Actividad 1.2 Normalidad univariada

Ricardo Kaleb Flores Alfonso

2024-09-20

0) Se importan las librerias

0.1) Se importa la base de datos

```
data(cars)
```

1) Pruebas de normalidad

1.1) Prueba de Anderson Darling

Dada la hipotesis que:

- H_0 := Los datos siguen una distribución normal
- H_A := Los datos no siguen una distribución normal

```
ad.test(cars$dist)
```

```
##
##  Anderson-Darling normality test
##
## data:  cars$dist
## A = 0.74067, p-value = 0.05021
```

Dado el valor de $p = 0.05021 > 0.05$ no hay suficiente evidencia para concluir que los datos no siguen una distribución normal. Esto significa que los datos en distancia siguen una distribución normal. Sin embargo dado que apenas pasa la prueba, es posible que no sea una distribución normal.

```
ad.test(cars$speed)
```

```
##
##  Anderson-Darling normality test
##
## data:  cars$speed
## A = 0.26143, p-value = 0.6927
```

Dado el valor de $p = 0.0.6927 > 0.05$ no hay suficiente evidencia para concluir que los datos no siguen una distribución normal. Esto significa que los datos en velocidad siguen una distribución normal.

1.2) Prueba de Kolmogorov Smirnov

Dada la hipotesis que:

- H_0 := Los datos siguen una distribución normal
- H_A := Los datos no siguen una distribución normal

```r
lillie.test(cars$dist)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cars$dist
## D = 0.12675, p-value = 0.04335
```

Dado que el valor de $p = 0.04335 < 0.05$ significa que hay evidencia suficiente para rechazar la hipotesis nula, por lo que los datos no siguen una distribución normal. Comparado con el resultado de la prueba de Anderson-Darling, se concluye que los datos no siguen una distribución normal.

```r
lillie.test(cars$speed)
```

