Fecha: Junio 06, 2023

Pruebas de Hipótesis en Ciencia de Datos

Roberto de J. Alfaro López[1]
[1]Universidad Autónoma de Querétaro, MX

Correspondiente al autor: Roberto de J. Alfaro López (e-mail: robertoalfaro14@hotmail.com).

RESUMEN Las pruebas de hipótesis son un componente fundamental en la ciencia de datos y en la estadística, pues se utilizan para tomar decisiones basadas en los datos recopilados. En este documento se presentan varios ejemplos de pruebas de hipótesis aplicados al dataset "*DailyDelhiClimateTrain*" para aquellas pruebas que requerían atributos numéricos, y al dataset "*Salary_Data*" para aquellas que requerían atributos categóricos. Además de los ejemplos, también se da una breve explicación sobre en qué consiste la prueba, sus ecuaciones y algunos casos de uso. Se podrá observar la utilidad de estos métodos, así como algunas de sus limitaciones.

PALABRAS CLAVE Distribución, Pruebas de Hipótesis, Correlación, Estacionariedad, Valor P.

I. OBJETIVO

Aplicar en una base de datos las diversas pruebas de hipótesis otorgadas por el profesor, dar un ejemplo de cada una, explicar en qué consiste la prueba, mostrar sus ecuaciones y mencionar algunos casos de uso.

II. INTRODUCCIÓN

Actualmente, la ciencia de datos se ha convertido en un pilar fundamental para la toma de decisiones. La capacidad de analizar y extraer conocimiento de grandes volúmenes de datos es esencial en una amplia gama de campos. En este contexto, las pruebas de hipótesis juegan un papel crucial, proporcionando un marco formal para evaluar ciertas afirmaciones y tomar decisiones basadas en los datos que se posean.

Las pruebas de hipótesis son un conjunto de procedimientos estadísticos que permiten tomar decisiones objetivas sobre las propiedades de una población basándose en una muestra de datos. Estas pruebas se basan en la formulación de dos hipótesis contrapuestas: la hipótesis nula, que representa una situación de no cambio o no efecto, y la hipótesis alternativa, que representa la situación de cambio o efecto que estamos interesados en detectar.

El proceso de prueba de hipótesis implica el cálculo de una estadística de prueba, que se utiliza para decidir si rechazar o no la hipótesis nula. Esta decisión se toma en función del valor P, que es la probabilidad de obtener un resultado al menos tan extremo como el observado, suponiendo que la hipótesis nula es verdadera. Si el valor P es menor que un nivel de significancia predefinido, se rechaza la hipótesis nula en favor de la hipótesis alternativa.

Existen diferentes tipos de pruebas de hipótesis, cada una con sus propias suposiciones y aplicaciones. Las pruebas de normalidad, por ejemplo, se utilizan para verificar si un conjunto de datos sigue una distribución normal, una suposición clave en muchos modelos estadísticos. Las pruebas de correlación, por otro lado, se utilizan para evaluar la relación entre dos variables.

Por otro lado, en el contexto de las series de tiempo, las pruebas de estacionariedad son esenciales para determinar si las propiedades estadísticas de una serie son constantes en el tiempo. Por último, las pruebas de hipótesis estadísticas paramétricas y no paramétricas se utilizan para comparar las medias de dos o más grupos, dependiendo de si los datos cumplen o no ciertas suposiciones.

A pesar de su utilidad, las pruebas de hipótesis no están exentas de ciertas limitaciones. La elección del nivel de significancia, por ejemplo, es a menudo arbitraria y puede afectar los resultados de la prueba. Además, las pruebas de hipótesis no pueden proporcionar una medida de la magnitud del efecto o de la importancia práctica de los resultados.

En este artículo, se explorarán algunas de las pruebas de hipótesis más usadas en la ciencia de datos, se mostrarán sus fundamentos teóricos, sus diferentes tipos y aplicaciones, así como sus limitaciones.

