

PEC2_IE

Amelia Martínez Sequera

Junio 2020

Enunciados.

a) Se está interesado en conocer que factores están asociados al riesgo nutricional en ancianos, medidos mediante una escala que clasifica a los ancianos de 1 a 20, donde 20 es un alto riesgo y 1 bajo riesgo. Como posibles factores entre otros se mide el ser hombre o mujer, el comer sólo o acompañado y el beber o no vino en las comidas. Se dispone de 150 ancianos para resolver la situación. Se desea contrastar si por separado son factores de riesgo de malnutrición.

Se trata de variables cualitativas (categóricas) independientes: hombre/mujer(sexo), comer solo/no (solo), vino/no (vino),...dicotómicas, tomarían el valor 0 para la opción de menos riesgo, por ejemplo, comer acompañado, y 1 para la opción de riesgo, por ejemplo, comer solo. El predictor, en cambio, es una variable discreta cuantitativa.

Primero se puede estudiar gráficamente la distribución del riesgo nutricional, representar los datos mediante un histograma y superponer la curva de una distribución normal con la misma media y desviación estándar que muestran los datos. Si éstos se distribuyen de forma normal, se puede contrastar si existe relación entre el riesgo de malnutrición y cualquiera de las variables con un t test de datos no apareados.

Si no se distribuyen normalmente, se puede hacer un test no paramétrico de Wilcoxon-Mann-Whitney como alternativa (también conocido como Wilcoxon rank-sum test o u-test), o un test de Bootstrapping. Con el U-test no es necesario asumir que las muestras se distribuyen de forma normal pero, para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría...). Se debería comprobar antes la igualdad de varianza.

A medida que el número de observaciones aumenta, también lo hace la robustez del t-test frente a desviaciones de la normalidad. Por otro lado, el test de Mann-Whitney-Wilcoxon incrementa su sensibilidad a diferencias más allá de las medianas. Por ejemplo, aumenta el poder estadístico de detectar diferencias significativas en la probabilidad de que observaciones de un grupo superen a las del otro debido únicamente a diferencias en la dispersión de las poblaciones de origen y no porque sus medianas sean distintas. Nuestro objetivo no es identificar cualquier diferencia distribucional, por eso con este test se podrían obtener p-values que no se corresponden con la pregunta que el queremos responder.

Por este motivo, podría ser apropiado utilizar el test de Fisher, que es el test exacto para estudiar si existe asociación entre dos variables cualitativas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiriera la otra. En este caso el test de Fisher se aplicaría con una tabla de $2 \times K$ niveles, aunque los requerimientos de cálculo son altos. Se pueden dividir los niveles de riesgo (20) en 2 o 3 (alto/medio/bajo) para simplificar el modelo.

b) Se dispone de 45 mujeres obesas de las cuales 22 eran premenopáusicas y 23 postmenopáusicas. Se tomaron 3 medidas antropométricas como el índice de masa corporal, el índice cintura-cadera y cintura-muslo. Además se clasificó a las mujeres según su obesidad fuera androide o ginoide. Se está interesado en evaluar la relación entre la menopausia y cada uno de los índices de obesidad por separado. Así como si el tipo de obesidad se asocia con tener o no la menopausia.

Variables continuas independientes: IMC, cint-cad, cint-musl; variable cualitativa: obesidad (dicotómica). Grupos: Menopausia/No menopausia

(1) Para ver la relación menopausia y cada uno de los índices antropométricos se podría realizar un t-test para datos no apareados, tras verificar la normalidad y homocedasticidad de las variables, con un test de Shapiro Wilks o Kolmogorov-Smirnov (para la normalidad), o con un test de Levene, de Barlett o de Fligner (para la varianza). En el caso que no siguieran una distribución normal, realizaríamos el test de Mann-Whitney-Wilcoxon, que es una alternativa al t-test cuando las muestras no proceden de poblaciones con distribución normal (asimetría o colas) o porque tienen un tamaño demasiado reducido para poder afirmarlo.

Ambos test permiten realizar inferencia sobre la diferencia entre poblaciones, pero las hipótesis contrastadas no son las mismas. Mientras que el t-test siempre compara las medias de los grupos, el test de Mann-Whitney-Wilcoxon es menos específico. Si las distribuciones de las poblaciones subyacentes se diferencian únicamente en localización, entonces el test de Mann-Whitney-Wilcoxon compara medianas, para el resto de escenarios (poblaciones con distinta distribución, dispersión, asimetría...) contrasta si $P(X > Y) = P(Y > X)$.

Para el U-test no es necesario asumir que las muestras se distribuyen de forma normal, pero para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría...), e igualdad de varianza entre grupos (homocedasticidad).

#ejemplo:

```
# t.test(datos$IMC[datos$menopausa==0], datos$IMC[datos$menopausa==1], mu =  
0, var.equal = T, conf.level = 0.95)  
# wilcox.test(IMC~menopausia, data = datos, distribution = "exact", conf.  
int=0.95)
```

(2) Queremos saber si el tipo de obesidad está relacionado con la menopausia: dos variables independientes y cualitativas, con dos niveles cada una.

(2a) Test χ^2 de Pearson (test de independencia): es un test no paramétrico que se emplea para estudiar si existe asociación entre dos variables categóricas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable, cuando los datos son independientes. Se trata por lo tanto de una expansión del Z-test para dos proporciones cuando una de las variables estudiadas tiene dos o más niveles. En este caso, en el que ambas variables tienen dos niveles (tabla 2x2) ambos test χ^2 goodness of fit y Z-test para dos proporción son equivalentes. Cuantifica y resume cómo de distinto es el número de eventos observados en cada nivel con respecto al número esperado acorde con H_0 . Esto permite identificar si la desviación total es mayor que la que cabría esperar simplemente por azar. La distribución chi-cuadrado es siempre positiva, por lo que para calcular el p-value solo se tiene en cuenta la cola superior.

Si el número de observaciones esperadas para alguno de los niveles es igual o menor a 5 la aproximación por el test χ^2 no es buena.

Condiciones del test - Independencia: las observaciones de la muestra deben ser independientes unas de otras. - Muestreo aleatorio. - Tamaño de la muestra $< 10\%$ población. - Cada observación contribuye únicamente a uno de los niveles. - Tamaño: cada nivel debe tener al menos 5 eventos esperados (acorde a H_0) y el número de observaciones totales (n) > 30 . En caso de no cumplirse esta condición, el test pierde precisión. En este caso desconocemos la proporción de androide/ginoide, si alguna es inferior a 5 deberíamos usar el test de Fisher o hacer una simulación de p-value.

```
#chisq.test(tabla, simulate.p.value = TRUE, B = 5000)
```

(2b) También podríamos utilizar el test de Fisher, que es el test exacto para estudiar si existe asociación entre estas dos variables cualitativas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra. En este caso el test de Fisher se aplica para comparar dos variables categóricas con dos niveles cada una (tabla 2x2), pero también es posible utilizarlo con tablas 2xK niveles, aunque los requerimientos de cálculo son altos.

El test de Fisher es más preciso que sus equivalentes aproximados (test chi-square de independencia o G-test de independencia) cuando el número de eventos esperado por nivel es pequeño. Se recomienda utilizarlo siempre que sea posible (tiempo de computación) aunque para observaciones totales > 1000 los resultados de los test aproximados son muy parecidos.

Condiciones del test:

- Independencia: las observaciones de la muestra deben ser independientes unas de otras. - Muestreo aleatorio. - Tamaño de la muestra $< 10\%$ población. - Cada observación contribuye únicamente a uno de los niveles. - Las frecuencias marginales de columnas y filas tienen que ser fijas. Si esta condición no se cumple, el test de Fisher deja de ser exacto.

H_0 : Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

Ha: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

```
#fisher.test(x = tabla, alternative = "two.sided")
```

c) Se efectuó un ensayo aleatorio doble ciego para evaluar el efecto del inhibidor de la enzima conversiva de la angiotensina, trandolapril, sobre la neuropatía diabética, asignando a 23 pacientes el trandolapril y a otros 23 a un placebo. Como respuesta entre otras variables se midió la velocidad de conducción motora del nervio del perineo(m/s). Indica el análisis para comparar el nivel de dicha variable al finalizar el estudio.

Variable respuesta: Velocidad de conducción motora del nervio del perineo, cuantitativa y continua. 2 grupos: control y medicados. Datos no apareados.

