Capitolo 1

Ex 1.5

Use R to calculate the sum of the squares of all numbers from 1 to 100: $1^2 + 2^2 + \cdots + 99^2 + 100^2$.

sum(seq(100)^2) = 338350

Ex 1.7

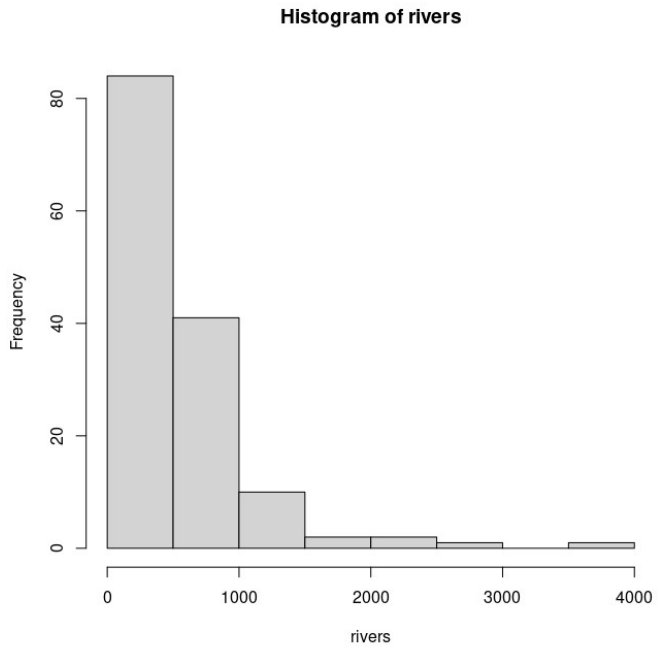
R has a built-in vector `rivers` which contains the lengths of major North American rivers.

a. Use `?rivers` to learn about the data set.
b. Find the mean and standard deviation of the rivers data using the base R functions `mean` and `sd` .
c. Make a histogram (`hist`) of the rivers data.
d. Get the five number summary (`summary`) of rivers data.
e. Find the longest and shortest lengths of rivers in the set.
f. Make a list of all (lengths of) rivers longer than 1000 miles.

a) ?rivers

b) mean(rivers) = 591.1844
sd(rivers) = 493.8708

c) hist(rivers)



d) summary(rivers) =

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
135.0	310.0	425.0	591.2	680.0	3710.0

e) max(rivers) = 3710
min(rivers) = 135

f) rivers[rivers>1000]
[1] 1459 1450 1243 2348 1171 3710 2315 2533 1306 1054 1270 1885 1100 1205 1038
[16] 1770

Ex 1.10

Consider the `mtcars` data set.

a. Which cars have 4 forward gears?
b. What subset of `mtcars` does `mtcars[mtcars$disp > 150 & mtcars$mpg > 20,]` describe?
c. Which cars have 4 forward gears and manual transmission? (Note: manual transmission is 1 and automatic is 0.)
d. Which cars have 4 forward gears or manual transmission?
e. Find the mean mpg of the cars with 2 carburetors.

a) mtcars[mtcars$gear==4,]

b) Auto con con più di 150 pollici cubi e che fanno più di 20 miglia per gallone

c) mtcars[mtcars$gear==4 & mtcars$am==1,]

d) mtcars[mtcars$gear==4 | mtcars$am==1,]

e) mean(mtcars[mtcars$carb == 2,]$mpg) = 22.4

Ex 1.11

Consider the `mtcars` data set.

a. Convert the `am` variable to a factor with two levels, `auto` and `manual`, by typing the following: `mtcars$am <- factor(mtcars$am, levels = c(0, 1), labels = c("auto", "manual"))`.
b. How many cars of each type of transmission are there?
c. How many cars of each type of transmission have gas mileage estimates greater than 25 mpg?

a) mtcars$am <- factor(mtcars$am, levels = c(0, 1), labels = c("auto", "manual"))

b) length(mtcars$am[mtcars$am == "manual"]) = 13
length(mtcars$am[mtcars$am == "auto"]) =19

c) length(mtcars$am[mtcars$am == "manual" & mtcars$mpg > 25]) = 6
length(mtcars$am[mtcars$am == "auto" & mtcars$mpg > 25]) = 0

Ex 1.14

This problem uses the package `Lahman`, which needs to be installed on your computer. The data set `Batting`, in the `Lahman` package contains batting statistics of all major league baseball players since 1871, broken down by season.

a. How many observations of how many variables are there?
b. Use the command `head(Batting)` to get a look at the first six lines of data.
c. What is the most number of triples (X3B) that have been hit in a single season?
d. What is the playerID(s) of the person(s) who hit the most number of triples in a single season? In what year did it happen?
e. Which player hit the most number of triples in a single season since 1960?

install.packages("Lahman")
library(Lahman)

a) ?Batting
"A data frame with 112184 observations on the following 22 variables."

b) head(Batting)

c) max(aggregate(x= Batting["X3B"], by = Batting["yearID"], FUN = sum)$X3B) = 1895

d) Batting[Batting$X3B == max(Batting$X3B),]$playerID = wilsoch01
Batting[Batting$X3B == max(Batting$X3B),]$yearID = 1912

e) Batting[Batting$X3B == max(Batting[Batting$yearID >= 1960,]$X3B) & Batting$yearID >= 1960,]$playerID = grandcu01

Ex 1.15

Consider the `bechdel` data set in the `fosdata` package.

a. How many movies in the data set pass the Bechdel test?
b. What percentage of movies in the data set pass the Bechdel test?
c. Create a table of number of movies in the data set by year.
d. Which year has the most movies in the data set?
e. How many different values are there in the `clean_test` variable?
f. Create a data frame that contains only those observations that pass the Bechdel test.
g. Create a data frame that contains all of the observations that do **not** have missing values in the `domgross` variable.

a) sum(bechdel$binary == "PASS") = 803

b) mean(bechdel$binary == "PASS")*100 = 44.760

c) table(bechdel$year)

d) max(table(bechdel$year)) = 129 film
ci sono 129 film per il 2010

e) length(unique(bechdel$clean_test)) = 5

f) bechdel[bechdel$binary == "PASS",]

g) na.omit(bechdel)

Capitolo 2

Ex 2.3

A hat contains slips of paper numbered 1 through 6. You draw two slips of paper at random from the hat, without replacing the first slip into the hat.

a. Write out the sample space S for this experiment.
b. Write out the event E, "the sum of the numbers on the slips of paper is 4."
c. Find $P(E)$.
d. Let F be the event "the larger number minus the smaller number is 0." What is $P(F)$?

a) S={(1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)}

b) E={(1, 3), (3, 1)}

c) P(E) = |E|/|S| = 2/30 = 1/15 = 0,067

d) L'evento F è impossibile poichè se la differenza tra I due numeri è zero, i numeri dovrebbero essere uguali, cosa impossibile dato che non viene reinserita la striscia di carta.

