

36 Esercizi del libro

Andrea Patrini (matr. 176907)

22 giugno 2025

Capitolo 1

Esercizio 1.9

Utilizzando il dataset integrato `state` (composto da variabili come `state.name`, `state.region` e `state.area`):

- Quali sono le possibili regioni degli stati? Quanti stati appartengono a ciascuna regione?
- Quali stati hanno un'area inferiore a 10,000 miglia quadrate?
- Quale stato ha il centro geografico più a sud? (Hint: usare `which.min`)

```
# Domanda a
cat("Domanda a \n")
regioni <- levels(state.region)
cat("Le possibili regioni sono:", paste(regioni, collapse = ", "), "\n\n")
conteggio_stati <- table(state.region)
cat("Numero di stati per regione:\n")
conteggio_stati
cat("\n")

# Domanda b
cat("Domanda b \n")
stati_piccoli <- state.name[state.area < 10000]
cat("Gli stati con area inferiore a 10,000 miglia quadrate sono:\n")
stati_piccoli
cat("\n")

# Domanda c
cat("Domanda c \n")
indice_sud <- which.min(state.center$y)
stato_piu_sud <- state.name[indice_sud]
cat("Lo stato con il centro geografico più a sud è:", stato_piu_sud, "\n")
cat("\n")
```

```
> Domanda a
> Le possibili regioni sono: Northeast, South, North Central, West
>
> Numero di stati per regione:
> state.region
```

```

> Northeast          South North Central          West
>           9          16          12          13
>
> Domanda b
> Gli stati con area inferiore a 10,000 miglia quadrate sono:
> [1] "Connecticut" "Delaware" "Hawaii" "Massachusetts"
> [5] "New Hampshire" "New Jersey" "Rhode Island" "Vermont"
>
> Domanda c
> Lo stato con il centro geografico più a sud è: Florida

```

Esercizio 1.10

Utilizzando il dataset `mtcars`:

- Quali auto hanno 4 marce avanti?
- Quale sottoinsieme di `mtcars` viene descritto da `mtcars[mtcars$disp > 150 & mtcars$mpg > 20,]`?
- Quali auto hanno 4 marce avanti e trasmissione manuale?
- Quali auto hanno 4 marce avanti O trasmissione manuale?
- Trovare la media dei `mpg` per le auto con 2 carburatori (`carb`)

```

# Domanda a
cat("Domanda a \n")
auto_4_marce_avanti <- mtcars[mtcars$gear == 4, ]
kable(auto_4_marce_avanti, caption = "Auto con 4 marce avanti")
cat("\n")

# Domanda b
cat("Domanda b \n")
sottoinsieme <- mtcars[mtcars$disp > 150 & mtcars$mpg > 20, ]
kable(sottoinsieme, caption = "Sottoinsieme con disp > 150 e mpg > 20")
cat("\n")

# Domanda c
cat("Domanda c \n")
auto_4_marce_manuale <- mtcars[mtcars$gear == 4 & mtcars$am == 1, ]
kable(auto_4_marce_manuale, caption = "Auto con 4 marce e trasmissione manuale")
cat("\n")

# Domanda d
cat("Domanda d \n")
auto_4_marce_o_manuale <- mtcars[mtcars$gear == 4 | mtcars$am == 1, ]
kable(auto_4_marce_o_manuale, caption = "Auto con 4 marce O trasmissione manuale")
cat("\n")

# Domanda e
cat("Domanda e \n")
media_mpg <- mean(mtcars[mtcars$carb == 2, "mpg"])
cat("La media dei mpg per le auto con 2 carburatori è:", media_mpg, "\n")

```

> Domanda a

Table 1: Auto con 4 marce avanti

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

>
> Domanda b

Table 2: Sottinsieme con disp > 150 e mpg > 20

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1

>
> Domanda c

Table 3: Auto con 4 marce e trasmissione manuale

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

>
> Domanda d

Table 4: Auto con 4 marce O trasmissione manuale

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

```
>
> Domanda e
> La media dei mpg per le auto con 2 carburatori è: 22.4
```

Esercizio 1.11

Utilizzando il dataset `mtcars`:

- Convertire la variabile `am` in `factor` con livelli "auto" e "manual"
- Quante auto ci sono per ogni tipo di trasmissione?
- Quante auto per tipo di trasmissione hanno un consumo maggiore di 25 mpg?

```
# Domanda a
# Creo una copia di mtcars per evitare problemi con altri codici
mtcars2 <- mtcars

# Domanda a:
cat("Domanda a \n")
mtcars2$am <- factor(mtcars2$am, levels = c(0, 1), labels = c("auto", "manual"))
cat("La variabile am è stata convertita in fattore:\n")
str(mtcars2$am)
cat("\n")

# Domanda b
cat("Domanda b \n")
conteggio_trasmissioni <- table(mtcars$am)
cat("\nNumero di auto per tipo di trasmissione:\n")
print(conteggio_trasmissioni)
cat("\n")
```

```
# Domanda c
cat("Domanda c \n")
auto_consumo_alto <- mtcars[mtcars$mpg > 25, ]
conteggio_consumo_alto <- table(auto_consumo_alto$am)
cat("\nNumero di auto con consumo > 25 mpg per tipo di trasmissione:\n")
print(conteggio_consumo_alto)
cat("\n")
```

```
> Domanda a
> La variabile am è stata convertita in fattore:
> Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
>
> Domanda b
>
> Numero di auto per tipo di trasmissione:
>
> 0 1
> 19 13
>
> Domanda c
>
> Numero di auto con consumo > 25 mpg per tipo di trasmissione:
>
> 1
> 6
```

Esercizio 1.12

Utilizzando il dataset `hot_dogs` dal pacchetto `fosdata`:

- Quante osservazioni e variabili ci sono? Quali tipi di variabili?
- Quali sono i tre tipi di hot dog presenti?
- Qual è il più alto contenuto di sodio (`sodium`) registrato?
- Qual è il contenuto calorico medio per gli hot dog di manzo (`Beef`)?

```
# Carica il dataset hot_dogs
data(hot_dogs)

# Domanda a
cat("Domanda a: \n")
num_osservazioni <- nrow(hot_dogs)
num_variabili <- ncol(hot_dogs)
cat("Numero di osservazioni:", num_osservazioni, "\n")
cat("Numero di variabili:", num_variabili, "\n")
cat("Tipi di variabili:")
str(hot_dogs)
cat("\n")

# Domanda b
cat("Domanda b: \n")
```

```

tipi_hot_dog <- unique(hot_dogs$type)
cat("I tre tipi di hot dog presenti sono:\n")
print(tipi_hot_dog)
cat("\n")

# Domanda c
cat("Domanda c: \n")
max_sodio <- max(hot_dogs$sodium, na.rm = TRUE)
cat("Il più alto contenuto di sodio registrato è:", max_sodio, "mg\n")
cat("\n")

# Domanda d
cat("Domanda d: \n")
hot_dog_manzo <- hot_dogs[hot_dogs$type == "Beef", ]
media_calorie <- mean(hot_dog_manzo$calories, na.rm = TRUE)
cat("Il contenuto calorico medio per gli hot dog di manzo è:", media_calorie, "calorie\n")

```

```

> Domanda a:
> Numero di osservazioni: 54
> Numero di variabili: 3
> Tipi di variabili: 'data.frame': 54 obs. of 3 variables:
> $ type : Factor w/ 3 levels "Beef","Meat",...: 1 1 1 1 1 1 1 1 1 1 ...
> $ calories: int 186 181 176 149 184 190 158 139 175 148 ...
> $ sodium : int 495 477 425 322 482 587 370 322 479 375 ...
>
> Domanda b:
> I tre tipi di hot dog presenti sono:
> [1] Beef Meat Poultry
> Levels: Beef Meat Poultry
>
> Domanda c:
> Il più alto contenuto di sodio registrato è: 645 mg
>
> Domanda d:
> Il contenuto calorico medio per gli hot dog di manzo è: 156.85 calorie

```

Esercizio 1.13

Utilizzando il dataset `DrinksWages` dal pacchetto `HistData`:

- Quante osservazioni e variabili ci sono? Quali tipi di variabili?
- La variabile `wage` contiene lo stipendio medio per ogni professione. Quale professione ha lo stipendio più basso?
- La variabile `n` contiene il numero di lavoratori intervistati per ogni professione. Calcolare il numero totale di lavoratori intervistati
- Calcolare il salario medio di tutti i lavoratori intervistati moltiplicando `wage * n` per ciascuna professione, sommando e dividendo per il numero totale dei lavoratori intervistati

```

data(DrinksWages)
# Domanda a
cat("Domanda a \n")
num_osservazioni <- nrow(DrinksWages)
num_variabili <- ncol(DrinksWages)
cat("Numero di osservazioni:", num_osservazioni, "\n")
cat("Numero di variabili:", num_variabili, "\n")
cat("Tipi di variabili:\n")
str(DrinksWages)
cat("\n")

# Domanda b
cat("Domanda b \n")
professione_stipendio_min <- as.character(DrinksWages[which.min(DrinksWages$wage), "trade"])
cat("La professione con lo stipendio più basso è:", professione_stipendio_min, "\n")
cat("\n")

# Domanda c
cat("Domanda c \n")
totale_lavoratori <- sum(DrinksWages$n)
cat("Il numero totale di lavoratori intervistati è:", totale_lavoratori, "\n")
cat("\n")

# Domanda d
cat("Domanda d \n")
salario_totale <- sum(DrinksWages$wage * DrinksWages$n)
salario_medio <- salario_totale / totale_lavoratori
cat("Il salario medio di tutti i lavoratori intervistati è:", salario_medio, "\n")
cat("\n")

