Prueba de Bondad de Ajuste de Kolmogorov-Smirnov (KS)

Hipótesis a contrastar:

H_0: Los datos analizados siguen una distribución *M*.
H_1: Los datos analizados no siguen una distribución *M*.

Estadístico de contraste:

$$D = \sup_{1 \le i \le n} \left| \hat{F}_n(x_i) - F_0(x_i) \right|$$

donde:

- x_i es el i-ésimo valor observado en la muestra (cuyos valores se han ordenado previamente de menor a mayor).
- $\hat{F}_n(x_i)$ es un estimador de la probabilidad de observar valores menores o iguales que x_i.
- $F_0(x)$ es la probabilidad de observar valores menores o iguales que x_i cuando H_0 es cierta.

Así pues, D es la mayor diferencia absoluta observada entre la frecuencia acumulada observada $\hat{F}_n(x)$ y la frecuencia acumulada teórica $F_0(x)$, obtenida a partir de la distribución de probabilidad que se especifica como hipótesis nula.

Si los valores observados $\hat{F}_n(x)$ son similares a los esperados $F_0(x)$, el valor de *D* será pequeño. Cuanto mayor sea la discrepancia entre la distribución empírica $\hat{F}_n(x)$ y la distribución teórica , mayor será el valor de *D.*

Por tanto, el criterio para la toma de la decisión entre las dos hipótesis será de la forma:

$$\boxed{\begin{array}{l}\text{Si } D \le D_\alpha \Rightarrow \text{Aceptar } H_0 \\ \text{Si } D > D_\alpha \Rightarrow \text{Rechazar } H_0\end{array}}$$

donde el valor D_α se elige de tal manera que:

$$\boxed{\begin{array}{l}P(\text{Rechazar } H_0 / H_0 \text{ es cierta}) = \\ \quad = P(D > D_\alpha / \text{Los datos siguen la distribucion M}) = \alpha\end{array}}$$

siendo α el nivel de significación del contraste.

Para el cálculo práctico del estadístico *D* deben obtenerse:

$$D^+ = \max_{1 \le i \le n} \left\{ \frac{i}{n} - F_0(x_i) \right\}, \qquad D^- = \max_{1 \le i \le n} \left\{ F_0(x_i) - \frac{i-1}{n} \right\}$$

y a partir de estos valores:

$$D = \max \left\{ D^+, D^- \right\}$$

A su vez, el valor de D_α depende del tipo de distribución a probar y se encuentra tabulado. En general es de la forma:

$$D_\alpha = \frac{c_\alpha}{k(n)}$$

donde c_α y $k(n)$ se encuentran en las tablas siguientes:

c_α	α		
Modelo	**0.1**	**0.05**	**0.01**
General	1.224	1.358	1.628
Normal	0.819	0.895	1.035
Exponencial	0.990	1.094	1.308
Weibull n=10	0.760	0.819	0.944
Weibull n=20	0.779	0.843	0.973
Weibull n=50	0.790	0.856	0.988
Weibull n=∞	0.803	0.874	1.007

DISTRIBUCIÓN QUE SE CONTRASTA	*k(n)*
General. Parámetros conocidos.	$k(n) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$
Normal	$k(n) = \sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}$
Exponencial	$k(n) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$
Weibull	$k(n) = \sqrt{n}$

Ejemplo 1:

Determinar si los valores de la primera columna se conforman a una distribución normal:

Y	Y-ordenados	Orden	F	Z	Fo	D+	D-
6.0	1.9	1	0.1	-1.628	0.051	0.049	0.051
2.3	2.3	2	0.2	-1.332	0.091	0.109	-0.009
4.8	3.3	3	0.3	-0.592	0.276	0.024	0.076
5.6	3.4	4	0.4	-0.518	0.302	0.098	0.002
4.5	4.5	5	0.5	0.296	0.616	-0.116*	0.216*
3.4	4.5	6	0.6	0.296	0.616	-0.016	0.116
3.3	4.8	7	0.7	0.518	0.698	0.002	0.098
1.9	4.8	8	0.8	0.518	0.698	0.102	-0.002
4.8	5.6	9	0.9	1.11	0.867	0.033	0.067
4.5	6.0	10	1.0	1.406	0.920	0.080	0.020

(media: 4.1 varianza: 1.82)

$$D_\alpha = \frac{0.895}{\sqrt{10} - 0.01 + \frac{0.85}{\sqrt{10}}} = \frac{0.895}{3.42} = 0.262$$

Como el valor D = 0.216 < 0.262, no se rechaza H_0 y se acepta que los datos se distribuyen normalmente.

