Inferencia Estadística | Tarea 03

Aguirre Calzadilla César Miguel

04 de octubre de 2024

Códigos

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

Problema 1

Sean $\mu \in \mathcal{R}$ y $\beta > 0$, considera la función $F(t)$. 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.

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

Comenzaremos justificando por qué la Distribución Logística es, valga la redundancia, una función de distribución. Para ello, recordemos que se deben cumplir los siguientes puntos:

- $F(t)$ es monótona no decreciente.
- $\lim_{x\to-\infty} F(t) = 0$ y $\lim_{x\to+\infty} F(t) = 1$
- $F(t)$ es continua.

Entonces, para el primer punto, tenemos que $F(t) = \frac{1}{1+e^{-\alpha/\beta}}$ con $\alpha = (t - \mu)$. Cuando α crece, entonces $F(t)$ también crece, pues el exponencial es negativo; esto significa que $e^{-\alpha/\beta}$ disminuye cada que $\alpha \uparrow$. Por lo tanto, la función es monótona no decreciente.

Ahora, para el segundo punto:

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

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

Y para el tercer punto, la logistica es, en efecto, continua en todo su dominio. Tiene una traza similar a la de la función sigmoide. Gráficamente, se ve como muestra la figura (1).

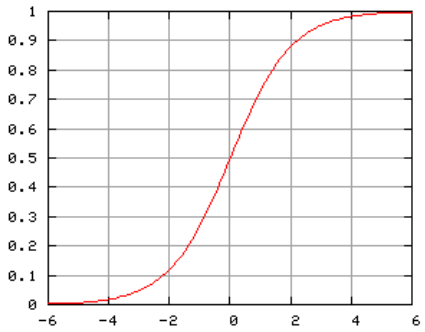


Figura 1: Gráfica de la función logística normalizada | Tomada de Wikimedia Commons.

Por lo tanto, es una función contínua en todo su dominio.

Ahora bien, calculemos la hazard de la Logística. Recordemos que la función de riesgo se define como:

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

Con $f(t)$ definida como la función de densidad de la logística $F(t)$. Para dar con $f(t)$, tenemos que derivar $F(t)$ tal que nos queda algo de la siguiente forma:

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

Procedemos por regla de la cadena y nos queda algo de la forma:

$$\frac{d}{dt}F(t) = -\frac{\frac{d}{dt}u(t)}{u^2(t)}$$

Con $u(t) = 1 + e^{-(t-\mu)/\beta}$.

Entonces:

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

Por lo tanto, la función hazard de la Distribución Logística quedaría de la siguiente manera:

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

Problema 2

Sea $X \sim Normal(0, 1)$, calcula los momentos pares e impares de X; es decir, calcula $E(X^p)$ para $p = 2k$ y $p = 2k - 1$, con $k \in \mathcal{N}$. 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.

Comencemos recordando que la función generadora de momentos de la Normal se ve la siguiente manera:

$$M_x(t) = E(e^{tx}) = e^{\frac{t^2}{2}} \quad \text{para} \quad X \sim Normal(0, 1)$$

Los momentos de X se obtienen derivando $M_x(t) = e^{\frac{t^2}{2}}$ respecto de t y evaluando en $t = 0$. Sabemos que el n-ésimo momento se verá como:

$$E(X^n) = \frac{d^n}{dt^n} M_x(t)|_{t=0}$$

Tenemos que para $E(X^{2k-1}) = \frac{d^{k-1}}{dt^{k-1}} M_x(t)|_t = 0$. Podemos ver la tendencia:

$$\frac{d}{dt} M_x(t) = e^{(t^2/2)} \cdot t$$

$$\frac{d^3}{dt^3} M_x(t) = (t^2 + 3)e^{(t^2/2)} \cdot t$$

$$\frac{d^5}{dt^5} M_x(t) = (t^4 + 10t^2 + 15)e^{(t^2/2)} \cdot t$$

$$\frac{d^7}{dt^7} M_x(t) = (t^6 + 21t^4 + 105t^2 + 105)e^{(t^2/2)} \cdot t$$

Podemos generalizar estos resultados con los polinomios de Hermite de orden $2n - 1$, tal que:

$$\frac{d^{2k-1}}{dt^{2k-1}} M_x(t) = H_{2k-1} e^{(t^2/2)} \cdot t$$

Por lo que, si siempre se cumple que $t = 0$, entonces $\frac{d^{2k-1}}{dt^{2k-1}} M_x(t) = 0$ para cualquier impar.

