R Notebook

1) Introducción

El método estándar para construir los intervalos de confianza para la regresión lineal (y en general para conocer su distribución) recae en el supuesto de normalidad:

- Los errores de la regresión están distribuidos normal
- El número de observaciones es suficientemente grande, en cuyo caso, el estimador se distribuye normal (Teorema Central del Límite)

Si se cumple el primer supuesto, entonces el estimador de β_1 se distribuye normal con la media y varianza vista en clase .

Al realizar regresión lineal es importante determinar estos supuestos se cumplen, por lo que existen distintas técnicas para lograrlo. En este curso veremos dos tipos:

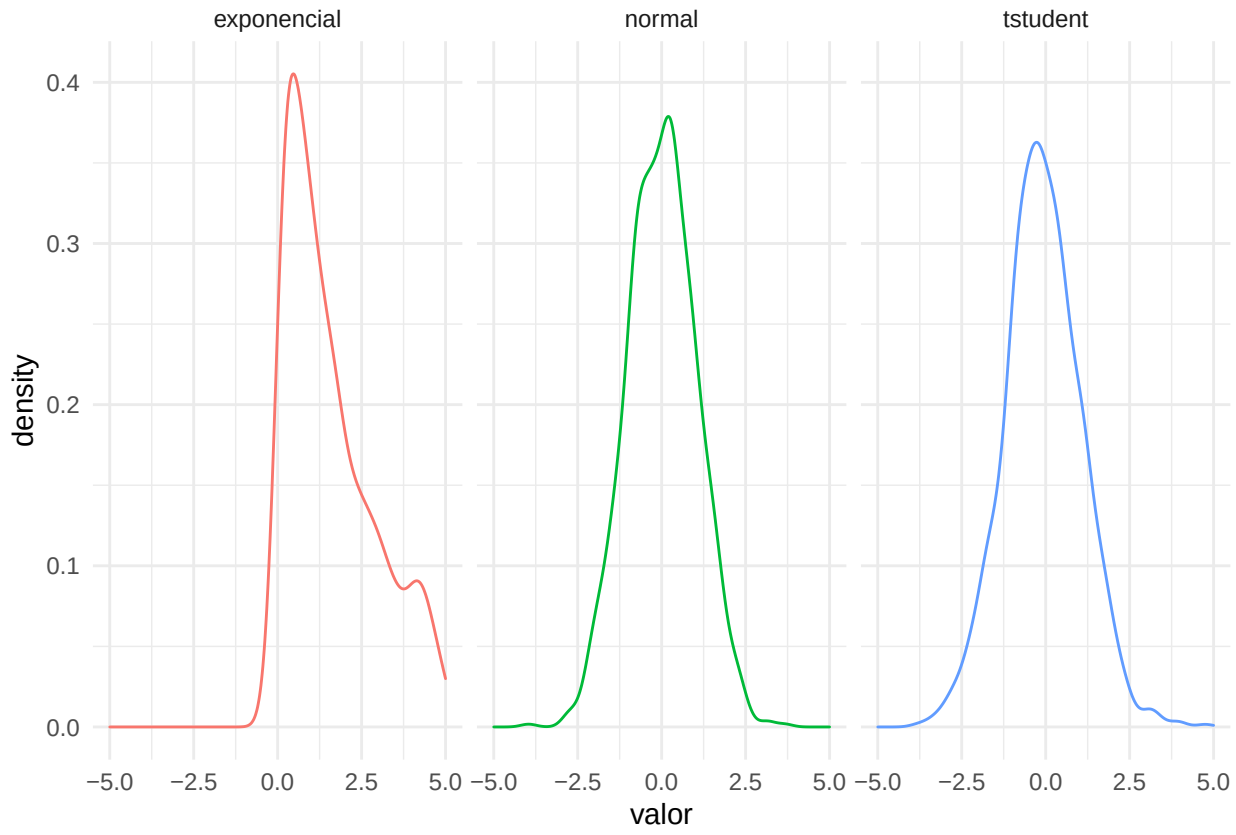
a) Métodos Gráficos
b) Pruebas de Hipótesis

2) Generación de datos:

