Análisis estadístico de la temperatura

Disponemos de dos ficheros de datos que recogen la temperatura corporal, medida en grados Fahrenheit, de una muestra de 65 hombres y otra de 65 mujeres.

Vamos a realizar un análisis descriptivo de la variable temperatura corporal masculina y femenina y comparar aspectos relevantes de ambas.

Resumen Estadístico

	male.Temperature	female.Temperature
Recuento	65	65
Promedio	97,8908	98,4569
Mediana	97,9	98,5
Moda		98,8
Desviación Estándar	0,668656	0,713479
Coeficiente de Variación	0,683063%	0,724662%
Mínimo	96,2	96,7
Máximo	99,3	100,6
Rango	3,1	3,9
Cuartil Inferior	97,5	98,0
Cuartil Superior	98,4	98,9
Rango Intercuartílico	0,9	0,9
Sesgo Estandarizado	-1,20754	0,141046
Curtosis Estandarizada	-0,266099	1,49237

Esta tabla muestra el resumen estadístico para las dos variables que estamos estudiando. En ella se incluyen medidas de tendencia central, de dispersión y de forma. Podemos observar que la muestra masculina tiene menor media muestral que la femenina, siendo además tanto el mínimo como el máximo de esta primera menor que el de la segunda. La muestra femenina tiene un rango mayor. Ambas tienen una desviación típica baja que refleja que los datos están agrupados en torno a la media. En la muestra femenina el valor que más se repite es el 98,8, mientras que en la masculina observamos que no hay una sola moda y mirando los datos del fichero vemos que los valores 98,2, 97,9 y 97,5 son los que más se repiten y lo hacen por igual.

Debemos prestar atención a los valores del sesgo estandarizado y la curtosis estandarizada. Sabemos que si el sesgo menor que -2 presenta asimetría a la izquierda mientras que si es mayor que 2 presenta asimetría a la derecha. Si la curtosis es mayor que 2 sabemos que la distribución es más picuda y por el contrario si es menor que -2 es más parecida a la uniforme. En este caso, ambas variables tienen el sesgo estandarizado y la curtosis estandarizada entre -2 y 2 lo que refleja que las muestras provienen de una distribución normal. En la tabla también hemos incluido la mediana, que es el valor por debajo del cual se encuentra el 50% de las observaciones del fichero, y el primer y tercer cuartil porque más adelante nos serán útiles estos datos.

Se puede elaborar la tabla de frecuencias a partir de los datos de la muestra. La tabla de frecuencias la hemos realizado tomando 19 intervalos iguales en el rango de la muestra. Las frecuencias muestran el número de datos en cada intervalo. Con los intervalos y las frecuencias se construye el histograma, el cual analizaremos más adelante. La tabla también muestra la frecuencia relativa, que es la proporción en cada intervalo.

Tabla de Frecuencias para male. Temperature

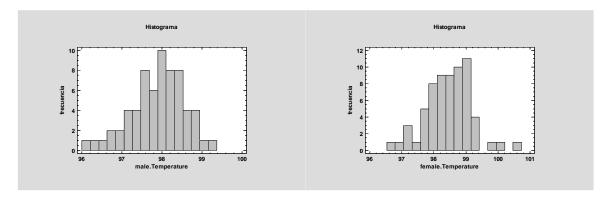
Clase	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa
	menor o igual	96	0	0,0000
1	96	96,2105	1	0,0154
2	96,2105	96,4211	1	0,0154
3	96,4211	96,6316	1	0,0154
4	96,6316	96,8421	2	0,0308
5	96,8421	97,0526	2	0,0308
6	97,0526	97,2632	4	0,0615
7	97,2632	97,4737	4	0,0615
8	97,4737	97,6842	8	0,1231
9	97,6842	97,8947	6	0,0923
10	97,8947	98,1053	10	0,1538
11	98,1053	98,3158	8	0,1231
12	98,3158	98,5263	8	0,1231
13	98,5263	98,7368	4	0,0615
14	98,7368	98,9474	4	0,0615
15	98,9474	99,1579	1	0,0154
16	99,1579	99,3684	1	0,0154
17	99,3684	99,5789	0	0,0000
18	99,5789	99,7895	0	0,0000
19	99,7895	100,0	0	0,0000
	mayor de	100	0	0,000

