Tarea Estadística

Naira Carruccio Villada

Ejercicio 1

a)

Obtenemos las medidas de centralización y de dispersión.

Table 1: Medidas de centralización

Grupo	Media	Mediana	Moda
1	123.4250	124.5	131
2	127.4833	126.0	121

Table 2: Medidas de dispersión

Grupo	Mínimo	Máximo	Rango.interc	Desviación.típica	Varianza	CV
1	105	142	11.5	9.026705	9.026705	0.0731351
2	96	154	19.4	13.888773	13.888773	0.1089458

La representatividad de la media en el conjunto de datos viene dado por el Coeficiente de Variación (CV), ya que muestra el número de veces que la desviación típica contiene a la media. Cuánto mayor es el coeficiente de variación, más valores dispersos habrá y por tanto, su media es menos representativa.

Como podemos observar en la Tabla 2, el Grupo 1 tiene un menor valor del Coeficiente de Variación con un 7.31% respecto al del Grupo 2 que tiene un 10,89%, igual en ambos la media es representativa, al ser menores del 30%. Sin embargo, la media más representativa entre los grupos es la perteneciente al Grupo 1, ya que contiene menor dispersión.

b)

Table 3: Estudio de la simetría y la curtosis en el Grupo 2

Periodo	Simetría	Curtosis
Antes	-0.1332	-0.6514
Después	-0.1889	0.2408

Existe una asimetría negativa en los dos periodos ya que su coeficiente es menor que cero. Además, podemos comentar que esta ligera asimetría puede ser debida por la presencia de valores atípicos por la izquierda que desplazan la media hacia ese lado, siendo la media menor que la mediana.

En cuánto a la curtosis, en el primer periodo sigue una distribución Platicúrtica siendo más aplanada que la normal ya que su coeficiente es menor que cero. En el periodo siguiente, la distribución es Leptocúrtica siendo más apuntada de lo normal porque el coeficiente es mayor que 0.

c)

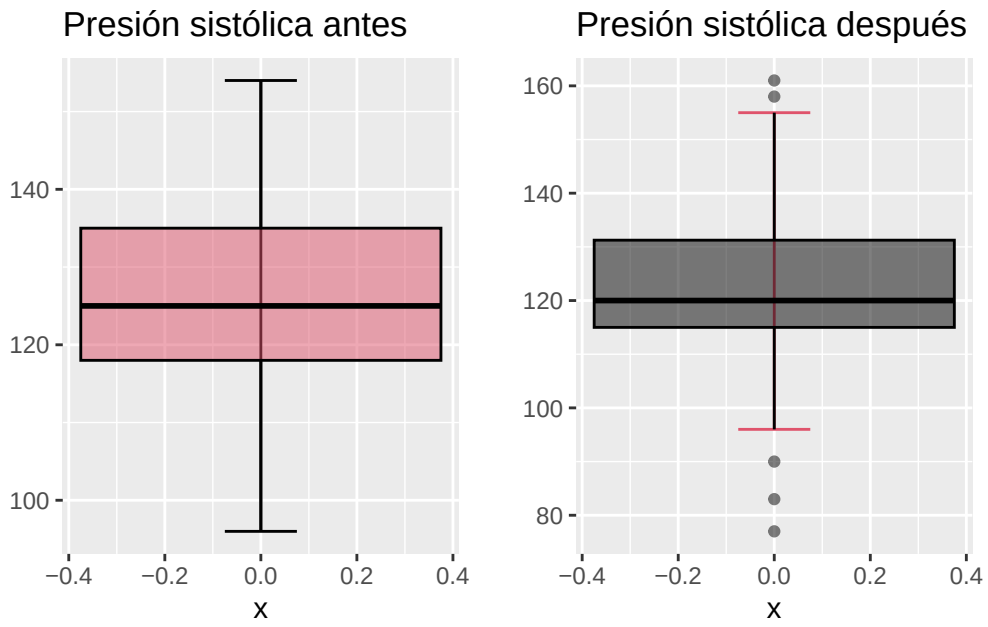
Table 4: Cuartiles de ambos periodos

Periodo	Primer.cuartil	Mediana	Tercer.cuartil	Rango.intercuartil
Antes	118	125	135.00	17.00
Después	115	120	131.25	16.25