Lo primero es comprobar que la variable cumple las condiciones de independencia, normalidad y homocedasticidad en ambos grupos con un test de Shapiro Wilks o Kolmogorov-Smirnov (para la normalidad), o con un test de Levene, de Barlett o de Fligner (para la varianza). Entonces realizaríamos un t Test convencional, comparando las medias entre control y pacientes medicados para datos no apareados. La hipótesis nula sería que las medias del grupo medicado y el grupo control son iguales. En R:

```
#t.test(datoscontrol,datosmedicados,alternative = "two.sided", mu = 0, var.equal = TRUE, conf.level = 0.95)
```

Si el valor de p es menor de 0.05 rechazaríamos la hipótesis nula de que las medias son iguales. También se puede hacer un contraste unilateral utilizando "less" (hipótesis alternativa del tipo $<$) y "greater" (hipótesis alternativa del tipo $>$).

Si se encuentran evidencias significativas de que los datos no proceden de poblaciones con distribución normal, se podría mirar la mediana, ya que por el tipo de datos, seguramente presentan una distribución simétrica. Entonces se puede utilizar un test no paramétrico basado en la mediana (Mann-Whitney-Wilcoxon test) o un test de Bootstrapping. Otra opción sería estudiar si los datos anómalos son excepciones que se pueden excluir del análisis.

Condiciones necesarias del test de Mann-Whitney-Wilcoxon:

- Los datos tienen que ser independientes.
- Los datos tienen que ser ordinales o bien se tienen que poder ordenar de menor a mayor.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales. Pero, para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría...).
- Igualdad de varianza entre grupos (homocedasticidad).

Al igual que ocurre con muchos test no paramétricos, el test de Mann–Whitney–Wilcoxon es menos potente que el t-test (tienen menos probabilidad de rechazar la H_0 cuando realmente es falsa) ya que ignora valores extremos. En el caso de los t-test, al trabajar con medias, si los tienen en cuenta. Esto hace a su vez que el test de Mann–Whitney–Wilcoxon sea una prueba más robusta que los t-test. En concreto, la pérdida de potencia es del 5%.

Si el tamaño de cada grupo fuera mayor que 30 (condición que no se cumple en este caso) se puede considerar que el t-test sigue siendo suficientemente robusto, aunque es necesario mencionarlo en las conclusiones.

d) Para estudiar el posible efecto beneficioso de la meditación trascendental para el tratamiento de pacientes con síntomas de ansiedad, se decide administrar a un grupo de 20 pacientes un protocolo de este tipo de actividad para ser realizado diariamente. Se mide el nivel de ansiedad de cada individuo al inicio del tratamiento y a los seis meses, a través de un test que proporciona una puntuación cuantitativa entre 0 y 100 puntos como valoración de dicho nivel de ansiedad. Se quiere comparar los niveles de ansiedad entre el inicio y a los seis meses para ver si se han producido cambios importantes.

La variable respuesta es el nivel de ansiedad (puntuación de 0-100 en el test), que es cuantitativa discreta. Se realiza el test antes y después del tratamiento, por lo que disponemos de dos muestras apareadas.

Ya que se trata de una variable discreta y el tamaño no es muy grande (<30), se podría realizar un test no paramétrico prueba de los rangos con signo de Wilcoxon, también conocido como Wilcoxon signed-rank test, que permite comparar poblaciones cuando sus distribuciones (normalmente interpretadas a partir de las muestras) no satisfacen las condiciones necesarias para otros test paramétricos. Es una alternativa al t-test de muestras dependientes cuando las muestras no siguen una distribución normal (muestran asimetría o colas) o cuando tienen un tamaño demasiado reducido para poder determinar si realmente proceden de poblaciones normales, como en este caso.

A la hora de elegir entre t-test o Wilcoxon signed-rank test, existen dos situaciones en las que, a priori, se puede recomendar utilizar un Wilcoxon signed-rank test antes que un t-test:

- Si el tamaño de las muestras es suficientemente grande para determinar (por métodos gráficos o contrastes de hipótesis) que la distribución de las poblaciones a comparar no es de tipo normal, en tal caso, los t-test no son adecuados, por lo que mejor emplear un Wilcoxon signed-rank test (Bootstrapping, regresión cuantílica, o test de permutación también podrían ser otras alternativa).
- Si el tamaño de las muestras no permite determinar con seguridad si las poblaciones de las que proceden se distribuyen de forma normal, y no se dispone de información que pueda orientar sobre la naturaleza de las poblaciones de origen (estudios anteriores, que sea un tipo de variable que se sabe que se distribuye casi siempre de forma normal...), entonces es más apropiado el

Wilcoxon signed-rank test ya que no requiere asumir la normalidad de las poblaciones.

Condiciones para la prueba de los rangos con signo de Wilcoxon:

- Los datos tienen que ser dependientes.
- Los datos tienen que ser ordinales, se tienen que poder ordenar de menor a mayor o viceversa.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales. Pero sea cual sea el tipo de distribución de las diferencias, tiene que ser simétrica.

A pesar de considerarse el equivalente no paramétrico del t-test, el Wilcoxon signed-rank test trabaja con medianas, no con medias. Tiene menos poder estadístico (menor probabilidad de rechazar la hipótesis nula cuando realmente es falsa) ya que ignoran valores extremos. En el caso de los t-test, al trabajar con medias, sí los tienen en cuenta. Esto a su vez, hace que el Wilcoxon signed-rank test sean una prueba más robusta que el t-test.

e) Se estudia un grupo de 33 pacientes afectados de Carcinoma hepatocelular, un grupo de 22 afectados únicamente de cirrosis y un grupo control de 31 donantes de sangre. Se determina la actividad celular NK frente a la línea celular K562 marcada con CR mediante la prueba de corta duración de citotoxicidad y el número de células CK en los tres grupos. (Nota la mediana de actividad celular en los hombres es de 39 unidades líticas / 10^7 de linfocitos y la mediana es de 28 y la media del número de células es de 178 y la mediana de 163)

Disponemos de 3 grupos (33 carcinoma hepatocel., 22 cirrosis, 31 control), y de 2 variables cuantitativas continuas: activ cel NK ($\mu=39 \times 10^{-7}$, mediana= 28×10^{-7}) y número de cel CK ($\mu=178$, mediana=163). Son muestras independientes.

Primero se debe comprobar la normalidad de las distribuciones de cada población (test Shapiro-Wilks, por ejemplo), o la igualdad de varianzas (homocedasticidad), aunque con los datos que nos facilitan ya vemos que el caso del recuento de células CK la mediana es un buen descriptor de esta variable.

Como la muestra es bastante grande, podemos usar un test paramétrico anova para comparar medias, ya que tenemos 3 grupos.

Si las poblaciones no siguieran una distribución normal o presentaran heterocedasticidad, realizaríamos la prueba de Kruskal-Wallis, también conocido como test H, que es la alternativa no paramétrica al test ANOVA de una vía para datos no pareados. Es una extensión del test de Mann-Whitney para más de dos grupos. Se trata por lo tanto de un test que emplea rangos para contrastar la hipótesis de que las muestras han sido obtenidas de una misma población. A diferencia del ANOVA en el

que se comparan medias, el test de Kruskal-Wallis contrasta si las diferentes muestras están equidistribuidas y que por lo tanto pertenecen a una misma distribución (población). Bajo ciertas simplificaciones puede considerarse que el test de Kruskal-Wallis compara las medianas.

H0: todas las muestras provienen de la misma población (distribución).

HA: Al menos una muestra proviene de una población con una distribución distinta.

Condiciones:

- No es necesario que las muestras que se comparan provengan de una distribución normal.
- Homocedasticidad: dado que la hipótesis nula asume que todos los grupos pertenecen a una misma población y que por lo tanto tienen las mismas medianas, es requisito necesario que todos los grupos tengan la misma varianza. Se puede comprobar con representaciones gráficas o con los test de Levene o Bartlett.
- Misma distribución para todos los grupos: la distribución de los grupos no tiene que ser normal pero ha de ser igual en todos (por ejemplo que todos muestren asimetría hacia la derecha).

En la mayoría de la bibliografía consultada consideran que el ANOVA es bastante robusto a la falta de normalidad sobre todo con muestras medianas o grandes. Solo recomiendan el uso del test de Kruskal-Wallis cuando las poblaciones a comparar sean claramente asimétricas, se cumpla que todas lo sean en la misma dirección y que la varianza sea homogénea. Si la varianza no es homogénea el test adecuado es un ANOVA con corrección de Welch. En los casos en los que se trata de datos puramente ordinales el ANOVA no es una opción.

```
#kruskal.test(activ celNK ~ grupo, data = datos)
#kruskal.test( numerocel ~ grupo, data = datos)
```

Ejercicio práctico.