```

```

> Domanda a
> Numero di osservazioni: 70
> Numero di variabili: 6
> Tipi di variabili:
> 'data.frame': 70 obs. of 6 variables:
> $ class : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
> $ trade : Factor w/ 70 levels "baker","barman",...: 38 10 25 55 36 44 68 34 14 11 ...
> $ sober : int 1 1 2 1 2 9 8 3 0 12 ...
> $ drinks: int 1 10 1 5 0 8 2 5 7 23 ...
> $ wage : num 24 18.4 21.5 21.2 19 ...
> $ n : int 2 11 3 6 2 17 10 8 7 35 ...
>
> Domanda b
> La professione con lo stipendio più basso è: factory worker
>
> Domanda c
> Il numero totale di lavoratori intervistati è: 604
>
> Domanda d
> Il salario medio di tutti i lavoratori intervistati è: 24.59782

```

Esercizio 1.15

Utilizzando il dataset `bechdel` dal pacchetto `fosdata`:

- Quanti film superano il test Bechdel?
- Quale percentuale di film supera il test?
- Crea una `table` dei film suddivisi per anno
- Quale anno ha più film?
- Quanti valori diversi contiene `clean_test`?
- Creare un data frame con film che passano il test Bechdel
- Creare un data frame che contenga tutte le osservazioni che non presentano valori mancanti nella variabile `domgross`

```
# Domanda a
cat("Domanda a \n")
film_superano_test <- sum(bechdel$clean_test == "ok")
cat("Numero di film che superano il test Bechdel:", film_superano_test, "\n")
cat("\n")

# Domanda b
cat("Domanda b \n")
percentuale_superano_test <- mean(bechdel$clean_test == "ok") * 100
cat("Percentuale di film che superano il test Bechdel:", round(percentuale_superano_test, 2), "%\n")
cat("\n")

# Domanda c
cat("Domanda c \n")
tabella_film_per_anno <- table(bechdel$year)
cat("Tabella dei film suddivisi per anno:\n")
print(tabella_film_per_anno)
cat("\n")

# Domanda d
cat("Domanda d \n")
anno_piu_film <- names(which.max(tabella_film_per_anno))
cat("L'anno con più film è:", anno_piu_film, "\n")
cat("\n")

# Domanda e
cat("Domanda e \n")
valori_clean_test <- unique(bechdel$clean_test)
num_valori_clean_test <- length(valori_clean_test)
cat("Numero di valori diversi in clean_test:", num_valori_clean_test, "\n")
cat("\n")

# Domanda f
cat("Domanda f \n")
film_passano_test <- bechdel[bechdel$clean_test == "ok", ]
cat("Data frame con i film che passano il test Bechdel:\n")
print(head(film_passano_test))
```

```

cat("\n")

# Domanda g
cat("Domanda g \n")
film_senza_mancanti <- bechdel[!is.na(bechdel$domgross), ]
cat("Data frame con osservazioni senza valori mancanti in domgross:\n")
print(head(film_senza_mancanti)) # Mostra solo le prime righe per brevità
cat("\n")

> Domanda a
> Numero di film che superano il test Bechdel: 803
>
> Domanda b
> Percentuale di film che superano il test Bechdel: 44.76 %
>
> Domanda c
> Tabella dei film suddivisi per anno:
>
> 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985
>   1    5    3    5    7    5    8    7    8    5   14    9   14    5   16   10
> 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001
>   10   14   19   14   15   13   20   16   26   36   42   51   62   56   63   64
> 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
>   80   64   81  100   90   73  101  124  129  124   86   99
>
> Domanda d
> L'anno con più film è: 2010
>
> Domanda e
> Numero di valori diversi in clean_test: 5
>
> Domanda f
> Data frame con i film che passano il test Bechdel:
>   year      imdb      title      test clean_test binary  budget
> 2  2012 tt1343727      Dredd 3D ok-disagree      ok  PASS 4.5e+07
> 8  2013 tt2194499    About Time ok-disagree      ok  PASS 1.2e+07
> 9  2013 tt1814621      Admission      ok      ok  PASS 1.3e+07
> 11 2013 tt1800241  American Hustle ok-disagree      ok  PASS 4.0e+07
> 12 2013 tt1322269 August: Osage County      ok      ok  PASS 2.5e+07
> 13 2013 tt1559547  Beautiful Creatures      ok      ok  PASS 5.0e+07
>   domgross  intgross      code budget_2013 domgross_2013  intgross_2013
> 2   13414714  40868994 2012PASS    45658735    13611086    41467257
> 8   15323921  87324746 2013PASS    12000000    15323921    87324746
> 9   18007317  18007317 2013PASS    13000000    18007317    18007317
> 11 148430908 249484909 2013PASS    40000000    148430908    249484909
> 12  37304874  50304874 2013PASS    25000000    37304874    50304874
> 13 19452138  55940671 2013PASS    50000000    19452138    55940671
>   period_code decade_code
> 2             1           1
> 8             1           1
> 9             1           1
> 11            1           1
> 12            1           1
> 13            1           1

```

```

>
> Domanda g
> Data frame con osservazioni senza valori mancanti in domgross:
>   year      imdb      title      test clean_test binary  budget
> 1 2013 tt1711425    21 &amp; Over    notalk    notalk  FAIL 1.30e+07
> 2 2012 tt1343727      Dredd 3D    ok-disagree      ok  PASS 4.50e+07
> 3 2013 tt2024544 12 Years a Slave notalk-disagree    notalk  FAIL 2.00e+07
> 4 2013 tt1272878      2 Guns    notalk    notalk  FAIL 6.10e+07
> 5 2013 tt0453562      42      men      men  FAIL 4.00e+07
> 6 2013 tt1335975    47 Ronin      men      men  FAIL 2.25e+08
>   domgross  intgross      code budget_2013 domgross_2013 intgross_2013
> 1 25682380 42195766 2013FAIL    13000000    25682380    42195766
> 2 13414714 40868994 2012PASS    45658735    13611086    41467257
> 3 53107035 158607035 2013FAIL    20000000    53107035    158607035
> 4 75612460 132493015 2013FAIL    61000000    75612460    132493015
> 5 95020213 95020213 2013FAIL    40000000    95020213    95020213
> 6 38362475 145803842 2013FAIL    225000000    38362475    145803842
>   period_code decade_code
> 1           1           1
> 2           1           1
> 3           1           1
> 4           1           1
> 5           1           1
> 6           1           1

```

Capitolo 2

Esercizio 2.2

Considera un esperimento in cui si lanciano due dadi e si sottrae il valore minore dal maggiore (ottenendo 0 in caso di pareggio):

- a. Qual è la probabilità di ottenere 0? Numero totale di esiti possibili:

$$6 \times 6 = 36$$

Casi favorevoli (dadi uguali):

$$(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6) \Rightarrow 6 \text{ casi}$$

Probabilità teorica:

$$P(\text{Differenza} = 0) = \frac{6}{36} = \frac{1}{6} \approx 0.1667$$

- b. Qual è la probabilità di ottenere 4?

Casi favorevoli:

$$(1, 5), (5, 1), (2, 6), (6, 2) \Rightarrow 4 \text{ casi}$$

Probabilità teorica:

$$P(\text{Differenza} = 4) = \frac{4}{36} = \frac{1}{9} \approx 0.1111$$

```

simulazioni <- 10000

dado1 <- sample(1:6, simulazioni, replace = TRUE)
dado2 <- sample(1:6, simulazioni, replace = TRUE)

differenze <- abs(dado1 - dado2)

# a.
prob_0 <- mean(differenze == 0)
cat("Probabilità simulata differenza 0:", round(prob_0, 4), "\n")

# b.
prob_4 <- mean(differenze == 4)
cat("Probabilità simulata differenza 4:", round(prob_4, 4), "\n")

> Probabilità simulata differenza 0: 0.1627
> Probabilità simulata differenza 4: 0.1096

```

Esercizio 2.9

Stimare la probabilità di ottenere esattamente 3 teste quando si lanciano 7 monete.

La probabilità di ottenere esattamente 3 teste in 7 lanci di moneta è data dalla **formula binomiale**:

$$P(X = 3) = \underbrace{\binom{7}{3}}_{\text{Combinazioni}} \cdot \underbrace{(0.5)^3}_{\text{Teste}} \cdot \underbrace{(0.5)^4}_{\text{Croci}}$$

Dove:

1. $\binom{7}{3}$ = Numero di modi per scegliere 3 successi (teste) su 7 prove:

$$\binom{7}{3} = \frac{7!}{3!(7-3)!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$$

2. $(0.5)^3$ = Probabilità di ottenere 3 teste
3. $(0.5)^4$ = Probabilità di ottenere 4 croci

Calcolo finale:

$$P(X = 3) = 35 \cdot (0.5)^7 = 35 \cdot \frac{1}{128} = \frac{35}{128} \approx 0.2734$$

```

simulazioni <- 10000

# Genera 7 lanci di moneta per ogni simulazione (1 = Testa, 0 = Croce)
lanci <- replicate(simulazioni, {
  sum(sample(0:1, 7, replace = TRUE))
})

prob_simulata <- mean(lanci == 3)

cat("Probabilità teorica:", 35/128, "~ (circa)", round(35/128, 4), "\n")
cat("Probabilità simulata:", prob_simulata, "~ (circa)", round(prob_simulata, 4))

```

> Probabilità teorica: 0.2734375 ~ (circa) 0.2734
> Probabilità simulata: 0.2785 ~ (circa) 0.2785

Esercizio 2.17

Un mazzo standard di carte ha 52 carte (4 per ogni valore: 2,3,4,5,6,7,8,9,10,J,Q,K,A). Nel blackjack: il giocatore pesca due carte e somma il loro valore. L'asso vale 11 (o 1), mentre K, Q, J valgono 10.

a. Blackjack significa pescare un asso e una carta che vale 10. Qual'è la probabilità di avere un blackjack?

