Statistica

Giovanna Jona Lasinio

Dipartimento di Scienze Statistiche
Università di Roma "La Sapienza"
email: giovanna.jonalasinio@uniroma1.it
materiale didattico: **(STATECO)**
http://elearning2.uniroma1.it/course/index.php?categoryid=641

Ecobiologia e Scienze Naturali

Outline 1

- Introduzione: elementi di statistica di base (media, varianza, indici di posizione, distribuzioni di frequenza). Grafici e statistiche riassuntive.
- Distribuzioni di probabilità: la distribuzione normale e sua rilevanza nella modellizzazione dei dati biologici. La distribuzione t di Student
- Il test t: principi del test, test t a due campioni con varianze uguali, test t a due campioni con varianze diverse. Test t per misure accoppiate.
- Analisi della varianza: cenni teorici, distribuzione F di Fisher-Snedecor, relazioni tra t ed F.
- Elementi di disegno degli esperimenti per l'analisi della varianza - disegni fattoriali - disegni fattoriali a due fattori - disegni fattoriali a più di due fattori

Outline 2

- Regressione lineare e correlazione: la correlazione di Pearson, il modello lineare con una variabile indipendente. Cenni sulla regressione multipla e sulla regressione non lineare.
- Test del chi-quadro: quando usarlo e quando non usarlo, uso del test come verifica della bontà di adattamento, uso del test nelle tabelle a doppia entrata. Cenni sui test non parametrici.
- Introduzione alle tecniche di analisi multivariata: analisi in componenti principali.
- Iscriversi al corso online: http://elearning.sta.uniroma1.it/moodle/ password: botanica2010

Testi di riferimento:
H. van Emden (2008) *Statistics for terrified biologists Blackwell* Publishing
CAST http://cast.massey.ac.nz/collection_public.html
Software: R package URL http://www.r-project.org/

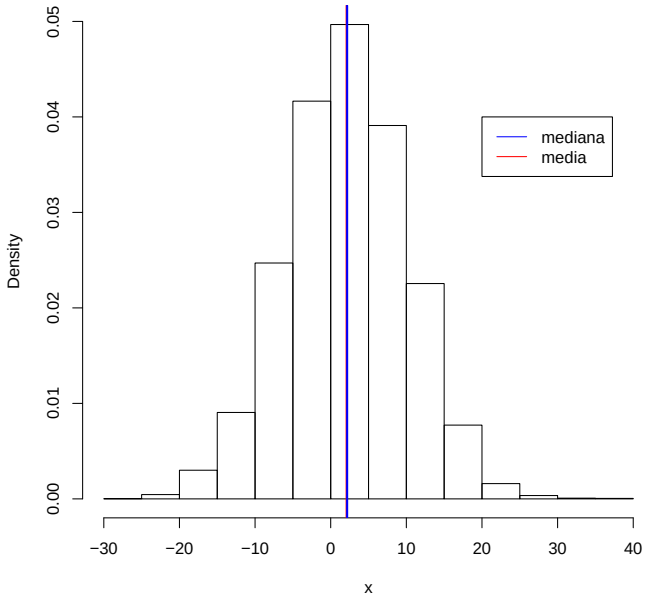
Scopo

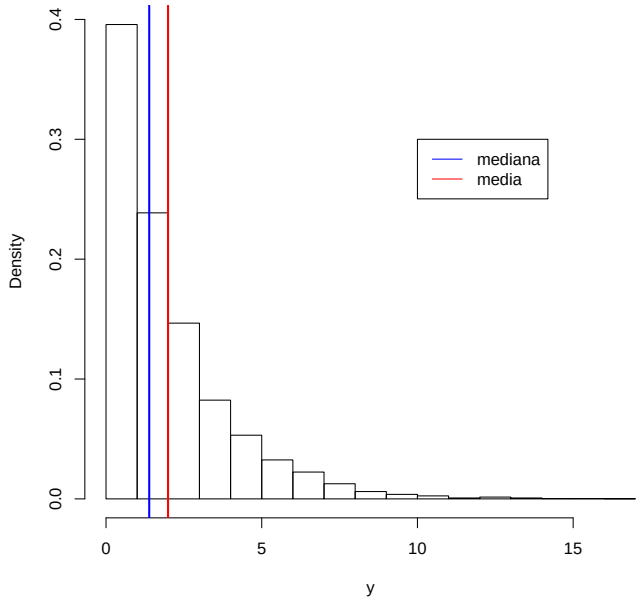
- Lo scopo generale del corso è di rendervi *utenti consapevoli*
- Quindi non viene richiesta la conoscenza dei dettagli tecnici (a parte poche e semplici cose), bensì la comprensione e consapevolezza delle ipotesi, delle assunzioni e del significato che sottende la tecnica ststistica
- Si vuole rendervi capaci di leggere correttamente un grafico, scegliere la tecnica *giusta* per ottenere una specifica informazione dai dati e implementare quest'ultima in R.

Elementi di Statistica di Base

- Il primo passo di una qualunque analisi statistica è **chiarire quale sia o siano le domande a cui si vuole rispondere**
- Ogni tecnica statistica risponde ad una domanda, ad una richiesta di informazione
- Ad esempio:
  1. Quale è un valore centrale attorno a cui oscillano i miei dati? $\rightarrow$ *La Media*: descrive il valore attorno al quale *oscillano* i dati, se immaginiamo i dati come un corpo rigido è il baricentro dell'oggetto
  2. Sotto (o sopra) quale valore si trova il 50% dei dati? $\rightarrow$ *La Mediana*: è quel valore che divide la distribuzione a metà (50% delle osservazione alla sua sinistra e alla sua destra) è il centro della distribuzione.
  3. Quanto sono dispersi i dati (quanto sono variabili)? una possibile risposta la fornisce *La Deviazione standard*: è la distanza euclidea media dalla media, quindi è la distanza media dal baricentro dell'oggetto.

