Inferencia Estadística | Tarea 04

Naikary Paloma Martinez Velázquez
Aguirre Calzadilla César Miguel

26 de octubre de 2024

Códigos

Todo el código escrito para esta tarea será anexado en un archivo de RStudio. Dentro se encuentran las rutinas escritas para la tarea así como comentarios sobre las mismas.

Problema 1

Resuelva lo siguiente:

a) Sea $X \sim$ Exponencial(β). Encuentre $P(|X - \mu_X| \geq k\sigma_X)$ para $k > 1$. Compare esta probabilidad con la que obtiene de la desigualdad de Chebyshev.

b) Sean $X_1, \ldots, X_n \sim$ Bernoulli(p) y $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Usando las desigualdades de Chebyshev y Hoeffding, acote $P(|\bar{X}_n - p| \geq \epsilon)$. Demuestre que para n grande la cota de Hoeffding es más pequeña que la cota de Chebyshev. ¿En qué beneficia esto?

Análisis del Problema

Para abordar la primera parte de este problema podemos modelarlo usando la distribución exponencial $X \sim$ Exponencial(β), cuya función de densidad es:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x \geq 0$$

Sabemos que para una distribución exponencial, la media y la varianza están dadas por:

$$\mu_X = \beta, \quad \sigma_X^2 = \beta^2$$

Queremos calcular la probabilidad $P(|X - \mu_X| \geq k\sigma_X)$, que equivale a encontrar $P(|X - \beta| \geq k\beta)$ para $k > 1$.

Para resolver esto, descomponemos la probabilidad en dos partes, teniendo en cuenta que la función de supervivencia de una variable exponencial es $P(X \geq x) = e^{-\frac{x}{\beta}}$.

Finalmente, compararemos este resultado con la cota proporcionada por la desigualdad de Chebyshev, la cual es:

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

Como la segunda parte del problema involucra una muestra de variables aleatorias de Bernoulli. Sabemos que $\mathbb{E}[X_i] = p$ y que la varianza de cada X_i es $\text{Var}(X_i) = p(1-p)$.

Queremos acotar la probabilidad $P(|\bar{X}_n - p| \geq \epsilon)$, que representa la probabilidad de que la media muestral $\bar{X}_n$ se desvíe de p por más de ϵ.

Aplicaremos dos métodos: - Chebyshev, que utiliza la varianza para proporcionar una cota general. - Hoeffding, que es una cota más precisa para variables acotadas como las Bernoulli.

Solución:

a)

La probabilidad que deseamos encontrar es $P(|X - \beta| \geq k\beta)$, que puede escribirse como:

$$P(|X - \beta| \geq k\beta) = P(X \geq \beta(1 + k)) + P(X \leq \beta(1 - k))$$

Sin embargo, dado que $X \geq 0$, la probabilidad $P(X \leq \beta(1 - k)) = 0$ para $k > 1$. Por lo tanto, solo necesitamos calcular $P(X \geq \beta(1 + k))$.

La función de supervivencia de la distribución exponencial nos dice que:

$$P(X \geq \beta(1 + k)) = e^{-(1+k)}$$

Por lo tanto, la probabilidad es:

$$P(|X - \beta| \geq k\beta) = e^{-(1+k)}$$

Comparación con la Desigualdad de Chebyshev

La desigualdad de Chebyshev establece que:

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

Comparando el resultado exacto $e^{-(1+k)}$ con $\frac{1}{k^2}$, observamos que para valores grandes de k, la cota de Chebyshev es mucho más grande que el valor exacto, lo que indica que Chebyshev es una aproximación muy conservadora.

Para la distribución exponencial, la probabilidad exacta decrece exponencialmente, mientras que la cota de Chebyshev solo decrece de forma cuadrática. Esto demuestra que la desigualdad de Chebyshev es bastante conservadora, especialmente para valores grandes de k.

b)

Cota con Chebyshev

Sabemos que la varianza de la media muestral es:

$$\text{Var}(\bar{X}_n) = \frac{p(1 - p)}{n}$$

Usando la desigualdad de Chebyshev, obtenemos:

$$P(|\bar{X}_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}$$

Cota con Hoeffding

La desigualdad de Hoeffding establece una cota más ajustada para variables acotadas, y en este caso nos da:

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Para n grande, la cota de Hoeffding decrece exponencialmente, mientras que la cota de Chebyshev solo decrece a una tasa de $\frac{1}{n}$. Esto hace que la cota de Hoeffding sea mucho más ajustada para muestras grandes, lo cual es particularmente útil cuando se necesitan límites más estrictos para garantizar la precisión.

La cota de Hoeffding es más precisa que la de Chebyshev en el caso de muestras grandes, lo que beneficia en situaciones donde es crucial controlar la probabilidad de grandes desviaciones.

Problema 2

Resuelva lo siguiente:

Parte I

Sean $X_1, ..., X_n$ *Bernoulli*(p). Sea $\alpha > 0$ y defina:

$$\epsilon_n = \sqrt{\frac{1}{2n} log(\frac{2}{\alpha})}$$

Sea $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Defina $C_n = (\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n)$. Use la desigualdad de Hoeffding para demostrar que:

$$P(p \in C_n) \geq 1 - \alpha$$

Diremos que C_n es un $(1-\alpha)$ intervalo de confianza para p. En la práctica, se trunca el intervalo, de tal manera que no vaya debajo del 0 o arriba del 1.