Tabla de Frecuencias para female. Temperature

Clase	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa
	menor o igual	96	0	0,0000
1	96	96,2632	0	0,0000
2	96,2632	96,5263	0	0,0000
3	96,5263	96,7895	1	0,0154
4	96,7895	97,0526	1	0,0154
5	97,0526	97,3158	3	0,0462
6	97,3158	97,5789	1	0,0154
7	97,5789	97,8421	5	0,0769
8	97,8421	98,1053	8	0,1231
9	98,1053	98,3684	9	0,1385
10	98,3684	98,6316	9	0,1385
11	98,6316	98,8947	10	0,1538
12	98,8947	99,1579	11	0,1692
13	99,1579	99,4211	4	0,0615
14	99,4211	99,6842	0	0,0000
15	99,6842	99,9474	1	0,0154
16	99,9474	100,211	1	0,0154
17	100,211	100,474	0	0,0000
18	100,474	100,737	1	0,0154
19	100,737	101,0	0	0,0000
	mayor de	101	0	0,0000

Como hemos dicho, podemos representar los histogramas, los cuales nos van a ser más útiles para extraer conclusiones.

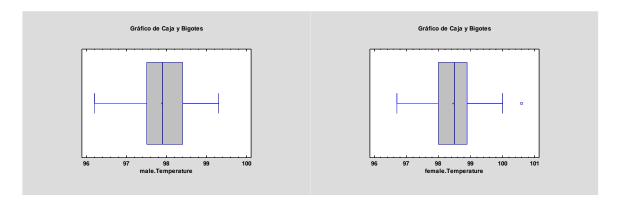
Histogramas



En el de la temperatura masculina podemos observar una ligera asimetría a la izquierda lo que intuíamos porque el valor del sesgo estandarizado es -1,20754. La curtosis estandarizada es negativa pero muy próxima a cero, lo que podemos interpretar en el gráfico, y refleja el hecho de que sigue una distribución normal.

Del histograma de la temperatura femenina podemos intuir que la distribución apenas presenta asimetría lo cual podemos asegurar por la proximidad a 0 del sesgo estandarizado. También podemos ver que esta distribución va a ser más picuda que la de la temperatura masculina por tener curtosis estandarizada positiva de valor 1,49237, pero aun así, como hemos comentado antes, dentro de los parámetros de la distribución normal. Observamos también la presencia de algunos valores alejados del resto pero este aspecto vamos a comentarlo mejor en el gráfico de cajas y bigotes.

Gráficos de cajas y bigotes



Sabemos que las tres líneas verticales que conforman la caja son el primer cuartil, la mediana y el tercer cuartil y las dos de los extremos son Q1-1.5RIC y Q3+1.5RIC, siendo RIC el recorrido intercuartílico. La mediana es el valor por debajo del cual se encuentra el 50% de las observaciones del fichero, y en ambos casos está muy próximo a la media.

Se observa que en el gráfico de la muestra masculina la distancia de la caja al primer bigote es mayor que la de la caja al segundo bigote, lo que refleja la ligera asimetría a la izquierda que veníamos comentando, mientras que en la muestra femenina esta diferencia considerable no se da y por eso el hecho de que apenas presente asimetría.

En el gráfico de la temperatura corporal femenina observamos que hay un dato de la muestra, el correspondiente al máximo, por encima del bigote marcado por Q3+1.5RIC, y esto nos informa sobre la existencia de un dato atípico.

A continuación, vamos a mostrar una gráficas obtenidas con R sobre el estudio de la distribución empírica de la muestra tanto en hombres como mujeres.

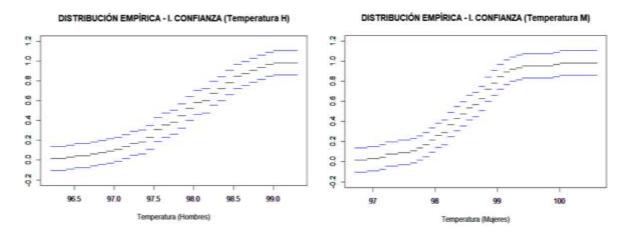


Fig. 1. Fig. 2. En negro aparece la distribución empírica asociada a la muestra de la temperatura en los hombres y mujeres respectivamente y en azul los intervalos de confianza al 95% aproximados.

