Tarea de Estadística

Juan Armario Muñoz

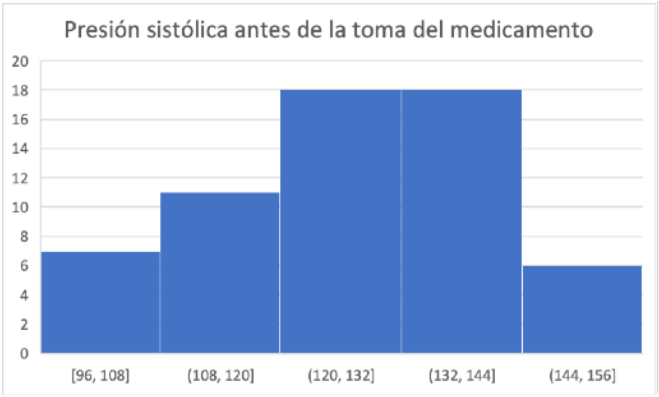
a) Obtener, usando algún programa estadístico, las medidas de centralización y dispersión para cada uno de los dos grupos de control de la variable (grupo 1 y grupo 2) que mide la presión sistólica antes de la toma del medicamento e indica si la media en cada uno de estos grupos puede considerarse representativa a partir de los datos obtenidos.

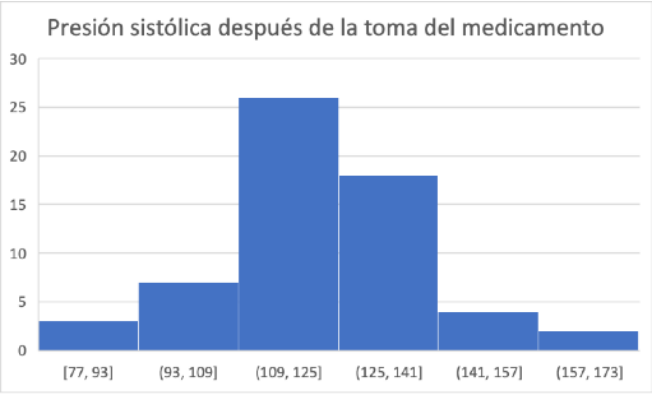
	Grupo 1
Media	123,425
Error típico	1,42724744
Mediana	124,5
Moda	131
Desviación estándar	9,026705393
Varianza de la muestra	81,48141026
Curtosis	-0,45691104
Coeficiente de asimetría	-0,138262431
Rango	37
Mínimo	105
Máximo	142
Suma	4937
Cuenta	40

	Grupo 2
Media	127,4833333
Error típico	1,793032918
Mediana	126
Moda	121
Desviación estándar	13,88877326
Varianza de la muestra	192,8980226
Curtosis	-0,602056286
Coeficiente de asimetría	-0,136599095
Rango	58
Mínimo	96
Máximo	154
Suma	7649
Cuenta	60

La media en ambos casos es representativa, ya que el coeficiente de variación es muy cercano a 0. Para el caso del grupo 1 es igual a 0,07313515, en torno al 7.3% de variación en nuestros datos. Para el caso del grupo 2 es igual a 0,1089458, aproximadamente el 10% de dispersion de los datos.

b) Estudiar la simetría y la curtosis del nivel de presión sistólica en los pacientes del segundo grupo para cada una de las mediciones de la hipertensión que aparecen en la tabla.





Como podemos observar en ambos gráficos para los niveles de presión sistólica en los pacientes del grupo 2, ambas muestras son prácticamente simétricas o ligeramente asimétricas. Resultado que podemos corroborar ya que ambos coeficientes de simetría, obtenidos en el apartado anterior, son, aunque negativos, muy cercanos a 0.

