Statistica ed elementi di probabilità

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

E 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:

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

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

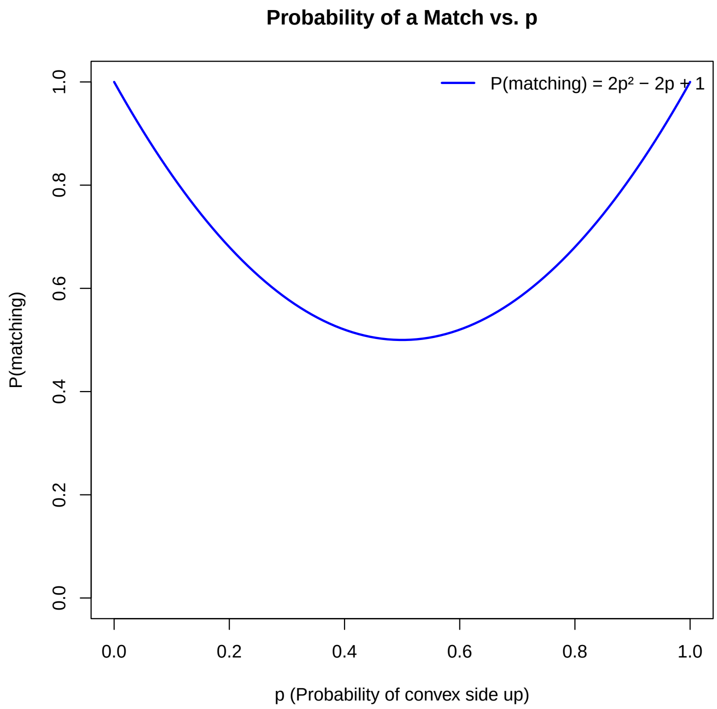
- La probabilità che cada sul lato non convesso(concave $= C_c$) è $\overline{C} = 1 - p$.

I possibili risultati sono:

- Entrambe sul lato convessi: $P(C_x C_x) = p \cdot p = p^2$,
- 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:

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

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%.

E 3.36

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

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

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:

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

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:

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

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:

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

E 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.

Spiegazione di Stand-up Maths.

E 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:

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

D. **Esponenziale**: il tempo tra gli arrivi dei clienti segue una distribuzione esponenziale, poiché gli arrivi avvengono a un tasso costante di 5 clienti all'ora($\lambda = 5$):

$$\text{Media} = \frac{1}{\lambda} = \frac{1}{5} \text{ ore} = 12 \text{ minuti}$$

Troviamo la probabilità che si debba attendere meno di 10 minuti($X < 10$):
La distribuzione cumulativa CDF è:

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

$$P(X < 10) = 1 - e^{-5 \cdot \frac{10}{60}} \approx 0.5$$

E. **Geometrica**: il numero di fallimenti prima della prima testa in una serie di lanci di moneta segue una distribuzione geometrica. Le prove sono indipendenti e si conta il numero di fallimenti prima del primo successo(testa).
Troviamo la media(probabilità di successo $p = 0.5$):

$$E(X) = \frac{1-p}{p} = \frac{1-0.5}{0.5} = 1$$

La probabilità che $P(X \leq 3)$:

$$P(X \leq 3) = (0.5)^1 + (0.5)^2 + (0.5)^3 = 0,875$$

F. **Normale**: la temperatura corporea tipica segue una distribuzione normale, poiché è il risultato di molti piccoli fattori indipendenti che contribuiscono al risultato.
La media è(fonte):

$$98.6F$$

Deviazione standard:

$$\sigma = 0.7F$$

E 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 $X + Y$.