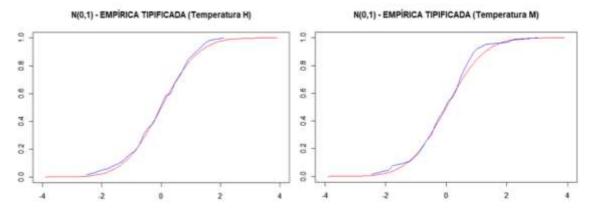


Fig. 3. Fig. 4. En rojo aparece la distribución de una normal N (0,1) y en azul el polígono de frecuencias relativas acumuladas de muestra de la temperatura en los hombres y mujeres respectivamente tipificada con la media y desviación muestral.



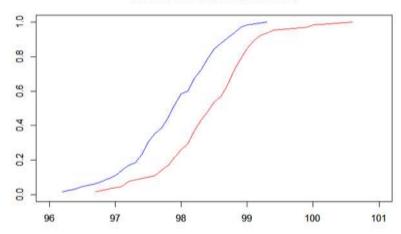


Fig. 5. En rojo aparece el polígono de frecuencias relativas acumuladas de muestra de la temperatura en las mujeres tipificada con la media y desviación muestral y en azul el polígono de frecuencias relativas acumuladas de muestra de la temperatura en los hombres tipificada con la media y desviación muestral.

Las figuras 1 y 2 nos permiten afirmar que, con una confianza del 95%, dado un punto x, la imagen de este por la función de distribución que desconocemos estará en la correspondiente franja azul.

Las figuras 3 y 4 indican que usar una normal para modelizar las temperaturas es una buena opción. Cabe destacar en la figura 4 que cuando la variable tipificada está en torno a 1 esta se acumula con mayor fuerza que la normal N(0, 1), esto podría darnos a pensar que el suponer una desviación típica ligeramente menor a la muestral explicaría mejor el suceso si suponemos que se rige por una normal.

Por último, la figura 5 muestra que la media de los hombres es menor que la de las mujeres al estar el polígono de frecuencias desplazado a la izquierda. Además, muestra que en los hombres tenemos datos por debajo del mínimo de las mujeres y que en las mujeres tenemos datos por encima del máximo de los hombres.

Intervalos de confianza del 95,0%

	Media	Error Est.	Límite Inferior	Límite Superior
male.Temperature	97,8908	0,0829366	97,7251	98,0565
female.Temperature	98,4569	0,0884962	98,2801	98,6337

	Sigma	Límite Inferior	Límite Superior
male.Temperature	0,668656	0,570219	0,808491
female.Temperature	0,713479	0,608443	0,862688

Como ya hemos comentado, podemos aproximar ambas variables por una distribución normal, y además tipificar a una N (0,1), y dar así un intervalo de confianza aproximado al 95%. La interpretación de estos intervalos es que podemos establecer con una confianza del 95% que la media y la desviación típica verdadera se encuentra en algún lugar entre los valores del intervalo obtenido en cada caso.

Para el cálculo de estos intervalos, Statgraphics emplea las fórmulas que hemos estudiado para la media (con desviación típica desconocida) y para la desviación típica. Estas son,

$$I = [\overline{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}] \text{ en el caso de la media y } I = [\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\frac{1-\alpha}{2}}}}] \text{ en el caso de la desviación típica.}$$

Javier Castellano Soria, Ignacio Fernández Sánchez-Pascuala y Fermín González Pereiro

En este caso es interesante dar un intervalo de confianza para la diferencia de medias. Primero vamos a ver el del cociente de varianzas para ver si suponemos varianzas iguales o distintas. El intervalo que sale es [0,694464; 1,86666] y como el 1 pertenece podemos suponer varianzas iguales. Entonces el intervalo para la diferencia de medias al 95% es [0,32617; 0,806137]. Para el cálculo de este sabemos que se emplea la fórmula $[\overline{X_1} - \overline{X_2} \pm t_{n+m-2,\alpha/2}S_c\sqrt{\frac{1}{n} + \frac{1}{m}}$ siendo $S_c^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$. En este caso se tiene n=m=65.