Ex 2.17

A standard deck of cards has 52 cards, four each of 2,3,4,5,6,7,8,9,10,J,Q,K,A. In blackjack, a player gets two cards and adds their values. Cards count as their usual numbers, except Aces are 11 (or 1), while K, Q, J are all 10.

a. "Blackjack" means getting an Ace and a value 10 card. What is the probability of getting a blackjack?
b. What is the probability of getting 19? (The probability that the sum of your cards is 19, using Ace as 11)

Use R to simulate dealing two cards, and compute these probabilities experimentally.

a) Per calcolare la probabilità di fare Blackjack devo calcolare:
P(estraggo una carta che vale 11 per prima) * P(estraggo una carta che vale 10 per seconda) + P(estraggo una carta che vale 10 per prima) * P(estraggo una carta che vale 11 per seconda)
= 4/52 * 16/51 + 16/52 * 4/51 = 32/663 = 0.048
deck = rep(c(2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 10, 10, 11), 4)
mean(replicate(100000, sum(sample(deck, 2, replace=FALSE)))==21) = 0.04754

b) Per calcolare la probabilità di fare 19 devo calcolare:
P(estraggo una carta che vale 11 per prima) * P(estraggo una carta che vale 8 per seconda) + P(estraggo una carta che vale 10 per prima) * P(estraggo una carta che vale 9 per seconda) +

P(estraggo una carta che vale 9 per prima) * P(estraggo una carta che vale 10 per seconda) +
P(estraggo una carta che vale 8 per prima) * P(estraggo una carta che vale 11 per seconda)
= 4/52 * 4/51 + 16/52 * 4/51 + 4/52 * 16/51 + 4/52 * 4/51 = 40/663 = 0,060
deck = rep(c(2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 10, 10, 11), 4)
mean(replicate(100000, sum(sample(deck, 2, replace=FALSE)))==19) = 0.06022

Ex 2.18

Deathrolling in World of Warcraft works as follows. Player 1 tosses a 1000-sided die. Say they get x_1. Then player 2 tosses a die with x_1 sides on it. Say they get x_2. Player 1 tosses a die with x_2 sides on it. This pattern continues until a player rolls a 1. The player who loses is the player who rolls a 1. Estimate via simulation the probability that a 1 will be rolled on the 4th roll in deathroll.

definisco la funzione dell'esperimento:

```
dr = function() {
        n=1000;
        i=0;
        while(n!=1){
                n=sample(n, 1)
                i=i+1
        }
        i
}
```

ripeto l'esperimento 1000 volte e calcolo la probabilità che il risultato sia 4
mean(replicate(1000, dr())==4) = 0.043

Ex 2.30

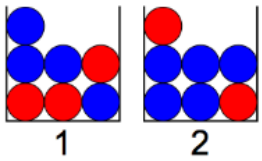
Suppose there is a new test that detects whether people have a disease. If a person has the disease, then the test correctly identifies that person as being sick 99.9% of the time (the *sensitivity* of the test). If a person does not have the disease, then the test correctly identifies the person as being well 97% of the time (the *specificity* of the test). Suppose that 2% of the population has the disease. Find the probability that a randomly selected person has the disease given that they test positive for the disease.

A = avere la malattia
B = essere positivi al test
Regola di Bayes
P(A|B) = P(B|A) * P(A) / P(B)

P(B|A) = probabilità che il test sia positivo avendo la malattia = 99.9%
P(A) = 2%
P(B) = P(B|A)*P(A) + P(B|$_A$)P($_A$)
= 0.999*0.02 + (1-97%)*(1-0.02) = 0.04938
P(A|B) = 0.405

Ex 2.31

Suppose that there are two boxes containing marbles.



Box 1 contains 3 red and 4 blue marbles. Box 2 contains 2 red and 5 blue marbles. A single die is tossed, and if the result is 1 or 2, then a marble is drawn from box 1. Otherwise, a marble is drawn from box 2.

a. What is the probability that the marble drawn is red?
b. What is the probability that the marble came from box 1 given that the marble is red?

a) A = la biglia sarà stata estratta dalla scatola 1
B = la biglia sarà stata estratta dalla scatola 2
C = la biglia estratta è rossa (scatola 1)
D = la biglia estratta è rossa (scatola 2)
P(A) = 1/3
P(B) = 2/3
P(C) = 3/7
P(D) = 2/7

F = la biglia estratta è rossa
P(F) = P(A)*P(C) + P(B)*P(D) = 1/3

b) P(A|F) = P(A∩F)/P(F)
A∩F = la biglia è stata estratta della scatola 1 ed è rossa
P(A∩F) = P(C)*P(A) = 1/7
P(A|F) = 3/7

Ex 2.36

A box contains 5 red marbles and 5 blue marbles. Six marbles are drawn without replacement.

a. How many ways are there of drawing the 6 marbles? Assume that getting all 5 red marbles and the first blue marble is different than getting all 5 red marbles and the second blue marble, for example.
b. How many ways are there of drawing 4 red marbles and 2 blue marbles?
c. What is the probability of drawing 4 red marbles and 2 blue marbles?

a) Per calcolare il numero di combinazioni possibili per 6 biglie basta sommare le combinazioni possibili con 1, 2, 3, 4 e 5 biglie rosse. Non possono esserci combinazioni con 0 biglie rosse poiché servirebbero 6 biglie blu (ne abbiamo 5) e neanche con 6 biglie rosse (ne abbiamo 5).
Combinazioni con n biglie rosse = $\binom{6}{n}$

$|S| = \binom{6}{1}+\binom{6}{2}+\binom{6}{3}+\binom{6}{4}+\binom{6}{5}=62$

b) $|A| = \binom{6}{2} = 15$

c) $P(A) = |A|/|S| = 15/62 = 0.2419355$

Capitolo 3

Ex 3.3

Let X be a discrete random variable with probability mass function given by

$$p(x) = \begin{cases} C/4 & x = 0 \\ C/2 & x = 1 \\ C & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of C that makes p a valid probability mass function.

La somma delle probabilità deve essere pari a 1 quindi:
C/4 + C/2 + C = 1
C = 4/7

Ex 3.15

Let k be a positive integer and let X be a random variable with pmf given by $p(x) = 1/k$ for $x = 1, \ldots, k$ and $p(x) = 0$ for all other values of x. Find $E[X]$.