Los cuartiles son los valores que dividen a la distribución en cuatro partes iguales. El primer cuartil representa que el 25% de los datos de los pacientes son menores o iguales a tener una presión sistólica de 118 antes de tomar el medicamento. Y después de 60 minutos de tomar el medicamento, de 115. En el segundo cuartil o mediana, antes de tomar el medicamento el 50% de los datos son menores o iguales a una presión sistólica de 125. Y después de tomar el medicamento, es de 120. En el tercer cuartil, observamos que antes de tomar el medicamento el 75% de los datos son menores o iguales a una presión sistólica de 135. Mientrás que después de tomar el medicamento, son menores o iguales a 131.25.

El rango intercuartil es la diferencia entre el tercer cuartil y el primero. El 50% de los datos se encuentran entre 118 y 135 antes de tomar el medicamento. Y después de tomar el medicamento, entre 115 y 131.25.

Boxplots



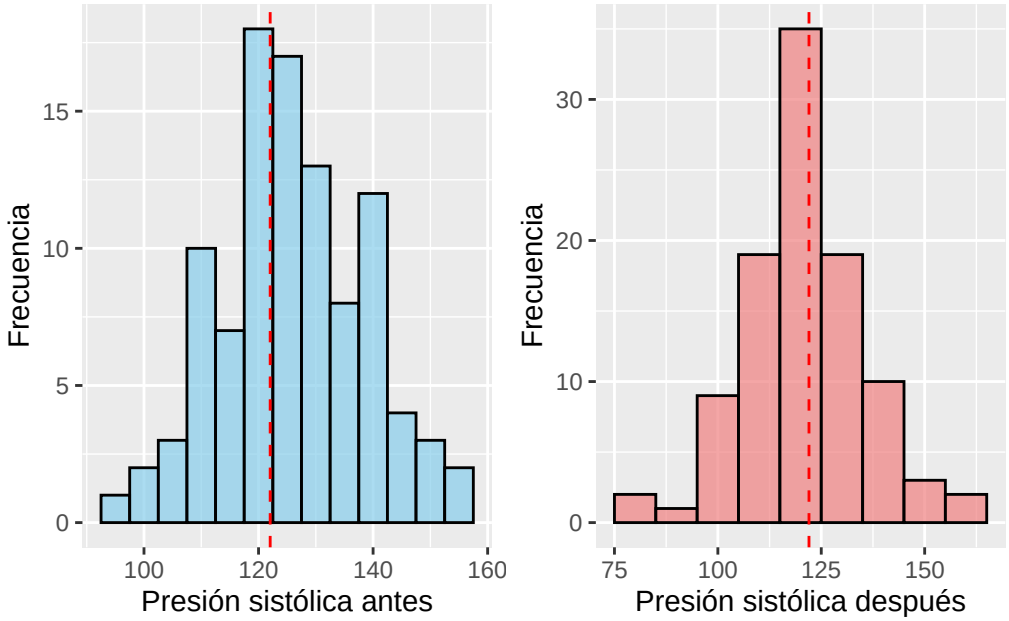
Al observar los diagramas de cajas, podemos determinar que en la variable presión sistólica anterior no hay valores atípicos, es decir, no hay ningún valor que se quede fuera del límite inferior (92.5) y del límite superior (160.5). Sin embargo, en la variable presión sistólica después hay valores atípicos que se quedan fuera del límite inferior (90.625) y del límite superior (155.625).Siendo exactos, los valores atípicos en el límite inferior pueden ser 90, 82, 76. Y en el límite superior pueden ser 158 y 160.8.

d)

Análisis de normalidad mediante gráficos:

En primer lugar, observaremos el histograma de ambas variables, presión sistólica antes y después.

