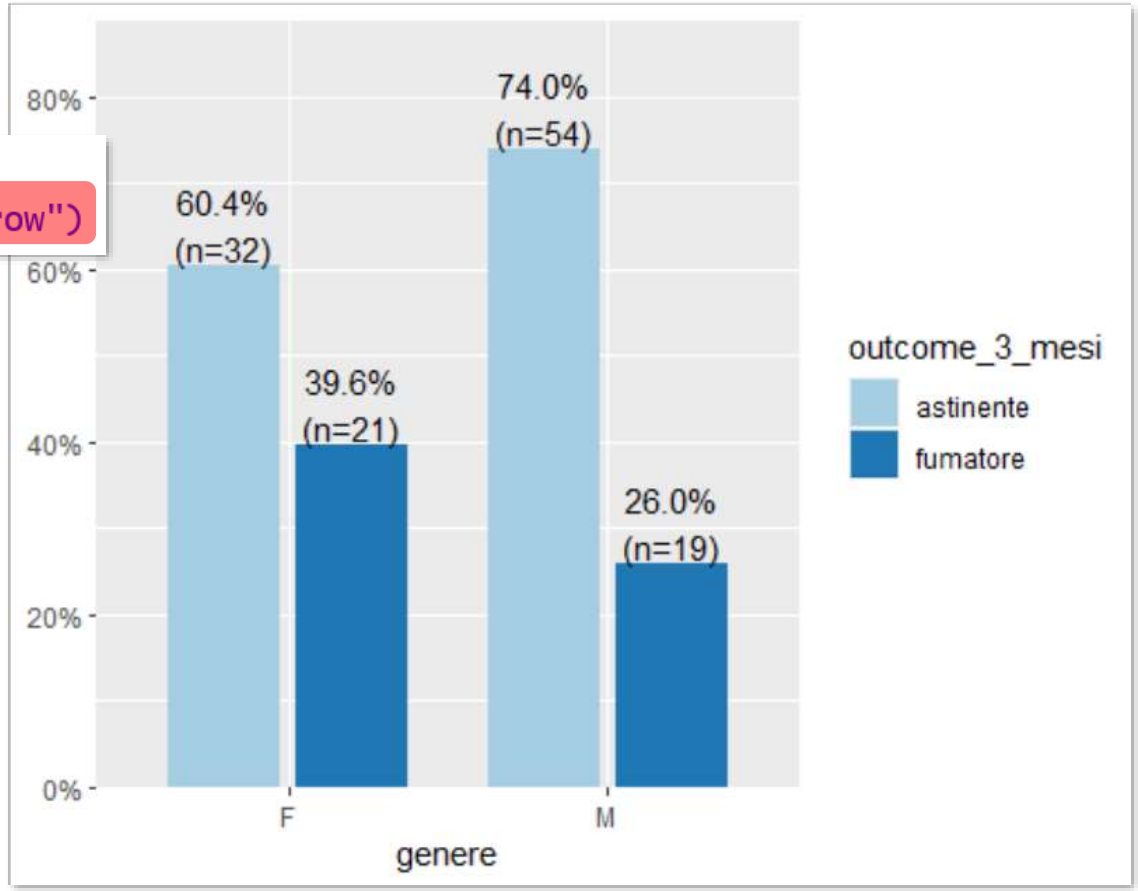
```
ES.h(p1 = .740, .260)
[1] 1.001309
```

Usate la variabile che indica lo ***status del paziente*** *dopo un anno dalla fine* ***del trattamento***

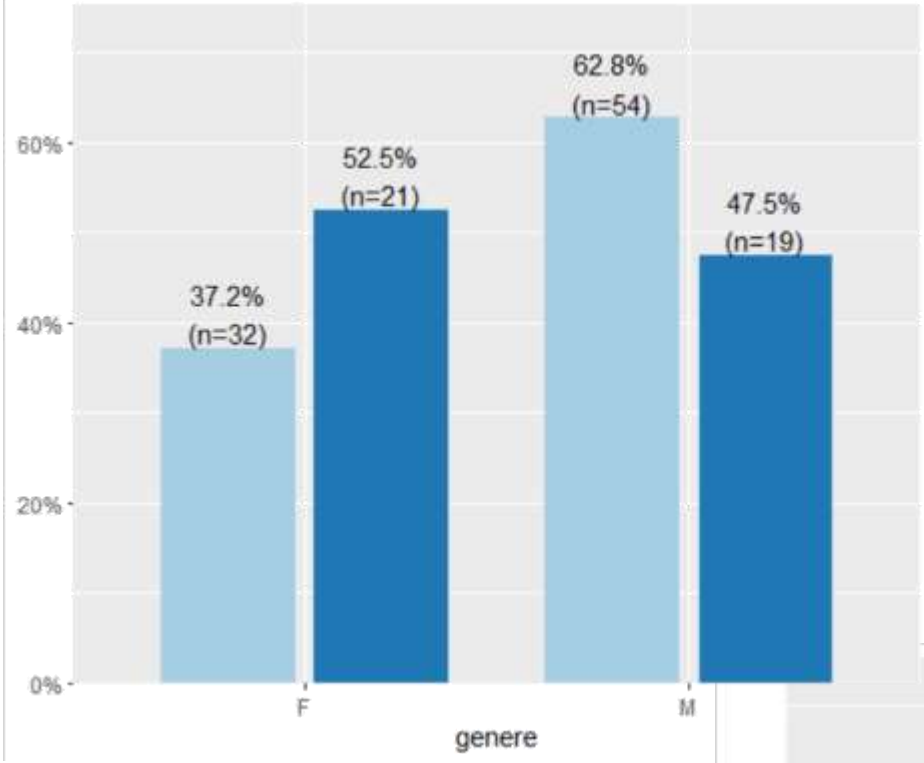
(fumo$outcome_12_mesi) per sapere se ***l'apparente vantaggio degli uomini*** *nell'essere astinenti a tre mesi si* ***mantenga*** *anche a lungo termine.*

La funzione `plot_xtab (x= variabile in riga, grp= variabile in colonna)` di `sjPlot` produce eleganti barplot delle percentuali; `margin= "row"/"col"/"cell"` definisce se entro riga / **entro colonna (default)** / sul totale:

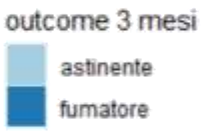
```
plot_xtab(x = fumo$genere,
grp=fumo$outcome_3_mesi, margin = "row")
```

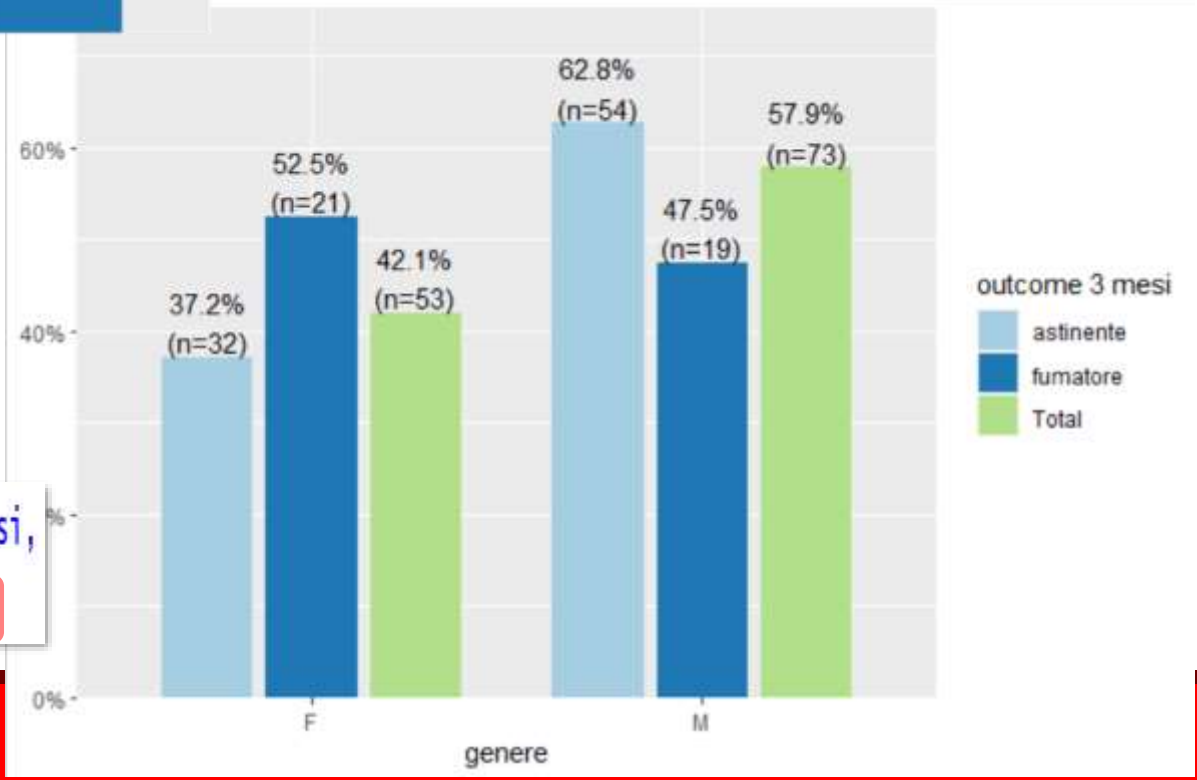


Quando è chiesto `margin= "col"`, sono visualizzate anche le barre dei marginali di riga: se non le volete, indicate `show.total=FALSE`.



