UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Ciencias

Inferencia a través del Análisis Gráfico de Datos

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

PRESENTA:

JAVIER AGUIRRE GUTIÉRREZ

DIRECTOR DE TESIS:

MAT. MARGARITA ELVIRA CHÁVEZ CANO

2014

1. Datos del alumno
Aguirre
Gutiérrez
Javier
55 29 93 82
Universidad Nacional Autónoma de
México
Facultad de Ciencias
Actuaría
306555655

2. Datos del tutor
Mat
Margarita Elvira
Chávez
Cano

3. Datos del sinodal 1
Act
Francisco
Sánchez
Villarreal

4. Datos del sinodal 2
Dra
Ruth Selene
Fuentes
García

5. Datos del sinodal 3
Dr
Fernando
Blatazar
Larios

6. Datos del sinodal 4
Act
Harim
García
Lamont

7. Datos del trabajo escrito
Inferencia a través del Análisis Gráfico de Datos
108 p
2014

Inferencia a través del Análisis Gráfico de Datos

Javier Aguirre Gutiérrez

2014

A mis padres: Micaela y Javier,
a mi familia,
a todos mis queridos amigos y maestros,
que han sido la inspiración en mi vida.

Índice General

Lista de Gráficas

Capítulo 1

Introducción

Una gráfica es generalmente aceptada como una de las formas más claras y efectivas de presentar datos. Muchos escritos hablan sobre la presentación gráfica, sobre el método apropiado que debe ser usado para algún propósito en específico y sobre las ventajas y desventajas de diversos enfoques en la materia.

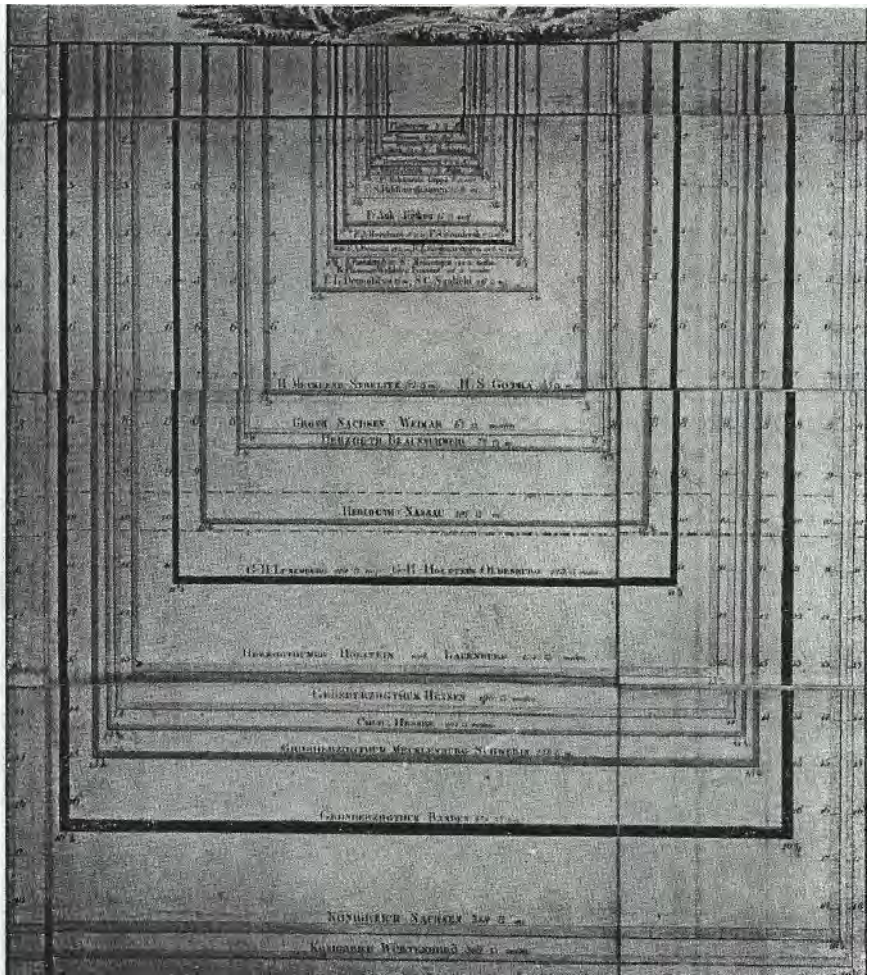
Ha habido poca curiosidad acerca de los origines históricos de la representación gráfica de datos; es imposible establecer que un método gráfico dado haya sido introducido por alguien o en algún momento en específico, sin embargo, se puede decir por quién, cierto método fue usado y la época en que lo investigó.

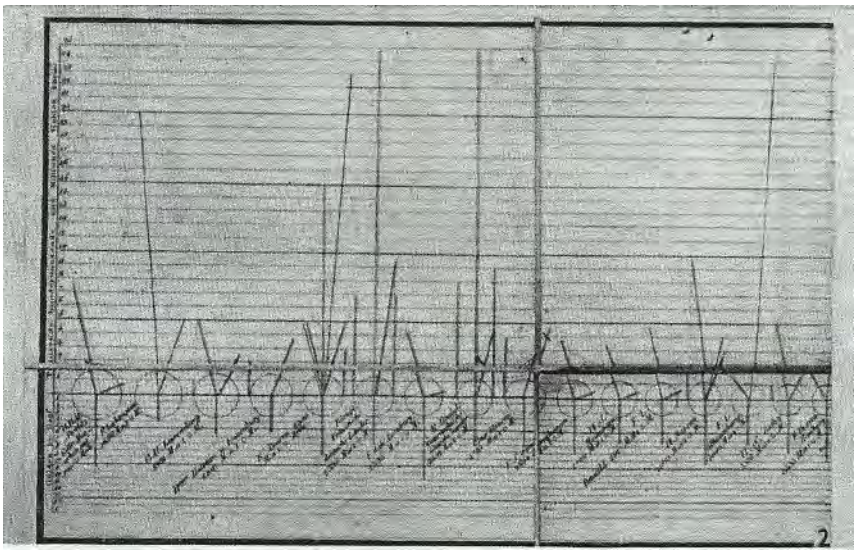
El desarrollo de la técnica en la que se usan coordenadas cartesianas, en el que un eje representa el tiempo, data de hace unos pocos siglos, aunque la representación espacial del tiempo se originó mucho antes. Existen casos en los que se registra el movimiento de las estrellas, específicamente, las inclinaciones de las orbitas planetarias siendo graficadas como función del tiempo, la música escrita representa uno de los primeros ejemplos del uso de series de tiempo.

La idea de usar coordenadas para determinar la ubicación de un punto en el espacio data al menos de los Griegos, aunque no fue hasta la época de Descartes que los matemáticos desarollaron sistemáticamente la idea.

Uno de los primeros usarios de la representación gráfica en Estadística fue A.F.W. Crome, quien usó la representación geométrica como ayuda para la enseñanza; Crome era un Geógrafo antes que un Estadístico, sus trabajos son descriptivos y aunque *Allgemeine Deutsche Biografie* lo llama un "pionero en la Estadística", seguramente usa la palabra en su acepción original (como datos que se refieren a estados). Entonces, el adjetivo estadístico, que aparece en casi todos los títulos de sus trabajos,

indica que ciertos datos numéricos son dados. En el trabajo *Über die Größe und Bevölkerung der sämtlichen europäischen Staaten (1785)*, Crome da una detallada descripción sobre datos geográficos de la mayor parte de los estados Europeos, y para resaltar la importancia de los datos de forma más clara, Crome realiza su *Großen Karte* que es similar a la gráfica que se presenta a continuación, dónde el área de cada uno de los estados fue tratado como un cuadrado porporcional al área. Entonces, se dibujan los cuadrados uno dentro del otro, de una manera tal como para mantener los lados verticales de los cuadrados, que son proporcionales a la raíz cuadrada del área representada, en escala. En realidad, Crome usó una representación espacial de magnitud, no magnitudes lineales que se mueven a través del tiempo.



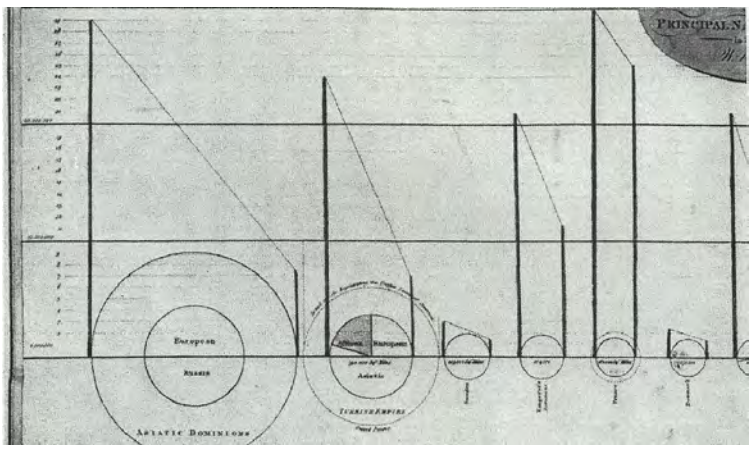


En 1820, Crome publicó la gráfica anterior, la cual consiste en una serie de círculos, los cuales representan la densidad de población en el estado al que se refiere, siendo la densidad inversamente proporcional al área del circulo. Las medias tangentes y los radios extendidos denotan el total de la población y el ingreso nacional (las líneas son dibujadas en distintos colores para poder diferenciarlas).

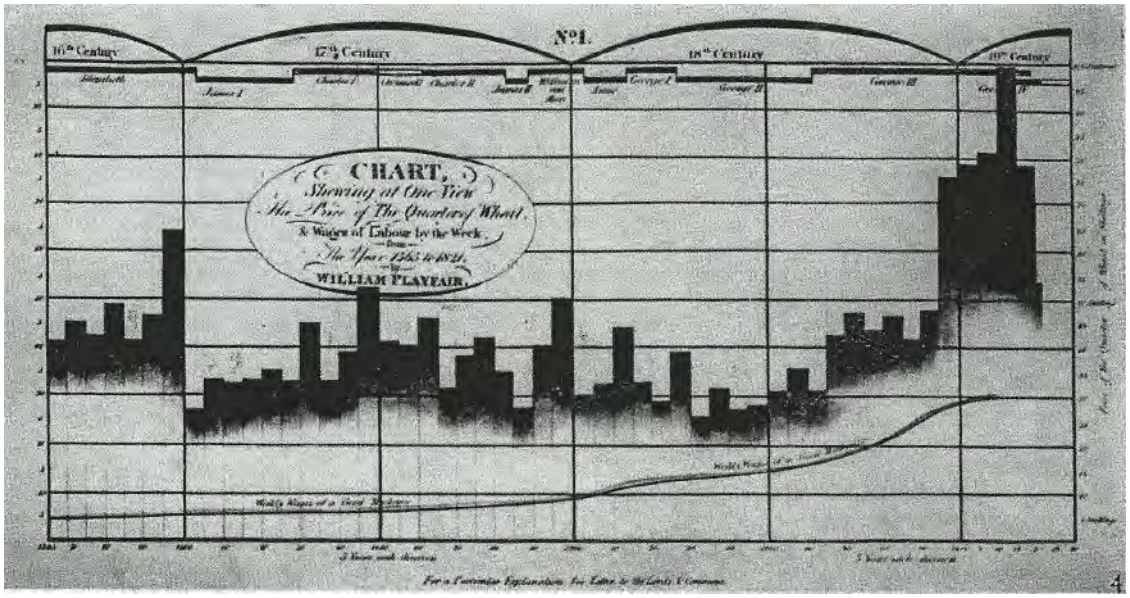
Este método es quizá más preciso que si la población y el ingreso nacional hubieran sido graficadas a escala, este método pudo haber sido un intento por realizar una gráfica de barras.

Los trabajos económicos de William Playfair (nacido en 1759) fueron ilustrados con histogramas, diagramas de pay. El escribió principalmente en Economía descriptiva, utilizó gráficas de Series de Tiempo, y círculos para representar magnitudes espaciales.

El objetivo de Playfair en publicar sus gráficas, era en el sentido de hacer Estadística, presumiblemente en el sentido de datos relacionados con estados.



Playfair publicó la gráfica anterior en 1801 (*Statistical Breviary*), antes que el trabajo de Crome, en él, muestra a los estados europeos como círculos, y en algunos casos, son subdividos como gráficas de pay. La línea del lado izquierdo del círculo, representa la población en millones, y la del lado derecho, representa ingresos nacionales en millones de libras. La pendiente que trata de unir a éstas dos trata de mostrar el radio aproximado de una con otra.



La gráfica anterior, muestra que dos series han sido superpuestas una sobre la otra, una como gráfica y la otra como histograma (las dos series aquí son los salarios semanales y el precio de un cuarto de trigo.)

Parece ser posible que la originalidad de Playfair recae no en el tipo de diagramas que utilizó, sino en las aplicaciones en Estadística descriptiva y Economía. William Playfair si no es el inventor de la representación gráfica como la conocemos, él fue quién la introdujo a la Estadística.

El objetivo de la presente tesis es la de estudiar las características de una muestra aleatoria y poder realizar inferencias acerca de ésta a través de un Análisis Gráfico, pudiendo de este modo ajustarle una distribución de probabilidad a la muestra para así poder realizar inferencias en cuanto a su comportamiento en las colas, simetría, cuantiles, así como la función muestral que estimará a la distribución teórica.

En el primer capítulo, se propone a la Función de Distribución Empírica como estimador de la función de distribución teórica, se describen algunas propiedades de ésta y se elaboran bandas de confianza para la función de distribución, utilizando el Teorema de Glivenko-Cantelli y la Desigualdad de Dvoretzky-Kiefer-Wolfowitz.

En el segundo capítulo, se investiga la Simetría de un conjunto de datos y se describen las principales técnicas para su correcta evaluación.

En el tercer capítulo, se describe la cola de la distribución que sigue un conjuntos de datos, el papel semilogarítmico en la evaluación de la cola y las principales funciones de distribución que son de interés para este enfoque.

En el cuarto capítulo, se evalua el Ajuste de la distribución completa, se realiza una comparación gráfica de la función de distribución teórica con la función de distribución empírica a través de una doble gráfica y el estudio de las gráficas Cuantil-Cuantil.

En el quinto capítulo, se estudian las gráficas de probabilidad y la relación lineal de algunas funciones de distribución para poder hacer una correcta evaluación sobre la función de distribución que siguen los datos.

Finalmente, se realiza una aplicación utilizando las técnicas gráficas estudiadas en la tesis.

Capítulo 2

Función de Distribución empírica

El propósito de muestrear alguna distribución es hacer inferencias acerca de la distribución muestrada, la cual se supone al menos en parte desconocida.

Una pregunta que surge es: ¿Por qué no estimar a la función de distribución desconocida?

La respuesta a la pregunta anterior es que podemos estimar a la función de distribución acumulativa desconocida usando la función de distribución acumulativa empírica (o muestral), la cual es una función de las estadísticas de orden.

Definición. *Función de distribución acumulativa muestral (o empírica).*
Sea $X_1, \ldots, X_n$ una muestra aleatoria (v.a.i.i.d.) de una distribución con función de distribución acumulativa $F(\cdot)$. Sean $X_{(1)} \leq X_{(2)} \leq, \ldots, \leq X_{(n)}$ las correspondientes estadísticas de orden.
La función de distribución acumulativa muestral, denotada por $F_n(x)$ es definida por:

$$F_n(x) = \frac{\#(X_i \leq x)}{n} \qquad \textit{(Número de } X_i \textit{ menores o iguales a } x\textit{).}$$

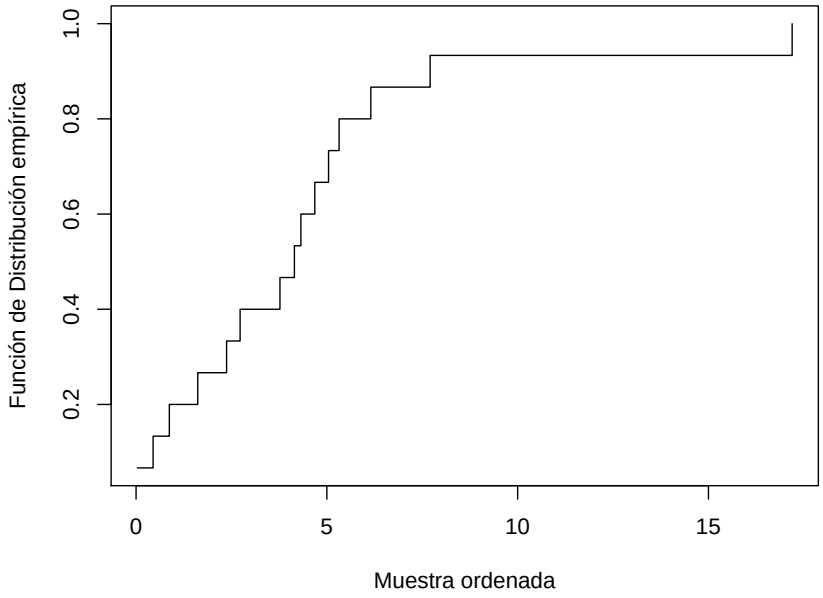
La definición formal es la siguiente:

$$F_n(x) = \begin{cases} 0 & \text{si } x < X_{(1)}, \\ \dfrac{i}{n} & \text{si } X_{(i)} \leq x < X_{(i+1)} \text{ con } i = 1, 2, ..., n-1, \\ 1 & \text{si } X_{(n)} \leq x. \end{cases}$$

Por lo que $F_n(x)$ es una función escalonada, obtenida de los datos; pues conforme x aumenta su valor, la función $F_n(x)$ realiza un salto (hacia arriba) de altura $\frac{1}{n}$ (si no hay valores repetidos) conforme cada observación de la muestra es alcanzada.

La función de distribución empírica tiene las siguientes características:

- Para x fija, $F_n(x)$ es una estadística, ya que es una función de la muestra.
- $F_n(x)$ registra la proporción de observaciones menores o iguales a x.



Gráfica 2.1: *Función de distribución acumulativa muestral de una muestra de tamaño* 15 *proveniente de una distribución Exponencial*(1/4) .

Definición. *Función indicadora.*
Sea Ω cualquier conjunto, con puntos w, sea A cualquier subconjunto de Ω . La función indicadora de A, denotada por $\mathbb{1}_A(\cdot)$ se define como:

$$\mathbb{1}_A(w) = \begin{cases} 1 & \text{si } w \in A, \\ 0 & \text{si } w \notin A. \end{cases}$$

Entonces podemos reescribir la función de distribución muestral como:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i), \text{ donde } X_1, \ldots, X_n \text{ es la m.a. de alguna distribución}$$

y

$$\mathbb{1}_{(-\infty,x]}(X_i) = \begin{cases} 1 & \text{si } X_i \in (-\infty, x], \\ 0 & \text{si } X_i \notin (-\infty, x]. \end{cases}$$

Sea $Z_i = \mathbb{1}_{(-\infty,x]}(X_i)$, es decir, la v.a. nos dice si X_i fue menor o igual a x,

$W = \sum_{i=1}^{n} Z_i$, es decir, la v.a. que cuenta el número de X_i's con $i = 1, \ldots, n$ que fueron menores o iguales a x.

Para cada muestra $X_1, \ldots, X_n$ consideramos a $\{X_j \leq x\}$ como el evento éxito y $\{X_j > x\}$ como el evento fracaso.

Entonces W es el número de éxitos en n realizaciones del experimento (consideramos un experimento a la acción de observar si X_i fue menor o igual a x y de serlo, contarlo como 1).

La probabilidad de que ocurra un éxito es la misma para cada $X_i,\ i = 1, \ldots, n$ pues son variables aleatorias idénticamente distribuidas:

$$\mathbb{P}[X_j \leq x] = F_X(x) = p.$$

El éxito o fracaso de la j-ésima realización del experimento es independiente del resultado de cualquier otra realización del experimento, pues X_j es independiente a X_s, con $j \neq s$, por lo que:

$$W \sim Binomial(n, p) \quad \text{con } p = F_X(x) \ \text{ y } \ q = 1 - p,$$

es decir

$$\mathbb{P}[W = k] = \binom{n}{k} p^k q^{n-k} \mathbb{1}_{\{0,1,\ldots,n\}}(k).$$

Debido a que:

$$\{W = k\} = \left\{n\frac{W}{n} = k\right\} = \left\{n\overline{W} = k\right\} = \left\{\overline{W} = \frac{k}{n}\right\}$$

y como

$$\overline{W} = F_n(x),$$

tenemos que:

$$\mathbb{P}\left[F_n(x) = \frac{k}{n}\right] = \binom{n}{k} p^k q^{n-k} \mathbb{1}_{\{0,1,\ldots,n\}}(k).$$

Por lo tanto, $nF_n(X) \sim Binomial(n, p)$.

Teorema. *Sea $F_n(x)$ la función de distribución acumulativa muestral, de una m.a. de tamaño n, proveniente de $F(\cdot)$, entonces:*

$$\mathbb{P}\left[F_n(x) = \frac{k}{n}\right] = \binom{n}{k} \left[F_n(x)\right]^k \left[1 - F_n(x)\right]^{n-k} \mathbb{1}_{\{0,1,\ldots,n\}}(k).$$

De donde:

$$\mathbb{E}\left[F_n(x)\right] = \mathbb{E}\left[\overline{W}\right] = \mathbb{E}\left[\frac{W}{n}\right] = \frac{1}{n}\mathbb{E}\left[W\right]$$

y debido a que $W \sim Binomial(n, p)$

$$= \frac{1}{n}\left(np\right) = p = F_X(x),$$

es decir, la función de distribución muestral es un estimador insesgado.

También:

$$\begin{aligned} Var[F_n(x)] = Var\left[\frac{W}{n}\right] &= \frac{1}{n^2} Var[W] \\ &= \frac{1}{n^2}\Big(np\left(1-p\right)\Big) = \frac{1}{n}p(1-p) \\ &= \frac{1}{n}F_X(x)\Big(1 - F_X(x)\Big). \end{aligned}$$

Teorema del Límite Central. *Sean $X_1, \ldots X_n$ v.a.i.i.d. con n entero positivo, con media $\mu_X < \infty$ y varianza $\sigma_X^2 < \infty$, entonces, para cada z:*

$$\lim_{n\to\infty} F_{Z_n}(z) = \mathbb{P}[Z_n \leq x] \longrightarrow \Phi(z),$$

donde:

$$Z_n = \frac{\Big(\overline{X_n} - \mathbb{E}[\overline{X_n}]\Big)}{\sqrt{Var(\overline{X_n})}},$$

$\Phi(\cdot)$ *es la función de distribución acumulativa normal estándar,*

$$\overline{X_n} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

La prueba del teorema puede encontrarse en [12].

