UNIVERSIDAD NACIONAL DE COLOMBIA- SEDE MEDELLÍN

ANÁLISIS DE DATOS AMBIENTALES-TRABAJO 1

4.6
Muy bien :)

Juan Pablo Ramírez Monsalve CC. 1036670881
Diana Camila Arroyave Romero CC. 1152457186

En el presente trabajo se busca caracterizar el comportamiento de una serie de datos mediante del uso de diversos parámetros estadísticos. La serie analizada consta de datos de temperatura media mensual de Colombia, que abarca el período comprendido entre enero de 1930 hasta diciembre de 2015, para un total de 1032 datos.

Se realizó la gráfica de dicha serie con el fin de realizar un primer acercamiento, analizando características que puedan ser de interés así como las posibles causas del comportamiento presentado. La gráfica se presenta a continuación

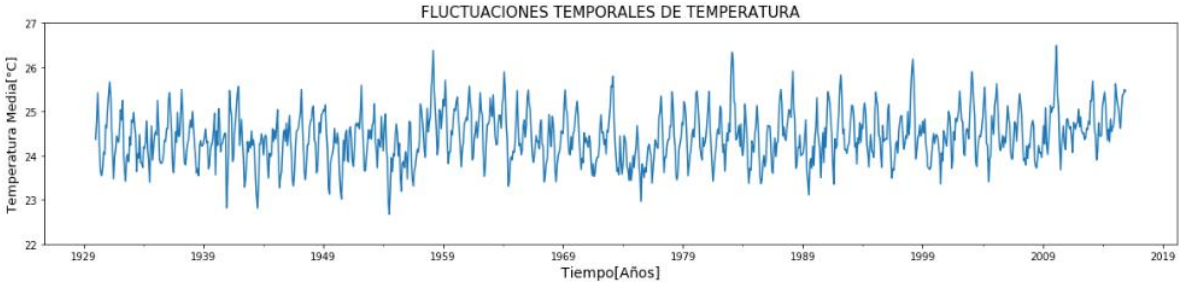

Figura 1. Serie de datos

Nota: *Si bien los datos son mensuales, se optó por una representación anual de las abscisas para una mejor visualización de la evolución de la serie (recurrente en las demás gráficas).*

Se puede observar un comportamiento con cierta periodicidad, presentando valores extremos de manera regular a causa del ciclo climatológico de temperatura del país, en el cual se presentan temporadas cálidas o frías en meses particulares del año (las cuales ocurren en periodos diferentes dependiendo de la región del país).

Ojo con la redacción y el lenguaje técnico

Dentro de la gráfica anterior cabe destacar la tendencia que se da entre los años de 1955 a 1959, presentando inicialmente una ligera tendencia decreciente seguida de una creciente; esto último puede deberse a la presencia de los fenómenos de "La Niña" y "El Niño" que se registraron en la región 3.4 durante el intervalo de tiempo mencionado; encontrándose así episodios fríos (La Niña) durante los años de 1954 a 1956, seguido de un aumento significativo en las temperaturas, con un episodio cálido (El Niño) entre 1957 a 1959 aproximadamente.

A continuación se presenta una vista en detalle del período de 1950 a 1965:

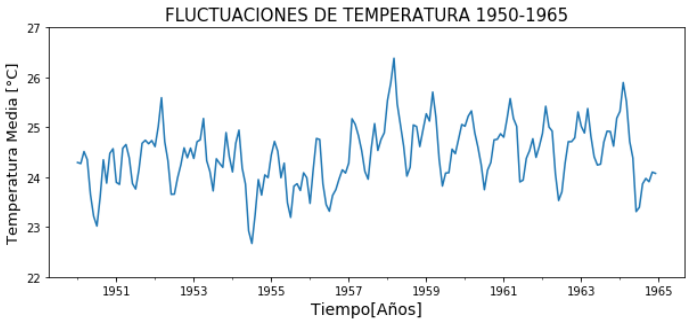

Figura 2. Vista a detalle 1950-1965

FUNCIÓN DE DISTRIBUCIÓN ACUMULADA

La función de distribución acumulada para el caso de estudio permite determinar el rango de temperaturas sobre el cual es más probable encontrar un dato registrado en la serie. Con esto se busca observar el comportamiento de los eventos extremos mencionados anteriormente, y si estos presentan altas o bajas probabilidades de ocurrencia.

