Los peces y el mercurio: análisis estadístico

Inteligencia Artificial avanzada para la ciencia de datos - TC3006C

Grupo 101

Portafolio de Implementación - Módulo 1

Cristofer Becerra Sánchez - A01638659

***Resumen*—Debido a que los metales pesados como el mercurio son altamente tóxicos en los seres humanos, la presencia de estos en animales de ecosistemas acuáticos, en particular en peces y mariscos, representa una amenaza a la salud pública. El presente reporte analiza un conjunto de datos de 53 lagos en Florida. Se intenta determinar los diferentes factores que influyen en la concentración de mercurio en el tejido de los peces a partir de preguntas como: ¿la edad de los peces es un factor significativo en la concentración de mercurio? Se implementaron dos modelos estadísticos para la investigación, a saber, el Análisis de Varianza (ANOVA) de un solo factor, y la prueba de hipótesis. Se encontró que, en promedio, el nivel máximo de mercurio encontrado en los peces de los lagos de Florida excede los límites establecidos por la FAO. El otro hallazgo fue que la edad de los peces, categorizados como jóvenes y adultos, no es un factor relevante en la concentración de mercurio en los peces de los lagos.**

***Index Terms*—Mercurio, Metilmercurio, peces, lagos, ANOVA, prueba de hipótesis**

I. Introducción

La presencia de metales pesados en animales de ecosistemas acuáticos, en particular en peces y mariscos, se ha convertido en una cuestión importante en la actualidad. El tema cobra relevancia debido a que metales como el mercurio son altamente tóxicos en los seres humanos, por lo que altas concentraciones de estos metales en el cuerpo representan una amenaza contra la salud humana; esto lo convierte en materia de salud pública.

Se presenta un conjunto de datos de un estudio en 53 diferentes lagos en Florida (III) en el que se intenta determinar los diferentes factores que influyen en la concentración de mercurio en el tejido de los peces. Una de las variables observadas fue el nivel máximo de mercurio en cada grupo de peces del lago; otra variable que se cree a priori relevante es la edad de los peces, representada como un estado binario de joven (0) y adulto/maduro (1).

Dado que la norma internacional CAC/GL 7-1991 del Codex Alimentarius regida por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) determina que nivel máximo de mercurio en los peces es de 0.5 mg CH_3Hg/kg (con excepción de peces depredadores), es imperativo demostrar si las muestras de los lagos investigados cumplen con este estándar o no [1].

Así entonces, el presente trabajo pretende elucidar en las siguientes preguntas de investigación:

1. ¿La media del nivel máximo de mercurio en los peces de los lagos de Florida será mayor al nivel permitido de concentración de mercurio (0.5 mg Hg/kg) según regulación internacional?
2. ¿La edad de los peces es un factor significativo en la concentración de mercurio?

Se implementarán dos modelos estadísticos para contestar las preguntas de investigación, a saber, la prueba de hipótesis y el análisis de varianza (ANOVA). Los resultados se desglosan a detalle en la siguiente sección (II).

II. Resultados y análisis

II-A. Prueba de hipótesis

Para contestar la primera pregunta de investigación,

1. ¿La media del nivel máximo de mercurio en los peces de los lagos de Florida será mayor al nivel permitido de concentración de mercurio (0.5 mg Hg/kg) según regulación internacional?,

debe notarse que se trata de una inferencia del valor promedio de mercurio (máximo) de la población –en este caso, la población de los peces de los lagos de Florida–, por lo que inmediatamente se identifica el modelo estadístico para contestarla: la prueba de hipótesis. Este modelo se escoge debido a que, a partir de las muestras obtenidas, puede infersise una media poblacional que representa el la concentración máxima de mercurio promedio *real* encontrada en los peces de los lagos del estado de Florida; es decir, una generalización del nivel máximo, en promedio.