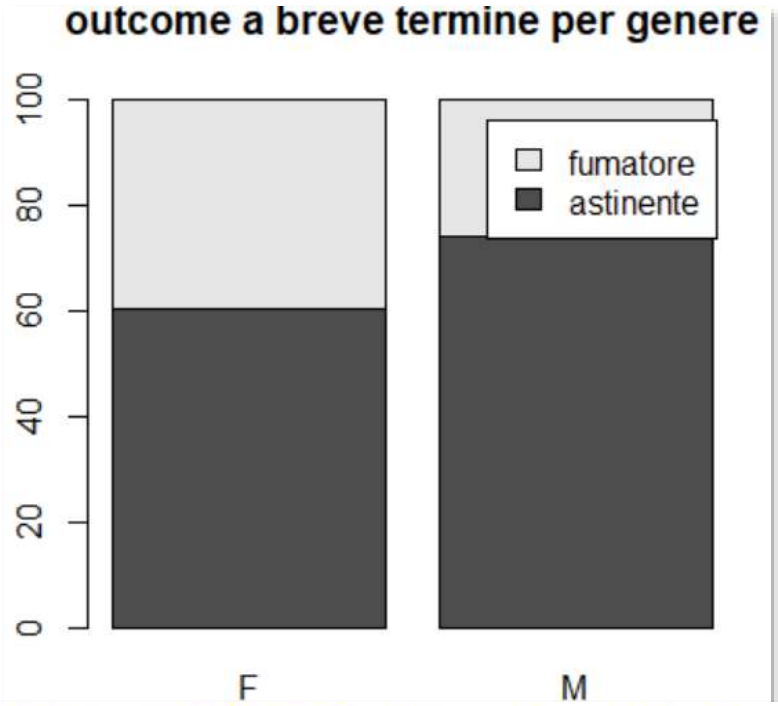

```
plot_xtab(x = fumo$genere, grp=fumo$outcome_3_mesi,
  legend.title= "outcome 3 mesi", margin = "col",
  show.total = FALSE)
```



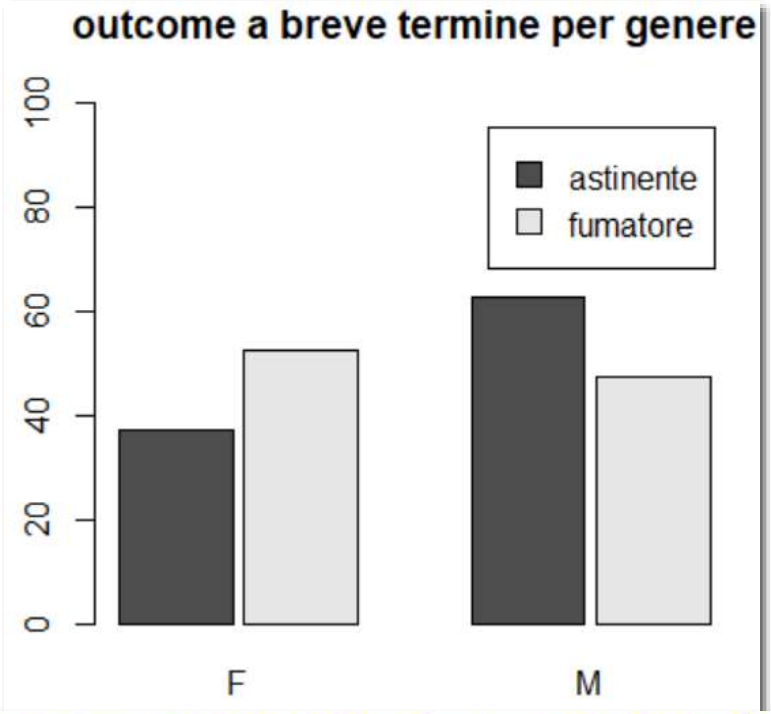



```
plot_xtab(x = fumo$genere, grp= fumo$outcome_3_mesi,
  legend.title="outcome 3 mesi", margin = "col")
```

Barplot si applica anche a tabelle $r \times c$: `barplot(table(x1, x2))` o `barplot(prop.table(table(x1,x2),margin=1 o 2)`: `beside=TRUE, space=c(tra le barre, tra le categorie)` sono argomenti opzionali che cambiano il layout del plot. In `space`, di default lo spazio tra le barre è `=0` (adiacenti), tra le categorie è `=1`.




```
barplot(prop.table(table(fumo$outcome_3_mesi, fumo$genere),
margin = 2)*100, main="outcome a breve termine per genere",
ylim=c(0,100), legend=TRUE)
```




```
barplot(prop.table(table(fumo$outcome_3_mesi, fumo$genere),
margin = 1)*100, main="outcome a breve termine per genere",
ylim=c(0,100), legend=TRUE, beside = TRUE, space = c(.1,1))
```

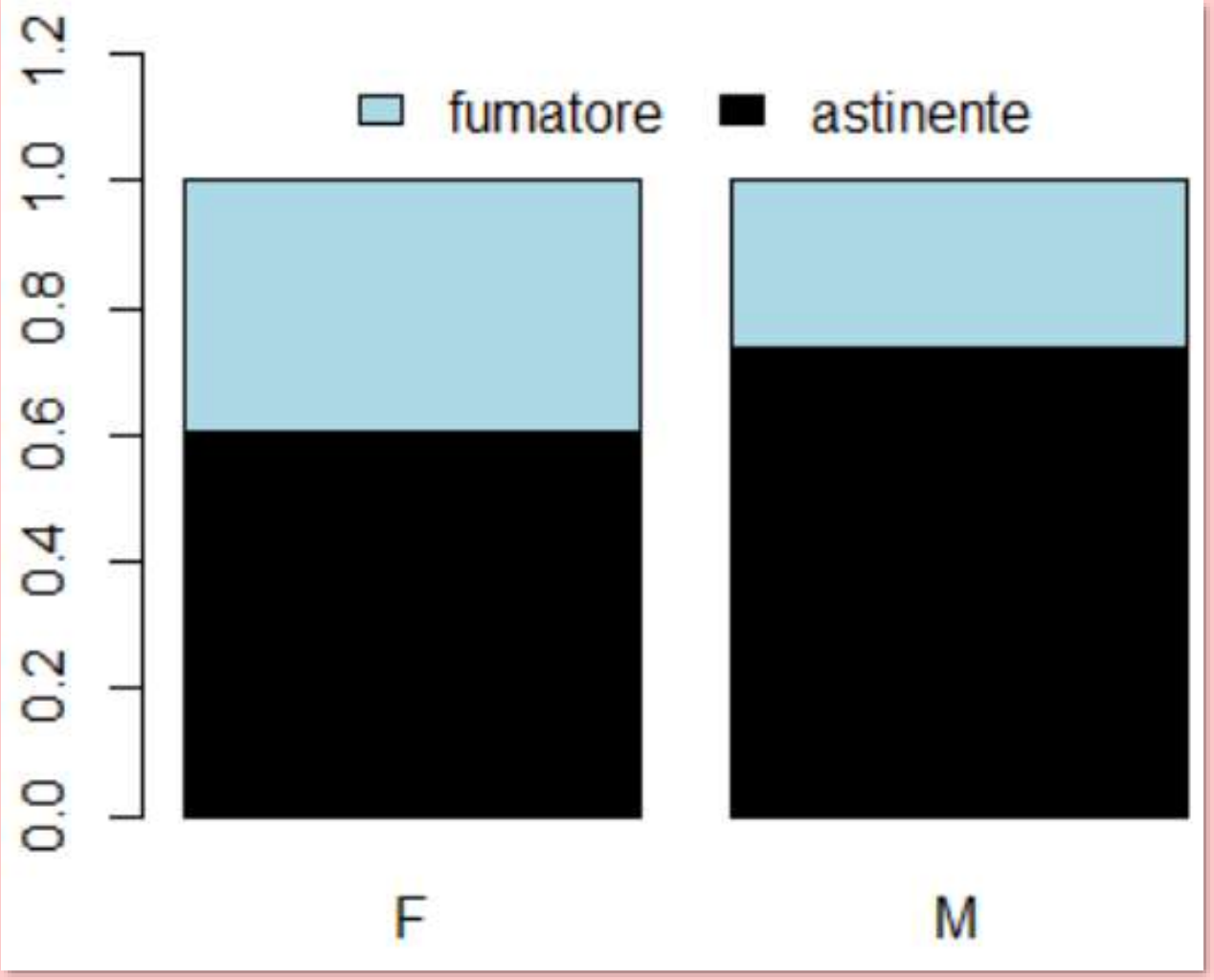
La legenda: un elemento potenzialmente fastidioso

Si può aggiungere una legenda a un qualsiasi grafico con la funzione `legend` (dopo averlo costruito), ma `barplot` consente di inserirla direttamente tra i propri argomenti con `legend= TRUE`. Bisogna evitare che si sovrapponga alle barre, per cui può essere utile aggiungere anche `args.legend= list(,)` in cui si può indicare `x= "top"`/ `"topright"`/ `"topleft"`/ `"bottom"`/ `"bottomright"`/ `"bottomleft"`, oltre a:

- Bordo e sfondo della legenda: `bty=` ; per non visualizzarli, `bty="n"`
- Colore del bordo dei quadratini: `border= "color"`
- Grandezza dei caratteri: `cex.names=` per le categorie in X, `cex.names=` per le voci della legenda
- Colonne in cui disporre la legenda: `ncol=` ; di default è `ncol= 1`
- `ecc…`

