Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Método extendido de ajuste de distribuciones

T E S I S

que para obtener el título de:

actuario

P R E S E N T A :

José Carlos Del Valle López

tutor

Act. Hercilio Barragán Anzures

Facultad de Estudios Superiores Acatlán, Edo. Mx., 2019

Agradecimientos

A mi mamá María del Carmen López Sansalvador, por ser la mujer más sobrealiente, excelsa, fuerte y maravillosa de este mundo, que me ha apoyado incondicionalmente, dado todo el amor, paciencia, cariño y ahora la mayor de las herencias: mi carrera universitaria.

A mi familia, mi tio Efraín por brindarme su apoyo en todo momento, a mi tía Lili por aconsejarme sabiamente e impulsarme desde donde esté, a mi hermano José Cecilio por siempre ser un ejemplo de perseverancia y superación, a mis abuelitos por su gran amor, a mis primos, quienes me han dado tantos momentos de felicidad y a todos con quienes he contado incondicionalmente y han estado siempre pendientes de mi.

A mi escuela, la UNAM que me ha dado la más grande oportunidad de la vida, a los mejores amigos, las mejores experiencias, los mejores maestros y aprendizajes invaluables, la que me llena de orgullo y que honraré y agradeceré eternamente.

A mis presentes maestros: Hercilio Barragán Anzures, Miguel Ángel Chávez, Gustavo Fuentes Cabrera a quienes respeto y admiro, que no solo me dieron una enseñanza sino nuevas formas de ver la vida.

A mis antiguos maestros Rodrigo Gámez Manzo, Leonardo Rebollo Pantoja, Fernando Muñoz Razo, Pablo Ruíz Murillo y Guadalupe Sumano Durán, gracias los ejemplos y el aprendizaje que me brindaron, por su pasión por enseñar, por su desempeño como docentes, como profesionales y como personas.

A mis amigos, Luis Orozco Córdoba, Jesús González Moreno, Alan Sánchez, Ruth Lara Castelán y a todos con quienes siempre he contado, en los buenos y malos momentos.

Índice general

Índice de figuras

Capítulo 1

La descomposición como solución

1.1. Introducción

Divide et impera

-Julio Cesar

Se avecinaba el año 1333, el imperio de Constantinopla habría de ser atacado por tres naciones enemigas, siendo superados cinco a uno, el consejero del emperador solicitó una audiencia para responder ante tal amenaza, al siguiente día de decidió liberar a los esclavos de la primera nación opositora, repitiendo esto durante siete días, diciéndoles que llevaran el mensaje: "Su deuda ha sido saldada", la siguiente semana los esclavos de la segunda nación fueron liberados, llevando soldados ocultos entre ellos cuya etnia de origen era compartida.

Al llegar los primeros a su ciudad natal las otras dos naciones aliadas que planeaban el ataque lo notaron y a la llegada de la última caravana fueron acusados de traición y atacados.

La tensión aumentó entre los aliados, los soldados ocultos planearon un ataque durante el cual anunciaban astutamente la traición, en el acto, los soldados arremetieron contra los viajeros quedando solo sus huesos.

Sus fuerzas se habían diezmado y al haberse enfrentado entre sí, al ejército de la gran Constantinopla le tomó poco tiempo sobreponerse contra sus enemigos.

Divide y vencerás, en la guerra, en la ciencia y en la vida, implica enfrentar

y resolver un problema complejo separándolo en componentes más sencillas tantas veces como sea necesario, por ejemplo:

En economía aplicamos la descomposición para realizar el análisis de series de tiempo, pues en el proceso separamos la tendencia, la estacionalidad y el ruido blanco, a cada parte aplicamos pruebas distintas que nos dan un mayor entendimiento de lo que sucede a través del tiempo.

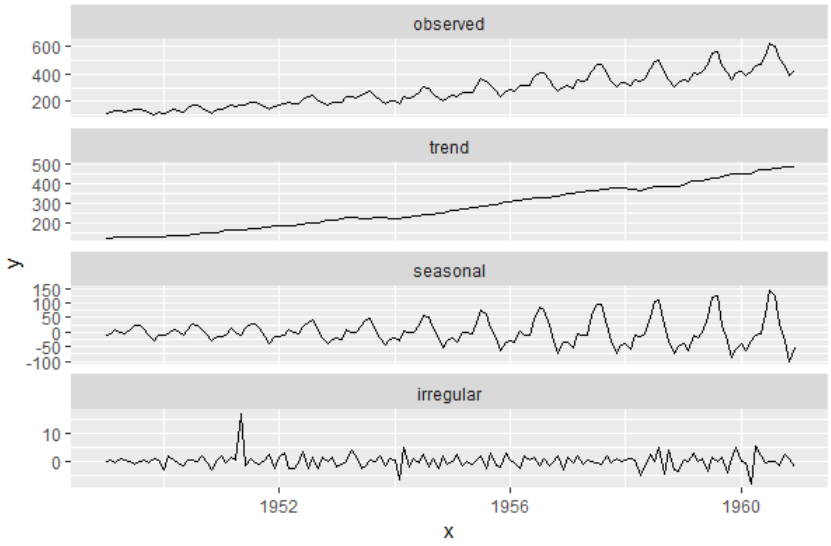


Figura 1.1: Ejemplo de una serie de tiempo en cada una de sus componentes (Fuente data set AirPassengers).

En el análisis multivariado y ciencias computacionales, aplicamos algoritmos de clasificación como Máquina Vector Soporte y K-Medias, para entender cómo puede dividirse cierta muestra y perfilar cada uno de los grupos obtenidos.

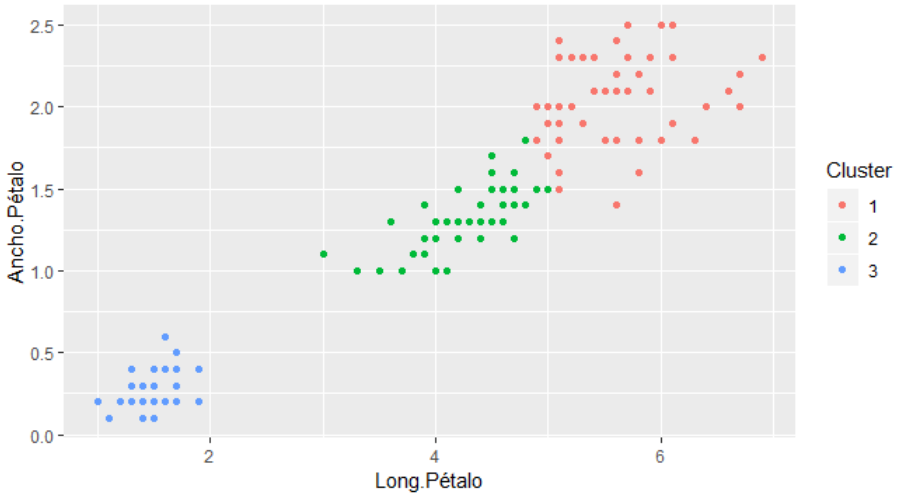


Figura 1.2: Ejemplo del algoritmo K-Medias en las medidas del tallo y pétalo de las flores (Fuente data set Iris).

En la ingeniería se emplean Series de Fourier para analizar una onda de sonido por medio de su descomposición en ondas más simples, dándonos un espectro más amplio de lo que sucede.

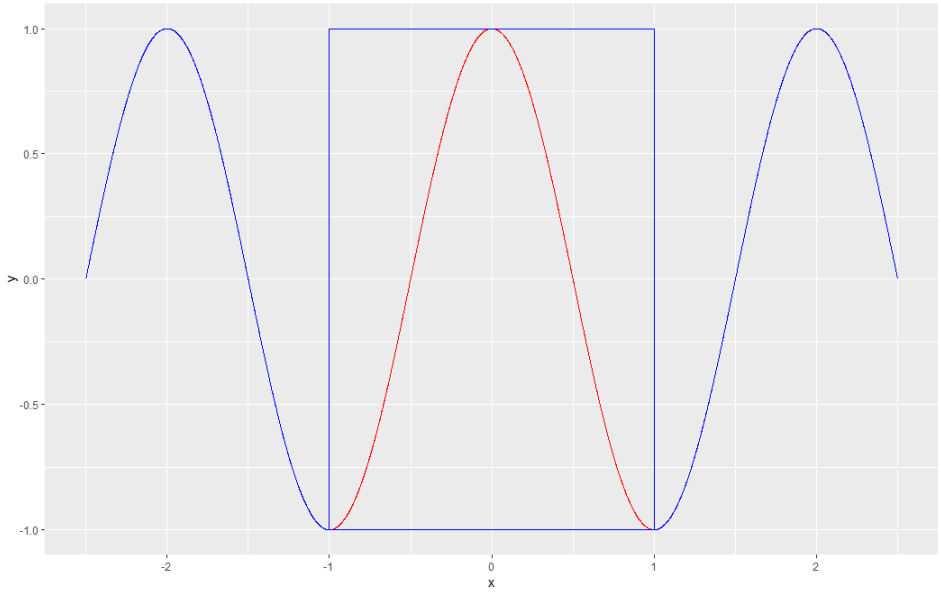


Figura 1.3: Ejemplo de descomposición de una función de distribución.

Si observamos cuidadosamente en este último ejemplo, al aislar una sección de onda veremos la forma de una función de distribución desplazada en el plano, a partir

de esta idea fue que nació este escrito, empezando por conocer métodos de descomposición y emplearlos para separar una distribución en componentes más simples de la forma más óptima y añadiendo el factor aleatoriedad.

Algunos ejemplos en donde observamos este tipo de comportamientos:

Biología: Distribución de las tasas de crecimiento entre los individuos.

Seguros: Identificación de siniestros con una alta suma asegurada.

Finanzas: Valor de una divisa por intervalos estacionales.

Tiempo de una transacción por tarjeta de crédito para la identificación de fraudes

Pricing: Estimación del valor de un producto por medio de la distribución de sus precios (este efecto aumenta cuando hay un incremento brusco en los precios de cierto producto).

Física: La cantidad de luz en el espectro visible, indica la composición química del astro observado, la distribución de dicha luz puede ser representada mediante una variable aleatoria.

Al pensarlo con detenimiento, contamos con las herramientas teóricas necesarias para realizar este análisis y medir la diferencia que existe entre los datos observados y el ajuste a un modelo paramétrico, puesto que en la práctica, si estos fallan, procedemos a calcular estadísticos con modelos no paramétricos o a trabajar bajo hipótesis que empíricamente no se cumplen.

Esta nueva visión es una propuesta al análisis de outliers, a la separación en panza y cola de una función de densidad o una nueva forma de análisis ante la incógnita de no poder ajustar cierta distribución nuestros datos.

Empezaremos por describir qué es una variable aleatoria y casos discretos y continuos, posteriormente, veremos funciones de distribución multimodales cuya forma sugiere una descomposición en partes más simples.

Se verán pruebas de bondad y ajuste para estimar la distribución y parámetros de una muestra observada y se emplearán indicadores para definir si una distribución debe ser o no separada y diversos métodos para clasificar estos grupos.

Habiendo definido lo anterior se describirá una propuesta para ajustar todo tipo de distribuciones y se definirá una forma de evaluarla.

Capítulo 2

Descripción de variables aleatorias

2.1. Introducción

Para entender el concepto de variable aleatoria, introduciremos primero los conceptos de experimento aleatorio y espacio muestral.

Consideremos un experimento cuyo resultado depende completamente del azar, es decir, es desconocido; a este suceso le llamaremos experimento aleatorio, por ejemplo:

1. El resultado de lanzamiento de un dado.

2. El resultado de una canica al girar en la ruleta.

3. El resultado del lanzamiento de una moneda.

Y un espacio muestral es el conjunto de todos los posibles resultados de un experimento aleatorio, siendo los de nuestros ejemplos:

1. S = {1, 2, 3, 4, 5, 6}.

2. S = {1, 2 , ..., 38}.

3. S = {"Águila", "Sol"}.

Al realizar un experimento, nos interesamos en los resultados obtenidos al hacerlo una y otra vez, por ejemplo:

1. La suma del lanzamiento de dos dados.

2. El número de veces que cae en la casilla 38.

3. La cantidad de Águilas después de n lanzamientos.

Estas cantidades o números de interés determinados por un experimento aleatorio son a las que denominamos variables aleatorias.

2.2. Distribuciones Discretas

Decimos que una variable aleatoria X tiene distribución discreta si el rango de X es numerable, es decir, el espacio muestral contiene una cantidad contable de elementos. En la mayoría de los casos dicho rango corresponde a $\mathbb{N}\ U\ \{0\}$. Algunos ejemplos son: el número de reclamaciones que recibe una compañía aseguradora, la cantidad de clientes que consumirán un producto determinado o la cuenta del número de personas que desarrollan una enfermedad.

Definimos la función de masa de probabilidades de la variable aleatoria X como: $f(x) = P(X = x)$ es decir, la probabilidad de que X tome el valor x.

Esta función para el espacio muestral $S = \{x_1, x_2, ...\}$ debe cumplir con:

1. $$f(x_i) \geq 0 \quad \forall i \in S$$

   .

2. $$\sum_{i=1}^{\infty} f(x_i) = 1$$

   .

Gráficamente puede verse de la siguiente forma:

$$P(X = 1) = .19, P(X = 2) = .23, P(X = 3) = .31, P(X = 4) = .27.$$

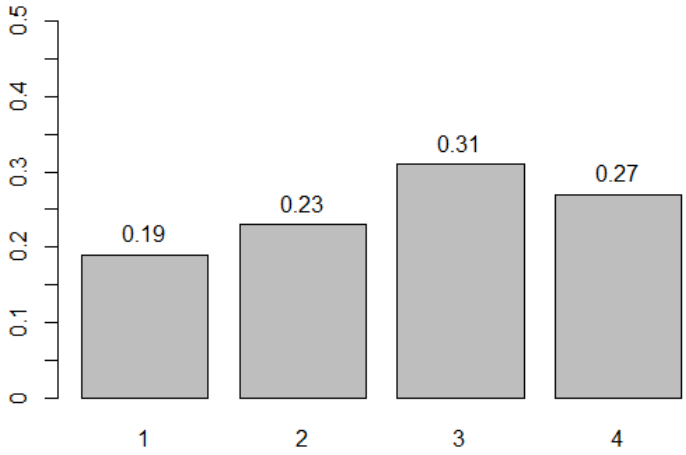


Figura 2.1: Función de Probabilidad Discreta.

La función de distribución acumulada o función de distribución de la variable aleatoria X es representada mediante $F(x) = P(X \leq x)$ significando la probabilidad de que la variable aleatoria X sea menor o igual a un valor dado x.

Definiremos también la esperanza o valor esperado de la variable aleatoria X como $E(X) = \sum_{i=1}^{n} x_i f(x_i)$ y representa el valor promedio que se obtendría al repetir el experimento. A continuación, describiremos una de estas funciones y sus características.

Distribución Poisson.

Sea X una variable aleatoria, decimos que X se distribuye Poisson o:

$X \sim Poisson(\lambda)$ con parámetro $\lambda > 0$.

Si su función de densidad se define como:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \text{ para } x = \{0, 1, 2, ...\}.$$

La esperanza de una variable aleatoria Poisson se define como:

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) \text{ para } X \sim Poisson(\lambda).$$

$$E(X) = \sum_{i=1}^{n} x_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{(x_i-1)!} = \lambda \sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i-1}}{(x_i-1)!}.$$

Si realizamos el cambio de variable $z_i = x_i - 1$ obtenemos que:

$$\sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i-1}}{(x_i-1)!} = \sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{z_i}}{(z_i)!} = 1.$$

por lo tanto,

$$E(X) = \lambda.$$

Y la varianza puede definirse como:

$$V(X) = E(X^2) - E^2(X).$$

Dado que ya conocemos el valor de $E(X) = \lambda$, la incógnita reside en $E(X^2)$, y empleando la fórmula para la esperanza:

$$E(X^2) = E(X^2 - X + X) = E((X(X-1) + X)) = E(X(X-1)) + E(X).$$

$$E(X(X-1)) = \sum_{i=1}^{n} x_i(x_i-1)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \lambda^2 \sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^{(x_i-2)}}{(x_i-2)!}.$$

Aplicamos nuevamente un cambio de variable con $z_i = x_i - 2$ y llegaremos a que

$$E(X(X-1)) = \lambda^2$$

entonces,

$$E(X^2) = \lambda^2 + \lambda$$

por lo tanto,

$$V(X) = \lambda^2 + \lambda \ - \lambda^2 = \lambda.$$

La función generadora de momentos o función generatriz de momentos de una variable aleatoria X se define como:

$$M_x(t) = E(e^{tX}), t \in \mathbb{R}.$$

Para este caso

$$M_x(t) = \sum_{i=1}^{n} e^{tx_i} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-\lambda} \sum_{i=1}^{n} \frac{(\lambda e^t)^{x_i}}{x_i!}.$$

Si recordamos el desarrollo en serie de la función exponencial, veremos que

$$e^t = \sum_{x=0}^{\infty} \frac{t^x}{x!}, \forall t \in \mathbb{R}.$$

En este punto, hay que recalcar que para la esperanza

$$\sum_{i=0}^{\infty} x_i f(x_i) = \sum_{i=0}^{n} x_i f(x_i) + \sum_{i=n+1}^{\infty} x_i f(x_i).$$

en dónde,

$$\sum_{i=n+1}^{\infty} x_i f(x_i) = 0.$$

dado que todos los valores de $f(x_i) = 0$ para valores que estén fuera del espacio muestral, entonces, al utilizar la forma en serie de la función exponencial llegamos a:

$$M_x(t) = e^{-\lambda} e^{\lambda e^t} = exp(\lambda(e^t - 1)).$$

Visualización de una variable aleatoria $X \sim Poisson(1)$:

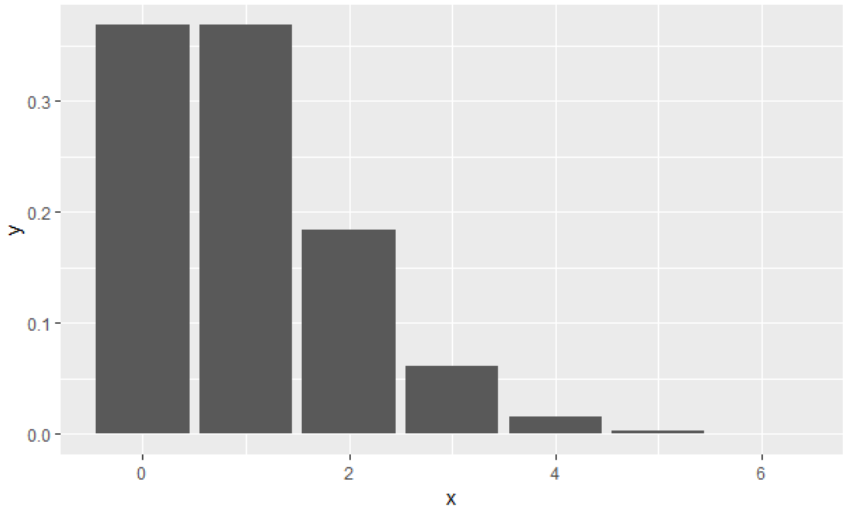


Figura 2.2: Distribución Poisson.

Ejemplo de código en R para la función generadora de números aleatorios, función de distribución y función de densidad respectivamente:

```
>set.seed(31109)
>rpois(n = 1,lambda = 2)
[1] 2

>ppois(q = 1,lambda = 1)
[1] 0.7357589

>dpois(x = 1,lambda = 3)
[1] 0.1493612
```

Ejemplificaremos entonces el uso de esta distribución:

Una empresa observa el número de clientes que requieren de sus servicios durante 9 horas, el número de clientes siguie una distribución Poisson y se sabe que el promedio que reciben es de 15.

¿Cuál es la probabilidad de que lleguen 20 clientes?

¿Cuál es la probabilidad de que no lleguen clientes en la primera hora?

¿Cuál es la función que describe que llegue al menos un cliente en cualquier can-

tidad de horas?

Primero debemos observar que el valor λ corresponde tanto a la esperanza como a la varianza de la distribución, entonces:

$$f\left(x\right) = \frac{e^{-15}15^{x}}{x!}.$$

Ahora ya somos capaces de responder la primera pregunta:

$$f\left(20\right) = \frac{e^{-15}15^{20}}{20!} = 0.04181.$$

Supongamos que los clientes llegan de forma regular durante esas 9 horas, es decir, que esperamos una media de 5 clientes al cabo de 3 horas, por lo que deberemos tratar con una distribución con parámetro $\lambda_k = 5$, donde k es el número de clientes que habrán llegado a las 3 horas, así, nuestra función de densidad toma la siguiente forma:

$$f\left(x\right) = \frac{e^{-5}5^{x}}{x!}.$$

entonces, la probabilidad de que no lleguen clientes es:

$$f\left(0\right) = \frac{e^{-5}5^{0}}{0!} = e^{-5} = 0.00674.$$

La última pregunta sugiere el siguiente planteamiento: $P(X > 0)$, es decir, la probabilidad de que llegue al menos un cliente, sin embargo, podemos plantearlo como $1 - P(X = 0)$, es decir, que tome todos los valores exceptuando el 0, y sustituyendo, obtenemos:

$$1 - f(0) = 1 - \frac{e^{-\lambda}\lambda^{0}}{0!} = 1 - e^{-\lambda}.$$

Y entonces, la función que buscamos con $\lambda =$ número de horas es:

$$f(\lambda) = 1 - e^{-\lambda}.$$

2.3. Distribuciones Continuas

Estas variables aleatorias son definidas de forma distinta, pues no es posible calcular la probabilidad puntual de la misma, es decir, $P(X = a)$, no obstante, es posible definir la probabilidad acumulada hasta cierto valor como $P(X \leq a)$, en otras palabras, la probabilidad de que el valor X sea menor a a. Por ejemplo, que el valor de los montos de la reclamación de una póliza sea menor a cierta cantidad, o que la suma de las ventas de un producto en un día sea mayor a un monto o que el tiempo promedio de vida de una persona sea mayor a tantos años y a diferencia de una distribución discreta, la esperanza se define como:

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

Distribución Exponencial.

Sea X una variable aleatoria, decimos que X se distribuye Exponencial o:

$X \sim exp(\lambda)$ con parámetro $\lambda > 0$ si su función de distribución acumulada se define como:

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{e.o.c} \end{cases}$$

es decir, su función de densidad se define como:

$$f(x) = P(X = x) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{e.o.c} \end{cases}.$$

Esperanza:

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

Que para este ejemplo resulta:

$$E(X) = \int_{-\infty}^{0} x * 0dx + \int_{0}^{\infty} x\lambda e^{-\lambda x}dx = \int_{0}^{\infty} x\lambda e^{-\lambda x}dx.$$

$$= [-xe^{-\lambda x}]_0^\infty + \int_0^\infty \lambda e^{-\lambda x} dx = 0 + [\frac{-1}{\lambda} e^{-\lambda x}]_0^\infty = 0 + \frac{1}{\lambda}.$$

Por lo tanto,

$$E(X) = \frac{1}{\lambda}.$$

Varianza:

La definición de la varianza en función de la esperanza es análoga a la de las variables discretas.

$$V(X) = E(X^2) - E^2(X).$$

Por lo que deberemos hallar el valor de $E(X^2)$ que por definición es:

$$E(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = [-x^2 e^{-\lambda x}]_0^\infty + \int_0^\infty 2xe^{-\lambda x} dx.$$
$$= 0 + \frac{2}{\lambda} \int_0^\infty xe^{-\lambda x} dx = \frac{2}{\lambda} E(X) = \frac{2}{\lambda^2}.$$

entonces,

$$V(X) = E(X^2) - E^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2}.$$

por lo tanto,

$$V(X) = \frac{1}{\lambda^2}.$$

En cuanto a la función generadora de momentos la definición se conserva:

$$M_x(t) = E(e^{tX}), t \in \mathbb{R}.$$

Para este caso:

$$M_x(t) = \int_0^\infty exp(tx) \lambda e^{-\lambda x} dx = \lambda \int_0^\infty exp((t - \lambda)x) dx.$$

$$\lambda[\frac{exp((t-\lambda)x)}{t-\lambda}]_0^\infty = \frac{\lambda}{\lambda - t}.$$

Hay que destacar que para que el valor de la ecuación se haga 0 cuando x tienda a infinito, se debe cumplir que $t - \lambda < 0$, de esa forma el factor al evaluar en infinito se convierte en 0.

A continuación, podemos ver una representación gráfica de una variable aleatoria $X \sim exp(1)$.

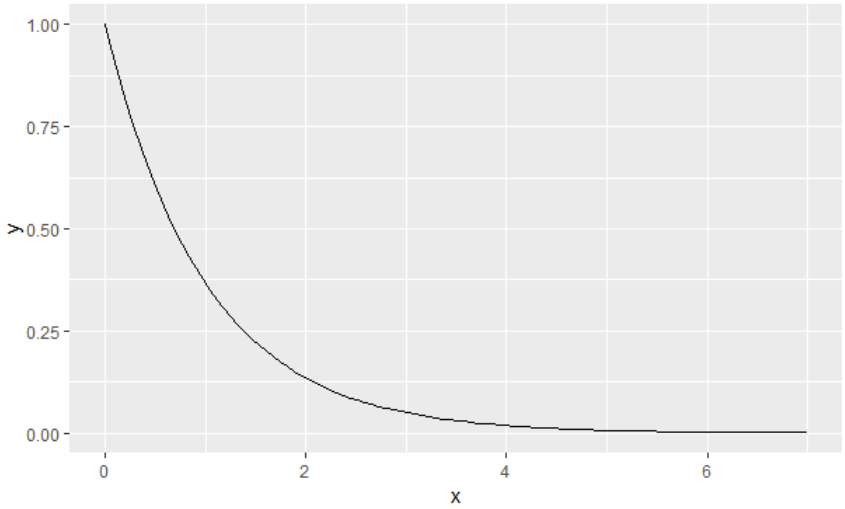


Figura 2.3: Distribución Exponencial.

Algunas aplicaciones para esta distribución son: El tiempo que tarda una máquina en proceso de producción en crear un objeto es de 9 minutos, supongamos que este sigue una distribución exponencial.

¿Cuál es la probabilidad de que la máquina tarde menos de 7 minutos?.

¿Cuál es la probabilidad de que la máquina tarde más de 15 minutos?.

Aquí también observamos que el parámetro $1/\lambda$ corresponde con la media de la distribución, es decir, la función de distribución acumulada se define como:

$$F(x) = \begin{cases} 1 - e^{-x/9}, & \text{si } x \geq 0 \\ 0 & \text{e.o.c} \end{cases}$$

Ahora podemos obtener las repuestas que buscábamos:

$$F(7) = \begin{cases} 1 - e^{-7/9}, & \text{si } x \geq 0 \\ 0 & \text{e.o.c} \end{cases} = 0.5405742.$$

Y análogamente:

$$1 - F(15) = \begin{cases} e^{-15/9}, & \text{si } x \geq 0 \\ 0 & \text{e.o.c} \end{cases} = 0.1888756.$$

2.4. Distribuciones Bimodales

Ya sentadas las bases, podemos entrar en un terreno de estudio más escabroso en el que la distribución de los datos parecería estar compuesta por la suma o composición de dos o más variables aleatorias, que tienen la peculiaridad de no ser fácilmente representadas mediante funciones de dos o menos parámetros y que tienen un mayor número máximos locales o modas que cambian el orden ascendente o descendente de la misma.

En la naturaleza es difícil hallar objetos de estudio o comportamientos que se ajusten a una varibale aleatoria si no se trata de un experimento controlado, es frecuente que nos lleguemos a encontrar con comportamientos más atípicos a los que pareciera que no se le puede ajustar una distribución pero que resulta necesario hacerlo.

