Estudio inferencial de Características y Proporciones: Un Estudio Comparativo entre automóviles y camionetas.

Autores: Mathias Gómez; Geiler Chable.

1. Introducción.

El presente estudio, detalla el análisis y la inferencia estadística realizada sobre un conjunto de datos vehiculares con el objetivo principal de evaluar las características, proporciones y rendimiento entre dos categorías principales: automóviles y camionetas.

1.1. Descripción de la base de datos.

La base de datos utilizada para el proyecto consiste en un registro tabular de 50 observaciones (50 filas) y, a su vez cuenta con 5 columnas que corresponden a una serie de variables descritas a continuación:

Tabla 1. Descripción de las variables de la base de datos.

Nombre de la Variable	Tipo de Dato	Descripción
Fabricante	Categórica (Nominal)	Marca o fabricante del vehículo.
Tipo de Vehículo	Categórica (Nominal)	Clasificación principal, diferenciando entre **valores 0 y 1, para Automóvil y Camioneta, respectivamente.**
Motor	Numérica (Continua)	Capacidad del motor.
Potencia	Numérica (Continua)	Potencia del motor (caballos de fuerza).
Rendimiento	Numérica (Continua)	Se refiere a las millas por galón que rinde el vehículo.

Fuente: Elaboración propia.

1.2. Objetivos del análisis.

Afirmar a partir de las pruebas estadísticas realizadas, si las características generales de los automóviles y las camionetas en esta muestra son comparables o si difieren significativamente.

- Determinar si existen diferencias estadísticamente significativas entre el rendimiento promedio entre el grupo de automóviles y el grupo de camionetas.
- Evaluar si la proporción de automóviles es estadísticamente igual a la proporción de camionetas dentro de la muestra.
- Determinar si existen variables que proporcionan información similar o equivalente.

2. Metodología.

La metodología se basa en un enfoque inferencial, utilizando un conjunto de técnicas estadísticas específicas para abordar los objetivos planteados.

2.1. Gestión de Datos y Redundancia.

2.1.1. Detección de valores atípicos.

El tratamiento de valores extremos es crucial para garantizar la robustez de las pruebas. Se utilizó el método del **Rango Intercuartílico (IQR).**

El valor atípico x_i es aquel que cae fuera del intervalo definido por los cuartiles (Q_1 y Q_3) y el rango intercuartílico (IQR). Podemos definir IQR como:

$$\boldsymbol{IQR = Q_3 - Q_1}$$

Para identificar las observaciones atípicas se establece como tales a aquellas que:

- $x_i < LI$, dónde: $Límite\ inferior\ (LI) = Q_1 - 1.5 * IQR.$
- $x_i > LS$, dónde: $Límite\ superior\ (LS) = Q_3 + 1.5\ x * IQR.$

Herramientas gráficas como los **gráficos de dispersió**n pueden ayudar a comprender mejor la distribución de la variable y la participación de los valores atípicos en el mismo.

2.1.2. Análisis de Variables Similares.

Para evaluar si variables numéricas proporcionan información similar, se utiliza el **Coeficiente de Correlación de Pearson** $(\boldsymbol{r})$.

Mide la fuerza y dirección de una relación lineal entre dos variables aleatorias $(X\ y\ Y)$.

$$pX, Y = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Su estimador muestral se tiene:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

El objetivo fue identificar variables con fuerte relación $(|\boldsymbol{r}| > \boldsymbol{0.70})$**,** para diagnosticar multicolinealidad indicando ser redundantes. **Mapas de calor** también son utilizados para representar de manera más gráfica estos resultados.

2.2. Inferencia para variables de estudio.

2.2.1. Proporción de muestras.

Para determinar si la proporción de una muestra es igual a otra, se formula una prueba de hipótesis para la diferencia de proporciones independientes**.**

Se aplica una **prueba Z para una diferencia de dos Proporciones.** Teniendo que:

- $\boldsymbol{H_0}: \boldsymbol{p_1 = p_2}$
- $\boldsymbol{H_a}: \boldsymbol{p_1 \neq p_2}$

Para el caso aplicado, teniendo proporción de Automóviles (p_A) y proporción de Camionetas(p_C). Podemos decir que:

- $\boldsymbol{H_0}: \boldsymbol{p_A - p_C = 0}$
  $\boldsymbol{H_a}: \boldsymbol{p_A - p_C \neq 0}$

Donde**,**

$$\boldsymbol{Z = \frac{(\widehat{p_A} - \widehat{p_C}) - 0}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_A} + \frac{1}{n_C})}}}$$

Donde $\boldsymbol{n_1\ y\ n_2\ representan\ los\ tamaños\ de\ muestra}$.
Se rechaza la hipótesis nula si el valor p de la distribución Normal Estándar es menor o igual a la significancia $\alpha = \boldsymbol{0.05}$.

$$\boldsymbol{Si\ p \leq 0.05 \Rightarrow Rechazar\ H_0}$$

2.2.2. Inferencia para Comparación de Medias.

- Para contrastar la hipótesis de igualdad de medias entre dos poblaciones independientes y normalmente distribuidas $(\mu_1 y\ \mu_2)$, Se aplica la Prueba t de Student de Welch (asumiendo varianzas diferentes).