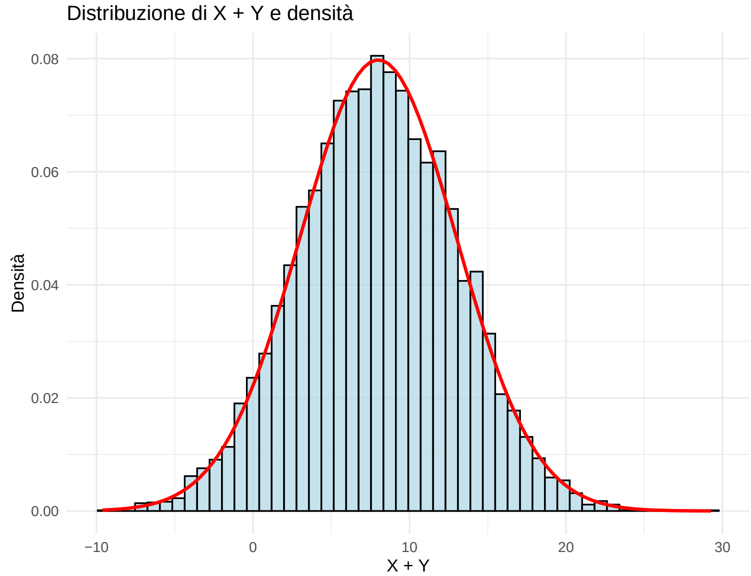
c. What are the mean and standard deviation of $5X - Y/2$?

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 $X + Y$:

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

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

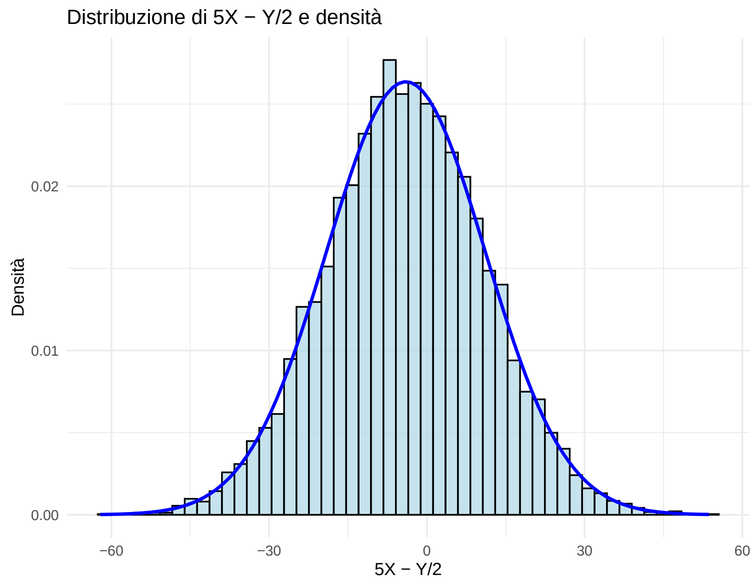


B.

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.

E 5.36

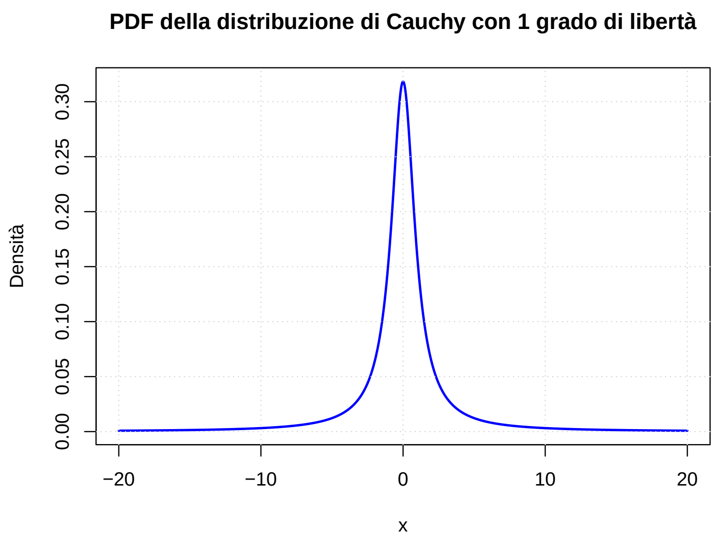
In this exercise, we investigate the importance of the assumption of finite mean and variance in the statement of the Central Limit Theorem. Let $X_1, \ldots, X_n$ be iid random variables with a t distribution with one degree of freedom, also called the Cauchy distribution. You can sample from such a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

