PROBABILIDADES Y ESTADÍSTICA (C)

Práctica 6

Los ejercicios de esta práctica se deben resolver usando R u otro software.
Los archivos de datos se encuentran en la página de la materia.

1. Una fábrica de alfajores tiene dos sedes: una en Quilmes y la otra en Pilar. Cada sede empaqueta sus alfajores en cajas con 4 unidades. Denotemos con X a la variable aleatoria que indica el número de alfajores defectuosos en una caja y con Y a la variable aleatoria que indica la sede de la que proviene la caja, Quilmes($Y = 0$) o Pilar ($Y = 1$).

   En el siguiente link podrá acceder a datos que se obtienen al examinar cajas al azar producidas por la fábrica. Es decir, realizaciones del vector (X, Y).

   Supongamos que se seleccionaron $n = 500$ cajas al azar producidas por la fábrica.

   Ingrese $n = 500$ y número de libreta, para obtener *sus datos*.

   Estimar

   a) la probabilidad de que una caja provenga de la sede Quilmes.

   b) la probabilidad de que una caja sea producida por la sede Quilmes y tenga 3 alfajores defectuosos.

   c) la funcion de probabilidad puntual conjunta del vector (X, Y).

   d) la esperanza y la varianza de X.

   e) la probabilidad de que una caja producida por la sede Quilmes tenga 3 alfajores defectuosos.

   f) la probabilidad de que una caja producida por la empresa y elegida al azar tenga 3 alfajores defectuosos (es decir $X = 3$).

   g) la probabilidad de que una caja con 3 alfajores defectuosos haya sido producida por la sede Quilmes.

   h) la probabilidad de que una caja con 3 alfajores defectuosos haya sido producida por la sede Pilar.

   Comandos de R:

   ```
   alfajores <- read.table("alfajores.txt", header = TRUE)
   table(alfajores)
   ```

2. Se quiere estudiar la distribución de la duración (en horas) de las lámparas producidas por una fábrica. Se observa la duración de 27 lámparas elegidas al azar de la producción de dicha fábrica y se obtienen los siguientes valores (también pueden encontrarlos en el archivo `lamparas.txt`).

26.43	33.58	65.86	29.18	5.92	13.29	13.54	64.78	56.11
23.60	33.39	100.32	28.04	29.63	2.41	3.17	11.99	6.47
23.59	17.96	32.27	2.09	57.43	15.31	42.85	1.68	49.61

a) Estimar la probabilidad de que una lámpara producida por esta fábrica dure más de 30 horas.

b) Implementar y graficar la función de distribución empírica de este conjunto de datos.

c) Completar: Estos datos permiten estimar que el 90 % de las lámparas producidas por esta fábrica dura más de ..... horas y el 10 % dura menos de ..... horas.

3. El archivo **graduados.txt**, contiene los promedios obtenidos en su carrera de grado de 30 inscriptos en el programa de postgrado del Departamento de Ingeniería Industrial e Investigación Operativa de la Universidad de Berkeley, California.

   a) Calcular la media muestral y la mediana muestral.

   b) Calcular el desvío estándar muestral y la distancia intercuartil.

   c) Realice un histograma con los datos y superponga la curva de una densidad normal con los parámetros que considere pertinentes.

   d) Realice un boxplot con este conjunto de datos. ¿Cuáles son sus características más sobresalientes? ¿Cómo relaciona lo observado en los gráficos con los valores estimados de media y mediana obtenidos en a)? ¿Hay outliers?

   e) ¿Qué distribución cree que tienen estos datos?

   f) Superponga en el histograma la curva de una densidad apropiada con los parámetros que considere pertinentes. Explore el comando `density` en R.

   g) ¿Qué otro gráfico conoce que le permitiría verificar si su conjetura es razonable?

   Algunos comandos de R:

```
graduados <- scan("graduados.txt")
hist(graduados, prob = TRUE)
boxplot(graduados)
```

4. La siguiente tabla contiene valores de población, en cientos de miles, de las 10 ciudades más pobladas de 4 países en el año 1967. Estos datos se encuentran en el archivo **ciudades.txt**.

Argentina	**EEUU**	**Holanda**	**Japón**
29.66	77.81	8.68	110.21
7.61	35.50	7.31	32.14

6.35	24.79	6.02	18.88
4.10	20.02	2.64	16.38
3.80	16.70	1.75	13.37
2.75	9.39	1.72	11.92
2.70	9.38	1.51	10.71
2.69	8.76	1.42	7.80
2.51	7.63	1.31	7.70
2.44	7.50	1.29	7.00

a) Construir en paralelo, para facilitar la comparación, un boxplot para los datos de cada país e identificar los puntos extremos en cada uno de ellos.

