Godinez Bravo Diego
Inferencia Estadística
Tarea 3

CIMAT

1. Sean $\mu \in \mathbb{R}$ y $\beta > 0$. Considera la siguiente función

$$F(t) = \frac{1}{1 + e^{-(t-\mu)/\beta}} \qquad t \in \mathbb{R}$$

Justifica que F es una función de distribución. La función F anterior define a la denominada distribución $Logistica(\mu, \beta)$. Calcula la función de riesgo (hazard) asociada a la distribución $Logistica(\mu, \beta)$.

Solución

Para demostrar que F es una función de distribución, debemos comprobar que se cumplen las siguientes propiedades:

i Sea a y b tal que $a, b \in \mathbb{R}$, con $a \leq b$; entonces se debe cumplir que $F(a) \leq F(b)$.

ii Dado que $F(a)$ es una probabilidad, el valor de la función de distribución se encuentra siempre entre 0 y 1. Por lo tanto:

$$\lim_{a \to +\infty} F(a) = \lim_{a \to +\infty} P(X \leq a) = 1$$
$$\lim_{a \to -\infty} F(a) = \lim_{a \to -\infty} P(X \leq a) = 0$$

iii F es continua por la derecha, es decir:

$$\lim_{\epsilon \downarrow 0} F(a + \epsilon) = F(a)$$

De manera que cualquier función que satisfaga estas propiedades es la función de distribución de alguna variable aleatoria.

i Sean t_1 y $t_2 \in \mathbb{R}$ y $t_1 \leq t_2$; entonces:

$$F(t_1) = \frac{1}{1 + e^{-(t_1-\mu)/\beta}}$$

$$F(t_2) = \frac{1}{1 + e^{-(t_2-\mu)/\beta}}$$

Dado que $t_1 \leq t_2$, entonces se debe cumplir que $F(t_1) \leq F(t_2)$:

$$\frac{1}{1 + e^{-(t_1-\mu)/\beta}} \leq \frac{1}{1 + e^{-(t_2-\mu)/\beta}}$$

De ser así entonces:

$$e^{-(t_1-\mu)/\beta} > e^{-(t_2-\mu)/\beta}$$

Por lo que:

$$-(t_1 - \mu)/\beta > -(t_2 - \mu)/\beta$$

Multiplicando ambos lados de la desigualdad por -1 tenemos que:

$$(t_1 - \mu)/\beta < (t_2 - \mu)/\beta$$

Multiplicando por β y sumando μ en ambos lados;

$$t_1 < t_2$$

De esta manera comprobamos que para todo t_1, t_2 tal que $t_1 < t_2$, se cumple $F(t_1) < F(t_2)$.

ii Al calcular el límite de la función $F(x)$ cuando $t \to +\infty$,

$$\lim_{t\to+\infty} F(t) = \lim_{t\to+\infty} \frac{1}{1+e^{-(t-\mu)/\beta}}$$

$$\frac{1}{1+\underbrace{e^{-\infty}}_{0}} = \frac{1}{1} = 1$$

La función de distribución tiende a uno cuando $t \to +\infty$.

$$\lim_{t\to+\infty} F(t) = 1$$

Por otro lado,

$$\lim_{t\to-\infty} F(t) = \lim_{t\to-\infty} \frac{1}{1+e^{-(t-\mu)/\beta}}$$

$$\frac{1}{\underbrace{1+e^{\infty}}_{\infty}} = \frac{1}{\infty} \approx 0$$

Por lo que la función de distribución tiende a cero cuando $t \to -\infty$.

$$\lim_{t\to-\infty} F(t) = 0$$

iii Sea $\epsilon > 0$, entonces:

$$\lim_{\epsilon\to 0} F(t+\epsilon) = \lim_{\epsilon\to 0} \frac{1}{1+e^{-((t+\epsilon)-\mu)/\beta}}$$

Cuando $\epsilon \to 0$,

$$\lim_{\epsilon\to 0} \frac{1}{1+e^{-((t+\epsilon)-\mu)/\beta}} = \frac{1}{1+e^{-(t-\mu)/\beta}} = F(t)$$

Por lo tanto:

$$\lim_{\epsilon\to 0} F(t+\epsilon) = F(t)$$

De esta manera comprobamos que la función $F(t)$ **es una función de distribución**.

La función de riesgo o tasa de falla, se encuentra definida de la siguiente manera:

$$h(t) = \frac{f(t)}{R(t)}$$

y dado que $R(t)$ es la probabilidad de que un 'arterfacto' falle despúes del tiempo t, entonces:

$$R(t) = P(T \geq t) = P(T > t)$$

la cual puede expresarse de la siguiente manera:

$$R(t) = 1 - F(t)$$

De manera que la función de riesgo se puede definir de la siguiente manera:

$$h(t) = \frac{f(t)}{1-F(t)}$$

Obtenemos $f(t)$ al derivar la función $F(t)$.

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}\frac{1}{1+e^{-(t-\mu)/\beta}}$$

Sea $u = \frac{1}{1+e^{-(t-\mu)/\beta}}$, aplicamos la regla de la cadena:

$$f(t) = \frac{d}{dt} = \frac{d}{du}\frac{du}{dt}$$

Calculamos $\frac{d}{du}\frac{1}{u}$:

$$\frac{d}{du}\frac{1}{u} = \frac{d}{du}u^{-1} = -u^{-2}$$

Por lo tanto:

$$\frac{d}{du} = \frac{1}{-u^2}$$

Por otro lado derivamos $\frac{du}{dt}$ tomando a $v = -(t-\mu)/\beta$, y aplicando la regla de la cadena:

$$\frac{du}{dt}1 + e^v = \frac{du}{dv}\frac{dv}{dt}$$

En este caso tenemos que:

$$\frac{du}{dv}1 + e^v = e^v$$

y:

$$\frac{dv}{dt} - (t-\mu)/\beta = -\frac{1}{\beta}(t-\mu) = \frac{-t}{\beta} + \frac{\mu}{\beta} = -\frac{1}{\beta}$$

De esta manera tenemos que:

$$\frac{du}{dt} = e^v - \frac{1}{\beta} = e^{-(t-\mu)/\beta} - \frac{1}{\beta}$$

Finalmente:

$$\frac{d}{dt} = \underbrace{\frac{1}{-u^2}}_{\frac{d}{du}} \cdot \underbrace{e^{-(t-\mu)/\beta} - \frac{1}{\beta}}_{\frac{du}{dt}}$$