  Hipótesis:

  $$H_0: \mu_1 - \mu_2 = 0\ vs\ H_{a:}\mu_1 - \mu_2 \neq 0$$

  Estadístico de prueba (t):

$$t = \frac{(\underline{X_1} - X_2)}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}}$$

Donde $\bar{x_i}$ y $s^2{}_i$ son la media y la varianza muestrales, y ni es el tamaño de la muestra i. Se consideran grados de libertad estimados por el software estadístico.

- La **estimación de Intervalos de Confianza (IC)** del 95% $(1 - \alpha)$, para la media del rendimiento de cada grupo:

$$IC = \bar{X} \pm t_{a/2,v} * \frac{s}{\sqrt{n}}$$

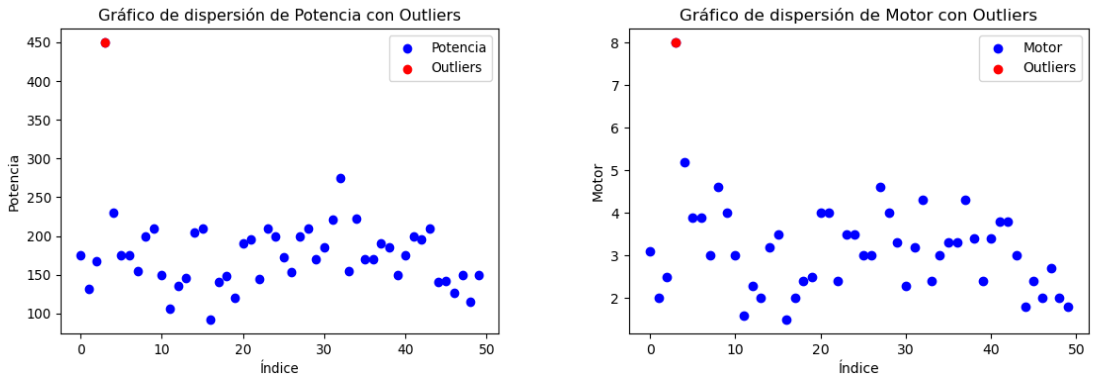
Donde $t\alpha/2, v$ es el valor crítico de la distribución t con v grados de libertad conforme a la fórmula de Satterthwaite.
La ausencia de solapamiento de los intervalos sirve como evidencia de la diferencia estadística entre medias.

3. Resultados.

3.1. Análisis de valores atípicos.

Para la identificación de valores atípicos (*outliers*), se elaboró un gráfico de dispersión para cada variable. Este análisis inicial reveló la presencia de valores atípicos significativos en las variables "Motor" y "Potencia", los cuales se desviaban del comportamiento habitual observado en la muestra de la categoría *Automóviles*.

Figura 1. Gráfico de dispersión con outliers para variables "Motor" y "Potencia".



Fuente: Creación propia en Python.

La detección de *outliers*, confirmada tanto visualmente con los gráficos como cuantitativamente con el Rango Intercuartílico, justificó la eliminación de la fila 4 de la base de datos. Esta acción fue necesaria para optimizar la muestra y obtener resultados de análisis inferencial más precisos.

3.2. Análisis diferencia de medias para variables de estudio.

El proceso de análisis comenzó con una evaluación de las variables de los vehículos para determinar la existencia de similitudes o patrones característicos entre ellos. Como paso inicial, se realizó un análisis descriptivo, el cual arrojó la siguiente información:

Tabla 2. Estadística descriptiva de las variables de estudio.

Tipo Vehículo	Motor (mean)	Motor (std)	Potencia (mean)	Potencia (std)	Rendimiento (mean)	Rendimiento (std)
Automóvil	2.8	0.88	168.59	40.85	25.51	2.75
Camioneta	3.38	0.77	176.95	29.01	20.04	2.85

Fuente: Creación propia en Python.

