BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

POSGRADO EN CIENCIAS MATEMÁTICAS

# SIMULACIÓN DE DATOS FALTANTES POR MEDIO DE ESTIMACIONES POR MÉTODO KERNEL

Tesis

PARA OBTENER EL GRADO DE:

**MAESTRO EN CIENCIAS MATEMÁTICAS**

PRESENTA:

LIC. JUAN ANTONIO VAZQUEZ MORALES

DIRECTORES DE TESIS:

DRA. HORTENSIA JOSEFINA REYES CERVANTES
DR. BULMARO JUÁREZ HERNÁNDEZ

PUEBLA, PUEBLA. 2020

**DRA. LIDIA AURORA HERNÁNDEZ REBOLLAR**
**SECRETARIA DE INVESTIGACIÓN Y**
**ESTUDIOS DE POSGRADO, FCFM-BUAP**
**P R E S E N T E:**

Por este medio le informo que el C:

JUAN ANTONIO VÁZQUEZ MORALES

estudiante de la Maestría en Ciencias (Matemáticas), ha cumplido con las indicaciones que el Jurado le señaló en el Coloquio que se realizó el día 7 de octubre de 2020, con la tesis titulada:

*Simulación de Datos Faltantes por medio de estimaciones por método Kernel*

Por lo que se le autoriza a proceder con los trámites y realizar el examen de grado en la fecha que se le asigne.

A T E N T A M E N T E.
H. Puebla de Z. a 22 de octubre de 2020

***DRA. PATRICIA DOMÍNGUEZ SOTO***
**COORDINADORA DEL POSGRADO**
**EN MATEMÁTICAS.**

Facultad
de Ciencias
Físico Matemáticas

Av. San Claudio y 18 Sur, edif. FM1
Ciudad Universitaria, Col. San
Manuel, Puebla, Pue. C.P. 72570
01 (222) 229 55 00 Ext. 7550 y 7552

*A la memoria del Dr. Iván Hernández Orzuna.*

*A mi familia, amigos y profesores.*

# Agradecimientos

Agradezco a mis padres, Isela y Constantino, y a mi abuelita Concepción por su apoyo moral para alcanzar este objetivo. A mis hermanas Rocio y Lucero. A mi sobrino Julian que me hace reir y hacerme recordar que uno siempre está aprendiendo.

A mis amigos y compañeros de la facultad, América, David, Julio, Lizbeth, Patricia, Roque, Silvia, Solehyr y Yasmín por su ayuda, comprensión y compañía.

A mis asesores de tesis, la Dra. Hortensia Josefina Reyes Cervantes y Dr. Bulmaro Juárez Hernández, gracias por su apoyo.

A mis sinodales, Dra. María de Lourdes Sandoval Solís, Dra. Gladys Linares Fleites, Dr. Hugo Adán Cruz Suárez y Dr. Francisco Solano Tajonar Sanabria por haber aceptado revisar este trabajo y aportar su conocimiento para la mejora del mismo.

Al Consejo de Ciencia y Tecnología (CONACYT) y la Benemérita universidad Autónoma de Puebla por el apoyo brindado.

Todas las personas mayores fueron al principio niños.
(Aunque pocas de ellas lo recuerdan)

# Introducción

En los últimos años, la contaminación ambiental provocada por mano del ser humano ha sido un tema de gran estudio, en especial los efectos en la salud. En la contaminación fotoquímica[1] el ozono es considerado principalmente el componente más tóxico de la mezcla. Los estudios han revelado que las altas concentraciones de ozono, tienen efecto relacionados con el sistema respiratorio, como la disminución de la función pulmonar o agravamiento del asma como se menciona en [2], por lo que su estudio es importante.

En el caso del monitoreo de los niveles de ozono, tanto como en otras áreas del conocimiento, existe pérdida de datos por muy diversas situaciones que el investigador no puede controlar, como son, falta de apoyo económico, los individuos han muerto o se cambiaron de sitio geográfico, desperfectos en los dispositivos de medición, etc. La pérdida de información existe por diversas situaciones que algunas veces el investigador no puede controlar y esto puede afectar la efectividad de las estimaciones.

Los primeros casos heurísticos para trabajar los datos faltantes, como el análisis de datos completos, o la imputación mediante la media

[1] Contaminación principalmente procedente de las reacciones de los hidrocarburos y los óxidos de nitrógeno, estimuladas por la luz solar intensa y el incremento de la temperatura.

(véase [5]) tienen algunos inconvenientes. Para el primer método, sus consecuencias radican en la cantidad de información que se pierda al descartar los casos faltantes, lo que se traduce en sesgo y falta de precisión en las estimaciones, en especial en el caso que los datos faltantes no sean aleatorios. Para el segundo método, que consiste en sustituir los datos faltantes por la media de los datos completos, tiende a subestimar el sesgo de los datos. También hay otros métodos de imputación simple o basados en la verosimilitud (véase [5]), sin embargo no existe un método definitivo que en lo general de buenos resultados.

Por otro lado, el método kernel utiliza un conjunto de datos provenientes de una función de distribución continua, univariada y desconocida, para aproximar a esta función. Los kerneles son funciones que se asocian a cada uno de los datos, y así, la suma ponderada de estas funciones es una aproximación a la función de densidad desconocida, como se explica en [16]. Este método ya es aceptado por la comunidad científica, y las investigaciones actuales se están enfocando a la elección del ancho de banda, el cual es un parámetro de importancia, por lo que, en el capítulo 1, se explica uno de los métodos clásicos para tal propósito.

Por esta razón, se desarrolla una metodología que use las estimaciones por método kernel para completar los datos faltantes, de tal manera, que concuerden lo mejor posible con los datos observados. En el capítulo 2 se explican diversos experimentos, el primero, sin tomar en cuenta como actúa el ancho de banda a la simulación de los datos faltantes, que de acuerdo a los gráficos realizados, se llega a la conclusión de que la metodología planteada es útil para simular los datos faltantes. Después, se realiza un segundo experimento, y con ayuda de una prueba de hipótesis, se decide que los resultados obtenidos son satisfactorios. Ya que la metodología es factible, se formulan dos algoritmos que resumen

la metodología para simular los datos faltantes.

Después se utilizan los datos de los niveles de ozono en la estación del Pedregal, situada en la ciudad de México. Por la naturaleza de los datos, es claramente que estos no son muestras independientes, pero se puede pensar, que los máximos cada 2 semanas si lo son, además que los niveles máximos son de interés, ya que como se dijo al inicio, sus altos niveles están relacionados con el agravamiento de problemas respiratorios, los resultados que se tienen se ven en el capítulo 3.

Hay que señalar que el trabajo es en esencia construir programas en un lenguaje de alto nivel para realizar experimentos controlados o simulaciones, por lo que se trabaja con ciertas distribuciones, las cuales tienen distintas características, con idea de analizar diferentes escenarios, ya que se quiere un método general. Al final del trabajo, se tiene la información para que los Algoritmos 2.1 y 2.2 se agreguen a los métodos para trabajar con datos faltantes, objetivo que durante la experimentación y el ejemplo de aplicación dan buenos resultados.

Para la mejor comprensión de este escrito, se necesita conocimientos de probabilidad, estadística, análisis numérico, teoría de la medida y programación, por lo que al final se incluye el Apéndice A, donde se enuncias definiciones y resultados necesarios para el trabajo desarrollado, aunque si ya se tiene conocimientos en estas áreas, puede omitir su lectura.

# Índice general

# Capítulo 1

# Método kernel

En probabilidad se trabaja con distribuciones de variables aleatorias continuas, las cuales tienen una función de densidad $f(x)$.

Si $f(x)$ no coincide con un modelo conocido, entonces el problema radica en encontrar la función de densidad de la variable $X$ a partir de una muestra $X_1, \ldots, X_n$, independientes e idénticamente distribuidas. Para tal objetivo, el método de kernel, utiliza un conjunto de datos $x_1, \ldots, x_n$ de una distribución continua(univariada) y desconocida para aproximar a la función de densidad.

## 1.1. Construcción de los kerneles

El primer objetivo es dar la aproximación por método kernel, a la cual se denotará por $\hat{f}(x)$. Para este propósito, se inicia por definir que es un kernel (véase [16]).

**Definición 1.1** (Kernel)**.** *Un kernel es una función* $K : \mathbb{R} \to \mathbb{R}$ *que cumple:*

*a)* $K(x) \in [0, \infty)$, $x \in [-1, 1]$,

*b)* $K(x) = 0$, $x \notin [-1, 1]$,

*c)* $K(x) = K(-x)$,

*d)* $\int_{-1}^{1} K(x)dx = 1$,

*e)* $\int_{-1}^{1} xK(x)dx = 0$ *y*

*f)* $\int_{-1}^{1} x^2K(x)dx > 0$.

De la Definición 1.1, las propiedades a), b), y d) dan como resultado que $K$ es una función de densidad, la propiedad c) es la definición de función par, es decir, que $K$ es simétrica respecto al eje $x$, la propiedad e) dice que la media de la variable asociada a $K$ es cero, y por último, la propiedad f) es la del segundo momento de la variable aleatoria asociada a $K$, la cual es finita.

## Parametrización de kerneles

En la Definición 1.1, $K$ es no negativa en el intervalo $[-1, 1]$, pero esto puede modificarse mediante un parámetro.

Sea $h > 0$, el kernel parametrizado en $h$ es

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right), \quad x \in [-h, h].$$

Esta modificación mantiene las propiedades *a)* a *f)* de la Definición 1.1, pero ahora sobre el intervalo $[-h, h]$, como se muestra a continuación.

**Nota:** A $h$ se le conoce como el **ancho de banda** de $K$.

**Lema 1.1.** *Si $K$ es un kernel y $h > 0$, entonces para*

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right), \quad x \in [-h, h],$$

*se tiene que,*

*a)* $K_h(x) \in [0, \infty)$, $x \in [-h, h]$,

*b)* $K_h(x) = 0$, $x \notin [-h, h]$,

*c)* $K_h(x) = K_h(-x)$,

*d)* $\int_{-h}^{h} K_h(x)dx = 1$,

*e)* $\int_{-h}^{h} xK_h(x)dx = 0$ *y*

*f)* $\int_{-h}^{h} x^2K_h(x)dx > 0$.