- **Carte favorevoli:**

- Assi: 4
- Carte da 10: 16 (4 dieci, 4 J, 4 Q, 4 K)

$$\text{Combinazioni favorevoli} = 4 \times 16 = 64$$

Combinazioni totali di 2 carte:

$$\binom{52}{2} = \frac{52 \times 51}{2} = 1326$$

Probabilità:

$$P(\text{Blackjack}) = \frac{64}{1326} = \frac{32}{663} \approx 0.0483 \quad (4.83\%)$$

b. Calcolare la probabilità di ottenere una somma di 19 (usando l'asso come 11).

- **Combinazioni possibili:**

- **Asso (11) + otto:**

$$4 (\text{Assi}) \times 4 (8) = 16$$

- **nove + dieci:**

$$4 (9) \times 16 (10/J/Q/K) = 64$$

- **Probabilità:**

$$\text{Combinazioni totali} = 16 + 64 = 80$$

$$P(\text{Somma } 19) = \frac{80}{1326} \approx 0.0603$$

```

simulazioni <- 10000

valori <- rep(c(2:10, "J", "Q", "K", "A"), each = 4)

# a. Blackjack (Asso + carta da 10)
blackjack <- replicate(simulazioni, {
  mano <- sample(valori, 2)
  ("A" %in% mano) && (sum(mano %in% c(10, "J", "Q", "K"))) == 1
})
prob_a <- mean(blackjack)

# b. Somma 19 (Asso = 11)
somma19 <- replicate(simulazioni, {
  mano <- sample(valori, 2)
  # Converti le carte in valori numerici
  valori_numerici <- sapply(mano, function(carta) {
    if (carta == "A") 11
    else if (carta %in% c("J", "Q", "K")) 10
    else as.numeric(carta)
  })
  sum(valori_numerici) == 19
})
prob_b <- mean(somma19)

cat("a. Probabilità simulata Blackjack:", round(prob_a, 4), "\n")
cat("b. Probabilità simulata somma 19:", round(prob_b, 4))

```

```

> a. Probabilità simulata Blackjack: 0.0485
> b. Probabilità simulata somma 19: 0.0582

```

Esercizio 2.20

Si lanciano due dadi:

- a. Qual è la probabilità che la somma sia esattamente 10?

Coppie favorevoli:

(4, 6), (5, 5), (6, 4) \Rightarrow 3 casi

$$P(\text{Somma} = 10) = \frac{3}{36} = \frac{1}{12} \approx 0.0833$$

- b. Qual è la probabilità che la somma sia almeno 10?

Coppie favorevoli:

(4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6) \Rightarrow 6 casi

$$P(\text{Somma} \geq 10) = \frac{6}{36} = \frac{1}{6} \approx 0.1667$$

- c. Qual è la probabilità che la somma sia esattamente 10, sapendo che la somma varrà almeno 10?
 $P(\text{Somma} = 10 \mid \text{Somma} \geq 10)$:

$$P = \frac{P(\text{Somma} = 10)}{P(\text{Somma} \geq 10)} = \frac{3/36}{6/36} = \frac{1}{2} = 0.5$$

```

simulazioni <- 10000

dado1 <- sample(1:6, simulazioni, replace = TRUE)
dado2 <- sample(1:6, simulazioni, replace = TRUE)
somme <- dado1 + dado2

# a.
prob_a <- mean(somme == 10)

# b.
prob_b <- mean(somme >= 10)

# c.
prob_c <- mean(somme == 10) / mean(somme >= 10)

cat("a. Probabilità teorica (10):", round(1/12, 4), "\n")
cat("a. Probabilità simulata (10):", round(prob_a, 4), "\n\n")

cat("b. Probabilità teorica (>=10):", round(1/6, 4), "\n")
cat("b. Probabilità simulata (>=10):", round(prob_b, 4), "\n\n")

cat("c. Probabilità teorica (10 | >=10):", 0.5, "\n")
cat("c. Probabilità simulata (10 | >=10):", round(prob_c, 4))

```

```

> a. Probabilità teorica (10): 0.0833
> a. Probabilità simulata (10): 0.0848
>
> b. Probabilità teorica (>=10): 0.1667
> b. Probabilità simulata (>=10): 0.1666
>
> c. Probabilità teorica (10 | >=10): 0.5
> c. Probabilità simulata (10 | >=10): 0.509

```

Esercizio 2.32

Si hanno 10 scatole numerate da 0 a 9. La scatola i -esima contiene i biglie rosse e $9-i$ biglie blu. Esperimento: Si sceglie una scatola a caso, si estraggono (con reimmissione) 3 biglie rosse consecutive

a. Qual è la probabilità che una quarta estrazione dalla stessa scatola sia rossa? La probabilità condizionata è data da:

$$P(\text{Rossa} \mid 3 \text{ Rosse}) = \frac{\sum_{i=0}^9 \left(\frac{i}{9}\right)^4}{\sum_{i=0}^9 \left(\frac{i}{9}\right)^3} = \frac{15333}{18225} \approx 0.8413 \quad (84.13\%)$$

b. Qual è la probabilità di aver scelto la scatola 9?

$$P(\text{Scatola } 9 \mid 3 \text{ Rosse}) = \frac{\left(\frac{9}{9}\right)^3}{\sum_{i=0}^9 \left(\frac{i}{9}\right)^3} = \frac{729}{2025} = \frac{9}{25} = 0.36 \quad (36\%)$$

Esercizio 2.34

Quanti modi ci sono di ottenere 4 teste lanciando 10 monete, assumendo che la quarta testa sia uscita al decimo lancio?

Per ottenere 4 teste in 10 lanci con la **4^a testa al decimo lancio**:

1. Nei primi 9 lanci devono esserci **esattamente 3 teste**
2. Il decimo lancio deve essere necessariamente testa

$$\text{Combinazioni} = \binom{9}{3} \times \frac{9!}{3!6!} = 84$$

```
simulazioni <- 10000

risultati <- replicate(simulazioni, {
  primi_9 <- sample(c("T", "C"), 9, replace = TRUE)
  sum(primi_9 == "T") == 3
})

frequenza_relativa <- mean(risultati)

cat("Frequenza relativa (verifica):", frequenza_relativa, "\n")
cat("Proporzione attesa (9C3/2^9):", choose(9, 3) / (2^9), "\n")
```

```
> Frequenza relativa (verifica): 0.1609
> Proporzione attesa (9C3/2^9): 0.1640625
```

Capitolo 3

Esercizio 3.1

Sia X una variabile aleatoria discreta con funzione di massa di probabilità:

$$p(x) = \begin{cases} \frac{1}{4} & \text{se } x = 0 \\ \frac{1}{2} & \text{se } x = 1 \\ \frac{1}{8} & \text{se } x = 2 \\ \frac{1}{8} & \text{se } x = 3 \end{cases}$$

- a. Verificare che p sia una funzione di massa di probabilità valida. Una funzione di massa di probabilità (PMF) è valida se:

1. $p(x) \geq 0$ per tutti gli x .
2. La somma di tutte le probabilità è 1.
3. **Non-negatività**: test positivo
4. **Somma delle probabilità**:

$$\frac{1}{4} + \frac{1}{2} + \frac{1}{8} + \frac{1}{8} = \frac{2+4+1+1}{8} = \frac{8}{8} = 1$$

Conclusione: $p(x)$ è una PMF valida.

- b. Calcolare $P(X \geq 2)$.

$$P(X \geq 2) = P(X = 2) + P(X = 3) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4} = 0.25$$

c. Calcolare $P(X \geq 2 \mid X \geq 1)$. Usiamo la formula della probabilità condizionata:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Dove:

- $A = X \geq 2$
- $B = X \geq 1$

Poiché $X \geq 2 \subseteq X \geq 1$, si ha $A \cap B = A$.

Calcoliamo:

1. $P(B) = P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{1}{4} = \frac{3}{4}$

2. $P(A \cap B) = P(X \geq 2) = \frac{1}{4}$

$$P(X \geq 2 \mid X \geq 1) = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \approx 0.3333$$

d. Calcolare $P(X \geq 2 \cup X \geq 1)$.

$$P(X \geq 2 \cup X \geq 1) = P(X \geq 1) = \frac{3}{4} = 0.75$$

Esercizio 3.9

Si scelga un numero intero tra 0 e 999 con tutti i numeri ugualmente probabili. Qual è il numero atteso di cifre nel numero scelto?

Il numero atteso di cifre è calcolato come:

$$E[\text{Cifre}] = \sum (\text{Cifre} \times P(\text{Cifre}))$$

- **1 cifra:** 0-9 (10 numeri)
- **2 cifre:** 10-99 (90 numeri)
- **3 cifre:** 100-999 (900 numeri)

Probabilità:

$$P(1 \text{ cifra}) = \frac{10}{1000} = 0.01$$

$$P(2 \text{ cifre}) = \frac{90}{1000} = 0.09$$

$$P(3 \text{ cifre}) = \frac{900}{1000} = 0.90$$

Calcolo valore atteso:

$$E[\text{Cifre}] = (1 \times 0.01) + (2 \times 0.09) + (3 \times 0.90) = 2.89$$

Esercizio 3.18

Steph Curry ha una precisione del 91% nei tiri liberi. Decide di tirare fino al primo errore. Qual è la probabilità che effettui esattamente 20 tiri liberi (incluso quello sbagliato)?