Con respecto a la curtosis, podemos observar a simple vista, que ambos casos siguen una distribución muy similar a la distribución normal.. Con los datos obtenidos en el apartado a, con valores muy cercanos a 0 pero negativos, ambas distribuciones son un poco más aplanadas que la distribución normal.

c) Indicar para cada una de las variables relacionadas con la medición de la presión sistólica que aparecen en el fichero el valor de los cuartiles y su significado y obtener el box- plot (diagrama de cajas) correspondiente. Estudiar la presencia de valores atípicos

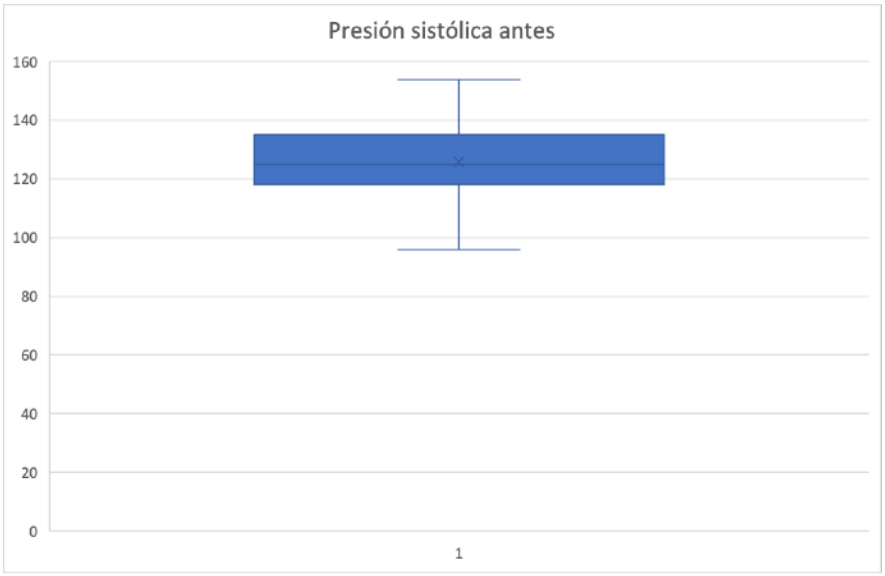
Un diagrama de caja (también, diagrama de caja y bigotes o box plot) es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, se muestran a simple vista la mediana y los cuartiles de los datos, y también pueden representarse sus valores atípicos.

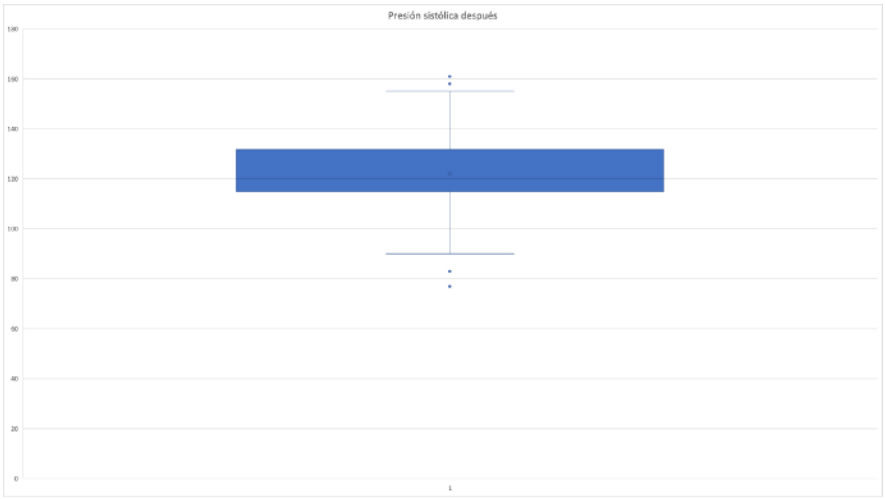
Los cuartiles son valores que dividen una distribución de datos en cuatro partes iguales. Es decir, el primer cuartil (Q1) representa el valor que separa el 25% inferior de los datos del 75% superior, el segundo cuartil (Q2) es el valor que separa el 50% inferior del 50% superior (también conocido como mediana), y el tercer cuartil (Q3) es el valor que separa el 75% inferior del 25% superior de los datos.