Contraste de hipótesis

Usando el lenguaje de programación R hemos sometido a varias hipótesis bilaterales la media y la desviación típica obteniendo intervalos (con una precisión de 0.001) que nos indican que si la hipótesis nula (un determinado valor del parámetro) pertenece a este, entonces no se rechazaría. Los resultados son los siguientes:

- IntervaloMediaTemperaturaHombres = [97.726, 98.055]
- IntervaloMediaTemperaturaMujeres = [98.281, 98.632]
- IntervaloDesviaciónHombres = [0.571, 0.807]
- IntervaloDesviaciónMujeres = [0.601, 0.862]

Estos valores eran de esperar al haber analizado previamente los intervalos de confianza al 95%.

Análisis estadístico de las pulsaciones

Ahora disponemos de dos ficheros tomados de la misma muestra que recogen los datos de las pulsaciones por minuto de los individuos.

Vamos a realizar un análisis descriptivo de la variable pulsaciones por minuto (P.P.M) masculina y femenina y comparar, al igual que en la variable anterior.

Resumen Estadístico

	P.P.M Hombres	P.P.M Mujeres
Recuento	65	65
Promedio	73,3692	74,1538
Mediana	73,0	76,0
Moda	78,0	79,0
Desviación Estándar	5,87518	8,10523
Coeficiente de Variación	8,00769%	10,9303%
Mínimo	58,0	57,0
Máximo	86,0	89,0
Rango	28,0	32,0
Cuartil Inferior	70,0	68,0
Cuartil Superior	78,0	80,0
Rango Intercuartílico	8,0	12,0
Sesgo Estandarizado	-0,168581	-0,964889
Curtosis Estandarizada	-0,528807	-1,17937

En este caso también se da que la muestra masculina tiene menor media muestral que la femenina, podría ser que exista una correlación entre temperatura y P.P.M, pero no es motivo de estudio en este curso. Además, la muestra femenina tiene un rango bastante superior, lo que induce (generalmente) una mayor desviación típica, como podemos observar.

Llama la atención que el valor más frecuente en ambas muestras sea muy próximo en hombres y mujeres.

Por otro lado, también ambas variables tienen el sesgo y la curtosis estandarizados entre -2 y 2 lo que significa que se puede suponer que las muestras provienen de una distribución normal. Además, se puede observar que en la muestra masculina la mediana y la media están muy próximas, por ello se da un sesgo estandarizado (coeficiente de simetría) muy cercano al 0, sin embargo, en la muestra femenina se separan algo más y el sesgo es superior en valor absoluto (más asimétrica por la izquierda).

También es lógico aparentemente que al ser el rango mayor en la muestra femenina también lo sea el rango intercuartílico.

A partir de la muestra podemos elaborar las tablas de frecuencias:

Tabla de Frecuencias para P.P.M Hombres

Clase	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa
	menor o igual	56	0	0,0000
1	56	58,1053	1	0,0154
2	58,1053	60,2105	0	0,0000
3	60,2105	62,3158	0	0,0000

4	62,3158	64,4211	4	0,0615
5	64,4211	66,5263	3	0,0462
6	66,5263	68,6316	5	0,0769
7	68,6316	70,7368	8	0,1231
8	70,7368	72,8421	9	0,1385
9	72,8421	74,9474	9	0,1385
10	74,9474	77,0526	7	0,1077
11	77,0526	79,1579	8	0,1231
12	79,1579	81,2632	4	0,0615
13	81,2632	83,3684	5	0,0769
14	83,3684	85,4737	1	0,0154
15	85,4737	87,5789	1	0,0154
16	Mayor de	87,5789	0	0,0000