Reemplazando u y realizando la multiplicación:

$$\frac{d}{dt} = \frac{e^{-(t-\mu)/\beta} - \frac{1}{\beta}}{-(1+e^{-(t-\mu)/\beta})^2} = \frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2}$$

Entonces:

$$f(t) = \frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2}$$

De esta manera podríamos obtener la función de riesgo:

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{\frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2}}{1 - \frac{1}{1+e^{-(t-\mu)/\beta}}}$$

Simplificando el denominador:

$$h(t) = \frac{\frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2}}{\frac{1+e^{-(t-\mu)/\beta}-1}{1+e^{-(t-\mu)/\beta}}} = \frac{\frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2}}{\frac{e^{-(t-\mu)/\beta}}{1+e^{-(t-\mu)/\beta}}}$$

$$= \frac{\frac{e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta)})^2}}{\frac{e^{-(t-\mu)/\beta}}{1+e^{-(t-\mu)/\beta}}} = \frac{1+e^{-(t-\mu)/\beta} \cdot e^{-(t-\mu)/\beta}}{\beta(1+e^{-(t-\mu)/\beta})^2 \cdot e^{-(t-\mu)/\beta}}$$

$$= \frac{(1+e^{-(t-\mu)/\beta})}{\beta(1+e^{-(t-\mu)/\beta})^2} = \frac{1}{\beta(1+e^{-(t-\mu)/\beta})}$$

De esta manera encontramos que la función de riesgo se encuentra definida como:

$$h(t) = \frac{1}{\beta(1+e^{-(t-\mu)/\beta})} = \frac{1}{\beta}F(t)$$

2. Sea $\sim X\ Normal(0,1)$. Calcula los momentos pares e impares de X; es decir, calcula $E(X^p)$, para $p = 2k$ y $p = 2k-1$, con $k \in \mathbb{N}$. **Nota**: Conviene considerar la diferencia entre par e impar para facilitar la cuenta. Usa la función generadora de momentos e investiga el concepto de doble factorial.

Solución

El valor esperado se encuentra definido de la siguiente manera:

$$E(X) = \int_{-\infty}^{\infty} x\ fx(x)\ dx \quad v.a.\ continua$$

Partimos de dicha definición para calcular los momentos pares ($p = 2k$) e impares ($2k-1$):

$$E(X) = \int_{-\infty}^{\infty} x^p\ fx(x)\ dx = \int_{-\infty}^{\infty} x^p\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= \int_{-\infty}^{0} x^p\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \int_{0}^{\infty} x^p\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Calculamos los momentos impares; $p = 2k-1$.

$$E[X^{2k-1}] = \int_{-\infty}^{0} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \int_{0}^{\infty} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

cuando $p = 2k-1$, x toma un valor negativo, y por propiedades de funciones funciones pares e impares entonces:

$$\int_{-\infty}^{0} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = -\int_{0}^{\infty} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Por lo tanto:

$$E[X^{2k-1}] = -\int_{0}^{\infty} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \int_{0}^{\infty} x^{2k-1}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

De esta manera los **momentos impares** estan definidos como:

$$E[X^{2k-1}] = 0$$

Por otro lado, cuando $p = 2k$:

$$E[X^{2k}] = \int_{-\infty}^{0} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

cuando $p = 2k$, x toma un valor positivo, y por propiedades de funciones funciones pares e impares entonces:

$$\int_{-\infty}^{0} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Por lo tanto:

$$E[X^{2k}] = \int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx + \int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= 2\int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

De esta manera los **momentos pares** estan definidos como:

$$E[X^{2k-1}] = 2\int_{0}^{\infty} x^{2k}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Desarrollando la ecuación podemos expresarla de la siguiente manera:

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k} e^{-\frac{1}{2}x^2} dx$$

Sea $t = \frac{1}{2}x^2$, $x = \sqrt{2t}$ entonces $\frac{dt}{dx} = x$, $dx = \frac{dt}{x}$

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k} e^{-\frac{1}{2}x^2} \frac{dt}{x} = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k} e^{-\frac{1}{2}x^2} dt x^{-1}$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k-1} e^{-\frac{1}{2}x^2} dt$$

Reemplazando y manipulando algebraicamente tenemos:

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty \sqrt{2t}^{2k-1} e^{-t} dt$$

$$= \frac{2}{\sqrt{2}\sqrt{\pi}} \int_0^\infty ((2t)^{\frac{1}{2}})^{2k-1} e^{-t} dt = \frac{2}{\sqrt{2}\sqrt{\pi}} \int_0^\infty (2t)^{k-\frac{1}{2}} e^{-t} dt$$

$$= \frac{2^1}{\sqrt{2}\sqrt{\pi}} \int_0^\infty (2)^{k-\frac{1}{2}} (t)^{k-\frac{1}{2}} e^{-t} dt = \frac{2^{k+\frac{1}{2}}}{\sqrt{2}\sqrt{\pi}} \int_0^\infty (t)^{k-\frac{1}{2}} e^{-t} dt$$

$$= \frac{2^{k+\frac{1}{2}} \cdot 2^{-\frac{1}{2}}}{\sqrt{\pi}} \int_0^\infty (t)^{k-\frac{1}{2}} e^{-t} dt = \frac{2^k}{\sqrt{\pi}} \int_0^\infty (t)^{k-\frac{1}{2}} e^{-t} dt$$

De esta manera obtenemos:

$$E[X^{2k-1}] = \frac{2^k}{\sqrt{\pi}} \int_0^\infty (t)^{k-\frac{1}{2}} e^{-t} dt$$

Rescatando la definición de la función Gamma (función factorial generalizada), presente en el modelo Gamma, tenemos:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

Sea $\alpha = k - \frac{1}{2}$, entonces:

$$= \frac{2^k}{\sqrt{\pi}} \Gamma(k + \frac{1}{2})$$

Considerando algunas de las propiedades de la función gamma:

** Sea n un entero positivo:

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

** Para valores no negativos de n se tiene:

$$\Gamma(\frac{1}{2} + n) = \frac{(2n)!}{4^n n!} \sqrt{\pi} = \frac{(2n-1)!!}{2^n} \sqrt{\pi}$$

Por lo tanto:

$$\Gamma(\frac{1}{2} + k) = \frac{(2k)!}{4^k k!} \sqrt{\pi} = \frac{(2k-1)!!}{2^k} \sqrt{\pi}$$