Modo alternativo de realizar la prueba de Kolmogorov Smirnov.

La toma de la decisión en el contraste anterior puede llevarse a cabo también mediante el empleo del p-valor asociado al estadístico D observado. El p-valor se define como:

$$\text{p-valor} = P\left(D > D_{obs} \,/\, H_0 \text{ es cierta}\right)$$

Si el p-valor es grande significa que, siendo cierta la hipótesis nula, el valor observado del estadístico *D* era esperable. Por tanto no hay razón para rechazar dicha hipótesis. Asimismo, si el p-valor fuera pequeño, ello indicaría que, siendo cierta la hipótesis nula, era muy difícil que se produjera el valor de *D* que efectivamente se ha observado. Ello obliga a poner muy en duda, y por tanto a rechazar, la hipótesis nula. De esta forma, para un nivel de significación α, la regla de decisión para este contraste es:

Si p-valor $\geq \alpha \Rightarrow$ Aceptar H_0
Si p-valor $< \alpha \Rightarrow$ Rechazar H_0

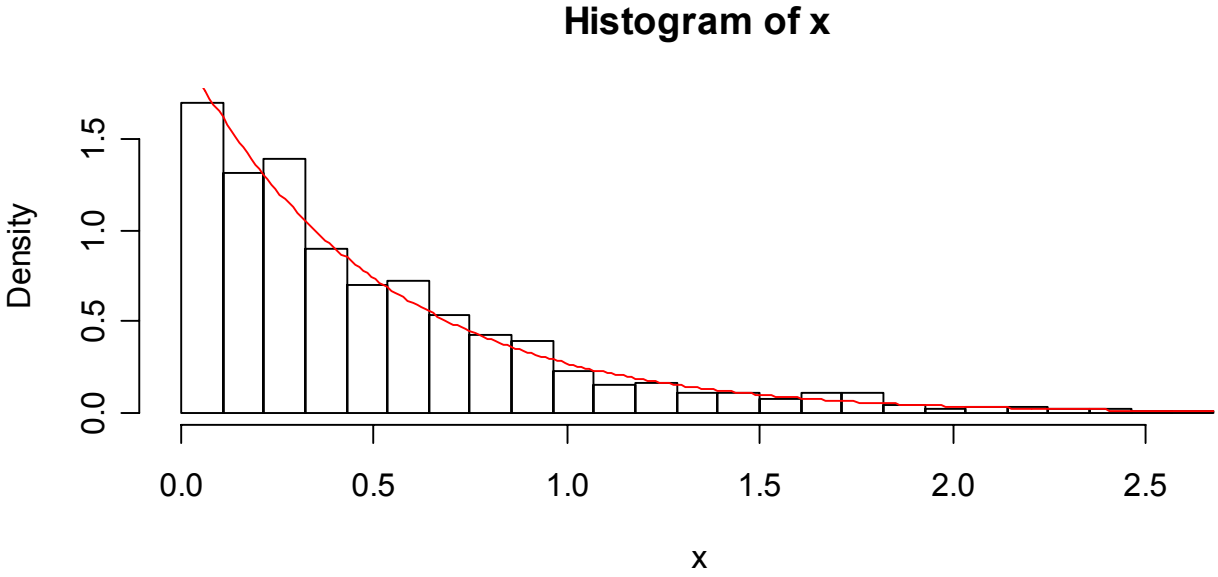
Obviamente, la obtención del p-valor requiere conocer la distribución de *D* bajo la hipótesis nula y hacer el cálculo correspondiente. En el caso particular de la prueba de Kolmogorov Smirnov, la mayoría de los paquetes de software estadístico realizan este cálculo y proporcionan el p-valor directamente.

