Corso di Statistica – Introduzione e Capitolo 2

Giovanni Marchetti

6 marzo 2019

Organizzazione del Corso

Organizzazione del corso

- Questo è il corso di Statistica 2019 CdL Economia B009
- lettere A-C B018993 B000319
- Docente/Esercitatore: **Giovanni Marchetti**
- 9 crediti (6 ore a settimana per 12 settimane)

Orario

Mercoledì	8:30 - 10:00	D5/003
Giovedì	12:00 - 13:30	D5/003
Venerdì	8:30 - 10:00	D5/003

Libro di testo

Facciamo i Capitoli 1-10 del libro

- **Newbold, Carlson, Thorne**. *Statistica, Principi e Metodi*, Pearson Prentice Hall. C'è una versione Ebook.
- vecchie edizioni OK
- Useremo **MYLAB**
- Programma: capitoli da 2 a 10. Spiegazioni su Moodle

Esame

- **Scritto** (20 domande di vario tipo)
- **Orale** (1 domanda di teoria)

Appelli

- 18 giugno 2019
- 16 luglio 2019
- 3 settembre 2019
- 9 dicembre 2019

Raccomandazione

Lo scritto non è facile, richiede impegno e ha un peso elevato sul risultato dell'esame.

Ma, **per esperienza**, tra le due strategie

1. Saper risolvere tutti gli esercizi dei compiti passati meccanicamente
2. Capire i concetti e fare un numero sufficiente di esercizi

dà **risultati molto migliori la seconda**!

Cose da fare subito

- Iscriversi al corso su MOODLE `e-l.unifi.it`. **Non c'è password**!
- Scaricare i lucidi del Capitolo 2. I lucidi vengono messi nella cartella **Lucidi** all'inizio della settimana.
- I lucidi sono una guida alla **lettura del libro**

Cose da fare dopo la lezione

- Procuratevi il libro e **leggete le parti fatte a lezione**.
- Procuratevi una **calcolatrice** (non usate il telefonino) e imparate a usarla. Meglio se permette di vedere le espressioni.
- **Fare alcuni esercizi** dopo ogni lezione.

Ricevimento

- **Prima o dopo la lezione**
- **Per email** a `giovanni.marchetti@unifi.it`, risposta garantita.
- Ricevimento al **Dipartimento di Statistica e Informatica**, Prendere appuntamento per email.

Capitolo 2

Sommario

- Dati
- Variabili e unità
- Classificazione delle variabili
  - qualitativa/quantitativa
  - discreta/continua
- Spoglio e distribuzioni di frequenza
- Visualizzazione

Dati quantitativi discreti

Voti allo scritto di Statistica di settembre 2018.

```
 8 18 21 20 28 20 15 18 20  8 22 12 19 30 29
16 20 19 20  2 30 16 14 13  8 30 15 14 23 21
30 14 22  7 30 22 19 16 21 18  8 30 15  5  9
30 18  9  9 16 25 27 30 11 18 14 10 21  4 11
17 27 12  8 11  6
```

- **Unità di studio**: studenti
- Rilevato il voto (**variabile**)
- La variabile è **quantitativa discreta**
- Il tempo è il momento dello scritto.

Dati quantitativi continui

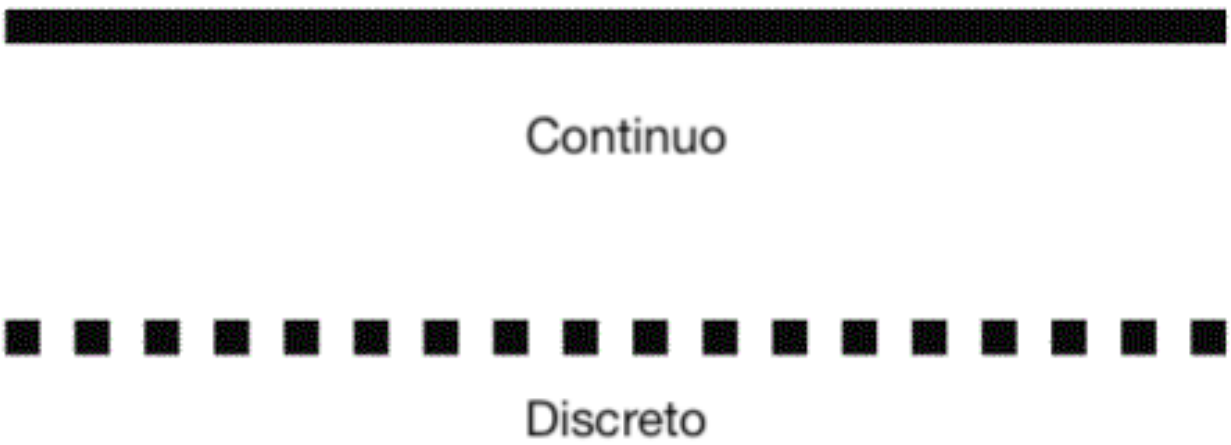
Ore di studio di 20 studenti (Libro Esercizio 2.33)

```
3.5 2.8 4.5 6.2 4.8 2.3 2.6 3.9 4.4 5.5
5.2 6.7 3.0 2.4 5.0 3.6 2.9 1.0 2.8 3.6
```

- **Unità**: 20 studenti
- Rilevato un tempo in ore (**variabile**)
- La variabile è **quantitativa continua**
- Il tempo è un giorno prima dell'esame

Differenza discreto/continuo

- Le variabili **discrete** sono misurate con un numero **intero**
- Le variabili **continue** sono misurate con un **numero reale**



Dati qualitativi nominali

Domanda di un prodotto di 750 clienti (Libro Esempio 2.9)

Tipo	Numero di clienti
Attrezzi	215
Legname da costruzione	215
Vernici	170
Altro	150
Totale	750

- **Unità**: 750 clienti
- Tipo di acquisto (**variabile**)
- La variabile è **qualitativa nominale**
- Osservazioni su varie unità in un certo intervallo di tempo

Variabile **qualitativa** = variabile **categorica**

Variabile qualitativa ordinale

Soddisfazione di un campione di dipendenti (Libro Eserc. 2.11)

Soddisfazione	Numero dipendenti
Molto soddisfatto	29
Abbastanza soddisfatto	55
Indifferente	5
Abbastanza insoddisfatto	20
Molto insoddisfatto	9
Totale	118

- **Unità** 118 dipendenti
- Grado di soddisfazione (**variabile**)
- La variabile è **qualitativa ordinale**
- Osservazioni su varie unità a un certo tempo

Serie storica

Immatricolati a Economia e Commercio negli AA dal 2008 al 2017

08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17
242	268	304	412	314	271	216	222	310

- **Unità**: il Corso di laurea
- Osservazioni su 9 AA (**tempi**)
- immatricolati (**variabile**)
- La variabile è **quantitativa discreta**

Spoglio e distribuzione di frequenze

- Quando si analizzano dati di una variabile osservati su n unità è comune fare una operazione di spoglio
- Per **spoglio** si intende la classificazione delle unità in gruppi determinati dalle **modalità** della variabile

Voti all'orale di Statistica

```
25 28 30 28 28 26 26 26 25 24 27 27

24 | 0          (1)
25 | 00         (2)
26 | 000        (3)
27 | 00         (2)
28 | 000        (3)
29 |            (0)
30 | 0          (1)
```

Spoglio e distribuzione di frequenze

Si definisce una tabella di modalità e frequenze

- assolute
- relative

	Frequenze		
Voti	assolute	relative	percentuali
24	1	1/12	8.3
25	2	2/12	16.7
26	3	3/12	25.0
27	2	2/12	16.7
28	3	3/12	25.0
29	0	0/12	0.0
30	1	1/12	8.3
Totale	12	1	100

Classi di modalità

Quando la variabile è **continua** si fanno delle **classi di modalità**

```
3.5 2.8 4.5 6.2 4.8 2.3 2.6 3.9 4.4 5.5
5.2 6.7 3.0 2.4 5.0 3.6 2.9 1.0 2.8 3.6
```

Diagramma ramo-foglia

```
1 | 0
2 | 346889
3 | 05669
4 | 458
5 | 025
6 | 27
```

Tabella di frequenze

Tempo di studio di 20 studenti

	Frequenze		
Tempo	assolute	relative	percentuali
$1 \vdash 2$	1	0.05	5
$2 \vdash 3$	6	0.30	30
$3 \vdash 4$	5	0.25	25
$4 \vdash 5$	3	0.15	15
$5 \vdash 6$	3	0.15	15
$6 \vdash 7$	2	0.10	10
Totale	20	1.00	100

Nota: il valore 3.0 va a finire nella terza classe non nella seconda!

