Econometría Aplicada I

Tarea 1

Statistics for Starbucks

El **objetivo** de esta tarea es utilizar las técnicas estadísticas y de probabilidad revisadas en el curso que permitan realizar un **análisis estadístico** aplicado a partir de un conjunto de datos que contiene toda la información nutricional del menú de la conocida marca `Starbucks`.

Para realizar la lectura del archivo denominado `starbucks.csv` desde el sitio de Internet donde se encuentra disponible, utilice la función en `read_csv` del paquete `readr` desde el siguiente sitio:
`https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-12-21/starbucks.csv`

Se disponen de 1,147 observaciones para 15 variables diferentes. En la última página de este archivo encontrará la descripción de las variables de las que dispone . Puede ver la fuente de la información del menú de Starbucks en:

`https://www.behance.net/gallery/58743971/Starbucks-Menu-Infographic-Design`

así mismo, puede revisar un ejemplo de una visualización de la información de este conjunto de datos en `R` en el siguiente sitio:

`https://www.elenadudukina.com/post/tidytuesday-wk52-2021/2021-12-21-tidy-tuesday-starbucks/`

0. Instrucciones Generales

A partir de la base de datos definida, resuelva cada uno de los siguientes ejercicios. No olvide agregar las interpretaciones a sus resultados cuando sea necesario. Deberán cargar en el vínculo habilitado para ello en el sitio Canvas del curso dos archivos:

1. Un primer archivo de texto con sus respuestas. Este archivo debe ser auto-suficiente. Esto quiere decir que con revisar este archivo debe ser posible calificar su tarea en su totalidad. Para este archivo pueden utilizar el formato de su preferencia (por ejemplo: pdf, LATEX, Word o incluso hojas escritas a mano y escaneadas).

2. El segundo archivo es un archivo de soporte que genere todos sus resultados. Este sí es un archivo de programación que al ejecutarse replicaría los resultados que se reportan en tu primer archivo. Pueden elegir el *software* de su preferencia, puede ser un R-script, Do-File o similar, aunque las recomendaciones que aquí se comparten son para el lenguaje `R`.

La fecha de entrega es el próximo **jueves 14 de septiembre** hasta las 23:59 horas. Se admiten entregas extemporáneas en el mismo vínculo pero con una penalización en su nota por cada día de retraso.

1. Análisis Descriptivo.

Inicie realizando un análisis descriptivo del conjunto de datos. Esto siempre debe ser el punto de partida de cualquier análisis estadístico pues permite identificar la composición del conjunto de datos así como la relevancia de la información recabada.

a. Realice un análisis descriptivo de los datos de manera general utilizando las funciones de alguno de los siguientes paquetes que lo hacen de forma automática: `skimr`, `DataExplorer`, `GGally`, `SmartEDA`. Elija el que considere más adecuado y observe con detalle los resultados del reporte que ha obtenido.

b. Construya una tabla en la que reporte los estadísticos descriptivos básicos (media, mediana, desviación estándar, estadísticos de posición) para las variables: `Serv Size_mL`, `Calories`, `Total Fat_g`, `Cholesterol_mg`, `Sugar_g`, `Caffeine_mg`.

c. Para las mismas variables, ahora realice un análisis gráfico que incluya (1) el histograma, (2) boxplot, (3) diagrama de dispersión entre las variables.

d. Construya un gráfico adecuado que permita identificar el resultado de incluir crema batida en su bebida sobre el total de calorías, grasa, colesterol o azúcar.

e. Redacte puntualmente cinco conclusiones relevantes que ha obtenido de este análisis descriptivo.

2. Estimación puntual

El análisis descriptivo de las variables nos ha dado una idea del tipo de distribución que éstas pueden llegar a tener. Ahora estimaremos puntualmente los parámetros de la posible distribución que pueden llegar a tener.

a. Utilice la función `fitdistr` del paquete `MASS` para estimar los parámetros de la función más adecuada para las variables: `Sugar_g` y `Caffeine_mg` mediante el método de máxima verosimilitud. Considere que no solo puede estimar una normal, la función tiene varias opciones para distribuciones no normales. Note que estos estimadores coincidirán con los que puede derivar de manera analítica en cada caso.

b. Construya una tabla que resuma los estimadores que obtuvo para cada parámetro de la distribución de probabilidad.

c. Grafique el histograma para cada una de estas variables en el que se sobreponga la distribución que ha estimado. Utilice también un gráfico cuantil-cuantil para verificar que el ajuste de la distribución empírica con la teórica es lo suficientemente adecuado.