1.Los datos.

```
library(foreign)
set.seed(12345)
dades140<-diabetes[sample(1:nrow(diabetes),140,replace=FALSE),]
View(dades140)
```

2.Las preguntas.

a) Comprueba la normalidad de las variables tiempo de seguimiento (TEMPSVIU), presión arterial sistólica(SBP). Muestra e interpreta los resultados.

Los análisis de normalidad, también llamados contrastes de normalidad, tienen como objetivo analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica. Pueden diferenciarse tres estrategias: las basadas en representaciones gráficas, en métodos analíticos y en test de hipótesis.

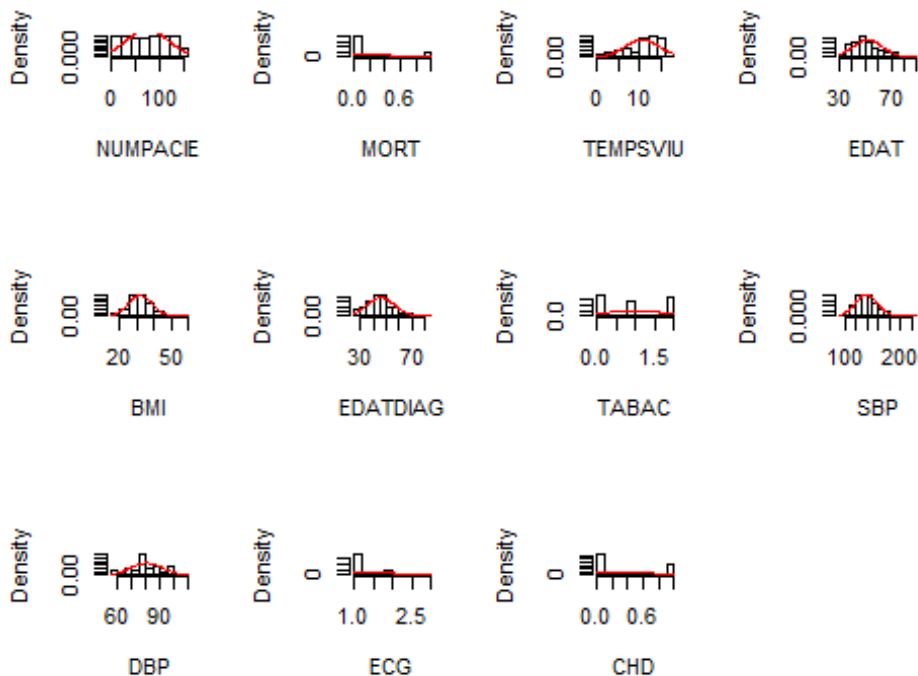
Histograma.

```
mvn(data=dades140, mvnTest="royston", univariatePlot="histogram")

## $multivariateNormality
##      Test      H      p value MVN
## 1 Royston 360.6621 4.176991e-71 NO
##
## $univariateNormality
##      Test Variable Statistic  p value Normality
## 1 Shapiro-Wilk NUMPACIE      0.9531      1e-04      NO
## 2 Shapiro-Wilk MORT          0.4461 <0.001      NO
## 3 Shapiro-Wilk TEMPSVIU      0.9206 <0.001      NO
## 4 Shapiro-Wilk EDAT          0.9571      2e-04      NO
## 5 Shapiro-Wilk BMI           0.9610      5e-04      NO
## 6 Shapiro-Wilk EDATDIAG      0.9666      0.0017      NO
## 7 Shapiro-Wilk TABAC         0.7741 <0.001      NO
## 8 Shapiro-Wilk SBP           0.9701      0.0036      NO
## 9 Shapiro-Wilk DBP           0.9755      0.0127      NO
## 10 Shapiro-Wilk ECG          0.5795 <0.001      NO
## 11 Shapiro-Wilk CHD          0.5994 <0.001      NO
##
## $Descriptives
##      n      Mean      Std.Dev Median  Min   Max   25th   75th
## NUMPACIE 140  74.8071429 43.4300722   75.5  1.0 149.0  36.750 112.250
## MORT      140   0.1642857 0.3718651    0.0  0.0   1.0   0.000   0.000
## TEMPSVIU 140  10.6350000 4.0081917   11.6  0.0  16.9   7.825  13.925
## EDAT      140  51.9571429 11.4721254   50.0 31.0  86.0  43.750  59.000
## BMI       140  31.9514286 6.9328380   31.6 18.2  59.7  26.575  35.725
## EDATDIAG 140  45.7714286 10.7089305   45.0 26.0  81.0  38.000  52.000
## TABAC     140   0.9714286 0.8561488    1.0  0.0   2.0   0.000   2.000
## SBP       140 138.8285714 19.9336824  138.0 98.0 222.0 125.500 150.500
## DBP       140  80.0571429 11.1959470   80.0 58.0 108.0  74.000  88.000
## ECG       140   1.3214286 0.5909127    1.0  1.0   3.0   1.000   2.000
## CHD       140   0.3428571 0.4763686    0.0  0.0   1.0   0.000   1.000
##
##      Skew      Kurtosis
## NUMPACIE -0.016370791 -1.23774333
## MORT      1.792673306  1.22256287
## TEMPSVIU -0.829053015 -0.19870570
## EDAT      0.714341588  0.08271588
## BMI       0.833222143  1.39641893
## EDATDIAG  0.690806813  0.79755585
## TABAC     0.053909569 -1.64339111
## SBP       0.703609707  1.30314693
## DBP      -0.002091442 -0.46785696
```



```
## ECG      1.646033114  1.58622566
## CHD      0.655040709 -1.58206947
```



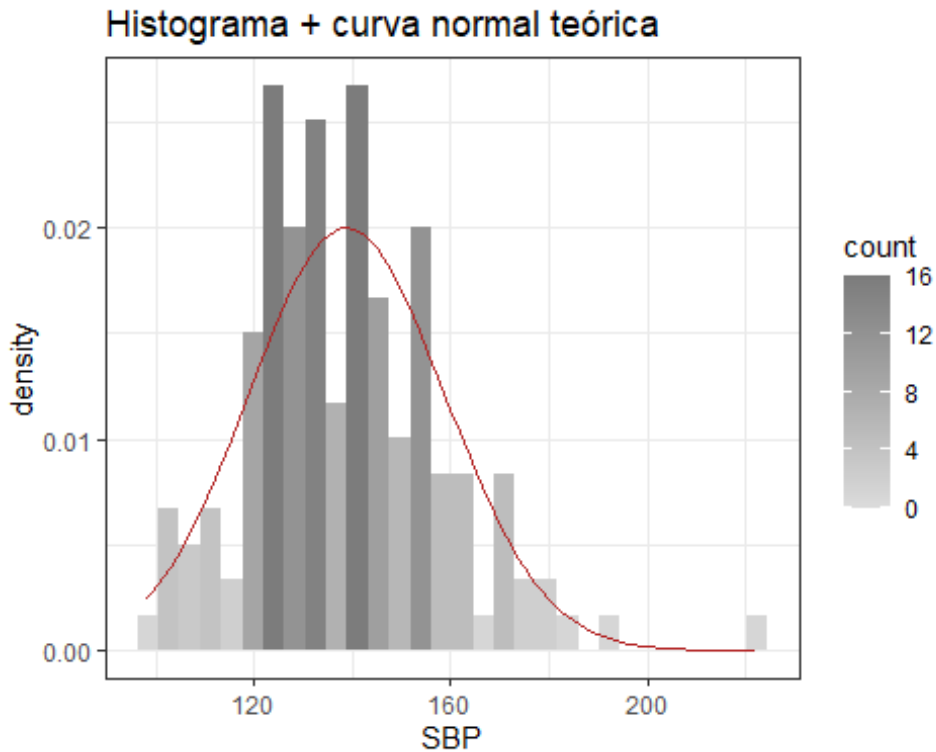
Al observar la normalidad puede suceder que, si los datos tienen una distribución normal multivariante, entonces, cada una de las variables tiene una distribución normal univariante, pero lo opuesto no tiene que ser verdad.

