Pontificia Universidad Católica de Valparaíso

Escuela de Ingeniería Industrial

***Data Science* Aplicada**

**Entregable 4**

***Modelo/metodología de data Science.***

**Integrantes Equipo**

Andrés Arenas Rodríguez. Mauricio Huerta Aguiar

Indy Navarro Vidal. Pablo Zúñiga Carvajal.

Juan Saavedra Jeria Víctor Leiva Sanchez

**Definición y supuesto de la metodología**

En la entrega anterior, gracias al análisis exploratorio de datos desarrollado, se tuvo un primer acercamiento al objetivo que se está estudiando, planteado en la pregunta de investigación:

*¿Es Santiago representativo de las tendencias políticas de todo el territorio nacional?*

Gracias a ello, fue posible obtener información preliminar acerca del comportamiento de los votos en las distintas regiones del territorio nacional, encontrando regiones que se asemejan al comportamiento de la región Metropolitana, como lo fue la región de Valparaíso. Cabe mencionar que todo este análisis se realizó por medio de diversas tablas de datos junto a diversos gráficos generados por Excel. Es por este motivo que se requiere la utilización de alguna metodología que permita cuantificar la semejanza entre las regiones.

Para poder relacionar el comportamiento de las regiones, se decide utilizar el estadístico chi-cuadrado ya que, a través de un test de hipótesis, permite comparar muestras de población, y entregar un valor numérico que permite cuantificar la semejanza entre las poblaciones.

El análisis se realiza comparando el comportamiento entre la región Metropolitana, con las distintas regiones, respecto a las distintas elecciones en los años 2013 -2017.

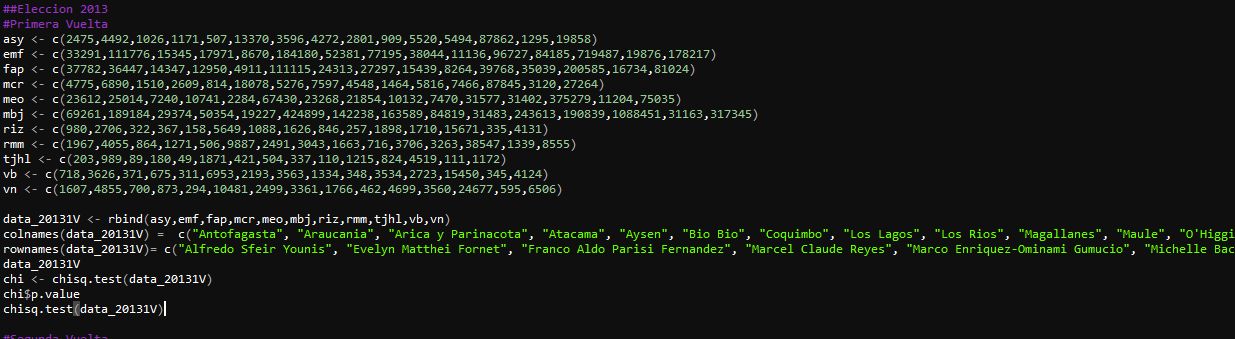
El primer paso para poder aplicar el estadístico chi-cuadrado es plantear la hipótesis, la cual se busca comprobar, expresado de la siguiente forma:

**Hipótesis Nula:** Existe homogeneidad en la proporción de votos para cada región/ Para cada voto la proporción entre las regiones son iguales.

**Hipótesis Alternativa:** No existe homogeneidad en la proporción de votos para cada región. / Para al menos un voto la proporción en cada región no es igual

**Análisis de datos**

En un principio al comenzar con el análisis de datos, se consideró hacer uso del estadístico chi-cuadrado. Para comenzar el análisis estadístico por medio del software R se procedió a generar un primer tratamiento y adaptación del *dataset,* para ello se separaron los resultados por cada proceso de elección, y a su vez, estos fueron separados por su vuelta correspondiente, lo anterior tiene lugar, dado que una primera vuelta posee una mayor cantidad de candidatos que para el caso de la segunda vuelta, a continuación se presenta una imagen del tratamiento de datos realizados, específicamente para la elección del año 2013 correspondiente a su primera vuelta:



**Figura 2.1:** Generación de archivo para test chi-cuadrado

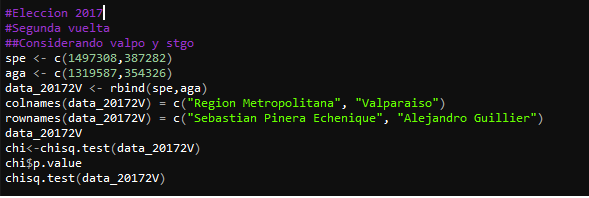
Una vez que los datos fueron cargados se realizó el test chi-cuadrado para cada uno de los procesos de elecciones, y para cada una de las vueltas correspondientes a cada proceso eleccionario. Para lo anterior, el objeto generado mediante el software R correspondió a una matriz, que clasifica a cada uno de los candidatos correspondiente a la elección y vuelta que participa, además del total de votos que obtuvo el candidato clasificado por cada región. Luego de lo anterior, y con las matrices construidas a cada una de estas se les realizó el test chi cuadrado para cada uno de los casos.

**Interpretación de resultados**

Como se mencionó en el apartado anterior, luego de la construcción de las matrices de las elecciones y vueltas correspondientes, y realizar el test de chi-cuadrado se procedió a comprar el valor p (*pvalue*) obtenido para cada proceso realizado, correspondiente a los procesos de elección del año 2013 y 2017 además de diferenciar por primera y segunda vuelta.

En primera instancia realizado el test para la primera vuelta correspondiente tanto a las elecciones del año 2013 como el año 2017, los valores obtenidos para el estimador del *pvalue* en ambos casos tomó un valor extremadamente pequeño y cercano a cero (*pvalue < 2.2e-16*) para ambos casos, algo similar ocurrió para ambos procesos en el caso de la segunda vuelta obteniendo un *pvalue* de igual forma cercano a cero.

Si bien, lo anterior en primera instancia y de manera directa refuta la hipótesis nula definida para el test planteado anteriormente, se realizó una segunda comprobación, con la finalidad de comprobar el objetivo anterior, la cual en esta oportunidad correspondió en realizar una comparación entre la *región metropolitana* y la *región de Valparaiso* considerando para este caso, solamente los resultados obtenidos en la segunda vuelta, esto a razón, de simplificar el análisis para comprobar los valores obtenidos. A continuación, se presenta la realización de este análisis correspondiente utilizando la segunda vuelta del año 2017:



**Figura 3.1:** Generación de archivo para test chi-cuadrado

En función de lo mencionado anteriormente, se comprobó nuevamente que los valores obtenidos en este caso eran ínfimos y a pesar de no ser *cero* su valor es despreciable (1,75e-46), por ende, se puede comprobar que la hipótesis nula planteada al inicio de este informe se refuta.

**Propuestas para trabajos futuros**

Tras el trabajo y los resultados arrojados, se ha observado que hay una clara refutación de la hipótesis nula en todos los casos, en favor de sus respectivas hipótesis alternativa, esto claramente demuestra de que hay pruebas suficientes para pensar en que no existe homogeneidad en las proporciones en cada región con respecto a sus votaciones.

Estas pruebas iniciales podrían llegar a responder en contra de la idea principal de que Santiago pareciera ser representativo de las votaciones de cada región, pero como se ha mencionado anteriormente, este ha sido el inicio del trabajo, donde se ha aplicado este test ampliamente usado en estadística para pruebas de homogeneidad e independencia, y considerando una sola variable la cual corresponde a las votaciones, en esta misma línea, se podrá considerar la utilización de otras pruebas de hipótesis, y el uso de otros test. Luego, existe una idea de incluir otras variables cuantitativas que permitan enriquecer dicho estudio, como, por ejemplo, cantidad de habitantes por región/comuna respectiva obtenidos de una fuente oficial como, por ejemplo, INE o SERVEL.

Además, se está considerando evaluar los aspectos más relevantes en las tendencias políticas de Santiago utilizando CHAID, y el método de k-means para realizar clusterización del conjunto de observaciones, por lo que podría permitir realizar una comparación de las regiones desde otra perspectiva, diferentes a las planteadas anteriormente.

**Bibliografía**