·3.10 ·8.50
·3.36

Statistica ed Elementi di Probabilità

Autore: Bilotti Alessandro
Matricola: 206409
Data: 10/01/2025

Per ogni simulazione è stato usato `set.seed(123)`, per riproducibilità.

Esercizio 1.1

Let `x <- c(1,2,3)` and `x <- c(6,5,4)`. Predict what will happen when the following pieces of code are run. Check your answer.

a. `x*2`

b. `x*y`

c. `x[1]*y[2]`

A. `[1] 2 4 6`

B. `[1] 6 10 12`

C. `[1] 5`

Esercizio 1.3

Determine the values of the `vector` vec after each of the following commands is run.

a. `vec <- 1:10`

b. `vec <- 1:10 * 2`

c. `vec <- 1:10 ^ 2`

d. `vec <- 1:10 + 1`

e. `vec <- 1:(10 * 2)`

f. `vec <- rep(c(1,1,2), times = 2)`

g. `vec <- seq(from = 0, to = 10, length.out = 5)`

A. `[1] 1 2 3 4 5 6 7 8 9 10`

B. `[1] 2 4 6 8 10 12 14 16 18 20`

C. `[1] 1 4 9 16 25 36 49 64 81 100`

D. `[1] 2 3 4 5 6 7 8 9 10 11`

E. `[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20`

F. `[1] 1 1 2 1 1 2`

G. `[1] 0 2.5 5 7.5 10`

Esercizio 1.4

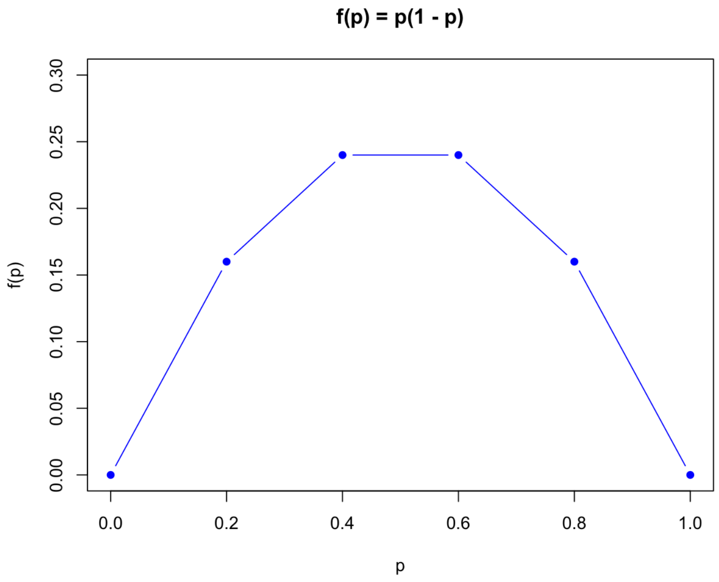
In this exercise, you will graph the function $f(p) = p(1-p)$ for $p \in [0,1]$.

a. Use `seq` to create a vector p of numbers from 0 to 1 spaced by 2.

b. Use `plot` to plot the p in the x coordinate and $p(1-p)$ in the y coordinate. Read the help page for `plot` and experiment with the `type` argument to find a good choice for this graph.

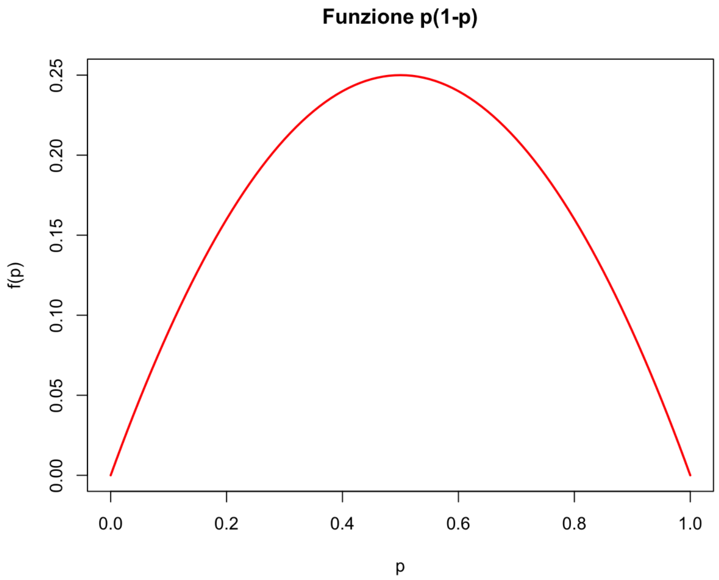
c. Repeat, but with creating a vector p of numbers from 0 to 1 spaced by 0.01.

A. `[1] 0.0 0.2 0.4 0.6 0.8 1.0`



B.

Statistica ed Elementi di Probabilità

C. 

~~D. [1] 1 4 9 16 25 36 49 64 81 100~~

E 1.6

Let x be the vector obtained by running the R command `x <- seq(from = 10, to = 30, by = 2)`.

a. What is the lenght of x? (By length, we mean the number of elements in the vector. This can be obtained using the `str` function or the `lenght` function.)

b. What is `x[2]`?

c. What is `x[1:5]`?

d. What is `x[1:3*2]`?

e. What is `x[1:(3*2)]`?

f. What is `x > 25`?

g. What is `x[x > 25]`?

h. What is `[-1]`?

i. What is `x[-1:-3]`?

A. `[1] 11`

B. `[1] 12`

C. `[1] 10 12 14 16 18`

D. `[1] 12 16 20`

E. `[1] 10 12 14 16 18 20`

F. `[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE`

G. `[1] 26 28 30`

H. `[1] 12 14 16 18 20 22 24 26 28 30`

I. `[1] 16 18 20 22 24 26 28 30`

Esercizio 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 geard or manual transmission?

e. Find the mean mpg of the cars with 2 carburetors.

A.
```
> row.names(mtcars)[mtcars$gear == 4]
[1] "Mazda RX4"      "Mazda RX4 Wag"  "Datsun 710"     "Merc 240D"
"Merc 230"    "Merc 280"        "Merc 280C"      "Fiat 128"
[9] "Honda Civic"    "Toyota Corolla" "Fiat X1-9"      "Volvo 142E"
```

B. Descrive il sottoinsieme in cui `disp`> 150 e `mpg`> 20:

```
> mtcars[mtcars$disp > 150 & mtcars$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
```

C.
```
> row.names(mtcars)[mtcars$gear == 4 & mtcars$am == 1]
    [1] "Mazda RX4"      "Mazda RX4 Wag"  "Datsun 710"     "Fiat 128"
    "Honda Civic"      "Toyota Corolla" "Fiat X1-9"      "Volvo 142E"
```

D.
```
> row.names(mtcars)[mtcars$gear == 4 | mtcars$am == 1]
[1] "Mazda RX4"      "Mazda RX4 Wag"  "Datsun 710"     "Merc 240D"
"Merc 230"      "Merc 280"        "Merc 280C"      "Fiat 128"
[9] "Honda Civic"    "Toyota Corolla" "Fiat X1-9"      "Porsche 914-2"
"Lotus Europa"    "Ford Pantera L" "Ferrari Dino"   "Maserati Bora"
[17] "Volvo 142E"
```

E.
```
> mean(mtcars$mpg[mtcars$carb == 2])
[1] 22.4
```

E 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 `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?

A.
```
> dim(Batting)
[1] 113799     22
```

B.

```
   playerID yearID stint teamID lgID  G AB R H X2B X3B HR RBI SB CS BB SO IBB HBP SH SF GIDP
1 aardsda01   2004     1    SFN   NL 11  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
2 aardsda01   2006     1    CHN   NL 45  2 0 0   0   0  0   0  0  0  0  0   0   0  1  0    0
3 aardsda01   2007     1    CHA   AL 25  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
4 aardsda01   2008     1    BOS   AL 47  1 0 0   0   0  0   0  0  0  0  1   0   0  0  0    0
5 aardsda01   2009     1    SEA   AL 73  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
6 aardsda01   2010     1    SEA   AL 53  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
```

C.
```
> max(Batting$X3B)
[1] 36
```

D.
```
> Batting %>% filter(X3B == max(X3B)) %>% select(playerID, yearID, X3B)
   playerID yearID X3B
1 wilsoch01   1912  36
```

E.
```
> Batting %>% filter(yearID >= 1960) %>% filter(X3B == max(X3B)) %>% select(playerID, yearID
   playerID yearID X3B
1 grandcu01   2007  23
```

Statistica ed Elementi di Probabilità

Esercizio 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 the 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. $P(E) = \frac{\text{Numero esisti positivi}}{\text{Numero esiti totale}} = \frac{3}{30} = \frac{1}{10}$

D. $P(F) = 0$

Esercizio 2.5

Suppose the proportion of M&M's by color is:

Color	**Probability**
Yellow	0.14
Red	0.13
Orange	0.20
Brown	0.12
Green	0.20
Blue	0.21

a. What is the probability that a randomly selected M&M is not green?

b. What is the probability that a randomly selected M&M is red, orange, or yellow?

c. Estimate the probability that a random selection of four M&M's will contain a blue one.

d. Estimate the probability that a random selection of six M&M's will contain all six colors.