Altre definizioni

Scale di misura

- dati su **scala di intervallo**. Esempio: la temperatura in gradi Celsius
- dati su **scala di rapporto**. Esempio: Una distanza tra due città. Presenza di uno **zero** che significa **assenza del carattere**

Dati di stato e di flusso

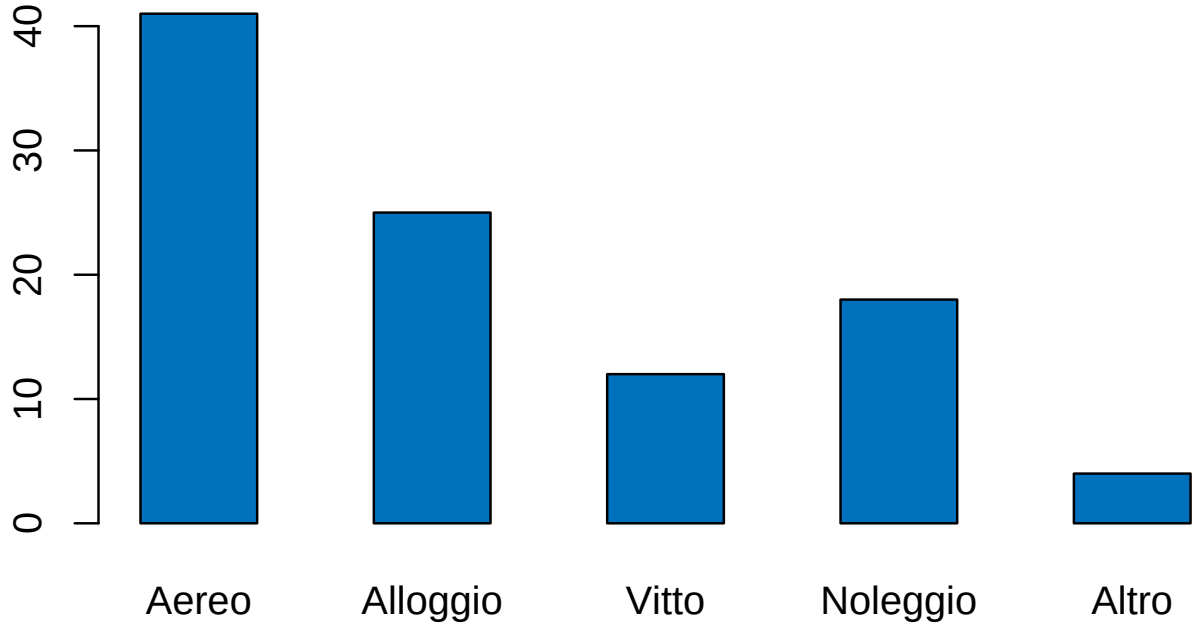
- **Dati di stato**: si possono misurare in un **istante di tempo** Esempio: la popolazioe al primo di gennaio del 2019
- **Dati di flusso**: si possono misurare solo in un **intervallo di tempo**. Esempio: il numero di nati a gennaio del 2019

Visualizzazione

Grafici per variabili categoriche

Esercizio 2.9 - Spese per trasferte (in %)

Aereo	Alloggio	Vitto	Noleggio	Altro	Totale
41	25	12	18	4	100



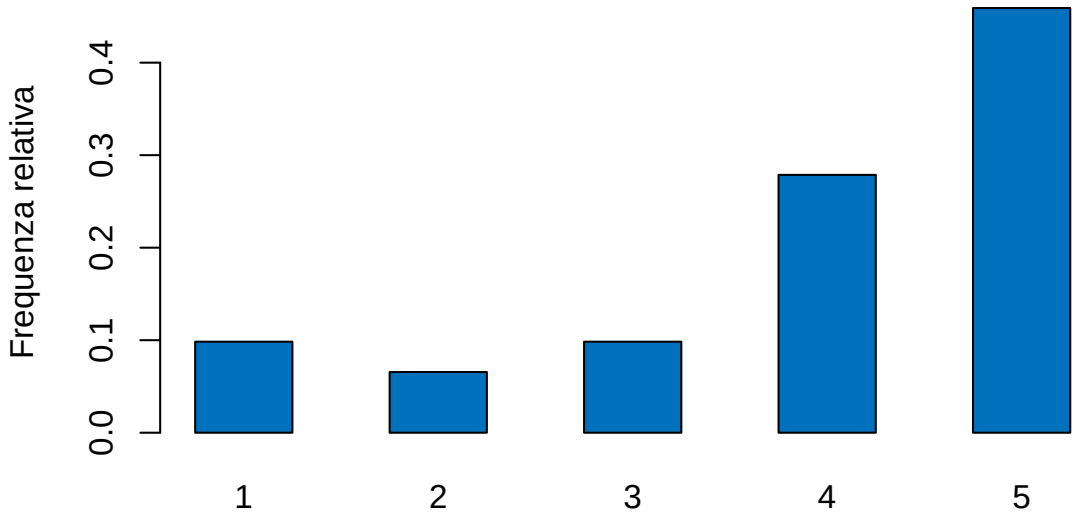
Usare i diagrammi a barre

Per favore non fate i grafici a torta!

Saltare i diagrammi di Pareto

Grafici per variabili discrete

- Usare i **diagrammi a barre**
- Rating di un prodotto su Amazon



Rappresentazione di dati continui

Si usa l'**istogramma**

- Si definiscono classi di **ampiezza uguale** che siano non sovrapposte e che contengano tutti i dati
- Si costruiscono dei rettangoli con BASE = classe e **AREA = frequenza**

NOTA: I rettangoli sono affiancati.