Previo a comenzar con los métodos gráficos y pruebas de hipótesis se generan datos provenientes de distintas distribuciones:

```
# La función rnorm genera muestras aleatorias normales. Tiene como parámetros:
# n=número de muestras,
# mean = media de la distribucion normal
# sd = desviacion estándar de la distribución normal

# rt genera números aleatorios que provienen de una distribución t-student

# rexp genera número aleatorios que provienen de distribución exponencial

library(tidyverse)
set.seed(532)


datos <-
  data.frame(
  normal = rnorm(n = 1000, mean = 0, sd = 1),
  tstudent = rt(n = 1000, df = 10),
  exponencial = rexp(n = 1000, rate = 0.5)
)

datos %>%
  gather(variable, valor) %>%
```

```
ggplot(aes(x =valor, color = variable))+
geom_density()+
facet_wrap(~variable)+
theme_minimal()+
theme(
  legend.position = "none"
)+
xlim(-5, 5)
```



Antes de comenzar, podemos observar que la distribución Normal y la t-student se asemejan, las dos son distribuciones simétricas, pero la distribución t-student tiene mayor varianza:

```
var(datos$normal)
```

```
## [1] 1.047178
```

```
var(datos$tstudent)
```

```
## [1] 1.310563
```

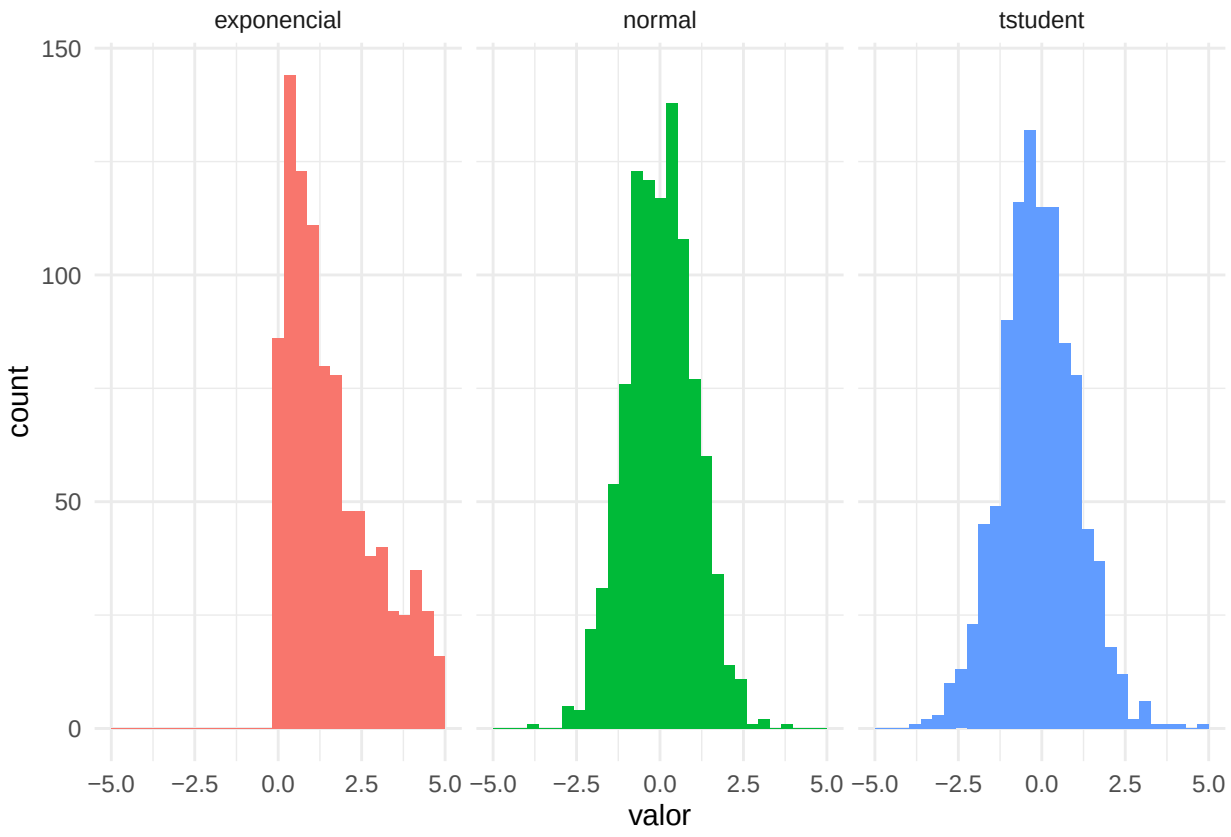
A simple vista a veces se vuelve un poco complicado poder determinar solo con la gráfica que distribución es normal y cual no.