Si usa la **distribuzione geometrica**:

$$P(X = k) = p^{k-1} \cdot (1 - p)$$
$$P(X = 20) = (0.91)^{19} \cdot 0.09 = 0.015$$

Esercizio 3.22

Siano X e Y variabili aleatorie con $E[X] = 2$ e $E[Y] = 3$.

a. Calcolare $E[4X + 5Y]$

Si usa la **linearità del valore atteso**:

$$E[4X + 5Y] = 4E[X] + 5E[Y] = 4 \cdot 2 + 5 \cdot 3 = 8 + 15 = 23$$

b. Calcolare $E[4X - 5Y + 2]$.

$$E[4X - 5Y + 2] = 4E[X] - 5E[Y] + 2 = 4 \cdot 2 - 5 \cdot 3 + 2 = 8 - 15 + 2 = -5$$

Esercizio 3.28

In un esperimento con due lanci di una moneta equilibrata dove abbiamo che:

- $X = 1$ se il primo lancio è testa, 0 altrimenti
- $Y = 1$ se il secondo lancio è testa, 0 altrimenti
- $Z = 1$ se entrambi i lanci sono uguali, 0 altrimenti

Dimostrare che:

a. X e Y sono indipendenti

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y$$
$$P(X = 1) = \frac{1}{2}$$
$$P(Y = 1) = \frac{1}{2}$$
$$P(X = 1, Y = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

b. X e Z sono indipendenti

Dimostrazione analoga al punto (a).

c. Y e Z sono indipendenti

Dimostrazione analoga al punto (a).

d. X , Y , e Z non sono mutualmente indipendenti

Controesempio:

$$P(X = 1, Y = 1, Z = 1) = \frac{1}{4} \neq \frac{1}{8} = P(X = 1)P(Y = 1)P(Z = 1)$$

Esercizio 3.33

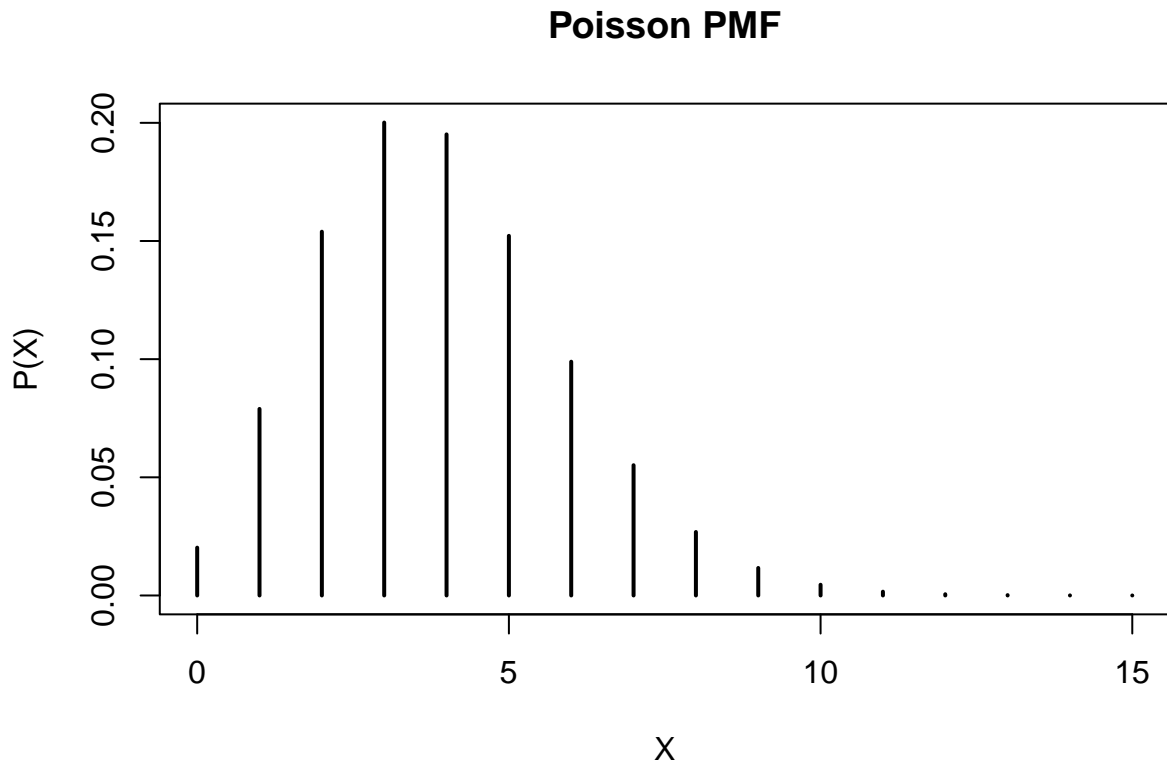
Sia $X \sim \text{Poisson}(\lambda = 3.9)$.

- Creare un grafico della funzione di massa di probabilità (PMF) di X .
- Determinare l'esito più probabile di X .
- Trovare a tale che $P(a \leq X \leq a + 1)$ sia massimizzata.
- Trovare b tale che $P(b \leq X \leq b + 2)$ sia massimizzata.

```
lambda <- 3.9
x <- 0:15

pmf <- dpois(x, lambda)

plot(0:15, pmf, type = "h", lwd = 2, main = "Poisson PMF",
     xlab = "X", ylab = "P(X)")
```



```

# b.
moda <- which.max(pmf) - 1 # Correzione indexing
cat("b. Moda di X:", moda, "\n")

# c.
prob_a <- ppois(x + 1, lambda) - ppois(x - 1, lambda)
a_ottimale <- x[which.max(prob_a)]
cat("c. a ottimale:", a_ottimale, "(P =", round(max(prob_a), 4), ")\n")

# d.
prob_b <- ppois(x + 2, lambda) - ppois(x - 1, lambda)
b_ottimale <- x[which.max(prob_b)]
cat("d. b ottimale:", b_ottimale, "(P =", round(max(prob_b), 4), ")\n")

```

```

> b. Moda di X: 3
> c. a ottimale: 3 (P = 0.3952 )
> d. b ottimale: 2 (P = 0.5492 )

```

Capitolo 4

Esercizio 4.2

Sia X una variabile aleatoria con pdf:

$$f(x) = \begin{cases} Cx^2 & 0 \leq x \leq 1 \\ C(2-x)^2 & 1 \leq x \leq 2 \end{cases}$$

Trovare la costante C .

$$\int_0^2 f(x)dx = \int_0^1 Cx^2dx + \int_1^2 C(2-x)^2dx = 1$$

Calcolo del primo integrale:

$$\int_0^1 Cx^2dx = C \left[\frac{x^3}{3} \right]_0^1 = C \cdot \frac{1}{3}$$

Calcolo del secondo integrale (con sostituzione $u = 2 - x$):

$$\int_1^2 C(2-x)^2dx = C \int_1^2 u^2du = C \left[\frac{u^3}{3} \right]_1^2 = C \cdot \left(\frac{2}{3} - \frac{1}{3} \right) = C \cdot \frac{1}{3}$$

Risultato:

$$\frac{C}{3} + \frac{C}{3} = \frac{2C}{3} = 1 \quad \Rightarrow \quad C = \frac{3}{2}$$

Esercizio 4.6

Sia X una variabile aleatoria con pdf:

$$f(x) = \begin{cases} 3(1-x)^2 & 0 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

a. Verificare che f sia una pdf valida.

Condizioni necessarie:

1. $f(x) \geq 0 \quad \forall x$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

Calcolo dell'integrale:

$$\int_0^1 3(1-x)^2 dx = 3 \int_0^1 (1-2x+x^2) dx = 3 \left[x - x^2 + \frac{x^3}{3} \right]_0^1 = 1$$

Conclusioni:

Entrambe le condizioni sono soddisfatte $\rightarrow f(x)$ è una pdf valida.

b. Trova il valore atteso e la varianza di X .

$$E[X] = \int_0^1 x \cdot 3(1-x)^2 dx = \frac{1}{4}$$
$$Var(X) = E[X^2] - (E[X])^2 = \frac{3}{80} \approx 0.0375$$

c. Trovare $P(X \leq 1/2)$.

$$\int_0^{0.5} 3(1-x)^2 dx = 3 \left[x - x^2 + \frac{x^3}{3} \right]_0^{0.5} = 3 \cdot \left(\frac{1}{2} - \frac{1}{4} + \frac{1}{27} \right) = \frac{3}{2} - \frac{3}{4} + \frac{1}{8} = \frac{7}{8} = 0.875$$

d. Trovare $P(X \leq 1/2 \mid X \geq 1/4)$.

$$P(1/4 \leq X \leq 1/2) = \int_{1/4}^{1/2} 3(1-x)^2 dx = 3 \left[x - x^2 + \frac{x^3}{3} \right]_{1/4}^{1/2} = \frac{19}{64}$$

$$P(X \geq 1/4) = \int_{1/4}^1 3(1-x)^2 dx = 3 \left[x - x^2 + \frac{x^3}{3} \right]_{1/4}^1 = \frac{27}{64}$$

$$P(A|B) = \frac{P(1/4 \leq X \leq 1/2)}{P(X \geq 1/4)} = \frac{19/64}{27/64} = \frac{19}{27} \approx 0.7037$$

Esercizio 4.7

Se $Var(X) = 3$, qual è $Var(2X + 1)$?