Note que este teorema dice que $\overline{X_n}$ se distribuye asintóticamente como una Normal con media μ_X y varianza $\frac{\sigma_X^2}{n}$.

Retomando la notación anterior al Teorema:

Notemos que $F_n(x) = \overline{W}$ es la media muestral de las variables aleatorias $Z_1, \ldots, Z_n$ con
$Z_i = \mathbb{1}_{(-\infty,x]}(X_i), \quad i = 1, \ldots, n.$
De acuerdo al Teorema del Límite Central tenemos para cada x fija:

$$\mathbb{P}\left[\frac{F_n(x) - \mathbb{E}[F_n(x)]}{Var[F_n(x)]} < z\right] \longrightarrow \Phi(z)$$

y como:

$$\mathbb{E}[F_n(x)] = F(x) \quad \text{y} \quad Var[F_n(x)] = \frac{F(x)\Big(1 - F(x)\Big)}{n},$$

se tiene que:

$$F_n(x) \text{ se distribuye asintóticamente } N\left(F(x), \frac{F(x)\Big(1 - F(x)\Big)}{n}\right).$$

Las siguientes definiciones y teoremas resultan de utilidad:

Definición. *Una sucesión de variables aleatorias* $\{X_n\}_{n\geq 1}$ *converge en distribución (o converge débilmente) a una variable aleatoria* X *(se le denota por* $X_n \xrightarrow{d} X$*) si:*

$$\mathbb{P}[X_n \leq x] \to \mathbb{P}[X \leq x] \quad \begin{array}{l} \textit{cuando } n \to \infty \\ \textit{en todos los puntos } x \\ \textit{donde } F_X(x) \textit{ es continua.} \end{array}$$

Definición. *Una sucesión de variables aleatorias* $\{X_n\}_{n\geq 1}$ *converge en probabilidad a una variable aleatoria* X *(se le denota por* $X_n \xrightarrow{p} X$*) si:*

$$\textit{Para todo } \epsilon > 0, \quad \mathbb{P}\left[\mid X_n - X \mid > \epsilon\right] \to 0 \quad \textit{cuando } n \to \infty.$$

Definición. *Una sucesión de variables aleatorias* $\{X_n\}_{n\geq 1}$ *converge casi seguramente a una variable aleatoria* X *(se le denota por* $X_n \xrightarrow{c.s.} X$*) si:*

$$\mathbb{P}\left[\lim_{n\to\infty} \mid X_n - X \mid = 0\right] = 1.$$

Definición. *Una sucesión de variables aleatorias* $\{X_n\}_{n\geq 1}$ *converge a una variable aleatoria* X *en media cuadrática (se le denota por* $X_n \xrightarrow{m.c.} X$*) si:*

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0.$$

Se mencionan a continuación algunas relaciones entre los tipos de convergencia de variables aleatorias mencionados previamente.

$$\text{Si} \quad X_n \xrightarrow{\text{c.s.}} X, \quad \text{entonces} \quad X_n \xrightarrow{\text{p}} X.$$

$$\text{Si} \quad X_n \xrightarrow{\text{p}} X, \quad \text{entonces} \quad X_n \xrightarrow{\text{d}} X.$$

La prueba de estas afirmaciones puede encontrarse en [8].

Teorema. *Desigualdad de Markov. Sea* X *una v.a. no negativa y supongamos que* $\mathbb{E}[X]$ *existe. Entonces:*

$$\textit{Para todo} \quad t > 0, \qquad \mathbb{P}[X > t] \leq \frac{\mathbb{E}[X]}{t}.$$

La prueba de esta desigualdad puede encontrarse en [8].

Teorema. *Desigualdad de Chebyshev. Sea* $\mathbb{E}[X] = \mu$ *y* $Var[X] = \sigma^2$*. Entonces:*

$$\mathbb{P}\left[\mid X - \mu \mid \geq t\right] \leq \frac{\sigma^2}{t^2} \quad y \quad \mathbb{P}\left[\mid Z \mid \geq k\right] \leq \frac{1}{k^2},$$

donde:

$$Z = \frac{X - \mu}{\sigma}.$$

Demostración.

$$\begin{aligned}\mathbb{P}[\mid X-\mu \mid \geq t] &= \mathbb{P}[\mid X-\mu \mid^2 \geq t^2] \text{ y usando la desigualdad de Markov} \\ &\leq \frac{\mathbb{E}[\mid X-\mu \mid^2]}{t^2} \\ &= \frac{\sigma^2}{t^2}.\end{aligned}$$

Análogamente

$$\begin{aligned}\mathbb{P}[\mid Z \mid \geq t] &= \mathbb{P}[\mid Z \mid^2 \geq t^2] \\ &\leq \frac{\mathbb{E}[\mid Z \mid^2]}{t^2} = \frac{\mathbb{E}[Z^2]}{t^2} \\ &= \frac{\sigma^2}{t^2}.\end{aligned}$$

Sea $t = k\sigma$, entonces:

$$\mathbb{P}[\mid Z \mid \geq k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

□

Retomando la notación anterior a las definiciones y teoremas: De acuerdo a la desigualdad de Chebyshev tenemos:

$$\begin{aligned}\mathbb{P}[\mid F_n(x) - p \mid > \epsilon] &\leq \frac{Var[F_n(x)]}{\epsilon^2} \\ &= \frac{p(1-p)}{n}\frac{1}{\epsilon^2}.\end{aligned}$$

Observemos donde alcanza el valor máximo $p(1-p)$:

$$\begin{aligned}&\text{Sea } f(p) = p(1-p), \\ &\frac{\mathrm{d}f}{\mathrm{d}p} = 1-2p, && \text{igualando a 0, obtenemos:} \\ &1-2p = 0, && \text{despejando a } p\text{, obtenemos que:} \\ &1/2 = p. && \text{El cual es un punto crítico.}\end{aligned}$$

Evaluémoslo en la segunda derivada

$$\frac{\mathrm{d}^2 f}{\mathrm{d}p^2} = -2 < 0. \qquad \text{Por lo que tenemos una concavidad hacia abajo, es decir,un máximo.}$$

Evaluando el punto crítico en la función, tendremos el valor máximo de ésta:

$$f(\frac{1}{2}) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4} \text{ para todo } p,$$

entonces, tenemos que:

$$\mathbb{P}[\mid F_n(x) - p \mid > \epsilon] \leq \frac{1}{4n\epsilon^2}.$$

Aplicando límites en ambos lados de la desigualdad:

$$\lim_{n\to\infty} \mathbb{P}[\mid F_n(x) - p \mid > \epsilon] \leq \lim_{n\to\infty} \frac{1}{4n\epsilon^2},$$

entonces:

$$\lim_{n\to\infty} \mathbb{P}[\mid F_n(x) - p \mid > \epsilon] \longrightarrow 0,$$

es decir, $F_n(x)$ converge en probabilidad a $F(x)$ cuando $n \longrightarrow \infty$.

Lo anterior también se puede notar analizando el Error Cuadrático Medio (ECM):

$$ECM(\widehat{\theta}) = Var(\widehat{\theta_n}) + sesgo^2(\widehat{\theta_n}), \qquad \text{donde: } sesgo(\widehat{\theta_n}) = \mathbb{E}[\widehat{\theta_n}] - \theta,$$
$$se(\widehat{\theta_n}) = \sqrt{Var\left(\widehat{\theta_n}\right)}.$$

Tenemos que:

$$\begin{aligned} sesgo\Big(F_n(x)\Big) &= \mathbb{E}[F_n(x)] - F(x) \\ &= F(x) - F(x) = 0. \\ Var\Big(F_n(x)\Big) &= \frac{F(x)\Big(1 - F(x)\Big)}{n}; \quad se\Big(F_n(x)\Big) = \sqrt{Var\Big(F_n(x)\Big)}, \end{aligned}$$

entonces:

$$ECM\Big(F(x)\Big) \xrightarrow{n\to\infty} 0 \text{ y por lo tanto, } F_n(x) \xrightarrow{\text{p}} F(x),$$

es decir, $F_n(x)$ es consitente.

Se consideran los siguientes Teoremas para el resultado anterior:

Teorema. *Si el sesgo* $\longrightarrow 0$ *y se* $\longrightarrow 0$ *cuando* $n \longrightarrow \infty$*, entonces:*
$\widehat{\theta_n}$ *es consistente, i.e.* $\widehat{\theta_n} \xrightarrow{p} \theta$.

Demostración.

$$\text{Si el } sesgo(\widehat{\theta_n}) \longrightarrow 0,$$
$$se(\widehat{\theta_n}) \longrightarrow 0,$$

entonces:

$$ECM(\widehat{\theta}) = Var(\widehat{\theta_n}) + sesgo^2(\widehat{\theta_n}) \longrightarrow 0,$$

entonces:

$$ECM(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta_n} - \theta)^2] \longrightarrow 0,\ n \longrightarrow 0,$$

i.e.

$$\widehat{\theta_n} \xrightarrow{\text{m.c.}} \theta, \quad \text{entonces} \quad \widehat{\theta_n} \xrightarrow{\text{p}} \theta.$$

□

Teorema. *Sea* $\{X_n\}_{n\geq 1}$ *una sucesión de v.a.*

$$Si \ \ X_n \xrightarrow{m.c.} X, \quad entonces \quad X_n \xrightarrow{p} X.$$

Demostración. Supongamos que $X_n \xrightarrow{\text{m.c.}} X$.
Sea $\epsilon > 0$, entonces usando la desigualdad de Chebyshev:

$$\begin{aligned}\mathbb{P}[\mid X_n - X \mid > \epsilon] &= \mathbb{P}\left[\mid X_n - X \mid^2 > \epsilon^2\right] \\ &\leq \frac{\mathbb{E}\left[(X_n - X)^2\right]}{\epsilon^2} \longrightarrow 0 \qquad \text{cuando } n \longrightarrow \infty\end{aligned}$$

$$\text{Por lo tanto} \quad \mathbb{P}[\mid X_n - X \mid > \epsilon] \longrightarrow 0.$$

□

La siguiente desigualdad es de particular interés.

Teorema. *Desigualdad de Hoeffding.*
Sea $Y_1, \ldots, Y_n$ *una m.a. tal que* $\mathbb{E}[Y_i] = \mu$ *y* $a \leq Y_i \leq b$, *entonces:*

$$Para \ todo \ \ \epsilon > 0, \qquad \mathbb{P}[\mid \overline{Y_n} - \mu \mid \geq \epsilon] \leq 2exp\left\{\frac{-2n\epsilon^2}{(b-a)^2}\right\}.$$

Entonces, usando dicha desigualdad en la función de distribución muestral y como $0 \leq Z_i \leq 1$, tenemos que:

$$\text{Para toda } \epsilon > 0, \qquad \mathbb{P}[\mid F_n(x) - F(x) \mid > \epsilon] \leq 2exp\{-2n\epsilon^2\}.$$

Todo lo que se ha visto anteriormente fue para cada $x \in \mathbb{R}$ fija; si estamos interesados en estimar $F_n(x)$ para toda x (en vez de para x fija), entonces estamos interesados en ver que tan cerca se encuentra $F_n(x)$ de $F(x)$ sobre todos los valores de x, entonces los siguientes resultados son de interés:

Teorema. *Teorema de Glivenko-Cantelli*
Sea $X_1, \ldots, X_n \sim F$*, entonces:*

$$\sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid \xrightarrow{c.s.} 0.$$

La prueba del teorema puede encontrarse en [8].
Dicho teorema establece que con probabilidad uno, la convergencia de $F_n(x)$ a $F(x)$ es uniforme en x.

Dado que estamos estimando una función de distribución desconocida $F(x)$, la forma teórica de hacer la comparación entre $F(x)$ y $F_n(x)$ se basa en el Teorema de Glivenko-Cantelli, de donde podemos obtener una banda de confianza para $F(x)$ basada en $F_n(x)$.

Supongamos que $F(x)$ es una función continua conocida. Entonces es posible calcular el valor de $\mid F_n(x) - F(x) \mid$ para cualquier valor de x. Basta considerar los puntos extremos (derecho e izquierdo) de un intervalo para determinar las magnitudes extremas que alcanza $\mid F_n(x) - F(x) \mid$ en ese intervalo. Es posible suponer que $\mid F_n(x) - F(x) \mid$ alcanza su máximo valor en este intervalo en el punto extremo izquierdo dado que $F_n(x)$ es constante sobre cada intervalo de la forma $x_i \leq x < x_{i+1}$ y dado que el punto derecho no está incluido en el intervalo.
Así que es necesario reemplazar máximo por supremo para estudiar que tan grande puede ser esta función. Estamos interesados por lo tanto en la función:

$$D_n = \sup_x \mid F_n(x) - F(x) \mid .$$

Se sigue ahora que será suficiente evaluar $\mid F_n(x) - F(x) \mid$ en los puntos extremos de cada intervalo y entonces se escoge el valor más grande de éstos para el valor de D_n.

Como $F_n(x)$ es una función de la muestra aleatoria, D_n es una variable aleatoria. Una característica importante de esta variable aleatoria es que su distribución no

depende de $F(x)$ excepto que $F(x)$ es una función continua. A partir de lo anterior, construiremos una banda de confianza para $F(x)$.

D_n no depende de la naturaleza de F ya que $F(X)$ tiene distribución uniforme sobre el intervalo $(0, 1)$, y el valor de D_n será el mismo si se trabaja con X o con $F(X)$. Como resultado podemos decir que se puede estudiar la distribución de D_n muestreando de la distribución uniforme o muestreando de la distribución original.

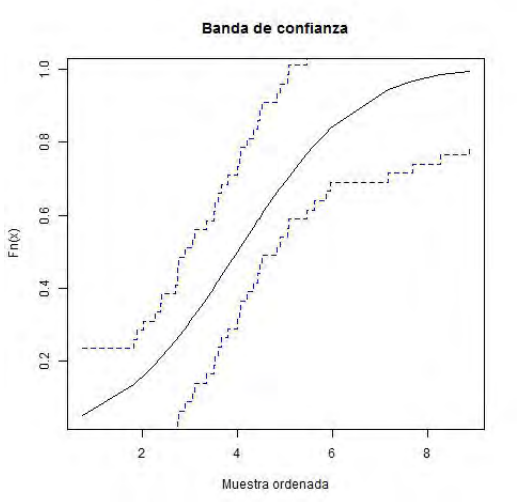
Supongamos que tenemos la distribución de D_n, entonces es posible encontrar un valor denotado por D_n^α tal que:

$$1 - \alpha = \mathbb{P}[D_n \leq D_n^\alpha],$$

entonces:

$$\begin{aligned} 1 - \alpha &= \mathbb{P}[\sup_x \mid F_n(x) - F(x) \mid \leq D_n^\alpha] \\ &= \mathbb{P}[-D_n^\alpha \leq F_n(x) - F(x) \leq D_n^\alpha, \quad \text{para todo } x] \\ &= \mathbb{P}[F_n(x) - D_n^\alpha \leq F(x) \leq F_n(x) + D_n^\alpha, \quad \text{para todo } x]. \end{aligned}$$

Esta última igualdad muestra que con probabilidad $1 - \alpha$ la función de distribución desconocida $F(x)$ está dentro de la banda determinada por las dos funciones escalonadas $F_n(x) - D_n^\alpha$ y $F_n(x) + D_n^\alpha$. La amplitud de esta banda de confianza sirve como medida de precisión de F_x como un estimador de $F(x)$.



Gráfica 2.2: *Banda de confianza usando el Teorema de Glivenko-Cantelli para $F(x)$ con un 95 % de confianza de una muestra $Normal(4, 4)$ de tamaño 40.*

Teorema. *Desigualdad de Dvoretzky-Kiefer-Wolfowitz (D.K.W.)*
Sea $X_1, \ldots, X_n$ una m.a. de F, entonces:

Para todo $\epsilon > 0$ *y cualquier* $n > 0$

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid > \epsilon\right] \leq 2exp\{-2n\epsilon^2\}.$$

Los detalles de esta desigualdad pueden encontrarse en [6].

Usando la desigualdad de D.K.W. podemos obtener otra banda de confianza para $F(x)$:

$$\begin{aligned}
\mathbb{P}\left[\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid > \epsilon\right] &\leq 2exp\{-2n\epsilon^2\} \\
1 - \mathbb{P}\left[\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid > \epsilon\right] &> 1 - 2exp\{-2n\epsilon^2\} \\
\mathbb{P}\left[\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid \leq \epsilon\right] &> 1 - 2exp\{-2n\epsilon^2\}.
\end{aligned}$$

Sea

$$\alpha = 2exp\{-2n\epsilon^2\}, \qquad \text{despejando a } \epsilon \text{ tenemos:}$$
$$\epsilon = \sqrt{\frac{1}{2n}log(\frac{2}{\alpha})},$$

entonces:

$$\mathbb{P}\left[\forall x \in \mathbb{R} \quad \mid F_n(x) - F(x) \mid \leq \sqrt{\frac{1}{2n}log(\frac{2}{\alpha})}\right] > 1 - \alpha,$$

entonces:

$$\mathbb{P}\Big[L(x) \leq F(x) \leq U(x), \quad \text{para toda } x \in \mathbb{R}] > 1 - \alpha,$$

con

$$\begin{aligned}
L(x) &= F_n(x) - \epsilon, \\
U(x) &= F_n(x) + \epsilon.
\end{aligned}$$

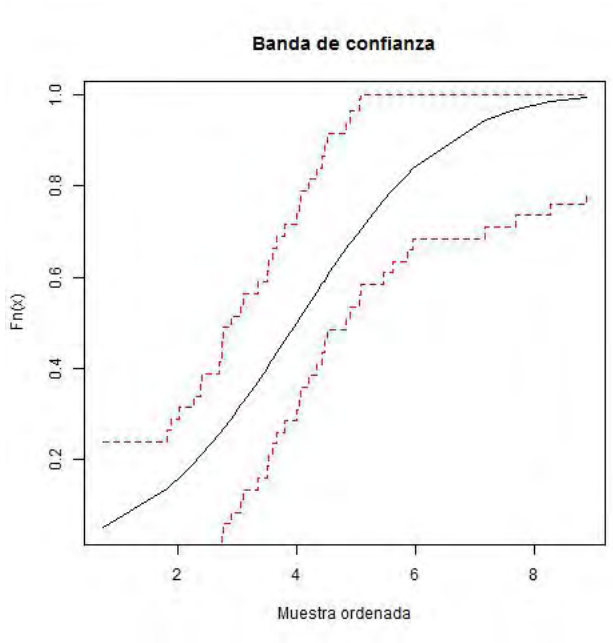
Tomando en consideración que $F(x) \in [0, 1]$, podemos reescribir las cotas de la banda de confianza con:

$$\begin{aligned}
L(x) &= max\{F_n(x) - \epsilon, 0\}, \\
U(x) &= min\{F_n(x) + \epsilon, 1\},
\end{aligned}$$

entonces para cualquier F y para todo n,

$$\mathbb{P}[L(x) \leq F(x) \leq U(x), \quad \text{para todo} \quad x \in \mathbb{R}] > 1 - \alpha.$$

Esta última igualdad muestra que con probabilidad $1 - \alpha$, la función de distribución desconocida $F(x)$ está dentro de la banda determinada por las dos funciones $L(x)$ y $U(x)$.



Gráfica 2.3: *Banda de confianza D.K.W. para* $F(x)$ *con un* 95 % *de confianza de una muestra* $Normal(4, 4)$ *de tamaño* 45.

Capítulo 3

Investigación de la Simetría

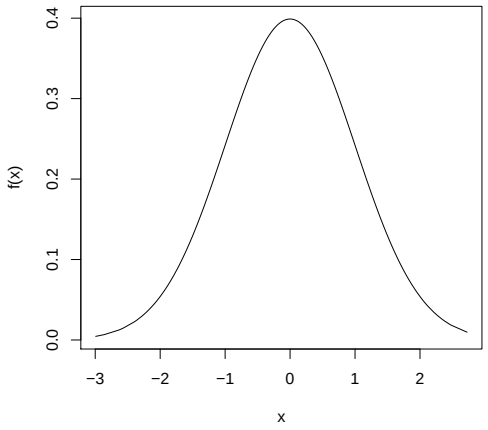
Una distribución o un conjunto de datos se dice que es simétrico si al graficar su función de densidad y si consideramos una recta paralela al eje de las ordenadas que pasa por la media de la función de densidad, en el lado izquierdo de la distribución aparece una "imagen espejo" del lado derecho de la distribución.

El sesgo es una medida sobre la falta de simetría; las distribuciones pueden ser sesgadas positivamente o negativamente.

- Un sesgo positivo (a la derecha) es aquel donde "la cola a la derecha de la mediana es más larga que la cola que se encuentra a la izquierda", donde "los valores de la cola derecha se encuentran más alejados de la mediana que los de la cola izquierda", y "los valores se encuentran concentrados en la cola izquierda".
- Un sesgo negativo (a la izquierda) es aquel donde "la cola a la izquierda de la mediana es más larga que la cola que se encuentra a la derecha", "los valores se encuentran concentrados en la cola derecha", etc.

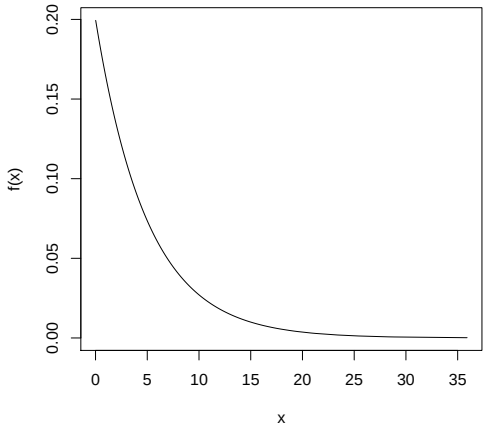
Algunos ejemplos que ilustran los conceptos anteriores son:

1. La distribución Normal (Simétrica).



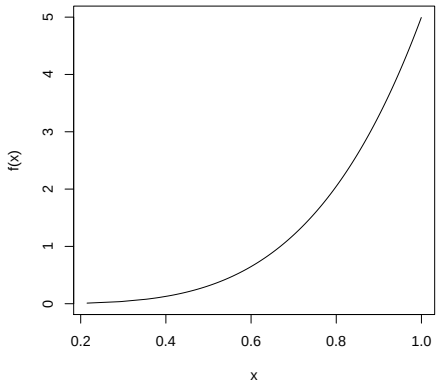
Gráfica 3.1: *Función de densidad de una distribución* $Normal(0, 1)$.