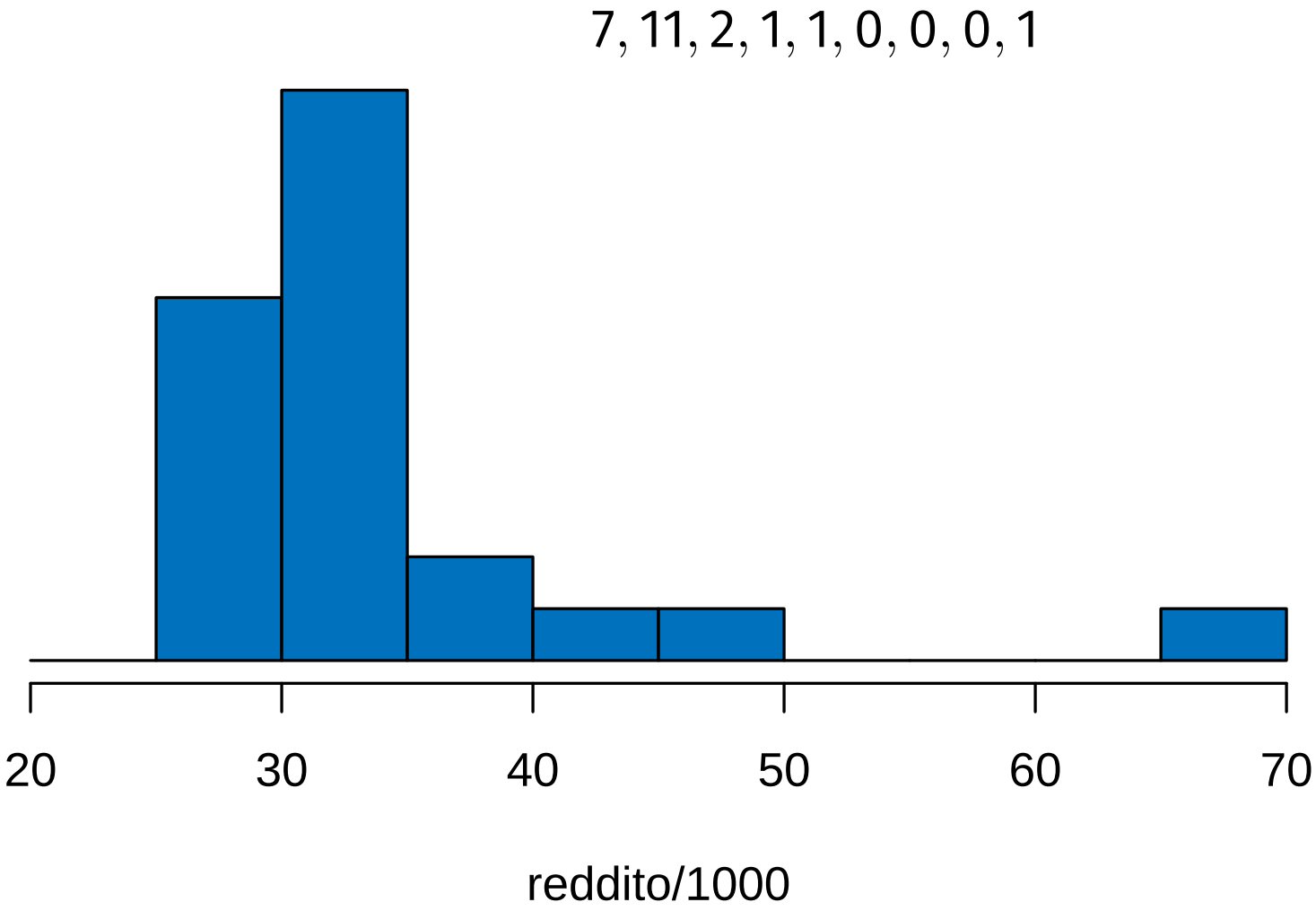
Distribuzione dei redditi

Si classificano 23 persone a seconda del reddito (in migliaia di Euro annui). Dati **Income.xlsx**

Classi	(1) Frequenza	(2) Ampiezza	(1)/(2) Densità
25 ⊣ 30	7	5	1.4
30 ⊣ 35	11	5	2.2
35 ⊣ 40	2	5	0.4
40 ⊣ 45	1	5	0.2
45 ⊣ 50	1	5	0.2
50 ⊣ 55	0	5	0.0
55 ⊣ 60	0	5	0.0
60 ⊣ 65	0	5	0.0
65 ⊣ 70	1	5	0.2
Totale	23		

Istogramma

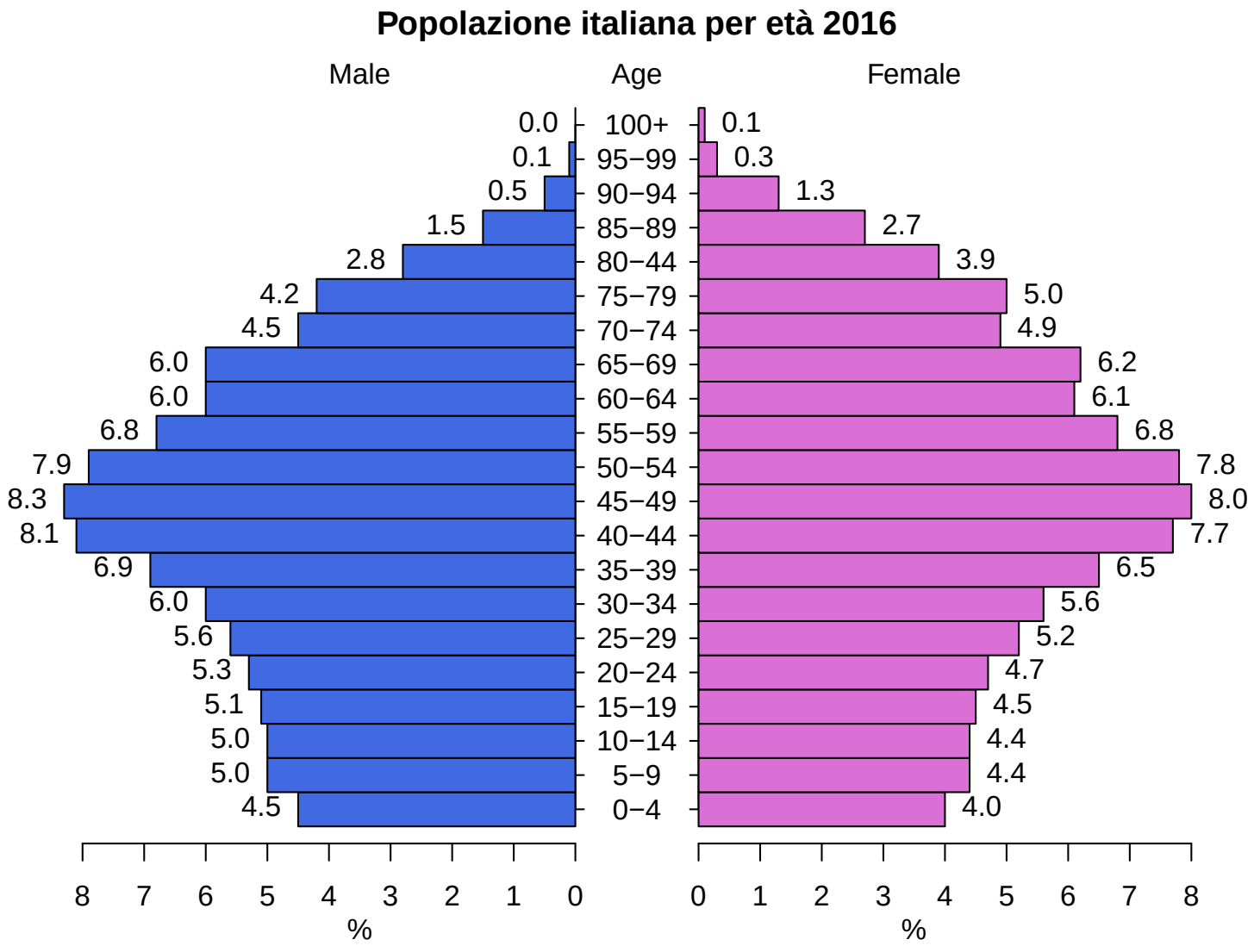
- Le altezze dei rettangoli sono le densità
- Le aree dei rettangoli sono le frequenze



Istogrammi e diagrammi a barre

- Istogrammi: per dati **continui**
- Diagrammi a barre: per dati **discreti** o **categorici**

Piramide della popolazione



Regole per la scelta del numero di classi

- Meglio valutare a occhio
- Il libro fornisce una tabella (difficile da ricordare)
- Una regola semplice: arrotonda