III. MARCO TEÓRICO

En esta sección, se presentará toda la teoría que hay detrás de las diferentes pruebas de hipótesis, así como también algunos casos de uso en donde se pueden aplicar. Las pruebas de hipótesis que se usaron fueron 16 y son las siguientes: prueba de Shapiro-Wilk, prueba de D'Agostino K^2, prueba de Anderson-Darling, coeficiente de correlación de Pearson, correlación de rangos de Spearman, correlación de rangos de Kendall, prueba de Chi-cuadrado, Dickey-Fuller aumentado, Kwiatkowski-Phillips-Schmidt-Shin, prueba t de Student, prueba t de Studen emparejada, prueba de análisis de varianza (ANOVA), prueba U de Mann-Whitney, prueba de rangos con signo de Wilcoxon, prueba H de Kruskal-Wallis y por último, prueba de Friedman.

A. Prueba de Shapiro-Wilk

Es una prueba estadística que se utiliza para determinar si un conjunto de datos sigue una distribución normal. La hipótesis nula (H0) de la prueba, es que los datos siguen una distribución normal. Por lo tanto, si el valor p resultante de la prueba es menor que el nivel de significancia (generalmente 0.05), se rechaza la hipótesis nula y se concluye que los datos no siguen una distribución normal.
Se calcula utilizando la siguiente ecuación:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Esta prueba se suele usar en muchos análisis para corroborar que la distribución de un atributo sea normal. Esto es útil porque te permite, posteriormente, elegir de mejor manera los métodos estadísticos más aptos para el dataset (esto porque ciertas técnicas requieren que los datos sigan una distribución normal, como la prueba t Student). Cabe mencionar que esta prueba funciona mejor para muestras pequeñas, por lo que, si esto no se cumple, será mejor utilizar otros métodos.

B. Prueba de K^2 de D'Agostino's

La prueba D'Agostino's K^2 es una prueba estadística que se utiliza para determinar si un conjunto de datos sigue una distribución normal. La hipótesis nula (H0) de la prueba, es que los datos siguen una distribución normal.
Esta prueba se basa en una combinación de la asimetría (skewness) y la curtosis (kurtosis) de los datos. La asimetría mide la simetría de la distribución, mientras que la curtosis mide la "pesadez" de las colas de la distribución.
La prueba D'Agostino's K^2 se calcula utilizando la siguiente ecuación:

$$K^2 = Z_{skewness}^2 + Z_{kurtosis}^2$$

Donde $Z_{skewness}$ es el valor de la asimetría de los datos y $Z_{kurtosis}$ el valor de la curtosis de los datos.
Los casos de uso para esta prueba son similares a la anterior, con la diferencia de que se adapta mejor a muestras más grandes, pues su potencia aumenta con el tamaño de la muestra, esto provoca que se utilice principalmente en grandes conjuntos de datos.

C. Prueba de Anderson-Darling

La prueba Anderson-Darling es una prueba estadística que se utiliza para verificar si una muestra de datos sigue una distribución normal. La hipótesis nula (H0) de la prueba Anderson-Darling es que los datos siguen la distribución especificada, en este caso, una distribución normal.
La prueba Anderson-Darling se calcula utilizando la siguiente ecuación:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln\left(Z_{(i)}\right) + \ln\left(1 - Z_{(n+1-i)}\right)\right)$$

Donde $Z_{(i)}$ son los datos ordenados y transformados a la distribución acumulada de probabilidad de la distribución normal y n es el tamaño de la muestra.
Se usa cuando se requiere verificar que la distribución de una muestra sea normal y además se deseen detectar desviaciones de la normalidad en las colas de distribución; esto debido a que esta prueba es especialmente sensible en los extremos de la distribución (colas).

D. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida estadística que calcula la fuerza y la dirección de la relación lineal entre dos variables continuas. Es paramétrica, los supuestos que considera es que las observaciones en cada muestra siguen una distribución normal y que tienen la misma varianza.
El coeficiente de correlación de Pearson se calcula utilizando la siguiente ecuación:

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

Cuando r es cercano a 1 se concluye que hay una fuerte correlación positiva y si es cercano a -1 una correlación negativa, la falta de relación se indica con 0.
Esta técnica se utiliza cuando se requiere analizar la relación entre dos atributos, específicamente su relación lineal, pues cabe el caso en el que la relación entre dos variables sea curvilínea (o de otra forma) en tales casos se preferirán otros métodos.