Una vez identificado el modelo estadístico a utilizar, se procede a plantear las hipótesis a probar, que toma la forma

$$H_0 : \ \mu_{\text{Hg-max}} = 0{,}5\text{mg Hg/kg} \tag{1}$$

$$H_a : \ \mu_{\text{Hg-max}} > 0{,}5\text{mg Hg/kg} \tag{2}$$

donde $\mu_{\text{Hg-max}}$ representa la media poblacional del nivel máximo de mercurio encontrado en los peces de los lagos. La hipótesis alternativa se escoge deliberadamente el caso de que la media poblacional sea mayor que el límite establecido por las regulaciones internacionales debido a que, si se debe rechazar la hipótesis nula, se tendría un severo problema que puede representar un riesgo a la salud pública. En otras palabras, es deseable que el nivel máximo encontrado por la prueba de hipótesis no supere o, en el peor de los casos, que sea igual que el máximo reglamentario.

Antes de proceder directamente con la prueba de hipótesis, se debe hacer una prueba de bondad de ajuste de los datos a una distribución normal; es decir, una prueba de normalidad. Esto con la intención de usar la distribución adecuada para la

Cristofer Becerra Sánchez, pertence al Tec de Monterrey Campus Monterrey, Monterrey, Nuevo León, Mexico, C.P. 64849

prueba de hipótesis. Como en cualquier prueba de bondad de ajuste, las hipótesis toman la forma

$$H_0 : \ f(x) = f_0(x) \tag{3}$$
$$H_a : \ f(x) \neq f_0(x) \tag{4}$$

donde $f(x)$ es la distribución de las medias de nivel máximo de mercurio, y $f_0(x)$ es una distribución normal.Se realiza una prueba de normalidad de Shapiro-Wilk con la biblioteca SciPy. Se tomará un nivel de significancia de $\alpha = 0{,}05$ para todas las pruebas de hipótesis del presente reporte. La prueba de Shapiro-Wilk retornó un p-value $p = 0{,}0467 < \alpha$ por lo que se rechaza la hipótesis nula. Por tanto, se concluye con una confianza del 95 % que la distribución de los datos no es una normal. Esto se comprueba gráficamente con un QQ-plot (ver figura REF).

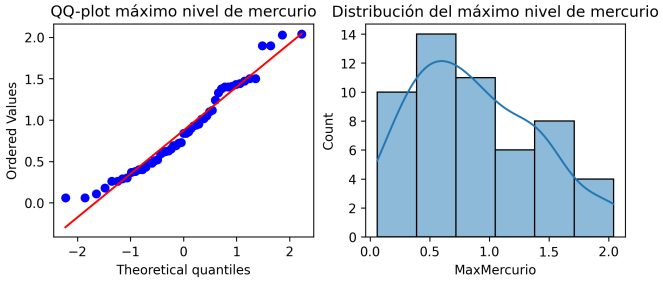

Figura 1. (a) QQ-plot de la muestra del nivel máximo de mercurio; el eje x representa los valores de los cuantiles teóricos que debería seguir una distribución normal, mientras que el eje y representa los cuantiles observados a partir de a muestra. (b) Histograma de la muestra del nivel máximo de mercurio; se grafica la curva de densidad sobre las barras que representan la frecuencia de los datos en el intervalo.

Puede observarse que la distribución no es una normal debido su sesgo, porque los datos tienen un ajuste relativamente razonable a los niveles teóricos; esto se debe a la naturaleza de la variable aleatoria, ya que en la práctica el nivel mínimo de mercurio no puede ser menor que cero, y los niveles no son muy altos como para desplazar la distribución lo suficiente como para que, en efecto, siga una normal.

Lo anterior cobra sentido al observar que el sesgo de los datos es de 0.4921 y tiene una curtosis de -0.5149; con esto se determina que la distribución tiene un sesgo a la derecha importante, además de ser una distribución platicúrtica. Por estas razones, además de que se tiene una muestra $n > 30$ con una desviación estándar poblacional desconocida, se optará por realizar la prueba de hipótesis con una distribución t de Student.

Se procede a computar la prueba de hipótesis con la distribución t de Student utilizando la función ttest_1samp() de SciPy. La media poblacional a probar es aquella de la regulación internacional (0.5 mg Hg/kg), y se especifica la hipótesis alternativa pertinente para que la función regrese el p-value correcto. Se obtuvo un estadístico t de prueba de $t_0 = 5{,}2229$, con su respectivo p-value de $p = 1{,}5714 \times 10^{-6}$; es notable que el p-value es considerablemente menor que el nivel de significancia por lo cual se rechaza la hipótesis nula. Se visualiza le resultado junto con la regla de decisión de la prueba utilizando el estadístico de prueba en la figura 2.