Reemplazando tenemos que:

$$= \frac{2^k}{\sqrt{\pi}} \frac{(2k-1)!!}{2^k} \sqrt{\pi}$$

Finalmente:

$$E[X^{2k-1}] = (2k-1)!!$$

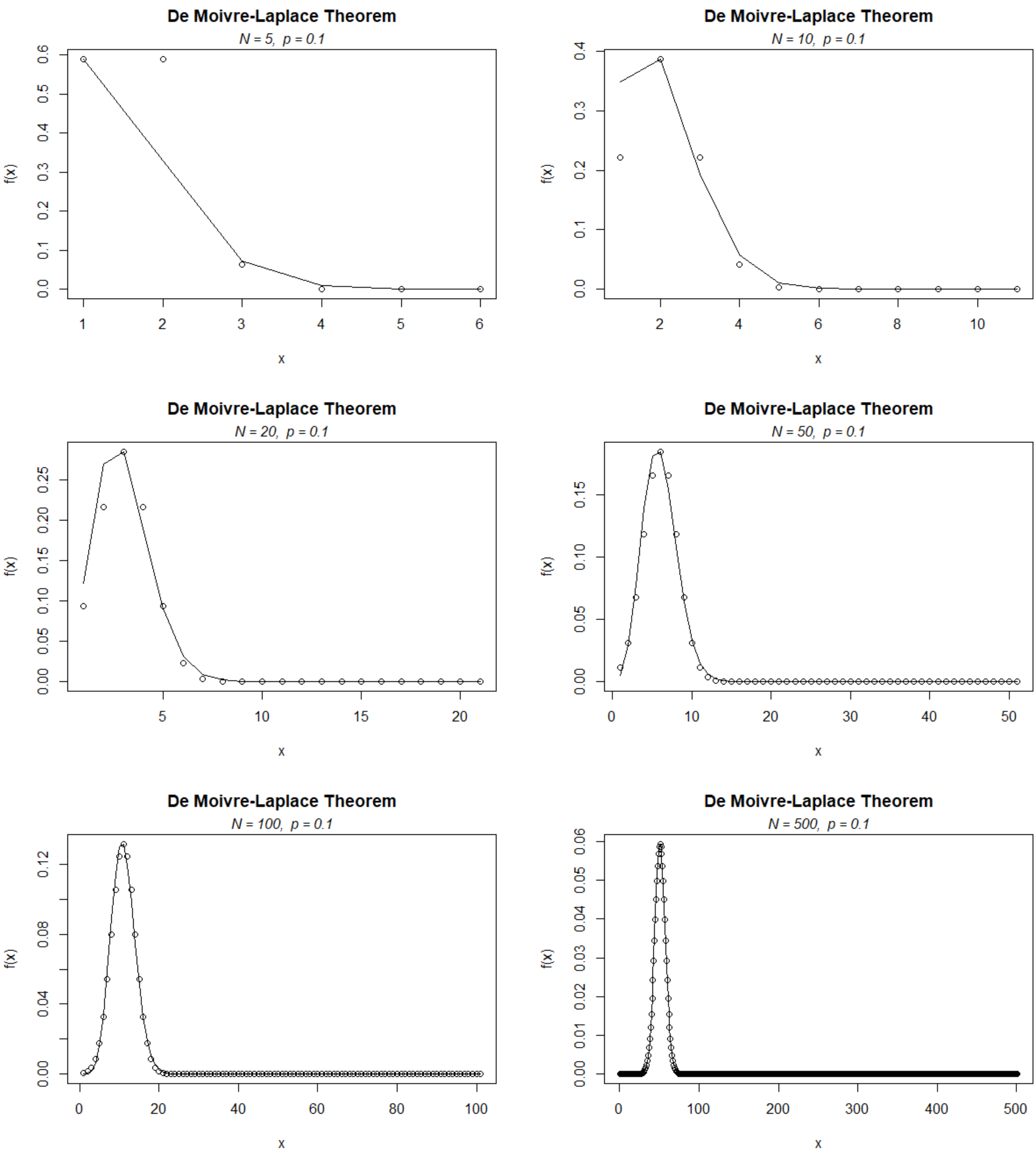
definición del doble factorial

3. En este ejercicio visualizaremos el Teorema de Moivre-Laplace (TML). Para p = 0.1 y A = 5, 10, 20, 50, 100, 500, grafica lo siguiente:

a) Sobre la misma figura, grafica la función de masa g(x) de una distribución $Binomial(n, p)$ y una la función de densidad f(x) de una distribución $Normal(np, npq)$, para todo $n \in A$ (i.e. presenta las 6 figuras).