A. $P(\text{not green}) = 1 - P(\text{green}) = 1 - 0.20 = 0.80$

B. $P(\text{red, orange, yellow}) = P(\text{red}) + P(\text{orange}) + P(\text{yellow}) =$
$= 0.13 + 0.20 + 0.14 = 0.47$

C. La probabilità che un M&M **non** sia blu è:
$P(\text{not blue}) = 1 - 0.21 = 0.79$

La probabilità che nessuno dei 4 estratti sia blu è: $P(\text{nessuno blue}) = 0.74^4$

La probabilità che almeno uno dei 4 M&M sia blu è il complemento: $P(\text{almeno uno blue}) = 1 - 0.79^4 \approx 0.61$

```
> mm <- c(0.14, 0.13, 0.20, 0.12, 0.20, 0.21)
> mm_colors <- c("yellow", "red", "orange", "brown", "green", "blue")
> mean(replicate(100000, sum(sample(mm_colors, 4, replace = TRUE, prob = mm) == "blue") > 0))
[1] 0.61162
```

D. La probabilità che un M&M abbia uno dei sei colori specifici è:

$$P(\text{tutti colori}) = P(\text{yellow}) \cdot P(\text{red}) \cdot P(\text{orange}) \cdot P(\text{borwn}) \cdot P(\text{green}) \cdot P(\text{blue})$$

$$P(\text{tutti colori}) = 0.14 \cdot 0.13 \cdot 0.20 \cdot 0.12 \cdot 0.20 \cdot 0.21 \approx 0.000092$$

```
> mean(replicate(100000, length(unique(sample(mm_colors, 6, replace = TRUE, prob = mm))) == 6))
[1] 0.01349
```

Esercizio 2.11

Suppose a die is tossed repeatedly, and the cumulative sum of all tosses seen is maintained. Estimate the probability that the cumulative sum ever is exactly 20. (Hint: the function `cumsum` computes the cumulative sums of a vector.)

```
> prove <- 10000
> results <- replicate(prove, {
+   lanci <- sample(1:6, size = 1000, replace = TRUE)
+   cumsum_lanci <- cumsum(lanci)
+   any(cumsum_lanci == 20)
+ })
> prob <- mean(results)
> prob
[1] 0.2833
```

Esercizio 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.

```
> dr <- replicate(10000, {
+   x1 <- sample(1:1000, 1)
+   x2 <- sample(1:x1, 1)
+   x3 <- sample(1:x2, 1)
+   x4 <- sample(1:x3, 1)
+   return(x4 == 1 & x3 > 1)
+ })
> mean(dr)
[1] 0.0463
```

Esercizio 2.29

Ultimate frisbee players are so poor they don't own coins. So, team captains decide which team will play offense first by flipping frisbees before the start of the game. Rather than flip one frisbee and call a side, each team captain flips a frisbee and one captain calls whether the two frisbees will land on the same side, or on different sides. Presumably, they do this instead of just flipping one frisbee because a frisbee is not obviously a fair coin - the probability of one side seems likely to be different from the probability of the other side.

a. Suppose you flip two fair coins. What is the probability they show different sides?

b. Suppose two captains flip frisbees. Assume the probability that a frisbee lands convex side up is p. Compute the probability (in terms of p) that the two frisbees match.

c. Make a graph of the probability of a match in terms of p

d. One Reddit user flipped a frisbee 800 times and found that in practice, the convex side lands up 45% of the time. When captains flip, what is the probability of "same"? What is the probability of "different"?

e. What advice would you give to an ultimate frisbee team captain?

f. Is the two-frisbee flip better than a single-frisbee flip for deciding the offense?

A.
- Ogni moneta ha probabilità $\frac{1}{2}$ di testa(H) o croce(T).
- Spazio campionario HH, HT, TH, TT, ogniuno con probabilità $\frac{1}{4}$.

La probabilità di HT, TH è: 1/4 1/4

$$P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

B. Sia p la probabilità dle fresbee che cada con il lato convesso(C_x) in su.

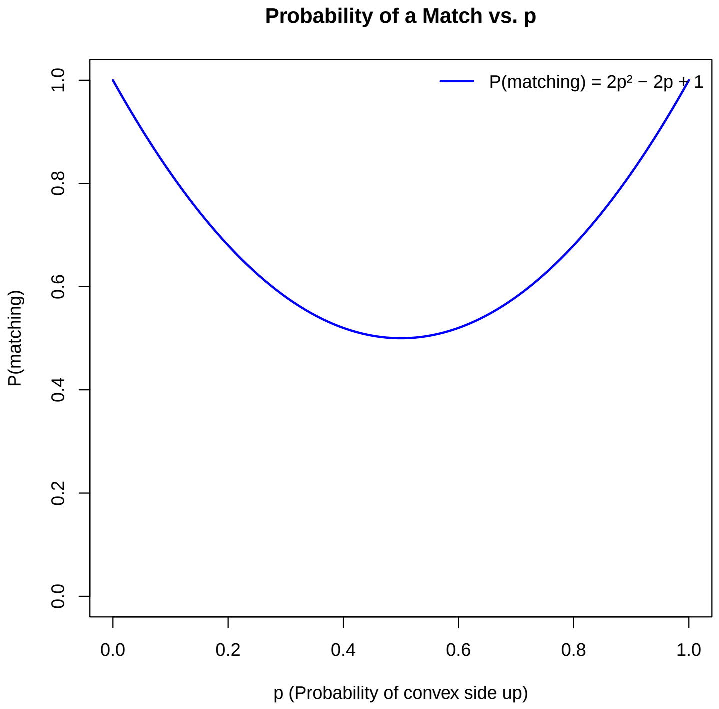
- La probabilità che cada sul lato non convesso(concavo $= C_c$) è $\overline{C} = 1 - p$.

I possibili risultati sono:

- Entrambe sul lato convesso: $P(C_x C_x) = p \cdot p = p^2$,
- Uno convesso, uno concavo: $P(C_x) + P(C_c) = 2 \cdot p \cdot (1 - p)$,
- Entrambe sul lato concavo: $P(C_c C_c) = (1 - p) \cdot (1 - p) = (1 - p)^2$.

La probabilità che i due fresbee abbiano lo stesso risultato è:

$$P(\text{matching}) = P(C_x C_x) + P(C_c C_c) = p^2 + (1 - p)^2$$



C.

D. Con $p = 0.45$, la probabilità che entrambe cadano sullo stesso lato è:

$$P(\text{matching} = \overset{C_x}{(0.45)^2} + (1 - \overset{C_c}{0.45})^2 = 0.505 = 50.5\%$$

E. Il metodo del "double fresbee flip" sembra avere un bias verso il "matching" (probabilità $P(\text{matching}) = 50.5\%, P(\overline{\text{matching}}) = 49.5\%$), quindi se la squadra vuole giocare prima in attacco, dovrebbe chiamare per i due fresbee che atterrano sullo stesso lato.

F. Conoscendo la probabilità p un singolo lancio è vantaggioso($P(C_x) = 0.45$, quindi chiamiamo il lato concavo). Lanciare 2 fresbee avvicina la probabilità a 50%. $P \to 50\%$

LGN $m \to \infty$

Esercizio 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. $\binom{10}{6} = 210$

B. $\binom{5}{4} \cdot \binom{5}{2} = 50$

C. $\frac{50}{210} \approx 0.238$

Esercizio 3.1

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

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

a. Verify that p is a valid probability mass function.

b. Find $P(X \geq 2)$.

c. Find $P(X \geq 2 | X \geq 1)$.

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

A. $1/4 + 1/2 + 1/8 + 1/8 = 1$

B. $P(X \geq 2) = 1/4 = \frac{1}{8} + \frac{1}{8}$

C. $P(X \geq 2 | X \geq 1) = 1/3 = \dfrac{P(X \geq 2 \cap X \geq 1)}{P(X \geq 1)} = \dfrac{1/4}{3/4} = \dfrac{1}{3}$

D. $P(X \geq 2 \cup X \geq 1) = P(X \geq 1) = 3/4$

Esercizio 3.10

In the summer of 2020, the U.S. was considering pooled testing of COVID-19. This problem explores the math behind pooled testing. Since the availability of tests is limited, the testing center proposes the following pooled testing technique:

- Two samples are randomly selected and combined. The combined sample is tested.
- If the combined sample tests negative, then both people are assumed negative.
- If the combined sample tests positive, then both people need to be retested for the disease.

Suppose in a certain population, 5% of the people being tested for COVID-19 actually have COVID-19. Let X be the total number of tests that are run in order to test two randomly selected people.

a. What is the pmf of X?

b. What is the expected value of X?

c. Repeat the above, but imagine that three samples are combined, and let Y be the total number of tests that are run in order to test three randomly selected people. If the pooled test is positive, then all three people need to be retested individually.

d. If your only concern is to minimize the expected number of tests given to the population, which technique would you recommend?