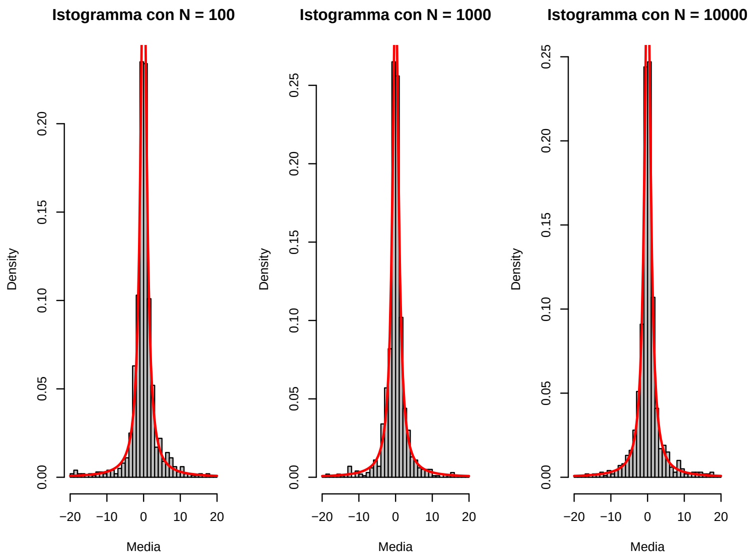
d. Does the Central Limit Theorem hold for this distribution?



A.

B.

```
> cat("Sample mean for N=100:", sample_means[1], "\n")
Sample mean for N=100: 4.133642
> cat("Sample mean for N=1000:", sample_means[2], "\n")
Sample mean for N=1000: 10.54364
> cat("Sample mean for N=10000:", sample_means[3], "\n")
Sample mean for N=10000: -1.501553
```



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.

E 7.6

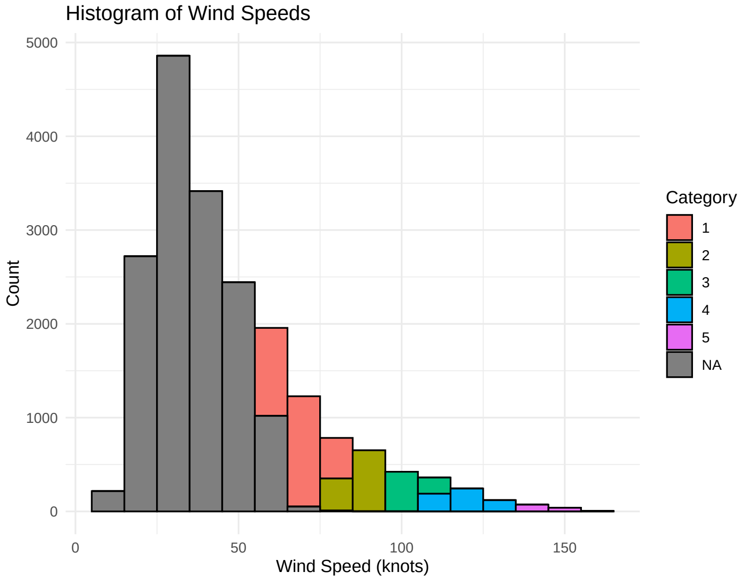
The data set `storms` is included in the `dplyr` package. It contains information about 425 tropical storms in the Atlantic.

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.