Solución:
Para resolver, hay que definir primero la desigualdad de Hoeffding.

Theorem. *Sean $Y_1, Y_2, .., Y_n$ variables aleatorias independientes tal que $E(Y_i) = 0$, con $a_i \leq Y_i \leq b_1$, y sea $\epsilon > 0$, entonces, para cada $t > 0$ se cumple que:*

$$P\Big(\sum_{i=1}^{n} Y_i \geq \epsilon\Big) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2 \frac{(b_i * a_1)^2}{8}}$$

En particular, si $X_1, X_2, .., X_n$ *Bernoulli*(p) independientes, entonces para cada $\epsilon > 0$ se cumple que:

$$P\Big(|\bar{X}_n - p| > \epsilon\Big) \leq 2e^{-2n\epsilon^t}$$

Donde $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Entonces, ya tenemos definida la media muestral $\hat{p}_n$ como $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Y se nos pide demostrar que el intervalo $C_n = (\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n)$ con $\epsilon_n = \sqrt{\frac{1}{2n} log(\frac{2}{\alpha})}$ es un intervalo de confianza para p.
Queremos que:

$$P(p \in C_n) \geq 1 - \alpha$$

.

Entonces, para nuestro caso, las variables X_i *Bernoulli*(p) toman valores en [0, 1], por lo que podemos aplicar la desigualdad de Hoeffding con $a_i = 0$ y $b_i = 1$ (son los extremos). Entonces, nuestra desigualdad se reduce a encontrar:

$$P(|\hat{p}_n - p| \geq \epsilon_n) \leq 2e^{(-2n\epsilon_n^2)}$$

Ahora, con $\epsilon_n = \sqrt{\frac{1}{2n}log(\frac{2}{\alpha})}$, la sección derecha de la desigualdad se ve como:

$$\begin{aligned} 2e^{(-2n\epsilon_n^2)} &= 2e^{(-2n)(\frac{1}{2n}log(\frac{2}{\alpha}))} \\ &= 2e^{(-log(\frac{2}{\alpha}))} \\ &= 2e^{(log(\frac{\alpha}{2}))} \\ &= 2\frac{\alpha}{2} = \alpha \end{aligned}$$

Por lo tanto:

$$P(|\hat{p}_n - p| \geq \epsilon_n) = P(p \notin C_n) \leq \alpha$$

Entonces:

$$P(p \in C_n) \geq 1 - \alpha$$

Esto último ya que $P(p \in C_n) + P(p \notin C_n) = 1$.

Parte II

Sea $\alpha = 0.05$ y $p = 0.4$. Mediante simulaciones, realice un estudio para ver qué tan a menudo el intervalo de confianza contiene a p (la cobertura). Haga esto para $n = 10, 50, 100, 250, 500, 1000, 2500, 5000, 10000$. Grafique la cobertura vs. n.

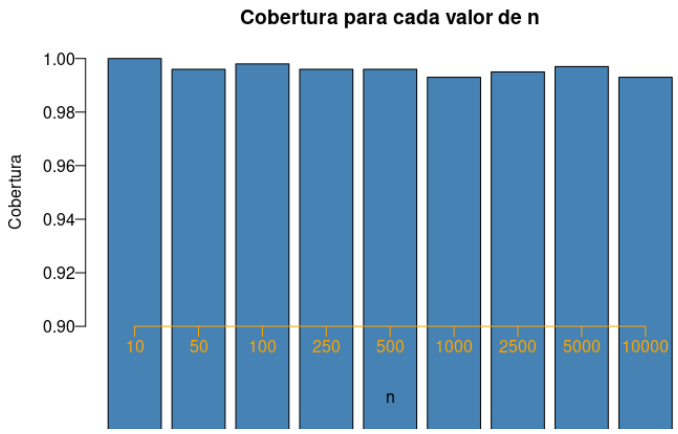
Este problema nos pide evaluar la cobertura del intervalo de confianza para la proporción de $p = 0.4$, utilizando un nivel de insignificancia $\alpha = 0.05$. Esto implica que, en teoría, el intervalo de confianza debería contener el valor real de p en aproximadamente el 95 % de las simulaciones.

Los resultados mostrados en la tabla siguiente demuestran que el intervalo de confianza incluye a p con una gran frecuencia, incluso para tamaños muy pequeños de muestreo. Sin embargo, a medida que n va creciendo, se pueden notar algunas fluctuaciones en la cobertura, pequeños, pero detectables. Esto puede deberse a que se trata de simulaciones, pero se necesitaría un análisis más profundo para determinarlo.

n	**Cobertura**
10	1.000
50	0.997
100	0.995
250	0.993
500	0.995
1000	0.992
2500	0.995
5000	0.997
10000	0.994

Cuadro 1: Cobertura para diferentes valores de n

El gráfico asociado a la tabla es el siguiente.



Parte III

Graficar la longitud del intervalo contra n. Suponga que deseamos que la longitud del intervalo sea menor que 0.05. ¿Qué tan grande debe ser n?

Encontramos a través de la simulación que el valor mínimo de m necesario para que la longitud del intervalo sea menor que 0.05 es de 5000.