Utilizziamo le proprietà della varianza:

1. $Var(aX + b) = a^2 Var(X)$

2. La varianza è invariante rispetto alle traslazioni (b non influenza il risultato)

$$\text{Var}(2X + 1) = 2^2 \cdot \text{Var}(X) = 4 \cdot 3 = 12$$

```
n <- 10000

X <- rnorm(n, mean = 0, sd = sqrt(3))

Y <- 2 * X + 1

varianza_simulata <- var(Y)

cat("Varianza teorica:", 12, "\n")
cat("Varianza simulata:", round(varianza_simulata, 4))
```

```
> Varianza teorica: 12
> Varianza simulata: 11.8975
```

Esercizio 4.10

Sia X una variabile aleatoria normale con media $\mu = 1$ e deviazione standard $\sigma = 2$.

a. Calcolare $P(a \leq X \leq a + 2)$ per $a = 3$

Standardizziamo:

$$Z = \frac{X - \mu}{\sigma}$$

Calcoliamo gli estremi:

$$z_1 = \frac{3 - 1}{2} = 1, \quad z_2 = \frac{5 - 1}{2} = 2$$

Quindi:

$$P(3 \leq X \leq 5) = P(1 \leq Z \leq 2) = \Phi(2) - \Phi(1) \approx 0.9772 - 0.8413 = 0.1359$$

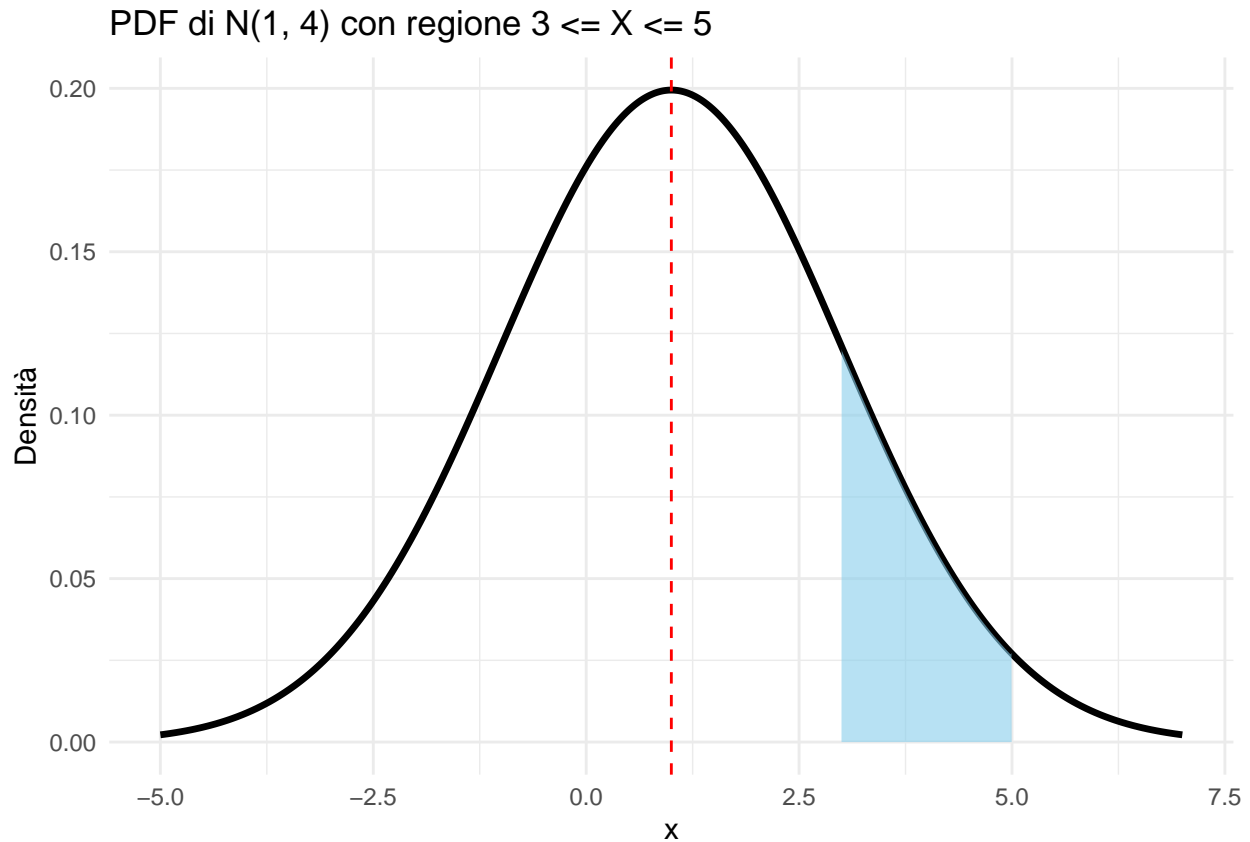
b. Disegnare il grafico della pdf di X ed evidenziare la regione corrispondente al punto (a).

```
# Parametri della distribuzione
mu <- 1
sigma <- 2

x_vals <- seq(-5, 7, length.out = 1000)
dens_df <- data.frame(
  x = x_vals,
  y = dnorm(x_vals, mean = mu, sd = sigma)
)

# Subset per l'area da ombreggiare
shade_df <- subset(dens_df, x >= 3 & x <= 5)
```

```
ggplot(dens_df, aes(x = x, y = y)) +
  geom_line(linewidth = 1.2) +
  geom_area(data = shade_df, aes(x = x, y = y), fill = "skyblue", color = "black", alpha = 0.6) +
  geom_vline(xintercept = mu, linetype = "dashed", color = "red") +
  labs(title = "PDF di N(1, 4) con regione 3 <= X <= 5", y = "Densità", x = "x") +
  theme_minimal()
```



c. Trovare il valore a che massimizza $P(a \leq X \leq a + 2)$.

La probabilità è massima quando l'intervallo $[a, a + 2]$ è **centrato sulla media** $\mu = 1$.

il valore ottimale di a è :

$$a + 1 = \mu \Rightarrow a = \mu - 1 = 0$$

Esercizio 4.11

I voti di un esame seguono una distribuzione normale con media $\mu = 80$ e deviazione standard $\sigma = 5$.

a. Probabilità che uno studente ottenga più di 85

$$Z = \frac{85 - 80}{5} = 1$$

$$P(X > 85) = 1 - \Phi(1) \approx 1 - 0.8413 = 0.1587$$

b. Probabilità che almeno 4 studenti (su 10) ottengano 85 o più

$$Y \sim \text{Bin}(n = 10, p = 0.1587)$$

$$P(Y \geq 4) = 1 - P(Y \leq 3) = 1 - \sum_{k=0}^3 \binom{10}{k} (0.1587)^k (1 - 0.1587)^{10-k}$$

```
p <- 0.1587
n <- 10

prob_Y_ge_4 <- 1 - pbinom(3, size = n, prob = p)
prob_Y_ge_4
```

```
> [1] 0.05979829
```

Esercizio 4.16

Confrontare la cdf e la pdf di una variabile casuale esponenziale con tasso $\lambda = 2$ con la cdf e la pdf di una variabile casuale esponenziale con $\lambda = 0.5$.

La pdf di una variabile casuale esponenziale con tasso λ è data da:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{per } x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

La cdf di una variabile casuale esponenziale con tasso λ è data da:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{per } x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

```
x <- seq(0, 5, length.out = 500)

# Calcolo delle PDF
pdf_lambda_2 <- dexp(x, rate = 2)
pdf_lambda_05 <- dexp(x, rate = 0.5)

# Calcolo delle CDF
cdf_lambda_2 <- pexp(x, rate = 2)
cdf_lambda_05 <- pexp(x, rate = 0.5)

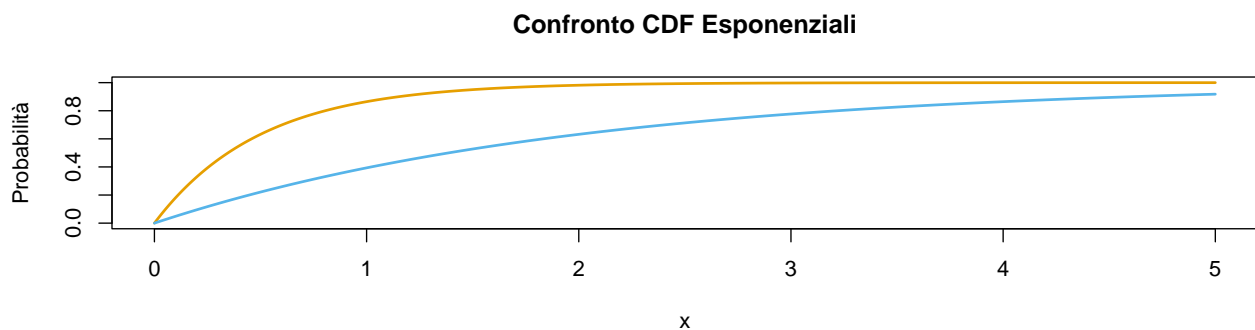
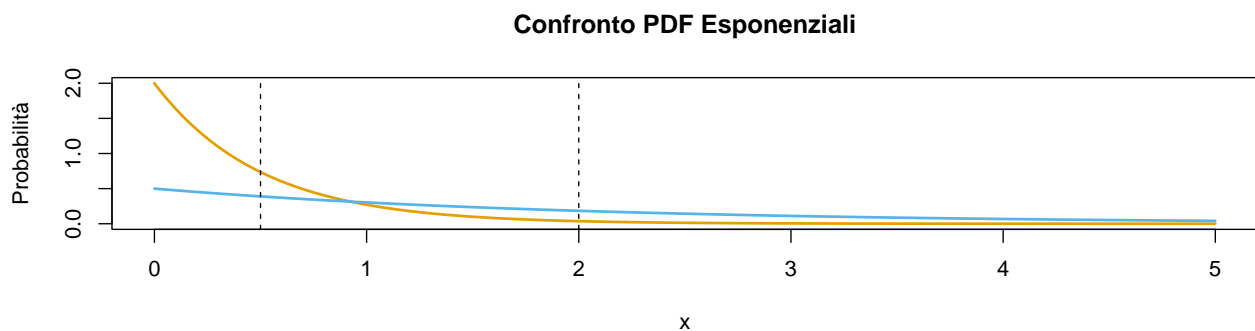
# Layout 2x1
par(mfrow = c(2, 1))
```

```

# Grafico PDF
plot(x, pdf_lambda_2, type = "l", col = "#E69F00", lwd = 2,
     main = "Confronto PDF Esponenziali",
     xlab = "x", ylab = "Probabilità",
     xlim = c(0, 5), ylim = c(0, max(pdf_lambda_2, pdf_lambda_05)))
lines(x, pdf_lambda_05, col = "#56B4E9", lwd = 2)
abline(v = c(0.5, 2), col = "black", lty = 2) # Senza l'alpha

# Grafico CDF
plot(x, cdf_lambda_2, type = "l", col = "#E69F00", lwd = 2,
     main = "Confronto CDF Esponenziali",
     xlab = "x", ylab = "Probabilità",
     xlim = c(0, 5), ylim = c(0, 1))
lines(x, cdf_lambda_05, col = "#56B4E9", lwd = 2)

