Métodos Estadísticos de la Ingeniería

U.E.G.I.

Trabajo en Grupo Final

2024-2025

Departamento:

Matemática Aplicada

Titulación:

Grado en Informática de Gestión y Sistemas de Información

2º Curso (1º Cuatrimestre)

Fecha: 04-12-2024

Fernández Molano, Iker

Horas Laguna, Urko

Rodríguez García, Eneko

Hermosilla Martinez-Moro, Gabriel

ÍNDICE

1. INTRODUCCIÓN

En este estudio estadístico se han tomado diferentes vehículos, desde un coche hasta un avión, los cuales, respectivamente, dan valores a las variables: coste (€), contaminación CO_2 (kg/h), dimensiones (m^2), índice de accidentes (0-100), coste por cliente (€), Nº de pasajeros, capacidad de carga (kg) y velocidad (km/h).

Por medio de estos datos y otras fuentes, así como, la DGT, se han obtenido datos estadísticos de interés. Estos resultados tienen como objetivo representar situaciones reales en términos de coste, impacto ambiental, seguridad, capacidad y eficiencia.

Como herramienta de cálculo de estadísticos hemos utilizado **R** y su extensión **RStudio.** Este programa implementa diversas funciones que facilitan la obtención de resultados, que a mano serían tediosos o complejos de calcular.

Por último, como práctica para lo aprendido en el aula, hemos tratado de abarcar todos los aspectos que se han impartido en el temario. Es por esto, que el informe está separado por temas.

2. ESTADÍSTICA DESCRIPTIVA

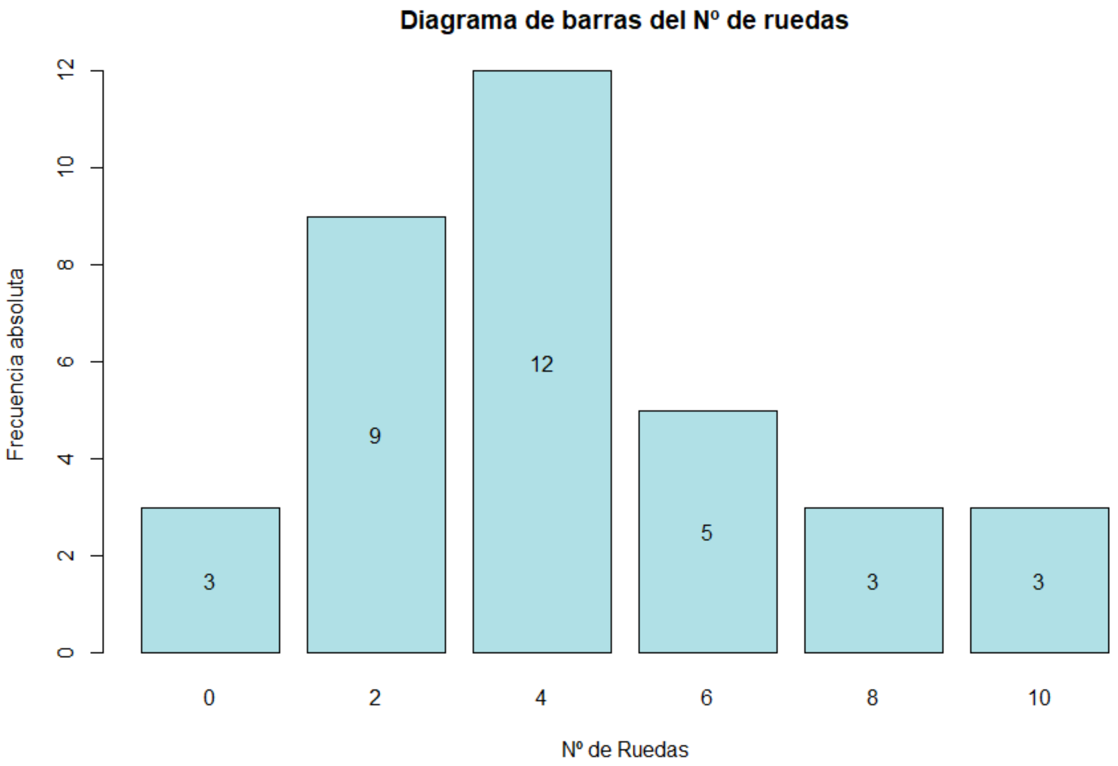
Se han calculado las medidas estadísticas de tendencia central (media, mediana y moda) para todas las variables presentes en el fichero de datos. Se omitieron los valores faltantes (NA) para garantizar la obtención de resultados precisos. Asimismo, los datos obtenidos serán muy útiles a lo largo del estudio. Las conclusiones calculadas son los siguientes:

VARIABLE	MEDIA	MEDIANA	MODA
Coste (€)	5027933,2857	50000	30000
Contaminación CO_2 (kg/h)	44,1737	12	0
Dimensiones (m^2)	198,7429	11	40
Índice de accidentes (0-100)	50,8571	50	50
Coste por cliente (€)	25,6323	3	3
Nº de pasajeros	124,8286	4	2
Capacidad de carga (kg)	9707,3810	500	50
Velocidad (km/h)	114,8571	90	25
Nº de ruedas	4,2857	4	4

Posteriormente, se ha elegido una de las variables, en este caso el número de ruedas, para hacer uso de las funciones de **R** que permiten obtener una tabla de frecuencias. Algunas de las funciones usadas han sido, *table* (Frecuencia absoluta), *prop.table* (Frecuencia relativa) y *cumsum* (Frecuencias acumuladas). Los datos han sido recopilados en un *data frame*.

Nº de Ruedas	Frec. Abs. (n_i)	Frec. Rel. (f_i)	Frec. Abs. Acu. (N_i)	Frec. Rel. Acu. (F_i)
0	3	0,0857	3	0,0857
2	9	0,2571	12	0,3428
4	12	0,3429	24	0,6857
6	5	0,1429	29	0,8286
8	3	0,0857	32	0,9143
10	3	0,0857	35	1

Una vez obtenida la tabla de frecuencias, se ha utilizado la función *barplot* para representar la frecuencia absoluta en un diagrama de barras. En este se identifica una tendencia hacia que lo vehículos tengan 4 o 2 ruedas.



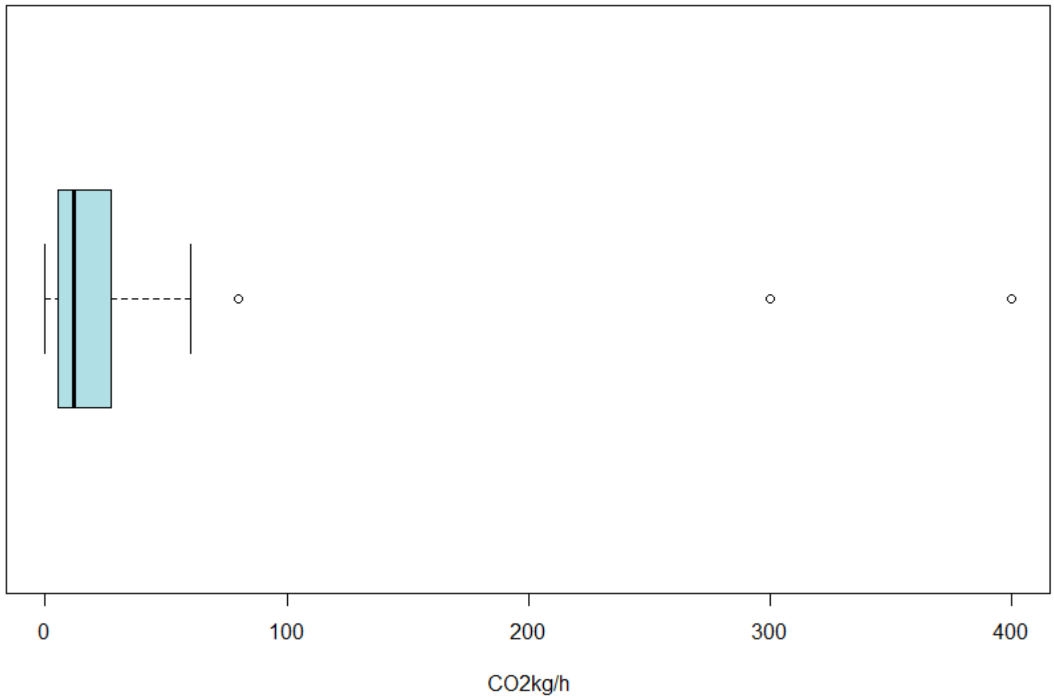
Por último, por medio de la función *boxplot*, se ha representado el diagrama de caja respecto la contaminación. Previamente, se ha calculado el rango intercuartílico (IQR), con el fin de obtener los datos atípicos, es decir, datos que distan del resto.