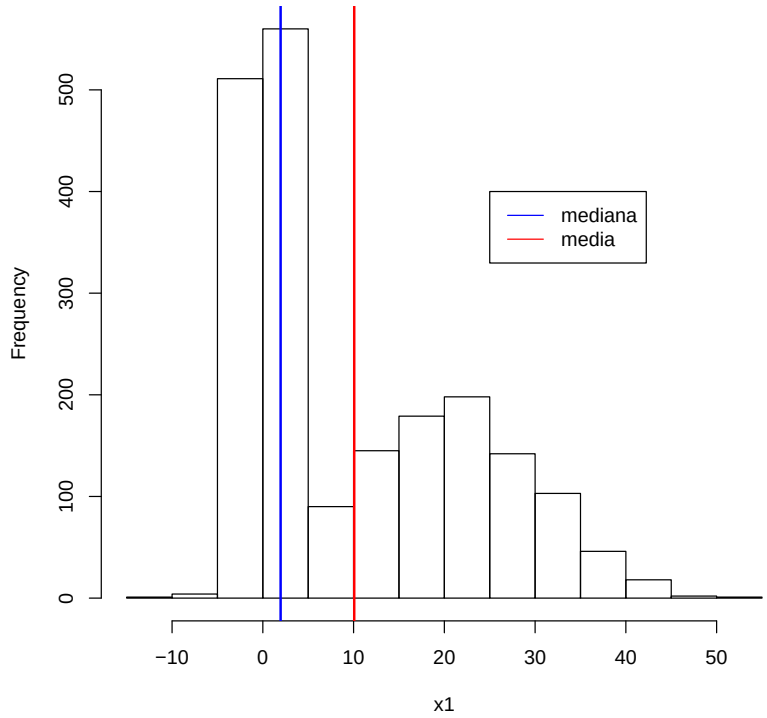
- In generale ci interessa costruire delle grandezze che diano informazioni sintetiche ed esaurienti
- Non è quasi mai sufficiente fornire una sola grandezza alla volta, né solo dei numeri. Ad esempio:





$\mu = 2.11$ $\sigma = 7.91$ asim$= 0$ $\mu = 1.99$ $\sigma = 1.98$ asim$= 1.9$

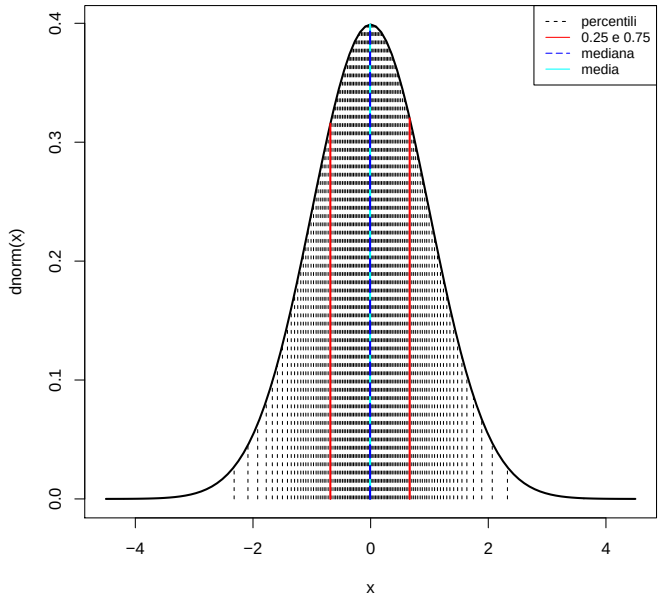
Quando la media funziona davvero male: esempio teorico



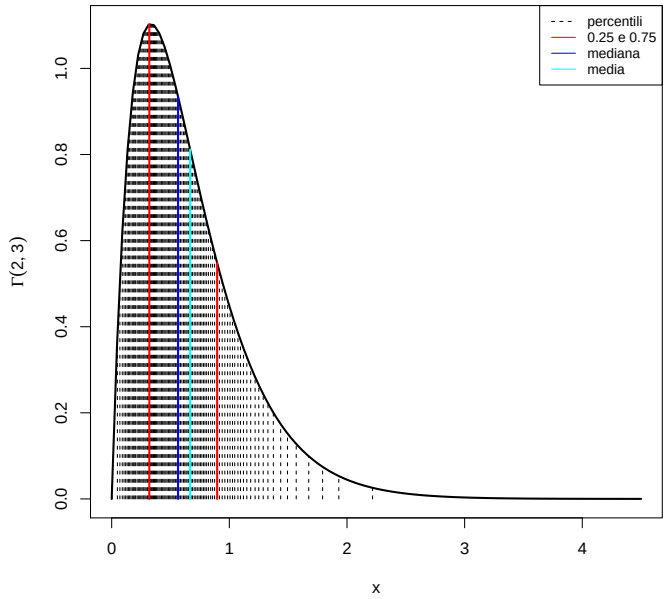
$$\mu = 9.85 \; \sigma = 12.05$$

Percentili

- Sono i valori della variabili d'interesse X che ne dividono la distribuzione in 100 parti uguali
- la mediana è il 50 esimo percentile, il massimo è il 100esimo percentile, il minimo il percentile di ordine 0