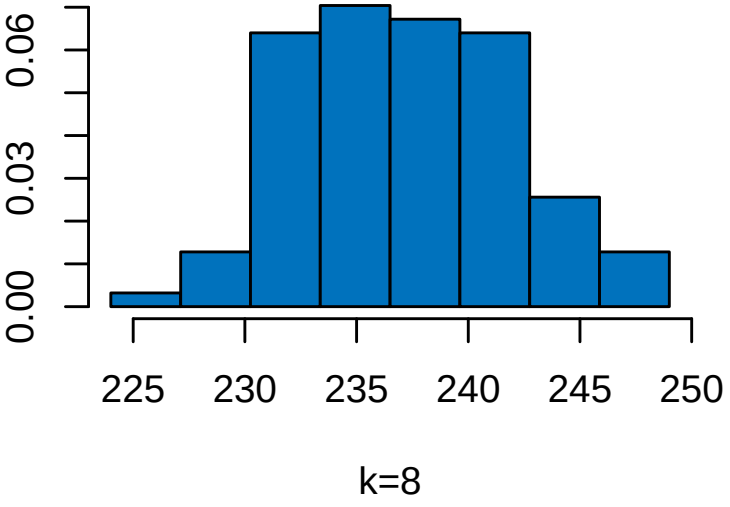
$$2\sqrt[3]{\text{n. osservazioni}}$$

Conoscendo il numero di classi k l'ampiezza di classe è

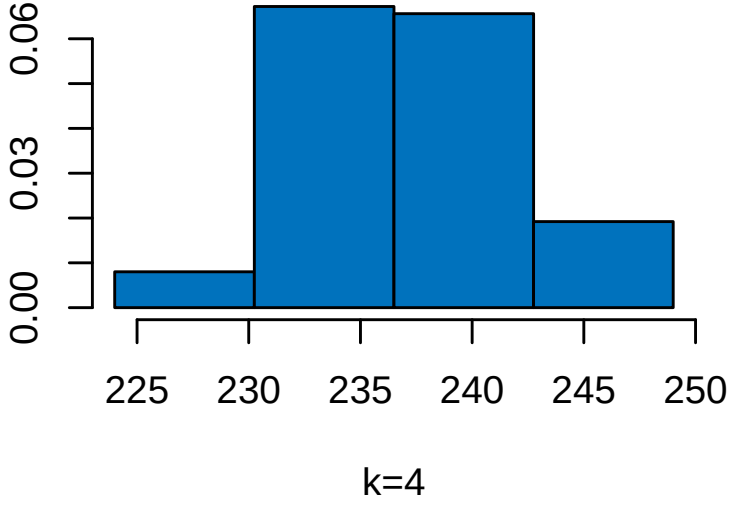
$$\text{ampiezza} = \frac{\max - \min}{k}$$

Esercizio 2.38 – Creme solari

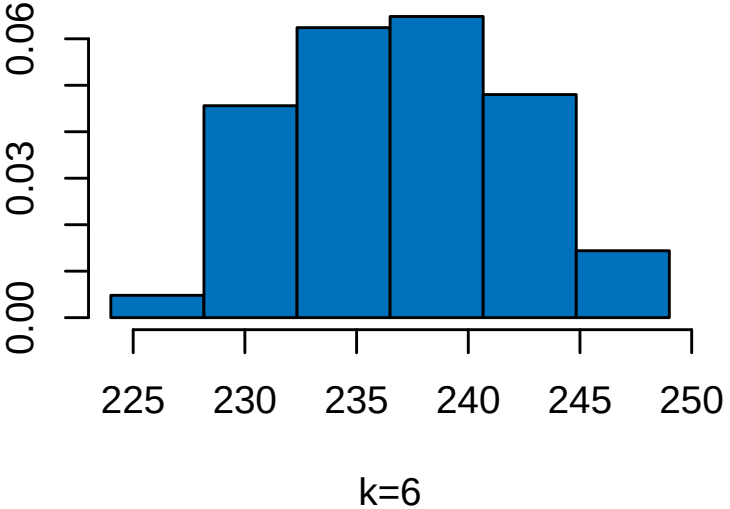
Contenuto reale di 100 flaconi da 237 ml (**Sun.xlsx**)



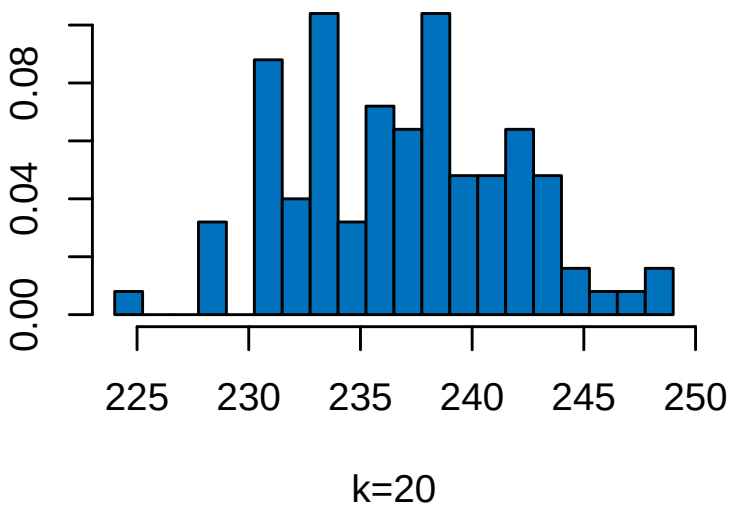
k=8



k=4



k=6



k=20

Distribuzioni di frequenze cumulate

Rating di un prodotto (numero di stelle)

Stelle	Frequenze	**Frequenze cumulate**
1	6	6
2	4	10
3	6	16
4	17	33
5	28	61
Totale	61	

Frequenze relative cumulate

Rating di un prodotto (numero di stelle)

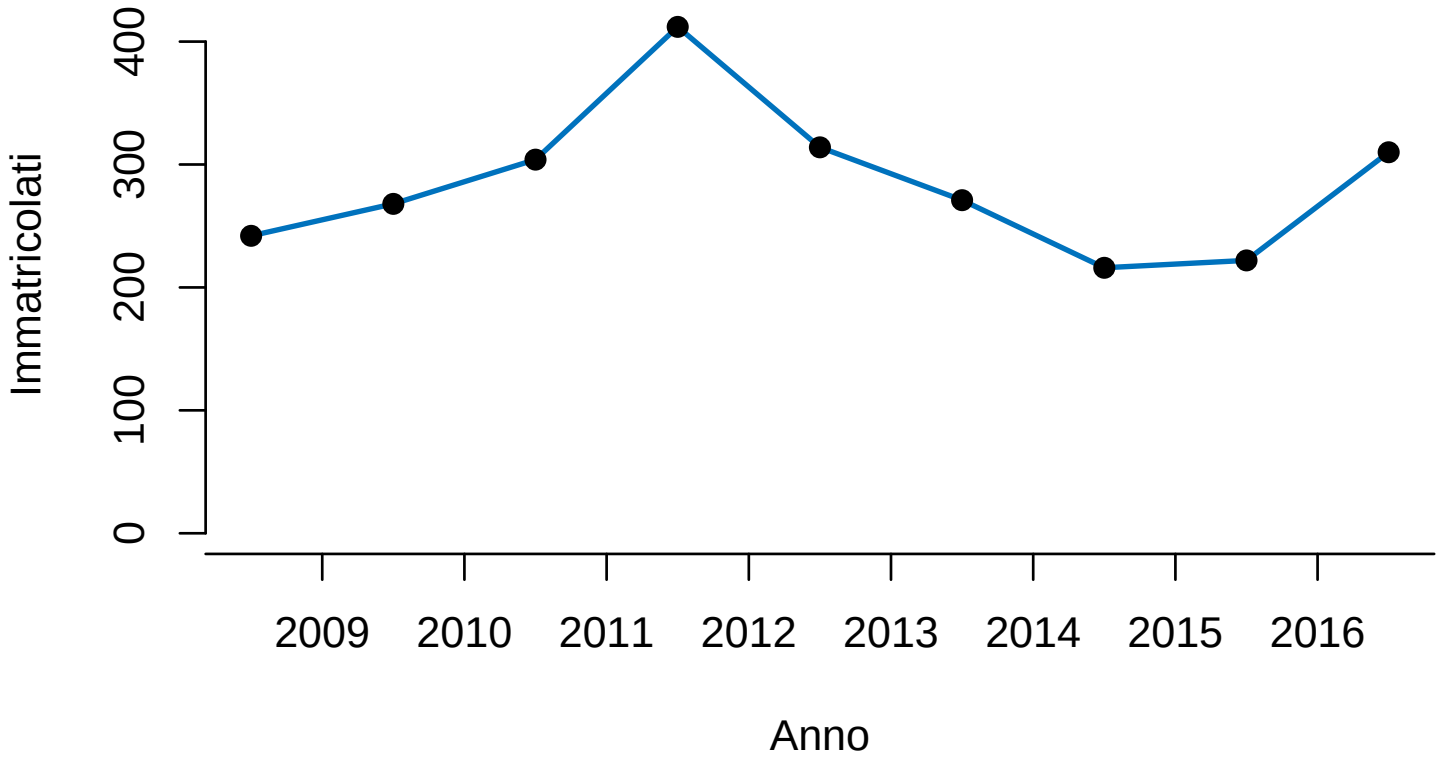
Stelle	%	% cumulate
1	9.8	9.8
2	6.6	16.4
3	9.8	26.2
4	27.9	54.1
5	45.9	100.0
Totale	100.0	

Quindi il 54.1 % dei clienti ha giudicato il prodotto meritevole di un numero di stelle sotto 5.