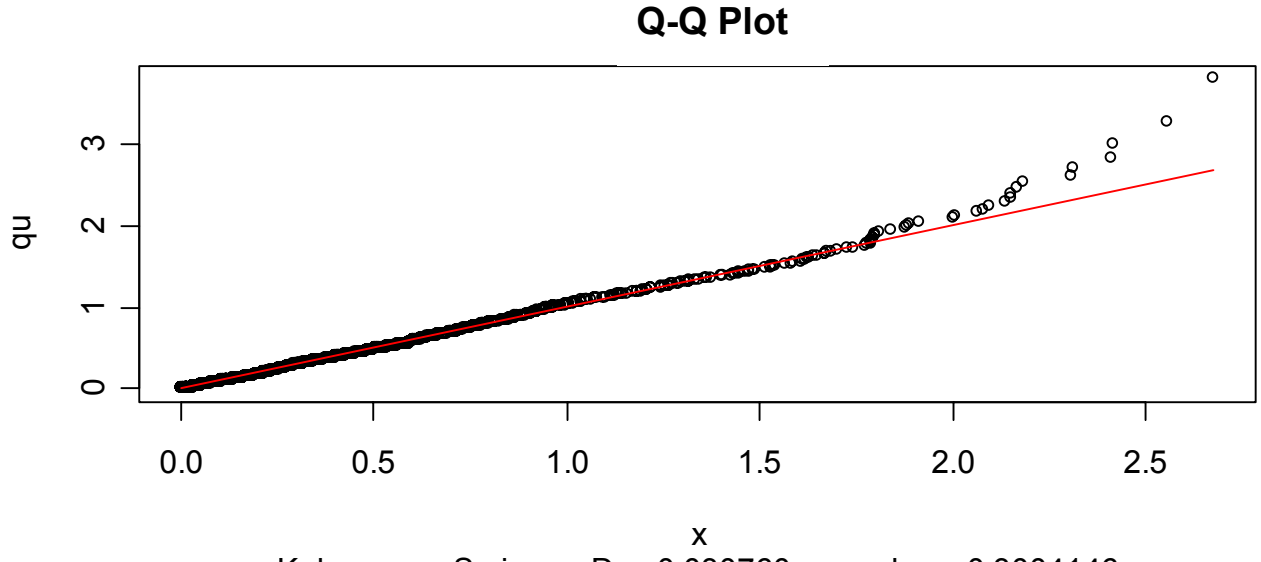
Ejemplo 2:

En los siguientes ejemplos se han simulado datos con distribución exponencial o normal, contrastándose en todos los casos si puede aceptarse que los datos siguen distribución exponencial. Se ha acompañado al contraste con el histograma de los datos y el gráfico Q-Q Plot (gráfico cuantil-cuantil: se representan los cuantiles de la distribución teórica supuesta frente a los cuantiles de la distribución empírica. En un buen ajuste, la gran mayoría de estos puntos deberían situarse sobre la recta y=x)

