UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MAESTRÍA EN ANÁLISIS Y VISUALIZACIÓN DE DATOS

Análisis e Interpretación de Datos

Actividad grupal. Definición de un problema estadístico: modelización y propuesta de soluciones

Presenta

Abraham Sandoval Altamirano

Daniel Isaí Yáñez Torres

Guillermo Gómez Sánchez

Pablo de Jesús Ramírez Hernández

*Todos participaron activamente en cada uno de los aspectos de la realización del trabajo.

Nombre del profesor

RAUL VALENTE RAMIREZ VELARDE

Ciudad de México, agosto del 2023

Estudio con modelos estadísticos sobre la relación entre el salario de trabajadores del área de ciencia de datos y el tamaño de la empresa donde laboran entre 2020 y 2023.

Abraham Sandoval Altamirano, Daniel Isaí Yáñez Torres, Guillermo Gómez Sánchez, Pablo de Jesús Ramírez Hernández.

Resumen

Haciendo uso de una base de datos actual se estudia la relación que existe entre el salario percibido por los profesionistas dedicados al área ciencia de datos y el tamaño de la empresa en la que ejercen, se hace uso de varios modelos estadísticos con la finalidad de determinar dicha conexión entre estos conceptos, así como entender las dependencias existentes entre ellos.

Palabras clave: Ciencia de datos, científico de datos, salario, empresa, modelo estadístico.

Summary

Using a current database, the relationship between the salary received by professionals dedicated to the data science area and the size of the company in which they work is studied, using various statistical models in order to determine said connection between these concepts, as well as understanding the dependencies between them.

Keywords: Data science, data scientist, salary, company, statistical model.

INTRODUCCIÓN

El objetivo del presente trabajo es reconocer, entender y establecer la relación entre el salario de las personas del área de ciencia de datos y el tamaño de la empresa donde se desempeñan como profesionistas.

En la actualidad, la ciencia de datos es importante porque permite tomar mejores decisiones, mejorar la eficiencia y crear nuevos productos y servicios. Se utiliza en una amplia gama de industrias, incluyendo la salud, la banca, la logística y el marketing. Por tal motivo el número de personas dedicadas a esta área se encuentra en constante crecimiento.

Para el estudio presente se usarán dos modelos estadísticos, entre ellos Kruskal-Wallis, al descartar el uso de ANOVA, el cual es una prueba no paramétrica que se utiliza para comparar tres o más grupos de datos. La prueba asume que los datos son independientes y que se distribuyen de forma similar en todos los grupos. Este modelo se utiliza cuando la distribución de los datos no es normal o cuando el tamaño de las muestras es pequeño.

En este proceso se calcula utilizando un estadístico llamado H. El estadístico H es una medida de la diferencia entre las medias de los grupos. Si el valor de H es significativo, entonces se puede concluir que al menos dos de los grupos son diferentes.

La prueba se interpreta utilizando una tabla de valores p. El valor p es la probabilidad de obtener un valor de H al menos tan grande como el

observado, si los grupos fueran iguales. Si el valor p es menor que un nivel de significación preestablecido, entonces se puede concluir que los grupos son diferentes.

Otro modelo utilizado en este trabajo es el contraste de hipótesis, es una técnica utilizada para tomar decisiones sobre afirmaciones o suposiciones acerca de una población, basándose en información de una muestra de datos. Consiste en formular dos hipótesis, la hipótesis nula (H0), que representa la afirmación o suposición inicial, y la hipótesis alternativa (H1), que representa la suposición contraria.

Se recopila información de la muestra y se utiliza para calcular un estadístico de prueba que medirá la discrepancia entre los datos observados y lo que se esperaría bajo la hipótesis nula. Con esta información, se calcula el valor p, que representa la probabilidad de obtener resultados tan extremos o más extremos que los observados, asumiendo que la hipótesis nula es verdadera.

Finalmente, se compara el valor p con un umbral predefinido llamado nivel de significancia (generalmente 0.05). Si este valor es mayor que el valor p, entonces existe suficiente evidencia estadística para concluir que se rechaza la hipótesis nula.