Il valore atteso E[X] di una variabile aleatoria X e massa di probabilità p(x) si calcola:
$E[X] = \sum_x x * p(x)$
nel nostro caso p(x) = 1/k per 1 < x < k, quindi:
$\sum_{n=1}^{k} n * \frac{1}{k} = \frac{1}{k} * \frac{k*(k+1)}{2} = \frac{k+1}{2}$

Ex 3.19

In October 2020, the YouTuber called “Dream” posted a speedrun of Minecraft and was accused of cheating.

In Minecraft, when you trade with a piglin, the piglin gives you an ender pearl 4.7% of the time. Dream got 42 ender pearls after 262 trades with piglin.

a. If you trade 262 times, what is the expected number of ender pearls you receive?
b. What is the probability of getting 42 or more ender pearls after 262 trades?

When you kill a blaze, you have a 50% chance of getting a blaze rod. Dream got 211 blaze rods after killing 305 blazes.

c. If you kill 305 blazes, what is the expected number of blaze rods you receive?
d. What is the probability of getting 211 or more blaze rods after killing 305 blazes?
e. Do you think Dream was cheating?

a) La funzione di distribuzione è una binomiale con n=262 e p = 0.047, quindi E[X] = np
il valore atteso come numero di ender perl è 262*0.047 = 12.314
Si può calcolare per simulazione con:
mean(rbinom(100000, size=262, prob=0.047))

b) $P(X \geq 42) = \sum_{x=42}^{\infty} P(X = x) = \sum_{x=42}^{\infty} \binom{262}{x} (0.047)^x (1-0.047)^{262-x}$
sum(dbinom(42:1000000, 262, 0.047)) = 4.6×10^{-12}
uso un numero arbitrariamente grande al posto dell’infinito

c) La funzione di distribuzione è una binomiale con n=305 e p = 0.50, quindi E[X] = np
il valore atteso come numero di blaze rod è 305*0.50 = 152.5
Si può calcolare per simulazione con:
mean(rbinom(100000, size=305, prob=0.50))

d) $P(X \geq 211) = \sum_{x=211}^{\infty} P(X = x) = \sum_{x=211}^{\infty} \binom{305}{x} (0.50)^x (1-0.50)^{305-x}$
sum(dbinom(211:1000000, 305, 0.50)) = 8.8×10^{-12}
uso un numero arbitrariamente grande al posto dell’infinito

e) Stava molto probabilmente barando, poiché, per completare la sfida, era necessario che queste due condizioni si verificassero infieme. La probabilità che questo avvenga è il prodotto tra le due probabilità calcolate al punto b) e al punto d). Una probabilità con un ordine di grandezza di 10^-24 si può considerare pressochè impossibile.

Ex 3.22

Let X and Y be random variables such that $E[X] = 2$ and $E[Y] = 3$.

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

a) E[4X + 5Y] = 4E[X] + 5E[Y] = 4*2 + 5*3 = 23
b) E[4X – 5Y + 2] = 4E[X] – 5E[Y] +E[2] = 4*2 - 5*3 + 2 = -5

Ex 3.32

In an experiment[22] to test whether participants have absolute pitch, scientists play notes and the participants say which of the 12 notes is being played. The participant gets 1 point for each note that is correctly identified, and 3/4 of a point for each note that is off by a half-step. (Note that if the possible guesses are 1:12, then the difference between 1 and 12 is a half-step, as is the difference between any two values that are 1 apart.)

a. If the participant hears 36 notes and randomly guesses each time, what is the expected score of the participant?
b. If the participant hears 36 notes and randomly guesses each time, what is the standard deviation of the score of the participant? Assume each guess is independent.

a) Per ogni nota il partecipate ha 1/12 di probabilità di indovinare la nota corretta (valore 1) e 2/12 di probabilità di indovinare le due note adiacenti (valore 0.75 ciascuna). Dunque il valore atteso per una sola nota è 1*1/12 + 0.75*2/12 = 5/24 = 0.208
Moltiplicando questo valore per le 36 note di ha come valore atteso 36*5/24 = 7.5

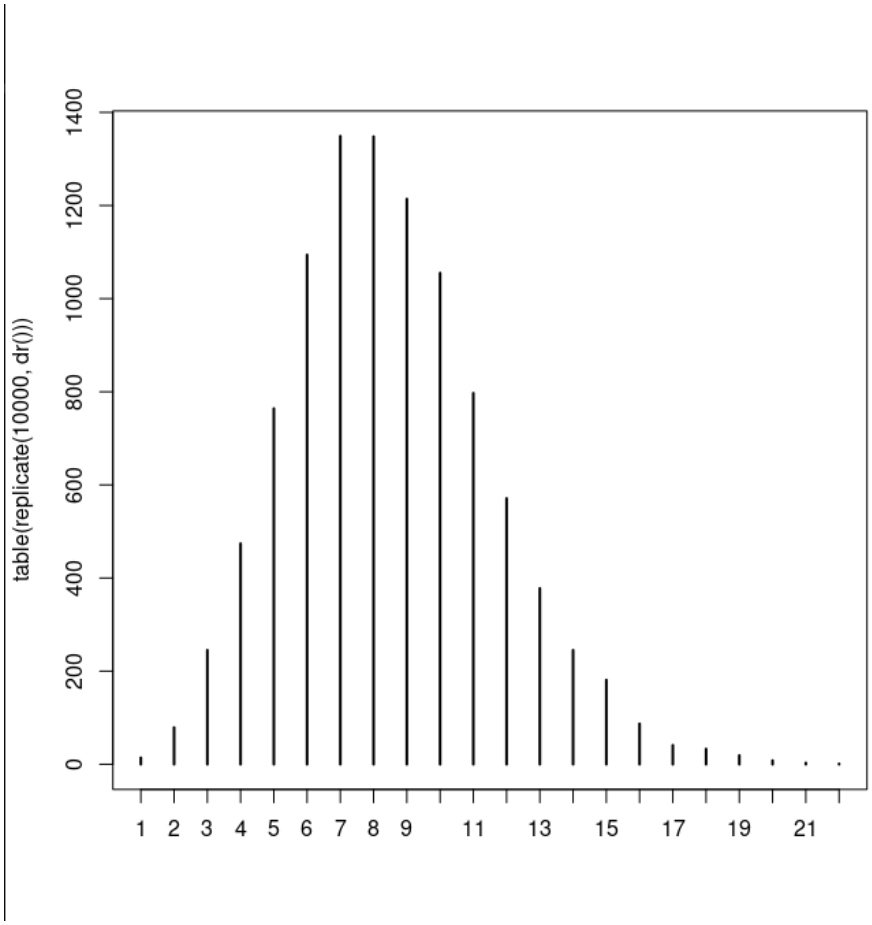
b) $Var(X) = E[X^2] - E[X]^2$
$sd(X) = \sqrt{Var(X)}$
$E[X^2] = 1^2 * \frac{1}{12} + 0.75^2 * \frac{2}{12} = \frac{17}{96} = 0.177$
$E[X]^2 = 0.208^2 = 0.043$
$Var(X) = 0.177 - 0.043 = 0.134$
$sd(X) = \sqrt{0.134} = 0.366$