Simulación de datos con distribución exponencial

n=1000

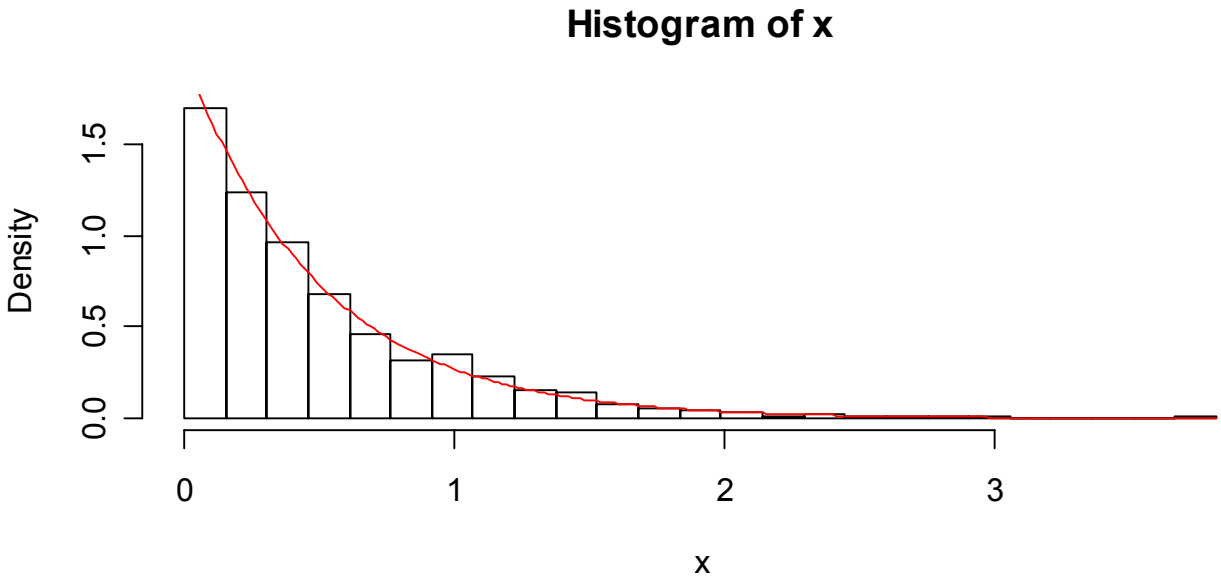


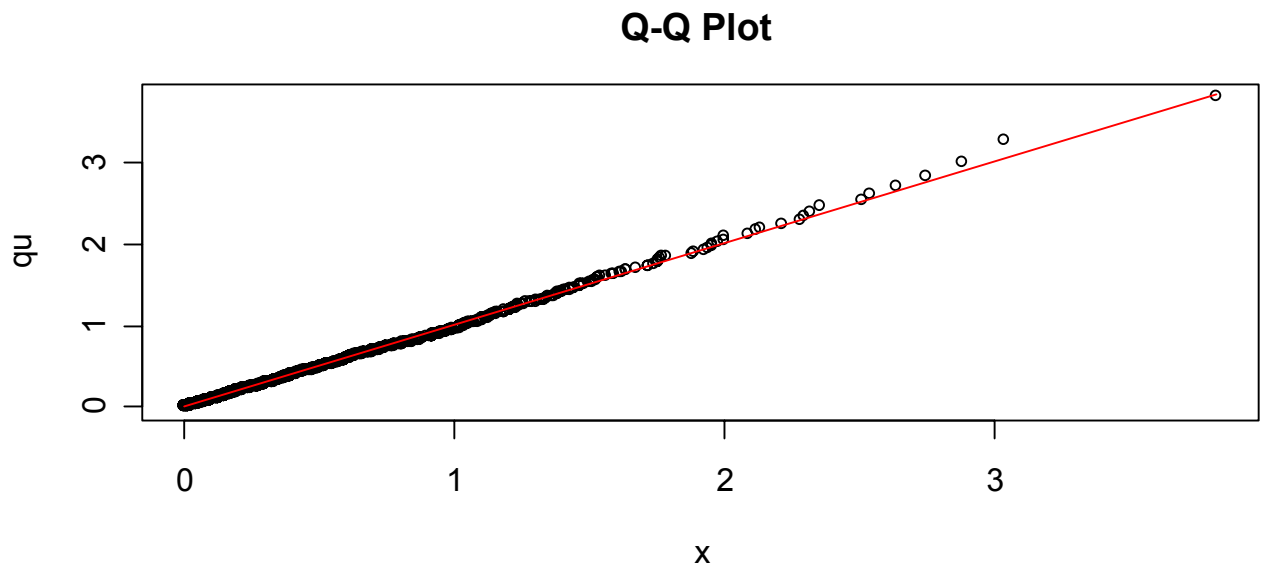


Kolmogorov Smirnov D = 0.030760 p-valor = 0.3004146

Simulación de datos con distribución exponencial:

n=1000

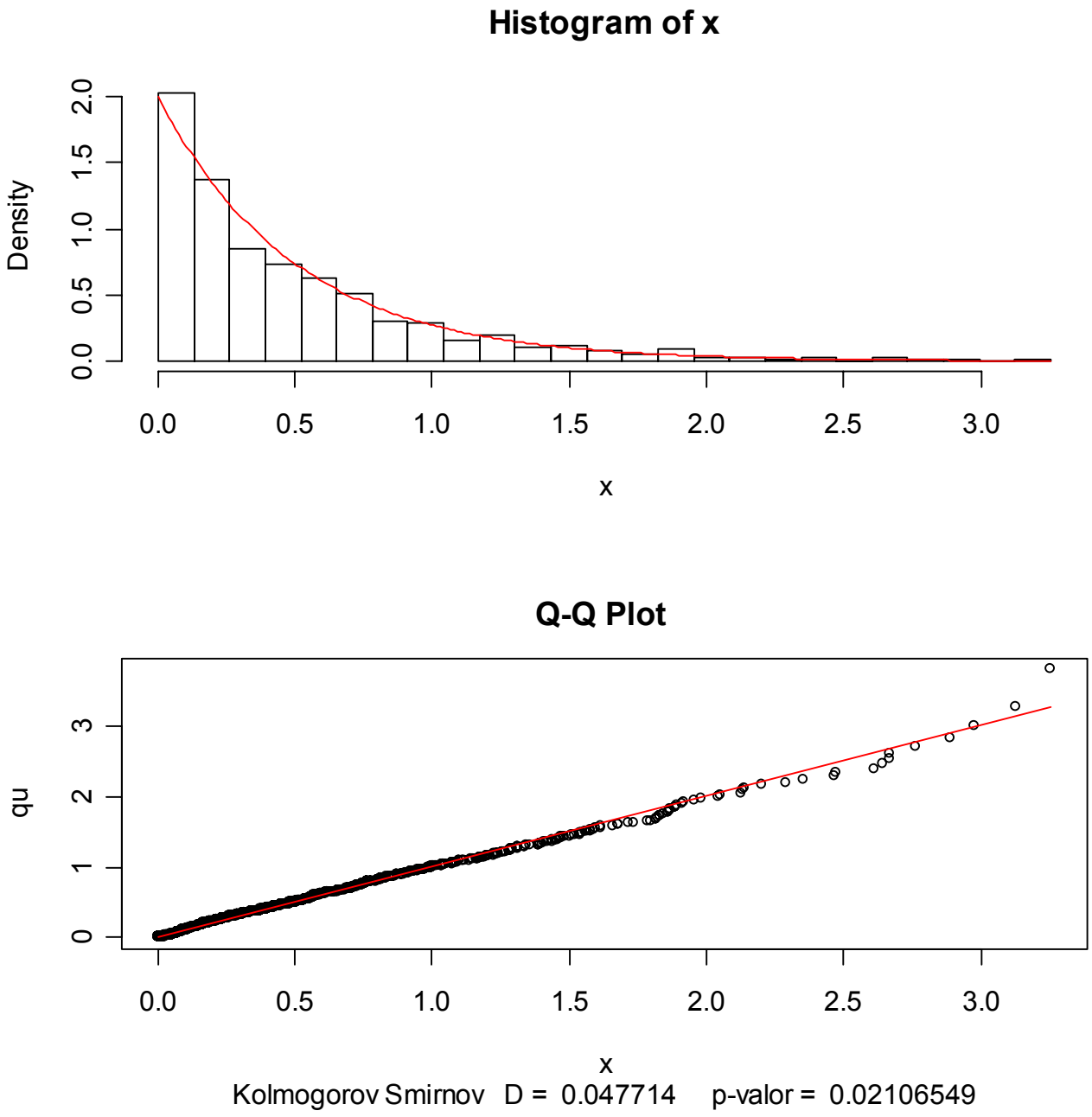




Kolmogorov Smirnov D = 0.018285 p-valor = 0.8917437

Simulación de datos con distribución exponencial

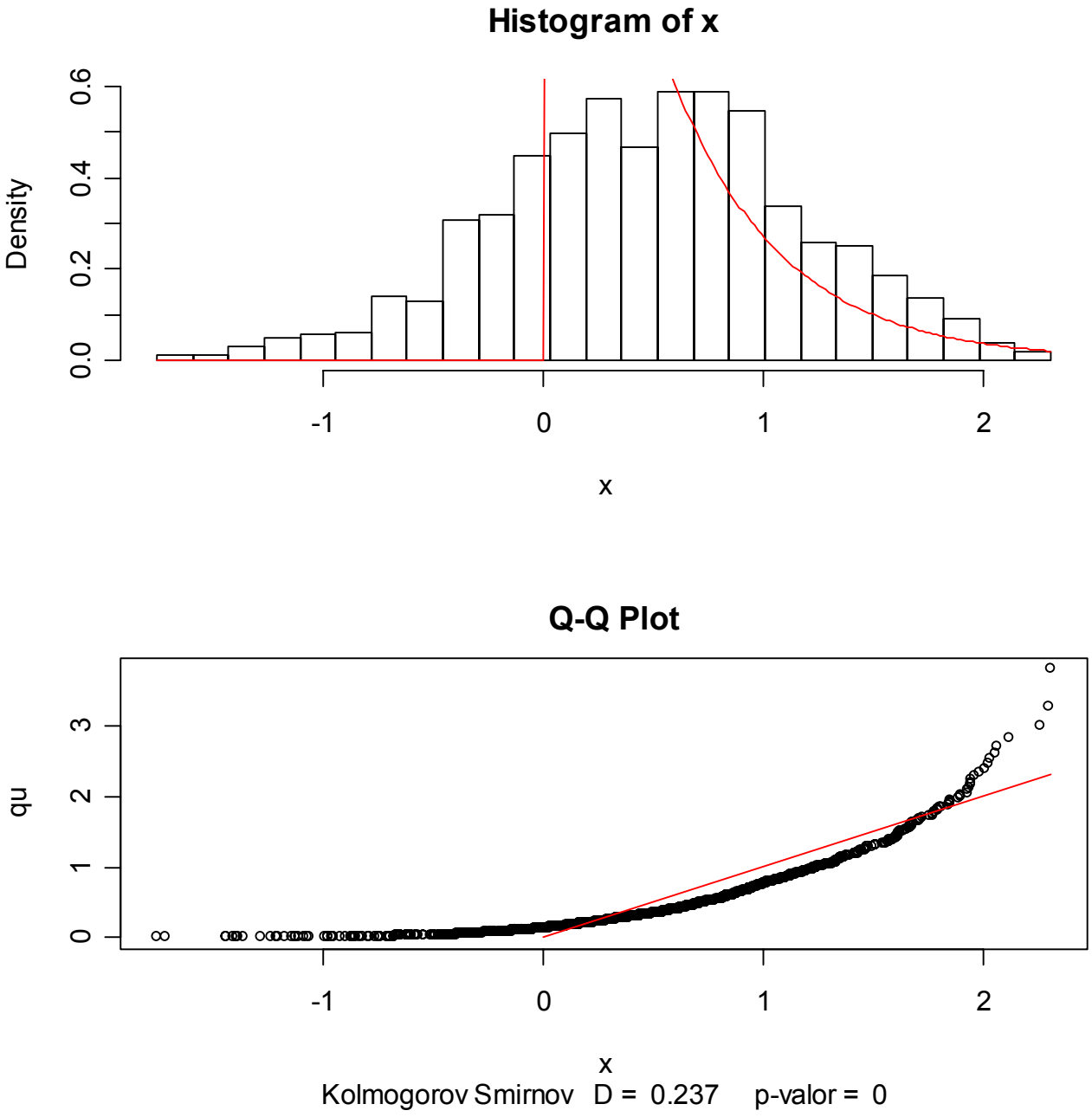
n=1000



Nótese que, en este caso, aunque los datos se han generado realmente con distibución exponencial, el p-valor conduce a rechazar que ésta sea la distribución de los datos.