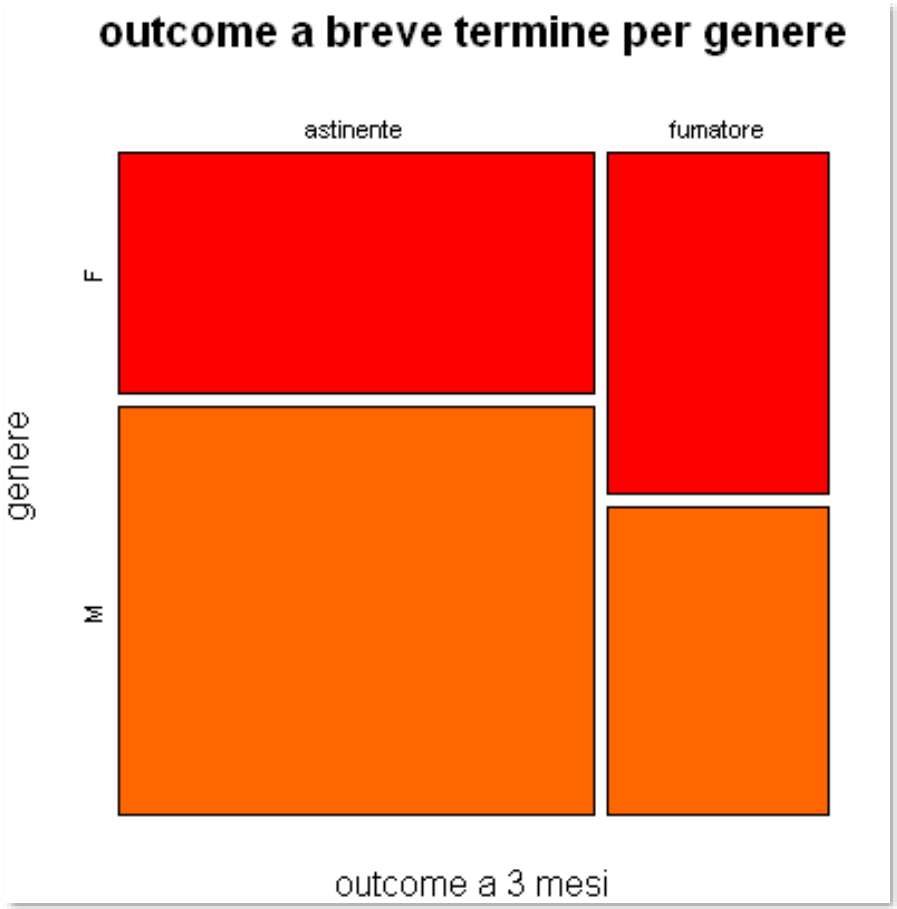
Potete anche costruire il vettore desiderato dei colori di ogni barra e indicarlo nell'argomento `col=` del `barplot`:

```
colori_legenda<-c("black", "light blue")
barplot(prop.table(table(fumo$outcome_3_mesi,  fumo$genere),2),
col= colori_legenda, legend= TRUE, args.legend = list(x="top",
ncol=2, bty="n"), ylim= c(0,1.2))
```



Un'altra rappresentazione grafica per l'associazione tra variabili categoriali è il `mosaicplot(table(x1, x2))`:

```
mosaicplot(table(fumo$genere,fumo$outcome_3_mesi), col=rainbow(15), xlab="genere",
 ylab="outcome", main="associazione genere e status a tre mesi")
```




```
round(prop.table(table(fumo$genere,
 fumo$outcome_3_mesi))*100,1)
     astinente fumatore
F         25.4     16.7
M         42.9     15.1
```

L'area di ogni rettangolo è **proporzionale alla probabilità condizionale** della cella, ovvero al **numero di frequenze osservate** che contiene rispetto al **totale**.

VERIFICARE IPOTESI SULL'ASSOCIAZIONE TRA DUE VARIABILI CATEGORIALI, **DATI INDIPENDENTI***

***In tutti i disegni con casi indipendenti, un soggetto è conteggiato in una sola cella.**

La predominanza dei maschi tra gli astinenti a tre mesi **potrebbe essere solo un caso**:

H_0: la **forma della distribuzione** dei dati è **rettangolare**: tutte le categorie si manifestano con la medesima frequenza → la diversità delle frequenze è solo una **fluttuazione casuale** → le variabili sono **indipendenti, cioè non associate in popolazione** (essere maschio o femmina **non cambia** la probabilità di riuscire a smettere di fumare) → nel mosaic plot i rettangoli **avrebbero un'area identica**.

H_1: la **forma della distribuzione** dei dati **non** è **rettangolare**: categorie di eventi si presentano con frequenze non casualmente diverse → le variabili sono **associate in popolazione**: essere maschio o essere femmina **cambia**, in meglio o in peggio, la probabilità di essere astinente o fumatore dopo tre mesi.

Dobbiamo determinare se le differenze tra le frequenze riscontrabili nella tabella di contingenza (o nel mosaic plot) siano **fluttuazioni casuali** di quella che è in realtà una distribuzione rettangolare in popolazione, o se esprimano una reale associazione tra le variabili in popolazione

Sono molti i modelli statistici per **attribuire un *p*-value** al dato sotto condizione di H_0 per avere un aiuto sulla decisione da prendere, nonché gli **indicatori di effect size / forza dell'associazione**.

Vedremo solo i principali strumenti:

odds ratio, **test** del **chi quadrato** e **test di Fisher**.

Odds e odds ratio

Soprattutto per verificare ipotesi **sull'efficacia di trattamenti** (**outcome favorevole**), o sull'azione di fattori di **rischio** nel manifestarsi di diverse patologie (**outcome sfavorevole**), si **rapporta la proporzione dei casi di un gruppo G_1 che presentano l'outcome** indagato (ad esempio, astinenti del gruppo counseling) **con la proporzione dei casi di un gruppo G_2 che presentano lo stesso outcome** (astinenti del gruppo vareniclina): .

Il rapporto è un odds

L'**odds a favore** del verificarsi di un evento A è dato dal **rapporto tra la probabilità che l'evento A si verifichi e la probabilità che si verifichi l'evento non atteso $\neg A$: $P(A) \,/\, P(\neg A)$.**

L'odds contro il verificarsi di un evento A è dato dal rapporto tra la probabilità che si verifichi **l'evento non atteso $\neg A$** e la **probabilità che si verifichi l'evento atteso A**: $P(\neg A)) / P(A))$

Contiamo le frequenze di un outcome atteso ($A - successo$) e di un outcome non atteso ($\neg A$ – $non\ successo$) in due gruppi (**sperimentale** e **controllo**).

	Successo	Non successo
Sperimentale	a	b
Controllo	c	d

Odds $G_{sperimentale}$: $P(A) \rightarrow a/(a+b)$ in rapporto alla $P(\neg A) \rightarrow b/(a+b)$

Odds $G_{controllo}$: $P(A) \rightarrow c/(c+d)$ in rapporto alla $P(\neg A) \rightarrow d/(c+d)$

Semplificando:

$$odds_{sper} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{a+b} \times \frac{a+b}{b} = \frac{a}{\cancel{a+b}} \times \frac{\cancel{a+b}}{b} = \frac{a}{b}$$

$$odds_{crl} = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{c+d} \times \frac{c+d}{d} = \frac{c}{\cancel{c+d}} \times \frac{\cancel{c+d}}{d} = \frac{c}{d}$$

Il **range** dei possibili valori di un odds **va da 0** (cella vuota al numeratore) **a infinito** (cella vuota al denominatore)

```
0/27; 27/0
[1] 0
[1] Inf
```

Il **rapporto tra due odds** si definisce **odds ratio (OR)**:

$$OR = \frac{odds_{G1}}{odds_{G2}}$$

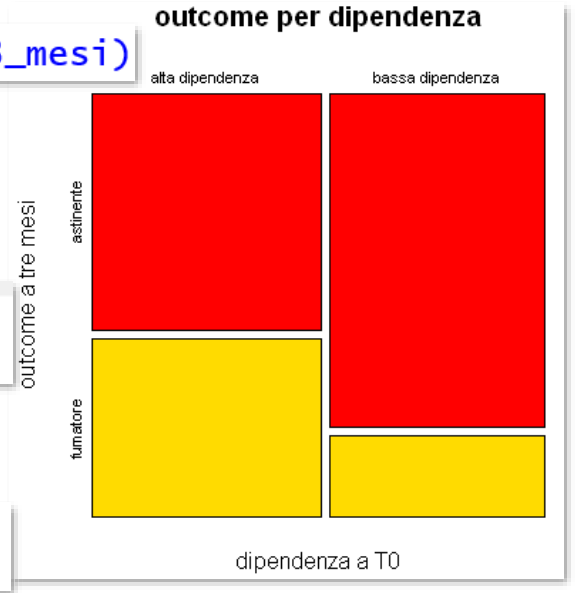
È una **MISURA DI RISCHIO**, che esprime **quanto è più (o meno) probabile che si verifichi A nel gruppo G_1 rispetto a quanto è probabile nel gruppo G_2**. Come gli odds, **l'OR varia da 0** ($odds = 0$ al numeratore) a infinito ($odds = 0$ al denominatore).

- **$OR = 1$: la probabilità dell'evento atteso in G_1 è uguale a quella riscontrata nel G_2**: appartenere a un gruppo o all'altro **non cambia la probabilità del successo** (H_0)
- **$OR > 1$: la probabilità dell'evento atteso in G_1 è maggiore di quella rilevata nel G_2 (H_1)**
- **$OR < 1$: la probabilità dell'evento atteso in G_1 è minore di quella rilevata nel G_2 (H_1)**

È possibile **assegnare un $p - value$ all'OR** per stabilire se sia una **casuale fluttuazione da $H_0 = 1$**: tra poco lo faremo nel **test di Fisher**, nella **regressione logistica** useremo il metodo della **massima verosimiglianza**. In entrambi i casi, si usa la **distribuzione di probabilità χ^2**.