3) Métodos gráficos

Los métodos gráficos tienen ciertas ventajas, entre ellas que conocemos la distribución de los datos, los valores que toman, la frecuencia de cada rango de valores, etc. Entre sus desventajas es que la decisión de determinar normalidad se deja a la persona que lo esté analizando.

3.1) Histograma El histograma representa la información en forma de barras, donde la superficie de cada barra implica la frecuencia de cada valor representado. Al igual que la gráfica de la sección 2, es facil identificar si tiene sesgo la distribución, pero se vuelve complicado saber si tiene la varianza es equivalente a la de una distribución normal.

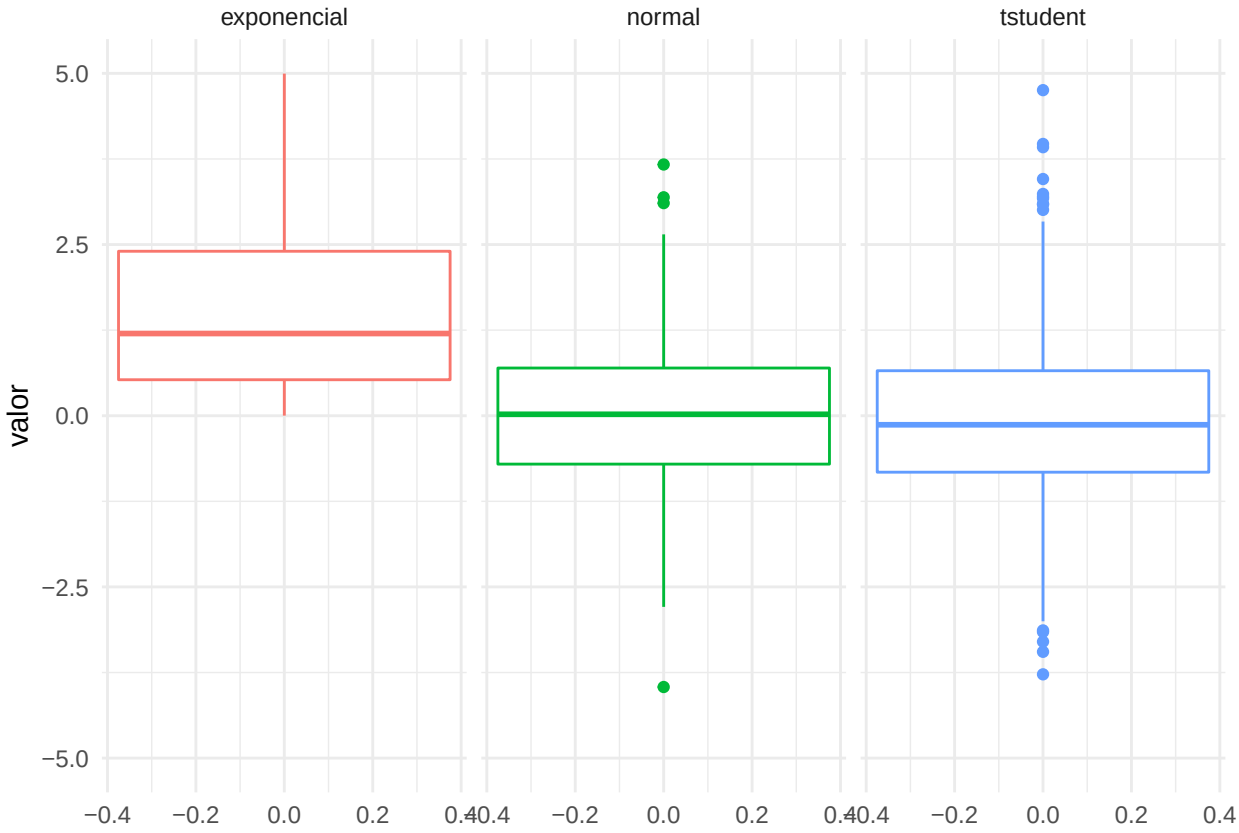
```
datos %>%
  gather(variable, valor) %>%
  ggplot(aes(x = valor, fill = variable))+
  geom_histogram()+
  facet_wrap(~variable)+
  theme_minimal()+
  theme(legend.position = "none")+
  xlim(-5,5)
```