Para el caso de la presión antes de la toma del medicamento, el valor de los cuartiles es el siguiente:
Q1= 118
Q2= 125
Q3= 135

Y Para el caso de la presión después de la toma del medicamento, el valor de los cuartiles es el siguiente:
Q1= 115
Q2= 120
Q3= 131.25

A continuación, con la ayuda del software Excel, calcularemos el diagrama de bigotes para la obtención de valores atípicos. En la presión sistólica antes no obtenemos ningún valor atípico. Sin embargo en la presión sistólica después, como indica el gráfico, si obtenemos valores atípicos.





d) Estudiar la normalidad de los datos de las variables relacionadas con la medición de la presión sistólica.

Para estudiar la normalidad de los datos de las variables relacionadas con la medición de la presión sistólica utilizando el test de Shapiro-Wilk a mano, sigue estos pasos:

Ordena los datos de menor a mayor.
Calcula la media y la desviación estándar de los datos.
Calcula los estadísticos W y W' del test de Shapiro-Wilk.
Interpreta los resultados.

Para el caso de la presión sistólica antes de la toma del medicamento, obtenemos:

Media (μ): 126.02
Desviación estándar (σ): 12.42
Calculamos W con la expresión: $$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

W= 37.29
W'= 1607.96

Comparando W' con la tabla de valores críticos del test de Shapiro-Wilk para n=100 obtenemos un valor crítico de aproximadamente 0.935.

Como W' > 0.935, no rechazamos la hipótesis nula de que los datos provienen de una distribución normal.
Por lo tanto, los datos parecen seguir una distribución normal según el test de Shapiro-Wilk

Para el caso de la presión sistólica después de la toma del medicamento, obtenemos:

Media (μ): 120.175
Desviación estándar (σ): 13.043
Calculamos W con la expresión: $$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

W= 0.9779
W'= –0.228

Comparando W' con la tabla de valores críticos del test de Shapiro-Wilk para n=100 obtenemos un valor crítico de aproximadamente 0.935.

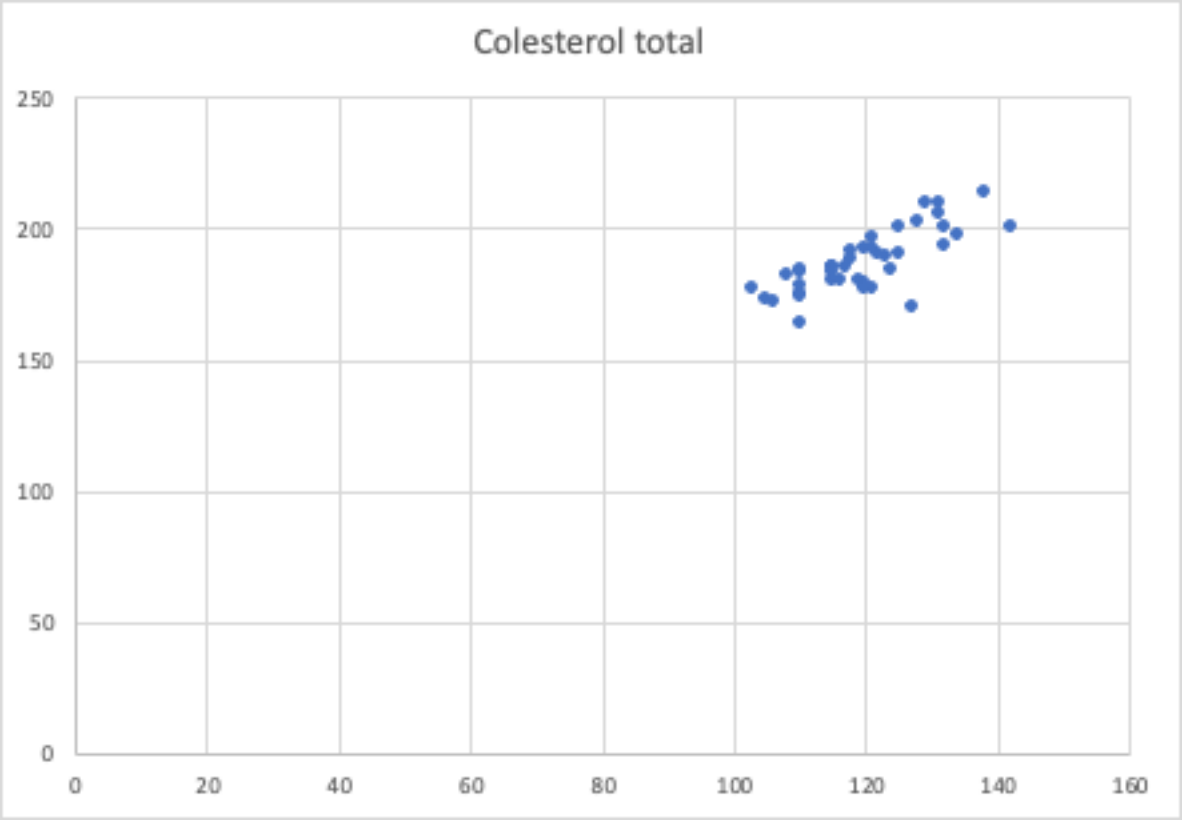
Como W' > 0.935, no rechazamos la hipótesis nula de que los datos provienen de una distribución normal.