Verifichiamo se la **probabilità a 3 mesi di smettere** di fumare per chi a T_0 aveva una **forte dipendenza** è la **stessa probabilità per chi aveva una ridotta dipendenza:**

```
dipendenza<-table(fumo$Fagerstrom_categorie, fumo$outcome_3_mesi)
round(prop.table(dipendenza,1)*100,1)
                  astinente fumatore
alta dipendenza        56.9     43.1
bassa dipendenza       80.3     19.7
```



Calcoliamo gli odds:

```
dipendenza
                  astinente fumatore
alta dipendenza          37       28
bassa dipendenza         49       12
```

```
(odds_alta<-37/28)
[1] 1.321429
```

```
(odds_bassa<-49/12)
[1] 4.083333
```

In **entrambi** i gruppi, la **probabilità di essere astinenti è maggiore** di quella di non ricavare beneficio ($odds > 1$), ma **quella del gruppo di pazienti con bassa dipendenza è oltre 3 volte più grande**: il rapporto $^{1.3}/_{4.1}$ è appunto **l'OR.**

Non ci sono funzioni di base per l'OR, ma è facile calcolarlo:

```
(odds_ratio<-odds_alta / odds_bassa)
[1] 0.3236152
```

L'OR si può **calcolare anche direttamente dalle frequenze**, come **prodotto incrociato**:

	Successo	Non successo
Sperimentale	a	b
Controllo	c	d

Infatti: $OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \times \frac{d}{c} = \frac{a \times d}{b \times c}$

OR: il **prodotto** dei **successi** in G_1 per gli **insuccessi** in G_2 ($a \times d$) in **rapporto** al **prodotto** dei **successi** in G_2 per gli **insuccessi** in G_1 ($c \times d$).

$$OR = \frac{a \times d}{b \times c}$$

Possiamo quindi calcolarlo così:

```
(OR<-(dipendenza[1,1]*dipendenza[2,2])/
    (dipendenza[1,2]*dipendenza[2,1]))
[1] 0.3236152
```

... o così:

```
dipendenza
                  astinente fumatore
alta dipendenza          37       28
bassa dipendenza         49       12
```

```
(OR<-(37*12)/(28*49))
[1] 0.3236152
```

*La probabilità di essere astinenti nel gruppo Alta dipendenza è di quasi **tre volte minore** della probabilità di essere astinenti nel gruppo bassa Dipendenza.*

```
1/OR
[1] 3.09009
```

Si può stabilire la **significatività** dell'OR (H_0: $OR = 1$) e calcolarne il ***CI***.

l'OR, però, segue una **distribuzione log-normale**, con forte **asimmetria destra**, che può essere **normalizzata facendone il logaritmo**.

$$z_{OR} = \frac{\log(OR)}{SE_{\log(OR)}}$$

In questo modo si può usare la distribuzione normale z:

$$SE_{\log(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

```
ES_ln<-sqrt(1/37+1/28+1/49+1/12)
```

```
log(OR)/ES_ln
[1] -2.76504
```

	astinente	fumatore
alta dipendenza	37	28
bassa dipendenza	49	12

```
pnorm(-2.76504,0,1, lower.tail = TRUE)
[1] 0.002845791
```

L'associazione è significativa

Per il CI:

$$CI_{\log(OR)} = \log(OR) + z_{\alpha/2} SE_{\log(OR)}$$

$$CI_{OR} = e^{\log(OR) + z_{\alpha/2} SE_{\log(OR)}}$$

Esponenziale: inverso della funzione logaritmica, in R **exp**

```
UL<-exp(log(OR)+(1.96*ES_ln))
LL<-exp(log(OR)-(1.96*ES_ln))
round(c(LL, OR, UL),3)
[1] 0.145 0.324 0.720
```

```
.324-.145
[1] 0.179
```

```
.720-.324
[1] 0.396
```

I limiti del CI NON sono simmetrici attorno all'OR (questo vale per **qualsiasi** proporzione) e **non contengono 1**.

Tranquilli, basta usare `DescTools`: `Desc(table)` ci dà OR e il suo CI, oltre alle descrittive e alla significatività dell'associazione con i test che vedremo subito.

Se servono solo OR e CI, basta `OddsRatio(table, interval=.95)`.

```
Desc(table(fumo$Fagerstrom_categorie, fumo$outcome_3_mesi),plotit=FALSE)
------------------------------------------------------------
table(fumo$Fagerstrom_categorie, fumo$outcome_3_mesi) (table)
Summary:
n: 126, rows: 2, columns: 2
Pearson's Chi-squared test (cont. adj):
  X-squared = 6.912, df = 1, p-value = 0.008562
Fisher's exact test p-value = 0.007021
McNemar's chi-squared = 5.1948, df = 1, p-value = 0.02265
                     estimate lwr.ci upr.ci'
odds ratio              0.324  0.145  0.720
```

```
OddsRatio(table(fumo$Fagerstrom_categorie, fumo$outcome_3_mesi), conf.level = .95)
odds ratio     lwr.ci     upr.ci
 0.3236152  0.1454518  0.7200103
```

Il dataframe contiene anche ***l'outcome a lungo termine****, cioè a distanza di un anno dalla fine del trattamento:* ***fumo$outcome_12_mesi.*** *Calcolate gli odds e l'OR per questo outcome rispetto alla gravità di dipendenza: quali conclusioni potete trarre?*

Anche la ***depressione*** *può essere una motivazione per fumare: calcolate gli odds e l'OR degli outcome, sia a breve sia a lungo termine, considerando la variabile categoriale* ***fumo$Zung_categorie e ricategorizzando i livelli dal fattore in modo da renderlo dicotomico*** *(“****depresso****” e “****non depresso****)”: quali conclusioni potete trarre?*

Il test χ^2 a due vie o test d'indipendenza

Con casi **indipendenti**, per verificare l'ipotesi di associazione in popolazione useremo il **test del χ^2 a due vie** (Pearson, 1900; Fisher, 1922). La **logica** e la **formula** del test sono **uguali** a quelle del χ^2 a una via : **confrontare le frequenze osservate** (O) con quelle **attese** (A) :

$$\chi^2_{(r-1)(c-1)} = \sum \frac{(O-A)^2}{A} \qquad A = \frac{marginale_{riga} \times marginale_{colonna}}{N}$$

Esiste un'associazione tra genere e outcome? Uomini e donne hanno la stessa probabilità di smettere di fumare?

```
(osservate<-table(fumo$genere, fumo$outcome_3_mesi))
   astinente fumatore
F         32        21
M         54        19
```

```
OddsRatio(osservate)
[1] 0.5361552
```

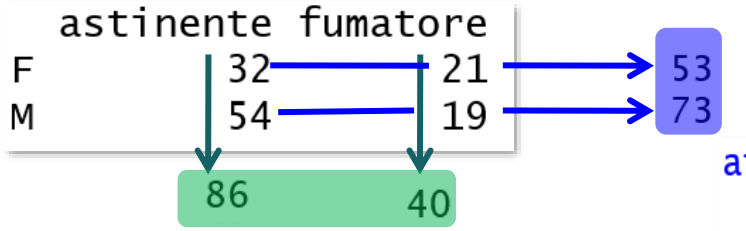
Calcoliamo A; **prima i marginali di riga e colonna** con `margin.table(table)`, che rendiamo matrici per il successivo calcolo:

```
marginali_riga<-as.matrix(margin.table(osservate,1))
```

```
marginali_riga
  [,1]
F   53
M   73
```

```
marginali_colonna
          [,1]
astinente   86
fumatore    40
```

```
marginali_colonna<-as.matrix(margin.table(osservate,2))
```



$$A = \frac{marginale_{riga} \times marginale_{colonna}}{N}$$

```
attese<-cbind(marginali_colonna[1]*marginali_riga/126,
   marginali_colonna[2]*marginali_riga/126)
colnames(attese)<-c("astinente", "fumatore")
```

```
attese
  astinente fumatore
F   36.1746  16.8254
M   49.8254  23.1746
```

```
osservate
  astinente fumatore
F        32       21
M        54       19
```

```
scarti<-(as.matrix(osservate)-attese)
scarti
       astinente  fumatore
F -4.174603   4.174603
M  4.174603 -4.174603
```

```
sum(scarti)
[1] 0
```

Ora gli scarti **al quadrato divisi per le A**: la loro somma è la statistica χ^2: $\chi^2_{(r-1)(c-1)} = \sum \frac{(O-A)^2}{A}$

```
scarti2<-(as.matrix(osservate)-attese)^2/attese
scarti2
       astinente  fumatore
F     0.4817554 1.0357742
M     0.3497676 0.7520004
```

```
(chi2<-sum(scarti2))
[1] 2.619298
```

probabilità di $\chi^2 \geq 2.61$, per $df = 2 - 1 \times 2 - 1$, sotto condizione di H_0:

```
pchisq(q =chi2,df= 1,lower.tail = FALSE)
[1] 0.1055711
```

Confermiamo H_0: **non c'è associazione** in popolazione tra genere ed outcome

SE l'associazione è significativa, ulteriori informazioni si ottengono **dalle celle degli scarti**

Residui di cella standardizzati r_{st}: come **punti z**, il **segno** indica se nella cella ci sono **più** o **meno osservazioni di quelle attese** in base al caso; il **valore assoluto** indica **quali** celle hanno **più contribuito** all'associazione.

$$r_{st_{ij}} = \frac{O_{ij} - A_{ij}}{\sqrt{A_{ij}}}$$

Residui di cella standardizzati corretti r_{adj}: ogni r_{st}, **diviso per la variabilità di tutti i residui**, si distribuisce come **un quantile z di una distribuzione normale standardizzata**:

$$r_{ad_{ij}} = \frac{O_{ij} - A_{ij}}{\sqrt{A_{ij}\left(1 - \frac{marginale_{ri}}{N}\right)\left(\frac{marginale_{cj}}{N}\right)}}$$

Quindi possiamo usare i r_{adj} per sapere quali O sono significativamente maggiori o minori delle A: basta ricordarsi $z = |1.96|$ per H_1 bidirezionale.