*Demostración.* Sea $h > 0$ y $K$ un kernel.

a) Sea $x \in [-h, h]$, entonces $-1 \leq \frac{x}{h} \leq 1$, así, por el inciso a) de la Definición 1.1, se tiene que $K_h(x) \geq 0$.

b) Si $x \notin [-h, h]$, entonces $\frac{x}{h} \notin [-1, 1]$, por lo que $K_h(x) = 0$.

c) Es trivial.

d) Tomando $u = \frac{x}{h}$,

$$\begin{aligned}\int_{-h}^{h} K_h(x)dx &= \int_{-h}^{h} \frac{1}{h} K\left(\frac{x}{h}\right) dx \\ &= \int_{-1}^{1} K(u)du \\ &= 1.\end{aligned}$$

e) Tomando $u = \frac{x}{h}$,

$$\begin{aligned}\int_{-h}^{h} xK_h(x)dx &= \int_{-h}^{h} \frac{1}{h} xK\left(\frac{x}{h}\right) dx \\ &= h\int_{-1}^{1} uK(u)du \\ &= 0.\end{aligned}$$

f) Tomando $u = \frac{x}{h}$, entonces

$$\begin{aligned}\int_{-h}^{h} x^2K_h(x)dx &= \int_{-h}^{h} \frac{1}{h} x^2K\left(\frac{x}{h}\right) dx \\ &= \int_{-1}^{1} (hu)^2K(u)du \\ &= h^2\int_{-1}^{1} u^2K(u)du > 0.\end{aligned}$$

□

## Traslación de kernel

En la Definición 1.1, el kernel esta centrado en 0, pero se puede centrar en cualquier otro punto $t \in \mathbb{R}$. El kernel parametrizado en $h$ y centrado en $t$ esta dado por

$$K_h(x) = \frac{1}{h} K\left(\frac{x - t}{h}\right), \quad x \in [t - h, t + h].$$

Ahora las condiciones de la Definición 1.1 son análogas para un kernel parametrizado en $h$ y centrado en $t$.

**Lema 1.2.** *Sea $K$ un kernel, $t \in \mathbb{R}$ y $h > 0$, entonces para*

$$K_h(x) = \frac{1}{h} K\left(\frac{x - x_i}{h}\right), \quad x \in [x_i - h, x_i + h],$$

*se cumple que*

*a)* $K_h(x) \in [0, \infty)$, $x \in [x_i - h, x_i + h]$,

*b)* $K_h(x) = 0$, $x \notin [x_i - h, x_i + h]$,

*c)* $K_h(x + x_i) = K_h(-x + x_i)$,

*d)* $\int_{x_i - h}^{x_i + h} K_h(x)dx = 1$,

*e)* $\int_{x_i - h}^{x_i + h} xK_h(x)dx = x_i$ *y*

*f)* $\int_{x_i - h}^{x_i + h} x^2 K_h(x)dx > 0$.

*Demostración.* Similar al Lema 1.1. □

## 1.2. Construcción de funciones de densidad de probabilidad

Sea $X$ una variable aleatoria continua. Supóngase la muestra aleatoria $x_1, x_2, \ldots, x_n$. El objetivo es obtener un estimador $\hat{f}(x)$ de la función $f(x)$ a partir de esta muestra. Para tal objetivo se explicará el método de kernel.

### El método kernel

Un kernel no es mas que una función de densidad. Si se coloca un kernel en cada uno de los datos de la muestra, la suma ponderada de estas funciones será una función de densidad. Esta suma es una función continua, que capta la influencia de los datos cercanos.

**Definición 1.2** (Estimación por kernels)**.** *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria e independiente, entonces la* ***estimación por kernel*** *es*

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \tag{1.1}$$

El ancho de banda $h$ es el parámetro de suavizado de $\hat{f}(x)$. Si $h$ es pequeño, más concentrada está la construcción del kernel en cada punto $x_i$. Si $h$ es muy grande, mayor la influencia e interacción del kernel hacia los puntos cercanos.

Cuando $h \to 0$, la contribución de cada kernel estará concentrada en cada punto $x_i$, así $\hat{f}(x)$ sera una función de masa. Si $h \to \infty$, $\hat{f}(x)$ se aplana en un solo cúmulo y con mayor dispersión.

### 1.2.1. Ejemplo del funcionamiento del método kernel

Para dar un ejemplo de como funciona el método kernel, se generan 50 datos de una distribución $Gamma(3, 0.5)$, cuya gráfica de su función de densidad se encuentra en la Figura 1.1.

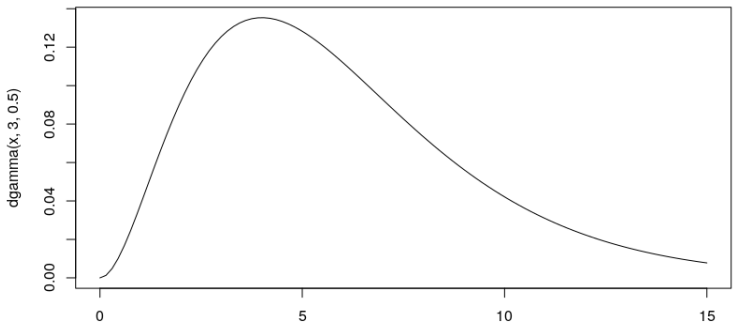


**Figura 1.1:** Función de densidad de una $Gamma(3, 0.5)$

En la Figura 1.2, que corresponde a la utilización de un kernel normal[1], se puede ver que si se aproxima a la forma de la Gamma correspondiente, y se obtiene un buen trabajo para $h = 1.5$ y $h = 2$.

Para la aproximación utilizando el kernel de Epanechnikov[2], la Figura

[1]El kernel normal se define por

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Este kernel es especial, pues su rango son los $\mathbb{R}$.

[2]El kernel Epanechnikov se define por

$$K(x) = \frac{3}{4}(1 - x^2), \qquad x \in [-1, 1].$$

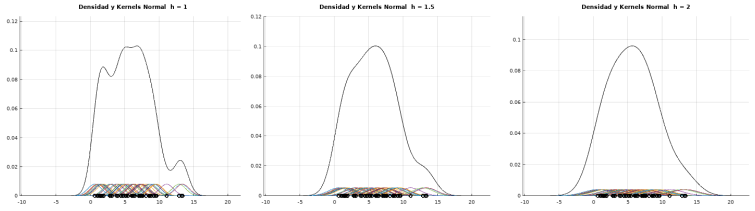


**Figura 1.2:** Estimación con 50 datos de una $Gamma(3, 0.5)$.

1.3 sugiere que debe usarse más datos para una mejor aproximación, aunque se observa que si $h$ aumenta, la aproximación mejora.

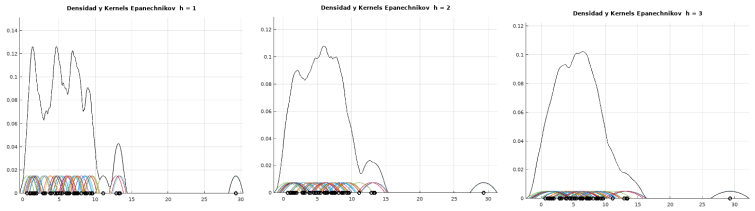


**Figura 1.3:** Estimación con 50 datos de una $Gamma(3, 0.5)$.

Para el kernel Rectangular[3], se observa que para un $h = 2$ la aproximación $\hat{f}(x)$ empieza a tomar la forma de la función de densidad deseada, y que si aumenta, la densidad aproximada se acerca mejor, como se observa en la Figura 1.4.

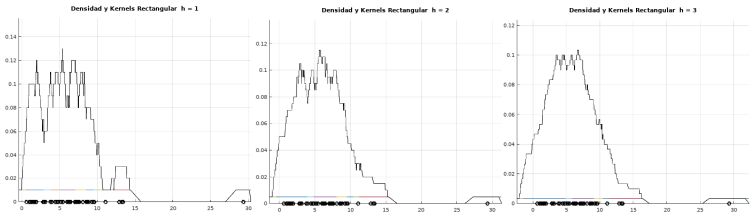


**Figura 1.4:** Estimación con 50 datos de una $Gamma(3, 0.5)$.

Como se observa en cada caso, un kernel da una medida alrededor de cada $x_i$, y $h$ es la longitud donde actúa el kernel, así como se tiene en la ecuación 1.1, la suma ponderada de cada kernel parametrizado y

[3]El kernel rectangular se define por

$$K(x) = \frac{1}{2}, \qquad x \in [-1, 1].$$

centrados en $x_i$ dan como resultado la aproximación por método kernel $\hat{f}(x)$.

## 1.3. Eficiencia del estimador $\hat{f}(x)$ con respecto a $f(x)$

Ahora es importante saber como medir que tan eficiente es la estimación $\hat{f}(x)$, para ello se da la definición siguiente.

**Definición 1.3.** *Se define las medidas siguientes*

- *Sesgo del estimador $\hat{f}(x)$ con respecto a $f(x)$ como*
$$B[\hat{f}(x)] = E[\hat{f}(x)] - f(x).$$
- *Varianza del estimador $\hat{f}(x)$ con respecto a $E[\hat{f}(x)]$ como*
$$V[\hat{f}(x)] = E[\hat{f}(x) - E[\hat{f}(x)]]^2.$$
- *Error cuadrático medio como*
$$ECM[\hat{f}(x)] = E[\hat{f}(x) - f(x)]^2.$$
*Si se desarrolla el cuadrado y de acuerdo a las definiciones anteriores, entonces*
$$ECM[\hat{f}(x)] = B^2[\hat{f}(x)] + V[\hat{f}(x)].$$
- *El error cuadrático medio integrado como*
$$ECMI[\hat{f}(x)] = \int_{\mathbb{R}} ECM[\hat{f}(x)]. \tag{1.2}$$

El siguiente Teorema exhibe que el método kernel efectivamente proporciona una aproximación a la función de densidad deseada.

**Teorema 1.1.** *Sea $x_1, \ldots, x_n$ una muestra de una variable aleatoria con densidad $f(x)$ continua, y sea $\hat{f}(x)$ la aproximación por método kernel, entonces*
$$\lim_{n\to\infty} ECM[\hat{f}(x)] \to 0,$$
*para toda $x \in \mathbb{R}$.*

*Demostración.* Véase [10]. □

Si se desea tener más información sobre las medidas de eficiencia de $\hat{f}(x)$, véase [6] o [17].

## 1.4. Elección del ancho de banda

Actualmente, el método kernel es aceptado y la investigación sobre este, están basadas en la elección del ancho de banda $h$. A continuación se mostrará un método para la elección de dicho parámetro. Para empezar se definen las funciones siguientes:

- Para cualquier kernel $K$ se tiene que

$$\mu_2(K) = \int z^2 K(z) dz.$$

- Para cualquier función cuadrado integrable, se tiene que

$$R(g) = \int g(x)^2 dx.$$

Con estas definiciones, se puede entonces explicar uno de los primeros selectores[4] propuestos, el selector directo plug-in. Por el resto de la del capítulo, se denotará por $\hat{f}(x, h)$ a la aproximación por método kernel con ancho de banda $h$.

### 1.4.1. Selector directo Plug-In (DPI)

El selector Plug-In se basa en la ecuación (1.2). El enfoque habitual es encontrar una aproximación de ECMI utilizando la técnica de expansión de la serie Taylor, así, una aproximación de (1.2) es

$$AECMI[\hat{f}(\cdot, h)] = (nh)^{-1} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 \Psi_4,$$

donde para cualquier número $r$ par,

$$\Psi_r = \int_{-\infty}^{\infty} f^{(r)}(x) f(x) dx.$$

Es claro que no es fácil calcular la última expresión, por lo que se utiliza su estimador

$$\hat{\Psi}_r(g_r) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i; g_r) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{g_r}^{(r)}(X_i - X_j)$$

[4] Algoritmos para la elección del ancho de banda.

donde $g_r$ es un ancho de banda provisional y $L$ un kernel provisional[5], por lo que $L_{g_r}$ es un kernel parametrizado en $g_r$.