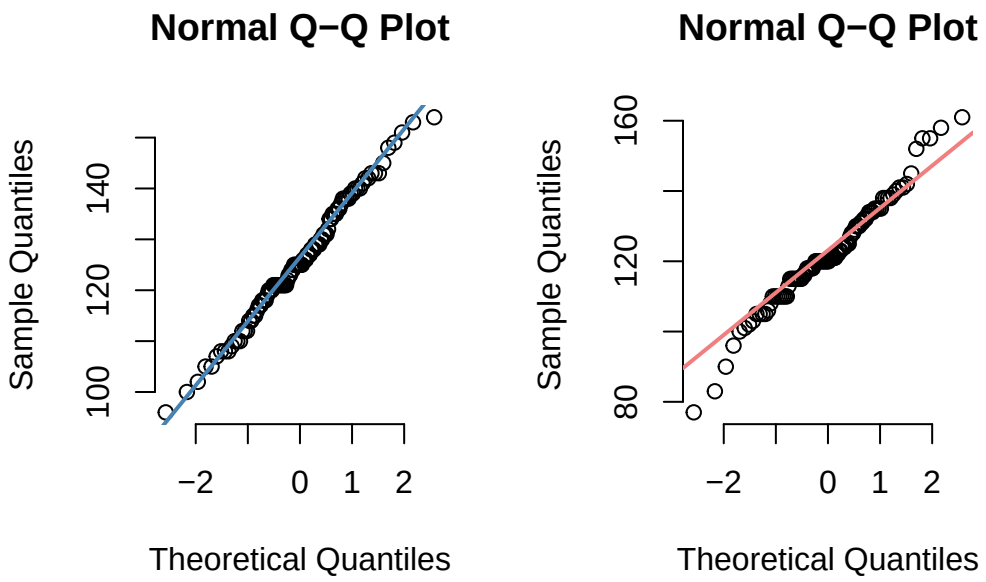
Histogramas de la presión sistólica antes y después



A primera vista, podemos deducir que el histograma de presión sistólica antes no es simétrico. Es decir, los datos no son simétricos alrededor de la media, la línea roja del histograma, y no tiene ningún patrón discernible por su forma aleatoria. Sin embargo, en el histograma de la presión sistólica después, los datos son más simétricos alrededor de la media, tiene una forma de campana. Además, solo coincide en la presión sistólica despúes que el centro del histograma coincida con la media. En cuánto a la dispersión, observamos que la presión sistólica antes tiene más datos alejados de la media que cerca de ella. Y en la presión sistólica después sucede de la manera contraria.

Una distribución normal tiende a ser simétrica en torno a la media, tiene el pico de datos en el centro del histograma y tiene más datos cerca de la media y menos a medida que te alejas. Por tanto, al comparar los histogramas, podemos deducir que es menos probable que la presión sistólica antes tenga una distribución normal. Y es más probable que la presión sistólica despuíes tenga una distribución normal.

QQ plot presión sistólica antes y después



El gráfico cuantil-cuantil, nos ayuda a poder comparar los cuantiles observados con los esperados de una distribución normal. Al alinearse los puntos aproximadamente en la línea recta, sugiere que hay una distribución normal. Por tanto, podemos sugerir que ambas variables siguen una distribución normal siendo la presión sistólica antes más notoria.

Contraste de hipótesis

Para concluir, el análisis de normalidad utilizamos el contraste de hipótesis.

Table 5: Tests de Normalidad : Presión sistólica antes

	Test	Statistic	Pvalue
W	Shapiro-Wilk	0.9919231	0.8156294
D	Kolmogorov-Smirnov	0.0537389	0.9349136

Table 6: Tests de Normalidad : Presión sistólica después

	Test	Statistic	Pvalue
W	Shapiro-Wilk	0.9788826	0.1087456
D	Kolmogorov-Smirnov	0.0878265	0.4234303