Esta función nos devuelve un resumen descriptivo de los datos. Observamos que en estas dos variables la mediana y la media son muy similares. También nos proporciona el test de normalidad de Shapiro-Wilk para cada variable. La hipótesis de este test es que los datos son normales. Por el valor de p value rechazamos la hipótesis, ninguna de las variables sigue una distribución normal.

Gráficamente, vemos que TEMPSVIU tiene una distribución sesgada a la izquierda, mientras que SBP tiene una distribución aproximadamente normal. Vamos a ver esta última con más detalle:

```
ggplot(data = dades140, aes(x = SBP)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(dades140$SBP),
                           sd = sd(dades140$SBP))) +
  ggtitle("Histograma + curva normal teórica") +
  theme_bw()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

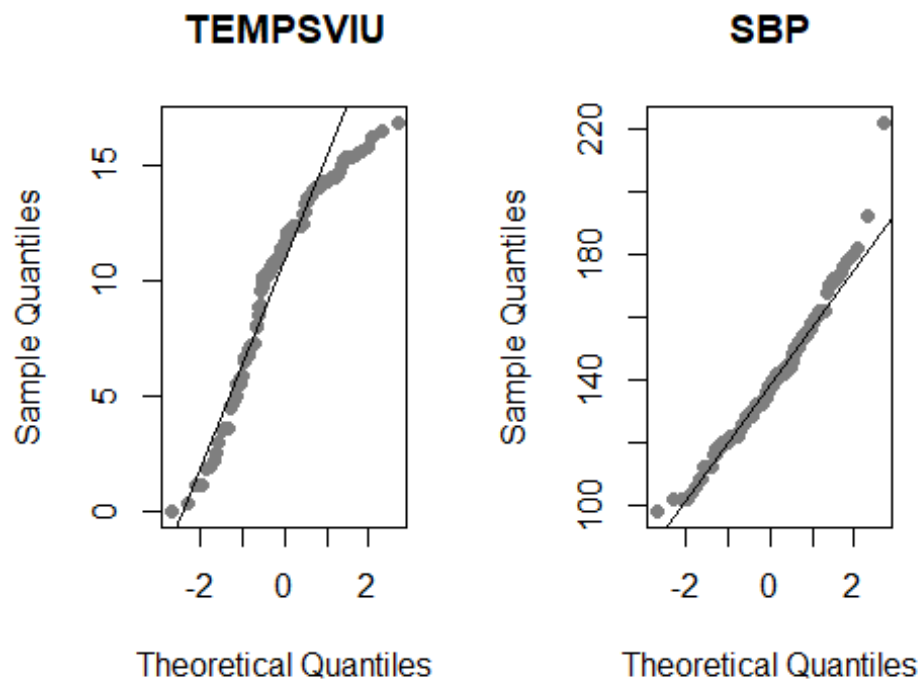


Los valores no se ajustan totalmente a la curva teórica.

QQ plot.

Cada eje se refiere a los cuantiles de distribuciones de probabilidad a comparar, donde uno de los ejes indica cuantiles teóricos (cuantiles hipotéticos) y el otro indica los cuantiles observados. Si los datos observados cumplen con la distribución hipotética (normalidad), los puntos en el gráfico Q-Q se encontrarán aproximadamente en la línea $y = x$.

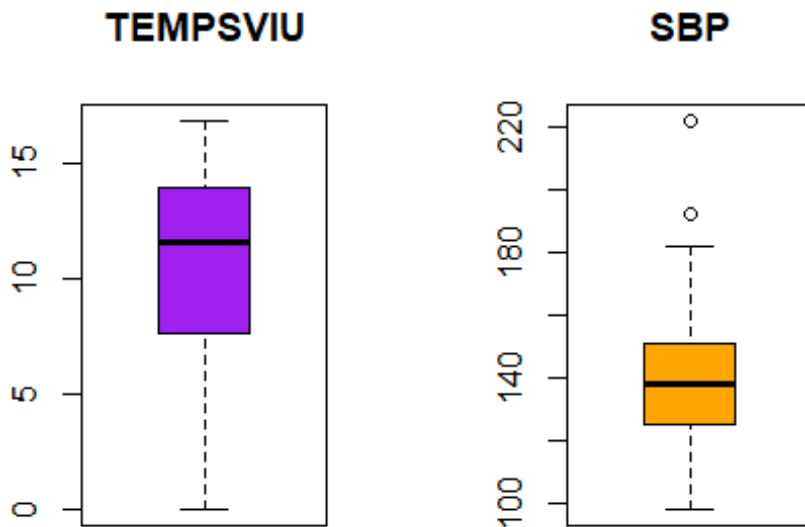
```
par(mfrow=c(1,2))
qqnorm(dades140$TEMPSVIU, pch=19, col = "gray50", main = "TEMPSVIU")
qqline(dades140$TEMPSVIU)
qqnorm(dades140$SBP, pch=19, col = "gray50", main = "SBP")
qqline(dades140$SBP)
```



Se puede ver en la gráfica que hay algunas desviaciones de la línea recta y esto indica posibles desviaciones de una distribución normal, especialmente en TEMPSVIU.

Boxplot.

```
par(mfrow=c(1,2))
boxplot(dades140$TEMPSVIU, col="purple", main="TEMPSVIU")
boxplot(dades140$SBP, col="orange", main="SBP")
```



Los diagramas de caja son bastante simétricos, sobretodo en SBP. Se podrían eliminar de los datos los valores extremos(outliers).

Tests de normalidad univariantes.

A continuación, se muestran los test de hipótesis más empleados para analizar la normalidad de los datos. En todos ellos, se considera como hipótesis nula que los datos sí proceden de una distribución normal y como hipótesis alternativa que no lo hacen. El p-value de estos test indica la probabilidad de obtener una distribución como la observada si los datos proceden realmente de una población con una distribución normal.

Test de Shapiro-Wilks. Este test se emplea para contrastar normalidad cuando el tamaño de la muestra es menor de 50 (no es nuestro caso). Para muestras grandes es equivalente al test de kolmogorov-Smirnov (Lilliefors).

```
mvn(data = dades140, univariateTest = "SW", desc = TRUE)

## $multivariateNormality
##           Test           Statistic           p value Result
## 1 Mardia Skewness 602.162399364036 1.14616024134213e-24    NO
## 2 Mardia Kurtosis 3.74514514658921 0.000180289511982235    NO
## 3           MVN           <NA>           <NA>           NO
##
## $univariateNormality
##           Test Variable Statistic   p value Normality
## 1 Shapiro-Wilk NUMPACIE    0.9531    1e-04         NO
```

```
## 2 Shapiro-Wilk MORT 0.4461 <0.001 NO
## 3 Shapiro-Wilk TEMPSVIU 0.9206 <0.001 NO
## 4 Shapiro-Wilk EDAT 0.9571 2e-04 NO
## 5 Shapiro-Wilk BMI 0.9610 5e-04 NO
## 6 Shapiro-Wilk EDATDIAG 0.9666 0.0017 NO
## 7 Shapiro-Wilk TABAC 0.7741 <0.001 NO
## 8 Shapiro-Wilk SBP 0.9701 0.0036 NO
## 9 Shapiro-Wilk DBP 0.9755 0.0127 NO
## 10 Shapiro-Wilk ECG 0.5795 <0.001 NO
## 11 Shapiro-Wilk CHD 0.5994 <0.001 NO
```

```
##
```

```
## $Descriptives
```

```
##          n      Mean   Std.Dev Median   Min    Max   25th   75th
## NUMPACIE 140  74.8071429 43.4300722   75.5   1.0  149.0  36.750 112.250
## MORT      140   0.1642857  0.3718651    0.0   0.0   1.0   0.000  0.000
## TEMPSVIU 140  10.6350000  4.0081917   11.6   0.0  16.9   7.825  13.925
## EDAT      140  51.9571429 11.4721254   50.0  31.0   86.0  43.750  59.000
## BMI       140  31.9514286  6.9328380   31.6  18.2   59.7  26.575  35.725
## EDATDIAG 140  45.7714286 10.7089305   45.0  26.0   81.0  38.000  52.000
## TABAC     140   0.9714286  0.8561488    1.0   0.0   2.0   0.000  2.000
## SBP       140 138.8285714 19.9336824  138.0  98.0  222.0 125.500 150.500
## DBP       140  80.0571429 11.1959470   80.0  58.0  108.0  74.000  88.000
## ECG       140   1.3214286  0.5909127    1.0   1.0   3.0   1.000  2.000
## CHD       140   0.3428571  0.4763686    0.0   0.0   1.0   0.000  1.000
```