El punto de partida del DPI es

$$h_{AECMI} = \left( \frac{R(K)}{\mu_2(K)^2 \hat{\Psi}_4(g_4) n} \right)^{1/5}, \tag{1.3}$$

la cual no puede calcularse de manera directa, ya que depende de $g_4$. Wand y Jones en [17]mencionan que $g_r$ puede ser calculada a partir de

$$g_{r,AECMI} = \left( \frac{2K^{(4)}(0)}{-\mu_2(K)\Psi_{r+2}(g_{r+2}) n} \right)^{1/(r+3)}. \tag{1.4}$$

Como se observa, $g_{r,AECMI}$ depende de $\Psi_{r+2}(g_{r+2,AECMI})$, con $r$ par.

El DPI clásico calcula $\Psi_8$ de acuerdo a la siguiente fórmula (válida para $r$ par),

$$\Psi_r^{NS} = \left( \frac{(-1)^{r/2} r!}{(2\hat{\sigma})^{r+1} (r/2)! \pi^{1/2}} \right)^{1/9}, \tag{1.5}$$

donde $\hat{\sigma}$ es la estimación de la desviación estándar $\sigma$.

A continuación, el procedimiento DPI se resume en el Algoritmo 1.1.

**Algoritmo 1.1** Algoritmo del selector DPI.

**Entrada:** Datos $X_1, \ldots, X_n$, kernel $K$ y kernel provisional $L$.
**Salida:** El ancho de banda $\hat{h}_{DPI,2}$.

1: Estimar $\Psi_8$ usando $\Psi_8^{NS}$(fórmula (1.5)).
2: Estimar $\Psi_6$ usando $\hat{\Psi}_6(g_6)$, donde $g_6$ se calcula por (1.4).
3: Estimar $\Psi_4$ usando $\hat{\Psi}_4(g_4)$, donde $g_4$ se calcula por (1.4).
4: Estimar $\hat{h}_{DPI,2}$ usando (1.3).

Para más detalles de la obtención de las ecuaciones y análisis del algoritmo, véase[17] y para aspectos computacionales [6].

[5]En la practica, usualmente $L = K$, donde $K$ es el kernel elegido.

# Capítulo 2

# Simulación de los datos faltantes

Una forma de simular datos usando kerneles, es la de asignar un cierto peso de acuerdo a la distancia de los datos disponibles más cercanos. Este método puede ser observado en [9].

En el trabajo de Herrera [7], se plantea usar las estimaciones por método de kernel para simular datos faltantes, el inconveniente es que no hace una explicación precisa, además de lo poco que se explica no coincide con los resultados obtenidos. Por esta razón se desarrolla una metodología con esta idea.

## 2.1. Metodología planeada

Para trabajar con la base de datos con datos faltantes se hacen las siguientes observaciones.

1. A partir de los datos disponibles, se obtiene una aproximación a la función de densidad continua (se denotará por $\hat{f}$), con ancho de banda y kernel fijos. En esta aproximación no se debe ocupar el kernel normal, ya que el dominio de la función obtenida debe ser acotado para el siguiente paso.

2. Se discretiza la función obtenida en 1. Para esto se elige el tamaño del rango de la variable discreta (un número finito $N$).

El proceso consiste en particionar el rango de $\hat{f}$ en intervalos del mismo tamaño, y tomar los elementos del rango como el punto medio, a estos elementos se les denota por $x_i$, $i = 1, \ldots, N$. La probabilidad que se le asigna a cada $x_i$, será el número $p_i \geq 0$ tal que cumple que la

$$\begin{aligned} &\textit{Longitud del intervalo} \times p_i \\ &\qquad = \textit{Probabilidad del intervalo considerado.} \end{aligned}$$

3. Un dato faltante en la base de datos, se sustituye por un valor aleatorio $x$, que es un valor aleatorio simulado de la variable obtenida en 2.

**Nota.** De acuerdo a las propiedades de la base de datos en la que se trabaja, se puede añadir más pasos u observaciones a la metodología anterior, por ejemplo se puede reemplazar los números negativos simulados por ceros, en caso de, que la base solo tenga datos no negativos.

## 2.2. Ejemplo de la metodología

Para ver si la metodología propuesta es viable, se propone la experimentación siguiente:

1. Se generan 1500 datos aleatorios de la variable aleatoria $100X$, donde $X \sim Beta(3, 2)$.

2. Se elige una muestra aleatoria de tamaño $M$ de los datos en 1 y son eliminados.

3. Se utiliza la metodología planteada para un ancho de banda $h$ y rango $N$ de la variable discretizada por algún kernel fijo.

Los parámetros utilizados son:

- Los datos eliminados $M = 50,\ 100$.
- Ancho de banda $h = 0.2,\ 0.8$
- Rango de la variable discreta $N = 250$.

- Kernel: Arco coseno[1], cuártico[2], Epanechnikov, rectangular o triangular[3].

Con la experimentación, se observa lo siguiente:

Para la eliminación de los 50 datos, en todos los kerneles con ancho de banda $h = 0.2$ o $h = 0.8$, se tiene que la curva de distribución y los histogramas son muy parecidos a los de los datos originales, como se observa en las Figuras 2.1 y 2.2. Hay que señalar que parece ser que los puntos de inflexión en la curva de densidad disminuyen cuando el ancho de banda aumenta (es decir, $h = 0.8$).

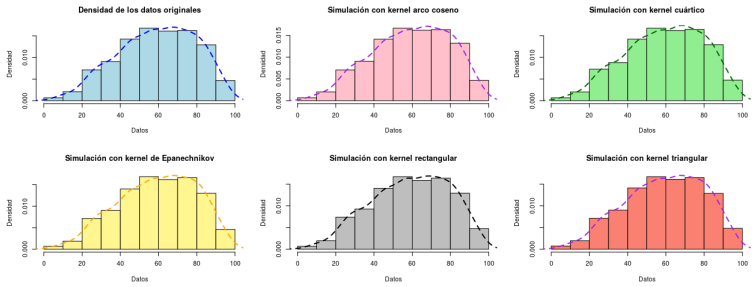


**Figura 2.1:** Prueba eliminando 50 datos y ancho de banda $h = 0.2$.

Para la eliminación de los 100 datos, se empieza a notar una diferencia en los histogramas, aunque parece conservar la misma forma como se observa en las Figuras 2.3 y 2.4, por lo que las curvas de densidad aún parecen conservar la forma, pero con más puntos de inflexión visibles. Al igual que en el caso de la eliminación de 50 datos, al aumentar el $h$ los puntos de inflexión parecen disminuir, aunque en menor manera al observado cuando se eliminan 50 datos.

En cuanto a los estadísticos que se comparan son la media y la varianza, siendo los valores 59.07452 y 411.9199 los valores de la muestra

[1]El kernel arco coseno se define por

$$K(x) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right), \qquad x \in [-1, 1].$$

[2]El kernel cuártico se define por

$$K(x) = \frac{15}{16}(1 - x^2)^2, \qquad x \in [-1, 1].$$

[3]El kernel triangular se define por

$$K(x) = 1 - |x|, \qquad x \in [-1, 1].$$

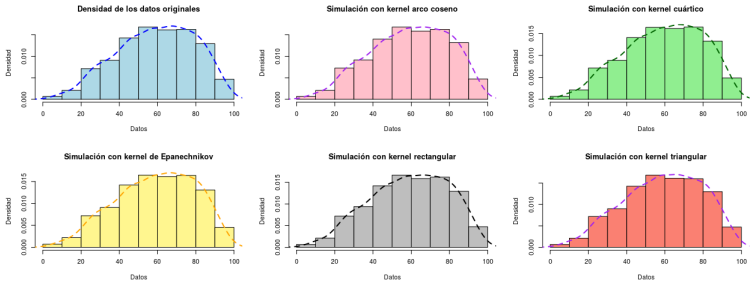


**Figura 2.2:** Prueba eliminando 50 datos y ancho de banda $h = 0.8$.

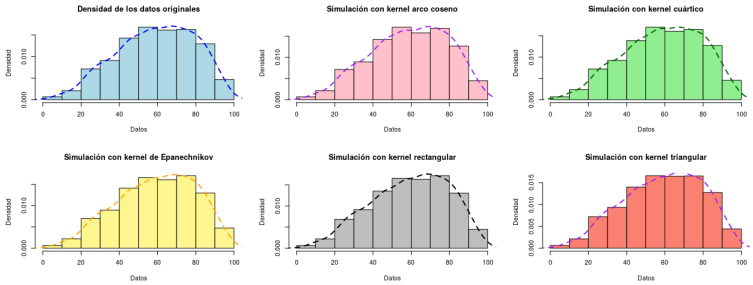


**Figura 2.3:** Prueba eliminando 100 datos y ancho de banda $h = 0.2$.

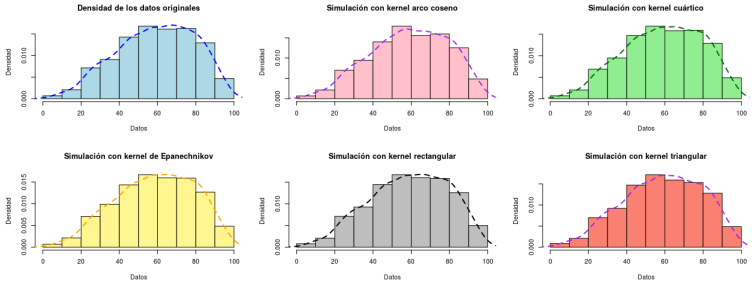


**Figura 2.4:** Prueba eliminando 100 datos y ancho de banda $h = 0.8$.

original respectivamente. Se observa que para las simulaciones donde se eliminaron 50 datos y se utiliza $h = 0.2$, la media en las distintas pruebas, varían alrededor del valor original, ya sea por décimas o centésimas. En cuanto las varianzas, estas varían en su mayoría por décimas. Esta información es observada en la Tabla 2.1.

