Segundo Parcial Estadistica

(1937)

Autores:

Valentín Marcial
Facundo Loser
Nicolle Rosatti
Joaquín Baldevenito
Leonardo Campos

Representante:

Nicolle Rosatti

3 de noviembre de 2023 - ESTADÍSTICA

1. OBJETIVOS:

El objetivo principal de este trabajo es la resolución completa de los ejercicios asignados, los cuales consisten en el análisis estadístico sobre una muestra de “Datos de clima y de uso vehicular en autopista Lugones, altura ex-ESMA sentido A, en la Ciudad de Buenos Aires, Argentina, desde el 1 de enero de 2023, hasta el 31 de marzo de 2023”. y sobre la temperatura ambiente en la Patagonia

2. MATERIALES:

Para la realización de este trabajo se utilizó la muestra “datosclima.csv” , en donde los datos fueron analizados usando el lenguaje de programación R, el informe fue realizado por la web de Documentos de Google, y los cálculos representado por “mathcha.io/editor” .

3. METODOLOGÍA:

3.1. Visualización de los Datos

Para aquellos datos del tipo cuantitativo se utilizaron histogramas y diagramas de caja para mayor visualización de los datos. Para aquellos datos cualitativos se utilizaron diagramas de barra y gráfico de torta , y para la comparación entre datos se utilizó diagrama de dispersión

3.2. Descripciones de los Datos

Los datos cuantitativos fueron tomados con la función “summary” para obtener los 5 datos principales de cada uno, también se utilizó un función auxiliar “outliers” para tomar sus valores atípicos, en cambio los datos cualitativos se describieron mediante la función “table”.

Problema 1

Se sabe que la temperatura ambiente, X, en una región de la Patagonia sigue una distribución $N(0,\sigma^2)$

El objetivo de esta investigación es construir un estimador de σ.

a) Mostrar que el estimador de máxima verosimilitud de σ es:

$$\hat{\theta}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

Respuesta:

$$L(\sigma) = \prod_{i=1}^{n} f(x_i\ ,\ \sigma)$$

$$l(\sigma) \quad = \quad \ln(L(\sigma)) \quad = \quad \ln(\prod_{i=1}^{n} f(x_i\ ,\ \sigma))$$

Reemplazamos $f(x_i\ ,\ \sigma)$ por la funcion de densidad de la normal

$$\ln(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{1}{2}*\left(\frac{x_i-\mu}{\sigma}\right)^2})$$

$$= \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{1}{2}*\left(\frac{x_i-\mu}{\sigma}\right)^2})$$

Como $\mu=0$ por Hipotesis:

$$= \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{1}{2}*\left(\frac{x_i}{\sigma}\right)^2})$$

$$= \sum_{i=1}^{n} \ln\left(e^{-\frac{1}{2}*\left(\frac{x_i}{\sigma}\right)^2} \right) - \ln(\sqrt{2\pi\sigma^2})$$

{Propiedad de Logaritmo}

$$= \sum_{i=1}^{n} -\frac{1}{2}*\left(\frac{x_i}{\sigma}\right)^2 - \ln(\sqrt{2\pi\sigma^2})$$

{Cancelativa del Logaritmo}

$$= -\frac{1}{2\sigma^2} * \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \ln(\sqrt{2\pi\sigma^2})$$

{Distributiva de la sumatoria}

$$= -\frac{1}{2\sigma^2} * \sum_{i=1}^{n} x_i^2 - n * \ln(\sqrt{2\pi\sigma^2})$$

Calculo la derivada de L(σ):

$$\mathbf{L(\sigma)'} = \frac{1}{\sigma^3} * \sum_{i=1}^{n} x_i^2 - \frac{n}{\sigma}$$

Calculamos los puntos criticos de L(σ):

$$\frac{1}{\sigma^3} * \sum_{i=1}^{n} x_i^2 - \frac{n}{\sigma} = 0$$

$$\frac{1}{\sigma^3} * \sum_{i=1}^{n} x_i^2 = \frac{n}{\sigma}$$

$$\sum_{i=1}^{n} x_i^2 = \frac{n}{\sigma} * \sigma^3$$

$$\frac{1}{n} * \sum_{i=1}^{n} x_i^2 = \sigma^2$$

$$\sqrt{\frac{1}{n} * \sum_{i=1}^{n} x_i^2} = \widehat{\sigma}_n$$

Encontramos que $\widehat{\sigma}_n$ es un punto critico de L(σ).

Verificaremos que $\widehat{\sigma}_n$ es un maximo de L(σ):

$$\mathbf{L(\sigma)''} = \frac{-3*\sum_{i=1}^{n} x_i^2}{\sigma^4} + \frac{n}{\sigma^2}$$

Probaremos suponiendo que L(σ)" < 0 llegando a una proposicion verdadera:

$$\frac{-3*\sum_{i=1}^{n} x_i^2}{\sigma^4} + \frac{n}{\sigma^2} < 0$$

$$\frac{-3*\sum_{i=1}^{n} x_i^2 + n*\sigma^2}{\sigma^4} < 0$$

$$-3*\sum_{i=1}^{n} x_i^2 + n*\sigma^2 < 0$$

Como n>0 y σ^2> 0 entonces n*σ^2>0 y por lo tanto -n*σ^2<0 :

$$-3*\sum_{i=1}^{n} x_i^2 < -n*\sigma^2 < 0$$

por transitividad de la relacion <:

$$-3*\sum_{i=1}^{n} x_i^2 < 0$$

$$\sum_{i=1}^{n} x_i^2 > 0$$

Esto es una proposicion verdadera ya que es una sumatoria de terminos positivos porque $x_i^2 > 0$

Por lo tanto se concluye que el EMV(σ)= $\sqrt{\frac{1}{n}*\sum_{i=1}^{n} x_i^2}$

b) Asumiendo que n = 30, hallar una estimación de E(σ_n) y del error estándar del estimador, basada en R = 500 réplicas de una N(0, 1) y con una semilla igual a 2.

Respuesta:

> #Cómputo de R=500 cuantiles basadas en muestras con n=0

El estimador de desvío poblacional es:

mean(vec)

0.9949266

El error estándar es:

sd(vec)

0.1312959

(los cálculos correspondientes completos fueron realizados en R)

c) Denotemos con σ $sombrerito_{n,i}$ al valor del estimador para la i-ésima réplica, i = 1, . . . , R.

¿Es la distribución empírica de los valores estimados aproximadamente simétricos? ¿Presenta valores atıpicos? (dar los valores en caso de existir).