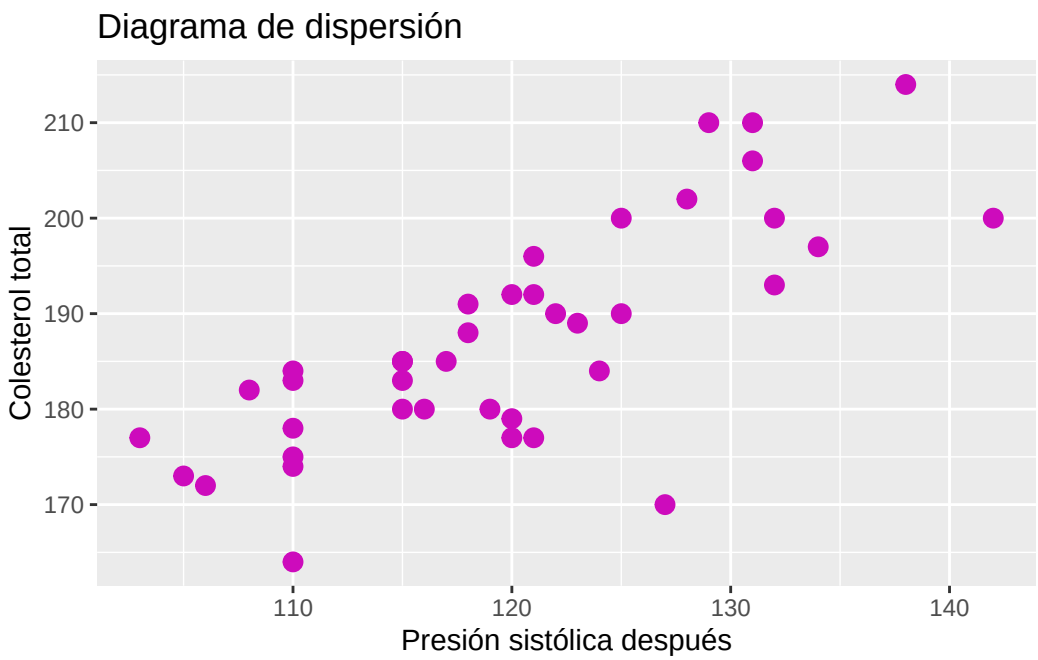
El test de Shapiro-Wilk y Kolmogorov-Smirnov mantienen algunas diferencias que pueden inferir en el resultado. El test de Shapiro-Wilk es más sensible para detectar desviaciones de la normalidad y es especializado para muestras pequeñas. Sin embargo, el test de Kolmogorov-Smirnov se ve menos afectado por el tamaño de la muestra y se limita el resultado a trabajar con valores idénticos.

Podemos determinar que ambas variables no rechazarían la hipótesis de tener una distribución normal, ya que sus P-valores superan el nivel de significación (0.05) en los dos tests estudiados. Es notoria la diferencia del P-valor en las variables, en la presión sistólica antes se acepta con 0.815/0.934 y en la presión sistólica después con 0.108/0.423.

En el caso, si dieran respuestas diferentes a la hipótesis, nos decantaríamos por el test de Shapiro-Wilk aunque sea para muestras pequeñas y la muestra de nuestro análisis sea 100, existen valores idénticos en nuestros datos que limitan el resultado del test de Kolmogorov-Shapiro. Además, el test de Kolmogorov-Smirnov es poco potente y pierde sensibilidad. Para que aumente su potencia se puede usar la correción de Lillefors.

Ejercicio 2

a)



Coef.Correlación
0.7671351

Como podemos ver en el Diagrama de dispersión, observamos una correlación alta positiva ya que los puntos de la gráfica tienen una tendencia hacia la derecha. Además, al interpretar el coeficiente de correlación (0.767) podemos afirmar que la presión sistólica despúes de la toma del medicamento y el colesterol total tienen una correlación positiva y fuerte.

b)

Modelo a estimar:

$$PRESIONDESP = b_1 * colesterol + b_0 + error$$

Estimamos el modelo que explica la Presión sistólica después en función del colesterol total.

```
Call:
lm(formula = `Presión sistólica después` ~ `Colesterol total`,
    data = Grupo1)

Residuals:
    Min      1Q  Median      3Q     Max
-10.717  -3.740  -1.705   3.640  17.536

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.16987   15.45740   0.399    0.692
`Colesterol total`  0.60761    0.08242   7.372 7.76e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.066 on 38 degrees of freedom
Multiple R-squared:  0.5885,    Adjusted R-squared:  0.5777
F-statistic: 54.34 on 1 and 38 DF,  p-value: 7.765e-09
```

$$PRESIONDESP = 0.61 * colesterol + 6.17$$

$$PRESIONDESP = 0.61 * 105 + 6.17 = 70.22$$

La estimación para un paciente joven cuyo colesterol total es 105 mg/Dl de la Presión sistólica despues de ingerir el medicamento es de 70.22.