E. Correlación de rangos de Spearman

La correlación de rangos de Spearman, es una medida estadística que calcula la fuerza y la dirección de la relación monotónica entre dos variables. Es una medida de correlación no paramétrica.
Se calcula utilizando la siguiente ecuación:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

El resultado es un valor entre -1 y 1 inclusive, donde 1 indica una correlación positiva perfecta, -1 una correlación negativa perfecta, y 0 ninguna correlación monotónica.

La correlación de rangos de Spearman es una medida de correlación no paramétrica, lo que significa que no asume ninguna distribución específica de los datos. Por lo tanto, es especialmente útil cuando los datos no cumplen las suposiciones de las pruebas paramétricas, como la normalidad.

F. Correlación de Rangos de Kendall

Es una medida estadística que calcula la relación entre dos variables. Es una prueba de correlación no paramétrica. La correlación de rangos de Kendall se calcula utilizando la siguiente ecuación:

$$\tau = \frac{(n_c - n_d)}{0.5n(n-1)}$$

Donde *nc* es el número de pares concordantes, *nd* el número de pares discordantes y *n* el número de observaciones. Un par es concordante si las dos variables cambian en la misma dirección (es decir, ambas aumentan o ambas disminuyen) y es discordante si las dos variables cambian en direcciones opuestas (es decir, una aumenta y la otra disminuye).
Es importante mencionar que esta prueba es no paramétrica. Además, a diferencia de la correlación de rangos de Spearman, la correlación de rangos de Kendall tiene en cuenta la magnitud de la diferencia entre los rangos, lo que puede ser útil en ciertas situaciones.

G. Prueba Chi-Cuadrado

La prueba de Chi Cuadrado es una prueba estadística no paramétrica que se utiliza para determinar si hay una asociación significativa entre dos variables categóricas. La hipótesis nula (H0) de la prueba de Chi Cuadrado es que las variables son independientes, es decir, no están asociadas. La fórmula para calcularlo es la siguiente:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Donde O_i son las frecuencias observadas y E_i son las frecuencias esperadas bajo la hipótesis nula.
Esta prueba es especialmente útil cuando los datos son categóricos y no cumplen las suposiciones de las pruebas paramétricas. Sin embargo, la prueba de Chi Cuadrado tiene ciertas suposiciones propias, como que las observaciones son independientes y que las frecuencias esperadas no son demasiado pequeñas.

H. Dickey-Fuller aumentado

La prueba de Dickey-Fuller Aumentada (ADF) es una prueba estadística que se utiliza para determinar si una serie temporal es estacionaria. La hipótesis nula (H0) de la prueba ADF es que la serie temporal tiene una raíz unitaria, es decir, no es estacionaria.

La prueba ADF se calcula utilizando una regresión de mínimos cuadrados ordinarios del siguiente modelo:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta \mathrm{y}_{t-1} + \cdots + \delta_p \Delta \mathrm{y}_{t-p} + \varepsilon_t$$

ADF es una prueba paramétrica y tiene ciertas suposiciones propias, como que los errores son independientes y están distribuidos normalmente. Si estas suposiciones no se cumplen, los resultados de la prueba ADF pueden ser engañosos.

I. Prueba Kwiatkowski-Phillips-Schmidt-Shin

La prueba de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) es una prueba estadística que se utiliza para determinar si una serie temporal es estacionaria alrededor de una media o una tendencia lineal. Una serie temporal estacionaria es aquella en la que las propiedades estadísticas, como la media y la varianza, son constantes a lo largo del tiempo.
La hipótesis nula (H0) de la prueba KPSS es que la serie temporal es estacionaria alrededor de una media o una tendencia lineal. La prueba KPSS se basa en la regresión lineal. Divide una serie en tres partes: una tendencia determinista (βt), un paseo aleatorio (rt) y un error estacionario (εt), con la ecuación de regresión:

$$x_t = r_t + \beta t + \varepsilon_1$$

KPSS es una prueba paramétrica y al igual que Dickey-Fuller, tiene las suposiciones propias de que los errores son independientes y están distribuidos normalmente.