Por lo tanto, los datos parecen seguir una distribución normal según el test de Shapiro-Wilk

Ejercicio 2

Con los datos del fichero anterior, se quiere estudiar la relación existente entre la presión sistólica después de la toma del medicamento y el colesterol total del paciente en los pacientes jóvenes (grupo 1)

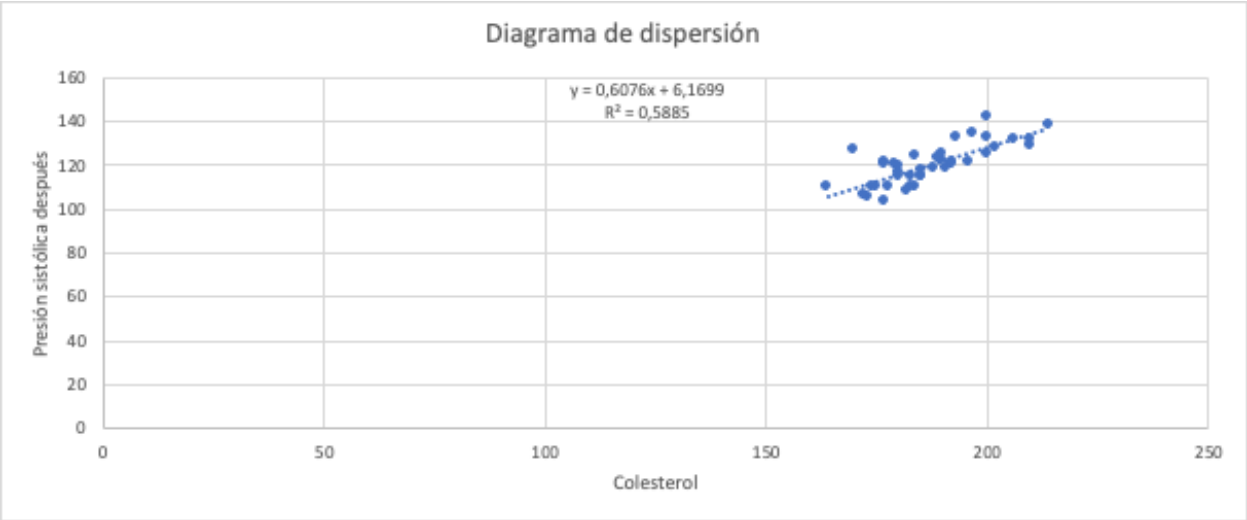
a)Estudiar la relación lineal existente entre estas dos variables de estudio.



