



Tecnológico de Monterrey

Actividad:

Reporte Final: Los peces y el mercurio, Parte 2

Módulo:

Módulo 5: Estadística e Inteligencia artificial avanzada para ciencia de datos

Grupo:

TC3007C.501

Nombre:

Franco Quintanilla Fuentes - A00826953

Maestra:

Blanca R. Ruiz Hernández

Fecha de entrega:

2 de diciembre de 2022

Resumen	3
Introducción	3
Análisis de resultados	4
Análisis de normalidad	4
1. Prueba de Normalidad de Mardia	5
2. Prueba de Normalidad de Anderson Darling	5
Análisis de componentes principales (PCA)	8
Conclusión	10
Referencias	11
Anexos	11
Repositorio de Github	11

Resumen

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud, por lo que realizamos un estudio estadístico con el fin de determinar las variables que influyen a la contaminación por mercurio. Para poder responder las preguntas, utilizamos métodos estadísticos como: Regresión lineal simple, y Análisis de la Varianza (ANOVA). En donde llegamos a los resultados que el PH es un factor que ayuda a reducir la contaminación de mercurio en los peces, y que la concentración de mercurio no varía entre las edades de los peces.

Introducción

En la investigación de (OCU, 2021) se menciona que el mercurio se libera al medio ambiente a través de procesos naturales, el cual se presenta en el suelo, el agua y la atmósfera. El problema empieza cuando el humano aporta grandes cantidades de mercurio al medio ambiente a través de distintos procesos. Como repercusión, las grandes concentraciones de mercurio llegan al agua, por ende a los animales, que en nuestro caso a considerar, los peces.

A nosotros nos interesa hacer un estudio estadístico, ya que el mercurio puede ocasionar efectos tóxicos en tanto órganos como sistemas, algunos ejemplos son:

- El sistema nervioso
- Los riñones
- El hígado
- Los órganos reproductivos

Los daños más peligrosos son los neurotóxicos, ya que sus efectos repercuten sobre el desarrollo neuronal, y los períodos de exposición durante el embarazo. Por eso, nuestro estudio se basó en una pregunta base, la cual fue desglosando 2 preguntas consecuentes.

- ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Las preguntas subyacentes de la pregunta base, son las que más énfasis tienen a la hora de preocuparnos por la salud humana.

1. ¿En qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?
2. ¿Cómo te ayudan los componentes principales a abordar este problema?

Con estas preguntas podemos empezar a realizar nuestro análisis estadístico, para poder sacar resultados educados y sustentados con estadística.

Análisis de resultados

Análisis de normalidad

Para empezar el análisis de normalidad, lo primero que vamos a hacer es visualizar el comportamiento de los datos mediante el histograma de los mismos, en donde podemos ver que ninguna de las variables parece tener un comportamiento normal, aunque hay algunos que pueden tener una tendencia como el PH, o el máximo de la concentración de mercurio.

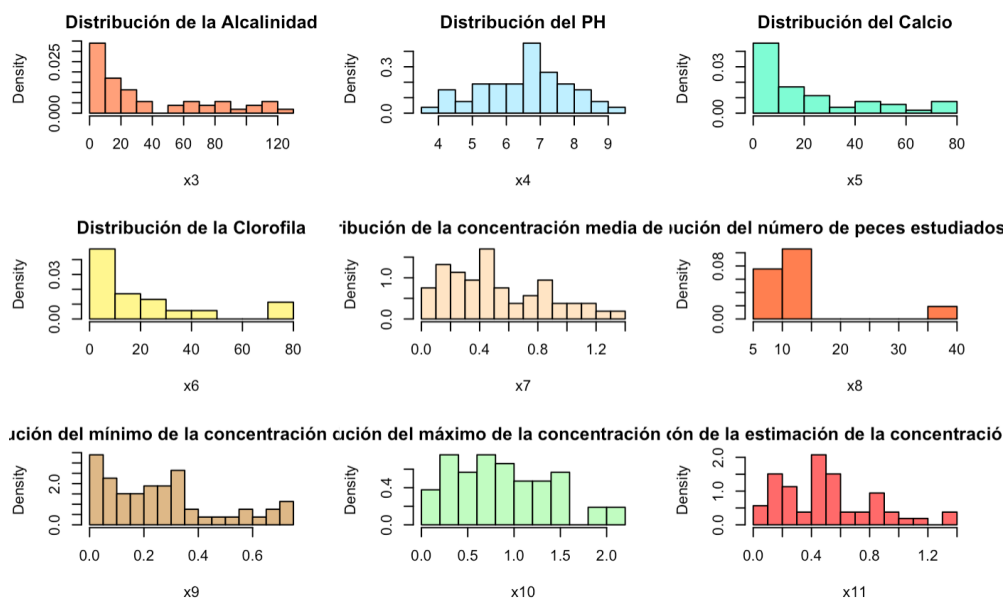


Figura 1: Distribución de las variables.