A continuación se presenta la gráfica de la función de distribución acumulada

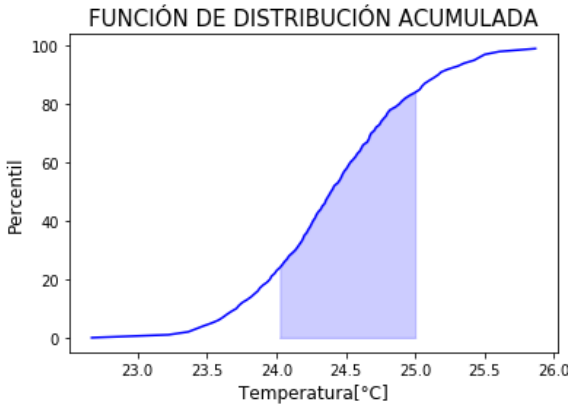

Figura 3. Función distribución acumulada

Se observan temperaturas mínimas y máximas de 22.5°C y 26°C respectivamente, obteniéndose una variación entre eventos extremos de aproximadamente 3°C. Cabe destacar que la variación entre 24°C y 25°C se encuentra entre el percentil 25 y el percentil 85, esto último nos indica que hay una probabilidad de 25 % de encontrar valores menores a los 24°C y de 15 % de encontrar temperaturas superiores a los 25°C, siendo estos rangos de valores los correspondientes a los eventos extremos significativos para períodos fríos y cálidos respectivamente. Resaltar que esto último no implica que estos eventos con poca probabilidad de ocurrencia se deban a fenómenos poco comunes o inusuales que generen aumentos o descensos repentinos en la temperatura; todo esto, partiendo de que en el caso analizado, estos eventos ocurren en meses puntuales del año (oleadas periódicas de frío o calor), y la cantidad de estos es significativamente menor respecto a la cantidad de meses en los cuales se tienen temperaturas medias entre los 24°C y 25°C.

Para visualizar mejor la distribución de los datos dentro de la serie se presenta el histograma de la misma.

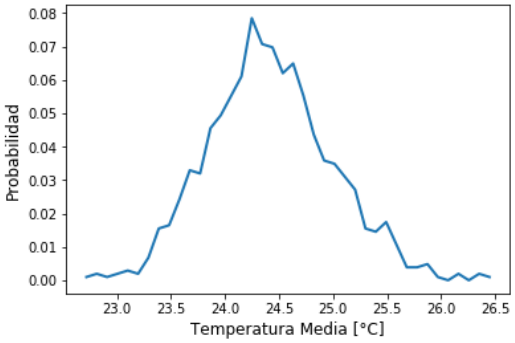

Figura 4. Histograma serie de temperatura mensual

Se observa una forma similar a la distribución normal o Gaussiana, donde al igual que en la función de distribución acumulada se tienen las mayores probabilidades de ocurrencia para el rango de temperaturas de entre 24°C y 25°C.

Para determinar si efectivamente la serie de datos presenta una distribución normal, se hizo el Test de Shapiro-Wilk, donde se obtuvo los siguientes:

$$Estadístico\ de\ prueba\ =\ 0.9959155321121216$$

$$Valor\ p\ =\ 0.00781430397182703$$

Ya que el *valor p* es mucho menor que el estadístico de prueba, se tiene suficiente evidencia para no rechazar la hipótesis de normalidad y por tanto se asume que se distribuyen normalmente.

NOTA: *Para la elaboración del histograma se hizo uso de un número de clases igual a 40, buscando eliminar el ruido dentro de la gráfica, pero evitando la pérdida de información en la misma.*

Si bien la distribución de la serie analizada es gaussiana, conviene estudiar si presenta alguna variabilidad con el tiempo; es decir, si el comportamiento de ésta es dependiente del intervalo de tiempo en el cual se evalúe y del tamaño del mismo. Con esto se busca una mejor comprensión de las fluctuaciones de temperatura a lo largo de los años para ver si el comportamiento actual es representativo del comportamiento que pudo tener en el pasado, y en caso de no ser así, observar la evolución del mismo.