Además, construimos la siguiente tabla con la longitud de los intervalos para diferentes valores de n.

n	**Longitud del Intervalo**
10	0.85893882
50	0.38412912
100	0.27162030
250	0.17178776
500	0.12147229
1000	0.08589388
2500	0.05432406
5000	0.03841291
10000	0.02716203

Cuadro 2: Longitud del intervalo para diferentes valores de n

El gráfico asociado a esta información es el siguiente.

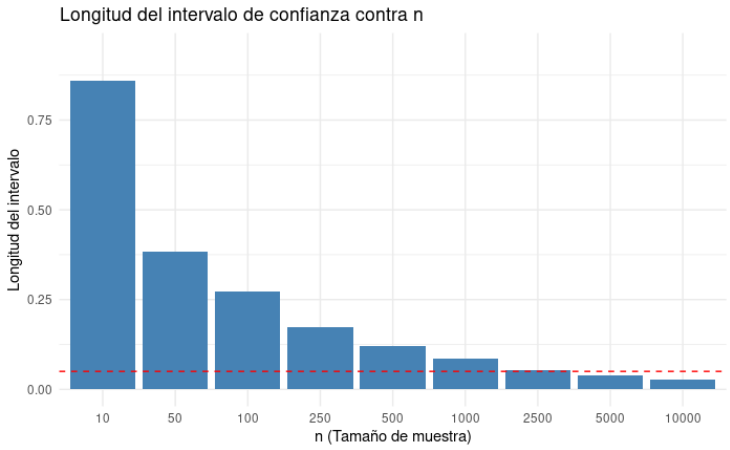


Figura 1: Caption

Podemos obtener algunas conclusiones de estos resultados. Para empezar, que el intervalo de confianza disminuye de manera significativa a medida que aumenta el tamaño de la muestra de n. Este comportamiento es de esperarse, pues a mayor número de observaciones, la estimación de la proporción debe ser más precisa. Esto resulta en

un intervalo de confianza más estrecho.

Asimismo, según la condición de un intervalo de longitud menor que 0.05, podemos ver que cuando $n = 5000$, la longitud del intervalo se encuentra debajo del umbral, alcanzando solo 0.03841291. Por lo tanto, será necesario una muestra de 5000 para cumplir con la condición.

Problema 3

Considera el problema 05 de la tarea 03. Usando la desigualad de Dvoretzky-Kiefer-Wolfowitz, escriba una función en R que calcule y grafique una región de confianza par ala función de distribución empírica. La función debe tomar como parámetros al conjunto de datos que se usan para construir la función de distribución empírica.

Análisis del Problema

En primer lugar, las bandas de confianza proporcionan un intervalo en el que es probable que caigan los puntos del gráfico Q-Q, dado un nivel de confianza $1 - \alpha$. La banda de confianza se construye de la siguiente manera:

$$C(n, \alpha) = \pm \frac{c_\alpha}{\sqrt{n}}$$

Donde n es el tamaño de la muestra, y c_α es un valor crítico para el nivel de confianza deseado. El valor crítico depende del nivel de confianza $1 - \alpha$. Existen tablas definidas para visualizar fácilmente distintos niveles de confianza comúnmente utilizados, pero este puede calcularse con la aproximación de Dvoretzky-Kiefer-Wolfowitz:

$$c_\alpha = \sqrt{\frac{1}{2n} ln\Big(\frac{2}{\alpha}\Big)}$$

La desigualdad de Dvoretzky-Kiefer-Wolfowitz establece que con probabilidad $1 - \alpha$, la distancia máxima entre la FDE $F_n(x)$ y la verdadera función de distribución $F(x)$ está acotada por:

$$\sup_x |F_n(x) - F(x)| \leq \epsilon$$

donde $\epsilon = \sqrt{\frac{\log(2/\alpha)}{2n}}$, con n el tamaño de la muestra y α el nivel de significancia. La región de confianza para $F_n(x)$ se define entonces por los límites superior e inferior dados por $F_n(x) + \epsilon$ y $F_n(x) - \epsilon$.

En este contexto, la desigualdad de Hoeffding también es útil aquí, y se relaciona con este análisis, ya que proporciona límites superiores para las desviaciones de la función de distribución empírica respecto a la verdadera función de distribución.

Para los valores $\alpha = 0.05 \Rightarrow c_\alpha = 1.358$, y para $\alpha = 0.01 \Rightarrow c_\alpha = 1.627$. Con esta información, podemos agregar las bandas de confianza al gráfico Q-Q.

La aproximación de Dvoretzky-Kiefer-Wolfowitz es fundamental para calcular regiones de confianza en este contexto, permitiendo visualizar los intervalos donde es probable que se encuentre la verdadera función de distribución.

