Máster Universitario en Ciencia de Datos
Fundamentos para el Análisis de Datos y la Investigación.
Prácticas y Tareas 2. Curso 2023-24. Prof. JM Sarabia

1. Simulacion y comando sample

1. Elegir muestras con el comando de `R sample`, con y sin reemplazamiento, a partir de diversas estructuras de datos.

2. Un inversor conoce que la variación del precio de un activo varía de acuerdo a la siguiente distribución de probabilidad:

X	-2	-1	0	1	2
$P(X = x)$	0.15	0.25	0.3	0.2	0.1

   Simular una muestra del precio de 1.000 activos mediante `sample`. Obtener el precio medio, la varianza y los coeficientes de asimetría y curtosis a partir de la muestra simulada.

3. Se lanzan dos dados regulares. Obtener la distribución de probabilidad de la suma de los resultados. Obtener la distribución de la suma mediante simulación usando `sample`.

2. Modelo normal

1. Haciendo uso del diagrama QQ plot y del test de Shapiro-Wilk, estudiar la normalidad de las variables `wage` del paquete ISLR. Estudiar las transformaciones $\log(x)$, $1/x$ y $\sqrt{x}$. Estudiar la normalidad de las variables del fichero de datos `Swiss`.

2. Estudiar la normalidad de las variables de rendimientos del fichero `Cotizaciones2020.tex`.

3. Las ventas semanales V de un producto siguen una distribución normal con media 200 y dt 30.

   a) Hallar: $P(V < 180)$, $P(V > 220)$, $P(190 < V < 230)$,

   b) Ventas máximas correspondientes al 15 por ciento de las menores ventas semanales

   c) Ventas mínimas correspondientes al 87 por ciento de las mayores ventas semanales

   d) Simular una muestra de 1000 ventas semanales.

4. El rendimiento anual de una inversión se supone que sigue una distribución Normal. El rendimiento medio previsto es del 10 % y la volatilidad (desviación típica) del 20 %.

   a) Hallar la probabilidad de que el rendimiento anual sea negativo.

   b) Probabilidad de que la inversión tenga una rendimiento anual superior al 15 %.

   c) Hallar la probabilidad anterior por simulación.

   d) Representar gráficamente las funciones de densidad y de distribución del rendimiento.

   e) Obtener los deciles del rendimiento.

3. Distribuciones en el muestreo

1. A partir de una función de densidad uniforme, obtener la distribución en el muestreo de la media muestral, a partir de diversos tamaños de muestra con `m=10.000` replicaciones.

2. Haciendo uso del dataset `cars`, obtener la distribución en el muestreo de la mediana de la variable `speed`, mediante el método bootstrap.

4. Clasificación mediante Naive Bayes

1. Haciendo uso de la base de datos `iris` queremos predecir la variable `Species` mediante el método Naive Bayes, haciendo uso de las cuatro variables explicativas.

   a) Por medio de la función `naiveBayes` del paquete `e1071` predecir la variable `Species`. Obtener medias pos grupo y explicar como funciona en este caso el método Naive Bayes

   b) Obtener las predicciones el modelo en términos de probabilidades

   c) Obtener la matriz de confusión y el porcentaje de aciertos del método

2. Por medio de la base de datos `Wage` (del paquete ISLR) se desea predecir los niveles del logaritmo del salario en función de las variables disponibles.

   a) Convertir en un factor con 2 niveles la variable log-salario.

   b) Predecir el factor anterior mediante Naive Bayes usando los predictores del dataset. Realizar el mismo ejercicio con 4 niveles del salario