Con el fin de tener un número de datos significativos que brinde mayor precisión al cálculo de cada uno de los histogramas, y evitar la aparición de comportamientos poco representativos en el resultado, se optó por la evaluación de estos en periodos de 10 años, con un rezago de un año entre uno y otro, para un total de 912 histogramas que se presentan de manera conjunta en la imagen a continuación.

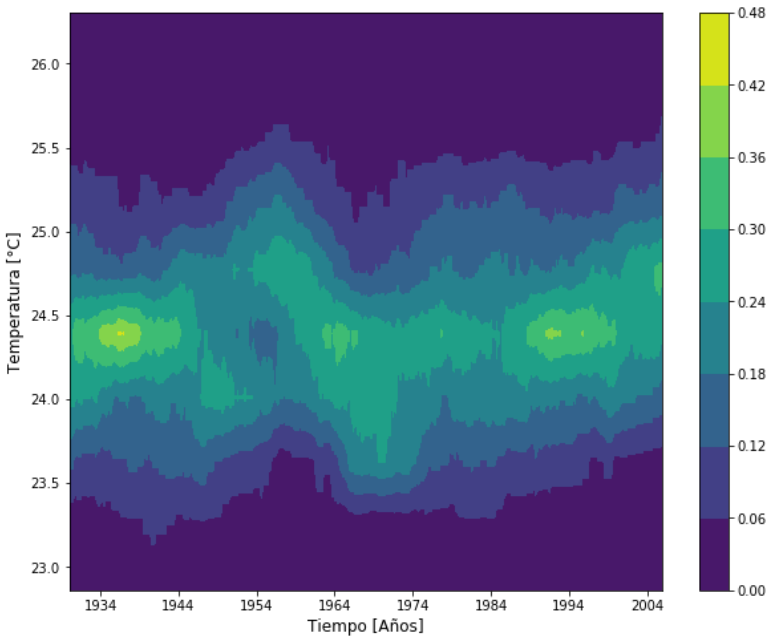

Figura 5. Variabilidad temporal de histogramas

Nuevamente se observan las mayores concentraciones de datos entorno a las temperaturas del tercio medio, observándose poca acumulación respecto a valores cercanos a los 23.5°C y los 25°C, y siendo prácticamente nulos en los valores más extremos de 23°C y 26°C. De lo anterior se puede concluir que si bien a lo largo del tiempo los histogramas presentan ligeras variaciones respecto a la forma de su distribución, siempre mantuvo de manera general la observada en la figura 4; estas variaciones mencionadas son producto de las fluctuaciones en los datos, propias de cada intervalo de 10 años analizados, las cuales pueden modificar los parámetros estadísticos como la media, mediana, varianza, etc.

Bajo el análisis anterior se procede a analizar dichos parámetros para observar su variabilidad en el tiempo.

La importancia de la inclusión de la estadística no paramétrica dentro del análisis de la serie radica en que no se puede definir o establecer a priori que esta siga una distribución conocida, ya que son los mismos datos observados los que determinan esto. Para el caso puntual analizado, a pesar de haber determinado el grado de similitud que presenta la distribución de los datos respecto a la distribución normal, los momentos no paramétricos siguen teniendo relevancia dentro del análisis, ya que algunos de ellos deben presentar un comportamiento característico basado en propiedades de la distribución, como puede ser la simetría que presenta respecto a los valores cercanos a los 24.5°C.