```
##          Skew   Kurtosis
```

```
## NUMPACIE -0.016370791 -1.23774333
## MORT      1.792673306  1.22256287
## TEMPSVIU -0.829053015 -0.19870570
## EDAT      0.714341588  0.08271588
## BMI       0.833222143  1.39641893
## EDATDIAG  0.690806813  0.79755585
## TABAC     0.053909569 -1.64339111
## SBP       0.703609707  1.30314693
## DBP      -0.002091442 -0.46785696
## ECG       1.646033114  1.58622566
## CHD       0.655040709 -1.58206947
```

```
shapiro.test(dades140$TEMPSVIU)
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data:  dades140$TEMPSVIU
```

```
## W = 0.92059, p-value = 5.056e-07
```

```
shapiro.test(dades140$SBP)
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data:  dades140$SBP
## W = 0.97006, p-value = 0.003616
```

La hipótesis nula del test es que la población sigue una distribución normal. A un nivel de significación del 5%, rechazamos esta hipótesis.

Test de Kolmogorov-Smirnov y modificación de Lillefors. El test de Kolmogorov-Smirnov permite estudiar si una muestra procede de una población con una determinada distribución (media y desviación típica), no está limitado únicamente a la distribución normal. A pesar de que continuamente se alude al test Kolmogorov-Smirnov como un test válido para contrastar la normalidad, esto no es del todo cierto. El Kolmogorov-Smirnov asume que se conoce la media y varianza poblacional, lo que en la mayoría de los casos no es posible. Esto hace que el test sea muy conservador y poco potente. Para solventar este problema, se desarrolló una modificación del Kolmogorov-Smirnov conocida como test Lilliefors. El test Lilliefors asume que la media y varianza son desconocidas, estando especialmente desarrollado para contrastar la normalidad. Es la alternativa al test de Shapiro-Wilk cuando el número de observaciones es mayor de 50.

```
mvn(data = dades140, mvnTest = "royston", univariateTest = "Lillie", desc = TRUE)
```

```
## $multivariateNormality
##      Test      H      p value MVN
## 1 Royston 360.6621 4.176991e-71 NO
##
## $univariateNormality
##
##      Test      Variable Statistic      p value Normality
## 1 Lilliefors (Kolmogorov-Smirnov) NUMPACIE      0.0671 0.1272 YES
## 2 Lilliefors (Kolmogorov-Smirnov) MORT      0.5064 <0.001 NO
## 3 Lilliefors (Kolmogorov-Smirnov) TEMPSVIU      0.1496 <0.001 NO
## 4 Lilliefors (Kolmogorov-Smirnov) EDAT      0.1138 1e-04 NO
## 5 Lilliefors (Kolmogorov-Smirnov) BMI      0.0728 0.0661 YES
## 6 Lilliefors (Kolmogorov-Smirnov) EDATDIAG      0.0887 0.009 NO
## 7 Lilliefors (Kolmogorov-Smirnov) TABAC      0.2503 <0.001 NO
## 8 Lilliefors (Kolmogorov-Smirnov) SBP      0.0834 0.0185 NO
## 9 Lilliefors (Kolmogorov-Smirnov) DBP      0.0878 0.0102 NO
## 10 Lilliefors (Kolmogorov-Smirnov) ECG      0.4496 <0.001 NO
## 11 Lilliefors (Kolmogorov-Smirnov) CHD      0.4213 <0.001 NO
##
## $Descriptives
##      n      Mean      Std.Dev      Median      Min      Max      25th      75th
## NUMPACIE 140 74.8071429 43.4300722 75.5 1.0 149.0 36.750 112.250
## MORT 140 0.1642857 0.3718651 0.0 0.0 1.0 0.000 0.000
## TEMPSVIU 140 10.6350000 4.0081917 11.6 0.0 16.9 7.825 13.925
## EDAT 140 51.9571429 11.4721254 50.0 31.0 86.0 43.750 59.000
## BMI 140 31.9514286 6.9328380 31.6 18.2 59.7 26.575 35.725
```

```
## EDATDIAG 140 45.7714286 10.7089305 45.0 26.0 81.0 38.000 52.000
## TABAC 140 0.9714286 0.8561488 1.0 0.0 2.0 0.000 2.000
## SBP 140 138.8285714 19.9336824 138.0 98.0 222.0 125.500 150.500
## DBP 140 80.0571429 11.1959470 80.0 58.0 108.0 74.000 88.000
## ECG 140 1.3214286 0.5909127 1.0 1.0 3.0 1.000 2.000
## CHD 140 0.3428571 0.4763686 0.0 0.0 1.0 0.000 1.000
## Skew Kurtosis
## NUMPACIE -0.016370791 -1.23774333
## MORT 1.792673306 1.22256287
## TEMPSVIU -0.829053015 -0.19870570
## EDAT 0.714341588 0.08271588
## BMI 0.833222143 1.39641893
## EDATDIAG 0.690806813 0.79755585
## TABAC 0.053909569 -1.64339111
## SBP 0.703609707 1.30314693
## DBP -0.002091442 -0.46785696
## ECG 1.646033114 1.58622566
## CHD 0.655040709 -1.58206947
```

```
lillie.test(x = dades140$TEMPSVIU)
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dades140$TEMPSVIU
## D = 0.14965, p-value = 3.772e-08
```