Ahora, para el caso de los pares.
Tenemos que:

$$\frac{d^2}{dt^2} M_x(t=0) = e^{(t^2/2)} \cdot (1+t)^2 = 1$$

$$\frac{d^4}{dt^4} M_x(t=0) = e^{(t^2/2)} \cdot (t^4 + 6t^2 + 3) = 3$$

$$\frac{d^6}{dt^6} M_x(t=0) = e^{(t^2/2)} \cdot (t^6 + 15t^4 + 45t^2 + 15) = 15$$

$$\frac{d^8}{dt^8} M_x(t=0) = e^{(t^2/2)} \cdot (t^8 + 28t^6 + 420t^2 + 105) = 105$$

Esta tendencia nos puede recordar, gracias al hint, que la derivada par de la generatriz de momntos de esta función será de la forma:

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

Es decir, esto es el doble factorial de todos los imapres hasta $2k-1$.

Problema 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:

Primero, es buena idea explicar un poco acerca del Teorema de De Moivre-Laplace. Este teorema, el cual es un caso especial del Teorema del Límite Central, establece una relación directa entre la Distribución Normal y la Distribución Binomial. Eso sí, bajo ciertas condiciones.

El teorema muestra que la función de masa de probabilidad $g(x)$ de una cantidad aleatoria de éxitos observados tras n ensayos de Bernoulli independientes, con probabildiad p de éxito, converge a la función de densisdad de probabildiad de la Normal con esperanza np y varianza $\sqrt{np(1-p)}$, mientras n siga creciendo.

Dicho teorema está establecido de la siguiente manera:

$$\binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} \qquad \text{con} \quad p+q=1, \quad p,q>0$$

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

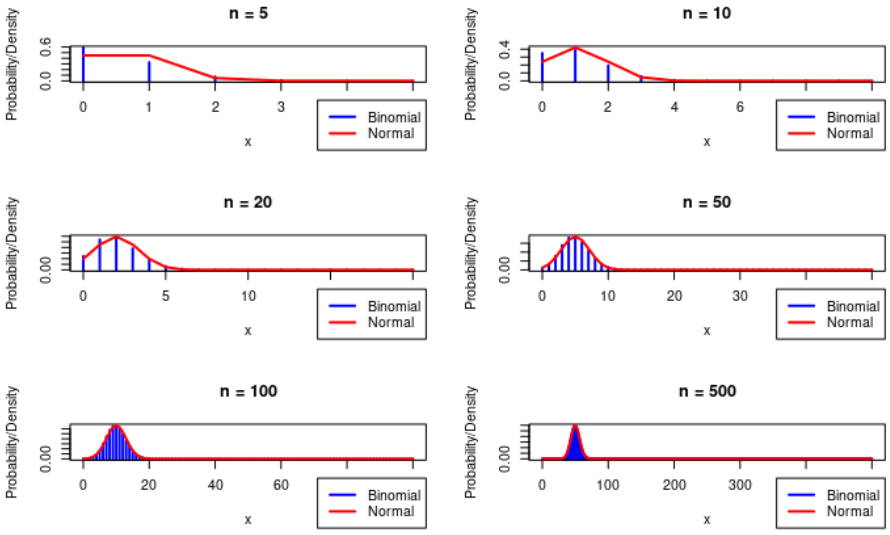


Figura 2: Función de masa | p=0.1

Podemos notar como a medida que n crece, la Binomial se aproxima cada vez mejor a la Normal. Esto ya está establecido de entrada por el TML. A medida que n crece, la función de masa binomial se aproximará cada vez más a la curva normal, lo que visualiza el Teorema de Moivre-Laplace.