J. Prueba t de Student

La prueba t de Student es una prueba estadística que se utiliza para determinar si la diferencia entre las medias de dos grupos es estadísticamente significativa. La hipótesis nula (H0) de la prueba t de Student es que las medias de los dos grupos son iguales. Se calcula utilizando la siguiente ecuación:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Estra prueba permite evaluar si hay diferencias significativas en algún aspecto, entre dos grupos diferentes pero relacionados, por lo que tiene aplicación en muchas áreas. Esta técnica asume que los datos siguen una distribución normal y que las varianzas de los dos grupos son iguales.

K. Prueba t de Student emparejada

Es una versión de la prueba t de Student, pero se utiliza cuando los grupos que se están estudiando, están formados por los mismos individuos, es decir, cuando las observaciones no son independientes. La hipótesis nula (H0) de la prueba t emparejada es que la diferencia media entre los pares de observaciones es cero. La ecuación que la rige es la siguiente:

$$t = \frac{\bar{\bar{d}}}{s_d / \sqrt{n}}$$

Donde *d* es la media de las diferencias entre pares de observaciones, *sd* su desviación estándar y *n* es el número de pares de observaciones.

Se usa como el t Student común, pero en casos en los que cada individuo en una muestra aparece también en la otra.

L. Prueba de Análisis de Varianza (ANOVA)

El Análisis de Varianza (ANOVA) es una prueba estadística que se utiliza para determinar si existen diferencias significativas entre las medias de dos o más grupos independientes. La hipótesis nula (H0) de la prueba ANOVA es que todas las medias de los grupos son iguales. Su fórmula es la que sigue:

$$F = \frac{MSB}{MSW}$$

Donde *MSB* es la media de los cuadrados entre los grupos, que se calcula como la suma de los cuadrados entre los grupos dividida por el número de grados de libertad entre los grupos y *MSW* es la media de los cuadrados dentro de los grupos, que se calcula como la suma de los cuadrados dentro de los grupos dividida por el número de grados de libertad dentro de los grupos.

La prueba ANOVA es una prueba paramétrica, asume que los datos de cada grupo siguen una distribución normal y que las varianzas de los grupos son iguales.

M. Prueba U de Mann-Whitney

La prueba U de Mann-Whitney, es una prueba estadística no paramétrica que se utiliza para determinar si hay una diferencia significativa entre las distribuciones de dos grupos independientes.

La hipótesis nula (H0) de la prueba U de Mann-Whitney es que las distribuciones de los dos grupos son iguales. Se calcula usando la siguiente ecuación:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Donde n_1 y n_2 son los tamaños de los dos grupos y R_1 es la suma de los rangos de las observaciones del primer grupo.

N. Prueba de rangos con signo de Wilcoxon

La prueba de rangos con signo de Wilcoxon es una prueba estadística no paramétrica que se utiliza para determinar si hay una diferencia significativa entre las distribuciones de dos grupos emparejados o relacionados.

La hipótesis nula (H0) de la prueba de rangos con signo de Wilcoxon es que las diferencias entre los pares de observaciones provienen de una distribución simétrica alrededor de cero. Se calcula con la siguiente ecuación:

$$W^+ = \sum_{z_i > 0} R_i$$

La prueba de rangos con signo de Wilcoxon tiene ciertas suposiciones propias, como que las observaciones son independientes y que las diferencias entre los pares de observaciones provienen de una distribución simétrica alrededor de cero.

O. Prueba de Kruskall-Wallis

Al igual que las dos pruebas anteriores, es no paramétrica y se utiliza para determinar si hay una diferencia significativa entre las distribuciones de dos o más grupos independientes (aquí la diferencia con otras medidas, pues puede comparar varios grupos).

La hipótesis nula (H0) de la prueba H de Kruskal-Wallis es que todas las distribuciones de los grupos son iguales. La prueba H de Kruskal-Wallis se calcula utilizando la siguiente ecuación:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(n+1)$$

Donde *n* es el número total de observaciones, *g* es el número de grupos, R_i es la suma de los rangos de las observaciones en el grupo i, n_i es el número de observaciones en el grupo i.