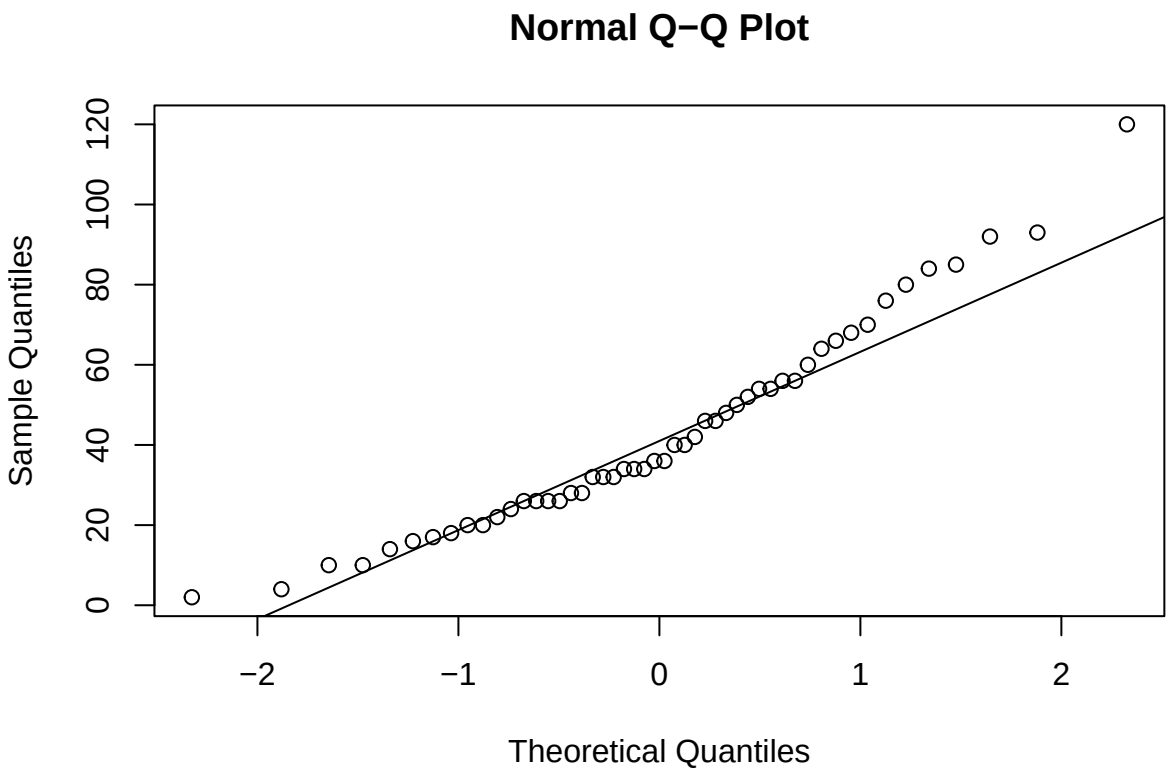
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cars$speed
## D = 0.068539, p-value = 0.8068
```

Dado que el valor de $p = 0.8068 > 0.05$ entonces no hay evidencia suficiente para rechazar la hipotesis nula, por lo que se asume que los datos tienen una distribución normal.

2) Elabora el QQPlot de cada variable

2.1) QQPlot de distancia

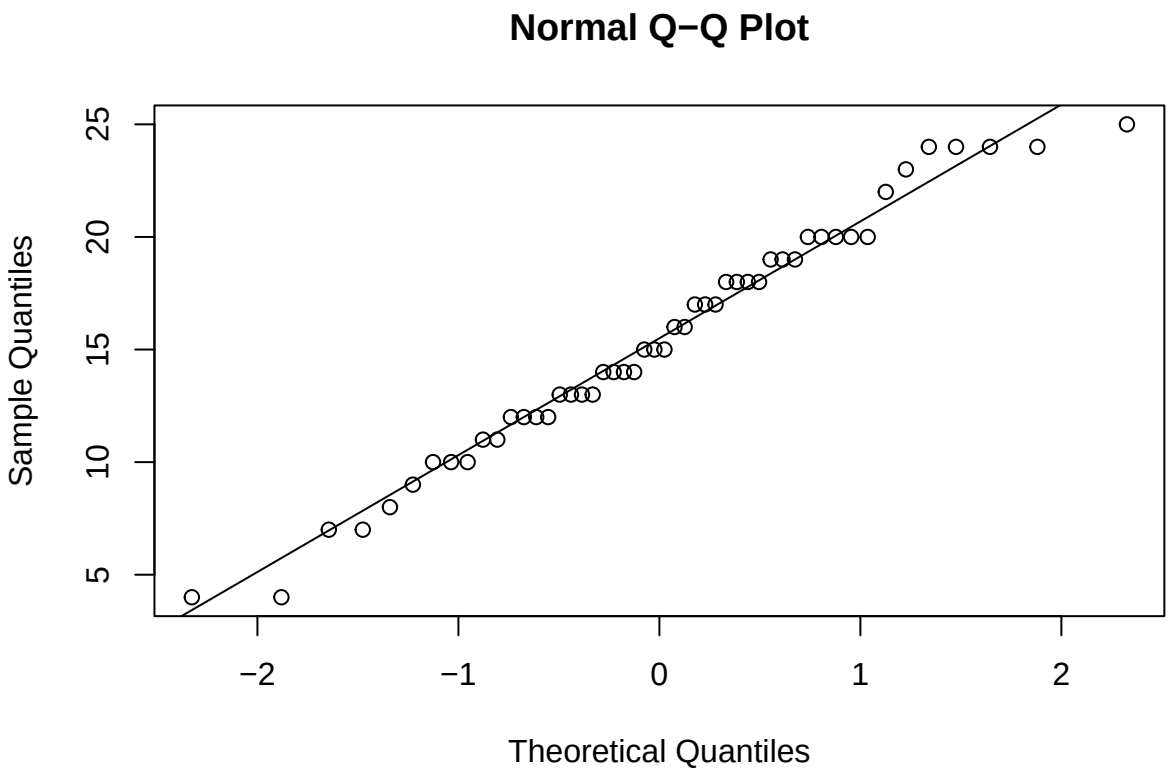
```r
qqnorm(cars$dist)
qqline(cars$dist)
```



Con este gráfico podemos observar como los datos de distancia siguen una distribució normal en la parte central, sin embargo los valores extremos se desvían significativamente. Esto sugiere la presencia de colas largas, lo cual afecta el análisis de valores extremos

2.2) QQPlot de velocidad