En la Tabla 2.2, se observa que las medias en las distintas pruebas

| | Arco coseno | Cuártico | Epanechnikov | Rectangular | Triangular |
|---|---|---|---|---|---|
| Media | 59.22896 | 59.17007 | 59.26351 | 59.03398 | 59.19494 |
| Diferencia absoluta | 0.15444 | 0.09555 | 0.18899 | 0.04054 | 0.12042 |
| Var | 412.3889 | 412.5736 | 408.4308 | 413.8236 | 412.6009 |
| Diferencia absoluta | 0.469 | 0.6537 | 3.4891 | 1.9037 | 0.681 |

**Tabla 2.1:** Estadísticos para la experimentación $h = 0.2$ y eliminando 50 valores.

oscilan alrededor del valor original, sin embargo, las medias parecen aumentar con respecto al caso anterior, ya que estas difieren por unidades a la original.

| | Arco coseno | Cuártico | Epanechnikov | Rectangular | Triangular |
|---|---|---|---|---|---|
| Media | 59.10845 | 59.32436 | 58.98606 | 58.91586 | 59.01785 |
| Diferencia absoluta | 0.03393 | 0.24984 | 0.08846 | 0.15866 | 0.05667 |
| Var | 415.4717 | 416.3047 | 416.329 | 415.6368 | 413.8145 |
| Diferencia absoluta | 3.5518 | 4.3848 | 4.4091 | 3.7169 | 1.8946 |

**Tabla 2.2:** Estadísticos para la experimentación $h = 0.8$ y eliminando 50 valores.

Para el caso donde se eliminan 100 datos, la Tabla 2.3 muestra los resultados cuando $h = 0.2$, las medias siguen siendo muy cercanas a la original, pero las medias empiezan a variar por unidades. Parece ser que la disminución de información influye en las varianzas.

| | Arco coseno | Cuártico | Epanechnikov | Rectangular | Triangular |
|---|---|---|---|---|---|
| Media | 59.02089 | 58.86333 | 59.34646 | 59.38363 | 58.99367 |
| Diferencia absoluta | 0.05363 | 0.21119 | 0.27194 | 0.30911 | 0.08085 |
| Var | 410.696 | 415.7778 | 411.3565 | 409.4465 | 409.6959 |
| Diferencia absoluta | 1.2239 | 3.8579 | 0.5634 | 2.4734 | 2.224 |

**Tabla 2.3:** Estadísticos para la experimentación $h = 0.2$ y eliminando 100 valores.

Los resultados de aumentar $h$ a 0.8 en el caso anterior, se presentan en la Tabla 2.4. Las medias no se ven afectadas de manera significativa, pero parece que al aumentar el $h$, estas están por debajo de la original. En cuanto a las varianzas, estas siguen variando en su mayoría por unidades.

Aunque gráficamente, el método es viable, no es prueba suficiente para dar por válido que el método funcione, por consiguiente, se va a

| | Arco coseno | Cuártico | Epanechnikov | Rectangular | Triangular |
|---|---|---|---|---|---|
| Media | 58.86993 | 58.9857 | 58.73032 | 58.84344 | 58.79469 |
| Diferencia absoluta | 0.20459 | 0.08882 | 0.3442 | 0.23108 | 0.27983 |
| Var | 411.0654 | 410.6154 | 415.4838 | 415.6053 | 414.9499 |
| Diferencia absoluta | 0.2659 | 1.3045 | 3.5639 | 3.6854 | 3.03 |

**Tabla 2.4:** Estadísticos para la experimentación $h = 0.8$ y eliminando 100 valores.

plantear una prueba estadística para ver si hay evidencia de que funcione o no el método.

## 2.3. Prueba de la metodología

Con ayuda de la prueba Kolmogorov-Smirnov explicada en la Sección A.7, se hace un experimento, en el cual se busca verificar que el método planteado anteriormente, en efecto, da una buena simulación de los datos. Dicha prueba se hace a un nivel de significancia de $\alpha = 0.05$.

### 2.3.1. Discretizando la aproximación por método kernel

El experimento se desarrolla de la manera siguiente:

1. Se simulan 1500 datos de una variable aleatoria asociada a una función de distribución $F(x)$.

2. A partir de los datos generados, se crea la estimación por kernel $\hat{f}(x)$, tomando el ancho de banda $h$ como el dado por el selector directo plug-in, con los kerneles rectangular, Epanechnikov, cuártico y triweight[4], con los códigos en [18].

3. Como se hizo al inicio del capítulo, se discretiza $\hat{f}(x)$, usando intervalos de $M = 100$, 500 y 1000.

[4]El kernel triweight se define por

$$K(x) = \frac{35}{32}(1 - x^2)^3, \qquad x \in [-1, 1].$$

4. Después de haber obtenido la discretización, se simulan 1500 datos con dicha variable, este número se toma debido a que las bases de datos del medio ambiente tienen un número de datos cercanos a 1500.

5. Se calcula el $p$-valor de la prueba de hipótesis:

$$H_0 : \text{Los datos tienen una distribución } F(x).$$
$$\text{vs}$$
$$H_1 : \text{Los datos no tienen una distribución } F(x).$$

En la Tabla 2.5 se utiliza $f(x)$ como la función de densidad de una $Beta(3, 2)$, cuyo rango es finito, por lo que se espera que $\hat{f}(x)$ sea una buena aproximación a $f(x)$. Se observa que en su mayoría la prueba de hipótesis es aceptada a un nivel de significancia $\alpha = 0.05$. Como se esperaba, al tomar más valores de $M$, la hipótesis es aceptada por todos los kerneles, aunque en 3 casos, la hipótesis es rechazada.

| | KERNEL | | | |
|---|---|---|---|---|
| | Rectangular | Epanechnikov | Cuártico | Triweight |
| $h$ | 0.08096936 | 0.103014 | 0.1220371 | 0.138579 |
| $p$-valor para $M = 100$ | 0.008589 | 0.3811 | 0.009352 | 0.6535 |
| $p$-valor para $M = 500$ | 1 | 0.0407 | 0.3061 | 0.6287 |
| $p$-valor para $M = 1000$ | 0.6474 | 0.1086 | 0.09922 | 0.3946 |

**Tabla 2.5:** Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra $Beta(3, 2)$.

En el caso de la densidad $Gamma(3, 0.5)$, cuyo rango son los números positivos, la experimentación con $M = 100$, las pruebas fueron rechazadas, como se observa en la Tabla 2.6, sin embargo, con los otros parámetros, la hipótesis es aceptada, lo que nos permite formular que si el rango de la discretización aumenta, los resultados son mejores.

En la Tabla 2.7 se muestran los $p - values$ obtenidos del experimento con una densidad $Gumbel(3, 4)$. Se observa que la hipótesis nula se rechaza en los casos donde se discretiza la variable en un rango de $M = 100$, la mitad en donde $M = 500$. En este caso es fácil pensar que si el número del rango en que se discretiza la variable aumenta, se obtiene una mejor aproximación a $f(x)$.

En el último caso, se trabaja con la densidad $Weibull(1, 2/3)$. En este, el experimento da información negativa, ya que en todo momento, la prueba de hipótesis es rechazada, como se observa en la Tabla 2.8.

| | KERNEL | | | |
|---|---|---|---|---|
| | Rectangular | Epanechnikov | Cuártico | Triweight |
| $h$ | 1.130752 | 1.43861 | 1.70427 | 1.935281 |
| $p$-valor para $M = 100$ | 0.002666 | 0.002353 | 0.01432 | 0.03368 |
| $p$-valor para $M = 500$ | 0.442 | 0.4958 | 0.11 | 0.8366 |
| $p$-valor para $M = 1000$ | 0.5356 | 0.1149 | 0.1549 | 0.7328 |

**Tabla 2.6:** Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra $Gamma(3, 0.5)$.

| | KERNEL | | | |
|---|---|---|---|---|
| | Rectangular | Epanechnikov | Cuártico | Triweight |
| $h$ | 1.750271 | 2.2268 | 2.63801 | 2.995588 |
| $p$-valor para $M = 100$ | 0.05419 | 0.002279 | 0.04846 | 0.06553 |
| $p$-valor para $M = 500$ | 0.2569 | 0.07 | 0.001438 | 0.04282 |
| $p$-valor para $M = 1000$ | 0.34 | 0.5499 | 0.05375 | 0.3512 |

**Tabla 2.7:** Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra $Gumbell(3, 4)$.

Analizando la información en conjunto, se formula lo siguiente:

1. Para densidades de cola pesada como lo son la Gumbell y Weibull, con pocos datos (se piensa que 1500 son pocos), parece no dar buenos resultados, ya que recordemos que la teoría de kernel, entre más datos, mejor la aproximación, entonces para ellas se necesitan más datos.

2. Entre mayor sea la discretización el método funciona mejor. Esto se debe a que al haber menor distancia entre los elementos del dominio de la función, más parecida sera a una una función continua.

3. Aunque el kernel Epanechnikov es el kernel más estudiado, los kerneles más útiles son el kernel Rectangular y Triweight.

Los resultados permiten pensar que la discretización es una buena opción excepto para las distribuciones de cola pesada, por ejemplo distribución Gumbel o Weibull, esto se debe a los pocos datos utilizados, ya que entre más datos, mejor funciona la aproximación por método kernel.

| | KERNEL | | | |
|---|---|---|---|---|
| | Rectangular | Epanechnikov | Cuártico | Triweight |
| $h$ | 0.1065738 | 0.1355896 | 0.1606282 | 0.182401 |
| $p$-valor para $M = 100$ | $7.876X10^{-5}$ | 0.0007375 | 0.005991 | $8.056X10^{-8}$ |
| $p$-valor para $M = 500$ | 0.05436 | 0.01336 | 0.01059 | 0.09265 |
| $p$-valor para $M = 1000$ | 0.02518 | 0.02696 | 0.07134 | $9.579X10^{-5}$ |

**Tabla 2.8:** Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra $Weibull(1, 2/3)$.

### 2.3.2. Simulando la aproximación por método kernel

Ahora, se hace un experimento parecido al anterior, pero en vez de discretizar $\hat{f}(x)$, se simula la variable asociada a dicha densidad. El experimento se desarrolla de la manera siguiente:

1. Se simulan 1500 datos de una función de densidad $f(x)$ con su respectiva función de distribución $F(x)$.

2. A partir de los datos generados, se crea la estimación por kernel $\hat{f}(x)$, tomando el ancho de banda $h$ como el dado por el selector directo plug-in, con los kerneles rectangular, Epanechnikov, cuártico, triweight y normal, con los códigos en [18].

3. Se simulan 1500 datos de la variable aleatoria asociada a $\hat{f}(x)$.

4. Se calcula el $p$-valor de la prueba de hipótesis:

$$H_0 : \text{Los datos tienen una distribución } F(x).$$
$$\text{vs}$$
$$H_1 : \text{Los datos no tienen una distribución } F(x).$$

En la Tabla 2.9, se ven los resultados del experimento cuando la densidad es una $Beta(3, 2)$, en todos los kerneles utilizados la hipótesis nula es aceptada. Se observa el experimento actual, es igual de factible que el caso discreto, cuando $M = 1000$.

En el caso de la densidad $Gamma(3, 0.5)$, la hipótesis nula es aceptada en todos los casos, como se muestra en la Tabla 2.10. Se tiene que trabajar con la forma continua, es igual de efectivo que trabajar con su discretización cuando $M = 500$.

| Kernel | Rectangular | Epanechnikov | Cuártico | Triweight | Normal |
|---|---|---|---|---|---|
| $h$ | 0.08096936 | 0.103014 | 0.1220371 | 0.138579 | 0.0465259 |
| $p$-valor | 0.9814 | 0.6279 | 0.1167 | 0.6094 | 0.6757 |

**Tabla 2.9:** Prueba de Kolmogorov-Smirnov para el método propuesto con simulación de $\hat{f}(x)$, para una muestra $Beta(3, 2)$.

| Kernel | Rectangular | Epanechnikov | Cuártico | Triweight | Normal |
|---|---|---|---|---|---|
| $h$ | 1.130752 | 1.43861 | 1.70427 | 1.935281 | 0.6498362 |
| $p$-valor | 0.07608 | 0.6175 | 0.5359 | 0.06692 | 0.3287 |

**Tabla 2.10:** Prueba de Kolmogorov-Smirnov para el método propuesto con simulación de $\hat{f}(x)$, para una muestra $Gamma(3, 0.5)$.