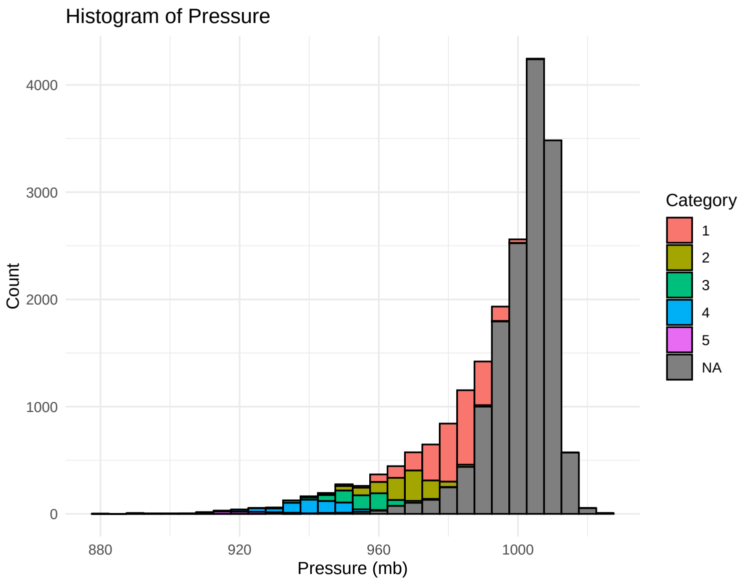
b. Repeat part (a) but make a histogram of the `pressure` variable. You should observe that high category storms have low pressure.

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.

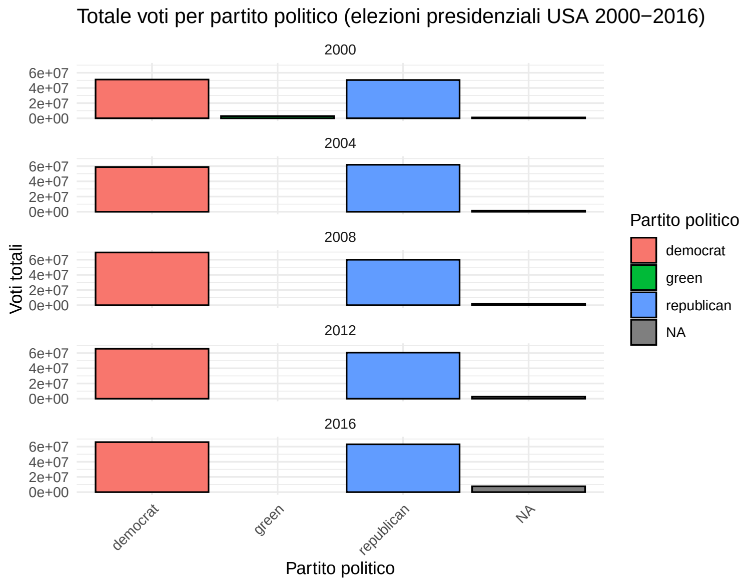
C. **Wind Speeds**: l'istogramma delle velocità del vento mostra una distribuzione verso destra, con la maggior parte delle tempeste che hanno velocità del vento piuttosto basse. Man mano che la categoria aumenta, le velocità del vento tendono a essere più elevate.

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.

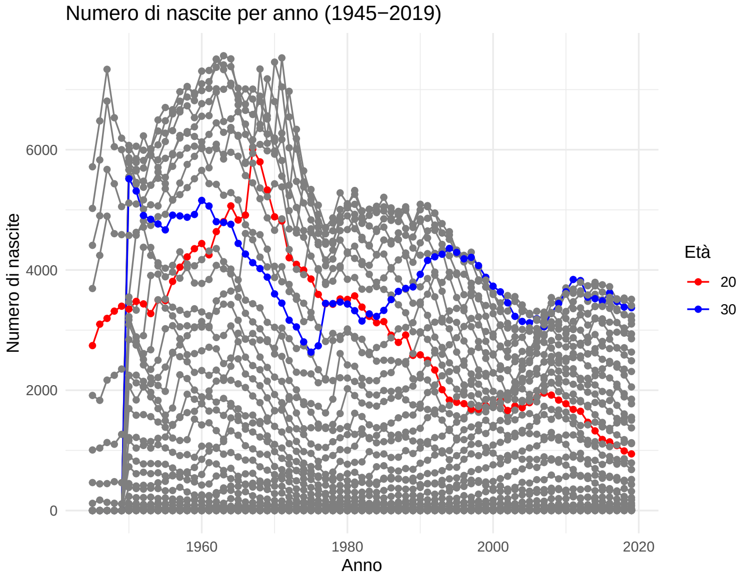
E 7.15

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.