3. Estimación por intervalo

Ahora estamos interesados en analizar la cantidad de energía medida en calorías (KCal) de cada café. Utilizaremos estimaciones por intervalo para el valor promedio de las calorías de cada bebida. Para estas estimaciones, considere que la distribución de esta variable es normal.

a. Construya un intervalo de confianza al 95 % para la media de la variable `Calories`. Interprete el resultado que está obteniendo.

b. Ahora construya el mismo intervalo de confianza al 95 % para la media de la variable `Calories` pero separada para cada tamaño de bebida (variable `Size`). Interprete sus resultados

c. Construya un gráfico que permita comparar estos intervalos de confianza que ha obtenido. No olvide que en este gráfico debe aparecer la estimación puntal y los límites del intervalo de confianza para cada tamaño de la bebida.

d. Escriba una conclusión de este análisis entre el número de calorías y el tamaño de la bebida.

4. Pruebas de Hipótesis

Por salud, los clientes prefieren cuidar la ingesta de calorías, grasa y azúcar en sus bebidas. Ahora se le pide probar la hipótesis de que el uso de los diferentes tipos de leche en la bebida (`Milk`) tiene un efecto significativo sobre estas variables (`Calories, Total_Fat_g, Sugar_g`). Para cada uno de estos tres caso, elija los subgrupos de tipo de leche que considere más adecuados a probar.

a. Plantea la prueba de hipótesis relevante a probar. Recuerde que no siempre la hipótesis alternativa es que el parámetro sea diferente entre grupos.

b. Realice la prueba utilizando una significancia del 5 %. Utilice el estadístico adecuado y concluya, aquí no utilice el valor-p o un intervalo de confianza.

c. Ahora calcule el valor-p exacto para esta prueba y concluya.

d. Interprete correctamente el resultado de esta prueba. Para cada uno de los tres casos, escriba una conclusión puntual a partir de los resultados obtenidos.

e. ¿Cree que existe alguna otra variable relevante que deba considerar y que permita concluir algún efecto significativo sobre la ingesta de calorías, grasa o azúcar en la bebida que no sea sólo el tipo de leche de la bebida? Justifique su respuesta.

5. Estimación por remuestreo

El menú que se nos comparte, nos muestra una posible relación inversa significativa entre cafeína (`Caffeine_mg`) y calorías (`Calories`) en las bebidas, una muy buena noticia para los amantes del café.

a. Calcule el coeficiente de correlación lineal entre estas variables y construya su diagrama de dispersión que incluya el resultado de la regresión lineal simple y su intervalo de confianza.

b. Para probar la significancia de este coeficiente de correlación nos interesaría contar con una estimación de su desviación estándar. Utilice la técnica de *Bootstrap* para estimar su varianza. Utilice dos tamaños de submuestras diferentes (el 50 % de los datos y el total de los datos) con 1,000 repeticiones. Compare el resultado de sus estimaciones.

c. Construya el histograma con los resultados de la simulación de este ejercicio.

d. Realice las mismas estimaciones pero utilizando la técnica *Jacknife*.

e. Con la estimación de la varianza que obtuvo por el método de *Bootstrap* con 1,000 repeticiones con el total de la muestra, realice la prueba de hipótesis de que el coeficiente de correlación entre cafeína y calorías en las bebidas es negativa.

Descripción de la base de datos: starbucks.csv

Información nutricional de bebidas de Starbucks Coffee Company

Variable	**Clase**	**Descripción**
Product_Name	Texto	Product Name
Size	Texto	Size of drink (short, tall, grande, venti)
Milk	Cualitativa	Milk Type: type of milk used 0 = none 1 = nonfat 2 = 2 % 3 = soy 4 = coconut 5 = whole
Whip	Cualititativa	Whip added or not (binary 0/1)
Serv_Size_mL	Cuantitativa	Serving size in ml
Calories	Cuantitativa	KCal
Total_Fat_g	Cuantitativa	Total fat grams
Saturated_Fat_g	Cuantitativa	Saturated fat grams
Trans_Fat_g	Cuantitativa	Trans fat grams
Cholesterol_mg	Cuantitativa	Cholesterol mg
Sodium_mg	Cuantitativa	Sodium milligrams
Total_Carbs_g	Cuantitativa	Total Carbs grams
Fiber_g	Cuantitativa	Fiber grams
Sugar_g	Cuantitativa	Sugar grams
Caffeine_mg	Cuantitativa	Caffeine in milligrams

Fuente: https://www.behance.net/gallery/58743971/Starbucks-Menu-Infographic-Design