Resultados

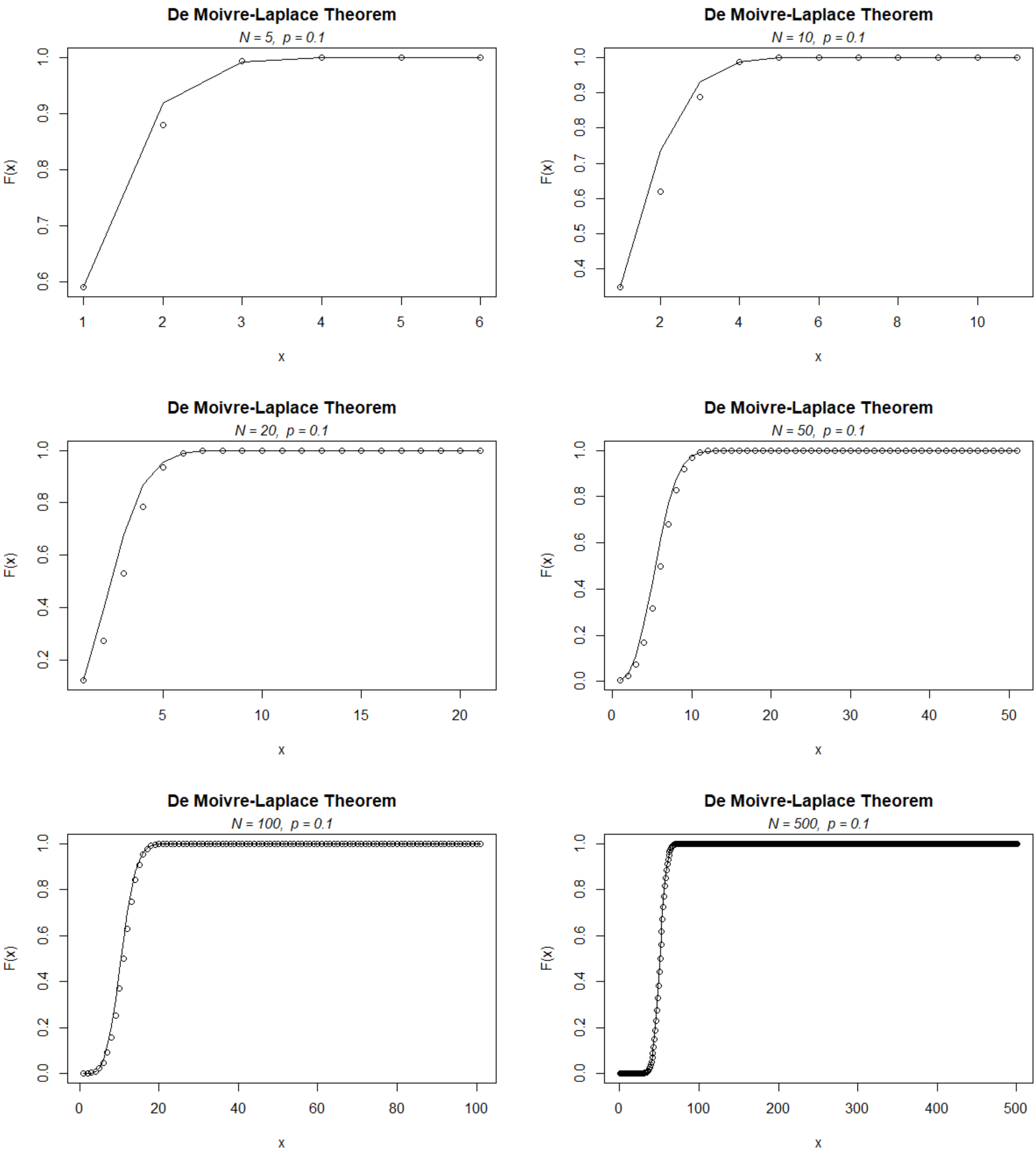
Función de Probabilidad de Masa (PMF)



b) Haz lo mismo que en el inciso anterior pero ahora para las funciones de distribución acumuladas de las binomiales y normales anteriores.

Resultados

Función de Distribución Acumulada (CDF)



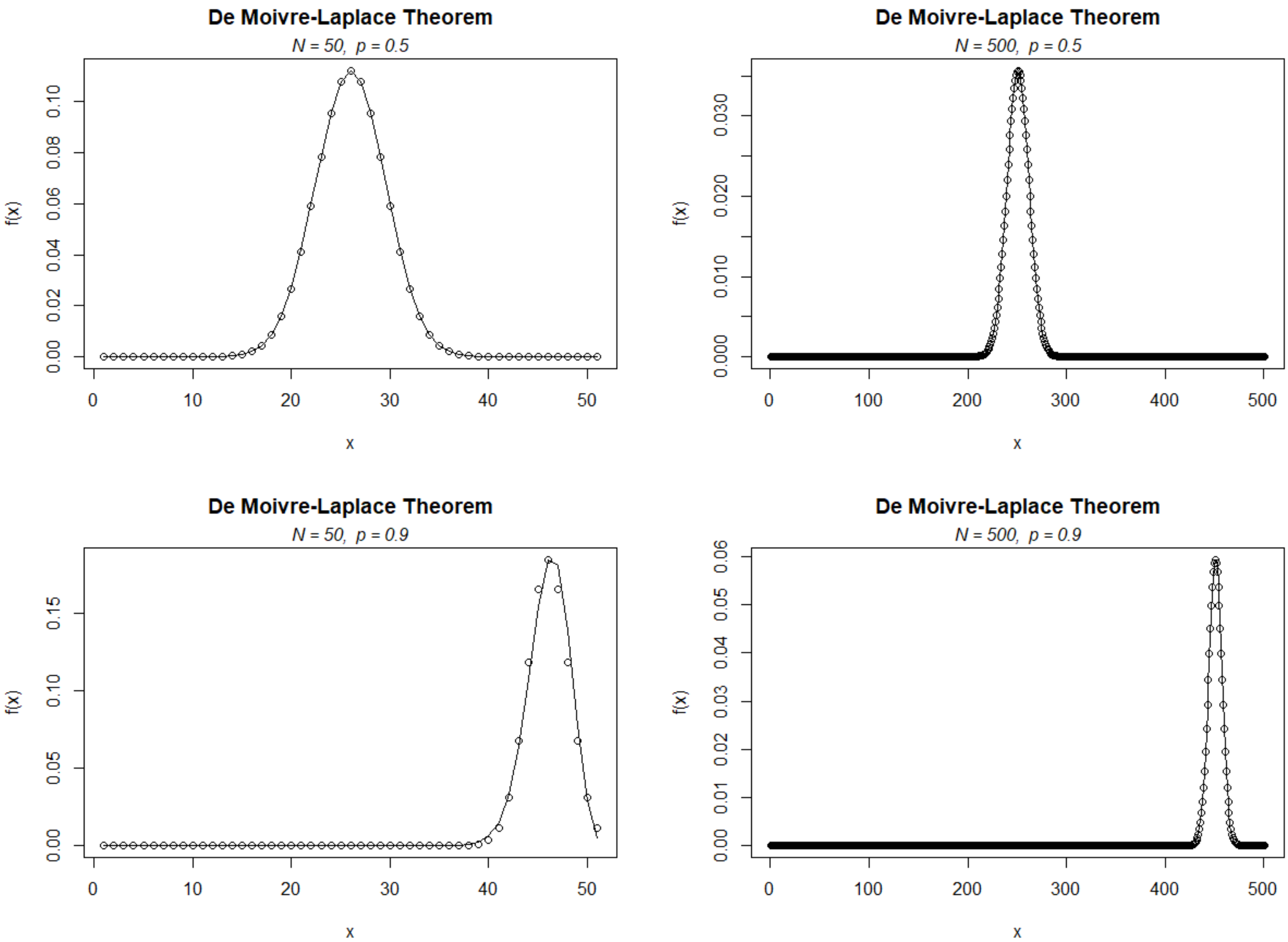
c) ¿Cuál es la relación entre las figuras anteriores y el TML? ¿Cambia el resultado si uno toma p = 0.5, 0.9?