E 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.

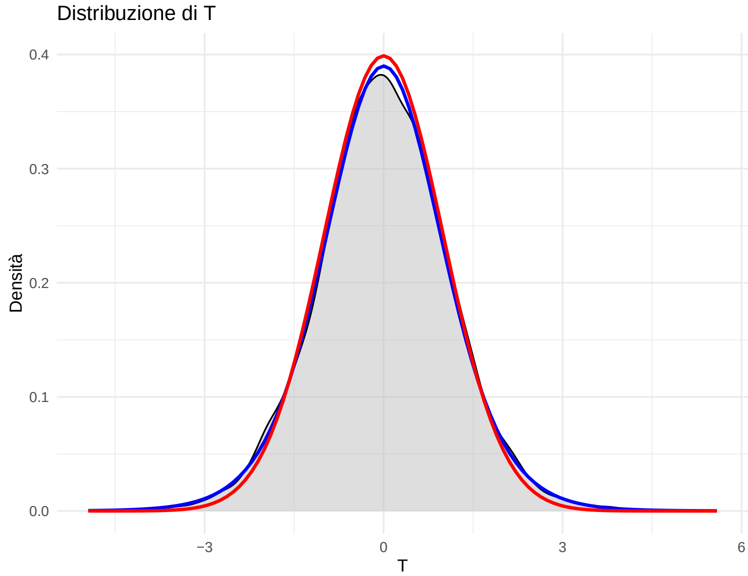


E 8.1

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.



E 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 $x_1, \ldots, x_{20}$ 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 $x_1, \ldots, x_{n-1}$ 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

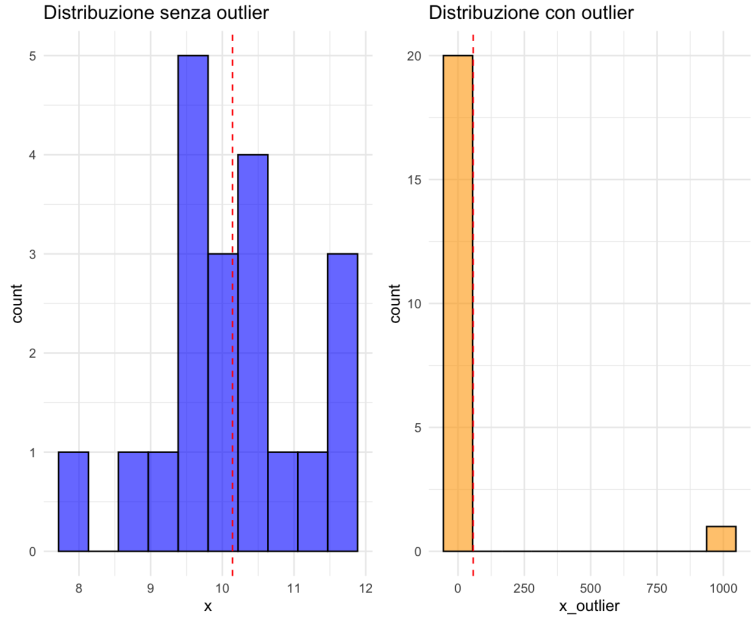
$$\lim_{x_n\to\infty} \frac{s^2}{x_n^2} = \frac{1}{n}$$

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 $x_n \to \infty$?



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 $\alpha = 0.05$.