Per non calcolare r_{st} e r_{adj} con le matrici, usiamo un trucco: un **test inferenziale può essere salvato come oggetto** di classe `htest`, `lm`, `glm`; sono tutte **liste**, che contengono (`oggetto_test$`) varie informazioni, tra cui appunto i residui di cella.

Salviamo come oggetto `chisq.test(X1, X2)`, o `chisq.test(table)`:

```
modello_chi<-chisq.test(fumo$genere,fumo$outcome_3_mesi, correct = FALSE)
```

```
modello_chi
        Pearson's Chi-squared test
data:  fumo$genere and fumo$outcome_3_mesi
X-squared = 2.6193, df = 1, p-value = 0.1056
```

$$r_{st_{ij}} = \frac{O_{ij} - A_{ij}}{\sqrt{A_{ij}}}$$

```
modello_chi$
modello_chi$statistic
modello_chi$parameter
modello_chi$p.value
modello_chi$method
modello_chi$data.name
modello_chi$observed
modello_chi$expected
modello_chi$residuals
modello_chi$stdres
```

```
modello_chi$residuals
                fumo$outcome_3_mesi
fumo$genere      astinente   fumatore
          F -0.6940860   1.0177299
          M  0.5914116 -0.8671796
```

r_{st}

```
scarti/sqrt(attese)
     astinente   fumatore
F -0.6940860   1.0177299
M  0.5914116 -0.8671796
```

```
modello_chi$stdres
                fumo$outcome_3_mesi
fumo$genere      astinente   fumatore
          F -1.618424   1.618424
          M  1.618424 -1.618424
```

r_{adj}

In nessuna cella ci sono $r_{adj} \geq 1.96$:

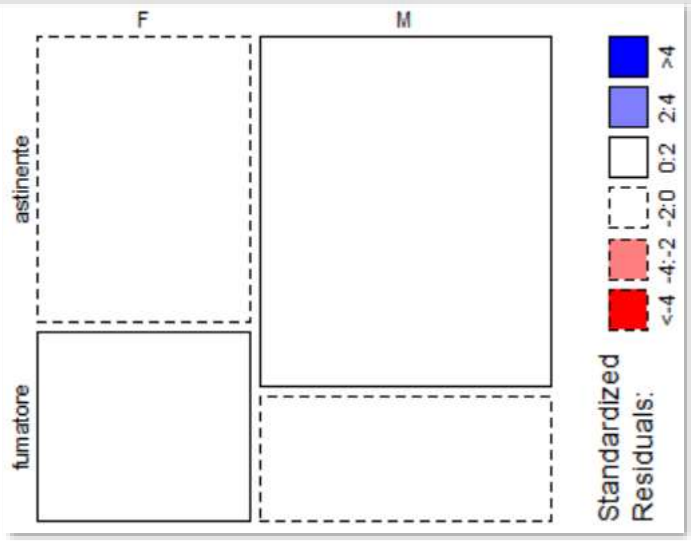
Tutti gli elementi dell'output fanno parte della lista:

```
modello_chi$method
[1] "Pearson's Chi-squared test"
modello_chi$data.name
[1] "fumo$genere and fumo$outcome_3_mesi"
```

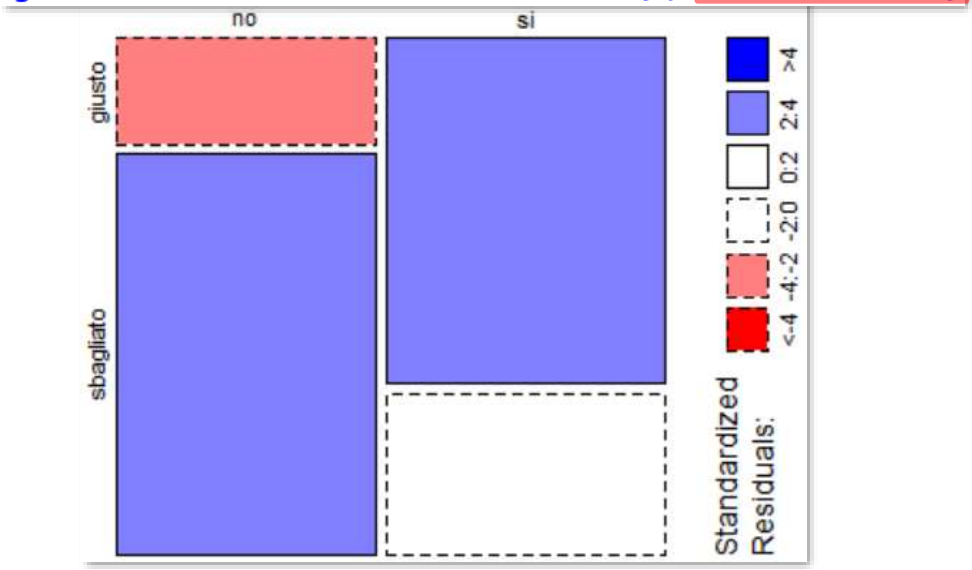
```
modello_chi$statistic
X-squared
 2.619298
modello_chi$parameter
df
 1
modello_chi$p.value
[1] 0.1055712
```

Indicando shade=TRUE nel mosaicplot, **le tessere sono colorate a seconda dell'intensità dei residui standardizzati**: **bianche**: r_{st} compresi tra |0| e |2|; **azzurre e blu**: r_{st} positivi da 2 a ≥ 4; **rosa e rosse**: r_{st} negativi da -2 a ≤ 4.

```
mosaicplot(osservate), shade = TRUE)
```




```
mosaicplot(table(gatti$vive_con_gatto,
gatti$riconosce_miao_isolamento), shade = TRUE)
```




```
modello_chi$residuals
              fumo$outcome_3_mesi
fumo$genere     astinente    fumatore
          F -0.6940860   1.0177299
          M  0.5914116  -0.8671796
```

```
miao<-chisq.test(gatti$vive_con_gatto,
   gatti$riconosce_miao_isolamento)
miao$residuals
gatti$vive_con_gatto     giusto sbagliato
                  no -2.238831  2.048511
                  si  2.155366 -1.972142
```

Può servire per visualizzare quale cella pesi di più nel determinare l'associazione e la sua direzione, ma, ***non essendo visualizzati i r_{adj}, non per la significatività*** *del residuo di cella.*

Correzione per la continuità (**Yates**), per tabelle $2x2$:

$$\chi^2_{corretto} = \sum \frac{(|O - A| - 0.5)^2}{A}$$

Le O non possono variare per unità frazionarie: uno scarto $O - A$, anche quando $= |1|$, "apparirà" grande **quando le O sono piccole**: la somma di grandi residui crea χ^2 **grandi**, probabilmente **afflitti da errori di I tipo**. La correzione riduce il χ^2, diminuendo la probabilità di errori di I tipo, ma aumentando **l'errore di II tipo** (Howell, 2006). Quindi, anche se R offre l'opportunità di calcolare la correzione (`correct=TRUE`), **ignoratela pure**

```
chisq.test(fumo$genere,fumo$outcome_3_mesi, correct = FALSE)
        Pearson's Chi-squared test
data:  fumo$genere and fumo$outcome_3_mesi
X-squared = 2.6193, df = 1, p-value = 0.1056
```

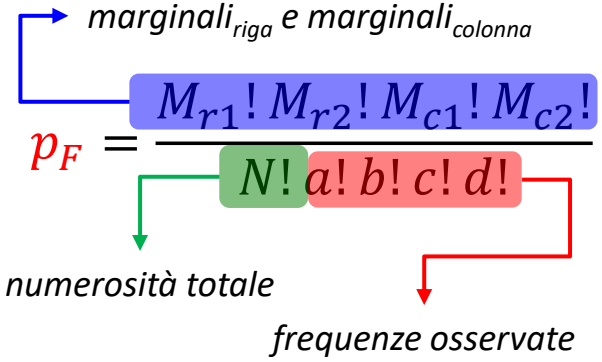
```
chisq.test(fumo$genere,fumo$outcome_3_mesi, correct=TRUE)
        Pearson's Chi-squared test with Yates' continuity correction
data:  fumo$genere and fumo$outcome_3_mesi
X-squared = 2.0294, df = 1, p-value = 0.1543
```

Il test di Fisher o della probabilità esatta

Il **test della probabilità esatta di Fisher** risolve un **grave problema** del test χ^2 : con campioni relativamente grandi, la distribuzione χ^2 si approssima a una distribuzione di probabilità χ^2, ma con **piccoli campioni l'approssimazione è inadeguata**, rendendo inaffidabili i $p - value$.

In tabelle $2x2$ il test χ^2 è interpretabile se tutte le $A > 5$; in tabelle $r \times c$, **le celle con $A < 5$ non devono essere $> 20\%$ delle celle della tabella** (e comunque il test perde in potenza), e **in nessuna** cella deve esserci una $A < 1$ (Howell, 2006).

Se in una tabella 2×2 il requisito è violato, meglio usare il **test esatto di Fisher:** è un **metodo per stimare la probabilità esatta** del dato usando il calcolo **combinatorio**.



p_F si distribuisce secondo la distribuzione **ipergeometrica**; è la somma della probabilità di ottenere la disposizione delle O e di ogni altra disposizione che dia **uguale o maggiore evidenza dell'associazione tra X_1 e X_2**, tenuti fissi i M_r e M_c.