Al estudiar en el gráfico de dispersion, la relación entre la presión sistólica, después de la toma del medicamento, y el valor del colesterol, podemos ver como estas tienen una relación directa o positiva. Al aumentar una aumenta la otra.

Esto queda corroborado al estudiar el coeficiente de correlación, $r = \frac{S_{xy}}{S_x S_y} = 82.27/9.03 * 13.89 = 0.656$, confirmando que la relación entre ambas variables es alta.

b) Obtener un modelo lineal que explica la presión sistólica del paciente joven a los 60 minutos de ingerir el medicamento en función de su colesterol total y realizar la estimación para un paciente del grupo 1 (joven) cuyo colesterol total es 105 mg/Dl



Para obtener un modelo lineal que explique la presión sistólica de un paciente en función de su colesterol, he optado por calcularlo con Excel a través del gráfico de dispersión y agregando la línea de tendencia lineal.

La ecuación obtenida ha sido: $y = 0.6076x + 6.1699$, donde y es la presión sistólica que varia dependiendo del colesterol que sería nuestra x.

La pendiente de esta recta, $b_1 = \frac{S_{xy}}{S_x^2}$, mide la variación promedio en Y por unidad adicional de cambio en X. La ordenada en el origen, $b_0 = \overline{Y} - \frac{S_{xy}}{S_x^2}\ \overline{X}$, valor medio estimado de Y cuando la variable independiente X toma el valor 0. Ambas, se podrían haber obtenido resolviendo las ecuaciones anteriormente mostradas.

La estimación de la presión sistólica para un paciente, cuyo colesterol tras haber ingerido el medicamento, es de 105 mg/Dl, correspondería a la siguiente expresión, en la que x tomaría el valor de 105:

$$\hat{y}(105) = 69.9679$$

Por lo tanto la presión sistólica sería de 69.97 mm Hg.

c) ¿Qué tanto por ciento de la presión sistólica del paciente joven a los 60 minutos de ingerir el medicamento no queda explicado por el anterior modelo? ¿Cómo podrías mejorar el modelo?

Al calcular en Excel el gráfico de correlación y su ecuación, también mostramos el valor del coeficiente de correlación, R^2. Este valor indica el porcentaje de variación de la variable dependiente, en nuestro caso la presión sistólica después, que queda explicado por la relación lineal con la variable dependiente, en nuestro caso el colesterol.

Como podemos observar nuestro valor de $R^2 = 0.5885$, o lo que es igual, la variación de la presión sistólica queda explicada por la variación del colesterol en un 58.85%. Para aumentar este valor, deberíamos de introducir más variables en nuestra regresión, sin embargo tendríamos que estudiar su relación con modelos de regresión múltiple.

d) Si aumentásemos el colesterol de un paciente en 5 mg/Dl ¿Qué variación experimentaría su presión sistólica después de 60 minutos de ingerir el medicamento?

Si aumentásemos el colesterol de un paciente en 5 mg/Dl, la presión sistólica variaría según el factor correspondiente a la resolución de la siguiente ecuación:

$$\hat{y}(5) = 0.6076 * 5 \ + 6.1699 = 9.2079 \ mm \ Hg$$

Por lo tanto la presión sistólica aumentaría en 9.2 unidades, si el colesterol aumentase 5 mg/Dl.

Ejercicio 3

a) Se quiere estudiar si se puede admitir que la presión sistólica media en el momento de la ingestión de la población adulta (grupo 2) es 130 mm de Hg. Obtener el intervalo de confianza al 95% y al 99% para el nivel medio de presión sistólica antes de la toma del medicamento en el grupo de los adultos y posteriormente contesta a la cuestión planteada con los resultados obtenidos o mediante un contraste de hipótesis.

Utilizando la fórmula del intervalo de confianza:

$$IC(\mu) = \left[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

Donde $z_{\frac{\alpha}{2}}$ es el valor crítico, para un intervalo de confianza del 95%, el valor crítico es aproximadamente 1.96, y para un intervalo de confianza del 99%, el valor crítico es aproximadamente 2.576.