Estos casos, se suelen atender separando la distribución de los valores extremos (outliers) que impiden su ajuste y posteriormente añadirlos al análisis, proceso que no siempre es posible debido a la forma de la distribución, para ello, son introducidas las distribuciones bimodales o en algunos casos encontradas como variables aleatorias mixtas. El término bimodal proviene del sufijo bi que significa dos y modal que significa moda, y son aquellas distribuciones que tienen dos modas y que gráficamente tienen más de un valor máximo local tal y como se muestra a continuación:

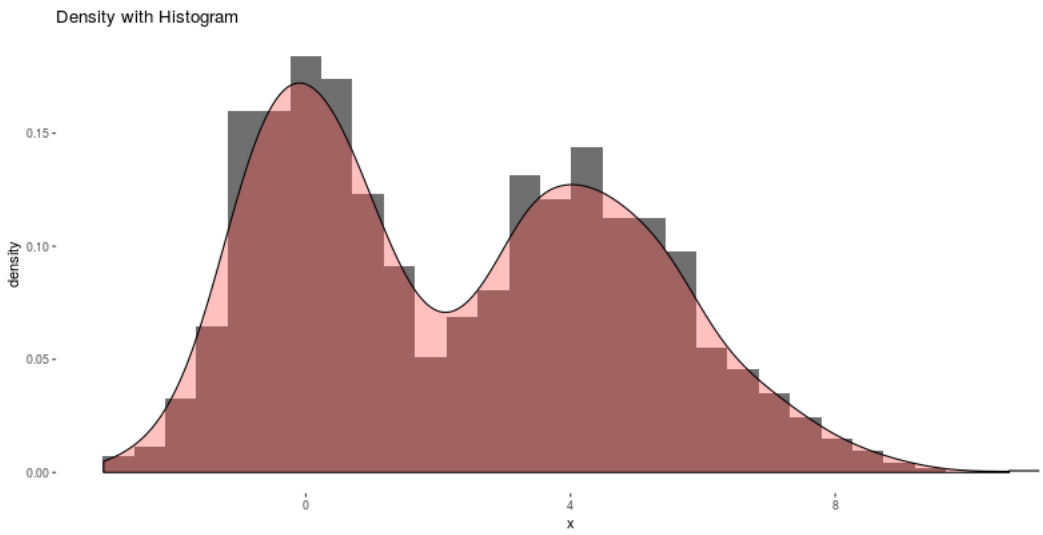


Figura 2.4: Distribución Bimodal

Un primer enfoque consiste en suponer que estamos observando una distribución cuya función de distribución está compuesta por dos funciones más simples.

$$F(x) = pF1(x) + (1-p)F2(x).$$

Para la gráfica anterior estamos uniendo una distribución Normal estándar y una función Gamma(9,2) en partes iguales (p=0.5), por lo que se puede sustituir de la siguiente forma:

$$F(x) = \frac{F1(x)}{2} + \frac{F2(x)}{2}.$$

El cálculo de dicha función de distribución no es para nada trivial, pues las componentes de la función toman los valores:

$$F1(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

$$F2(x) = \int_0^x \frac{u^{k-1} e^{-\frac{u}{\theta}}}{\theta^k \Gamma(k)} \quad du \text{ para } x > 0 \text{ y } k, \theta > 0.$$

Es por ello que debemos asegurarnos que no hemos hallado una distribución que cumpla con las hipótesis de las pruebas de bondad y ajuste, pues una distribución

bimodal requiere de un cálculo analítico complejo y que puede no tener una expresión analítica explícita, tal y como se muestra en el ejemplo anterior.

2.4.1. Distribución Beta-Normal

En la literatura es común encontrar funciones bimodales compuestas por dos distribuciones continuas, es el caso de la Normal-Normal, Beta-Normal, Weibull-Normal, entre otras, los nombres de dichas distribuciones indican la forma o distribución de cada componente asociada a las modas observadas o en este caso una composición de funciones.

Sea F(x) la función de distribución acumulada de la variable aleatoria X. La función de distribución acumulada para una clase generalizada de distribuciones para X puede definirse de la siguiente forma:

$$G(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{F(x)} t^{\alpha-1}(1-t)^{\beta-1}dt \text{ con } \alpha > 0, \beta < \infty.$$

La función de densidad de la clase generalizada de distribuciones G(x) es

$$g(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} F(x)^{\alpha-1}(1-F(x))^{\beta-1}F'(x).$$

Donde F(x) es la función de distribución acumulada de una variable aleatoria normal.

Entonces, decimos que la variable aleatoria X se distribuye Beta-Normal cuando su función de distribución acumulada es:

$$BN(\alpha,\beta,\mu,\sigma) = G(\Phi(\frac{x-\mu}{\sigma})) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Phi(\frac{x-\mu}{\sigma})^{\alpha-1}(1-\Phi(\frac{x-\mu}{\sigma}))^{\beta-1}\sigma^{-1}\phi(\frac{x-\mu}{\sigma}).$$

Esta distribución se caracteriza por tener cuatro parámetros que juntos describen la localización, la escala y la forma.

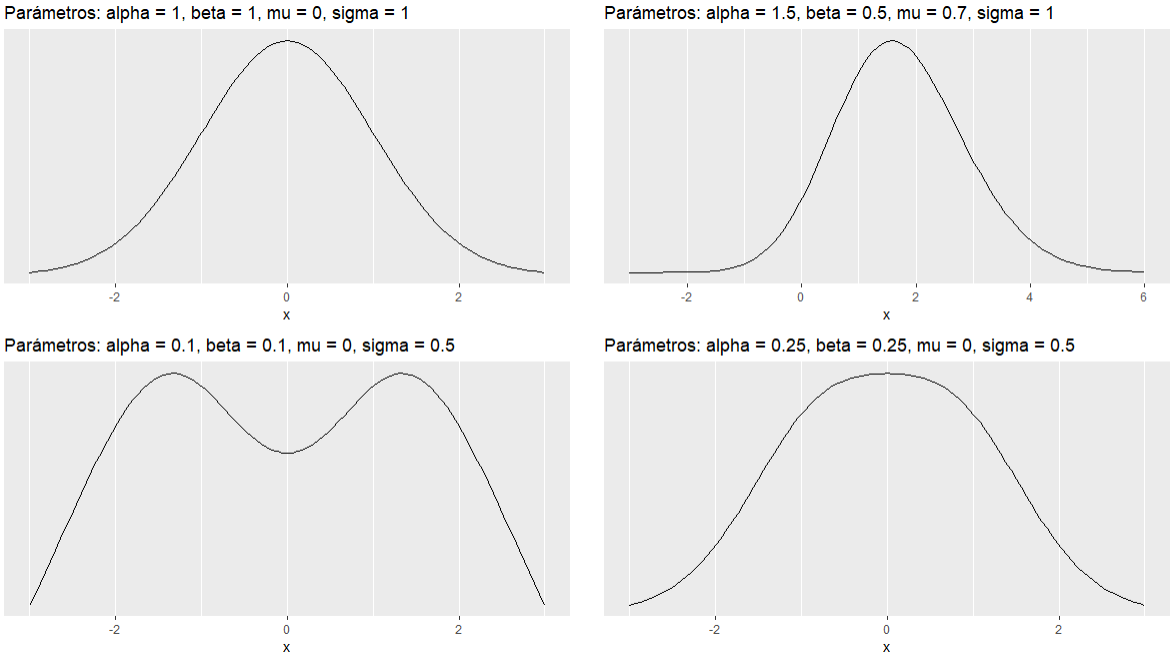


Figura 2.5: Beta-Normal simétrica.

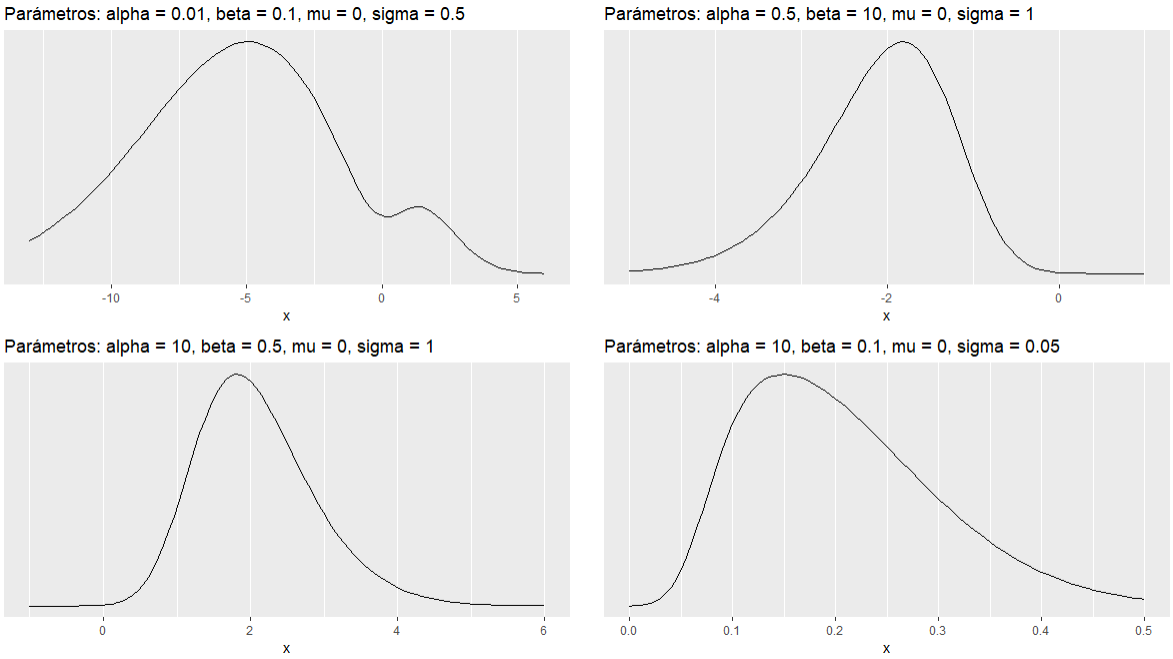


Figura 2.6: Beta-Normal asimétrica.

Ya que hemos tenido una breve visión de nuestra distribución, realizaremos el cálculo de la esperanza. Dado que es una expresión con un gran número de parámetros, definiremos la esperanza para algunos valores de α y β

Una segunda forma de escribir la distribución beta-normal es la siguiente:

$$\frac{dG_0}{dx} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\{\Phi(\frac{x-\mu}{\sigma})\}^{\alpha-1}\{1-\Phi(\frac{x-\mu}{\sigma})\}^{\beta-1}\sigma^{-1}\phi(\frac{x-\mu}{\sigma}).$$

Definimos

$$G_{i+1}(x) = exp((x-\mu)^2/2\sigma^2)\frac{dG_i}{dx}\forall i = 0,1,2,...$$

Entonces,

$$G_1(x) = exp((x-\mu)^2/2\sigma^2)\frac{dG_0}{dx}.$$

$$G_2(x) = exp((x-\mu)^2/2\sigma^2)\{\frac{d^2G_0}{dx^2} + \frac{x-\mu}{\sigma^2}\frac{dG_0}{dx}\}.$$

$$\iff \int_{-\infty}^{\infty}\frac{d^2G_0}{dx^2}dx + \int_{-\infty}^{\infty}\frac{(x-\mu)}{\sigma^2}\frac{dG_0}{dx}dx = \int_{-\infty}^{\infty}(exp(-(x-\mu))^2/\sigma^2)G_2(x)dx.$$

Dado que la primera integral en el lado izquierdo es cero, obtenemos

$$E(X) = \int_{-\infty}^{\infty}x\frac{dG_0}{dx}dx = \mu + \sigma^2\int_{-\infty}^{\infty}(exp(-(x-\mu)^2/2\sigma^2)\frac{dG_1(x)}{dx})dx.$$

$$I_n(a) = \int_{-\infty}^{\infty}\Phi(a(x-\mu)+\mu)^n \exp\left(-(x-\mu)^2/\sigma^2\right)dx.$$

Cuando n es un número impar de la forma $n = 2m+1$ el integrando de la función anterior es también impar y podemos expresarlo de la siguiente manera:

$$\int_{-\infty}^{\infty}\left\{\Phi[a(x-\mu)+\mu] - \frac{1}{2}\right\}^n \exp\left(-(x-\mu)^2/\sigma^2\right)dx = 0$$

Utilizando el teorema Binomial en el resultado anterior obtenemos:

$$\int_{-\infty}^{\infty} \sum_{i=0}^{2m+1} (-1)^i (\Phi(a(x-\mu)+\mu))^{2m+1-i} \left(\frac{1}{2}\right)^i \begin{pmatrix} 2m+1 \\ i \end{pmatrix} \exp\left(-(x-\mu)^2/\sigma^2\right) dx = 0.$$

Gacias a la propiedad de convergencia uniforme, podemos reescribir la integral de la suma como la suma de las integrales.

$$\sum_{i=0}^{2m-1} \int_{-\infty}^{\infty} (-1)^i (\Phi[a(x-\mu)+\mu])^{2m+1-i} \left(\frac{1}{2}\right)^i \begin{pmatrix} 2m+1 \\ i \end{pmatrix} \exp\left(-(x-\mu)^2/\sigma^2\right) dx = 0.$$

Entonces,

$$\begin{aligned} I_{2m+1}(a) &= \int_{-\infty}^{\infty} (\Phi(a(x-\mu)+\mu))^{2m+1} \exp\left(-(x-\mu)^2/\sigma^2\right) dx \\ &= \sum_{i=1}^{2m+1} (-1)^{i+1} (\Phi(a(x-\mu)+\mu))^{2m+1-i} \left(\frac{1}{2}\right)^i \begin{pmatrix} 2m+1 \\ i \end{pmatrix} \cdot \\ &\quad \exp\left(-(x-\mu)^2/\sigma^2\right) \\ &= \sum_{i=1}^{2m+1} (-1)^{i+1} 2^{-i} \begin{pmatrix} 2m+1 \\ i \end{pmatrix} I_{2m+1-i}(a) \end{aligned}$$

De esta última fórmula, podemos obtener las soluciones para $n = 1$ y $n = 3$ (recordemos que es para n impar).

$$I_1(1) = \frac{1}{2} I_0(1) = \frac{\sigma\sqrt{\pi}}{2}.$$

$$I_3(1) = \frac{3}{2} I_2(1) - \frac{1}{4} I_0(1).$$

Cuando n es par, no hay una forma analítica concreta de $I_n(1)$ para $n > 2$ Así pues, obtendremos el valor para $n = 2$ derivando $I_n(a)$ respecto a a

$$\begin{aligned} I_2'(a) &= \int_{-\infty}^{\infty} 2\Phi(a(x-\mu)+\mu)\Phi'(a(x-\mu)+\mu)\exp\left(-(x-\mu)^2/\sigma^2\right)dx \\ &= \int_{-\infty}^{\infty} 2\Phi(a(x-\mu)+\mu)\left(\frac{x-\mu}{\sigma\sqrt{2\pi}}\right)\exp\left(-\left(a^2+2\right)(x-\mu)^2/2\sigma^2\right)dx \end{aligned}.$$

Integraremos por partes con

$u = \Phi(a(x-\mu)+\mu)$

$dv = (x-\mu)\exp\left(-\left(a^2+2\right)(x-\mu)^2/2\sigma^2\right)$

Así obtenemos

$$\begin{aligned} I_2'(a) &= \int_{-\infty}^{\infty} 2\Phi(a(x-\mu)+\mu)\Phi'(a(x-\mu)+\mu)\exp\left(-(x-\mu)^2/\sigma^2\right)dx \\ &= \frac{a}{\left(a^2+2\right)\pi}\int_{-\infty}^{\infty}\exp\left(-\left(a^2+1\right)(x-\mu)^2/\sigma^2\right)dx \\ &= \frac{a\sigma}{\left(a^2+2\right)\sqrt{a^2+1}\sqrt{\pi}} \end{aligned}.$$

Por lo tanto,

$$I_2(a) = \int_{-\infty}^{\infty}\frac{a\sigma}{\left(a^2+2\right)\sqrt{a^2+1}\sqrt{\pi}} = \frac{\sigma}{\sqrt{\pi}}\arctan\sqrt{a^2+1}.$$

Obtendremos los valores para $\alpha = 4$ y $\beta = 2$.

$$\frac{dG_0}{dx} = \frac{\Gamma(6)}{\Gamma(4)\Gamma(2)}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^3\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}\sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)$$

también,

$$G_1(x) = \frac{20}{\sigma\sqrt{2\pi}}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^3\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}.$$

Entonces,

$$\begin{aligned}\frac{dG_1}{dx} =& \frac{60}{\sigma\sqrt{2\pi}}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^2\sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)\\ &-\frac{80}{\sigma\sqrt{2\pi}}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^3\sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)\end{aligned}.$$

Ahora poseemos todas las herramientas para el cálculo de $E(X)$:

$$E(X) = \mu + \sigma^2\int_{-\infty}^{\infty}\exp\left(-(x-\mu)^2/2\sigma^2\right)\frac{dG_1}{dx}.$$

Y utilizando la definición de $I_n(a)$:

$$\begin{aligned}=&\mu + \frac{30}{\pi}\int_{-\infty}^{\infty}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^2\exp\left(-\left(\frac{x-\mu}{\sigma}\right)^2\right)dx\\ &-\frac{40}{\pi}\int_{-\infty}^{\infty}\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^3\exp\left(-\left(\frac{x-\mu}{\sigma}\right)^2\right)dx\\ =&\mu + \frac{30}{\pi}I_2(a) - \frac{40}{\pi}I_3(a)\\ =&\mu + \frac{30\sigma}{\pi\sqrt{\pi}}\arctan\sqrt{2} - \frac{40}{\pi}\left[\frac{3}{2}I_2(a) - \frac{1}{4}I_0(a)\right]\\ =&\mu + \frac{30\sigma}{\pi\sqrt{\pi}}(\arctan\sqrt{2}) - \frac{60\sigma}{\pi\sqrt{\pi}}(\arctan\sqrt{2}) + \frac{10}{\pi}I_0(a)\\ =&\mu + \frac{30\sigma}{\pi\sqrt{\pi}}(\arctan\sqrt{2}) - \frac{60\sigma}{\pi\sqrt{\pi}}(\arctan\sqrt{2}) + \frac{10\sigma}{\sqrt{\pi}}\\ =&\mu + \frac{10\sigma}{\sqrt{\pi}} - \frac{30\sigma}{\pi\sqrt{\pi}}\arctan\sqrt{2}\end{aligned}.$$

2.4.2. Región de bimodalidad

Como vimos en las gráficas, la distribución beta-normal puede ser bimodal para ciertos valores de los parámetros α y β. La solución analítica de α y β para saber si la distribución es bimodal en muchos casos no puede ser resuelta algebraicamente, no obstante, existen indicadores como el coeficiente de bimodalidad que veremos más adelante que nos pueden ayudar a tratar este problema. A continuación, veremos una representación de la región en donde los parámetros vuelven bimodal a nuestra distribución, recordemos que μ y σ tienen la misma interpretación que en una variable aleatoria normal, el posicionamiento de la media y la magnitud de la desviación estándar.

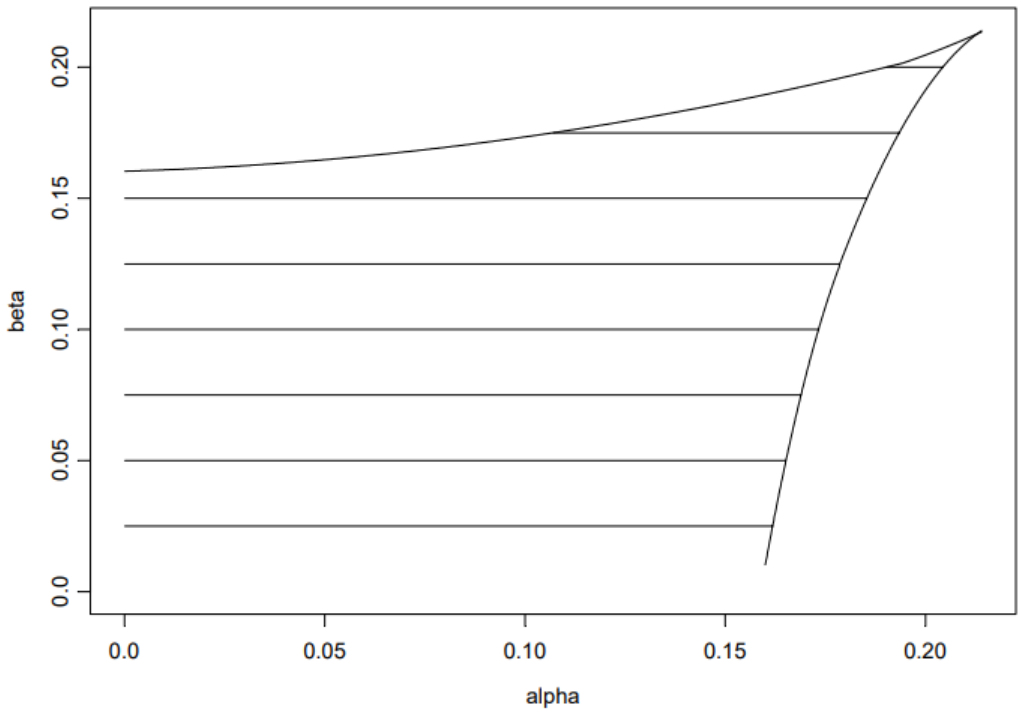


Figura 2.7: Región de Bimodalidad para BN(α,β,0,1) [FAMOYE, LEE & EUGENE].

2.4.3. Multimodalidad

Al igual que la bimodalidad, el término proviene de las palabras multi y moda y es una generalización de la definición de bimodalidad, pudiendo haber dos o más modas dentro de la distribución.

La cantidad de parámetros de una distribución multimodal son un arma de doble filo, pues ayudan a explicar el comportamiento de un conjunto de datos específicos pero al añadir más información esta puede perder su capacidad de generalizar y hacer más compleja nuestra distribución.

Capítulo 3

Estimación y Ajuste

3.1. Introducción

Hay muchas formas de aproximar una variable aleatoria a observaciones que vemos en la naturaleza, muestras de características humanas y en general, conjuntos de datos que puedan ser medidos o contados, en esta sección, exploraremos los enfoques clásicos que nos han ayudado a tener una mayor certidumbre sobre los parámetros que definen a estas variables aleatorias.

3.2. Estimación de parámetros

Antes de pensar en asignar una distribución a nuestros datos, es necesario conocer el valor de ciertos estadísticos de nuestra muestra que funjan como parámetros para la distribución que deseamos ajustar.

3.2.1. Máxima Verosimilitud

La estimación por máxima verosimilitud (EMV) es un método para estimar los parámetros de una muestra observada y ajustarlos a una función de probabilidad.

Sea $(x_1, ..., x_n)$ un vector aleatorio cuya distribución depende del parámetro desconocido θ.

La función de verosimilitud del vector $(x_1, ..., x_n)$ es:

$$L(\theta) = f_{X_1,...,X_n}(x_1, ..., x_n; \theta).$$

si $X_1, ..., X_n$ son independientes, entonces,

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

y si son idénticamente distribuidas:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

El objetivo de este método es encontrar el valor de θ que maximice la función de verosimilitud, que representa la distribución conjunta de nuestro vector aleatorio, a este valor se le llama ëstimador de máxima verosimititudÿ será el valor que le habremos de asignar al parámetro θ.

Recordemos que para encontrar el máximo de una función se debe hallar la derivada de esa función e igualarla a 0, también es válido aplicar transformaciones a la función siempre y cuando estas sean crecientes no afectando el valor máximo de la función, entonces podemos buscar los siguientes resultados:

$$\frac{\partial(L(\theta))}{\partial\theta} = \frac{\partial(\prod_{i=1}^{n} f(x_i; \theta))}{\partial\theta} = 0.$$

De igual forma,

$$\frac{\partial(log(\prod_{i=1}^{n} f(x_i; \theta)))}{\partial\theta} = \frac{\partial(\sum_{i=1}^{n} log(f(x_i; \theta)))}{\partial\theta} = 0.$$

A esta función también se le llama Log Verosimilitud, veamos entonces el siguiente ejemplo:

Sea $(x_1, ..., x_n)$ una m.a proveniente de una distribución $X \sim N(\mu, \sigma)$, independientes e idénticamente distribuidas, ¿cuál es el EMV para el parámetro μ?

La función de densidad de una distribución normal es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

entonces,

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}.$$

Aplicaremos la derivada a la función de Log Verosimilitud para obtener el estadístico, obteniendo:

$$L(\mu) = (\frac{1}{\sigma\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2).$$

entonces,

$$log(L(\mu)) = -nlog(\sigma\sqrt{2\pi}) - (\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2).$$

Ahora derivemos respecto al parámetro que deseamos, en este caso μ e igualemos a 0

$$\frac{\partial log(L(\mu)))}{\partial\theta} = 0.$$

$$log(L(\mu)) = -(\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2\,(x_i-\mu)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0.$$

entonces,

$$\sum_{i=1}^{n}(x_i-\mu) = -n\mu + \sum_{i=1}^{n}(x_i) = 0.$$

que sucede si y solo si,

$$\mu = \frac{\sum_{i=1}^{n}(x_i)}{n} = \bar{X}.$$

Por lo tanto, el EMV para el parámetro μ es la media de la distribución, que gráficamente se puede ver de la siguiente forma.

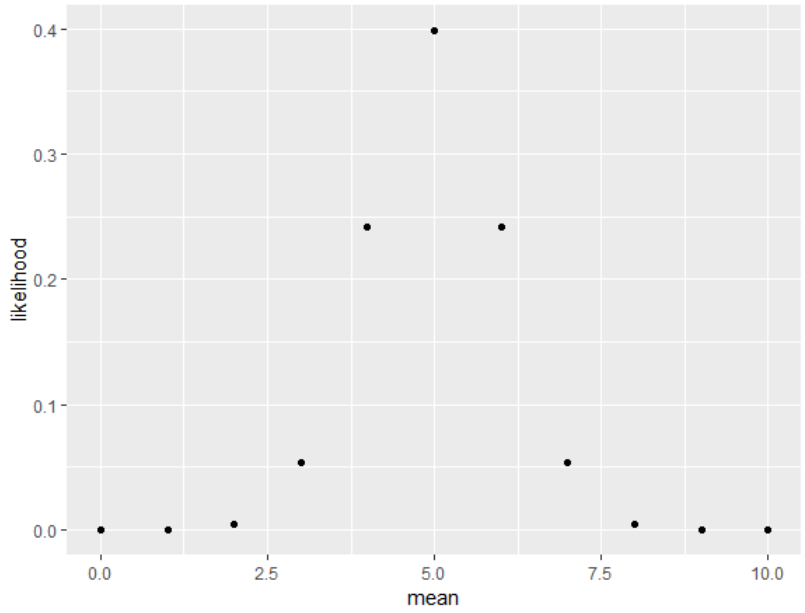


Figura 3.1: Estimador de máxima verosimilitud para una distribución Normal(0,1).

3.2.2. Método de momentos

Si bien el método de Máxima Verosimilitud nos ofrece certidumbre sobre el posible valor del parámetro desconocido θ, también existe una forma más intuitiva por medio de la función generadora de momentos de una función de densidad.

Para abordar este método, es necesario recordar dos definiciones importantes,

Sea X una v.a y sea k un entero mayor que 0. El k-ésimo momento de x, si existe, es el número obtenido de $E(X^k)$, también se le llama momento poblacional.

