Parte A: Estadística Descriptiva

El RMS Titanic fue en su momento el mayor barco de pasajeros del mundo, hundiéndose en su viaje inaugural de Southampton a Nueva York en el año 1912. En el evento fallecieron 1514 de las 2223 personas que iban a bordo, entre tripulación y pasajeros.

En el presente práctico se trabajará con el conjunto de datos titanic, que figura en el data.frame titanic.txt. El conjunto de datos es un clásicos de las competencias de "Machine Learning", donde se busca determinar un mecanismo de clasificación que, en función de diversos variables de cada pasajero, prediga si el pasajero sobrevivió o no a la catástrofe. Las variables del conjunto de datos son:

survival: supervivencia (0 No, 1 Sí).

pclass: clase del pasajero (1,2 o 3).

sex: sexo del pasajero ("male", "female").

age: edad del pasajero.

sibsp: cantidad de hermanos y cónyuges (totalizado) embarcados (número entero).

parch: cantidad de padres e hijos (totalizado) embarcados (número entero).

ticket: código del boleto (texto).

fare: tarifa del pasaje (número real).

embarked: puerto de embarque (S= Southampton, Q=Queenstown, C = Cherbourg)

1. Leer el conjunto de titanic.txt teniendo en cuenta que en la primera línea del archivo figura el nombre de las variables y los datos se encuentran separados por tabulaciones (sep="\t"), y asígnelo al data.frame titanic.
2. Inspeccionar los primeros casos del archivo.
3. Establecer el número de variables y de casos.
4. Realizar un attach de titanic.
5. Inspeccionar los nombres de las variables de titanic e identificar de qué tipo de variable se trata cada una de ellas.
6. Decidir usando un método gráfico si el sexo del pasajero fue importante para su supervivencia.
7. Decidir usando un método gráfico si la edad del pasajero fue importante para su supervivencia.
8. Decidir usando un método gráfico si la clase del pasajero fue importante para su supervivencia.
9. ¿Tiene sentido asumir que la variable que indica la tarifa tiene distribución normal? ¿Por qué? ¿Puede decidir de antemano quién es mas grande si la media o la mediana? Obtener estas dos medidas de centralidad y computar la media 0.1-podada.

10. Hay algunos pasajeros a los que les vieron la cara cuando compraron el boleto. Obtener sus nombre. ¿Sobrevivieron?

11. Decidir usando un método gráfico si la tarifa del pasajero fue importante para su supervivencia.

12. Respecto a la relación entre la edad y la tarifa podemos pensar que las personas más jóvenes tenían menos dinero y por ende compraron los tiquetes más baratos. ¿Puede confirmar esto en base a los datos?

13. Obtener las medidas de dispersión vistas en clase para la variable edad. Únicamente en base a esto, ¿se puede decir algo de la normalidad de esta variable?

14. Para los mayores de 10 años, calcular la diferencia entre la suma de las tarifas pagadas por hombres y la suma de las tarifas pagadas por mujeres. Hacer lo mismo para mayores de 0.5 años, 15 años, 20 años, 40 años y 60 años. Hacer este ejercicio usando la menor cantidad de copy-paste posible.

Parte B: Algo de Programación

Los monos de Shakespeare

Un resultado simpático y bastante fácil dice que si uno pone a un mono a tipear al azar, con probabilidad 1 eventualmente va a escribir las obras completas de Shakespeare (obviamente el resultado vale para cualquier secuencia finita de letras). El resultado es todavía más fuerte: la esperanza del tiempo que tarda el mono en escribir esta secuencia de letras es finita (probar esto no es difícil, pueden intentarlo). El objetivo de este ejercicio es estimar esta esperanza empíricamente.

Como tenemos recursos de tiempo y memoria limitados, vamos a suponer que nuestro lenguaje tiene solo dos letras: A y B.

1. Hacer una función que tenga como parámetro a una palabra (secuencia de letras) w, simule a un mono escribiendo al azar y devuelva el primer momento en que aparece escrita la palabra w.

2. Replicar la función anterior $m = 1000$ veces y estimar el tiempo esperado que el mono necesita para escribir cada una de las palabras AA, AB, AAA, AAB, ABA, $BABA$ y $BABAB$. ¿Es cierto que el tiempo esperado únicamente depende de de longitud de la palabra? Si se notan diferencias para palabras de la misma longitud, ¿a qué podría deberse?

3. Para una palabra fija, el tiempo de espera es una variable aleatoria. ¿Puede suponerse normal? En caso negativo, ¿a qué familia cree que pertenece?