Stimiamo l'associazione tra **genere e outcome a 12 mesi** nel solo gruppo che ha seguito il counseling: `counseling ← subset(fumo, fumo$terapia=="counseling")`

```
(dodici<-table(counseling$genere,counseling$outcome_12_mesi))
      astinente fumatore
  F           3         3
  M           5         2
```

```
chi12<-chisq.test(dodici, correct=FALSE)
Warning message:
In chisq.test(dodici, correct = FALSE) :
  L'approssimazione al Chi-quadrato potrebbe essere inesatta
```

Il campione è molto piccolo, e il warning di `chisq.test` ci informa della violazione del requisito:

```
chi12$expected
   astinente fumatore
F  3.692308 2.307692
M  4.307692 2.692308
```

Usiamo `fisher.test(X1, X2)` o `fisher.test(table)`: di default dà OR (`or=TRUE`) e il suo CI (`confint=.95`)

```
fisher.test(dodici)
        Fisher's Exact Test for Count Data
data:  dodici
p-value = 0.5921
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.02225065 6.32392290
sample estimates:
odds ratio
 0.4303728
```

Il $p-value$ è superiore alla soglia $\alpha = .05$ e nel CI del $95\% CI$ è compreso il valore $H_0: OR = 1$

A differenza del test χ^2, il test di Fisher consente di verificare ipotesi **monodirezionali**: $OR > 1$ o $OR > 1$ (`alternative= "greater" / "less"`).

```
fisher.test(dodici, alternative = "g")
        Fisher's Exact Test for Count Data
data:  dodici
p-value = 0.9138
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.03388761       Inf
sample estimates:
odds ratio
 0.4303728
```

*L'UL del CI è il **massimo teorico** del range di variazione dell'OR = in[de]finito*

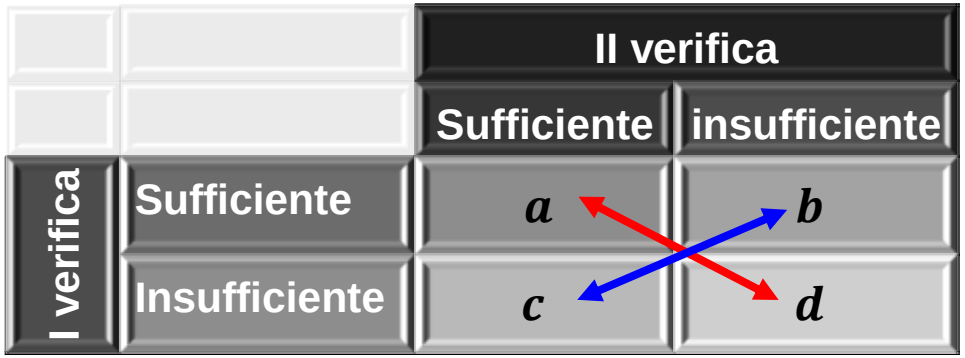
```
fisher.test(dodici, alternative = "l")
        Fisher's Exact Test for Count Data
data:  dodici
p-value = 0.4126
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 4.473644
sample estimates:
odds ratio
 0.4303728
```

*Il LL del CI è il **minimo teorico** del range di variazione dell'OR = 0*

Il test si può applicare a tabelle $r \times c$ grazie all'estensione di **Freeman e Halton** (1951), anche se il test χ^2, se i requisiti sono rispettati, è più potente e preferibile. R non ha problemi ad applicare `fisher.test` a tabelle più grandi di 2×2.

VERIFICARE IPOTESI SULL'ASSOCIAZIONE TRA DUE VARIABILI APPAIATE - **DATI DIPENDENTI**

L'indipendenza dei dati **non** è rispettata nei disegni **within subjects**: nei disegni **longitudinali** si conteggia lo stesso individuo (almeno) in due condizioni: alla **baseline** e a T_1.



$a \leftrightarrow d$: **diagonale risultati coerenti**

$b \leftrightarrow c$: **diagonale risultati incoerenti**

H_0: i soggetti cambiano o restano uguali in maniera **casuale** → in ogni cella c'è una numerosità simile, e la diversità delle frequenze è solo una **fluttuazione casuale** → le variabili sono **indipendenti, cioè non associate in popolazione.**

H_1: i soggetti cambiano o restano uguali in maniera **non casuale** → le frequenze **si addensano attorno a una delle due diagonali**, e nell'altra ci sono pochi o nessun caso → le variabili sono **dipendenti, cioè associate in popolazione.**

Per tabelle 2×2, il **test di McNemar** consente di attribuire un p-value ai dati sotto condizione di H_0, utilizzando la distribuzione di probabilità χ^2 per $df = 1$:

Corso di formazione su tema sociale		T1	
		Pro	Contro
T0	Pro	a	b
	Contro	c	d

Si basa sulla diagonale delle celle dei i casi **incoerenti**:

$$\chi^2_M = \frac{(b-c)^2}{b+c}$$

Per tabelle $r \times c$, si usa il **test di McNemar-Bowker**, che valuta la **simmetria dei dati attorno alla diagonale delle frequenze coerenti**: **H_0 è verificata tramite chi ha cambiato categoria.**

Corso di formazione su tema sociale		T1		
		Pro	?	Contro
T0	Pro	x_{11}	x_{12}	x_{13}
	?	x_{21}	x_{22}	x_{23}
	Contro	x_{31}	x_{32}	x_{33}

Differenza tra le O a **destra** della diagonale e quelle che a **sinistra** occupano la posizione simmetrica

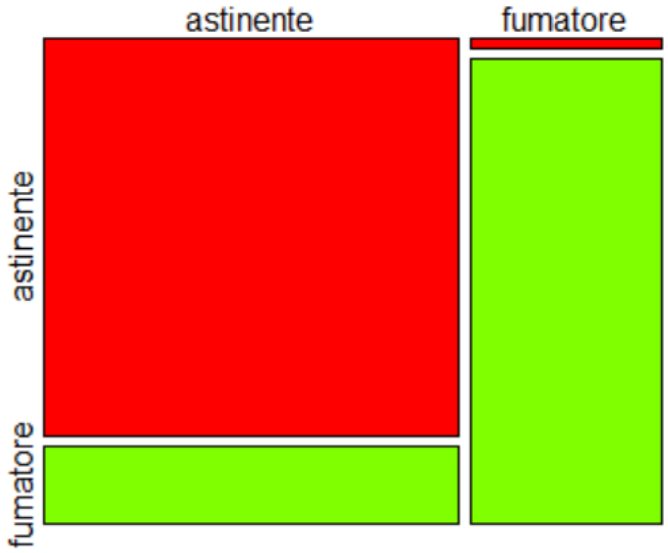
$(X_{12} \leftrightarrow X_{21}, X_{13} \leftrightarrow X_{31}, X_{23} \leftrightarrow X_{33}$

$$\chi^2_{MB} = \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

Totale delle O a **destra** e a **sinistra** della diagonale

`mcnemar.test(X1,X2, correct=TRUE/FALSE)` dà il test di McNemar/McNemar-Bowker.

*C'è associazione tra gli **outcome a 3 e a 12 mesi**? Chi era astinente a breve termine è rimasto tale a distanza di un anno? E chi era fumatore?*




```
(outcome<-table(fumo$outcome_3_mesi, fumo$outcome_12_mesi))
           astinente fumatore
astinente         72        14
fumatore           1        39
```

***L'associazione sembra buona**: astinenti e fumatori a 3 mesi restano perlopiù immutati a 12 mesi. Tra gli incoerenti, sembra più facile riprendere a fumare dopo aver smesso, che fare il contrario.*

$$\chi^2_M = \frac{(b-c)^2}{b+c}$$

```
(mcnemar<-(14-1)^2/(14+1)
[1] 11.26667
```

```
pchisq(11.267,df = 1,lower.tail = FALSE)
[1] 0.0007891116
```

```
mcnemar.test(outcome, correct = FALSE)
        McNemar's Chi-squared test
data:  outcome
McNemar's chi-squared = 11.267, df = 1, p-value = 0.0007891
```

***Rifiutiamo H_0**: esiste una'associazione non casuale tra l'outcome a breve e a lungo termine.*

LIMITI DEI TEST DI ASSOCIAZIONE E COEFFICIENTI D'INTENSITÀ

Tutti i test basati sulla distribuzione χ^2 sono **molto potenti per grandi *N*** (aumenta il rischio di errore α), quanto **poco potenti nel caso di piccoli *N*** (aumenta il rischio di errore β).

Creiamo due distribuzioni bivariate: otteniamo AB_2 *moltiplicando per due le O in ogni cella di* AB*. I due campioni hanno* ***diverse frequenze assolute, cioè diverso N*** *($N_A = 100$, $N_B = 200$)****, ma le frequenze relative sono identiche in ogni cella****.*

```
(AB<-table(A, B))
     B
      b1 b2
A a1  15 35
  a2  30 20
```

```
prop.table(AB)
          B
           b1   b2
A a1 0.15 0.35
  a2 0.30 0.20
```

```
(AB_2<-AB*2)
     B
      b1 b2
A a1  30 70
  a2  60 40
```

```
prop.table(AB_2)
          B
           b1   b2
A a1 0.15 0.35
  a2 0.30 0.20
```

```
chisq.test(AB, correct = FALSE)
        Pearson's Chi-squared test
data:  AB
X-squared = 9.0909, df = 1, p-value = 0.002569
```

```
chisq.test(AB_2, correct = FALSE)
        Pearson's Chi-squared test
data:  AB_2
X-squared = 18.182, df = 1, p-value = 2.008e-05
```

Il χ^2 di AB_2 è il doppio del χ^2 di AB, e il suo il $p - value$ è più lontano dalla soglia α:

Il **quantile χ^2 aumenta quando lo scarto tra O e A è moltiplicato per una quantità costante**, anche se le *f* restano uguali come % sul totale e nei loro rapporti.

```
(OR_AB<-(15*20)/(35*30))
[1] 0.2857143
```

```
(OR_AB_2<-(30*40)/(70*60))
[1] 0.2857143
```

Tabelle con N **più ridotti**, anche con uguali proporzioni nelle celle, rendono **più probabile accettare H_0**. Creiamo CD, con f_r identiche a quelle di AB e AB_2 e *f* inferiori ($N_c = 40$):

```
C<-c(rep("c1",20), rep("c2",20))
D<-c(rep("d1",6), rep("d2",14), rep("d1",12), rep("d2",8))
```

```
(CD<-table(C,D))
    D
     d1 d2
C c1  6 14
  c2 12  8
```

```
prop.table(CD)
     D
      d1   d2
C c1 0.15 0.35
  c2 0.30 0.20
```

```
chisq.test(CD, correct = FALSE)
        Pearson's Chi-squared test
data:  CD
X-squared = 3.6364, df = 1, p-value = 0.05653
```

```
(OR_CD<-(6*8)/(14*12))
[1] 0.2857143
```

È quindi necessario associare al test **indici ponderati per N**, che stimino **l'intensità dell'associazione**, con un **range di variazione tra 0 (indipendenza) e 1 (perfetta associazione)**: ne vediamo solo alcuni, affiancando alle procedure di calcolo le funzioni di DescTools.

