Ejercicios 2

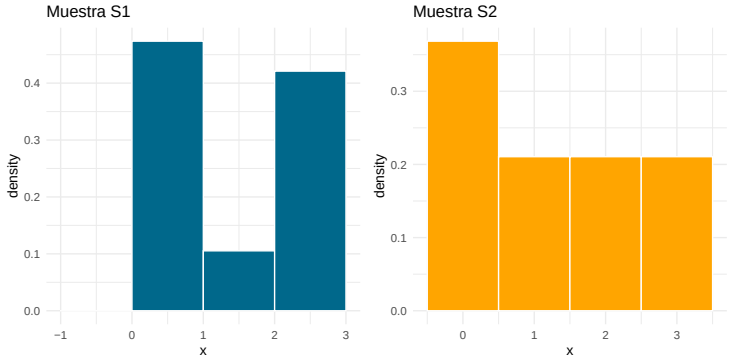
Rodrigo Zepeda

6 de febrero de 2020

1. Análisis Exploratorio de Datos II

1. Determina si es posible concluir a partir de la información dada justificando matemáticamente la conclusión o dando un contraejemplo donde se demuestre que la conclusión puede ser falsa.

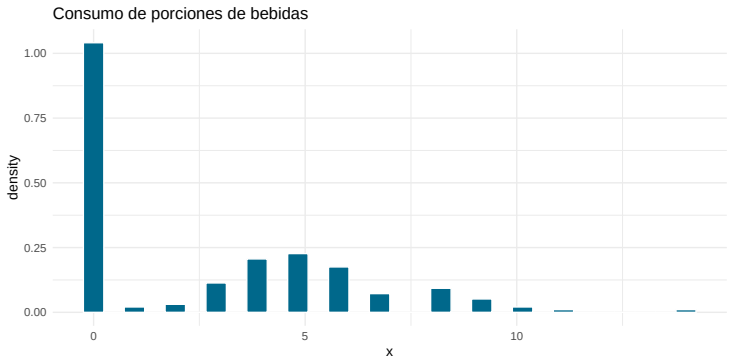
   a) Los histogramas de la siguiente figura corresponden a dos muestras $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ distintas:



   b) Si la temperatura mediana de la Ciudad de México en Febrero es de 13°C y la forma de convertir de Celsius a Farenheiht (°F) es a través de la fórmula $^\circ F = \frac{9}{5} \cdot {}^\circ C + 32$, entonces la temperatura mediana de la ciudad de México en Farenheiht es 55.4°F.

   c) Si un valor es identificado como *outlier* entonces dicho valor es un error de medición y no proviene de la misma distribución que el resto de las x_k en la muestra.

   d) Se midió el consumo de bebidas alcohólicas en distintas personas y se cuantificaron los vasos (250 ml) al día consumidos en las variables $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$. El histograma de dichas variables es como sigue:



1) El valor de $x_n = 14$ es un outlier.
2) Un estimador robusto de la media es la media truncada por ambos lados al 5 %.

e) Si $\mathcal{S} = \{x_1\}$ la función de distribución acumulada empírica coincide con la de una variable aleatoria Y dada por $Y = x_1 \cdot Z$ donde $\mathbb{P}(Z = z) = 1$

f) Si $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ y $F_n(x)$ es la distribución acumulada empírica asociada, entonces la media $\mathbb{E}[X]$ de una variable aleatoria $X \sim F_n(x)$ coincide con la media muestral $\mu_S = \sum_i x_i/n$

2. Sea $\hat{f}_h(x)$ la densidad kernel asociada a $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ con un núcleo $K(u) \geq 0$ y $h > 0$. Demuestra:

a) $\hat{f}_h(x)$ es una densidad de probabilidad.

b) Si $n = 1 = h$ y el Kernel es Gaussiano, $\hat{p}_h(x)$ corresponde a una normal con media x_1.

c) Suponiendo que $x_1 < x_2 < \ldots < x_n$ y determina la media de una variable aleatoria X que se distribuye con densidad $\hat{p}_h(x)$ bajo:

1) Kernel rectangular
2) Kernel epanechnikov
3) Kernel gaussiano

3. Sea H la función de histograma definida en clase con ancho de banda h. Determina $\lim_{h\to\infty} H_h$.

4. **¿Verdadero o falso? Justifica** ¿Es un histograma un estimador kernel?

5. **¿Verdadero o falso? Justifica** Un histograma es una función de masa de probabilidad.

6. Demuestra que para $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ la función de distribución empírica $F_n(x)$ es una función de distribución acumulada; es decir:

a) $\lim_{x\to-\infty} F_n(x) = 0$

b) $\lim_{x\to\infty} F_n(x) = 1$

c) $\lim_{x\to x_0^+} F_n(x) = F_n(x_0)$ (continua por la derecha)

d) $\lim_{x\to x_0^-} F_n(x)$ existe

e) $F_n(x)$ es no decreciente.

7. La tabla 1 muestra datos observados del PIB de un país en billones de dólares: Ajusta una línea $\ell(x) = ax+b$ y una parábola $q(x) = a\cdot x^2+b\cdot x+c$

Año	**PIB**
2000	0.5
2005	1.2
2010	1.5
2015	2.1

Cuadro 1: Los datos observados

para obtener la mejor línea (resp. parábola) que ajuste esos puntos. ¿Qué valor de PIB se espera para el 2020 bajo cada uno de los dos modelos?

8. Ante una muestra $\mathcal{S} = \{(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$ demuestra que si se desea ajustar la línea $y = \beta_0 + \beta_1 t$ por mínimos cuadrados a los datos, se tiene que:

$$\beta_0 = \bar{y} - \beta_1 \bar{t}$$

donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$ es el promedio de t. Y que además:

$$\beta_1 = \frac{(t_1-\bar{t})(y_1-\bar{y}) + (t_2-\bar{t})(y_2-\bar{y}) + \cdots + (t_n-\bar{t})(y_n-\bar{y})}{(t_1-\bar{t})^2 + (t_2-\bar{t})^2 + \cdots + (t_n-\bar{t})^2}$$

9. Demuestra que si se tiene una muestra de tamaño n con $x_1 < x_2 < \ldots < x_n$, entonces existe un polinomio de grado $n-1$ que ajusta dichos datos a la perfección.