Recordamos que:

$$IQR = Q3 - Q1$$
$$[L, H] = [Q3 - 1'5 \cdot IQR,\ Q3 + 1'5 \cdot IQR]$$

Los valores atípicos son aquellos que no se encuentran entre L (Límite inferior) y H (Límite superior). Los cuartiles, por otro lado, han sido calculados con la función *quantile* de **R**.

Hemos obtenido como resultado, que el barco de pasajeros, el barco comercial, el avión comercial y el ferry contaminan considerablemente más que el resto de los vehículos. Este resultado, tiene sentido, pues son los vehículos de mayor tamaño y consumo entre los analizados.


0
100
200
300
400
CO2kg/h

3. COMBINATORIA Y PROBABILIDAD

Se ha querido buscar una relación entre la velocidad y el índice de accidentes. Es por eso, que se ha calculado la probabilidad de que un vehículo cualquiera tenga una velocidad mayor que la media (114,8571 km/h) dado que el índice de accidentes es mayor que la media de todos los vehículos (50,8571). Primero, se definen las premisas.

A : "Índice de accidentes mayor que la media".

B : "Velocidad mayor que la media".

Se busca:

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

Según la regla de Laplace:

P(A)=$\frac{n^{\underline{o}}\ casos\ favorables}{n^{\underline{o}}\ casos\ posibles} = \frac{25}{35} \approx 0{,}7143 = 71{,}43\%$

El número de casos favorables se ha calculado con la función *subset,* resultando en el subconjunto de vehículos que tenían un índice de accidentes mayor que la media.

P$(A \cap B) \approx 0{,}1714 = 17{,}14\%$

P(B | A) ≈ $0{,}24$ = 24%

Aunque los vehículos con un índice de accidentes superior a la media tienen algo más de probabilidad (24%) de tener una velocidad superior a la media, esta probabilidad no es lo suficientemente alta como para afirmar con certeza que existe una conexión directa entre la velocidad y los accidentes. Es probable que otros factores influyan en el aumento del índice de accidentes.

Proporción de vehículos de gran tamaño, de más de 6 ruedas

Además, se ha investigado la proporción de vehículos con más de 6 ruedas. Este análisis es relevante en contextos de seguridad y diseño de infraestructura vial.

La proporción de vehículos con más de 6 ruedas es aproximadamente 15%, lo que indica que un cierto porcentaje de vehículos son de tipo más grande, como camiones, autobuses o metros. Son datos para tener en cuenta ya que producen grandes emisiones de CO2, pero al ser varios de ellos transportes públicos, no reflejan correctamente esa producción de CO2 ya que deberíamos tener en cuenta el número de pasajeros.