Ex 3.40

Deathrolling in World of Warcraft works as follows. Player 1 tosses a 1000-sided die. Say they get x_1. Then player 2 tosses a die with x_1 sides on it. Say they get x_2. Player 1 tosses a die with x_2 sides on it. The player who loses is the player who first rolls a 1.

a. Estimate the expected total number of rolls before a player loses.
b. Estimate the probability mass function of the total number of rolls.
c. Estimate the probability that player 1 wins.

a) Per questo esercizio riutilizziamo la funzione dr() definita nell'esercizio 2.18.
Ripetiamo l'esperimento un numero arbitrariamente grande di volte e facciamo la media di tutti i valori ottenuti.
$mean(replicate(100000, dr())) \approx 8.500$
b) plot(table(replicate(10000, dr())), type='h')



c) Il giocatore 1 vince se il giocatore 2 perde, quindi quando la funzione dc() dà in output un valore pari
mean(replicate(100000, dr())%%2==0) = 0.49804

Capitolo 4

Ex 4.4

Provide an example of a pdf f for a random variable X such that there exists an x for which $f(x) > 1$. Is it possible to have $f(x) > 1$ for all values of x?

$f(x) = 2 \; per \; 0 \leq x \leq \frac{1}{2}$, 0 altrimenti

In questa funzione $f(0) > 1$ quindi la condizione è soddisfatta.
Per rispettare la condizione che l'integrale della funzione sia uguale a 1, non è possibile che tutti i valori di f(x) siano maggiori di 1.

Ex 4.7

If $\mathrm{Var}(X) = 3$, what is $\mathrm{Var}(2X + 1)$?

$Var(X) = 3$
$Var(2X+1) = E[(2X+1)^2] - E[2X+1]^2$
$E[4X^2+4X+1] - (2E[X]+1)^2$
$4E[X^2]+4E[X]+1-4E[X]^2-4E[X]-1$
$4E[X^2]-4E[X]^2 = 4Var(X) = 4*3 = 12$

Ex 4.11

Suppose that scores on an exam are normally distributed with mean 80 and standard deviation 5, and that scores are not rounded.

a. What is the probability that a student scores higher than 85 on the exam?
b. Assume that exam scores are independent and that 10 students take the exam. What is the probability that 4 or more students score 85 or higher on the exam?

a) P(X >85) = 1 – F(85) = 1 – pnorm(85, 80, 5) = 0.159

b) Scrivo la funzione test che simula 10 esami e ritorna true se 4 o più riportano un punteggio superiore a 85

```
test() = function() {
    a = rnorm(10, 80, 5)
    length(a[a>85]) >= 4
}
```

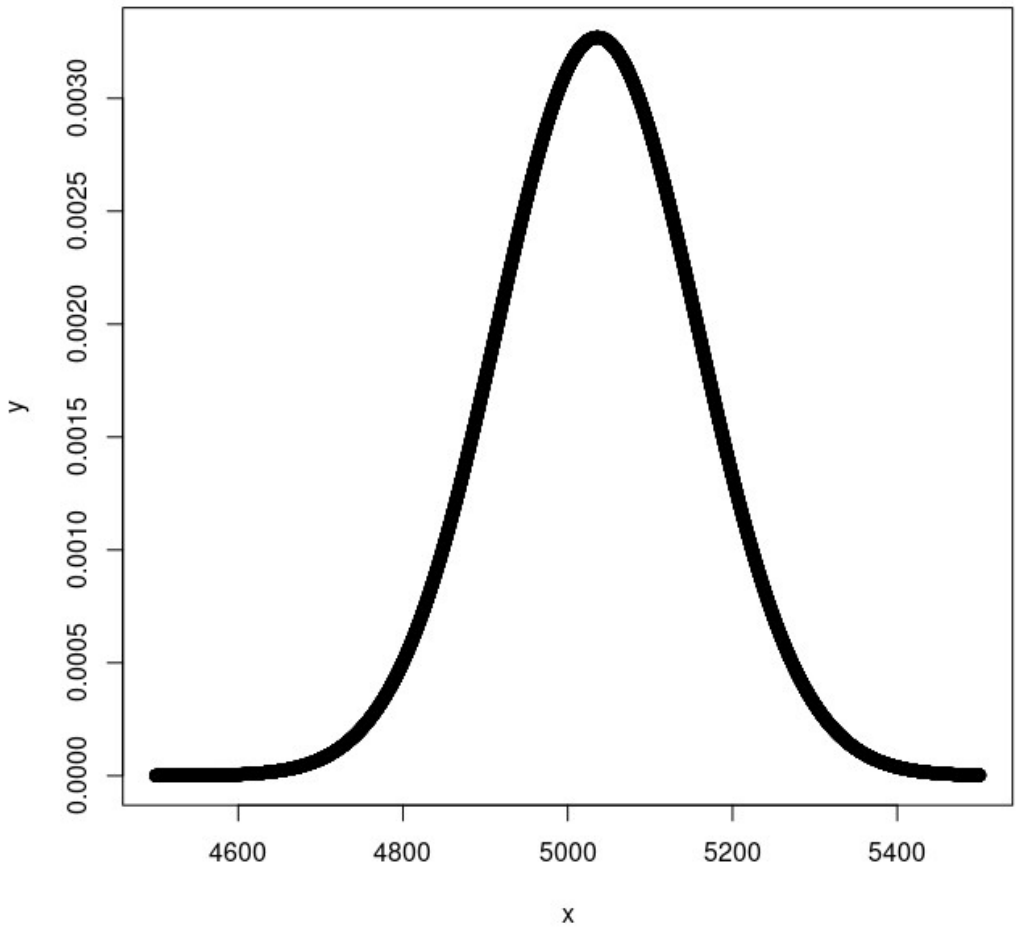
replico questo test 10000 volte e calcolo la probabilità che sia un successo
mean(replicate(10000, test())) = 0.0614

Ex 4.12

Climbing rope will break if pulled hard enough. Experiments show that 10.5 mm dynamic nylon rope has a mean breaking point of 5036 lbs with a standard deviation of 122 lbs. Assume breaking points of rope are normally distributed.

a. Sketch the distribution of breaking points for this rope.
b. What proportion of ropes will break with 5000 lbs of load?
c. At what load will 95% of all ropes break?

a) x = seq(4500, 5500, length = 10000)
y=dnorm(x, 5036, 122)
plot(x, y)



b) P(X <= 5000) = F(5000) = pnorm(5000, 5036, 122) = 0.384
c) P(x) = 0.95
x = qnorm(0.95, 5036, 122) = 5236.672

Ex 4.19

Suppose the time to failure (in years) for a particular component is distributed as an exponential random variable with rate $\lambda = 1/5$. For better performance, the system has two components installed, and the system will work as long as either component is functional. Assume the time to failure for the two components is independent. What is the probability that the system will fail before 10 years have passed?