A. $p(x) = \begin{cases} 0.9025 & x = 1 \\ 0.0975 & x = 3 \end{cases}$

B. $E[X] = 1 \cdot 0.9025 + 3 \cdot 0.0975 = 1.195$

C.
```
> sim_3_campioni <- replicate(n_simulazioni, {
+   test <- sample(c("positivo", "negativo"), 3, replace = TRUE,
                   prob = c(prob_pos, 1 - prob_pos))
+   if (any(test == "positivo")) {
+     return(4)  # Quattro test
+   } else {
+     return(1)  # Un test
+   }
+ })
> pmf_3_campioni <- table(sim_3_campioni) / n_simulazioni
> print(pmf_3_campioni)
sim_3_campioni
     1      4
0.8602 0.1398
> expected_3_campioni <- mean(sim_3_campioni)
> print(expected_3_campioni)
[1] 1.4194
```

D. $E[2] \leq E[3]$, quindi si consiglia la tecnica a 2 campioni.

Esercizio 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 possibilità di ottenere una ender pearl tramite uno scambio con i piglin è $p = 0.047 = 4.7\%$, su 262 scambi eseguiti. Il numero previsto di ender pearl è quindi:

$$E[\text{ender pearl}] = n \cdot p = 262 \cdot 0.047 = 12.314 \approx 12.3\%$$

B. Troviamo ora la probabilità di ottenere almeno 42 ender pearl dopo 262 scambi utilizzando la Distribuzione Binomiale, per ottenere la probabilità che su n prove indipendenti si abbiano x successi:

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

La probabilità di ottenere 42 o più ender pearl in 262 scambi è:

$$P[X \geq 42] = \sum_{x=42}^{262} \binom{262}{x} \cdot 0.047^x \cdot (1-0.047)^{262-x} = 4.6^{-12}$$

C. Uccisi 305 con probabilità $p = 0.5 = 50\%$ di ottenere una blaze road, ci aspettimo di averne circa:

$$E[\text{blaze road}] = n \cdot p = 305 \cdot 0.5 = 152.5$$

D. La probabilità di ottenere 211 blaze road su 305 uccisioni è:

$$P[X \geq 211] = \sum_{x=211}^{305} \binom{305}{x} \cdot 0.5^x \cdot (1-0.5)^{305-x} = 8.8^{-12}$$

E. I risultati di Dream sono altamente improbabili per entrambi gli eventi. La probabilità di ottenere 42 o più perle di ender e 211 o più bacchette di blaze è praticamente nulla se il gioco non è stato, in qualche modo, modificato. È quindi ragionevole concludere che Dream abbia barato nella speedrun del gioco.

Esercizio 3.24
Roll two ordinary dice and let Y be their sum.

a. Compute the pmf for Y exactly.

b. Compute the mean and standard deviation of Y.

c. Check that the variance of Y is twice the variance for the roll of one die.

A.
```
> lanci <- replicate(10000, sum(sample(1:6, 2, replace = TRUE)))
> pmf <- table(lanci) / length(lanci)
> pmf
lanci
     2      3      4      5      6      7      8      9     10     11     12
0.0272 0.0599 0.0820 0.1158 0.1405 0.1595 0.1392 0.1096 0.0834 0.0562 0.0267
```

B.
```
~~> mean(lanci)~~
~~[1] 6.9732~~
> mean(lanci)
[1] 6.9732
> var(lanci)
[1] 5.874869
> sd(lanci)
[1] 2.423813
```

C.
```
> lanci_dado <- sample(1:6, 10000, replace = TRUE)
> var_dado <- var(lanci_dado)
> var_dado
[1] 2.910484
> 2 * var_dado
[1] 5.820968
```

Esercizio 3.30
Suppose that 55% of voters support Proposition A.

a. You poll 200 voters. What is the expected number that support the measure?

b. What is the margin of error for your poll (two standard deviations)?

c. What is the probability that your poll claims that Proposition A will fail?

d. How large a poll would you need to reduce your margin of error to 2%?

A. $E[X] = n \times p = 200 \times 0.55 = 100$

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

B. Margine di errore: $2 \times \sigma$. Quindi:

$$\sigma = \sqrt{\frac{0.55 \times (1 - 0.55)}{200}} \approx 0.0352$$

$$0.0352 \times 2 = 0,0704 \approx 7\%$$

C. Probabilità che il sondaggio affermi che la Proposta A fallirà:

```
> pbinom(99,200,.55)
[1] 0.06807525
```

D. Per ridurre il margine di errore a 2% dobbiamo trovare:

$$2 \times \sqrt{\frac{0.55 \times (1 - 0.55)}{n}} = 0.02$$

Per trovare n usiamo:

```
> p <- 0.55
> MOE <- 0.02
> (p * (1 - p)) / (MOE / 2)^2
[1] 2475
```

$$\frac{P(1-p)}{\left(\frac{0.2}{2}\right)^{2}}$$

Esercizio 3.36

As stated in the text, the pdf of a Poisson random variable $X \sim Pois(\lambda)$ is

$$p(x) = \frac{1}{x!}\lambda^x e^{-\lambda}, \quad x = 0, 1, ...$$

Prove the following:

a. p is a pdf. (You need to show that $\sum_{x=0}^{\infty} p(x) = 1$.)

b. $E(X) = \lambda$. (Show that $\sum_{x=0}^{\infty} xp(x) = \lambda$.)

c. $Var(X) = \lambda$. (Compute $E[X(X-1)]$ by summing the infinite series $\sum_{x=0}^{\infty} x(x-1)p(x)$. Use $E[X(X-1)] = E[X^2] - E[X]$ and Theorem 3.9 to finish.)

Theorem 3.9:

$$Var(X) = E[X^2] - E[X]^2$$

A. Per verificare che p sia una PDF dobbiamo verificare che la somma di tutti i valori di $p(x)$ sia uguale a 1.
Sostituiamo:

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}$$

Facendo attenzione si riconosce essere una "estensione" della serie di Taylor per e^{λ}:

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \quad \text{ovvero} \quad e^x = \sum_{x=0}^{\infty} \frac{x^n}{n!} \quad \forall x$$

Quindi:

$$\sum_{x=0}^{\infty} p(x) = \underset{1 \cdot 1}{e^{-\lambda} \cdot e^{\lambda}} = 1 \quad \forall x$$

B. Calcoliamo il valore atteso $E(X) = \lambda$:

$$E(X) = \lambda = \sum_{x=0}^{\infty} xp(x)$$

Sostituiamo $p(x)$ con $\frac{\lambda^x e^{-\lambda}}{x!}$:

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!}$$

Semplifichiamo:

$$E(X) = \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!}$$

Spostiamo l'indice della sommatoria a 0 usando $y = x - 1$ per semplificare:

$$E(X) = \sum_{y=0}^{\infty} \frac{\lambda^{y+1} e^{-\lambda}}{y!}$$

Portiamo fuori λ con:

$$E(X) = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$$

Come prima si semplifica grazie a Taylor:

$$E(X) = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

$$\lambda \; 1 \cdot 1$$

C. Calcoliamo $Var(X) = \lambda$:
Usando il teorema 3.9:

$$Var(X) = E[X^2] - E[X]^2$$

Abbiamo già trovato $E[X] = \lambda$, quindi calcoliamo $E[X^2]$, semplifichiamo:

$$E(X^2) = E(X(X-1)) + E(X)$$

Calcoliamo $E(X(X-1))$:

$$E(X(X-1)) = \sum_{x=0}^{\infty} x(x-1)p(x) = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x e^{-\lambda}}{x!}$$

Semplifichiamo i termini:

$$x(x-1) \cdot \frac{\lambda^x}{x!} = \frac{\lambda^x}{(x-2)!}$$

Ottenendo:

$$E(X(X-1)) = e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!}$$

Semplifichiamo la sommatoria come prima utilizzando $y = x - 2$:

$$E(X(X-1)) = \sum_{y=0}^{\infty} \frac{\lambda^{y+2}}{y!}$$

Troviamo λ^2:

$$E(X(X-1)) = \lambda^2 e^{-\lambda} \sum_{y=0}^{\inf} \frac{\lambda^y}{y!}$$

Come prima si semplifica grazie a Taylor:

$$E(X(X-1)) = \lambda^2 e^{-\lambda} \cdot e^{\lambda} = \lambda^2$$

$$\lambda^2 \; 1 \cdot 1$$

Statistica ed Elementi di Probabilità

Esercizio 4.1

Let X be a random variable with pdf given by $f(x) = 2x$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise.

a. Find $P(X \geq 1/2)$.

b. Find $P(X \geq 1/2|X \geq 1/4)$.