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

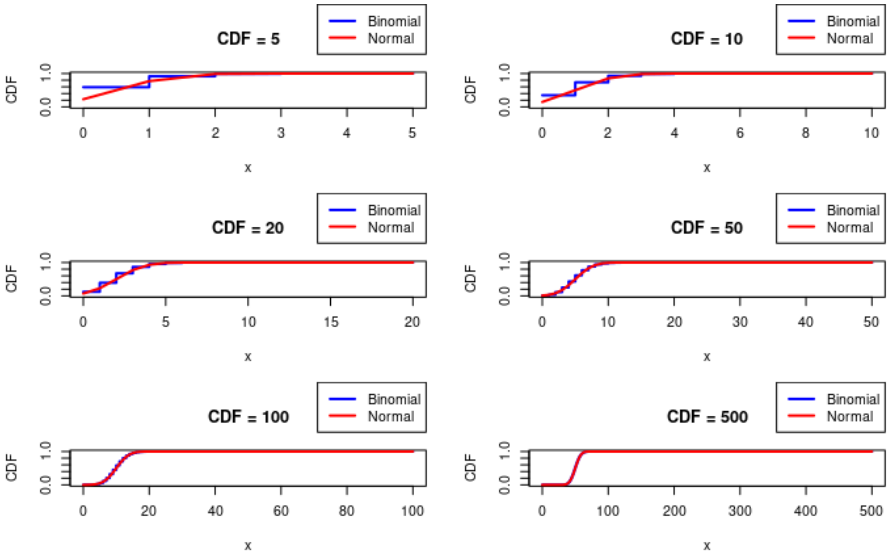


Figura 3: CDF | p=0.1

c) Cuál es la relación entre as figuras anteruires y el TML? Cambia el resultado si uno toma p=0.5 o p=0.9?