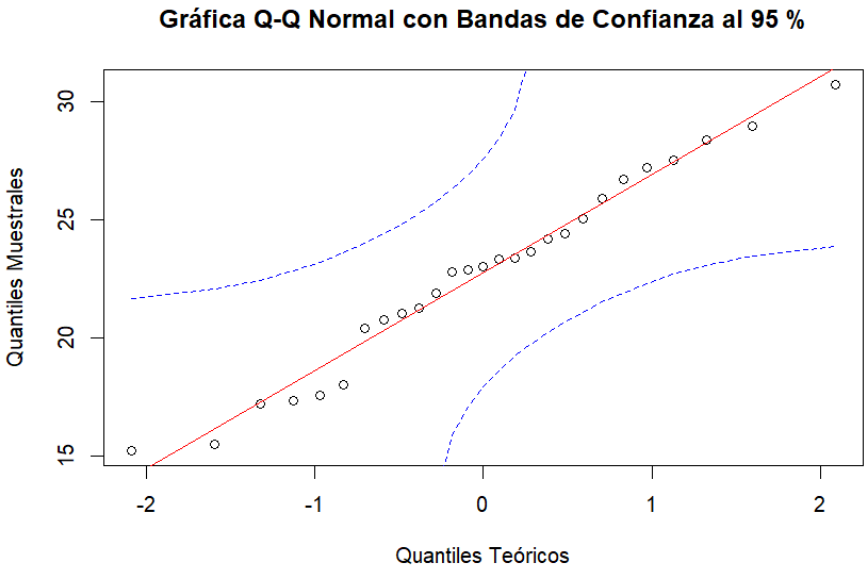


Figura 2: Bandas de confianza para $\alpha = 0.05$

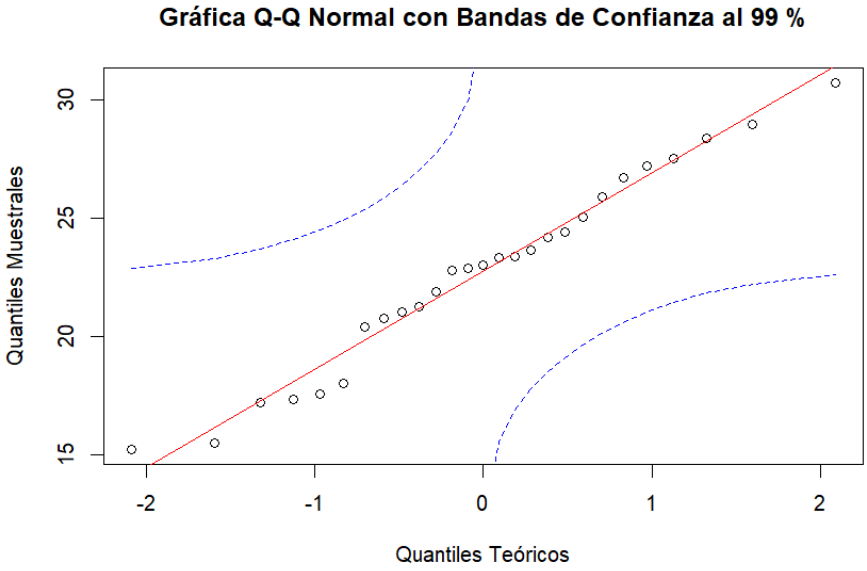


Figura 3: Bandas de confianza para $\alpha = 0.01$

Problema 4

En este ejercicio, se revisará la estimación de densidades.

a) Escriba una función en R que estime una densidad por el método de kernels. La función deberá recibir al punto x donde se evalúa el estimador, al parámetro de suavidad h, al kernel que se utilizará en la estimación y al conjunto de datos.

b) Cargue en R el archivo `"Tratamiento.csv"`, el cual contiene la duración de los períodos de tratamiento (en días) de los pacientes de control en un estudio de suicidio. Utilice la función del inciso anterior para estimar la densidad del conjunto de datos para $h = 20, 30, 60$. Grafique las densidades estimadas. ¿Cuál es el mejor valor para h? Argumente su respuesta.

c) En el contexto de la estimación de densidades, escriba una función en R que determine el ancho de banda que optimiza el Integrado del Error Cuadrático (ISE). Grafique la densidad con el ancho de banda óptimo para el conjunto de datos de `"Tratamiento.csv"`.

Análisis del Problema

La estimación de densidades es un método utilizado en estadística para estimar la función de densidad de probabilidad de una variable aleatoria. En este documento, se abordará la estimación de densidades por el método de kernels, que permite suavizar los datos y proporcionar una estimación no paramétrica de la función de densidad de probabilidad.

Estimación de Densidad por el Método de Kernels

El objetivo de esta sección es estimar la densidad de una variable continua utilizando el método de kernels. Este método se basa en la siguiente fórmula:

La estimación de la densidad de una variable aleatoria X en un punto x se calcula mediante la siguiente expresión:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

donde:

- $\hat{f}(x)$ es la densidad estimada en el punto x.
- n es el número de observaciones.
- h es el ancho de banda (parámetro de suavidad).
- K es el kernel utilizado.
- X_i son los datos observados.

Los kernels más comunes son:

- Kernel Gaussiano:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{2}$$

- Kernel Uniforme:

$$K(u) = \begin{cases} \frac{1}{2} & \text{si } |u| \leq 1 \\ 0 & \text{en otro caso} \end{cases} \tag{3}$$

- Kernel Epanechnikov:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{si } |u| \leq 1 \\ 0 & \text{en otro caso} \end{cases} \tag{4}$$

Para la segunda parte del problema cargaremos los datos desde el archivo `Tratamiento.csv` y utilizaremos la función de estimación de densidad para calcular y graficar la densidad para diferentes valores de h.

Se graficarán las densidades estimadas para diferentes valores de h (20, 30, 60).

Por último, el ancho de banda que optimiza el Integrado del Error Cuadrático (ISE) puede determinarse mediante la siguiente fórmula:

$$h_{opt} = \left(\frac{4}{d+2}\right)^{\frac{1}{5}} \hat{\sigma} n^{-\frac{1}{5}} \tag{5}$$

donde:

- d es la dimensión de los datos (en este caso, $d = 1$).
- $\hat{\sigma}$ es una estimación de la desviación estándar de los datos.
- n es el número de observaciones.