A. Controllimo che la pdf sia normalizzata:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Calcolando:

$$\int_0^1 2x\ dx = [x^2]_0^1 = 1^2 - 0^2 = 1$$

La pdf è normalizzata, quindi procediamo trovando $P(\geq 1/2)$:

```
pdf <- function(x) {
  ifelse(x >= 0 & x <= 1, 2 * x, 0)
}
# P(X >= 1/2)
p_x_greater_than_1_2 <- integrate(pdf, 1/2, 1)$value
cat("P(X >= 1/2):", p_x_greater_than_1_2, "\n")
```

B. Calcoliamo ora $P(X \geq 1/2|X \geq 1/4)$:

```
> # P(X >= 1/4)
> p_x_greater_than_1_4 <- integrate(pdf, 1/4, 1)$value
> cat("P(X >= 1/4):", p_x_greater_than_1_4, "\n")
P(X >= 1/4): 0.9375
>
> # P(X >= 1/2 | X >= 1/4)
> p_x_given_x_1_4 <- p_x_greater_than_1_2 / p_x_greater_than_1_4
> cat("P(X >= 1/2 | X >= 1/4):", p_x_given_x_1_4, "\n")
P(X >= 1/2 | X >= 1/4): 0.8
```

Esercizio 4.8

For each of the following functions, decide whether the function is a valid pdf, a valid cdf or neither.

a. $h(x) = \begin{cases} 1 & 0 \leq x \leq 2 \\ -1 & 2 \leq x \leq 3 \\ 0 & otherwise \end{cases}$

b. $h(x) = sin(x) + 1$

c. $h(x) = \begin{cases} 1 - e^{-x^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$

d. $h(x) = \begin{cases} 2xe^{-x^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$

A. Può assumere valori negativi, -1 se $2 \leq x \leq 3$, quindi non è valida.

B. $sin(x)$ Non converge, quindi non può essere pdf.

C. È una cdf: monotona non decrescente e limiti:

- $\lim_{x\to-\infty} h(x) = 0$
- $\lim_{x\to+\infty} h(x) = 1$

D. È una pdf:

- $h(x) \geq 0 \forall x$
- $\int_0^{+\infty} 2xe^{-x^2} = 1$

Esercizio 4.10
Let X be a normal rv with mean 1 and standard deviation 2.

a. Find $P(a \leq X \leq a + 2)$ when $a = 3$.

b. Sketch the graph of the pdf of X, and indicate the region that corresponds to your answer in the previous part.

c. Find the value of a such that $P(a \leq X \leq a + 2)$ is the largest.

A. Usiamo a CDF della normale per normalizzare:

$$P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 3)$$

Normalizziamo con:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 1}{2}$$

Sostituiamo con i valori:

$$Z_3 = \frac{3 - 1}{2} = 1$$

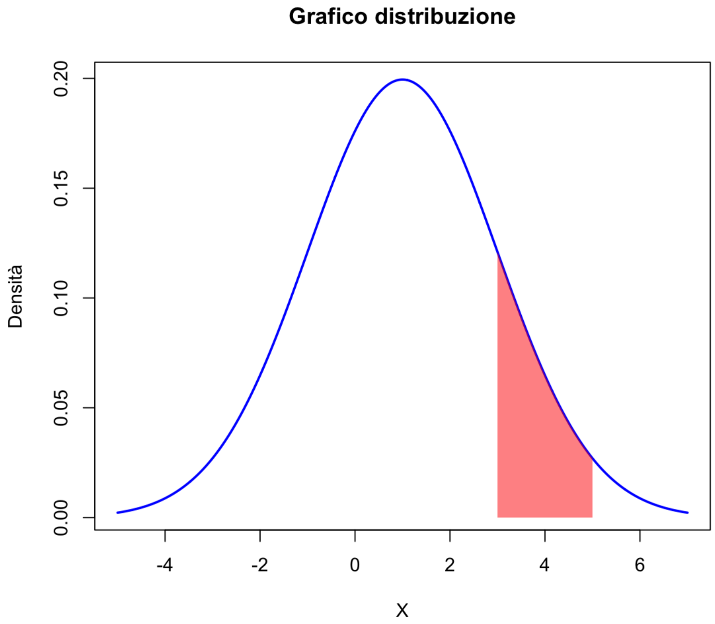
$$Z_5 = \frac{5 - 1}{2} = 2$$

Usando le tabelle troviamo i valori:

$$\Phi(1) \approx 0.84134 \quad \Phi(2) \approx 0.97725$$

Quindi:

$$P(3 \leq X \leq 5) \approx 0.9772 - 0.8413 = 0.1359$$



B.

C. Per trovare il valore che massimizza a possiamo usare:

```
x <- seq(-10, 10, by=0.1)
probs <- pnorm(x + 2, mean=1, sd=2) - pnorm(x, mean=1, sd=2)
max_a <- x[which.max(probs)]
print(max_a)
[1] 0
```

Esercizio 4.19

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?

Abbiamo una distribuzione esponenziale:

$$P(T > t) = e^{-\lambda t} = e^{-t/5}$$

Sappiamo che il sistema fallisce solo se entrambi i componenti falliscono entro 10 anni.

Definiamo i tempi di fallimento come T_1 e T_2. E stimiamo la probabilità che falliscano entro i 10 anni:

$$P(T_1 \leq 10 \cup T_2 \leq 10) = P(T_1 \leq 10) \times P(T_2 \leq 10)$$

Sapendo che $T_1 \sim Exp(1/5)$, la sua CDF è:

$$P(T_i \leq 10) = 1 - P(T_i > 10) = 1 - e^{-10/5} = 1 - e^{-2}$$

Quindi la probabilità che entrambi falliscano è:

$$P(T_1 \leq 10) \times P(T_2 \leq 10) = (1 - e^{-2})^2$$

Semplificando:

$$(1 - 0.1353)^2 = (0.8647)^2 \approx 0.75 = 75\%$$

Esercizio 4.22

For each of the following descriptions of a random variable, indicate whether it can best be modeled by binomial, geometric, Poisson, uniform, exponential, or normal. Answer the associated questions. Note that not all of the experiments yield random variables that are exactly of the type listed above, but we are asking about reasonable modeling.

a. Let Y be the random variable that counts the number of sixes which occur when a die is tossed 10 times. What type of random variable is Y? What is $P(Y = 3)$? What is the expected number of sixes? What is $Var(Y)$?

b. Let U be the random variable which counts the number of accidents which occur at an intersection in one week. What type of random variable is U? Suppose that, on average, 2 accidents occur per week. Find $P(U = 2), E(U)$ and $Var(U)$. Suppose a stop light has a red light that lasts for 60 seconds, a green light that lasts for 30 seconds, and a yellow light that lasts for 5 seconds. When you first observe the stop light, it is red. Let X denote the time until the light turns green. What type of rv would be used to model X? What is its mean?

c. Customers arrive at a teller's window at a uniform rate of 5 per hour. Let X be the length in minutes of time that the teller has to wait until they see their first customer after starting their shift. What type of rv is X? What is its mean? Find the probability that the teller waits less than 10 minutes for their first customer.

d. A coin is tossed until a head is observed. Let X denote the total number of tails observed during the experiment. What type of rv is X? What is its mean? Find $P(X \leq 3)$.

e. Let X be the recorded body temperature of a healthy adult in degrees Fahrenheit. What type of rv is X? Estimate its mean and standard deviation, based on your knowledge of body temperatures.

A. **Binomiale**: la variabile Y conta i numeri di successi in un determinato numero di lanci. Ogni lacio è indipendente dagli altri, e ci sono due soli risultati: 6(successo) o $\overline{6}$(non 6, insuccesso).

$$P(Y = x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}$$

$$P(Y = 3) = \binom{10}{3} \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^7 = 0.155$$

Per il numero previsto di 6:

$$E(Y) = n \cdot p = 10 \cdot \frac{1}{6} \approx 1.67$$

Per la varianza:

$$Var(Y) = n \cdot p \cdot (1-p) = 10 \cdot \frac{1}{6} \cdot \frac{5}{6} \approx 1.39$$

B. **Poisson**: la distribuzione di Poisson viene utilizzata per contare eventi che accadono in modo indipendente a un tasso medio costante. Con $\lambda = 2$ incidenti medi a settimana:

$$P(U = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(U = 2) = \frac{2^2 e^{-2}}{2!} \approx 0.27$$

Per il valore atteso:

$$E(U) = \lambda = 2$$

Per la varianza:

$$Var(U) = \lambda = 2$$

C. **Esponenziale**: il tempo fino a quando il semaforo diventa verde segue una distribuzione esponenziale, poiché l'evento (il semaforo che diventa verde) accade a un tasso costante e il tempo tra eventi successivi è senza memoria.

La media di una v.a. esponenziale con tasso $\lambda = 1/\text{tempo medio}$ è:

$$\text{Media}: \frac{1}{\lambda}$$

Esercizio 4.28

This problem was reported to be a Google interview question. Suppose you have a stick of length one meter. You randomly select two points on the stick, and break the stick at those two places. Estimate the probability that the resulting three pieces of stick can be used to form a triangle.

```
> n_sim <- 100000
> res <- replicate(n_sim, {
+   x <- runif(1)
+   y <- runif(1)
+   if (x > y) {
+     temp <- x
+     x <- y
+     y <- temp
+   }
+   # disequazioni triangolari
+   cond1 <- (y > 1 - y)
+   cond2 <- (1 - y + x > y - x)
+   cond3 <- (1 - x > x)
+   cond1 & cond2 & cond3
+ })
> probabilità <- mean(res)
> print(probabilità)
[1] 0.24806
```

Statistica ed Elementi di Probabilità

Esercizio 5.1

Let Z be a standard normal random variable. Estimate via simulation $P(Z^2 < 2)$.

```
n_sim <- 100000
z_samples <- rnorm(n_sim)
z_squared <- z_samples^2
prob <- mean(z_squared < 2)
print(prob)
```

Esercizio 5.8

Consider an experiment where 20 balls are placed randomly into 10 urns. Let X denote the number of urns that are empty.

a. Estimate via simulation the pmf of X.

b. What is the most likely outcome?

c. What is the least likely outcome that has positive probability?