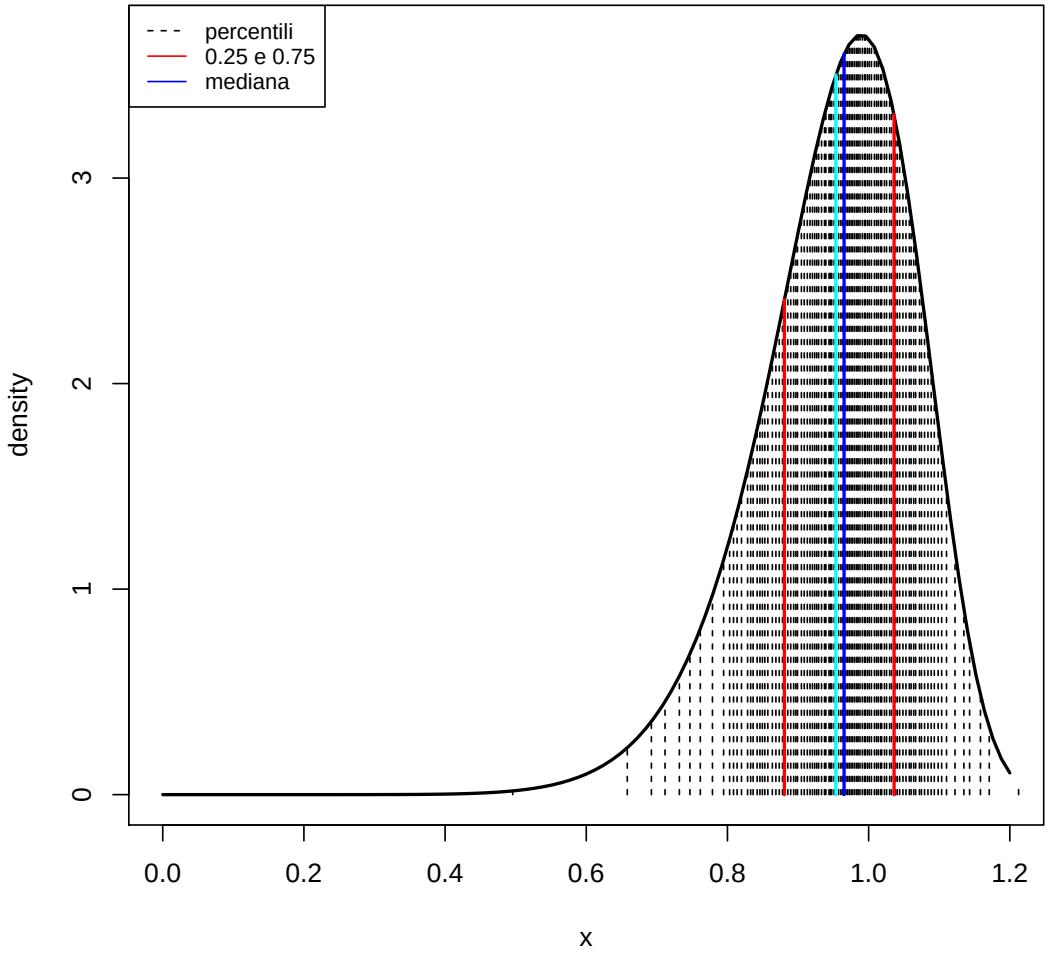


Distribuzione simmetrica
asimmetria= 0



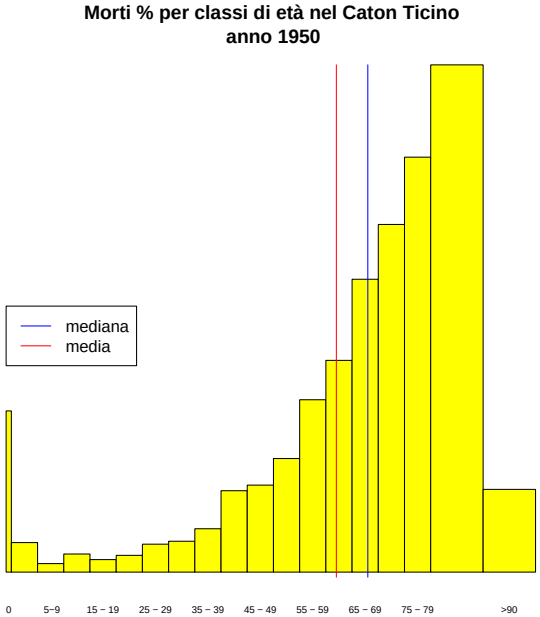
Distribuzione asimmetrica
asimmetria= 1.5 positiva

Percentili

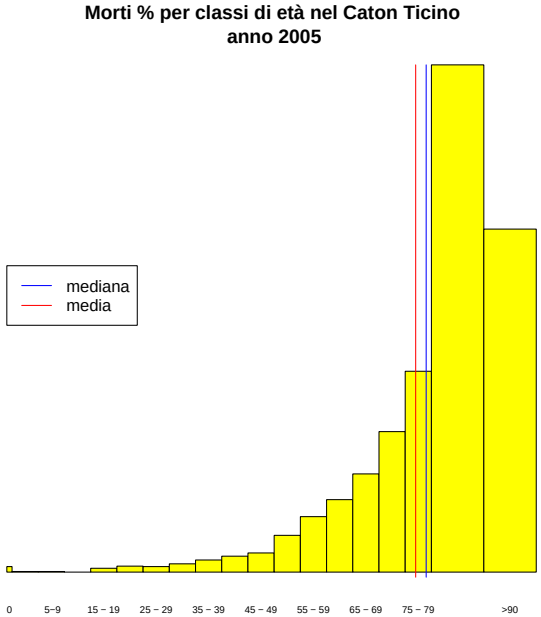


asimmetria$= -0.49$ negativa

Quando la media funziona davvero male: esempio reale distribuzione dei morti per età nel Canton Ticino



1950



2005

- Passando dal 1950 al 2005, oltre a uno spostamento verso destra della zona in cui si concentrano le età di morte (testimoniato dall'aumento sia della media che della mediana), possiamo osservare un maggiore addensamento dei dati verso le età più anziane.
- Questa percezione intuitiva può essere precisata considerando l'intervallo in cui si colloca il 50% centrale dei dati, ossia i dati che vanno dal 25esimo al 75esimo percentile

Indice	1950	2005
25%	52	72
mediana	64	80
75%	76	85
media	63	78

- L'ampiezza di questo intervallo, è detta *distanza interquartile* distanza tra il valore che delimita il primo quarto dei dati da quello che ne delimita l'ultimo quarto, nel nostro esempio passa da 19 a 13.
- La distanza interquartile, (IQR -IntraQuartile Range), è l'indice di dispersione d'uso più generale