La estadística descriptiva (media y desviación estándar) muestra valores distintos para *Automóvil* y *Camioneta*. No obstante, para validar si esta diferencia es estadísticamente significativa y no producto del azar, es indispensable realizar una prueba de hipótesis. Asumiendo la posibilidad de varianzas poblacionales desiguales (heterogeneidad), la prueba seleccionada es la prueba t de Welch con hipótesis nula “las medias de las muestras son iguales” y, en contraparte, la hipótesis alternativa “las medias de las muestras no son iguales”.

Tabla 3. Pruebas de hipótesis por variable Motor.

$H_0: \mu_{automóviles} = \mu_{camionetas}$ $H_1: \mu_{automóviles} \neq \mu_{camionetas}$	
t - estadístico	-1.319
valor p	0.1938
Decisión	No rechazar H0
Conclusión	No hay diferencia significativa en la media de motor

Fuente: Creación propia en Python.

Tabla 4. Pruebas de hipótesis por variable Potencia.

$H_0: \mu_{automoviles} = \mu_{camionetas}$ $H_1: \mu_{automóviles} \neq \mu_{camionetas}$	
t - estadístico	0.12
valor p	0.9048
Decisión	No rechazar H_0
Conclusión	No hay diferencia significativa en la media de potencia

Fuente: Creación propia en Python.

Tabla 5. Pruebas de hipótesis por variable Rendimiento.

$H_0: \mu_{automóviles} = \mu_{camionetas}$ $H_1: \mu_{automóviles} \neq \mu_{camionetas}$	
t - estadístico	5.944
valor p	0
Decisión	Rechazar H_0
Conclusión	Hay diferencia significativa en la media de Rendimiento

Fuente: Creación propia en Python.

3.3. Prueba de hipótesis para proporciones.

Por otra parte, se considera realizar una prueba de hipótesis a la proporción de las muestras, considerando como hipótesis nula “las proporciones son iguales” e hipótesis alternativa “las proporciones son desiguales”.

Tabla 6. Pruebas de hipótesis por proporciones.

$H_0: p_{automóviles} = p_{camionetas}$ $H_1: p_{automóviles} \neq p_{camionetas}$	
Prueba z	1.0102
p - valor	0.3124
Decisión	No rechazar H_0
Conclusión	No hay evidencia para afirmar que hay proporciones diferentes.

Fuente: Creación propia en Python.

3.4. Inferencia estadística para variable Rendimiento (intervalo de confianza).

Tras realizar las pruebas de hipótesis para la igualdad de medias, se determinó que la hipótesis nula fue rechazada únicamente para la variable "Rendimiento". Este resultado indica una diferencia estadísticamente significativa en el rendimiento promedio entre *automóviles* y *camionetas*. Para cuantificar y visualizar la magnitud de esta diferencia, se procederá a calcular los intervalos de confianza correspondientes para la media de cada categoría de vehículo, y así establecer conclusiones más robustas.

Tabla 7. Estimación de intervalo de confianza variable Rendimiento.

Tipo de Vehículo	Media Muestral Rendimiento	IC al 95%
Automóviles (n=27)	25.52	[24.43, 26.61]
Camionetas (n=22)	20.05	[18.78, 21.31]
Estadístico t = 6.7870, p-valor = 0.000		

Fuente: Creación propia en Python.

Con ello, podemos observar que con 95% de confianza, el rendimiento de los automóviles, podrían tomar valores de 24.43 a 26.61 millas por galón, mientras que el de las camionetas

suele ser menor, con valores de 18.78 a 21.31 millas por galón con el mismo nivel de confianza.

3.5. Inferencia estadística para variable Rendimiento.

Para determinar si existen variables que proporcionan información similar, podemos calcular el coeficiente de Correlación de Pearson entre las variables numéricas (Motor, Potencia y Rendimiento). Obtenemos los siguientes resultados:

Tabla 8. Matriz de coeficiente de Pearson para variables numéricas.

	Motor	Potencia	Rendimiento
Motor	1.000000	0.800412	-0.745561
Potencia	0.800412	1.000000	-0.507775
Rendimiento	-0.745561	-0.507775	1.000000

Fuente: Creación propia en Python.

Principales conclusiones:
La correlación entre **Motor y Potencia** es del r = 0.800412. Esto demuestra una fuerte correlación positiva, lo que significa que a medida que el tamaño del motor aumenta, la potencia también aumenta. Esto implica que ambas variables, están midiendo en gran medida la misma característica del vehículo (capacidad de desempeño). Por lo tanto, se considera **redundante**.

Además, la correlación entre Motor - Rendimiento es fuertemente negativa, mientras que Potencia - Rendimiento es moderadamente negativa, confirmando que las variables de desempeño son influyentes en la eficiencia.

4. Conclusiones.

4.1. ¿Se puede afirmar que los automóviles y las camionetas tienen características similares?