A.
```
> simulation <- function(n_balls, n_urns) {
+   balls_in_urns <- sample(1:n_urns, size=n_balls, replace=TRUE)
+   empty_urns <- sum(!1:n_urns %in% balls_in_urns)
+   return(empty_urns)
+ }
> n_simulations <- 10000
> n_balls <- 20
> n_urns <- 10
> empty_urns_simulations <- replicate(n_simulations, simulation(n_balls, n_urns))
> pmf <- table(empty_urns_simulations) / n_simulations
> print(pmf)
empty_urns_simulations
     0      1      2      3      4      5 
0.2101 0.4387 0.2761 0.0679 0.0068 0.0004 
```

B.
```
> most_likely_outcome <- as.integer(names(pmf)[which.max(pmf)])
> most_likely_outcome
[1] 1
```

C.
```
> least_likely_outcome <- as.integer(names(pmf)[which.min(pmf[pmf > 0])])
> least_likely_outcome
[1] 5
```

Statistica ed Elementi di Probabilità

Esercizio 5.16

Let X and Y be independent normal random variables with means $\mu_X = 0, \mu_Y = 8$ and standard deviations $\sigma_X = 3$ and $\sigma_Y = 4$.

a. What are the mean and variance of $X + Y$?

b. Simulate the distribution of $X + Y$ and plot it. Add a normal pdf to your plot with mean and standard deviation to match the density of $X + Y$.

c. What are the mean and standard deviation of $5X - Y/2$?

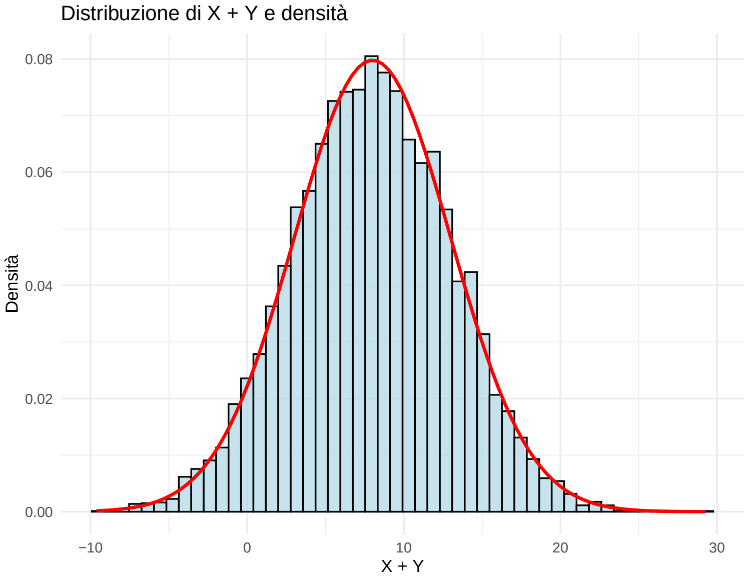
d. Simulate the distribution of $5X - Y/2$ and plot it. Add a normal pdf to your plot with mean and standard deviation to match the density of $5X - 2Y$.

A. Media e varianza di $X + Y$:

$$\text{Media}: \mu_X + \mu_Y = 0 + 8 = 8$$

$$\text{Varianza}: Var(X) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y = 3^2 + 4^2 = 9 + 16 = 25$$

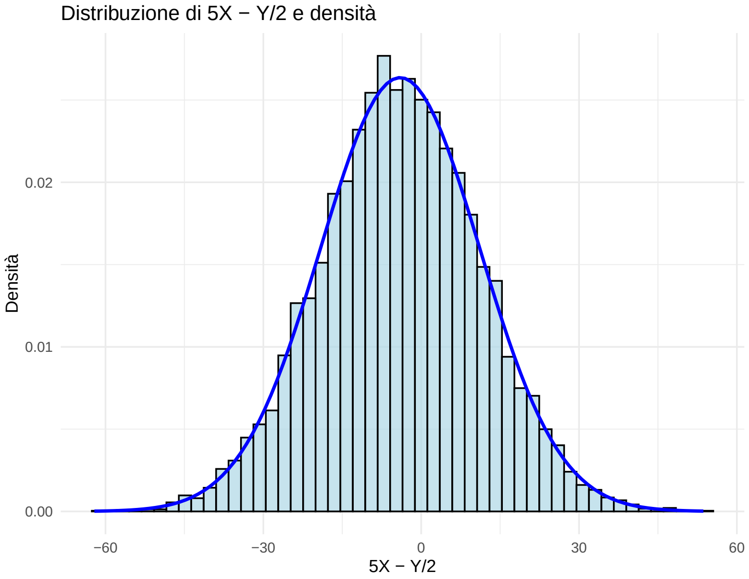
B.



C. Media e variazna di $5X - Y/2$:

$$\text{Media}: 5\mu_X - \frac{\mu_Y}{2} = 5 \cdot 0 - \frac{8}{2} = -4$$

$$\text{Varianza}: Var(X) = \sigma^2_{5X - \frac{\mu_Y}{2}} = 5^2\sigma^2_X + \left(\frac{1}{2}\right)^2 \sigma^2_Y = 25(3^2) + \frac{1}{4}(4^2) = 225 + 4 = 229$$



D.

Esercizio 5.30
Let $X_1, ..., X_n$ be independent uniform rvs on the interval $(0, 1)$.

a. What are the mean μ and the sd σ of a uniform rv on the interval $(0, 1)$?

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

A.

$$\mu = \frac{a+b}{2} = \frac{0+1}{2} = \frac{1}{2}$$

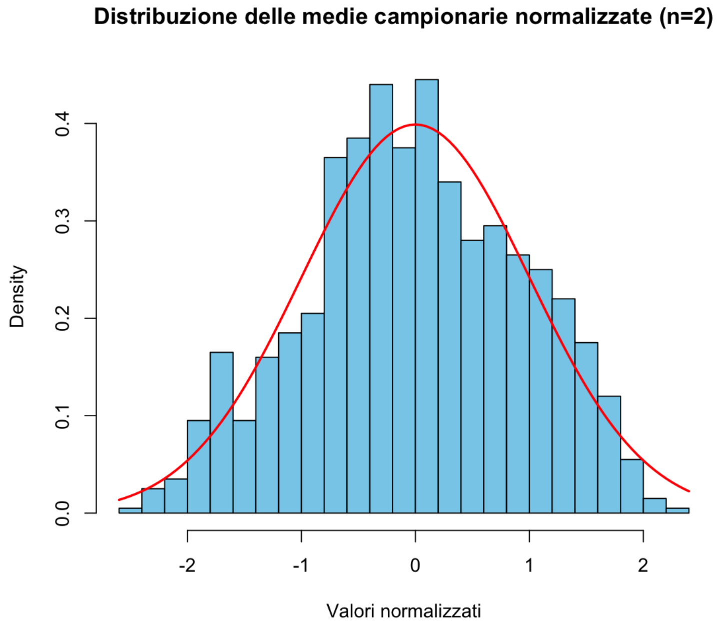
$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(1-0)^2}{12} = \frac{1}{12}$$

Quindi:

$$\sigma = \sqrt{\frac{1}{12}} \approx 0.2887$$

B.

```
> sample_size <- 2
> n <- 1000
> simulate_sample_mean <- function(n, sample_size) {
+   sample_means <- replicate(n, {
+     sample_data <- runif(sample_size)
+     sample_mean <- mean(sample_data)
+     (sample_mean - 0.5) / (sqrt(1/12) / sqrt(sample_size))
+   })
+   return(sample_means)
+ }
```



Esercizio 5.36

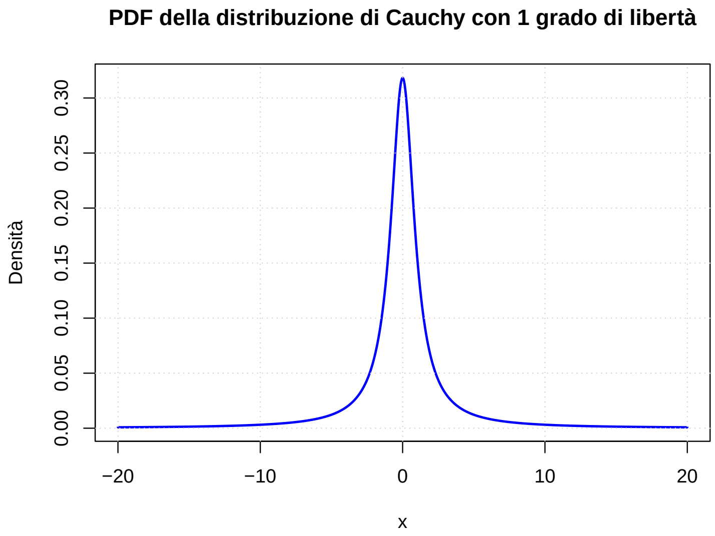
In this exercise, we investigate the importance of the assumption of finite mean and variance in the statement of the Central Limit Theorem. Let $X_1, \ldots, X_n$ be iid random variables with a t distribution with one degree of freedom, also called the Cauchy distribution. You can sample from such a t random variable using `rt(N, df = 1)`.

a. Use `dt(x, 1)` o plot the pdf of a t random variable with one degree of freedom.

b. Confirm for $N = 100, 1000, 10000$ that `mean(rt(N, 1))` does not give consistent results. This is because $\int_{-\infty}^{+\infty} |x| dt(x, 1) dx = \infty$, so the mean of a t random variable with 1 degree of freedom does not exist.

c. Estimate by simulation the pdf of $\overline{X}$ for $X = 100, 1000, 10000$. To visualize this distribution, use a histogram with `breaks = c(-Inf, -20:20, Inf)` and `xlim = c(-20,20)`. Check by adding a curve that $\overline{X}$ has the t distribution with 1 df no matter what N you choose.