P. Prueba de Friedman

Parecido a la anterior, permite detectar diferencias en las distribuciones de tres o más grupos emparejados (aquí la diferencia). La hipótesis nula (H0) de la prueba de Friedman es que las distribuciones de los grupos emparejados son iguales. Su ecuación es la siguiente:

$$H = \frac{12}{Nk(k+1)} \sum R_j^2 - 3N(k+1)$$

Donde *N* es el número de pares de observaciones, *k* es el número de grupos y *Rj* la suma de los rangos de las observaciones en el grupo *j*.

IV. MATERIALES Y MÉTODOS

Las 16 pruebas se llevaron a cabo sobre dos datasets, el primero fue “DailyDelhiClimateTrain” el cual contiene datos sobre la temperatura, humedad, velocidad del viento y presión media de dicha ciudad en cada día del año. Por su parte para aquellas pruebas que requerían datos categóricos, se usó el dataset “Salary_Data” que contiene datos sobre diversas personas en el aspecto laboral, como su género, edad, nivel educativo, área/puesto de trabajo, años de experiencia y salario.

V. RESULTADOS

Antes de realizar los ejemplos para cada una de las pruebas de hipótesis, se procedió a visualizar la información necesaria para entender los resultados que cada una de estas pruebas arrojaría. Esta información necesaria fue la media, varianza y desviación estándar de cada atributo, el valor mínimo y máximo de cada uno, las distribuciones y las series de tiempo. Todo esto se realizó sólo para el dataset del clima. A continuación, se muestra esta información:

TABLE I
MEDIA, DESVIACIÓN ESTÁNDAR Y VARIANZA

Atributo	Media	Desviación Estándar	Varianza
meantemp	25.495521	7.348103	53.994614
humidity	60.77170292	16.769652	281.221237
Wind_speed	6.802209	4.561602	20.808214
meanpressure	1011.104548	180.231668	32483.454272

TABLE II
MÍNIMO Y MÁXIMO DE CADA ATRIBUTO

Atributo	Mínimo	Máximo
meantemp	6.0	38.7142
humidity	13.4285	100.0
Wind_speed	0.0	42.22
meanpressure	-3.0416	7679.3333