```



```

# Layout originale
par(mfrow = c(1, 1))

```

Capitolo 5

Esercizio 5.2

Siano X e Y variabili aleatorie esponenziali indipendenti con **tasso 3**. Sia $Z = \max(X, Y)$.

- Stimare tramite simulazione $P(Z < 1/2)$
- Stimare media e deviazione standard di Z

Sappiamo che $X, Y \sim \text{Exp}(\lambda = 3)$ indipendenti. Definiamo:

$$Z = \max(X, Y)$$

$$P(Z < z) = P(X < z, Y < z) = P(X < z) \cdot P(Y < z)$$

La funzione di ripartizione della variabile esponenziale è:

$$P(X < z) = 1 - e^{-\lambda z}$$

Quindi:

$$P(Z < z) = (1 - e^{-3z})^2$$

```
lambda <- 3
n_sim <- 10000

Z_values <- replicate(n_sim, {
  X <- rexp(1, rate = lambda)
  Y <- rexp(1, rate = lambda)
  max(X, Y)
})

# Calcolare la probabilità stimata di Z < 1/2
prob_Z_less_1_2 <- mean(Z_values < 1/2)

# Media e deviazione standard di Z
mean_Z <- mean(Z_values)
sd_Z <- sd(Z_values)

cat("Probabilità P(Z < 1/2):", prob_Z_less_1_2, "\n")
cat("Media di Z:", mean_Z, "\n")
cat("Deviazione standard di Z:", sd_Z, "\n")
```

```
> Probabilità P(Z < 1/2): 0.6124
> Media di Z: 0.4926815
> Deviazione standard di Z: 0.3681469
```

Esercizio 5.8

Si consideri un esperimento dove 20 palline vengono distribuite casualmente in 10 urne. Sia X il numero di urne vuote.

- Stimare tramite simulazione la pmf di X
- Determinare l'esito più probabile

```

# Parametri
n_balls <- 20
n_urns <- 10
n_sim <- 10000

X_values <- replicate(n_sim, {
  balls <- sample(1:n_urns, n_balls, replace = TRUE)
  sum(table(factor(balls, levels = 1:n_urns)) == 0)
})

# Distribuzione di probabilità (PMF)
pmf_X <- table(X_values) / n_sim

cat("PMF di X:\n")
print(pmf_X)

# Esito più probabile
most_probable_outcome <- as.numeric(names(pmf_X)[which.max(pmf_X)])
cat("\nL'esito più probabile è:", most_probable_outcome, "\n")

```

```

> PMF di X:
> X_values
>      0      1      2      3      4      5
> 0.2112 0.4317 0.2789 0.0697 0.0084 0.0001
>
> L'esito più probabile è: 1

```

Esercizio 5.15

Siano X e Y variabili aleatorie uniformi indipendenti su $[0, 1]$. Sia $Z = \max(X, Y)$.

- Tracciare la pdf di Z
- Determinare quale è la probabilità maggiore tra:

$$P\left(0 \leq Z \leq \frac{1}{3}\right)$$

$$P\left(\frac{1}{3} \leq Z \leq \frac{2}{3}\right)$$

Sappiamo che se $X \sim U(0, 1)$ e $Y \sim U(0, 1)$ indipendenti, la CDF $Z = \max(X, Y)$ è:

$$F_Z(z) = P(\max(X, Y) \leq z) = P(X \leq z, Y \leq z) = P(X \leq z) \cdot P(Y \leq z) = z^2$$

La PDF di Z è la derivata della CDF :

$$f_Z(z) = \frac{d}{dz} F_Z(z) = 2z, \quad \text{per } z \in [0, 1]$$

Calcoliamo le probabilità:

$$P\left(0 \leq Z \leq \frac{1}{3}\right) = \int_0^{1/3} 2z \, dz = [z^2]_0^{1/3} = \left(\frac{1}{3}\right)^2 - 0^2 = \frac{1}{9}$$

$$P\left(\frac{1}{3} \leq Z \leq \frac{2}{3}\right) = \int_{1/3}^{2/3} 2z \, dz = [z^2]_{1/3}^{2/3} = \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9} - \frac{1}{9} = \frac{3}{9} = \frac{1}{3}$$

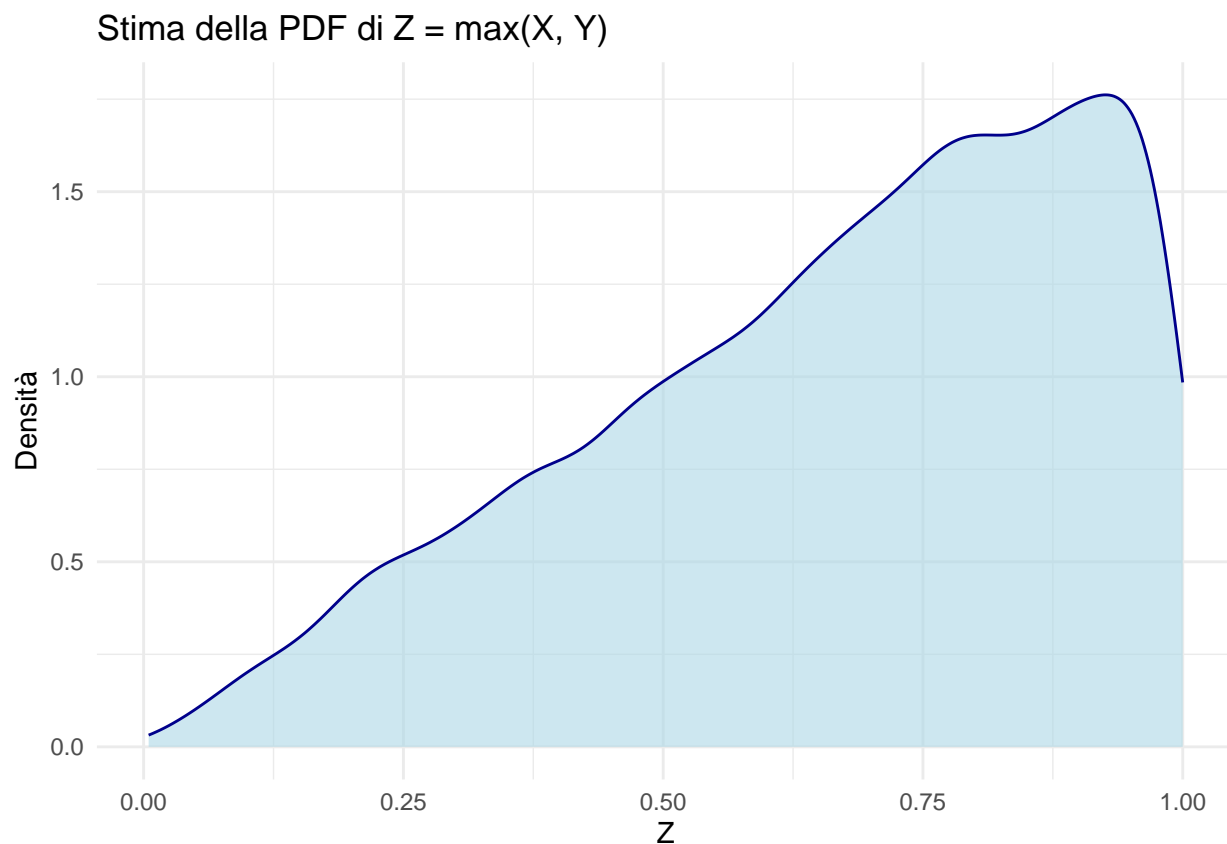
La probabilità maggiore è la seconda.

```
n_sim <- 10000

Z_values <- replicate(n_sim, {
  X <- runif(1)
  Y <- runif(1)
  max(X, Y)
})

df <- data.frame(Z = Z_values)

ggplot(df, aes(x = Z)) +
  geom_density(fill = "lightblue", color = "darkblue", alpha = 0.6) +
  labs(title = "Stima della PDF di Z = max(X, Y)",
       x = "Z",
       y = "Densità") +
  theme_minimal()
```



Esercizio 5.31

Siano X_1, \dots, X_n variabili esponenziali indipendenti con tasso $\lambda = 10$.

