PROBABILIDADES Y ESTADÍSTICA (C)

Práctica 6

Los ejercicios 2 a 6 de esta práctica se deben resolver usando R u otro software.
Los archivos de datos se encuentran en la página de la materia.

1. Consideremos $X_1, \ldots, X_n$ una muestra de una población cualquiera. Sean $\bar{X}$ y $\tilde{X}$ la media y la mediana muestral, respectivamente.

   a) 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 $\tilde{Y}$?

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

   c) Sea S_X^2 la varianza muestral correspondiente a la muestra. Demostrar que:

      i) $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2$.
      ii) Si $Y_i = X_i + c$, con c constante, entonces $S_Y^2 = S_X^2$.
      iii) Si $Y_i = cX_i$, con c constante, entonces $S_Y^2 = c^2 S_X^2$.

   d) Sean $E(X_i) = \mu$ y $\text{med}(X_i) = \tilde{\mu}$.

      i) ¿Para qué valores de c se minimiza $\sum_{i=1}^n (X_i - c)^2$?
      (Sugerencia: derivar con respecto a c).

      ii) Usando el ítem anterior decidir cuál de estas dos cantidades es más pequeña:

$$\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 \qquad \text{ó} \qquad \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 .$$

      iii) ¿Para qué valores de c se minimiza $\sum_{i=1}^n |X_i - c|$?
      (Sugerencias: 1: ¿Se puede usar la misma técnica que en i)? 2: Para fijar ideas, comience con una muestra de tamaño $n = 3$).

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

   d) ¿Qué distribución cree que tienen estos datos?

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

3. 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 box-plot para los datos de cada país e identificar los puntos extremos en cada uno de ellos.

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

4. 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?

5. Con el fin de determinar cuál sería un mejor suplemento dietario, se realizó una comparación de la retención de dos formas de hierro: Fe2+ y Fe3+. Los investigadores dividieron aleatoriamente a 36 ratas en dos grupos de igual número. A un grupo se le suministró en forma oral una concentración de 1.2 millimolar de Fe2+ y al otro grupo se le suministró la misma concentración de Fe3+. Al cabo de cierto tiempo se realizó un conteo en cada rata para determinar el porcentaje de hierro retenido. Estos datos se encuentran en el archivo **hierro.txt**.

   a) Realizar los boxplots y los qq-plots de los porcentajes obtenidos para cada grupo. En base a estos gráficos, ¿es razonable suponer que cada uno de los conjuntos de datos provienen de una distribución normal? ¿Por qué?

b) En una segunda etapa, los investigadores transformaron los datos aplicando la función logaritmo natural (ln) a cada una de las observaciones. Repetir el análisis anterior con los datos transformados. En base a la información obtenida, ¿es razonable suponer que cada uno de los conjuntos de datos transformados se distribuyen según una distribución normal? ¿Por qué?

6. El archivo **cpu.txt** contiene tiempos de CPU (en segundos) correspondientes a 1000 trabajos enviados por una consultora. Para este conjunto de datos:

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

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

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

   d) ¿Cree que los datos tienen distribución normal?

   e) ¿Qué medida de posición considera más apropiada para describir el centro de los datos?