3.2 Boxplot Los Boxplot nos permite conocer las principales estadísticas de los datos: mínimo, primer cuartil, mediana, tercer cuartil, máximo, datos atípicos. Con estos datos podemos determinar si la distribución de los datos es simétrica y tambien darnos una idea de que tanto varian. Por ejemplo, en la siguiente gráfica podemos notar:

- Los datos exponenciales no provienen de una distribución simétrica
- Es dificil determinar si la varianza es adecuada para una normal, o bien tiene colas más pesadas como la t-student

```
datos %>%
  gather(variable, valor) %>%
  ggplot(aes(x = valor, color = variable))+
  geom_boxplot()+
  facet_wrap(~variable)+
  coord_flip()+
  theme_minimal()+
  theme(legend.position = "none")+
  xlim(-5,5)
```

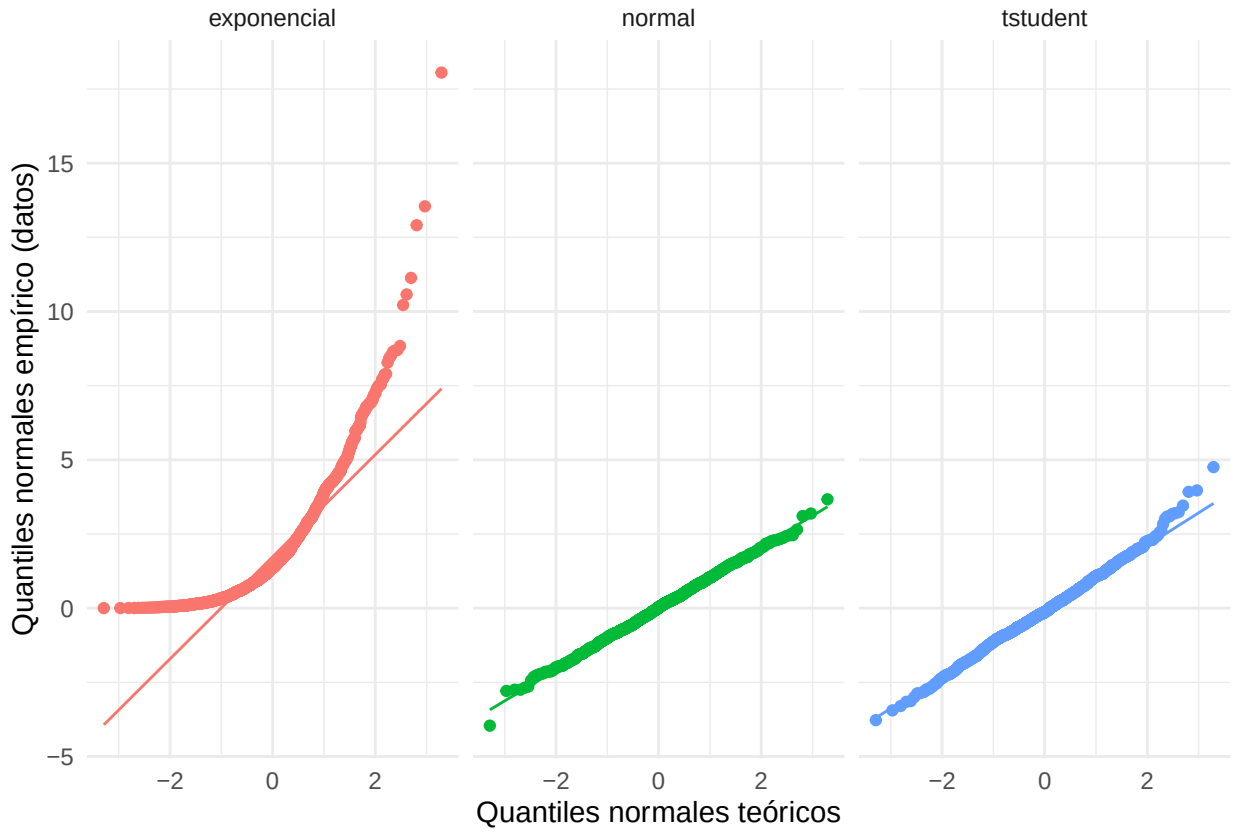


