Tipología y ciclo de vida de los datos: Heart Attack Analysis & Prediction dataset

Autores : Soulaiman el Hamri y Eloy Pérez González

Enero 2023

Contents

1. Descripción del dataset

El dataset elegido es:

https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

Es un juego de datos que contiene la información de diferentes parámetros que se han medido al realizar un análisis cardíaco. Estos datos se han obtenido de personas de diferente rango de edades y sexo.

Las preguntas que se pretenden responder son: ¿Qué sexo tiene más colesterol? ¿Cómo afecta el aumento de los niveles del colesterol a la frecuencia cardíaca?

2. Integración y selección de los datos de interés a analizar

En este apartado vamos a explorar el conjunto de datos con el que vamos a trabajar y seleccionar los datos que nos serán útiles para estudiar.

El primer paso es cargar el dataset con los datos originales:

```
# Cargamos el juego de datos heart_in.csv
path = '../data/heart_in.csv'
heartAttackData <- read.csv(path)
```

A continuació verificamos la estructura del juego de datos. Vemos el número de columnas que tenemos y ejemplos del contenido de las filas.

```
dim(heartAttackData)
```

```
## [1] 303  14
```

```
str(heartAttackData)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(heartAttackData)
```

```
##   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
## 1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
```

```
## 2  37   1  2    130  250   0       1       187    0     3.5   0   0     2       1
## 3  41   0  1    130  204   0       0       172    0     1.4   2   0     2       1
## 4  56   1  1    120  236   0       1       178    0     0.8   2   0     2       1
## 5  57   0  0    120  354   0       1       163    1     0.6   2   0     2       1
## 6  57   1  0    140  192   0       1       148    0     0.4   1   0     1       1
```

Podemos ver en el data frame que tenemos 303 observaciones y un total de 14 variables. Por cada variable podemos ver su tipo y a continuación un pequeño subconjunto de datos como ejemplo. Hemos podido comprobar que las variables contenidas en el fichero y sus tipos se corresponden a las que se han cargado.

A continuación se explica el significado de cada variable:

Variable	Descripción
age	Edad del paciente.
sex	Sexo del paciente. En la web donde se ha obtenido el dataset no se especifica el significado de cada valor del sexo (hombre o mujer) • Valor 0: Se desconce. • Valor 1: Se desconoce.
cp	Tipo de dolor en el pecho. • Valor 1: angina típica. • Valor 2: angina atípica. • Valor 3: dolor no anginoso. • Valor 4: asintomático.
trtbps	Presión arterial en reposo (en mm Hg).
chol	Colesterol en mg/dl obtenido a través del sensor BMI.
fbs	(Azúcar en sangre en ayunas > 120 mg/dl) (1 = verdadero; 0 = falso).
restecg	Resultados electrocardiográficos en reposo. • Valor 0: normal. • Valor 1: tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV). • Valor 2: mostrar hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes.
thalachh	Frecuencia cardíaca máxima alcanzada.
exng	Angina inducida por el ejercicio (1 = sí; 0 = no).
oldpeak	Pico anterior.
slp	Pendiente.
caa	Número de vasos grandes (0-3).
thall	Desconocido.
output	Es el target. • 0= menos posibilidades de ataque al corazón. • 1= más posibilidades de ataque al corazón.

Los hechos numéricos que se quieren estudiar serán los siguientes:

- sex, chol y thalachh.

Vamos a realizar la subselección de los datos a estudiar:

```
# Seleccionamos las variables sex, chol y thalachh
myvars <- c("sex", "chol", "thalachh")
subHeartAttackData <- heartAttackData[myvars]
```

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

A continuación vamos a procesar los datos para elminar los que sean erróneaos o redundantes. También vamos a ver si podemos eliminar algunos atributos de los datos.

Primeramente, vamos a ver si hay valores vacíos o nulos:

```
# Buscamos campos nulos y vacíos
print("Valores nulos:")
```

```
## [1] "Valores nulos:"
```

```
colSums(is.na(subHeartAttackData))
```

```
##      sex     chol thalachh
##        0        0        0
```

```
print("Campos vacíos:")
```

```
## [1] "Campos vacíos:"
```

```
colSums(subHeartAttackData == "")
```

```
##      sex     chol thalachh
##        0        0        0
```

Podemos ver que el subdataset no contiene ningún campo nulo o vacío.

3.2. Identifica y gestiona los valores extremos.

A continuación vamos a identificar y gestionar los valores extremos. Para detectar dichos valores extremos o *outliers,* se va a obtener los rangos de los valores maximos y minimos de cada variable y a continuación se va a representar los datos mediante graficos de cajas (*boxplots*).

Primero empezamos mirando el rango de valores maximos y minimos de cada variable:

```
library(ggplot2)

range(subHeartAttackData$sex)
```

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

```
range(subHeartAttackData$chol)
```

```
## [1] 126 564
```

```
range(subHeartAttackData$thalachh)
```

```
## [1]  71 202
```

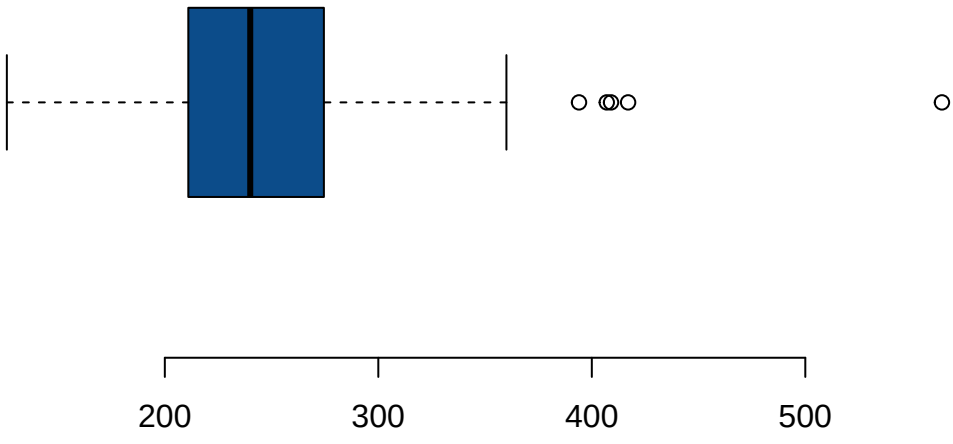
Sabemos que sex es una variable categórica, vamos a ver si sus valores numéricos tiene un significado categórico definido.

```
unique(subHeartAttackData$sex)
```