Respuesta:

-Como se puede observar la distribución empírica de los datos es aproximadamente simétrica:

Figura N º1: Histograma estimador del desvío muestral

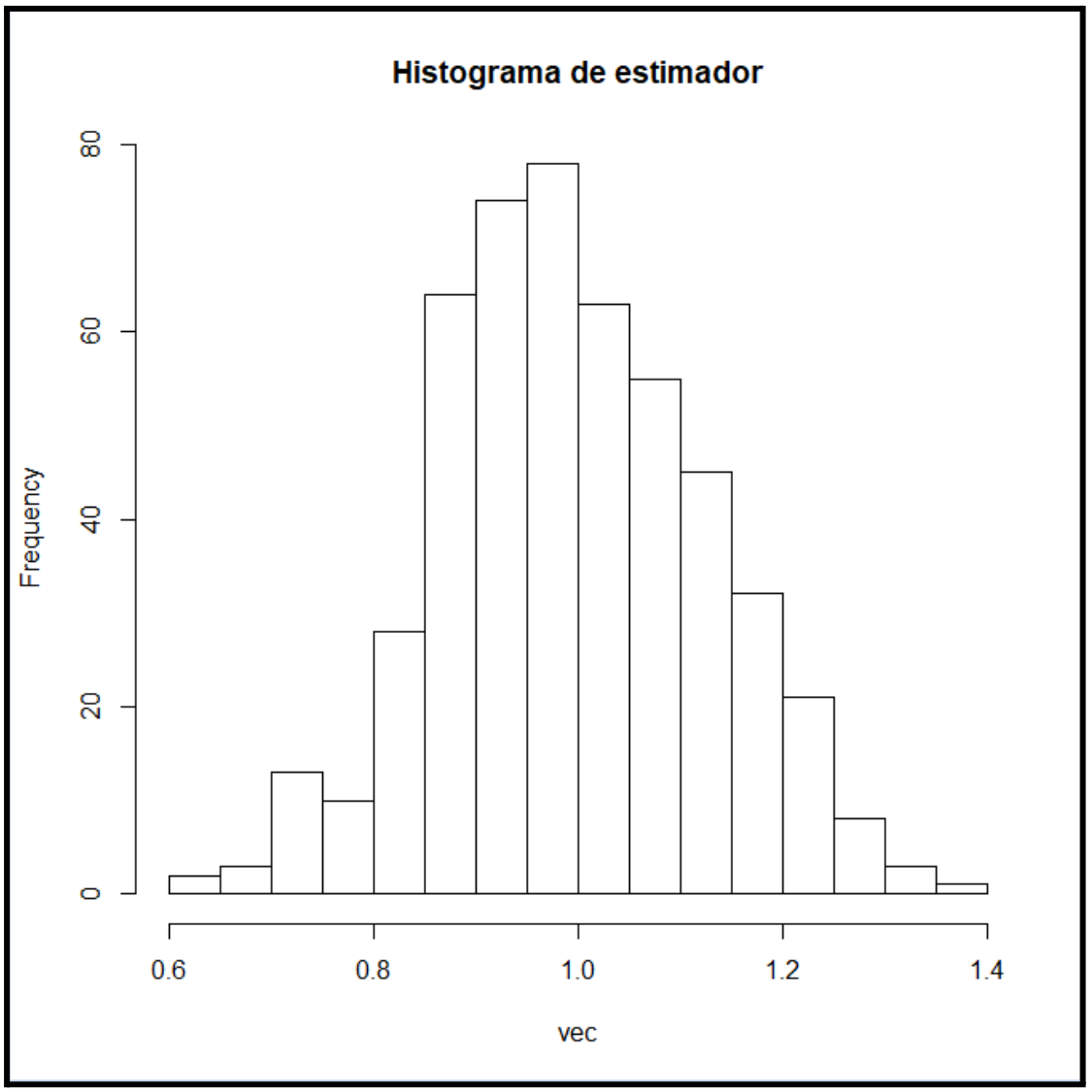
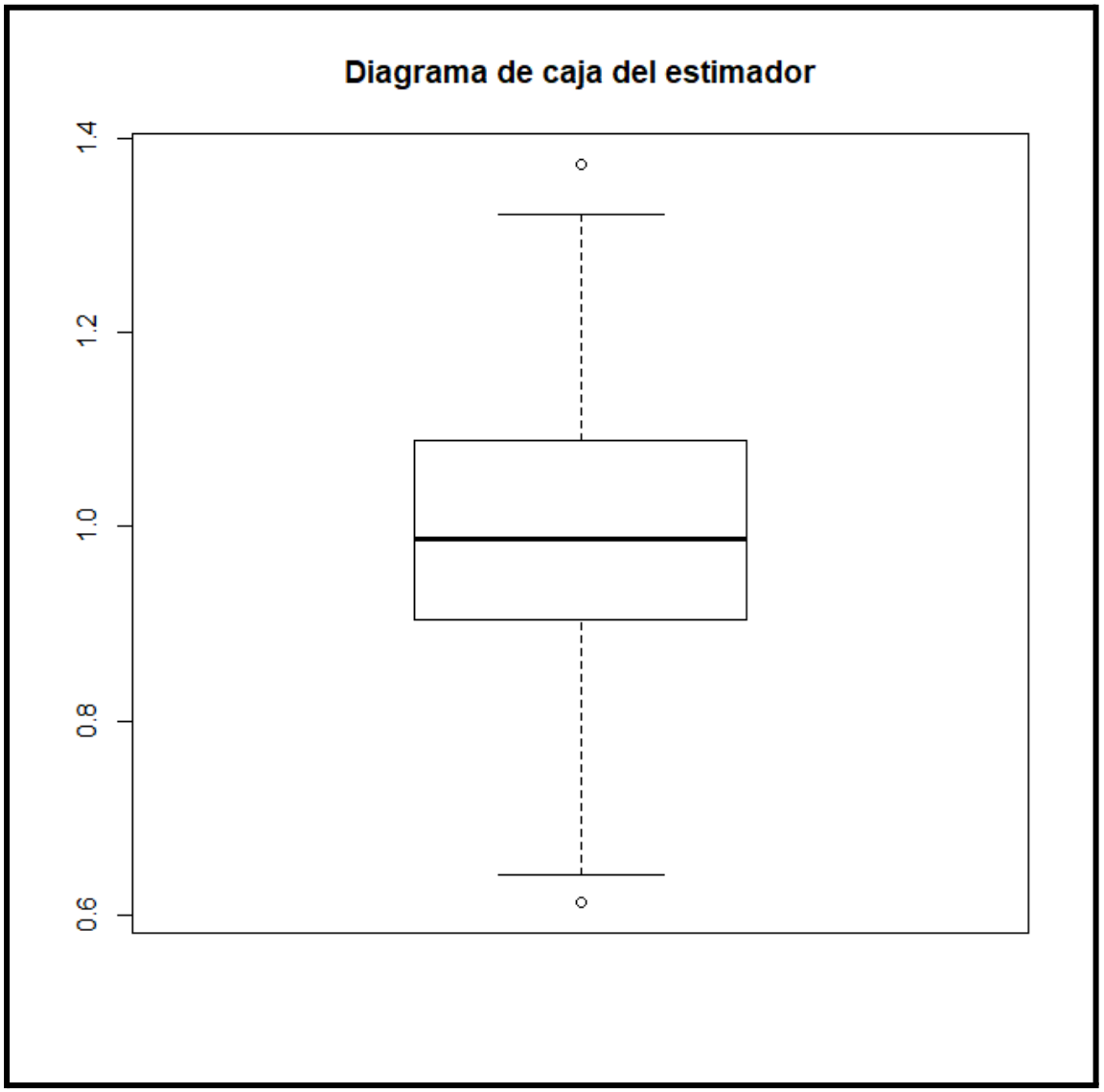


Figura Nº2: Diagrama de caja del estimador del desvío muestral



-En este diagrama de caja se denotan 2 valores atípicos.

en R:

> outliers(vec)

[1] 0.6132335

[2] 1.3735876

Aclaración: La función "outliers" definida en R toma un vector con valores y retorna los valores atípicos, es decir aquellos que se encuentran más alejados que los bigotes. Esto es: aquellos valores que se encuentran más lejos que 1.5 veces el rango intercuartílico.