c)

El coeficiente de determinación R cuadrado, indica el porcentaje de variación de la variable dependiente Presión sistólica despúes que se explica por la relación lineal con la variable independiente Colesterol total. Como podemos observar en la tabla anterior, hay dos tipos de R cuadrado, el R cuadrado y el ajustado. Si interpretamos el R cuadrado sin ajustar nos da que la variable colesterol total explica un 58.85% la presión sistólica después. No obstante, el r cuadrado sin ajustar tiende a sobreestimar el modelo si hay predictores irrelevantes. Por ello, usaremos el R cuadrado ajustado que elimina los predictores irrelevantes siendo una bondad de ajuste más precisa, donde la variable colesterol total explica un 57.78% la presión sistólica después.

d)

El valor del coeficiente de Colesterol Total en el modelo de regresión lineal es de 0.61, este valor indica que por cada aumento adicional del colesterol total la Presión sistólica después aumentaría 0.61 mmHg. Por tanto, si se aumenta el colesterol de un paciente en 5mg/DL, la presión sistólica aumentaría en proporción a esta ecuación:

$$PRESIONDESP = 0.61 * 5 + B_0$$

Es decir, la Presión sistólica después aumentaría un 3.05 mmHg.

Ejercicio 3

a)

En este apartado, el método que utilizaremos en la inferencia estadística es el intervalo de confianza para la media de una población con distribución normal.
Primero, obtenemos el intervalo de confianza al 95% y al 99% para el nivel medio de presión sistólica antes de la toma del medicamento.

Intervalo de confianza al 95%

```
    One Sample t-test

data:  Grupo2$`Presion sistólica antes`
t = -1.4036, df = 59, p-value = 0.1657
alternative hypothesis: true mean is not equal to 130
95 percent confidence interval:
 123.8955 131.0712
sample estimates:
mean of x
 127.4833
```

Intervalo de confianza al 99%

```
    One Sample t-test

data:  Grupo2$`Presion sistólica antes`
t = -1.4036, df = 59, p-value = 0.1657
alternative hypothesis: true mean is not equal to 130
99 percent confidence interval:
 122.7107 132.2560
sample estimates:
mean of x
 127.4833
```

Tras los resultados, podemos observar que para el intervalo de confianza al 95% es (123.8955,131.0712). Es decir, la media se encuentra entre 123.8955 y 131.0712 al 95%. Por tanto, la media propuesta podría ser válida ya que 130 mm de Hg entra dentro del intervalo. De hecho, al realizar el contraste de hipótesis observamos que el p-valor (0.1657) es mayor del nivel se significación (0.05). Por tanto, aceptaríamos la hipótesis nula donde la media de la presión sistólica antes de la toma del medicamento para la población adulta puede tomar valor de 130 mm de Hg.

En cuanto, al intervalo de confianza al 99% es (122.7107,132.2560), la media se encuentra dentro de este intervalo. Por ende, la media propuesta de 130 mm de Hg está dentro del intervalo. Hecho que se respalda con el contraste de hipótesis ya que el P-valor (0.1657) es mayor que el nivel de significación (0.01), no rechazando la hipótesis nula.

b)

Para poder observar si la presión sistólica media es distinta dependiendo de la edad, realizamos un intervalo de confianza de diferencia de medias al 95%.