Una vez calculado el ancho de banda óptimo, se puede graficar la densidad estimada.

Resultados e intepretación

b)

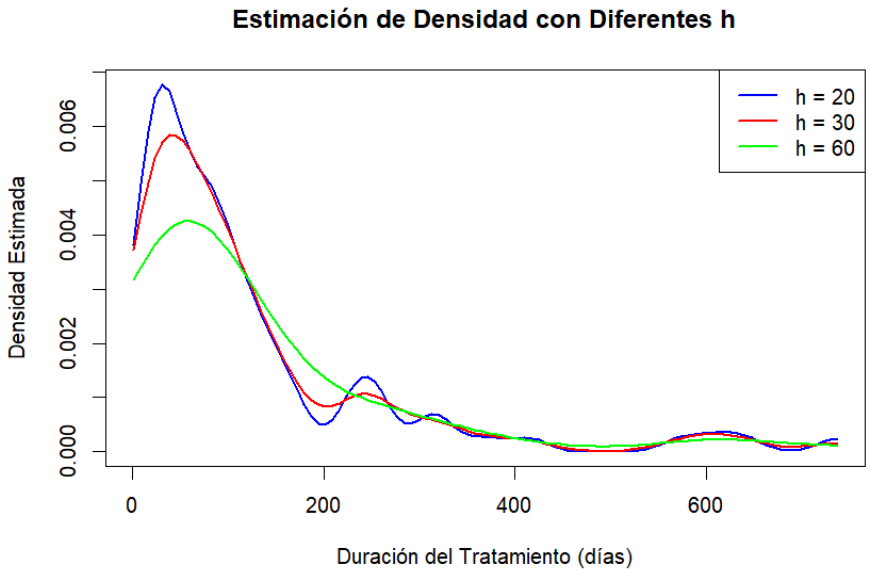


Figura 4: Estimación de densidad con diferentes anchos de banda

Basándonos gráfica, podemosinterpretar lo siguiente:

- $h = 20$ (línea azul): Este es el ancho de banda menor de los tres valores y podemos ver que muestra más detalles en la densidad estimada. También pueden observarse picos y variaciones significativas, lo que indica que existen fluctuaciones locales en los datos.
- $h = 30$ (línea roja): En este valor de ancho de banda se suavizan los detalles. Dicho valor de h parece capturar bien la estructura de los datos y puede ser buena opción para tener un balance entre capturar detalles finos y aplicar una suavización adecuada.
- $h = 60$ (línea verde): Este es el valor de ancho de banda más grande y es el que genera la densidad más suavizada. Esto resulta útil para observar las tendencias generales, auqnue podríamos estar perdiendo detalles importantes en la estructura de los datos.

El valor óptimo de h depende del equilibrio entre capturar detalles importantes y suavizar adecuadamente la curva de densidad. En este caso, el valor $h = 30$ parece ser una opción más adecuada, ya que captura las características principales de la distribución sin agregar demasiado ruido, como ocurre con $h = 20$, ni perder detalles, como con $h = 60$.

c)

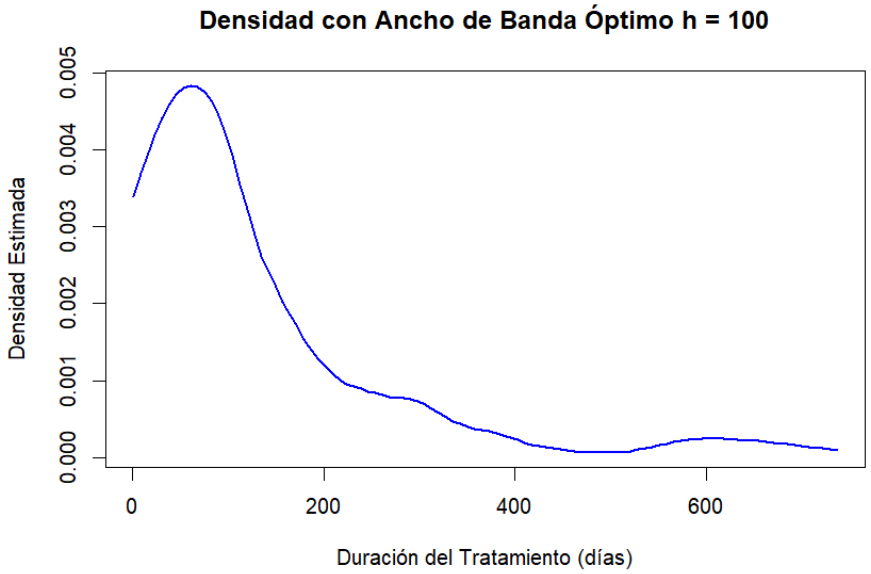


Figura 5: Densidad con el ancho de banda que optimiza el ISE

La Figura 5 muestra la densidad estimada de la duración de los períodos de tratamiento, utilizando el ancho de banda óptimo $h = 100$ obtenido al minimizar el Integrado del Error Cuadrático (ISE).