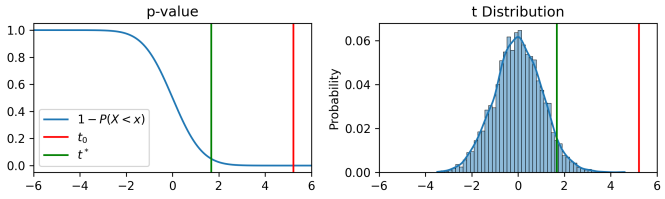

Figura 2. (a) P-value en función del estadístico de prueba t_0; la línea verde (izquierda) representa el valor t determinado por el nivel de significancia α, mientras que la línea roja (derecha) representa el valor del estadístico de prueba obtenido t_0. (b) Función de densidad de probabilidad de una distribución t de Student con los grados de libertad correspondiente al tamaño de la muestra del máximo de mercurio; de igual manera se grafica el valor que determina la región de rechazo y el estadístico de prueba obtenido.

Es evidente que el estadístico de prueba t_0 está dentro de la zona de rechazo. Por lo tanto, se concluye con un nivel de significancia del 95 % que la media del nivel máximo de mercurio de los lagos de Florida es mayor al nivel máximo de mercurio establecido por reglamentos internacionales. Finalmente se grafica el intervalo de confianza de este resultado en la figura 3

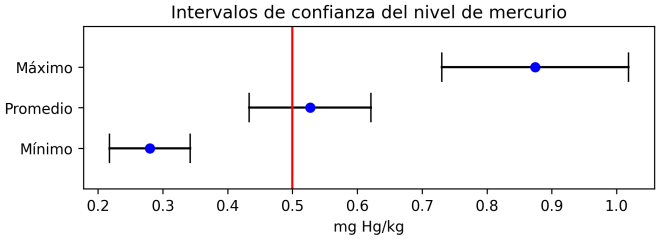

Figura 3. Intervalos de confianza de los diferentes vectores con valores de mercurio. *Mínimo* representa la muestra de los valores mínimos de mercurio encontrados en los peces de un lago, *Promedio* representa la muestra de los promedios de nivel de mercurio de cada lago, y *Máximo* es la muestra de los valores máximos de mercurio registrados para cada lago.

Se observa que el nivel máximo de mercurio reglamentario se encuentra fuera del intervalo de confianza del 95 % del máximo. Por esa razón puede decirse que la media poblacional es mayor que el máximo nivel permitido. Sin embargo, nótese que el máximo reglamentario se encuentra *dentro* del intervalo de confianza del promedio, por lo cual, ni siquiera puede asegurarse que el nivel medio de mercurio de los peces de los lagos estudiados sea menor a este valor.

II-B. ANOVA de la Edad

El segundo modelo estadístico a implementar en el análisis es el ANOVA; este nos permitirá identificar si una variable categórica, en este caso la edad del pez (joven o maduro) es un factor significativo en el nivel de mercurio. Además, esta herramienta nos ayudará a elucidar la pregunta de investigación más general sobre los factores principales que influyen en la concentración de mercurio en los peces. Se utilizará la función anova_stat() de BioInfoKit para realizar el ANOVA. El modelo toma la forma

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{5}$$

y, recordando la prueba de hipótesis del análisis de varianza,

$$H_0 : \ \mu_1 = \mu_2 \tag{6}$$
$$H_a : \ \mu_1 \neq \mu_2 \tag{7}$$

donde μ_1 es la media de las observaciones de mercurio promedio en los lagos con peces jóvenes, y μ_2 es la media de mercurio en la muestra de lagos con peces maduros. El análisis de varianza y todas las respectivas pruebas de hipótesis realizadas en el análisis se realizarán con un nivel de significancia de $\alpha = 0{,}05$. Los resultados del ANOVA se despliegan en la siguiente tabla.