Il primo coefficiente di intensità di associazione è stato il **coefficiente di contigenza C di Pearson**:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

non corretto

Dovrebbe variare da 0 (nessuna associazione) a 1, ma il suo il **limite superiore** ($\sqrt{(k-1)/k}$; k: il **numero minore** tra r e c) **tende a (ma non raggiunge) 1 solo per tabelle molto grandi** (almeno 5×5). Ad esempio:

C_{max} **se** $k = 2$	C_{max} **se** $k = 3$	C_{max} **se** $k = 5$	C_{max} **se** $k = 12$
`sqrt((2-1)/2)`	`sqrt((3-1)/3)`	`sqrt((5-1)/5)`	`sqrt((12-1)/12)`
`[1] 0.7071068`	`[1] 0.8164966`	`[1] 0.8944272`	`[1] 0.9574271`

Sakoda propone la correzione C_{adj} per estendere fino a 1 il range di C, indipendentemente dalla grandezza della tabella:

$$C_{adj} = \frac{C}{C_{max}}$$

```
chi_AB<-chisq.test(AB, correct = FALSE)
```

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

```
sqrt(chi_AB$statistic/(chi_AB$statistic+100))
0.2886751
```

C non corretto: associazione debole

$$C_{adj} = \frac{C}{C_{max}}$$

```
sqrt(chi_AB$statistic/(chi_AB$statistic+100)) / 0.7071068
0.4082483
```

C corretto: associazione moderata

`ContCoef(x1,x2, correct=TRUE/FALSE)` di `DescTools` dà C o C_{adj}:

```
ContCoef(A,B,correct=FALSE)
[1] 0.2886751

ContCoef(A,B,correct=TRUE)
[1] 0.4082483

ContCoef(AB,correct=TRUE)
[1] 0.4082483
```

Cohen (1977) propone di usare C (non corretto) per **stimare quale sia l'apporto di N alla significatività del test χ^2: indice w.**

$$w = \sqrt{\frac{C^2}{1 - C^2}}$$

effetto di N: $w \leq .30$ *small*; $w > .30 \leq .50$ *medium*; $w > .50$ *large*

```
C_AB<-ContCoef(A,B,correct=FALSE)
w_AB<-sqrt(C_AB^2/(1-(C_AB^2)))
w_AB
[1] 0.3015113
```

Per la tabella AB, la significatività del test deve poco a N e molto di più all'effettiva associazione tra le due variabili A e B

Con **dati realmente dicotomici** si può usare il **coefficiente phi** di Pearson (φ, coefficiente di **correlazione punto-tetracorica**): media geometrica delle differenze tra le proporzioni del fattore nelle righe e il fattore nelle colonne; si può calcolare dal χ^2 **non corretto** del test:

```
(AB<-table(A, B))
   B
    b1 b2
A a1 15 35
  a2 30 20
```

$$\varphi = \frac{|(ad) - (bc)|}{\sqrt{r_1 \times r_2 \times c_1 \times c_2}}$$

```
abs(((15*20)-(35*30)))/sqrt(50*50*45*55)
[1] 0.3015113
```

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

```
sqrt(chi_AB$statistic/100)
0.3015113
```

Possiamo usare `Phi(x1, x2)` o `Phi(table(x1,x2)` di `DescTools`

```
Phi(AB)
[1] 0.3015113
```

*Avete notato che coincide con **w**?*

Il suo **range di variazione va da 0 a** $\sqrt{(min(r, c) - 1)}$. In tabelle 2×2 il limite superiore è $= \sqrt{(2-1)} = 1$, ma in **tabelle più grandi è** > 1 (in una tabella 3×3: $\sqrt{(3-1)} = 1.4$).

Quindi, in tabelle *rxc* usiamo il **coefficiente V di Cramér** (1946):

$$V_C = \sqrt{\frac{\varphi}{min(r-1, c-1)}}$$

$$V_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Però, il coefficiente V **approssima ai limiti teorici solo se i marginali di riga e colonna sono uguali** (disegno bilanciato); altrimenti, il **limite inferiore è > 0** e quello **superiore <1.**

In `CramerV(x1, x2 o table)`, con `correct=TRUE` si applica una **correzione** a questo bias (**Bergsma**, 2013), e si può specificare anche il suo CI, con `conf.level= alfa`.

```
CramerV(AB)
[1] 0.3015113
```

```
CramerV(AB, correct=TRUE, conf.level = .95)
 Cramer V    lwr.ci    upr.ci
0.2857143 0.1054848 0.4975080
```

Il **coefficiente Q di Yule** (1900: Q in omaggio a Quetelet) si ricava dall'**OR**: si interpreta in valore assoluto.

$$Q = \frac{(ad) - (bc)}{(ad) + (bc)} = \frac{OR - 1}{OR + 1}$$

Varia da 0 (associazione assente) a 1, **indipendentemente dai marginali (**Warren, 2008).

```
(AB<-table(A, B))
    B
     b1 b2
A a1 15 35
  a2 30 20
```

```
OR<-(15*20)/(35*30)
(OR-1)/(OR+1)
[1] -0.5555556
```

`YuleQ(x1, x2)` o `YuleQ(table)` di `DescTools` riporta solo il coefficiente:

```
YuleQ(AB)
[1] -0.5555556
```

Alcuni dettagli potenzialmente utili…

Desc(table) fornisce il test χ^2 per tabelle di qualsiasi dimensione; per visualizzare la correzione di Yates in tabelle 2×2, va aggiunto **verbose=3**: da **1** a **3**, gli output sono sempre più ricchi. Dà il test di Fisher e il test di McNemar **solo per tabelle 2×2**; per tabelle $r \times c$, oltre al test χ^2, dà test che non fanno parte del nostro programma.

```
Desc(osservate, verbose = 3)
-------------------------------------------------------
osservate (table)
Summary:
n: 126, rows: 2, columns: 2
Pearson's Chi-squared test:
  X-squared = 2.6193, df = 1, p-value = 0.1056
Pearson's Chi-squared test (cont. adj):
  X-squared = 2.0294, df = 1, p-value = 0.1543
Fisher's exact test p-value = 0.1232
McNemar's chi-squared = 13.653, df = 1, p-value = 0.0002199
                estimate lwr.ci upr.ci'
odds ratio         0.536  0.251  1.145
rel. risk (col1)   0.816  0.631  1.055
rel. risk (col2)   1.522  0.914  2.535
rel. risk (row1)   0.709  0.474  1.060
rel. risk (row2)   1.322  0.918  1.903
```

	astinente	fumatore	Sum
freq	32	21	53
perc	25.4%	16.7%	42.1%
p.row	60.4%	39.6%	.
p.col	37.2%	52.5%	.
freq	54	19	73
perc	42.9%	15.1%	57.9%
p.row	74.0%	26.0%	.
p.col	62.8%	47.5%	.
freq	86	40	126
perc	68.3%	31.7%	100.0%
p.row	.	.	.
p.col	.	.	

Ne accenniamo…

Altri coefficienti di associazione e correlazione …

	estimate	lwr.ci	upr.ci'
Phi Coeff.	0.1442	-	-
Contingency Coeff.	0.1427	-	-
Cramer V	0.1442	0.0000	0.3188
Goodman Kruskal Gamma	-0.3020	-0.6469	0.0430
Kendall Tau-b	-0.1442	-0.3196	0.0312
Stuart Tau-c	-0.1325	-0.2943	0.0292
Somers D C|R	-0.1360	-0.3017	0.0298
Somers D R|C	-0.1529	-0.3422	0.0364
Pearson Correlation	-0.1442	-0.3112	0.0315
Spearman Correlation	-0.1442	-0.3112	0.0315
Lambda C|R	0.0000	0.0000	0.0000
Lambda R|C	0.0377	0.0000	0.2672
Lambda sym	0.0215	0.0000	0.1532
Uncertainty Coeff. C|R	0.0165	-0.0235	0.0566
Uncertainty Coeff. R|C	0.0152	-0.0217	0.0520
Uncertainty Coeff. sym	0.0158	-0.0226	0.0542
Mutual Information	0.0149	-	-

Dovremo comunque gestire a parte i r_{adj}

Un'occhiata al **rischio relativo**: negli studi prospettici esprime **l'incidenza**, cioè la **proporzione di nuovi casi tra i soggetti esposti a un fattore di rischio** rispetto a quelli **non esposti**.