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

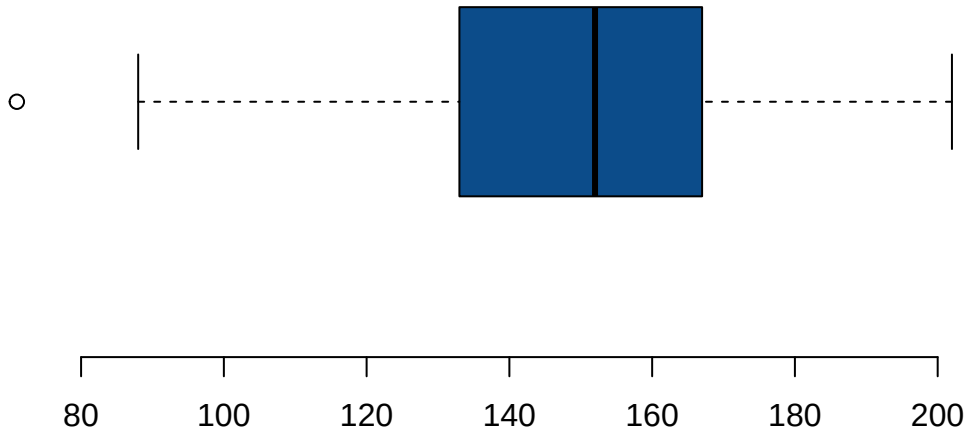
Ahora nos falta por ver si hay valores extremos en chol y thalachh. En caso que hayan, se ha decidido eliminar las observaciones de dichos valores extremos ya que se consideran valores erróneos.

```
#chol
g_caja<-boxplot(subHeartAttackData$chol, col="#0c4c8a", frame.plot=F, horizontal = TRUE)
```



```
#Eliminamos outliers:
subHeartAttackData<-subHeartAttackData[!(subHeartAttackData$chol %in% g_caja$out),]
```

```
#thalachh
g_caja<-boxplot(subHeartAttackData$thalachh, col="#0c4c8a", frame.plot=F, horizontal = TRUE)
```




```
#Eliminamos outliers:
subHeartAttackData<-subHeartAttackData[!(subHeartAttackData$thalachh %in% g_caja$out),]
```

Tras eliminar los outliers comprobamos con cúantas observaciones nos quedamos:

```
dim(subHeartAttackData)
```

```
## [1] 297   3
```

Se puede comprobar que se han perdido 6 filas de datos, es decir, alrededor del 1,9%, lo que no supone un problema.

Por último, vamos a extraer el datset con los datos finales que se van a analizar.

```
write.csv(subHeartAttackData, "../data/heart_out.csv")
```

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar

Para los análisis se separaran los datos en dos grupos, en función del sexo. Para cada grupo observaremos los valores de colesterol y frecuencia cardíaca máxima y compararemos las medias, para ver si existen diferencias entre las poblaciones.

Por otro lado intentaremos probar la relación entre colesterol y frecuencia cardíaca mediante una regresión lineal.

Para empezar crearemos una etiqueta categórica para cada sexo, que es útil para ciertos tipos de test y representaciones visuales:

```
subHeartAttackData$sex_name[subHeartAttackData$sex == 0] = "sex0"
subHeartAttackData$sex_name[subHeartAttackData$sex == 1] = "sex1"
```

Seguidamente separamos el dataset en dos en función del sexo:

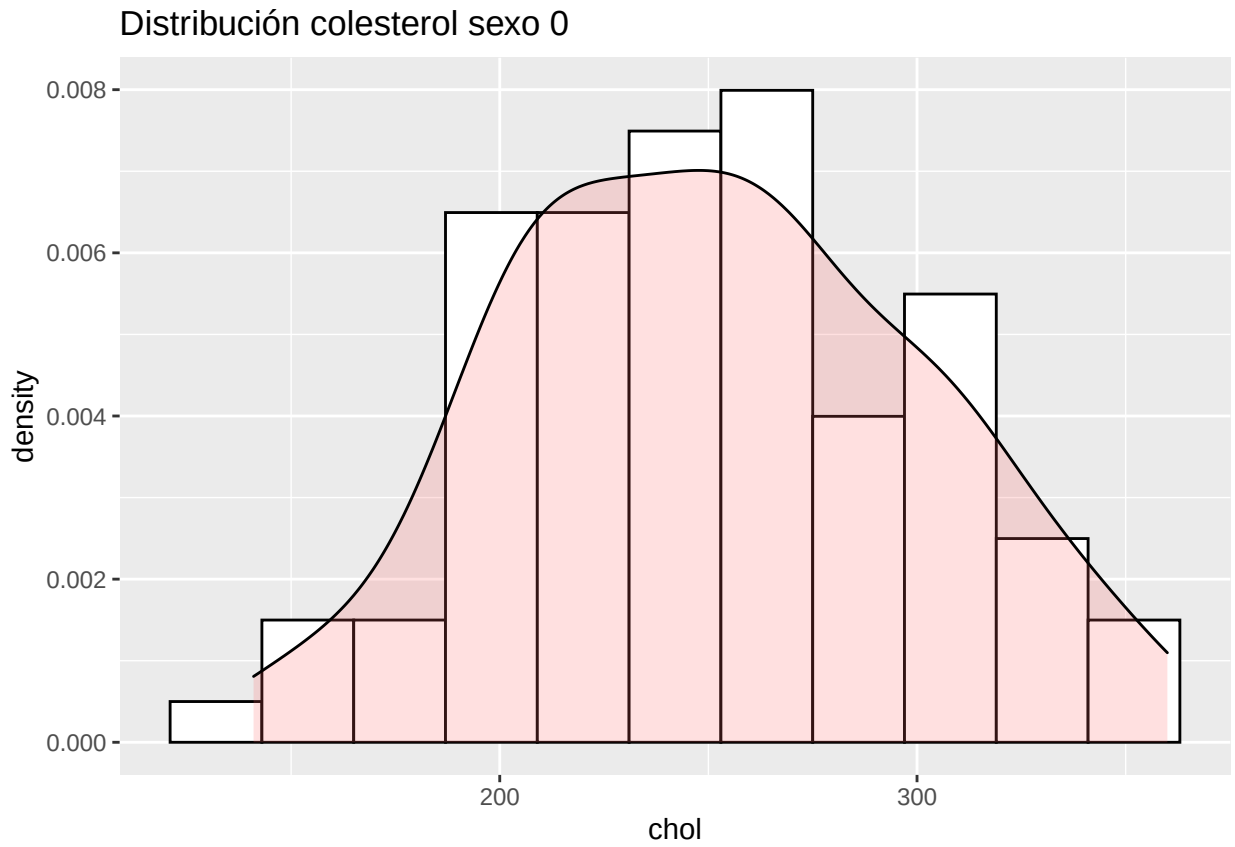
```
sex0Data = subHeartAttackData[subHeartAttackData$sex == 0,]
sex1Data = subHeartAttackData[subHeartAttackData$sex == 1,]
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

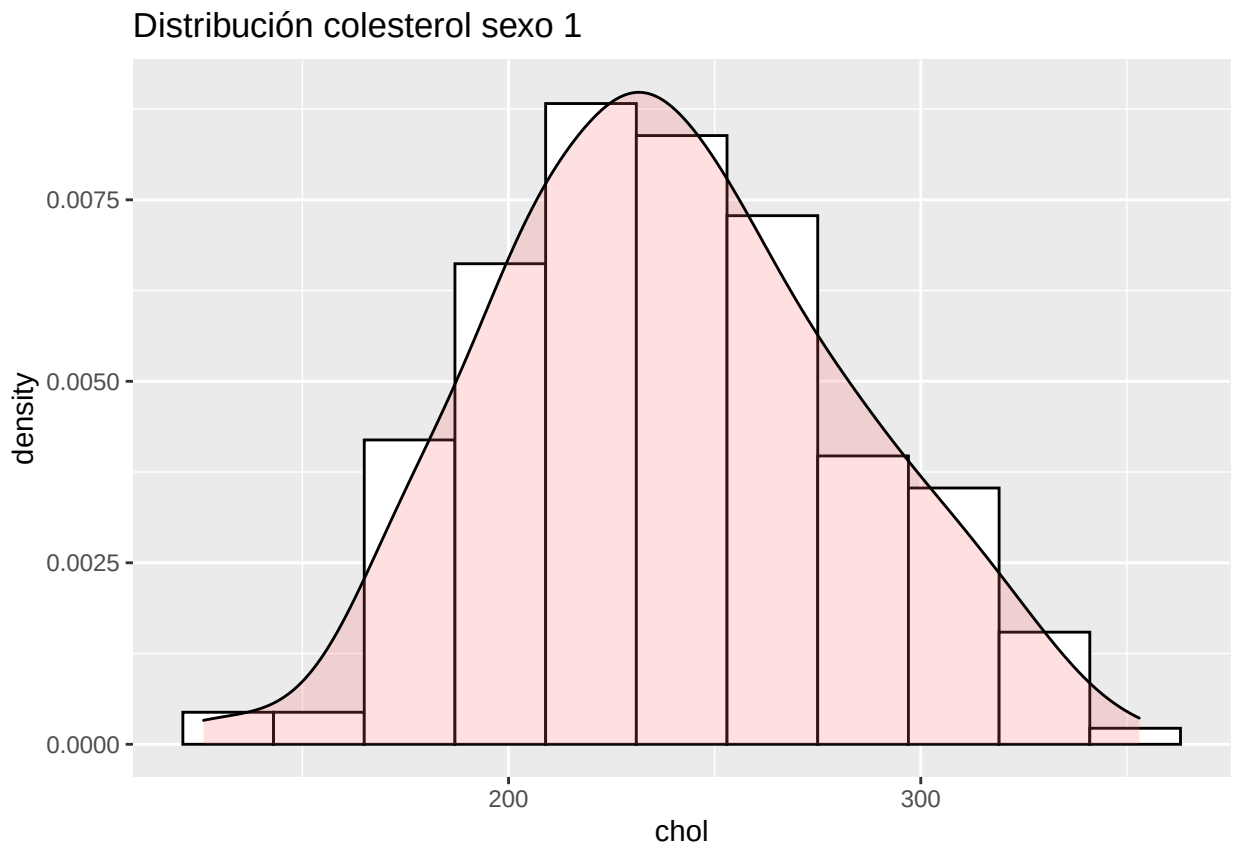
4.2.1. Normalidad y homoscedasticidad en colesterol

Vamos a representar en primer lugar la distribución del colesterol para cada sexo:

```
ggplot(sex0Data, aes(x=chol)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=22, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  ggtitle("Distribución colesterol sexo 0")
```




```
ggplot(sex1Data, aes(x=chol)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=22, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  ggtitle("Distribución colesterol sexo 1")
```



Observamos que en ambos casos parece seguirse una distribución similar a la normal. Procedemos a aplicar la prueba de Shapiro-Wilk para comprobar si efectivamente estamos ante una distribución que podemos considerar normal.

```
shapiro.test(sex0Data$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sex0Data$chol
## W = 0.98791, p-value = 0.57
```

```
shapiro.test(sex1Data$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sex1Data$chol
## W = 0.99362, p-value = 0.5218
```

Con un p-valor superior a 0.5 en ambos casos, no rechazamos la hipótesis nula y aceptamos que la distribución del colesterol es normal para ambos sexos.

Debido a que los grupos siguen una distribución normal respecto de las medidas de colesterol, aplicamos el test de Levene para comprobar si podemos considerar que tienen varianzas similares:

```r
library(car)
leveneTest(chol ~ sex_name, data = subHeartAttackData)
```

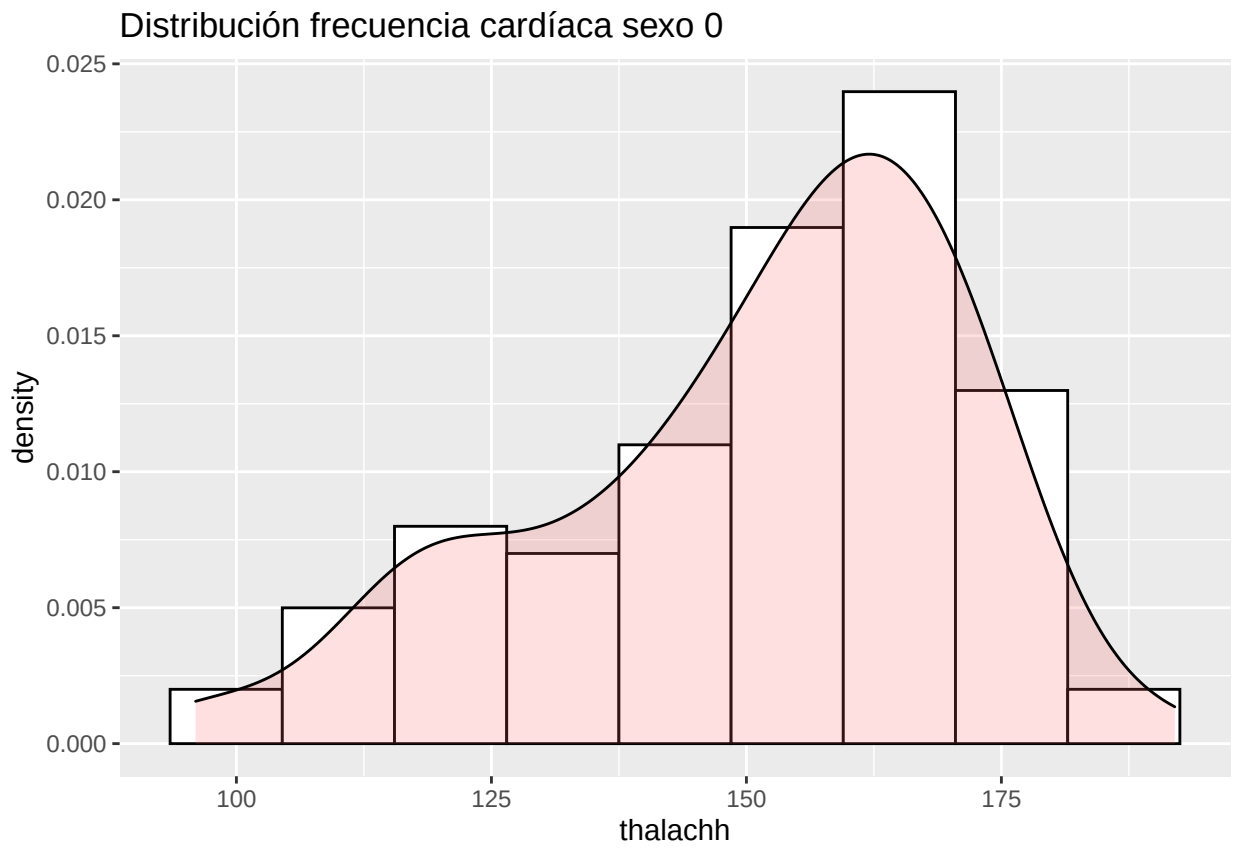
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  3.0947 0.07958 .
##       295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comprobamos que el p-valor de la prueba de Levene es superior a 0.05, por lo que no se rechaza la hipótesis nula y se considera que las muestras tienen varianzas similares.

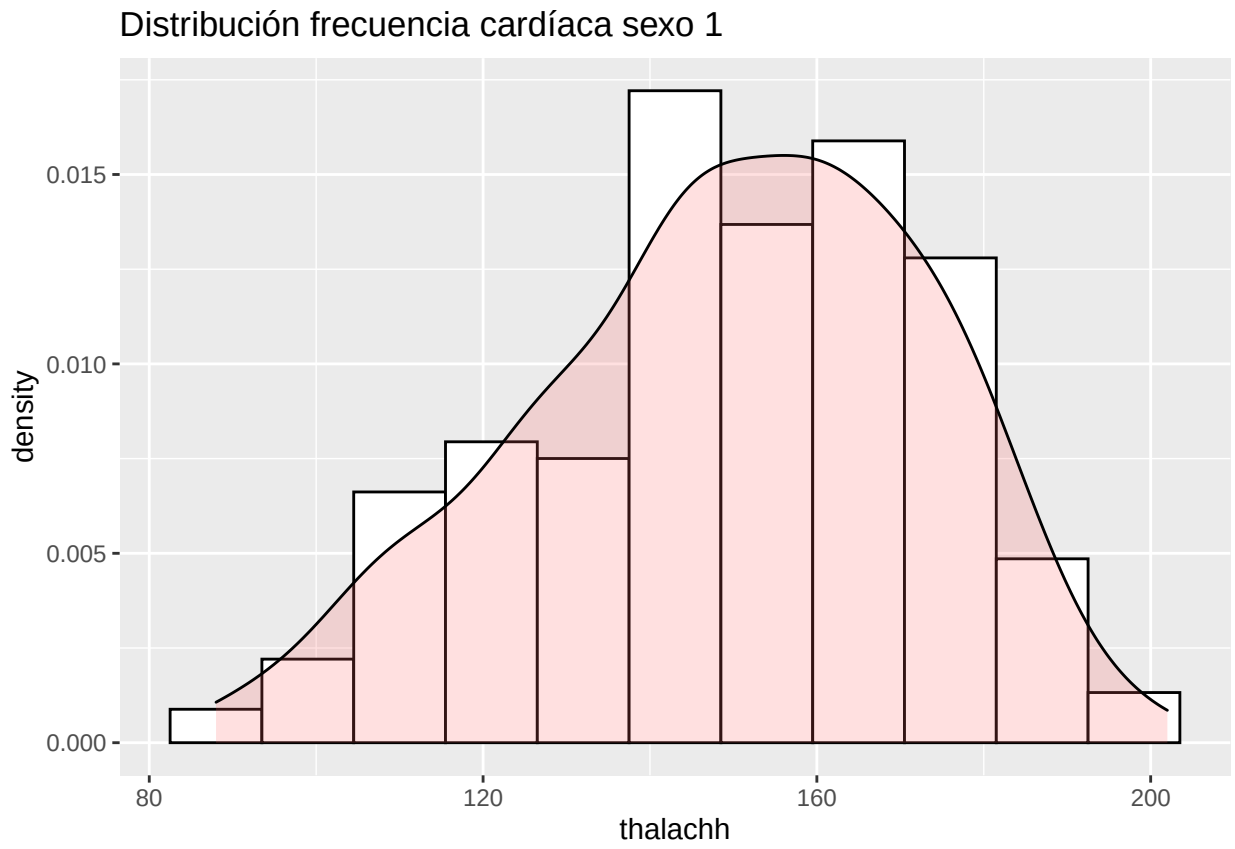
4.2.2. Normalidad y homoscedasticidad en frecuencia cardíaca

Vamos ahora con los datos de frecuencia cardíaca, los graficamos de similar manera que el colesterol:

```r
ggplot(sex0Data, aes(x=thalachh)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=11, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  ggtitle("Distribución frecuencia cardíaca sexo 0")
```



```
ggplot(sex1Data, aes(x=thalachh)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=11, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  ggtitle("Distribución frecuencia cardíaca sexo 1")
```



Apreciamos en ambos casos que una cola es sensiblemente más alargada que la otra, por lo que es posible que no nos encontremos ante una distribución normal. Aplicamos la prueba de Shapiro-Wilk para comprobarlo:

```
shapiro.test(sex0Data$thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sex0Data$thalachh
## W = 0.94619, p-value = 0.0009105
```

```
shapiro.test(sex1Data$thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sex1Data$thalachh
## W = 0.98244, p-value = 0.01133
```

Con un contundente p-valor cercano al cero, rechazamos la hipótesis nula y entendemos que las distribuciones no siguen un patrón normal.

Los test de Shapiro nos indica que no se sigue una distribución normal en ninguno de los grupos para la frecuencia cardíaca, por lo que aplicamos el test no paramétrico de Fligner-Killeen para conocer si las varianzas de los grupos son similares:

```
fligner.test(thalachh ~ sex_name, data = subHeartAttackData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  thalachh by sex_name
## Fligner-Killeen:med chi-squared = 3.0415, df = 1, p-value = 0.08116
```

Con un p-valor de 0.08 no rechazamos la hipótesis nula y consideramos que los grupos tienen varianzas similares.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos

4.3.1. Comparación de medias de colesterol por sexo

En primer lugar, aplicaremos un análisis estadístico descriptivo para conocer cada uno de los grupos. En primer lugar vamos a ver en que rangos se distribuyen los datos.