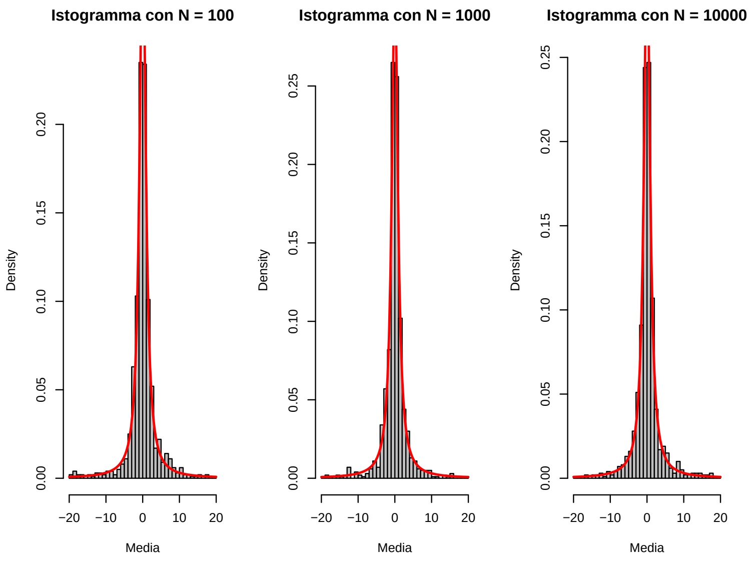
d. Does the Central Limit Theorem hold for this distribution?



A.

B.

```
> N_values <- c(100, 1000, 10000)
> means <- sapply(N_values, function(N) mean(rt(N, df = 1)))
> means
[1]  0.26725462 -1.98145108 -0.00170821
```



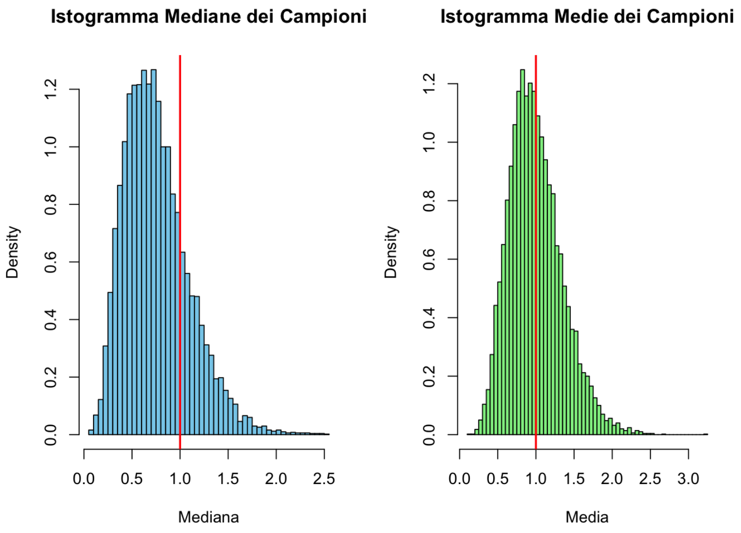
C.

D. Il Teorema del Limite Centrale non vale per la distribuzione di Cauchy, poiché la distribuzione della media campionaria non si avvicina a una distribuzione normale. La distribuzione della media campionaria rimane piuttosto simile alla distribuzione di Cauchy ”originale” anche per un vasto numero di campioni.

Esercizio 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.

```
> lambda <- 1
> sample_size <- 8
> n_simulations <- 10000
>
> medians <- numeric(n_simulations)
> means <- numeric(n_simulations)
>
> for (i in 1:n_simulations) {
+   sample_data <- rexp(sample_size, rate = lambda)
+   medians[i] <- median(sample_data)
+   means[i] <- mean(sample_data)
+ }
```



Esercizio 6.1

The built-in data set `iris` is a data frame containing measurements of the sepals and petals of 150 iris flowers. Convert this data to a tibble with new variables `Sepal.Area` and `Petal.Area` which are the product of the corresponding length and width measurements.

```
> library(tibble)
> library(dplyr)
> iris_tibble <- as_tibble(iris) %>%
+   mutate(
+     Sepal.Area = Sepal.Length * Sepal.Width,
+     Petal.Area = Petal.Length * Petal.Width
+   )
> print(iris_tibble)
# A tibble: 150 × 7
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal.Area Petal.Area
          <dbl>       <dbl>        <dbl>       <dbl> <fct>        <dbl>      <dbl>
 1          5.1         3.5          1.4         0.2 setosa        17.8       0.28
 2          4.9         3            1.4         0.2 setosa        14.7       0.28
 3          4.7         3.2          1.3         0.2 setosa        15.0       0.26
 4          4.6         3.1          1.5         0.2 setosa        14.3       0.3
 5          5           3.6          1.4         0.2 setosa        18         0.28
 6          5.4         3.9          1.7         0.4 setosa        21.1       0.68
 7          4.6         3.4          1.4         0.3 setosa        15.6       0.42
 8          5           3.4          1.5         0.2 setosa        17         0.3
 9          4.4         2.9          1.4         0.2 setosa        12.8       0.28
10          4.9         3.1          1.5         0.1 setosa        15.2       0.15
```

Esercizio 6.4

This question uses the DrinksWages from the HistData package. This data, gathered in 1910, was a survey of people working in various trades (bakers, plumbers, goldbeaters, etc.). The trades are assigned class values of A, B, or C based on required skill. For each trade, the number of workers who drink (`drinks`), the number of sober workers (`sober`), and wage information (`wage`) was recorded. There is also a column $n = drinks +$ sober which is the total number of workers surveyed for each trade.

a. Compute the mean wages for each class, A, B, and C.

b. Find the three trades with the highest proportion of drinkers. Consider only trades with 10 or more workers in the survey.

A.
```
> library(HistData)
> library(dplyr)
> data("DrinksWages")
> mean_wages <- DrinksWages %>%
+   group_by(class) %>%
+   summarise(mean_wage = mean(wage, na.rm = TRUE))
> print(mean_wages)
# A tibble: 3 × 2
  class mean_wage
  <fct>     <dbl>
1 A          20.3
2 B          27.4
3 C          34.0
```

B.
```
> top_drinkers <- DrinksWages %>%
+   filter(n >= 10) %>%
+   mutate(prop_drinkers = drinks / n) %>%
+   arrange(desc(prop_drinkers)) %>%
+   select(trade, prop_drinkers) %>%
+   head(3)
> print(top_drinkers)
    trade prop_drinkers
1  cabmen     0.9090909
2 tailors     0.7894737
3   mason     0.7727273
```

Esercizio 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.
```
> library(Lahman)
> library(dplyr)
> data("Batting")
> most_bats_no_HR <- Batting %>%
+   filter(HR == 0) %>%
+   group_by(playerID) %>%
+   summarise(total_AB = sum(AB, na.rm = TRUE)) %>%
+   arrange(desc(total_AB)) %>%
+   head(1)
> print(most_bats_no_HR)
# A tibble: 1 × 2
  playerID  total_AB
  <chr>        <int>
1 slaglji01     4341
```

B.
```
> latest_year <- max(Batting$yearID, na.rm = TRUE)
> active_players <- Batting %>%
+   filter(yearID == latest_year) %>%
+   select(playerID) %>%
+   distinct()
> most_bats_active_no_HR <- Batting %>%
+   filter(HR == 0, playerID %in% active_players$playerID) %>%
+   group_by(playerID) %>%
+   summarise(total_AB = sum(AB, na.rm = TRUE)) %>%
+   arrange(desc(total_AB)) %>%
+   head(1)
> print(most_bats_active_no_HR)
# A tibble: 1 × 2
  playerID  total_AB
  <chr>        <int>
1 strawmy01      725
```

Esercizio 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.

```
> library(dplyr)
> data("storms")
> most_common_name <- storms %>%
+   count(name, sort = TRUE) %>%
+   filter(n == max(n))
> print(most_common_name)
# A tibble: 1 × 2
  name       n
  <chr>  <int>
1 Bonnie   328
```

B.

```
> most_common_year <- storms %>%
+   count(year, sort = TRUE) %>%
+   filter(n == max(n))
> print(most_common_year)
# A tibble: 1 × 2
   year     n
  <dbl> <int>
1  2005   873
```

C.

```
> second_strongest_storm <- storms %>%
+   group_by(name) %>%
+   arrange(desc(wind)) %>%
+   summarise(second_max_wind = nth(wind, 2, order_by = wind), .groups = "drop") %>%
+   filter(!is.na(second_max_wind)) %>%
+   arrange(desc(second_max_wind))
> strongest_second_storm <- second_strongest_storm %>%
+   slice(1)
> print(strongest_second_storm)
# A tibble: 1 × 2
  name  second_max_wind
  <chr>           <int>
1 Greta              50
```

Esercizio 6.36

Consider the `scotland_births` data set in the `fosdata` package. This data gives the number of births by the age of the mother in Scotland for each year from 1945-2019. This data is in wide format. (Completion of this exercise will be helpful for Exercise 7.28.)

a. Convert the data into long format with three variable names: `age, year and births`, where each observation is the number of `births` in `year` to mothers that are `age` years old.

b. Convert the year to integer by removing the `x` and using `as.integer`.

c. Which year had the most babies born to mothers 20-years-old or younger?