Habiendo calculado la media muestra, 123.48, y la desviación estándar, 9.026, y teniendo como tamaño de la muestra, 60.

Intervalo de confianza al 95%: $= 123.48 \pm 1.96 \times \frac{9.026}{\sqrt{60}} = 123.48 \pm 2.34$

Intervalo de confianza al 95% = (121.14, 125.82)

Intervalo de confianza al 99%: $= 123.48 \pm 2.576 \times \frac{9.026}{\sqrt{60}} = 123.48 \pm 3.06$

Intervalo de confianza al 99% = (120.42, 126.54)

Ya que 130 no está dentro del intervalo, no podemos admitir ese valor como media de la presión sistólica en el momento de la ingestion de la población adulta.

Comprobaremos dicho resultado con un contraste de hipótesis.

Planteemos el contraste de hipótesis: $H0$: $\mu = 130$ $H1$: $\mu \neq 130$
El valor estadístico de contraste bajo la hipótesis nula sería:

$$Z = \frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{n}} = \frac{123.48 - 130}{9.026/\sqrt{60}} = 5.60$$

En este caso, como el nivel de significación $\alpha = 0.05$, tenemos que la región de aceptación será (-1,96,1,96) ya que 1,96 es el valor crítico de la normal que le corresponde a ese nivel de significación.

Dado que el valor de nuestro estadístico es menor que cualquier nivel de significancia comúnmente utilizado (como 0.05 o 0.01), rechazamos la hipótesis nula. Esto significa que hay suficiente evidencia estadística para concluir que la presión sistólica media en el momento de la ingestión de la población adulta no es igual a 130 mmHg.

b) Obtener el intervalo de confianza al 95% para la diferencia de medias en la presión sistólica entre adultos y jóvenes después de la ingestión del medicamento. ¿Se puede concluir que después de la ingesta del medicamento la presión sistólica media de la población es distinta dependiendo de la edad?

Utilizando la ecuación del calculo del intervalo de confianza para la diferencia de medias:

$$\overline{x_1} - \overline{x_2} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Donde; para adultos, la media es 127,48, la desviación estándar es 13,89, y tamaño de la muestra es 60. Para jovenes, la media es de 123,43, desviación estándar 9,027, y tamaño de la muestra 40.

$$(127.48 - 123.43) \pm 1.984 \times \sqrt{\frac{13.89^2}{60} + \frac{9.027^2}{40}} = 4.05 \pm 0.463$$

Por lo tanto, el intervalo de confianza al 95% para la diferencia de medias en la presión sistólica entre adultos y jóvenes después de la ingesta del medicamento es aproximadamente (3.587, 4.513) Para determinar si la presión sistólica media de la población es distinta dependiendo de la edad, debemos observar si este intervalo incluye el valor 0.

Como este intervalo no incluye el valor de cero, podemos concluir que la diferencia en las medias de presión sistólica entre adultos y jóvenes después de la ingesta del medicamento es estadísticamente significativa. Es decir, hay una diferencia significativa en la presión sistólica media entre estos dos grupos de edad después de tomar el medicamento.

Por lo tanto, podemos concluir que después de la ingesta del medicamento, la presión sistólica media de la población es distinta dependiendo de la edad, siendo mayor en adultos que en jóvenes.

c) Se quiere estudiar la proporción de la población con una presión sistólica inicial igual o superior a 130 mm de Hg (prehipertensión). A partir de la muestra del fichero (tomando todos los datos de presión sistólica antes de la toma del medicamento) obtener un intervalo de confianza al 99% de la proporción de la población con hipertensión y contrastar la hipótesis que el porcentaje de la población con presión sistólica superior o igual a 130 mm de Hg es 0,30 con nivel de significación del 5%.

Tomaremos como proporción muestral 0.35, ya que los valores de la muestra que cumplen el criterio de se mayor o igual a 130, son 35 y el tamaño de la muestra es de 100. Al dividir dichos valores obtenemos la proporción muestral.