```
qqnorm(cars$speed)
qqline(cars$speed)
```



Con este gráfico podemos observar como los datos de velocidad siguen una distribució normal en la mayoria de los casos.

3) Calcula el sesgo y curtosis de cada variable

3.1) Sesgo y curtosis de distancia

```
print(paste("Sesgo de la distancia:",skewness(cars$dist)))
```

```
## [1] "Sesgo de la distancia: 0.782483517311497"
```

```
print(paste("Curtosis de la distancia:",kurtosis(cars$dist)-3))
```

```
## [1] "Curtosis de la distancia: 0.248018657170519"
```

El valor de sesgo obtenido para distancia es positivo, lo que significa que existe una mayor cantidad de datos del lado derecho del promedio. Sin embargo la curtosis obtenida es mayor a 0, esto se interpreta como una curva leptocurtica, lo cual indica que hay más datos en las colas comparado con una distribución normal. Esto nos da una mayor cantidad de datos extremos. ## 3.2) Sesgo y curtosis de velocidad

```
print(paste("Sesgo de la velocidad:",skewness(cars$speed)))
```

```
## [1] "Sesgo de la velocidad: -0.113954770128283"
```

```
print(paste("Curtosis de la velocidad:",kurtosis(cars$speed)-3))
```

```
## [1] "Curtosis de la velocidad: -0.577147423943737"
```

El valor de sesgo obtenido para velocidad es negativo cercano a 0, lo que significa que exisen más datos del

lado izquierdo del promedio, sin embargo esto no afecta tanto al modelo. Sin embargo la curtosis obtenida es menos que 0, esto se interpreta como una curva platicurtica, lo cual indica que hay menos datos en las colas comparado con una distribución normal. Esto nos da una menor cantidad de datos extremos.

4) Calcula la media, mediana y rango para cada variable

4.1) Calculo de medida, mediana y rango medio de distancia

```
print(paste("Mediana de la distancia:",median(cars$dist)))
```

```
## [1] "Mediana de la distancia: 36"
```

```
print(paste("Media de la distancia:", mean(cars$dist)))
```

```
## [1] "Media de la distancia: 42.98"
```

```
print(paste("Rango medio de la distancia:", (max(cars$dist) - min(cars$dist))/2))
```

```
## [1] "Rango medio de la distancia: 59"
```

La mediana de 36 representa el valor tipico de distancia, dado que el promedio es 42.98 significa que existen datos con valores altos que mueven la media hacia arriba, esto genera sesgo hacia la derecha. El rango de 59 significa que hay un mayor rango en el que se distribuyen los datos.

4.2) Calculo de medida, mediana y rango de velocidad