Sobre la muestra analizada de 49 registros (tras la eliminación de un *outlier*), la prueba de hipótesis no detectó una diferencia estadísticamente significativa en la media de las variables "Motor" y "Potencia" entre automóviles y camionetas. Esto sugiere que, para esta población, ambos tipos de vehículos comparten características promedio similares en cuanto a su motorización.

En contraposición, se encontró que la variable "Rendimiento" sí es significativamente diferente. Esta disparidad nos permite concluir que, a pesar de la igualdad en la media de motor y potencia, el rendimiento es una cualidad distintiva que provoca que los dos tipos de vehículos difieran hasta cierto punto dentro de la muestra estudiada.

4.2. En esta muestra, ¿Se puede afirmar que la proporción de automóviles es igual a la proporción de camionetas?

A pesar de que la proporción muestral observada para la muestra de los automóviles ($\frac{27}{49}$) es numéricamente superior a la muestra de las camionetas ($\frac{22}{49}$), el valor p de la prueba de hipótesis es superior al nivel de significancia (0.05). Por lo tanto, no existe evidencia estadística suficiente para rechazar la hipótesis nula de que las proporciones poblacionales son iguales. Es decir, desde un punto de vista estadístico, ambos grupos comparten la misma proporción.

4.3. Realice inferencia estadística para el rendimiento que tienen los automóviles y el rendimiento que tienen las camionetas. ¿Qué se puede afirmar de los resultados obtenidos?

Los intervalos de confianza al 95% confirman la disparidad inferencial: el rendimiento de los automóviles se proyecta entre 23.82 y 26.44 millas por galón, mientras que el de las camionetas se encuentra consistentemente en un intervalo de 18.78 a 21.31 millas por galón. Por lo que, podemos concluir que, es muy poco probable (casi imposible) que un automóvil y una camioneta tengan el mismo rendimiento, siendo el de los automóviles mayor que el de las camionetas.

4.4. ¿Hay variables que proporcionan valor similar o equivalente?

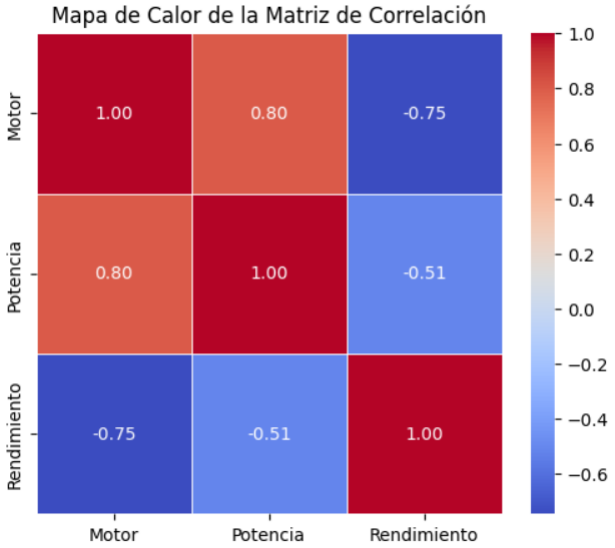
Sí, se puede afirmar que hay variables que proporcionan un valor medio similar o equivalente entre las categorías de vehículos.

Variables Similares (Equivalencia): Las pruebas de hipótesis no encontraron diferencias estadísticamente significativas en las medias de "Motor" y "Potencia" entre los automóviles y las camionetas (no se rechazó la H0). Esto implica que, desde una perspectiva poblacional, el nivel promedio de motor y potencia es similar para ambos grupos.

Variables Diferentes: En contraparte, la prueba de hipótesis para la variable "Rendimiento" sí mostró un rechazo a la hipótesis nula. Esto significa que el rendimiento promedio es estadísticamente diferente entre los dos tipos de vehículos, siendo esta la variable clave que los distingue.

Esto se comprueba también con el mapa de calor, que muestra de manera más visual el estadístico de Pearson.

Figura 2. Mapa de calor para estadístico de Pearson.



Fuente: Creación propia en Python.

El mapa de calor no sugiere que las variables son similares o equivalentes, solamente que las magnitudes de sus correlaciones son muy parecidas. En donde ambas tienen un valor de correlación fuerte.

Tabla 9. Interpretación por valor de correlación.

Variables	Valor Correlación	Conclusión
Motor - Potencia	0.8	Relación fuerte donde ambas variables se mueven en la misma dirección (positiva)
Motor - Rendimiento	-0.75	Relación fuerte donde las variables se mueven en direcciones opuestas.

Fuente: Creación propia en Python.

Por tanto, el grado de asociación es similar (ambas relaciones son muy fuertes), no las variables que las componen.