STATISTICA E ANALISI DEI DATI

Capitolo 2 – Grafici di frequenza e contingenza

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

DISTRIBUZIONI DI FREQUENZA

- Per quanto riguarda la **distribuzione di frequenza**, consideriamo una variabile X e indichiamo con $z_1, z_2, \ldots, z_k$ le **modalità distinte** da essa assunte
  - Se la variabile X:
    - è qualitativa le modalità indicano delle qualità distinte degli individui
    - è quantitativa le modalità sono dei numeri reali distinti
- Consideriamo poi un campione $(x_1, x_2, \ldots, x_n)$ costituito da n osservazioni di X
  - Se indichiamo con n_i il numero di volte in cui ciascuna osservazione z_i è presente nel campione, ossia la frequenza assoluta con cui essa appare nel campione, l'insieme $\{(z_i, n_i), i = 1, 2, \ldots, k\}$ si chiama **distribuzione di frequenza**

DISTRIBUZIONI DI FREQUENZA

- Per quanto riguarda la **distribuzione di frequenza**, consideriamo una variabile X e indichiamo con $z_1, z_2, \ldots, z_k$ le **modalità distinte** da essa assunte
  - Se la variabile X:
    - è qualitativa le modalità indicano delle qualità distinte degli individui
    - è quantitativa le modalità sono dei numeri reali distinti
- Consideriamo poi un campione $(x_1, x_2, \ldots, x_n)$ costituito da n osservazioni di X
  - Se indichiamo con n_i il numero di volte in cui ciascuna osservazione z_i è presente nel campione, ossia la frequenza assoluta con cui essa appare nel campione, l'insieme $\{(z_i, n_i), i = 1, 2, \ldots, k\}$ si chiama **distribuzione di frequenza**

4	4	4	2	1	0	1	3	2	3	3	2	1	2	0	1	4	3	1	0	0	4	1	1	2	0	1	1	3	4	4	0	3	3	4	4	0
4	0	1	1	1	2	3	1	3	1	3	0	3	0	3	2	0	2	4	3	1	2	0	2	3	2	3	3	1	1	3	1	1	2	3	0	4
3	4	3	2	4	1	3	3	1	4	3	2	3	2	1	4	4	2	4	4	1	3	2	0	1	3	4	4	3	4	4	3	1	1	4	0	3
4	0	2	3	0	2	4	4	2	0	0	1	3	4	3	3	2	1	1	3	3	3	4	2	3	3	2	2	2	3	4	3	0	4	1	3	2
3	2	3	1	4	4	4	3	2	3	0	0	4	4	0	1	4	0	1	0	0	4	1	3	3	4	2	2	0	4	0	1	1	3	4	3	3
1	2	4	2	4	3	0	0	2	3	2	2	4	3	4	1	3	3	4	3	2	1	1	0	0	2	4	2	1	2	0	1	4	0	0	1	1
3	4	2	0	3	1	0	1	1	0	1	2	4	4	2	3	3	0	3	4	2	2	4	1	0	0	3	0	1	4	2	4	0	4	1	1	1
0	0	4	0	0	0	0	0	0	1	4	0	0	0	1	3	1	4	1	2	1	2	3	3	4	0	4	4	1	3	0	4	3	2	3	3	2
1	1	1	0	1	4	3	4	2	2	3	0	0	2	4	3	4	1	0	1	3	3	0	3	4	1	3	2	4	3	4	3	0	4	3	0	1
2	4	4	0	1	4	3	4	1	1	0	3	2	0	2	1	4	1	2	0	2	1	0	4	2	3	1	3	0	4	0	0	4	0	4	4	2
2	1	0	1	2	1	1	4	0	1	2	2	2	4	1	0	2	0	1	2	3	4	3	4	3	1	1	0	4	1	4	1	1	4	2	1	0

Numero ordini z_i	Frequenze assolute n_i
0	78
1	87
2	68
3	86
4	88
Totale	**407**

Source

DISTRIBUZIONI DI FREQUENZA

- Esempio: La **distribuzione di frequenza** di variabili quantitative continue:
  - Prendiamo adesso in esame una variabile continua come il fatturato mensile di 100 aziende
  - Supponiamo inoltre che tali fatturati siano espressi in migliaia di euro e che sia stati raggruppati in classi
    - Allora, la tabella di frequenza è di questo tipo:

Fatturato mensile $[z_i - z_{i+1} [$	Frequenze assolute n_i
0 - 10	30
10 - 20	15
20 - 40	25
40 - 70	20
70 - 100	10
Totale	**100**

Source

Source

DISTRIBUZIONI DI FREQUENZA

- Esempio: La **distribuzione di frequenza** di variabili quantitative continue:
  - Prendiamo adesso in esame una variabile continua come il fatturato mensile di 100 aziende
  - Supponiamo inoltre che tali fatturati siano espressi in migliaia di euro e che sia stati raggruppati in classi
    - Allora, la tabella di frequenza è di questo tipo:

Fatturato mensile $[z_i - z_{i+1} [$	Frequenze assolute n_i
0 - 10	30
10 - 20	15
20 - 40	25
40 - 70	20
70 - 100	10
Totale	**100**

Source

- Esempio: La **distribuzione di frequenza** di variabili qualitative:
  - Immaginiamo di rilevare il colore degli occhi in una popolazione
  - In questo caso la variabile colore degli occhi assume le modalità marroni, azzurri, neri e verdi e ciascuna di queste compare con una certa frequenza come mostra la tabella qui sotto:

Colore occhi z_i	Frequenze assolute n_i
Marroni	18
Azzurri	45
Neri	89
Verdi	12
Totale	**164**

Source

FREQUENZA RELATIVA

- Se non esistono dati mancanti, la somma delle frequenze assolute è sempre uguale alla numerosità del campione, ossia $n = n_1 + n_2 + \ldots + n_k$
- La **frequenza relativa** di un evento è il rapporto tra la frequenza assoluta di quell'evento e il numero totale di osservazioni o prove
  - Viene espressa come una frazione o una percentuale, e indica la proporzione con cui un determinato evento si verifica rispetto al totale

$$Frequenza\ relativa = \frac{Numero\ totale\ di\ osservazioni}{Frequenza\ assoluta\ dell'evento}$$

- O in altri termini:

$$f_i = \frac{n_i}{n} \qquad (i = 1,2, \ldots, k)$$

- Senza dati mancanti la somma delle frequenze relative è sempre unitaria, ossia $f_1 + f_2 + \ldots + f_k = 1$

DISTRIBUZIONI DI FREQUENZA

- Per una variabile qualitativa, in R la costruzione di una distribuzione di frequenza viene effettuata utilizzando la funzione `table()`
- La funzione table() in R crea una **tabella di contingenza** che mostra la frequenza (**assoluta**) degli elementi unici di un vettore o la combinazione di livelli di più variabili
  - **Definizione**: Una **tabella di contingenza** è una tabella che riassume la distribuzione congiunta di due o più variabili categoriali, mostrando le frequenze (assolute o relative) con cui si verificano le combinazioni dei livelli delle variabili

Colore occhi z_i	Frequenze assolute n_i
Marroni	18
Azzurri	45
Neri	89
Verdi	12
Totale	**164**

Source

DISTRIBUZIONI DI FREQUENZA

- Ad esempio, consideriamo un campione di 36 elementi che costituiscono delle osservazioni di una variabile qualitativa *people* con quattro modalità (bambino, giovane, adulto, anziano):

```
> people <- c(rep("bambino",8),rep("giovane",3),rep("adulto",4), rep("giovane",9),rep("anziano",6),
rep("bambino",2),rep("adulto",4))
> people
 [1] "bambino" "bambino" "bambino" "bambino" "bambino" "bambino" "bambino" "bambino"
 [9] "giovane" "giovane" "giovane" "adulto"  "adulto"  "adulto"  "adulto"  "giovane"
[17] "giovane" "giovane" "giovane" "giovane" "giovane" "giovane" "giovane" "giovane"
[25] "anziano" "anziano" "anziano" "anziano" "anziano" "anziano" "bambino" "bambino"
[33] "adulto"  "adulto"  "adulto"  "adulto"
> table(people)
people
 adulto anziano bambino giovane
      8       6      10      12
```

- Il comando `rep()` è usato per ripetere più volte le varie modalità nel vettore
- La funzione `table()` riporterà le modalità della variabile qualitativa che presentano un valore di frequenza assoluta diverso da zero.
  - Si nota che `table(people)` ordina le modalità della variabile qualitativa uomo in ordine alfabetico.

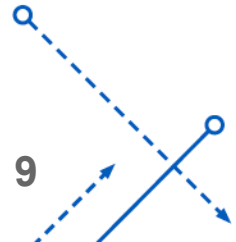
DISTRIBUZIONI DI FREQUENZA

- Se si vuole ottenere la **distribuzione delle frequenze relative** basta utilizzare il comando:

```
> table(people)/length(people)   #calcola le frequenze relative
people
   adulto   anziano   bambino   giovane
0.2222222 0.1666667 0.2777778 0.3333333
```

- Se si prende in esame una variabile ordinabile, bisogna considerarla come un **fattore** e ordinare i livelli di questo fattore prima di costruire la distribuzione di frequenza
  - Riferendosi all'esempio precedente si ha:

```
> people_ordered <- ordered(people, levels=c("bambino","giovane","adulto","anziano"))
> table(people_ordered)
people_ordered
bambino giovane  adulto anziano
     10      12       8       6
```



FREQUENZA ASSOLUTA CUMULATA

- La **frequenza assoluta cumulata** rappresenta il numero totale di osservazioni che hanno un valore inferiore o uguale a un determinato valore in un insieme di dati
  - In altri termini: È la somma delle frequenze assolute fino a quel punto nella distribuzione.
  - È definita come segue:

$$N_i = n_1 + n_2 + \cdots + n_i \qquad (i = 1,2, \ldots, k)$$

- La frequenza assoluta cumulata fornisce un'idea di quante osservazioni si sono accumulate fino a un certo valore
  - Supponiamo di considerare un campione di osservazioni con le seguenti frequenze assolute

Valore	Frequenza Assoluta
1	2
2	5
3	8
4	10

Valore	Frequenza Assoluta Cumulata
1	2
2	2 + 5 = 7
3	7 + 8 = 15
4	15 + 10 = 23

FREQUENZA RELATIVA CUMULATA

- La **frequenza relativa cumulata** è la proporzione (o percentuale) di osservazioni che hanno un valore inferiore o uguale a un determinato valore
  - In altri termini: È la somma delle frequenze relative fino a quel punto nella distribuzione
  - È definita come segue:

$$F_i = f_1 + f_2 + \cdots + f_i \qquad (i = 1,2, \ldots, k)$$

- Continuando l'esempio precedente se il totale delle osservazioni è 25, la frequenza relativa cumulata sarà:

Valore	Frequenza Assoluta
1	2
2	5
3	8
4	10

Valore	Frequenza Relativa Cumulata
1	$\frac{2}{25} = 0.08$
2	$\frac{7}{25} = 0.28$
3	$\frac{15}{25} = 0.60$
4	$\frac{25}{25} = 1.00$

DISTRIBUZIONI DI FREQUENZA

- Per calcolare le frequenze cumulate si deve utilizzare la funzione `cumsum()` che permette di calcolare le somme cumulate degli elementi di un vettore. Riferendoci all'esempio precedente, calcoliamo queste frequenze:

```
> cumsum(table(people))
 adulto anziano bambino giovane
      8       14       24       36
> cumsum(table(people_ordered))
bambino giovane  adulto anziano
     10       22       30       36
```

Frequenze Assolute Cumulate

```
> cumsum(table(people_ordered)/length(people_ordered))
  bambino   giovane    adulto   anziano
0.2777778 0.6111111 0.8333333 1.0000000
> cumsum(table(people)/length(people))
   adulto   anziano   bambino   giovane
0.2222222 0.3888889 0.6666667 1.0000000
```

Frequenze Relative Cumulate

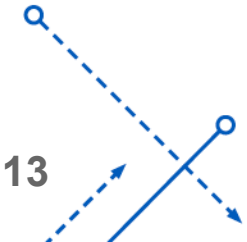
FREQUENZE DI INTERVALLI

- Le frequenze assolute e relative possono essere calcolate anche per variabili quantitative sempre che il numero di modalità distinte sia ben definito (es. i voti dal 18 al 30)
  - Spesso si preferisce **raccogliere le informazioni in classi** e calcolare le frequenze assolute o relative per classi, ossia le frequenze con cui gli elementi del vettore cadono nelle diverse classi.
  - `cut()` raggruppa i dati relativi ad un vettore in intervalli elencando nel parametro `breaks` gli estremi degli intervalli (aperti a sinistra e chiusi a destra)
    - Se desideriamo ottenere intervalli chiusi a sinistra e aperti a destra occorre specificare in `cut()` l'opzione `right = FALSE`
  - Calcoliamo la frequenza assoluta con cui un assegnato insieme di voti è suddiviso nelle quattro classi:

```
> # Creazione del vettore con i voti degli studenti
> votiStudenti <- c(rep(18,6), rep(21,8), rep(25,10), rep(28,6),
+                   rep(30,7), rep(25,2), rep(24,6), rep(25,4))
> votiStudenti
> sort(votiStudenti)
 [1] 18 18 18 18 18 18 21 21 21 21 21 21 21 21 24 24 24 24 24 24 25 25 25 25 25 25 25 25 25 25 25 25 25 25 28 28 28 28 28 28 30 30 30 30 30 30 30
```

```
> table(cut(votiStudenti, breaks = c(17, 21, 24, 27, 31)))

(17,21] (21,24] (24,27] (27,31]
     14       6      16      13
```



FREQUENZE DI INTERVALLI

- Le frequenze assolute e relative possono essere calcolate anche per variabili quantitative sempre che il numero di modalità distinte sia ben definito (es. i voti dal 18 al 30)
  - Spesso si preferisce **raccogliere le informazioni in classi** e calcolare le frequenze assolute o relative per classi, ossia le frequenze con cui gli elementi del vettore cadono nelle diverse classi.
  - `cut()` raggruppa i dati relativi ad un vettore in intervalli elencando nel parametro `breaks` gli estremi degli intervalli (aperti a sinistra e chiusi a destra)
    - Se desideriamo ottenere intervalli chiusi a sinistra e aperti a destra occorre specificare in `cut()` l'opzione `right = FALSE`
  - Calcoliamo la frequenza assoluta con cui un assegnato insieme di voti è suddiviso nelle quattro classi:

```
> # Creazione del vettore con i voti degli studenti
> votiStudenti <- c(rep(18,6), rep(21,8), rep(25,10), rep(28,6),
+                   rep(30,7), rep(25,2), rep(24,6), rep(25,4))
> votiStudenti
> sort(votiStudenti)
 [1] 18 18 18 18 18 18 21 21 21 21 21 21 21 21 24 24 24 24 24 24 25 25 25 25 25 25 25 25 25 25 25 25 25 25 28 28 28 28 28 28 30 30 30 30 30 30 30

> sort(votiStudenti)
 [1] 18 18 18 18 18 18 21 21 21 21 21 21 21 21 24 24 24 24 24 24 25 25 25 25 25 25 25 25 25 25 25 25 25 25 28 28 28 28 28 28 30 30 30 30 30 30 30
```

```
> table(cut(votiStudenti, breaks = c(17, 21, 24, 27, 31)))

(17,21] (21,24] (24,27] (27,31]
     14       6      16      13
```

```
> table(cut(votiStudenti, breaks = c(17, 21, 24, 27, 31), right=FALSE))

[17,21) [21,24) [24,27) [27,31)
      6       8      22      13
```

FREQUENZE NEI DATI

- **Gestione dei dati mancanti:** Quando ci sono valori mancanti, si può usare la frequenza assoluta delle categorie per decidere come imputare i valori mancanti
  - Ad esempio, si può sostituire i valori mancanti con la categoria più frequente (moda) nei dati categoriali
- **Ribilanciamento dei dataset sbilanciati:** Nei problemi di classificazione binaria o multiclass, le frequenze relative delle classi vengono usate per identificare le classi sbilanciate
  - Applicando tecniche come il downsampling (ridurre il numero di campioni della classe maggioritaria) o upsampling (replicare i campioni della classe minoritaria)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

FREQUENZE NEI DATI

- **Encoding di variabili categoriali**: Le frequenze relative delle categorie possono essere utilizzate per trasformare variabili categoriali in valori numerici, un processo chiamato target encoding
  - Questo è particolarmente utile per variabili con molte categorie, dove il semplice one-hot encoding potrebbe risultare inefficiente.

ID	Quartiere	Prezzo
1	A	300,000
2	B	450,000
3	A	350,000
4	C	600,000
5	B	500,000
6	C	550,000

ID	Quartiere (encoded)	Prezzo
1	325,000	300,000
2	475,000	450,000
3	325,000	350,000
4	575,000	600,000
5	475,000	500,000
6	575,000	550,000

FREQUENZE NEI DATI

```
> encoding <- df %>% group_by(Quartiere) %>% summarise(media_prezzo = mean(Prezzo))
>
> encoding
# A tibble: 3 × 2
  Quartiere media_prezzo
  <chr>            <dbl>
1 A               325000
2 B               475000
3 C               575000
> df_encoded <- df %>% left_join(encoding, by = "Quartiere")
>
> df_encoded
  ID Quartiere Prezzo media_prezzo
1  1         A 300000       325000
2  2         B 450000       475000
3  3         A 350000       325000
4  4         C 600000       575000
5  5         B 500000       475000
6  6         C 550000       575000
```

- group_by(Quartiere)
  - Raggruppa i dati per la colonna Quartiere
  - Crea gruppi separati per ogni valore unico nel quartiere

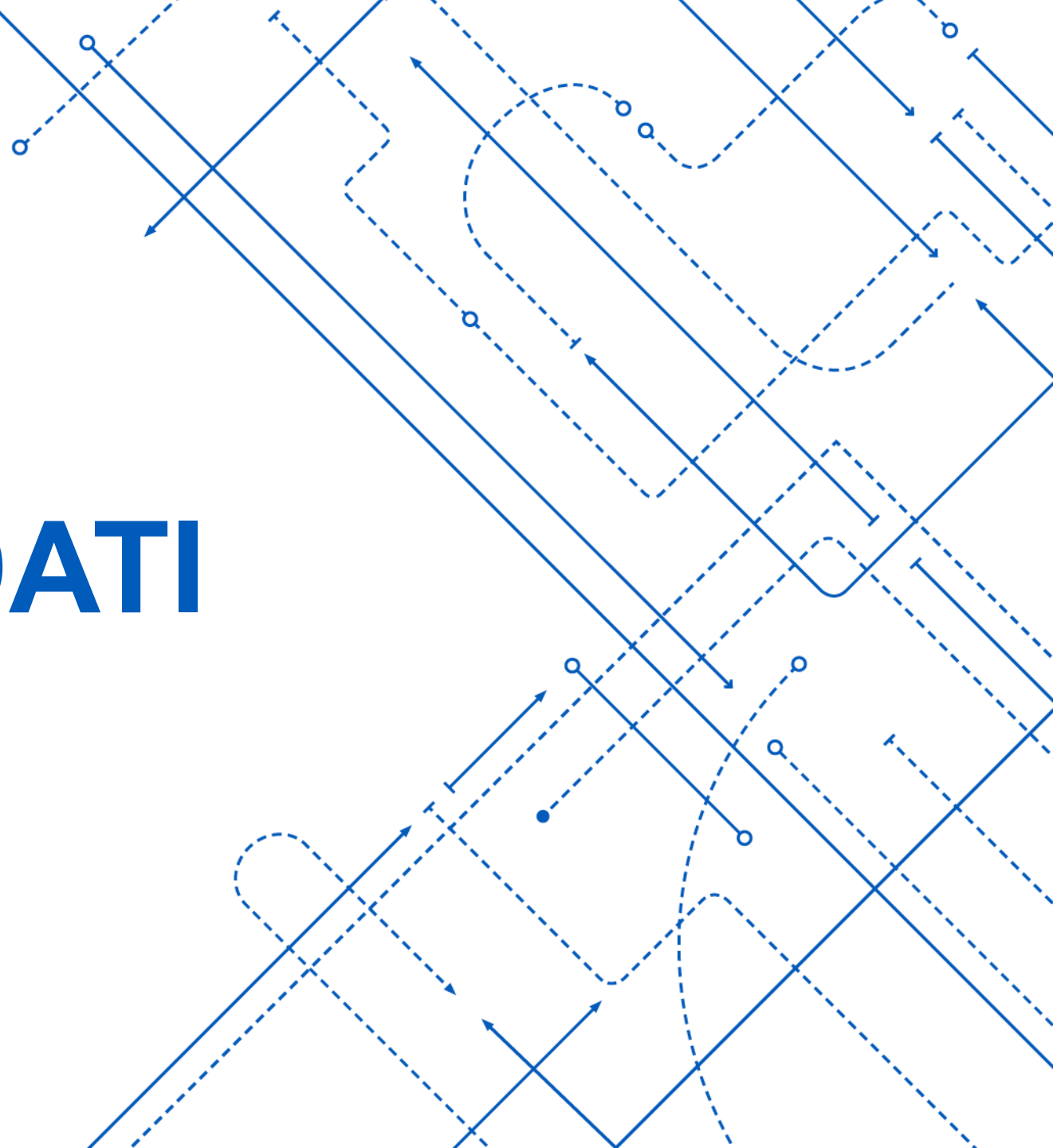
- summarise(media_prezzo = mean(Prezzo))
  - summarise(): Crea un nuovo dataframe riassuntivo
  - media_prezzo = mean(Prezzo): Calcola la media della colonna Prezzo per ogni gruppo
  - Crea una nuova colonna chiamata media_prezzo