El intervalo buscado será:

$$IC(p) = \left[\hat{p} - z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$$

$$IC(p) = 0.35 \pm 2.576 \times \sqrt{\frac{0.35 \times (1 - 0.35)}{100}} = 0.35 \pm 0.12295$$

Por lo tanto, el intervalo de confianza al 99% para la proporción de la población con presión sistólica igual o superior a 130 mmHg, es aproximadamente (0.227, 0.473). Es decir, entre el 22.7% y el 47.3%. Como 0.30 está dentro de los valores del intervalo calculado podemos confirmar la hipótesis propuesta en el enunciado del problema.

Ahora, para contrastar la hipótesis nula *H*0: p = 0.30 contra la hipótesis alternativa *H*1: p ≠ 0.30 con un nivel de significación del 5%, podemos utilizar una prueba de hipótesis de proporciones z, utilizando la fórmula:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.35 - 0.30}{\sqrt{\frac{0.30 \times 0.70}{100}}} = 1.091$$

Usando una tabla de valores z o un software estadístico, encontramos que el valor crítico de z es aproximadamente 1.96 para un nivel de significancia del 5%.
Como 1.091 no es mayor que 1.96 no podemos rechazar la hipótesis nula.
*H*0: p = 0.30. Por lo tanto, no hay suficiente evidencia para afirmar que la proporción de la población con presión sistólica igual o superior a 130 mmHg es diferente de 0.30 con un nivel de significancia del 5%. No podemos rechazar que el 30% de la población tenga una presión sistólica de 130 mmHg.

d) (VOLUNTARIO) Por último, se quiere estudiar la eficacia del medicamento en la población adulta. ¿Existe variación significativa de la presión sistólica después de la toma del medicamento en la población del grupo 2? Plantea el correspondiente contraste de hipótesis considerando un nivel de significación del 5%. Ayuda: Para contestar a la pregunta has de considerar la series de datos obtenidas a partir de las diferencias entre la presión sistólica antes de la toma y la presión sistólica al cabo de 60 minutos en el grupo de los adultos (contraste de muestras emparejadas).

Para responder a si existe variación significativa de la presión sistólica después de la toma del medicamento en la población del grupo 2, primeramente deberemos obtener los datos de la diferencia de la presión sistólica antes y después de la ingesta y a continuación obtendremos la media y la desviación estándar de dichos datos. Finalizando con un contraste de hipótesis, haciendo uso del estadístico bilateral de igualdad, bajo la hipótesis nula y sabiendo que dicho estadístico sigue una distribución t-Student con n-1 grado de libertad. En nuestro caso 60-1=59.

Para la media obtenemos un valor de -4,03333333 y para la desviación estándar 8,25641336. Y para el grupo 2 sabemos que el tamaño de la muestra es de 60 individuos.

Hipótesis nula (Ho): No hay diferencia significativa en la presión sistólica después de tomar el medicamento en la población del grupo 2.

Ho: $\mu d = 0$ (donde μd representa la media de las diferencias)

Hipótesis alternativa (H1): Hay una diferencia significativa en la presión sistólica después de tomar el medicamento en la población del grupo 2.

H1: $\mu d \neq 0$

La ecuación para calcular el estadístico t en este caso es:

$$T = \frac{\overline{D} - D}{\frac{s_d}{\sqrt{n}}} = \frac{-4.0333333}{\frac{8.25641336}{\sqrt{60}}} = -1.86$$

Ahora, necesitamos encontrar el valor crítico de t para un nivel de significación del 5% y 59 grados de libertad. Luego, compararemos este valor crítico con el valor calculado de t para tomar una decisión. Para un nivel de significancia del 5% y 59 grados de libertad, el valor crítico de t es aproximadamente ±2.000.

Dado que el valor calculado de t es -1.86, que está dentro del rango de -2.000 a +2.000, no podemos rechazar la hipótesis nula. Por lo tanto, no hay evidencia suficiente para concluir que hay una diferencia significativa en la presión sistólica después de tomar el medicamento en la población del grupo 2.