Podemos observar que existe un pico pronunciado en la densidad para valores cercanos a 0 días, lo cual sugiere que una cantidad significativa de pacientes en el grupo de control tuvo un tratamiento de corta duración. Esto podría indicar que muchos de los tratamientos en el estudio duraron poco tiempo.

Después de este pico, la densidad decrece rápidamente a medida que aumenta la duración del tratamiento. Esto implica que es menos común que los pacientes tengan tratamientos de larga duración, con una frecuencia que disminuye a medida que el tiempo se extiende. Para duraciones superiores a los 200 días, la densidad permanece baja y estable, sugiriendo que solo una pequeña proporción de pacientes mantuvo el tratamiento durante periodos prolongados.

Dicha gráfica sugiere que el ancho de banda $h = 100$ permite una estimación de la densidad que suaviza adecuadamente la curva, evitando el sobreajuste, lo cual sería evidente en curvas con picos irregulares si el valor de h fuera menor.

Problema 5

Considera el siguiente experimento en dos etapas: primero, se escoge un punto X con distribución uniforme (0,1); después se elige un punto Y con distribución uniforme $(-X, X)$. El vector aleatorio (X, Y) representa el resultado del experimento. ¿Cuál es la densidad condicional de X dada Y?

Sabemos que $X\ Unif(0, 1)$, por lo que cumple con que $0 < x < 1$. Ahora, Y también se distribuye de manera uniforme, pero sobre $(-X, X)$, por lo que tenemos algo de la forma: $-x < y < x$. Entonces, el "soporte" de la densidad de probabilidad conjunta de $f(x, y)$ está definido para $0 < x < 1$ y $-x < y < x$.

Con ello, sabemos que $Y\ Unif(-x, x)$, y cuando tenemos una variable aleatoria Y con distribución uniforme sobre un intervalo $[a, b]$, la densidad de probabilidad es constante, definida como:

$$f(y) = \frac{1}{b-a} \quad para \ \ a < y < b$$

De ese modo, podemos construir $f_{X|Y}(y|x)$:= la densidad de Y dado $X = x$. Por lo tanto:

$$f_{X|Y}(y|x) = \frac{1}{x-(-x)} = \frac{1}{2x} \quad para \quad -x < y < x$$

Ahora bien, la densidad marginal de $X\ Unif(0, 1)$ será $f_x(x) = 1$ para $0 < x < 1$. Por lo tanto:

$$f(x, y) = f_{Y|X}(Y|X) \cdot f_x(x) = \frac{1}{2x} \cdot 1 = \frac{1}{2x}$$

Lo anterior, para $0 < x < 1$ y $-x < y < x$. Para cualquier otro par (x, y), se debe cumplir con que $f(x, y) = 0$. Entonces, finalmente, la densidad de $f(x, y)$ será:

$$f(x, y) = \begin{cases} \frac{1}{2x} & 0 < x < 1 \ \ \& \ \ -x < y < x \\ 0 & cualquier\ otro\ caso \end{cases}$$

Podemos demostrar que nuestra $f(x, y)$ construida es, en efecto, una densidad de probabilidad válida si confirmamos que:

(a) $f(x, y) \geq 0 \forall (x, y)$

(b) La integral de $f(x, y)$ es igual a 1

Entonces, para **(a)**: basta con recurrir a que $0 < x < 1$ y que $-x < y < x$ para confirmar que $f(x, y) > 0$. Por lo tanto, se cumple **(a)**.

Para **(b)**: tenemos que se debe cumplir que:

$$\int_0^1 \int_{-x}^{x} f(x, y)\, dydx = 1$$

Por lo tanto, tendremos que desarrollar:

$$I = \int_0^1 \int_{-x}^{x} f(x,y)\, dydx = \int_0^1 \left(\int_{-x}^{x} \frac{1}{2x}\, dy \right) dx = \int_0^1 \frac{1}{2x} \left(\int_{-x}^{x} dy \right)\, dx$$

$$I = \int_0^1 \frac{1}{2x} \left(\int_{-x}^{x} dy \right)\, dx = \int_0^1 \frac{1}{2x} \Big(y\Big)_{-x}^{x}\, dx = \int_0^1 \frac{1}{2x} \cdot (2x)\, dx = \int_0^1 dx = 1$$

Por lo tanto, se cumple con que la integral de $f(x,y) = 1$ y podemos decir que la función propuesta es una densidad válida.

Ahora, se nos pide obtener la probabilidad marginal de Y. Esto es: $f_y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$.

Como $f(x,y) = \frac{1}{2x}$ para $0 < x < 1$ y $-x < y < x$. Además, sabemos que $-x < y < x \Leftrightarrow |y| < x$, y, dadas las condiciones, se debe cumplir con que $|y| < x < 1$. Dicho rango será el que defina los límites de nuestra integral para la probabilidad marginal de Y, i.e., $f_y(y)$. Entonces:

$$f_y(y) = \int_{|y|}^{1} \frac{1}{2x}\, dx = \frac{1}{2} \int_{|y|}^{1} \frac{dx}{x} = \frac{1}{2} ln(x) \Big|_{|Y|}^{1} = \frac{1}{2} \Big(ln(1) - ln(|y|) \Big)$$

$$\Rightarrow f_y(y) = -\frac{1}{2} ln(|y|) \quad para \quad -1 < y < 1$$

Fuera del intervalo, $f_y(y) = 0$.