```
print(paste("Mediana de la velocidad:",median(cars$speed)))
```

```
## [1] "Mediana de la velocidad: 15"
```

```
print(paste("Media de la velocidad:", mean(cars$speed)))
```

```
## [1] "Media de la velocidad: 15.4"
```

```
print(paste("Rango medio de la velocidad:", (max(cars$speed) - min(cars$speed))/2))
```

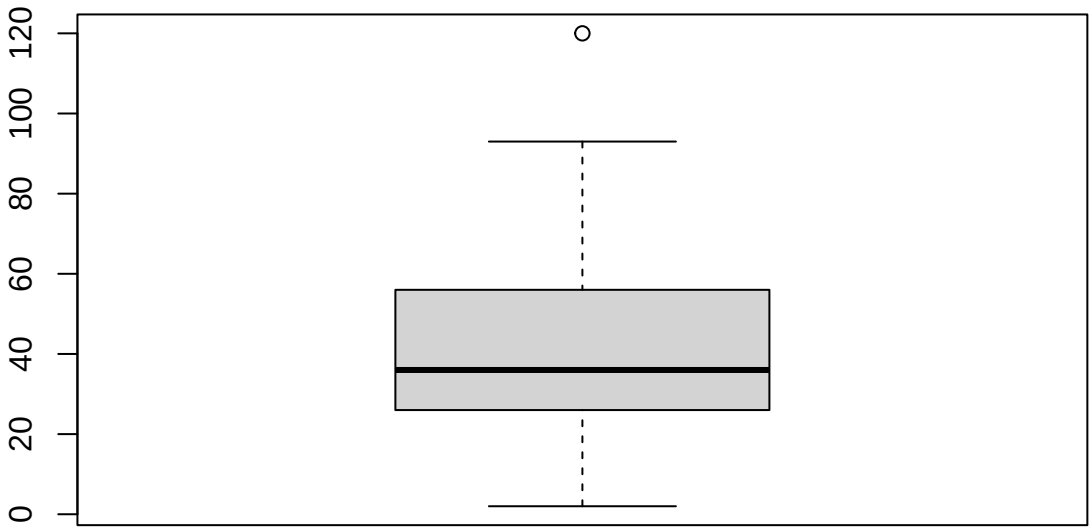
```
## [1] "Rango medio de la velocidad: 10.5"
```

Con un valor de 15 como mediana y 15.4 como media se presenta que la mayoria de los datos se distribuyen de manera simetrica, asi como un valor de 10.5 como rango media muestra una menor variabilidad que los datos de distancia. A pesar de esto los datos estaran ligeramente sesgados hacia la derecha.

5) Elabora los gráficos de caja y bigotes

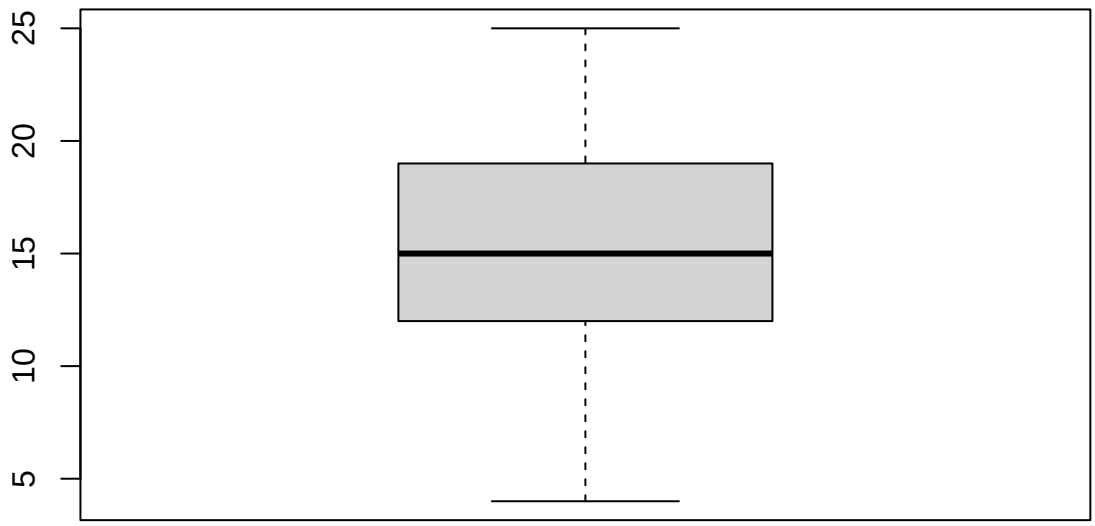
5.1) Grafico de distancia

```
boxplot(cars$dist)
```



El gráfico de caja y bigotes de la distancia muestra lo ya analizado en el punto previo de manera gráfica. Se observa como existe un gran rango donde se encuentran los datos, e incluso hay datos que se salen de este rango. Por el gráfico se comprueba que existe una mayor cantidad de datos del lado derecho de la media.
5.2 Grafico de velocidad