Intervalo de confianza diferencia de medias

```
    Welch Two Sample t-test

data:  Grupo1$`Presión sistólica después` and Grupo2$`Presión sistólica después`
t = -1.3211, df = 94.344, p-value = 0.1897
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.885275  1.785275
sample estimates:
mean of x mean of y
   119.90    123.45
```

El intervalo de diferencia de medias en la presión sistólica después entre adultos y jóvenes es (-8.885,1.7852). Si las medias fueran iguales, la diferencia sería cero. Por tanto, como 0 está dentro del intervalo podemos argumentar que no hay diferencia significativa entre las medias de los adultos y jóvenes. En otras palabras, la edad no influye significativamente en la presión después de la ingesta del medicamento.

c)

Primero calcularemos la proporción de la población con una presión sistólica inicial igual o superior a 130 mm de Hg (prehipertensión)

```r
prop_130 <- mean(HIPERTENSION$`Presion sistólica antes` >= 130)
prop_130
```

```
[1] 0.35
```

El 35% es la proporción de la muestra que tiene una presión sistólica inicial igual o superior a 130 mm de Hg.

Intervalo de confianza al 99%, proporción de 0.35.

```
    1-sample proportions test with continuity correction

data:  35 out of 100, null probability 0.5
X-squared = 8.41, df = 1, p-value = 0.003732
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.2356764 0.4837232
sample estimates:
   p
0.35
```

El intervalo de confianza al 99% de la proporción de la población con hipertensión es de (0.2357, 0.4837). Es decir, se estima al 99% que la proporción de población con hipertensión está entre el 23.57% y el 48.37%. Por tanto, el 35% de la proporción de población está dentro del intervalo de confianza y podría ser válido.

Contraste de hipótesis e intervalo de confianza al 95% para una proporción del 30%

```
    1-sample proportions test with continuity correction

data:  35 out of 100, null probability 0.3
X-squared = 0.96429, df = 1, p-value = 0.3261
alternative hypothesis: true p is not equal to 0.3
95 percent confidence interval:
 0.2591235 0.4525560
sample estimates:
   p
0.35
```

Al realizar el contraste de hipótesis si la proporción de población con hipertensión fuera del 30%, la aceptaríamos ya que el P-valor ,0.3261, es mayor que 0.05. Es decir, sería posible que la proporción de población con hipertensión sea un 30%.

En este caso, el intervalo de confianza al 95% de la proporción de la población con hipertensión es (0.2591, 0.4525). Por ende, la proporción de población con presión sistólica superior o igual a 130mm de Hg está entre el 25.91% y el 45.25%. Donde el 30% de proporción que propone el enunciado, está dentro del intervalo.

d)

Para poder determinar la eficacia del medicamento en la población adulta después de la toma del medicamento. Realizaremos el contraste de igualdad de medias para datos emparejados. Es decir, observaciones pareadas donde se extraen dos muestras no independientes con el mismo tamaño de muestra de dos poblaciones normales. En este caso, las muestras se extraen de la misma población adulta (Grupo 2) de las diferentes presiones sistólicas.

Por tanto, realizamos el contraste y el intervalo sobre la diferencia de medias.

$$d = \mu_1 - \mu_2$$

Contraste e intervalo de diferencia de medias al 95% de ambas presiones sistólicas

```
    Paired t-test

data:  Grupo2$`Presion sistólica antes` and Grupo2$`Presión sistólica después`
t = 3.784, df = 59, p-value = 0.0003631
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 1.900476 6.166191
sample estimates:
mean difference
       4.033333
```

Tras los resultados del contraste de hipótesis, observamos que el P-valor es 0.0003631 y al ser menor que el nivel de significación de 0.05, rechazaríamos la hipótesis nula. Siendo la hipótesis nula:

$$\mu_1 = \mu_2$$

Donde no hay diferencia entre las medias de las diferentes presiones. Es decir, existe una diferencia significativa entre la presión sistólica tomada antes del medicamento y la presión sistólica tomada después del medicamento. El intervalo del confianza al 95% de la diferencia media entre ambas presiones es de (1.900, 6.166). Por ello, podemos deducir por el intervalo que la diferencia media de ambas presiones es positiva. En otras palabras, en cuanto a las medias, la presión sistólica después del medicamento es mayor que la presión sistólica antes del medicamento. En consecuencia, el medicamento no es eficaz ya que no se ha reducido la presión sistólica después de la toma del medicamento.