4. VARIABLE ALEATORIA DISCRETA Y CONTINUA

En el año 2022, según los datos del Ministerio de Industria y Turismo, en el País Vasco, 762.209 vehículos acudieron a la revisión a la Inspección Técnica de Vehículos (ITV). Según los datos, aproximadamente el 15,48% de los vehículos no pasaron la inspección. Se quiere saber cuál es la probabilidad de que el año siguiente, si se espera el mismo número de vehículos, más de 120.000 vehículos no pasen la inspección.

Se considera la variable aleatoria:

X : "Número de vehículos que no pasan la ITV"

S (Éxito) : "El vehículo no pasa la ITV" → $P(S) = 0{,}1548$

F(Fracaso) : "El vehículo pasa la ITV" → $P(F) = 1 - P(S) = 0{,}8452$

El experimento aleatorio consiste en repetir 762.209 veces la misma prueba de Bernoulli.

X~B($X \sim B(762209,\ 0'1548)$

$$P(X > 120000) = 1 - P(X \leq 119999) = \ 1 - \sum_{i=0}^{119999} 119999\mathbb{C}i \cdot P(S)^{i} \cdot P(F)^{119999-i}$$

n, 762.209, es un número demasiado grande, por lo que se optará por aproximar la binomial a otro método de distribución, la distribución normal. Porque se cumple que:

$$n \cdot P(S) = 117990 > 5$$
$$n \cdot P(F) = 644219 > 5$$
$$n = 762209 > 30$$
$$X \sim N\left(\mu = n \cdot P(S) = 117990,\ \sigma\ =\ \sqrt{n \cdot P(S) \cdot P(F)} \approx 315'7928\right)$$

$$P\big(X > 120000 \,\big|\, X \sim B(762209{,}0'1548)\big) \cong$$
$$\cong\ P\big(X > 119999'50 \,\big|\, Z \sim N(117990{,}315'7928)\big) =$$

En este primer paso, corregimos por continuidad. Posteriormente, se tipifica.

$$= P\left(Z = \frac{X-\mu}{\sigma} > \frac{119999'50 - 117990}{315'7928} \approx 6'3635\right) = P\big(Z > 6'3635 \,\big|\, Z \sim N(0{,}1)\big)$$

$$= 1 - P\big(Z < 6'3635 \big| Z \sim N(0{,}1)\big) = 1 - pnorm(6.3635)$$
$$=$$
$$= pnorm(6.3635,\ lower.tail = FALSE) \approx 9'8604 \cdot 10^{-11}$$

La probabilidad de que más de 120.000 vehículos no superen la inspección es muy baja. Está suficientemente cerca de la media como para no ser 0, pero suficientemente lejos de la misma para ser casi "imposible". Por lo que, lo más seguro es que en el año 2023, no se superen los 120.000 vehículos sin superar la inspección.

Además, según la DGT, la media de contaminación producida por los vehículos registrados es de 44’1737kg/h. Se quiere saber cuál es la probabilidad de que en un día un vehículo contamine como mínimo 280kg/h y menos de 300kg/h.

Calcularíamos, en primer lugar, todos los datos relevantes para la resolución del problema, a partir de la información dada en el enunciado:

$$n = length(Conta\min a\,cion) = 35$$

$$\mu = 44{,}1737 \;\rightarrow\; round(Medias[2], 4)$$

$$S^2 = 8581{,}5850 \;\rightarrow\; var(Conta\min a\,cion,\ na.rm = TRUE)$$

$$\sigma = 91{,}3039 \;\rightarrow\; round\left(\sqrt{\frac{(n-1)\cdot S^2}{n}},\ 4\right)$$

Se considera la variable aleatoria:

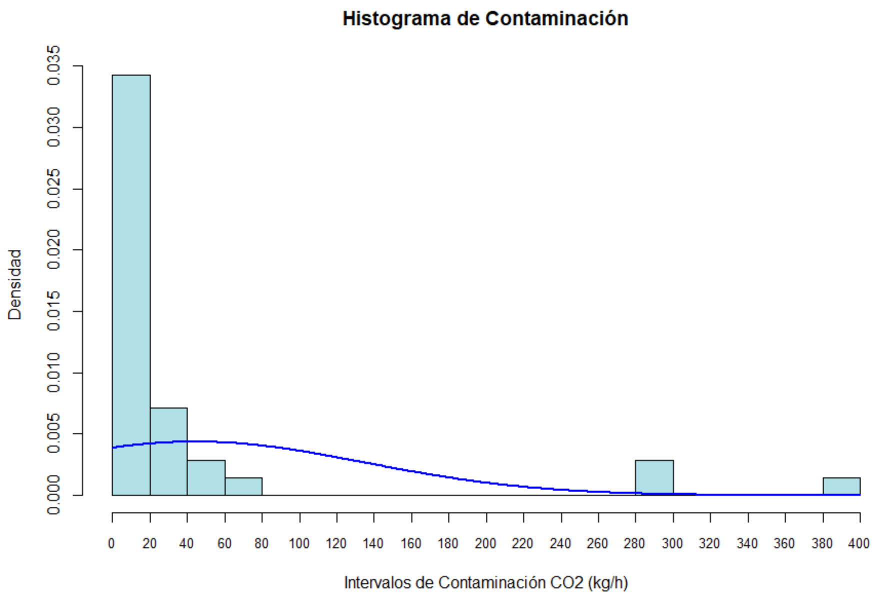
Y: "Contaminación de CO2, en kg, en una hora, producida por cualquier vehículo”

$$Y \sim N(\mu = 44'1737,\ \sigma = 91'3039)$$

Debemos resolver: $P(280 < X < 300) = P(X < 300) - P(X < 280)$. Para ello:

$$pnorm(300{,}44'1737{,}91'3039) - pnorm(200{,}44'1737{,}91'3039) \approx 0'0024 = 0'24\%$$

La probabilidad es muy baja, pocos vehículos contaminan entre 280 y 300 kg de CO_2 por hora. Además, realizaremos un histograma que resalte adecuadamente la relación entre intervalos de contaminación y densidad.



Por otro lado, los precios de los vehículos han crecido durante los últimos años. El sueldo medio mensual en España en el año 2023 fue de 2273€. Sabiendo esto, ¿cuál es la probabilidad de que un español promedio, pueda comprar un vehículo de 4 ruedas o menos con el sueldo íntegro de 3 meses?

Se considera la variable aleatoria:

X: "Precio del vehículo, en euros (€)"

Esta variable, sigue una distribución normal de media μ y desviación típica σ , las cuales se calcularán más adelante.

$$X \sim N(\mu,\ \sigma)$$

Este dinero se obtiene de multiplicar por 3 el sueldo medio:

$$dinero = 3 \cdot 2273 = 6819 \ \text{€}$$

Para poder comprar el vehículo, el precio del vehículo deberá ser menor o igual a este dinero.

Ahora calcularemos la media de los precios de los vehículos utilizando los comandos de R para poder acceder a nuestros datos.

Considerando que se ha definido la constante $Coste < -datos\$Coste.$, seleccionaremos los vehículos que tienen 4 o menos ruedas utilizando

$$ruedas = subset(datos,\ NRuedas \leq 4)$$

A continuación, almacenaremos solamente los datos con la variable

$$precios = ruedas\$Coste$$

y calcularemos la media y la desviación típica de la siguiente manera:

$$\mu = mean(precios, na.rm = TRUE) = 2258861$$

Para la desviación típica, calcularemos la varianza a partir de la cuasi varianza y la desviación típica será la raíz cuadrada de la varianza:

$$cvar.precios = var(precios,\ na.rm = TRUE)$$

$$n = length(precios) = 24$$

$$\sigma = \sqrt{\frac{(n-1) \cdot cvar.precios}{n}} = 6588264$$

Por lo tanto, la distribución quedaría: $X \sim N(2258861{,}6588264)$.

