Trabajo Práctico Nº 1

Análisis Exploratorio de Datos - Maestría en Estadística Aplicada

Año 2020

Problema 1

Datos

En el año 2015 se llevó a cabo una investigación a campo en la ciudad de Rosario para evaluar diferentes aspectos relacionados con la salud infantil en menores de 5 años. Durante el relevamiento se encuestaron a las madres que asistieron con sus niños a algunos centros de salud de la ciudad. Cada madre respondía cuestionarios acerca de: aspectos de salud del niño, de condiciones de las viviendas que habitaban y de la situación socio-económica del hogar del niño. Se realizaron además, mediciones antropométricas y de laboratorio sobre los niños, bajo consentimiento informado de las madres.

El objetivo primario de dicha encuesta era obtener información que permita mejorar la situación sanitaria asociada a esa población vulnerable. Además, se habían planteado además una multiplicidad de objetivos secundarios, entre ellos, estudiar la situación de anemia por deficiencia de hierro en los niños, lo cual podría estar relacionado con una mala alimentación, entre otros motivos. Para ello se registró la hemoglobina (Hb) en gr/dl de cada niño. En particular, interesa detectar si hay diferencias en el nivel de Hb entre los efectores o centros de salud, lo cual permitiría identificar poblaciones de niños más susceptibles a la enfermedad de la anemia.

Los datos se encuentran en el archivo `anemia.csv`.

Consigna

Aplicar las herramientas de análisis exploratorio univariado discutidas en el material de estudio para dar respuesta a los objetivos mencionados, estudiando la distribución de la variable hemoglobina en sangre en función de la suplementación por hierro y del centro de salud. Para finalizar con el análisis, emplear los siguientes métodos inferenciales que permiten llevar adelante las comparaciones pertinentes:

- ANOVA
- Kruskal-Wallis
- Bootstrapping

Tener en cuenta que todas las técnicas mencionadas descansan en ciertos supuestos para que su aplicación sea válida. Mencionar cuáles son y si los mismos se cumplen o no, con o sin ayuda de transformaciones. Aplicar todas las técnicas independientemente de la verificación o no de los supuestos, *con fines didácticos*.

Instrucciones para bootstrapping

Bootstrapping es una técnica no paramétrica que permite estimar distribuciones muestrales de estadísticos, y así hacer inferencia, basándose en el remuestreo con reposición de los datos recolectados. Si bien es una metodología con mucho y complejo desarrollo teórico y con diversas variantes que se encuentran ya implementadas en los distintos softwares estadísticos, vamos a realizar nuestra propia versión simplificada. Esto añade a la tarea un pequeño ejercicio de programación, en el caso de no sentirse cómodos en la solución del mismo pueden consultar a los docentes y les indicaremos qué paquete y función de R pueden utilizar.

Efron y Tibshirani, dos de los principales teóricos del *bootstrap* propusieron en su libro *An Introduction to Bootstrap* el siguiente algoritmo para la comparación de dos medias:

224 HYPOTHESIS TESTING WITH THE BOOTSTRAP

Algorithm 16.2

Computation of the bootstrap test statistic for testing equality of means

1. Let $\hat{F}$ put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, 2, \ldots n$, and $\hat{G}$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{x}$, $i = 1, 2, \ldots m$, where $\bar{z}$ and $\bar{y}$ are the group means and $\bar{x}$ is the mean of the combined sample.
2. Form B bootstrap data sets $(\mathbf{z}^*, \mathbf{y}^*)$ where $\mathbf{z}^*$ is sampled with replacement from $\tilde{z}_1, \tilde{z}_2, \cdots \tilde{z}_n$ and $\mathbf{y}^*$ is sampled with replacement from $\tilde{y}_1, \tilde{y}_2, \ldots \tilde{y}_m$.
3. Evaluate $t(\cdot)$ defined by (16.5) on each data set,

$$t(\mathbf{x}^{*b}) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\bar{\sigma}_1^{2*}/n + \bar{\sigma}_2^{2*}/m}}, \quad b = 1, 2, \cdots B. \qquad (16.6)$$

4. Approximate $\mathrm{ASL}_{\mathrm{boot}}$ by

$$\widehat{\mathrm{ASL}}_{\mathrm{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t_{obs}\}/B, \qquad (16.7)$$

where $t_{obs} = t(\mathbf{x})$ is the observed value of the statistic.