2. La distribución Exponencial (con sesgo positivo).



Gráfica 3.2: *Función de densidad de una distribución* $Exponencial(1/5)$.

3. La distribución Beta(5,1) (con sesgo negativo).



Gráfica 3.3: *Función de densidad de una distribución* $Beta(5,1)$.

Definición. *Percentil.*
Sea $0 < p < 1$. *El* $100p-$*ésimo percentil de una muestra, es el valor que la divide de tal manera que el* $p\,\%$ *de los valores se encuentran debajo y el* $(100-p)\,\%$ *de los valores permanecen arriba.*

- El percentil 25 es conocido como el primer cuartil.
- El percentil 50 es conocido como el segundo cuartil o mediana.
- El percentil 75 es conocido como el tercer cuartil.

$$\underbrace{\underbrace{X_{(1)}, X_{(2)}, \ldots, X_{(m-1)}}_{25\,\%}\ P_{25\,\%}\ X_{(m)}, \ldots, X_{(p-1)}}_{50\,\%}\ P_{50\,\%}\ \underbrace{X_{(p)}, \ldots, X_{(q-1)}}_{75\,\%}\ P_{75\,\%}\ \underbrace{X_{(q)}, \ldots, X_{(n)}}_{25\,\%}.$$

Es común referirse en un contexto estadístico al $100p-$ésimo percentil como el cuantil de orden p. El cuantil de orden p de la distribución de una variable aleatoria X, es un valor ψ_p tal que $\mathbb{P}[X \leq \psi_p] \geq p$ y $\mathbb{P}[X \geq \psi_p] \geq 1-p$.

Entonces, dado un conjunto de valores $x_1, \ldots, x_n$ para poder definir a los cuantiles para cualquier fracción p necesitamos:

- Ordenar los valores de tal forma que se obtenga:

  $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, donde $x_{(i)},\ i = 1, \ldots, n$
  es la i-ésima estadística de orden de la muestra.

- Tomamos a las estadísticas de orden como los cuantiles que corresponden a las fracciones p_i:

$$p_i = \frac{i-1}{n-1}, \qquad i = 1, \ldots n. \tag{3.1}$$

Para definir al cuantil que corresponde a la fracción p, usamos interpolación lineal entre los 2 p_i más cercanos. Si p se encuentra a una fracción f alejada de p_i, definimos el $p-$ésimo cuantil como:

$$Q(p) = (1-f)Q(p_i) + fQ(p_{i+1}).$$

A la función Q se le conoce como la función cuantil.

Si una distribución es simétrica, entonces es natural pensar que se acumula la misma probabilidad en $F(Q_p)$ con $0 < p < \frac{1}{2}$ y $1 - F(Q_{1-p})$; o que la probabilidad entre la mediana y el cuantil Q_p con $0 < p < \frac{1}{2}$ (se encuentra debajo de la media) es la misma que la de la mediana y el cuantil Q_{1-p}.

Entonces en la gráfica de la función de distribución $F(x)$, la distancia en el eje de las abcisas entre la mediana y cualquier cuantil Q_p debajo de la mediana es igual a la distancia entre la mediana y el cuantil Q_{1-p}.

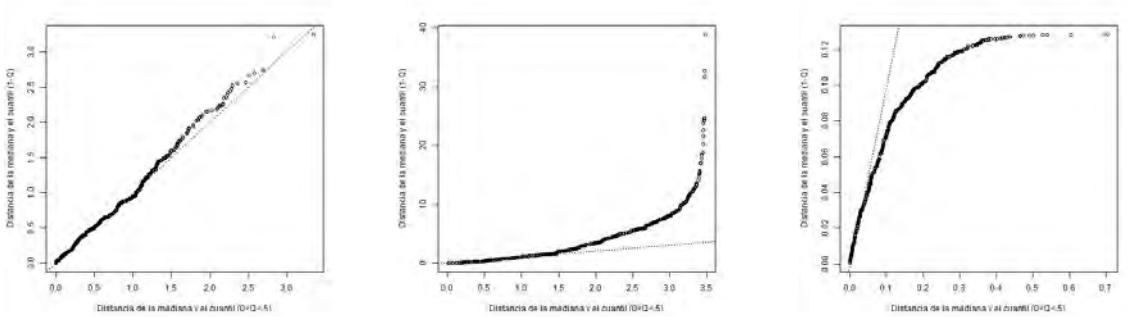
Supongamos que tenemos una colección de observaciones $x_1, \ldots, x_n$, diremos que éstos se distribuyen simétricamente si su función cuantil satisface:

$$Q(.5) - Q(p) = Q(1-p) - Q(.5) \qquad \text{con } 0 < p < .5, \tag{3.2}$$

es decir que el $p-$ésimo cuantil se encuentra a la misma distancia debajo de la mediana como el $(1-p)-$ésimo cuantil se encuentra arriba de ella.

Entonces para verificar la simetría de un conjunto de observaciones, podemos escoger entre las siguientes opciones:

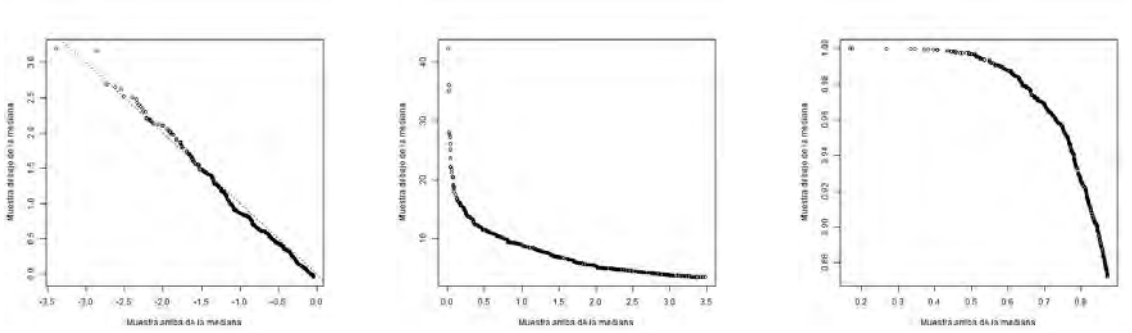
1. Graficar los valores $Q(1-p_1)-Q(.5),\ldots,Q(1-p_{\frac{n}{2}})-Q(.5)$ contra los valores de $Q(.5)-Q(p_1),\ldots,Q(.5)-Q(p_{\frac{n}{2}})$ y si los valores graficados caen en la línea $y=x$, entonces, $x_1,\ldots,x_n$ se distribuyen simétricamente.



(a) *Distribución* $Normal(0,1)$ **(b)** *Distribución* $Exponencial(1/5)$ **(c)** *Distribución* $Beta(5,1)$

Gráfica 3.4: Gráficas sobre la simetría que ilustran el punto 1.

2. Graficar los valores $Q(1-p_1),\ldots,Q(1-p_{\frac{n}{2}})$ contra los valores de $Q(p_1),\ldots,Q(p_{\frac{n}{2}})$ y si los valores graficados caen en la línea $y=-x$, entonces la muestra observada se distribuye simétricamente.

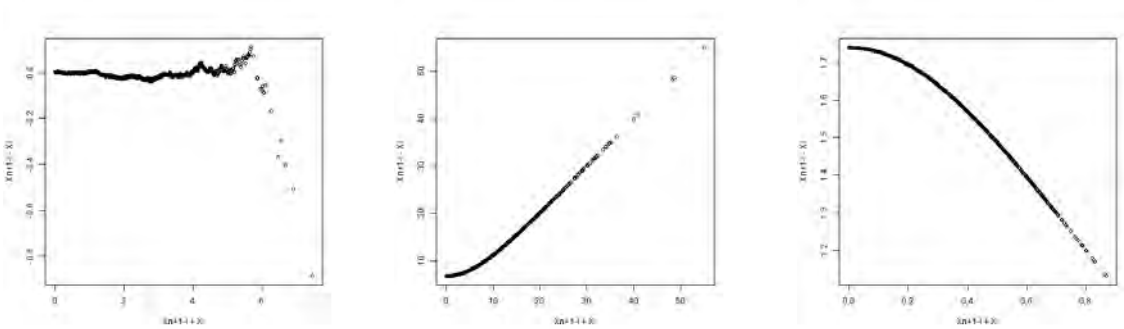


(a) *Distribución* $Normal(0,1)$ **(b)** *Distribución* $Exponencial(1/5)$ **(c)** *Distribución* $Beta(5,1)$

Gráfica 3.5: Gráficas sobre la simetría que ilustran el punto 2.

Lo anterior es equivalente a graficar la mitad superior de las observaciones ordenadas contra la inferior, i.e., $x_{(n+1-i)}$ vs $x_{(i)}$ con $i \leq \frac{n}{2}$; una pendiente negativa que excede la unidad en valor absoluto indica un sesgo positivo, una pendiente negativa menor a la unidad en valor absoluto indica un sesgo negativo.

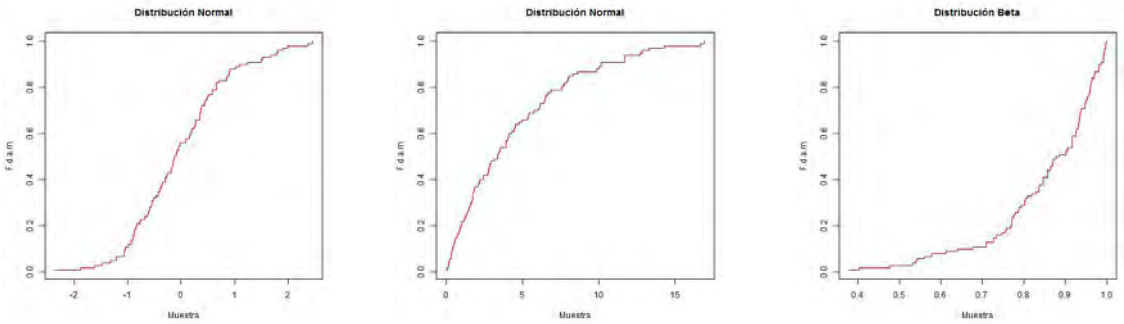
3. Graficar las sumas $x_{(n+1-i)} + x_{(i)}$ contra las diferencias $x_{(n+1-i)} - x_{(i)}$, que producirán una configuración horizontal para una distribución simétrica.



(a) *Distribución* $Normal(0, 1)$ **(b)** *Distribución* $Exponencial(1/5)$ **(c)** *Distribución* $Beta(5, 1)$

Gráfica 3.6: Gráficas sobre la simetría que ilustran el punto 3.

4. Graficar la función de distribución acumulativa muestral contra la muestra, esta gráfica es una gráfica de los cuantiles si se toma a la muestra ordenada como los cuantiles y a la función de distribución empírica como las fracciones que le corresponden a los cuantiles.



(a) *Distribución* $Normal(0, 1)$ **(b)** *Distribución* $Exponencial(1/5)$ **(c)** *Distribución* $Beta(5, 1)$

Gráfica 3.7: Gráficas sobre la simetría que ilustran el punto 4.

La relación (3.2) es reflejada en esta gráfica. Si la distribución tiene un sesgo positivo, la porción de la f.d.a.m. para $\frac{i}{n}$ valores cercanos a uno, será usualmente larga y plana en comparación con el resto de la f.d.a.m. Similarmente, si la distribución tiene un sesgo negativo, la porción larga y plana se encontrará en $\frac{i}{n}$ valores menores que .1 .

Capítulo 4

Evaluación de la cola

En ocasiones el interés no radica en describir la distribución completa de una variable sino en describir una o ambas colas de la distribución. Por ejemplo la *Environmental Protection Agency* se interesa en hacer frecuentemente inferencias y en emitir las normas relativas a las altas concentraciones de diversos contaminantes. En tales situaciones es más importante entender el comportamiento de las colas que si se ajusta a una distribución por completo. Aunque un modelo particular puede describir adecuadamente la mayor parte de la distribución, éste sería inservible para predecir máximos o valores extremos si el modelo falla en los cuantiles superiores. También, un modelo que no es lo suficientemente preciso en una gran parte de los datos, todavía puede ser útil para predecir valores extremos si éste describe adecuadamente el comportamiento de los cuantiles superiores. Bryson (1974)(Véase [2]) presentó una técnica gráfica para tratar con la evaluación del comportamiento de las colas de una distribución. El desarrollo que se presenta a continuación es una adaptación de Curran y Frank (1975)(Véase [4]).

Siendo específicos, nuestro interés recae en evaluar el comportamiento de la cola derecha; matemáticamente, esto es equivalente a evaluar el "espesor" de la cola derecha o encontrar un modelo matemático que "ajuste" a la cola derecha.

Se considera que el modelo matemático más conveniente es la distribución exponencial, ya que se considera la base para discriminar las colas.
Su función de densidad de probabilidad es:

$$f(x) = \lambda exp(-\lambda x), \qquad x > 0, \quad \lambda > 0.$$

Su función de distribución acumulativa es:

$$F(x) = 1 - exp(-\lambda x).$$

De donde, al despejar $-\lambda x$, tenemos:

$$log(1 - F(x)) = -\lambda x. \tag{4.1}$$

La igualdad anterior implica que si $1 - F(x)$ es graficado contra x en papel semilogarítmico, la gráfica resultante será una línea recta.

Por lo anterior, es conveniente usar a la distribución exponencial como la distribución de referencia y comparar otras distribuciones con ella.

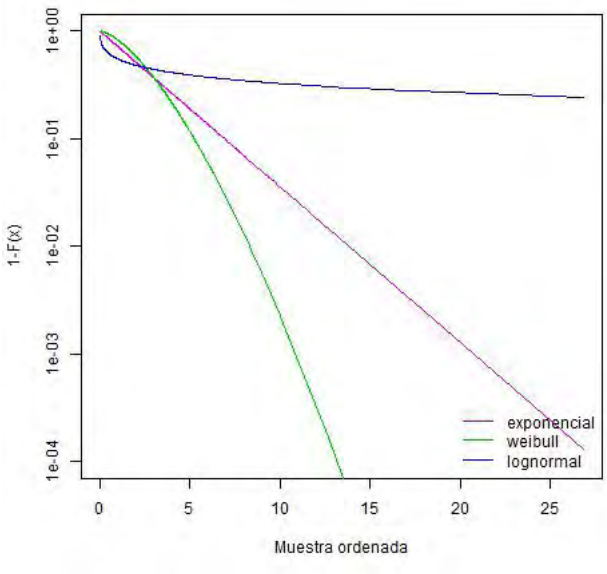
Las distribuciones Weibull y Lognormal son usualmente las dos distribuciones de mayor interés potencial para este tipo de problema.

La Weibull de 2 parámetros tiene como función de densidad de probabilidad y función de distribución acumulativa respectivamente a:

$$f(x) = r\lambda(\lambda x)^{(r-1)}exp(-\lambda x)^r, \qquad \lambda > 0,\ r > 0,\ x > 0.$$
$$F(x) = 1 - exp(-\lambda x)^r.$$

La distribución lognormal tiene como *f.d.p.* y *f.d.a.*:

$$f(x) = \frac{1}{x\sigma(2\pi)^{\frac{1}{2}}}exp\frac{-(logx - \mu)^2}{2\sigma^2}, \qquad \mu \in \mathbb{R},\ \sigma^2 > 0,\ x > 0.$$
$$F(x) = \Phi(\frac{logx - \mu}{\sigma}), \qquad x > 0, \text{ donde } \Phi(.) \text{ es la función de distribución normal estándar.}$$



Gráfica 4.1: *Colas de distintas distribuciones*

Notemos en la gráfica que la distribución exponencial produce una línea recta, la distribución lognormal se curva hacia arriba y la distribución Weibull (con $r > 1$) se curva hacia abajo. Una distribución que se curva hacia abajo se dice que tiene "colas ligeras", la cual tiene una función de densidad de probabilidad cuya cola superior se acerca a cero más rápido que la cola de la exponencial y por lo tanto es menos probable que produzca valores altos.
Por otra parte, aquella que se curva hacia arriba, se dice que tiene "colas pesadas". Una distribución con una cola pesada tiene una función de densidad de probabilidad cuya cola superior se acerca a cero con menos rapidez que la cola de la distribución exponencial, en otras palabras, tiene mayor probabilidad de producir altos valores.

Si éstos son modelos para la cola superior, un análisis de la gráfica de $(1 - F_n(x))$ contra X_i (las observaciones ordenadas) usualmente indicarán cual es el modelo apropiado.

El papel semilogarítmico usado en la gráfica anterior, es un papel de 4 ciclos. Tres cuartos del eje vertical le concierne solo los puntos superiores de 10^{-1} a 10^{-4} de la distribución ($.9 \leq F(x) \leq .9990$).
Dicha gráfica se concentra casi exclusivamente en el 10 % superior de la distribución.

Sin embargo, contiene en el eje vertical al resto de la distribución y graficar toda la distribución puede causar confusiones al tratar de juzgar el ajuste de la cola superior.

En general, los puntos para los cuales $F_n(x) \leq .5$ no deberían ser graficados, debido a que de esta manera se asegura el análisis de la cola superior eliminando la parte de la distribución que no concierne a las colas. Para muestras pequeñas como 100, D'Agostino encuentra conveniente usar el papel semilogarítmico de 2 ciclos, definiendo a $F_n(x)$ como:

$$F_n(x) = \frac{\#(X_j \leq x) - .5}{n} \tag{4.2}$$

y graficar sólo los datos de $F_n(x) \geq .5$.

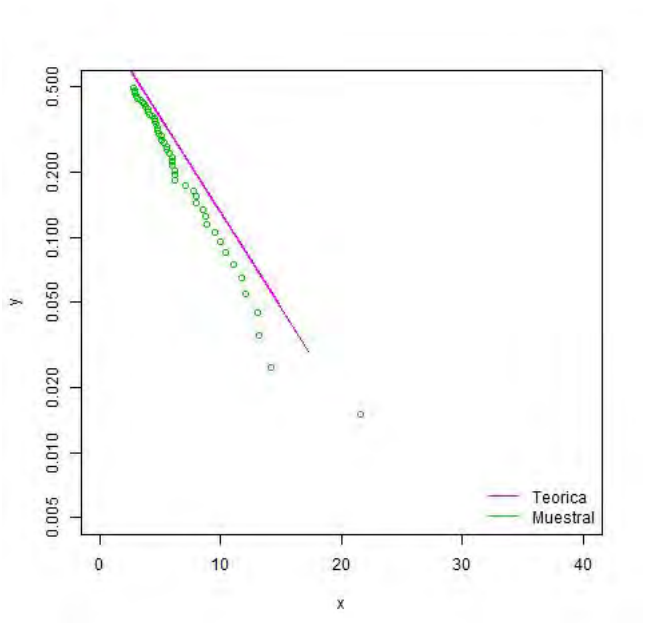
La ecuación (4.2) se define de esa forma, debido a que se realiza un ajuste de corrección por continuidad; consideramos la observación muestral que se encuentra la mitad en el grupo inferior y la mitad en el grupo superior.
Ejemplo: Si $n = 9$, la media muestral es la 5^{a} observación más grande y se considera esta observación en 2 partes, una en la mitad inferior y otra en la mitad superior.
Ejemplo 1:

Consideremos un conjunto de datos observados con distribución exponencial con parámetro $\lambda = \frac{1}{4}$. Los puntos representan los datos observados de $1 - F_n(x)$, con $F_n(x) \geq .5$, los cuales tienden a asemejar una línea recta. Si la distribución exponencial es un modelo adecuado para los datos observados, entonces la línea recta para la distribución exponencial teórica debería ajustarse a los datos observados.

Para obtener la línea recta teórica, se necesita al parámetro λ, el cual puede ser estimado a partir de (4.1), es decir:

$$\frac{-log(1 - F(x))}{x} = \lambda,$$

donde x representa cualquier dato observado. En particular se puede tomar el operador esperanza de los datos observados y la esperanza de los $1 - F(x)$ como valores representativos de dichas variables, y a partir de ahí, ajustar el parámetro λ.

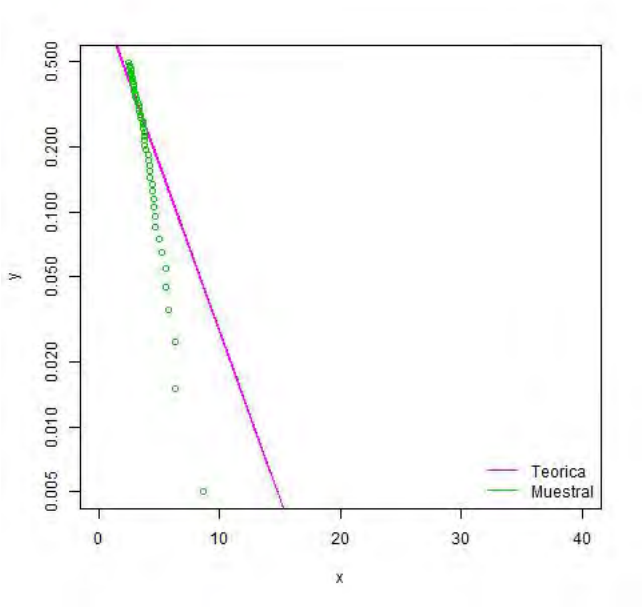


Gráfica 4.2: *Cola de la muestra exponencial con parámetro $\lambda = \frac{1}{4}$ y su respectiva cola teórica.*

Se obtiene en la simulación una $\lambda = 0.2036271$ para la línea teórica y se infiere del ejercicio, que la distribución exponenical es un modelo apropiado que se ajuste de manera adecuada para todos los datos observados, a excepción de los últimos 2. (Véase la Figura 4.2).

Es importante notar que la bondad de ajuste de los datos observados recae en la distancia horizontal de los datos observados a la línea teórica, no en las distancias verticales.

Ejemplo 2:
Consideremos un conjunto de datos observados con distribución $Weibull(2, 3)$. Usando una $\hat{x}$ y $1 - \widehat{F(x)}$ con el método propuesto en el ejemplo anterior, obtenemos una $\lambda = .3588979$, con la cual al graficar la línea recta teórica, ésta se encuentra por arriba de la mayor parte de los datos.



Gráfica 4.3: *Cola de la muestra* $Weibull(2, 3)$ *y su respectiva cola teórica, según el método propuesto en el ejemplo anterior.*

Al usar distintos valores $F_n(x)$, se infiere que el modelo de la distribución exponencial no es apropiado para ajustar a la muestra, y hay que tener bajo consideración que los datos observados tienen una cola más ligera que la distribución exponencial.