Para p = 0.5

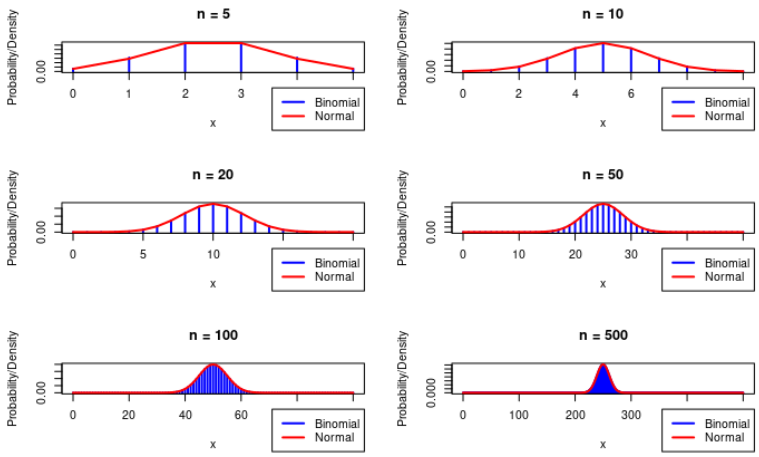


Figura 4: Función de masa | p=0.5

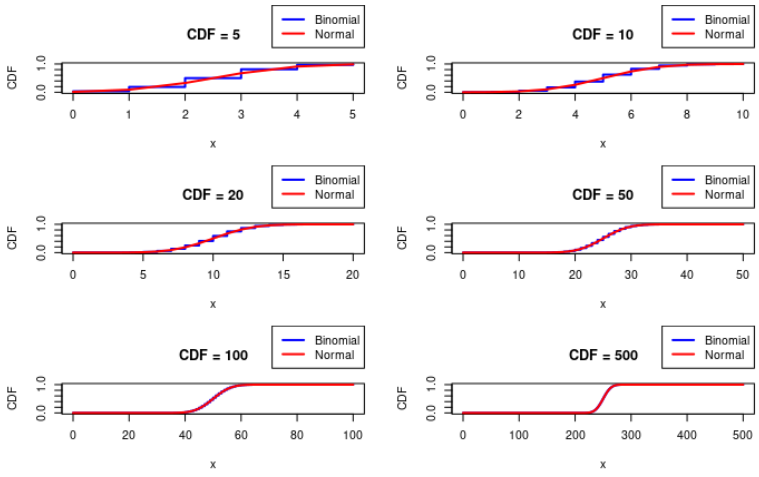


Figura 5: CDF | p=0.5

Para p = 0.9

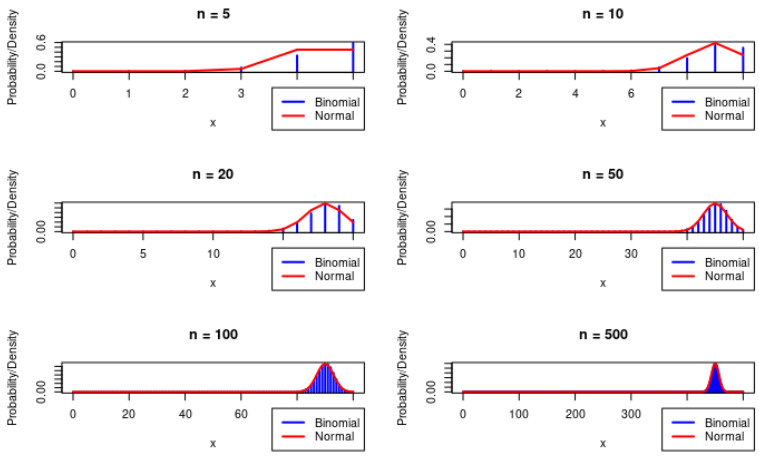


Figura 6: Función de masa | p=0.9

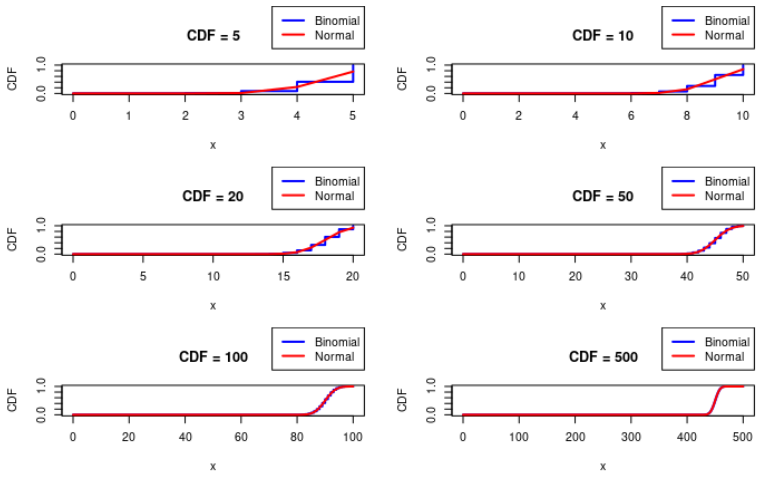


Figura 7: CDF | p=0.9

Sí hay cambio, de hecho, podemos ver como para $p = 0.1$ la función de masa está cargada hacia la izquierda para todos los n, caso contrario sucede para $p = 0.9$, donde $g(x)$ está cargada hacia la derecha. En el caso de $p = 0.5$, la función de masa se mantiene centrada, con media a la mitad de nuestro intervalo.

Este comportamiento sucede también con la CDF, es muy notorio como para $p = 0.5$ el punto de inflexión ocurre a la mitad del intervalo, mientras que para $p = 0.1$ y $p = 0.9$, este ocurre más cargado a la izquieda o a la derecha, respectivamente.

Problema 4

Una partícula se encuentra en 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 izquierdaes 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 "caminata aleatoria en una dimensión".

Sea X_n la posición de la partícula después de n saltos. Podemos expresar X_n como la suma de los movimientos individuales de x_i. Entonces:

$$X_n = \sum_{i=1}^{n} x_i$$

Aquí, x_i es una variable aleatoria que toma el valor de (-1) si la partícula va a la izquierda y $(+1)$ si va a la derecha, por lo que $x_i = \pm 1$. Ambas con probabilidad p y $1 - p$ respectivamente.

Por lo tanto, la esperanza $E(x_i)$ se vería como:

$$E(x_i) = (-1)p + (1)(1 - p) = 1 - 2p$$

Esto sucede por ser la suma ponderada de los valores posibles multiplicados por la probabilidad. Entonces, la esperanza de la partícula tras n saltos es la suma de las esperanzas de cada salto:

$$E(X_n) = \sum_{i=1}^{n} E(x_i) = n \cdot (E(x_i)) = n \cdot (1 - 2p)$$

La varianza de X_n, a su vez, será la suma de las varianzas de cada salto, pues cada salto será independiente. Por lo que tendremos los siguiente:

$$V(X_n) = \sum_{i=1}^{n} V(x_i) = n \cdot V(x_i)$$

En el caso de la varianza, sabemos que: $V(X) = E(X^2) + (E(X))^2$. Por lo que podemos desarrollar para encontrar el valor de $V(x_i)$, pues además sabemos que $x_i = \pm 1$, así:

$$\Rightarrow E(x_i^2) = (-1)^2 p + (1)^2(1 - p) = p + 1 - p = 1$$

$$\Rightarrow (E(x_i))^2 = (1 - 2p)^2 = 1 - 4p + 4p^2$$

$$\Rightarrow E(x_i^2) - (E(x_i))^2 = 1 - (1 - 4p + 4p^2) = 4p - 4p^2 = 4p(1 - p)$$