La idea de este algoritmo es recrear las observaciones disponibles en un escenario donde es cierta la hiptósis nula de igualdad de medias, remuestrear a partir de la misma generando una distribución empírica de la estadística de la prueba y luego comparar el valor observado de la estadística en la muestra original con dicha distribución empírica. Generalizar este procedimiento para el caso de más de dos grupos y emplearla en el análisis de datos.

Se debe tener en cuenta que esta versión simplificada de *bootstrapping* también descansa en el supuesto de que las distribuciones presentan similar dispersión. Aplicar el método aún si esto no es así, ya que no buscamos profundizar en aspectos teóricos del *bootstrap* (que dan para un curso en sí mismo) sino introducir su existencia y principios generales.

Problema 2

Para cada una de las canciones disponibles en su servicio de streaming, Spotify registra una serie de variables sobre características generales (artista, duración, disco, etc.) y sobre aspectos que tratan de describir características musicales o sobre la *intención* de la canción. Este tipo de información está disponible en el portal de Spotify para desarrolladores y se puede descargar mediante R con el paquete Rspotify.

El siguiente cuadro presenta una breve descripción de dichas variables, conocidas como *song features* (puede consultarse más información acá):

Nombre original	Traducción	Tipo de variable	Interpretación
energy	energia	numérica	Intensidad y actividad (de 0 ->1), ej: heavy metal cercano a 1
key	tonalidad	factor	Tonalidad: 0 = Do, 1 = C#, 2 = Re, y así sucesivamente
loudness	volumen	numérica	Volumen promedio en decibeles
mode	modo	factor	Modo: Mayor/Menor
speechiness	hablado	numérica	Detecta la presencia de palabra hablada (0->1), ej: un podcast tendría puntaje alto
acousticness	acústico	numérica	Detecta si es acústica (0->1)
instrumentalness	instrumental	numérica	Instrumental (0->1), más de 0.5 es considerado instrumental
liveness	en vivo	numérica	Público presente (0->1), ej: un recital debería tener valor cercano a 1
valence	positividad	numérica	Positividad (0->1) valores cercanos a 1 son alegres, eufóricos, y valores cercanos a 0 son tristes o de ira
tempo	tempo	numérica	Tempo promedio (pulsos/minuto)
duration_ms	duración	numérica	Duración en milisegundos
time_signature	tiempo compás	entera	Cantidad de pulsos por compás

Para este problema, hemos seleccionado 6 discos que recorren la trayectoria como solista del gran músico argentino Charly García. Los datos se encuentran en el archivo `charly.txt`. El objetivo es analizar si las *song features* presentan comportamientos similares a lo largo de la obra de Charly, señalando patrones constantes en su autoría, o si hay discos que se diferencian del resto, indicando la existencia de modos de composición particulares en cada producción.

Para esto, vamos a reconocer que no estamos ante un caso de análisis demasiado formal con definiciones sobre las que descansan las herramientas estadísticas clásicas (¿cuál es la población?, ¿cómo fueron seleccionadas las muestras?, ¿son aleatorias?, ¿qué significaría hacer inferencia?, ¿tiene sentido hablar de independencia entre las observaciones?, etc.). De todos modos, vamos a lanzarnos a hacer un análisis descriptivo, imaginando que tal vez las canciones incluidas en un disco pueden pensarse como una muestra que representa la intencionalidad creativa del autor en el período de la publicación del mismo. Por lo tanto, no es necesario que finalicen su análisis con el empleo de herramientas inferenciales, pero sí debemos tener cuidado que descripciones que pretendan concluir sobre diferencias en posición son válidas sólo si las distribuciones son similares.

Pueden recurrir a otras técnicas gráficas que les resulten de interés.