Sin embargo, puede ser explicado por el modelo exponencial que incorpora un desplazamiento:

$$f(x) = \lambda exp(-\lambda(x-\theta)), \qquad \lambda > 0, \quad x > 0.$$

$$\begin{aligned} F(x) = \mathbb{P}[X \leq x] &= \int_{\theta}^{x} \lambda \exp(-\lambda(\mathrm{t}-\theta))\,\mathrm{d}t \\ &= -exp(-\lambda(t-\theta))\Big|_{\theta}^{x} \\ &= -exp(-\lambda(x-\theta)) + 1. \end{aligned}$$

Para obtener el valor de $\hat{\lambda}$, llegamos a la expresión:

$$F_n(x) = 1 - exp(-\lambda(x-\theta)), \qquad \text{despejando } x\text{, tenemos:}$$
$$x = \frac{-log(1-F_n(x))}{\lambda} + \theta.$$

La idea es proponer dos valores de x distintos (o dos valores de $F_n(x)$) para producir un sistema de ecuaciones para estimar $\hat{\lambda}$ y $\hat{\theta}$. Las ecuaciones son:

$$\frac{-log(1-F_n(x_1))}{\lambda} + \theta = x_1. \tag{4.3}$$
$$\frac{-log(1-F_n(x_2))}{\lambda} + \theta = x_2. \tag{4.4}$$

Despejando θ de (4.4) y sustituyéndolo en (4.3) tenemos:

$$x_1 = \frac{-log(1-F_n(x_1))}{\lambda} + x_2 + \frac{log(1-F_n(x_2))}{\lambda}$$
$$= \frac{1}{\lambda} log\Big(\frac{1-F_n(x_2)}{1-F_n(x_1)}\Big) + x_2.$$

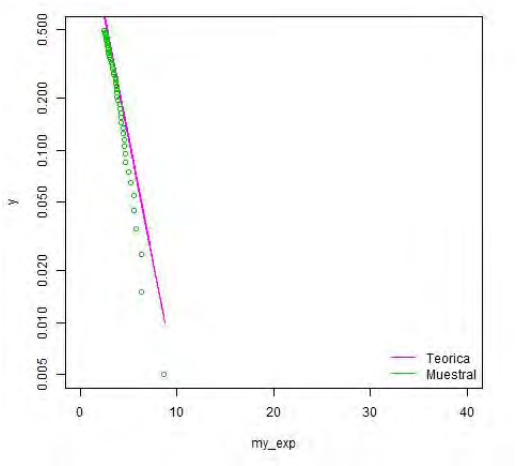
y despejando $\hat{\lambda}$ tenemos:

$$\hat{\lambda} = \frac{log\Big(\frac{1-F_n(x_2)}{1-F_n(x_1)}\Big)}{x_1 - x_2},$$

por lo que:

$$\hat{\theta} = x_2 + \frac{log(1-F_n(x_2))}{\hat{\lambda}}.$$

Entonces, proponiendo el valor $x_2 = 2.7814$ y $F_n(x_2) = .495$, obtenemos $\hat{\lambda} = 0.6502678$ y $\hat{\theta} = 1.700003$.



Gráfica 4.4: *Cola de la muestra $Weibull(2,3)$ y su respectiva cola teórica, según el nuevo método propuesto.*

Entonces, se puede inferir que la distribución exponenical desplazada es un modelo apropiado que se ajusta de manera adecuada para los datos observados.

El análisis realizado en los parámetros anteriores, puede ser fácilmente modificado para analizar la cola inferior de la distribución, graficando $F_n(x)$ y $F(x)$ respectivamente contra las observaciones, en papel semilogarítmico.

Capítulo 5

Evaluación del Ajuste de la distribución completa

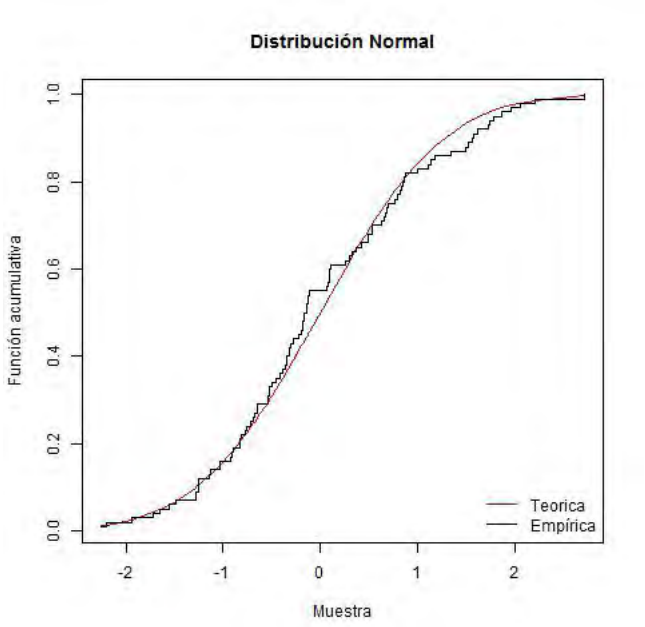
La idea de este capítulo es poder decir que función de distribución se ajusta a la muestra de interés, al compararla (distribucionalmente) con otra muestra, de la cual en principio se sabe que función de distribución tiene, a través de algunas técnicas (gráficas) como una doble gráfica y la gráfica cuantil-cuantil.

5.1. Doble gráfica

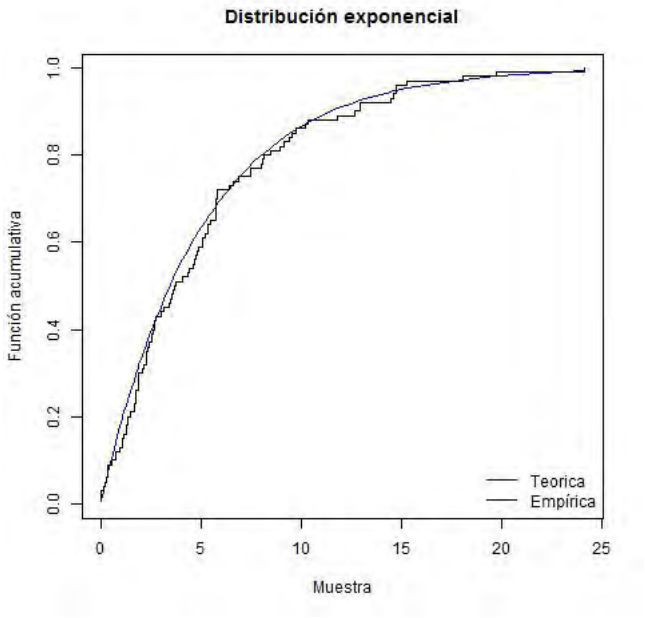
La función de distribución acumulativa empírica puede ser usada para evaluar que tan bien una distribución particular se ajusta al conunto completo de datos. El procedimiento consiste en hacer una doble gráfica, la función de distribución acumulativa empírica y la función de distribución de la distribución hipotética.

Si los valores de los parámetros de la distribución hipotética no se encuentran especificados, éstos tienen que ser estimados a partir del conjunto de datos bajo análisis tal como el método de momentos ó máxima verosimilitud y usarlos para graficar la distribución hipotética.

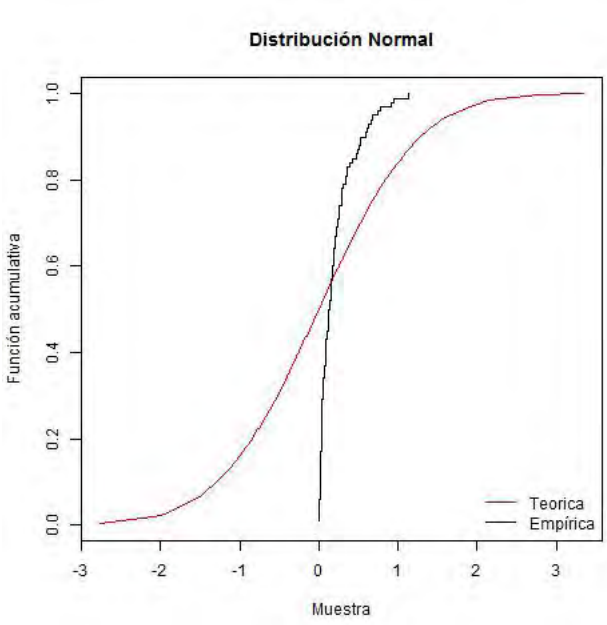
A continuación se presentan algunos ejemplos que ilustran este procedimiento.



Gráfica 5.1: *Muestra y función de distribución teórica de una distribución* $Normal(0, 1)$.



Gráfica 5.2: *Muestra y función de distribución teórica de una distribución* $Exponencial(1/5)$.



Gráfica 5.3: *Muestra proveniente de una distribución* $Normal(0,1)$ *y función de distribución teórica de una distribución* $Exponencial(1/5)$.

Usualmente este procedimiento informal es el primer paso de un análisis más elaborado que incluye técnicas numéricas formales.

5.2. Histograma

Sea $X_1, \ldots, X_n$ una muestra aleatoria, proveniente de F$(\cdot)$.
Sea $(a, b]$ con $a, b \in \mathbb{R}$, un intervalo que contiene a todos los X_i, es decir:

$$a < X_{(1)} < X_{(n)} < b.$$

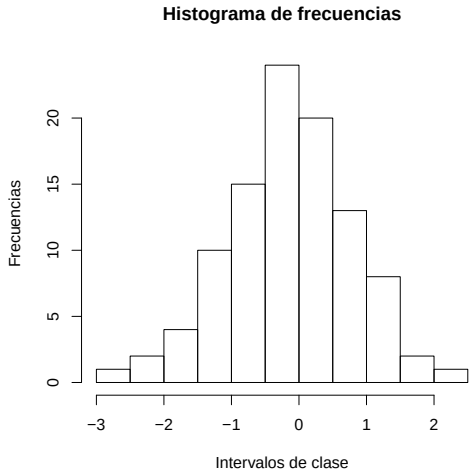
Se realiza una partición de dicho intervalo en un número m de intervalos de "clase" (que no se superponen), cada uno de ellos con un ancho igual a $w = \dfrac{(b-a)}{m}$, por lo que el $k-$ésimo intervalo de clase, J_k, es: $(a + (k-1)w, a + kw]$.
Denotemos a n_k como el número de elementos de la muestra que caen en el intervalo de clase k, es decir:

$$n_k = \sum_{i=1}^{n} \mathbb{1}_{J_k}(X_i).$$

Un histograma es una gráfica en forma de barras, donde las bases de las barras son los intervalos de clase y la altura de cada barra es el número de elementos que caen en cada intervalo de clase.

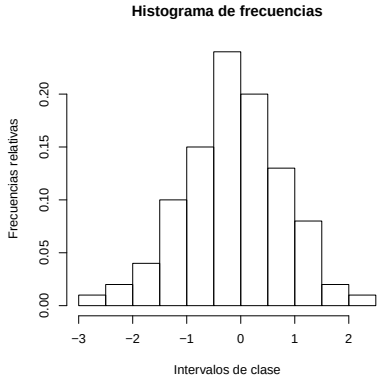
Tipos de Histogramas:

1. Histograma de frecuencias.- La altura de las barras del histograma muestra el número de observaciones que cayeron en el intervalo de clase respectivamente, es decir, sobre cada intervalo J_k, es dibujada una barra vertical de tamaño igual a n_k (frecuencia).



Gráfica 5.4: *Histograma de una muestra de tamaño* 100 *proveniente de una distribución* $Normal(0, 1)$.

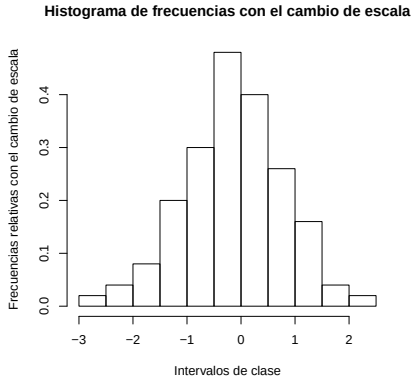
2. Histograma de frecuencias relativas.- Difiere con el histograma de frecuencias solo en que la altura de la barra sobre cada intervalo de clase J_k es n_k/n, donde $n = \sum_{k=1}^{m} n_k$.



Gráfica 5.5: *Histograma de frecuencias relativas de una muestra de tamaño* 100 *proveniente de una distribución* $Normal(0, 1)$.

Si redefinimos la altura de las barras de tal forma de que el área de la barra sobre J_k es la frecuencia relativa n_k/n, tendríamos como consecuencia que el área total dentro de las barras del histograma es 1. Entonces, modificando la altura de las

barras, de esta forma, las alturas serán: $J_k = \frac{n_k}{wn}$ entonces el área de la barra para cada k será $w\frac{n_k}{wn} = \frac{n_k}{n}$.



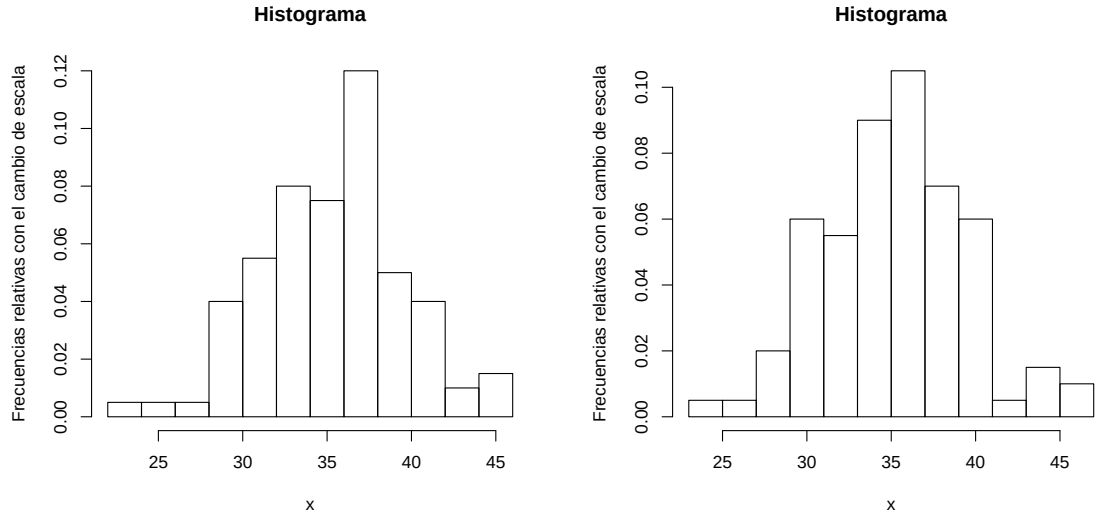
Gráfica 5.6: *Histograma de frecuencias con el cambio de escala, de una muestra de tamaño* 100 *proveniente de una distribución* $Normal(0, 1)$.

Como consecuencia, podemos definir la siguiente función:

$$f_h = \begin{cases} \dfrac{n_k}{nw} & \text{si } x \in J_k, \\ 0 & \text{si } x \notin (a, b], \end{cases}$$

la cual es una función de densidad en el histograma. (Los valores de una variable aleatoria X con función de densidad f_h, caerán en el intervalo J_k con probabilidad n_k/n y el valor de X es generado uniformemente entre $a + (k-1)w$ y $a + kw$).

La ventaja de los histogramas es que proveen una forma de ver la distribución general de un conjunto de valores, sin embargo, la forma del histograma depende del conjunto particular de intervalos de clase que fueron escogidos para éste.



Gráfica 5.7: Gráficas donde se ilustra la dependencia del histograma sobre los intervalos de clase que son elegidos, a partir de una muestra de tamaño 100 proveniente de una distribución $Normal(35, 4)$.

Cuando el tamaño de la muestra es muy grande y cuando el número de las barras que componen el histograma es grande (aunque dicho número es menor que el tamaño de muestra), esta función de densidad es muy parecida a la función de densidad verdadera (que es generalmente desconocida) de la v.a. X.

Supongamos que la verdadera función de densidad de una muestra aleatoria $X_1, \ldots, X_n$ para un número grande n es f y $w = \frac{(b-a)}{m}$ es pequeño; notemos que el área bajo la verdadera función de densidad f sobre J_k es:

$$\int_{a+(k-1)w}^{a+kw} f(x)\,\mathrm{d}x.$$

Las áreas sobre J_k bajo el histograma con el cambio de escala y bajo la densidad verdadera son necesariamente muy cercanas con alta probabilidad (cuando w es pequeña y n es grande), ésto se debe a la *Ley fuerte de los Grandes Números.*

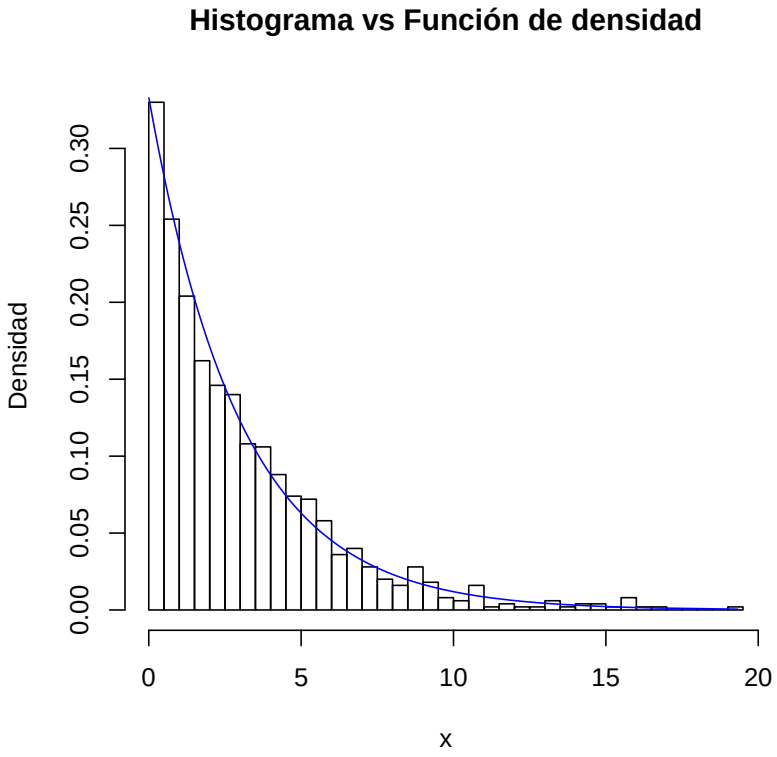
Teorema. *Ley fuerte de los grandes números.*
Sea $X_1, \ldots, X_n$ una muestra aleatoria con media μ. Entonces:

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{c.s.} \mu.$$

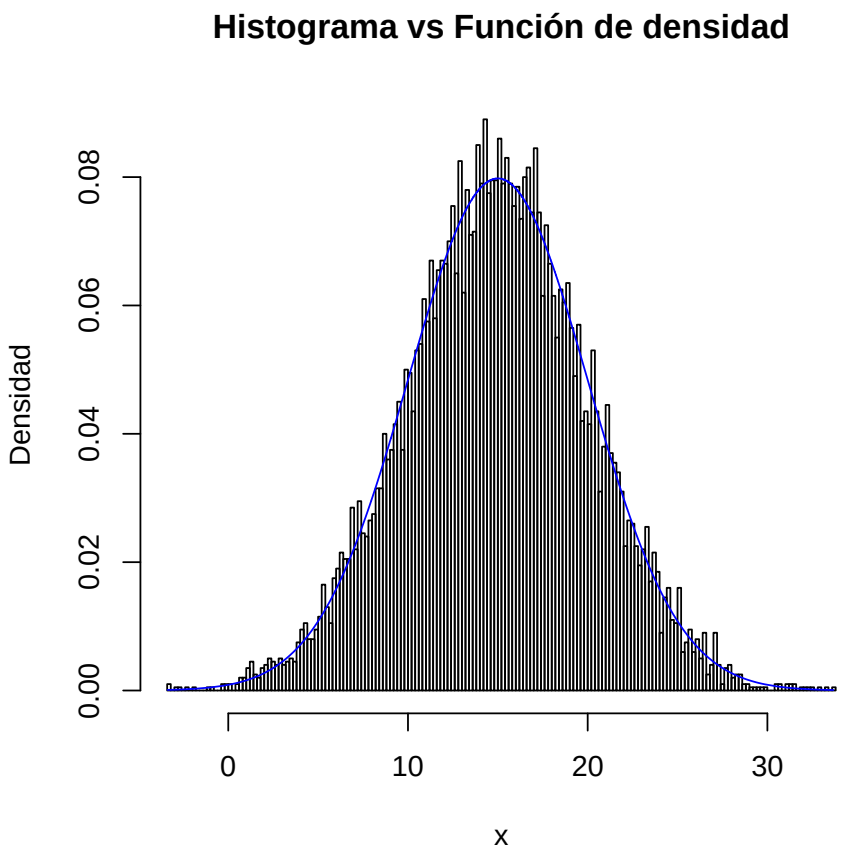
Si las X_i son funciones indicadoras de algún evento A bajo repeticiones independientes de un experimento aleatorio, entonces, $\frac{\sum_{i=1}^{n} X_i}{n}$ es la frecuencia relativa de A en las primeras n repeticiones, y por la ley fuerte de los grandes números tenemos que

con probabilidad 1, las frecuencias relativas convergen a la probabilidad de A ($\mathbb{P}[A]$), por ser una suma de variables aleatorias independientes con distribución Bernoulli.

En nuestro caso, el evento A es $\{X_i \in J_k\}$, y dado que la frecuencia relativa es n_k/n (el área de la barra en el histograma con el cambio de escala), ésta converge con probabilidad 1 a $\mathbb{P}[X_i \in J_k] = \int_{a+(k-1)w}^{a+kw} f(x)\,\mathrm{d}x$ (el área bajo la verdadera función de densidad).