Simulación de datos con distribución normal

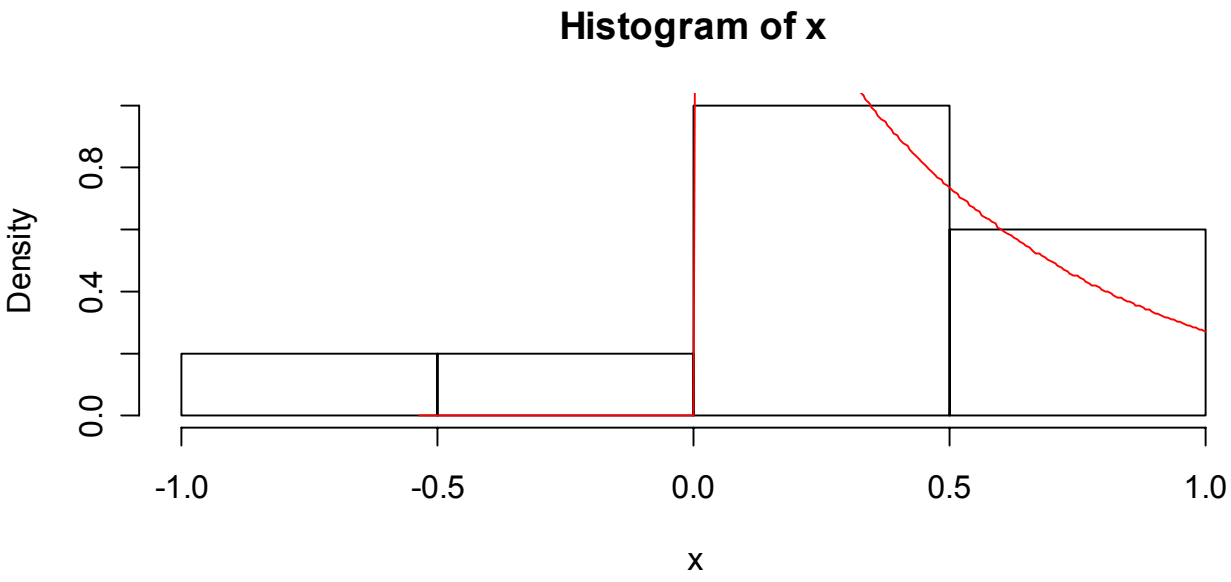
n=1000

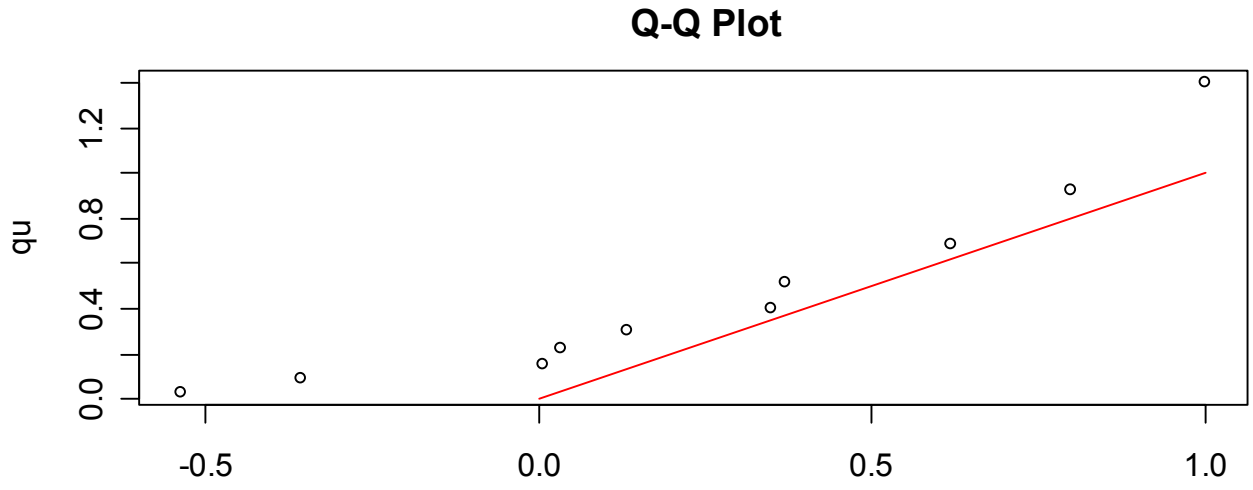


En este caso, obviamente se rechaza que la distribución sea exponencial, cosa que además se ve claramente en los gráficos.