Problema 2

a) Estadística descriptiva.

Presentar un breve informe tomando como sugerencia los siguiente items:

i) Describir, a través de resúmenes descriptivos y visualizaciones adecuadas, las variables flujo vehicular, temperatura media, precipitaciones, día de la semana, tipo de día.

Respuesta:

Figura Nº3: Histograma para variable de Flujo Vehicular

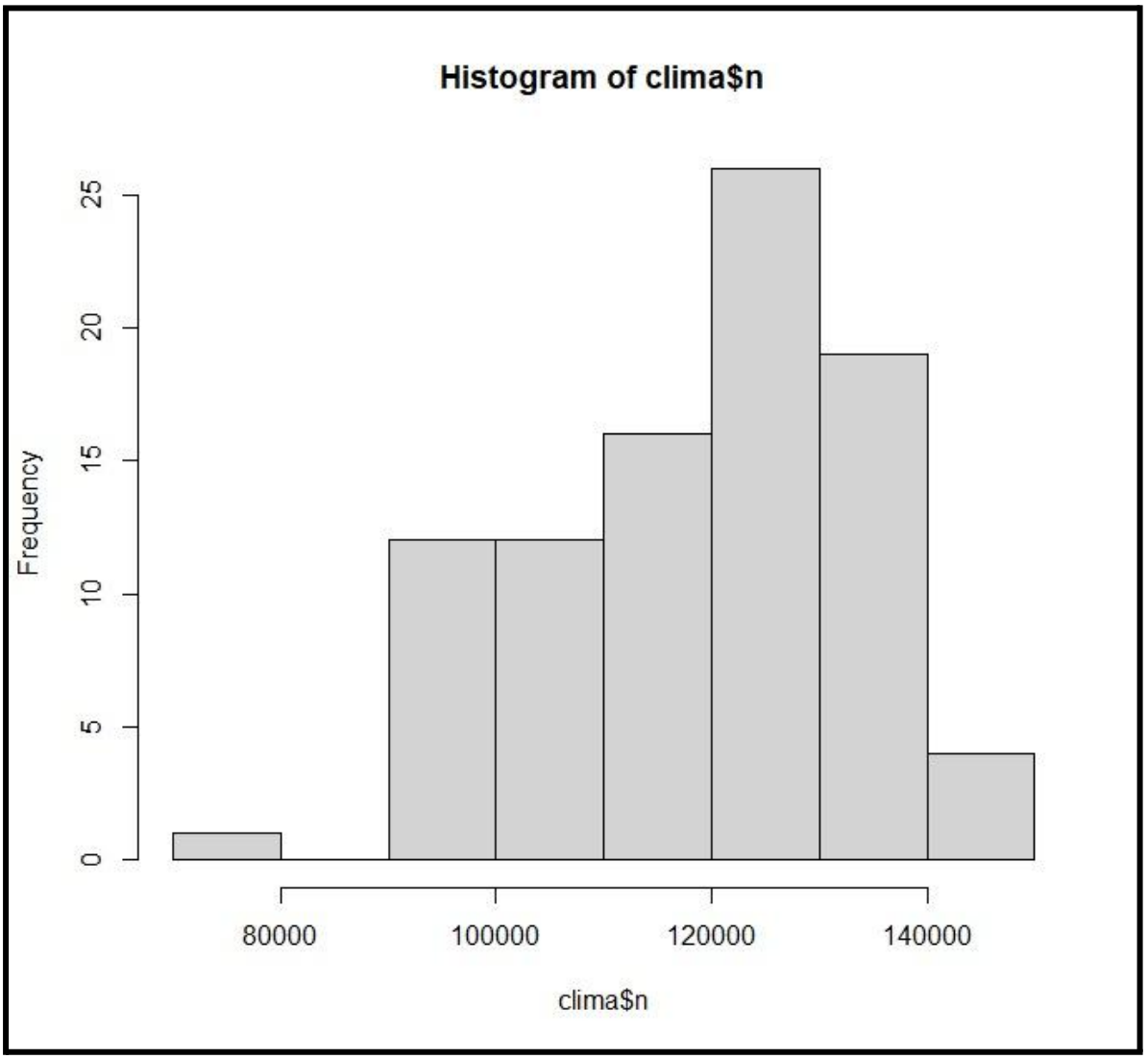
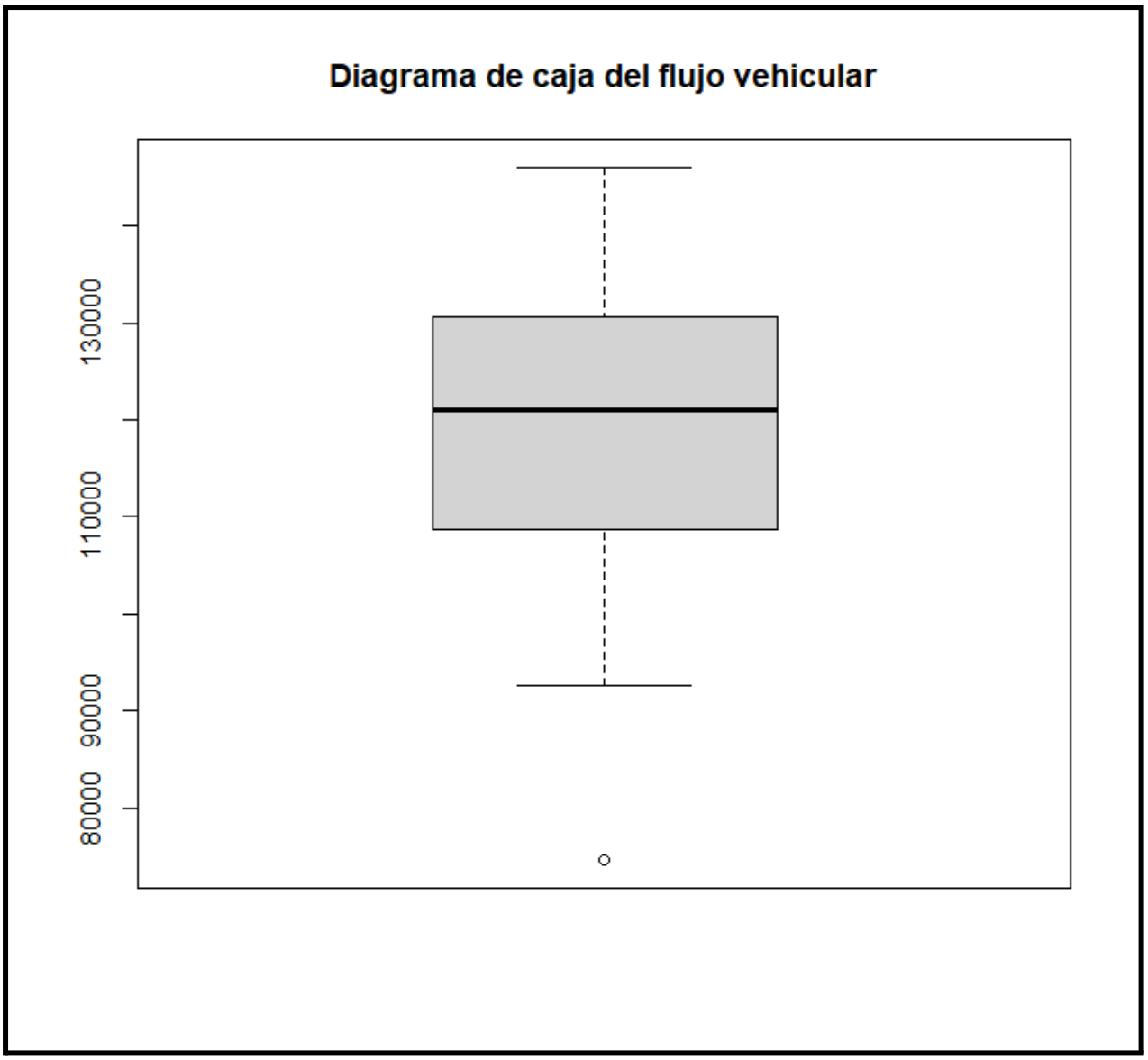