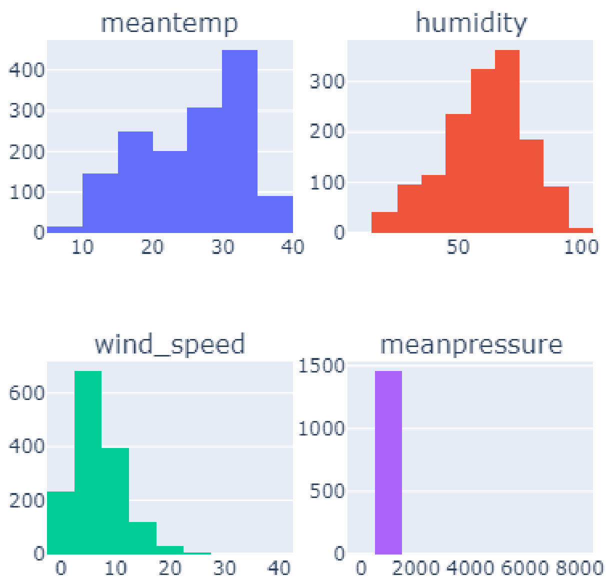

FIGURA 1. **Distribuciones de los atributos para el dataset del clima.**

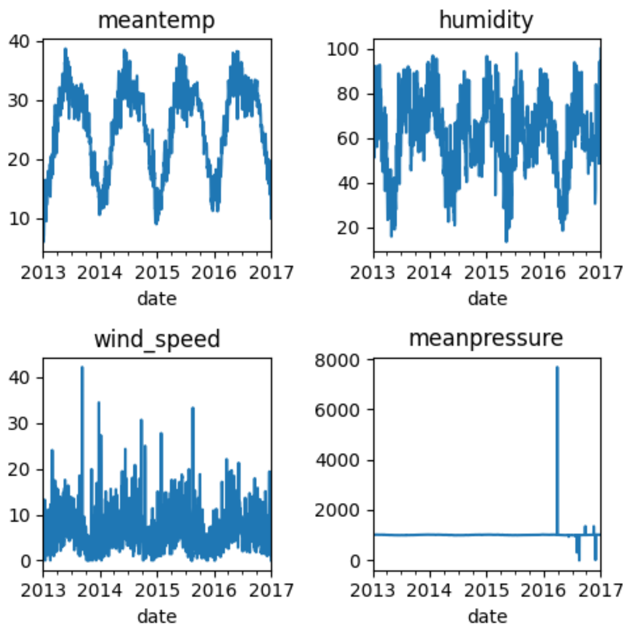

FIGURA 2. **Cambio de los datos de los atributos a través del tiempo.**

A. Prueba de Shapiro-Wilk

Esta prueba permite conocer si la distribución de un atributo es normal. En este caso se evaluó el atributo "*humidity*" el cual se puede observar visualmente que podría cumplir con esta característica, después de correr el código se obtuvieron los siguientes valores:

stat	0.985
p	0.000

Imprimiendo en pantalla que es "Probablemente No Gaussiana".

B. Prueba de K^2 de D'Agostino's

Para esta prueba se analizó el atributo *"humidity"*, al correr el test se obtuvo lo siguiente:

stat	35.517
p	0.000

Por lo que se obtuvo como resultado "Probablemente No Gaussiana".

C. Prueba de Anderson-Darling

Esta es la última prueba de normalidad que se usó. Se examinó el mismo atributo que las anteriores. Por la naturaleza de la prueba no se muestran los resultados en una tabla.

Se obtuvo un valor "stat = 5.562" y el resultado en pantalla fue "Probablemente No Gaussiana al 1.0%.

D. Coeficiente de correlación de Pearson

Ahora se comienzan con las pruebas de correlación. Se examinó la relación entre la columna "*meantemp*" y "*humidity*", esto debido a que eran los atributos que poseían las varianzas más parecidas lo cual es uno de los supuestos de este método. Los resultados fueron:

stat	-0.572
p	0.000

Lo que significa "Probablemente Dependiente". Esto tiene sentido dado que estas variables suelen cambiar con cierta frecuencia si la otra lo hace, hablando de los datos climáticos.

E. Correlación de rangos de Spearman

Los resultados para esta prueba usando los atributos "*meantemp*" y "*humidity*" fueron:

stat	-0.576
p	0.000

Estos resultados indican "Probablemente Dependiente".

F. Correlación de Rangos de Kendall

En esta prueba se usaron los mismos atributos que las dos anteriores. Los resultados arrojados se muestran a continuación:

stat	-0.411
p	0.000

Lo que significa "Probablemente Dependiente".

G. Prueba Chi-Cuadrado

Esta es la última de las pruebas de correlación aplicadas. Sin embargo, el dataset que se usó fue diferente, pues Chi-Cuadrada está hecha para atributos categóricos. Tal dataset fue "*Salary_Data*". Para este se examinaron los atributos "*Gender*" y "*Job_Title*" con el fin de revisar si había relación entre el trabajo desempañado y el sexo de la persona. Los resultados de la prueba fueron los siguientes.

stat	1536.325
p	0.000

Lo cual indica que hay cierta dependencia entre ambos atributos.

H. Dickey-Fuller aumentado

Aquí comienzan las pruebas de estacionariedad. En este caso se hizo la prueba con el atributo "*meantemp*". El resultado de esta prueba fue:

stat	0.188
p	0.100

Esto significa que "La serie temporal probablemente es estacionaria".

I. Prueba Kwiatkowski-Phillips-Schmidt-Shin

Al igual que en la anterior, se usó el atributo "*meantemp*". Los resultados fueron:

stat	0.188
p	0.100
Valores críticos	{'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}

Lo que arroja "La serie temporal probablemente es estacionaria".