Para la densidad de una $Gumbel(3, 4)$, en tres de los cinco kerneles, se acepta la hipótesis nula, como se observa en la Tabla 2.11. Para esta distribución, parece dar mejores resultados la discretización cuando $M = 1000$.

| Kernel | Rectangular | Epanechnikov | Cuártico | Triweight | Normal |
|---|---|---|---|---|---|
| $h$ | 1.750271 | 2.2268 | 2.63801 | 2.995588 | 1.00587 |
| $p$-valor | 0.02375 | 0.03311 | 0.3756 | 0.1091 | 0.3148 |

**Tabla 2.11:** Prueba de Kolmogorov-Smirnov para el método propuesto con simulación de $\hat{f}(x)$, para una muestra $Gumbel(3, 4)$.

Para el último caso, los datos en la Tabla 2.12, hipótesis nula se acepta en el caso del kernel cuártico, al igual que con la variable discretizada con $M = 1000$.

Analizando la información anterior, se observa que:

- Simular la distribución que se obtiene del método kernel parece funcionar muy bien, pero el inconveniente es el tiempo de ejecución de los códigos, ya que esta simulación se hace con el método de aceptación-rechazo.

- El método parece efectivo y mejor que el caso de discretización.

- Funciona mejor con densidades con rango acotado, y en el caso de una densidad de cola pesada, como lo son la Weibull y Gumbel, se necesitaría mas datos.

| Kernel | Rectangular | Epanechnikov | Cuártico | Triweight | Normal |
|---|---|---|---|---|---|
| $h$ | 0.1065738 | 0.1355896 | 0.1606282 | 0.182401 | 0.06124732 |
| $p$-valor | 0.01347 | 0.03546 | 0.06024 | 0.009711 | 0.005966 |

**Tabla 2.12:** Prueba de Kolmogorov-Smirnov para el método propuesto con simulación de $\hat{f}(x)$, para una muestra $Weibull(1, 2/3)$.

## 2.4. Algoritmo para la simulación de datos faltantes

Con los datos obtenidos, se tiene evidencia de que la propuesta es viable para ser incluido en los métodos para trabajar con datos faltantes, por lo que ya se puede formular el algoritmo para simulación de datos faltantes.

Para el algoritmo, se requiere de los datos con los que se quiere trabajar que tienen datos faltantes, donde se utiliza la aproximación $\hat{f}(x)$. Las ideas utilizadas hasta el momento se resume en el Algoritmo 2.1.

**Algoritmo 2.1** Imputación de datos faltantes por método de kernel.

**Entrada:** Datos en vector $DATOS$ y el kernel $K$.
**Salida:** Datos en vector $DATOS$.

1: Se crea una copia de $DATOS$, la cual es llamada $D$;
2: Se eliminan las entradas de $D$ donde se tiene datos faltantes;
3: Se obtiene el ancho de banda $h$ utilizando el selector DPI, usando el kernel $K$ y $D$;
4: Se obtiene la aproximación por método kernel $\hat{f}(x)$ con ancho de banda $h$, datos $D$ y kernel $K$;
5: Para cada entrada de $DATOS$ que sea un dato faltante, reemplazar por una simulación de una variable aleatoria $X$ con función de densidad $\hat{f}(x)$.

El Algoritmo 2.1 fue utilizado en el apartado 2.3.2, e implementado en R. La variable aleatoria se simula utilizando una idea derivada del Teorema A.1, e ilustrada en el Ejemplo A.2.

Aunque el algoritmo da buenos resultados, el inconveniente radica en que se hacen operaciones no siempre útiles, esto debido que se usa la idea del Ejemplo A.2.

Otra idea que se planteo durante las pruebas fue la de discretizar la función $\hat{f}(x)$, con idea de evitar cálculos innecesarios, dividiendo su rango en $M$ partes iguales, esta idea se resume en el Algoritmo 2.2.

**Algoritmo 2.2** Imputación de datos faltantes por discretizar aproximación por método de kernel.

**Entrada:** Datos en vector $DATOS$, el kernel $K$ y el entero $M$.
**Salida:** Datos en vector $DATOS$.

1: Se crea una copia de $DATOS$, la cual es llamada $D$;
2: Se eliminan las entradas de $D$ donde se tiene datos faltantes;
3: Se obtiene el ancho de banda $h$ utilizando el selector DPI, usando el kernel $K$ y $D$;
4: Se obtiene la aproximación por método kernel $\hat{f}(x)$ con ancho de banda $h$, datos $D$ y kernel $K$;
5: Se crea matriz $V$ de tamaño $2 \times M$;
6: Se obtiene $maxi = \text{máx}\{D\} + h$;
7: Se obtiene $mini = \text{mín}\{D\} - h$;
8: Se calcula $Paso = (maxi - mini)/(M + 1)$;
9: **Para** $i = 1$ hasta $M$ **hacer**
10: Se calcula $V[1, i] = mini + i \times PASO - PASO/2$, es decir, un punto medio;
11: Se calcula $V[2, i] = P(mini + (i - 1) \times PASO < X < mini + i \times PASO)$, es decir, la probabilidad;
12: **Fin Para**
13: Se eliminan las columnas de $V$ donde la segunda entrada de la columna es 0.
14: Para cada entrada de $DATOS$ que sea un dato faltante, reemplazar por una simulación de una variable aleatoria $V$ con función de densidad $\hat{f}(x)$.

En el Algoritmo 2.2, los pasos 9-12 son solo para obtener una discretización de la variable aleatoria asociada a $\hat{f}(x)$, siendo la primera fila el rango de la variable y la segunda fila las probabilidades. En el paso 13, se hace para la mejor implementación del Lema A.1.

Es claro que el Algoritmo 2.2 es más rápido que el Algoritmo 2.1, pero el problema radica en que se simulan datos de un conjunto finito, para simular una variable aleatoria con dominio no finito, y en ciertos subconjuntos del recorrido de $\hat{f}(x)$ es cero. Por tal motivo, la aplicación de los algoritmos podría tener resultados dudosos, más si se tiene una pequeña cantidad de datos en relación con el dominio de $f(x)$.

# Capítulo 3

# Ejemplo de aplicación. Niveles máximos de ozono

En este capítulo, se utiliza el Algoritmo 2.2 para simular los niveles máximos de ozono, de la estación Pedregal de la ciudad de México, obtenidos en la página de la Red Automática de Monitoreo Atmosférico recuperados de [12]. Los datos corresponden del 1 de enero de 1995 a la 1:00 horas al 1 de enero de 2020 a las 0:00 horas, tomando como dato las partes por millón de $O_3$.

## 3.1. Estimación de datos dudosos

Se puede pensar que los máximos tomados cada 2 semanas son independientes, con esta idea, se toman los máximos cada 2 semanas, sin omitir los datos faltantes, con lo que se obtienen 652 datos. Los datos omitidos son aquellos que en las 2 semanas consideradas, los datos faltantes son más del $100p\,\%$, donde $p = 0.5,\ 0.25$ y $0.1$, los datos restantes son considerados datos completos u originales.

Para el cálculo del ancho de banda $h$, se consideran solo los datos completos, y con los códigos en [18], dichos anchos de banda se encuentran en la Tabla 3.1. Se observa que para kernel, al haber menor número de datos, el ancho de banda aumenta, esto se debe que al tener una menor cantidad de datos, se necesita tener más información de los puntos

vecinos.

| Kernel | Rectangular | Epanechnikov | Cuártico | Triweight |
|---|---|---|---|---|
| $h$ para $p = 0.5$ | 16.85445 | 21.44324 | 25.40305 | 28.84638 |
| $h$ para $p = 0.25$ | 16.9088 | 21.51239 | 25.48496 | 28.9394 |
| $h$ para $p = 0.1$ | 17.48886 | 22.25037 | 26.35922 | 29.93216 |

**Tabla 3.1:** Estimación del ancho de banda $h$ para distintos $p$.

Una vez calculados, se realiza la simulación de los datos faltantes, esto se hace discretizando la aproximación por método de $\hat{f}(x)$, con un número esperado de intervalos $M = 1000$, y se sustituyen los datos faltantes por una simulación de dicha variable, y como el valor simulado puede no ser entero, se trunca este valor. Lo anterior se realiza con los kerneles rectangular, Epanechnikov, cuártico y triweight.

## 3.2. Resultados

En la Tabla 3.2, se muestran los resultados obtenidos cuando la porción de datos faltantes, es mayor a $p = 0.5$, por lo que fueron omitidos solamente 6 datos. De acuerdo a los datos, el valor mínimo, máximo, media y 1er cuartil no se modifican, permanecen iguales a los datos completos en la simulación con cualquier kernel, lo cual nos dice que a una pequeña cantidad de datos incompletos, las estadísticas son muy cercanas.

| | Mín | 1er Cuar. | Mediana | Media | 3er Cuar. | Máx |
|---|---|---|---|---|---|---|
| Original | 66.0 | 127.0 | 153.0 | 167.2 | 204.0 | 349.0 |
| K. Rectangular | 66.0 | 127.0 | 153.0 | 167.2 | 204.0 | 349.0 |
| K. Epanechnikov | 66.0 | 127.0 | 153.0 | 167.5 | 204.5 | 349.0 |
| K. Cuártico | 66.0 | 127.0 | 153.0 | 167.2 | 204.0 | 349.0 |
| K. Triweight | 66.0 | 127.0 | 153.0 | 167.5 | 204.5 | 349.0 |

**Tabla 3.2:** Cálculo de las estadísticas para los datos originales y los datos obtenidos con la simulación de datos, con un $p = 0.5$. Se toma $q_3$ como el tercer cuartil.

En la Figura 3.1 se observa los histogramas de los datos originales, así como de los datos simulados con la metodología planteada, los cuales a simple vista no parecen diferentes. En la Figura 3.2, se colocan las

curvas de densidad aproximadas dadas por el paquete R, con lo que se observa una mínima diferencia entre ellas.