Sea $x_1, ..., x_n$ una m.a. de la distribución $f(x;\theta)$ y sea k un entero mayor que 0, definimos al k-ésimo momento muestral como la variable aleatoria.

$$\frac{\sum_{i=1}^{n}\left(x_i^k\right)}{n}.$$

Este método consiste en igualar los momentos muestrales y los poblacionales y resolver el sistema de ecuaciones generado para los parámetros que se desean obtener, igualando tantas ecuaciones como número de parámetros a calcular.

$$E(X^k) = \frac{\sum_{i=1}^{n} \left(x_i^k\right)}{n}.$$

Veamos ahora un ejemplo:

Sea X una v.a. con parámetro desconocido $\theta > 0$ y función de densidad

$$f(x) = \begin{cases} \theta x^{\theta-1}, & \text{si } x \in (0,1) \\ 0, & \text{e.o.c} \end{cases}$$

Dado que solo deseamos obtener un parámetro, será necesario obtener el primer momento y resolver la siguiente ecuación para $k = 1$.

$$E(X) = \frac{\sum_{i=1}^{n} (x_i)}{n}.$$

Es sencillo ver que

$$E(X) = \int_0^1 x\theta x^{\theta-1} = \theta[\frac{x^{\theta+1}}{\theta+1}]_0^1 = \frac{\theta}{\theta+1}.$$

entonces, $\frac{\widehat{\theta}}{\widehat{\theta}+1} = \bar{X}$ y despejando el parámetro desconocido $\widehat{\theta}$, llegamos a

$$\widehat{\theta} = \frac{\bar{X}}{\bar{X}+1}.$$

3.3. Pruebas de Bondad y Ajuste

Ya que sabemos maneras de estimar los parámetros para una distribución que queramos ajustar, es necesario aplicar pruebas de bondad y ajuste para corroborar la certidumbre de nuestra distribución, estas pruebas describen que tan bien se ajusta una muestra observada respecto a un modelo teórico, para ello deberemos emplear contraste de hipótesis.

3.3.1. Pruebas de hipótesis

Una hipótesis es una afirmación expuesta a ser o no rechazada acerca de una característica de nuestros datos y generalmente basada en observaciones reales, por ejemplo, que la media muestral es igual a 0 lo cual se denota como $H_0 : \mu = 0$

Es imperativo remarcar que una hipótesis está en constante verificación, por lo que no se puede estar completamente convencido de que esta es acepada dada la naturaleza aleatoria de nuestros datos. Ahora recapitularemos brevemente las pruebas de hipótesis.

Hipótesis nula

Ésta se denota como H_0, es la proposición que se desea rechazar, en la que se declara un valor y se contrasta con un parámetro o estimador de nuestra muestra observada, por ejemplo:

$$H_0 : \mu = \mu_0.$$

$$H_0 : \mu \geq \mu_0.$$

Hipótesis Alternativa

La Hipótesis Alternativa, se denota como H_1, esta se puede verificar en base a la evidencia de la muestra y su región debe abarcar el complemento de nuestra hipótesis nula siendo:

$$H_0 : \mu \neq \mu_0.$$

$$H_0 : \mu < \mu_0.$$

Las hipótesis alternativas del anterior ejemplo respectivamente

Tipos de error

Podemos encontrar cuatro posibles situaciones dados los resultados que arrojan las hipótesis representadas en el siguiente cuadro:

	H_0 Verdadera	H_0 Falsa
Rechazamos H_0	Error Tipo I P(ET I) = α	Decisión Correcta
No Rechazamos H_0	Decisión Correcta	Error Tipo II P(ET II) = β

La Probabilidad de cometer un Error Tipo I se conoce como Nivel de Significancia, se denota como α y es el tamaño de la región de rechazo, este corresponde al conjunto de valores tales que, si la prueba estadística cae dentro de este rango, decidimos rechazar la Hipótesis Nula

3.3.2. Kolmogorov-Smirnov

Este test nos ayudará a ver las diferencias entre dos distribuciones de probabilidad distintas para determinar si tienen o no la misma distribución planteando las siguientes hipótesis:

$$H_0 : F_X(x) = F_Y(x), \forall x \in R.$$

$$H_1 : F_X(x) \neq F_Y(x).$$

En este caso, nos interesa no rechazar la hipótesis nula, es decir, que el p.valor sea mayor a α, al que le asignaremos el valor de 0.05, este valor no es absoluto y no hay un criterio estricto para definir si es o no suficiente para rechazar la hipótesis nula.

Primero necesitamos definir la distribución empírica Sea $x_1, ..., x_n$ una m.a., la distribución empírica se define como

$$F_e(x) = \frac{\#\{i|X_i \leq x\}}{n}.$$

Es decir, la proporción de valores observados menores o iguales a x

Por ejemplo, supongamos que tengo los datos observados $x_1 = 3, x_2 = 5, x_3 = 1$, estos los ordeno de menor a mayor $x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 5$, entonces,

$$F_e(x) = \begin{cases} 0, & \text{si } x < 1 \\ 1/3, & \text{si } 1 \leq x < 3 \\ 2/3, & \text{si } 3 \leq x < 5 \\ 1, & \text{e.o.c} \end{cases}$$

El estadístico de Kolmogorov-Smirnov se define para toda $x \in R$ como

$$D = max(F_e(x) - F(x)).$$

para distribuciones discretas y para continuas como:

$$D = sup(F_e(x) - F(x)).$$

Esta diferencia se ve representada a continuación

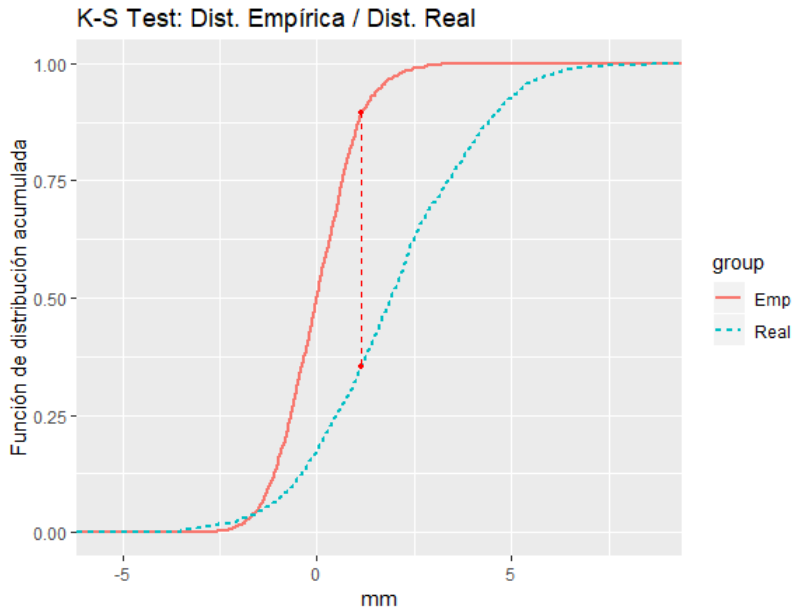


Figura 3.2: Estadístico de KS para dos distribuciones N(0,1) y N(2,2).

Podemos considerar las diferencias que se encuentran por encima o por debajo de la distribución, denotándolas como:

$$D^+ = sup(F_e(x) - F(x)).$$

$$D^- = sup(F(x) - F_e(x)).$$

y con ello redefinir el estadístico D:

$$D = max\{\frac{j}{n} - F(x_{(j)}), F(x_{(j)}) - \frac{j-1}{n}\}.$$

Posteriormente deberemos revisar que el valor D para calcular el p valor de la siguiente manera:

$$p = P_F(D \geq d).$$

También podemos ver el estadístico $D_{n,\alpha}$ el cual, si resulta ser mayor a D podemos decir que se rechaza la hipótesis nula.

Esto significa que el p.valor dependerá de la distribución de D y de F que se esté planteando.

Veamos entonces el siguiente ejemplo:

Sea X una muestra aleatoria con los siguientes valores (esta muestra será utilizada para ejercicios posteriores y ya se encuentra ordenada para facilitar los cálculos).

$X = (-2.736, -2.445, -1.816, -1.607, -1.358, -1.087, -0.853, -0.832, -0.818,$
$-0.721, -0.613, -0.567, -0.439, -0.402, -0.258, -0.217, -0.16, -0.105, -0.006,$
$0.133, 0.242, 0.308, 0.33, 0.362, 0.683, 0.791, 0.794, 1.209, 1.284, 1.339).$

La obtención de dicha muestra se realizó mediante el código:

```
set.seed(31109); X<-sort(rnorm(30,0,1))
```

¿Proviene dicha muestra de una distribución Normal(0,1)?

Para este ejemplo el hecho de tener un primer valor muy bajo al igual que el tamaño de la muestr afectará el resultado final, sin embargo, procederemos con los cálculos.

Muestra	F. AC	i/N	$Fn(X)$	Diferencias
-2.736	1	0.033	0.00311	0.030
-2.445	2	0.067	0.00723	0.059
-1.816	3	0.100	0.03467	0.065
-1.607	4	0.133	0.05398	0.079
-1.358	5	0.167	0.08729	0.079

Y la máxima diferencia D es:

0.362	24	0.800	0.641	0.159

$D = 0.159$.

Y dado que $D_{30,0.05} = 0.2417$, entonces $D < D_{n,\alpha}$, por lo tanto, no rechazamos la hipótesis nula con un nivel de confianza del 5 %, y entonces, la muestra pertenece a una distribución normal con un nivel de confianza del 95 %.

3.3.3. Anderson-Darling

La prueba de Anderson-Darling es otra forma de contrastar valores observados respecto a una distribución y que comparte las hipótesis de la prueba de Kolmogorov-Smirnov:

$$H_0 : F_X(x) = F_Y(x), \forall x \in R.$$

$$H_1 : F_X(x) \neq F_Y(x).$$

Es decir, nos ayudará a determinar si una muestra observada proviene de cierta distribución de tamaño N.

El estadístico de prueba se define como:

$A^2 = -N - S$.

Donde N es el tamaño de muestra y S:

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [ln(F(Y_i)) + ln(1 - F(Y_{N+1i}))].$$

Con F como la función de distribución acumulada y las Y_i ordenadas ascendentemente.

Utilizando la muestra aleatoria del ejercicio anterior, obtendremos los siguientes resultados, mostraremos los 5 primeros renglones:

No.Obs	De menor a mayor	F Teórica	Ln F
1	-2.7355	0.0031	-5.7719
2	-2.4454	0.0072	-4.9290
3	-1.8161	0.0346	-3.3618
4	-1.6073	0.0539	-2.9190
5	-1.3576	0.0872	-2.4384

Dado que la segunda componente de S corresponde a $N+1-i$, los cálculos serán análogos a colocar los valores de forma descendente, obteniendo entonces:

De mayor a menor	CDF Teórica $F(Yn1-i)$	$LN(1-Y)$	Suma
0.2539	0.6002	-0.91685	-6.6888
0.2452	0.5968	-0.90851	-17.5127
0.2374	0.5938	-0.90099	-21.3139
0.2277	0.5900	-0.89175	-26.6752
0.2197	0.5869	-0.88422	-29.9043

Los estadísticos resultantes son los siguientes:

$$A^2 = 1.3475.$$

Para la distribución Normal contamos con el estadístico de prueba ajustado con un nivel de confianza de 5 % (que queremos que sea mayor a nuestro estadístico A^2).

$$(1+\frac{4}{N}-\frac{25}{N^2})A_N^2 = 1.519943.$$

Si vemos los niveles de confianza del estadístico para una distribución Normal(0,1), con un nivel de confianza de 1 %, obtenemos:

$$(1 + \frac{4}{N} - \frac{25}{N^2})A_N^2 = 1.029.$$

Por lo tanto, tenemos una certeza de al menos 95 % de que nuestros datos provengan de una distribución Normal(0,1) pero no suficiente para tener una certeza del 99 %.

3.4. Aplicación de transformaciones

Otra forma de extender nuestro universo de distribuciones ajustables es agregando parámetros de escala y desplazamiento a aquellas distribuciones que están limitadas por su función indicadora y aplicando la función inversa de esos parámetros al momento de simular, por medio de una transformación lineal $T(X) = a * X + b$.

Un ejemplo de una de distribución que al aplicarle una transformación lineal conserva el kernel es la Cauchy, pues solo son modificados sus parámetros, lo que nos indica que estos representan desplazamiento y escala.

Por ejemplo, sea X una v.a. continua con distribución Cauchy(1,0).

¿Como se distribuirá $a * X + b$?

Si $X \sim Cauchy$, entonces su función de densidad está dada por:

$$f(x) = (\pi\gamma[1 + (\frac{x - x_0}{\gamma})^2])^{-1}.$$

y al sustituir,

$$f(x) = (\pi[1 + (x)^2])^{-1}.$$

Para aplicar la transformación lineal, deberemos hacer uso de la técnica de la transformación, primero definimos $Y = g(X) = aX + b$, esta transformación al ser lineal, no afecta el rango en el que se encontraba definida, es decir, la variable Y puede tomar valores en los reales.

Definimos $X = g^{-1}(X) = \frac{Y-b}{a}$ y el jacobiano como $\frac{\partial X}{\partial Y} = \frac{1}{a}$.

Entonces, la función de densidad de Y es:

$$f_Y(y) = f_X(g^{-1}(X))|\frac{\partial X}{\partial Y}| = (\pi a[1 + (\frac{Y-b}{a})^2])^{-1}.$$

Que sigue una distribución Cauchy con parámetros a y b, de escala y posición respectivamente. No siempre es posible tener esta particularidad con distribuciones que tengan un solo parámetro, pensemos por ejemplo en una institución financiera que recibe ingresos debido a un producto que ofrecen, sin embargo, hay ocasiones en las que dicho producto no genera uilidades, la forma en la que se gana y se pierde el dinero es igual a una distribución exponencial como se muestra a continuación:

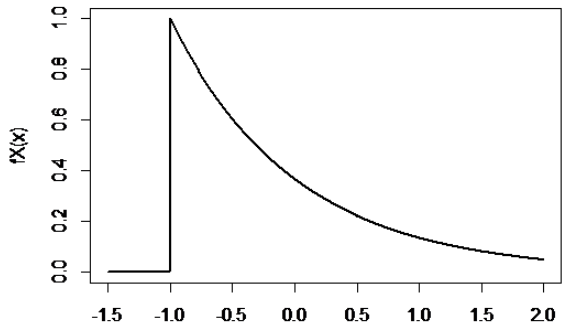


Figura 3.3: Distribución exponencial fuera de su rango usual.

Es importante considerar nuevos parámetros de los que podamos disponer, por lo que se aconseja realizar las siguientes transformaciones a partir de la función indicadora que acompañe a la distribución.

Si $x > 0$, como el caso de la distribución exponencial, se debe realizar el ajuste sobre:

$$Y = \frac{X - min(X)}{max(X) - min(X)}.$$

si algún valor de X se encuentra por debajo del 0.

Si $x \in [0, 1]$ como el caso de la distribución Beta o la Kumaraswamy.

$$Y = X + min(X) + \epsilon$$

si todos los valores de X son mayores que 0. Dadas las estructuras de programación actuales hay que definir el valor de ϵ como un número suficientemente pequeño en este caso 10^{-15}, cuya única función sea ubicar los valores por encima del cero sin que afecte el valor de los parámetros estimados).

Capítulo 4

Métodos de Clasificación y Multimodalidad

4.1. Introducción

Dentro de nuestro universo de datos, podemos contar con observaciones inusuales, también llamados outliers y es de esperar que alguna de estas afecte negativamente el ajustar una distribución y debamos recurrir a modelos no paramétricos para obtener información a partir de nuestros datos como el valor de la media o la distancia intercuartílica y deberemos dar un tratamiento especial a estas observaciones por separado.

Las pruebas de bimodalidad o multimodalidad van estrechamente relacionadas a la separación y clasificación de una distribución, por lo que habrá una transición natural a estas pruebas a través de los primeros métodos de clasificación.

4.2. Clasificación de outliers

Supongamos entonces que tenemos la siguiente información que representa la estatura en metros de un grupo de personas:

$$X = \{1.70, 1.56, 1.52, 1.53, 1.24, 1.45, 1.75, 1.63, 1.70, 1.38, 1.67, 1.41, 1.65, 1.54, 1.49, 2.11, 2.07\}.$$

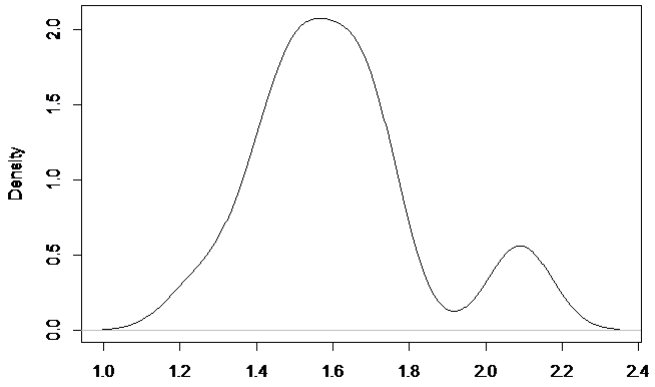


Figura 4.1: Distribución de estaturas.

Es de esperar que la distribución de estos datos sea normal, por otro lado, contamos con dos observaciones atípicas.

Este tipo de problemas suele suceder porque al momento de capturar la información esta no es homogénea, dado que se omitieron variables como la edad o el sexo y que de poder clasificarse, las pruebas anteriores pasarían sin problema, observando distribuciones distintas para adultos y niños, otra forma de atacar el problema es incrementando el tamaño de muestra, probando con distribuciones de colas pesadas como Cauchy o transformar los datos para probar con nuevas distribuciones como Beta o Kumaraswamy tal y como se mencionó en el capítulo anterior.

Observemos entonces cómo se ajusta una distribución normal con y sin outliers.

Debemos también tener presente la definición de partición:

Una partición es un subconjunto de X que contiene la unión de elementos no vacíos de X tal que cada elemento x en X está contenido en solamente uno de estos subconjuntos (i.e., X es la unión disjunta de estos subconjuntos).

Se dice que una familia de subconjuntos P es una familia de particiones si y solo si se cumplen las siguientes tres condiciones:

1. El conjunto vacío no forma parte de la familia P. (Esta condición es una formalidad).

2. La unión de los elementos en P es igual a X (equivalentemente podemos decir que la familia P cubre a X.): $\bigcup_{A \in P} A = X$.

3. La intersección de cualesquier dos elementos de P es vacío:

   $(\forall A, B \in P)\ A \neq B \implies A \cap B = \emptyset$.

   La forma de distinguir y separar estas observaciones también ha sido afrontada en el pasado mediante los siguientes métodos:

4.2.1. Método z

Este método nos ayuda a identificar el valor a partir del cual podemos considerar una observación como atípica, para ello es necesario conocer la distribución de nuestros datos o en su defecto el valor de sus percentiles, determinar un nivel de aceptación α y un estadístico por ejemplo la media.

Dependiendo de esta información será el análisis que deberemos realizar, en caso de no conocer nuestra distribución, emplearemos el delimitador 1-α que indicará el percentil en donde se realizará la partición de los datos.

En caso de tener una distribución conocida ya no sería necesario realizar el análisis, no obstante, esto también depende del p.valor a partir del cual rechazamos la hipótesis nula y no se descarta la posibilidad de aumentarlo en caso de separar la distribución, primero buscaremos la forma de homologarla por medio de la estandarización, es decir, un método que nos permita transformar una muestra de tal forma que la distribución asociada tenga parámetros más simples y que pueda ser fácilmente estimada.

Así, podemos definir el valor de z como:

$$z = \frac{x_i - \mu}{\sigma}.$$

Lo primero que se realiza es la sustracción de la media para posicionar nuestra distribución en el valor 0, lo que nos permite ver la simetría si es que la hay, poste-

riormente se divide entre la desviación estándar para que pueda ser escalada o más aproximable a una distribución normal.

Los valores de z son pensados originalmente para variables que se distribuyan como una normal y nos permite comparar unidades que no se encuentran en la misma escala.

Para otras distribuciones, el valor de z se verá afectado por la función indicadora de la distribución, por ejemplo, la distribución Beta se encuentra definida entre el 0 y el 1, por lo que habría que restar el valor mínimo y dividir entre la diferencia del máximo y el mínimo para que pueda ser ajustada al espacio correspondiente.

Ahora veamos una tabla de probabilidades para la distribución normal estándar.

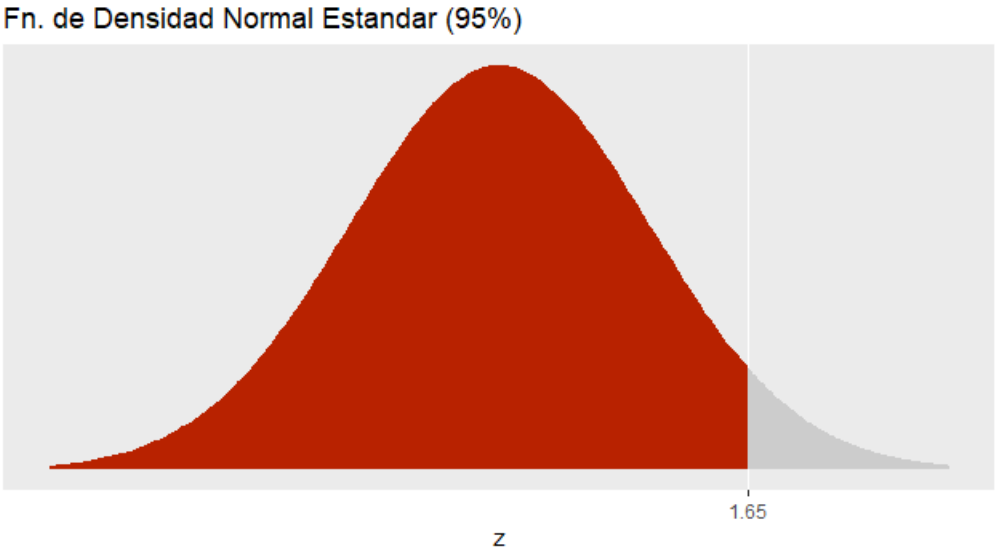


Figura 4.2: Distribución Normal estándar con un α del 5 %.

Tabla 4.1: Tabla de valores de la distribución normal estándar.

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1	0,8413	0,8437	0,8461	0,8484	0,8508	0,8531	0,85543	0,8576	0,85993	0,86214
1,1	0,8643	0,8665	0,8686	0,8707	0,8728	0,8749	0,87698	0,879	0,881	0,88298
1,2	0,8849	0,8868	0,8887	0,8906	0,8925	0,8943	0,89617	0,8979	0,89973	0,90147
1,3	0,9032	0,9049	0,9065	0,9082	0,9098	0,9114	0,91309	0,9146	0,91621	0,91774
1,4	0,9192	0,9207	0,9222	0,9236	0,9250	0,9264	0,92785	0,9292	0,93056	0,93189
1,5	0,9331	0,9344	0,9357	0,9369	0,9382	0,9394	0,94062	0,9417	0,94295	0,94408
1,6	0,9452	0,9463	0,9473	0,9484	0,9495	0,9505	0,95154	0,9525	0,95352	0,95449
1,7	0,9554	0,9563	0,9572	0,9581	0,9590	0,9599	0,9608	0,9616	0,96246	0,96327
1,8	0,9640	0,9648	0,9656	0,9663	0,9671	0,9678	0,96856	0,96926	0,96995	0,97062
1,9	0,9712	0,9719	0,9725	0,9732	0,9738	0,9744	0,975	0,9755	0,97615	0,9767
2	0,9772	0,9777	0,9783	0,9788	0,9793	0,9798	0,9803	0,9807	0,98124	0,98169

Ahora, utilizando el valor α que decidimos seremos capaces de definir a partir de qué valor podemos considerar que una observación es atípica.

Por ejemplo, para un nivel de confianza $(1-\alpha)\,\%$ del $95\,\%$, buscamos en la tabla aquel valor más cercano a 0.95, en las columnas se indica el primer dígito y el primer valor decimal y las columnas indican el valor centesimal que habrá que sumar y entonces, al ser la columna 1.6 y la columna 0.05, podemos decir que el valor a partir del cual consideramos que una observación es atípica es de 1.65, a estos valores los llamaremos puntos de corte, pues separan en dos o más particiones a nuestra distribución.

Es sencillo observar que entre más atípico sea el valor o menor sea α el valor x a partir del cual consideramos atípica una observación será mayor.

Para el caso de distribuciones multimodales la obtención analítica de la función de distribución como vimos anteriormente, requiere de cierta complejidad o no cuenta con una expresión analítica, por lo que se debe recurrir a aproximaciones por medio de métodos numéricos para estimar los estadísticos que deseamos contrastar.

4.2.2. Rango intercuartílico

Esta es una medida de dispersión de los datos, la cual nos ayudará a determinar un rango a partir del cual una observación se considera outlier.

Definimos el rango intercuartílico (RI) como la diferencia entre el primer y el tercer cuartil, es decir,

$$RI = Q_3 - Q_1.$$

Y podemos decir que una observación es un outlier cuando se encuentra fuera del siguiente rango:

$$Rango = (Q_1 - 1.5RI, Q3 + 1.5RI).$$

Este método es comúnmente usado en los diagramas de cajas, veamos pues un ejemplo con nuestra muestra de datos:

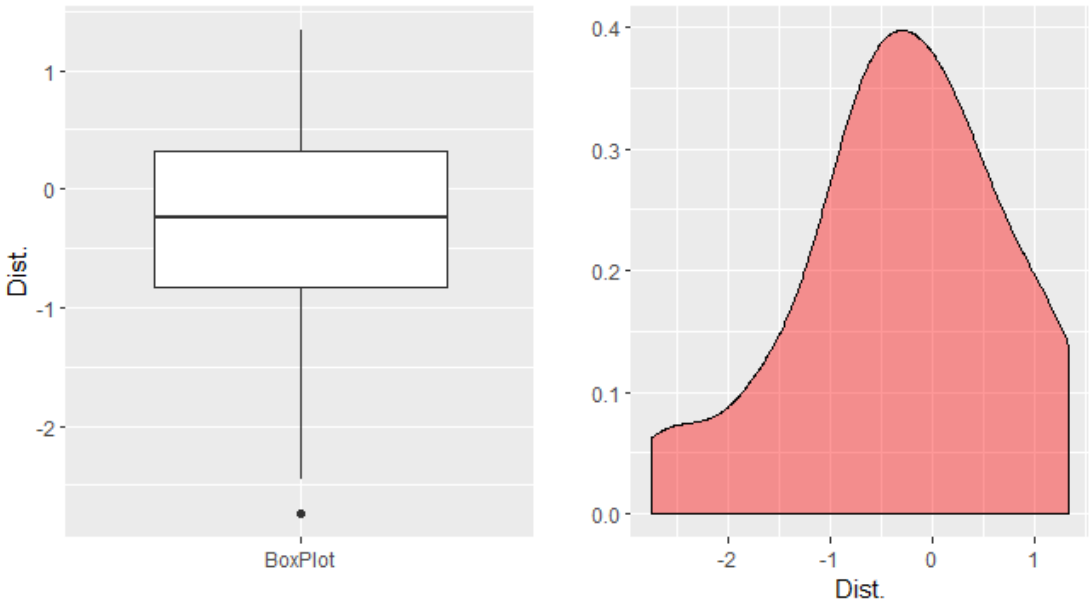


Figura 4.3: Boxplot y Distribución Normal de la semilla 31109.

Primer cuantil $= -0.828$ Tercer cuantil $= 0.324$ Rango intercuartílico $= 1.152$

Por lo que nuestra región sin outliers es

$$\{X | X \notin (-2.56, 1.48)\}.$$

Es decir, el único outlier corresponde a la observación -2.74, notemos también que nuestros puntos de corte son $-2.56 y 1.48$.

4.2.3. Otros métodos de clasificación

Otra forma de clasificar estos datos es por medio de modelos de aprendizaje supervizado y uno no supervisado, los cuales nos ofrecen un criterio para separar un conjunto de puntos en k-grupos.

K-medias

Este método corresponde a la clase algoritmos de aprendizaje no supervisado, es decir, busca semejanzas en los datos sin tener una predicción como objetivo.

Esta requiere de un único parámetro: el número de grupos en los que se separará la muestra. Para este caso este número será igual o menor al número de máximos locales en nuestra función de distribución, pues pensamos en que cada máximo representa la presencia de una distribución. Lo ejemplificaremos con un modelo bivariado, pues se extiende de forma natural a dimensiones más grandes y se aprecia mejor el resultado que en una sola dimensión.