J. Prueba t de Student

A continuación, se aplicaron las pruebas de hipótesis estadísticas paramétricas. Para t de Student se trabajó con los atributos "*meantemp*" y "*humidity*", pues son los que poseen distribuciones más similares. Se obtuvo lo siguiente:

stat	-73.670
p	0.000

O sea que tienen "Probablemente Diferente Distribución".

K. Prueba t de Student emparejada

Para esta prueba igualmente se usó "meantemp" y "humidity". Los valores calculados fueron:

stat	-61.812
p	0.000

Por lo que poseen "Probablemente Diferente Distribución".

L. Prueba de Análisis de Varianza (ANOVA)

Esta es la última de las pruebas de hipótesis estadísticas paramétricas. Se usaron los mismos atributos que en las últimas dos, y los resultados son:

stat	42852.883
p	0.000

Por lo que tienen “Probablemente Diferente Distribución”.

M. Prueba U de Mann-Whitney

Por último, se muestran los resultados de las pruebas de hipótesis estadísticas no paramétricas. Para esta se usaron los atributos “*meantemp*” y “*humidity*”. Los resultados se muestran a continuación:

stat	86147.000
p	0.000

Esto indica “Probablemente es Diferente Distribución”.

N. Prueba de rangos con signo de Wilcoxon

Se usaron los mismos atributos para esta prueba, se obtuvo lo siguiente:

stat	14236.500
p	0.000

Lo que indica “Probablemente Diferente Distribución”.

O. Prueba de Kruskall-Wallis

Igual se examinaron los mismos atributos, los valores obtenidos fueron:

stat	1853.077
p	0.000

Esto es indicador de “Probablemente Diferente Distribución”.

P. Prueba de Friedman

Por último, en esta prueba se examinaron los cuatro atributos “*meantemp*”, “*humidity*”, “*wind_speed*”, “*meanpressure*”. Los resultados fueron:

stat	4239.155
p	0.000

Esto indica que los atributos poseen “Probablemente Diferente Distribución”.

VI. CONCLUSIONES

En el presente documento se dieron los fundamentos teóricos y algunos ejemplos para diversas pruebas de hipótesis que abarcaban: pruebas de normalidad, pruebas de correlación, pruebas de estacionariedad, pruebas de hipótesis estadísticas paramétricas y pruebas de hipótesis estadísticas no paramétricas.

Con respecto a las pruebas de normalidad, se examinó el atributo “humidity”, que visualmente se veía más parecido a una distribución normal; sin embargo, los resultados arrojaron que este no lo era. Probablemente esto se puede explicar por el hecho de que las pruebas toman en cuenta ciertos valores que no se ven en las gráficas y que provocan que la distribución esté sesgada.

Con respecto a las pruebas de correlación en las primeras pruebas se usó el dataset del clima y en la última (Chi-Cuadrado) el dataset de salarios. Con respecto al primer dataset se examinó la relación entre el atributo “*meantemp*” y “*humidity*”, los resultados fueron parecía existir una correlación negativa. Por su parte para el dataset de salarios se usaron los atributos “*Job_Title*” y “*Gender*”, se mostró que había cierta dependencia entre ambos.

Por su parte, los resultados de las pruebas de estacionariedad, ambos arrojaron que la serie parecía ser estacionaria, esto parece tener algo de sentido dado que el atributo probado fue “meantemp”, y si se revisa la serie temporal (figura 2) los patrones se repiten cada año, lo cual parece ser la causa del resultado.

Por último, se revisaron las pruebas de hipótesis estadísticas paramétricas y no paramétricas. En todas se usaron los atributos “meantemp” y “humidity” y arrojan el mismo resultado de “Probablemente Diferente Distribución”, lo cual tiene sentido si se visualizan los histogramas de la figura 1. En el último test se examinaron todos los atributos y se obtuvo la misma respuesta.

BIBLIOGRAFÍA

[1] M. A. Aceves Fernandez, Inteligencia Artificial para Programadores con Prisa. Independently Published, 2021.

[2] F. Berzal, Redes Neuronales & Deep Learning, vol. 1, 2019.