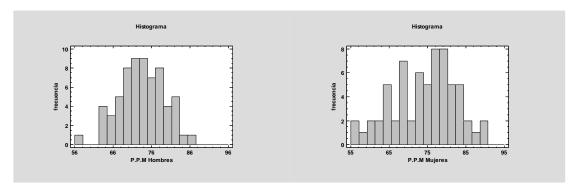
Tabla de Frecuencias para P.P.M Mujeres

Clase	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa
	menor o igual	55	0	0,0000
1	55	57,1053	2	0,0308
2	57,1053	59,2105	1	0,0154
3	59,2105	61,3158	2	0,0308
4	61,3158	63,4211	2	0,0308
5	63,4211	65,5263	5	0,0769
6	65,5263	67,6316	2	0,0308
7	67,6316	69,7368	7	0,1077
8	69,7368	71,8421	2	0,0308
9	71,8421	73,9474	6	0,0923
10	73,9474	76,0526	5	0,0769
11	76,0526	78,1579	8	0,1231
12	78,1579	80,2632	8	0,1231
13	80,2632	82,3684	5	0,0769
14	82,3684	84,4737	5	0,0769
15	84,4737	86,5789	2	0,0308
16	86,5789	88,6842	1	0,0154
17	88,6842	90,7895	2	0,0308
18	Mayor de	90,7895	0	0,0000

Otra forma de representar las tablas de frecuencias es mediante los histogramas, donde estas medidas mencionadas en el resumen pueden ver de forma más explícita:

Javier Castellano Soria, Ignacio Fernández Sánchez-Pascuala y Fermín González Pereiro

Histogramas

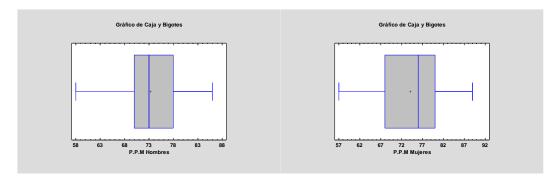


Hemos realizado la representación del histograma partir de 19 intervalos entre 55-95 en el caso de las mujeres y 56-96 en el caso de los hombres.

En la de las P.P.M femenina podemos observar una asimetría a la izquierda superior a la de la masculina, que es bastante simétrica, debido al sesgo y también es más plana (en el histograma de las mujeres el eje vertical solo tiene hasta 8), debido a la curtosis, aunque en ambos casos tampoco es muy notorio, recordemos que ambas se pueden suponer que provienen de una normal.

También vemos que en la gráfica masculina se concentran más valores alrededor de la media mientras que en la femenina menos, por eso la masculina tiene una varianza inferior.

Gráficos de cajas y bigotes



Mediante la gráfica de cajas y bigotes vemos representado de una mejor forma los cuartiles, rango, rango intercuartílico...

Si nos fijamos, el rango intercuartílico es mayor en las mujeres, como preveíamos, por tanto, concentra menos datos alrededor de la media lo que se relaciona con una mayor varianza en este caso. Además, la diferencia entre la mediana y el tercer cuartil en la gráfica de la muestra femenina es mínima, lo que significa que concentra muchos valores en ese intervalo, lo que explica el sesgo de -0,96.

Cabe observar que en el primer caso la media está por encima de la mediana, pero muy próxima, lo que explica una curtosis muy cercana al 0, sin embargo, en el segundo es al revés.

A continuación, se muestran una serie de gráficas sobre el estudio de la distribución empírica de la muestra:

Gráficas Distribución Empírica

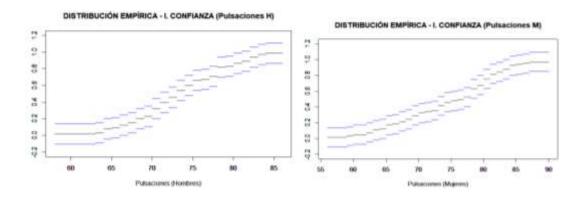


Fig. 6. En negro aparece la distribución empírica asociada a la muestra de las pulsaciones en los hombres y en azul los intervalos de confianza al 95%

Fig. 7. En negro aparece la distribución empírica asociada a la muestra de las pulsaciones en las mujeres y en azul los intervalos de confianza al 95%

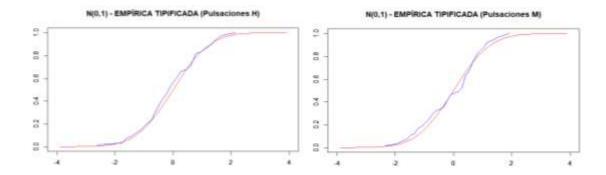


Fig. 8. En rojo aparece la distribución de una normal N(0,1) y en azul el polígono de frecuencias relativas acumuladas de muestra de las pulsaciones en los hombres tipificada con la media y desviación muestral.