Para poder hacer un análisis más profundo de nuestros datos, vamos a realizar distintas pruebas de normalidad.

1. Prueba de Normalidad de Mardia

Para la prueba de normalidad de Mardia se utilizó la librería de R de **MVN**, en donde los resultados fueron los siguientes:

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	184.544953319842	1.65571079235445e-09	NO
Mardia Kurtosis	1.9860226693287	0.0470308068283103	NO
MVN	NA	NA	NO

Figura 2: Resultados de la prueba de normalidad de Mardia en R.

Como podemos observar en este caso, no pasan la prueba de normalidad de Mardia, esto en base a los resultados de la curtosis y el sesgo que presentan.

2. Prueba de Normalidad de Anderson Darling

Para el caso de la prueba de normalidad de Anderson Darling se utilizó la librería de R de **MVN**, con el comando de **univariateNormality**, en donde los resultados fueron los siguientes:

Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
Alcalinidad	3.6725	<0.001	NO
PH	0.3496	0.4611	YES
Calcio	3.9790	<0.001	NO
Clorofila	4.7492	<0.001	NO
Min_Conc	1.8380	1e-04	NO
Max_Conc	0.6585	0.081	YES
Est_Conc	0.8640	0.0248	NO

Figura 3: Resultados de la prueba de normalidad de Anderson Darling en R.

En este caso, los resultados nos dicen que nuestras conjeturas fueron correctas, que tanto el **PH** como el **Máximo de la concentración de mercurio** tienen un comportamiento normal, ya que pasaron el test, y las demás variables no tienen ese comportamiento.

Con estas variables, volvemos a crear un DataFrame con solamente esas variables para enfocarnos en las únicas que tienen un comportamiento normal. Con esto podemos graficar los respectivos plots para ver el comportamiento bivariado de ambas variables.

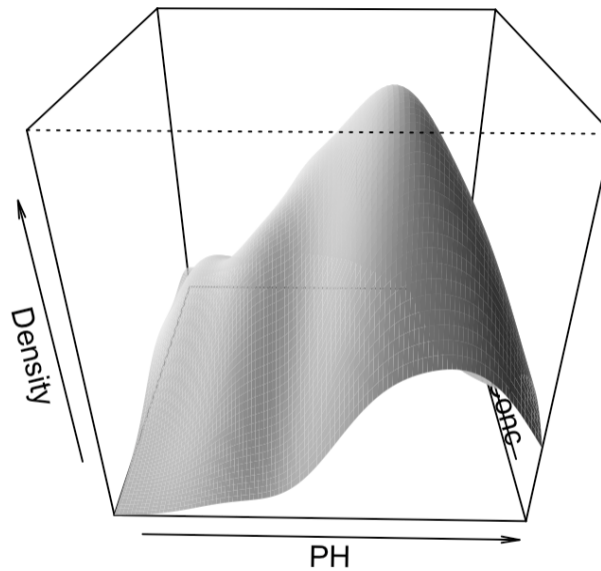


Figura 4: Perspective Plot entre el PH y la Máxima Concentración de Mercurio.

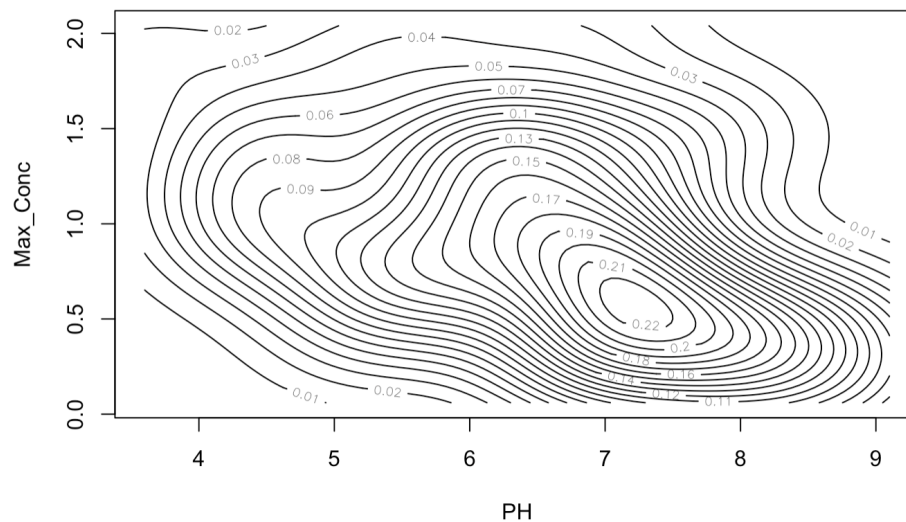


Figura 5: Contour Plot entre el PH y la Máxima Concentración de Mercurio.