P(X <= 10) = probabilità che un componente si rompa prima di 10 anni
P(X <= 10) = F(10) = pexp(10, rate = 1/5) = 0.865
A noi interessa sapere la probabilità che l'intero sistema si rompa, dunque che entrambi i componenti si rompano. Dobbiamo moltiplicare le due probabilità.
F(10)^2 = (pexp(10, rate = 1/5))^2 = 0.748

Ex 4.27

Suppose that you have two infinite, horizontal parallel lines that are one unit apart. You drop a needle of length 1/2 so that its center between the two lines is uniform on $[0, 1]$, and the angle that the needle forms relative to the parallel lines is uniform on $[0, \pi]$.

Estimate the probability that the needle touches one of the parallel lines, and confirm that your answer is approximately $1/\pi$.

Calcolo un vettore di 10000 elementi casuali tra 0 e pi e ne faccio il seno e successivamente moltiplico per 25
sinrho = sin(sample(1000, 10000, replace = TRUE)*(pi/1000))*0.25
Calcolo un vettore di 100 elementi casuali tra 0 e 1
x = sample(1000, 10000, replace = TRUE)/1000
Creo un dataframe con questi due record
df = data.frame(sinrho, x)
Seleziono solo i record che toccano una delle due linee parallele
df_touch = df[df$x + df$sinrho >= 1 | df$x - df$sinrho <=0,]
Conto i record che soddisfano le richieste e divido per il numero totale di record per ottenere la probabilità
nrow(df_touch)/10000 = 0.322

La probabilità teorica è di $\frac{1}{\sqrt{pi}}$ = 0.318 ed è simile a quella trovata sperimentalmente

Capitolo 5

Ex 5.2

Let X and Y be independent exponential random variables with rate 3. Let $Z = \max(X, Y)$ be the maximum of X and Y.

a. Estimate via simulation $P(Z < 1/2)$.
b. Estimate the mean and standard deviation of Z.

a) mean(replicate(100000, max(rexp(1, 3), rexp(1, 3)))<=0.5)
b) mean(replicate(100000, max(rexp(1, 3), rexp(1, 3))))
sd(replicate(100000, max(rexp(1, 3), rexp(1, 3))))

Ex 5.13

Suppose there are two candidates in an election. Candidate A receives 52 votes and Candidate B receives 48 votes. You count the votes one at a time, keeping a running tally of who is ahead. At each point, either A is ahead, B is ahead, or they are tied. Let X be the number of times that Candidate B is ahead in the 100 tallies.

a. Estimate the pmf of X and plot it.
b. Estimate $P(X > 50)$.

a) Creo un vettore contenente tutti i voti (-1 significa voto per A, 1 per B)
votes = rep(c(-1, 1), c(52, 48))

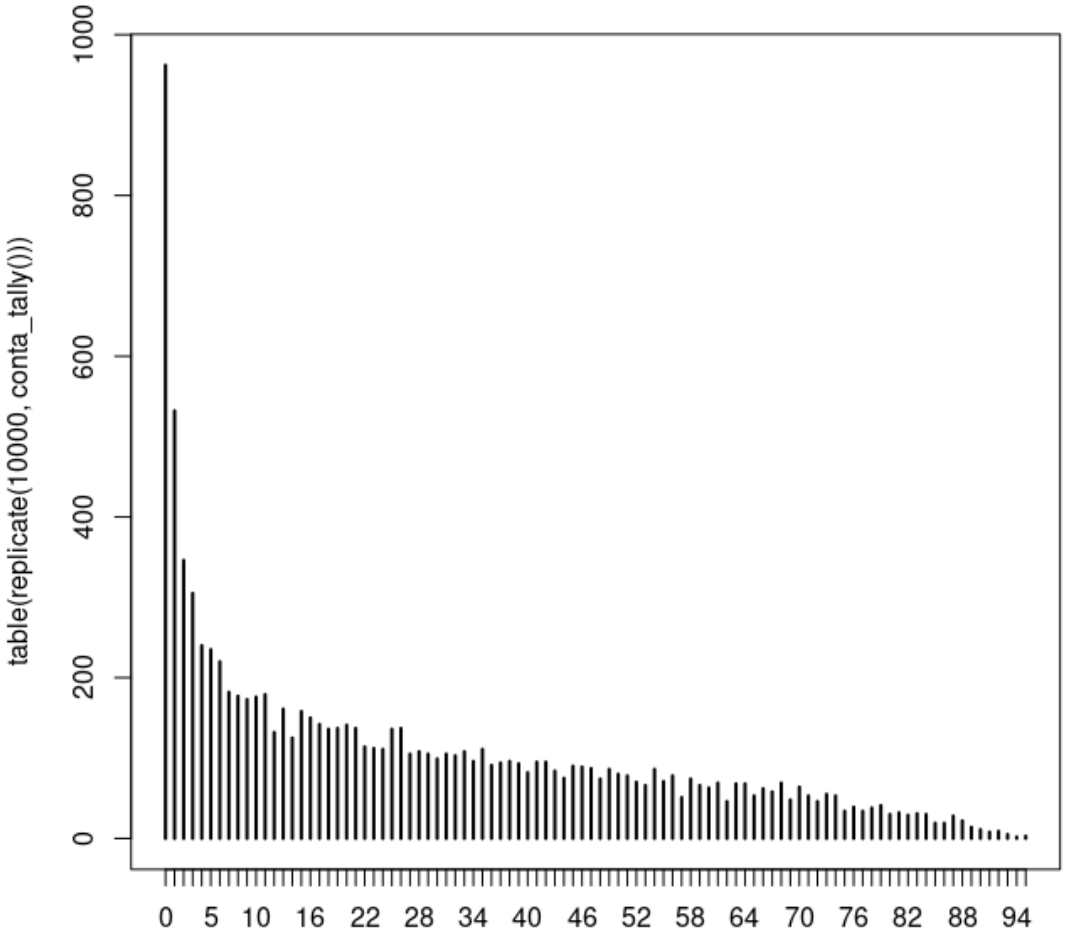
Creo un vettore, per ora vuoto, che conterrà in ogni record un numero positivo se è in vantaggio B, un numero negativo se è in vantaggio A e 0 in caso di pareggio (ovvero contiene nell'elemento n la somma di tutti i voti da i a n)
tallies = c()