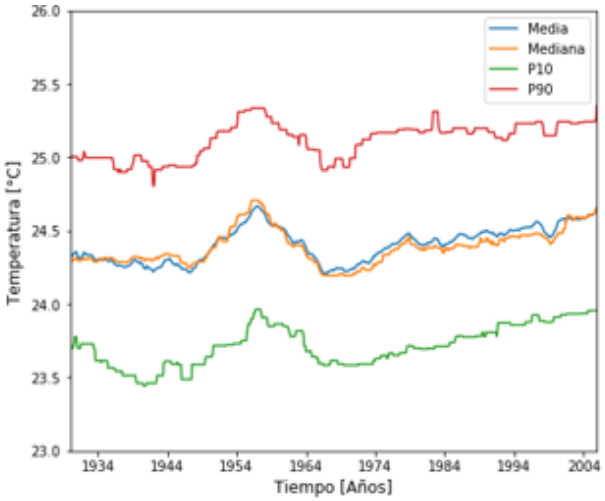

Figura 6. Media, Mediana, P10 y P90

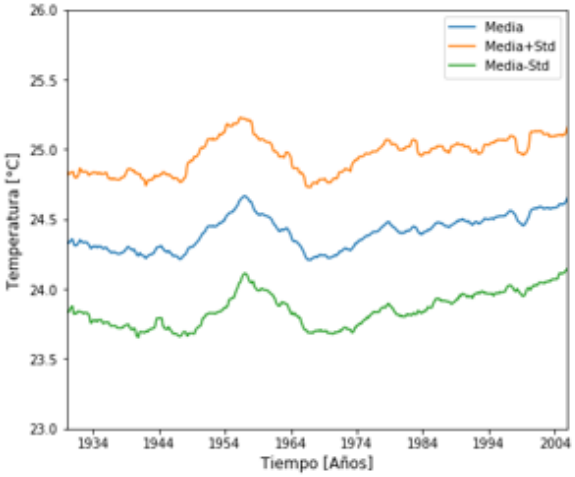

Figura 7. Media y desviación estándar

En ambas figuras se nota nuevamente la agrupación de la mayoría de datos respecto al intervalo de temperatura de 24°C a 25°C aproximadamente, mostrando que la probabilidad de encontrar valores mayores a 25.5°C o menores a 23.5°C es del 10 % (figura 6), y permitiendo ver que la dispersión de los datos respecto a la media es del orden de 0.5°C (tanto hacia arriba como hacia abajo), lo que corresponde a una dispersión del 15 % respecto a la diferencia máxima posible de 3°C (figura 7).

Se destaca la similitud entre los valores de la media y la mediana, esto se debe a lo mencionado anteriormente, ya que estos momentos estadísticos tienden a ser iguales en cuanto mayor es la simetría que presenta la distribución. Respecto a esto cabe mencionar que si bien la media y la mediana son iguales en una distribución Gaussiana, no es exclusivo de la misma, ya que se pueden encontrar distribuciones simétricas que cumplan esto sin ser Gaussianas o aproximarse a estas.

No, habla de la presencia de outliers en la serie y la cantidad de datos de la misma

Por último se resalta el crecimiento de los momentos estadísticos alrededor del año 1959, correspondiente al ya mencionado fenómeno del niño que se desarrolló en este periodo de tiempo y que generó un incremento significativo en las temperaturas medias mensuales registradas.

Tras analizar el comportamiento tanto de la media como la mediana, es necesario realizar un análisis a profundidad de la dispersión que pueden presentar los datos respecto a estas, para esto se estudia el comportamiento que presenta la desviación estándar y el rango intercuartil (IQR).

Ojo con la leyenda, además dada tu distribución no me convencen estos resultados

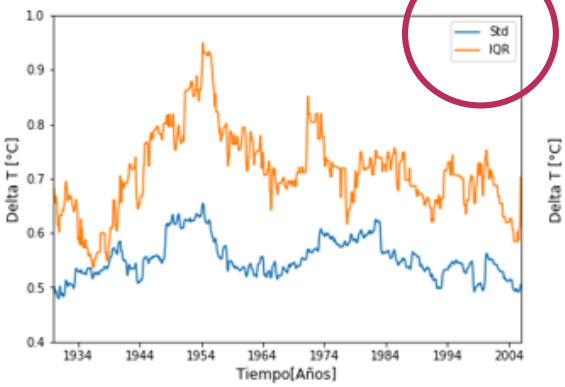

Figura 8. Coeficiente de asimetría y Yule Kendall

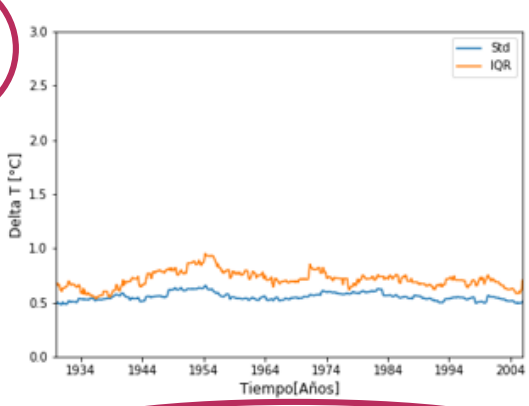

Figura 9. Vista comparativa con variación máxima

???