Por esto, el boxplot nos permite disernir si la distribución está sesgada, pero no es sencillo determinar si la varianza corresponde a una distribución normal.

3.3 qqplot (quantile-quantile plot) La intucion detrás del qqplot es que los cuantiles de nuestros datos (quantil empírico) deben estar "en linea" perfecta con los quantiles teóricos de una distribución normal. Para facilitar en entendimiento, veamos las siguientes gráficas:

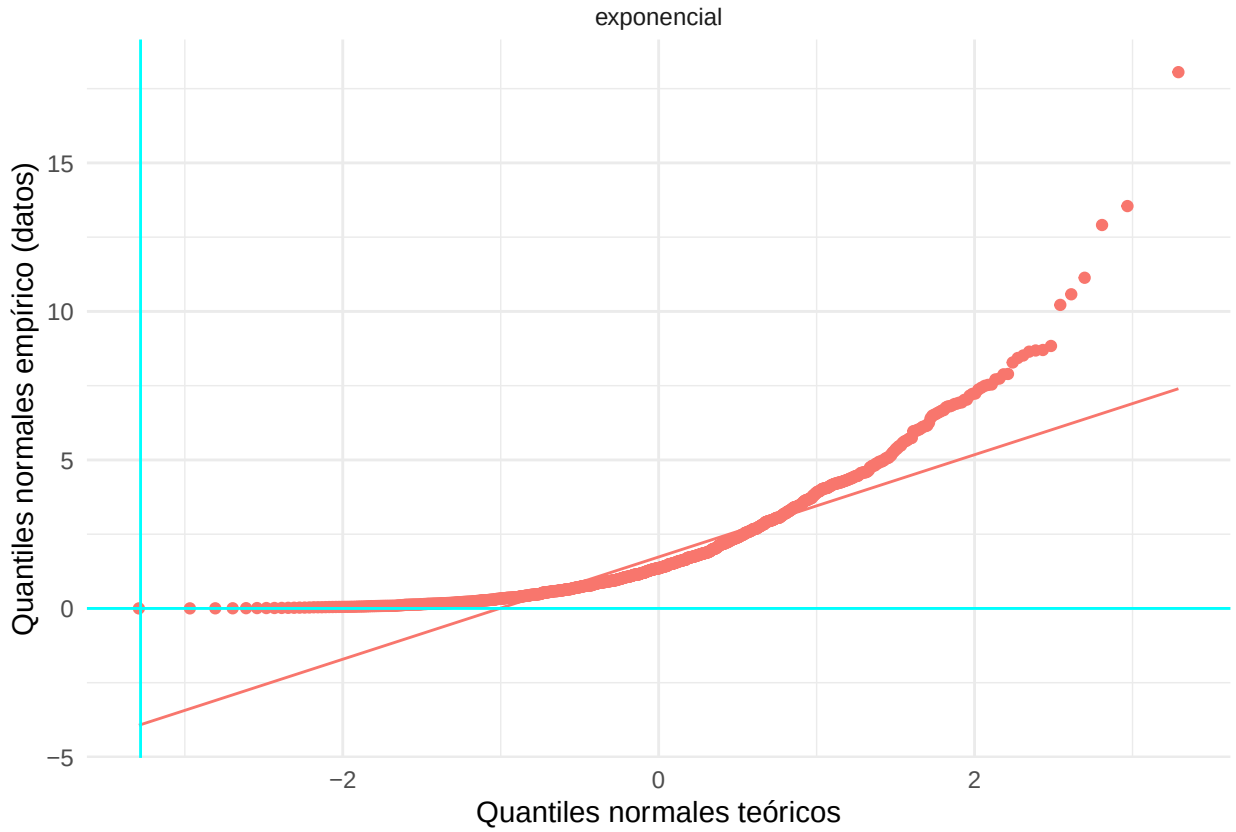
```
datos %>%
  gather(variable, valor) %>%
  ggplot(aes(sample = valor, color = variable))+
  stat_qq() + stat_qq_line()+
  facet_wrap(~variable)+
```

```
  theme_minimal()+
  theme(legend.position = "none")+
  labs(x = "Quantiles normales teóricos",
       y = "Quantiles normales empírico (datos)")
```