Ora definisco una funzione che calcola il numero di volte in cui è in vantaggio B:

```
conta_tally = function(){
    random_votes = sample(votes) #randomizza i voti
    #calcola tutti i record del vettore tallies
    for(x in 1:100) {
        tallies = append(tallies, sum(random_votes[1:x]))
    }
    #calcola quante volte appare un numero positivo (vantaggio B)
    length(tallies[tallies > 0])
}
```

Eseguo l'esperimento 1000 volte e faccio il grafico della densità.
plot(density(replicate(1000, conta_tally())))

b) mean(replicate(10000, conta_tally()) >50) = 0.200


table(replicate(10000, conta_tally()))
0
200
400
600
800
1000
0 5 10 16 22 28 34 40 46 52 58 64 70 76 82 88 94

Ex 5.24

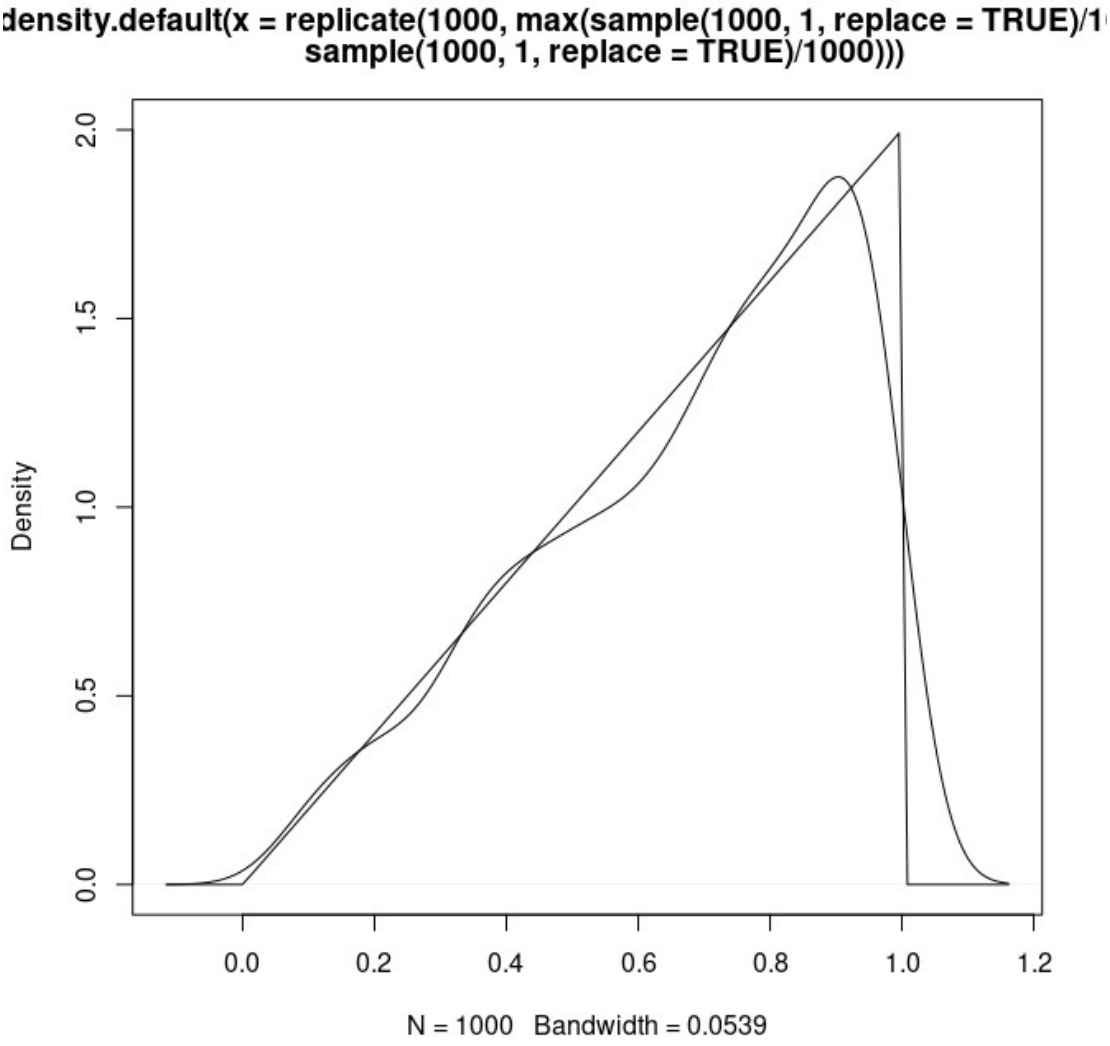
The *beta distribution* plays an important role in Bayesian statistics. It has two parameters, called *shape parameters*.

Let X and Y be independent uniform rvs on the interval $[0, 1]$. Estimate via simulation the pdf of the maximum of X and Y, and compare it to the pdf of a beta distribution with parameters `shape1 = 2` and `shape2 = 1`. (Use `dbeta()`.)

Prendo due numeri casuali tra 0 e 1 e ne calcolo il massimo, ripeto l'esperimento 1000 volte e infine faccio il grafico della densità
plot(density(replicate(1000, max(sample(1000, 1, replace=TRUE) /1000, sample(1000, 1, replace=TRUE) /1000))), ylim = c(0, 2))

Aggiungo la distribuzione beta con shape1 = 2 e shape 2 = 1
curve(dbeta(x, shape1 = 2, shape2 = 1), add= TRUE)

Si può notare che la distribuzione sperimentale si attiene relativamente fedelmente alla curva teorica