El primer paso del algoritmo es colocar k puntos de forma aleatoria dentro de nuestros datos a los que denominamos centroides. Posteriormente se calculan las distancias de todos los puntos respecto a los centroides, se utiliza la norma euclidiana por defecto y a cada observación le es asociado el centroide más próximo. Una vez hecho esto, se desplazan los centroides al centro masa o gravedad de los puntos y se repite el proceso.

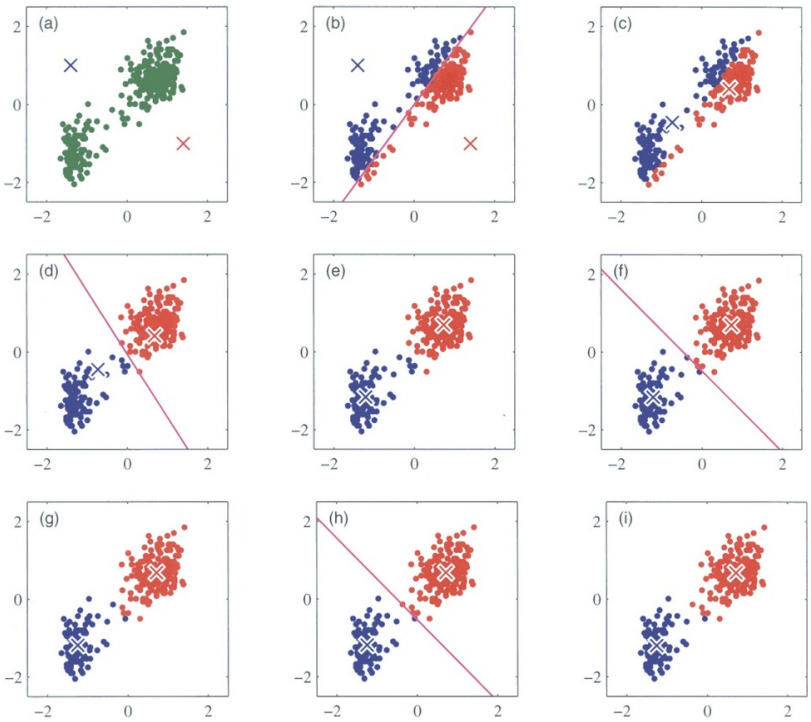


Figura 4.4: Ilustración del algoritmo de K-medias de [Bishop 2006].

Notemos a continuación que en cada iteración los centroides se desplazan y las observaciones pueden o no cambiar el centroide que tengan asociado, terminando el proceso una vez que estos dejen de desplazarse.

Una de sus desventajas es que si el número de grupos es grande es posible que arroje resultados distintos, no obstante, se recomienda un número no mayor a 5 grupos, pues podría presentarse sobreajuste dentro de nuestra función final.

La aplicación de los métodos para la detección de outliers en este caso nos puede ayudar a separar la distribución, una de estas aplicaciones es el cálculo de siniestros cuya suma asegurada es demasiado grande.

Ejemplo de obtención de los puntos de corte:

4.3. Pruebas de multimodalidad

Una prueba intuitiva es la observación de la función de densidad empírica (obtenida a través de los datos), no obstante, la creación de la misma recae en un método gráfico, en la estimación del kernel de la distribución o en el conteo de los puntos críticos y resulta no ser un método completamente objetivo.

Una distribución de este estilo provee de datos importantes acerca de la distribución, como que la media no es necesariamente un parámetro de máxima verosimilitud, que la muestra de datos no es homogénea, que las observaciones pueden venir de dos distribuciones empalmadas o que puede haber un error en los instrumentos de medición.

Para el ejemplo generado se utilizaron las siguientes distribuciones empalmadas:

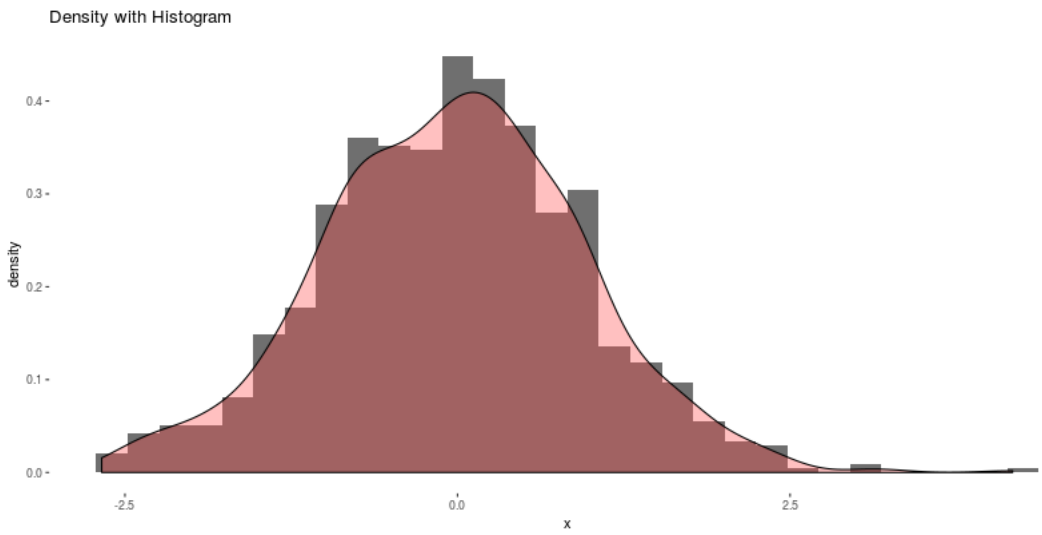


Figura 4.5: Distribución Normal(0,1).

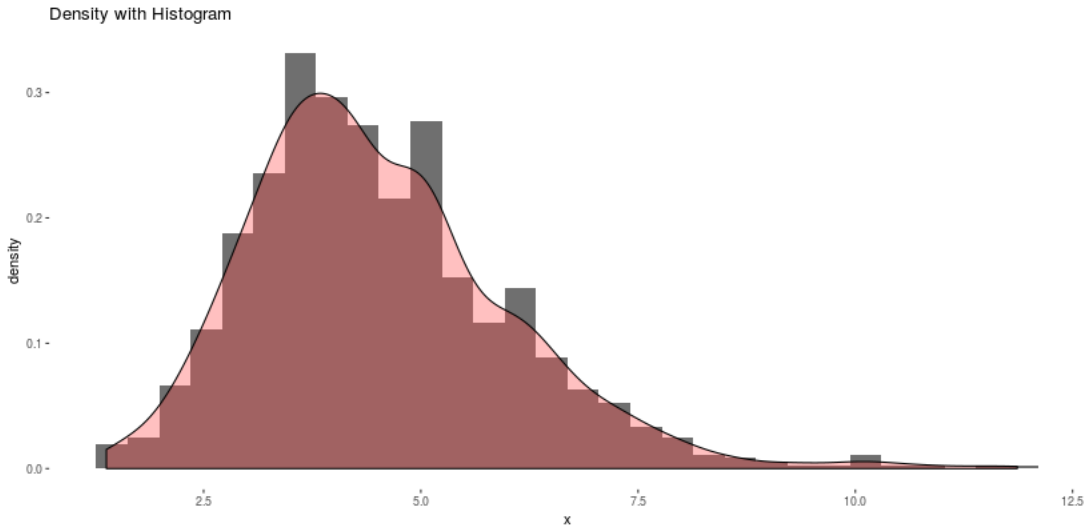


Figura 4.6: Distribución Gamma(9,2).

Existen pruebas para la detección de la multimodalidad en una función de distribución como el exceso de masa y ancho de banda crítico.

4.3.1. Ancho de banda crítico y número de raíces

Uno de los métodos no paramétricos que podemos utilizar es la estimación estadística del Kernel, éste es una función que cumple con las mismas características

que una función de densidad y que indica la forma que tendrá cada punto de nuestra distribución.

1.-

$$K(x) \geq 0 \quad \forall i \in S.$$

2.-

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Cada observación de nuestra muestra tendrá la misma función kernel asociada, sin embargo, debemos asumir un supuesto distinto para el kernel que estemos utilizando, por ejemplo, el de tipo gaussiano implica que la distribución final se compondrá de la suma de variables aleatorias normales, es decir, incluso si nuestra función final es de cola pesada o triangular, al usar esta distribución asumimos que estas se forman de la suma de un conjunto de distribuciones normales como se muestra en la siguiente gráfica.

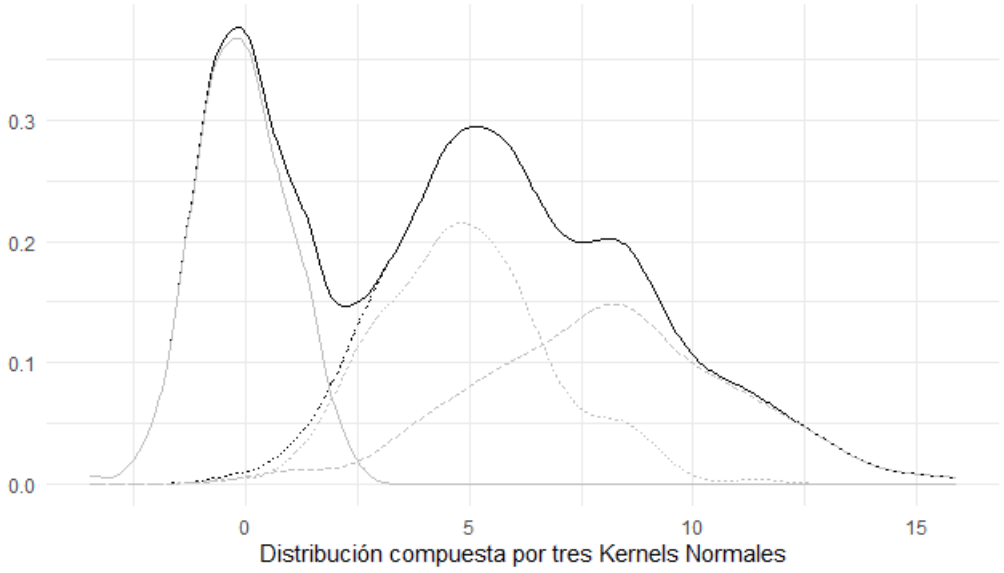


Figura 4.7: Kernel Gaussiano.

Debido a este problema, existen otros tipos de kernel para estimar la función de densidad, particularmente, para una distribución multimodal o cuya densidad parece alejarse de una distribución normal, se puede también emplear el método Improved

Sheather-Jones (ISJ). Dependiendo del tipo de kernel será la forma final de nuestra curva, algunos ejemplos son:

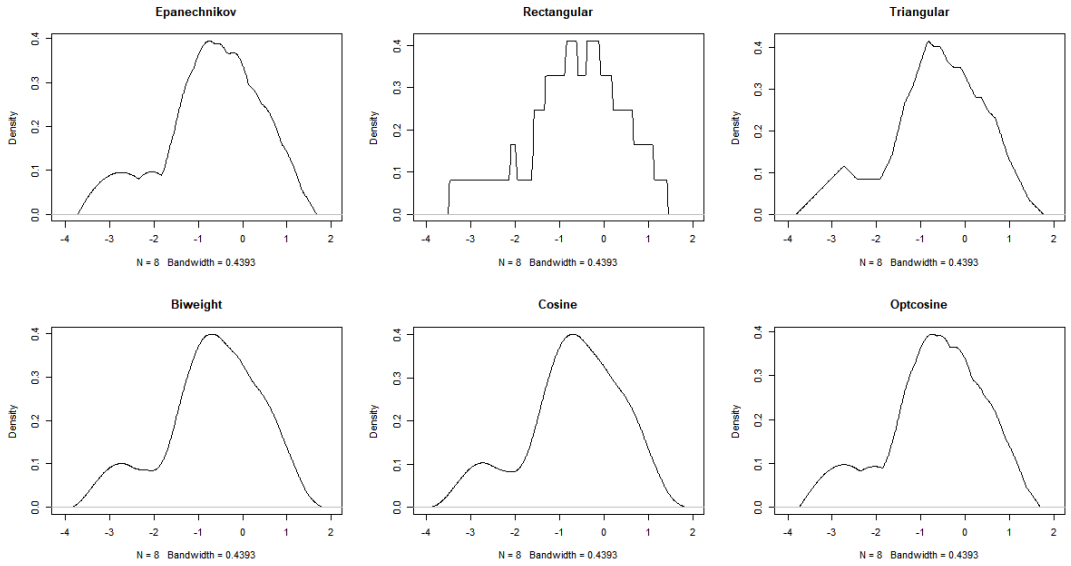


Figura 4.8: Tipos de Kernel.

El problema a resolver consiste en tener un entendimiento mayor de nuestra información y en el caso de no hallar una función de densidad teórica que se ajuste a nuestros datos, emplear la estimación del Kernel con el requerimiento de cierto número de supuestos o restricciones que no veríamos en un modelo paramétrico.

El estadístico asociado es:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}).$$

Donde h es nuestro parámetro de afinamiento al cual denominamos ancho de banda, éste determinará la forma de nuestra función kernel de la siguiente forma:

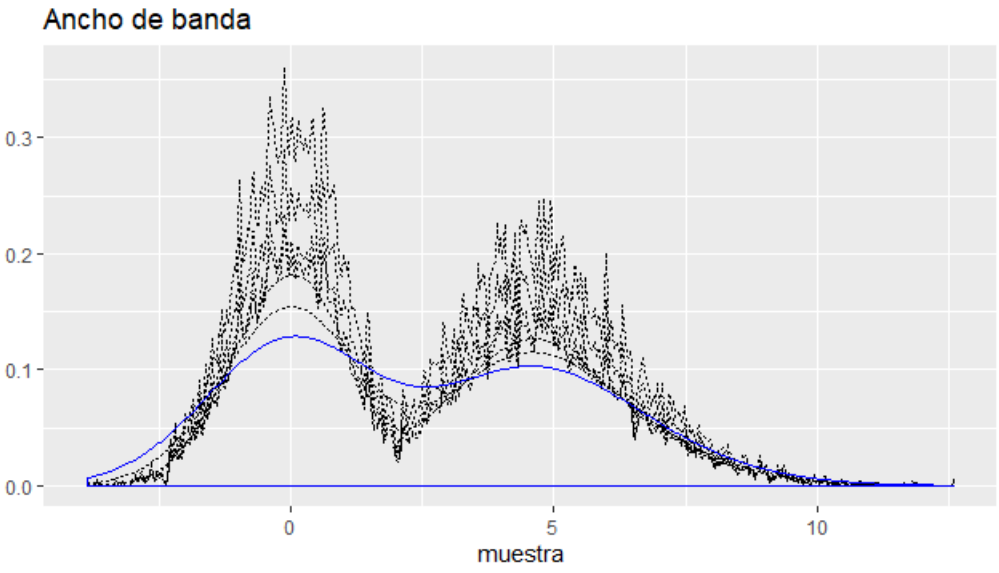


Figura 4.9: Función Normal-Gamma con diferentes anchos de banda.

Como podemos observar, tanto el número de modas como la forma de nuestra distribución son dependientes del ancho de banda, así como la forma del kernel.

La curva en color azul representa la verdadera forma de nuestra distribución, nosotros buscamos aquella h que minimice el error cuadrático medio de la integral (MISE), es decir, la h tal que nuestra distribución se aproxime más a su forma real, es decir,

$$\text{MISE}(h) = \text{E}\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right].$$

Para un número de modas $k \in N$, el ancho de banda crítico es el menor de los anchos de banda tal que la densidad del kernel tiene al menos k modas, entonces:

$$h_k = inf\{h : M(\widehat{f_h}) \geq k\}.$$

En donde $M(f_h)$ es el número de modas de f_h y f_h representa el estimador de la densidad del kernel de una muestra aleatoria $X = (X_1, ..., X_n)$.

Conteo de puntos críticos

Ya que contamos con nuestra función kernel definida, podemos encontrar la cantidad de máximos locales que tiene la función.

Conceptualmente, el cálculo de máximos y mínimos de una función se realiza por medio de la derivada, y dada la naturaleza del problema y que no contamos con una expresión analítica explícita para esta función procederemos a realizar el cálculo de forma numérica utilizando el kernel generado por la muestra.

Inicialmente, es sencillo el cálculo de esta función, solo debemos elegir una función kernel y una vez generada la función de densidad contar la cantidad de veces que la distribución pasa de ser creciente a decreciente, es decir, la identificación de máximos y mínimos en R:

Crearemos una muestra bimodal

```
set.seed(31109);c(rnorm(30   ,0   ,1) ,rnorm(30,4) )->a
```

Obtenemos el Kernel de la función:

```
Dens<-density(a)\$y
```

Conteo de las veces que cambia de ser TRUE a FALSE.

```
TF<-c(Dens<c(Dens[-1],0))
```

Se ven tres máximos y mínimos, lo cual coincide con los dos puntos máximos que tenemos y un mínimo local.

```
Cambios<-sum(TF!=c(TF[-1],FALSE))
```

Es decir, pasa de ser creciente a decreciente, decreciente a creciente y nuevamente vuelve a ser decreciente, significa que la regla que buscamos solo para contar los máximos locales será:

```
Maxloc<-(Cambios+1)/2
```

Realizando el ejercicio anterior sobre la muestra que tratamos llegamos a tres modas, lo que podemos utilizar para la separación de los grupos dando, los valores máximos obtenidos como los centroides del algoritmo de k medias o por consiguiente también

es posible separar la muestra en particiones que sean definidas en los valores donde la función pasa de ser decreciente a creciente.

Numéricamente debemos encontrar los puntos en donde nuestro kernel pasa de ser creciente a decreciente, también cabe señalar que no tenemos distribuciones con asíntotas, pues nuestros datos son finitos, por lo que los puntos máximos de nuestra función están bien definidos.

Una de las ventajas de la solución numérica es que esta no incrementa su complejidad si cambiamos el ancho de banda de nuestro kernel, en cambio, la solución analítica implica un incremento significativo en la complejidad de la función.

¿Solución numérica o solución analítica?

Recordemos que para la obtención de una solución analítica deberemos asumir que la función kernel es una representación fiel de nuestros datos, sin embargo, el mayor problema de dicho método es que asumimos que nuestra distribución ya tiene una asociada una función, que es lo que estamos tratando de encontrar.

La solución numérica no necesita tener una función explícita asociada, más allá de la estimación del kernel.

4.3.2. Coeficiente de bimodalidad

Esta prueba se remite a distribuciones con dos modas, no obstante, nos ofrece un criterio confiable a partir de dos características conocidas para las distribuciones: kurtosis (Ku) y asimetría u oblicuidad, mejor conocida como skewness (Sk).

Como breve preámbulo, recordemos que la kurtosis es una medida para la forma de la distribución de una variable aleatoria y podemos estimarlo mediante la siguiente fórmula:

$$\mathrm{Ku}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^2}.$$

La asimetría como su nombre lo indica es una medida que indica que tan simétrica es una distribución de probabilidad respecto a su media. El valor de asimetría

puede ser positivo o negativo, o indefinido y se calcula mediante el tercer momento estandarizado de la variable aleatoria, es decir:

$$\mathrm{Sk}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}}.$$

El coeficiente de asimetría será calculado mediante la siguiente fórmula:

$$b = \frac{Sk^2 + 1}{Ku}.$$

Como nos hemos percatado, una distribución bimodal no siempre tiene una expresión algebraica para ciertos parámetros de la función y su complejidad es muchas veces el freno para la realización de cálculos de estadísticos o de una función generadora de momentos, es por ello que utilizaremos una muestra aleatoria obtenida a través de una distribución bimodal.

En caso de tener un número finito de observaciones, surge una segunda fórmula para el coeficiente de bimodalidad:

$$b = \frac{Sk^2 + 1}{eKu + \frac{3(n-1)^2}{(n-2)(n-3)}}.$$

En donde n es el tamaño de la muestra la muestra y Sk y eKu la asimetría de la muestra eKu es el exceso de kurtosis de la muestra.

Debido a su forma (en donde la moda no es un valor definido) La distribución uniforme será nuestro parteaguas para definir si una distribución es bimodal o no por medio del coeficiente. El valor del mismo para la distribución uniforme es $5/9$. Los valores superiores a $5/9$ pueden indicar bimodalidad o multimodalidad. El valor se encuentra entre 0 y 1, alcanzando el valor de 1 solo para la distribución Bernoulli, pues solo hay dos valores distintos.

Es posible obtener valores mayores a $5/9$ para distribuciones unimodales que tengan mucho sesgo, es decir, que tengan muchos valores en la cola de la distribución.

Definiremos primero la obtención de los indicadores a partir de nuestra muestra finita de datos.

$$Sk = \frac{\sqrt{n(n-1)}}{n-2} Sk.$$

Donde Sk también representa la asimetría de la distribución (para una muestra finita de datos)

$$eKu = (n-1)\frac{(n+1)Ku - 3(n-1)}{(n-2)(n-3)} + 3.$$

Veamos el siguiente ejemplo: La muestra es obtenida a partir del siguiente código en R:

```
set.seed(31109);X<-c(rnorm(700),rgamma(700,9,2))
```

Cabe resaltar que dicho código no representa una operación o transformación entre las funciones Normal y Gamma utilizadas, sino la unión de dos muestras concatenadas entre sí.

El coeficiente de asimetría (CA) cobra sentido a partir de los siguientes valores:

Si CA<0: la distribución tiene una asimetría negativa, puesto que la media es menor que la moda.

Si CA=0: la distribución es simétrica.

Si CA>0: la distribución tiene una asimetría positiva, ya que la media es mayor que la moda.

Veamos entonces los siguientes resultados:

Primero probaremos con una distribución uniforme de tamaño $10,000$ con $X = 1, 2, ..., 9999, 10000$, $n = 10,000$, $E(X) = 5000.5$.

$$Sk = \frac{1/10000[(1-5000.5)^3 + (2-5000.5)^3 + ...]}{1/10000[(1-5000.5)^2 + (2-5000.5)^2 + ...]^{3/2}} \frac{\sqrt{10000(9999)}}{9998}.$$

$$eKu = (9999)\frac{(10001)\frac{\mathrm{E}\left[(X-5000.5)^4\right]}{(\mathrm{E}[(X-5000.5)^2])^2} - 3(9999)}{(9998)(9997)} + 3 = -1.201501.$$

Y finalmente,

$$b = \frac{0^2 + 1}{eKu + \frac{3(9999)^2}{(9998)(9997)}} = 0.5557409.$$

Entre más incrementemos el tamaño de la muestra, el valor tenderá a $5/9$.

Ahora veamos los resultados de las pruebas para las siguientes distribuciones.

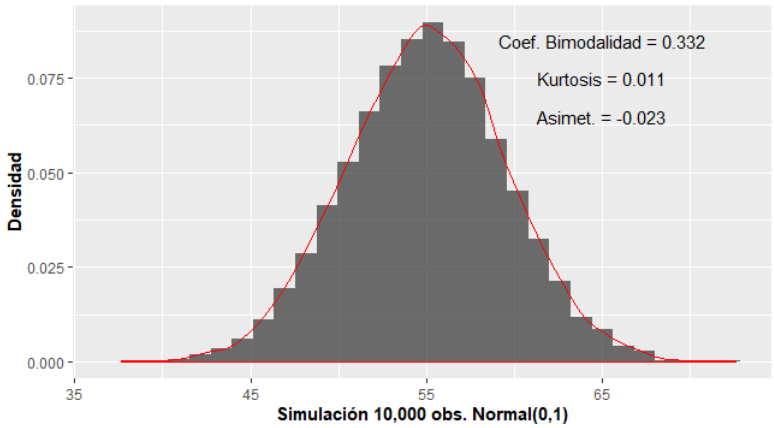


Figura 4.10: Coeficiente de bimodalidad, Kurtosis y Asimetría Normal.

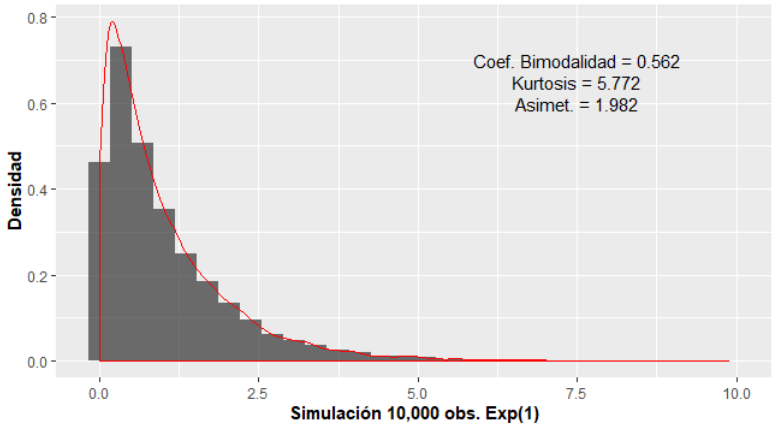


Figura 4.11: Coeficiente de bimodalidad, Kurtosis y Asimetría Normal-Exponencial.

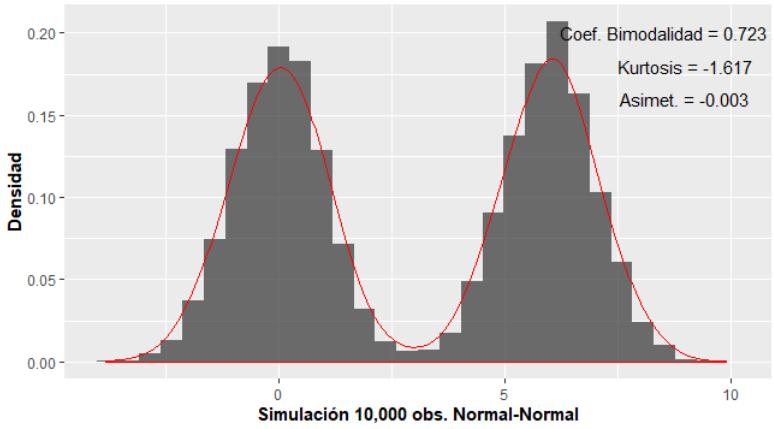


Figura 4.12: Coeficiente de bimodalidad, Kurtosis y Asimetría Normal-Normal.

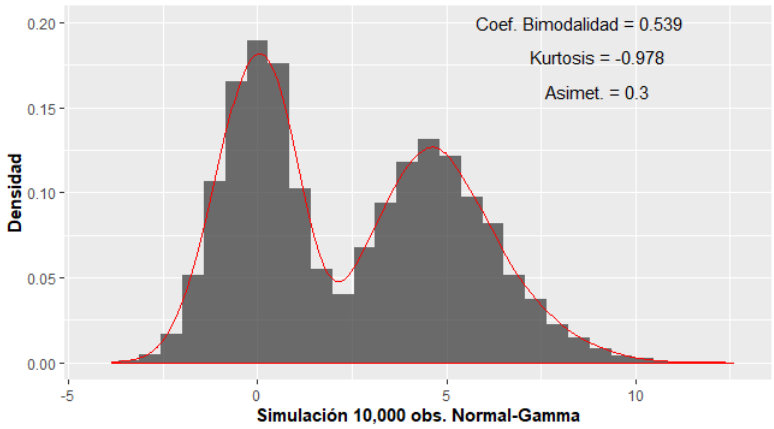


Figura 4.13: Coeficiente de bimodalidad, Kurtosis y Asimetría Normal-Gamma.

Con el coeficiente de bimodalidad observamos un nuevo fenómeno, es de interés resaltar que las distribuciones bimodales generadas no necesariamente superaron el valor de 5/9 en contraste con la distribución exponencial que se encontró muy cerca de dicho valor, esto nos indica que el indicador cumple su función para muestras cuya separación entre modas son grandes pero que no es sensible al conteo de las mismas, es decir, debemos hacer caso al indicador si el valor se encuentra más próximo a uno, no obstante, no podemos decir que la muestra no se ajusta a una distribución bimodal si el valor es inferior a 5/9, tal y como se muestra en la distribución Normal-Gamma.