```
summary(sex0Data$chol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   141.0   213.5   249.0   251.6   288.0   360.0
```

```
cat("Var: ", var(sex0Data$chol), "\n")
```

```
## Var:  2415.09
```

```
cat("Sd: ", sd(sex0Data$chol))
```

```
## Sd:  49.14357
```

```
summary(sex1Data$chol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   208.0   234.5   239.3   268.5   353.0
```

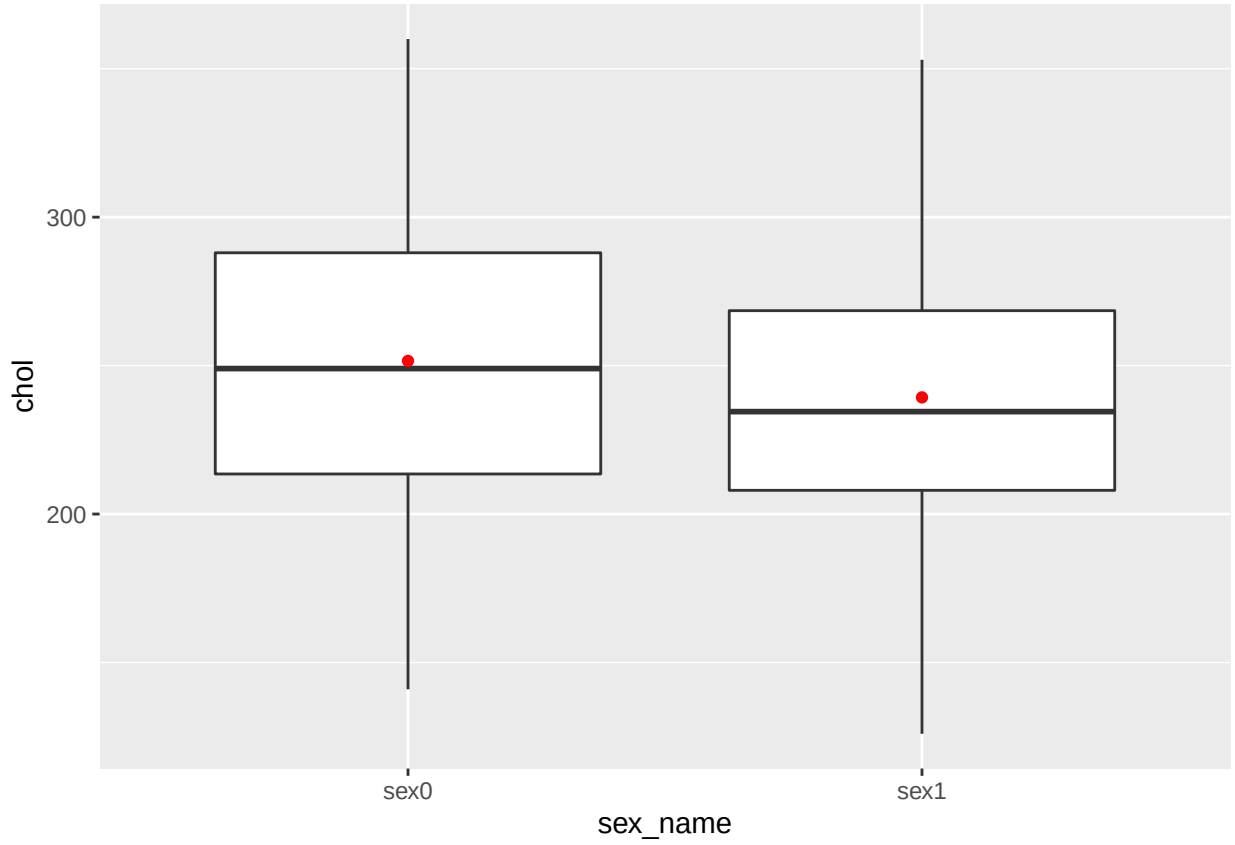
```
cat("Var: ", var(sex1Data$chol), "\n")
```