- La misura della variabilità dei dati è un'informazione fondamentale.
- Esistono diversi strumenti che possiamo utilizzare:
- **Range** o campo di variazione : minimo e massimo valore della serie di dati
- **Total Deviation** o Devianza totale: se abbiamo n dati x la cui media è $\bar{x}$, la devianza totale è data dalla somma di tutte le differenze in valore assoluto $|x - \bar{x}|$

  $$\sum |x - \bar{x}|$$

  . Il valore di questa misura è però molto influenzato dal numero di osservazioni quindi...
- **Mean Deviation** o devianza media

  $$\frac{\sum |x - \bar{x}|}{n}$$

- **La varianza**: quando la calcoliamo in teoria è semplicemente la distanza euclidea media dalla media aritmetica

$$\frac{\sum(x-\bar{x})^2}{n}$$

- La varianza campionaria è solo leggermente diversa da un punto di vista del calcolo

$$\frac{\sum(x-\bar{x})^2}{n-1}$$

  in parole però diremo che è la somma delle differenze dalla media al quadrato divisa per i *gradi di libertà*.

- Il concetto di gradi di libertà è molto importante in statistica. Nel caso della varianza campionaria dobbiamo tener conto che stiamo usando tutto il campione per calcolare la media e che questa media poi la usiamo nel calcolo della varianza.

- In questo modo *fissiamo (vincoliamo)* l'insieme di dati che usiamo per la varianza ad avere quella media.

Altre misure di dispersione

- Spesso, ad esempio se si vogliono confrontare dati analoghi ma espressi con unità di misura diverse, può essere comodo ricorre all'indice di dispersione noto come coefficiente di variazione e indicato con CV =deviazione standard/media. Lo si usa, ovviamente, se media $\neq 0$.
- Un altro indice utile è il coefficiente di asimmetria, pari a la media dello scarto cubico (dalla media), divisa per il cubo della deviazione standard.
- Se i dati sono simmetrici rispetto alla media l'indice è nullo; se hanno una coda verso destra è positivo; se l'hanno verso sinistra, l'indice è negativo.

Natura dei dati e misure statistiche: che tipo di dati e che tipo di misure

- **Quantitativi**: massimo contenuto informativo. Su di essi è possibile calcolare medie aritmetiche, geometriche, varianze ecc.
- **Qualitativi ordinati** (ad esempio il titolo di studio). Indici di posizione (mediana, percentili)
- **Qualitativi sconnessi** (colore degli occhi). Distribuzioni di frequenza

- Come valutiamo la dispersione se abbiamo dei dati qualitativi?
- Una possibilità è usare IQR che mi da informazioni limitate ed ha senso solo per dati almeno ordinati.
- Immaginiamo di avere osservato una caratteristica con K modalità (categorie). Possiamo allora distinguere due situazioni estreme:
  1. Massima eterogeneità: tutte le modalità hanno eguale frequenza ($n_i = N/K$)
  2. Minima eterogeneità (massima omogeneità): tutte le unità statistiche presentano la stessa modalità (ad esempio $n_i = 0$, $i = 1, 2, \ldots, h-1, h+1, \ldots, K$, $n_h = N$)

- Possiamo costruire delle misure di eterogeneità:

Indice di Gini	$G = 1 - \sum_{k=1}^{K} \frac{n_i}{N}$	Varia tra 0 e $1 - \frac{1}{K}$
Indice di Gini Relativo	$G_R = \frac{G}{1-1/K}$	Varia tra 0 e 1
Entropia	$H = -\sum_{i=1}^{K} \frac{n_i}{N} \log\left(\frac{n_i}{N}\right)$	Varia tra 0 e $\log(K)$
Entropia Relativa	$H_R = \frac{H}{\log(K)}$	Varia tra 0 e 1