Debemos tomar más de un criterio para la detección de la multimodalidad que en conjunto nos indiquen si muestra observada proviene de alguna distribución con

más de una moda. También es válida una confirmación visual (si el problema o el tiempo de cómputo lo permiten), pues los criterios vistos en esta sección son también empleados para la creación de histogramas.

Capítulo 5

Construcción del modelo

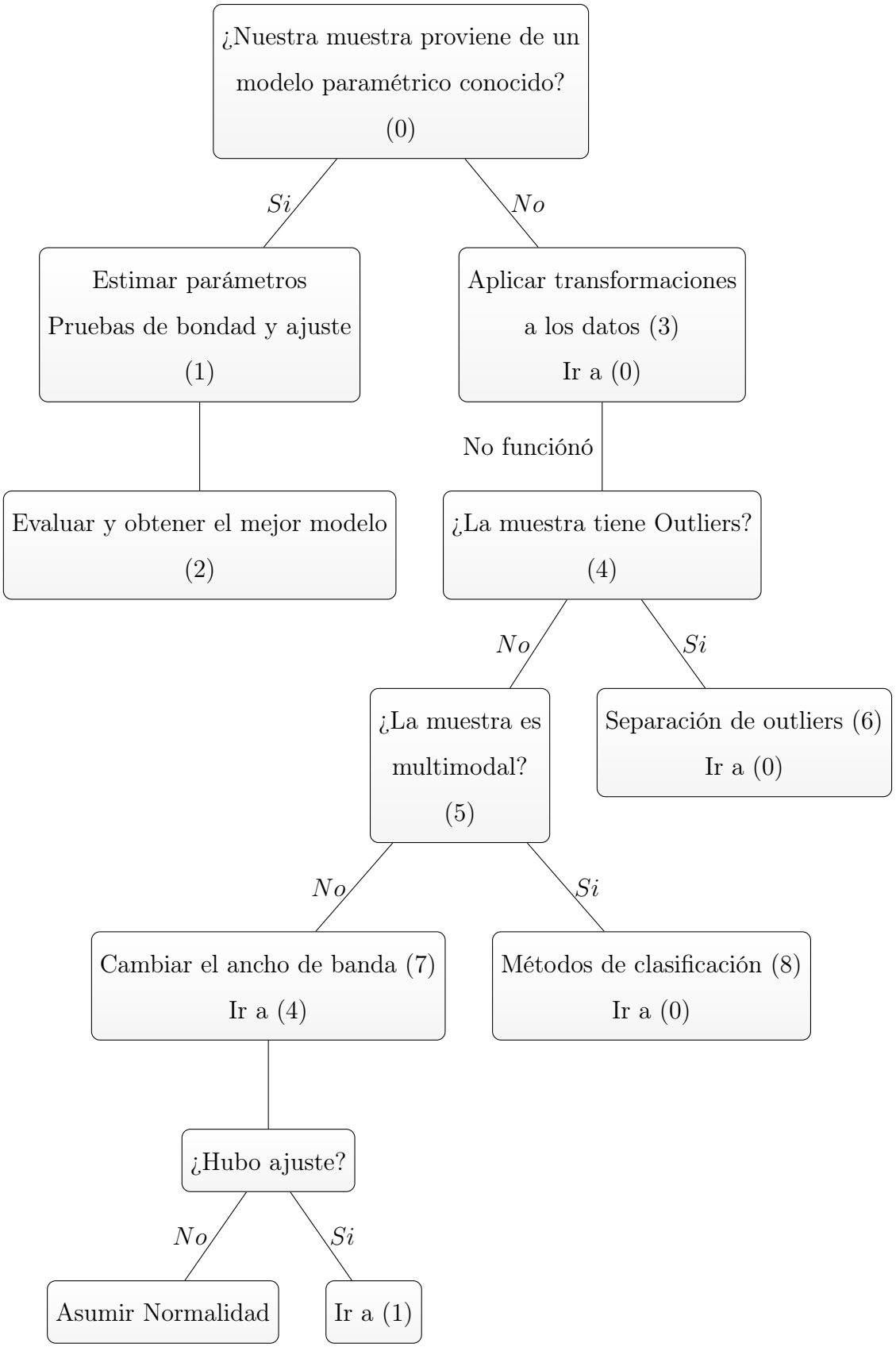
5.1. Introducción

Ahora que hemos definido metodologías para estimar parámetros, ajustar distribuciones, separar outliers, clasificar observaciones y pruebas de multimodalidad es momento de aplicarlo.

Pensemos en el caso de una compañía de seguros, en dicho caso la solución es separar las observaciones de los outliers dado que el volumen de estos es relativamente bajo y ajustar la distribución, la idea detrás de este método consiste en generalizar dicho proceso, incluyendo el caso en el que los outliers dejen de serlo para convertirse en otra distribución, es decir, que de un conjunto de datos al que las distribuciones usuales fueron rechazadas y que ahora decidimos clasificar nuestras observaciones obtendremos conjuntos de datos independientes que de tal forma que al ser nuevamente examinados para una nueva prueba de bondad y ajuste pasen las pruebas y entonces tratar a la distribución total como un conjunto de distribuciones con una proporción determinada.

El siguiente árbol de decisión nos ayudará a entender cuál es la forma de aplicar estos conocimientos para separar y tratar una muestra aleatoria.

Partimos de una muestra X, partiendo del supuesto de que dicha muestra no tiene valores desconocidos.


¿Nuestra muestra proviene de un modelo paramétrico conocido? (0)
Si
No
Estimar parámetros Pruebas de bondad y ajuste (1)
Aplicar transformaciones a los datos (3) Ir a (0)
No funciónó
Evaluar y obtener el mejor modelo (2)
¿La muestra tiene Outliers? (4)
No
Si
¿La muestra es multimodal? (5)
Separación de outliers (6) Ir a (0)
No
Si
Cambiar el ancho de banda (7) Ir a (4)
Métodos de clasificación (8) Ir a (0)
¿Hubo ajuste?
No
Si
Asumir Normalidad
Ir a (1)

Veamos entonces la información que nos ofrece cada camino a continuación:

Recuadro 1.- Dos posibles resultados: Máxima Verosimilitud y Método de Momentos, el criterio que puede usarse es: Tomar los parámetros del primer o segundo método.

Recuadro 1.- Dos posibles resultados Kolmogorov-Smirnov, Anderson-Darling Qua pase al menos un Test, o que deba pasar ambos.

Recuadro 2.- Un solo resultado Criterio de Akaike.

Recuadro 3.- Tres posibles resultados Desplazar la muestra a valores mayores que 0, estandarizarla o ajustarla dentro del intervalo (0,1).

Recuadros 4 y 6.- Dos posibles resultados Rango intercuartílico, Método Z Separar Outliers mediante el primer o segundo método.

Recuadro 5.- Dos posibles resultados Contar el número de modas y criterio de bimodalidad Si pasa el criterio de bimodalidad los resultados de K medias serán más acertados.

Recuadro 7.- Un solo resultado en el que separamos.

Recuadro 8.- Al menos dos resultados posibles K medias con K igual al número de modas Contar el número de modas por medio del número de máximos que tenga la función.

5.2. Variedad de casos

A partir del árbol de decisión antes observado, podemos darnos cuenta de que para cada recuadro contamos con al menos una metodologías de decisión y el aplicar cada una de ellas nos puede llevar a distintas soluciones, es decir, que una de las particularidades de este método es que podemos llegar a distintos resultados dependiendo de los criterios que tomemos en cada punto del mismo y dada la naturaleza del algoritmo K-medias, es posible que los centroides ajusten nuevos grupos, por ello se recomienda plantar una semilla o colocar un generador de números aleatorios controlado para generar resultados reproducibles.

Es posible que existan más criterios y pruebas a incluir en cada recuadro, sin

embargo, aquí se verán aquellas analizadas dentro del presente escrito.

Los tenores de este procedimiento incluyen aplicar transformaciones, distintos métodos de estimación de parámetros y diversos test de ajuste de distribuciones.

Una sutil observación es que el proceso de separación de la distribución se aplica de forma recursiva hasta que todas las separaciones tengan una distribución asignada que haya pasado las pruebas.

El éxito de este modelo depende de la cantidad de grupos en los que se haya separado la distribución, es decir, la cantidad de veces que fue rechazada la hipótesis de las pruebas de bondad y ajuste.

Definimos el nivel de profundidad como la cantidad de recursiones que se ha aplicado el algoritmo. Es de esperar que, a menor nivel de profundidad, es decir, menor número de cortes mejor será el modelo, pues cada separación incluye una nueva distribución a ser ajustada y entonces, una mayor número de parámetros.

Veremos entonces que si el nivel de profundidad es 1 quiere decir que tendremos al menos dos grupos en los que será separada la distribución y quiere decir la distribución puede asociarse a una variable aleatoria bimodal o en su defecto una variable que no se haya ajustado a las distribuciones probadas por el modelo.

Si el nivel de profundidad es dos, significa que tendremos al menos tres grupos en los que se habrá separado la distribución, por lo que podemos definir la siguiente relación:

$$\text{nivel de profundidad} < min\{\#\ \text{Grupos}\}.$$

Si el nivel de profundidad es mayor o igual a tres significa que serán ajustadas al menos cuatro distribuciones a nuestro modelo y tendremos ante nosotros una distribución multimodal.

Debido a ello es indispensable identificar la mejor forma de separar los datos, pensando en una métrica relacionada con el criterio de Akaike, en el que se castiga el número de parámetros, siendo para este caso penalizar el modelo por el número de distribuciones ajustadas.

El modelo ideal es aquel en el que no es necesario aplicar cortes, es decir, con nivel de profundidad de 0 o, en otras palabras, el modelo estándar.

La realidad no puede presentar problemas más complejos como el fallo en los instrumentos de medición o falsos positivos que no siempre podremos medir por lo que siempre hay que analizar los resultados con esceptisismo y tratar nuestras observaciones anticipadamente.

5.3. Comparativa entre modelos

Serán tomadas las siguientes consideraciones: Nuestros datos no tienen valores nulos o vacíos. Que el criterio de bimodalidad de una distribución no supere el valor de 5/9 no significa que esta no lo sea, mientras tanto, un valor cercano a 1 indica que la distribución debe ser separada. Serán probadas las siguientes distribuciones: Exponencial, Gamma, Log-Normal, Weibull, Normal, Cauchy, T y Beta. El p.valor será aplicable para los test de Kolmogorov-Smirnov y Anderson-Darling, tomado dos criterios distintos, el primero será si pasó uno de ellos o si pasó ambos y se tomará un $\alpha = 0.05$ por defecto. Se realizará con 500 muestras distintas de tamaño 5000 cada una para asegurar la consistencia de los resultados y que sean generalizables.

Trabajaremos con los datos generados a partir de la siguiente expresión en el software estadístico R:

Primero generaremos muestras de tamaño $1,000$ a partir de los siguientes modelos:

Normal-Weibull.

Código:

```
R<-c(rnorm(500,0,1/4),rweibull(500,5,3/4))
```

Ahora analicemos una muestra con distribuciones más separadas entre sí.

Exponencial-Weibull.

Código:

```
R<-c(rexp(500,4),rweibull(500,5,3))
```

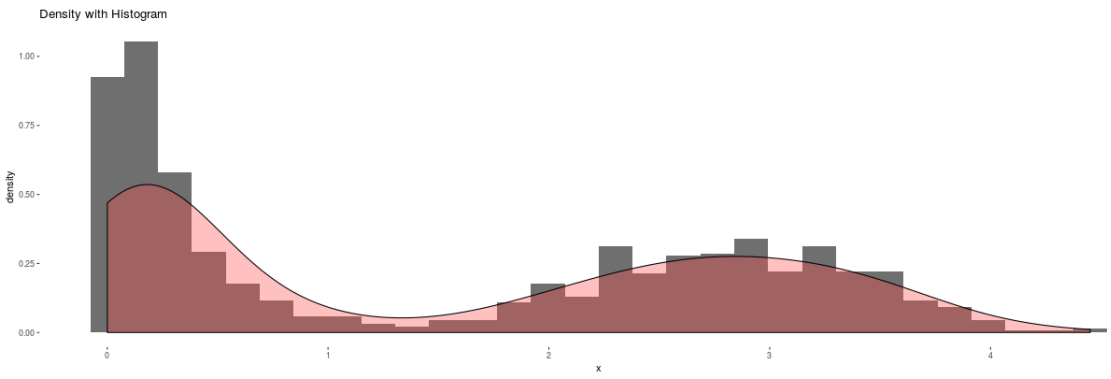


Figura 5.1: Distribución de la variable generada R.

Para el planteamiento de la hipótesis nula en general, realizaremos el planteamiento de la misma para cada una de las distribuciones, es decir, $H_0norm = F_X(x)$ donde X se distribuye normal, $H_0gamma = F_X(x)$ donde X se distribuye gamma, ..., $H_0t = F_X(x)$ donde X se distribuye t-student.

Y finalmente, nuestra hipótesis general:

$$H_0 : H_0norm \cup H_0gamma \cup \ ... \ \cup H_0t.$$

Esto quiere decir, que si alguna de estas hipótesis se cumple será suficiente para decir que nuestros datos se ajustan a una distribución conocida.

Y diremos que nuestra distribución está compuesta por un subconjunto finito de distribuciones conocidas y se puede expresar como una distribución mixta.

$$F(x) = \sum_{i=1}^{n} w_i \, P_i(x).$$

En donde $w_i = 1$ donde w_i representan los pesos que ocupan cada uno de los subconjuntos y $P_i(x)$ son las funciones de distribución asociadas.

Paso (1) Estimación de parámetros para distribuciones continuas del tipo: Exponencial, Beta, Gamma, Log-Normal, Normal, Weibull, Cauchy y T.

Resultados de las pruebas de bondad y ajuste:

Previamente, se transformará la muestra a partir de la función indicadora de la distribución, es decir, se aplicarán las transformaciones de la sección 3.3.4.

Tabla 5.1: Resultados pruebas de Bondad y Ajuste.

Dist	método	AD_p.v	KS_p.v	Param2	Param1
exp	mle	6.00E-07	0	0.666	NA
exp	mme	6.00E-07	0	0.666	NA
gamma	mle	6.00E-07	0	0.686	0.457
gamma	mme	6.00E-07	0	1.269	0.845
lnorm	mle	6.00E-07	0	-0.477	1.724
lnorm	mme	6.00E-07	0	0.116	0.762
weibull	mle	6.00E-07	0	0.802	1.356
norm	mle	6.00E-07	0	1.502	1.333
norm	mme	6.00E-07	0	1.502	1.333
cauchy	mle	6.00E-07	0	0.915	1.039
beta	mme	6.00E-07	2.33E-09	0.484	0.916

Paso 3 Como podemos observar, ninguno de los métodos aplicados a las distribuciones no fue rechazado, por lo que podemos decir que todas las hipótesis fueron rechazadas y entonces, nuestra distribución no proviene de ningún modelo paramétrico conocido.

Dado que las transformaciones fueron hechas para ajustar la distribución a las funciones indicadoras de la función, podemos ir al siguiente paso.

Ahora procederemos a realizar pruebas de identificación de outliers en la distribución y realizaremos nuevamente las pruebas.

Utilizando el método Z con los percentiles del 97.5 % igual a 3.741 y del 2.5 % que corresponde al valor 0.012 excluyendo el 5 % de nuestra distribución y de ajustar, recalcularemos esa proporción aplicando el mismo método.

Tabla 5.2: Pruebas de Bondad y Ajuste sin outliers.

Dist	método	AD_p.v	KS_p.v	Param2	Param1
exp	mle	6.32E-07	0	0.67	NA
exp	mme	6.32E-07	0	0.67	NA
gamma	mle	6.32E-07	0	0.774	0.519
gamma	mme	6.32E-07	0	1.325	0.888
lnorm	mle	6.32E-07	0	-0.37	1.514
lnorm	mme	6.32E-07	0	0.119	0.75
weibull	mle	6.32E-07	0	0.87	1.403
norm	mle	6.32E-07	0	1.492	1.297
norm	mme	6.32E-07	0	1.492	1.297
cauchy	mle	6.32E-07	0	0.965	1.052
beta	mme	6.32E-07	4.84E-08	0.391	0.582

Era de esperar, dada la forma de la distribución que la muestra principal no ajustara, pues el criterio de separación por medio de este método no necesariamente corresponde a la presencia de un conjunto de observaciones separadas, y solo se presenta un cambio en los parámetros estimados.

En contraste, veamos el método del rango intercuartílico, que toma outliers que se encuentren fuera del rango $(-2.370, 5.341)$, no obstante, no hay ningún valor que se encuentre fuera de dicho rango en la muestra y obtendríamos los mismos resultados que encontramos al inicio.

Como siguiente paso, revisaremos el criterio de bimodalidad, el cual toma el valor 0.746, valor superior a 9/5 lo que sugiere que nuestra muestra es multimodal.

Dados estos resultados, procederemos a realizar la clasificación de los datos, para ello emplearemos el método k medias, obteniendo el valor de k por medio del conteo de puntos máximos en el kernel generado.

Recordemos que una gráfica no representa una prueba, y aunque nos puede servir como guía, si deseamos replicar un proceso una gran cantidad de veces, puede que no contemos con el tiempo necesario para observar todas las gráficas generadas con distintos kernels, por lo que es preciso que las pruebas sean computacionalmente realizables y estén analíticamente justificadas.

Ahora realizaremos nuevamente el proceso para cada subconjunto de observaciones generado.

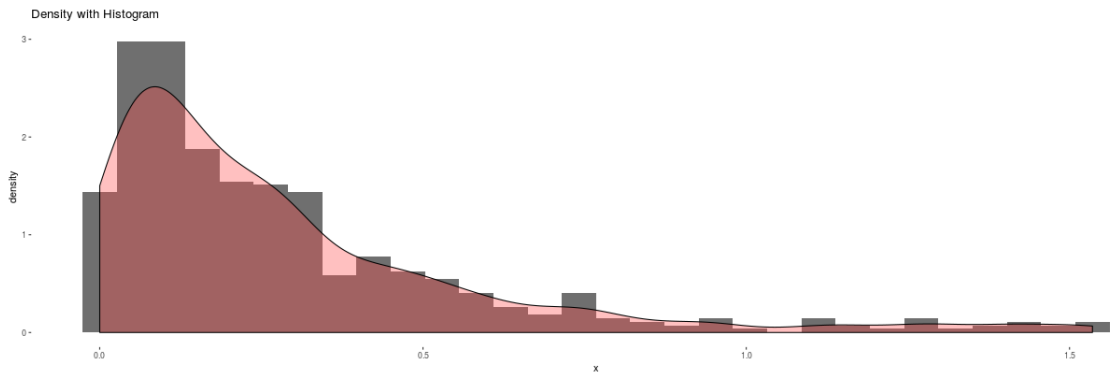


Figura 5.2: Separación del primer subconjunto con distribución desconocida.

Dist	method	AD_p.v	KS_p.v	Par1	Par2	Par21	Par22
exp	mle	0.474	0.295	3.476	NA	0.000	1.000
exp	mge	0.546	0.812	3.622	NA	0.000	1.000
gamma	mle	0.458	0.287	1.008	3.504	0.000	1.000
gamma	mge	0.530	0.879	1.024	3.739	0.000	1.000
lnorm	mle	0.017	0.055	-1.818	1.229	0.000	1.000
lnorm	mge	0.038	0.535	-1.717	1.140	0.000	1.000
weibull	mle	0.516	0.336	0.987	0.286	0.000	1.000
weibull	mge	0.516	0.858	1.011	0.275	0.000	1.000
norm	mle	0.000	0.000	0.288	0.303	0.000	1.000
norm	mge	0.000	0.000	0.217	0.198	0.000	1.000
cauchy	mle	0.000	0.000	0.160	0.114	0.000	1.000
cauchy	mge	0.000	0.000	0.208	0.123	0.000	1.000
beta	mge	0.000	0.314	0.911	4.501	0.000	1.535

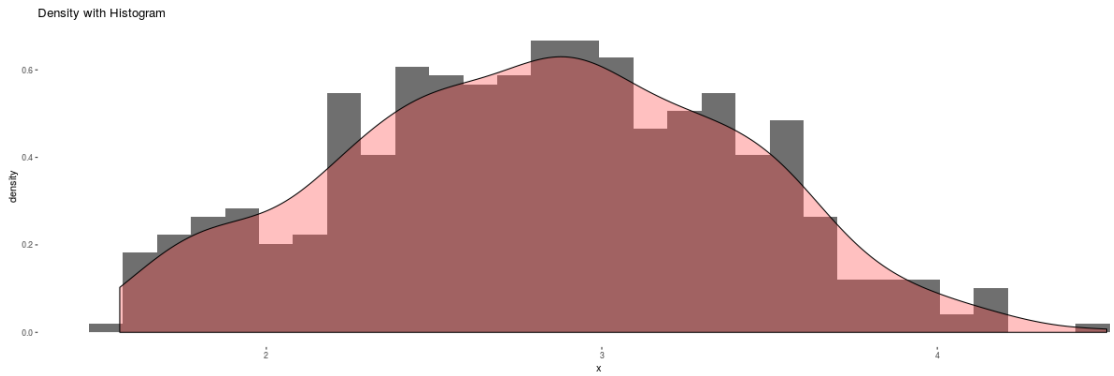


Figura 5.3: Separación del segundo subconjunto con distribución desconocida.

Dist	method	AD_p.v	KS_p.v	Par1	Par2	Par21	Par22
exp	mle	0.000	0.000	0.356	NA	0.000	1.000
exp	mme	0.000	0.000	0.356	NA	0.000	1.000
gamma	mle	0.104	0.246	22.195	7.894	0.000	1.000
gamma	mme	0.073	0.253	23.317	8.294	0.000	1.000
lnorm	mle	0.023	0.064	1.011	0.217	0.000	1.000
lnorm	mme	0.009	0.064	1.013	0.205	0.000	1.000
weibull	mle	0.585	0.641	5.395	3.047	0.000	1.000
norm	mle	0.570	0.813	2.811	0.582	0.000	1.000
norm	mme	0.570	0.813	2.811	0.582	0.000	1.000
cauchy	mle	0.000	0.001	2.827	0.377	0.000	1.000
beta	mme	0.000	0.757	8.142	4.904	0.000	4.505

Vemos resultados favorables para ambas pruebas en ambas muestras teniendo como mejores candidatas las siguientes distribuciones:

Primera muestra:

Distribuciones $gamma(1.024, 3.739)$, $exp(3.622)$ y $weibull(1.011, 0.275)$, siendo la primera aquella con la menor región de rechazo, con un p.valor de 0.879, no obstante, la distribución exponencial nos ofrece un modelo más simple con un pequeño costo del p.valor.

Segunda muestra:

Distribuciones $normal(2.811, 0.582)$, $beta(8.142, 4.904)$ (transformada) y finalmente $weibull(5.395, 3.047)$, siendo nuevamente la primera aquella con la menor región de rechazo, con un p.valor de 0.813 aunado a que es una distribución normal.

Una de las observaciones que podemos destacar es la adición de dos nuevos campos, que indican parámetros de desplazamiento y escala, es decir, la aplicación de una transformación visto de la siguiente forma.

$$Y = X * Par22 + Par21.$$

En donde la única distribución que ameritó dicho cambio fue la beta, pues recordemos que se encuentra definida en el intervalo $[0, 1]$, por lo que al aplicar la transformación, ahora es capaz de generar nuevos valores fuera de dicho rango.

Una de las desventajas de aplicar una transformación lineal a una variable alea-

toria que se encuentra acotada es que esta seguirá estando acotada por un intervalo finito y la muestra generada por dicha distribución no podrá rebasar los mínimos y máximos generados la muestra real observada, por lo que se debe restringir el uso de las mismas a observaciones que sabemos que se encuentran acotadas en la realidad o que sabemos han alcanzado su mínimo y máximo, por ejemplo, la temperatura en kelvins alcanzando un mínimo en -273 o el número total de siniestros reportados por una compañía aseguradora, siendo el máximo igual al número de pólizas vigentes que hayan emitido.

Debido a que ambas muestras pasaron las pruebas de bondad y ajuste, no rechazamos nuestra hipóteis nula (pues ambas tuvieron un p.valor mayor a 0.05) y entonces, nuestra hipótesis general se cumple y decimos que nuestra muestra se distribuye $Z = (X + Y)/2$ en donde X y Y tienen las siguientes distribuciones asociadas.

Tabla 5.3: Combinaciones posibles para Z.

X	Y
normal(2.811,0.582)	gamma(1.024,3.739)
normal(2.811,0.582)	exp(3.622)
normal(2.811,0.582)	weibull(1.011,0.275)
beta(8.142,4.904)	exp(3.622)
beta(8.142,4.904)	weibull(1.011,0.275)
beta(8.142,4.904)	gamma(1.024,3.739)
weibull(5.395,3.047)	weibull(1.011,0.275)
weibull(5.395,3.047)	gamma(1.024,3.739)
weibull(5.395,3.047)	exp(3.622)

Cada una de las columnas indica una solución válida para el ajuste de nuestra distribución, en donde los pesos w_i corresponden al valor de $\frac{1}{2}$, pues cada distribución ocupa el 50 % de nuestra muestra.

Recordemos que nuestra muestra provenía de un modelo $X+Y$ donde X de distribuye $weibull(, 5, 3)$ Y Y como $exp(4)$, y al contrastar los resultados, encontramos un a combinación similar donde X se distribuye $weibull(5.39, 3.04)$ y Y como $exp(3.62)$, distribuciones y parámetros que sin duda se asemejan a los reales.

Una de las características de este procedimiento es que a mayor cantidad de modas encontradas o cortes realizados, mayor será el número de soluciones posibles y al tener

distintos posibles resultados, merece la pena preguntarnos cuál de ellos es el mejor de ellos.

5.4. Valuación del modelo

Es preciso contar con un criterio de evaluación al ejecutar este procedimiento, pues la complejidad y capacidad computacional pueden ser optimizadas dependiendo de dicho valor, éste se ponderará por medio de una medida originada por medio del criterio de Akaike, considerando cada distribución como parte de una única función con parámetros correspondientes a los de cada función de distribución, de tal forma que se obtenga el mejor modelo.

5.4.1. Criterio de Akaike

Akaike describe su metodología como una forma de valorar la diferencia entre la exactitud y la complejidad de un modelo, tomando como criterio de complejidad el número de parámetros y la exactitud como el valor de máxima verosimilitud después de aplicar la función logaritmo por medio del cual obtenemos un criterio para elegir entre dos o más modelos, el criterio de Akaike (AIC) es definido como:

$$AIC = 2k - 2\ln(L).$$

En donde k es el número de parámetros de los que se compone nuestro modelo y L es máximo valor obtenido por la función de verosimilitud. Es entonces que surge la interrogante: ¿cómo obtener un criterio de evaluación adecuado para cualquier modelo?

Notemos dos cosas importantes, el criterio de Akaike al estar basado en la función de verosimilitud y retorna valores en los reales y dado que este representa la exactitud de nuestro modelo deberemos hacer un símil con nuestro proceso, es decir, emplear una medida de complejidad (el p.valor en este caso) y retornar valores en R, añadiendo la particularidad de que sea creciente y que retorne el valor 0 al evaluarla en .05, pues

es el valor más empleado para determinar el rechazo o la aceptación de una prueba y segundo, no es necesario cambiar el criterio de complejidad, pues es sencillo conocer el número de parámetros que estamos utilizando.

La función propuesta es:

$$f(x) = tan(\pi * (x - 1/2)) - c.$$

donde c corresponde a $f(0.05)$, es decir, $c = tan(\pi * (.05 - 1/2))$.

Recordemos que la correspondencia exacta entre la complejidad y la precisión en el criterio de Akaike es que debe incrementarse significativamente (de forma exponencial) la precisión para compensar la asignación de un nuevo parámetro, análogamente, se debe incrementar significativamente un p.valor (tangencialmente) para justificar la inclusión de un nuevo parámetro.