Gráfica 5.8: *Histograma de frecuencias relativas escaladas, con* 35 *intervalos de clase de una muestra de tamaño* 100 *proveniente de una distribución Exponencial*(1/3) *con la función de densidad verdadera sobrepuesta.*



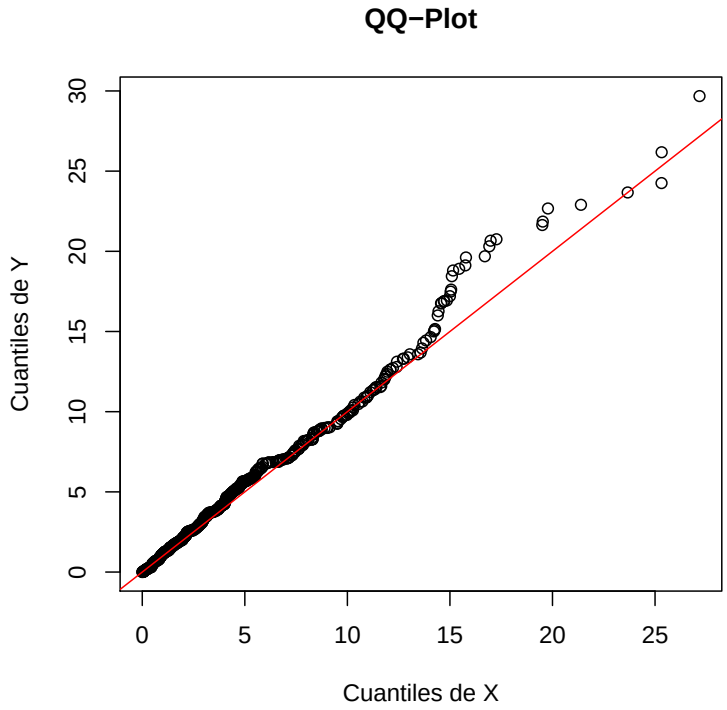
Gráfica 5.9: *Histograma de frecuencias relativas con cambio de escala, con* 200 *intervalos de clase de una muestra de tamaño* 10000 *proveniente de una distribución* $Normal(15, \sqrt{5})$ *con la función de densidad verdadera sobrepuesta.*

5.3. Gráficas Cuantil-Cuantil (QQ-Plots)

Si X y Y son variables aleatorias idénticamente distribuidas, entonces, éstas tendrán los mismos cuantiles, por lo que al graficar los cuantiles de X contra los cuantiles de Y, ésta producirá la linea recta $y = x$; a esta gráfica se le llama Gráfica Cuantil-Cuantil o QQ-Plots.

Entonces, si se tienen 2 muestras, $x_1, \ldots, x_n$ (la muestra de interés) y $y_1, \ldots, y_m$ (muestra auxiliar) y suponemos que éstas tienen la misma distribución, además, si se obtienen las fracciones muestrales p_i (Véase el concepto relacionado a la ecuación (3.1)) tomando como referencia aquel tamaño de muestra cuyo valor sea el más grande entre ambas muestras, y posteriormente, si con estas fracciones muestrales se calculan los cuantiles correspondientes a cada muestra para después al graficar los cuantiles asociados a la segunda muestra contra los cuantiles asociados a la primera

muestra, éstos caerán aproximadamente sobre la linea recta $y = x$.

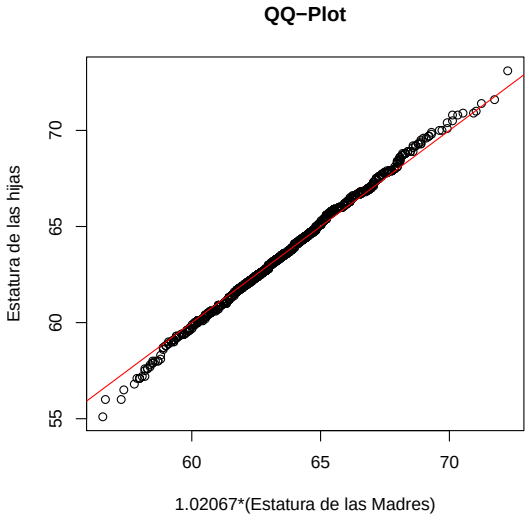


Gráfica 5.10: *QQ-Plot de una muestra Exponencial con* $\lambda = 1/5$ *de tamaño* 500.

En caso de conocer la distribución de la que proviene la muestra $y_1, \ldots, y_m$, sabremos que dicha distibución será la distribución de la muestra $x_1, \ldots, x_n$.

Una propiedad interesante de las QQ-Plots es que si Y es una función lineal de X, entonces la correspondiente QQ-Plot producirá una línea recta pero con pendiente y/o ordenada al origen distinta. Si caen sobre una recta con pendiente de 45 grados pero con distinta ordenada al origen, tendrán un traslado en el parámetro de localización; si cambia la pendiente cambiará en la desviación estándar.

(a) *QQ-Plot de las Estaturas de las hijas contra las Estaturas de las Madres, se observa que no caen los cuantiles en la recta* $y = x$.

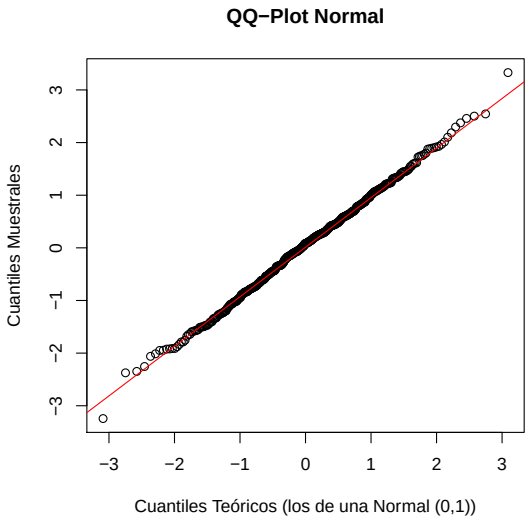


(b) *QQ-Plot de las Estaturas de las hijas contra las Estaturas de las Madres, se observa que caen los cuantiles en la recta* $y = x$, *además Estatura_hijas=* $1.02067\mathbf{x}$*Estatura_Madres.*

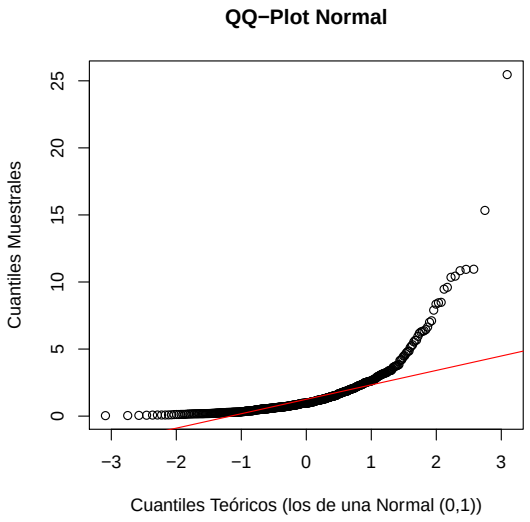
Gráfica 5.11: Gráfica donde se ilustra la invarianza lineal de las QQ-Plots, usando una muestra de estaturas de hijas y otra de estaturas de Madres.

Las principales ventajas que presenta este tipo de gráfica son que el ser humano puede detectar configuraciones lineales fácilmente y la propiedad de invarianza lineal que tiene, sin embargo, el costo de tener estas ventajas es que se necesita un mayor número de observaciones que cuando se hacen comparaciones un poco más simples.

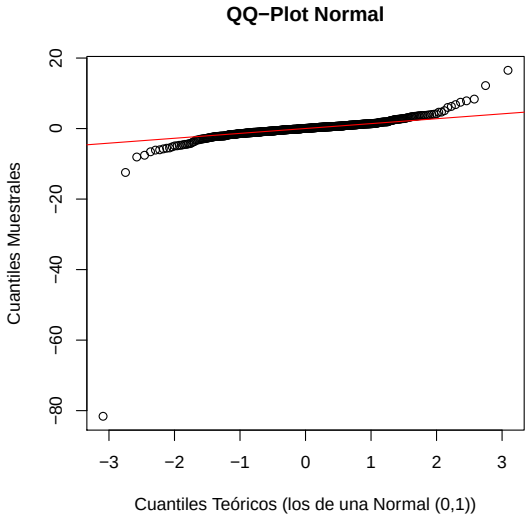
Usualmente, se toma como referencia la QQ-Plot Normal (aquella que usa como muestra auxiliar a una muestra que tiene función de distribución $Normal(0, 1)$), ya que al ser la distribución $Normal(0, 1)$ una distribución simétrica, permite observar el comportamiento de las colas (si son "ligeras" o "pesadas") así como el sesgo de la distribución de la muestra de interés, ya que las QQ-Plots tienden a enfatizar la estrucutra comparativa de la gráfica en las colas y a difuminar las distinciones en el centro de la gráfica (donde las probabilidades asociadas a los cuantiles son altas).



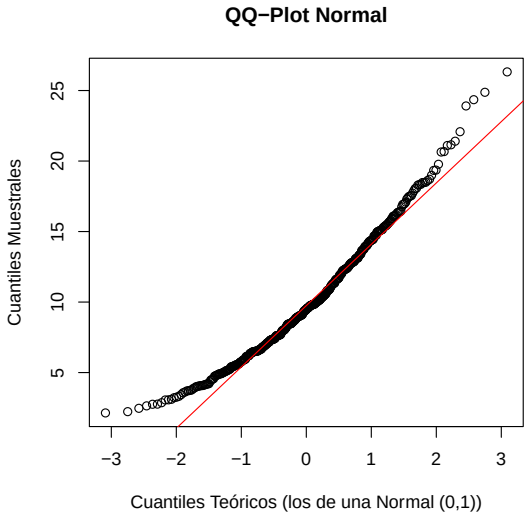
(a) *QQ-Norm de una muestra de una distribución* $Normal(0, 1)$.



(b) *QQ-Norm de una muestra de una distribución* $Lognormal(0, 1)$.



(c) *QQ-Norm de una muestra de una distribución* $t - Student$ *con* 2 *grados de libertad.*



(d) *QQ-Norm de una muestra de una distribución* $\chi^2_{(10)}$.

Gráfica 5.12: Gráfica donde se ilustran distintos tipos de QQ-Norm.

Capítulo 6

Gráficas de Probabilidad

El problema de usar la función de distribución muestral al tratar de juzgar visualmente la exactitud de una distribución hipotética específica se debe a la curvatura de las gráficas de la función de distribución muestral y función de distribución, ya que usualmente es muy difícil juzgar visualmente la cercanía de la gráfica de la función escalonada con la gráfica de la función de distribución.

Si se trata de tomar una decisión basada en una inspección visual, probablemente es más fácil juzgar si un conjunto de observaciones se desvían de la línea recta.

Una gráfica de probabilidad, es una gráfica de los datos observados que producirá una línea recta si la distribución hipotética es la verdadera distribución subyacente. La línea recta es obtenida de la transformación de la escala vertical de la gráfica de la función de distribución muestral a una escala que producirá exactamente una línea recta. Los parámetros de la distribución teórica, pueden ser estimados por la línea recta ajustada.

Recordemos que la gráfica de la función de distribución muestral está basada en graficar $F(x)$ contra x; para datos muestrales $F(x)$ es reemplazada por $F_n(x)$ y los valores de x graficados son los valores observados de la v.a. X.

En la gráficas de probabilidad, la función de distribución muestral $F_n(x)$ o $F_n(x_i)$ es usualmente definida como:

$$F_n(x_i) = p_i = \frac{i - .5}{n}, \qquad i = 1, \ldots, n$$

o

$$F_n(x_i) = p_i = \frac{i - c}{n - 2c + 1}, \qquad i = 1, \ldots, n, \quad \text{con} \;\; 0 \leq c \leq 1.$$

(Cuando $c = .5$, obtenemos la primera ecuación).

Para construir una gráfica de probabilidad necesitamos:

1. Elegir una distribución teórica F para la muestra.
2. Estimar la función de distribución teórica.
3. Graficar la función de distribución estimada o una función de ésta contra x
   (Si F es la verdadera función de distribución, la gráfica de probabilidad debería ser aproximadamente una línea recta).
4. Ajustar una línea recta a través de los puntos. La línea recta debe ser escogida de manera que se ajuste a la mayoría de los datos, pudiendo ignorar outliers o datos de validez dubitativa.
   (Los parámetros de la distribución teórica pueden ser estimados por la línea recta ajustada).

6.1. Relación lineal de algunas distribuciones

La relación lineal de algunas distribuciones es:

1. **Logística**

$$F(x) = [1 + exp\{\frac{-\pi(x-\mu)}{\sigma 3^{(1/2)}}\}]^{(-1)},$$

entonces, despejando a x, tenemos:

$$x = \sigma log\Big(\frac{1-F(x)}{F(x)}\Big)\frac{3^{(1/2)}}{\pi} + \mu.$$

La relación es lineal entre x y $log\Big(\frac{1-F(x)}{F(x)}\Big)\frac{3^{(1/2)}}{\pi}$.

Por lo tanto, la gráfica de probabilidad se realiza graficando $log\Big(\frac{1-p_i}{p_i}\Big)\frac{3^{(1/2)}}{\pi}$ contra x_i.

2. **Normal**

$F(x) = \Phi\Big(\frac{x-\mu}{\sigma}\Big)$, donde Φ es la función de distribución acumulativa $N(0,1)$.

Despejando a x, tenemos:

$$x = \mu + \sigma\Phi^{-1}\Big(F(x)\Big).$$

La relación es lineal entre x y $\Phi^{-1}\Big(F(x)\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $\Phi^{-1}\Big(p_i\Big)$ contra x_i.

3. **Lognormal**

$$F(x) = \Phi\Big(\frac{logx-\mu}{\sigma}\Big).$$

Despejando a $logx$, tenemos:

$$logx = \mu + \sigma\Phi^{-1}\Big(F(x)\Big).$$

La relación es lineal entre $logx$ y $\Phi^{-1}\Big(F(x)\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $\Phi^{-1}\Big(p_i\Big)$ contra $logx_i$.

4. **Valor Extremo (Gumbel)**

$$F(x) = 1 - exp\{-exp\Big(\frac{x-\mu}{\sigma}\Big)\},$$

despejando a x, tenemos:

$$x = \mu + \sigma log\Big(- log(1 - F(x))\Big).$$

La relación es lineal entre x y $log\Big(- log(1 - F(x))\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $log\Big(-log(1-p_i)\Big)$ contra x_i.

5. **Weibull**

$$F(t) = 1 - exp\{-\Big(\frac{t}{\theta}\Big)^{k}\},$$

despejando a $logt$, tenemos:

$$logt = log\theta + \frac{log\Big(- log\Big(1 - F(t)\Big)\Big)}{k}.$$

La relación es lineal entre $logt$ y $log\Big(- log\Big(1 - F(t)\Big)\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $log\Big(-log\Big(1-p_i\Big)\Big)$ contra $logx_i$.

6. **Uniforme**

$$F(x) = \frac{x-a}{b-a},$$

despejando a x, tenemos:

$$x = (b-a)F(x) + a.$$

La relación es lineal entre x y $F(x)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar p_i contra x_i.

7. **Exponencial**

$$F(x) = 1 - exp\{-\Big(\frac{x}{\theta}\Big)\},$$

despejando a x, tenemos:

$$x = -log\Big(1 - F(x)\Big)\theta.$$

La relación es lineal entre x y $-log\Big(1 - F(x)\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $-log\Big(1 - p_i\Big)$ contra x_i.

8. **Cauchy**

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctan\Big(\frac{x - \mu}{\sigma}\Big),$$

despejando a x, tenemos:

$$x = tan\Big(\pi\Big(F(x) - \frac{1}{2}\Big)\Big)\sigma + \mu.$$

La relación es lineal entre x y $tan\Big(\pi\Big(F(x) - \frac{1}{2}\Big)\Big)$.

Por lo tanto, la gráfica de probabilidad se obtiene al graficar $tan\Big(\pi\Big(p_i - \frac{1}{2}\Big)\Big)$ contra x_i.

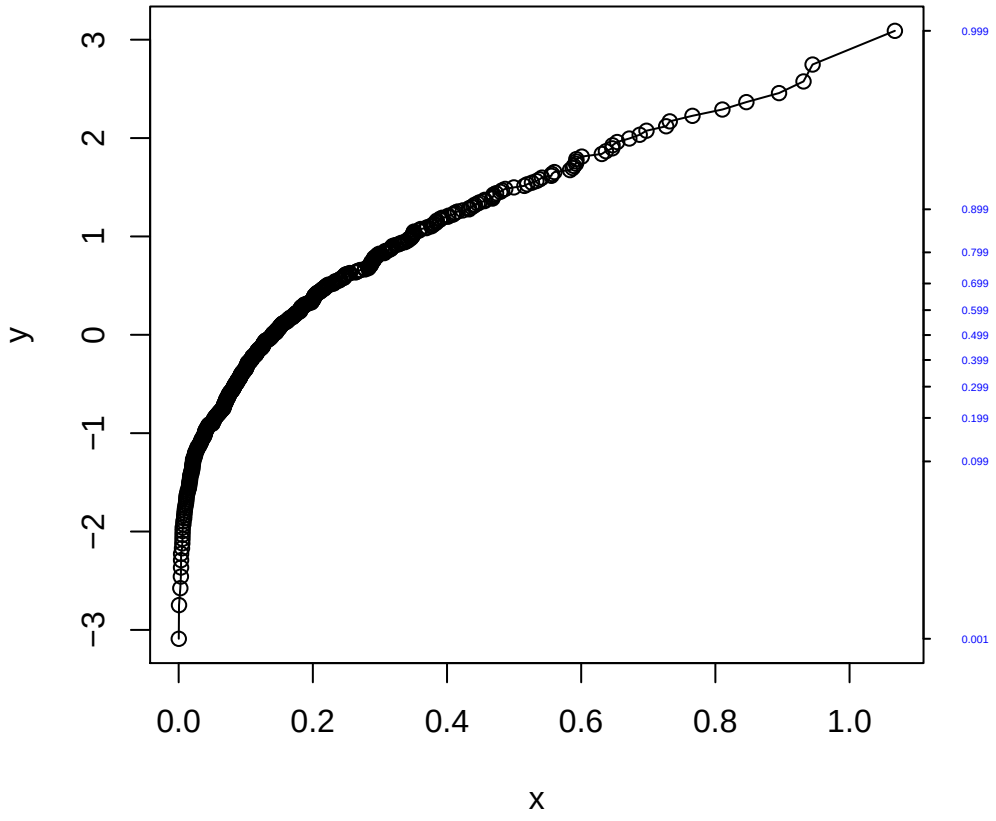
6.2. Formas de etiquetar a las ordenadas en la gráfica de probabilidad

Hay 2 formas de etiquetar el eje vertical:

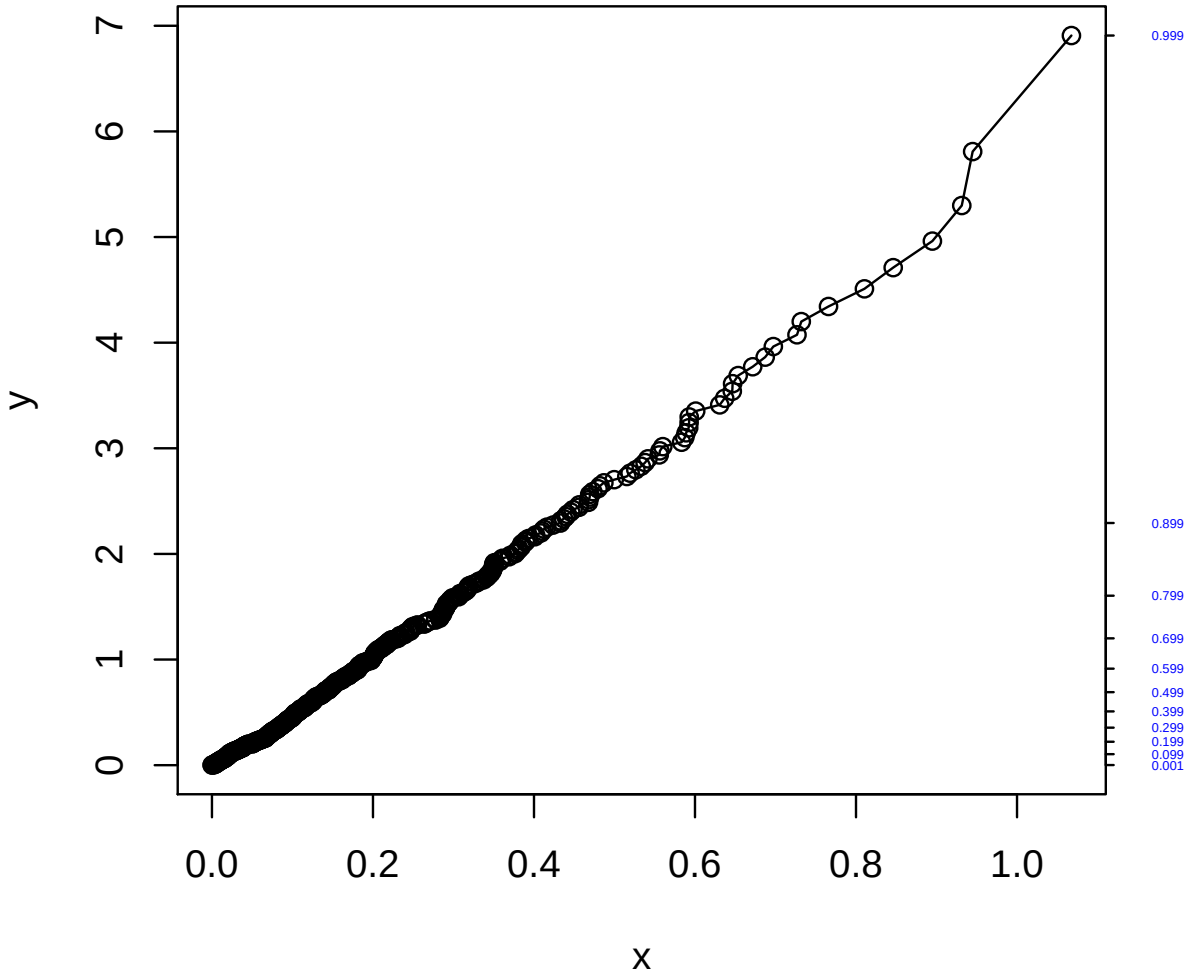
1. El primero, que quizás es el más informativo, es, etiquetar el eje en términos de $F_n(x)$ o $100F_n(x)$ el cual es el método más convencional.

2. En términos de los valores de la variable estandarizada z.

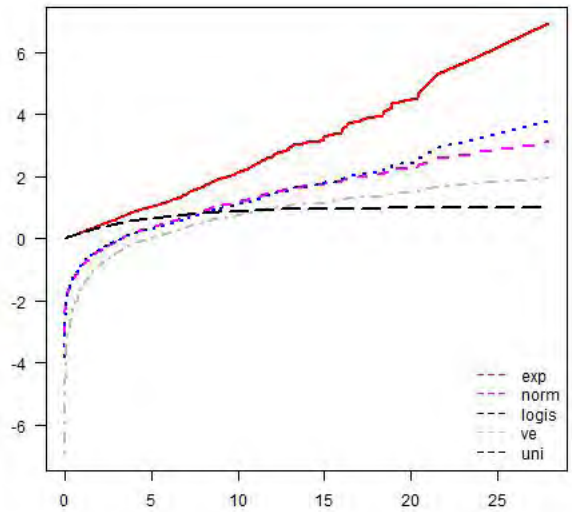
A continuación se presentan algunos ejemplos que ilustran el concepto de Gráfica de Probabilidad.