Grafici per serie storiche

- Le **serie storiche** sono dati osservati su **una unità** in **più tempi**
- Immatricolati al Corso di Laurea in EC di FI

08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17
242	268	304	412	314	271	216	222	310



Un caso di studio

http://local.disia.unifi.it/gmm/scuola

31 Dec - 6 Jan 2019

Average Daily Traffic

Page Views	Unique Visits	First Time Visits	Returning Visits
2.1 ▼	1.1 ▲	0.9	0.3 ▲
Prev Week	Prev Week	Prev Week	Prev Week
2.3	1	0.9	0.1

Daily Traffic Breakdown

	Page Views	Unique Visits	First Time Visits	Returning Visits
Mon	0	0	0	0
Tue	0	0	0	0
Wed	1	1	0	1
Thu	12	5	4	1
Fri	1	1	1	0
Sat	1	1	1	0
Sun	0	0	0	0
Total	**15**	**8**	**6**	**2**
Avg	2	1	1	0

Il data set

Dati di flusso: rilevazioni su 2119 giorni

Dati aggregati per settimana (297 settimane)

Settimana	Pagine	Visite	Prime visite	Ricorrenti
W18 2013	38	13	12	1
W19 2013	48	23	19	4
W20 2013	263	142	134	8
W21 2013	101	56	51	5
⋮	⋮	⋮	⋮	⋮
W50 2018	11	5	4	1
W51 2018	40	11	7	4
W52 2018	16	7	6	1
W01 2019	15	8	6	2

Unità – variabili

- Le **osservazioni** riguardano il sito web in 297 settimane
- Le **variabili** sono

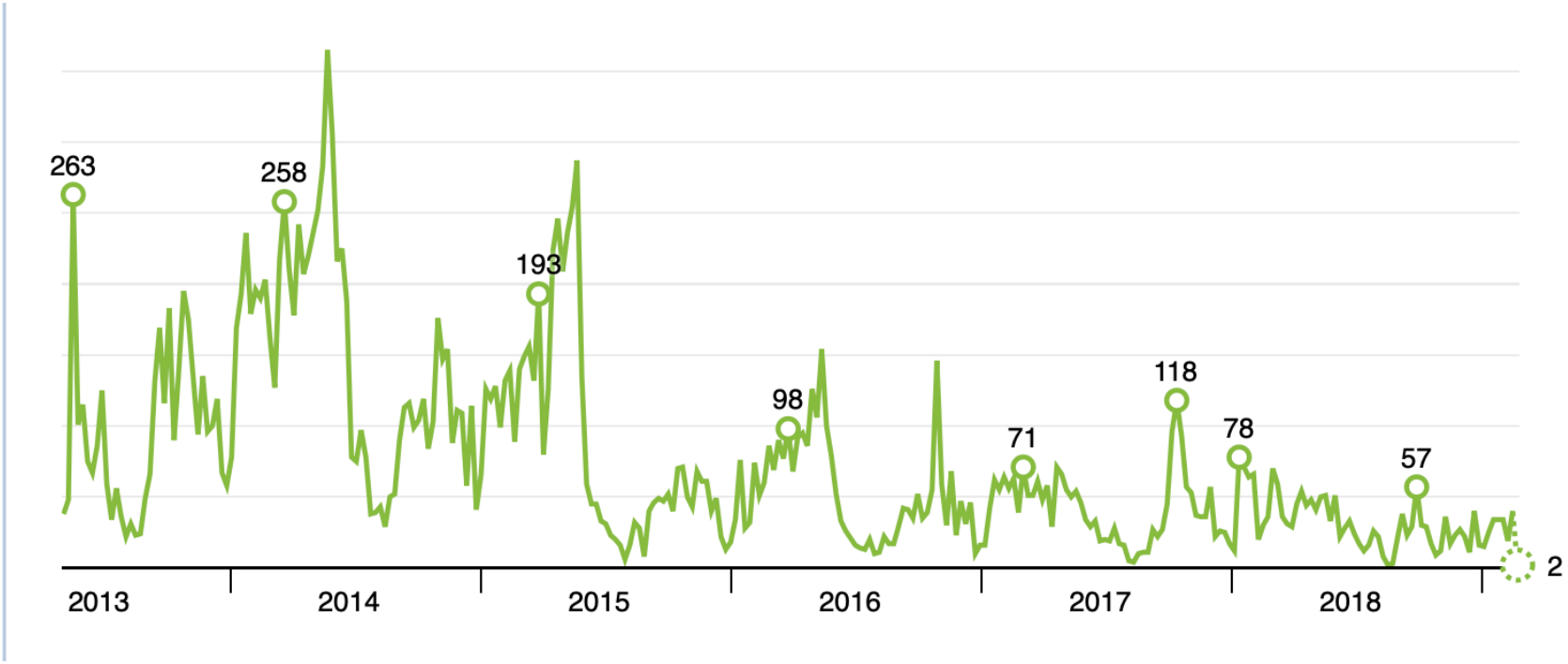
  X il n. di Pagine
  Y il n. di visite
  Z il n. di prime visite
  U il n. di visite ricorrenti

- X_t numero di pagine nella settimana t
- Relazione: $Y_t = Z_t + U_t$

Visualizzazione di una serie

I dati temporali costituiscono una **serie storica**

Pagine visitate



Variazione assoluta e relativa

Data una serie storica la variazione tra due valori a tempi diversi x_{prima} e x_{dopo} si possono misurare in due modi

- **variazione assoluta** = $x_{dopo} - x_{prima}$
- **variazione relativa** = $\dfrac{x_{dopo} - x_{prima}}{x_{prima}}$

La VR si esprime in percentuale.

Esempio

Numero di immatricolati

Anno Accademico	15-16	16-17
Immatricolati	222	310

$$VA = 310 - 222 = 88, \quad VR = \frac{88}{222} = 39.6\%$$

Rendimento di un fondo

Anno	01/01/2017	01/01/2019
Valore	15000	14800

$$VA = -200\,Euro, \quad VR = -200/15000 = -1.3\%$$

Relazioni tra variabili – Introduzione

Matrici di dati

- I dati **cross-sectional** sono ottenuti misurando **più variabili** sullo stesso insieme di unità **a un tempo fisso**.
- Un campione di voti di 10 studenti partecipanti a un test d'ingresso a Economia a FI. (Totale = 1128)

Maturità	Test	Sesso	Scuola
88	13.50	F	T
66	11.50	F	T
76	12.50	F	L
80	7.75	F	P
68	11.75	F	L
74	10.00	M	L
91	21.75	M	L
80	14.75	M	T
74	18.75	M	L
74	12.50	M	T

Variabili

Le variabili (le colonne della matrice) sono

- Voto alla Maturità
- Punteggio al test d'ingresso
- Sesso: Sesso (M,F)
- Scuola: Tipo di scuola (L = liceo, T = tecnico, P = professionale, A = altro)