En la gráfica aparece una línea y varios puntos. Cada uno de los puntos corresponde a un dato. ¿Cómo se lee esta gráfica? Agarremos un punto:

```
datos %>%
  gather(variable, valor) %>% filter(variable == "exponencial") %>%
  ggplot(aes(sample = valor, color = variable))+
  stat_qq() + stat_qq_line()+
  facet_wrap(~variable)+
  theme_minimal()+
  geom_hline(yintercept = 0, color = "cyan")+
  geom_vline(xintercept = -3.28, color = "cyan")+
  theme(legend.position = "none")+
  labs(x = "Quantiles normales teóricos",
       y = "Quantiles normales empírico (datos)")
```



El dato señalado es el valor empírico 0.0016. Se debe encontrar en que posición estaría si ordenaramos los datos de menor a mayor, en este caso es el mínimo, entonces es el dato 1 de 1000. Ahora, hay que encontrar a que cuantil corresponde la probabilidad 1/1000 de una distribución normal:

```
# el primer dato de los datos exponenciales, que quantil representa
# de datos normales
qnorm(1/1000) # -3.09
```

```
## [1] -3.090232
```

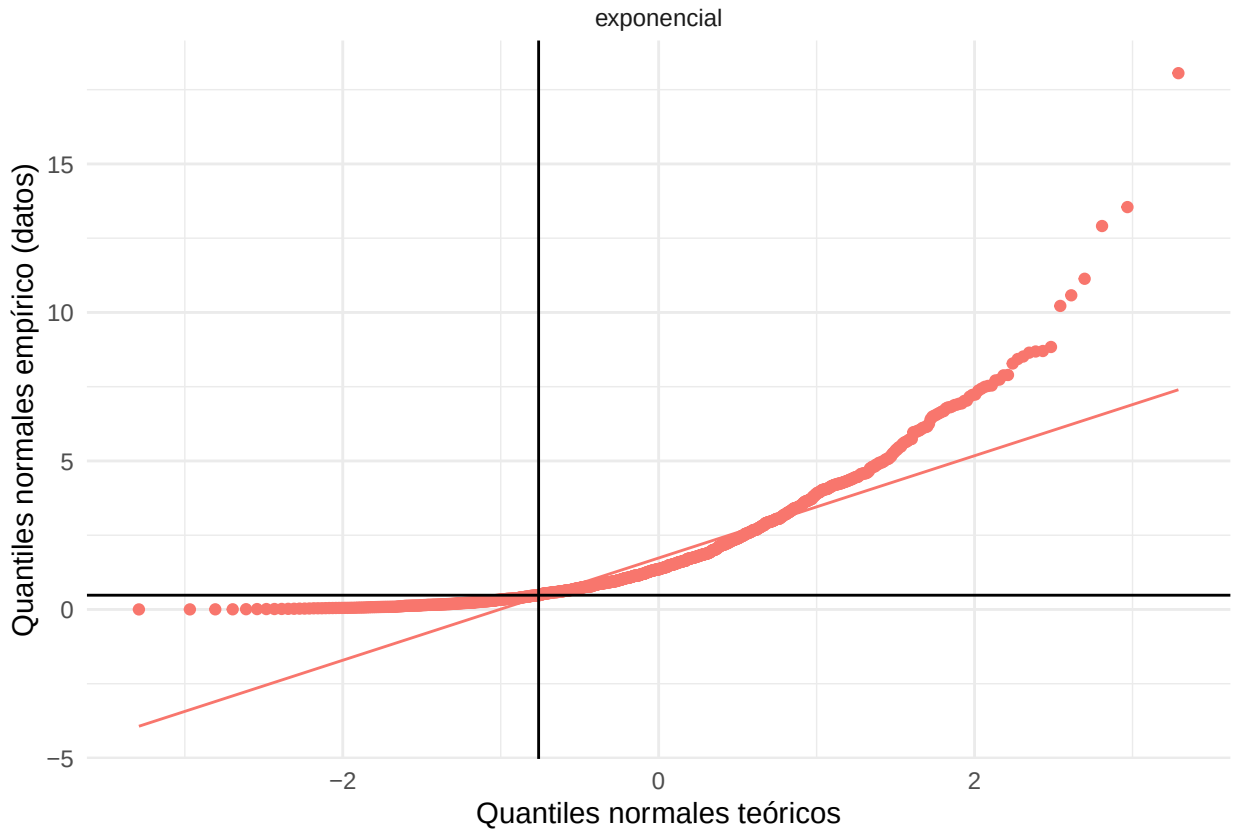
```
# coordenada: (-3.09, 0.0016)
```

De esta manera, podemos generar la coordenada (-3.09, 0.0016), el primer dato corresponde al cuantil de la normal y el segundo al dato empírico. Veamos otro ejemplo:

Tomemos un valor distinto al mínimo, por ejemplo ejemplo el valor 0.48, que es el número 225 de 1000 ordenado de menor a mayor:

```
datos %>%
  gather(variable, valor) %>% filter(variable == "exponencial") %>%
  ggplot(aes(sample = valor, color = variable))+
  stat_qq() + stat_qq_line()+
  facet_wrap(~variable)+
  theme_minimal()+
  geom_hline(yintercept = 0.48)+
```

```
  geom_vline(xintercept = -0.76)+