```
## Var:  1839.236
```

```r
cat("Sd: ", sd(sex1Data$chol))
```

```
## Sd:  42.88631
```

```r
ggplot(subHeartAttackData, aes(x = sex_name, y = chol)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", color="red", fill="red")
```



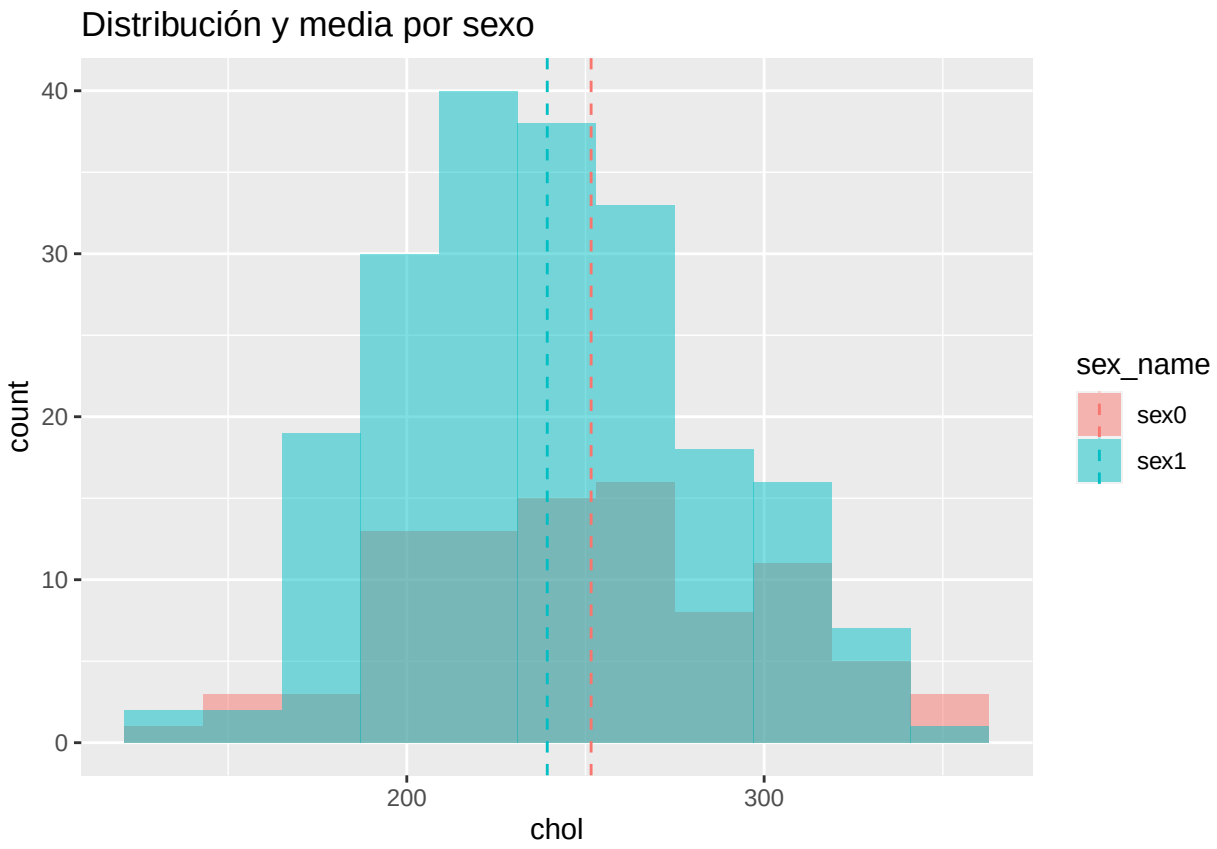
Apreciamos que para el sexo 0 tenemos tanto una media como una mediana superior a la del sexo 1, además de un rango intercuartílico mas amplio.

Vamos a hacer una visualización para ver que podemos extraer de la distribución y de la media:

```r
sex_means <- data.frame (
  sex_name = c("sex0", "sex1"),
  chol = c(
    mean(subHeartAttackData$chol[subHeartAttackData$sex_name == "sex0"]),
    mean(subHeartAttackData$chol[subHeartAttackData$sex_name == "sex1"])
  )
)


ggplot(subHeartAttackData, aes(x=chol, fill=sex_name)) +
  geom_histogram(binwidth=22, alpha=.5, position="identity") +
  geom_vline(
```

```
    data=sex_means, aes(xintercept=chol,  colour=sex_name),
    linetype="dashed", linewidth=1
  ) +
  ggtitle("Distribución y media por sexo")
```



Vemos que la media de colesterol parece ser superior para las personas del sexo 0. Lo comprobamos aplicando la prueba de t de Student:

```
t.test(chol ~ sex_name, data = subHeartAttackData, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  chol by sex_name
## t = 2.0622, df = 153.13, p-value = 0.02044
## alternative hypothesis: true difference in means between group sex0 and group sex1 is greater than 0
## 95 percent confidence interval:
##  2.425925      Inf
## sample estimates:
## mean in group sex0 mean in group sex1
##           251.5824           239.3010
```

Observamos que tenemos un p-valor menor que 0.05, por lo que rechazamos la hipótesis nula y podemos afirmar con un 95% de confianza que el colesterol es superior en personas del sexo 0.

4.3.2. Comparación de medias de frecuencia cardíaca por sexo

Vamos a examinar ahora los datos para la frecuencia cardíaca.

```
summary(sex0Data$thalachh)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    96.0   138.5   157.0   150.9   165.5   192.0
```

```
cat("Var: ", var(sex0Data$thalachh), "\n")
```

```
## Var:  422.4408
```

```
cat("Sd: ", sd(sex0Data$thalachh))
```

```
## Sd:  20.55336
```

```
summary(sex1Data$thalachh)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    88.0   132.0   151.0   149.3   168.0   202.0
```

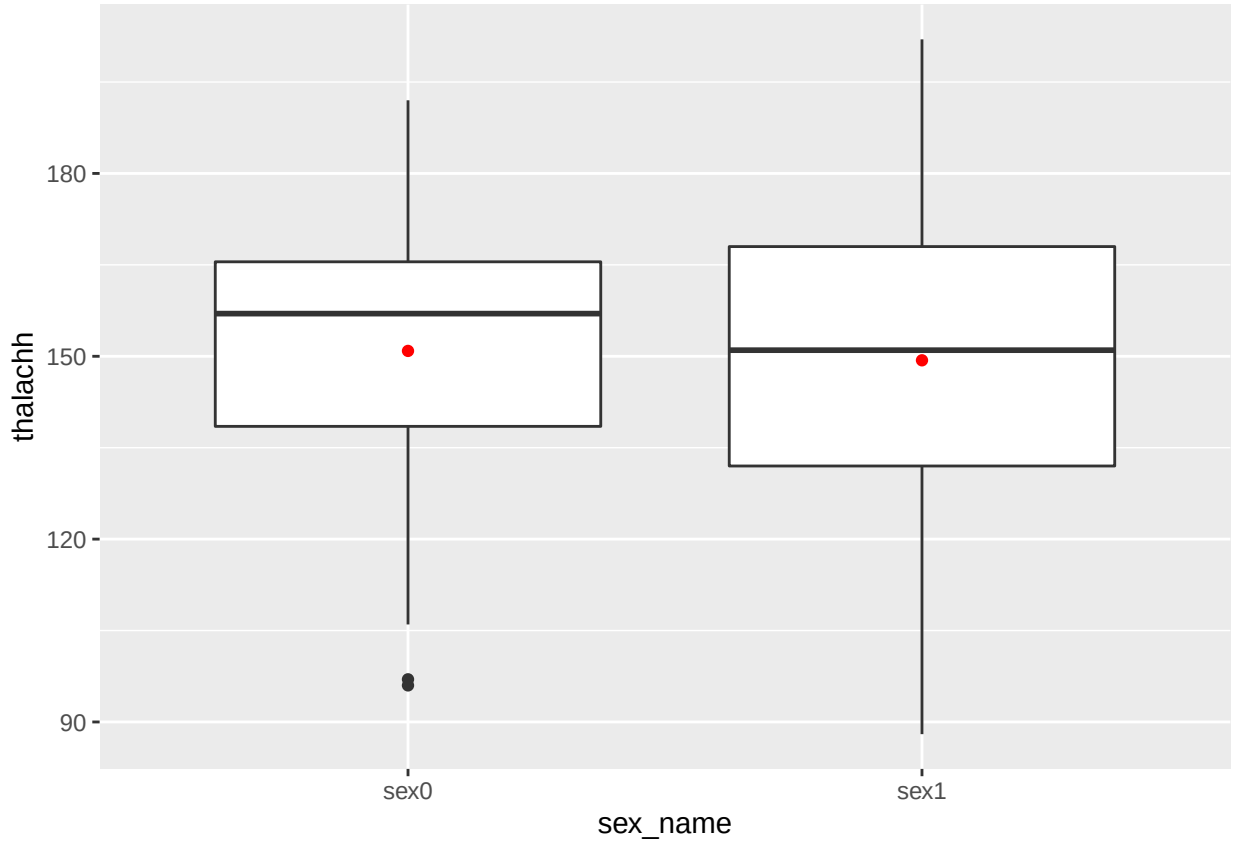
```
cat("Var: ", var(sex1Data$thalachh), "\n")
```