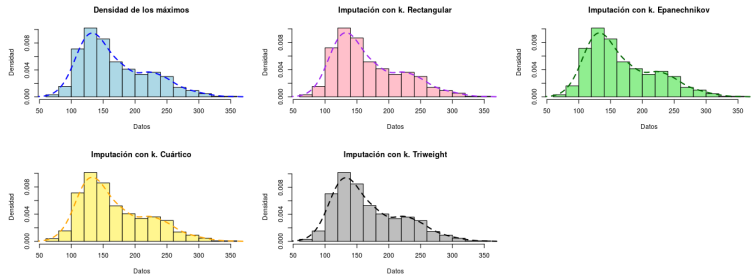


**Figura 3.1:** Histogramas y densidades para $p = 0.5$.

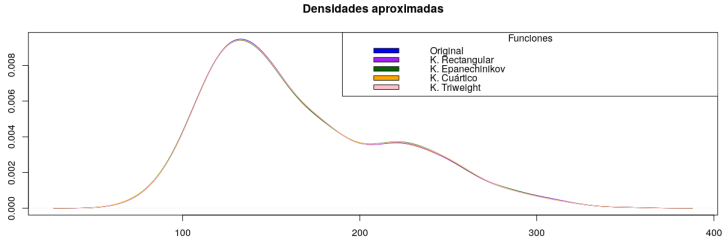


**Figura 3.2:** Comparación de las densidades obtenida para $p = 0.5$.

Por lo anterior, se puede pensar que el kernel que mejor funciona son el rectangular y cuártico, ya que estos son los que sus estadísticas coinciden con los datos considerados originales.

Para el caso donde $p = 0.25$, se omitieron un total de 26 datos faltantes, más de 3 veces la cantidad del caso anterior. Como se observa en la Tabla 3.3, los mínimos y máximos no cambian. Se empieza a observar diferencias en las otras estadísticas, que se esperaba al aumentar la cantidad de datos faltantes, además de observar que en la simulación, con ningún kernel se logró que las estadísticas coincidieran con los originales.

Por otro lado, en cuanto a los histogramas, haciendo una inspección a simple vista a la Figura 3.3, se puede observar el cambio de tamaño en algunas barras, excepto en los casos de los kerneles rectangular y triweight. En cuanto a las funciones de densidad aproximadas en la Figura 3.4, ya se empieza a observar una diferencia más clara entre curvas, pero las colas son parecidas.

| | Mín | 1er Cuar. | Mediana | Media | 3er Cuar. | Máx |
|---|---|---|---|---|---|---|
| Original | 66.0 | 127.0 | 154.5 | 168.3 | 207.0 | 349.0 |
| K. Rectangular | 66.0 | 127.0 | 153.0 | 167.6 | 206.2 | 349.0 |
| K. Epanechnikov | 66.0 | 127.8 | 155.0 | 168.7 | 207.0 | 349.0 |
| K. Cuártico | 66.0 | 127.0 | 154.5 | 168.4 | 207.0 | 349.0 |
| K. Triweight | 66.0 | 127.0 | 154.5 | 168.5 | 207.0 | 349.0 |

**Tabla 3.3:** Cálculo de las estadísticas para los datos originales y los datos obtenidos con la simulación, con un $p = 0.25$.

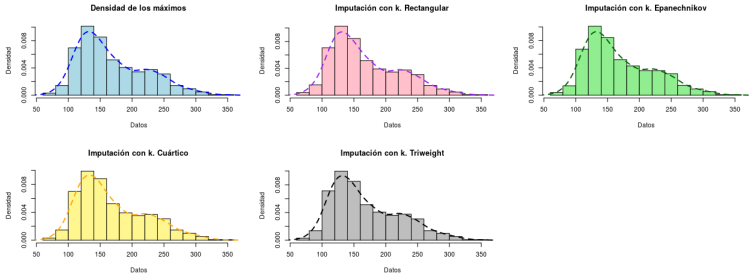


**Figura 3.3:** Histogramas y densidades para $p = 0.25$

Se puede decir que los kerneles que dan mejor resultado son el cuártico y el triweight, ya que son los que coinciden en sus estadísticas con los originales, a excepción de la media. A diferencia del caso anterior, el que da peores resultados, pero no necesariamente malos, parece ser el kernel rectangular, que es el que tiene mayor diferencia en 3 estadísticas.

En la Tabla 3.4, se observan las estadísticas para el caso donde $p = 0.1$, la cual se omiten 84 datos. La diferencia entre las estadísticas parece aumentar como se espera, pero en el caso del 3er cuartil, la diferencia en el caso del kernel rectangular y cuártico difieren en 3 unidades.

| | Mín | 1er Cuar. | Mediana | Media | 3er Cuar. | Máx |
|---|---|---|---|---|---|---|
| Original | 66.0 | 128.0 | 156.0 | 170.1 | 211.0 | 349.0 |
| K. Rectangular | 66.0 | 128.0 | 157.0 | 170.4 | 208.0 | 349.0 |
| K. Epanechnikov | 66.0 | 127.0 | 155.0 | 170.2 | 211.0 | 349.0 |
| K. Cuártico | 66.0 | 127.0 | 155.0 | 168.7 | 208.0 | 349.0 |
| K. Triweight | 66.0 | 128.0 | 156.0 | 170.5 | 211.0 | 349.0 |

**Tabla 3.4:** Cálculo de las estadísticas para los datos originales y los datos obtenidos con la simulación de datos, con un $p = 0.1$.

En la observación de los histogramas de la Figura 3.5 se ve diferencias en los tamaños de las barras, excepto con el kernel triweight. En la observación de las densidades aproximadas, en la Figura 3.6, no se observa la diferencia entre las colas de las densidades, pero si en el resto

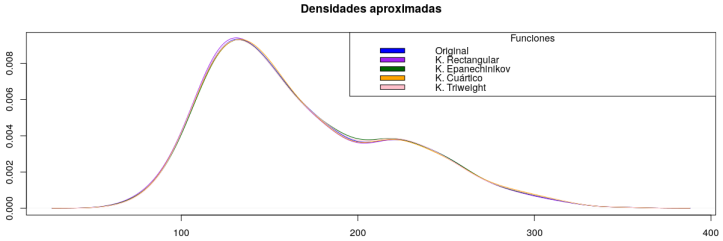


**Figura 3.4:** Comparación de las densidades obtenida para $p = 0.25$.

de la gráfica.

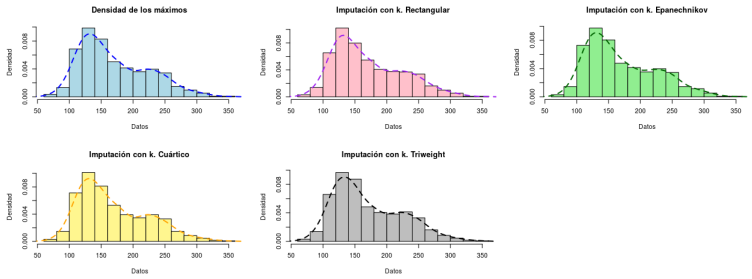


**Figura 3.5:** Histogramas y densidades para $p = 0.1$

Por lo anterior se podría decir que el kernel con mejores resultados resulta ser el triweight, y el peor ahora resulta ser el cuártico al tener más diferencia entre sus estadísticas con respecto a la original.

En general, se puede decir, que de acuerdo a las gráficas y las estadísticas, cualquier kernel proporciona una buena simulación de datos, ya que las estadísticas solo difieren por unas cuantas unidades, por ejemplo, en la Tabla 3.4, en el caso de usar un kernel rectangular, el cual la diferencia entre el 3er cuartil de los datos originales con respecto al obtenido con los datos simulados, tiene una diferencia de 3 unidades, el cual se puede pensar que es grande, pero se está hablando en partes por millón, lo cual solo es una pequeña diferencia.

Por tanto, la simulación de datos faltantes por medio del método kernel, es un buen método para trabajar cuando existen datos faltantes.

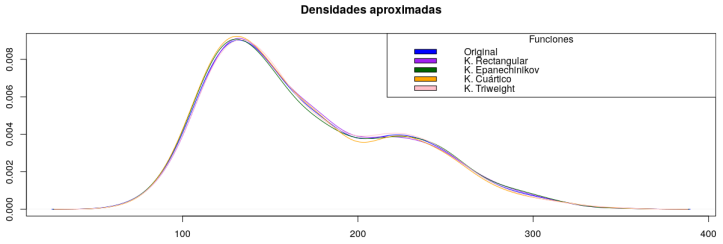


**Figura 3.6:** Comparación de las densidades obtenida para $p = 0.1$.

# Capítulo 4

# Conclusiones

En este trabajo, se desarrollo la idea de utilizar las estimaciones por método kernel, para simular los datos faltantes en bases de datos, por lo que se planteó una metodología. Se diseñan experimentos donde se ve la viabilidad del método propuesto, por lo que la metodología expuesta es aceptada como método para trabajar con datos faltantes. Por todo esto, las ideas se resumen en los Algoritmos 2.1 y 2.2.

Las bases en las que se recomienda aplicar los algoritmos, son aquellas en las que existe independencia entre los datos, es decir, uno elemento no depende de otro, ya que es un requerimiento para el método kernel. Se debe considerar que las observaciones provienen de una variable aleatoria, o bien, hacer un tratamiento a los datos, que permitan estudiar el fenómeno de interés, y así aplicar los algoritmos que se formulan en el capítulo 2.

Para el desarrollo de este trabajo, se crearon varias funciones en R para experimentar las ideas en el capítulo 2. Los resultados que se obtienen, indican que la metodología propuesta originalmente, que consiste en la discretización de la función $\hat{f}(x)$ obtenida a partir del método kernel, es factible, por lo que en trabajos posteriores, se planea estudiar el rango de la variable discretizada y el kernel, e incluso encontrar estadísticas para elegir la mejores opciones en kernel.

A lo largo del trabajo, se planteó la idea de que $\hat{f}(x)$ no fuera discretizada, por lo que en los códigos $R$, se crearon funciones para simular dicha función. Dicha propuesta, evita la elección del ancho de banda. Aunque la implementación, al simular valores aleatorios por el método de Aceptación-Rechazo y puesto que dicha función es para cualquier $\hat{f}(x)$. El

tiempo solo aumenta por el número de datos faltantes y el tamaño de muestra.

A la modificación de la metodología original, el experimento, dio resultados parecidos, por lo que también es viable utilizar el Teorema A.1 para simular $\hat{f}(x)$. Esta modificación es útil cuando se supone que los datos tienen un función de densidad. Por último, con esta idea, se tiene como trabajos futuros la mejora de los códigos, el saber como elegir el mejor kernel y construir algunos resultados teóricos que permitan generalizar resultados.

Es claro que al experimentar, $\hat{f}(x)$, tenga un rango mayor a la $f(x)$ si esta es de rango finito, esto por el ancho de banda obtenido, pero esto no fue especialmente un problema, ya que la probabilidad de que se obtuviera un número fuera del rango de la densidad $f(x)$, es relativamente pequeña, contrastando por el hecho de que nunca se obtuvo un dato simulado fuera del rango en los experimentos que se llevaron acabo.

En cuanto al ejemplo de aplicación, sobre los niveles máximo de ozono, se ve que es una buena opción para tratar con los datos faltantes, pues en las gráficas y estadísticas dan buenos resultados. Como conclusión, los datos que se simularon resultan ser buenos datos para sustituir la información que falta, por lo que, si se desea calcular cualquier estadística que requiera una gran cantidad de datos, el Algoritmo 2.2 es una herramienta útil para tal objetivo.

Cabe recalcar, que la experimentación dio los resultados pensados al inicio del trabajo, por lo que los Algoritmos 2.1 y 2.2 ya pueden ser agregados a los métodos para trabajar con datos faltantes, la cual es la principal aportación de esta tesis.