- Determinare media μ e deviazione standard σ
- Determinare la dimensione campionaria n necessaria per l'approssimazione normale

Una variabile esponenziale con parametro λ ha:

$$\mu = \frac{1}{\lambda} = \frac{1}{10} = 0.1, \quad \sigma = \frac{1}{\lambda} = 0.1$$

La media campionaria \bar{X}_n di n variabili esponenziali indipendenti ha media μ e deviazione standard:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{n}}$$

Secondo il **teorema del limite centrale**, la distribuzione di \bar{X}_n si approssima a una normale per n sufficientemente grande. Dal grafico notiamo che una approssimazione sia buona per $n \geq 30$.

```
# Parametri della distribuzione esponenziale
lambda <- 10
mu <- 1 / lambda
sigma <- 1 / lambda

sample_sizes <- c(5, 10, 30, 50, 100)
num_sim <- 10000

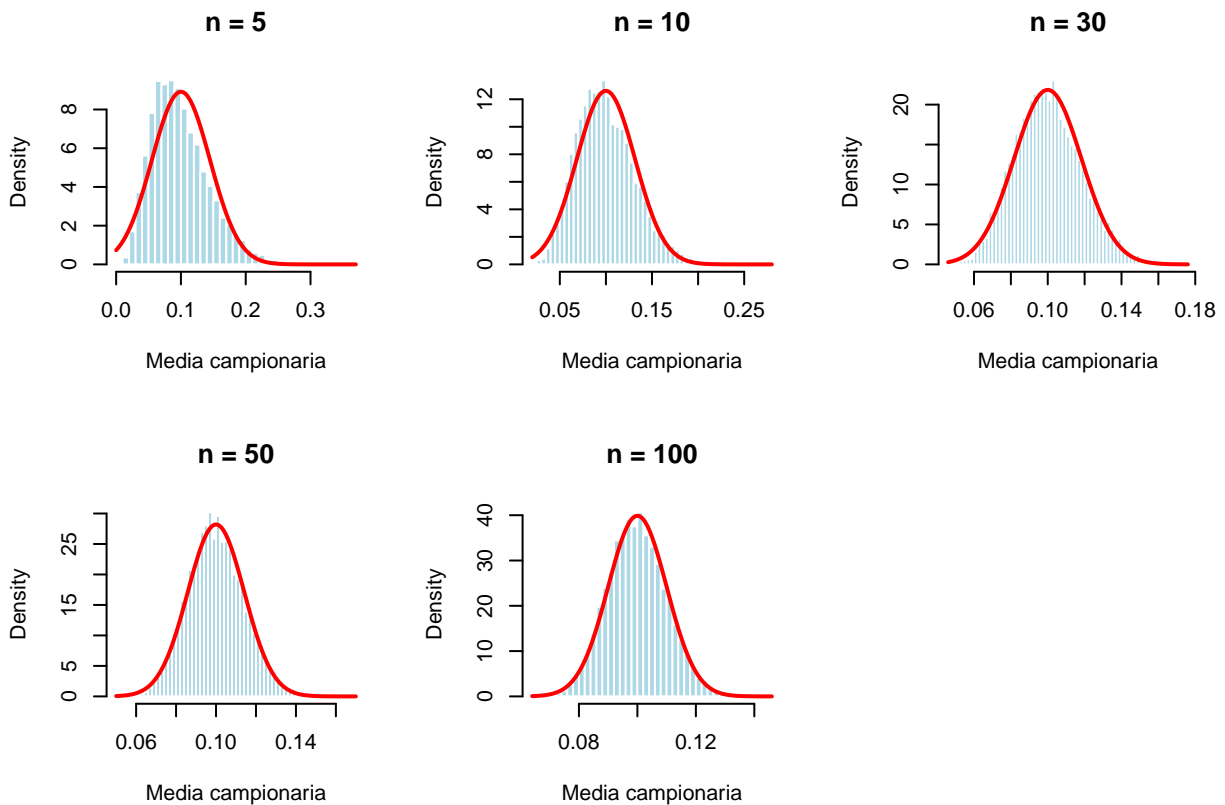
par(mfrow = c(2, 3)) # Layout per 5 grafici

for (n in sample_sizes) {
  sample_means <- replicate(num_sim, mean(rexp(n, rate = lambda)))

  hist(sample_means, breaks = 50, probability = TRUE,
        main = paste("n =", n), xlab = "Media campionaria",
        col = "lightblue", border = "white")

  curve(dnorm(x, mean = mu, sd = sigma / sqrt(n)),
        add = TRUE, col = "red", lwd = 2)
}

par(mfrow = c(1, 1)) # Reset layout
```



Esercizio 5.38

Siano X_1, \dots, X_8 variabili normali indipendenti con media 2 e deviazione standard 3.

Dimostrare tramite simulazione che $\frac{\bar{X}-2}{S/\sqrt{8}}$ segue una distribuzione t

Per un campione di dimensione n , la statistica T segue una distribuzione t con $n - 1$ gradi di libertà se il campione proviene da una distribuzione normale. In questo caso, con $n = 8$, i gradi di libertà sono 7.

La T è costruita utilizzando:

- \bar{X} , la media campionaria,
- S , la deviazione standard campionaria.

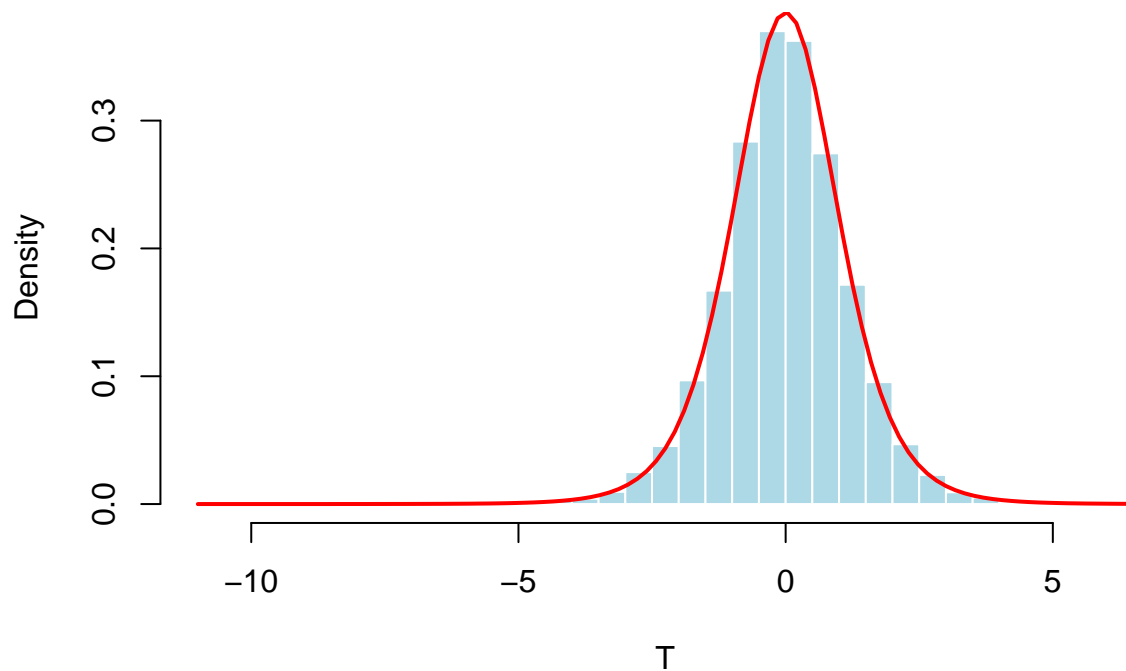
```
# Parametri della distribuzione normale
mu <- 2
sigma <- 3
n <- 8
n_sim <- 10000

T_values <- replicate(n_sim, {
  X <- rnorm(n, mean = mu, sd = sigma)
  X_bar <- mean(X)
  S <- sd(X)
  (X_bar - mu) / (S / sqrt(n))
})
```

```
# Istogramma della statistica T
hist(T_values, breaks = 50, probability = TRUE,
     col = "lightblue", border = "white",
     main = "Distribuzione della statistica T",
     xlab = "T")

# Curva teorica t di Student con n - 1 gradi di libertà
curve(dt(x, df = n - 1), add = TRUE, col = "red", lwd = 2)
```

Distribuzione della statistica T



Esercizio 5.43

Determinare tramite simulazione se $\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ è uno stimatore distorto per σ^2 quando $n = 10$

Il valore atteso dello stimatore $E[\sigma^2]$ dipende dalla relazione tra la varianza σ^2 e la media del campione μ .