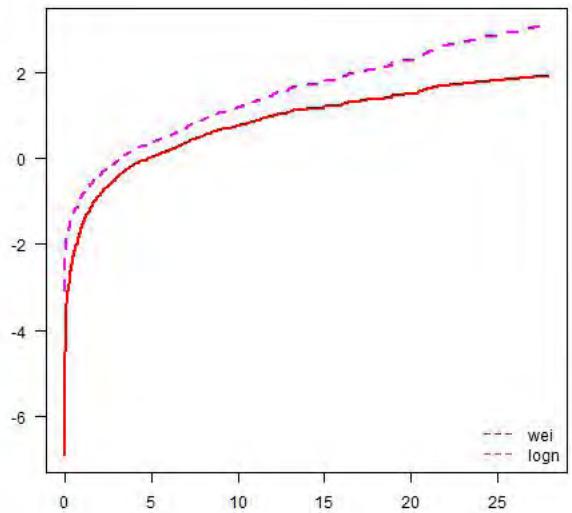
Gráfica 6.1: *Gráfica de probabilidad Normal de una muestra con distribución Exponencial*(5) *de tamaño* $n = 500$.



Gráfica 6.2: *Gráfica de probabilidad Exponencial de una muestra con distribución Exponencial*(5) *de tamaño* $n = 500$.



Gráfica 6.3: *Muestra Exponencial*$(1/5)$ *de tamaño* $n = 500$ *sometida a varias gráficas de probabilidad.*



Gráfica 6.4: *Muestra Exponencial*$(1/5)$ *de tamaño* $n = 500$ *sometida a varias gráficas de probabilidad en papel semilogarítmico.*

Se observa que al someter una muestra a la prueba gráfica, si ésta sigue la distribución correspondiente, se obtendrá una línea recta.

6.3. Medianas

La gráfica de probabilidad, puede ser usada para obtener el valor de los parámetros de la distribución si la forma de la distribución de la muestra es conocida o para obtener una impresión visual para afirmar si la muestra sigue una distribución.

La dificultad recae en que para cada punto $\Big(x_i, F_n(x_i)\Big)$, (resultando exactamente una línea recta), solamente la coordenada x_i es conocida, por lo que es usada una estimación de $F(x_i)$ $\Big(F(x_i) = p(i)\Big)$.

Hemos considerado diferentes estimaciones para $F(x_i)$, que solo dependen del orden de la muestra.

Consideremos:

- $p(i) = \dfrac{i}{n}$ que tiene la desventaja de que $p(n) = 1$, por lo que el punto $\Big(x_n, p(n)\Big)$ cae fuera de la gráfica de probabilidad.

- $p(i) = \dfrac{i-1}{n}$ que tiene la desventaja de que $p(1) = 0$, por lo que el punto $\Big(x_1, p(1)\Big)$ cae fuera de la gráfica de probabilidad.

Dichas objeciones pueden solucionarse usando:

$$p(i) = \frac{i - .5}{n}$$

o

$$p(i) = \frac{i}{n+1}. \tag{6.1}$$

Como $f_{X(i)} = \dfrac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i}$, entonces:
$X_{(i)} \sim Beta\Big(i, n-i+1\Big)$, entonces $\mathbb{E}[X_{(i)}] = \dfrac{i}{n+1}$, de donde se obtiene la aproximación de (6.1).

La mayoría de las estimaciones para $F(x_i)$, pueden ser expresadas de la forma:

$$p(i) = \frac{i-a}{n+b}, \qquad \text{donde a y b son funciones de } i \text{ y } n \text{ respectivamente.}$$

Por otra parte, sea $X_{(1)}, \ldots, X_{(n)}$ una muestra ordenada,
entonces:

$$F_{X_{(i)}}(x) = \mathbb{P}[X_{(i)} \leq x]. \tag{6.2}$$

El evento $\{X_{(i)} \leq x\}$ ocurre cuando ocurren i o más de los eventos $\{X_k \leq x\}$, con $k = 1, \ldots, n$. El número de ocurrencias de los eventos $\{X_k \leq x\}$, con $k = 1, \ldots, n$, tiene distribución Binomial con parámetros n y $p = F(x)$, ya que si tenemos: $Y_1, \ldots, Y_n$ v.a.i. con distribución Bernoulli, donde:

$$Y_k = \mathbb{1}_{(-\infty, x]}(X_k) = \begin{cases} 1 & \text{si } X_k \in (-\infty, x), \\ 0 & \text{si } X_k \notin (-\infty, x). \end{cases}$$

La probabilidad de éxito para cada Y_k es: $\mathbb{P}[X_k \leq x] = F(x)$, asociado con el valor 1.

Entonces, $\sum_{k=1}^{n} Y_k \sim Bin\Big(n, F(x)\Big)$. Por lo tanto:

$$\begin{aligned} F_{X_{(i)}}(x) &= \mathbb{P}[Y_1 + \ldots + Y_n \geq i] \\ &= \sum_{k=i}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}. \end{aligned}$$

Sea $y = F(x)$,

$$G_i(y) = G_{n,i}(y) = \sum_{k=i}^{n} \binom{n}{k} [y]^k [1 - y]^{n-k}.$$

Observemos que $G_i(y)$ es el residuo después de r términos de la serie de Taylor de la distribución Binomial.

Podemos ver la expresión de $G_i(y)$ como la función beta incompleta regularizada:

$$\begin{aligned} G_i(y) &= I_y(i, n - i + 1) \\ &= \frac{n!}{(i-1)!(n-i)!} \int_0^y u^{i-1}(1-u)^{n-i} \, \mathrm{d}u. \end{aligned}$$

Veamos la relación de la distribución Binomial con la distribución Beta incompleta. De la relación de Taylor con el residuo de forma integral:

$$f(a+h) = \sum_{j=0}^{r-1} \frac{h^j}{j!} f^{(j)}(a) + \int_0^1 \frac{h^r (1-t)^{r-1}}{(r-1)!} f^{(r)}(a+th) \, \mathrm{d}t.$$

Veamos la distribución Binomial

$$\begin{aligned} \text{Sea } a &= q, \\ h &= p, \\ f(x) &= x^n, \qquad\qquad \text{con } n \text{ entero positivo,} \end{aligned}$$

entonces:

$$\begin{aligned} (q+p)^n &= \sum_{j=0}^{r-1} \frac{p^j}{j!} n_{[j]} q^{n-j} + R_r \\ &= \sum_{j=0}^{r-1} \binom{n}{j} \frac{p^j}{q}^{n-j} + R_r, \end{aligned}$$

donde:

$$\begin{aligned} x_{[n]} &= x(x-1)\dots(x-n+1), \\ R_r &= \int_0^1 \frac{p^r(1-t)^{r-1}}{(r-1)!} n_{[r]}(q+tp)^{n-r}\,\mathrm{d}t. \end{aligned}$$

Sea

$$\begin{aligned} t &= 1 - \frac{x}{p}, \\ \mathrm{d}t &= 1 - \frac{-1}{p}\mathrm{d}x, \end{aligned}$$

entonces:

$$\begin{aligned}
R_r &= \int_p^0 \frac{p^r \left(\frac{x}{p}\right)^{r-1}}{(r-1)!} n_{[r]}(1-x)^{n-r}\left(\frac{-1}{p}\right)\mathrm{d}x \\
&= \int_p^0 \frac{p\,(x)^{r-1}}{(r-1)!}\frac{n!}{(n-r)!}(1-x)^{n-r}\left(\frac{-1}{p}\right)\mathrm{d}x \\
&= \int_p^0 -\left(\frac{(x)^{r-1}}{(r-1)!}\frac{n!}{(n-r)!}(1-x)^{n-r}\right)\mathrm{d}x \\
&= \int_0^p \frac{(x)^{r-1}}{(r-1)!}\frac{n!}{(n-r)!}(1-x)^{n-r}\,\mathrm{d}x \\
&= \frac{n!}{(r-1)(n-r)!}\int_0^p (x)^{r-1}\,(1-x)^{n-r}\,\mathrm{d}x \\
&= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)}B_p(r,n-r+1) \\
&= \frac{B_p(r,n-r+1)}{B(r,n-r+1)} \\
&= I_p(r,n-r+1) \\
&= I_q(n-r+1,r).
\end{aligned}$$

Análogamente, el residuo después de $r-1$ términos es: $I_p(r-1,n-r+2)$ Entonces el r-ésimo término es:

$$I_p(r,n-r+1)+I_p(r-1,n-r+2) = I_q(n-r+1,r)-I_q(n-r+2,r-1).$$

Observemos que:

$$\begin{aligned}
G_i(y) &= \sum_{k=i}^{n}\binom{n}{k}y^k(1-y)^{n-k} \\
&= 1-\sum_{k=0}^{i-1}\binom{n}{k}y^k(1-y)^{n-k} \\
&= 1-\sum_{k=n-i+1}^{n}\binom{n}{k}y^{n-k}(1-y)^k = 1-G_{n+1-i}(1-y).
\end{aligned}$$

Lo cual nos dice que si las observaciones no se encuentran ordenadas de manera ascendente, sino descendente, entonces, i será reemplazado por $(n+1-i)$, y por $1-y$ y G por $1-G$.

Entonces es claro afirmar que la relación simétrica existe en:

$$p(i) = 1-p(n+1-i).$$

Una aplicación del resultado anterior es la siguiente:

Proposición 1.

$$I_x(a,b) = 1 - I_{1-x}(b,a).$$

Demostración.

Notemos que:

$$G_i(y) = 1 - G_{n+1-i}(1-y)$$

y

$$\begin{aligned} G_i(y) &= I_y(i, n, n-i+1), \\ G_{n+1-i}(1-y) &= I_{1-y}(n-i+1, i), \end{aligned}$$

entonces:

$$I_y(i, n-i+1) = 1 - I_{1-y}(n-i+1, i).$$

$$\text{Por lo tanto,} I_x(a,b) = 1 - I_{1-x}(b,a).$$

□

Si y_i^* es la mediana de $y_{(i)} = F(x_{(i)})$,
de modo que:

$$\begin{aligned} G_i(y_i^*) &= \frac{1}{2}, \qquad (6.3) \\ 1 - G_{n+1-i}(1-y_i^*) &= \frac{1}{2} \qquad (1-y_i^* \text{ es la mediana de } y_{(n+1-i)}). \end{aligned}$$

Escribiendo y_i* como:

$$y_i* = \frac{i-a}{n+b},$$

entonces, tenemos:

$$\text{Med}(y_{(n+1-i)}) = 1 - y_i* = 1 - \frac{i-a}{n+b}. \qquad (6.4)$$

$$\text{Med}(y_{(n+1-i)}) = \frac{n+1-i-a}{n+b}. \qquad (6.5)$$

Igualando (6.4) y (6.5), tenemos:

$$\frac{n+b-i-a}{n+b} = \frac{n+1-i-a}{n+b},$$

despejando a b, tenemos:

$$\begin{aligned} b &= n+1-i-a-n+i-a \\ &= 1-2a. \end{aligned}$$

Entonces

$$y_i^* = \frac{i-a}{n+b} = \frac{i-a}{n+1-2a}.$$

Entonces reescribimos a (6.3) como:

$$G_i(y_i^*) = \sum_{k=i}^{n} \binom{n}{k} \left(\frac{i-a}{n+1-2a}\right)^k \left(1-\frac{i-a}{n+1-2a}\right)^{n-k} = \frac{1}{2}$$

o

$$G_i(y_i^*) = \sum_{k=0}^{i-1} \binom{n}{k} \left(\frac{i-a}{n+1-2a}\right)^k \left(1-\frac{i-a}{n+1-2a}\right)^{n-k} = \frac{1}{2}. \tag{6.6}$$

Buscamos una constante que aproxime a a.
Por ello tomamos el $\lim_{n\to\infty}$ en (6.6):

$$\lim_{n\to\infty} G_i(y_i^*) = \sum_{k=0}^{i-1} \frac{n!}{(n-k)!k!} \frac{(i-a)^k}{(n+1-2a)^k} \left(1-\frac{i-a}{n+1-2a}\right)^{n-k} = \frac{1}{2}. \tag{6.7}$$

Lema. *Sea $a_1, a_2, \ldots$ una sucesión de números que convergen a a, i.e.* $\lim_{n\to\infty} a_n = a$, *entonces:*

$$\lim_{n\to\infty} \left(1+\frac{a_n}{n}\right)^n = e^a.$$

Observemos que:

$$\lim_{n\to\infty} \left(1-\frac{i-a}{n+1-2a}\right)^{n-k} = e^{(-(i-a))},$$

entonces

$$\lim_{n\to\infty} G_i(y_i^*) = e^{(-(i-a))} \sum_{k=0}^{i-1} \frac{(i-a)^k}{k!} = \frac{1}{2}.$$

La cual es una Poisson con parámetro $(i-a)$.
Para distintos valores de i, podemos obtener distintos valores de a al resolver la ecuación, cuando:

1. $i = 4$,

$$\sum_{k=0}^{3} e^{(-(4-a))} \frac{(4-a)^k}{k!} = \frac{1}{2}$$

obtenemos:

$$a = .32794.$$

2. $i = 10$,

$$\sum_{k=0}^{9} e^{(-(10-a))} \frac{(10-a)^k}{k!} = \frac{1}{2}$$

obtenemos:

$$a = .33129.$$

3. $i = 30$ obtenemos: $a = .33267$.

4. $i = 50$ obtenemos: $a = .33294$.

5. $i = 100$ obtenemos: $a = .33314$.

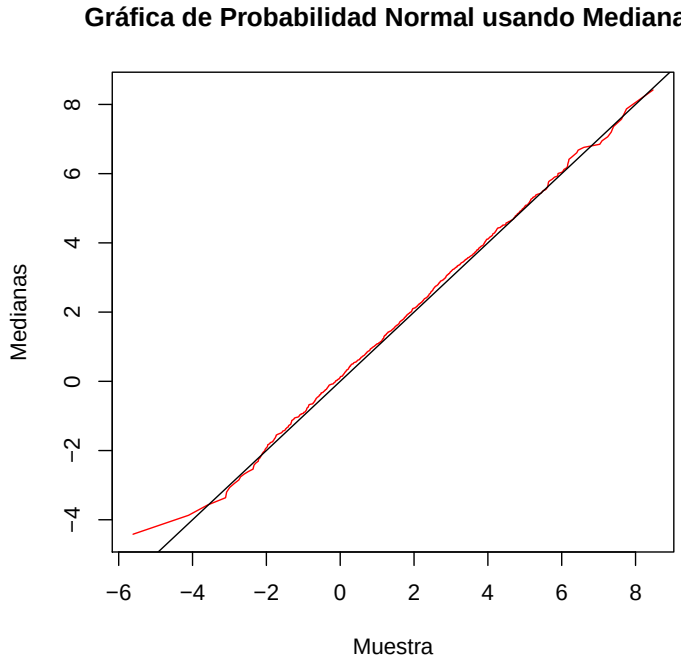
Por lo tanto, $\lim_{i\to\infty} e^{(-(i-a))} \sum_{k=0}^{i-1} \frac{(i-a)^k}{k!}$ nos da un valor de $a = \frac{1}{3}$.

Por lo que :

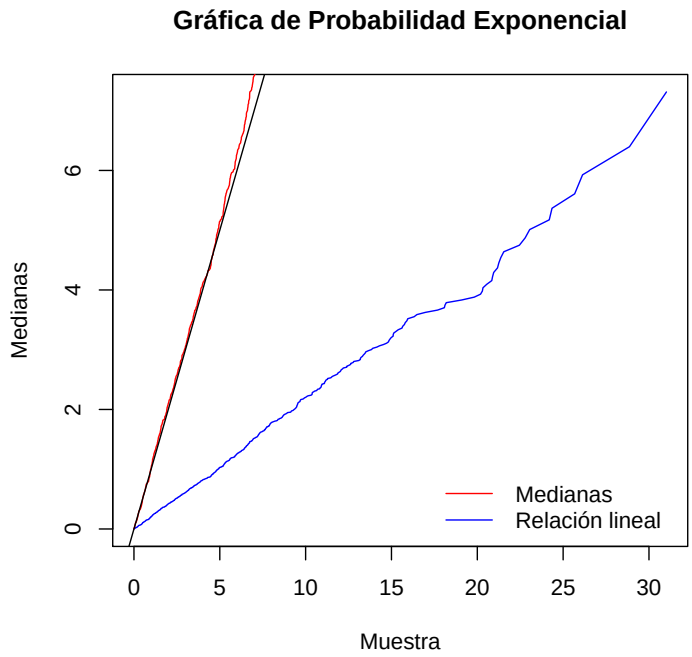
$$p(i) = y_i^* = \frac{i - \left(\frac{1}{3}\right)}{n + 1 - 2\left(\frac{1}{3}\right)}.$$

Es importante observar que significa y_i^* en (6.3). y_i^* es aquel valor que estimará a la función de distribución en el punto x_i, además es aquel valor donde el 50 % de los y_i con $i = 1, \ldots, n$ son menores que él y el 50 % son mayores; además, es aquel valor que al ser evaluado en su función de distribución se obtiene el valor $\frac{1}{2}$. La única distribución que cumple lo anterior es la $Uniforme(0, 1)$, entonces, si queremos obtener medianas M_i para una $Normal(a, b)$, basta aplicarle la inversa de dicha función de distribución a las medianas uniformes $m_i = y_i$, i.e. $M_i = \Phi^{(-1)}(m_i)$ (Por el Teorema de la Transformación Integral de Probabilidad).

Entonces, para obtener una gráfica de probabilidad, graficamos las medianas contra la muestra observada, ya que las medianas juegan el papel de la relación lineal que obteniamos previamente.



Gráfica 6.5: *Gráfica de probabilidad Normal usando medianas, usando una muestra* $Normal(2, 2)$ *de tamaño* $n = 1000$.



Gráfica 6.6: *Gráfica de probabilidad Exponencial usando medias, usando una muestra* $Exponencial(1/5)$ *de tamaño* $n = 1000$*, donde se compara el uso de Medianas y el uso de relaciones lineales.*

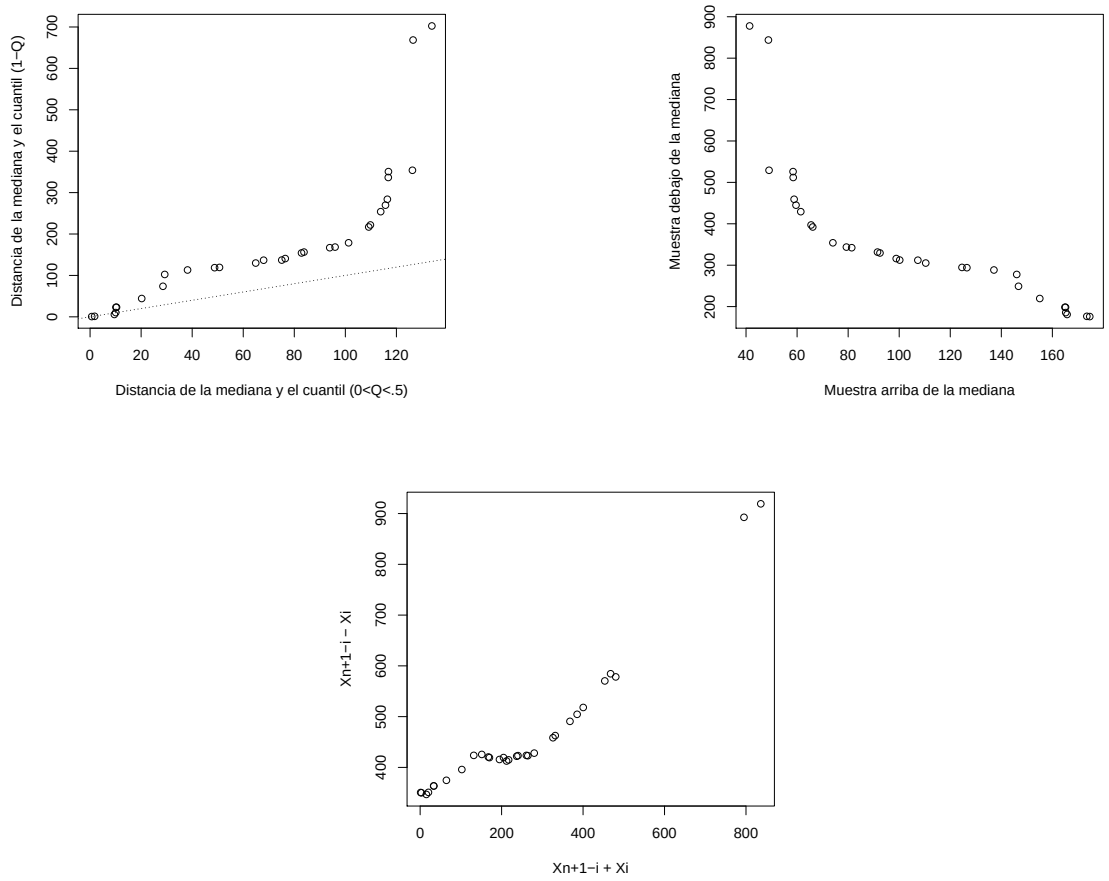
Capítulo 7

Aplicación

El objetivo de este capítulo es el de aplicar las técnicas gráficas descritas en este trabajo a una muestra, con la finalidad de poder observar la distribución de los datos, además de las ventajas y las limitaciones de usar técnicas gráficas para realizar inferencias.

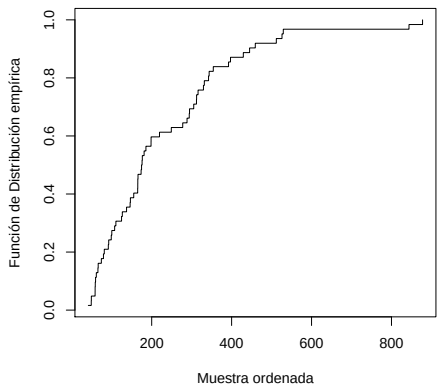
Los datos que se estudiarán a continuación representan la *LFreq* de 62 canciones distintas, es decir, la mediana de la ubicación del quinceavo pico más alto en el periodograma. Estos datos fueron obtenidos por la Dr. Di Cook, del Departamento de Estadística de la Universidad del Estado de Iowa (Véase [15]). Ella leyó cada pista de audio en el software de edición de música Amadeus II, la cortó y guardó los primeros 40 segundos, como un archivo WAV (WAV es un formato de audio desarrollado por Microsoft); estos archivos se leyeron en R utilizando el paquete *tuneR*, el cual convierte el documento de audio en datos numéricos. Los CDs de los que fueron tomadas dichas canciones, tienen 2 canales, izquierdo y derecho, y la variable en estudio, *LFreq*, fue calculada en ambos canales.