  theme(legend.position = "none")+
  labs(x = "Quantiles normales teóricos",
       y = "Quantiles normales empírico (datos)")
```




```
qnorm(225/1000) # -0.76
```

```
## [1] -0.755415
```

```
# coordenada: (0.48, -0.76)
```

Vemos que representa el cuantil 225/1000, que representa el quantil teórico normal de -0.04.

Recapitulando, cada punto en la gráfica representa una observación y la línea en cada facet representa como se comportarían las observaciones de una distribución normal. Si los puntos se ajustan a la linea, quiere decir que se asemeja a una normal.

- En el caso de datos normales (línea verde), los puntos coinciden casi al 100% con la línea, lo que nos lleva a pensar que efectivamente se distribuyen normal
- En el caso de los puntos azules, vemos que en centro se parecen los datos a la línea, pero en una de las colas no coinciden. ¿Qué implicaría?:

Por eso, de los métodos gráficos, el qqplot es de los que más utilidad nos generan, ya que no solo nos permite determinar si los datos tienen sesgo 0, si no además, que la varianza es equivalente a la varianza de una distribución normal.

4) Pruebas de Hipótesis

4.1) Prueba de hipótesis Jarque-Bera Los métodos anteriores son gráficos, pero ¿qué sucede si queremos una manera más estadística para llegar a una conclusión? R: aplicamos una prueba de hipótesis

La prueba de hipótesis que aplicamos recibe el nombre de Jarque-Bera, la cual tiene como H_0 : Los datos se distribuyen normal vs H_1 : Los datos no se distribuyen normal. El estadístico lo llamaremos JB:

$$JB = \frac{n}{6}(S_k^2 + \frac{1}{4}(K-3)^2)$$

Donde en este caso llamaremos S_k al coeficiente de asimetría y K a la curtosis.