A.

```
> library(tidyverse)
> library(fosdata)
> scotland_births_long <- scotland_births %>%
+   pivot_longer(
+     cols = starts_with("x"),
+     names_to = "year",
+     values_to = "births"
+   )
> scotland_births_long
# A tibble: 3,375 × 3
     age year  births
   <int> <chr>  <int>
 1    12 x1945      0
 2    12 x1946      0
 3    12 x1947      0
 4    12 x1948      0
 5    12 x1949      0
 6    12 x1950      0
 7    12 x1951      1
 8    12 x1952      0
 9    12 x1953      0
10    12 x1954      0
# i 3,365 more rows
# i Use ‘print(n = ...)‘ to see more rows
```

B.

```
> scotland_births_long <- scotland_births_long %>%
+   mutate(year = as.integer(str_remove(year, "x")))
> scotland_births_long
# A tibble: 3,375 × 3
     age  year births
   <int> <int>  <int>
 1    12  1945      0
 2    12  1946      0
 3    12  1947      0
 4    12  1948      0
 5    12  1949      0
 6    12  1950      0
 7    12  1951      1
 8    12  1952      0
 9    12  1953      0
10    12  1954      0
# i 3,365 more rows
# i Use ‘print(n = ...)‘ to see more rows
```

C.

```
> young_mothers <- scotland_births_long %>%
+   filter(age <= 20)
> most_births_year <- young_mothers %>%
+   group_by(year) %>%
+   summarize(total_births = sum(births)) %>%
+   filter(total_births == max(total_births))
> most_births_year
# A tibble: 1 × 2
   year total_births
  <int>        <int>
1  1967        15457
```

Esercizio 6.38

Suppose you sample five numbers from a uniform distribution on the interval $[0, 1]$. Use simulation to show that the expected value of the kth smallest of the five values is $\frac{k}{6}$. That is, the minimum of the five values has expected value 1/6, the second smallest of the values has expected value 2/6, and so on.

```
> n_simulations <- 10000
> n_values <- 5
> expected_values <- numeric(n_values)
> for (i in 1:n_simulations) {
+   random_values <- runif(n_values)
+   sorted_values <- sort(random_values)
+   expected_values <- expected_values + sorted_values
+ }
```

```
> expected_values <- expected_values / n_simulations
> expected_values
[1] 0.1660240 0.3307278 0.4959557 0.6624964 0.8317862
```

Vediamo che il valore minore è il primo, $\frac{1}{6} \approx 0.1660240$, il secondo minore $\frac{2}{6} \approx 0.3307278$, etc...

Esercizio 7.6

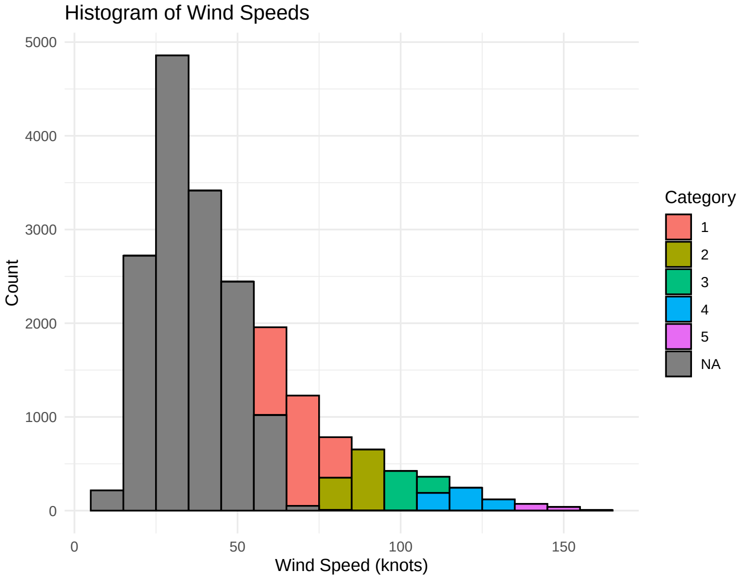
The data set `storms` is included in the `dplyr` package. It contains information about 425 tropical storms in the Atlantic.

a. Produce a histogram of the `wind` speeds in this data set. Fill your bars using the `category` variable so you can see the bands of color corresponding to the different storm categories.

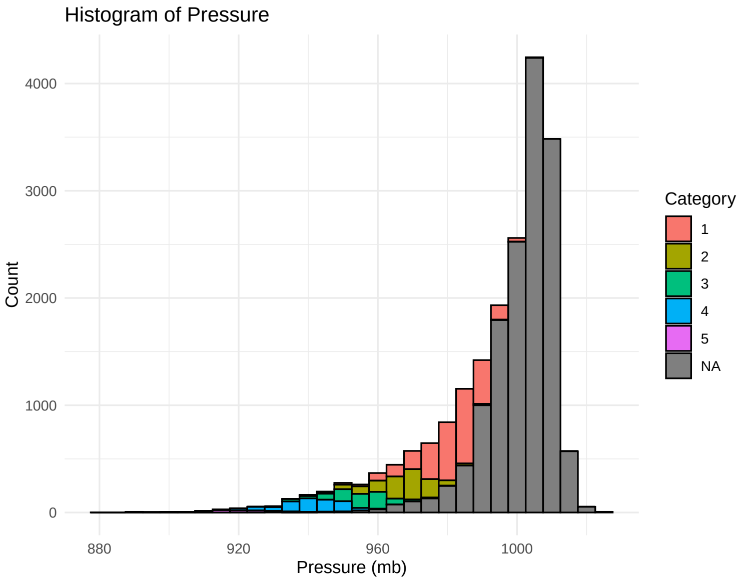
b. Repeat part (a) but make a histogram of the `pressure` variable. You should observe that high category storms have low pressure.

c. Describe the general shape of these two distributions.

d. What type is the `category` variable in this data set? How did that affect the plots?



A.



B.

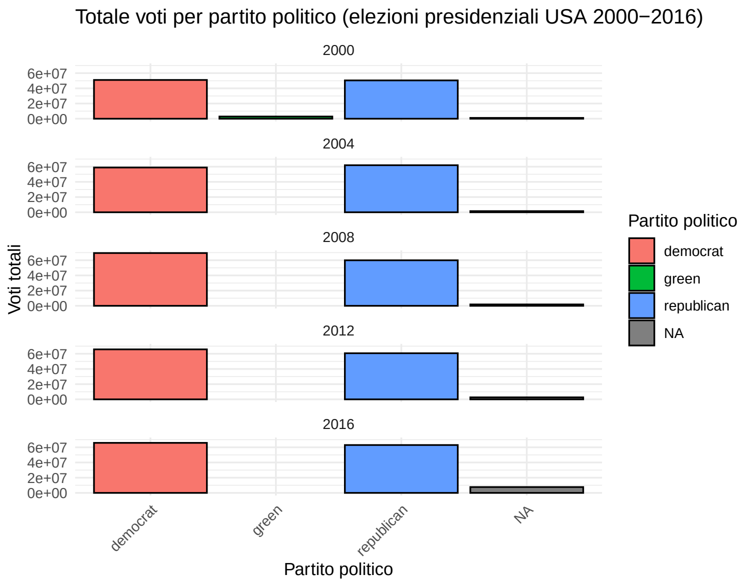
C. **Wind Speeds**: l'istogramma delle velocità del vento mostra una distribuzione verso destra, con la maggior parte delle tempeste che hanno velocità del vento piuttosto basse. Man mano che la categoria aumenta, le velocità del vento tendono a essere più elevate.

   Pressure: l'istogramma della pressione mostra una distribuzione verso sinistra, con la maggior parte delle tempeste che hanno pressioni piuttosto elevate. Le pressioni più basse sono invece associate a tempeste più forti.