STATISTICA E ANALISI DEI DATI

Capitolo 2 – Grafici di Frequenza

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

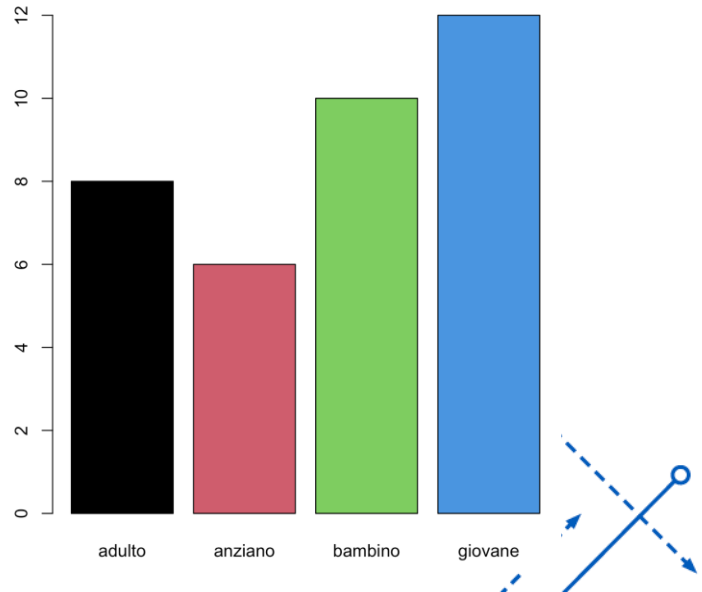


BARPLOT PER LE FREQUENZE

- Consideriamo una variabile qualitativa X e indichiamo con $z_1, z_2, \ldots, z_k$ le modalità distinte da essa assunte. Consideriamo poi un campione $x = (x_1, x_2, \ldots, x_n)$ costituito da n osservazioni di X
  - Disponiamo sull'asse orizzontale ed in modo equispaziato le modalità assunte da X e sull'asse verticale riportiamo le frequenze assolute o le frequenze relative
  - Tracciamo dei rettangoli (barre) centrati sulle modalità z_i tutti della stessa base e altezza pari alle frequenze (assolute o relative), ottenendo un grafico (o diagramma) a barre
  - Grafici di questo tipo sono di solito utilizzati per visualizzare i valori di una qualche quantità (ad esempio, la frequenza) per diverse modalità o categorie. In R si ottiene un grafico a barre utilizzando `barplot(table(x))`

```
> uomo<-c(rep("bambino",8),rep("giovane",3),rep("adulto",4),
+ rep("giovane",9),rep("anziano",6),rep("bambino",2),rep("adulto"
    ,4))
>
> barplot(table(uomo),col=1:4)
```

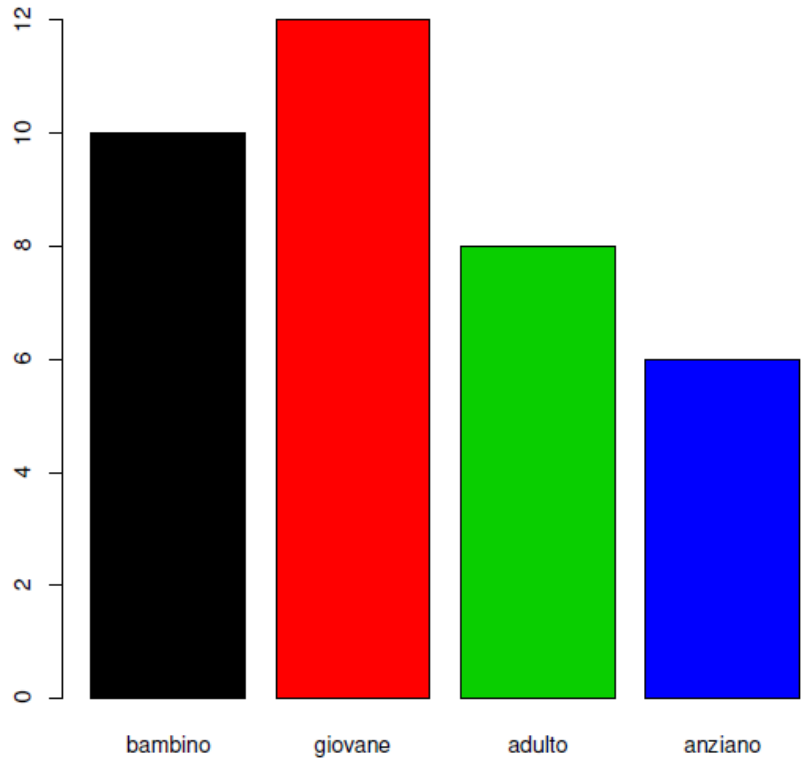
- Il parametro col = 1:4 permette di colorare i rettangoli in modo differente



BARPLOT PER LE FREQUENZE

- Da notare che le modalità della variabile uomo sono ordinate alfabeticamente.
  - Affinché le modalità siano ordinate in modo differente occorre trasformare il vettore uomo in un fattore uomo1.

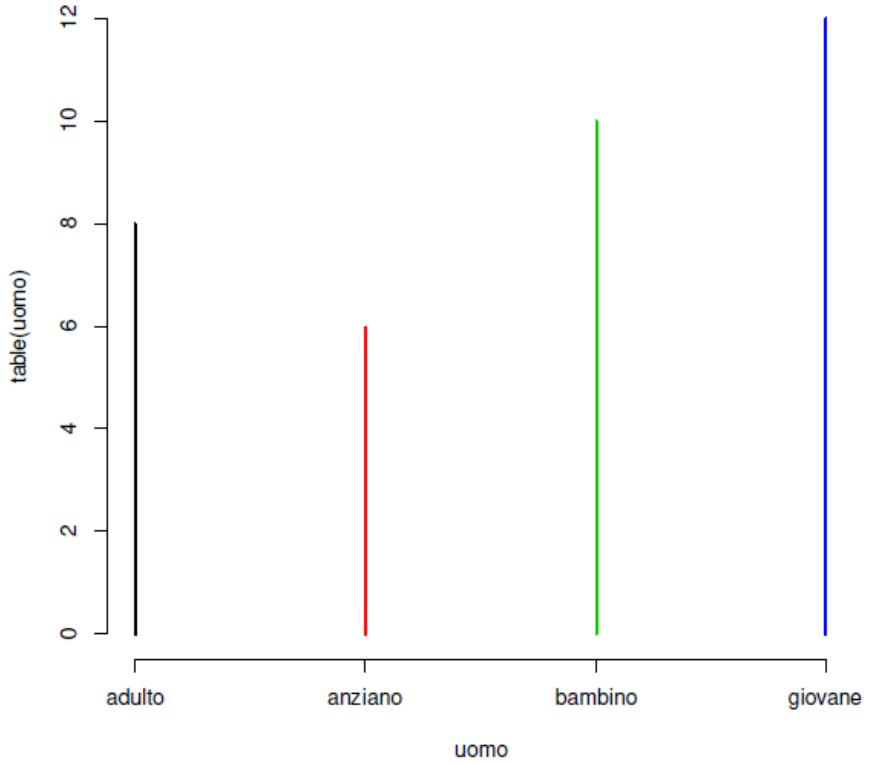
```
> uomo1<-ordered(uomo,levels=c("bambino","giovane","adulto","
    anziano"))
>
> barplot(table(uomo1),col=1:4)
```



BARPLOT PER LE FREQUENZE

- Per la variabile qualitativa *X* è possibile anche costruire un grafico a bastoncini utilizzando il comando `plot(table(x))`:

```
> plot(table(uomo),col=1:4)
```



- Come prima, possiamo ordinare le modalità:

```
> uomo1<-ordered(uomo,levels=c("bambino","giovane","adulto","
    anziano"))
> plot(table(uomo1),col=1:4)
```

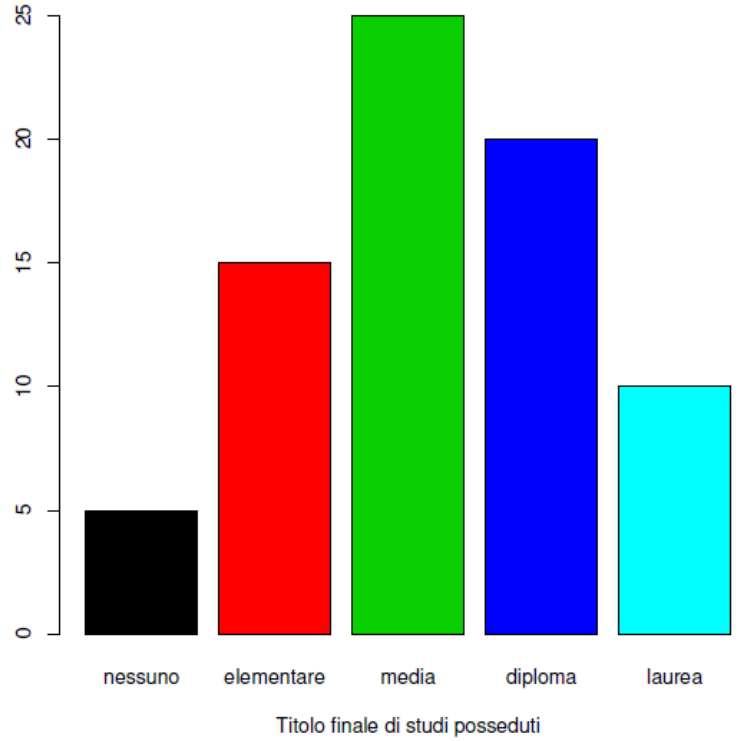
BARPLOT PER LE FREQUENZE

- Se si desidera un grafico a bastoncini che riporti le frequenze relative si può utilizzare il comando `plot(table(x)/length(x))`

```
> uomo1<-ordered(uomo,levels=c("bambino","giovane","adulto","
    anziano"))
>
> plot(table(uomo1)/length(uomo1),col=1:4)
```

- Si possono anche realizzare grafici a barre per variabili qualitative ordinabili utilizzando la funzione `plot()` invece della funzione `barplot()`. Ad esempio, se si considera il titolo finale di studio posseduto da un campione di 75 persone:

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))
>
> titoloFinale<-ordered(titolo,levels=c("nessuno","elementare",
+ "media","diploma","laurea"))
>
> plot(titoloFinale,xlab="Titolo finale di studi posseduti",col
    =1:5)
```



PIE CHART

- Un altro tipo di rappresentazione si ottiene mediante i **diagrammi a torta** che permettono di attribuire ciascuna modalità della variabile qualitativa in esame ad un settore circolare di un cerchio, la cui ampiezza è proporzionale alle frequenze
- I diagrammi a torta sono **quindi utili quando i dati non sono numerici ma categorici**.
- Un grafico a torta si costruisce tracciando un cerchio e suddividendolo in tanti settori circolari (fette o spicchi) quante sono le modalità distinte di dati
  - ogni settore ha un angolo al centro proporzionale alla frequenza (relativa o assoluta) della modalità corrispondente.
- Il sistema R sceglie il tipo di diagramma a torta con i diversi settori colorati differentemente utilizzando il comando pie(table(x)). Riferendoci all'esempio precedente:

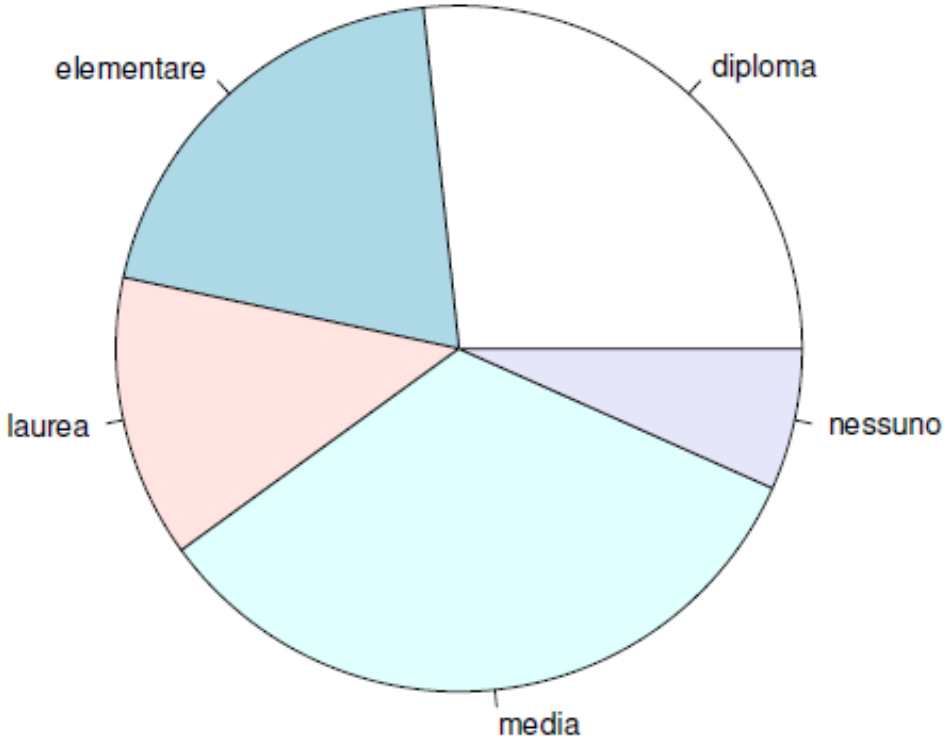
```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))
> pie(table(titolo))
```

PIE CHART

- Otteniamo quindi un diagramma suddiviso in 5 settori

- Si può anche scegliere un tratteggio particolare da utilizzare nei diagrammi a torta per tratteggiare differentemente i diversi settori utilizzando i comandi `density =` e `angle =`.

```
> pie(table(titolo),density=10,angle=15+10*(1:5), col=1:5)
```



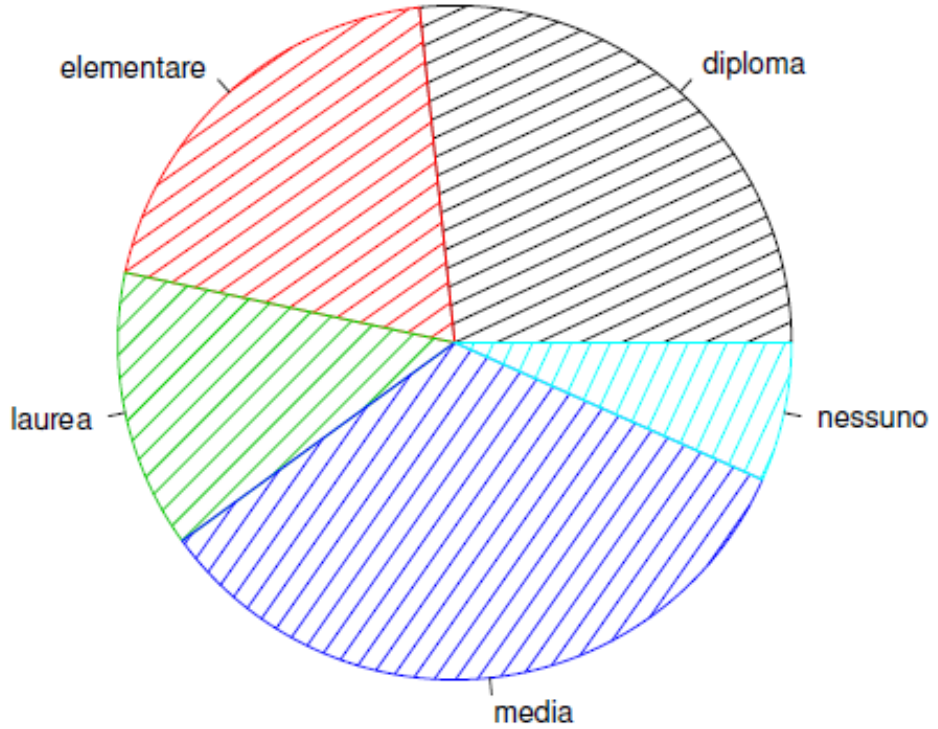
PIE CHART

- Otteniamo quindi un diagramma suddiviso in 5 settori

- Si può anche scegliere un tratteggio particolare da utilizzare nei diagrammi a torta per tratteggiare differentemente i diversi settori utilizzando i comandi `density =` e `angle =`.

```
> pie(table(titolo),density=10,angle=15+10*(1:5), col=1:5)
```

- Si nota che i cinque settori sono stati tratteggiati differentemente utilizzando delle linee ugualmente distanziate e inclinate opportunamente con angolazioni di 25, 35, 45, 55, 65 gradi.



ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

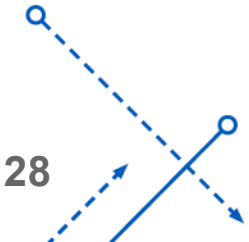
ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

- Usiamo `round()` per arrotondare i numeri



ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

- Usiamo `round()` per arrotondare i numeri
- `label()` definisce le etichette

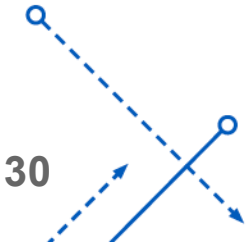
ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

- Usiamo `round()` per arrotondare i numeri
- `label()` definisce le etichette
- `paste()` serve a concatenare stringhe



ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

- Usiamo `round()` per arrotondare i numeri
- `label()` definisce le etichette
- `paste()` serve a concatenare stringhe
- `sep=` è il separatore tra le stringhe

ESEMPIO

- Costruiamo ora un diagramma a torta per il titolo di studio dal campione di 75 individui utilizzando i valori percentuali.

```
> titolo<-c(rep("nessuno",5),rep("elementare",15),rep("media",25),
+ rep("diploma",20),rep("laurea",10))