Resultados del ANOVA de la edad de los peces					
	gl	SC	SCM	F	$P(F < f)$
Edad	1	0.0715	0.0715	0.6102	0.4383
Residuo	51	5.9764	0.1172	NA	NA

Se obtiene un estadístico F bajo, con un p-value mucho mayor que α; esto indica que no hay evidencia suficiente para afirmar con una significancia del 95 % que las medias son diferentes, por lo tanto, se concluye con el mismo nivel de significancia que la concentración media de mercurio en los peces jóvenes es igual a la concentración en peces maduros.

Además, puede realizarse una prueba de Tukey para identificar los intervalos de confianza de las respectivas medias analizadas. Los resultados de dicha prueba se presentan en la siguiente tabla.

Prueba de Tukey					
Grupo 1	Grupo 2	Diff	p-adj	Inferior	Superior
Joven	Adulto	0.0939	0.4383	-0.1474	0.3352

La gráfica de los intervalos de confianza se ilustra en la figura 4. Se nota claramente el traslape en los intervalos de confianza, por lo tanto, la edad no es un factor relevante en la concentración media de mercurio en los peces.

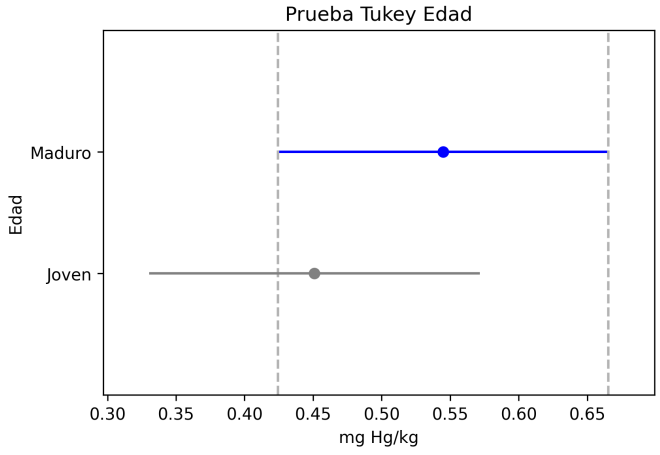

Figura 4. Intervalos de confianza de la media de ambos tratamientos, pez joven y pez adulto. El eje x no representa τ_i sino la concentración de mercurio, por lo cual la interpretación es directa.

Se nota que la media de los peces adultos (maduros) sí es mayor a la media de los peces jóvenes, ya que la primera tiene un valor cercano a 0.55 mg Hg/kg y la segunda se acerca más a los 0.45 mg Hg/kg. No obstante, el intervalo de confianza de ambas medias es extenso, tal que existe un traslape entre ellos, lo que permite concluir que no hay diferencia *significatiava* entre los tratamientos.

II-C. Verificación del ANOVA

Para que las conclusiones anteriores resulten verdaderas, se debe verificar que los supuestos del modelo. Se probarán las siguientes tres suposiciones del modelo:

1. Los residuos siguen una distribución normal
2. Las muestras provienen de distribuciones con varianzas idénticas
3. Los residuos del ANOVA son independientes

Así, pues, se procedió a realizar la comprobación de las suposiciones enunciadas con anterioridad.

II-C1. Normalidad de los residuos: Se comienza con una prueba de Shapiro-Wilk de normalidad para comprobar que los residuos del ANOVA efectuado siguen una distribución normal. Se obtuvo un estadístico de prueba $W = 0{,}9322$, que entrega un p-value de $p = 0{,}0049$. La prueba resulta en un p-value que obliga a rechazar la hipótesis nula que establece que los residuos del ANOVA, en efecto, siguen una distribución normal. Se comprueba esta aseveración mediante el QQ-plot de los residuos (ver figura 5).