Finalmente, se pide encontrar a la probabilidad condicional de X dado Y. Es decir, queremos ahora a $f_{X|Y}(X|Y)$. Ya sabemos que $PX = x|Y = y = f_{X|Y}(X|Y) = \frac{f(x,y)}{f_y(y)}$. Esto es:

$$f_{X|Y}(X|Y) = \frac{\frac{1}{2x}}{-\frac{ln|y|}{2}} = -\frac{2}{2xln|y|} = -\frac{1}{xln|y|} \quad para \quad |y| < x < 1$$

Fuera del intervalo, $f_{X|Y}(X|Y) = 0$

Problema 6

Cargue en R al conjunto de datos "Maíz.csv", el cual contiene el precio mensual de la tonelada de maíz y el precio de la tonelada de tortillas en USD. En este ejercicio tendrá que estimar los coeficientes de una regresión lineal simple.

a) **Calcule de forma explícita la estimación de los coeficientes vía mínimos cuadrados, y ajuste la regresión correspondiente. Concluya.**

Después de realizar un script para conseguir realizar por mínimos cuadrados un ajuste lineal, se encontró que la pendiente de la recta es de 0.46 unidades. La gráfica encontrada es la siguiente.

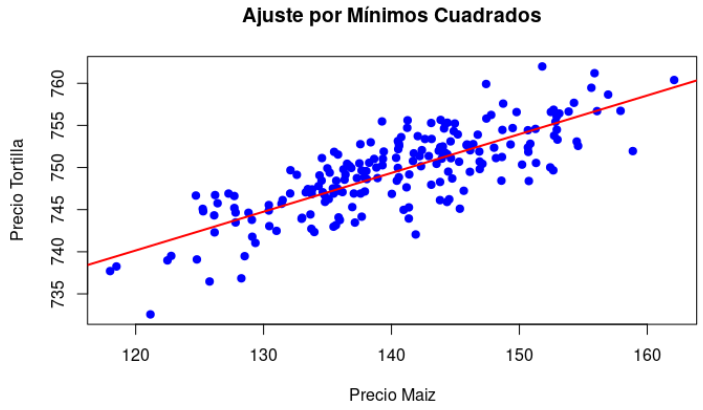


Figura 6: Caption

Nuestro código calcula los coeficientes de una regresión lineal simple, para modelar la relación entre el precio del maíz (x) y el precio de la tortilla (y). La idea es encontrar una recta tal que esta se ajuste de la mejor manera a los datos observados y registrados en el conjunto "Maiz.csc". De tal forma que podemos encontrar una recta que se ajuste a la tendencia de los datos, de la forma:

$$y = \beta_0 + \beta_1 x$$

donde β_0 es la pendiente de la recta (para este caso, $\beta_0 = 0.4600343$), que indica la tasa de cambio de y respecto a la variable x (es decir, del precio de la tortilla, respecto al del maíz). Por otra parte, β_1 es la intersección en el eje Y, que indica el valor de y cuando $x = 0$ (en nuestro caso, y = 684.9545 cuando x = 0).

Tanto por los valores encontrados, como por lo que se observa en la gráfica, es claro que la tendencia es creciente. Es decir, a mayor precio del maíz, mayor precio de la tortilla. La tendencia tiene sentido, tanto teóricamente, como en la práctica. Entonces, podemos confiar en que hay suficientes pruebas como para concluir un ajuste de tendencia lineal es bueno para este caso.

b) **Calcule de forma explícita la estimación de los coeficientes vía regresión no-paramétrica tipo kernel. (ver Nadaraya, E. A. (1964). "On Estimating Regression". Theory of Probability and its pplications. 9 (1): 141–2. doi:10.1137/1109020) y ajuste la regresi´on correspondiente. Concluya.**

Para abordar este problema, lo primero que debemos hacer es explicar un poco respecto al modelo de Nadaraya-Watson.

En el artículo: *"On Estimating Regressions"*, de E. A. Nadaraya, se muestra la estimación de la regresión cuando los parámetros que tenemos no siguen una forma analítica conocida. Es decir, cuando no podemos utilizar técnicas como los mínimos cuadrados para ajustar un modelo lineal simple. El análisis propone una alternativa estadística para estimar curvas de regresión a partir de los datos empíricos, utilizando una función de densidad, denominada: la función de núcleo $K(x)$, también conocida como Kernel.

El estimador de Nadaraya-Watson propone una ecuación para la estimación de la curva de regresión, la cual es la siguiente:

$$m_h(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}$$

Como se mencionó, $K(x)$ es la función Kernel que determina la "suavización" del estimador. Por otro lado, h es el parámetro de banda, el cual controla dicho nivel de suavidad. Este estimador está destinado a ponderar los valores medidos de Y_i, según qué tan cercano está de x, acorde a la medida del Kernel.

En el artículo, podemos leer como, bajo ciertas condiciones, el estimador es bastante consistente. Es decir, a medida que el tamaño de la muestra crece, el estimador converge a la verdadera curva de regresión. Tal propiedad vuelve al estimador de Nadaraya-Watson en uno bastante sólido. De igual forma, establece que, bajo ciertas condiciones, el estimador es asintóticamente normal. Esto significa que se distribuye de manera normal en el límite cuando el tamaño de la muestra tiende a infinito.