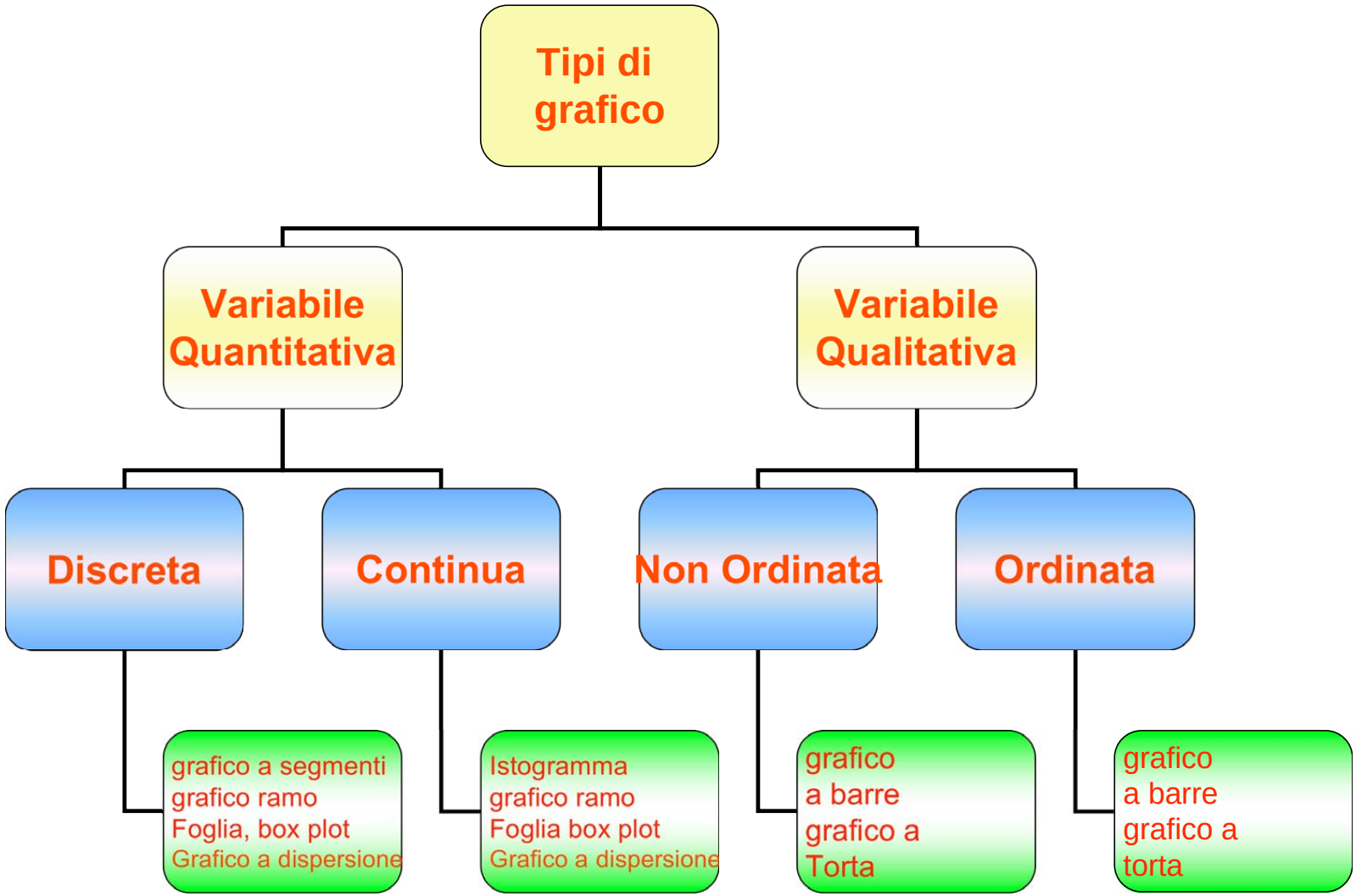
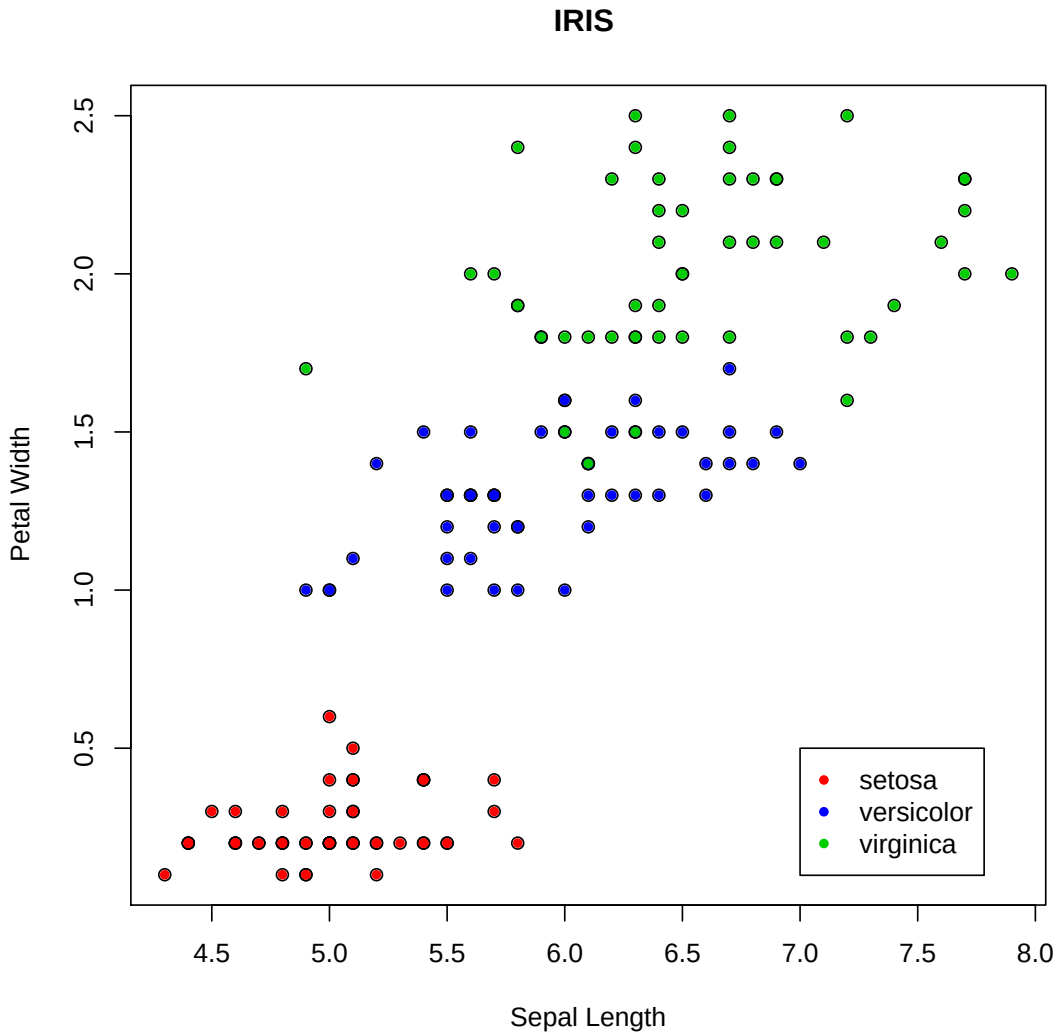
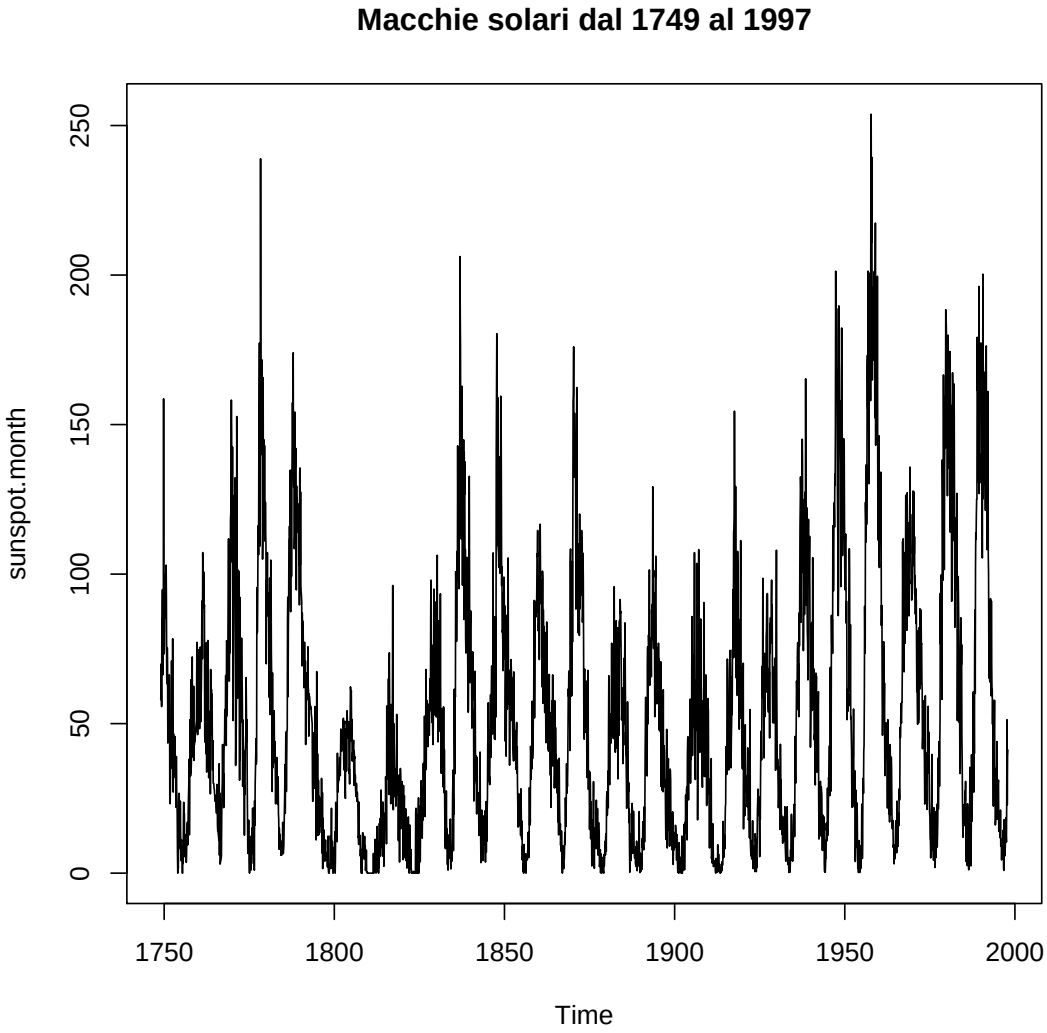