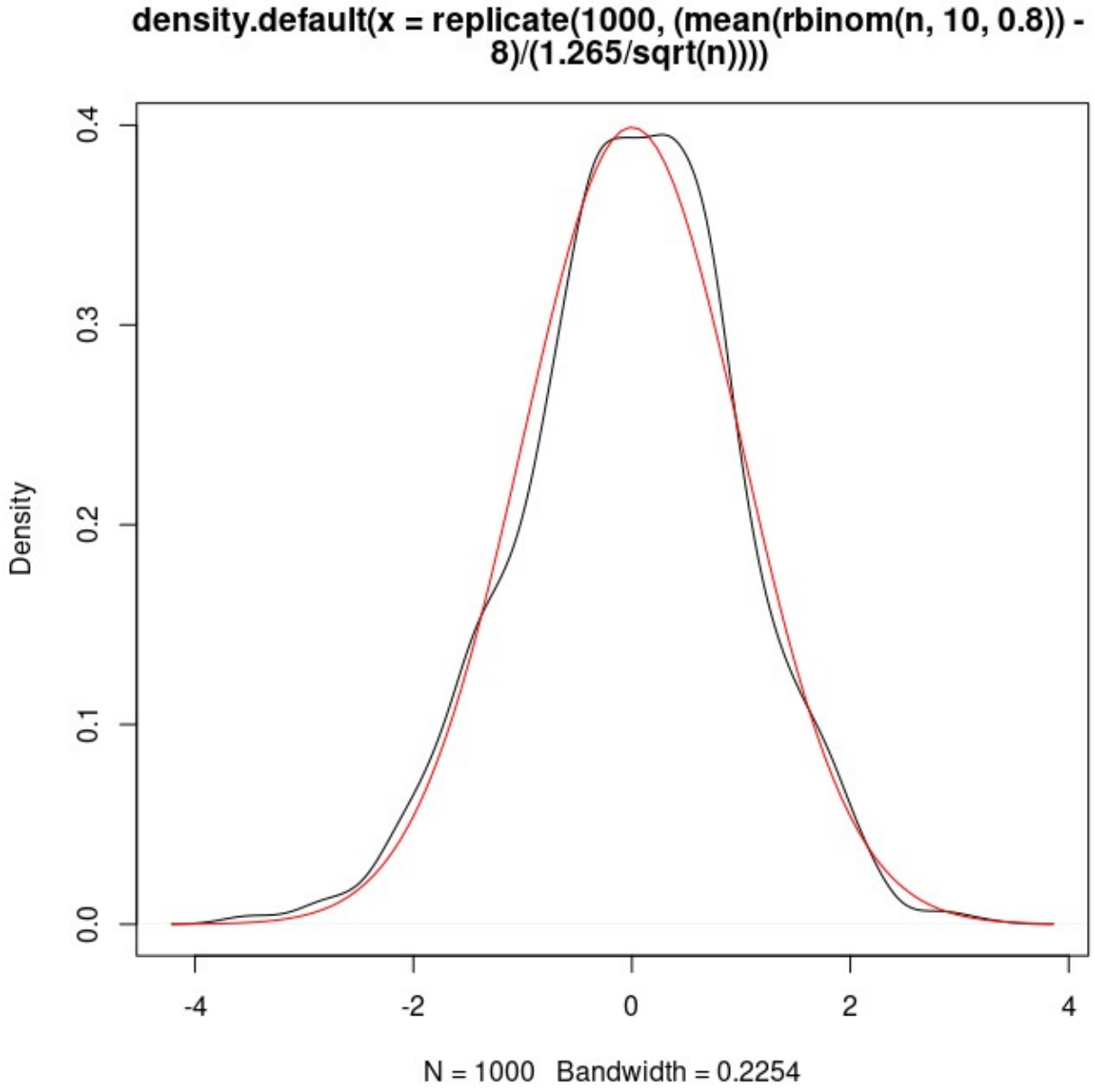
Ex 5.33

Let $X_1, \ldots, X_n$ be independent binomial rvs with $n = 10$ and $p = 0.8$.

a. What are the mean μ and the sd σ of the $\mathrm{Binom}(10, 0.8)$ distribution?

b. How large does n need to be before the pdf of $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is approximately that of a standard normal rv?

a) Seguendo la teoria:
mean(Binom(10, 0.8)) = np = 8
sd(Binom(10, 0.8)) = np(1-p) = 1.265
b) plot(density(replicate(1000, (mean(rbinom(n, 10, 0.8)) - 8) / (1.265/sqrt(n)))))
curve(dnorm(x), add = TRUE, col = "red")



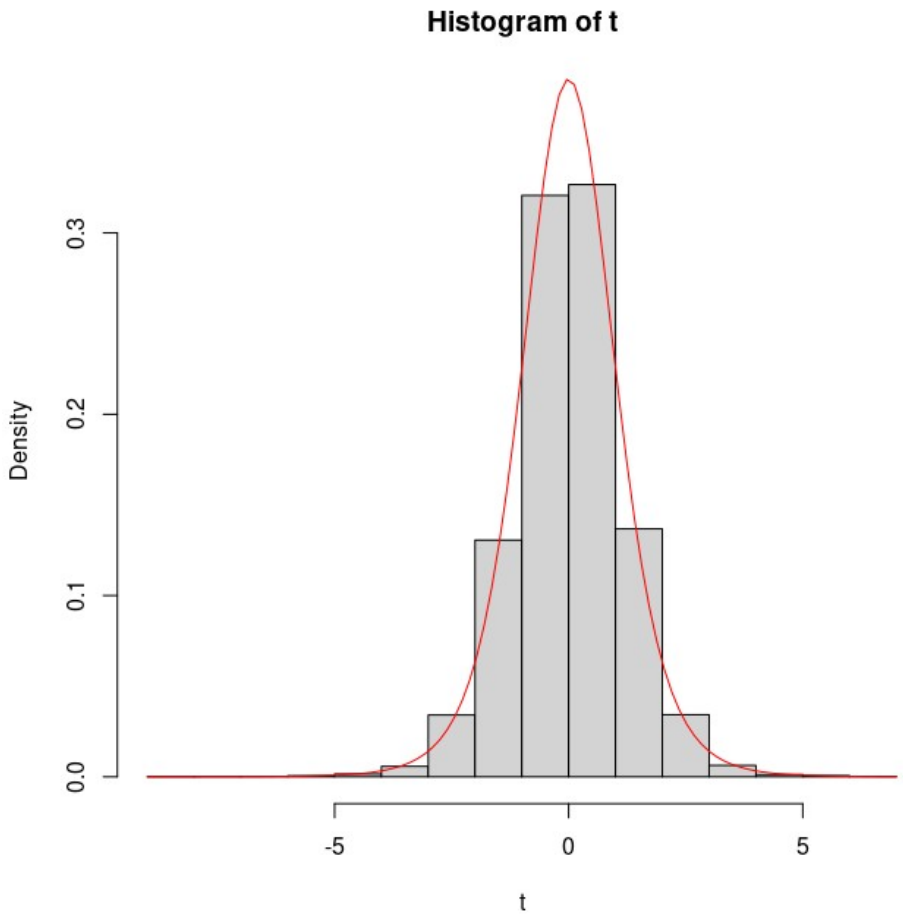
Ex 5.38

Let $X_1, \ldots, X_8$ be independent normal random variables with mean 2 and standard deviation 3. Show using simulation that

$$\frac{\overline{X} - 2}{S/\sqrt{8}}$$

is a t random variable with 7 degrees of freedom.

```
tData <- replicate(10000, {
        X <- rnorm(8, 2, 3)
        (mean(X) - 2) / (sd(X) / sqrt(8))
})
hist(tData,
        probability = TRUE,
        ylim = c(0, .37),
        main = "Histogram of t",
        xlab = "t"
)
curve(dt(x, df = 7), add = TRUE, col = "red")
```



Ex 5.42

Show through simulation that the median is a biased estimator for the mean of an exponential rv with $\lambda = 1$. Assume a random sample of size 8.