Figura Nº4: Diagrama de caja para variable de Flujo Vehicular



Resumen de los valores:

summary(clima$n) #sumario del flujo vehicular

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
74642	108951	121073	119348	130410	146072

-Mínimo: 74642 **-Primer Cuartil(25%): 108951** **-Media: 119348**

-Máximo: 146072 **-Mediana: 121073** **-Tercer cuartil(75%): 130410**

Figura Nº5: Histograma para variable de Temperatura media

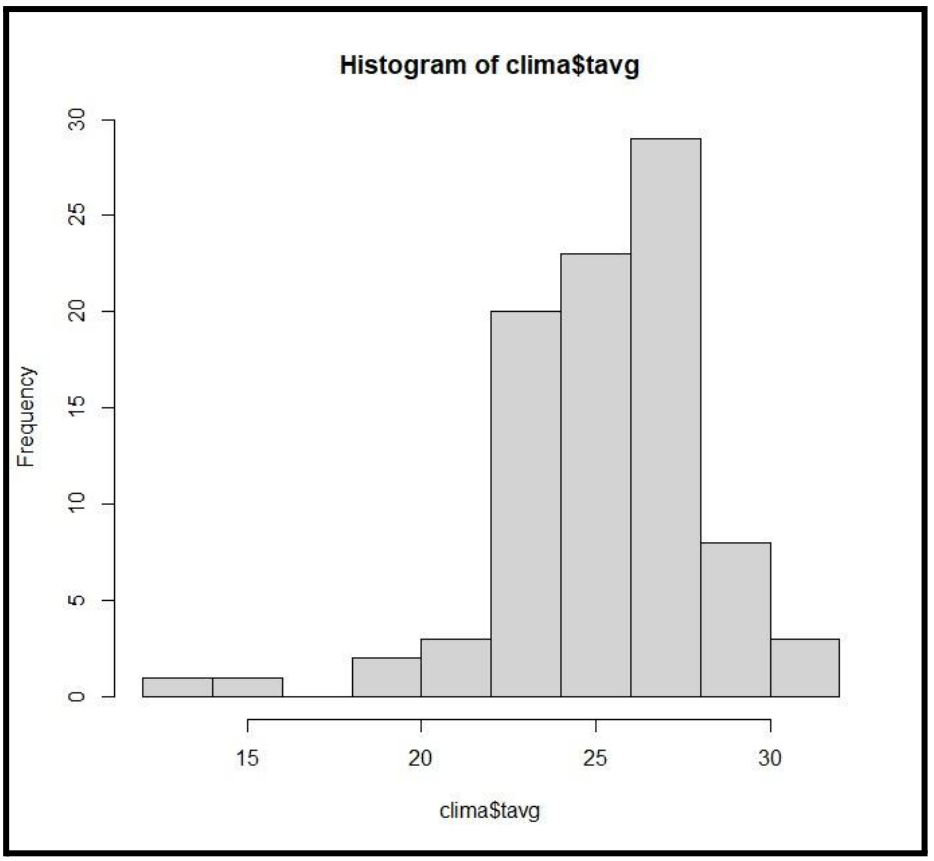
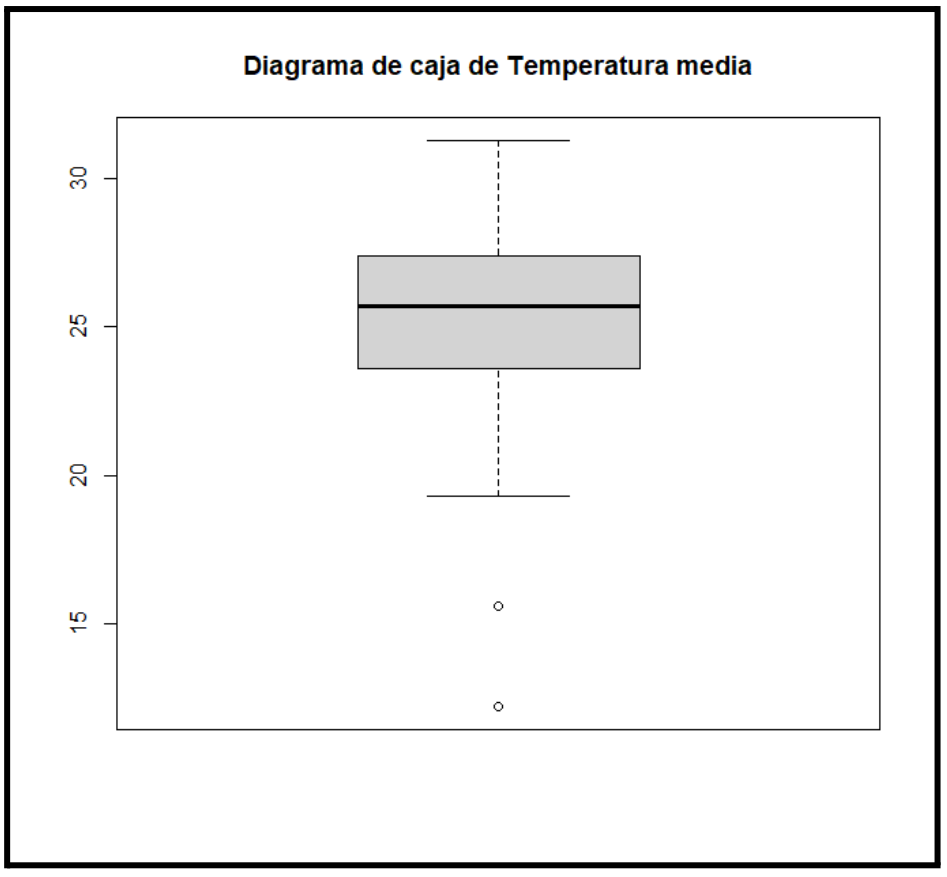


Figura Nº6: Diagrama de caja para variable de Temperatura media



Resumen de los valores:

summary(clima$tavg) #sumario de la temperatura media

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.20	23.65	25.70	25.36	27.38	31.30