> freqRel<-table(titolo)/length(titolo)
> freqRel # visualizza le frequenze relative
titolo
   diploma elementare     laurea      media    nessuno
0.26666667 0.20000000 0.13333333 0.33333333 0.06666667
>
> percentuali<-freqRel*100
> percentuali # visualizza le percentuali
titolo
   diploma elementare     laurea      media    nessuno
 26.666667  20.000000  13.333333  33.333333   6.666667
> sum(percentuali)
[1] 100
>
> perc<-round(percentuali)
> labelP<-c("diploma","elementare","laurea","media","nessuno")
> labelP<-paste(labelP,perc)
> labelP<-paste(labelP,"%",sep="")
> pie(percentuali,label=labelP, col=rainbow(length(labelP)),
+ main="Valori percentuali")
```

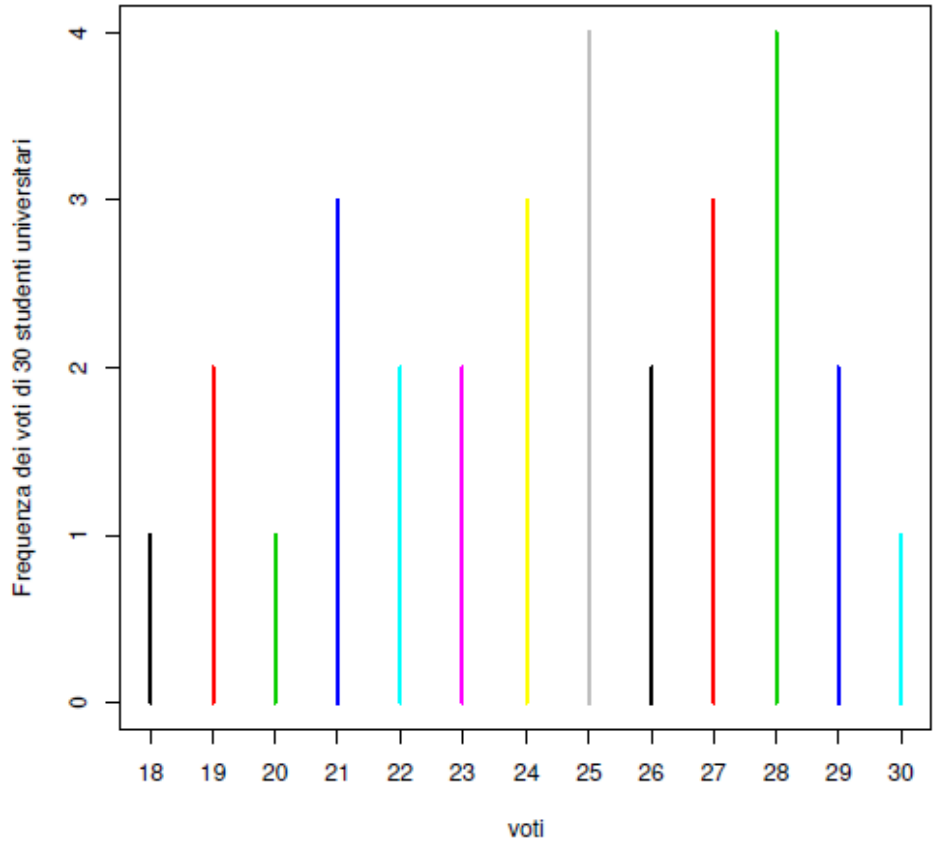
- Usiamo `round()` per arrotondare i numeri
- `label()` definisce le etichette
- `paste()` serve a concatenare stringhe
- `sep=` è il separatore tra le stringhe
- `rainbow()` è un modo alternativo per selezionare i colori

GRAFICI DI FREQUENZA E CONTINGENZA

- Per rappresentare invece correttamente la distribuzione di frequenza di una variabile quantitativa *X* occorre utilizzare il comando `plot(table(x))`.

- Esso produce un grafico a bastoncini in cui sull'asse orizzontale sono riportati i valori e sull'asse verticale le frequenze assolute dei valori distinti assunti nel vettore. Riferendosi ai voti di 30 studenti universitari si ha:

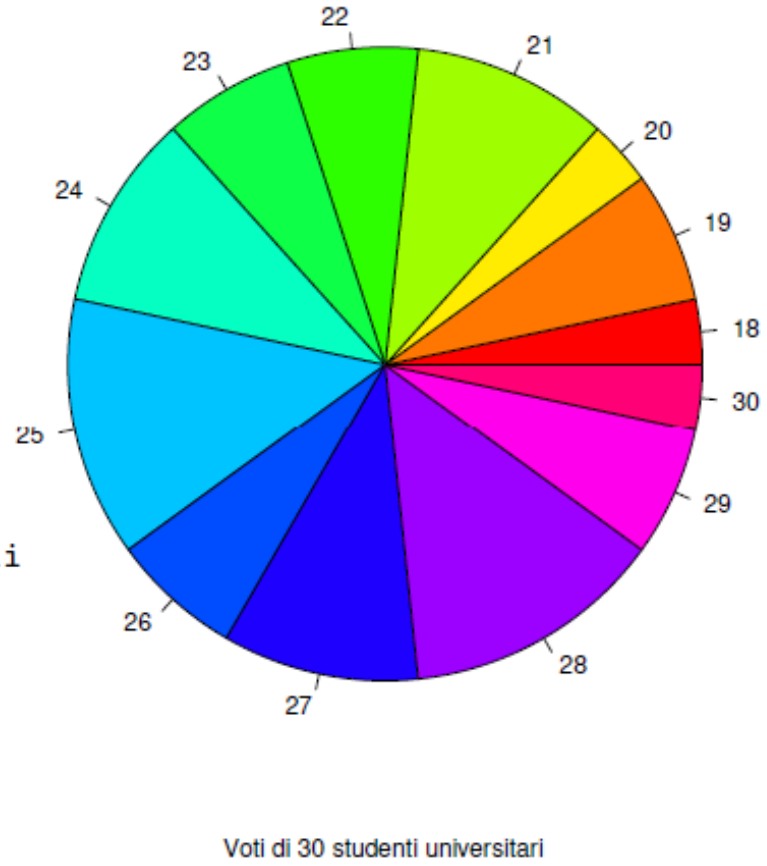
```
> voti<-c(18,19,20,30,29,28,21,22,23,27,26,25,24,25,
+ 26,24,23,22,27,28,21,24,25,25,27,19,21,28,29,28)
> plot(table(voti),ylab="Frequenza dei voti di 30 studenti
    universitari", col=1:13)
```



GRAFICI DI FREQUENZA E CONTINGENZA

- Per variabili quantitative si può anche considerare una rappresentazione tramite diagrammi a torta attraverso il comando `pie(table(x))`, dove *x* è un vettore o un fattore, anche se tale grafico non si rivela spesso utile.

```
> pie(table(voti), col = rainbow(13), radius = 0.9, xlab="Voti di
  30 studenti universitari")
```



STATISTICA E ANALISI DEI DATI

Capitolo 2 – Tabelle di contingenza

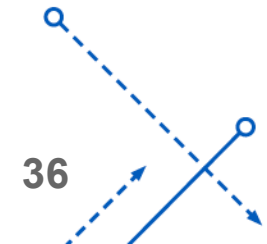
Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

Breve Recap

- **Variabili (osservabili)**
  - Variabili quantitative
  - Variabili qualitative
- **Modalità (distinte)**
  - Frequenza assoluta
  - Frequenza relativa
  - frequenza assoluta cumulata
  - frequenza relativa cumulata

- **Distribuzione di frequenza**
  - distribuzione di frequenza relativa
  - distribuzione di frequenza assoluta



Breve Recap

- **Frequenza congiunta**
  - Tabelle di contingenza (o tabella a due vie, o cross-tab)
  - Distribuzione di frequenza marginale
  - Frequenze relative congiunte
  - Frequenze relative marginali
  - Grafici per tabelle di contingenza

Frequenza Congiunta

Consideriamo una popolazione composta da «persone» di sesso diverso e un insieme di «bevande distinte».

Ipotizziamo di «contare» quante volte le persone di sesso X bevono caffe oppure tè!

In questo modo, possiamo rispondere a domande del tipo «Quante persone stanno insieme in due categorie specifiche?»

FREQUENZA CONGIUNTA

La definizione di frequenza congiunta ci permette di rispondere a domande come la precedente.

Definizioni (informali):

- la **frequenza congiunta assoluta** è il numero di individui che rientrano contemporaneamente in due categorie (es. donne e caffè).
- La **frequenza congiunta relativa** è la stessa quantità divisa per il totale degli individui (una percentuale).

TABELLA DI CONTINGENZA

Una *tabella di contingenza* **riassume la distribuzione congiunta di due o più variabili <u>categoriali</u>**, mostrando le frequenze (assolute o relative) con cui si **verificano le combinazioni dei livelli delle variabili**.

- Ci permette di «valutare» eventuali legami tra due variabili X e Y rilevate congiuntamente **<u>sugli stessi individui</u>**;
- Queste tabelle riassumono la distribuzione congiunta e permettono di valutare l'associazione tra le categorie

TABELLA DI CONTINGENZA

In una tabella di contingenza**, ogni cella contiene una frequenza congiunta.**

I totali di riga e di colonna sono le **frequenze marginali**.

	Tè	Caffè	Totale
Donne	30	20	50
Uomini	25	25	50
Totale	55	45	100

TABELLA DI CONTINGENZA

Domande per voi:

- Perché solo variabili **categoriali**?

- Perché è necessario che le variabili siano rilevate congiuntamente **sugli stessi individui?**

- Per le variabili **quantitative** che strumenti usiamo?

TABELLE DI CONTINGENZA

Definizioni:

- Sia X una variabile e indichiamo con $x_1, x_2, \ldots, x_h$ le modalità distinte da essa assunte e sia Y un'altra variabile e indichiamo con $y_1, y_2, \ldots, y_k$ le modalità distinte da essa assunte
- Consideriamo un campione $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ costituito da n osservazioni di (X, Y)
  - L'elemento n_{ij} in posizione (i, j) della tabella di contingenza, detto **frequenza congiunta**, **rappresenta il numero di volte in cui una particolare coppia di modalità** (x_i, yj) si presenta nel campione.

TABELLE DI CONTINGENZA

Definizioni:

- Dalla tabella è possibile ottenere la **distribuzione di frequenza marginale di X**:

$$n_{\cdot j} = \sum_{i=1}^{h} n_{i,j} \qquad (j = 1, 2, \ldots, k)$$

- Mentre la **distribuzione di frequenza marginale di Y è**:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{i,j} \qquad (i = 1, 2, \ldots, h)$$

TABELLE DI CONTINGENZA

X \ Y	w_1	w_2	$\dots$	w_j	$\dots$	w_k	
z_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1k}	$n_{1.}$
z_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2k}	$n_{2.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
z_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ik}	$n_{i.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
z_h	n_{h1}	n_{h2}	$\dots$	n_{hj}	$\dots$	n_{hk}	$n_{h.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.j}$	$\dots$	$n_{.k}$	n

FREQUENZE RELATIVE CONGIUNTE

- È anche possibile considerare **le frequenze relative** associate ad una tabella di contingenza.

- Parliamo in questo caso di **frequenze relative congiunte che misurano la proporzione di osservazioni** che appartengono simultaneamente a specifiche categorie di due variabili

FREQUENZE RELATIVE CONGIUNTE

- Le **frequenze relative congiunte** sono definite come:

$$f_{ij} = \frac{n_{ij}}{n} \qquad (i = 1, 2, \ldots, h; j = 1, 2, \ldots, k)$$

In pratic, normalizziamo il numero di osservazioni in ogni cella della tabella rispetto al totale delle osservazioni.

(E secondo voi, cosa diventa?)

Frequenze Relative Congiunte

X \ Y	w_1	w_2	$\dots$	w_j	$\dots$	w_k	
z_1	f_{11}	f_{12}	$\dots$	f_{1j}	$\dots$	f_{1k}	$f_{1.}$
z_2	f_{21}	f_{22}	$\dots$	f_{2j}	$\dots$	f_{2k}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
z_i	f_{i1}	f_{i2}	$\dots$	f_{ij}	$\dots$	f_{ik}	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
z_h	f_{h1}	f_{h2}	$\dots$	f_{hj}	$\dots$	f_{hk}	$f_{h.}$
	$f_{.1}$	$f_{.2}$	$\dots$	$f_{.j}$	$\dots$	$f_{.k}$	1

$$f_{ij} = \frac{n_{ij}}{n} \qquad (i = 1, 2, \dots, h; j = 1, 2, \dots, k)$$

$$f_{i,\cdot} = \sum_{j=1}^{k} f_{ij} = \sum_{j=1}^{k} \frac{n_{ij}}{n} = \frac{1}{n} \sum_{j=1}^{k} n_{ij} = \frac{n_{i\cdot}}{n}$$

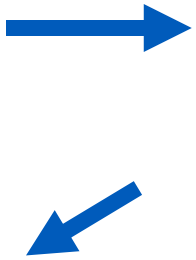
$$f_{\cdot,j} = \sum_{i=1}^{h} f_{ij} = \sum_{i=1}^{h} \frac{n_{ij}}{n} = \frac{1}{n} \sum_{i=1}^{h} n_{ij} = \frac{n_{\cdot j}}{n}$$

FREQUENZE RELATIVE CONGIUNTE

- Consideriamo una tabella di contingenza che riassume la relazione tra due variabili categoriali: Età (Giovane, Adulto, Anziano) e Abitudine al Fumo (Fumatore, Non Fumatore)
  - Le frequenze relative congiunte rappresentano la proporzione di persone in ciascuna combinazione di categorie (Età e Abitudine al Fumo) rispetto al totale delle osservazioni

$$P(\text{Età, Abitudine al Fumo}) = \frac{\text{Frequenza Assoluta}}{\text{Totale}}$$

	Fumatore	Non Fumatore	Totale
Giovane	15	35	**50**
Adulto	30	20	**50**
Anziano	5	20	**25**
Totale	**50**	**75**	**125**



	Fumatore	Non Fumatore	Totale
Giovane	$\frac{15}{125} = 0.12$	$\frac{35}{125} = 0.28$	**0.40**
Adulto	$\frac{30}{125} = 0.24$	$\frac{20}{125} = 0.16$	**0.40**
Anziano	$\frac{5}{125} = 0.04$	$\frac{20}{125} = 0.12$	**0.20**
Totale	**0.40**	**0.60**	**1**

In questo esempio:

- **Giovani fumatori** costituiscono il 12% della popolazione totale;
- **Giovani non fumatori** sono il 28% della popolazione totale
- **Adulti fumatori** sono il 24%, mentre **adulti non fumatori** sono il 16%, e così via

FREQUENZE RELATIVE MARGINALI

- Le **frequenze relative marginali** rappresentano la distribuzione di una singola variabile senza considerare l'altra variabile
  - Vengono calcolate come la somma delle frequenze relative congiunte per ciascuna categoria di una variabile, indipendentemente dalle categorie dell'altra variabile
- Le **frequenze relative marginali di *X*** sono invece indicate come:

$$f_{i,\cdot} = \sum_{j=1}^{k} f_{ij} = \sum_{j=1}^{k} \frac{n_{ij}}{n} = \frac{1}{n} \sum_{j=1}^{k} n_{ij} = \frac{n_{i\cdot}}{n}$$

- mentre quelle di *Y* sono definite come:

$$f_{\cdot,j} = \sum_{i=1}^{h} f_{ij} = \sum_{i=1}^{h} \frac{n_{ij}}{n} = \frac{1}{n} \sum_{i=1}^{h} n_{ij} = \frac{n_{\cdot j}}{n}$$

TABELLA DI CONTINGENZA

X \ Y	w_1	w_2	$\dots$	w_j	$\dots$	w_k	
z_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1k}	$n_{1\cdot}$
z_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
z_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
z_h	n_{h1}	n_{h2}	$\dots$	n_{hj}	$\dots$	n_{hk}	$n_{h\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot k}$	n

FREQUENZE RELATIVE MARGINALI

	Fumatore	Non Fumatore	Totale
Giovane	15	35	**50**
Adulto	30	20	**50**
Anziano	5	20	**25**
Totale	**50**	**75**	**125**

	Fumatore	Non Fumatore	Totale
Giovane	$\frac{15}{125} = 0.12$	$\frac{35}{125} = 0.28$	**0.40**
Adulto	$\frac{30}{125} = 0.24$	$\frac{20}{125} = 0.16$	**0.40**
Anziano	$\frac{5}{125} = 0.04$	$\frac{20}{125} = 0.12$	**0.20**
Totale	**0.40**	**0.60**	**1**

	Totale	Frequenza Relativa Marginale
Giovane	**50**	$\frac{50}{125} = 0.40$
Adulto	**50**	$\frac{50}{125} = 0.40$
Anziano	**25**	$\frac{25}{125} = 0.20$
Totale	**125**	**1**

	Totale	Frequenza Relativa Marginale
Fumatore	**50**	$\frac{50}{125} = 0.40$
Non Fumatore	**75**	$\frac{75}{125} = 0.60$
Totale	**125**	**1**

TABELLE DI CONTINGENZA

- In R per determinare le distribuzioni di frequenza bivariate il comando da utilizzare è sempre `table()`
  - Ad esempio, per un campione di 15 individui, consideriamo una tabella di contingenza supponendo che *X* individua il colore dei capelli (biondi, castani, neri, rossi) e *Y* il tipo di capelli (lisci, ondulati, ricci):