Como podemos observar en el plot del contorno, se apreciaria que los datos están centrados en que entre mayor PH tenga el agua, la concentración máxima de mercurio va a ir disminuyendo, aunque su comportamiento no es del todo

homogéneo, como podemos observar. Después de esto, vamos a buscar los datos influyentes, por lo que vamos a utilizar un grafico QQplot multivariado, que en este caso sería bivariado y lo hacemos de la siguiente manera.

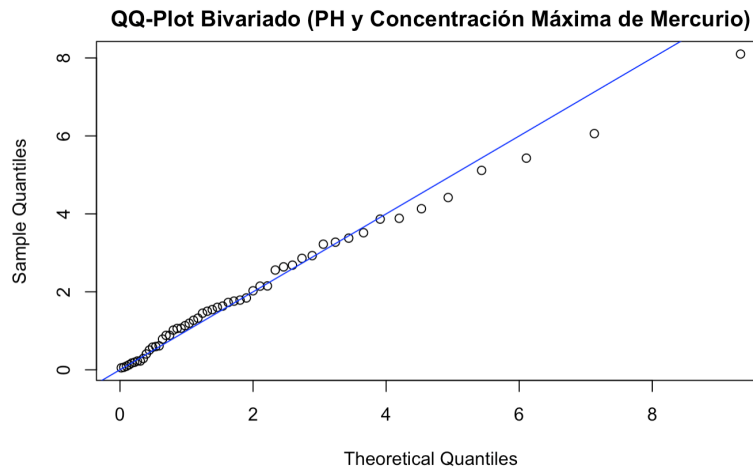


Figura 6: QQ-Plot Bivariado entre el PH y la Concentración Máxima de Mercurio.

Lo que nos dice el gráfico QQ plot bivariado, es que tiene un comportamiento con asimetría negativa es decir, que los datos están sesgados a la izquierda, por eso se comporta de la manera que sigue la tendencia normal, pero al final caen los datos. También podemos hacer uso de la misma librería para observar los datos atípicos, y podemos ver que solo 4 de esos datos son atípicos y que los demás se encuentran dentro de la distancia de Mahalanobis haciendo uso del **Chi-Square QQ-Plot**.

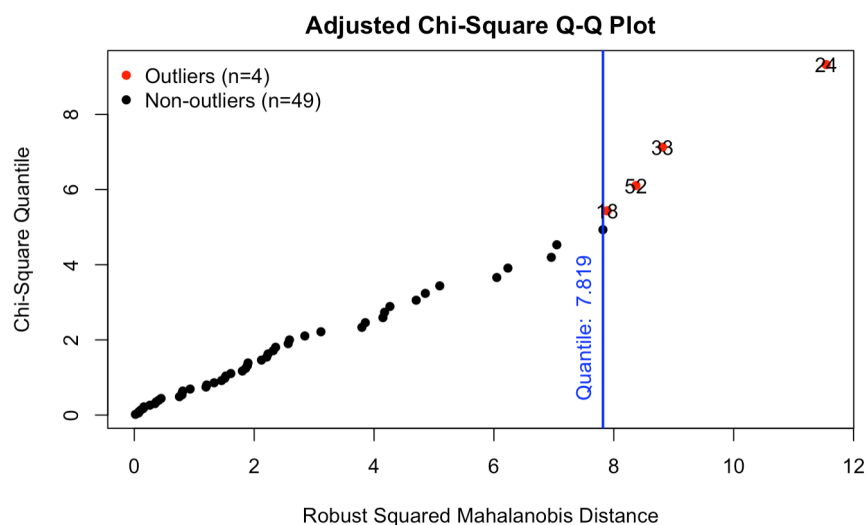


Figura 7: Adjusted Chi-Square QQ Plot.

Análisis de componentes principales (PCA)

Para el caso de PCA, lo primero que tenemos que hacer es sacar la matriz de correlación de nuestras variables.