En la figura 8 se observa que si bien se presenta una diferencia entre ellas, este tiene un valor máximo de aproximadamente 0.3°C, el cual puede considerarse pequeño en una serie que presenta fluctuaciones máximas de 3°C (figura 9). Este desfase puede deberse a la naturaleza de la distribución, ya que al ser simétrica y con la mayor concentración de datos cercanos a la media, la desviación estándar es menor (puesto que esta lo que hace es comparar el valor en la posición *i* respecto al valor de la media), mientras que el IQR al simplemente tomar los valores de los cuartiles 1 y 3 no toma en cuenta la centralización que presenta la serie.

Si bien a partir del histograma y la semejanza entre la media y la mediana dan un indicio de la simetría que presenta la serie, para determinar cómo se ha comportado esta simetría a lo largo del tiempo se recurre al cálculo del coeficiente de asimetría (medida paramétrica) y el coeficiente de Yule Kendall (medida no paramétrica), los cuales se encargan de medir cómo se agrupan los datos respecto a la media y la mediana respectivamente.

En ambos casos, un valor positivo del coeficiente en cuestión implica que se posee una asimetría positiva, es decir, una acumulación de los datos hacia la izquierda de la distribución, de manera análoga un valor negativo indica una acumulación hacia la derecha. Valores cercanos a cero indican simetría respecto a la media o la mediana dependiendo del parámetro evaluado.

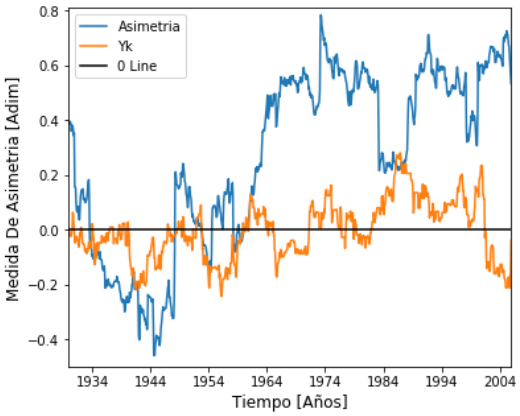

Figura 10. Medidas de asimetría

Esto tienen mucho más sentido

A simple vista no se puede concluir sobre la simetría de la distribución de los datos, porque aunque hay intervalos de tiempo donde coinciden y tanto el coeficiente de asimetría y el Yule Kendall son positivos o negativos, también los hay donde son contrarios.

Para determinar si la distribución es simétrica o asimétrica -en este segundo caso también la naturaleza de la asimetría- se procedió a tomar tres momentos de la serie: cuando ambos estadísticos son positivos, cuando ambos son negativos y cuando son contrarios, y analizarlos más detalladamente a partir de los histogramas

- **Ambos parámetros negativos**

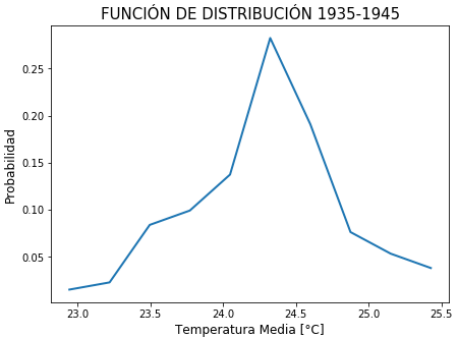

Figura 11. Histograma 1935-1945

El primer caso es cuando ambos son negativos lo que indica una asimetría negativa, para corroborar los resultados se encontraron los valores de la Media, Mediana y la Moda, de modo que

$$\text{Media} = 24.2947816794\ °C \quad \text{Mediana} = 24.3161\ °C \quad \text{Moda} = 24.6023\ °C$$

$$Media < Mediana < Moda$$

Lo que también indica asimetría negativa.