-Mínimo: 12.20 -Primer Cuartil(25%): 23.65 -Media: 25.36

-Máximo: 31.30 -Mediana: 25.70 -Tercer cuartil(75%): 27.38

Figura Nº7: Histograma de la variable Precipitación

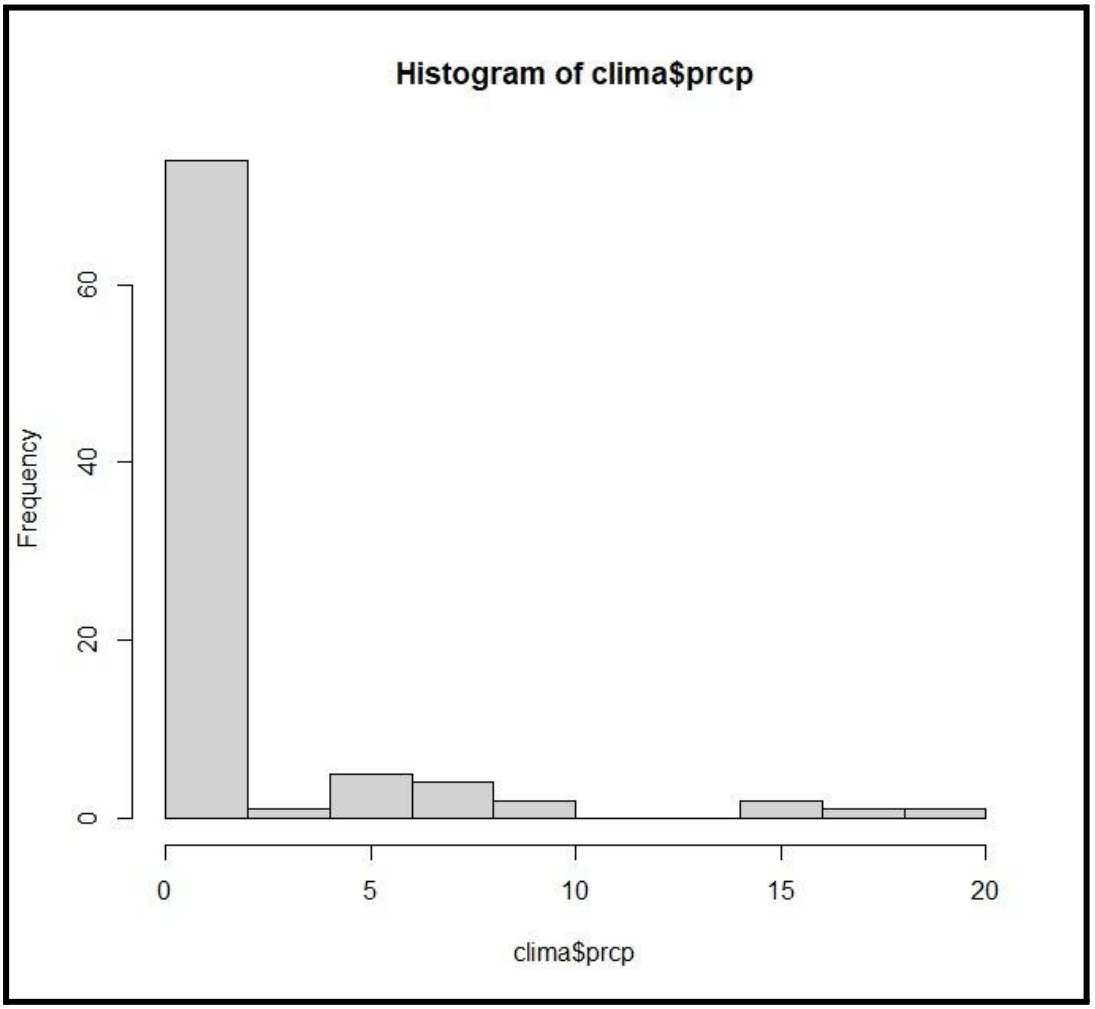
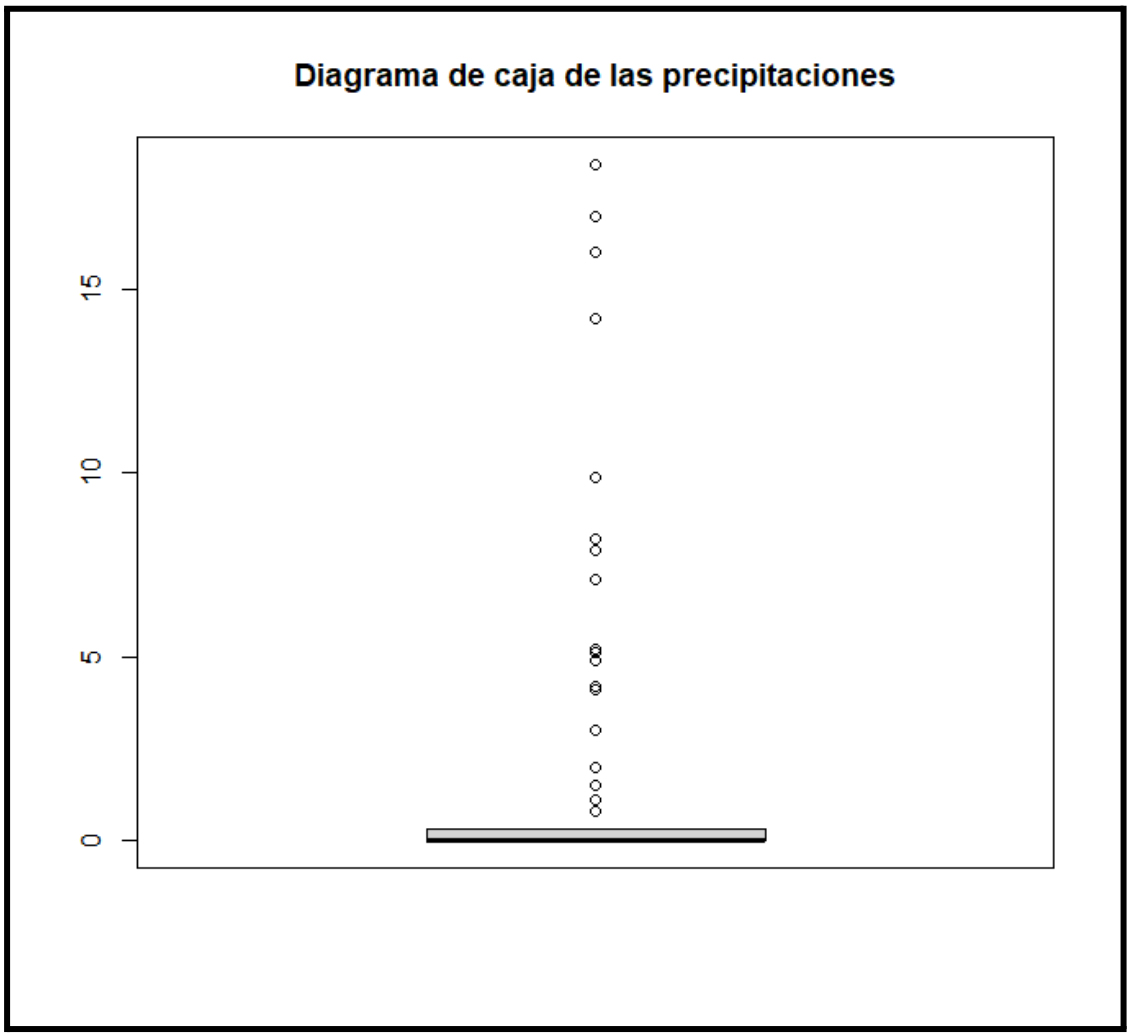


Figura Nº8: Diagrama de caja de la variable Precipitación



Resumen de los 5 valores:

summary(clima$prcp) #sumario de las precipitaciones

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
0.000	**0.000**	**0.000**	**1.630**	**0.275**	**18.400**