Tipi di grafico
Variabile Quantitativa
Variabile Qualitativa
Discreta
Continua
Non Ordinata
Ordinata
grafico a segmenti
grafico ramo
Foglia, box plot
Grafico a dispersione
Istogramma
grafico ramo
Foglia box plot
Grafico a dispersione
grafico
a barre
grafico a
Torta
grafico
a barre
grafico a
torta

Grafico a dispersione



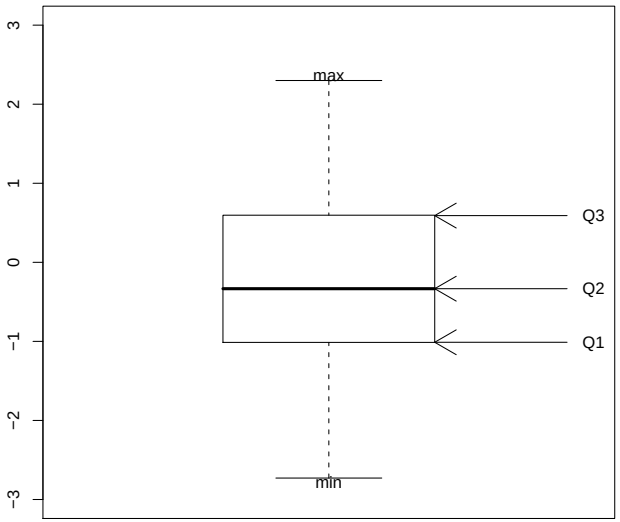
- Nel grafico a dispersione cerchiamo di individuare le relazioni tra due variabili quantitative.
- Se si dispongono lungo una retta, se si raggruppano o se insieme danno luogo ad altri pattern individuabili ad occhio nudo.
- Se una delle due variabili è il tempo ci permettono di visualizzare la *serie storica* degli eventi che stiamo studiando e capire che relazioni possono esistere tra le osservazioni.

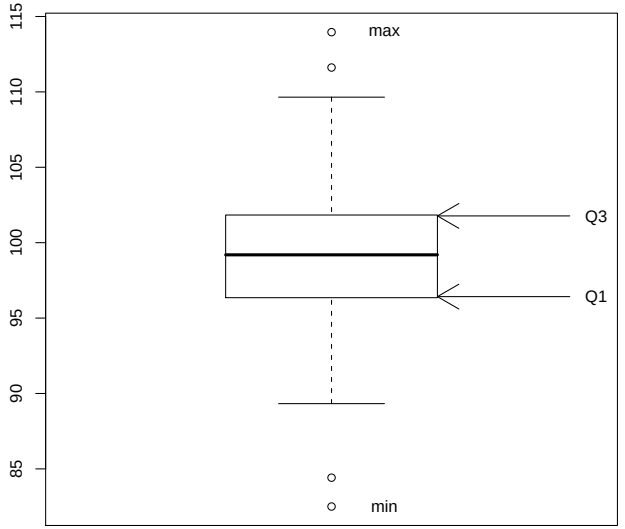
Grafico a dispersione, serie storica, è chiaramente visibile una periodicità con cicli sia secolari che di minor durata



Boxplot:

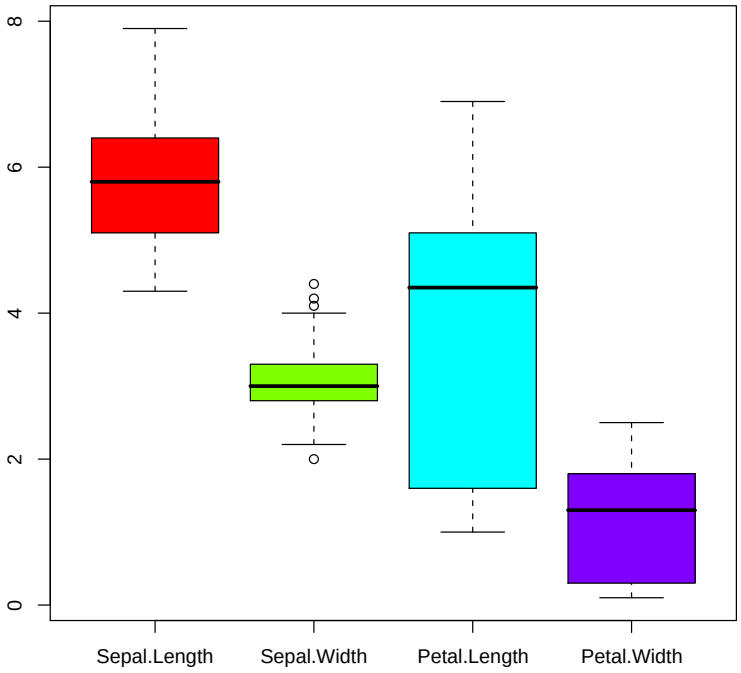
- Il box-plot (o semplicemente boxplot o anche box and whiskers plot, letteralmente: diagramma a scatola e baffi) è una forma di rappresentazione grafica che serve per descrivere in modo compatto la distribuzione di una variabile.
- È il disegno su un piano cartesiano di un rettangolo, i cui estremi sono il primo e terzo quartile (Q1 e Q3), è tagliato a metà da una linea che rappresenta la mediana (Q2). Il minimo della distribuzione viene indicato con (Q0), mentre il massimo con (Q4).
- Abitualmente vengono aggiunte due righe (detti anche baffi) corrispondenti ai valori distanti 1.5 volte la distanza interquartile (Q3-Q1) a partire rispettivamente dal primo dal terzo quartile. Alle volte vengono anche rappresentati nel grafico i valori che fuoriescono dall'intervallo delimitato dai due baffi come punti isolati (valori anomali)
- Pare che a John Wilder Tukey venne chiesto perch é nella determinazione dei valori adiacenti superiore ed inferiore fosse stata scelta una distanza limite dai quartili pari a 1.5 e lui avrebbe risposto perché 1 è poco e 2 troppo.