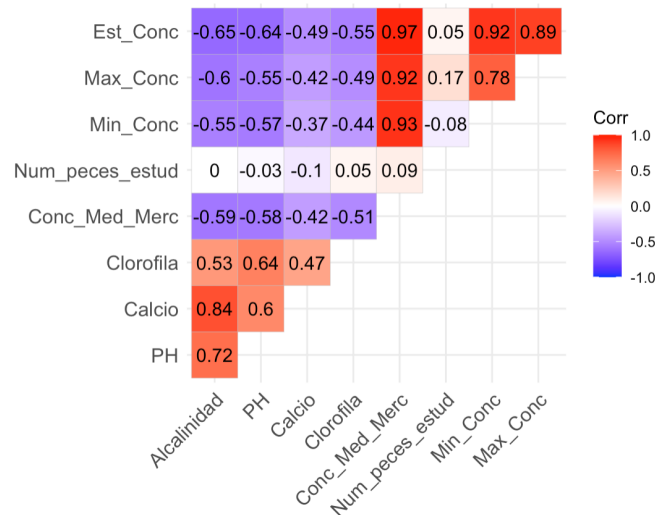


Figura 8: Plot de la matriz de correlación de las variables.

Ya desde aquí podemos hacer algunas inferencias en los datos y cuales son los componentes que más aportan y cuales no aportan, pero además de la pura matriz de correlación, vamos a hacer todo el análisis de componentes principales, lo que nos arrojó los siguientes resultados.

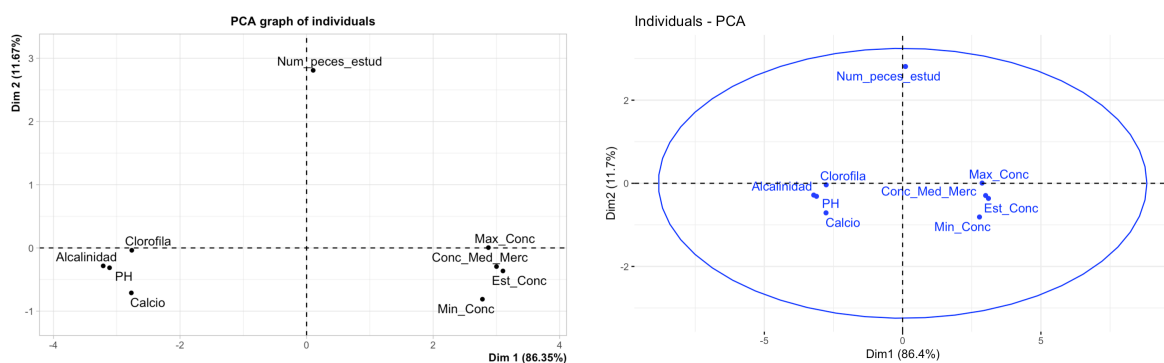


Figura 9: Gráfico de los componentes principales de las variables.

Como podemos ver en los diferentes plots pasados, lo que nos está representando es el comportamiento de las variables en base a la dimensión en la que se encuentran y su aportación a la misma.

En la dimensión 1, que es nuestro PCA 1, nos representa que ahí se encuentran la mayoría de los datos, y que las que aportan positivamente son las mismas variables de mercurio, y las que representan de manera negativa, son las demás, como la clorofila, el PH, el calcio y la alcalinidad. Esto nos quiere decir y comprobar lo que hemos estado analizando en todo este estudio, que los componentes del **PH, Calcio, Alcalinidad, y Clorofila** nos ayudan a disminuir la cantidad de Mercurio en los peces y en el agua de los lagos.

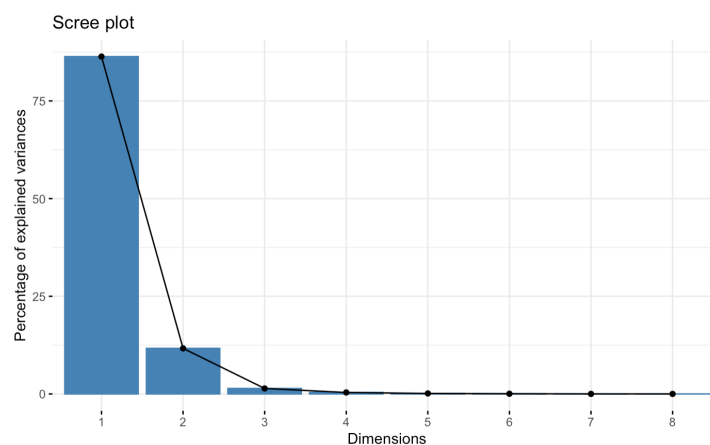


Figura 10: Scree Plot en la reducción de dimensiones del PCA.