```
## Var:  555.3474
```

```
cat("Sd: ", sd(sex1Data$thalachh))
```

```
## Sd:  23.56581
```

```
ggplot(subHeartAttackData, aes(x = sex_name, y = thalachh)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", color="red", fill="red")
```

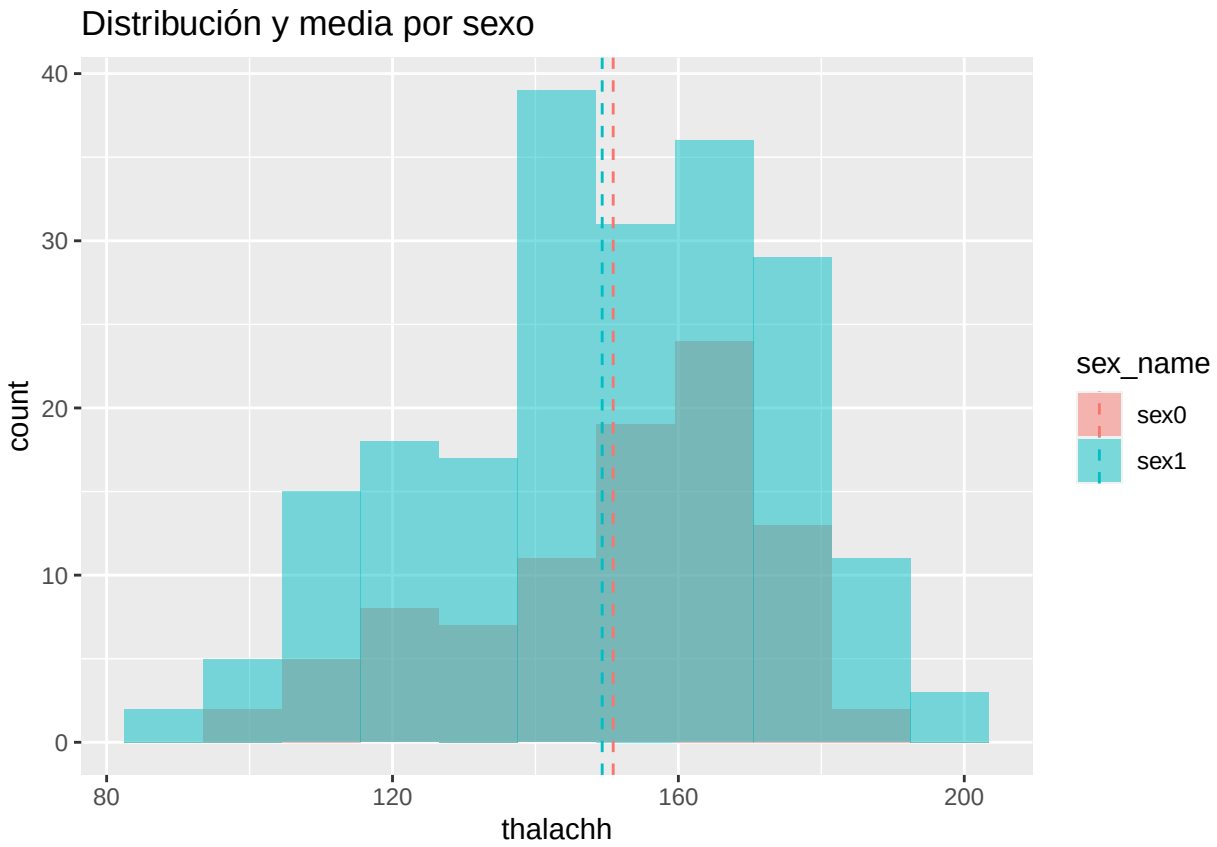


En este caso, podemos ver que la mediana del sexo 0 es superior y el rango intercuartílico es más estrecho. También se aprecia visualmente que la media del sexo 0 es superior, pero por muy poco.

Vamos a ver la media sobre la distribución:

```
sex_means <- data.frame (
  sex_name = c("sex0", "sex1"),
  thalachh = c(
    mean(subHeartAttackData$thalachh[subHeartAttackData$sex_name == "sex0"]),
    mean(subHeartAttackData$thalachh[subHeartAttackData$sex_name == "sex1"])
  )
)


ggplot(subHeartAttackData, aes(x=thalachh, fill=sex_name)) +
  geom_histogram(binwidth=11, alpha=.5, position="identity") +
  geom_vline(
    data=sex_means, aes(xintercept=thalachh,  colour=sex_name),
    linetype="dashed", linewidth=1
  ) +
  ggtitle("Distribución y media por sexo")
```



Como comentábamos antes, se aprecia que las medias de frecuencia cardíaca son bastante similares, siendo la del sexo 0 solo ligeramente superior.

Vamos a aplicar la prueba de Mann-Whitney (ya que no tenemos distribuciones normales) para ver si podemos considerar que la media de las poblaciones es distinta:

```
wilcox.test(thalachh ~ sex_name, data = subHeartAttackData, alternative="two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  thalachh by sex_name
## W = 9750.5, p-value = 0.5805
## alternative hypothesis: true location shift is not equal to 0
```

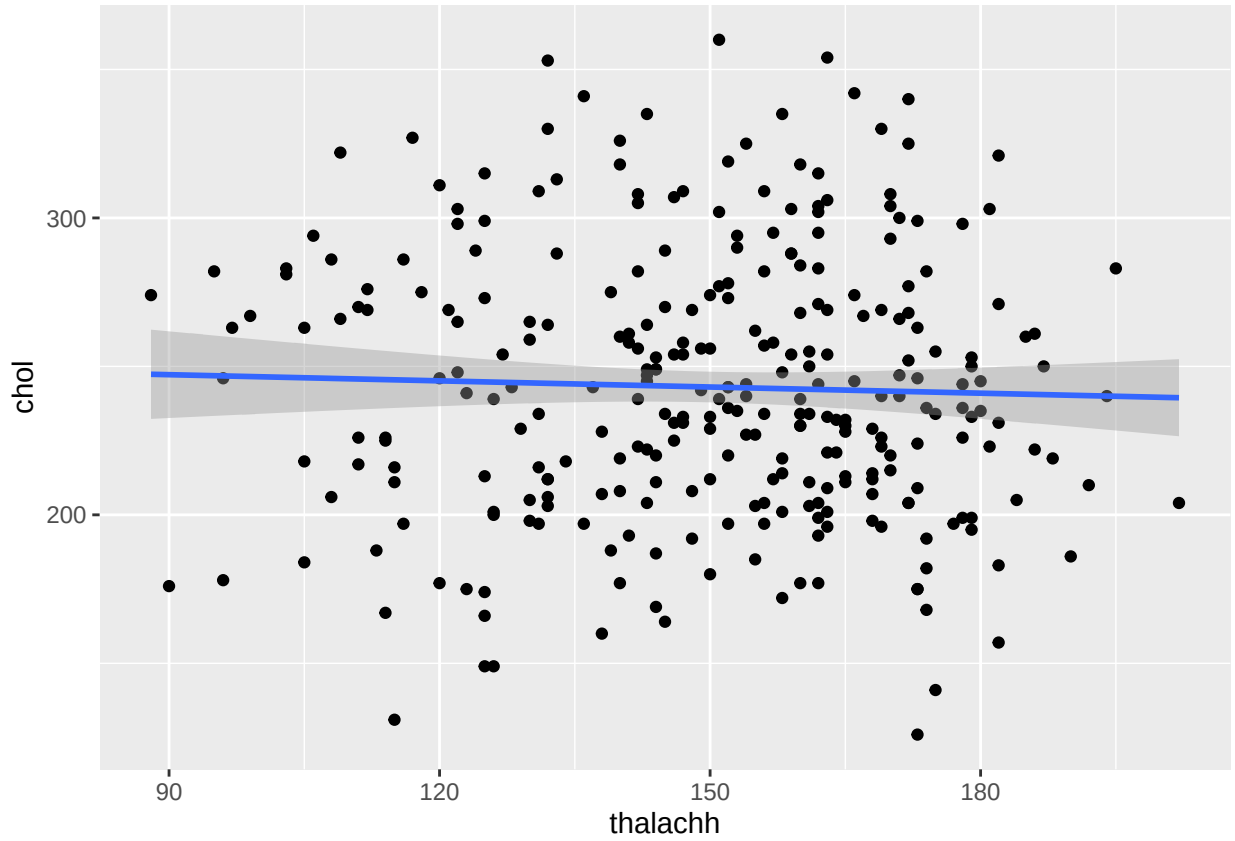
Con un p-valor superior a 0.05 no podemos rechazar la hipótesis nula, por lo que consideramos que la media de la frencuencia cardíaca máxima de las poblaciones de ambos sexos es similar.

4.3.3. Correlación entre colesterol y frecuencia cardíaca

Por último vamos a examinar si existe alguna relación entre el nivel de colesterol y la frecuencia cardíaca máxima alcanzada. Para ello emplearemos una regresión lineal.

Graficaremos en primer lugar la regresión, a ver que se puede observar.

```
ggplot(subHeartAttackData,aes(thalachh, chol)) +
  geom_point() +
  geom_smooth(method='lm')
```



De un simple vistazo ya nos es sencillo percibir que no parece que el colesterol este relacionado del forma alguna con la frecuencia cardíaca.

Vamos a examinar los valores obtenidos de la regresión con más detalle:

```
lm(thalachh ~ chol, data = subHeartAttackData)
```

```
##
## Call:
## lm(formula = thalachh ~ chol, data = subHeartAttackData)
##
## Coefficients:
## (Intercept)         chol
##   154.04727     -0.01743
```

```
summary(lm(thalachh ~ chol, data = subHeartAttackData))
```

```
##
## Call:
## lm(formula = thalachh ~ chol, data = subHeartAttackData)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.272 -15.896   3.226  17.560  51.508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.04727    7.21687  21.345   <2e-16 ***
## chol         -0.01743    0.02919  -0.597    0.551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 295 degrees of freedom
## Multiple R-squared:  0.001207,   Adjusted R-squared:  -0.002179
## F-statistic: 0.3563 on 1 and 295 DF,  p-value: 0.551
```

Podemos apreciar que el coeficiente de correlación es prácticamente 0, indicando que el colesterol y la frecuencia cardíaca no se encuentran para nada relacionados. Si observamos el coeficiente de determinación, vemos que del mismo modo el colesterol no explica la varianza de la frecuencia cardíaca.

Podemos concluir por tanto que el colesterol y la frecuencia cardíaca no están relacionados de una forma lineal.

5. Representación de los resultados a partir de tablas y gráficas

Se han representado a lo largo de los apartados 3 y 4.

6. Resolución del problema

Los resultados que hemos podido obtener són:

- Entre sexo 0 y sexo 1 las muestras del colesterol y frecuencia cardíaca tienen una varianza similar.
- Los sexo 0 tienen el colesterol más alto que sexo 1.
- Entre sexo 0 y sexo 1 no se aprecia diferencias significativas en la frecuencia cardíaca.
- No exisite relación lineal entre el colesterol y la frecuencia cardíaca.

A las preguntas pleantadas al inicio, se ha podido ver que las personas de sexo 0 sufren más de colesterol que las personas de sexo 1 y que no hay una relación aparente entre el colesterol y la frecuencía cardíaca.

7. Código

El código de la práctica es en R y se puede encontrar en el fichero hearthattack.Rmd del siguiente enlace a Github: https://github.com/SulaimanUOC/HeartAttack/tree/main/code

8. Vídeo

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/...),,) junto con enlace al repositorio Git entregado.

9. Contribuciones

Contribuciones	Firma
Investigación previa	Eloy y Soulaiman
Redacción de las respuestas	Eloy y Soulaiman
Desarrollo del código	Eloy y Soulaiman
Participación en el vídeo	Eloy y Soulaiman