Fig. 9. En rojo aparece la distribución de una normal N(0,1) y en azul el polígono de frecuencias relativas acumuladas de muestra de las pulsaciones en las muieres tipificada con la media v desviación muestral.

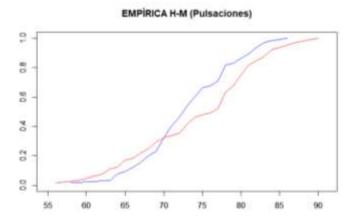


Fig. 10. En rojo aparece el polígono de frecuencias relativas acumuladas de muestra de las pulsaciones en las mujeres tipificada con la media y desviación muestral y en azul el polígono de frecuencias relativas acumuladas de muestra de las pulsaciones en los hombres tipificada con la media y desviación muestral.

Las figuras 6 y 7 nos permiten afirmar que, con una confianza del 95%, dado un punto x, la imagen de este por la función de distribución que desconocemos estará en la correspondiente franja azul.

Las figuras 8 y 9 indican que el uso de una normal para modelizar las pulsaciones por minutos es coherente pues en ambos casos, el polígono de frecuencias relativas acumuladas de la muestra tipificada no se separa mucho de la función de distribución de una normal N(0,1).

Por último, con la figura 10 observamos que en la franja de las 57 y 69 pulsaciones se acumulan más mujeres, pero, luego a partir de 70 pulsaciones se acumulan más hombres.

Intervalos de confianza del 95,0%

	Media	Límite Inferior	Límite Superior
P.P.M Hombres	73,3692	71,9134	74,825
P.P.M Mujeres	74,1538	72,1455	76,1622

	Sigma	Límite Inferior	Límite Superior
P.P.M Hombres	5,87518	5,01026	7,10385
P.P.M Mujeres	8,10523	6,912	9,80026

En este caso como ya hemos comentado, las variables se pueden aproximar por una normal, además la podemos tipificar a una N(0,1), donde podemos dar un intervalo de confianza aproximado del 95% para la media , y para la varianza comparándola mediante métodos con una chi cuadrado con n-1 grados libertad.

Además podemos calcular también el intervalo de confianza del 95% de la diferencia de medias. Para ver si suponemos varianzas iguales o distintas vemos que el intervalo para el cociente de varianzas es [0320483; 0,86143]. Como el 1 no pertenece no podemos suponer varianzas iguales. Por tanto calculamos el intervalo para la diferencia de medias suponiendo varianzas distintas, y este es [-3,790; 1,112].

Contraste de hipótesis

1. Bilaterales

Usando el lenguaje de programación R hemos sometido a varias hipótesis bilaterales la media y la desviación típica obteniendo intervalos (con una precisión de 0.001) que nos indican que si la hipótesis nula (un determinado valor del parámetro) pertenece a este, entonces no se rechazaría. Los resultados son los siguientes:

- IntervaloMediaPulsacionesHombres = [71.915, 74.824]
- IntervaloMediaPulsacionesMujeres= [72.734, 76.681]
- IntervaloDesviaciónPulsacionesHombres = [5.012, 7.103]
- IntervaloDesviaciónPulsacionesMujeres = [6.798, 9.636]

2. Unilaterales

Mediante las fórmulas proporcionadas en el curso sobre el contraste de hipótesis, deducidas por el teorema de Karlin-Rubin, dada la muestra vemos para el contraste de hipótesis cuya hipótesis nula es que la media sea menor o igual que un determinado valor, para varianza desconocida, se puede ver cuál es el valor de la media en el cual el p-valor es mayor que el nivel de significación, obteniendo así el intervalo en el que se acepta la hipótesis nula en tanto en hombres como mujeres:

- IntervaloMediaPulsacionesHombre= [72.153, ∞)
- IntervaloMediaPulsacionesMujeres=[72.476,∞)

•

Del mismo modo, poniendo como hipótesis nula que sea mayor o igual que un determinado valor, el intervalo de valores donde se acepta la hipótesis nula es el siguiente:

- IntervaloMediaPulsacionesHombre= (-∞,74,585]
- IntervaloMediaPulsacionesMujeres=(-∞,75.831]