Resultados

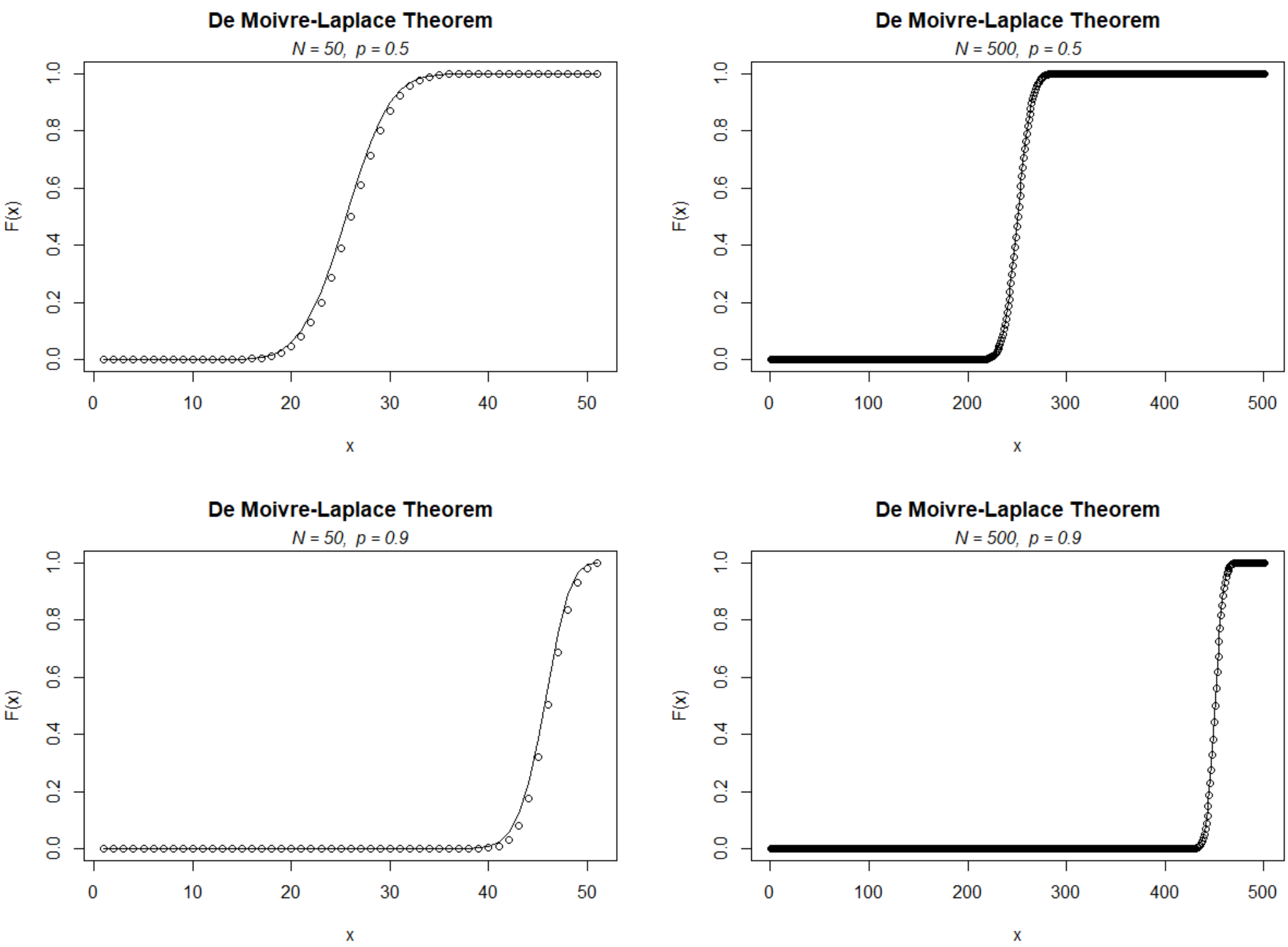
Como establece el Teorema de De Moivre-Laplace, una variable aleatoria X con distribución **binomial** $X \sim Binomial(n, p)$, definida en n pruebas independientes cada una con probabilidad de éxito p con $0 < p < 1$; converge a una distribución normal $X \sim Normal(np, \sqrt{np(1-p)})$, conforme n crece.

En los casos anteriores se realizaron las simulaciones con un valor $p = 0{,}1$. Sin embargo, al realizar las simulaciones modificando el valor p ($p = 0{,}5$; $p = 0{,}9$), observamos que la curva de distribución se sigue asemejando a una distribución normal conforme aumentamos el valor de n. La única diferencia, en ambos casos, es que el valor esperado $E(x) = np$ se ve desplazado.

Función de Probabilidad de Masa (PMF)



Función de Distribución Acumulada (CDF)



4. Una partícula se encuentra inicialmente en el origen de la recta real y se mueve en saltos de una unidad. Para cada salto, la probabilidad de que la partícula salte una unidad a la izquierda es p y la probabilidad de que salte una unidad a la derecha es 1 - p. Denotemos por X_n a la posición de la partícula después de n unidades. Encuentre $E(X_n)$ y $Var(X_n)$. Esto se conoce como una caminata aleatoria en una dimensión.

Solución

Dada la naturaleza del problema podemos considerar que la caminata es aleatoria en el sentido de que la opción de ir hacia adelante o hacia atrás en cada paso es aleatoria y completamente independiente de lo que haya sucedido con anterioridad.

Por lo tanto, definimos la camina aleatoria de la siguiente manera:

$$S_n = \sum_{k=1}^{n} X_k$$

donde S_n es la posición de la partícula despúes de n movimientos, y X_k es un conjunto de variables aleatorias independientes e idénticamente distribuidas.