Para investigaciones futuras, se plantea utilizar la misma idea de los kerneles para la simulación de datos faltantes, ahora con los datos de ozono por periodos de tiempo, ya que este tipo de datos tienden a repetirse año con año. También se quiere analizar más a fondo la teoría de kerneles, para ver si se trabaja con kerneles no simétricos se obtiene una mejor aproximación en los casos de la distribución Gumbell, o bien, utilizar ahora los kerneles para aproximar funciones de densidad de 2 o más dimensiones, y así poder simular datos en donde interfieren más de una variable.

# Apéndice A

# Conceptos y Resultados básicos

A continuación se presentan algunos conceptos y resultados necesarios para la comprensión de este trabajo. Para este apéndice, es necesario tener conocimientos básicos de conjuntos de teoría de conjuntos que pueden ser consultados en la sección 3 del capítulo 1 de [13], además de conocimientos de cálculo de limites que pueden ser consultados en [1].

## A.1. Espacio de probabilidad

**Definición A.1** (Espacio muestral)**.** *El espacio muestral, de un experimento aleatorio*[1] *es el conjunto de todos los posibles resultados del experimento y se le denota, generalmente, por* $\Omega$*. A un resultado particular del experimento se le denota por la letra* $\omega$*.*

**Definición A.2** ($\sigma$-álgebra)**.** *Una colección* $\mathscr{F}$ *de subconjuntos de un espacio muestral* $\Omega$ *es una* $\sigma$*-álgebra(sigma-álgebra) si cumple las tres condiciones siguientes:*

1. $\Omega \in \mathscr{F}$.

2. *Si* $A \in \mathscr{F}$ *entonces* $A^c \in \mathscr{F}$.

[1] Un experimento aleatorio es aquél en el que si se repite con las mismas condiciones iniciales no se garantiza los mismos resultados.

3. *Si* $A_1, A_2, \ldots \in \mathscr{F}$ *entonces* $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$.

*A la pareja* $(\Omega, \mathscr{F})$ *se le llama espacio medible y a los elementos de* $\mathscr{F}$ *se les llama eventos o conjuntos medibles.*

**Definición A.3** ($\sigma$-álgebra de Borel de $\mathbb{R}^n$)**.**

$$\mathscr{B}(\mathbb{R}^n) = \sigma\{\mathscr{B}(\mathbb{R}) \times \cdots \times \mathscr{B}(\mathbb{R})\}.$$

**Definición A.4** (Medida de probabilidad)**.** *Sea* $(\Omega, \mathscr{F})$ *un espacio medible. Una medida de probabilidad es una función* $P : \mathscr{F} \to [0, 1]$ *que satisface*

1. $P(\Omega) = 1$.
2. $P(A) \geq 0$, *para cualquier* $A \in \mathscr{F}$.
3. *Si* $A_1, A_2, \ldots \in \mathscr{F}$ *son ajenos dos a dos, esto es,* $A_n \cap A_m = \emptyset$ *para* $n \neq m$, *entonces* $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$.

Si se desea saber propiedades de la medida de probabilidad, se puede consultar [14].

**Definición A.5** (Espacio de probabilidad)**.** *Un espacio de probabilidad es una terna* $(\Omega, \mathscr{F}, P)$, *en donde* $\Omega$ *es un conjunto arbitrario,* $\mathscr{F}$ *es una* $\sigma$*-álgebra de subconjuntos de* $\Omega$, *y* $P$ *es una medida de probabilidad definida sobre* $\mathscr{F}$.

## A.2. Variables aleatorias

**Definición A.6** (Variable aleatoria)**.** *Una variable aleatoria es una transformación* $X$ *del espacio de resultados* $\Omega$ *al conjunto de números reales, es decir,*

$$X : \Omega \to \mathbb{R},$$

*tal que para cualquier número real* $x$,

$$\{\omega \in \Omega : X(w) \leq x\} \in \mathscr{F}.$$

**Proposición A.1.** *Sean* $X$ *y* $Y$ *variables aleatorias y* $c$ *una constante, entonces*

*a)* $cX$,

*b)* $X + Y$,

*c)* $XY$,

*d)* $X/Y$ *donde* $Y \neq 0$,

*e)* máx$\{X, Y\}$ *y*

*f)* mín$\{X, Y\}$

*son variables aleatorias.*

**Definición A.7** (Función de distribución)**.** *La función de distribución de una variable aleatoria* $X$ *es la función* $F : \mathbb{R} \to [0, 1]$*, definida por*

$$F(x) = P(X \leq x), \qquad x \in \mathbb{R}.$$

**Proposición A.2.** *Sea* $F(x)$ *la función de distribución de una variable aleatoria. Entonces*

1. $\lim_{x\to\infty} F(x) = 1$.
2. $\lim_{x\to-\infty} F(x) = 0$.
3. *Si* $x_1 \leq x_2$*, entonces* $F(x_1) \leq F(x_2)$.
4. $F(x)$ *es continua por la derecha, es decir,* $F(x+) = F(x)$.

*Demostración.* Véase [14]. □

**Definición A.8** (Variable aleatoria discreta)**.** *La variable aleatoria* $X$ *se llama discreta si su correspondiente función de distribución* $F(x)$ *es una función constante por pedazos. Sean* $x_1, x_2, \ldots$ *los puntos de discontinuidad de* $F(x)$*. En cada uno de estos puntos el tamaño de la discontinuidad es* $P(X = x_i) = F(x_i) - F(x_i-) > 0$*. A la función* $f(x)$ *que indica estos incrementos se le llama función de probabilidad de* $X$*, y se define como sigue*

$$f(x) = \begin{cases} P(X = x) & \text{si } x = x_1, x_2, \ldots \\ 0 & \text{otro caso.} \end{cases}$$

*La función de distribución se reconstruye de la forma siguiente*

$$F(x) = \sum_{u \leq x} f(u).$$

**Definición A.9** (Variable aleatoria continua)**.** *La variable aleatoria $X$ se llama continua si su correspondiente función de distribución es una función continua.*

**Definición A.10** (Variable aleatoria absolutamente continua)**.** *La variable aleatoria continua $X$ con función de distribución $F(x)$ se llama absolutamente continua, si existe una función no negativa e integrable $f$ tal que para cualquier valor de $x$ se cumple*

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

*En tal caso a la función $f(x)$ se le llama función de densidad de $X$.*

Para hablar de una variable aleatoria, es suficiente hacer referencia a su función de distribución o función de densidad, ya que de una se puede obtener la otra.

**Definición A.11** (Independencia de variables aleatorias)**.** *Se dice que las variables aleatorias $X$ y $Y$ son independientes si los eventos $(X \leq x)$ y $(Y \leq y)$ son independientes para cualesquiera valores reales de $x$ y $y$, es decir, si se cumple la igualdad*

$$P[(X \leq x) \cap (Y \leq y)] = P(X \leq x)P(Y \leq y).$$

**Definición A.12** (Esperanza)**.** *Sea $X$ una variable aleatoria discreta con función de probabilidad $f(x)$. La esperanza de $X$ se define como el número*

$$E[X] = \sum_{x} xf(x).$$

*suponiendo que esta suma es absolutamente convergente, es decir, cuando la suma de los valores absolutos es convergente. Por otro lado, si $X$ es continua con función de densidad $f(x)$, entonces la esperanza es*

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx,$$

*suponiendo que esta integral es absolutamente convergente, es decir, cuando la integral de los valores absolutos es convergente.*

**Definición A.13** (Varianza)**.** *Sea $X$ una variable aleatoria discreta con función de probabilidad $f(x)$. La varianza de $X$ se define como el número*

$$V[X] = \sum_{x} (x - \mu)^2 f(x).$$

*cuando esta suma es convergente y en donde $\mu$ es la esperanza de $X$. Para una variable aleatoria continua $X$ con función de densidad $f(x)$ se define*

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx,$$

*cuando esta integral es convergente.*

**Proposición A.3.** *Sean $X$ y $Y$ dos variables aleatorias con varianza finita y sea $c$ una constante. Entonces*

1. $V[X] \geq 0$.
2. $V[c] = 0$.
3. $V[cX] = c^2 V[X]$.
4. $V[X + c] = V[X]$.
5. $V[X] = E[X^2] - E^2[X]$.
6. *En general,* $V[X + Y] \neq V[X] + V[Y]$.
7. *Si $X$ y $Y$ son variables independientes, entonces*

$$V[X + Y] = V[X] + V[Y].$$

*Demostración.* Véase [13]. □

## A.3. Distribuciones de probabilidad

A continuación se mencionan las distribuciones utilizadas en este trabajo. Como es costumbre a la función de densidad se denotará por $f(x)$ y el símbolo $\sim$ se entiende que es "distribuida como".

### A.3.1. Distribución Beta

$X \sim Beta(a, b)$ con $a > 0$, $b > 0$.
$f(x) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, para $x \in (0, 1)$.[2]

[2] $\Gamma(x)$ se conoce como la función gamma, y se define por

$$\Gamma(x) = \int_0^{\infty} t^{z-1} e^{-t} dt.$$

$E[X] = \dfrac{a}{a+b}$.
$V[X] = \dfrac{ab}{(a+b+1)(a+b)^2}$.

### A.3.2. Distribución Gamma

$X \sim Gamma(a, s)$ con $a > 0$, $s > 0$.
$f(x) = \dfrac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}$, para $x > 0$.
$E[X] = as$.
$V[X] = as^2$.

### A.3.3. Distribución Gumbel

Esta distribución es típica en las distribuciones de valores extremos, y su análisis puede consultarse en [3]. Para este trabajo solo se necesita su función de distribución, la cual es

$$F(x) = e^{-e^{-(x-\mu)/\sigma}} \qquad \text{para } x \in \mathbf{R},$$

donde $\mu \in \mathbb{R}$ y $\sigma > 0$.

Para referirnos a una variable aleatoria con la distribución Gumbel, se utiliza la notación $X \sim Gumbel(\mu, \sigma)$.

### A.3.4. Distribución t

$X \sim t(n)$ con $n > 0$.
$f(x) = \dfrac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} (1 + x^2/n)^{-(n+1)/2}$, para $x \in \mathbb{R}$.
$E[X] = 0$.
$V[X] = \dfrac{n}{n-2}$, para $n > 2$.

---

Además para $n$ entero positivo se cumple que

$$\Gamma(n) = (n-1)!.$$

### A.3.5. Distribución Uniforme(continua)

$X \sim Unif[a,b]$ con $a < b$.
$f(x) = \dfrac{1}{b-a}$, para $x \in [a,b]$.
$E[X] = \dfrac{a+b}{2}$.
$V[X] = \dfrac{(b-a)^2}{12}$.

### A.3.6. Distribución Weibull

$X \sim Weibull(a,b)$ con $a > 0$, $b > 0$.
$f(x) = \dfrac{a}{b}\left(\dfrac{x}{b}\right)^{(a-1)} e^{-(x/b)^a}$, para $x > 0$.
$E[X] = b\Gamma\left(1 + \dfrac{1}{a}\right)$.
$V[X] = b^2\left(\Gamma\left(1 + \dfrac{2}{a}\right) - \Gamma\left(1 + \dfrac{1}{a}\right)^2\right)$.