```
> colore<-c("biondi","castani","neri","rossi","biondi","castani",
+ "neri","castani","biondi","biondi","castani","castani",
+ "castani","neri","rossi")
>
> tipo<-c("lisci","lisci","lisci","lisci","ondulati","ondulati",
+ "ondulati","ondulati","ondulati","ondulati","ondulati",
+ "ricci","ricci","lisci","ricci")
>
> table(colore,tipo)
         tipo
colore    lisci ondulati ricci
  biondi      1        3     0
  castani     1        3     2
  neri        2        1     0
  rossi       1        0     1
```

TABELLE DI CONTINGENZA

- Dopo aver creato la tabella della distribuzione di frequenza congiunta `nomeTable` utilizziamo le operazioni di slicing
  - Per estrarre la riga *i*-esima basta utilizzare il comando `nomeTable[i, ]`
  - Per estrarre la colonna *j*-esima basta utilizzare il comando `nomeTable[,j]`.
- Ad esempio, riferendoci all'esempio precedente, estraiamo la seconda riga e la terza colonna della tabella di contingenza:

```
> capelli<-table(colore,tipo)
>
> capelli[2,] # seconda riga della tabella
   lisci ondulati     ricci
       1         3         2
>
> capelli[,3] # terza colonna della tabella
 biondi castani     neri    rossi
      0       2        0        1
```

TABELLE DI CONTINGENZA

- I comandi `margin.table(nomeTable, 1)` e `margin.table(nomeTable, 2)` permettono rispettivamente di ottenere la distribuzione di frequenza marginale di *X* e la distribuzione di frequenza marginale di *Y*
  - Riferendosi all'esempio precedente calcoliamo le due distribuzioni di frequenza marginali:

```
> capelli<-table(colore,tipo)
>
> margin.table(capelli,1) # distribuzione marginale del colore dei
    capelli
colore
 biondi castani    neri   rossi
      4       6       3       2
>
> margin.table(capelli,2) # distribuzione marginale del tipo dei
    capelli
tipo
   lisci ondulati    ricci
       5       7       3
```

GRAFICI DI FREQUENZA E CONTINGENZA

Tabelle di contingenza

- In R, per calcolare la distribuzione delle frequenze relative congiunte si può utilizzare il comando `prop.table(nomeTabella)`
- Da questa si possono ricavare le distribuzioni delle frequenze relative marginali utilizzando `margin.table()`

```
> prop.table(capelli) # frequenze relative congiunte
          tipo
colore          lisci   ondulati      ricci
  biondi   0.06666667 0.20000000 0.00000000
  castani  0.06666667 0.20000000 0.13333333
  neri     0.13333333 0.06666667 0.00000000
  rossi    0.06666667 0.00000000 0.06666667
>
> freqCapelli <- prop.table(capelli)
>
> margin.table(freqCapelli, 1) # distribuzione delle frequenze
    relative marginali del colore dei capelli
colore
   biondi    castani      neri     rossi
0.2666667 0.4000000 0.2000000 0.1333333
>
> margin.table(freqCapelli, 2) # distribuzione delle frequenze
    relative marginali del tipo di capelli
tipo
    lisci  ondulati     ricci
0.3333333 0.4666667 0.2000000
```

GRAFICI DI FREQUENZA E CONTINGENZA

Tabelle di contingenza

- Conoscendo la distribuzione delle frequenze relative congiunte e la distribuzione delle frequenze relative marginali è possibile calcolare la **distribuzione delle frequenze relative di *Y* condizionata dalle modalità assunte da X**, così definita:

$$f(j|i) = \frac{f_{ij}}{f_{i.}} = \frac{n_{ij}}{n_{i.}} \qquad (i = 1, 2, \ldots, h; j = 1, 2, \ldots, k)$$

- Allo stesso modo, è possibile calcolare **distribuzione delle frequenze relative di *X* condizionata dalle modalità assunte da *Y***:

$$f(i|j) = \frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n_{.j}} \qquad (i = 1, 2, \ldots, h; j = 1, 2, \ldots, k)$$

- Una proprietà fondamentale delle frequenze relative condizionate è la seguente:

$$\sum_{j=1}^{k} f(j|i) = 1 \ (i = 1, 2, \ldots, h), \qquad \sum_{i=1}^{h} f(i|j) = 1 \ (j = 1, 2, \ldots, k).$$

GRAFICI DI FREQUENZA E CONTINGENZA

- Il comando `prop.table(nomeTabella, 1)` permette invece di ottenere la distribuzione delle **frequenze relative di *Y* condizionata dalle modalità assunte da *X***, mentre il comando `prop.table(nomeTabella, 2)` permette di ricavare la distribuzione delle frequenze relative di *X* condizionata dalle modalità assunte da *Y*.
- Riferendosi all'esempio precedente si ha:

```
> prop.table(capelli, 1) # distribuzione delle frequenze relative
    condizionate f(j|i)
          tipo
colore         lisci  ondulati     ricci
  biondi   0.2500000 0.7500000 0.0000000
  castani  0.1666667 0.5000000 0.3333333
  neri     0.6666667 0.3333333 0.0000000
  rossi    0.5000000 0.0000000 0.5000000
>
> prop.table(capelli, 2) # distribuzione delle frequenze relative
    condizionate f(i|j)
          tipo
colore         lisci  ondulati     ricci
  biondi   0.2000000 0.4285714 0.0000000
  castani  0.2000000 0.4285714 0.6666667
  neri     0.4000000 0.1428571 0.0000000
  rossi    0.2000000 0.0000000 0.3333333
```

GRAFICI DI FREQUENZA E CONTINGENZA

Grafici per tabelle di contingenza

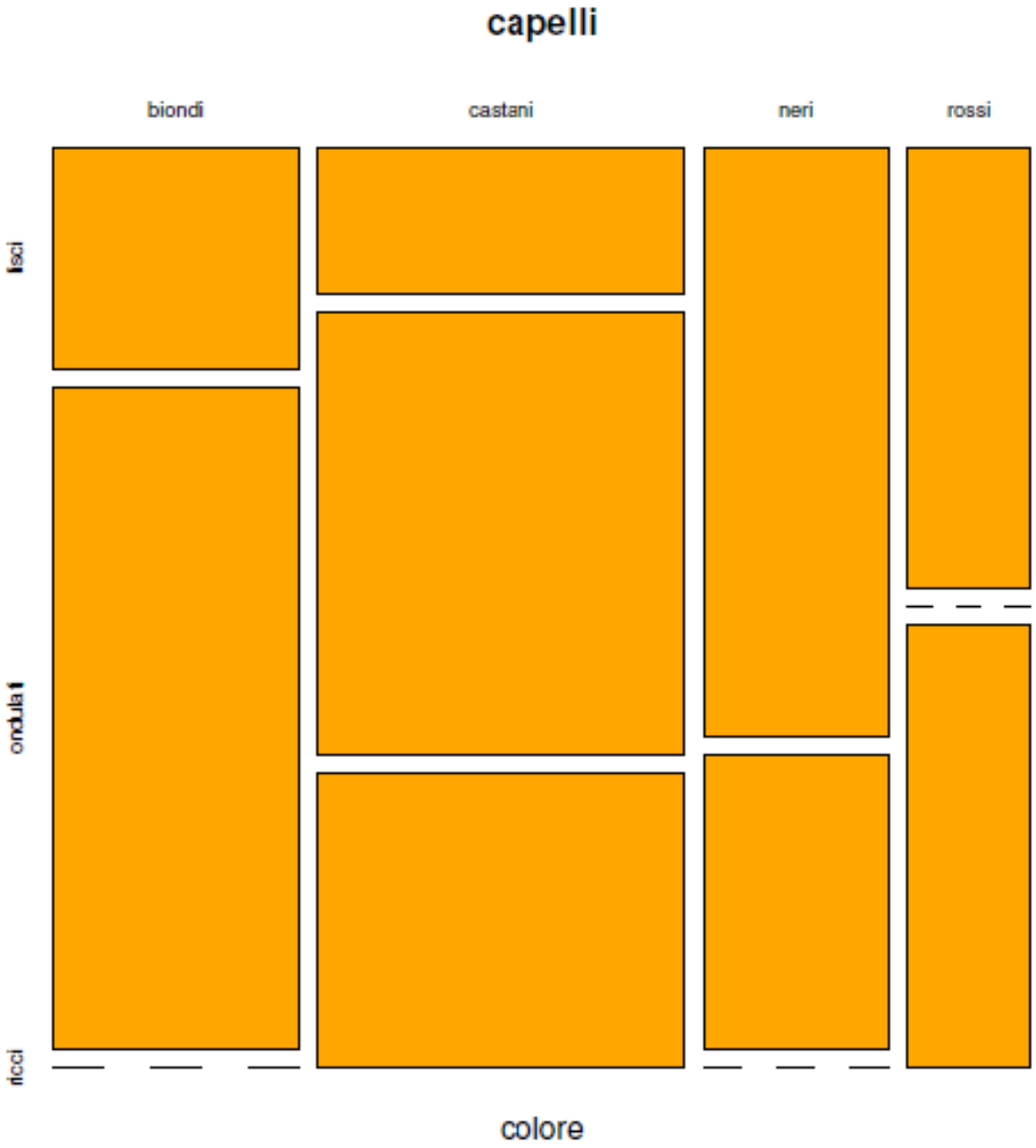
- Sia X una variabile di tipo qualitativo e indichiamo con $z_1, z_2, \ldots, z_h$ le modalità distinte da essa assunte e sia Y un'altra variabile e indichiamo con $w_1, w_2, \ldots, w_k$ le modalità distinte da essa assunte.

- Considerando un campione $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ costituito da n osservazioni di (X, Y), costruiamo la tabella di contingenza contenente nella posizione (i, j) la frequenza congiunta n_{ij} , che rappresenta il numero di volte in cui una particolare coppia di valori (z_i, w_j) si presenta nel campione.

- Le tabelle di contingenza possono essere rappresentate graficamente in vari modi.

GRA ONTINGENZA

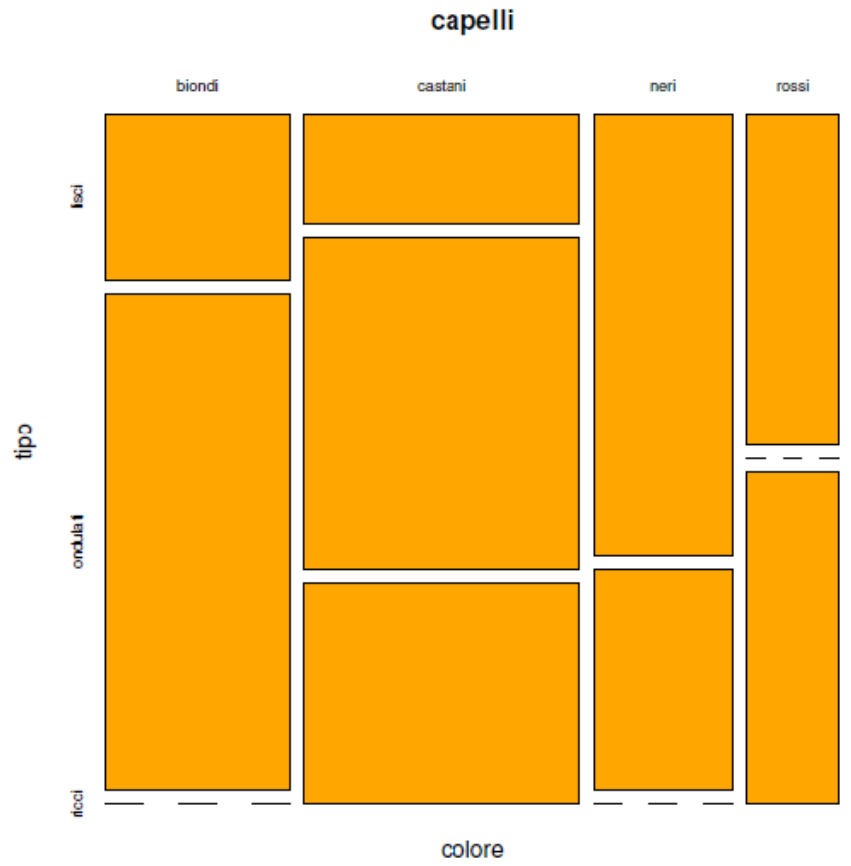
- Una tabella grafico con rettangoli è
- Il rettangolo posizione n relativo dell congiunta.
- In R, se si as rettangoli d plot(Z).
- Ad esempio un campion



GRAFICI DI FREQUENZA E CONTINGENZA

Grafici per tabelle di contingenza

```
> colore<-c("biondi","castani","neri","rossi","biondi","castani",
+ "neri","castani","biondi","biondi","castani","castani",
+ "castani","neri","rossi")
>
> tipo<-c("lisci","lisci","lisci","lisci","ondulati","ondulati",
+ "ondulati","ondulati","ondulati","ondulati","ondulati",
+ "ricci","ricci","lisci","ricci")
>
> table(colore,tipo)
          tipo
colore     lisci ondulati ricci
  biondi       1         3     0
  castani      1         3     2
  neri         2         1     0
  rossi        1         0     1
> capelli<-table(colore,tipo)
>
> plot(capelli,col="orange")
```

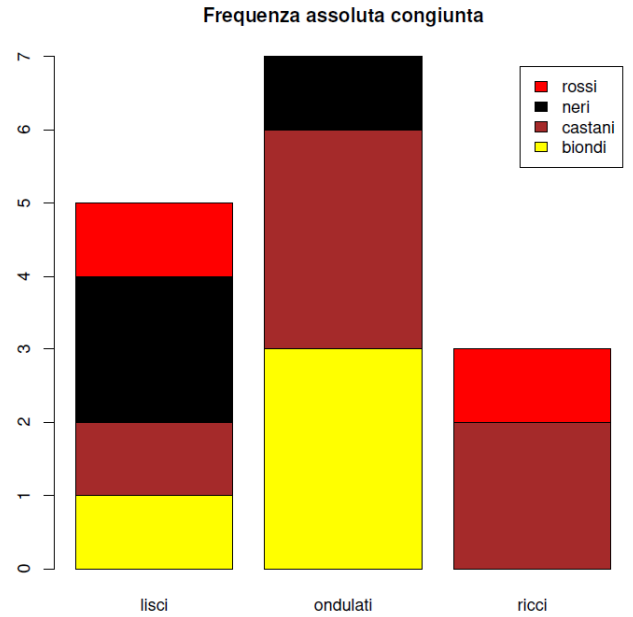


- Da notare come siano presenti solo 9 rettangoli pieni e 3 vuoti corrispondenti alle coppie la cui frequenza è nulla

GRAFICI DI FREQUENZA E CONTINGENZA

- La tabella di contingenza può essere rappresentata mediante un **grafico a barre sovrapposte (Stacked)**.
- Il numero di barre è pari al numero delle modalità delle colonne della tabella di contingenza
  - Inoltre, all'interno di ciascuna barra sono rappresentate una sopra l'altra in altezza **le frequenze di ciascuna modalità** delle righe della tabella di contingenza.
- Nel seguente codice, invece, sono state invertite le righe e le colonne della tabella di contingenza. Si noti come sull'asse delle ordinate siano indicate le frequenza assolute congiunte

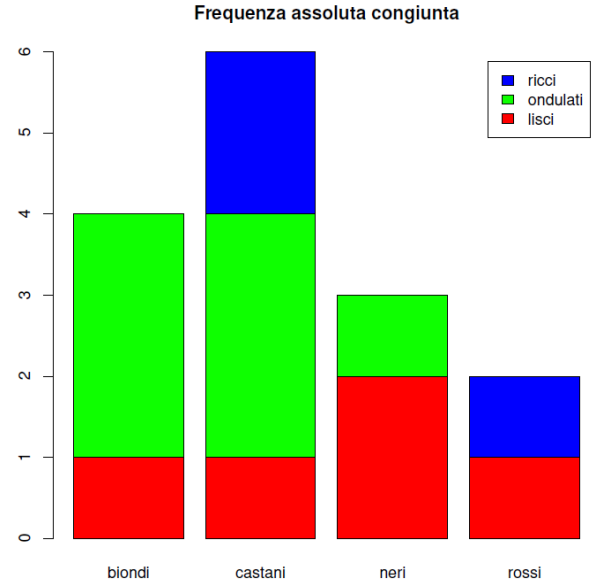
```
> capelli<-table(colore,tipo)
> barplot(capelli,main="Frequenza assoluta congiunta",
+ legend=c("biondi","castani","neri","rossi"),
+ col=c("yellow","brown","black","red"))
```



GRAFICI DI FREQUENZA E CONTINGENZA

- È preferibile rappresentare la tabella di contingenza mediante un **grafico a barre sovrapposte (Stacked)**.
- Il numero di barre è pari al numero delle modalità delle colonne della tabella di contingenza
  - Inoltre, all'interno di ciascuna barra sono rappresentate una sopra l'altra in altezza **le frequenze di ciascuna modalità** delle righe della tabella di contingenza.
- Nel seguente codice, invece, sono state invertite le righe e le colonne della tabella di contingenza. Si noti come sull'asse delle ordinate siano indicate le frequenza assolute congiunte