Ya que se respetan formalmente los conceptos concebidos por Akaike, procedemos a crear un Criterio de Ajuste (CA) por medio de la siguiente fórmula:

$$CA = 2k - tan(\pi(x - 1/2)) + tan(\pi(.05 - 1/2)).$$

Donde k es el número de parámetros de la distribución y p es el p.valor del ajuste obtenido.

En dónde k es el número de parámetros y p es el p.valor obtenido tras aplicar las pruebas de KS o AD.

Al tratarse de un modelo mixto, definiremos el Criterio de Ajuste Mixto como la suma de los criterios de ajuste de cada partición multiplicado por su vector de peso correspondiente a la proporción de observaciones medidas, al igual que la función (2.2.2), es decir,

$$CA_{Mixto} = \sum_{i=1}^{p} w_i * CA_i$$

Recordemos que el AIC puede arrojar valores negativos, por lo que es posible que, al añadir una partición y entonces más parámetros, mejore considerablemente

el modelo, de la misma forma penalizamos la inclusión de modelos con una mayor cantidad de parámetros.

Ya que ha sido definido, es posible hacer uso del criterio de Akaike para indicar cuál es el mejor modelo para cada partición de la distribución, esto, para beneficiar modelos con menor número de parámetros y dado que cuentan con un número de observaciones finitas, podemos decir que nuestro criterio ahora se encuentra bien definido y lo expresamos a través de la siguiente fórmula.

$$\text{AICc} = \text{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

Con n igual al número de observaciones por partición.

Veamos entonces el ejemplo con la variable con la que hemos estado trabajando, tal y como lo vimos en el capítulo 2, el estadístico de máxima verosimilitud para una variable aleatoria normal es $\bar{X}$, que para nuestro conjunto de datos es -0.318, $k = 2$ y $n = 30$, por lo tanto,

$$AIC = 2 * 2 - 2ln(-0.318) = 5.844$$

y

$$AICc = 5.844 + \frac{2 * 2^2 + 2 * 2}{30 - 2 - 1} = 6.288$$

Sin embargo, ya que sabemos cuál es el mejor modelo para cada uno de nuestros grupos, aun queda la incógnita sobre cómo elegir el mejor resultado de todos los posibles.

Habrá que considerar tres criterios:

La cantidad de parámetros de cada partición en los que se divide la muestra con su respectiva proporción, el AIC y p.valor de cada uno de ellos.

La forma en la que se verá beneficiado dicho indicador será por medio de un bajo número de distribuciones, y cada una de ellas con una baja cantidad de parámetros. También incluiremos la utilización de modelos con bajo CA y que hayan pasado las pruebas de bondad y ajuste con los valores más cercanos a 1.

Malinterpretación del modelo

La primera malinterpretación de los resultados puede relacionarse con la elección de las variables aleatorias que ajustaron a las observaciones, siendo posible obtener resultados distintos a partir de Anderson-Darling y Kolmogorov-Smirnov, la solución a dicho dilema será por medio del p.valor que arrojen las pruebas, es decir, el no rechazar la hipótesis que indica que pertenecen a cierta distribución, lo que implica obtener un p.valor mayor.

Como ejemplo veremos la siguiente muestra:

```
set.seed(31109);c(rexp(700,1))
```

Al aplicar la metodología descrita esperaríamos obtener una distribución exponencial como resultado, sin embargo, al revisar los resultados obtenemos:

Tabla 5.4: Resultados de pruebas de bondad y ajuste con diferentes estimaciones de parámetros.

Distribución	AD_p.v	KS_p.v	Param 1	Param 2	p.v. medio
exp	0.975	0.845	0.965	NA	0.910
exp	0.982	0.943	0.974	NA	0.963
gamma	0.977	0.839	1.006	0.971	0.908
gamma	0.989	0.951	1.012	0.989	0.970
weibull	0.970	0.857	0.996	1.035	0.914
weibull	0.987	0.953	1.007	1.026	0.970

El primer renglón corresponde a los parámetros obtenidos mediante el método de momentos y la segunda por medio de máxima verosimilitud.

La relación entre la distribución gamma y exponencial se hace notar al revisar el segundo parámetro de la función gamma pues es muy cercano a uno, lo que indica que es también una exponencial, no obstante, también tenemos la distribución weibull como candidata.

Es aquí en donde debemos tomar una decisión, sabemos que nuestra muestra proviene de una distribución exponencial con parámetro uno y, sin embargo, el p.valor medio de las dos pruebas de bondad y ajuste indican que la función weibull tiene un rango menor de rechazo. Pensando en la simplicidad, y en la poca diferencia

que arrojaron los p.valores, deberemos elegir entre perder verosimilitud pero ganar simplicidad en el modelo.

Veamos rápidamente el ajuste de la función weibull respecto a la muestra original.

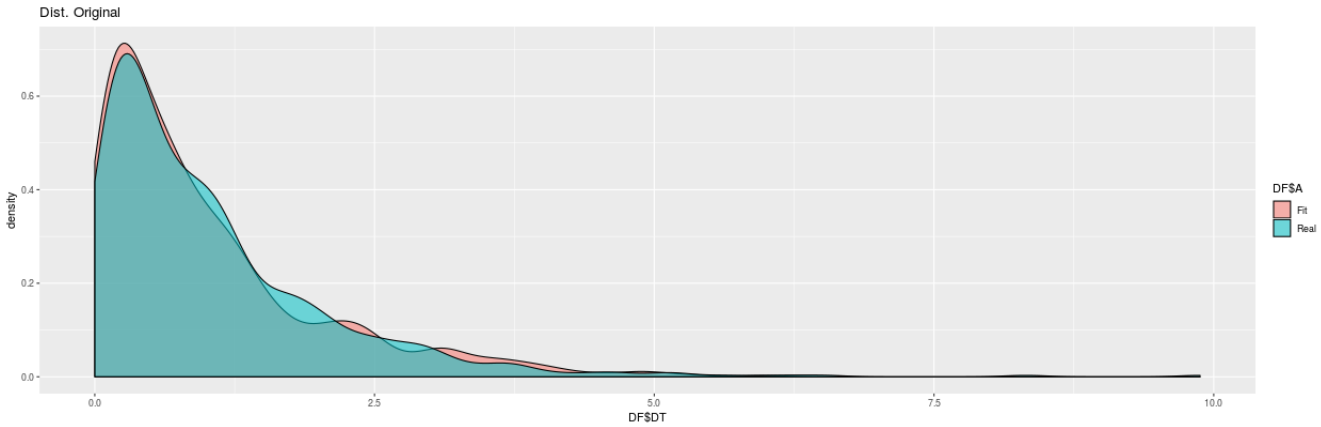


Figura 5.4: Muestra Exponencial con ajuste de distribución Weibull.

Un segundo error que puede presentarse es aprender a distinguir entre información que no ha sido previamente procesada y la identificación de una distribución multimodal.

Veamos como ejemplo la tabla de medidas de longitudes de las flores.

Si procediéramos con un ajuste inmediato, veríamos la presencia de una distribución multimodal.

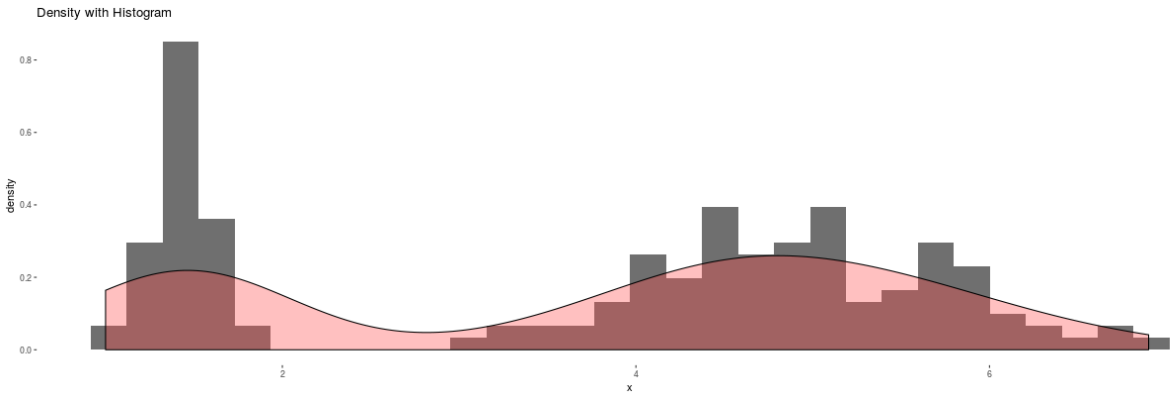


Figura 5.5: Ejemplo de la distribución del largo de un pétalo de tres especies distintas de flores.

No obstante, merece la pena observar que contamos con una variable categórica correspondiente al número de flores.

La mejor separación es aquella que ya está definida por las variables categóricas que acompañan a los datos, a esto lo denominaremos como separación natural de la información, pues al separar los datos por medio de esta variable es evidente que cada flor tiene una distribución ajustable que puede ser estudiada.

Veremos los resultados de esta separación al aplicar los test de ajuste Anderson-Darling y Kolmogorov-Smirnov.

Virginica:

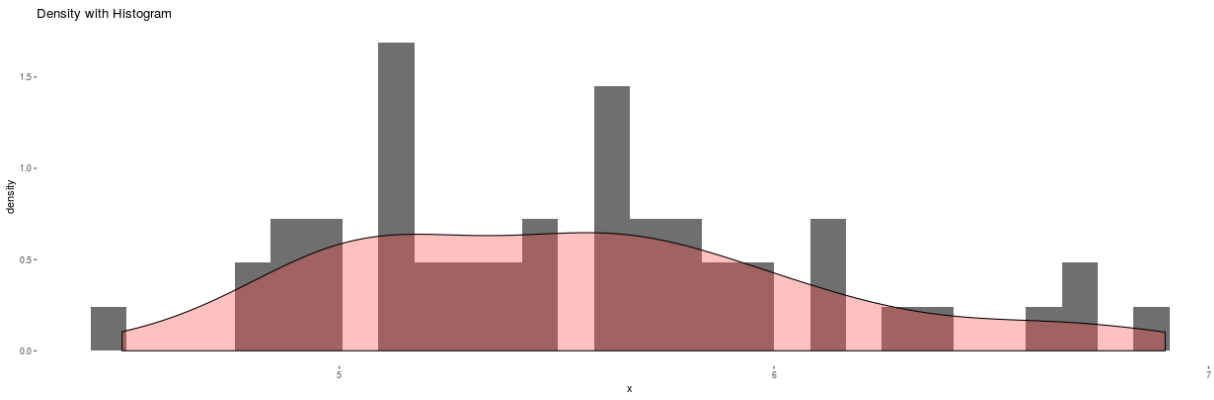


Figura 5.6: Ejemplo de la distribución del largo de un pétalo de la flor del tipo virginica.

Distribución asociada: lnorm(1.705, 0.101), Anderson-Darling: 0.836. Kolmogorov-Smirnov: 0.785.

Versicolor:

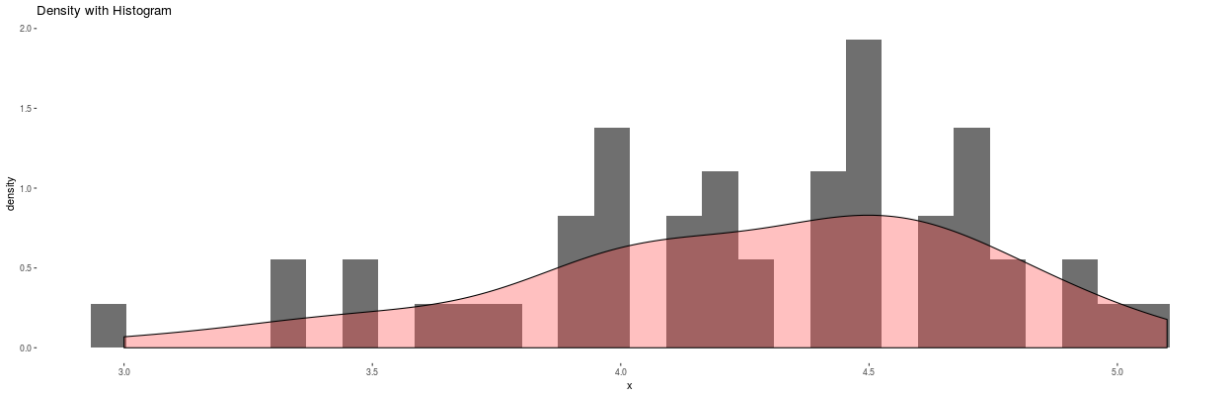


Figura 5.7: Ejemplo de la distribución del largo de un pétalo de la flor del tipo versicolor.

Distribución asociada: weibull(10.685, 4.469), Anderson-Darling: 0.982, Kolmogorov-Smirnov: 0.911.

Setosa:

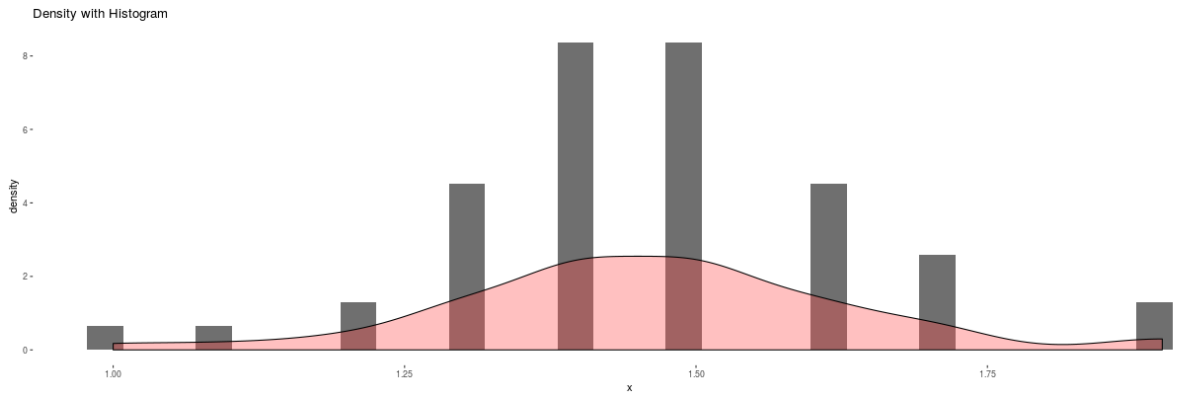


Figura 5.8: Ejemplo de la distribución del largo de un pétalo de la flor del tipo setosa.

Distribución asociada: gamma(86.105, 58.843), Anderson-Darling: 0.347, Kolmogorov-Smirnov: 0.320.

5.5. Ventajas

Una de las primeras ventajas de este modelo es que podemos diferenciar claramente cada componente de la distribución. Gana precisión al momento de estimar una distribución.

Es una generalización de los modelos contemporáneos y la solución no se basa en fuerza bruta, es decir, no es necesario calcular cientos de distribuciones distintas para elegir alguna que ajuste.

Se puede simplificar fácilmente a un modelo de ajuste clásico.

Es sencillo obtener más de un modelo multimodal que ajuste.

Puede ayudarnos a visualizar si la información que estamos evaluando necesita un análisis previo.

Amplía el panorama para una mejor evaluación de estadísticos locales y no uno global que pudiera tener un mayor error.

Hay más de una forma de escribir una variable aleatoria multimodal, podemos verla como una combinación lineal de dos o más distribuciones, por ejemplo:

$Y = Norm * I + Gamma * I$

En donde la función indicadora nos facilita el análisis del total de la distribución y al ser una suma, los estadísticos como la esperanza y varianza pueden ser estimados de la misma forma que en el método clásico.

Se pueden combinar distribuciones de colas pesadas, añadiendo más familias con las que es posible trabajar.

Se ofrece un modelo paramétrico para información a la que es casi imposible ajustar o cuyos p.valores son insuficientes.

5.6. Desventajas y Soluciones

Al tener un mayor número de cortes y parámetros, la función de distribución multimodal se ve afectada por el criterio de Akaike y el sesgo, por lo que se recomienda su uso buscando en principio distribuciones de un solo parámetro (como la exponencial), y un número pequeño de particiones para la distribución, es decir buscar que sea bimodal o multimodal.

Dependiendo del método de corte o división de la variable aleatoria se podrán obtener diferentes resultados para una sola muestra, para reducir el número de estos, se deben emplear diferentes tamaños de muestra y llegar a aquella que más se se parezca a la original y elegir las distribuciones con menor cantidad de parámetros.

Por el momento solo y está enfocado en modelos univariantes, al añadir más dimensiones al modelo, el proceso se vuelve más robusto, debiendo añadir más criterios para los algoritmos de separación e introducir cópulas para no perder la correlación entre marginales.

La definición formal de un modelo multimodal es difícil de describir debido a los problemas de distribuciones empalmadas entre sí.

Capítulo 6

Aplicaciones

6.1. Introducción

A continuación, veremos dos casos reales en los que este modelo representa una mejora en el análisis de la información observada.

Dado el objetivo de este escrito, tocaremos el análisis cualitativo y cuantitativo del problema, entrando más a fondo en las implicaciones y ventajas que tiene la aplicación de este modelo.

La principal ventaja de conocer la distribución de los datos es la capacidad de medir la probabilidad de ocurrencia de un evento o la obtención de sus estadísticos y otras medidas (media, mediana, percentiles, rango intercuartílico, entre otros) a partir de la misma sin necesidad de recurrir a una distribución empírica.

6.2. Ejemplo de Aplicación. Fraude de tarjetas de crédito

6.2.1. Introducción

Analizaremos un caso particular, este conjunto de datos contiene las transacciones realizadas con tarjetas de crédito en septiembre de 2013 por titulares europeos. La información cubre dos días en los que se presentaron 492 fraudes de 284,807 transaccio-

nes. El total de los datos está altamente desbalanceado, pues los fraudes representan el 0.172 % de todas las transacciones.

La tabla contiene solo variables numéricas y será de nuestro interés estudiar cada una de ellas y en particular su relación con transacciones fraudulentas y no fraudulentas. La información del usuario se encuentra protegida debido a la sensibilidad de la información. Las variables para este análisis serán Tiempo y Cantidad. El Tiempo contiene los segundos transcurridos entre cada transacción y la primera transacción en el conjunto de datos. La Cantidad indica el monto de la transacción y finalmente, la variable Clase toma el valor 1 en caso de fraude y 0 en caso contrario.

Para efectos de este trabajo, nos remitiremos a observar la distribución de las variables Tiempo, Cantidad pues ambas son ejemplos de variables aleatorias continuas y mientras que Clase al ser categórica nos ayudará con la realización del análisis exploratorio comparando los resultados de nuestro modelo con el modelo usual.

Obtendremos las distribuciones del tiempo y cantidad dado que la clase es 0, es decir, no hubo fraude y dado que la clase es 1.

A partir de los resultados responderemos a las preguntas: ¿existe una diferencia en distribución al cambiar la clase para ambas variables?, ¿qué distribución(es) componen a cada uno de los modelos?

Si hay diferencia en estas, ¿cuáles son los máximos locales para cada distribución y que se puede decir de estos?

Y finalmente, ¿cuáles fueron los problemas computacionales al enfrentar este ejemplo?

6.2.2. Análisis Exploratorio

Este análisis será realizado mediante el uso del software estadístico R y será documentado junto con el código respectivo. Primero leeremos nuestros datos y nos aseguramos que nuestra información esté completa, es decir, no haya presencia de datos vacíos.

```
data<-read.csv(paste0(getwd(),"/creditcard.csv"))
#Conteo de na's por fila
sapply(data,function(x) sum(ifelse(is.na(x) | is.nan(x),1,0)))
```

Veamos primero un breve resumen estadístico de la totalidad de la muestra.

Tabla 6.1: Estadísticos originales.

Tiempo	Monto
Min. : 0	Min. : 0.00
1stQu.: 54202	1stQu.: 5.60
Mediana : 84692	Mediana : 22.00
Media : 94814	Media : 88.35
3rdQu.:139321	3rdQu.: 77.17
Max. :172792	Max. :25691.16

Como podemos observar, la variable tiempo muestra una diferencia entre la Mediana y la moda y dado el volumen de datos, podemos descartar la distribución normal, por otro lado, todos los cuantiles presentan diferencias similares, por lo que al hacer la prueba de outliers con rango intercuartílico seguramente veremos pocos. Respecto a la variable monto, hay una clara diferencia entre el tercer cuantil y el máximo, y entonces, podemos esperar la presencia de outliers en nuestra distribución, de igual forma, vemos una notable diferencia entre mediana y moda, por lo que esperamos que la distribución esté cargada hacia la izquierda, pues la media es incluso mayor al tercer quantil.

También vemos que ambas muestras tienen su mínimo en el valor 0, por lo que no será necesario realizar transformaciones para ajustar distribuciones que se encuentren definidas en $\mathbb{R}$, pero será necesario sumar un épsilon para aquellas definidas en $\mathbb{R}^+$ como la exponencial, por ejemplo, veamos la distribución de las dos variables separando los fraudes (Clasificación 1) de la muestra original.

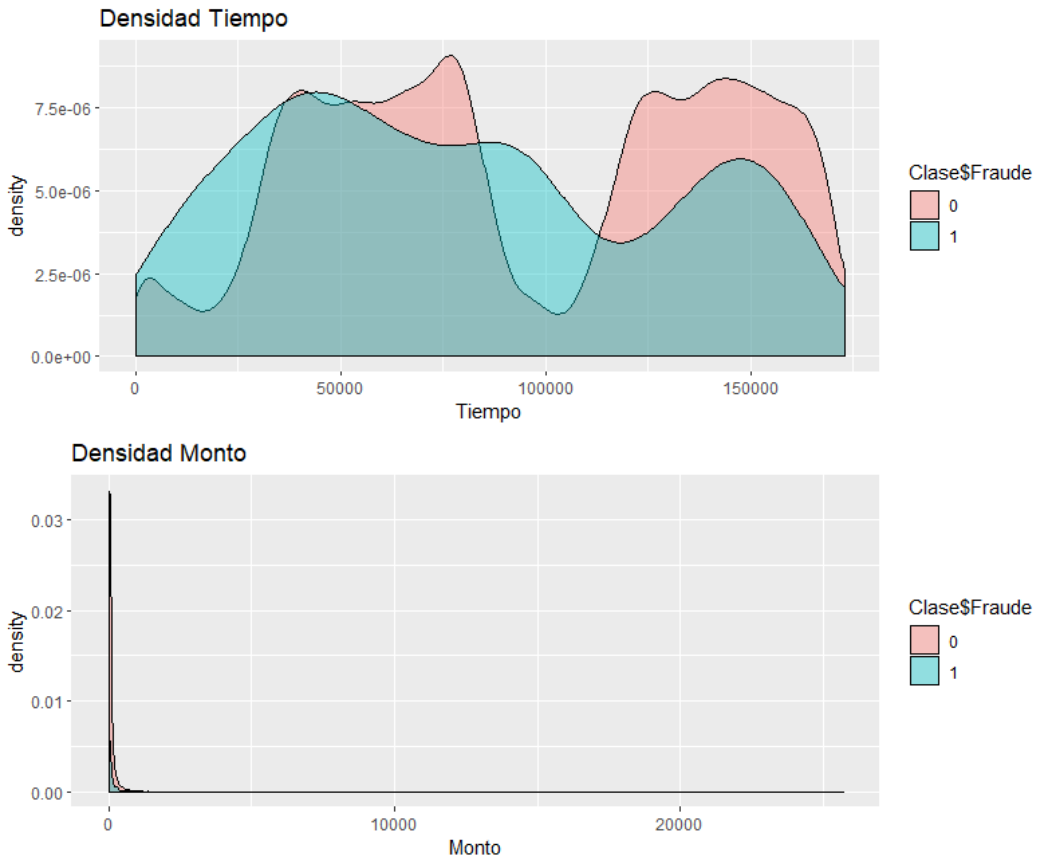


Figura 6.1: Distribuciones de las variables con y sin fraude.

Con este primer esbozo podemos decir que la variable tiempo muestra claramente una distribución bimodal o incluso multimodal para las transacciones que resultaron ser fraudulentas y las que no así como una notable diferencia entre los valores cercanos a 30 mil y 100 mil, por otro lado, la variable monto sugiere aplicar una transformación logarítmica o la sustracción de outliers empleando el método Z con el percentil del 99 %, para la transformación se aplicó el incremento de un valor épsilon igual a 0.001, es decir, $\mathcal{T}(x) = log(x + \epsilon)$, $\epsilon = 0.0001$ con el que nuestra transformación estará bien definida pues no tendrá valores en $-\infty$, entonces, veamos los resultados en la distribución a continuación.

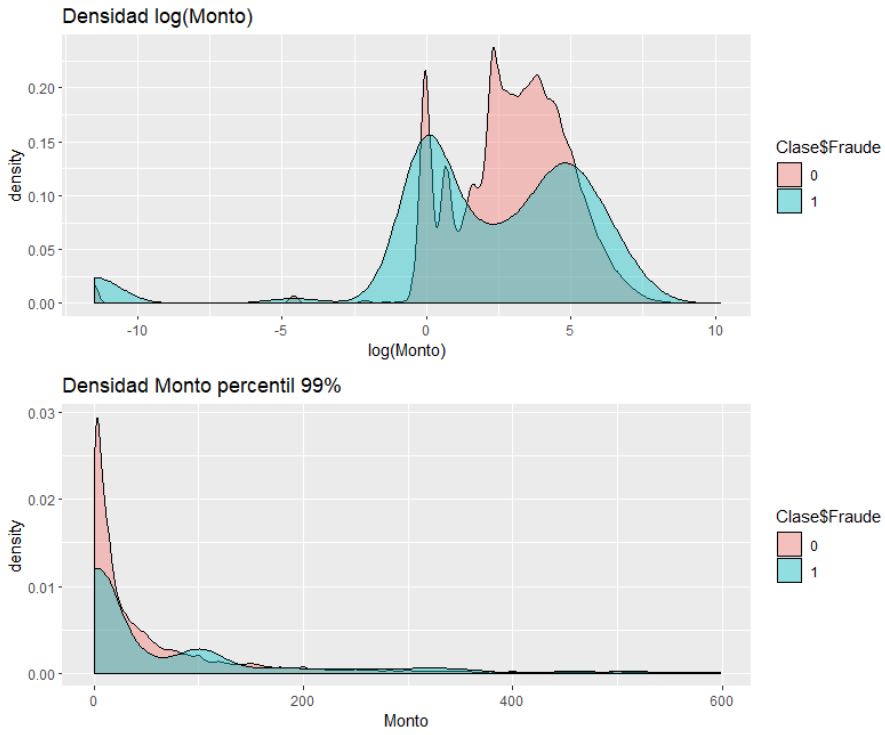


Figura 6.2: Distribuciones de la variable tiempo.

El logaritmo de la variable monto parece tener una distribución multimodal asociada para los dos tipos de clasificación, por lo que podríamos aplicar nuestro modelo, vemos también que al retirar el percentil del 99 % podemos tener una visión más reconocible de nuestra distribución, no obstante, no podemos reconocer una distribución log-normal o similares debido que la transformación T muestra más de un máximo local.

Con el objetivo de mostrar la presencia de observaciones atípicas y elegir qué transformación utilizaremos, veremos los diagramas de caja para los casos propuestos.

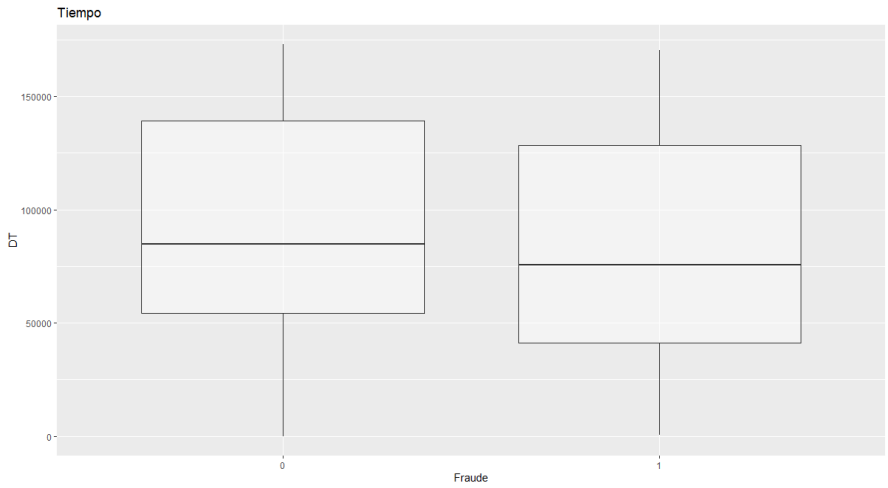


Figura 6.3: Fraude y no fraude (tiempo).