Finalmente:

$$V(X_n) = E(x_i^2) - (E(x_i))^2 = n \cdot \Big(4p(1 - p)\Big) = 4np(1 - p)$$

Por lo tanto:

$$V(X_n) = 4np(1 - p)$$
$$E(X_n) = n(1 - 2p)$$

Problema 5

El siguiente conjunto de datos contiene mediciones del diámetro de un agave, medido en decímetros, en distinras locaciones no cercanas. (El conjunto de datos se presenta en el código).

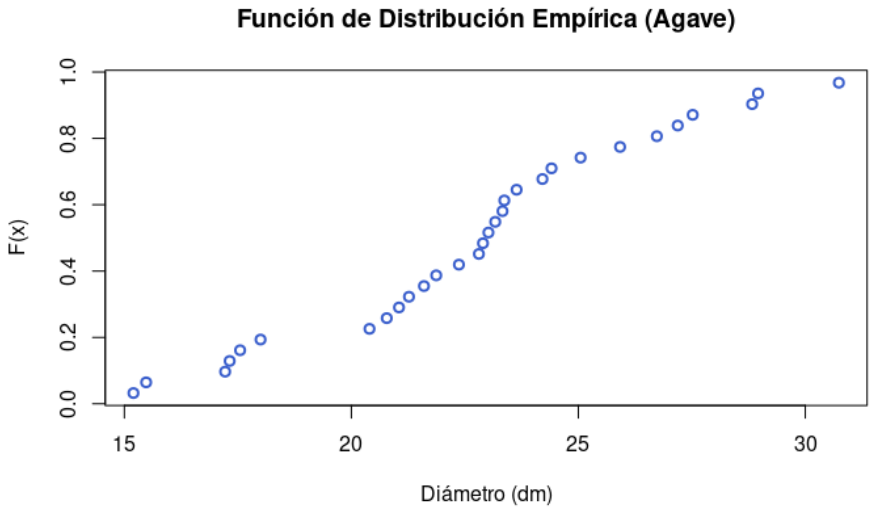
Antes que nada, sería bueno hablar un poco acerca de la Distribución Empírica. Se trata de una distribución basada en los datos que se observan, en este caso, los diámetros del agave, sin asumir ningún modelo previamente. Esta es una de las principales diferencias con una distribución teórica establecida, como la normal o la binomial. Es bastante útil para describir la probabilidad acumulada de un conjunto de datos observados.

Para un conjunto de datos $D = d_1, d_2, d_3, ..., d_n$ la función de distribución empírica $F_D(x)$ evalúa la proporción de los datos que son menor o iguales a un x seleccionado:

$$F_D(x) = \frac{\text{Número de elementos en } D \leq x}{n}$$

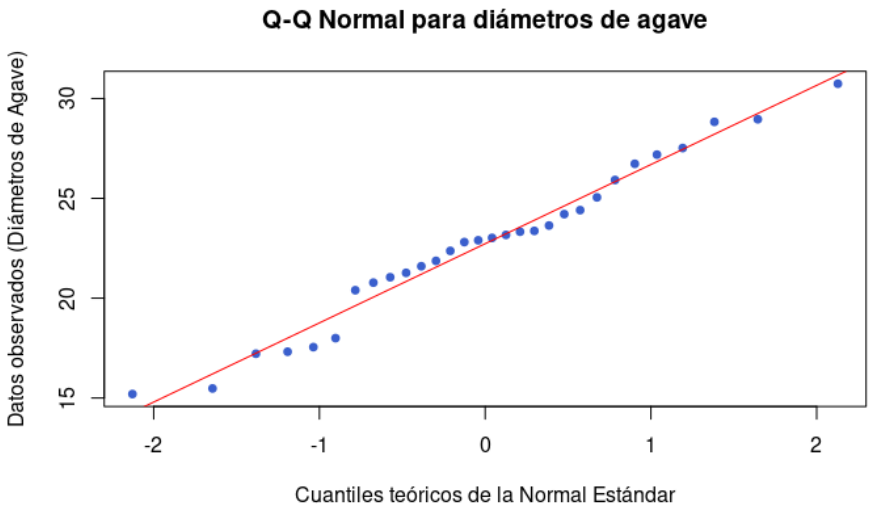
Esta función toma valores en el rango de [0,1], además de ser no decreciente con saltos en los puntos donde encontramos las observaciones de los datos. Cada salto tiene un incremento de magnitud $\frac{1}{n}$, con n como el tamaño del conjunto de datos.