- **Ambos parámetros positivos**

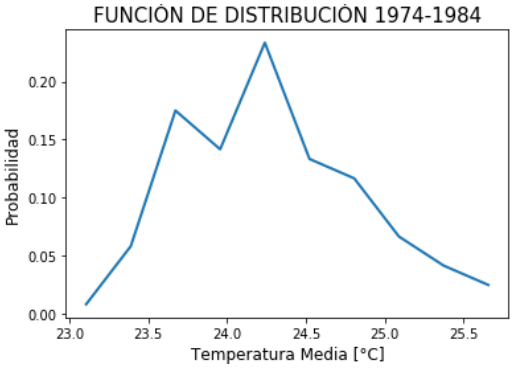

Figura 12. Histograma 1974-1984

El segundo caso es cuando ambos son positivos lo que indica una asimetría positiva, en cuanto a las medidas de tendencia central reafirman esto último:

$$\text{Media} = 24.2761966667\ °C \quad \text{Mediana} = 24.23215\ °C \quad \text{Moda} = 22.9624\ °C$$

????

$$Media > Mediana > Moda$$

- Coeficiente de asimetría positivo y Yule Kendall negativo

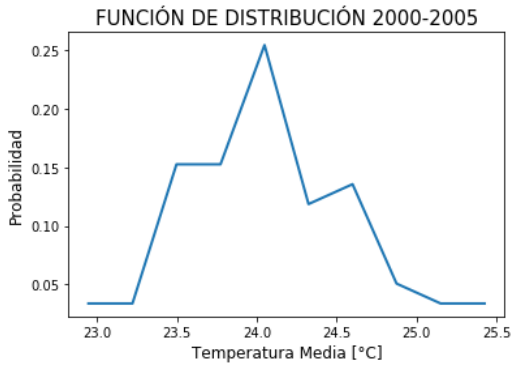

Figura 13. Histograma 200-2005

$$\text{Media} = 24.5110435754\ °C\ \text{Mediana} = 24.4658\ °C\ \text{Moda} = 24.2185\ °C$$

$$Media > Mediana > Moda$$

En el último caso la relación que presentan la media, la mediana y la moda coinciden con el coeficiente de asimetría diciendo que la distribución es asimétrica positiva.

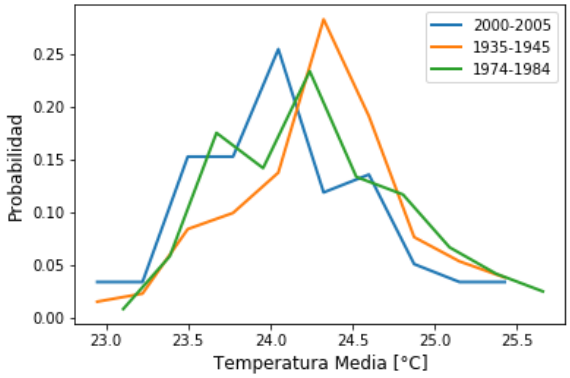

Figura 14. Histogramas superpuestos

Con base en las tres comparaciones, y por la naturaleza gaussiana de la distribución, se concluye que el parámetro que mejor representa el comportamiento de la serie en cuanto a la distribución de los datos es el coeficiente de asimetría. Independiente de esto, partiendo de los valores máximos y mínimos que toman ambos estadísticos [-0.4 , 0.8] se puede afirmar que la serie tiene elevado grado de simetría, lo que es consecuente con lo afirmado en el análisis de parámetros anteriores sobre la alta probabilidad de obtener valores de temperatura entre 24 °C y 25 °C, y una menor probabilidad para valores por fuera de este rango.

Por último se busca determinar cómo evoluciona la serie estudiada a lo largo del tiempo, particularmente se determinará si tiene alguna tendencia creciente, decreciente, o no presenta tendencia alguna; en nuestro caso, esto nos permitirá obtener información sobre si la temperatura media mensual de Colombia ha aumentado, disminuido o se ha mantenido estable desde la década de los 30.