		Sviluppo malattia	
		Sì	**No**
Esposizione al rischio	**Sì**	***a***	***b***
	No	***c***	***d***

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

In fumo i dati non sono prospettici, ma giusto per capire il calcolo:

	astinente	fumatore
alta dipendenza	37	28
bassa dipendenza	49	12

```
(37/(37+28))/(49/(49+12))
[1] 0.7086342
```

Il RR è pari a .709 per gli astinenti (colonna 1)

```
(28/(37+28))/(12/(49+12))
[1] 2.189744
```

Il RR è pari a 2.189 per gli astinenti (colonna 1)

Ecco i test di significatività per una **tabella *rxc***, oltre il solito test χ^2 :

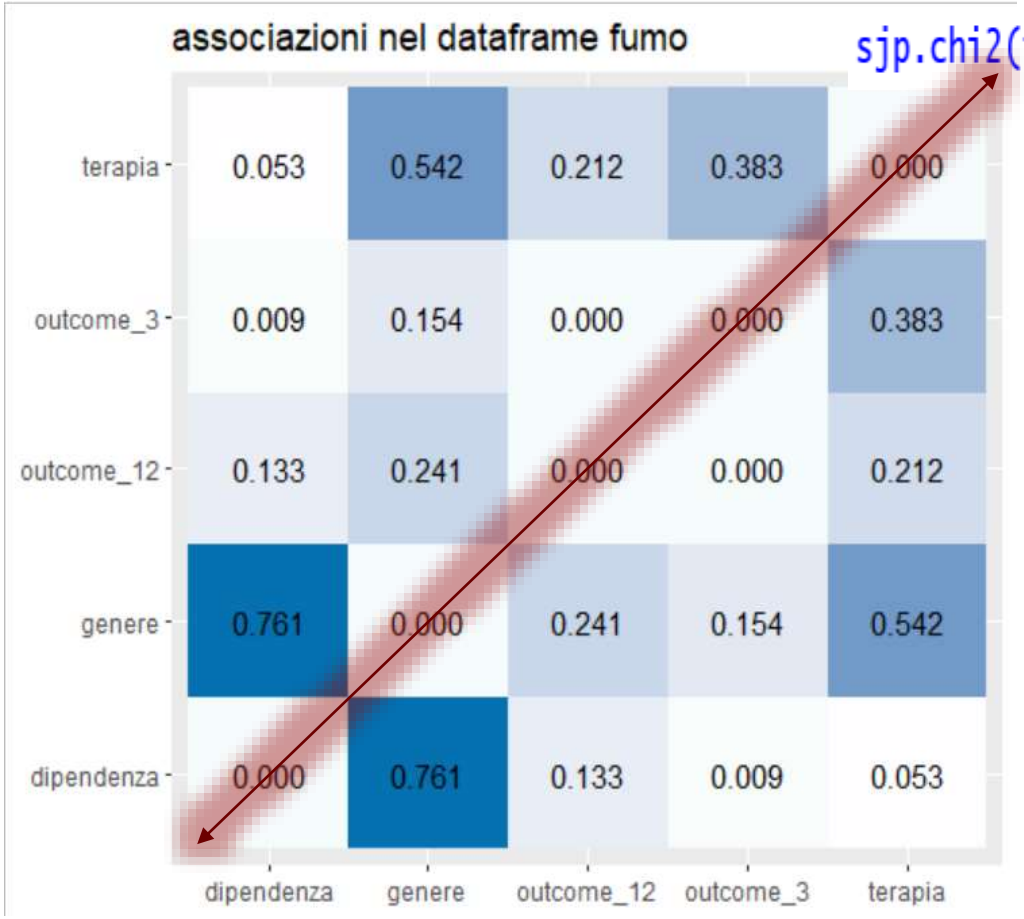
```
Desc(table(fumo$terapia, fumo$outcome_3_mesi))
-----------------------------------------------------
table(fumo$terapia, fumo$outcome_3_mesi) (table)

Summary:
n: 126, rows: 3, columns: 2
Pearson's Chi-squared test:
  X-squared = 1.9202, df = 2, p-value = 0.3829
Log likelihood ratio (G-test) test of independence:
  G = 1.8812, X-squared df = 2, p-value = 0.3904
Mantel-Haenszel Chi-squared:
  X-squared = 1.8471, df = 1, p-value = 0.1741
```

log-likelihood Ratio Test: lo faremo nella regressione logistica

***Generalized Cochran- Mantel - Haenszel test** (Mantel, 1963): stima l'associazione anche in presenza di stratificazioni, confrontando gli OR degli strati. Non fa parte del nostro programma*

Se dovete fare **molti test** χ^2 sugli stessi dati e volete individuare i soli test significativi per poi approfondirli, usate `sjp.chi2(data.frame(`x_1`, `$x_2 \ldots x_k$`)` di `sjPlot`: richiede il **dataframe delle sole variabili da associare**, e il suo **prodotto** è un **grafico: matrice di** $p - value$

```
fm<-data.frame(terapia= fumo$terapia, genere=fumo$genere, dipendenza= fumo$Fagerstrom_categorie,
outcome_3=fumo$outcome_3_mesi, outcome_12=fumo$outcome_12_mesi)
```




```
sjp.chi2(fm,title = "associazioni nel dataframe fumo")
```

$p - value$ *di ogni associazione a coppie,* ***ridondante attorno alla diagonale*** *in cui ogni variabile è associata con se stessa.*

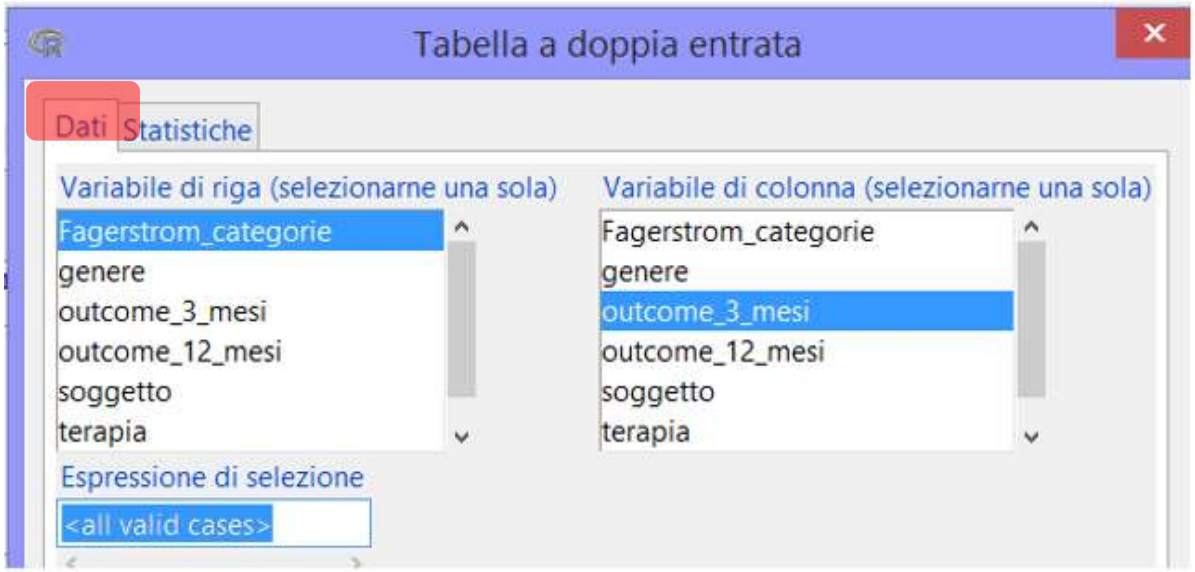
La **scala di colori** indica la graduazione nella lontananza dalla soglia di significatività, dal bianco (significativo) al tono più scuro.

Se volete calcolare i test di associazione con Rcommander:

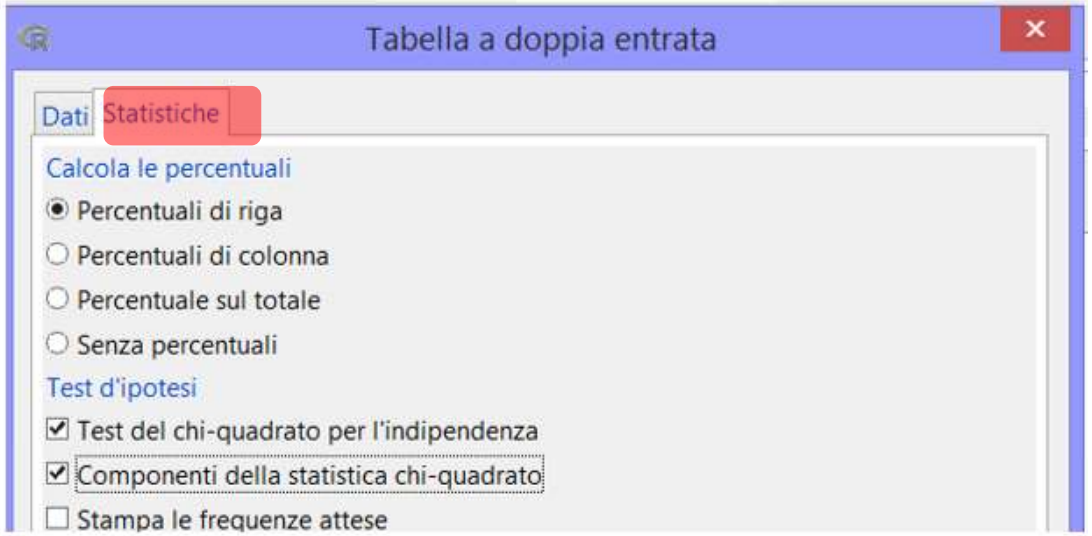
`Statistiche →`

`Tabelle di contingenza`

`→ Tabella a doppia entrata`



Mancano test di McNemar, res_{adj} e coefficienti di intensità dell'associazione.



Verificate la ***significatività*** *e l'intensità dell'associazione tra* ***gravità della dipendenza a T0 ($Fagerstrom_categorie)*** *e* ***l'outcome****, sia a* ***breve*** *sia a* ***lungo*** *termine.*

Recuperate il dataframe *gatti*:

- verificate l'ipotesi che vivere con un gatto faciliti il riconoscimento delle intenzioni comunicative del micio, usando le tre diverse situazioni sperimentali;
- verificate il luogo comune che vede le zitelle come particolarmente amanti dei gatti.