-Mínimo: 0.000 -Primer Cuartil(25%): 0.000 -Media: 1.630

-Máximo: 18.400 -Mediana: 0.000 -Tercer cuartil(75%): 0.275

Figura Nº9: Gráfico de torta de la variable Dia de la Semana

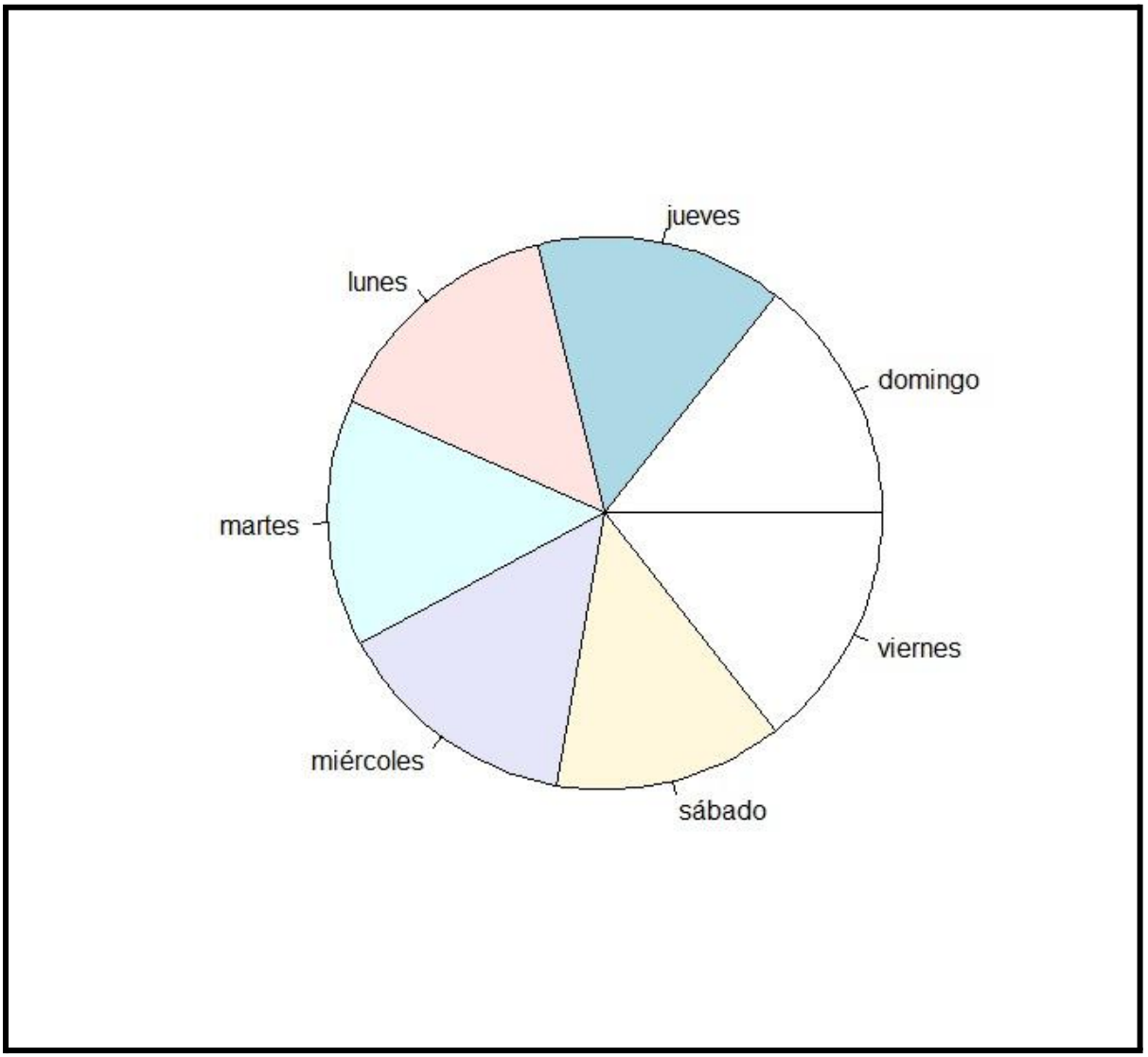


Tabla de frecuencia de la variable Dia de la Semana:

table(clima$dia)

domingo	jueves	lunes	martes	miércoles	sábado	viernes
13	13	13	13	13	12	13

Figura Nº10: Gráfico de torta de la variable Tipo de Dia

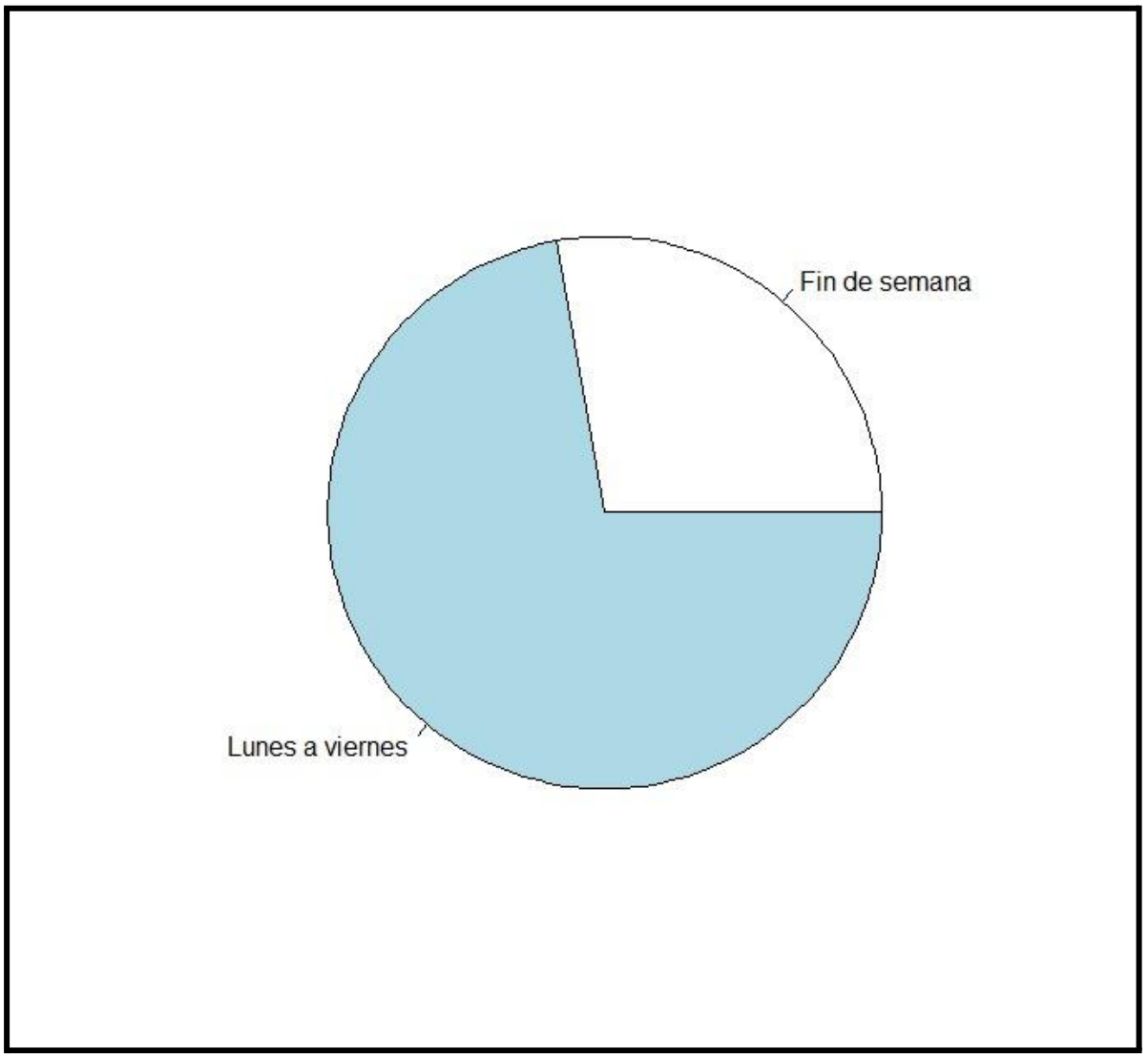


Tabla de frecuencia de la variable Tipo Dia:

table(clima$tipo_dia)

Fin de semana	**Lunes a viernes**
25	**65**

ii) A partir de lo hallado en el ıtem **i)**:

- ¿Cuál es el valor de precipitación tal que por debajo de él se encuentra el 25% de las observaciones?

Respuesta:

-El 25% de los valores de precipitación se encuentran por debajo de 0 mm, es decir, un cuarto de los días tomados no llovió

quantile(prcp, .25, na.rm = TRUE) (cuartil 25% / primer cuartil)

25%

0

- ¿Cuál es el flujo vehicular medio? ¿Cuál es la temperatura media máxima en el período de tiempo registrado?

Respuesta:

El flujo vehicular medio es de 119348 autos por dia, y la temperatura media máxima es de 31.3 ºC

#flujo vehicular medio

mean(clima$n)

119347.8

#Temperatura media máx en el periodo de tiempo registrado

max(clima$tavg)

31.3

◦ ¿Cuántos días hay sin precipitaciones?

Respuesta:

Hay 64 días sin precipitaciones

#días sin precipitaciones

f(clima$prcp)

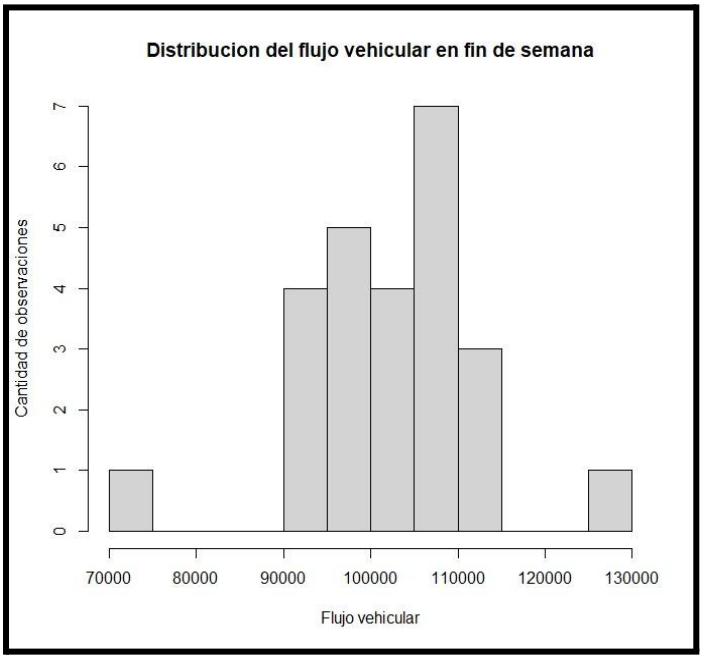
64

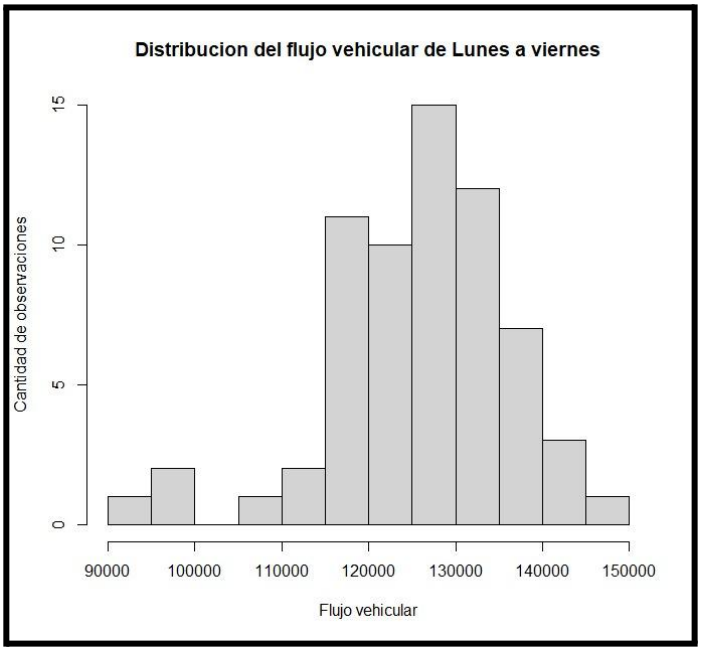
iii) ¿La distribución del flujo vehicular es similar o diferente según el tıpo de dıa? ¿hay valores atıpicos de flujo en cada uno de esos grupos (“Fin de semana” y “Lunes a viernes”)

Respuesta:

-La distribución flujo vehicular es similar en ambos tipos de días, la única diferencia es que la media es diferente

Figura Nº11 y Nº12 : Histogramas del flujo vehicular “Fin de Semana” / “Lunes a viernes”.





-Los valores atípicos de “Lunes a viernes” son: 93178, 95786, 98285 autos/dia y de “Fin de semana” solo 74642 autos/dia, en R:

#valores atípicos

outliers(clima$n[clima$tipo_dia == "Fin de semana"])

74642

outliers(clima$n[clima$tipo_dia == "Lunes a viernes"])