```
lillie.test(x= dades140$SBP)
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: dades140$SBP
## D = 0.083365, p-value = 0.01854
```

Descartamos que estas variables sean normales ($\alpha=0.05$).

El hecho de no poder asumir la normalidad influye principalmente en los test de hipótesis paramétricos (t-test, anova,...) y en los modelos de regresión. Las principales consecuencias de la no normalidad son: - Los estimadores mínimo-cuadráticos no son eficientes (de mínima varianza). - Los intervalos de confianza de los parámetros del modelo y los contrastes de significancia son solamente aproximados y no exactos.

El teorema del límite central requiere que la población o poblaciones de las que proceden las muestras tengan una distribución normal, no que la tengan las muestras. Si las muestras se distribuyen de forma normal, se puede aceptar que así lo hacen las poblaciones de origen. En el caso de que las muestras no se distribuyan de forma normal pero se tenga certeza de que las poblaciones de origen sí lo hacen, entonces, los resultados obtenidos por los contrastes paramétricos sí son válidos. El teorema del

límite central, permite reducir los requerimientos de normalidad cuando las muestras suficientemente grandes.

b) ¿Existen diferencias entre la presión sistólica y la existencia de antecedentes coronarios? ¿Y entre el tiempo de seguimiento y los antecedentes coronarios(CHD)? . Comprueba la existencia de estas relaciones. Muestra e interpreta los resultados.

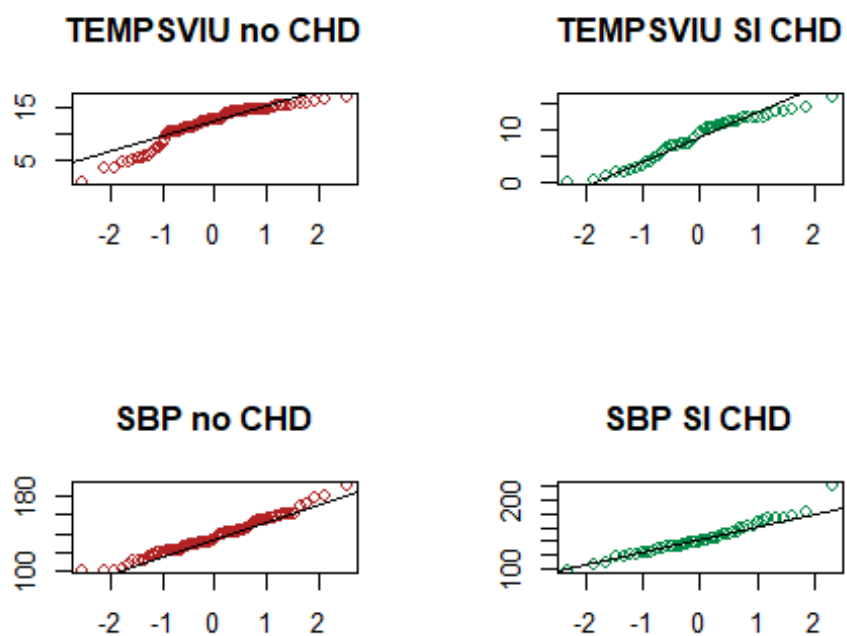
Presión sistólica y tiempo de seguimiento son variables cuantitativas continuas, en cambio, la variable antecedentes coronarios sólo toma los valores 0 y 1 (dicotómica). Contrastaremos si existen diferencias en la presión sistólica en caso de que haya o no antecedentes, y si existen diferencia en la media de tiempo de seguimiento según haya o no antecedentes.

Las condiciones para calcular intervalos de confianza o aplicar un test de hipótesis basados en la distribución T-student son las mismas que para el teorema del límite central:

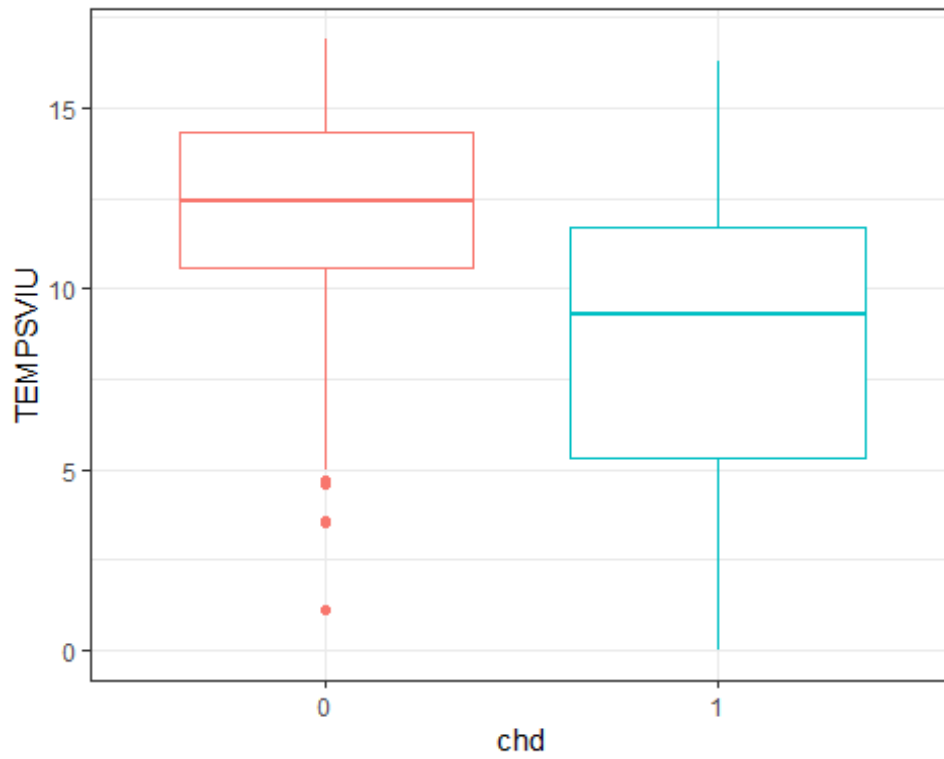
Independencia: Para ello, el muestreo debe ser aleatorio y el tamaño de la muestra inferior al 10% de la población. Las observaciones no son pareadas. Se puede afirmar que son independientes.

Normalidad:

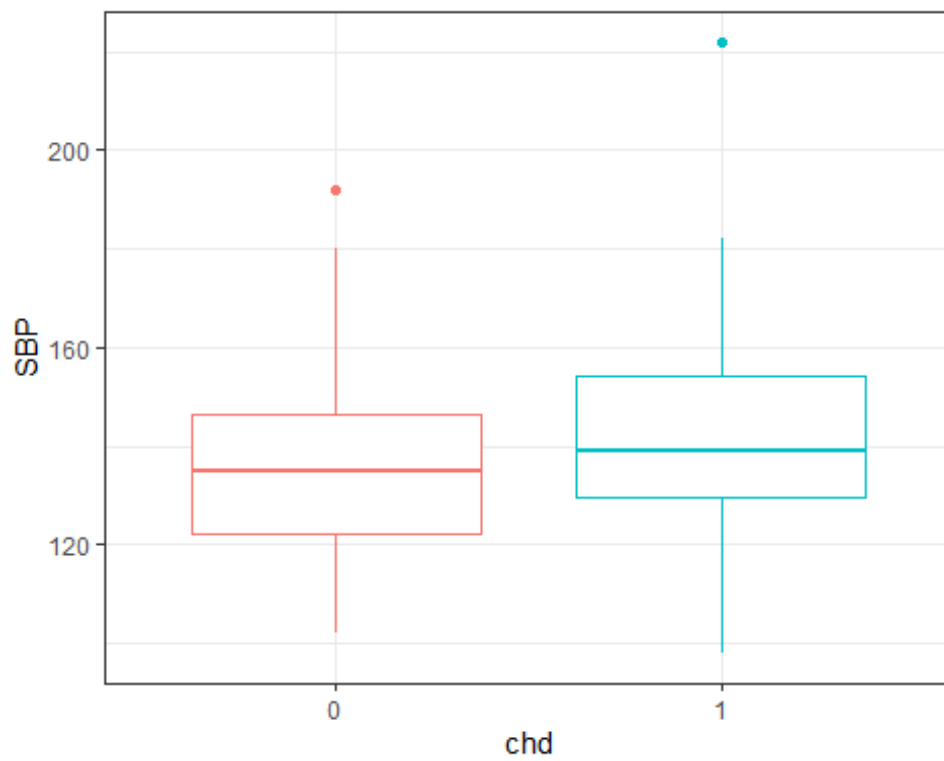
```
par(mfrow = c(2, 2))
qqnorm(dades140$TEMPSVIU[dades140$CHD == "0"], xlab = "", ylab = "",
      main = "TEMPSVIU no CHD", col = "firebrick")
qqline(dades140$TEMPSVIU[dades140$CHD == "0"])
qqnorm(dades140$TEMPSVIU[dades140$CHD == "1"], xlab = "", ylab = "",
      main = "TEMPSVIU SI CHD", col = "springgreen4")
qqline(dades140$TEMPSVIU[dades140$CHD == "1"])
qqnorm(dades140$SBP[dades140$CHD == "0"], xlab = "", ylab = "",
      main = "SBP no CHD", col = "firebrick")
qqline(dades140$SBP[dades140$CHD == "0"])
qqnorm(dades140$SBP[dades140$CHD == "1"], xlab = "", ylab = "",
      main = "SBP SI CHD", col = "springgreen4")
qqline(dades140$SBP[dades140$CHD == "1"])
```

```
chd<-as.factor(dades140$CHD)
ggplot(data = dades140) +
  geom_boxplot(aes(x = chd, y = TEMPSVIU, colour = chd)) +
  theme_bw() + theme(legend.position = "none")
```



```
ggplot(data = dades140) +  
  geom_boxplot(aes(x = chd, y = SBP, colour = chd)) +  
  theme_bw() + theme(legend.position = "none")
```



```

lillie.test(dades140$SBP[dades140$CHD == "0"])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dades140$SBP[dades140$CHD == "0"]
## D = 0.070785, p-value = 0.3081

lillie.test(dades140$SBP[dades140$CHD == "1"])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dades140$SBP[dades140$CHD == "1"]
## D = 0.10455, p-value = 0.2106

lillie.test(dades140$TEMPSVIU[dades140$CHD == "0"])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dades140$TEMPSVIU[dades140$CHD == "0"]
## D = 0.13296, p-value = 0.0003805

lillie.test(dades140$TEMPSVIU[dades140$CHD == "1"])

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dades140$TEMPSVIU[dades140$CHD == "1"]
## D = 0.13075, p-value = 0.03906

```

Existen evidencias significativas de que los datos no proceden de poblaciones con distribución normal. Sin embargo, existe simetría, lo cual indica que la mediana es una buena medida descriptiva. Además, el tamaño grande nos permite comparar medias, se puede considerar que el t-test sigue siendo suficientemente robusto. También podríamos utilizar un test no paramétrico basado en la mediana (Mann-Whitney-Wilcoxon test) o un test de Bootstrapping. Otra opción sería estudiar si los datos anómalos son excepciones que se pueden excluir del análisis.

Test de igualdad de varianzas:

```

fligner.test(TEMPSVIU~chd, data = dades140)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  TEMPSVIU by chd
## Fligner-Killeen:med chi-squared = 7.6743, df = 1, p-value = 0.005601

leveneTest(TEMPSVIU~chd, data = dades140, center = "median")

```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value Pr(>F)
## group  1  6.0713 0.01497 *
##      138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fligner.test(SBP~chd, data = dades140)

##
## Fligner-Killeen test of homogeneity of variances
##
## data: SBP by chd
## Fligner-Killeen:med chi-squared = 0.81933, df = 1, p-value = 0.3654

leveneTest(SBP~chd, data = dades140, center = "median")

## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value Pr(>F)
## group  1  1.3582 0.2459
##      138
```

En el caso de la presión sistólica, ninguno de los dos test encuentra evidencias significativas (para $\alpha = 0.05$) de que las variancias sean distintas entre ambas poblaciones. En cambio las varianzas no son iguales en el caso del tiempo de estudio. Cuando no se cumple esta condición de homocedasticidad, se puede emplear un Welch Two Sample t-test, que incorpora una corrección a través de los grados de libertad que compensa la diferencia de varianzas, con el inconveniente de que pierde poder estadístico.