En el caso del plot de codo, podemos ver como nuestro problema pasa de tener 8 dimensiones, a tener solo 1 dimensión, la cual tiene una combinación lineal de las distintas variables antes presentadas.

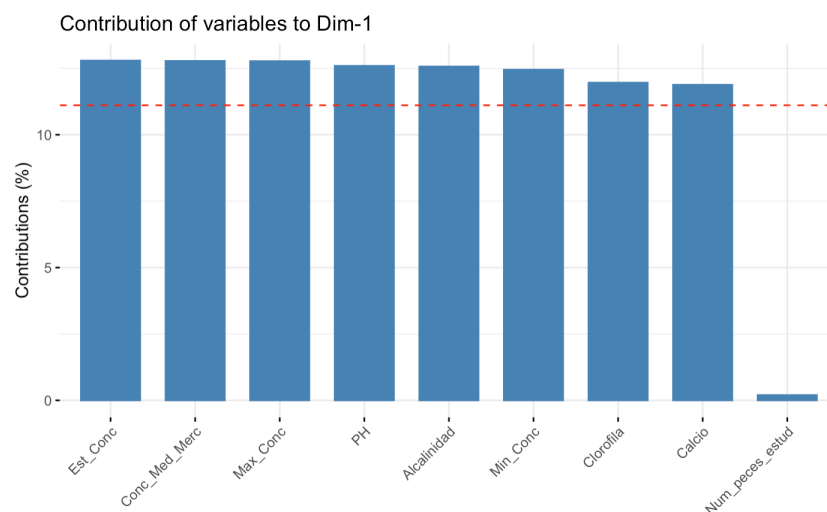


Figura 11: Contribuciones de las variables en la primera dimensión.

Como mencionamos anteriormente, nuestro problema se volvió de una sola dimensión, y podemos observar que la mayoría de las variables aportan muchísimo a ese nuevo componente principal, excepto la variable de número de peces estudiados, la cual podemos ver que es obsoleta y no aporta nada de información en nuestro componente.

	Dim.1	Dim.2
Alcalinidad	12.5702028	0.17197254
PH	12.5945715	0.10195279
Calcio	11.8816560	2.42957163
Clorofila	11.9611991	0.05379898
Conc_Med_Merc	12.7792349	0.55490106
Num_peces_estud	0.1975586	93.49955239
Min_Conc	12.4499927	2.66025508
Max_Conc	12.7721368	0.01992908
Est_Conc	12.7934475	0.50806646

Figura 12: Contribución de cada variable en las primeras 2 dimensiones del PCA.

Como podemos ver, cada variable nos da entre el 11% y 12% de la información a nuestro componente principal en la primera dimensión. Por otra parte, gracias a estos datos nos podemos dar cuenta que nuestro segundo componente, es decir nuestra dimensión 2, tiene el 93% de la información, por lo que acapara todo este componente.

Conclusión

Después de hacer todo el análisis estadístico en base a la concentración de mercurio en los peces, podemos observar que:

- El test de normalidad que más nos ayudó a obtener resultados del comportamiento normal de los datos, fue el de Anderson Darling ya que este tiende a ser más efectivo a la hora de detectar las desviaciones que se presentan en las colas de la distribución, además de que los test de normalidad se basan en la simetría y la curtosis para corroborar la misma.

- El Análisis de Componentes Principales nos ayudó para poder reducir la dimensión de nuestro problema, ya que al principio contábamos con 8 variables, que eso representa 8 diferentes dimensiones en las que las variables se pueden comportar, entonces, lo que hace el PCA, es hacer una combinación lineal de esas variables para poder reducir el tamaño de dimensiones y facilitar el procesamiento.
- En nuestro caso, podemos ver que los componentes del **PH, Calcio, Alcalinidad, y Clorofila** nos ayudan a disminuir la cantidad de Mercurio en los peces y en el agua de los lagos, mientras que el número de peces estudiados si afecta en el mismo análisis del estudio, pero en tan solo un 11.67% de la información.

Referencias

OCU. (2021, 6 abril). Mercurio en el pescado. www.ocu.org. Recuperado 14 de septiembre de 2022, de <https://www.ocu.org/alimentacion/alimentos/noticias/mercurio-en-pescado-un-problema-serio522454>

Anexos

Repositorio de Github

<https://github.com/francoquintanilla0/Merucrio-Peces-2>