Podemos decir que el estimador es útil para ajustar modelos cuando la relación entre las variables no parece del todo lineal, aunque, en este caso y por simplicidad, se está aplicando a un modelo que sí apunta a tener una tendencia lineal. Sin embargo, cuando queremos obtener patrones más complejos, el ajuste de Nadaraya-Watson resulta ser muy buena elección.

En un primer intento, se aplicó el Kernel Gaussiano para realizar el ajuste. Contrario al segundo intento, no se acomodaron los datos antes de realizar la estimación. Cada intento utilizó distintas bandas h, para notar como, a diferentes valores, el estimador se ajusta mejor o peor a los datos.

Prueba 01: datos sin ordenar para el estimador

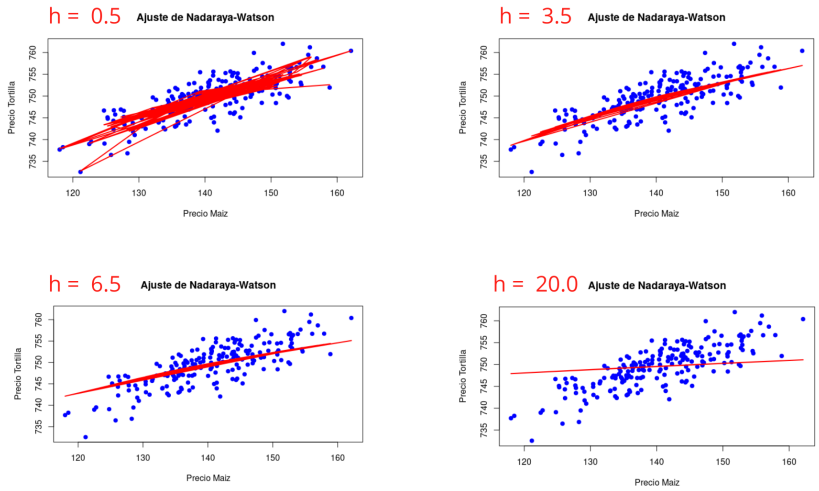


Figura 7: Caption

Prueba 02: datos ordenados para el estimador

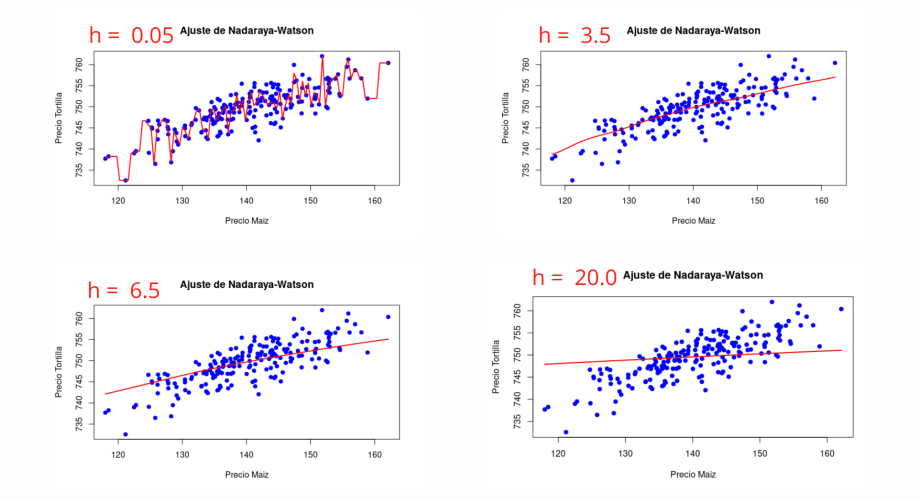


Figura 8: Caption

c) **Compare ambos resultados. ¿Qué diferencias observa?**

A través de ambas pruebas, podemos ver que la forma de ordenar los datos parece ser una muy buena idea para mejorar la suavidad de la tendencia de ajuste, sobre todo para h pequeñas. Observando el resultado para $h = 3.5$ en ambos paneles, se nota cómo al introducir de manera ordenada los datos al ajuste, el estiamdor se va ajustando más suavemente a la tendencia lineal del conjunto.

También, podemos notar como, con un h pequeña, existe sobre-ajuste (comunmente conocido como *overfitting*), en el cual el estimador intenta ajustarse "a la fuerza" a los datos. El *overfitting* genera esa curva curiosa, en la cual se recorre uno a uno los datos. Por otro lado, con $h = 20$, se presenta *underfitting*, el caso opuesto al anterior. Esto significa que el estimador no se acerca para nada a la tendencia de los datos, en este caso, quedándose como una línea casi horizontal.

Por otra parte, al comparar el ajuste con Kernel Gaussiano, comparado con el ajuste por mínimos cuadrados, son bastante parecidos cuando h oscila alrededor de 6 unidades. Sin embargo, presenta una ligera concavidad alrededor del punto $(x = 140, y = 750)$. Esto parecería indicar que la tendencia de los datos no es totalmente lineal. Pese a todo, no tenemos suficientes muestras como para determinar si, en un caso límite (con muchos más datos), se revele la concavidad de la estimación Nadaraya-Watson. Esto último no parece tener oportunidad de suceder, pues el ajsute lineal parece ser la mejor opción para los datos del precio del maiz-tortilla. Quizás para otro fenómeno, de una distinta naturaleza, esto sí tendría sentido.

A pesar de todo, creo que ambos ajustes se acercan bastante a una estimación adecuada.