Dada la función de valor esperado tenemos que:

$$E(X) = \sum_{i=1}^{n} X_i P(X = X_i)$$

Y si consideramos que X_k puede tomar los valores $\{1, -1\}$ con $P(X_k = 1) = 1 - p$ y $P(X_k = -1) = p$, entonces:

$$E(X_k) = \sum_{i=1}^{n} X_i P(X = X_i)$$

$$E(X_k) = 1(1-p) + (-1)p$$

$$= 1 - p - p$$

Entonces el valor esperado de la cada variable aleatoria X_k esta dado de la siguiente manera:

$$E(X_k) = 1 - 2p$$

Por lo tanto:

$$E(S_n) = E\Big(\sum_{k=1}^{n} X_k\Big)$$

$$= \sum_{k=1}^{n} E(X_k)$$

$$= n\ E(X_k)$$

De esta manera encontramos el **valor esperado**:

$$E(S_n) = n\ (1 - 2p)$$

Por otro lado, podemos encontrar la varianza de la siguiente manera:

$$V(S_n) = V\Big(\sum_{k=1}^{n} X_k\Big)$$

Al ser una suma de variables aleatorias **independientes**, entonces la varianza es la suma de las varianzas de cada variable aleatoria individual;

$$V(S_n) = \sum_{k=1}^{n} V(X_k)$$

Calculamos la varianza de la variable X_k:

$$V(X_k) = E[X_k^2] - E[X_k]^2$$

Por un lado tenemos que $E[X_k]^2$:

$$E[X_k]^2 = (1-2p)^2 = 1 - 4p + 4p^2$$

Mientras que $E[X_k^2]$:

$$E[X_k^2] = (1)^2(1-p) + (-1)^2 p = (1)(1-p) + (1)p = 1 - p + p = 1$$

Sustituyendo tenemos que:

$$V(X_k) = 1 - 1 + 4p - 4p^2) = 4p - 4p^2$$

Finalmente tenemos que la **varianza** esta dada de la siguiente manera:

$$V(S_n) = \sum_{k=1}^{n} V(X_k) = n(4p - 4p^2)$$

$$= 4np - 4np^2$$

$$= 4np(1-p)$$

Sea $q = 1 - p$, entonces:

$$V(S_n) = 4npq$$

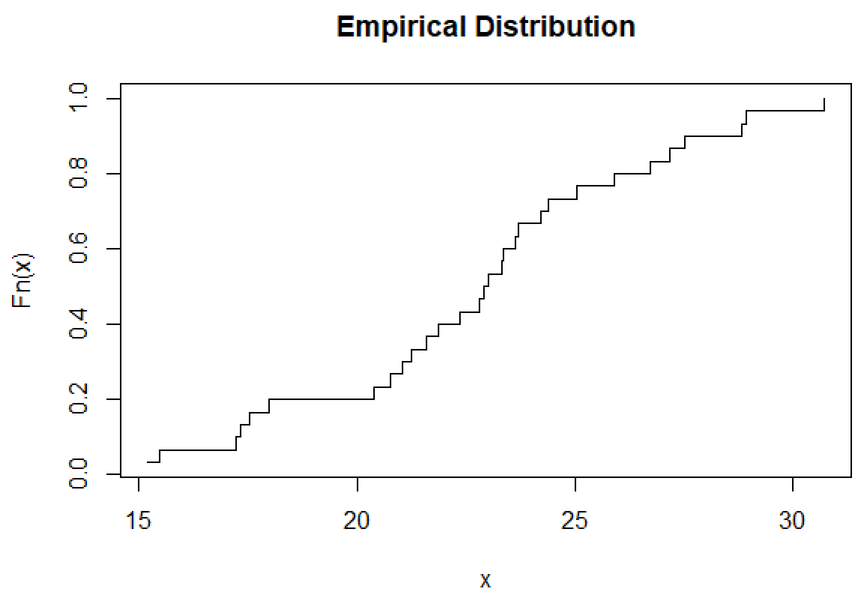
5. El siguiente conjuntos de datos contiene mediciones del diámetro de un agave, medido en decímetros, en distintas localizaciones no cercanas.

a) Escriba una función en R que calcule la función de distribución empírica para un conjunto de datos dado D. La función debe tomar como parámetros al valor x donde se evalúa y al conjunto de datos D. Utilizando esta función grafique la función de distribución empírica asociada al conjunto de datos de agave. Ponga atención a los puntos de discontinuidad. ¿Qué observa? **Nota**: Escriba la función mediante el algoritmo descrito en las notas de la clase; para este ejercicio no vale usar la funciones implementadas en R que hacen lo pedido.

La gráfica de la función de distribución empírica muestra los valores únicos presentes en el conjunto de datos, cada uno de ellos representado por un escalón, el cual además, indica la probabilidad de que una observación sea menor o igual que el valor en el cual se encuentra dicho escalón.

Podemos observar que los puntos de discontinuidad predominan entre el rango de valores $20 < x < 25$. Esto indica que las observaciones del conjunto de datos dado se concentran, en su mayoría, en este rango.

De manera general, la gráfica de la función de distribución empírica muestra como se distribuyen los valores en el conjunto de datos y cómo cambia la probabilidad acumulada a medida que recorres los valores contenidos en el conjunto de datos.



Función de Distribución Empírica asociada al conjunto de datos

b) Escriba una función en R que determine la gráfica Q-Q normal de un conjunto de datos. La función debe tomar como parámetro al conjunto de datos y deberá graficar contra el percentil estandarizado de la normal. Para poder comparar el ajuste más claramente, la función además deberá ajustar en rojo a la recta $sx+\bar{x}$ (s =desviación estándar muestral y $\bar{x}$ =media muestral). Usando esta función, determine la gráfica Q-Q normal. ¿Qué observa? **Nota**: La misma del inciso a).

Podemos notar que los puntos en el gráfico se encuentran, en su mayoría, alineados con la recta $sx+\bar{x}$, lo cual podría sugerir que los datos tienen una buena aproximación a la normalidad. Dado que los datos se desvían de la recta en el extremo izquierdo de la recta, podríamos considerar la idea de que la curva de distribución se encuentre sesgada hacía la izquierda.