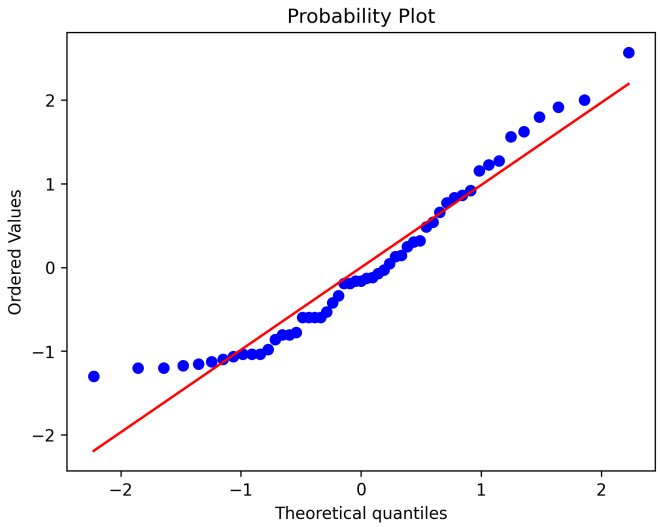

Figura 5. QQ-plot de los residuos del ANOVA de la edad de los peces. Es notable el pobre ajuste a una línea recta que presenta dicho conjunto de datos. En particular se distingue la cola izquierda, aunque la cola derecha también presenta una desviación considerable. Además, la parte central de la distribución toma la forma de una w, no de una línea recta.

Debido a que el modelo no cumple con los postulados del ANOVA, se concluye que ese análisis es inválido. Por esta razón es necesario tomar medidas para obtener conclusiones válidas a partir del modelo elegido.

II-D. ANOVA de la edad con transformación de Box-Cox

Un remedio para la normalidad de los residuos es realizar una transformación de Box-Cox,

$$f(x,\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases},$$

sobre la variable de respuesta –el nivel medio de mercurio de los peces en el lago respectivo–. Esto se debe a que la normalidad de los residuos depende de la normalidad de las muestras con los tratamientos a contrastar. Se encontró que el parámetro λ que maximiza la función de verosimilitud en este caso es de $\lambda = 0{,}4504$. Entonces, se vuelve a realizar el ANOVA con la variable dependiente transformada. Los resultados de este nuevo análisis se despliegan en la siguiente tabla.

ANOVA de la edad de los peces					
	gl	SC	SCM	F	$P(F < f)$
Edad	1	0.4307	0.4307	1.6654	0.2027
Residuo	51	13.1882	0.2586	NA	NA

Es posible observar que el p-value es menos de la mitad que la ocasión anterior, no obstante, aún así se tiene un valor mayor que α por lo que la conclusión anterior prevalece. Se realiza la misma prueba de Tukey, presentada en la tabla a continuación

Prueba de Tukey con Box-Cox					
Grupo 1	Grupo 2	Diff	p-adj	Inferior	Superior
Joven	Adulto	0.2304	0.2027	-0.128	0.5888

y se grafican los intervalos de confianza en la figura 6.

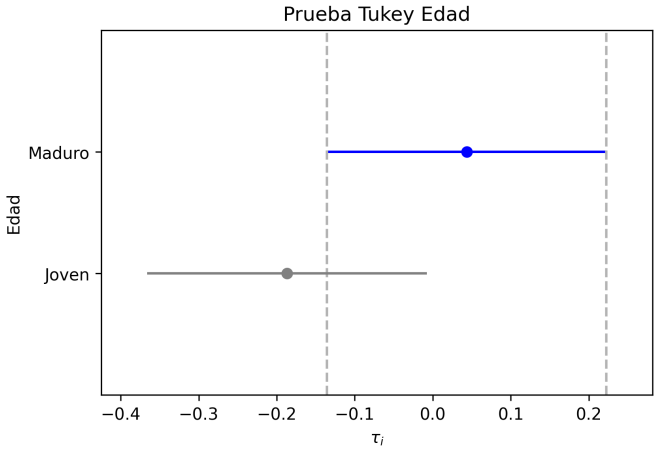

Figura 6. Intervalos de confianza de la media de ambos tratamientos, pez joven y pez adulto. En este caso sí se maneja τ_i en el eje x debido a la transformación de Box-Cox. Es posible llegar a la misma conclusión que la ocasión anterior: los intervalos se sobreponen por lo cual no hay diferencia significativa entre tratamientos.

Puede verse que la transformación afectó a los intervalos de manera importante, pues el intervalo de confianza de los peces maduros ya no incluye la media de los peces jóvenes. Sin embargo, los intervalos de confianza siguen traslapados por lo cual, junto con el estadístico F y su respectivo p-value, puede decirse con un nivel de significancia del 95 % que la edad de los peces no es un factor importante en la concentración de mercurio en los peces.