La media che ci aspettiamo per una distribuzione esponenziale con lambda=1 è 1/lambda, ovvero 1.
Ora calcoliamo il bias della mediana:
Prendiamo 8 sample di una distribuzione esponenziale con lambda=1 e calcoliamo la mediana, successivamente sottraiamo al risultato la media che ci aspettiamo.
Replichiamo questo esperimento tante volte.

```
mean(replicate(10000, {
  x <- rexp(8, 1)
  s <- median(x)
  (s - 1)
}))
```

La media di tutti questi esperimenti è: -0.243
Questa è un'approssimazione del bias.

Capitolo 6

Ex 6.2

Consider the `austen` data set in the `fosdata` package. This data frame contains the complete texts of *Emma* and *Pride and Prejudice*, with additional information which you can read about in the help page for the data set. Each of the following tasks corresponds to using a single `dplyr` verb.

a. Create a new data frame that consists only of the observations in *Emma*.
b. Create a new data frame that contains only the variables `word` , `word_length` and `novel` .
c. Create a new data frame that has the words in both books arranged in descending word length.
d. Create a new data frame that contains only the longest words that appeared in either of the books.
e. What was the mean word length in the two books together?
f. Create a new data frame that consists only of the distinct words found in the two books, together with the word length and sentiment score variables. (Hint: use `distinct`).

a) filter(austen, novel == "Emma")
b) select(austen, word, word_length, novel)
c) arrange(austen, desc(word_length))
d) slice_max(austen, word_length)
e) summarize(austen, Mean = mean(word_length))
f) distinct(austen, word, word_length, sentiment_score)

Ex 6.15

Exercises 6.11 – 6.16 all use the `Batting` data set from the `Lahman` package. This gives the batting statistics of every player who has played baseball from 1871 through the present day.

a. Which player has the most lifetime at bats without ever having hit a home run?
b. Which active player has the most lifetime at bats without ever having hit a home run? (An active player is someone with an entry in the most recent year of the data set).

a) Batting %>%
select(playerID, yearID, HR) %>%
group_by(playerID) %>%
mutate(years=max(yearID)-min(yearID)+1, HR=sum(HR)) %>%
filter(HR==0) %>%
select(playerID, years) %>%
distinct() %>%
ungroup() %>%
slice_max(years)
= moyerja01 (27 anni)

b) Batting %>%
select(playerID, yearID, HR) %>%

group_by(playerID) %>%
mutate(years=max(yearID)-min(yearID)+1, HR=sum(HR)) %>%
filter(HR==0, yearID == 2022) %>%
select(playerID, years) %>%
distinct() %>%
ungroup() %>%
slice_max(years)
= perezol01 (21 years)

Ex 6.25

Consider the `storms` data set in the `dplyr` package, from Example 6.5. Recall that `name` and `year` together identify all storms except Zeta (2005-2006).

a. Which name(s) was/were given to the most storms?
b. Which year(s) had the most named storms?
c. The second strongest storm named Lili had maximum wind speed of 100. Which name's second strongest storm in terms of maximum wind speed was the strongest among all names' second strongest storms? The `dplyr` function `nth` may be useful for doing this problem.

a) storms %>%
distinct(name, year) %>%
group_by(name) %>%
reframe(name, n=n()) %>%
distinct() %>%
slice_max(n=10, n)
= Ana, Bonnie, Claudette, Danielle, Earl (8 storms)

b) storms %>%
distinct(name, year) %>%
group_by(year) %>%
reframe(year, n=n()) %>%
distinct() %>%
slice_max(n=10, n)
= 2020 (20 storms)

c) storms %>%
group_by(name, year) %>%
mutate(wind=max(wind)) %>%
distinct(name, year, wind) %>%
group_by(name) %>%
arrange(name, desc(wind)) %>%
filter(wind==nth(wind, 2)) %>%
distinct(wind) %>%
ungroup() %>%
slice_max(wind)
= Felix (120)

Ex 6.30

Exercises 6.30 – 6.32 require the `babynames` data frame from the `babynames` package. Say that a name is popular if it was given to 1000 or more babies of a single sex. How many popular female names were there in 2015? What percentage of these popular female names ended in the letter 'a'?

```
babynames %>%
filter(year == 2015, sex=="F", n>= 1000) %>%
summarise(n=n())
```
= 325

Ex 6.31

Exercises 6.30 – 6.32 require the `babynames` data frame from the `babynames` package. Consider the `babynames` data set. Restrict to babies born in 2003. We'll consider a name to be gender neutral if the number of male babies given that name is within plus or minus 20% of the number of girl babies given that name. What were the 5 most popular gender neutral names in 2003?

```
filter(babynames, year==2003) %>%
group_by(name) %>%
mutate(ratio=first(n)/last(n), tot =first(n)+last(n)) %>%
filter(ratio <= 1.2, ratio >=0.8, ratio !=1) %>%
ungroup() %>%
slice_max(tot, n=10)
```
= Riley (8318), Peyton (3468), Jessie (1245), Devyn (608), Armani (523)

Ex 6.37

The data set `world_cup` from `fosdata` has the results of all games in the 2014 and 2015 FIFA World Cup soccer finals. From this data, create a data frame which has the total number of goals scored by each team in the 2015 World Cup. Your data frame should have only two variables, `team` and `goals`, and 24 rows, one for each team. Display the entire data frame in descending order of goals scored.

```
home = world_cup %>%
filter(competition == "2015 FIFA Women's World Cup") %>%
group_by(team_1) %>%
mutate(home_goals = sum(score_1)) %>%
select(team_1, home_goals) %>%
rename(team = team_1) %>%
distinct() %>%
ungroup()

away = world_cup %>%
filter(competition == "2015 FIFA Women's World Cup") %>%
group_by(team_2) %>%
mutate(away_goals = sum(score_2)) %>%
select(team_2, away_goals) %>%
rename(team = team_2) %>%
distinct() %>%
ungroup()

full_join(home, away) %>%
rowwise() %>%
mutate(goals = sum(home_goals, away_goals, na.rm=TRUE)) %>%
select(team, goals) %>%
arrange(desc(goals))
```