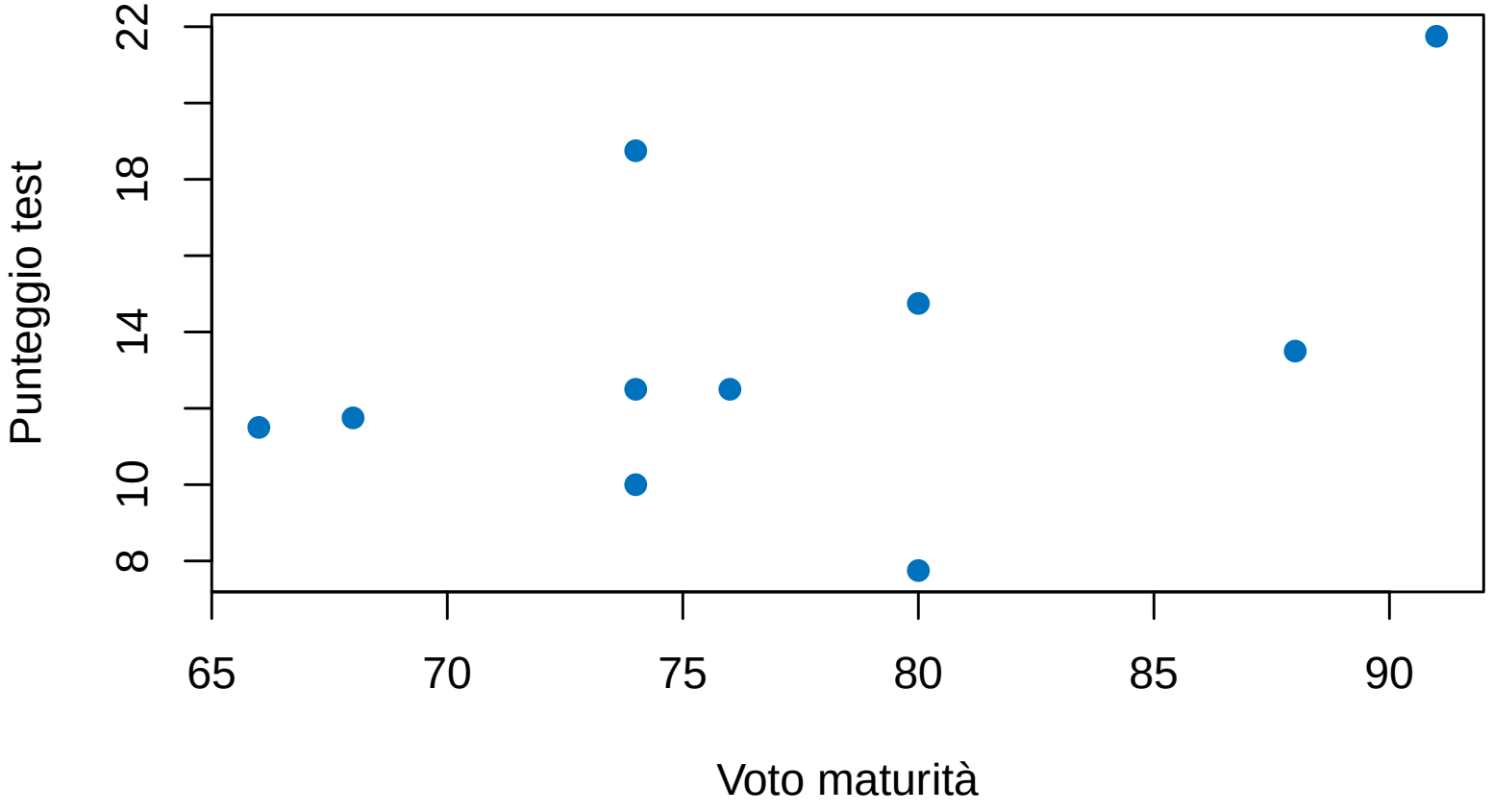
Diagramma di dispersione

Date due variabili quantitative, per esempio,

- X **Voto** alla maturità
- Y **Punteggio** al test d'ingresso

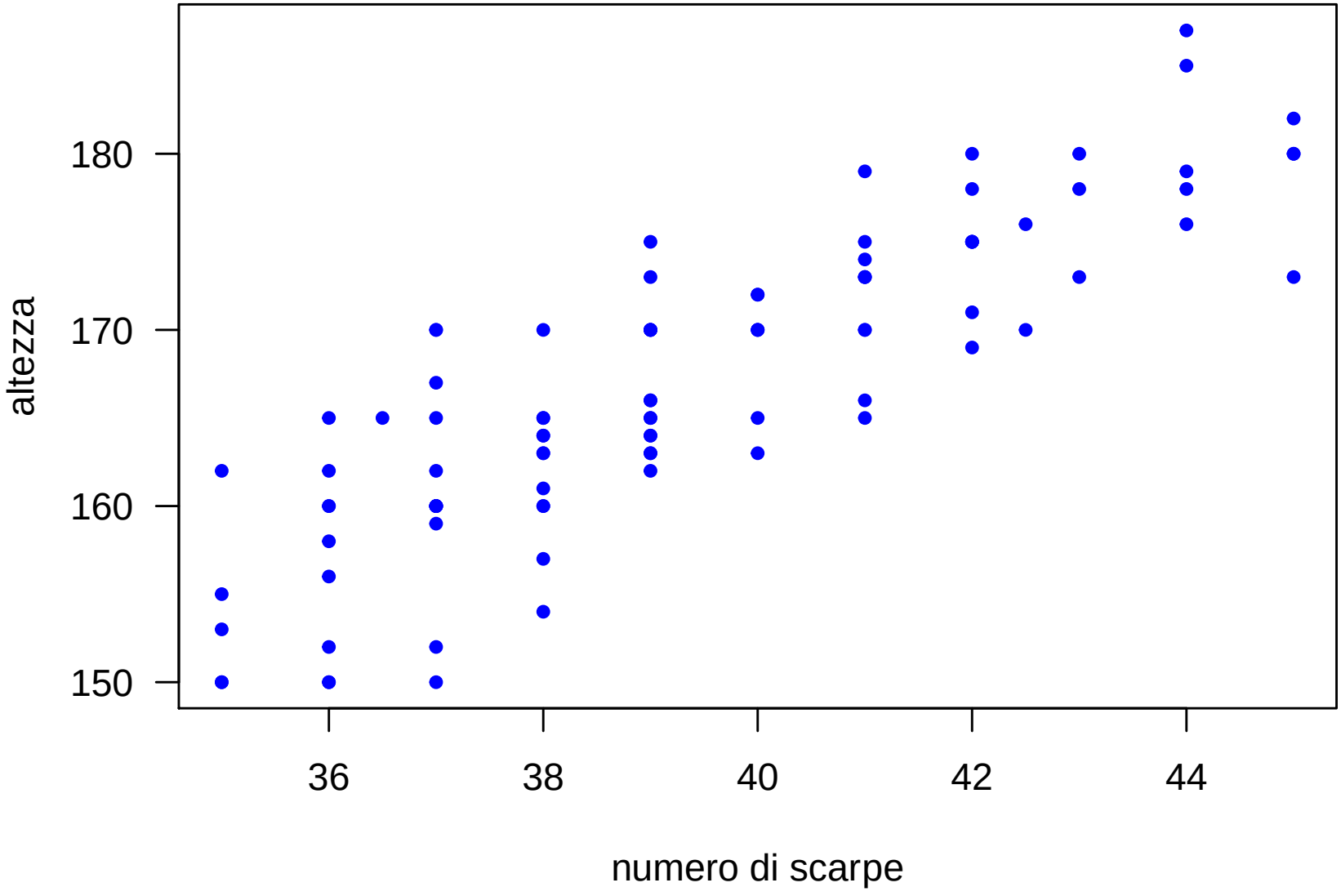
- Ci aspettiamo che all'aumentare del voto alla maturità aumenti il punteggio al test
- Un grafico cartesiano dei punti (x, y) permette di valutare l'andamento
- Si chiama comunemente **scatter**

Interpretazione



- La relazione dovrebbe essere crescente ma non è molto forte.
- La forza si deduce dall'**allineamento**