Lo stimatore corretto della varianza, che è non distorto, è $E[\sigma^2]_{\text{corretto}} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

```
# Parametri della distribuzione normale
mu <- 0          # media
sigma <- 1       # deviazione standard
n <- 10          # dimensione del campione
n_sim <- 10000   # numero di simulazioni

# Simulazione dello stimatore non corretto della varianza (divide per n, non n-1)
sigma_squared_hat <- replicate(n_sim, {
  X <- rnorm(n, mean = mu, sd = sigma)
  mean((X - mu)^2) # uso della vera media popolazioneale (bias)
})
```

```
# Calcolo del valore atteso dello stimatore
mean_sigma_squared_hat <- mean(sigma_squared_hat)

# Output dei risultati
cat("Valore atteso dello stimatore (media):", mean_sigma_squared_hat, "\n")
cat("Varianza vera sigma^2:", sigma^2, "\n")

> Valore atteso dello stimatore (media): 0.9994803
> Varianza vera sigma^2: 1
```

Capitolo 6

Esercizio 6.2

Il dataset “austen” contiene i testi completi di *Emma* e *Pride and Prejudice*:

1. Crea un nuovo dataframe che contenga solo le osservazioni in *Emma*.
2. Crea un nuovo dataframe che contenga solo le variabili *word*, *word_length* e *novel*.
3. Crea un nuovo dataframe che contenga le parole di entrambi i libri disposte in ordine decrescente per lunghezza della parola.
4. Crea un nuovo dataframe che contenga solo le parole più lunghe che sono apparse in uno dei due libri.
5. Qual è la lunghezza media delle parole nei due libri insieme?
6. Crea un nuovo dataframe che contenga solo le parole distinte trovate nei due libri, insieme alle variabili **word_length** e **sentiment_score**. (Suggerimento: usa *distinct*).

```
# 1. Filtra Emma
emma_df <- austen %>% filter(novel == "Emma")

# 2. Seleziona colonne
words_df <- austen %>% select(word, word_length, novel)

# 3. Ordina per lunghezza
sorted_words <- austen %>% arrange(desc(word_length))

# 4. Parole più lunghe
max_len <- sorted_words$word_length[1]

longest_words <- sorted_words %>%
  filter(word_length == max_len) %>%
  select(novel, word, word_length)

# 5. Lunghezza media
mean_length <- austen %>% summarise(mean_length = mean(word_length, na.rm = TRUE))

# 6. Parole uniche
```

```

unique_words <- austen %>% distinct(word, .keep_all = TRUE) %>% select(word, word_length, sentiment_score)

cat("1. Emma dataframe contiene ", nrow(emma_df), "osservazioni.\n")
cat("2. Dataframe con word, word_lenth, novel: \n")
print(head(words_df))
cat("3. Le parole sono ordinate per lunghezza e numero di parole più lunghe sono: \n")
print(head(sorted_words))
cat("4. Le parole più lunghe sono: \n")
print(longest_words)
cat("5. Lunghezza media delle parole nei due libri insieme: ", mean_length$mean_length, "\n")
cat("6. Parole distinte trovate nei due libri: ", nrow(unique_words), "\n")

```

```

> 1. Emma dataframe contiene 160419 osservazioni.
> 2. Dataframe con word, word_lenth, novel:
>   word word_length novel
> 1  emma           4  Emma
> 2   by            2  Emma
> 3  jane            4  Emma
> 4 austen           6  Emma
> 5 volume           6  Emma
> 6    i             1  Emma
> 3. Le parole sono ordinate per lunghezza e numero di parole più lunghe sono:
>   word sentence chapter word_length stop_word sentiment_score
> 1 conscience-stricken  5281      34         19      FALSE          0
> 2 respectable-looking  4334      43         19      FALSE          0
> 3 companionableness   220       2         17      FALSE          0
> 4 cheerful-tempered   1732     11         17      FALSE          0
> 5 unceremoniousness   1743     12         17      FALSE          0
> 6 manchester-street   5832     37         17      FALSE          0
>   novel
> 1      Emma
> 2 Pride and Prejudice
> 3      Emma
> 4      Emma
> 5      Emma
> 6      Emma
> 4. Le parole più lunghe sono:
>   novel word word_length
> 1      Emma conscience-stricken 19
> 2 Pride and Prejudice respectable-looking 19
> 5. Lunghezza media delle parole nei due libri insieme: 4.325518
> 6. Parole distinte trovate nei due libri: 9498 .

```

Esercizio 6.8

Quale film nel genere *Comedy/Romance* che è stato valutato almeno 50 volte ha la valutazione media più bassa? Quale ha la valutazione media più alta? (dataset: fodsata::movies)

```

high <- movies %>%
  group_by(title) %>%
  filter(genres == "Comedy|Romance" & n() > 50) %>%
  summarize(rating = mean(rating)) %>%
  slice_max(rating)

low <- movies %>%
  group_by(title) %>%
  filter(genres == "Comedy|Romance" & n() > 50) %>%
  summarize(rating = mean(rating)) %>%
  slice_min(rating)

cat("Film con la valutazione media più bassa: n")
print(low)
cat("Film con la valutazione media più alta: \n")
print(high)

```

```

> Film con la valutazione media più bassa: n# A tibble: 1 x 2
>   title                rating
>   <chr>                <dbl>
> 1 What Women Want (2000)  3.14
> Film con la valutazione media più alta:
> # A tibble: 1 x 2
>   title                                rating
>   <chr>                                <dbl>
> 1 Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001)  4.18

```

Esercizio 6.10

Quale utente ha dato la valutazione media più alta? (dataset: fodsata::movies)

```

# Trova l'utente con la valutazione media più alta
top_user <- movies %>%
  group_by(userId) %>%
  summarise(
    num_ratings = n(),
    avg_rating = mean(rating, na.rm = TRUE)
  ) %>%
  arrange(desc(avg_rating)) %>%
  slice_head(n = 1)

cat("Utente con valutazione media più alta: ", top_user$userId, " con valutazione media ", top_user$avg_rating)

```

```

> Utente con valutazione media più alta: 53 con valutazione media 5

```

Esercizio 6.29

1. Quanti frutti contengono la parola “berry” nel loro nome?
2. Alcuni di questi frutti contengono la parola “fruit” nel loro nome. Trova questi frutti e rimuovi la parola “fruit” per creare una lista di parole che possono essere usate come frutti. (Suggerimento: usa `str_remove`).

```
# Convento 'fruit' in un data frame con una colonna 'name'
fruit_df <- data.frame(name = fruit, stringsAsFactors = FALSE)

# 1.
berry_fruits <- fruit_df %>%
  filter(str_detect(name, "berry")) %>%
  nrow()

# 2.
cleaned_fruits <- fruit_df %>%
  filter(str_detect(name, "fruit")) %>%
  mutate(clean_name = str_remove(name, "fruit")) %>%
  select(clean_name)

cat("1. Numero di frutti con 'berry' nel nome: ", berry_fruits, "\n")
cat("2 Frutti a cui è stato rimosso la parola fruit:\n")
print(cleaned_fruits)
```

```
> 1. Numero di frutti con 'berry' nel nome: 14
> 2 Frutti a cui è stato rimosso la parola fruit:
>   clean_name
> 1      bread
> 2      dragon
> 3      grape
> 4      jack
> 5      kiwi
> 6    passion
> 7      star
> 8      ugli
```

Esercizio 6.34

Analisi del dataset “billboard” nel pacchetto `tidyr`:

1. Quale artista ha avuto il maggior numero di tracce nelle classifiche del 2000?
2. Quale traccia del 2000 ha trascorso il maggior numero di settimane al primo posto? (Questo richiede di ordinare i dati).

```
# 1.
top_artist_2000 <- billboard %>%
  filter(year(as.Date(date.entered)) == 2000) %>%
  count(artist, sort = TRUE) %>%
  slice_head(n = 1)
```

```
# 2.
top_track_2000 <- billboard %>%
  mutate(date.entered = as.Date(date.entered)) %>%
  filter(year(date.entered) == 2000) %>%
  pivot_longer(
    cols = starts_with("wk"),
    names_to = "week",
    values_to = "position",
    values_drop_na = TRUE
  ) %>%
  filter(position == 1) %>%
  count(track, sort = TRUE) %>%
  slice_head(n = 1)

cat("Artista con il maggior numero di tracce nel 2000: ", top_artist_2000$artist, "\n")
cat("Traccia con il maggior numero di settimane al #1 nel 2000: ", top_track_2000$track, "\n")

> Artista con il maggior numero di tracce nel 2000: Jay-Z
> Traccia con il maggior numero di settimane al #1 nel 2000: Independent Women Pa...
```

Esercizio 6.38

Supponiamo di campionare cinque numeri da una distribuzione uniforme sull'intervallo $[0,1]$. Usa la simulazione per mostrare che il valore atteso del k -esimo valore più piccolo dei cinque è $k/6$. Ovvero, il minimo dei cinque valori ha valore atteso $1/6$, il secondo più piccolo ha valore atteso $2/6$, e così via.

```
simulation_results <- replicate(10000, {
  samples <- runif(5)
  sort(samples)
}) %>%
  t() %>%
  as_tibble(.name_repair = "unique") %>%
  summarise(across(everything(), mean)) %>%
  pivot_longer(everything()) %>%
  mutate(expected = row_number()/6)
```

```
> New names:
> * ' ' -> '...1'
> * ' ' -> '...2'
> * ' ' -> '...3'
> * ' ' -> '...4'
> * ' ' -> '...5'
```

```
print(simulation_results)
```

```
> # A tibble: 5 x 3
>   name value expected
```

```
>      <chr> <dbl>      <dbl>
> 1 ...1  0.167      0.167
> 2 ...2  0.332      0.333
> 3 ...3  0.501      0.5
> 4 ...4  0.667      0.667
> 5 ...5  0.832      0.833
```