Verifiquemos la simetría de estos datos, utilizando las técnicas gráficas estudiadas en el Capítulo 3.



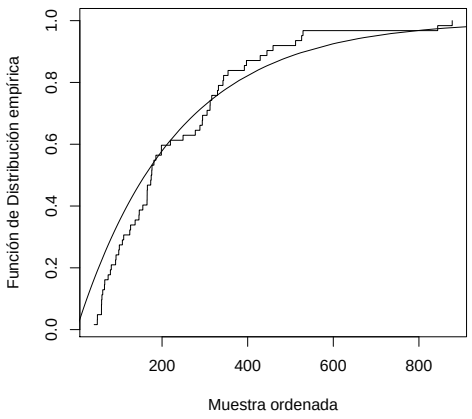
Gráfica 7.1: Gráficas sobre la simetría.

A pesar de no tener una muestra muy grande, las gráficas anteriores, nos sugieren que los datos no se distribuyen simétricamente, en éstas obtenemos una configuración similar a las que se obtienen con distribuciones que tienen sesgo a la derecha.



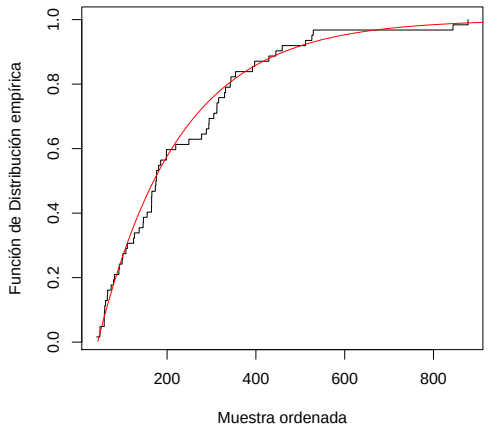
Gráfica 7.2: *Función de distribución empírica.*

La Función de distribución empírica tiene el comportamiento de distribución con sesgo a la derecha, como lo es la distribución Exponencial o Weibull.
Ya tenemos una idea sobre qué distribución puede tener nuestros datos. Sin embargo aún falta conocer los parámetros que tendría ésta.
Proponemos como posible distribución que siguen los datos a la distribución Exponencial con parámetro λ, y como estimador del parámetro de ésta, el estimador máximo verosímil, es decir, $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 0.004321632$.
En principio, evaluaremos el ajuste de esta distribución a nuestros datos a través de una doble gráfica.



Gráfica 7.3: *Muestra y función de distribución teórica Exponencial(.0043).*

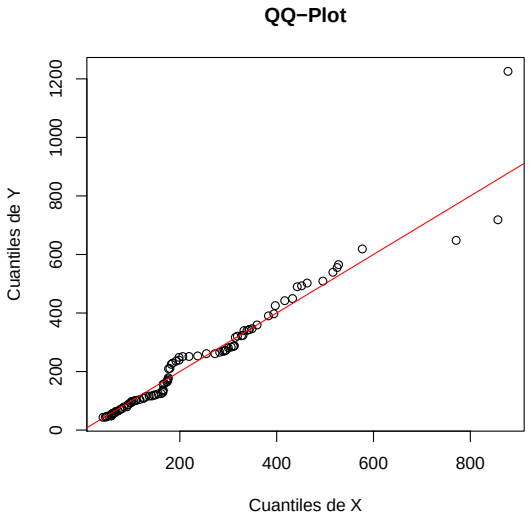
Observamos que la distribución Exponencial con $\hat{\lambda} = .00432$ se ajusta de buena forma a nuestros datos.
Quizá una distribución Exponencial desplazada se ajuste de mejor forma a nuestros datos, entonces, proponemos como posible distribución de los datos a la distribución Exponencial desplazada con parámetros λ, θ, y como estimadores de sus parámetros, a los estimadores por el método de momentos, es decir, $\hat{\lambda} = \frac{2n\frac{\sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2} = 0.005492374$ y $\hat{\theta} = \frac{ln(\hat{\lambda}) + ln(\frac{\sum_{i}^{n} x_i}{n})}{\hat{\lambda}} = 43.64735$.
Evaluemos el ajuste de esta distribución a nuestros datos a través de una doble gráfica.



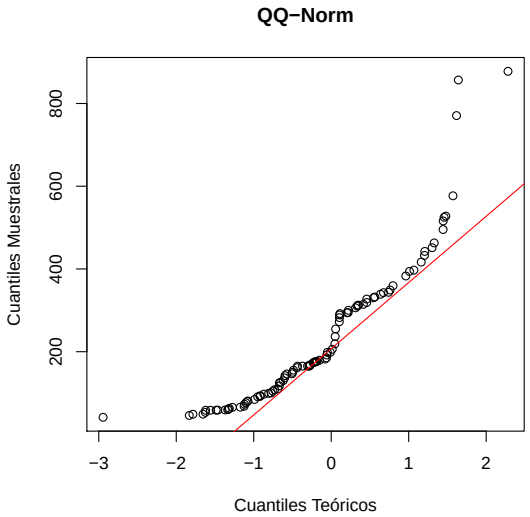
Gráfica 7.4: *Muestra y función de distribución teórica Exponencial desplazada(.0054,43.64).*

El ajuste de la distribución Exponencial como la Exponencial desplazada (con dichos parámetros) a nuestros datos según las dobles gráficas es muy similar, aunque parece que la distribución Exponencial desplazada lo hace de mejor forma.

Realicemos una QQ-Plot para ver si los cuantiles de los datos tienen los cuantiles de la distribución anterior.



(a) *Cuantiles de los datos contra cuantiles de la distribución Exponencial desplazada(.0054,43.64).*

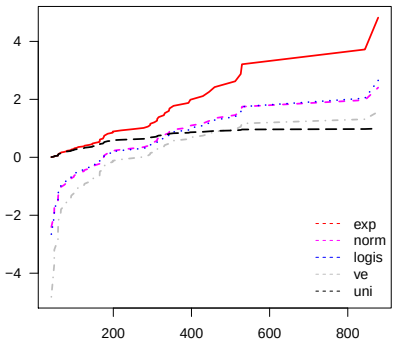


(b) *QQ-Norm de los datos.*

Gráfica 7.5: Gráficas Cuantil-Cuantil.

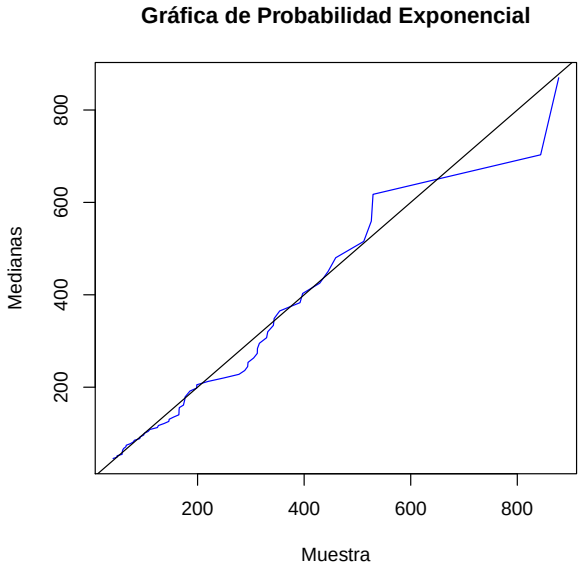
En la QQ-Plot, observamos que al graficar los cuantiles de los datos contra los

cuantiles de la distribución Exponencial desplazada(.0054,43.64), éstos caen aproximadamente sobre la línea $y = x$, lo que nos indica que ambas muestran provienen de la misma distribución, además, en la QQ-Norm, aparece una configuración de los puntos, similar al de distribuciones con sesgo a la derecha.



Gráfica 7.6: *Gráfica de Probabilidad.*

Observamos en la gráfica de probabilidad que aquella configuración que se asemeja más a una línea recta, es aquella cuya distribución hipotética es la distribución Exponencial (La gráfica de probabilidad de una distribución hipotética exponencial y una exponencial desplazada es la misma), lo que no indica que la distribución Exponencial desplazada es la verdadera distribución subyacente.



Gráfica 7.7: *Gráfica de Probabilidad utilizando medianas.*

En la gráfica de probabilidad anterior, se observa que los datos pertenecientes a la

cola de la distribución, caen burdamente en la línea recta $y = x$, lo cual pone en duda el ajuste completo de la distribución a los datos.

Hay evidencia suficiente para poder decir que los datos provienen de una distribución Exponencial desplazada con parámetros $\hat{\lambda} = .0054$ y $\hat{\theta} = 43.64$, sin embargo en algunas gráficas, resulta difícil poder evaluar correctamente el ajuste a la distribución, por ello, el análisis gráfico, es buen primer procedimiento, que además de conducir a un análisis más elborado con técnicas técnicas estadísticas más formales, permite llegar a éste con una perspectiva más clara del problema al investigador.

Comentarios Finales

En este trabajo se presenta una recopilación de las técnicas gráficas más importantes que son usadas para elaborar un análisis descriptivo de un conjunto de datos, cuya finalidad es ser un estudio que establezca las bases de estudios posteriores de otro tipo.

Se estima la función de distribución teórica a través de la función de distribución empírica, la cual, tiene distintas expresiones, sin embargo, no hay un criterio establecido que permita decir cual es mejor que otra, la elección de dicha expresión depende de la experiencia que tenga el investigador y las técnicas, métodos y estudios en los que se vaya a utilizar.

Las principales ventajas que presentan las gráficas Cuantil-Cuantil y las gráficas de probabilidad es que el ser humano puede detectar configuraciones lineales fácilmente, además la ventaja de usar gráficas de probabilidad (además de que producirán una línea recta si la distribución hipotética es la verdadera distribución subyacente) es que cuando se le ajusta a dicha configuración de datos una línea recta, se pueden obtener los parámetros de la función de distribución que se le ajusta, a través de la pendiente y la ordenada al origen de dicha recta. La diferencia de esta técnica con la que usa las medianas, es que cuando se usan las medianas para realizar las gráficas de probabilidad, se obtiene un análisis más rápido y sencillo de los datos, ya que éstos caerán forzosamente en la línea recta $y = x$ si la distribución hipotética es la verdadera distribución subyacente (aunque para poder realizarla, es necesario saber los parámetros de la distribución que se le ajustan a los datos).

Usualmente al realizar las gráficas anteriores, se observará que un modelo particular puede describir adecuadamente la mayor parte de los datos, sin embargo, es necesario realizar un análisis sobre las colas de la distribución, para entender su comportamiento y así poder concluir que los datos se ajustan a una distribución por completo.

La visualización de los datos, a través de gráficas, es un buen método para presentar

datos; también, le dan al investigador, un buen panorama respecto al tema de estudio o sobre las conclusiones a las que llega, usualmente, en el análisis gráfico que se trató en la tesis, resulta difícil poder evaluar correctamente el ajuste a la distribución, por ello, hay que poner atención a diversos factores, como lo son el tamaño de la muestra, la calidad de la muestra, así como la experiencia y el conocimiento del investigador, pues éstos influyen en la inferencia que se realizará.

Apéndices

Apéndice A

Generación de variables aleatorias

Teorema. *Si X es una variable aleatoria con función de distribución continua F (y F es invertible) y si $U = F(X)$, entonces U tiene distribución uniforme en el intervalo $(0, 1)$.*

Demostración.

$$\begin{aligned} F_U(u) &= \mathbb{P}[U \leq u] \\ &= \mathbb{P}[F(X) \leq u] && \text{y aplicando la función inversa} \\ &= \mathbb{P}[X \leq F^{-1}(u)] \\ &= F\Big(F^{-1}(u)\Big) = u \end{aligned}$$

Por lo tanto $U \sim Unif(0, 1)$.

□

Corolario. *Si U es una v.a. con distribución uniforme en el intervalo $(0, 1)$, la variable aleatoria $F^{-1}(U) = X$ tiene función de distribución F.*

Es decir, a partir de una variable aleatoria uniforme en el intervalo $(0, 1)$, podemos generar variables aleatorias de cualquier distribución.

Generación de muestras de variables aleatorias. *Método de la transformación inversa.*
Para generar valores de una v.a. X con función de distribución F, se utiliza el siguiente algoritmo:

1. *Generar un número aleatorio $U \sim Unif(0, 1)$.*

2. *Tomar* $X = F^{-1}(U)$.

Supongamos que queremos generar números aleatorios de la distribución exponencial desplazada, con parámetro λ, y $x > \theta$ entonces:

$$F(x) = \begin{cases} 1 - exp\Big(- \lambda(x - \theta)\Big) & x > \theta, \\ 0 & \text{e.o.c.} \end{cases}$$

Haciendo $U = F(X)$, con $U \sim Unif(0, 1)$, tenemos que:

$$\begin{aligned} U &= 1 - exp\Big(- \lambda(x - \theta)\Big) \\ (U - 1) &= -exp\Big(- \lambda(x - \theta)\Big) \\ log(1 - U) &= -\lambda(x - \theta) \\ \frac{-log(1 - U)}{\lambda} + \theta &= x, \end{aligned}$$

es una v.a. con distribución exponencial desplazada con parámetro λ.

Apéndice B

Propiedades de la función de distribucion exponencial desplazada

Tenemos que $f(x) = \lambda exp\{-\lambda(x-\theta)\}$. La función generadora de momentos es:

$$\begin{aligned} m_X(t) = \mathbb{E}[e^{Xt}] &= \int_0^\infty e^{xt}\lambda e^{-\lambda(x-\theta)}\,\mathrm{d}x \\ &= \lambda e^{\lambda\theta}\int_0^\infty e^{(t-\lambda)x}\,\mathrm{d}x \\ &= \lambda e^{\lambda\theta}\frac{1}{(t-\lambda)}e^{(t-\lambda)x}\Big|_0^\infty \text{ con t}<\lambda \\ &= e^{\lambda\theta}\frac{\lambda}{(\lambda-t)}. \end{aligned}$$

El primer momento es:

$$\begin{aligned} \mathbb{E}[X] = \frac{\mathrm{d}m_X(t)}{\mathrm{d}t}\Big|_{t=0} &= \frac{\lambda e^{\lambda\theta}}{(\lambda-t)^2}\Big|_{t=0} \\ &= \frac{e^{\lambda\theta}}{\lambda}. \end{aligned}$$

El segundo momento es:

$$\begin{aligned} \mathbb{E}[X^2] = \frac{\mathrm{d}^2m_X(t)}{\mathrm{d}t^2}\Big|_{t=0} &= \frac{2\lambda e^{\lambda\theta}}{(\lambda-t)^3}\Big|_{t=0} \\ &= \frac{2e^{\lambda\theta}}{\lambda^2}. \end{aligned}$$

Los estimadores por el método de momentos para λ y θ se obtienen de la siguiente forma:

Se igualan los momentos poblacionales con los momentos muestrales:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{e^{\lambda\theta}}{\lambda} \tag{B.1}$$

y

$$\frac{\sum_{i=1}^{n} x_i^2}{n} = \frac{2e^{\lambda\theta}}{\lambda^2}. \tag{B.2}$$

Despejando de la ecuación (B.1) a θ, tenemos:

$$\theta = \frac{log(\sum_{i=1}^{n} x_i) + log(\lambda)}{\lambda}.$$

Sustituyendo la expresión anterior en la ecuación (B.2), obtenemos:

$$\frac{\sum_{i=1}^{n} x_i^2}{n} = \frac{2e\lambda\left(\frac{log(\sum_{i=1}^{n} x_i) + log(\lambda)}{\lambda}\right)}{\lambda^2},$$

despejando de la ecuación anterior a λ, obtenemos:

$$\lambda = \frac{2n\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}.$$

Por lo tanto, los estimadores por el método de momentos para λ y θ son:

$$\hat{\lambda} = \frac{2n\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

y

$$\hat{\theta} = \frac{log(\sum_{i=1}^{n} x_i) + log(\hat{\lambda})}{\hat{\lambda}}.$$

Apéndice C

Códigos en R Project de las gráficas

Se crea la función de distribución empírica:

```
fdam<-function(muestra,x)
{
#muestra=sort(muestra) Si no le metemos la muestra ordenada


i=0
while(muestra[i+1]<=x)
{
i=i+1
if(i==length(muestra)-1)
{
break;
}
}

if(muestra[i+1]<=x)
{
i=i+1
}
return(i/length(muestra))
}
```

Se manda llamar la función de dsitribución empírica y se grafica dicha función contra la muestra ordenada

```
graph<-function(muestra)
{
muestra=sort(muestra)
grafica=matrix(0,length(muestra),2)
for(i in 1:length(muestra))
{
grafica[i,]=c(muestra[i],fdam(muestra,muestra[i]))
}
return(grafica)
}
muestra=rnorm(50,-2,2)
b=graph(muestra)
plot(b, type='s')
```

Función con la que se crean las banda de confianza usando el Teorema de Glivenko-Cantelli

```
KSm<-function(muestra,alpha,sigma)
{
muestra=sort(muestra)
grafica=matrix(0,length(muestra),5)

grafica[1,]=c(muestra[1],pnorm(muestra[1],alpha,sigma),
fdam(muestra,muestra[1]),0,0)
grafica[1,4]=abs(grafica[1,2]-grafica[1,3])
grafica[1,5]=abs(grafica[1,2]-0)
for(i in 2:length(muestra))
{
grafica[i,]=c(muestra[i],pnorm(muestra[i],alpha,sigma),
fdam(muestra,muestra[i]),0,0)
grafica[i,4]=abs(grafica[i,2]-grafica[i,3])
grafica[i,5]=abs(grafica[i,2]-grafica[i-1,3])
}
#return(max(grafica[,4],grafica[,5]))  ##Si quiero
# que me devuelva el Dn
```

```
return(grafica) ###Para las bandas de confianza
}

mu=4
sigma=2
muestra=rnorm(40,mu,sigma)
Smirnov=KSm(muestra,mu,sigma)
##Dnalpha observado
Dnalpha=.210
Smirnov=cbind(Smirnov,Smirnov[,3]+ Dnalpha,Smirnov[,3]-Dnalpha)

###Se crean las bandas de confianza
plot(sort(muestra),pnorm(sort(muestra),mu,sigma),
pch=19, xlab="Muestra ordenada",ylab="Fn(x)",
main="Banda de confianza",type='l')
points(Smirnov[,1], Smirnov[,6], type="s", lty=2,col="blue")
points(Smirnov[,1],Smirnov[,7], type="s", lty=2,col="blue")
```

Función con la que se crean las banda de confianza usando la desigualdad DKW.

```
graph<-function(muestra,alpha)
{
muestra=sort(muestra)
grafica=matrix(0,length(muestra),4)
epsilon=((1/(2*length(muestra)))*log(2/alpha))^(1/2)

for(i in 1:length(muestra))
{
grafica[i,]=c(muestra[i],fdam(muestra,muestra[i]),
max(fdam(muestra,muestra[i])-epsilon,0),
min(fdam(muestra,muestra[i])+epsilon,1))
}
return(grafica)
}

muestra=rnorm(4,2)
b=graph(muestra,.05)###Al ingresar alpha, genero una banda
###del 1-alpha
```

```
b
mu=4
sigma=2
plot(sort(muestra),pnorm(sort(muestra),mu,sigma),
pch=19, xlab="Muestra ordenada",ylab="Fn(x)", main="Banda de confianza",
type='l')
points(b[,1], b[,3], type="s", lty=2, col="red")
points(b[,1],b[,4], type="s", lty=2, col="red")
```

Gráficas de funciones de densidad y funciones de distribución

```
##################################función de Densida####################
########Normal
mu=0
var=1
x1=sort(rnorm(1000,mu,var))
x2=dnorm(x1,mu,var)
plot(x1,x2, type="l",xlab = "Muestra", ylab="Función de densidad")
########Exponencial
lambda=1/5
y1=sort(rexp(1000,lambda))
y2=dexp(y1,lambda)
plot(y1,y2, type="l",xlab = "Muestra", ylab="Función de densidad")
######Beta(5,1)
a=5
b=1
z1=sort(rbeta(1000,a,b))
z2=dbeta(z1,a,b)
plot(z1,z2,type="l",xlab = "Muestra", ylab="Función de densidad")
############################Función de Distribución###
#########Normal
x3=pnorm(x1,mu,var)
plot(x1,x3,type='l',xlab = "Muestra",
ylab = "Función de distribución acumulativa",
main="Distribución Normal",col="red")
#############Exponencial
y3=pexp(y1,lambda)
plot(y1,y3,type='l',xlab = "Muestra",
```

```
ylab = "Función de distribución acumulativa",
main="Distribución exponencial", col="blue" )
###########Beta(5,1)
z3=pbeta(z1,a,b,lower.tail = TRUE, log.p = FALSE)
plot(z1,z3, type='l',xlab = "Muestra",
 ylab = "Función de distribución acumulativa",
main="Distribución Beta(5,1)", col="green" )
```

Gráficas donde se muestra la simetría de la muestra

```
symgraph1=function(x)
{
n = length(x)
n2 = n %/% 2
muestra = sort(x)
mediana = median(x)
plot(mediana-muestra[1:n2], rev(muestra)[1:n2]-mediana,
xlab = "Distancia de la mediana y el cuantil (0<Q<.5)",
ylab = "Distancia de la mediana y el cuantil (1-Q)")
abline(a = 0, b = 1, lty = "dotted")
}
symgraph1(z)


symgraph2=function(x)
{
n = length(x)
n2 = n %/% 2
muestra = sort(x)
plot( muestra[1:n2] ,rev(muestra)[1:n2],
xlab = "Muestra arriba de la mediana", ylab = "Muestra debajo de la mediana")
abline(a = 0, b = -1, lty = "dotted")
}
symgraph2(z)

symgraph3=function(x)
{
n = length(x)
```

```
n2 = n %/% 2
muestra = sort(x)
plot(rev(muestra)[1:n2]-muestra[1:n2],rev(muestra)[1:n2]+muestra[1:n2] ,
xlab = "Xn+1-i + Xi", ylab = "Xn+1-i - Xi")

}
symgraph3(z2)
```