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 parámetro 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 funciones implementadas en R que hacen lo pedido.**



Podemos notar como la distribución empírica tiene saltos en los puntos donde hay datos, y los saltos son de magnitud $\frac{1}{n}$. Esto indica cómo la función de distribución se mantiene constante hasta el siguiente valor que provoca el salto. También podemos ver cómo la función mantiene una tendencia monótona no decreciente.

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 $sx + \bar{x}$ (con s como a desviación estándar muestral, y $\bar{x}$ como la media muestral). Usando esta función, determine la gráfica Q-Q Normal. ¿ Qué observa?**



Bueno, aquí estaría bien explicar un poco sobre qué nos está diciendo la gráfica. En este QQ-Plot, podemos ver puntos que representan los diámetros observados del agave frente a los cuantiles teóricos de una distribución normal estándar. La línea roja que atraviesa el gráfico es la representación de la recta de ajuste $xs + \bar{x}$, con s como la desviación estándar muestral y $\bar{x}$ la media muestral.

Además, en la región central del gráfico (cuando los cuantiles son cercanos a 0), los puntos se aproximan a la línea roja, esto indica que los diámetros de agave se ajustan bien a la distribución normal en el rango medio de valores. Asimismo, en los extremos izquierdo y derecho podemos observar como los puntos comienzan a desviarse de la línea roja. Esto es sobre todo notorio en la parte derecha, donde los diámetros son mayores a lo esperado por los datos. Sin embargo, pese a la dispersión, podemos asegurar que el comprotamiento de los datos presenta una tendencia normal bastante sólida.

c) **Añada a la función anterior (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?**

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

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

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

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

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

Las gráficas quedaron como se ven en la siguiente página.