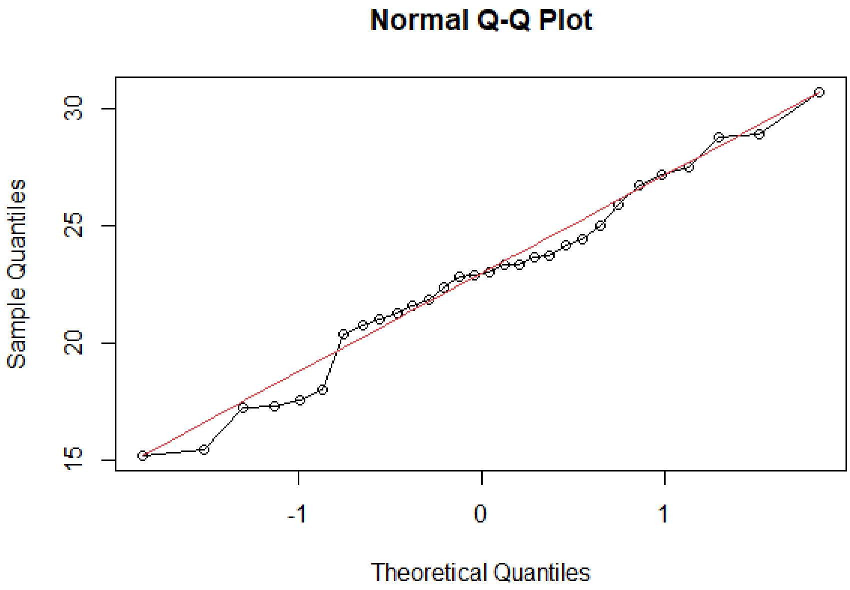
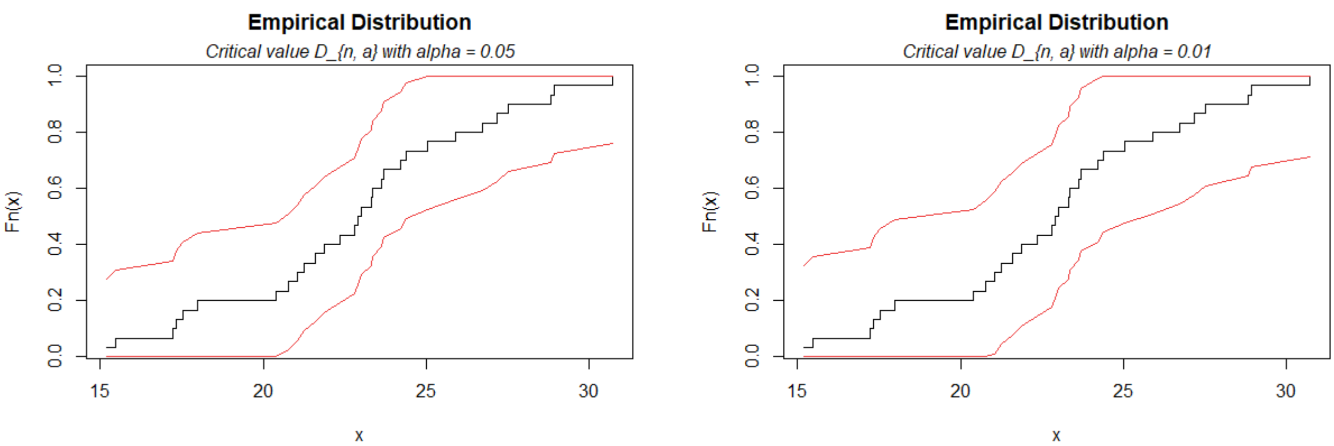


Gráfico Q-Q Normal

c) Añada a la función anteriores (función de distribución empírica y Q-Q normal) la opción de que grafiquen la banda de confianza, de cobertura $1-\alpha$, basada en el estadístico de Kolmogorov-Smirnov. La función debe tomar como parámetros al conjunto de datos y el nivel de confianza $1-\alpha$. Aplique esta función al conjunto de datos para un nivel de confianza $1-\alpha = 0{,}95, 0{,}99$. ¿Qué observa? **Nota**: Recurra a las notas sobre las bandas de confianza de los gráficos Q-Q normales que se incluyeron en la clase 10; no vale usar la funciones implementadas en R que hacen lo pedido. No es necesario entender a detalle la prueba de Kolmogoro-Smirnov, en este punto solo consideraremos su aspecto operacional; al final del curso, en una de las exposiciones finales, se presentará la prueba con detalle.

Las bandas de confianza se representan en ambos gráficos como líneas las cuales rodean los valores presentes en nuestro conjunto de datos. Dado que en cada uno de las gráficos las observaciones se encuentran dentro de las bandas de confianza, no hay evidencia significativa para sugerir que los datos no siguen una distribución normal, esto con un nivel de confianza $1-\alpha$ para $\alpha = 0{,}05$ y $\alpha = 0{,}01$.



Función de Distribución Empírica considerando intervalos de confianza basados en la prueba Kolmogorov Smirnov

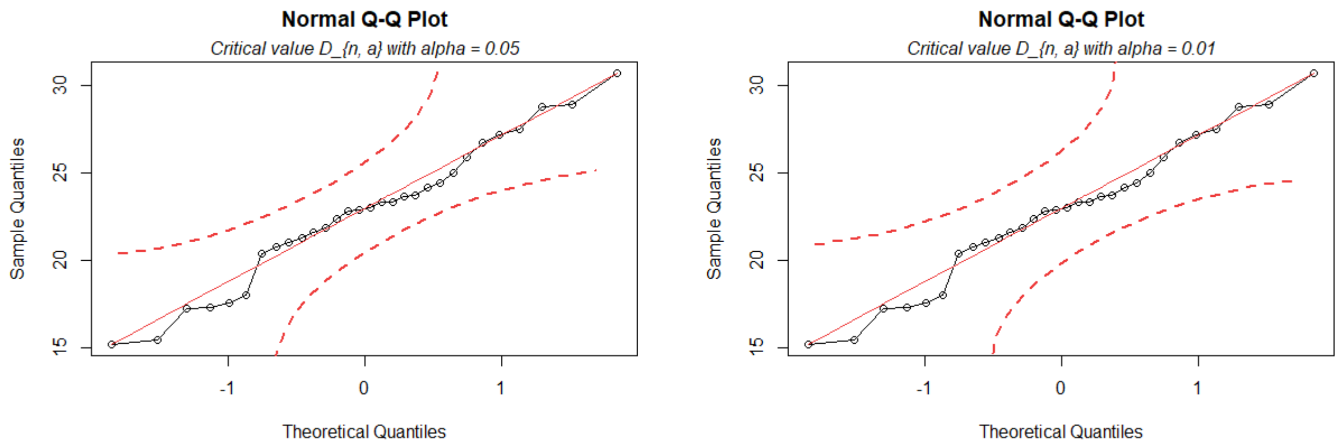


Gráfico Q-Q Normal considerando intervalos de confianza basados en la prueba Kolmogorov Smirnov

d) Escriba una función en R que determine el gráfico de probabilidad normal. La función debe tomar como parámetro al conjunto de datos. ¿Qué observa? **Nota**: La misma del inciso a).

El gráfico de normalidad es una herramienta útil para evaluar la normalidad de un conjunto de datos. Ya que no se observan desviaciones notables de los puntos generados con referencia a la línea de referencia, podemos afirmar con cierto grado de certeza que los datos se aproximan a una distribución normal.