Gráficas donde se evaluan las colas de las distribuciones Exponencial, Lognormal y Weibull

```
##Primero se asignan los datos
##Exponencial
n=1000
lambda=1/3
x=rexp(n,lambda)
x=sort(x)
x
y=1-pexp(x,lambda)
y

################Wiebull
shape=1.5 ##Tiene que ser mayor que 1
scale=3
x2=rweibull(n,shape, scale)
x2=sort(x2)
x2
y2=1-pweibull(x2,shape,scale, lower.tail = TRUE, log.p = FALSE)
y2

#################Lognormal
meanlog=1/2
sdlog=4
x3=rlnorm(n, meanlog, sdlog)
x3=sort(x)
x3
y3=1-plnorm(x, meanlog, sdlog, lower.tail = TRUE, log.p = FALSE)
y3
```

```
###Gráfica con la que se evaluan las colas de dichas distribuciónes
plot(x,y,log="y", type="l", ylim=c(.0001,1),col=14,
xlab="Muestra ordenada",ylab="1-F(x)")
lines(x2,y2,col=3)
lines(x3,y3,col=4)
legend("bottomright",legend =c("exponencial","weibull","lognormal"),lty =1 ,
col= c(14,3,4),xjust = 1, yjust = 1, bty = "n")
```

Primer ejemplo donde se evalua la cola de la muestra proveniente de la distribución exponecial con $\lambda = 1/4$

```
fdam<-function(muestra,x)
{
#muestra=sort(muestra) Si no le metemos la muestra ordenada
i=0
while(muestra[i+1]<=x)
{
i=i+1
if(i==length(muestra)-1)
{
break;
}
}
if(muestra[i+1]<=x)
{
i=i+1
}
return((i-.5)/length(muestra))
}

graph<-function(muestra)
{
muestra=sort(muestra)
grafica=matrix(0,length(muestra),2)
for(i in 1:length(muestra))
{
grafica[i,]=c(muestra[i],fdam(muestra,muestra[i]))
```

```
}
return(grafica)
}

muestra=rexp(100,1/4)
##################
b=graph(muestra)
b<-b[!(b[,2]<=.5),]###Elimino los Fn(x)<=.5
b[,2]=1-b[,2]       ###Obtengo el 1-Fn(x)
#b
###############
lambda=log(mean(b[,2]))/-mean(b[,1])  ##Estimacion del parametro de lambda
lambda

x=rexp(50,lambda) ###Yo la propongo
#x
y=1-pexp(x,lambda)
#y
plot(x,y,log="y", type="l", ylim=c(.005,.5),xlim=c(.005,40),col=14)
########
points(b,ylim=c(.005,.5),xlim=c(0,40),col=3)###La sentencia de ylim
acorta la
muestra propuesta por decreto, sino, tendriamos que dividir a la
muestra y
luego graficarla
legend("bottomright",legend =c("Teorica","Muestral"),lty =1,col= c(14,3),
xjust = 1, yjust = 1, bty = "n")
```

Segundo ejemplo donde se evalua la cola de la muestra proveniente de la distribución Weibull(2,3)

```
shape=2 ##Tiene que ser mayor que 1
scale=3
n=100
muestra=rweibull(n,shape, scale)
b=graph(muestra)
b<-b[!(b[,2]<=.5),]###Elimino los Fn(x)<=.5
b[,2]=1-b[,2]       ###Obtengo el 1-Fn(x)
```

```
b
###Análisis exponencial
lambda=log(mean(b[,2]))/-mean(b[,1])  #Estimacion del parametro de lambda
x=rexp(n,lambda) ###Yo la propongo
x
y=1-pexp(x,lambda)
y
plot(x,y,log="y", type="l", ylim=c(.005,.5),xlim=c(.005,40),col=14)
########
points(b,ylim=c(.005,.5),xlim=c(0,40),col=3)
legend("bottomright",legend =c("Teorica","Muestral"),lty =1,col= c(14,3),
xjust = 1, yjust = 1, bty = "n")


#########Se observa que su cola es más ligera  y la
##exponencial no ajusta a la muestra

#######GENERACION DE EXPONENCIALES DESPLAZADAS
gen_exp<-function(n, lambda,theta)
{

U <-runif(n)# valores uniformes
exponencial <- (-(1/lambda))*log(1-U)+theta# vector de
####valores exponenciales
return(exponencial)
}
my_exp<-gen_exp(n,lambda, theta)

#Graficando
hist (my_exp , freq=F)
x=seq(0 ,10 ,by=0.1)
curve(dexp(x , lambda ) ,add=TRUE, col=2, lty=3,lwd=3)
################################################
b
fx2=.495
x2=2.7814
lambda=log((1-fx2)/(mean(b[,2])))/(mean(b[,1])-x2)
```

```
theta=x2+log(.495)/lambda
#################
n=100
my_exp<-gen_exp(n,lambda, theta)

distrib<-function(x,lambda,theta)
{
z=1-exp(-lambda*(x-theta))
return(z)
}

distribucion=distrib(my_exp,lambda,theta)
y=1-distribucion
plot(my_exp,y,log="y", type="l", ylim=c(.005,.5),
xlim=c(.005,40),col=14)

points(b,ylim=c(.005,.5),xlim=c(0,40),col=3)###La sentencia de ylim
#acorta la
#muestra propuesta por decreto por decreto, sino, tendriamos
#que dividir a la
#muestra y luego plotearla
########
legend("bottomright",legend =c("Teorica","Muestral"),lty =1,
col= c(14,3),
xjust = 1,
yjust = 1, bty = "n")
```

Dobles gráficas de funciones de distribución teóricas y empíricas

```
########Función acumulativa Normal
a=0
b=1
x=rnorm(100,a,b)
x=sort(x)
y=pnorm(x,a,b)
plot(x,y,type='l',xlab = "Muestra", ylab = "Función acumulativa",
 main="Distribución Normal",col="red")
empirica=graph(x)
```

```
lines(empirica,type='s')
legend("bottomright",legend =c("Teorica","Empírica"),lty =1,
col= c("red","black"),xjust = 1, yjust = 1, bty = "n")

#######Función acumulativa exponencial
lambda=1/5
x=rexp(100,lambda)
x=sort(x)
y=pexp(x,lambda)
plot(x,y,type='l',xlab = "Muestra", ylab = "Función acumulativa",
main="Distribución exponencial", col="blue" )
empirica=graph(x)
lines(empirica,type='s')
legend("bottomright",legend =c("Teorica","Empírica"),lty =1,
col= c("blue","black"),xjust = 1, yjust = 1, bty = "n")
####Muestra exp, teorica normal
a=0
b=1
x=rnorm(100,a,b)
x=sort(x)
y=pnorm(x,a,b)
plot(x,y,type='l',xlab = "Muestra", ylab = "Función acumulativa",
 main="Distribución Normal",col="red")

lambda=5
x=rexp(100,lambda)
x=sort(x)
empirica=graph(x)
lines(empirica,type='s')
legend("bottomright",legend =c("Teorica","Empírica"),
lty =1,col= c("red","black")
xjust = 1, yjust = 1, bty = "n")
```

QQ-Plots

```
###Forma en que se graficaron las qq-plots
n=500
x=rexp(n,1/5)
```

```
m=500
y=rexp(n,1/5)
n = max(length(x), length(y))
p = (1:n - 1)/(n - 1)
qx = quantile(x, p)
qy = quantile(y, p)
plot(qx, qy, xlab="Cuantiles de X", ylab="Cuantiles de Y", main="QQ-Plot")
abline(0,1,col="red")
#####Si quiesieramos ahorrarnos las lineas anteriores
qqplot(x,y)
#############Ejemplos de estaturas de las mujeres
#####Se usa el paquete alr3, y los daros heights
names(heights)
attach(heights)
x=Mheight
y=Dheight
n = max(length(x), length(y))
p = (1:n - 1)/(n - 1)
qx = quantile(x, p)
qy = quantile(y, p)
plot(qx, qy, xlab="Estaturas de las Madres",
 ylab="Estaturas de las hijas",
main="QQ-Plot")
abline(0,1,col="red")
############
m=mean(quantile(y, seq(0, 1, length = n))/
quantile(x, seq(0, 1, length = n)))
x=m*x
y=Dheight
n = max(length(x), length(y))
p = (1:n - 1)/(n - 1)
qx = quantile(x, p)
qy = quantile(y, p)
plot(qx, qy, xlab="1.02067*(Estatura de las Madres)",
ylab="Estatura de las hijas", main="QQ-Plot" )
abline(0,1,col="red")
###############Diversas qq-plots con varias distribuciones
```

```
y=rnorm(500,0,1)#######Cuantiles teóricos
x=rexp(500,5)
qqplot(y,x)
abline(0,1)
######Normal
x=rnorm(500,0,1)
qqnorm(x,xlab="Cuantiles Teóricos (los de una Normal (0,1))",
ylab="Cuantiles Muestrales", main="QQ-Plot Normal")
qqline(x,col="red")
#######Exponencial
x=rexp(500, 1/5)
qqnorm(x,xlab="Cuantiles Teóricos (los de una Normal (0,1))",
ylab="Cuantiles Muestrales", main="QQ-Plot Normal")
qqline(x, col="red")
######
x=rlnorm(500, 0,1)
qqnorm(x,xlab="Cuantiles Teóricos (los de una Normal (0,1))",
ylab="Cuantiles Muestrales", main="QQ-Plot Normal")
qqline(x, col="red")
#######
######
x=rt(500, 2)
qqnorm(x,xlab="Cuantiles Teóricos (los de una Normal (0,1))",
ylab="Cuantiles Muestrales", main="QQ-Plot Normal")
qqline(x, col="red")
####
x=rchisq(500, 10)
qqnorm(x,xlab="Cuantiles Teóricos (los de una Normal (0,1))",
ylab="Cuantiles Muestrales", main="QQ-Plot Normal")
qqline(x, col="red")
```

Gráficas de probabilidad de una sola distribución

```
ploteo<-function(x,y,p)
{
plot(x,y,type="o")
i=1
y2=y[i]
```

```
p2=p[i]
while(i<=10)
{
y2=c(y2,y[(length(x)/10)*i])
p2=c(p2,p[(length(x)/10)*i])
i=i+1

}
axis(4, at=y2,labels=p2,col.axis="blue",las=2, cex.axis=0.3, tck=-.01)
}
ploteolog<-function(x,y,p)
{
plot(x,y,log="x",type="o")
i=1
y2=y[i]
p2=p[i]
while(i<=10)
{
y2=c(y2,y[(length(x)/10)*i])
p2=c(p2,p[(length(x)/10)*i])
i=i+1

}
axis(4, at=y2,labels=p2,col.axis="blue",las=2, cex.axis=0.3, tck=-.01)
}
exponencial<-function(x,p)
{
y=-log(1-p)
ploteo(x,y,p)

}
weibull<-function(x,p)
{
y=log(-log(1-p))
ploteolog(x,y,p)

}
```

```
normal<-function(x,p)
{
y=qnorm(p,0,1)
ploteo(x,y,p)

}
lognormal<-function(x,p)
{
y=qnorm(p,0,1)
ploteolog(x,y,p)
}
logistica<-function(x,p)
{
y=((3^(1/2))/pi)*log(p/(1-p))
ploteo(x,y,p)
}
valorextremo<-function(x,p)
{
y=log(-log(1-p))
ploteo(x,y,p)
}
uniforme<-function(x,p)
{
y=p
ploteo(x,y,p)
}
cauchy<-function(x,p)
{
y=tan(pi*(p-.5))
ploteo(x,y,p)
}

grafprob<-function(muestra,tipo)
{
x=sort(muestra)
p = (1:length(x) - .5)/length(x)
if(tipo=='exponencial')
```

```
{
exponencial(x,p)
}
if(tipo=='weibull')
{
weibull(x,p)
}
if(tipo=='normal')
{
normal(x,p)
}
if(tipo=='lognormal')
{
lognormal(x,p)
}
if(tipo=='logistica')
{
logistica(x,p)
}
if(tipo=='valorextremo')
{
valorextremo(x,p)
}
if(tipo=='uniforme')
{
uniforme(x,p)
}
if(tipo=='cauchy')
{
cauchy(x,p)
}



}
##Algunas corridas
n=500
muestra=rexp(n,5)
```

```
tipo='normal'
grafprob(muestra,tipo)



tipo='exponencial'
grafprob(muestra,tipo)
```

Gráficas de probabilidad de varias distribuciones simultáneamente

```
exponencial<-function(x,p)
{
y=-log(1-p)
#ploteo(x,y,p)
return(y)

}
weibull<-function(x,p)
{
y=log(-log(1-p))
#ploteolog(x,y,p)
return(y)

}
normal<-function(x,p)
{
y=qnorm(p,0,1)
#ploteo(x,y,p)
return(y)

}
lognormal<-function(x,p)
{
y=qnorm(p,0,1)
#ploteolog(x,y,p)
return(y)
}
logistica<-function(x,p)
{
```

```
y=((3^(1/2))/pi)*log(p/(1-p))
#ploteo(x,y,p)
return(y)
}
valorextremo<-function(x,p)
{
y=log(-log(1-p))
#ploteo(x,y,p)
return(y)
}
uniforme<-function(x,p)
{
y=p
#ploteo(x,y,p)
return(y)
}
cauchy<-function(x,p)
{
y=tan(pi*(p-.5))
#ploteo(x,y,p)
return(y)
}



grafprob2<-function(muestra)
{
x=sort(muestra)
p = (1:length(x) - .5)/length(x)



#wei=weibull(x,p)
#logn=lognormal(x,p)

distribuciones= data.frame (exp=exponencial(x,p),
norm=normal(x,p),logis=logistica(x,p),ve=valorextremo(x,p),
uni=uniforme(x,p))
#Quitar la cauchy
```

```
rangoy<- range(distribuciones)
rangox<-range(x)
plot.new()
plot.window(xlim=rangox,ylim=rangoy)
lines(x,distribuciones$exp,type="l", lty=1,lwd=2,col=2)
lines(x,distribuciones$norm,type="l", lty=2,lwd=2,col=14)
lines(x,distribuciones$logis,type="l", lty=3,lwd=2,col=4)
lines(x,distribuciones$ve,type="l", lty=4,lwd=2,col=16)
lines(x,distribuciones$uni,type="l", lty=5,lwd=2,col=17)
#lines(x,distribuciones$cau,type="l", lty=6,lwd=2,col=19)
legend("bottomright",legend =names(distribuciones),lty =2 ,
col= c(2,14,4,16,17,19),xjust = 1, yjust = 1, bty = "n")
axis(1); axis(2, las = 1);
box();



}

grafprob3<-function(muestra)
{
x=sort(muestra)
p = (1:length(x) - .5)/length(x)

distribuciones= data.frame (wei=weibull(x,p),logn=lognormal(x,p))

rangoy<- range(distribuciones)
rangox<-range(x)
plot.new()
plot.window(xlim=rangox,ylim=rangoy)
lines(x,distribuciones$wei,type="l", lty=1,lwd=2,col=2)
lines(x,distribuciones$logn,type="l", lty=2,lwd=2,col=14)
legend("bottomright",legend =names(distribuciones),lty =2 ,
col= c(2,14),xjust = 1, yjust = 1, bty = "n")
axis(1); axis(2, las = 1);
box();
```

```
}

n=500
muestra=rexp(n,1/5)

grafprob2(muestra)

grafprob3(muestra)

# Gráficas de probabilidad usando medianas

#######Medianas
normal<-function(muestra, mu,sigma)
{
x=sort(muestra)
p = (1:length(x) - (1/3))/(length(x)+1-2*(1/3))
y=qnorm(p,0,1)*sigma+mu
plot(x,y,type="l",col="red",
main="Gráfica de Probabilidad Normal usando Medianas",
xlab="Muestra",ylab="Medianas")
#abline(a=-mu/sigma,b=1/sigma, lty="dotted")
abline(0,1)

}
mu=2
sigma=2
n=1000
muestra=rnorm(n,mu,sigma)
normal(muestra,mu,sigma)

exponencial<-function(muestra,l)
{
x=sort(muestra)
p = (1:length(x) - (1/3))/(length(x)+1-2*(1/3))
y=-log(1-p)
plot(x,y,type="l",col="blue",
```

```
main="Gráfica de Probabilidad Exponencial usando Medianas",
 xlab="Muestra",ylab="Medianas")
y=-log(1-p)*(1/l)
lines(x,y,type="l",col="red")
abline(0,1)
#abline(a=0, b=l, lty="dotted")

}
n=1000
l=1/5
muestra=rexp(n,l)
exponencial(muestra,l)
```

Si se tienen muestras sin valores repetidos, para calcular la función de distribución acumulativa muestral se usa:

```
p = (1:length(muestra) - .5)/length(muestra)
p
```

Código de la Aplicación

```
music <- read.csv("http://streaming.stat.iastate.edu/~dicook
/multivariatelectures/data/music-plusnew-sub.csv", row.names=1)

x3=music$LFreq
symgraph1(x3)
symgraph2(x3)
symgraph2(x3)

fdam<-function(muestra,x)
{
#muestra=sort(muestra) Si no le metemos la muestra ordenada
i=0
while(muestra[i+1]<=x)
{
i=i+1
if(i==length(muestra)-1)
{
```

```
break;
}
}

if(muestra[i+1]<=x)
{
i=i+1
}
return(i/length(muestra))
}
graph<-function(muestra)
{
muestra=sort(muestra)
grafica=matrix(0,length(muestra),2)
for(i in 1:length(muestra))
{
grafica[i,]=c(muestra[i],fdam(muestra,muestra[i]))
}
return(grafica)
}
b=graph(x3)
plot(b, type='s', xlab="Muestra ordenada",
ylab="Función de Distribución empírica")
################
y=rexp(100,1/mean(x3))
sort(y)
lines(sort(y),pexp(sort(y),1/mean(x3)))
#######Exponencial desplazada
plot(b, type='s', xlab="Muestra ordenada",
ylab="Función de Distribución empírica")
lambda=2*length(x3)*(mean(x3))/sum(x3^2)
theta=(log(mean(x3))+log(lambda))/lambda
n=1000
gen_exp<-function(n, lambda,theta)
{

U <-runif(n)# valores uniformes
```

```
exponencial <- (-(1/lambda))*log(1-U)+theta#
return(exponencial)
}
my_exp<-gen_exp(n,lambda, theta)
distrib<-function(x,lambda,theta)
{
z=1-exp(-lambda*(x-theta))
return(z)
}

distribucion=distrib(my_exp,lambda,theta)
lines(sort(my_exp),sort(distribucion),col="red")
##################QQ-Plot
#####################################################
n=100
y=gen_exp(n,lambda, theta)
x=x3
n = max(length(x), length(y))
p = (1:n - 1)/(n - 1)
qx = quantile(x, p)
qy = quantile(y, p)
plot(qx, qy, xlab="Cuantiles de X",
 ylab="Cuantiles de Y", main="QQ-Plot")
abline(0,1,col="red")

################QQ-Norm
y=rnorm(100,0,1)#######Cuantiles teóricos
n = max(length(x), length(y))
qp = (1:n - 1)/(n - 1)
qx = quantile(x, p)
qy = quantile(y, p)
plot(qy,qx, xlab="Cuantiles Teóricos",
ylab="Cuantiles Muestrales", main="QQ-Norm")
qqline(x3, col="red")
#################Gráficas de probabilidad
muestra=x3
grafprob2(muestra)
```

```
##########Medianas
exponencialmed<-function(muestra,l,theta)
{
x=sort(muestra)
p = (1:length(x) - (1/3))
/(length(x)+1-2*(1/3))
y=-log(1-p)*(1/l)+theta
plot(x,y,type="l",col="blue",
main="Gráfica de Probabilidad Exponencial",
 xlab="Muestra",ylab="Medianas")
abline(0,1)

}
muestra=x3
l=lambda
theta=theta
exponencialmed(muestra,l,theta)
```

Bibliografía

[1] Benard, A y Bos-Levenbach, E. C., *Het uitzetten van waarnemingen op waarschijnlijkdeids-papier*, Statististica Neerlandica, Volume 7, 1953.

[2] Bryson, M.C., *Heavy tailed distributions: properties and test*, Technometrics 16, 61-68.

[3] Casella, G. y Berger, R.L., *Statistical Inference*. 2a Ed., Duxbury Press, Belmont CA. pp 226-230.

[4] Curran, T.C. and Frank, N.H., *Assessing the validity of the log-normal and model when predicting maximun air pollution concentrations*, Presented at the 68th Anual Meeting of the Air Pollution Control Asociation, Boston, Mass.

[5] D'Agostino, R.B., 1986, *Graphical Analysis*, Chapter 2 en D'Agostino, R.B. y Stephens, M.A., *In Goodness-of-Fit Techniques*, Marcel Dekker, New York.

[6] Dvoretzky, A., Kiefer, J., Wolfowitz, J., *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, 1956, Annals of Mathematical Statistics 27 (3): 642–669.

[7] Filliben, J.J., *The Probability Plot Correlation Coefficient Test for Normality*, Technometrics, Vol. 17, No. 1, pp. 111-117.

[8] Karr, A.F., *Probability*, New York, Springer-Verlang. pp. 122, 135-151, 206.

[9] Lee, Elisa Tee y Wang J.W., *Statistical Methods for Survival Data Analysis*, 3a ed, pp. 198-209.

[10] Mood, A. M., Graybill F. A. y Boes D. C., *Introduction to the Theory of Statistics*. 3a Ed., New York: McGrawHill, pp. 504-514.

[11] Nelson, W. , *Applied Life Analysis*, 1982, John Wiley & Sons. pp. 103-119

[12] Rincón, L., *Curso intermedio de probabilidad*, Las Prensas de Ciencias, Fac. de Ciencias, UNAM. pp. 347-351,„357-359.

[13] Royston Erica, *A note on the history of the graphical presentation of data*, Studies in the History of Probability and Statistics, Division of Research Techniques, Londond School of Economics and Political Science.

[14] Wilk, M.B. y Gnanadesikan, R., *Probability plotting methods for the analysis of data*, Biometrika, 1968, 55,1, p. 1.

[15] http://streaming.stat.iastate.edu/~dicook/multivariatelectures/data/music-plusnew-sub.csv

Ahora, cuando el bardo de esta vida alborea ante mí,
abandonaré la pereza, para la cual no hay tiempo en la vida,
entraré sin distracción en el camino del escuchar y el oír, la reflexión
y la contemplación, y la meditación,
haciendo que las perpeciones y la mente sean el mismo camino, y realizaré
los tres kayas: la mente iluminada;
ahora que he logrado esta oportunidad única de tener un cuerpo humano,
no hay tiempo en el camino para que la mente deambule distraída.