Ahora, quedaría calcular $P(X \leq 6819)$ mediante el comando $pnorm(6819, \mu, \sigma) \approx 0.3662408$

5. ESTIMACIÓN DE PARÁMETROS DE UNA DISTRIBUCIÓN

Calculamos intervalos de confianza al 99% para la media de la contaminación y para la proporción de vehículos que no pasan la ITV.

- **Intervalo de confianza para la media de la contaminación (kg/h)**: El intervalo de confianza para la media de la contaminación es:

  [3’8401kg/h,84’5073kg/h]

Esto significa que, con un 99% de confianza, la verdadera media de las emisiones de CO_2 de todos los vehículos está entre 3’8401kg/h y 84’5073kg/h

- **Intervalo de confianza para la proporción de vehículos que no pasan la ITV**: Con una proporción observada de 0.1548, el intervalo de confianza para la proporción de vehículos que no pasan la ITV es:

  [0’1537,0’1559]

Esto indica que, con un 99% de confianza, la verdadera proporción de vehículos que no pasan la ITV está entre el 15.37% y el 15.59%.

6. CONTRASTES DE HIPÓTESIS PARAMÉTRICOS

Realizamos pruebas de hipótesis para verificar si la media de la contaminación es mayor a 50 kg/h y si la proporción de vehículos que no pasan la ITV es diferente del 15%.

- **Prueba de hipótesis para la media**: Se probó si la media de las emisiones de CO2 es mayor que 50 kg/h. El valor p obtenido fue $\mathbf{p} = 0.7121$, lo que indica que aceptamos la hipótesis nula (que establece que la media es 50 kg/h). Por lo tanto, podemos concluir que la media de las emisiones es próxima a 50 kg/h.

  t.test(Contaminacion, mu=50, conf.level=099)

- **Prueba de hipótesis para la proporción de vehículos que no pasan la ITV**: Para verificar si la proporción de vehículos que no pasan la ITV es diferente del 15%, el valor p de la prueba fue $\mathbf{p} = 0.1548$. Como este valor es mayor que el nivel de significancia de 0.01, no rechazamos la hipótesis nula, lo que sugiere que la proporción de vehículos que no pasan la ITV no es significativamente diferente del 15%.

  prop.test(x=n1*P.S, n=n1, p=0.15, conf.level=0.99)

Este análisis proporciona una visión exhaustiva de las emisiones de CO2 de los vehículos, tanto desde una perspectiva descriptiva como inferencial. A través de las estadísticas descriptivas, las pruebas de hipótesis y los intervalos de confianza, se puede afirmar que las emisiones promedio de CO2 en los vehículos son significativamente mayores que 50 kg/h y que existe una notable variabilidad entre las emisiones de los diferentes vehículos. El análisis también sugiere que la proporción de vehículos que no pasan la ITV es compatible con un valor del 15%, y la relación entre la velocidad y los accidentes es moderadamente negativa. Las probabilidades calculadas permiten hacer inferencias sobre los niveles extremos de contaminación, proporcionando información valiosa para la formulación de políticas ambientales y de tráfico.

7. BIBLIOGRAFÍA

Dirección General de Tráfico. *Página oficial de la Dirección General de Tráfico*. Gobierno de España. Consultado el 18 de octubre de 2024. https://www.dgt.es/inicio/

Statista. *The Statistics Portal*. Consultado el 18 de octubre de 2024. https://www.statista.com/

Ministerio de Transportes, Movilidad y Agenda Urbana. *Observatorio del Transporte y la Logística en España (OTLE)*. Consultado el 2 de noviembre de 2024. https://otle.transportes.gob.es/

Ministerio de Industria, Comercio y Turismo. *Página oficial del Ministerio de Industria, Comercio y Turismo*. Consultado el 2 de diciembre de 2024. https://industria.gob.es/es-es/Paginas/Index.aspx