## A.4. Simulación de variables aleatorias

Actualmente la simulación de una variable aleatoria con cierta distribución esta programada en varios software como R. Por otro lado, existen variables aleatorias no tan comunes que no están programadas o bien, un investigador propone una nueva variable, por ello, se necesita métodos para simular variables aleatorias.

Todos los software o paquete estadístico ya tiene programada la simulación de una variable aleatoria uniforme, hecho que se utiliza para simular otras variables, como se mostrará a continuación.

**Definición A.14.** *Para una función no decreciente $F$ en $\mathbb{R}$, el inverso generalizado de $F$, $F^-$, es la función definida por*

$$F^-(u) = \text{ínf}\{x : F(x) \geq u\}.$$

**Lema A.1** (Transformación Integral de Probabilidad)**.** *Si $U \sim Unif[0,1]$ y $F$ una función de distribución, entonces la variable aleatoria $F^-(U)$ tienen la distribución $F$.*

*Demostración.* Véase [15]. □

Por lo tanto, de acuerdo al Lema anterior, para generar una variable aleatoria $X$ que tenga función de distribución $F$, es suficiente generar $U \sim Unif[0,1]$ y luego hacer la transformación $x = F^{-}(u)$.

**Ejemplo A.1.** *Si $X \sim Gumbel(\mu, \sigma)$, entonces $F(x) = e^{-e^{-(x-\mu)/\sigma}}$, entonces resolviendo para $x$ en $u = e^{-e^{-(x-\mu)/\sigma}}$ se tiene $x = \mu - \sigma \log(-\log(u))$. Por lo tanto, si $U \sim Unif[a,b]$, ls variable aleatoria $X = \mu - \sigma \log(-\log(U))$ tiene distribución Gumbel.*

**Teorema A.1** (Teorema fundamental de simulación)**.** *Simular $X$ con función de densidad $f(x)$ es equivalente a simular*

$$(X, U) \sim Unif\{(x,u) : 0 < u < f(x)\}.^{3}$$

*Demostración.* Véase [15]. □

**Ejemplo A.2.** *Sea $X$ una variable aleatoria con función de densidad $f(x)$ y $m > 0$ tal que $f(x) \leq m$, para toda $x \in \mathbb{R}$. Se supone que $f(x) = 0$ para $x \notin [a,b]$, donde $a < b$, entonces se puede simular el par $(Y,U) \sim Unif\{(y,u) : 0 < u < m\}$ y $U|Y \sim Unif(0,m)$, y tomando el par solo si $0 < u < f(y)$ es satisfecho. Esto da como resultado la distribución correcta del valor aceptado de $Y$, es decir $X$, porque*

$$\begin{aligned} P(X \leq x) &= P(Y \leq x | U < f(Y)) \\ &= \frac{\int_a^x \int_0^f (y) du dy}{\int_a^b \int_0^f (y) du dy} \\ &= \int_a^x f(y) dy. \end{aligned}$$

## A.5. Estadísticas

En esta sección, se analizan los conceptos de estadística que permiten una comprensión del trabajo realizado. Si se desea ver ejemplos de las definiciones, véase [11].

**Definición A.15** (Población)**.** *Una población representa la colección completa de elementos o resultados de la información buscada.*

[3] Vector uniforme en $\Omega = \{(x,u) : 0 < u < f(x)\}$, para mas información véase [15].

**Definición A.16** (Muestra)**.** *Una muestra constituye un subconjunto de una población, que contiene elementos o resultados que realmente se observan.*

**Definición A.17** (Muestra aleatoria)**.** *Una muestra aleatoria de tamaño n es una muestra elegida por un método en el que cada colección de n elementos de la población tiene la misma probabilidad de formar la muestra.*

**Definición A.18** (Media (muestral))**.** *Sea $X_1, \ldots, X_n$ una muestra. La media muestral es*

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

**Definición A.19** (Varianza (muestral))**.** *Sea $X_1, \ldots, X_n$ una muestra. La varianza muestral es la cantidad*

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{A.1}$$

**Lema A.2.** *Una fórmula equivalente a (A.1), que puede ser más fácil de calcular, es*

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

*Demostración.* Véase [8]. □

**Definición A.20** (Media (muestral))**.** *Si n números están ordenados del más pequeño al más grande:*

- *Si n es impar, la mediana muestral es el número en la posición* $\frac{n+1}{2}$.
- *Si n es par, la mediana muestral representa el promedio de los números en las posiciones* $\frac{n}{2}$ *y* $\frac{n}{2} + 1$.

**Definición A.21** (Cuartiles)**.** *Los cuartiles dividen la muestra tanto como sea posible en cuartos. Una muestra tiene tres de aquéllos. El método más simple cuando se calcula es el siguiente: Sea n el tamaño de la muestra. Ordene los valores de la muestra del más pequeño al más grande. Para encontrar el "primer cuartil", calcule el valor* $0.25(n+1)$. *Si éste es un entero, entonces el valor de la muestra en esa posición es el primer cuartil. Si no, tome entonces el promedio de los valores de la*

*muestra de cualquier lado de este valor. El "tercer cuartil" se calcula de la misma manera, excepto que se usa el valor* $0.75(n+1)$. *El "segundo cuartil"*[4] *usa el valor* $0.5(n+1)$.

## A.6. Prueba de hipótesis

Ahora se explicara de manera rápida que es una prueba de hipótesis. Ejemplos pueden ser consultados en [8] y [11].

**Definición A.22** (Hipótesis (estadística))**.** *Una hipótesis (estadística) es una aseveración sobre un modelo probabilístico. El procedimiento mediante el cual se juzga la factibilidad de la hipótesis es una prueba de hipótesis.*

**Definición A.23** (Hipótesis nula y alternativa)**.** *La hipótesis nula(denotada por* $H_0$*),es aquella que el investigador está dispuesto a sostener como plausible, a menos que la evidencia experimental en su contra sea sustancial. Su negación es llamada la hipótesis alternativa(denotada por* $H_1$*).*

**Definición A.24** (Región critica)**.** *La región critica es el conjunto de todos los puntos en el espació muestra tal que da el resultado de la decisión de rechazar la hipótesis nula.*

**Definición A.25** (Tipos de error)**.** *En una prueba de hipótesis pueden cometerse dos tipos de error. El llamado Error Tipo I consiste en rechazar una hipótesis nula que es cierta y el Error Tipo II consiste en no rechazar una hipótesis nula que es falsa. Las probabilidades de los errores respectivos se denotan por* $\alpha$ *y* $\beta$.

**Definición A.26** (Nivel de significancia)**.** *En un aprueba de hipótesis, el valor máximo de la probabilidad de Error Tipo I es llamado el nivel de significancia de la prueba. Se le llama, ocasionalmente, el tamaño de la prueba*[5].

**Definición A.27** ($p$-valor)**.** *El p-valor es el más pequeño nivel de significancia al cual la hipótesis puede ser rechazada para la muestra dada.*

[4] El segundo cuartil coincide con la mediana.

[5] El nivel de significancia que se elige en una prueba dependerá de la naturaleza del problema.

## A.7. Prueba de Kolmogorov-Smirnov

A continuación se mostrara una prueba estadística que no depende de una distribución, por lo que no se tiene un parámetro del cual se haga una inferencia, pero sigue las mismas ideas de una prueba de hipótesis explicadas en la sección anterior.

Sea $S(x)$ la función de distribución empírica basada en la muestra aleatoria $X_1, \ldots, X_n$, $F(x)$ es la función de distribución desconocida y $F^*(x)$ una función de distribución hipotética dada. La prueba de hipótesis de Kolmogorov-Smirnov es la siguiente:

$$\begin{gathered} H_0 : F(x) = F^*(x). \\ \text{vs} \\ H_1 : F(x) \neq F^*(x). \end{gathered}$$

El estadístico de prueba esta dada por

$$T = \sup_x |F^*(x) - S(x)|,$$

tomando como regla de decisión:

$$\text{Rechazar } H_0 \text{ si } D > t^n_{1-\alpha}{}^6,$$

o bien, si se calcula el $p$-valor $p$, se utiliza la regla de decisión:

$$\text{Rechazar } H_0 \text{ si } p < \alpha.$$

Para mas información véase [4].

[6] La cantidad $t^n_{1-\alpha}$ es aquel número tal que

$$P(X \leq t^n_{1-\alpha}) = 1 - \alpha,$$

donde $X \sim t(n)$.

# Referencias


[1] Arizmendi, H., Carrillo, A., and Lara, M. *Cálculo*, 2da ed. Instituto de matemáticas UNAM, 2016.

[2] Ballester, F. Contaminación atmosférica, cambio climático y salud. *Revista Española de Salud Pública 79*, 2 (2005), 159–175.

[3] Coles, S. *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.

[4] Conover, J. *Practical nonparametric statistics*, 3ra ed. John Wiley & Sons, Inc, 1999.

[5] García, J., Albaladejo, J., and Fernández, J. Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones. *Estadística española 48*, 162 (2006), 241–270.

[6] Gramacki, A. *Nonparametric kernel density estimation and its computational aspects*. Springer, 2018.

[7] Herrera, D. *Descripción de Modelos ARCH y GARCH: El caso de Netflix, Inc. Tesis de Licenciatura, BUAP*. 2017.

[8] Infante, S., and Zárate, G. *Métodos estadísticos: un enfoque interdisciplinario*, 2da ed. Trillas, 1986.

[9] Lee, H., and Kang, K. Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Advances in Meteorology 2015* (2015).

[10] Miñarro, A. Estimación no paramétrica de la función de densidad. *Documento de Trabajo. Universidad de Barcelona. Barcelona, España* (1998).

[11] Navidi, W. *Estadística para ingenieros y científicos. Ed. Me Graw Hill*. Interamericana. México, 2006.

[12] RAMA. Bases de datos - red automática de monitoreo atmosférico (rama). http://www.aire.cdmx.gob.mx/default.php?opc='27aKBh'27, 2011.

[13] Rincón, L. *Introducción a la probabilidad*. Universidad Nacional Autónoma de México, Facultad de Ciencias, 2014.

[14] Rincón, L. *Curso intermedio de probabilidad*. Universidad Nacional Autónoma de México, Facultad de Ciencias, 2015.

[15] Robert, C., and Casella, G. *Monte Carlo Statistical Methods*, 2da ed. Springer Texts in Statistics. Springer, 2004.

[16] Rodríguez, L. Construcción de kernels y funciones de densidad de probabilidad. *Departamento de Matemáticas, ESPOL*. (2013).

[17] Wand, M., and Jones, C. *Kernel smoothing*. Chapman and Hall/CRC, 1995.

[18] Wand, M., and Ripley, B. Functions for kernel smoothing supporting Wand & Jones(1995) (R package version 2.23-15). http://CRAN.R-project.org/package=KernSmooth, 1999.