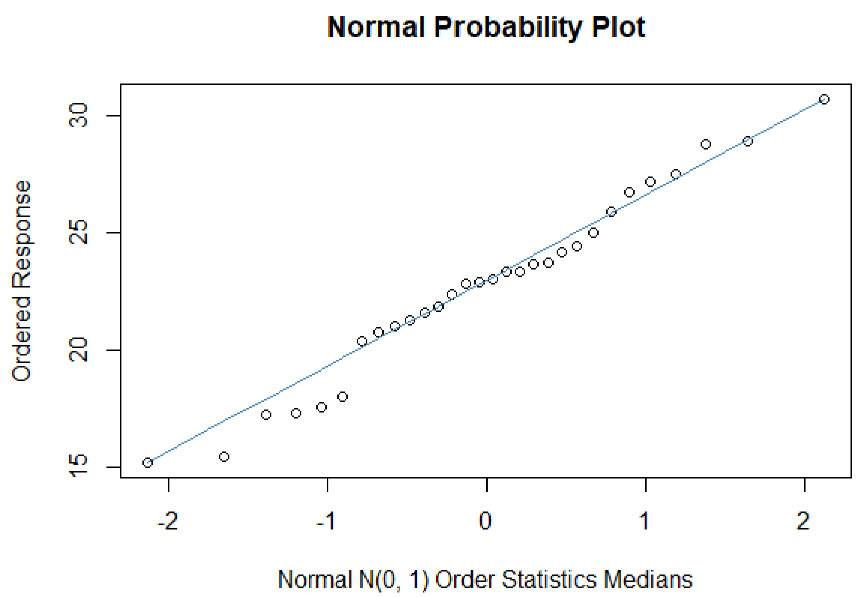


Gráfico Q-Q Normal

e) ¿Los datos anteriores se distribuyen normalmente? Argumente.

De manera general, podemos afirmar con cierto grado de certeza que los datos siguen una distribución normal. Esto basándonos en los gráficos que muestran los datos contenidos alrededor de las bandas de confianza generadas, basadas en la prueba Kolmogorov Smirnov con un nivel de confianza $1-\alpha$ para $\alpha = 0{,}05$ y $\alpha = 0{,}01$.

6. En este ejercicio se comprobará que tan buena es la aproximación dada por las reglas empíricas para algunas de las distribuciones estudiadas en la clase. Considerese las distribuciones $Unif(a = -3, b = 3)$, $Normal(0,1)$, $Exponencial(2)$, $Gamma(\alpha = 2, \beta = 1)$, $Gamma(\alpha = 3, \beta = 1)$, $Beta(\alpha = 2, \beta = 2)$, $Weibull(\alpha = 4, \beta = 1)$ y $Lognormal(\mu = 3, \sigma = 2)$.

a) Leer las reglas empirícas en https://en.wikipedia.org/wiki/68

b) Para cada una de las distribuciones anteriores, haga una tabla que muestre las probabilidades contenidas en los intervalos $(\mu - k\sigma, \mu + k\sigma)$, para k = 1, 2, 3. Utilice las fórmulas de las medias y varianzas contenidas en las notas para determinar μ y σ en cada caso. Puede usar R para determinar las probabilidades pedidas.

Resultados

Distribución	$k = 1$	$k = 2$	$k = 3$
Uniform $Unif(a = -3, b = 3)$	0.5773	1.0000	1.0000
Normal $N(0,1)$	0.6826	0.9544	0.9973
Exponential $Exp(2)$	0.8646	0.9502	0.9816
Gamma $Gamma(\alpha = 2, \beta = 1)$	0.7375	0.9533	0.9859
Gamma $Gamma(\alpha = 3, \beta = 1)$	0.7153	0.9558	0.9882
Beta $Beta(\alpha = 2, \beta = 2)$	0.6260	0.9838	1.0000
Weibull $Weibull(\alpha = 4, \beta = 1)$	0.6218	0.9591	0.9933
Lognormal $Lognormal(\mu = 3, \sigma = 2)$	0.9802	0.9912	0.9948

Tabla 1. Probabilidades teóricas contenidas dentro del intervalo $(\mu - k\sigma,\ \mu + k\sigma)$; con $k = 1, 2, 3$.u

c) En R, simule n = 1000 muestras de cada una de las distribuciones anteriores y calcule la media muestral $\bar{x}$ y la varianza muestral s^2 como se mencionó en la clase. En cada caso, calcule la proporción de observaciones que quedan en los intervalos $(\bar{x}-ks, \bar{x}+ks)$, para k = 1, 2, 3. Reporte sus hallazgos en una tabla como la del inciso anterior. ¿Qué tanto se parecen la tabla de este inciso y la del anterior?

Resultados

Distribución	$k=1$	$k=2$	$k=3$
Uniform $Unif(a=-3, b=3)$	0.5694	1.0000	1.0000
Normal $N(0,1)$	0.6826	0.9544	0.9973
Exponential $Exp(2)$	0.8485	0.9499	0.9814
Gamma $Gamma(\alpha=2, \beta=1)$	0.7239	0.9500	0.9843
Gamma $Gamma(\alpha=3, \beta=1)$	0.6839	0.9521	0.9863
Beta $Beta(\alpha=2, \beta=2)$	0.6241	0.9827	1.0000
Weibull $Weibull(\alpha=4, \beta=1)$	0.6892	0.9651	0.9995
Lognormal $Lognormal(\mu=3, \sigma=2)$	0.9557	0.9766	0.9849

Tabla 2. Probabilidades observadas contenidas dentro del intervalo $(\mu-k\sigma,\ \mu+k\sigma)$; con $k=1,2,3$. Simulación realizada con $n=1000$.

Comparando los resultados obtenidos en ambas tablas, podemos observar que los resultados de la simulación se asemejan a los resultados obtenidos en la **Tabla 1**. En el caso de la distribución Normal $N(0,1)$, el porcentaje de valores que se encuetran dentro de cada uno de los intervalos considerados coincide con lo establecido por la regla empírica para un conjunto de datos aproximadamente normal ($\sim$68% en el intervalo $\mu-\sigma,\ \mu+\sigma$; $\sim$95% en el intervalo $\mu-2\sigma,\ \mu+2\sigma$; y $\sim$99.7% en el intervalo $\mu-3\sigma,\ \mu+3\sigma$). De manera similar, en las distribuciones restantes las probabilidades contenidas en cada uno de los intervalos coinciden con las probabilidades teóricas obtenidas para los valores $k=1,2,3$.

Por lo tanto, podemos decir que las reglas empíricas son una buena aproximación para describir el comportamiento de las distribuciones en cuestión.