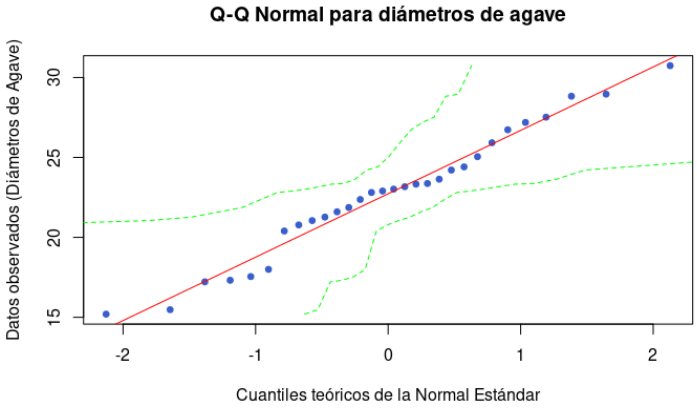


Figura 8: Bandas para $\alpha = 0.05$

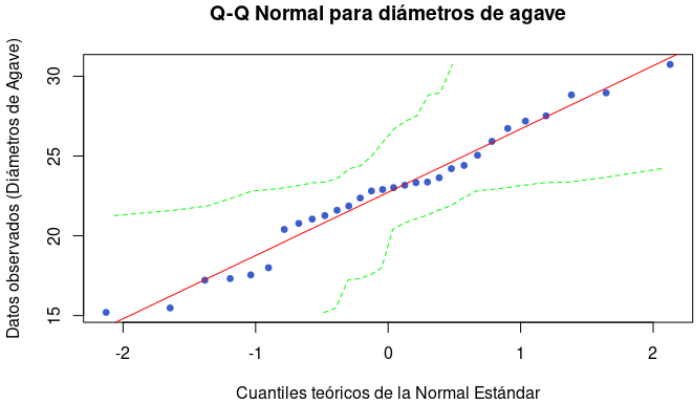


Figura 9: Bandas para $\alpha = 0.09$

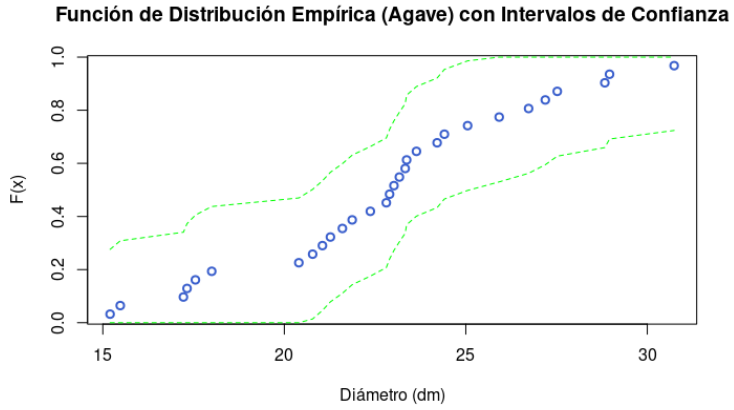


Figura 10: Bandas para $\alpha = 0.05$

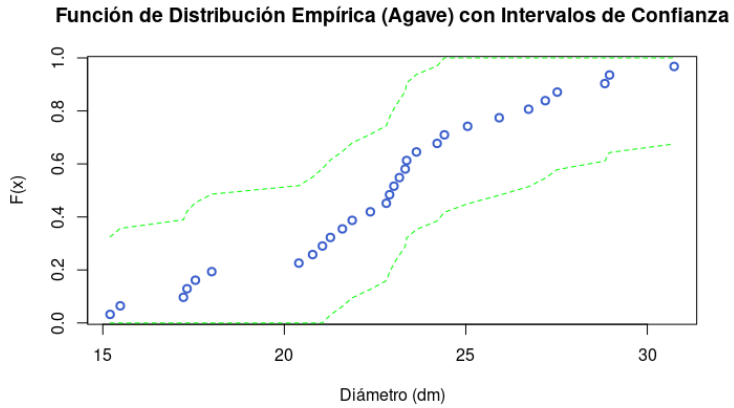


Figura 11: Bandas para $\alpha = 0.09$

Problema 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. Considere las distribuciones: $Unif(a = -3, b = 3)$**,** $Normal(0, 1)$**,** $Exp(2)$**,** $Gamma(\alpha = 2, \beta = 1)$**,** $Gamma(\alpha = 3, \beta = 1)$**,** $Beta(\alpha = 2, \beta = 2)$**,** $Weibull(\alpha = 4, \beta = 1)$ **y** $LogNormal(\mu = 3, \sigma = 2)$**.**

a) **Lee las reglas empíricas.**

También conocidas como regla 68-95-99.7, se trata de una especie de "atajo" o "truco" para recordar propiedades de las distribuciones en estadística y probabilidad. Principalmente se usa para recordar cómo se concentran los datos dentro de la Distribución Normal (DN), de hecho de ahí viene su el nombre de 68-95-99.7, pues para la DN podemos encontrarnos al 68 % de los datos dentro de una distribución estándar σ, al 95 % de los datos dentro de 2σ y al 99.5 % dentro de 3σ.

Esta propiedad tiene implicaciones directas. Por ejemplo, nuevamente para la DN, la mayoría de las observaciones cae dentro del rango $\mu \pm \sigma$, i. e., una desviación estándar por encima o por debajo de la media. Entonces, si seleccionamos un valor al azar dentro de nuestro conjunto de datos distribuido normalmente, hay un 68 % de probabilidad de que esté dentro de dicho rango.