```
boxplot(cars$speed)
```

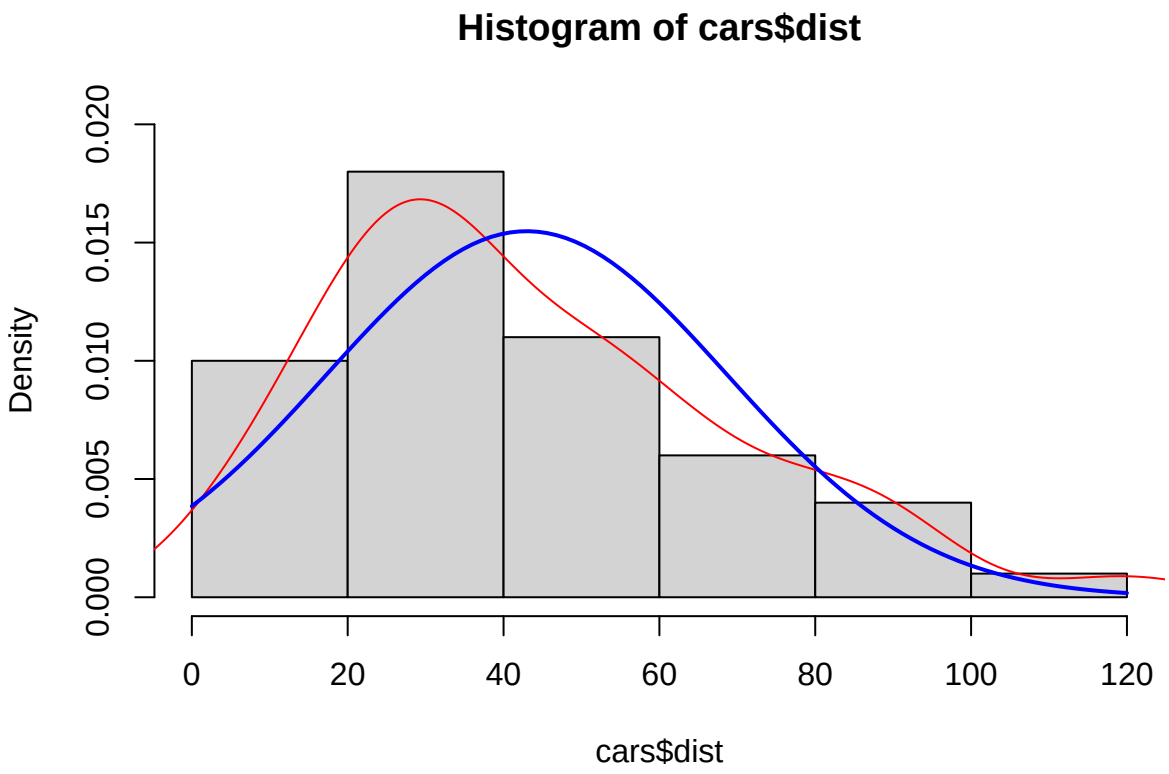


El gráfico de caja y bigotes de la velocidad muestra lo ya analizado en el punto previo de manera gráfica. Se observa como existe un rango relativamente igual del lado derecho e izquierdo del promedio donde se encuentran los datos, asi como la media y el promedio se encuentra casi en el mismo valor lo que significa que se tiene una distribución normal de datos.

6) Realiza el histograma y su distribución teórica

6.1) Histograma de distancia

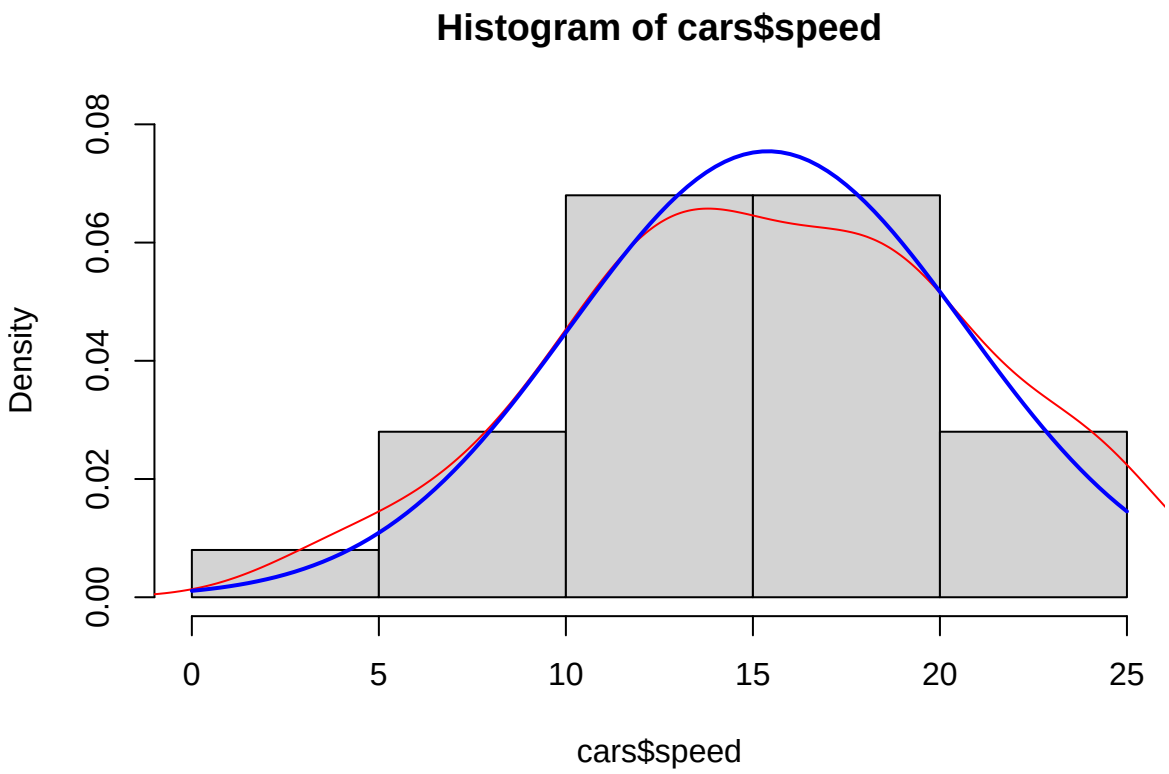
```
hist(cars$dist,freq = FALSE,ylim=c(0,0.02))
lines(density(cars$dist),col="red")
curve(dnorm(x,mean=mean(cars$dist),sd=sd(cars$dist)),from=0,to=120,add=TRUE,col="blue",lwd=2)
```



Con el histograma de distancia se confirma que la distribución de los datos no sigue una curva normal, hay una mayor cantidad de datos del lado derecho y los datos se distribuyen en un gran rango.

6.2) Histograma de velocidad