Para esto se hace uso del test de Mann Kendall, el cual es un test no paramétrico que permite analizar estas tendencias

Antes de iniciar el test se presentan una serie de recomendaciones que deben tenerse en cuenta al momento de realizar el análisis sobre el resultado obtenido

- **Los datos no se recopilan estacionalmente:** las tendencias ascendentes y descendentes de los datos pueden afectar el resultado del test.
- **Los datos no presentan ninguna covariable:** si se está analizando la influencia de algún factor en la variable presente en los datos, se debe garantizar que este es el único factor que afecta dicha variable, ya que la existencia de otro factor que presente correlación con la variable estudiada puede llevar a conclusiones erróneas al no tenerse en cuenta.
- **Sólo se tiene un dato por valor de tiempo:** Si se tienen varios datos registrados para el mismo tiempo se recomienda realizar el test usando un parámetro como la media o la mediana

El test de Mann Kendall se describe a continuación.

Se parte de una hipótesis nula (H_o) en la cual se plantea que la serie presenta un comportamiento aleatorio, y una hipótesis alternativa ($H_{\boldsymbol{\alpha}}$) la cual puede ser la existencia de una tendencia positiva, negativa o que simplemente presente tendencia. Tras la definición de estas se siguen los siguientes pasos

1. Estando los (n) datos organizados en el orden en que fueron recopilados se procese a hacer la comparación entre cada uno de ellos con los demás de la serie; es decir, se determina el signo de todas las posibles $n(n-1)/2\ x_j - x_k$ donde $j > k$
2. De cada una de estas diferencias se determina el signo de las mismas definimiento la siguiente función :

$$signo\ (x_j - x_k) = 1\ si\ x_j - x_k > 0$$
$$signo\ (x_j - x_k) = 1\ si\ x_j - x_k < 0$$
$$signo\ (x_j - x_k) = 0\ si\ x_j - x_k = 0$$

3. Calcular el número de diferencias positivas menos el número de diferencias positivas mediante la siguiente expresión:

$$S = \Sigma_{k-1}^{n-1} \Sigma_{j-k+1}^{n} signo(x_j - x_k)$$

4. Partiendo de que el número de datos es mayor a 10, se procede a calcular la varianza como sigue:

$$\mathrm{VAR}(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p-1}^{g} t_p(t_p-1)(2t_p+5)\right]$$

Siendo "g" la cantidad de datos cuyo valor se repiten y t_p el número de veces que se repite el número específico "p"

5. Calcular el estadístico de prueba Z_{MK} de la siguiente manera

$$Z_{MK} = \frac{S-1}{[Var(s)]^{\frac{1}{2}}} \quad si\ S > 0$$

$$Z_{MK} = \frac{S+1}{[Var(s)]^{\frac{1}{2}}} \quad si\ S < 0$$

$$Z_{MK} = 0 \quad si\ S = 0$$

6. Se define un nivel de significancia ($\boldsymbol{\alpha}$) para realizar la comparación del Z_{MK} con el Z de la distribución normal, este Z se define según la hipótesis alternativa planteada. Si $H_{\boldsymbol{\alpha}}$ se define como tendencia positiva o negativa se busca el Z de la normal correspondiente a $Z_{1-\alpha}$; en caso de que $H_{\boldsymbol{\alpha}}$ se defina como la existencia de tendencia el Z de la distribución normal se calcula como $Z_{1-\left(\frac{\alpha}{2}\right)}$

7. Realizar la comparación y análisis de hipótesis como sigue:

   Si $H_{\boldsymbol{\alpha}}$ es existencia de tendencia positiva, $H_{\boldsymbol{o}}$ es rechazado si se cumple que
   $$Z_{MK} \geq Z_{1-\alpha}$$

   Si $H_{\boldsymbol{\alpha}}$ es existencia de tendencia negativa, $H_{\boldsymbol{o}}$ es rechazado si se cumple que
   $$Z_{MK} \leq -Z_{1-\alpha}$$

   Si $H_{\boldsymbol{\alpha}}$ es existencia de tendencia, $H_{\boldsymbol{o}}$ es rechazado si se cumple que
   $$|Z_{MK}| \geq -Z_{1-\alpha/2}$$

Para nuestro caso puntual definimos la hipótesis nula y a hipótesis alternativa de la siguiente manera:

- $\boldsymbol{\alpha}$: 0.05
- H_o : La serie se distribuye de manera aleatoria
- $H_{\boldsymbol{\alpha}}$: La serie presenta tendencia positiva

Debido al comportamiento cuasi periódico que presenta la serie, se realizó el test de Mann Kendall para los parámetros estadísticos de media, mediana, percentil 90 y percentil 10, con el fin de disminuir los efectos de las tendencias propias que se observan en los datos en intervalos de tiempo mucho menores que el de la serie completa.