Relazione tra altezza e numero di scarpe



Su un campione di studenti di ambo i sessi

Associazione tra due variabili qualitative

Consideriamo per gli studenti presenti al test d'ingresso le variabili

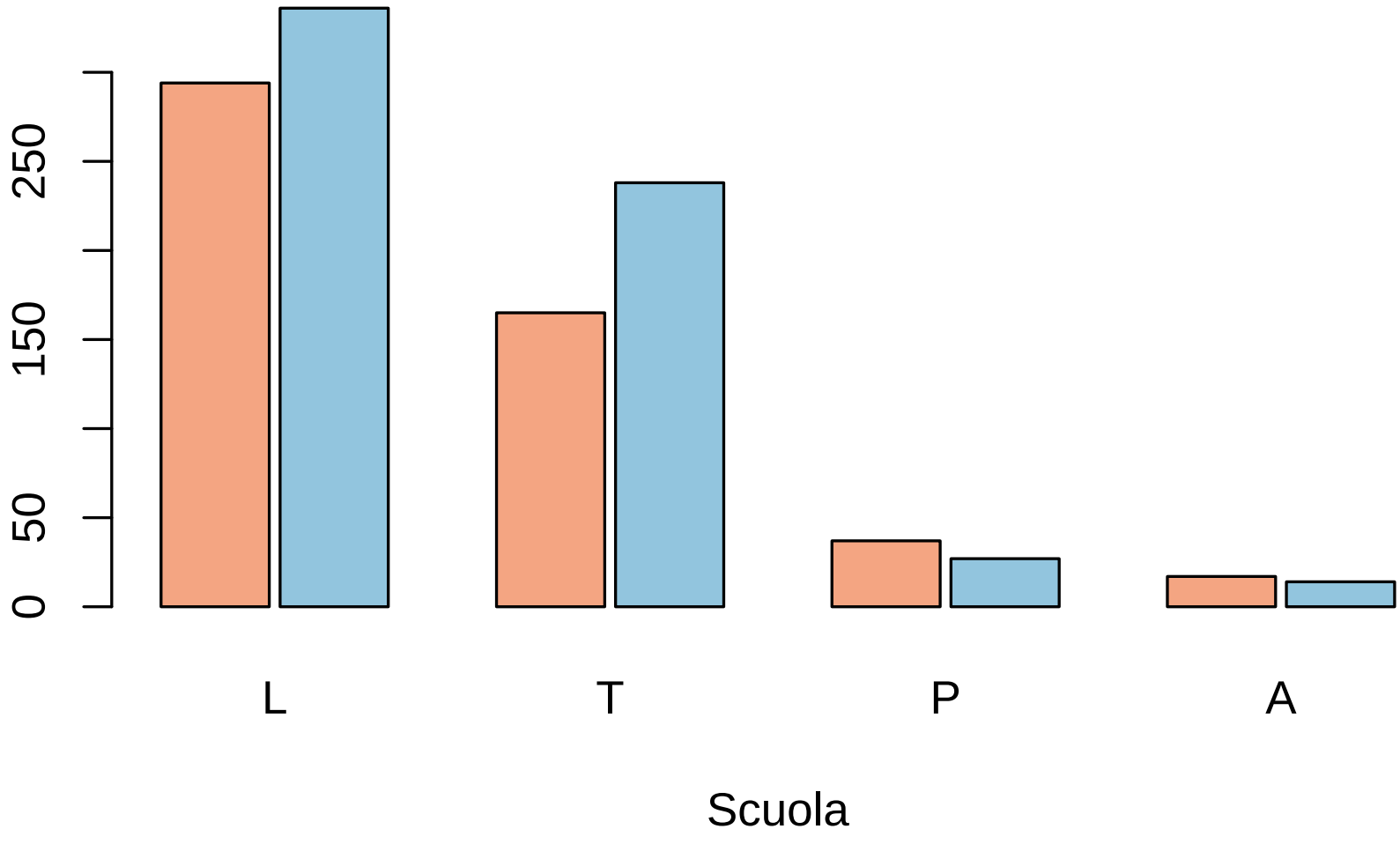
- Sesso (M,F)
- Tipo di scuola (L, T, P, A)

Lo spoglio deve essere fatto rispetto a tutte e due le variabili

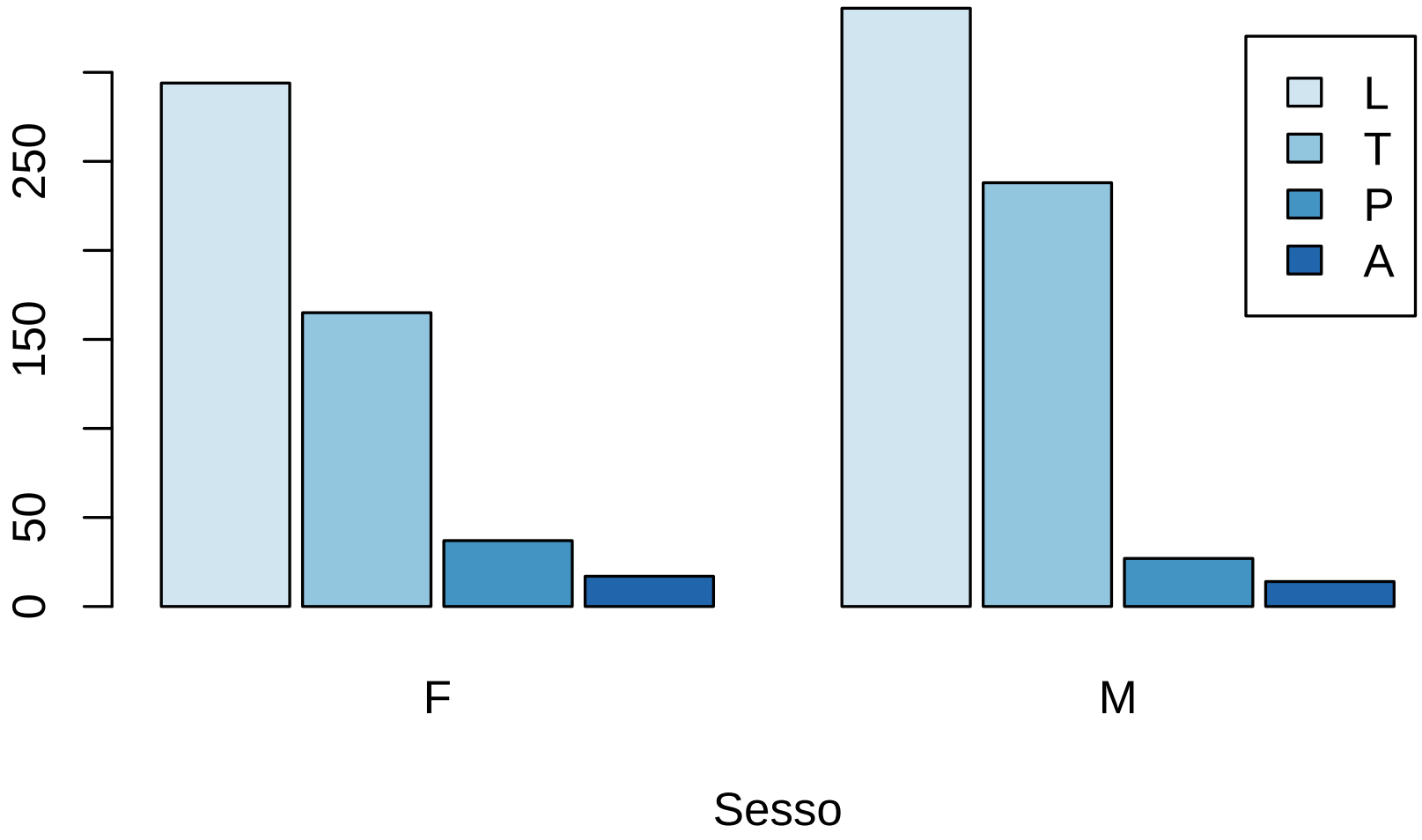
	Tipo di Scuola				
Sesso	L	T	P	A	Totale
F	294	165	37	17	513
M	336	238	27	14	615
Totale	630	403	64	31	1128

Si ottiene una **tabella a doppia entrata** 2×4.