II-E. Verificación del ANOVA con Box-Cox

Una vez más, es necesario comprobar las suposiciones fundamentales del modelo. Comenzando por la prueba de normalidad, se obtuvieron los siguientes resultados: un estadístico $W = 0{,}9762$, con un p-value correspondiente de $p = 0{,}3668$, que en esta ocasión es mucho mayor que el nivel de significancia α. Por esa razón sí se cumple la condición de normalidad de los residuos con la transformación efecutada. La gráfica QQ-plot de la fiugra 7 soporta la decisión de no rechazar la hipótesis nula.

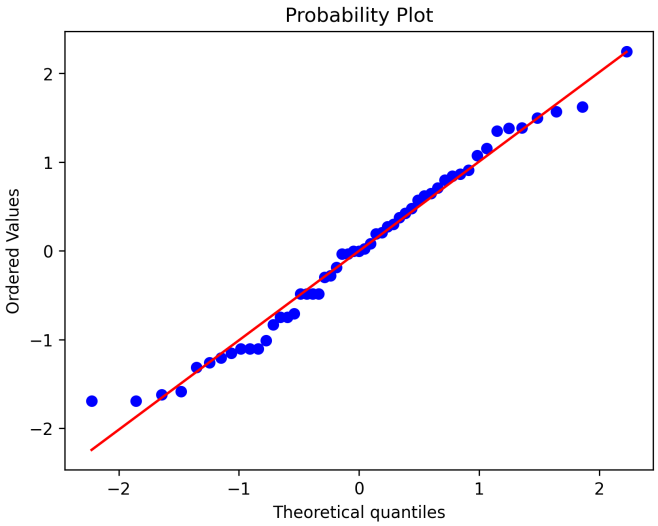

Figura 7. QQ-plot de los residuos del ANOVA de la edad de los peces tras la transformación de Box-Cox.

Tras la transformación se nota que existe un ajuste mucho más razonable a la línea recta, a pesar de que todavía presenta un ajuste pobre cerca de las colas.

Se procede a demostrar la suposición de que las muestras provienen de distribuciones con varianzas idénticas. Para esto se realiza una prueba de Levene; dicha prueba indicará si las varianzas de ambos tratamientos provienen de una varianza poblacional idéntica; es decir, las varianzas de las muestras son iguales de forma significativa.

La prueba de Levene realizada retorna un estadístico $W = 1{,}5723$, que resulta en un p-value de $p = 0{,}2156$. Sucede entonces que el p-value es mayor que el nivel de significancia α por cual que no se rechaza la hipótesis nula. Esto implica que las varianzas de las muestras son iguales con un nivel de significancia del 95 %, comprobando así la condición de homocedasticidad. Adicionalmente grafican los residuos contra los valores ajustados del ANOVA en la figura 8.

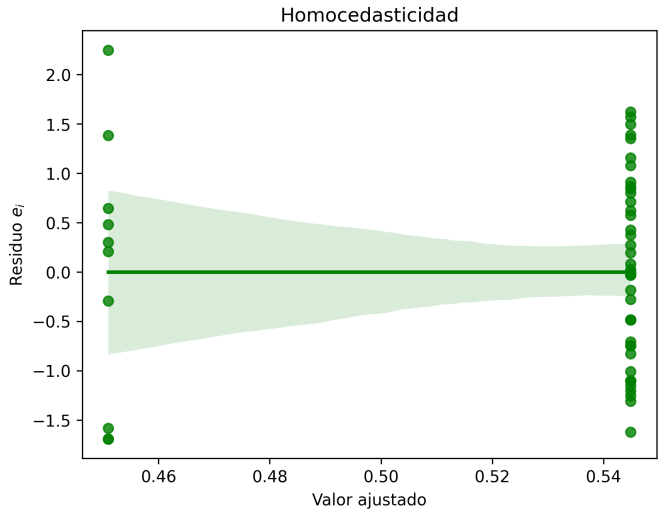

Figura 8. Gráfica de la varianza de los residuos; se grafica el valor ajustado de cada predicción contra el residuo correspondiente.