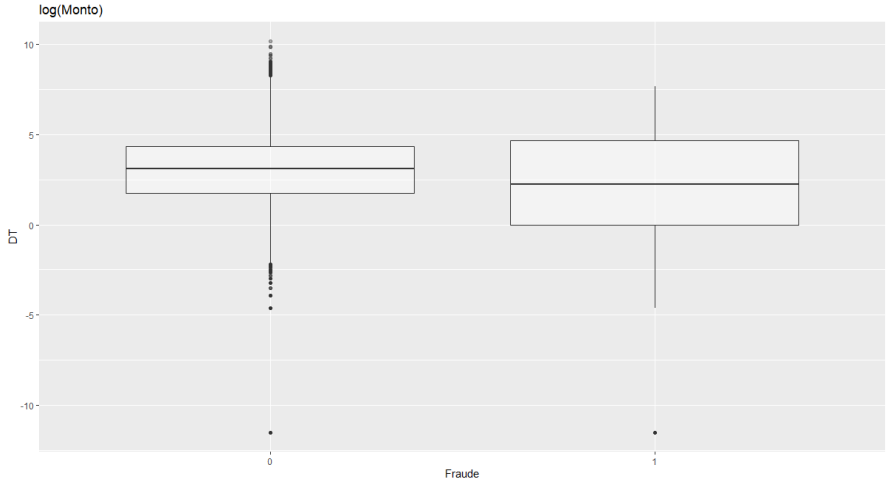


Figura 6.4: Fraude y no fraude (monto aplicando logaritmo).

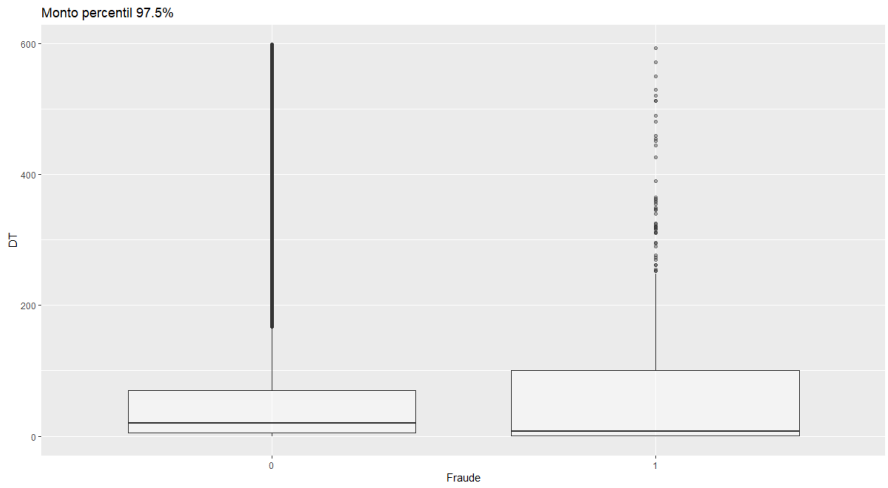


Figura 6.5: Fraude y no fraude (monto hasta el percentil 97.5 %).

Así como hay distribuciones cuyo valor esperado no existe (la distribución Cauchy es un ejemplo) también hay distribuciones multimodales con las que no hace sentido emplear diagramas de caja, pues los estadísticos observados no pueden describir la forma de la distribución, cabe aclarar que el caso en el que aplica es siempre y cuando las modas no se encuentran muy separadas.

También es posible que un outlier no sea coherente con la definición de observación atípica, pudiendo estar no en los extremos sino en el centro de la distribución multimodal, razón por la cual vemos tantas observaciones en el diagrama de caja del monto cuando podrían no serlo.

Ahora bien, procederemos a responder la primera pregunta. Al observar detenidamente, la transformación por medio del logaritmo elimina por completo la presencia de outliers para la variable de monto de los fraudes y reduce notablemente su presencia en donde no hubo fraude en comparación con el haber retirado el 2.5 % de la distribución en donde aun hay una gran presencia de estos, debido a ello, se optará por realizar el ajuste sobre la variable transformada por medio del logaritmo, no obstante, la forma de esta en la clase 0 es ciertamente errática en comparación con la generada por la del percentil, por lo que también se puede optar por aplicar el ajuste posteriormente a esta separación de outliers sin aplicar la transformación logarítmica.

Recordemos que para el valor 0 añadimos un épsilon, el valor de dicho épsilon

será aquel que no genere una nueva partición, teniendo total control sobre su valor tras aplicar logaritmo y no serán considerados como outliers. No obstante, al realizar una breve observación notamos que la probabilidad de que haya una transacción fraudulenta al haber pasado por este proceso es nula, al no haber registro de ninguna de ellas, lo cual es completamente coherente, pues no hay razón de hacer fraude para una transacción de valor 0.

Para establecer si existe una diferencia emplearemos el test de Kolmogorov-Smirnov (no es necesario aplicarlo a las transformaciones pues esta será aplicada a ambas clases).

Como resultado obtuvimos que la máxima de las diferencias en la función de distribución (el estadístico) cobró un valor de $D = 0.27119$ con ello, $p.valor < 2.2e-16$, valor muy cercano a 0 por lo que podemos rechazar la hipótesis nula y decir que ambas muestras provienen de una distinta distribución, de la misma manera rechazamos la hipótesis nula para la variable tiempo al haber obtenido $D = 0.16939$ y $p.valor = 1.15e-12$.

Ahora bien, procederemos a realizar el análisis de las distribuciones, para ello se ha diseñado la paquetería "FitUltD.[en] el software estadístico R, la cual reproduce el modelo antes definido en este escrito arrojando los siguientes resultados:

Función de distribución, densidad, cuantiles y generadora de números aleatorios de las variables que superaron, entcétera (más información en el apéndice del documento).

Para este ejercicio hemos de enfocarnos primero en la variable tiempo, como primer paso y para reducir tiempo de cómputo obtendremos una muestra aleatoria representativa del 5 % de nuestra variable, encontrando la mejor muestra por medio de la siguiente fórmula:

$$min_{D_i^+ \in D}\{D_i^+\}$$

Donde D es el conjunto de estadísticos de pruebas de KS obtenidas de diferentes semillas generadoras de números aleatorios, es decir, usaremos la muestra que tenga

la menor de las diferencias entre sus funciones de distribución.

Ahora aplicaremos el método diseñado para ajustar una distribución multimodal a nuestra muestra de datos con las siguientes características:

P.valor mínimo aceptable para decir que ha pasado las pruebas Kolmogorov-Smirnov y Anderson-Darling: 0.01.

Criterio para pasar las pruebas: No debe rechazar la hipótesis nula en ninguna de las pruebas.

Dado que hay números repetidos para una variable que conceptualmente es continua añadiremos un poco de ruido a la misma añadiendo números decimales al tiempo, es decir, sumaremos una v.a. Uniforme(-1, 1).

Esta suma afecta mínimamente a la distribución, pues esta maneja valores en las decenas de millar por lo que afectamos sus valores un 0.01 % con una distribución simétrica con media en 0.

Los resultados fueron los siguientes:

Para la muestra de las transacciones no fraudulentas:

Número de v.a's: 72. D = 0.0061283, p.value = 0.6884.

Min.	1stQu.	Mediana	Media	3rdQu.	Max.
13	54340	84605	94785	139204	172786
-67.58	54231.04	84659.9	94858.22	138884.09	173120.99

Para las transacciones fraudulentas (492):

Número de v.a's: 8. D = 0.02681, p.value = 0.8839.

Min.	1stQu.	Mediana	Media	3rdQu.	Max.
406	41242	75569	80747	128483	170348
-3614418	41256	73487	80435	127278	764240

En este caso vemos presencia de outliers que no hacen sentido, por lo que amerita conocer el tipo de distribuciones.

La presencia de la variable aleatoria Cauchy contribuye a la presencia de outliers al ser una distribución de cola pesada, por lo que volveremos a reproducir el análisis con la misma semilla, pero sin considerar esta variable aleatoria en el proceso y obteniendo los siguientes resultados:

Min.	1stQu.	Mediana	Media	3rdQu.	Max.
-9023	41242	73449	80583	126744	178981

Dist	AD_p.v	KS_p.v	Parm1	Parm2	estimateLL2	method
weibull	0.136754741	0.034017396	2.758646078	26844.85368	1	mle
weibull	0.03798159	0.056398658	2.778058736	29809.32831	1	mlg2
norm	0.312730006	0.116107734	24246.96386	9099.406396	1	mle
norm	0.312730006	0.116107734	24246.96386	9099.406396	1	mme
norm	0.368654804	0.324285092	24794.58659	9796.325795	1	mge
beta	7.23E-06	0.475624416	2.200719155	1.437593195	40086	mme
beta	7.23E-06	0.500954303	2.306060435	1.495076138	40086	mge
weibull	0.197959097	0.289337981	2.900170045	28042.82241	1	mge

Número de v.a's: 10. D = 0.020117, p.value = 0.9906.

Efectivamente, a pesar de haber incrementado el número de v.a's empleadas el estadístico se redujo y el p.valor incrementó considerablemente.

Vemos también la presencia de observaciones negativas y observaciones muy por encima del máximo de la muestra observada en la distribución, esto debido a que las distribuciones que se encuentran en los límites inferior y superior son Distribuciones normales, abarcando el 16.8 % y 16.4 % del total de la distribución respectivamente, afortunadamente contamos con más de una distribución que pasaron las pruebas de bondad y ajuste para escoger: El rango analizado corresponde a $[406, 40086]$.

Podemos elegir entre la distribución beta multiplicada por un factor de escala lo cual prevendría los casos negativos pero con el costo de un bajo estimador de Anderson-Darling, más distribuciones normales con las que se presentaría el mismo problema o la distribución $weibull(2.9001, 28042.8224)$.

El rango analizado corresponde a $[143354, 170348]$.

Para la distribución del límite superior realizamos el mismo análisis probando las siguientes distribuciones: gamma(489.729, rate = .003180806), norm(153888.408, 6934.835), lnorm(11.9437, .04.5257).

El valor máximo para la distribución del tiempo es 172,792 a partir de 284,807 observaciones y viendo que el valor máximo de nuestra distribución LogNormal para una simulación del mismo número de observaciones fue 190,823, veremos si la cola de la Gamma y Normal se ajusta más a la muestra, es decir, la cola se encuentra por

debajo de la distribución LogNormal.

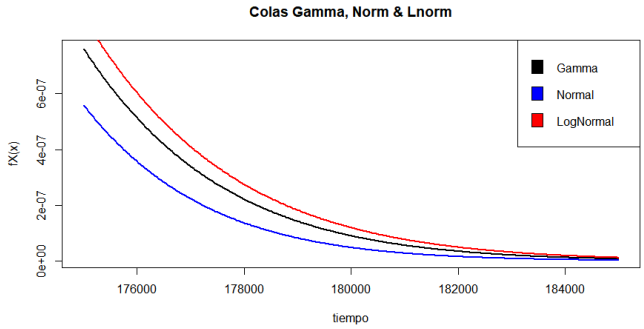


Figura 6.6: Comparativo Distribuciones.

En este caso también emplearemos una distribución que no genere una gran cantidad de outliers dada la naturaleza de nuestra muestra y a su vez que pase los criterios antes tomados, entonces, descartaremos las distribuciones: t, lognormal, normal, cauchy y beta.

Ahora, al proceder con el análisis estadístico podemos ver que los números obtenidos cobran mayor sentido que antes:

Min.	1stQu.	Median	Mean	3rdQu.	Max.
5330	41309	74917	81027	126257	171483

Y al aplicar las pruebas de bondad ajuste con validación cruzada obtenemos un p.valor medio de 0.9610299, es decir, perdemos .03 respecto al p.valor anterior pero ahora nuestros datos cobran un sentido real.

No rechazamos que la muestra generada por el modelo planteado y el tiempo tienen una distribución distinta y destacando que el p.valor se encuentra muy por encima de .05.

El primer resultado que podemos visualizar es la cantidad de variables aleatorias utilizadas a las que denominaremos componentes, en este caso hay 72 y 10 para las variables de tiempo sin y con fraude respectivamente, es decir, las muestras fueron divididas en 72 y 10 particiones, antes de proceder al análisis de estas, dado el amplio margen de no rechazo arrojado por Kolmogorov-Smirnov y al elevado número de componentes, merece la pena repetir el experimento cambiando los parámetros

iniciales, en este caso serán el p.valor mínimo aceptable y el número de clústers con el que separa el algoritmo k medias.

A partir de qué p.valor mínimo aceptable se sigue cumpliendo la prueba de k.s y se seguirán reduciendo el número de v.a's?

Probaremos entonces con los p.valores mínimos admisibles dentro del conjunto $A = \{.05, .025, 0.01, 0.005, 0.0025, 0.001, 0.0005, 0.00025, 0.0001, 0.00005, 0.000025, 0.00001, 0\}$.

El cambio en el número de clústers a 2 para los valores con los que hubo fraude arroja los siguientes resultados:

Número de variables: 3. D = 0.050813, p.value = 0.5491.

Min.	1stQu.	Median	Mean	3rdQu.	Max.
1991	43457	76453	81145	108814	174596

Dist.	Prop	ADp.v	KSp.v	Parm1	Parm2	method	Obs	L.Inf	L.Sup
weibull	0.537	0.013	0.093	2.079	50042	mlg2	264	406	82289
lnorm	0.195	0.096	0.052	11.461	0.081	mge	96	83934	118603
weibull	0.268	0.569	0.85	13	151703.5	mge	132	121238	170348

Ahora vemos una mejora muy significativa, pasando de 10 particiones a solamente 3, que a pesar de haber reducido el p.valor, se sigue rechazando la hipótesis nula con diferencia, veremos a continuación si un cambio en el p.valor sobre el modelo inicial o un cambio en ambos parámetros genera resultados similares.

Veamos la gráfica con solo las tres variables aleatorias:

P.valor inicial	Clústers	Número Comp.	Media P.valor
3,4	3	4	0.663
3,4	4	5	0.66
1	2	7	0.611
2,3,4	2	7	0.641
2	4	9	0.627
1,2	3	10	0.827

Vemos a continuación dos gráficos de la matriz de resultados completa aplicando validación cruzada para cada punto, es decir, realizando hasta 10 pruebas con muestras aleatorias distintas y mostrando la media de estas como el p.valor final y con las siguientes características:

1. A mayor altitud más exacto es el modelo

2. Entre más azul, más simple es el modelo

3. Ente más cerca se esté del origen mayor rango de error se le permitirá al modelo

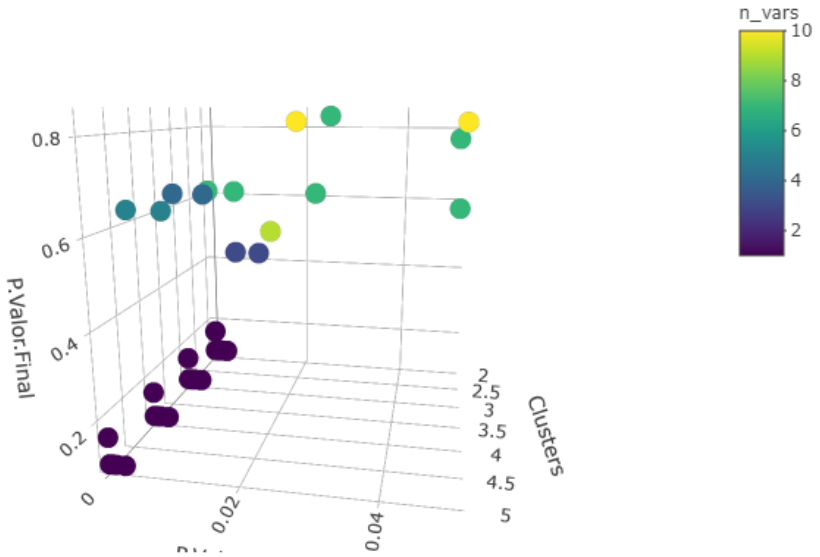


Figura 6.7: Matriz de Resultados ejes XZ.

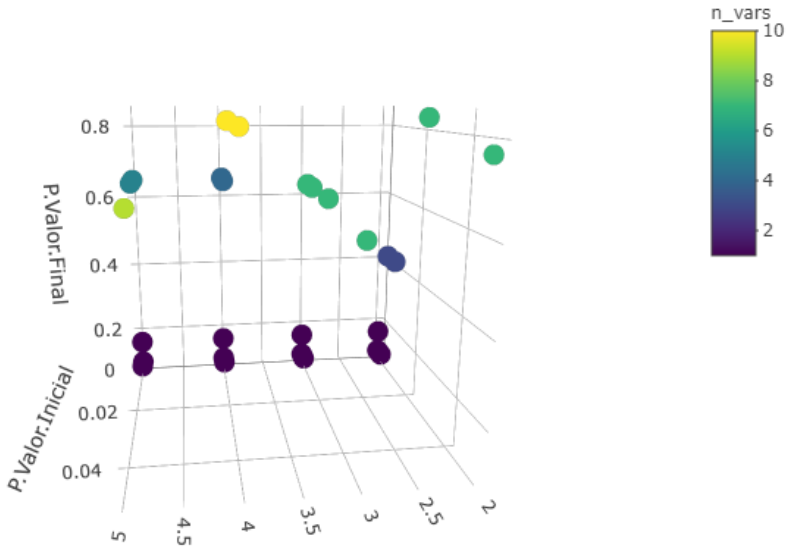


Figura 6.8: Matriz de Resultados ejes YZ.

Notemos que el color representa el número de variables o número de componentes. La matriz nos indica que distintas combinaciones de parámetros pueden generar los mismos resultados, en particular observamos:

1. La media del p.valor del modelo final obtenido no tiene una relación estrictamente creciente o decreciente con los parámetros iniciales.

2. Un menor número de clústers no afectará los resultados finales si el p.valor inicial es muy bajo, esto porque al ser menos estrictos con las pruebas AD y KS no hay necesidad de hacer otra partición, razón por la cual vemos el mismo color.

3. Los modelos con un mayor número de particiones generalmente tienen un mayor p.valor final.

4. El número total de clústers no parece añadir suficiente al p.valor, sin embargo, es seguro que aumenta en gran medida la complejidad del modelo.

Cabe destacar que la cantidad de observaciones afecta los resultados, no obstante, vemos un fenómeno bastante llamativo, pues a pesar de que el p.valor inicial disminuye

a valores menores a .02 (el p.valor con el que se decidió rechazar las pruebas AD y KS) aun así, nuestra muestra total sigue pasando las pruebas con un p.valor cercano a .20, es decir, el p.valor final es mayor a cualquiera de los p.valores iniciales en la vasta mayoría de los casos y con bastante diferencia. Pero no seamos ingenuos, pues en los valores extremos es claro que hay una mejora en el p.valor final al ir incrementando el p.valor inicial. Vemos pues el criterio de Akaike modificado para cada elemento de la matriz y decidamos cuales son los mejores parámetros iniciales para nuestro modelo.

Ahora que hemos encontrado la mejor forma de clasificar este tipo de distribuciones, se reproducirá el ajuste para los valores en donde no hubo fraude obteniendo la siguiente matriz:

P.Valor.Inicial	Clústers	n_vars	P.Valor(muestra)	P.Valor(total)	Criterio
0.025	2	48	0.99999957	0.4608896	89.80974
0.01	2	39	0.99992249	0.4549604	71.82870
0.005	2	33	0.99250649	0.4198035	59.94366
0.0025	2	29	0.99642590	0.4100457	51.97662
0.001	2	23	0.94920519	0.3587117	40.16177
5.00E-04	2	20	0.96005086	0.3215668	34.31398
0.00025	2	20	0.97136963	0.3623131	34.14797
1.00E-04	2	19	0.96434685	0.3353963	32.25499
5.00E-05	2	19	0.96968254	0.3593642	32.15925

De forma objetiva, el incremento del p.valor es de solamente un .03 entre el mejor y el peor modelo, no obstante, la complejidad es casi el doble, empleando nuestro criterio de elección del mejor modelo, podemos concluir que el modelo con 19 variables y un p.valor inicial de 0.00005, es el más óptimo y al revisarlo obtenemos el siguiente ajuste:

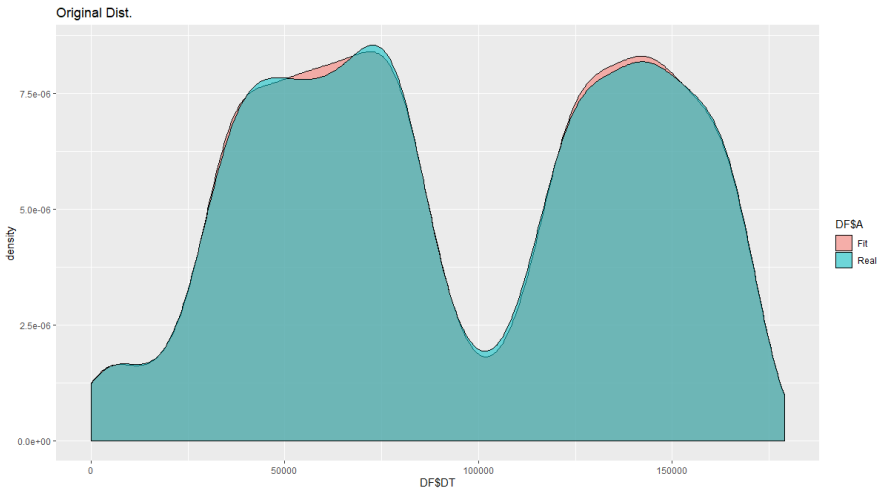


Figura 6.9: Ajuste de un modelo con 19 v.a. de transacciones no fraudulentas.

Recordemos que el ajuste del p.valor se está realizando con una muestra aleatoria representativa, es por eso que en la tabla los p.valores de la muestra son mayores a las del totalidad, sin embargo, ambos son suficientes para rechazar la hipótesis nula, aunado a ello, notamos que al ajustar 10,000 observaciones de la muestra aleatoria se incrementó también el número de variables y se redujo el p.valor, por lo que podemos decir que la cantidad de observaciones y la representatividad de la muestra juegan un papel importante en el modelo y podemos considerarlos como parámetros de entrada a ser evaluados.

Es de esperar que un modelo sobreajustado encaje muy bien con la muestra con la que fue creado, no obstante, puede que no esté generalizando como desearíamos.

¿Hay alguna forma de determinar los parámetros iniciales tales que maximicen el criterio de nuestro modelo?

6.3. Otros campos de aplicación

Las distribuciones multimodales no suelen aplicarse con frecuencia debido a la complejidad al calcular sus parámetros, estas se pueden manifestar como fenómenos de diversa índole, en la actualidad y en la práctica se suele acudir a la distribución empírica o distribuciones como poisson o binomial negativa para la frecuencia de los

siniestros y weibull para la severidad, creando modelos compuestos para distribuciones conocidas como forma de estimar el monto de reclamaciones futura.

Las funciones de distribución como Log-Normal, Weibull o Gamma para medir la severidad del siniestro son útiles de forma teórica, no obstante, llegan a explicar de forma muy escueta la realidad de las reclamaciones, es en este factor en donde entra este modelo como una nueva propuesta, pues existen reclamaciones cuya suma asegurada se encuentra muy por encima de la media esperada, lo que provoca que su cola deba estudiarse como una parte separada.

Capítulo 7

Conclusiones

El criterio principal por el que se debe empezar a tratar un conjunto de datos es por medio de sus variables discretas, pues los errores de ajuste se deben principalmente a que la información no ha sido tratada correctamente.

La separación de la información es una buena práctica aun si parece no haber razón para dividirla, pues, si se obtienen distribuciones distintas para cada partición, habremos reducido la complejidad del ejercicio, por el contrario, si poseen la misma distribución habremos obtenido una muestra representativa del total de la información, en caso de contar con un conjunto de datos de gran tamaño se puede optar por reducir dimensiones, tomar muestras representativas o priorizar solo las variables más significativas.

Las distribuciones multimodales también pueden significar el reconocimiento de un comportamiento atípico que parece estar influenciado por dos o más componentes más simples que no logramos medir por medio de otra variable, en la actualidad no se sabe si esto es debido a que los instrumentos de medición que poseemos no son capaces de reconocer todas las partes necesarias que separen los datos de forma natural o si se presentó un error o pérdida en los datos recabados.

Una de las principales desventajas es el sobreajuste, en caso de que una muestra muy grande sea dividida en una gran cantidad de bloques, esto indica que no será bueno generalizando, pues la cantidad de componentes que tendrá la función afectan muy fuertemente por medidas como el criterio de Akaike.

Antes de ejecutar el algoritmo, siempre hay que verificar que no haya variables categóricas que podamos usar a nuestro favor.

Debido a la cantidad de tiempo de cómputo y a la estructura de las funciones se decidió limitar el ajuste a variables aleatorias más empleadas, no obstante, es posible incorporar nuevas variables aleatorias.

La importancia del entendimiento de nuestros datos es también un factor clave, por ejemplo, al hablar de tiempo, no es admisible encontrar valores negativos o extremadamente grandes, en contraste, el monto permite una mayor flexibilidad, pues es posible que el monto de la comisión sea mayor al monto extraído y entonces, admitir valores negativos cercanos a cero.

Debido a ello, se realizaron diversas mejoras y consideraciones para este modelo, en donde es posible restringir la utilización de variables aleatorias que incumplan con estos requerimientos, tales como la inclusión de variables truncadas o el poder descartar la utilización de aquellas que tengan colas pesadas.

También es importante la corroboración de nuestros resultados contemplando distintos test durante el proceso, así como el empleo de técnicas como validación cruzada.

Apéndice A

Anexo Código en R

Se ha realizado un paquete en R exclusivo para este trabajo, disponible con el siguiente código: install.packages("FitUltD")

Este incluye dos funciones: FDist y FDistUlt.

FDist que ajusta una distribución a una muestra de datos dada con la ventaja de que ajusta la función indicadora de esta a la forma en la que están dados los datos, el código comentado se encuentra a continuación.

Parámetros:

Parámetro X: Muestra aleatoria observada.

Parámetro gen: Un entero positivo, indica la cantidad de números aleatorios que se generarán empleando la distribución resultante.

Parámetro Cont: TRUE, por defecto la distribución es considerada como continua.

Parámetro inputNA: En caso de haber valores perdidos, imputará el valor colocado en este parámetro.

Parámetro plot: FALSE. Si es TRUE, una gráfica de la distribución de los datos será retornada.

Parámetro $p.val_min$: 0.05 por defecto, es el valor mínimo del p.valor para las pruebas de Anderson-Darling y KS para no rechazar la hipótesis nula.

Parámetro crit: Un entero positivo para definir qué criterio se usará. Si es 1 muestra las distribuciones que no fueron rechazadas por las pruebas Anderson-Darling o Kolmogorov-Smirnov, en otros casos el criterio es que no deben ser rechazadas por

ambas pruebas.

Parámetro DPQR: Si es TRUE crea la función de distribución, densidad y función cuantiles con los nombres dfit, pfit y qfit respectivamente.

Resultado: Calcula el nombre de la distribución con sus respectivos parámetros, una función generadora de números aleatorios, una muestra aleatoria de dicha función, los resultados de las pruebas de Anderson-Darling y KS p.values, una gráfica entre la muestra real y la distribución asignada y una lista con las funciones de densidad, distribución y cuantiles.

