Máster en Big Data. Fundamentos matemáticos del análisis de datos.

Examen.

Fernando San Segundo

Curso 2019-20. Instrucciones actualizadas en fecha: 2020-10-08

Preliminares e instrucciones de entrega.

- Crea una carpeta para los ficheros del examen llamada:

  `examen_apellido_nombre`

  También puedes, si lo prefieres, crear esa carpeta como un proyecto de RStudio.

- Guarda el fichero `2019-Examen-MBDFME.Rmd` disponible en Moodle en esa carpeta. Es el fichero que usarás para introducir tus respuestas. Si tienes problemas para descargarlo desde Moodle abre un script vacío en el editor de RStudio y *eligiendo previamente la carpeta del examen como directorio de trabajo* copia este script y ejecútalo:

```
# Comprobación del directorio de trabajo
getwd()

examFile =  "2019-Examen-MBDFME.Rmd"

examURL = paste0("https://musing-stonebraker-3eeaa5.netlify.app/", examFile)

if(!file.exists(examFile)){
  download.file(examURL, examFile)
} else {
  warning(paste0("Cuidado: el fichero del enunciado", examFile, " ya existe en tu directorio de tra
}
```

- La carpeta del proyecto debe contener una subcarpeta llamada `datos`.

- Vamos a trabajar con una tabla de datos llamada `frami.csv`, que puedes descargar desde este enlace: frami.csv
  Precaución: no es exactamente la misma que hemos usado en el curso. Si ese enlace no funciona abre un script vacío en el editor de RStudio y *eligiendo previamente la carpeta del examen como directorio de trabajo* copia este script y ejecútalo:

```
# Comprobación del directorio de trabajo
getwd()

dataFile =  "frami.csv"

dataURL = paste0("https://musing-stonebraker-3eeaa5.netlify.app/", dataFile)

localFile = paste0("./datos/", dataFile)
```

```r
if(!dir.exists("datos")) warning("CUIDADO: No existe la carpeta datos en tu directorio de trabajo.'
if(!file.exists(localFile)){
  download.file(dataURL, destfile = localFile)
} else {
  warning(paste0("Cuidado: el fichero de datos", dataFile, " ya existe en tu carpeta de datos"))
}
```

- Cuando termines de responder a las preguntas del examen, comprime la carpeta de tu proyecto en formato `zip`, de manera que el archivo comprimido tenga el nombre de esa carpeta y extensión `.zip`. Enviáme por correo ese fichero comprimido a la dirección `fsansegundo@comillas.edu`.
- Se valorará adicionalmente el uso de funciones del tidyverse (dplyr, ggplot).

Apartado 1.

Lee el fichero de datos y crea un data.frame de R llamado `frami`, que usaremos para el resto del examen.

- C1: ¿Cuántas observaciones hay en la tabla? ¿Cuántas variables?
- C2: ¿Hay datos ausentes? ¿De qué tipo son las variables?
- C3: Antes de seguir adelante vamos a eliminar todas las observaciones que contienen datos ausentes. Asegúrate de que el conjunto de datos resultante se sigue llamando `frami`. ¿Cuántas observaciones han quedado en la tabla tras eliminar los datos ausentes?

Apartado 2.

- C4: Vamos a analizar la posible relación entre el factor género (variable `male`) y el factor fumador (variable `currentSmoker`). Haz un análisis mediante gráficos de esa relación. Calcula la tabla de probabilidad, con probabilidades marginales, asociada a la relación entre estos dos factores.
- C5: ¿Cuál es la probabilidad de que un fumador sea mujer?
- C6: ¿Y cuál es la probabilidad de que un hombre sea fumador?
- C7: ¿Cuál es la probabilidad de que una persona elegida al azar sea fumadora? ¿Y de que sea a la vez hombre y no fumador?
- C8: Si las dos *variables aleatorias* género `male` y condición de fumador `currentSmoker` fueran independientes se cumpliría exactamente:

$$P(\text{hombre y fumador }) = P(\text{hombre}) \cdot P(\text{fumador })$$

  Pero los datos que tenemos son muestrales, así que lo más que podemos esperar es una igualdad aproximada. ¿Se cumple esa igualdad aproximada en nuestros datos? ¿Qué opinas sobre la independencia de estas variables?
- C9: Si elegimos de forma independiente (con remplazamiento) 10 personas de esta muestra, ¿cuál es la probabilidad de que 4 de ellas sean fumadoras? ¿Y cuál es la probabilidad de que lo sean 4 o menos?
- OP1: *Opcional:* la prueba χ^2, que no hemos visto en clase, es un contraste de hipótesis sobre independencia de dos factores. La *hipótesis nula* de ese contraste es siempre *los factores son independientes*. Ejecuta en R:

```r
chisq.test(frami$male, frami$currentSmoker)
```

  Examina el resultado y saca conclusiones sobre la independencia de esos dos factores.