Para saber si una distribución de datos es normal, se puede utilizar la herramienta de Anderson-Darling, se basa en la comparación de los valores observados con los valores esperados según la distribución teórica.

Esta prueba busca evaluar si los datos se ajustan a una distribución específica mediante la comparación de estadísticos de ajuste, si el valor resultante de la prueba es menor que un umbral predefinido (nivel de significancia), se puede concluir que los datos provienen de la distribución especificada. En caso contrario, se rechaza esa hipótesis y se considera que los datos no siguen esa distribución.

METODOLOGIA Y RESULTADOS

Para ejecutar un ANOVA se debe tener una variable de respuesta continua y al menos un factor categórico con dos o más niveles. Los análisis ANOVA requieren datos de poblaciones que sigan una distribución aproximadamente normal con varianzas iguales entre los niveles de factores.

El ANOVA es la mejor técnica para utilizarse siempre que se cumplan sus condiciones, se procede a realizar un estudio sobre el comportamiento de los residuos. Para ello se usa la prueba de normalidad Lilliefors (Kolmogorov-Smirnov). Se usa un $\alpha = 0.05$ y se obtiene que el p-valor es 2.26*10-16, por lo tanto, se puede afirmar que los residuos no tienen distribución normal, como no se cumple esta condición se procede a hacer el análisis por medio del estudio Kruskal-Wallis.

Primero se cargan los datos de los diferentes archivos en diferentes variables. A continuación, se procede a combinar los datos en un solo dataframe y se crea el vector de factores, así como la variable relacionada al salario en dólares.

Al utilizar la prueba Kruskal-Wallis detecta que al menos una mediana no es igual a las otras.

```
> x = dunn.test(SalaryInUSD,CompanySize, method="bonferroni", list=T, alpha=0.05, altp=T)
  Kruskal-Wallis rank sum test

data: SalaryInUSD and CompanySize
Kruskal-Wallis chi-squared = 668.1047, df = 2, p-value = 0
```

Luego de eso se usa la prueba de Dunnet donde se compara un tratamiento control con todos los demás para ver si son iguales o no.

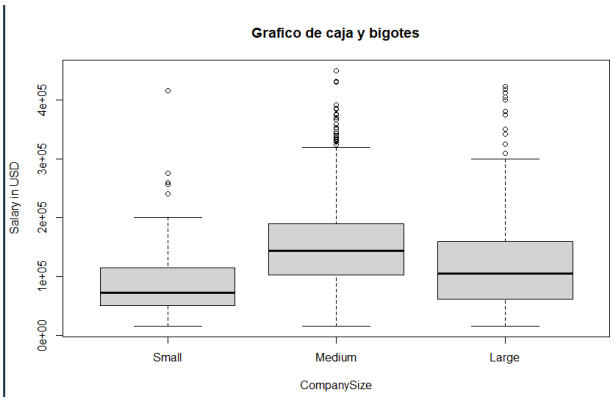
Al utilizar la prueba de Dunne con un alfa de 0.05 obtenemos que los p-valores son muy pequeños para verlos en el resumen de la prueba de Dunne y se obtuvieron mediante variables.

```
> x = dunn.test(SalaryInUSD,CompanySize, method="bonferroni", list=T, alpha=0.05, altp=T)
  Kruskal-Wallis rank sum test

data: SalaryInUSD and CompanySize
Kruskal-Wallis chi-squared = 668.1047, df = 2, p-value = 0


                  Comparison of SalaryInUSD by CompanySize
                                (Bonferroni)
Col Mean-|
Row Mean |      Large     Medium
---------+----------------------
  Medium |  -17.59854
         |    0.0000*
         |
   Small |   8.558608   20.34023
         |    0.0000*    0.0000*


List of pairwise comparisons: Z statistic (adjusted p-value)
-----------------------------------
Large - Medium : -17.59854 (0.0000)*
Large - Small  :  8.558608 (0.0000)*
Medium - Small :  20.34023 (0.0000)*
```

Los p valores para cada comparación son:

Comparaciones	p-value
Medium – Large	7.58e^-69
Small – Large	3.4e^-17
Small – Medium	1.7 e^-91

Como para cada comparación el p-valor es menor a 0.05 (la alfa establecida para el problema) se puede decir que el promedio del salario que recibe un trabajador en ciencia de datos es diferente para cada tamaño de empresa. Se puede utilizar el siguiente grafico de caja y bigotes para hacer un análisis gráfico.