Los resultados se presentan a continuación.

Parámetro	S	Var (S)	Z MK	Z normal
Serie	57355	122299937	5.1862	1.644854
Media	185200	84421762	20.1560	1.644854
Mediana	155451	84408714	16.9190	1.644854
P90	181877	84349223	19.8031	1.644854
P10	209336	84382110	22.6780	1.644854

Tabla1. Resultados test Mann Kendall

De la tabla 1 se puede observar lo mencionado anteriormente, ya que los parámetros estadísticos presentan valores de Z cercanos a 20, mientras que el valor de la Serie de datos completa es de aproximadamente 5, lo que indica que el cálculo de la última puede verse influenciado por la cuasi periodicidad que presentan los datos.

Partiendo de las hipótesis nula y alternativa planteadas para el análisis, se rechaza la hipótesis nula de aleatoriedad, por lo que se asume que la serie presenta tendencia creciente. De los parámetros analizados se tiene que la tendencia creciente de los percentiles 90 y 10 implica que posiblemente en un análisis a futuro se tenga que los valores extremos no se presenten entre 25°C y 26°C (Derecha) o 23°C y 24°C (Izquierda) sino que dichos intervalos se moverán a temperaturas mayores; esto ya se puede observar de manera leve en la figura 1 donde las temperaturas más bajas y las más altas a partir de la década del 90 presentan valores mayores que las correspondientes en años anteriores, consecuente con el calentamiento global.

CONCLUSIONES

Bajo los análisis hechos se determinó la tendencia que ha tenido la temperatura media mensual al país desde la década de 1930 hasta la actualidad, sin embargo como ya se mencionó, series que presentan tendencias propias en intervalos pequeños de tiempo como la estudiada deben tratarse de manera rigurosa para no caer en conclusiones erróneas, por lo que se recomienda realizar el test de Mann Kendall estacional con el fin de obtener un resultado con mayor grado de precisión.

La obtención de la distribución normal es consecuente con la variación mínima que presenta una variable como la temperatura en latitudes cercanas al ecuador, donde no se tienen cambios drásticos como en latitudes altas o bajas producto del cambio de estaciones durante el transcurso del año; por lo que se espera que mientras más alejada del ecuador sea la región de análisis, menos semejanza presentará la distribución de temperaturas respecto a la distribución normal.

Bibliografía

BIDEGAIN, M. D. (2011). *Análisis Estadístico de Datos Climáticos.* Recuperado el 06 de 09 de 2018, de Distribuciones de Probabilidad: http://meteo.fisica.edu.uy/Materias/Analisis_Estadistico_de_Datos_Climaticos/teorico_AEDC/Distribuciones_Probabilidad_2011.pdf

Bocanegra, J. E. (31 de 12 de 2007). *Modelo institucional del IDEAM sobre el efecto climático de los.* Recuperado el 06 de 09 de 2018, de

http://www.ideam.gov.co/documents/21021/440517/Modelo+Institucional+El+Ni%C3%B1o+-+La+Ni%C3%B1a.pdf/232c8740-c6ee-4a73-a8f7-17e49c5edda0

Mann-Kendall Test For Monotonic Trend. (s.f.). Recuperado el 06 de 09 de 2018, de https://vsp.pnnl.gov/help/index.htm#vsample/design_trend_mann_kendall.htm

MOLANO, J., BATISTA, J. (1967). *Calendario Climatológico Areonáutico Colombiano.* Recuperado el 06 de 09 de 2018, de https://www.sogeocol.edu.co/documentos/096_calen_climat.pdf

Mann Kendall Trend Test: Definition, Running the test. Recuperado el 06 de 09 de 2018, de http://www.statisticshowto.com/mann-kendall-trend-test/

Yule and Kendall Coefficient. Recuperado el 06 de 09 de 2018, de https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-32833-1_431

Climate Change Knowledge Portal. Recuperado el 20 de 09 de 20018, de http://sdwebx.worldbank.org/climateportal/index.cfm?page=downscaled_data_download&menu=historical