```
> capelliNew<-table(tipo,colore)
> barplot(capelliNew,main="Frequenza assoluta congiunta",
+ legend=c("lisci","ondulati","ricci"),
+ col=c("red","green","blue"))
```

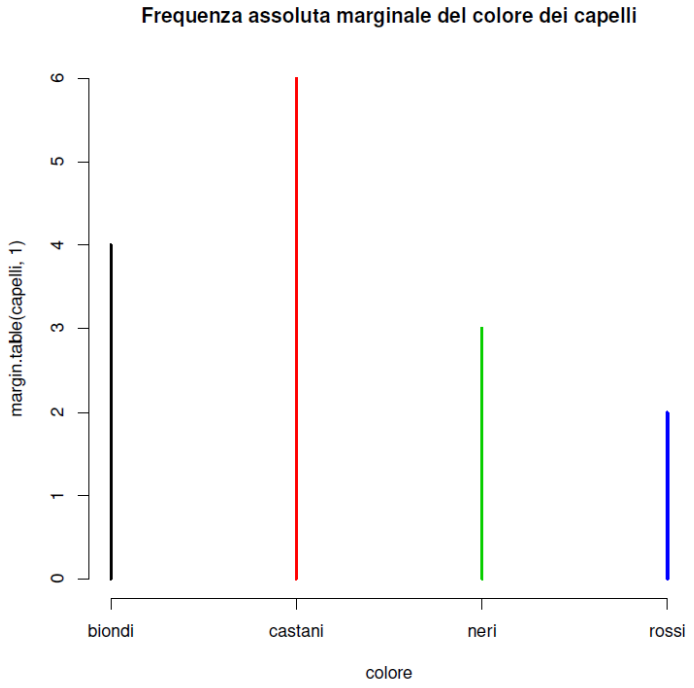


GRAFICI DI FREQUENZA E CONTINGENZA

Grafici per tabelle di contingenza

- **Grafici per le frequenze assolute marginali:** Per ottenere un grafico a barre della distribuzione marginale relativa al colore dei capelli basta utilizzare il comando:

```
> plot(margin.table(capelli,1),
+ main="Frequenza assoluta marginale del colore dei capelli",
+ col=1:4)
```



GRAFICI DI FREQUENZA E CONTINGENZA

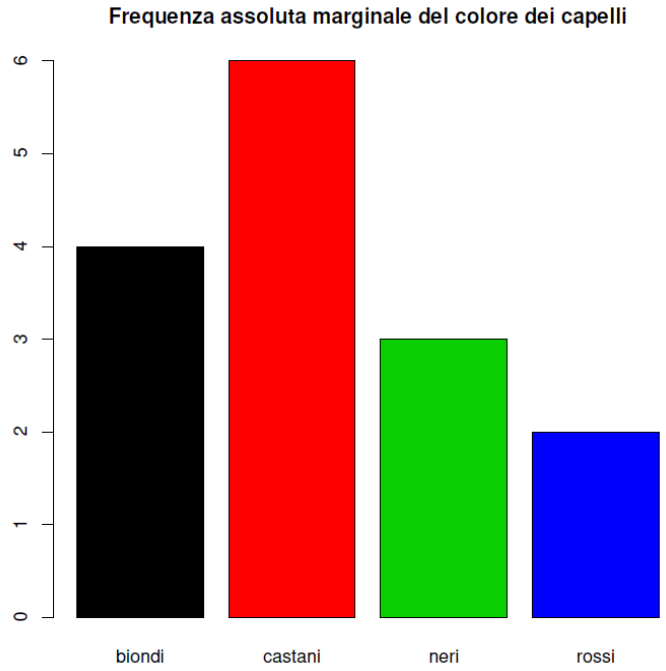
Grafici per tabelle di contingenza

- **Grafici per le frequenze assolute marginali:** Per ottenere un grafico a barre della distribuzione marginale relativa al colore dei capelli basta utilizzare il comando:

```
> plot(margin.table(capelli,1),
+ main="Frequenza assoluta marginale del colore dei capelli",
+ col=1:4)
```

- Oppure:

```
> barplot(margin.table(capelli,1),
+ main="Frequenza assoluta marginale del colore dei capelli",
```

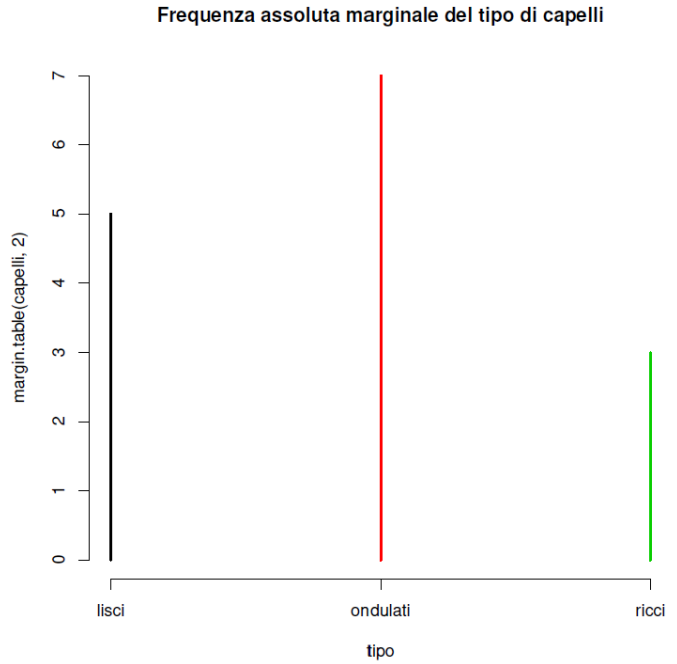


GRAFICI DI FREQUENZA E CONTINGENZA

Grafici per tabelle di contingenza

- Analogamente, per il grafico della **distribuzione marginale** relativa al tipo di capelli basta utilizzare il comando:

```
> plot(margin.table(capelli,2),
+ main="Frequenza assoluta marginale del tipo dei capelli",
+ col=1:3)
```



GRAFICI DI FREQUENZA E CONTINGENZA

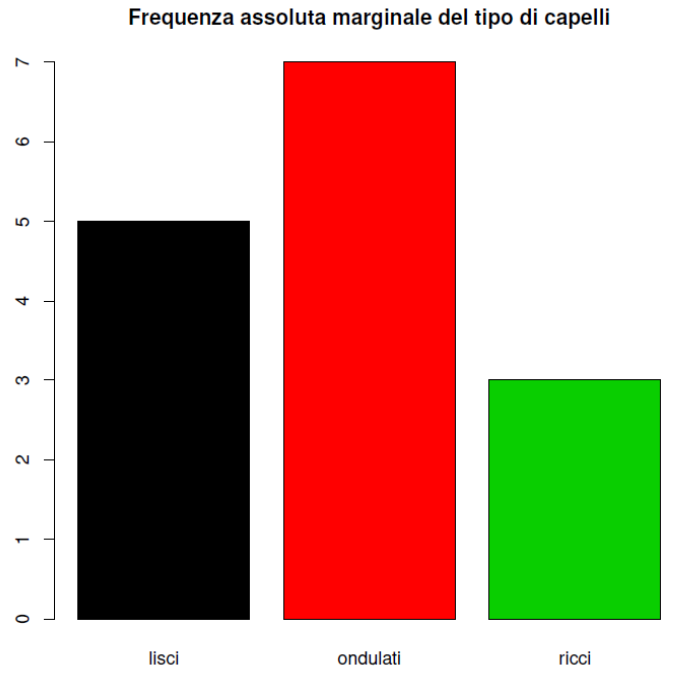
Grafici per tabelle di contingenza

- Analogamente, per il grafico della **distribuzione marginale** relativa al tipo di capelli basta utilizzare il comando:

```
> plot(margin.table(capelli,2),
+ main="Frequenza assoluta marginale del tipo dei capelli",
+ col=1:3)
```

- Oppure, per avere un grafico a barre:

```
> barplot(margin.table(capelli,2),
+ main="Frequenza assoluta marginale del tipo di capelli",
+ col=1:3)
```



GRAFICI DI FREQUENZA E CONTINGENZA

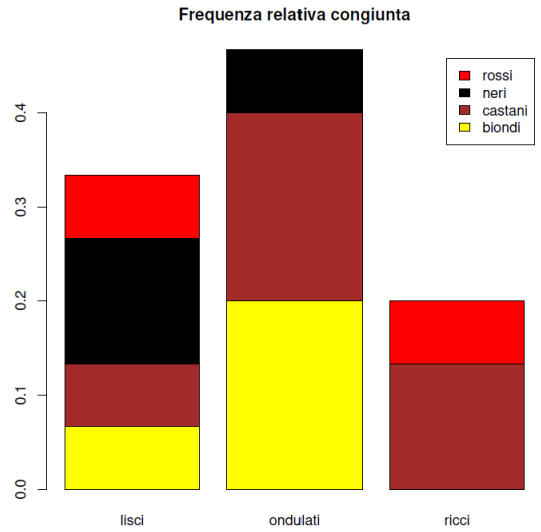
- **Grafici per le frequenze relative congiunte:** Consideriamo ora le distribuzioni le frequenze relative e generiamo un grafico a barre stacked:

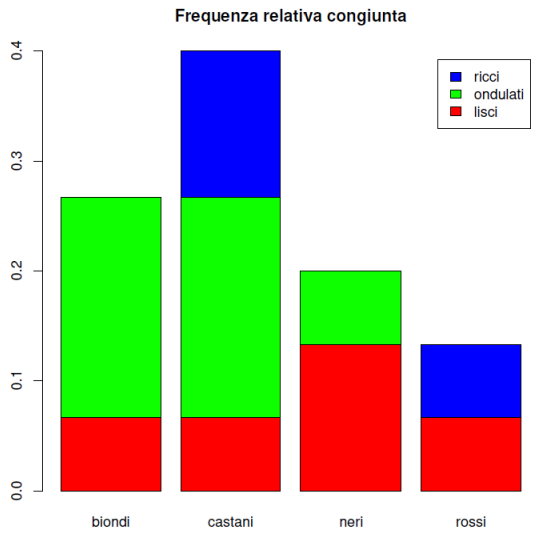
```
> capelli<-table(colore,tipo)
> freqCapelli<-prop.table(capelli)
> barplot(freqCapelli,main="Frequenza relativa congiunta",
+ legend=c("biondi","castani","neri","rossi"),
+ col=c("yellow","brown","black","red"))
```

- Analogamente, con il seguente codice otteniamo un altro barplot equivalente

```
> capelliNew<-table(tipo,colore)
> freqCapelliNew<-prop.table(capelliNew)
> barplot(freqCapelliNew,main="Frequenza relativa congiunta",
+ legend=c("lisci","ondulati","ricci"),
+ col=c("red","green","blue"))
```

- Si noti che in entrambi i grafici sull'asse delle ordinate sono indicate le frequenze relative congiunte





DOMANDE?