Para el segundo modelo estadístico se realizó la prueba de Anderson-Darling para saber si el comportamiento de los datos se atribuye a una distribución normal o no.

Para ello definimos la hipótesis como sigue:

•H_0: Los datos siguen una distribución normal

•H_1: Los datos no siguen una distribución normal.

Los datos obtenidos se muestran en la tabla 01.

Tabla 01.

Tamaño Empresa	Valor A	Valor p
Small	22.634	2.2e^-16
Medium	32.596	2.2e^-16
Large	16.024	2.2e^-16

Se considera una significancia de 0.05 para mantener baja la probabilidad de cometer un error tipo I. Con esta consideración podemos notar que la distribución de datos no tiene un comportamiento normal.

Luego, para determinar que los promedio por pares de tamaño de empresa son diferentes, se procede a utilizar la prueba Z. (Ya que se tienen más de 30 muestras por categoría).

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}.$$

Para ello consideramos las siguientes hipótesis.

• El promedio de salario de la población 1 es igual al promedio de salario de los científicos de datos en la población 2

• El promedio de salario de la población 1, es diferente al promedio de salario de los científicos de datos en la población 2.

Entendiendo población 1 y población 2, como cada uno de los pares de compañías (small-medium, small-large y medium-large). De esta prueba se obtuvo que.

Pares Población	p-value
Small – Medium	2.2e^-100
Small – Large	1.3e^-18
Medium – Large	1.4 e^-46

CONCLUSIONES

Con los resultados obtenidos, se puede concluir, en primera instancia, que el comportamiento de la base de datos no es normal, por tal motivo se decidió utilizar los métodos de Kruskal-Wallis y contraste de hipótesis por medio del uso de una prueba Z.

El uso de estos métodos estadísticos nos permitió llegar a las mismas conclusiones en cada caso de modelo utilizado. Lo segundo que se puede concluir es que el promedio de los salarios es diferente para cada tamaño de empresa, y como se puede observar en la gráfica de bigotes, sorpresivamente resulta interesante observar que el promedio de salario es mayor para las empresas de tamaño mediano, superando al promedio para las empresas grandes. Así como que los salarios más competitivos se encuentran dentro de las empresas de tamaño mediano y grande.

De igual forma se pudo experimentar el uso y funcionalidad de los dos modelos estadísticos planteados, y observar su fiabilidad al arrojar los resultados que permitieron llegar a las mismas conclusiones.

BIBLIOGRAFIA

1.3.5.14. Anderson-Darling Test. (s/f). Nist.gov. Recuperado el 6 de agosto de 2023, de https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

Anderson-Darling Test in R (Quick Normality Check). (2021, noviembre 9). R-Bloggers. https://www.r-bloggers.com/2021/11/anderson-darling-test-in-r-quick-normality-check/

Combining vectors. (s/f). R-tutor.com. Recuperado el 6 de agosto de 2023, de http://www.r-tutor.com/r-introduction/vector/combining-vectors

Johnson, R., & Kuby, P. (2012). Estadistica Elemental. Cengage Learning Editores.

Mean by factor by level. (s/f). Stack Overflow. Recuperado el 6 de agosto de 2023, de https://stackoverflow.com/questions/23395366/mean-by-factor-by-level

Roberts, M. J., & Russo, R. (2014). A student's guide to analysis of variance. Routledge.

RPubs - 19.2. Pruebas de hipótesis en R: Una media. (s/f). Rpubs.com. Recuperado el 6 de agosto de 2023, de https://rpubs.com/hllinas/R_Test_1Media

Sort matrix according to first column in R. (s/f). Stack Overflow. Recuperado el 6 de agosto de 2023, de https://stackoverflow.com/questions/14359726/sort-matrix-according-to-first-column-in-r

Extracción bases: https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023?resource=download