Rappresentazione grafica (1)



Rappresentazione grafica (2)

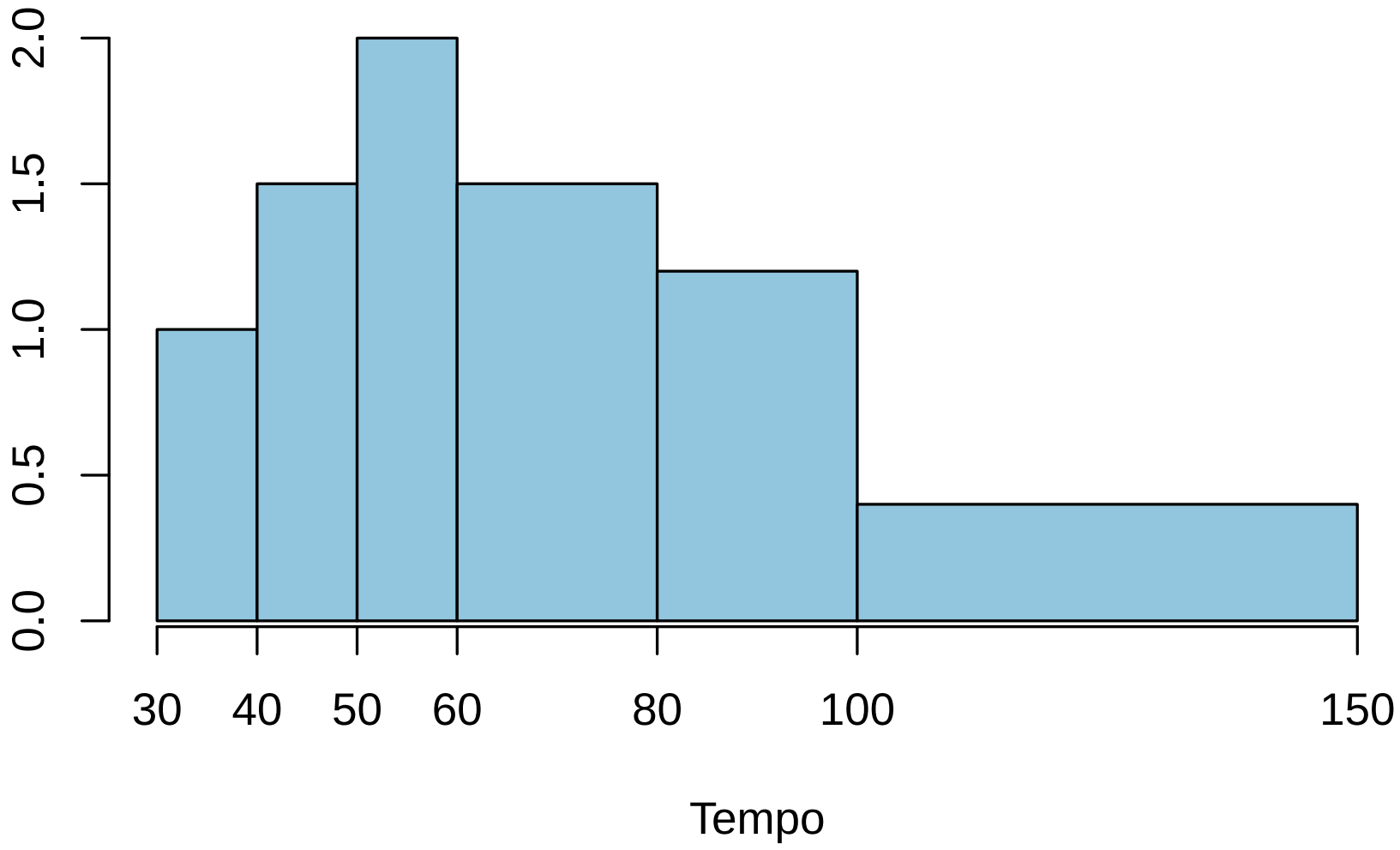


Errori nella presentazione dei dati

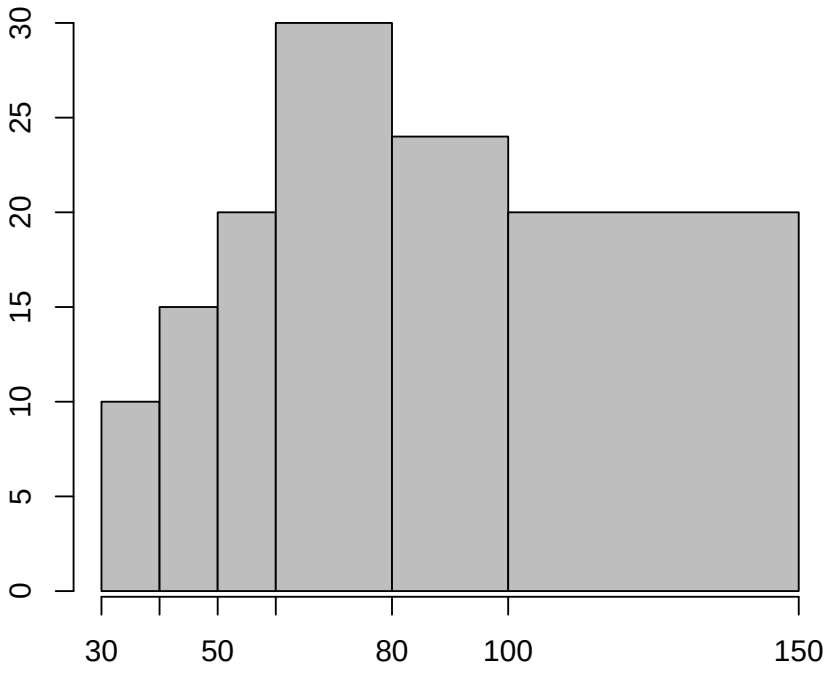
Esercizio 2.27 Costruire l'istogramma della distribuzione dei dipendenti seconda la variabile **Tempo** impiegato a finire un lavoro

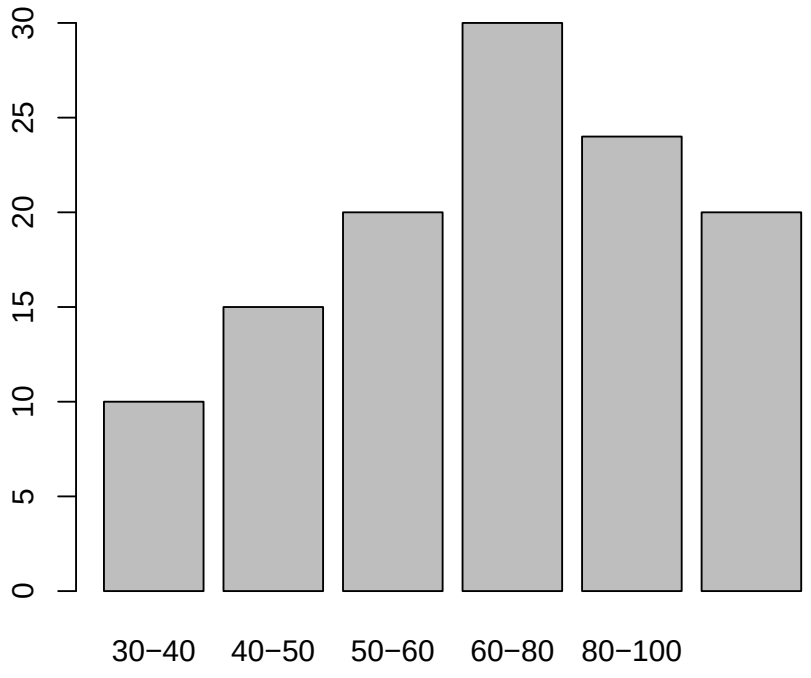
Tempo (sec)	# Dipendenti	Ampiezza	Densità
$30 \vdash 40$	10	10	1.0
$40 \vdash 50$	15	10	1.5
$50 \vdash 60$	20	10	2.0
$60 \vdash 80$	30	20	1.5
$80 \vdash 100$	24	20	1.2
$100 \vdash 150$	20	50	0.4
Totale	119		

Corretto



Sbagliati


0
5
10
15
20
25
30
30
50
80
100
150


0
5
10
15
20
25
30
30−40
40−50
50−60
60−80
80−100

Esercizi suggeriti

Esercizi

- **Libro**: 2.1, 2.9, 2.11, 2.14, 2.20, 2.32, 2.33, 2.34, 2.40, 2.41
- **Esercizi elementari** : Da 1.1 a 1.6

Leggete il **Riepilogo** e le **parole chiave**

Lasciate gli esercizi riepilogativi per il ripasso finale