Esto nos puede ayudar a realizar inferencias rápidas sobre el comportamiento de los datos, como conocer qué tantos de ellos tenemos alrededor de σ, 2σ o 3σ.

b) **Para cada una de las distribuciones anteriores, haga una tabla que muestre las probabilidad contenidas en los intervalos $(\mu - \sigma, \mu + \sigma)$, 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.**

Los resultados obtenidos por la regla empírica

Distribución	$k = 1$	$k = 2$	$k = 3$
Uniforme	0.5773503	1.0000000	1.0000000
Normal	0.6826895	0.9544997	0.9973002
Exponencial	0.8646647	0.9502129	0.9816844
Gamma21	0.7375188	0.9533779	0.9859151
Gamma22	0.7375188	0.9533779	0.9859151
Beta	0.9690942	1.0000000	1.0000000
Weibull	0.6717197	0.9571037	0.9991509
Lognormal	0.9802718	0.9912252	0.9948701

Cuadro 1: Valores de k para diferentes distribuciones

Esta tabla muestra los valores teóricos para cada una de las distribuciones establecidas. Podemos notar como el cálculo parece correcto al revisar los resultados de la Distribución Normal, destacados en verde.

Es notorio como la distribución parece haber un resultado extraño para el caso de $k = 1$, pues debería dar una probabilidad igual a 1 para todo su dominio. Esto parece estar sucediendo ya que el intervalo se sale de la función.

Algo curioso también sucede con la Distribución Gamma. Se realizaron dos pruebas extra, llamémoslas $Gamma(1, 2)$ y $Gamma(1, 300)$ y encontramos que los valores para $k = 1, 2, 3$ fueron, en ambos casos, $(0.8646647, 0.9502129, 0.9816844)$ respectivamente. Podemos notar en la tabla que $Gamma(2, 1)$ y $Gamma(2, 2)$ los valores para $k = 1, 2, 3$, fueron también los mismos para ambas funciones: $(0.7375188, 0.9533779, 0.9859151)$. Esto nos puede estar indicando que para la probabilidad Gamma, la localización de los datos que se distribuyen según esta función mantiene las mismas proporciones si se fija α o β.

Asimismo, para la función $Weibull(4, 1)$ podemos ver que mantiene una proporción muy similar con la $Normal(0, 1)$. Entonces prácticamente mantiene la regla 67-95-99.7.

Por otra parte, tanto la *Lognormal* como la *Beta* mantienen proporciones muy similares entre ellas. Esto no significa directamente que tengan un perfil similar, pero al menos sí nos deja ver que para estas función es difícil decir que cierta cantidad de datos caerá en σ, 2σ o 3σ.

c) **En R, simule $n = 1000$ muestras de cada una de las distribuciines anteriores y calcule la media muestral $\bar{x}$ y la vairanza muestral s^2 como se mencionó en clases. 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 anterior?**

Distribución	**Media**	**Varianza**	$k = 1$	$k = 2$	$k = 3$
Uniforme	0.0484	2.9568	0.589	1.000	1.000
Normal	-0.0010	0.9573	0.677	0.947	0.999
Exponencial	0.4662	0.2198	0.867	0.947	0.981
Gamma21	1.9841	1.9524	0.714	0.952	0.989
Gamma22	3.9342	7.4140	0.727	0.951	0.986
Beta	0.5135	0.0499	0.618	0.981	1.000
Weibull	0.9162	0.0692	0.668	0.957	1.000
Lognormal	140.5787	409251.3226	0.973	0.983	0.990

Cuadro 2: Media, varianza y proporciones para diferentes distribuciones

Esta tabla muestra los valores para $k = 1, 2, 3$ después de realizar 1000 simulaciones. Estos resultados son ciertamente parecidos a los de las reglas empíricas. Podemos ver en el Cuadro 3 la diferencia absoluta entre los resultados de la simulación y los de las teórucas.

Distribución	**k=1**	**k=2**	**k=3**
Uniforme	0.0116	0.0000	0.0000
Normal	0.0057	0.0075	0.0017
Exponencial	0.0023	0.0032	0.0007
Gamma(2,1)	0.0235	0.0014	0.0031
Gamma(2,2)	0.0105	0.0024	0.0001
Beta	0.3511	0.0190	0.0000
Weibull	0.0037	0.0001	0.0008
Lognormal	0.0073	0.0082	0.0049

Cuadro 3: Diferencias absolutas para las distribuciones en función de $k = 1, 2, 3$.

Para cada una de las distribuciones podemos ver que hay diversas magnitudes de diferencia. En general, parece que se mantienen bien excepto por algunos casos específicos, como en el caso $k = 1$ de la función *Beta*. Por lo que podemos concluir que las reglas empíricas en general pueden ser una aproximación confiable para ver cómo se acumulan los datos alrededor de diferentes σ.