Ejemplo:

```
FDist<-function(X, gen=1,Cont=TRUE, inputNA , plot=FALSE, p.val_
    min=.05, crit=2,DPQR=TRUE){
  #Crear variables que serán frecuentemente utilizadas como
      los valores únicos de X, la cardinalidad de estos y del
      conjunto X.
  uX<-unique(X)
  luX<-length(uX)
  lX<-length(X)
  #Por defecto elimina los valores NA, amenos que exista un
      valor a imputar.
  if(missing(inputNA)){X<-na.omit(X)}
  else{X<-ifelse(is.na(X),inputNA,X)}
  if(lX==0){
    return(NULL)
  }
  #Retira infinitos
  X<-X[X!=(-Inf) & X!=Inf]
  #Si hay un solo valor, eso equivale a simular una normal
      con desviación estándar 0.
  if (luX<2) {
```

```
    fun_g<-function(n=gen){return(rep(X[1],n))}
    return(list(paste0("norm(",X[1],",0)"),fun_g,rep(X[1],gen
        ),data.frame(Dist="norm",AD_p.v=1,KS_p.v=1,Parm1=X[1],
        Parm2=0,estimateLL1=0,estimateLL2=1,method="assumption
        ",PV_S=2,Obs=gen,Lim_inf=min(X),Lim_sup=max(X)),NULL))
  }
  #Si hay dos valores equivale a una distribución binomial.
     con parámetro p igual a la proporción del primer valor
     en X.
  if (luX==2) {
    X<-sort(X)
    p<-length(X[X==uX[2]])/lX
    gene<-stats::rbinom
    formals(gene)[1]<-lX
    formals(gene)[2]<-1
    formals(gene)[3]<-p
    distribu<-paste0("binom(",p,")")
    MA=gene(n = gen)
    if(plot){
      DF<-rbind(data.frame(A="Fit",DT=MA),
                data.frame(A="Real",DT=X))
      pl <- ggplot2::ggplot(DF,ggplot2::aes(x=DF$DT, fill=DF$A
          )) + ggplot2::geom_density(alpha=0.4) +ggplot2::
          ggtitle(distribu)
    }else{
      pl<-NULL
    }
    return(list(distribu,gene,MA[1:gen],data.frame(Dist="
        binom",AD_p.v=1,KS_p.v=1,Parm1=1,Parm2=p,estimateLL1
        =0,estimateLL2=1,method="assumption",PV_S=2,Obs=gen,
```

```
      Lim_inf=min(X),Lim_sup=max(X)),pl))
}
#Si la distribución de desea evaluar como continua pero hay
    valores repetidos, se a adirán centesimales obtenidos
   de una distribución uniforme (-1,1) para evitar
   conflicots con la ejecución de las pruebas sin que se
   afecte la distrubución.
if(lX!=luX & Cont){
  X<-jitter(X, amount = 1/100)
}
#A través de esta sección, si la variable "Nombres" no
   existe, asignará las distribuciones del paquete stats
   para que sean icluidas en las pruebass de bondad y
   ajuste.
if(!exists("Nombres")){
  Nombres=c("exp","pois","beta","gamma","lnorm","norm","
     weibull","nbinom","hyper","cauchy","binom","unif","t")
}
#Se creará una lista con las funciones de densidad,
   distribución, cuantiles y generadora de números
   aleatorios de las variables que se encuentran en el
   objeto "Nombres" así como si esta es continua o discreta
    y el dominio donde se encuentra bien definida.
lnbs<-length(Nombres);lnb<-(1:lnbs)
comb<-expand.grid(c("p","d","q","r"),Nombres);comb<-comb[
   order(comb[[1]]),]
func<-purrr::map2(as.character(comb[[2]]),as.character(comb
   [[1]]),~get(paste0(.y,.x)))
DIS<-list(Nombres=Nombres,
          p=func[lnb],
```

```
            d=func[lnb+lnbs],
            q=func[lnb+lnbs*2],
            r=func[lnb+lnbs*3],
            d_c=c(1,0,1,1,1,1,1,0,0,1,0,1,1),
            indicadora=c("0","0","01","0","0","R","0","0","0"
                ,"R","0","R","R")
)
#Se separarán por medio de su dominio, las que son mayores
    que 0, las que se encuentran en los reales y las que se
    encuantran en el intervalo (0,1).
DIS<-purrr::map(DIS,~subset(.x, DIS$d_c==as.numeric(Cont)))
DIS_0<-purrr::map(DIS,~subset(.x, DIS$indicadora=="0"))
DIS_R<-purrr::map(DIS,~subset(.x, DIS$indicadora=="R"))
DIS_01<-purrr::map(DIS,~subset(.x, DIS$indicadora=="01"))
if(sum(purrr::map_dbl(DIS_0,~length(.x)))==0){DIS_0<-NULL}
if(sum(purrr::map_dbl(DIS_R,~length(.x)))==0){DIS_R<-NULL}
if(sum(purrr::map_dbl(DIS_01,~length(.x)))==0){DIS_01<-NULL
    }
#Definimos parámetros de escala igual a 1 y deslazamiento
    igual a 0 así como un épsiloon.
bt<-X
despl<-0
escala<-1
eps<-1E-15
#Se crearán tres variables aleatorias transformadas, es
    decir, se le restará el valor mínimo y se dividirá entre
     la diferencia del máximo y el mínimo alterando los pará
    metros de escala y desplazamiento para poder recuperar
    la transformación, esto para que todas las
    distribuciones estén bien definidas en su dominio, el
```

```
    valor de épsilon se emplea para sumárselo a las
    variables que deben ser estríctamente mayores que 0.
if (sum(X<0)>0){
  if (sum(X<0)/lX<0.03){
    bt<-ifelse(X<0,eps,X)
    b_0<-bt
  }else{
    b_0<-bt-min(bt)+eps
    despl<- min(bt)
  }
}else{
  b_0<-bt
}
if(max(X)>1){
  escala<-max(bt)
  b_01<-(bt-despl)/(escala-despl)
}else{
  b_01<-bt
}
#Se creará la función que realiza el ajuste con los métodos
    de momentos y máxima verosimilitud de los paquetes
    fitdistplus y MASS.
fit_b<-function(bt,dist="",Cont.=Cont){
  if(is.null(dist)){return(NULL)}
  Disc<-!Cont
  aju<-list()
  if(!dist %in% DIS_01$Nombres){
    suppressWarnings(aju[[1]]<-try(fitdistrplus::fitdist(bt
        ,dist,method = "mle",discrete = Disc),silent = TRUE)
        )
```

```
    }
    suppressWarnings(aju[[2]]<-try(fitdistrplus::fitdist(bt,
        dist,method = "mme",discrete = Disc),silent = TRUE))
    suppressWarnings(aju[[3]]<-try(fitdistrplus::fitdist(bt,
        dist,method = c("mge"),discrete = Disc),silent = TRUE)
        )
    suppressWarnings(aju[[4]]<-try(MASS::fitdistr(bt,dist),
        silent = TRUE))
    if(!assertthat::is.error(aju[[4]])){aju[[4]]$distname<-
        dist}
    if(assertthat::is.error(aju[[1]]) & assertthat::is.error(
        aju[[2]]) &
        assertthat::is.error(aju[[3]]) & assertthat::is.error(
            aju[[4]])){
      return(list())
    }
    funciónales<-!purrr::map_lgl(aju,~assertthat::is.error(.x
        ))
    names(aju)<-c("mle","mme","mge","mlg2")
    aju<-aju[funciónales]
    return(aju)
}
#Ahora se aplicará el ajuste dependiendo del dominio de la
    distribución, es decir, tendremos tres casos (en los
    reales, mayores que 0 y entre 0 y 1).
suppressWarnings(try(aju_0<-purrr::map(DIS_0$Nombres,~fit_b
    (b_0,.x)),silent = TRUE))
suppressWarnings(try(aju_R<-purrr::map(DIS_R$Nombres,~fit_b
    (bt,.x)),silent = TRUE))
suppressWarnings(try(aju_01<-purrr::map(DIS_01$Nombres,~fit
```

```
    _b(b_01,.x)),silent = TRUE))
  #Se unen los resultados  se descartan los que estén vacíos.
  AAA<-list(aju_0,aju_R,aju_01)
  descate<-purrr::map(AAA,~length(.x))!=0
  AAA<-AAA[descate]
  bts<-list(b_0,bt,b_01)[descate]
  num<-0
  Compe<-data.frame()
  #Dependiendo de los parámetros obtenidos, se evaluará cada
      resultado empleando Kolmogorov-Smirnov y Anderson-
      Darling.
  for (aju_ls in 1:length(AAA)) {
    aju<-AAA[[aju_ls]]
    aju<-aju[purrr::map_lgl(aju,~length(.x)>0)]
    bs<-bts[[aju_ls]]
    #Para cada distribución de cada partición se realizan
        ambas pruebas.
    for (comp in 1:length(aju)) {
      if(length(aju)==0 ||length(aju[[comp]])==0){next()}
      for (ress in 1:length(aju[[comp]])) {
        num<-num+1
        if(length(aju[[comp]])!=0){evaluar<-aju[[comp]][[ress
            ]]
        }else{evaluar<-NULL}
        if (is.null(evaluar) | length(evaluar)==0 |
            c(NA) %in% evaluar$estimate | c(NaN) %in% evaluar
                $estimate) {next()}
        distname<-evaluar$distname
        method<-names(aju[[comp]])[[ress]]
        dist_pfun<-try(get(paste0("p",distname)),silent =
```

```
    TRUE)
  dist_rfun<-try(get(paste0("r",distname)),silent =
    TRUE)
  if(assertthat::is.error(dist_rfun)){next()}
  #Se extraen los parámetros de cada función y se
    reemlazan con los obtenidos por las estimaciones.
  argumentos<-formalArgs(dist_pfun)
  argumentos<-argumentos[argumentos %in% names(evaluar$
    estimate)]
  num_param<-length(argumentos)
  evaluar$estimate<-evaluar$estimate[names(evaluar$
    estimate) %in% argumentos]
  #Se aplican distintos casos dependiendo de la
    cantidad de parámetros de cada función.
  if(num_param==1){
    EAD<-try(AD<-ADGofTest::ad.test(bs,dist_pfun,
      evaluar$estimate[1]),silent = TRUE)
    KS<-try(stats::ks.test(bs,dist_pfun,evaluar$
      estimate[1]),silent = TRUE)
    if(assertthat::is.error(KS)){KS<-data.frame(p.value
      =NA)}
    if(assertthat::is.error(EAD)){next()}
    if(is.na(KS$p.value)){next()}
    Chs<-data.frame(p.value=0)
  }
  if(num_param==2){
    suppressWarnings(
      Err_pl<-try(AD<-ADGofTest::ad.test(bs,dist_pfun,
        evaluar$estimate[1],evaluar$estimate[2]),
        silent = TRUE))
```

```
    if (assertthat::is.error(Err_pl)) {
      Err_pl<-try(AD<-ADGofTest::ad.test(bs,dist_pfun,
          evaluar$estimate[1],,evaluar$estimate[2]),
          silent = TRUE)
    }
    KS<-try(stats::ks.test(bs,dist_pfun,evaluar$
        estimate[1],evaluar$estimate[2]),silent = TRUE)
    if(assertthat::is.error(KS)){KS<-data.frame(p.value
      =NA)}
    if(assertthat::is.error(Err_pl)){next()}
    if(is.na(KS$p.value)){next()}
    suppressWarnings(
      EE_Chs<-try(dst_chsq<-dist_rfun(length(bs),
          evaluar$estimate[1],evaluar$estimate[2]))
    )
    if(assertthat::is.error(EE_Chs) | prod(is.na(EE_Chs
        ))==1){
      dst_chsq<-dist_rfun(length(bs),evaluar$estimate
          [1],,evaluar$estimate[2])
    }
    Chs<-data.frame(p.value=0)
  }
  pvvv<-p.val_min
  if(all(is.na(KS$p.value))){
    crit<-AD$p.value>pvvv
  }else{
    if(crit==1){
      crit<-AD$p.value>pvvv | KS$p.value>pvvv
    }else{
```

```
      crit<-AD$p.value>(pvvv) & KS$p.value>(pvvv)
    }
  }
  if(crit){
    if(aju_ls %in% 3){
      estimate3=despl
      estimate4=escala
    }else if(aju_ls==1){
      estimate3=despl
      estimate4=1
    }else{
      estimate3=0
      estimate4=1
    }
    #Los resultados de cada ajuste que haya pasado las
       ruebas se acumulará dentro de una tabla.
    Compe<-rbind(Compe,data.frame(Dist=distname,AD_p.v=
       AD$p.value,KS_p.v=KS$p.value,
                                     Chs_p.v=Chs$p.value,
                                     Parm1=evaluar$
                                        estimate[1],Parm2=
                                        evaluar$estimate
                                        [2],
                                     estimateLL1=estimate3
                                        ,estimateLL2=
                                        estimate4,method=
                                        method
    ))
    }else{
    next()
```

```
        }

      }
    }
  }
  #En caso de no haber ninguna función que haya pasado la
     prueba se termina el algoritmo indicando que no hubo
     ajuste.
  if (nrow(Compe)==0) {
    warning("No fit")
    return(NULL)
  }
  Compe$PV_S<-rowSums(Compe[,2:4])
  Compe<-cbind(Compe,data.frame(Obs=gen,Lim_inf=min(X),Lim_
     sup=max(X)))
  #Se determina cuál fue el mejor ajuste a tarvés del p.valor
     de ambas pruebas, aquel que haya pasado ambas y en suma
     sea el número mayor.
  WNR<-Compe[Compe$PV_S %in% max(Compe$PV_S),][1,]
  distW<-WNR$Dist
  paramsW<-WNR[1,names(Compe)[startsWith(names(Compe),"estim"
     )]]
  paramsW<-paramsW[,!is.na(paramsW)]
  if(gen<=0){gen<-1}
  #Se crea la función generadora de números aleatorios que
     incluye la operación inversa de escala y deslazamiento
     para que no haya problema.
  generadora_r<-function(n=gen,dist=distW,params=paramsW){
    fn<-get(paste0("r",dist))
    formals(fn)[1]<-n
```

```
    for (pr in 1:(length(params)-2)) {
      formals(fn)[pr+1]<-as.numeric(params[pr])
    }
    fn()*params[,length(params)]+params[,length(params)-1]
  }
  #Si el parámetro DPQR es verdadero, obtiene también las
     funciones de densidad, distribución y de cuantiles.
  if(DPQR){
    generadoras<-function(x,tipo,dist=distW,params=paramsW){
      fn<-get(paste0(tipo,dist))
      formals(fn)[1]<-x
      for (pr in 1:(length(params)-2)) {
        formals(fn)[pr+1]<-as.numeric(params[pr])
      }
      class(fn)<-"gl_fun"
      fn
    }
    rfit<-generadora_r
    class(rfit)<-"gl_fun"
    pfit<-generadoras(1,"p")
    qfit<-generadoras(1,"q")
    dfit<-generadoras(1,"d")
  }
  #Se ejecutan las funciones para generar los resultados
     obteniendo una tabla con la unión de estos.
  MA<-generadora_r()
  paramsAUX<-c()
  paramsW2<-data.frame()
  for(cl in 1:nrow(paramsW)){
    paramsW2<-rbind(paramsW2,round(paramsW[1,],3))
```

```
  }
  if(paramsW2[,length(paramsW2)]!=1 | paramsW2[,length(
      paramsW2)-1]!=0){
    distribu<-paste0(WNR$Dist,"(",paste0(paramsW2[,1:(length(
        paramsW2)-2)],collapse = ",␣"),")*",paramsW2[,length(
        paramsW2)],"+",paramsW2[,length(paramsW2)-1])
  }else{
    distribu<-paste0(WNR$Dist,"(",paste0(paramsW2[,1:(length(
        paramsW2)-2)],collapse = ",␣"),")")
  }
  p<-c()
  #Se crea una gráfica con la muestra aleatoria generada y X,
      es decir, los datos reales.
  if(plot){
    DF<-rbind(data.frame(A="Fit",DT=MA),
              data.frame(A="Real",DT=X))
    p <- ggplot2::ggplot(DF,ggplot2::aes(x=DF$DT, fill=DF$A))
        + ggplot2::geom_density(alpha=0.4) +ggplot2::ggtitle(
        distribu)
  }
  #Se unen los resultados con mejores aproximaciones a la
      función real
  class(WNR)<-c(class(WNR),"data.frame.wnr")
  return(list(distribu,generadora_r,MA,WNR[,-4],p,list(rfit,
      pfit,dfit,qfit),Compe[,-4]))
}
```

La segunda función FDistUlt emplea la función anterior para ajustar una distribución y agregar los métodos de clasificación en caso de ser neceario, es decir, en caso de no haber ajuste separa la muestra. Por defecto utilizando el método de K-Medias y continúa de forma recursiva hasta hallar las distribuciones que simulan a la muestra.

Ajusta un conjunto de ovservaciones para determinar si provienen de una determinada distribución.

Parámetros:

Parámetro X: Muestra de observaciones a ser ajustadas

Parámetro n.obs: Un entero positivo, es el tamaño de la muestra aleatoria generada por la función.

Parámetro ref: Número de clústers que serán usados por el algoritmo k-medias para hacer la searación de la distribución. En caso contrario utiliza la clasificación de mclust por defecto.

Parámetro crt: Criterio para decidir si se pasan ambas pruebas (KS y AD) para rechazar la hipótesis nula o si solamente una es suficienete.

Parámetro plot: FALSE. Si es TRUE, genera una gráfica de la función de densidad total.

Parámetro subplot FALSE: Si es TRUE, genera una gráfica de cada una de las particiones.

Parámetro $p.val_min$: P. valor mínimo aceptable no rechazar la hipótesis nula.

Resultado: Una lista que contiene las funciones generadoras de números aleatorios de cada partición, una muestra aleatoria, una tabla con los resultados de las pruebas de KS y AD, las gráficas correspondientes y una tabal con todos los otros posibles resultados que hayan pasado las pruebas.

Ejemplo:

```
FDistUlt<-function(X,n.obs=length(X),ref="OP",crt=1,plot=
    FALSE,subplot=FALSE,p.val_min=.05){
  #Primero reviso si ref es un número, en caso de serlo
      revisar que no sea tan grande (1/3 de la muestra total),
       para poder ejecutar el algoritmo de k-medias sin
      errores.
  if(!is.numeric(ref)){}else{
    if(ref>length(X)/3){warning("Number of clusters must be
```

```
        less␣than␣input␣length/3")
      return(NULL)}}
  #Dada la naturaleza recursiva del algoritmo, definiremos
     una función internamente que realice el proceso de
     separación de clústers.
  desc<-function(X, fns=FALSE, ref.=ref, crt.=crt, subplot.=
     subplot, p.val_min.=p.val_min){
    eval<-function(X, fns.=fns, crt.=crt, subplot.=subplot, p.val
       _min.=p.val_min){
    #Se evalua la muestra en la función que da como resultado
       si pasó o no la prueba, en caso de no pasarla,
       regresa el valor NULL como respuesta.
      FIT<-FDist(X, length(X), crit = crt, plot = subplot, p.val_
         min=p.val_min)
      FIT
    }
    #Definimos la función que divide la muestra con base en
       los resultados de KS y AD de la función anterior.
    div<-function(X, ref.=ref){
      df<-data.frame(A=1:length(X), B=X)
      #El primer criterio de separación será si X contiene
         valores discretos y continuos, realizando pruebas
         con las respectivas distribuciones.
      Enteros<-X-floor(X)==0
      if(any(Enteros)){
        if(all(Enteros)){
          if(!is.numeric(ref)){
          #En caso de no hber colocdo un número de clústers
             específico, se utilizará la función Mclust para
             obtener el mejor número de gruos utilizando el m
```

```
            étodo del codo como criterio.
          mod1<-mclust::Mclust(X,modelNames=c("E", "V"))$
              classification
        #En caso de retornar un solo grupo, se volverá a
            hacer el análisis con dos grupos por defecto.
          if(length(table(mod1))==1){
            df$CL<-kmeans(df,2)$cluster
          }else{
            df$CL<-mod1
          }
        }else{
          df$CL<-kmeans(df,ref)$cluster
        }
      }else{
        df$CL<-ifelse(Enteros,1,2)
      }
    }else{
    #Se realizará el mismo procedimiento con números reales
        en lugar de enteros.
      if(!is.numeric(ref)){
        mod1<-mclust::Mclust(X)$classification
        if(length(table(mod1))==1){
          df$CL<-kmeans(df,2)$cluster
        }else{
          df$CL<-mod1
        }
      }else{
        df$CL<-kmeans(df,ref)$cluster
      }
    }
```

```
    #Se realizará la partición de la muestra empleando la
        variable clúster de k-medias como criterio.
    CLS<-purrr::map(unique(df$CL),~df[df$CL==.x,2])
    CLS
    return(CLS)
  }
  #Se aplicará la función anterior a X.
  suppressWarnings(EV<-eval(X,fns))
  #Si no hubo ajuste, es decir, si se regresa el valor null
      volverá a hacer nuevas particiones hasta que ajuste o
      hasta llegar a un tama o de muestra menor a 50, en
     cuyo caso asignará la distribución normal a los datos.
  if(is.null(EV)){
    if(length(X)>50){
      DV<-purrr::map(div(X),~desc(.x,fns))
      return(DV)
    }else{
    #Si los datos son NULL (no hubo ajuste) y el nama o de
        muestra no es mayor a 50 entonces asignará la
       distribución normal y todo lo que ello implica (los
       parámetros de dicha distribución serán la media y la
        desviación estándar de X).
      FN<-rnorm
      formals(FN)[1]<-length(X)
      formals(FN)[2]<-mean(X)
      formals(FN)[3]<-ifelse(length(X)==1,0,sd(X))
      dfw<-data.frame(Dist="norm",AD_p.v=1,KS_p.v=1,Parm1=
          mean(X),Parm2=sd(X),estimateLL1=0,estimateLL2=1,
          method="asumption",PV_S=2,Obs=length(X),Lim_inf=
          min(X),Lim_sup=max(X))
```

```
        class(dfw)<-c(class(dfw),"data.frame.wnr")
        return(list(paste0("norm(",mean(X),",",ifelse(length(
          X)==1,0,sd(X)),")"),FN,FN(),dfw))
      }
    }else{
      return(EV)
    }
  }
  #Ya que se han realizado todas las particiones y resultados
    , dado que la funión purrr::map retorna listas, estas se
     encontrarán anidadas dada la naturaleza recursiva del
    algoritmo, por lo que se creará una función que retorne
    cada elemento anidado de la lista a una nueva.
  FCNS<-desc(X)
  flattenlist <- function(x){
    morelists <- sapply(x, function(xprime) class(xprime)
      [1]=="list")
    out <- c(x[!morelists], unlist(x[morelists], recursive=
      FALSE))
    if(sum(morelists)){
      base::Recall(out)
    }else{
      return(out)
    }
  }
  superficie<-flattenlist(FCNS)
  #Se separarán los elementos de la lista única generada con
    base en sus características.
  #Primero se separan las funciones.
  FUN<-superficie[purrr::map_lgl(superficie,~"function" %in%
```

```
    class(.x))]
#Después se separan las funciones del tipo gl, es decir,
    funciones generadoras de números aleatorios.
Global_FUN<-superficie[purrr::map_lgl(superficie,~"gl_fun"
    %in%class(.x))]
#Después las de tipo texto, que contienen el nombre de las
    distribuciones utilizadas.
Dist<-unlist(superficie[purrr::map_lgl(superficie,is.
    character)])
#Después las de tipo gráfico en caso de haberlas.
PLTS<-superficie[purrr::map_lgl(superficie,ggplot2::is.
    ggplot)]
#Las de tipo tabla wnr, es decir, las tablas que contienen
    las distribuciones con los mejores p.valores.
dfss<-superficie[purrr::map_lgl(superficie,~"data.frame.wnr
    " %in%class(.x))]
#Finalmente las funciones tipo tabla, es decir, todos los
    posibles resultados generados que pasaron las pruebas AD
     y/o (dependiendo del criterio) KS.
dfss_all<-superficie[purrr::map_lgl(superficie,~is.data.
    frame(.x) & !"data.frame.wnr" %in%class(.x))]
#Se unen todas las tablas con las mejores distribuciones de
     cada partición para generar la distribución total.
PV<-do.call("rbind",dfss[purrr::map_lgl(dfss,~ncol(.x)==12)
    ])
Len<-MA<-c()
repp<-floor(n.obs/length(X))+1
#Se genera la muestra aleatoria con base en las funciones
    generadoras de números aleatorias.
for (OBS in 1:repp) {
```

```
    for (mst in 1:length(FUN)) {
      ljsd<-FUN[[mst]]()
      MA<-c(MA, ljsd)
      if(OBS==1){
        Len<-c(Len, length(ljsd)/length(X))
      }
    }
  }
  MA<-sample(MA, n.obs)
  #Crea una tabla con todas las distribuciones no vacías.
  pv1<-data.frame(Distribution=Dist[nchar(Dist)!=0], Dist_Prop
      =Len[nchar(Dist)!=0])
  p.v<-try(cbind(pv1,PV))
  if(assertthat::is.error(pv1)){p.v<-pv1}
  cp<-plt<-c()
  #Genera una gráfica (si es que aplica) de la diferencia en
      distribución entre la muestra aleatoria y los valores
      reales.
  if(plot){
    DF<-rbind(data.frame(A="Fit",DT=MA),
              data.frame(A="Real",DT=X))
    plt <- ggplot2::ggplot(DF, ggplot2::aes(x=DF$DT, fill=DF$A)
        ) + ggplot2::geom_density(alpha=0.55)+ggplot2::ggtitle
        ("Original Dist.")
    plt
  }
  #En caso de existir las graficas para cada partición las
      une en una sola.
  TPlts<-c()
  if(subplot){
```

```
    cp<-cowplot::plot_grid(plotlist = PLTS, ncol = floor(sqrt
        (length(PLTS))))
  }
  TPlts<-list(plt,cp)
  #Regresa los resultados en una única lista.
  return(list(unlist(FUN),MA,p.v,TPlts,Global_FUN,dfss_all))
}
```

Todos los gráficos que fueron hechos en R se encuentran disponibles en la siguiente dirección:

https://github.com/jcval94/Tesis

Aquellas gráficas que muestran una función de distribución o el ajuste de alguna de ellas se generó por medio de la siguiente aplicación en línea:

https://jcval94.shinyapps.io/FitUltD$_{S}hiny$/

Bibliografía

[1] B. W. Silverman. (1981). *Using Kernel Density Estimates to Investigate Multimodality.* Journal of the Royal Statistical Society, 99.

[2] Nornadiah Mohd Razali, Yap Bee Wah. (2011). *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests.* Journal of Statistical Modeling and Analytics, 33.

[3] Sonja Engmann, Denis Cousineau. (2011). *COMPARING DISTRIBUTIONS: THE TWO-SAMPLE ANDERSON-DARLING TEST AS AN ALTERNATIVE TO THE KOLMOGOROV-SMIRNOFF TEST.* Applied Quantitative Methods, 18.

[4] Sheldon Ross. (2010). *A FIRST COURSE IN PROBABILITY.* United States of America: PEARSON.

[5] Thomas R. Knapp. (2007). *Bimodality Revisited.* Journal of Modern Applied Statistical Methods, 6, 20.

[6] Jose Ameijeiras–Alonso, Rosa M. Crujeiras Alberto Rodríguez Casal. (2019). *Mode testing, critical bandwidth and excess mass.* Mathematical Analysis and Optimization, 54.

[7] B. W. Silverman. (1981). *Using Kernel Density Estimates to Investigate Multimodality.* Journal of the Royal Statistical Society, 99.

[8] Christopher M. Bishop. (2006). *Pattern Recognition and Machine Learning.* USA: Springer.

[9] Nicholas Eugene, Carl Lee Felix Famoye (2002): *BETA-NORMAL DISTRIBUTION AND ITS APPLICATIONS*, Communications in Statistics - Theory and Methods, 31:4, 497-512.