$$S_k = \frac{E(X-\mu)^2}{\sigma^3}$$
$$K = \frac{E(X-\mu)^4}{(E(X-\mu)^2)^2}$$

El estadístico JB se distribuye de como una $JB \sim \chi^2_2$ y siempre rechazamos para valores altos de la distribución.

```
library(tidyverse)
library(moments)
library(tseries)

# Jarque-Bera
nrows = nrow(datos)

# normal
skew_normal = (sum((datos$normal-mean(datos$normal))**3)/nrows) / (var(datos$normal)*(nrows-1)/nrows)**

kurt_normal =  (sum((datos$normal-mean(datos$normal))**4)/nrows) / (var(datos$normal)*(nrows-1)/nrows)*

JB_normal = nrows*((skew_normal^2)/6 + (kurt_normal-3)*(kurt_normal-3)/(24))
pchisq(JB_normal, lower.tail = F, df = 2)
```

4.1.1) Cálculo con fórmulas:

```
## [1] 0.9900528
```

```
# valorp = 0.99 -> No Rechazamos H0

# t-student
skew_t = (sum((datos$tstudent-mean(datos$tstudent))**3)/nrows) / (var(datos$tstudent)*(nrows-1)/nrows)*

kurt_t =  (sum((datos$tstudent-mean(datos$tstudent))**4)/nrows) / (var(datos$tstudent)*(nrows-1)/nrows)

JB_t = nrows*((skew_t^2)/6 + (kurt_t-3)*(kurt_t-3)/(24))

pchisq(JB_t, lower.tail = F, df = 2)
```

```
## [1] 0.0007907831
```

```
# valorp = 0.0007 -> Rechazamos H0

# exponencial
skew_exp = sum((datos$exponencial-mean(datos$exponencial))**3)/nrows / (var(datos$exponencial)*(nrows-1)

kurt_exp =  (sum((datos$exponencial-mean(datos$exponencial))**4)/nrows) / (var(datos$exponencial)*(nrows

JB_exp = nrows*((skew_exp^2)/6 + (kurt_exp-3)*(kurt_exp-3)/(24))
pchisq(JB_exp, lower.tail = F, df = 2)
```

```
## [1] 0
```

```
# valorp = 0 -> Rechazamos H0
```

```
# formulas para sesgo y kurtosis
skewness(datos$normal)
```

4.1.2) Cálculo con paquetes:

```
## [1] 0.004523786
```

```
kurtosis(datos$normal)
```

```
## [1] 3.01995
```

Jarque bera con función

```
# coinciden con valores de arriba
jarque.bera.test(datos$normal)
```

```
##
##  Jarque Bera Test
##
## data:  datos$normal
## X-squared = 0.019994, df = 2, p-value = 0.9901
```

```
jarque.bera.test(datos$tstudent)
```

```
##
##  Jarque Bera Test
##
## data:  datos$tstudent
## X-squared = 14.285, df = 2, p-value = 0.0007908
```

```
jarque.bera.test(datos$exponencial)
```

```
##
##  Jarque Bera Test
##
## data:  datos$exponencial
## X-squared = 2802.6, df = 2, p-value < 2.2e-16
```

5) Ejercicio:

Se usarán los datos de lecturas anteriores de house_rent. El modelo a ajustar es:

$$rent \sim size$$

. Es decir nos gustaría poder ajustar el precio solamente con el tamaño de la casa.

5.1) Ajusta la regresión lineal y determina si la β_0, β_1 y R^2 son significativas: Aunque $R^2 = 0.17$, vemos que el valorp es cercano a 0, por lo que podemos rechazar $H_0 : P^2 = 0$. Entonces si es significativa. La misma conclusión se puede derivar de β_0 y β_1.

Los residuos los podemos calcular:

5.2) Realiza un histograma para mostrar la distribución de los errores de la regresión Se puede observar que hay un ligero sesgo a la derecha, pero realmente con la gráfica no podemos concluir

5.3) Realiza un boxplot para mostrar la distribución de los errores de la regresión En esta gráfica es un poco más complicado determinar si hay un sesgo o incluso si corresponde a una varianza de una distribución normal

5.4) Realiza un qqplot para mostrar la distribución de los errores de la regresión Con la gráfica podemos observar que tiene colas más pesadas que una distribución normal, por lo que probablemente el supuesto de normalidad no se cumpla.

5.5) Realiza una prueba Jarque-Bera para mostrar la distribución de los errores de la regresión Por lo tanto rechazamos que los datos sean normales.