Comandos de R:

```
ciudades <- read.table("ciudades.txt", header = TRUE)
View(ciudades)
boxplot(ciudades)
```

b) Comparar los centros de cada población, sus dispersiones y su simetría. ¿Cuál es el país más homogéneamente habitado?

5. El archivo **ingresos.txt** contiene el ingreso mensual de un conjunto de 1000 trabajadores registrados de una ciudad, en miles de pesos.

a) ¿Cuál es el ingreso mínimo percibido por los trabajadores encuestados? Estime la proporción de los trabajadores de la ciudad que percibe el ingreso mínimo.

b) Estimar el ingreso mensual que se necesita para pertenecer al 10 % de trabajadores de la ciudad con ingresos más altos.

c) Calcular la media muestral, la mediana muestral y la media α-podada con $\alpha = 0{,}10$ (10 %).

d) Calcular el desvío estándar muestral y la distancia intercuartil.

e) Realizar un histograma y un boxplot. ¿Cuáles son las características más sobresalientes? ¿Hay outliers?

f) ¿Cree que los datos tienen distribución normal?

g) Discutir con un compañero las ventajas y desventajas de cada medida de posición para describir el centro de los datos.

6. Este ejercicio es para familiarizarse con el uso e interpretación de los QQ–plots.

a) Generar muestras de tamaño 25, 50 y 100 de una distribución normal. Construir QQ–plots para cada una de ellas. Repetir varias veces.

b) Repetir a) para una $\Gamma(5, \frac{1}{2})$.

c) Repetir a) para $Y = \dfrac{Z}{U}$ donde $Z \sim N(0,1)$ y $U \sim \mathcal{U}(0,1)$ independientes.

d) Repetir a) para una distribución uniforme.

e) Repetir a) para una distribución exponencial.

f) ¿Puede distinguir, en base a los QQ–plots, entre la distribución normal del ítem a) y las siguientes distribuciones que no son normales?

7. Generar un conjunto de $n = 10$ datos provenientes de una distribución cualquiera. Llamémoslos $x_1, \ldots, x_n$.

   a) Implementar y graficar la función ℓ_2 que, a cada c le asigna $\ell_2(c) = \sum_{i=1}^{n} (x_i - c)^2$. ¿Para qué valor o valores de c se minimiza la función? Experimentar con otros valores de n. Conjeturar una respuesta y luego hacer la demostración para probar su conjetura.

   b) Implementar y graficar la función ℓ_1 que ahora, a cada c le asigna $\ell_1(c) = \sum_{i=1}^{n} |x_i - c|$. ¿Para qué valor o valores de c se minimiza la función? Experimentar con varios valores de n y conjeturar una respuesta.

8. Consideremos $X_1, \ldots, X_n$ una muestra de una población cualquiera.

   a) Sean $\bar{X}$ y $\tilde{X}$ la media y la mediana muestral, respectivamente.

      i) Si se suma una constante c a cada uno de los X_i de la muestra, obteniéndose $Y_i = X_i + c$, ¿cómo se relacionan $\bar{X}$ con $\bar{Y}$ y $\tilde{X}$ con $\tilde{Y}$?

      ii) Si cada X_i es multiplicado por una constante c, obteniéndose $Y_i = cX_i$, responder a la pregunta planteada en (i).

   b) Sea $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ la varianza muestral correspondiente a la muestra. Demostrar que:

      i) Si $Y_i = X_i + c$, con c constante, entonces $S_Y^2 = S_X^2$.

      ii) Si $Y_i = cX_i$, con c constante, entonces $S_Y^2 = c^2 S_X^2$.

      iii) $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2$.

9. El conjunto de datos `departamentos.csv` contiene información sobre departamentos en venta en 2016 en la Ciudad de Buenos Aires. Fuente: `https://data.buenosaires.gob.ar`

   a) Grafique el precio en dólares vs la cantidad de metros cuadrados cubiertos. ¿Cómo describiría la relación?

   b) Estime la correlación entre las variables X = cantidad de metros cuadrados cubiertos e Y = precio en dólares.

   c) Estimar la recta de regresión de Y vs X, es decir, la recta que, para cada X, da el mejor predictor de Y basado en X, según el criterio del ECM.

   d) Superponer el gráfico de la recta de regresión estimada al grafico realizado en el ítem a).