```
hist(cars$speed,freq = FALSE, ylim=c(0,0.08))
lines(density(cars$speed),col="red")
curve(dnorm(x,mean=mean(cars$speed),sd=sd(cars$speed)),from=0,to=25,add=TRUE,col="blue",lwd=2)
```



Con el histograma de velocidad se confirma que la distribución de los datos sigue una curva normal, hay una mayor cantidad de datos cerca del promedio y los datos se distribuyen en un rango menor al obtenido en distancia

Conclusión

En este análisis de normalidad univariada aplicado al conjunto de datos cars, se usaron diversas pruebas estadísticas para evaluar la distribución de las variables distancia y velocidad.

Las pruebas de Anderson-Darling y Kolmogorov-Smirnov (Lilliefors) arrojaron resultados distintos para la variable distancia. Mientras que la prueba de Anderson-Darling no proporcionó suficiente evidencia para rechazar la hipótesis de normalidad, la prueba de Kolmogorov-Smirnov indicó lo contrario. Esto se debe a la sensibilidad de cada prueba a diferentes características de los datos, como la presencia de valores extremos. Para el caso de velocidad, ambas pruebas coincidieron en que los datos siguen una distribución normal.

Los QQPlots y gráficos de caja y bigotes complementaron el análisis al mostrar que los valores extremos de la distancia se desvían notablemente, lo que sugiere colas largas. Esto se reflejaa también en los valores de sesgo y curtosis, donde la variable distancia tuvo un sesgo positivo y una curtosis leptocúrtica, confirmando la presencia de más datos extremos en las colas. Por otro lado la velocidad mostró una distribución más simétrica y con menor presencia de valores extremos, apoyada por su sesgo y curtosis cercanos a 0.

Es por esto que aunque la variable velocidad presenta características más cercanas a una distribución normal, la distancia muestra desviaciones significativas en los extremos. Esto debe tomarse en cuenta en cualquier análisis donde se dependa de la suposición de normalidad, ya que influirá en la precisión de los resultados.