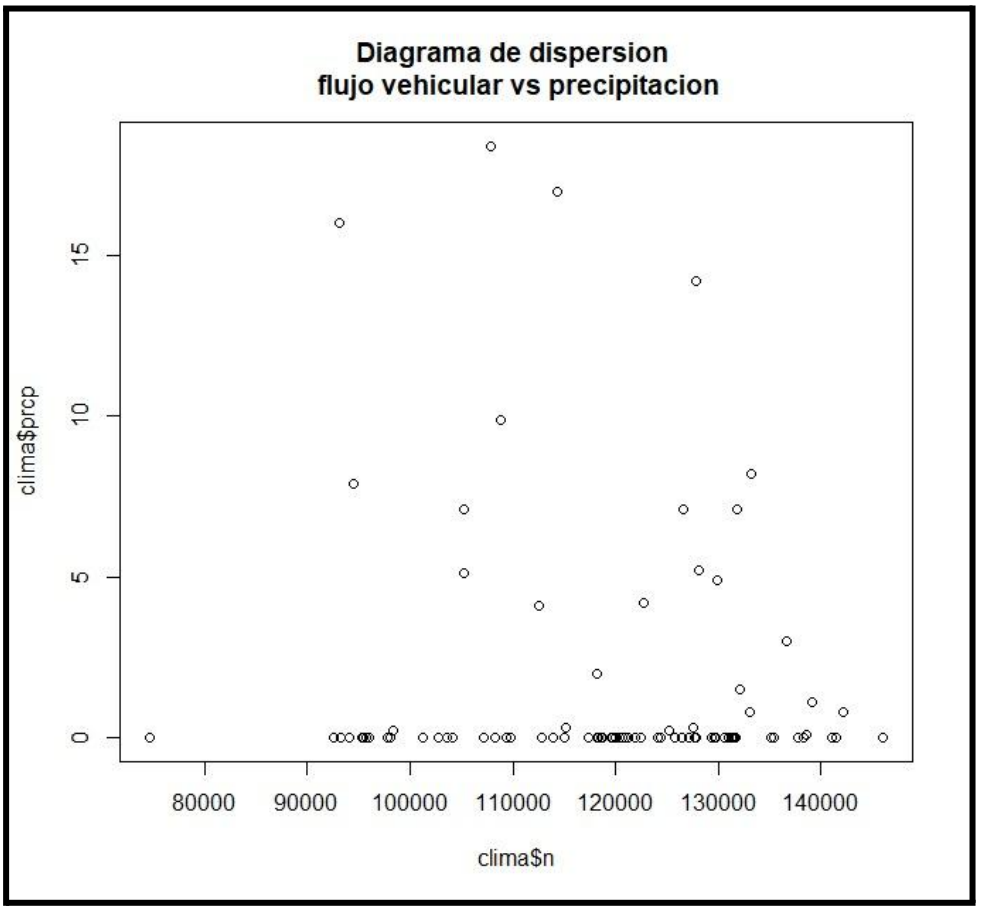
93178 95786 98285

iv) ¿Sugieren los datos la existencia de algún patrón de asociación lineal entre flujo vehicular y precipitación?

Respuesta:

Al revisar el diagrama de dispersión se puede decir que no hay relación entre flujo vehicular y precipitación

Figura Nº13: Diagrama de dispersión entre el flujo vehicular y la precipitación



v) Crear una nueva variable que agrupe los valores de flujo vehicular en “Alto” y “Bajo” según el flujo vehicular sea mayor a 120000 o menor o igual a ese valor, respectivamente. Dar una tabla de datos no agrupados para la nueva variable.

Respuesta:

```
#variable nueva segun el flujo

clima$tipo_flujo <- ifelse(clima$n > 120000, "Alto", "Bajo")

#tabla de datos no agrupados

table(clima$tipo_flujo)

Alto  Bajo

 49    41
```

b) Inferencia.

En cada inferencia dar detalladamente los pasos necesarios hasta establecer la conclusi´on en t´erminos

del problema.

i) Se sabe que la variable precipitación´on tiene una varianza $\sigma^2 = 25$ ¿Hay evidencias estadísticamente significativas para concluir que la precipitación media en toda la zona supera 0 mm, para un nivel de significación $\alpha = 0.01$?.

Respuesta: (en la siguiente página)

i)

$$H_0: \mu = 0 \text{ versus } H_A: \mu > 0 \text{ (Unilateral)}$$

$$\alpha = 0.01 \quad n = 90 \quad X\,barra_n = 1.63$$

Donde n fue calculado en R a través de obtener la cantidad de elementos de la base de datos en la variable Precipitación , y $X\,barra_n$ fue calculado a traves de tomar la media de la precipitacion en la base de datos.

Paso 1: "Estadistico de Pruebas":

$$T = \frac{X\,barra_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$t_{obs} = \frac{1.63 - 0}{\frac{5}{\sqrt{90}}} = 3.09$$

Paso 2: "Intervalo de Rechazo":

$$P(T \geq Z_0 | H_0) = 0.01$$
$$1\text{-}\Phi(Z_0) = 0.01$$
$$0.99 = \Phi(Z_0)$$
$$qnorm(0.99) = Z_0$$

Realizando este calculo en R Z_0 = 2.32
El intervalo de rechazo es [2.32, +∞)

Paso 3: "Resultados e Interpretacion":

Como t_{obs} = 3.09 y 3.09 $\in$ [2.32, +∞) entoces rechazamos H_0 y concluimos que hay evidencia estadisticamente significativa con una probabilidad de error de 0.01 de que la precipitacion media en toda la zona supera los 0mm.

ii) Construir un intervalo de confianza del 99 por ciento para la precipitación media en toda la zona.

ii) Sabemos que del inciso i) conocemos que n = 90, $X\ barra_n$ **= 1.63 y sabemos que** $\alpha = 0.01$ **por que** $1 - \alpha = 0.99$

$$1 - \alpha = 0.99$$
$$-0.99 + 1 = \alpha$$
$$0.01 = \alpha$$

Por definicion los extremos del intervalo de confianza vienen dado por:

$$X\ barra_n \pm Z_{\frac{\alpha}{2}} \cdot \sigma(X\ barra_n)$$

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = P(X \leq Z_{0.005}) = qnorm(1 - 0.005) = 2.57$

$$a = 1.63 - 2.57 * \frac{5}{\sqrt{90}} = 0.2754911$$

$$b = 1.63 + 2.57 * \frac{5}{\sqrt{90}} = 2.984509$$

Concluimos que c una confianza del %99 el intervalo [0.2754911 ; 2.984509] contiene a la precipitacion media en toda la zona.

iii) Para un nivel de significación del 10 por ciento, ¿es posible concluir que la proporción de flujo vehicular Alto en toda la zona es diferente de 1/2?.

iii) El problema plateado es un Test de Hipótesis de nivel aproximado para una proporción

$$H_0: p = \frac{1}{2} \text{ versus } H_A: p \neq \frac{1}{2} \text{ (Bilateral)}$$

$\hat{p}_n = \frac{49}{90}$ donde n = 90.

Paso 1: "Estadistico de Prueba"

$$T = \frac{\hat{p}_n - p_0}{\sqrt{\frac{p_0 \,.\, (1 - p_0)}{n}}} \approx N(0,1) \text{ si n es grande}$$

$$t_{obs} = \frac{\frac{49}{90} - 0.5}{\sqrt{\frac{0.5 \,.\, (1 - 0.5)}{90}}} = 0.8432740427$$

Paso 2: "Resultados basados en el p-valor e interpretación"

$$\text{p-valor} = 2 * P((T \geq 0.8432740427 \mid \text{vale } H_0)$$

$$\text{p-valor} = 2 * (1 - \Phi(0.8432740427))$$

$$\text{p-valor} = 0.3990752$$

como p-valor > 0.01 entonces no rechazamos H_0 y concluimos que no hay evidencia estadisticamente significativa para afirmar que la proporción de flujo vehicular alta en toda la zona es diferente de $\frac{1}{2}$.

Conclusión

A lo largo de este trabajo se aplicaron los contenidos aprendidos en la materia “Estadística”, a través de análisis estadísticos y técnicas, que fueron fundamentales para desarrollar los ejercicios propuestos y hacer una buena interpretación de datos y diagramas.
Aportamos resultados claves como: estimaciones, hallazgos y pruebas de hipótesis para comprender la varianza de la temperatura en la Patagonia y datos climáticos en distintos puntos de Buenos Aires.
Además, como entorno de software utilizamos R Studio, basándonos en R, lenguaje de programación, que es fuertemente utilizado para estadística y análisis de datos.