D. La variabile `category` è di tipo "numeric", ci aiuta a distinguere i vari valori nel grafico.

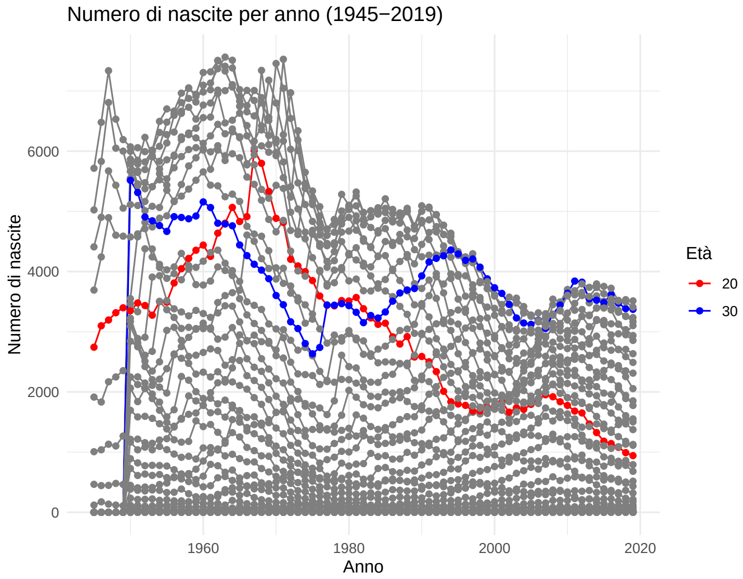
Esercizio 7.15

The `fosdata::pres_election` data set gives voting results from the 2000-2016 U.S. presidential elections. Produce five bar charts, one for each election, that show the total number of votes received by each political party. Use `facet_wrap` to put all five charts into the same visualization.



Esercizio 7.28

Consider the `scotland_births` data set in the `fosdata` package. This data set contains the number of births in Scotland by age of the mother for each year from 1945-2019. In Exercise 7.28, you converted this data set into long format.



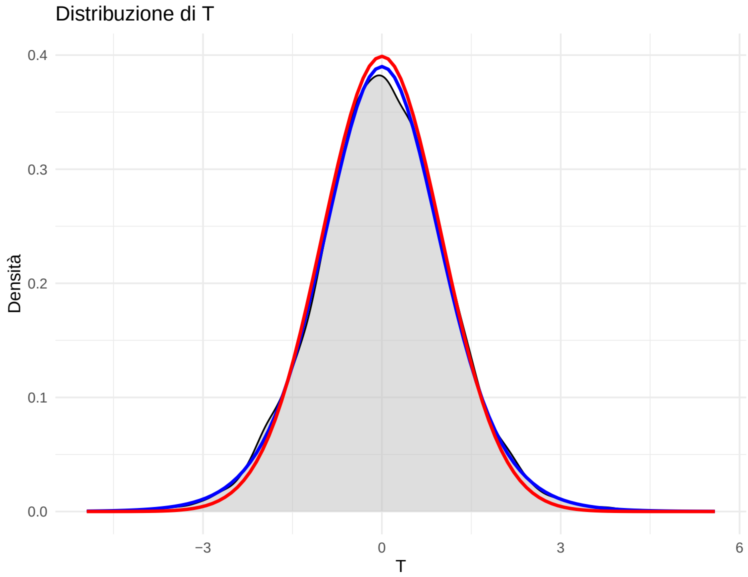
Statistica ed Elementi di Probabilità

Esercizio 8.1

Let $X_1, \ldots, X_n$ be independent normal random variables with mean 1 and standard deviation 3. Simulate 10000 values of

$$T = \frac{\overline{X} - 1}{S/\sqrt{12}}$$

and plot the density function of T . On your plot, add a curve in blue for t with 11 degrees of freedom. Also add a curve in red for the standard normal distribution. Confirm that the distribution of T is t with with 11 df.



Esercizio 8.20

Suppose that a dishonest statistician is doing a t-test of $H_0 : \mu = 0$ at the $\alpha = 0.05$ level. The statistician waits until they get the data to specify the alternative hypothesis. If $\overline{X} > 0$, then they choose $H_\alpha : \mu > 0$ and if $\overline{X} < 0$, they choose $H_\alpha < 0$. Suppose the statistician collects 20 independent samples and the underlying population is standard normal. Use simulation to confirm that the null hypothesis is rejected 10% of the time.

```
> simulation <- function(n_sim = 1000, n_samples = 20, alpha = 0.05) {
+   reject_count <- 0
+   for (i in 1:n_sim) {
+     X <- rnorm(n_samples, mean = 0, sd = 1)
+     sample_mean <- mean(X)
+     if (sample_mean > 0) {
+       t_test_res <- t.test(X, mu = 0, alternative = "greater")
+     } else {
+       t_test_res <- t.test(X, mu = 0, alternative = "less")
+     }
+     if (t_test_res$p.value < alpha) {
```

```
+       reject_count <- reject_count + 1
+     }
+   }
+   err <- reject_count / n_sim
+   return(err)
+ }
> rejected_tare <- simulation(1000, 20, 0.05)
> rejected_tare
[1] 0.099
```

0.099 $\approx$ 10%, conferma che l'approccio dello statistico disonesto porta a rifiutare l'ipotesi nulla in circa il 10% dei casi.

Esercizio 8.50

This problem explores how the $t-$test behaves in the presence of an outlier.

a. Create a data set of 20 random values $x_1, \dots, x_{20}$ with a normal distribution with mean 10 and sd 1. Replace x_{20} with the number 1000. Perform a $t-$test with $H_0 : \mu = 0$, and observe the value of t. It should be close to 1. Is the $t-$test able to find a significant difference between the mean of this data and 0?

   The next parts of this problem ask you to prove that the t-test statistic is always close to 1 in the presence of a large outlier.

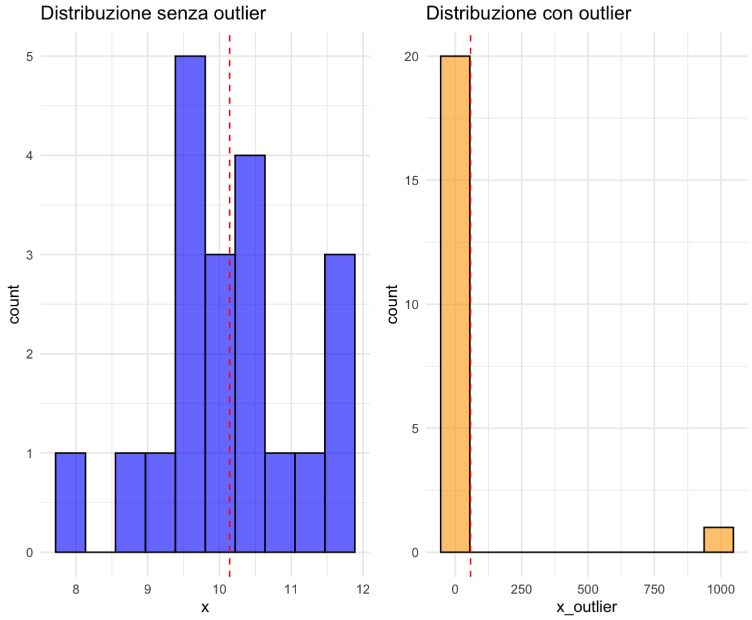
b. Assume that $x_1, \dots, x_{n-1}$ are not changing, but x_n varies. Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as usual. Show that
$$\lim_{x_n \to \infty} \frac{\overline{x}}{x_n} = \frac{1}{n}$$

c. Let the sample variance be $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ as usual. Show that
$$\lim_{x_n \to \infty} \frac{s^2}{x_n^2} = \frac{1}{n}$$

d. Finally show that
$$\lim_{x_n \to \infty} \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = 1$$
where μ_0 is any real number. (Hint: divide top and bottom by x_n , then use parts (b) and (c)).

e. What does this say about the ability of a t-test to reject $H_0 : \mu = \mu_0$ at the $\alpha = .05$ level as $x_n \to \infty$?

Statistica ed Elementi di Probabilità



A. La t-value è circa $1,20$ e il p-value è circa $0,245$. Questo significa che il t-test non rileva una differenza significativa tra la media del dataset e 0. L'outlier (1000) aumenta la deviazione standard, riducendo la capacità del test di individuare una differenza particolarmente significativa.

B. La media campionaria è definita come:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sapendo che $x_n \to \infty$, possiamo "portare fuori" x_n:

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n-1} x_i + x_n\right)$$

Semplifichiamo per x_n e otteniamo:

$$\lim_{x_n \to \infty} \overline{x} = \frac{x_n}{n}$$

C. La varianza campionaria è definita come:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Scomponiamo il termine della sommatoria $(x_i - \overline{x})^2$:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n-1}(x_i - \overline{x})^2 + (x_n - \overline{x})^2\right)$$

Sapendo che $x_n \to \infty$ il termine $(x_n - \overline{x})^2 > (x_i - \overline{x})^2$, quindi:

$$\lim_{x_n \to \infty} s^2 = \frac{x_n^2}{n-1}$$

Semplifichiamo con x_n^2 e otteniamo:

$$\lim_{x_n \to \infty} \frac{s^2}{x_n^2} = \frac{1}{n}$$

D. Sappiamo che il t-test è:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = 1$$

Semplifichiamo con x_n come ci è stato consigliato:

$$t = \frac{\frac{\overline{x}}{x_n} - \frac{\mu_0}{x_n}}{\frac{s}{x_n}/\sqrt{n}}$$

Da (b.): $\lim_{x_n \to \infty} \frac{\overline{x}}{x_n} = \frac{1}{n}$, da (c.): $\lim_{x_n \to \infty} \frac{s}{x_n} = \sqrt{\frac{1}{n}}$. Otteniamo:

$$\lim_{x_n \to \infty} \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = 1$$

E. Quando $x_n \to \infty$ notiamo che:

- $t-$test tende a 1;
- Siccome il valore $p > \alpha$, il $t-$test **non rigetta** l'ipotesi H_0 al livello $\alpha = 0.05$.