T-test:

Contrastamos si las medias de SBP son iguales para ambos grupos:

```
t.test(dades140$SBP[dades140$CHD==0], dades140$SBP[dades140$CHD==1], alternative = "two.sided", mu = 0, var.equal = TRUE, conf.level = 0.95)

##
## Two Sample t-test
##
## data: dades140$SBP[dades140$CHD == 0] and dades140$SBP[dades140$CHD =
= 1]
## t = -1.7293, df = 138, p-value = 0.08599
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -13.0624403 0.8740345
## sample estimates:
## mean of x mean of y
## 136.7391 142.8333
```

No podemos rechazar la hipótesis nula de que ambas medias son iguales, por el momento no podemos decir que existan diferencias entre los dos grupos de antecedentes coronarios (nivel de significación 5%).

Y si el tiempo de seguimiento es el mismo (de media) en caso que haya o no antecedentes:

```
t.test(dades140$TEMPSVIU[dades140$CHD==0], dades140$TEMPSVIU[dades140$CHD==1], mu = 0, var.equal = F, conf.level = 0.95)

##
## Welch Two Sample t-test
##
## data: dades140$TEMPSVIU[dades140$CHD == 0] and dades140$TEMPSVIU[dades140$CHD == 1]
## t = 4.9344, df = 79.302, p-value = 4.351e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.059500 4.844123
## sample estimates:
## mean of x mean of y
## 11.818478  8.366667
```

Aquí si que rechazamos la hipótesis nula, TEMPSVIU no es el mismo si se tienen o no antecedentes coronarios.

Test de Wilcoxon-Mann-Whitney como alternativa al t-test.

- Los datos tienen que ser independientes.
- Los datos tienen que ser ordinales o bien se tienen que poder ordenar de menor a mayor.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales. Pero, para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría...).
- Igualdad de varianza entre grupos (homocedasticidad).

```
wilcox.test(dades140$TEMPSVIU[dades140$CHD==0], dades140$TEMPSVIU[dades140$CHD==1], alternative = "two.sided", mu = 0, paired = FALSE, conf.int = 0.95)

##
## Wilcoxon rank sum test with continuity correction
##
## data: dades140$TEMPSVIU[dades140$CHD == 0] and dades140$TEMPSVIU[dades140$CHD == 1]
## W = 3325.5, p-value = 9.321e-07
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
```

```
## 2.000062 4.499951
## sample estimates:
## difference in location
## 3.299951

wilcox.test(dades140$SBP[dades140$CHD==0],dades140$SBP[dades140$CHD==1],a
lternative = "two.sided", mu = 0, paired = F, conf.level = 0.95)

##
## Wilcoxon rank sum test with continuity correction
##
## data: dades140$SBP[dades140$CHD == 0] and dades140$SBP[dades140$CHD =
= 1]
## W = 1881, p-value = 0.1514
## alternative hypothesis: true location shift is not equal to 0
```

Cuando hay ligaduras o ties, la función `wilcox.test()` no es capaz de calcular el p-value exacto por lo que devuelve una aproximación asumiendo que U (W) se distribuye de forma \sim normal. En estos casos, o cuando los tamaños muestrales son mayores de 20 y se quiere la aproximación por la normal, es más recomendable emplear la función `wilcox_test()` del paquete `coin`.

```
wilcox_test(TEMPSVIU~chd, data = dades140, distribution = "exact", conf.i
nt=0.95)

##
## Exact Wilcoxon-Mann-Whitney Test
##
## data: TEMPSVIU by chd (0, 1)
## Z = 4.9077, p-value = 4.407e-07
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
## 2.0 4.5
## sample estimates:
## difference in location
## 3.3

wilcox_test(SBP~chd, data = dades140, distribution = "exact", conf.int=0.
95)

##
## Exact Wilcoxon-Mann-Whitney Test
##
## data: SBP by chd (0, 1)
## Z = -1.437, p-value = 0.1515
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
## -12 2
## sample estimates:
## difference in location
## -6
```

La función devuelve el valor de Z, por lo que se puede calcular fácilmente el tamaño efecto

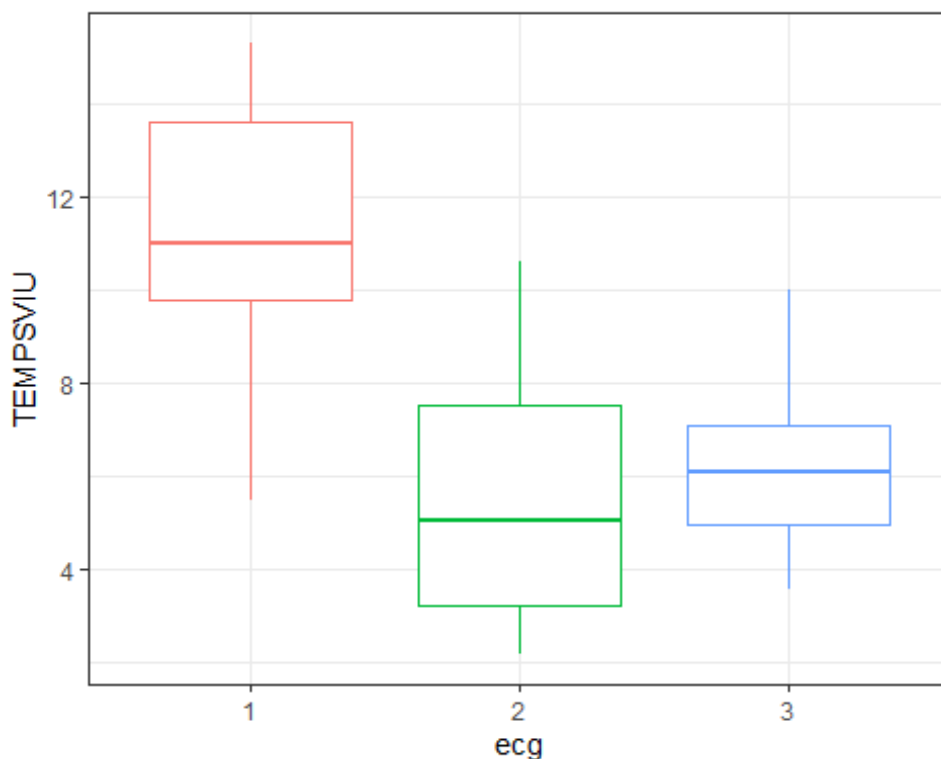
```
tamano_efectoTEMPSVIU <- 4.91/140  
tamano_efectoSBP <- -1.437/140
```

La diferencia entre las probabilidades de que observaciones de los individuos con antecedentes coronarios difieran a las de los individuos sin antecedentes no es significativa (p-value = 0.1515) en el caso de la presión sistólica. En cambio, sí lo es en el caso del tiempo de seguimiento (p-value=4.407e-07).

c) Se está interesado en observar la relación entre el tiempo de seguimiento (TEMPSVIU) hasta la muerte (MORT) y la alteración del electrocardiograma (ECG). Indica la estrategia de análisis para poder estudiar esta relación. Muestra e interpreta los resultados.

Suponemos que el estatus MORT=1 equivale a haber fallecido. Entendemos que, de momento, sólo nos interesan las observaciones de aquellos individuos que han fallecido, así que extraemos los datos correspondientes a éstos.