Simulación de datos con distribución normal

n=10



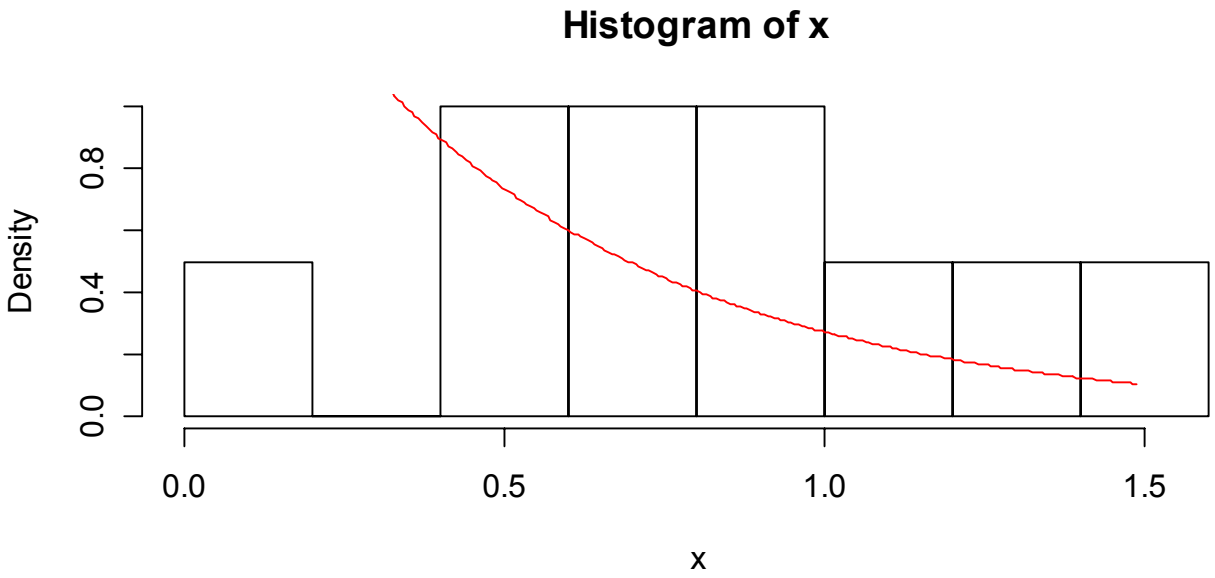


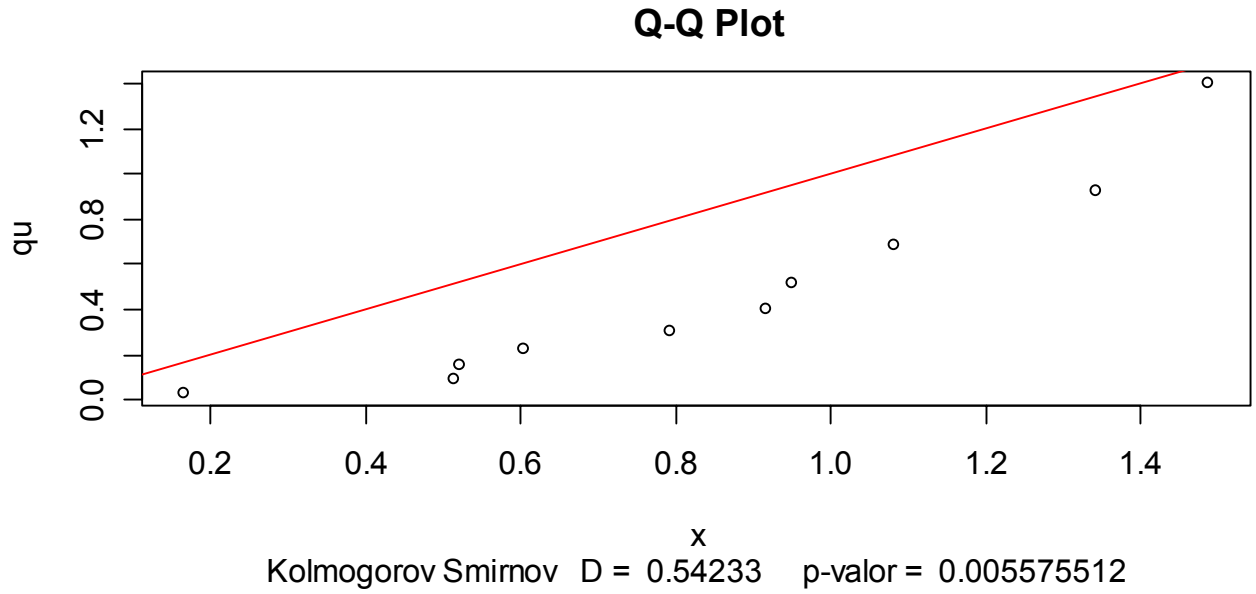
Kolmogorov Smirnov D = 0.33481 p-valor = 0.2122409

En este caso, aunque los datos se han generado con distribución normal, el contraste conduce a aceptar que siguen distribución normal. Ello se debe a que en general cuando hay poca información (en este caso sólo diez datos), la hipótesis nula tiende a no ser rechazada, salvo que haya una evidencia abrumadora en su contra.

Simulación de datos con distribución exponencial

n=10





Kolmogorov Smirnov D = 0.54233 p-valor = 0.005575512

Aquí ha ocurrido lo contrario al caso anterior; a pesar de que los datos son originalmente exponenciales, el contraste rechaza que lo sean