Resaltando que la visualización de regresión de Seaborn indica una línea constante en y además de que la dispersión de ambos conjuntos de datos es similar, se confirma gráficamente la prueba de Levene.

Se procede a compobar la condición de independencia de los residuos mediante una gráfica de los valores contra el número de residuo; es decir, el eje y comprende los residuos e_n, mientras que el eje x comprende el subconjunto de números naturales $n = \{1, 2, 3, ..., N\}$ donde N es el tamaño del vector de los residuos. Tal visualización se aprecia en la figura 9.

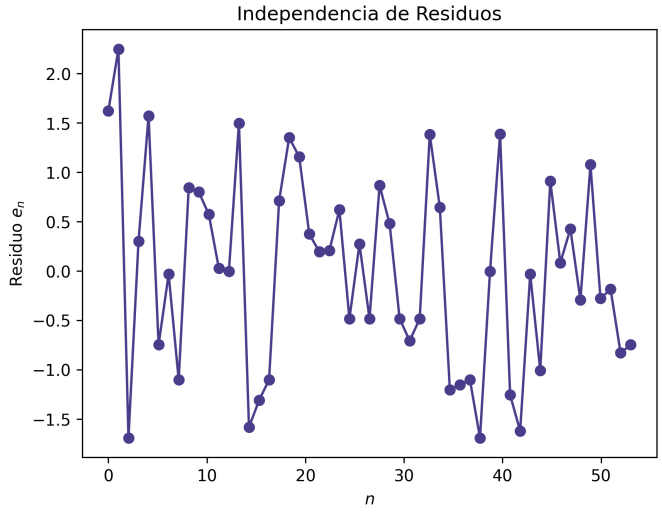

Figura 9. Gráfica del valor de los residuos con respecto al número de observación (elemento del conjunto de los residuos).

Es evidente que los residuos no presentan tendencias ya que oscilan de manera aleatoria entre valores positivos y negativos, de tal suerte que se comprueba gráficamente la condición de independencia.

Habiendo comprobado todas las condiciones del ANOVA, se puede por fin afirmar la conclusión con certeza: la concentración media de mercurio en los peces jóvenes es igual a la concentración en peces maduros con un nivel de significancia del 95 %.

III. Conclusión

El análisis anterior permite realizar varias conclusiones. Primero que nada, la prueba de hipótesis realizada con una distribución t de Student permitió inferir que, en promedio, el nivel máximo de mercurio encontrado en los peces de los lagos de Florida excede los límites establecidos por la FAO. Esto es una situación preocupante ya que no es posible garantizar que la pesca de estos lagos sea segura para su consumo; es decir, el consumir peces de estos lagos potencialmente representa un riesgo a la salud humana.

Encima de esta situación, al realizar los intervalos de confianza de las diferentes mediciones (mínimo, medio y máximo de la figura 3) de concentración de mercurio se observa que ni siquiera puede garantizarse que el nivel **medio** de mercurio en los peces sea menor al límite reglamentario; es posible realizar esta conclusión ya que el valor regulado se encuentra dentro del intervalo de confianza de los valores promedio de los lagos. Esto resulta aún más preocupante ya que, en promedio, es probable que uno esté consumiendo pescados con una concentración de metilmercurio justo al límite de los posibles efectos adversos a la salud.

En tanto a la pregunta sobre los factores que afectan estos niveles, puede afirmarse con una significancia del 95 % que la edad de los peces, entre jóvenes y adultos, no es un factor relevante. Es posible hacer la afirmación anterior a partir de los resultados obtenidos del ANOVA ya que se trata de la edad, una variable categórica, que puede interpretarse como un tratamiento del análisis de varianza.

Anexo

Enlace al repositorio de GitHub, *Mercurio*, con todos los archivos del proyecto (Jupyter Notebook, Notebook en formato .py, la base de datos utilizada, y el presente documento PDF): https://github.com/crisb-7/Mercurio

Referencias

[1] Food and Agriculture Organization of the United Nations. 2022. Report of the twenty-second session of the Codex Committee on fish and fishery products. [online] Recuperado de: https://www.fao.org/3/w1607e/w1607e00.htm el 16 de Septiembre del 2022.