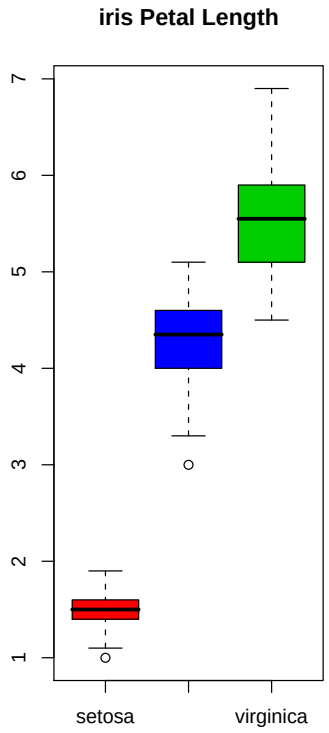


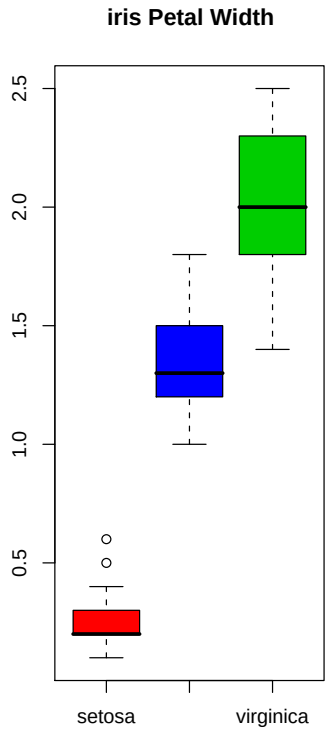


senza valori anomali

con valori anomali


0
2
4
6
8
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width


iris Petal Length
1
2
3
4
5
6
7
setosa
virginica

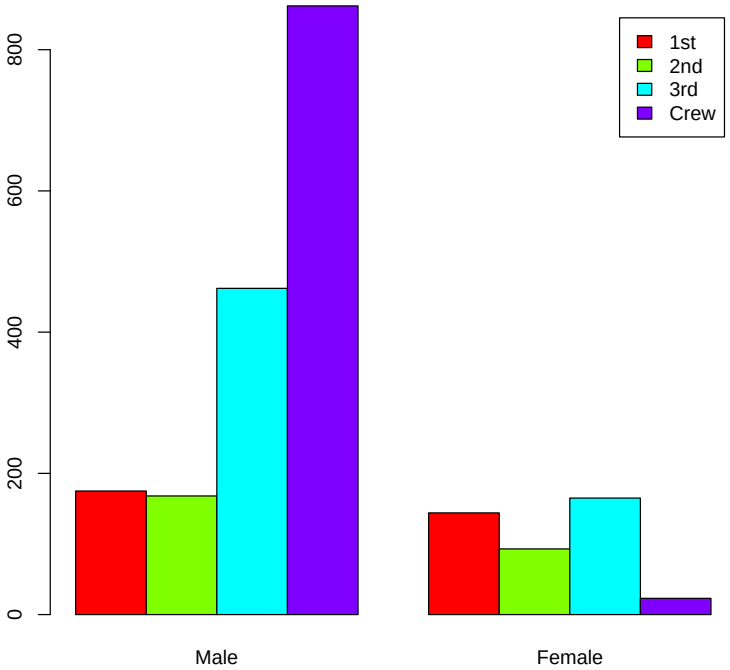

iris Petal Width
0.5
1.0
1.5
2.0
2.5
setosa
virginica

Dati Qualitativi: Adulti presenti sul Titanic per sesso e classe

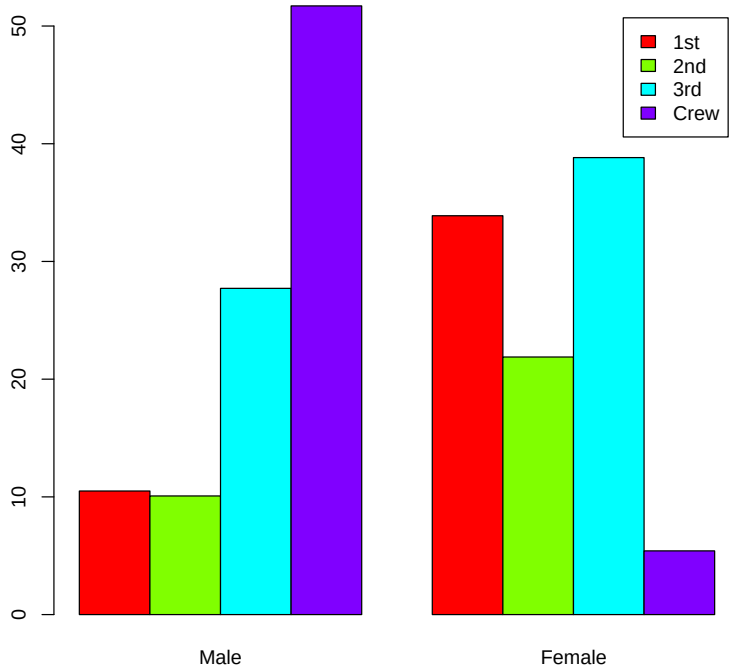
Classe	Male	Female
1st	175	144
2nd	168	93
3rd	462	165
Crew	862	23

Dati Qualitativi: grafici a barre

Adulti presenti sul Titanic distinti per sesso



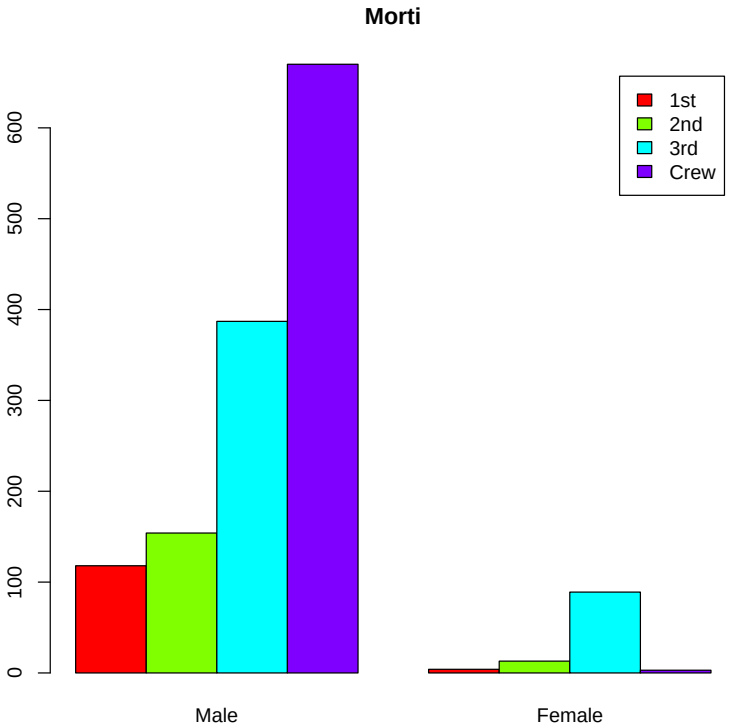
valori assoluti

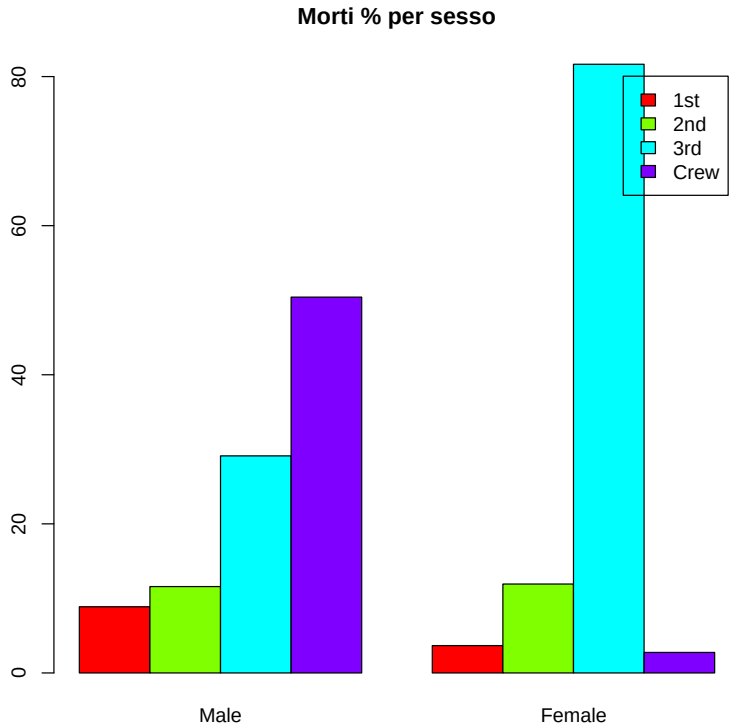


percentuali per sesso

Dati Qualitativi: Morti adulti del Titanic per sesso e classe

Classe	Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3





valori assoluti

percentuali per sesso

le percentuali sono costruite rispetto al totale dei morti per ciascun genere

Visualizziamo la tavola intera dei dati Titanic: Grafico a mosaico