Apartado 3.

- C10: En este apartado trabajamos con la variable `heartRate`, la frecuencia cardiaca. Haz un resumen numérico básico. ¿Es representativa la media? Represénta la variable gráficamente (puuedes usar más de un tipo de gráfico). Se valorará especialmente el uso de `ggplot`.
- C11: ¿Hay datos atípicos? ¿Cuántos? ¿Sabes qué posiciones (número de fila) ocupan en la tabla?
- C12: ¿Crees que está variable es aproximadamente normal? Justifica tu conclusión basándote en los gráficos del apartado anterior o en algún gráfico adicional.
- C13: *Asumiendo que es aproximadamente normal (sea cual sea tu conclusión del apartado anterior)* construye un intervalo de confianza al 95% para la media de esta variable,

Apartado 4

En las preguntas de este apartado se puede usar R básico o dplyr. Procurad hacer al menos una con `dplyr` y, en general, combinad varias técnicas para responder a las preguntas.

- C14: ¿Cuántas personas padecen diabetes? Para las personas que padecen diabetes y que tienen edades superiores a 50 años ¿cuál es su nivel medio de colesterol?
- C15: Haced una tabla cruzada en la que parezca el nivel medio de colesterol para cada combinación posible de las variables nivel de educación y género (`male`).

Apartado 5

- C16: Analicemos ahora la posible relación entre las variables presión sistólica `sysBP` y presión `diastólica diaBP`. Haz un diagrama de dispersión y un modelo de regresión lineal usando `sysBP` como variable respuesta y `diaBP` como variable explicativa. Añade la recta de regresión lineal al diagrama de dispersión.
- C17: ¿Crees que el modelo de regresión lineal es adecuado para describir la relación entre esas variables?
- C18: ¿Qué porcentaje de la variabilidad en `diaBP` se explica con ese modelo de regresión?
- C19: ¿Cuál es el valor de `diaBP` que predice el modelo para alguien con `sysBP` igual a 160? Se valora el uso de `predict`.
- C20: ¿Cuál es el residuo de la primera observación de la tabla?
- OP2: *Opcional:* vamos a pensar sobre la posible influencia del género en ese modelo de regresión. Pra ello vamos a construir dos modelos de regresión adicionales para las mismas variables (diaBP, sysBP). Pero en el primer modelo usamos solo observaciones de hombres y en el segundo modelo solo observaciones de mujeres. Representa las dos rectas en el mismo diagrama de dispersión que la recta del primer modelo (en total habrá tres rectas). Saca conclusiones sobre la posible influencia del género en la regresión.

Apartado 6

- C21: Pensemos en la variable nivel de colesterol `totChol`. Sospechamos que el nivel medio de colesterol de las personas de la muestra es mayor que 235. ¿Avalan los datos esta sospecha, con un nivel de significación del 95%?
- C22: Repite el análisis pero ahora solo con observaciones correspondientes a hombres. Y luego con observaciones de mujeres. ¿Qué piensas sobre la relación entre nivel de colesterol y el género?
- C23: Estudia la normalidad de la variable *edad*.
- OP3: *Opcional:* en el tema 5 hemos visto como usar el bootstrap para calcular un intervalo de confianza. Usa ese método para hacer un intervalo de confianza bootstrap al 95% para la edad en la muestra

completa. A la vista del apartado anterior ¿crees que habrá diferencia entre este intervalo bootstrap y el clásico? ¿cuál es preferible?

- C24: Divide la variable edad en 4 subintervalos de longitud 10 años y llama `franjaEdad` al resultado. Divide el colesterol en intervalos de 100 unidades desde 100 hasta 600 y llama `nivelCol` al resultado. Ahora estudia la psible relación entre `franjaEdad` y `nivelCol`.