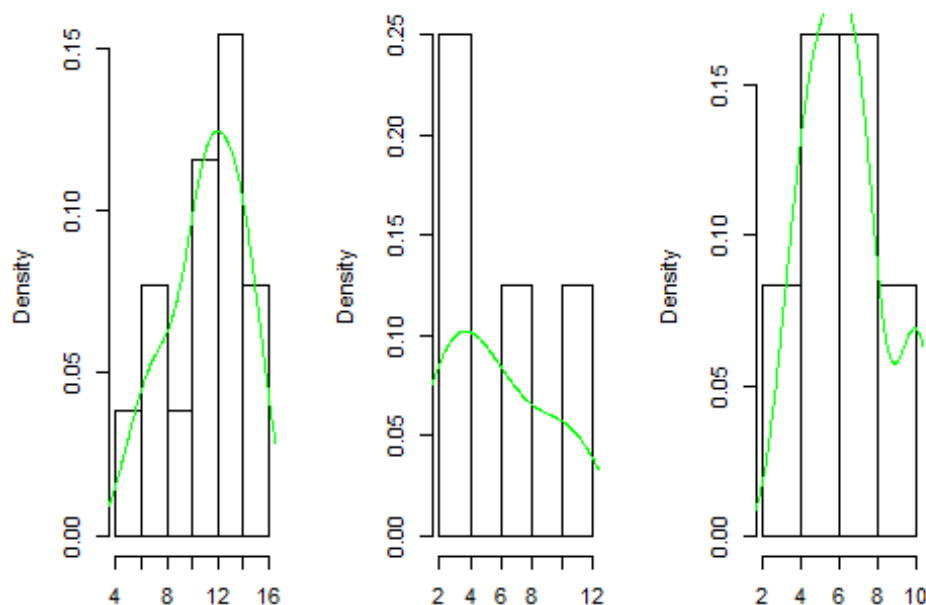
```
dadesRIP <- subset(dades140, dades140$MORT==1)  
dadesRIP$ecg <- as.factor(dadesRIP$ECG)  
ggplot(data = dadesRIP, mapping = aes(x = ecg, y = TEMPSVIU, colour = ecg)) +  
  geom_boxplot() +  
  theme_bw() +  
  theme(legend.position = "none")
```



Las cajas 2 y 3 son simétricas.

```
par(mfrow=c(1,3))
hist(dadesRIP$TEMPSVIU[dadesRIP$ecg==1], freq = F)
lines(density(dadesRIP$TEMPSVIU[dadesRIP$ecg==1]), col="green")
hist(dadesRIP$TEMPSVIU[dadesRIP$ecg==2], freq = F)
lines(density(dadesRIP$TEMPSVIU[dadesRIP$ecg==2]), col="green")
hist(dadesRIP$TEMPSVIU[dadesRIP$ecg==3], freq = F)
lines(density(dadesRIP$TEMPSVIU[dadesRIP$ecg==3]), col="green")
```

dadesRIP\$TEMPSVIU[dadesRIP\$ecg==1] dadesRIP\$TEMPSVIU[dadesRIP\$ecg==2] dadesRIP\$TEMPSVIU[dadesRIP\$ecg==3]



dadesRIP\$TEMPSVIU[dadesRIP\$ecg==1] dadesRIP\$TEMPSVIU[dadesRIP\$ecg==2] dadesRIP\$TEMPSVIU[dadesRIP\$ecg==3]

```
shapiro.test(dadesRIP$TEMPSVIU[dadesRIP$ecg==1])
```

```
##
## Shapiro-Wilk normality test
##
## data:  dadesRIP$TEMPSVIU[dadesRIP$ecg == 1]
## W = 0.95762, p-value = 0.7166
```

```
shapiro.test(dadesRIP$TEMPSVIU[dadesRIP$ecg==2])
```

```
##
## Shapiro-Wilk normality test
##
## data:  dadesRIP$TEMPSVIU[dadesRIP$ecg == 2]
## W = 0.94728, p-value = 0.6992
```

```
shapiro.test(dadesRIP$TEMPSVIU[dadesRIP$ecg==3])
```



```
##
## Shapiro-Wilk normality test
##
## data:  dadesRIP$TEMPSVIU[dadesRIP$ecg == 3]
## W = 0.96447, p-value = 0.8534
```

En la representación gráfica de los datos parece que los dos primeros grupos no se distribuyen de forma normal, pero el test de normalidad no encuentra evidencias de no normalidad.

Homocedasticidad:

```
leveneTest(TEMPSVIU ~ecg, data = dadesRIP, center = "median")
## Levene's Test for Homogeneity of Variance (center = "median")
##      Df F value Pr(>F)
## group 2  0.5913  0.563
##      20
```

No hay evidencias en contra de la homogeneidad de varianzas.

Podemos utilizar un test anova para contrastar la igualdad de medias de los tres grupos:

```
anova<- aov(TEMPSVIU~ecg, data=dadesRIP)
summary(anova)
##              Df Sum Sq Mean Sq F value Pr(>F)
## ecg           2  144.7    72.34    8.53 0.0021 **
## Residuals    20  169.6     8.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Existen diferencias significativas.

La opción no paramétrica es (aunque en este caso no sería necesaria): ### Test de Kruskal-Wallis:

```
kruskal.test(TEMPSVIU ~ecg, data = dadesRIP)
##
## Kruskal-Wallis rank sum test
##
## data:  TEMPSVIU by ecg
## Kruskal-Wallis chi-squared = 10.668, df = 2, p-value = 0.004824
```

El test encuentra significancia en la diferencia de al menos dos grupos.

Comparaciones post-hoc para saber que dos grupos difieren:

Existen diferentes métodos de corrección del nivel de significancia, entre ellos destacan el de Bonferroni (`p.adjust.method = bonferroni`), que es muy estricto, y el de Holm, este último parece ser más recomendado.

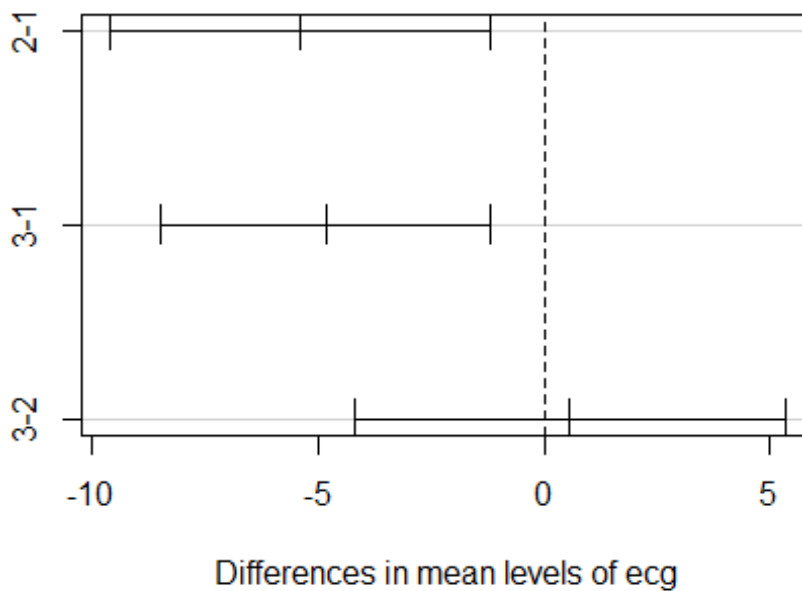
```
pairwise.wilcox.test(x = dadesRIP$TEMPSVIU, g = dadesRIP$ecg, p.adjust.me  
thod = "holm", pool.sd = TRUE, paired = FALSE, alternative = "two.sided")  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot c  
ompute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot c  
ompute  
## exact p-value with ties  
  
##  
## Pairwise comparisons using Wilcoxon rank sum test  
##  
## data: dadesRIP$TEMPSVIU and dadesRIP$ecg  
##  
## 1 2  
## 2 0.03 -  
## 3 0.02 0.67  
##  
## P value adjustment method: holm
```

Las comparaciones por pares encuentran diferencias significativas entre 1-2 y 1-3 ($\alpha=0.05$).

TukeyHSD(anova)

```
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##  
## Fit: aov(formula = TEMPSVIU ~ ecg, data = dadesRIP)  
##  
## $ecg  
## diff lwr upr p adj  
## 2-1 -5.390385 -9.602922 -1.177847 0.0110066  
## 3-1 -4.815385 -8.451607 -1.179162 0.0085505  
## 3-2 0.575000 -4.180710 5.330710 0.9498564  
  
plot(TukeyHSD(anova))
```

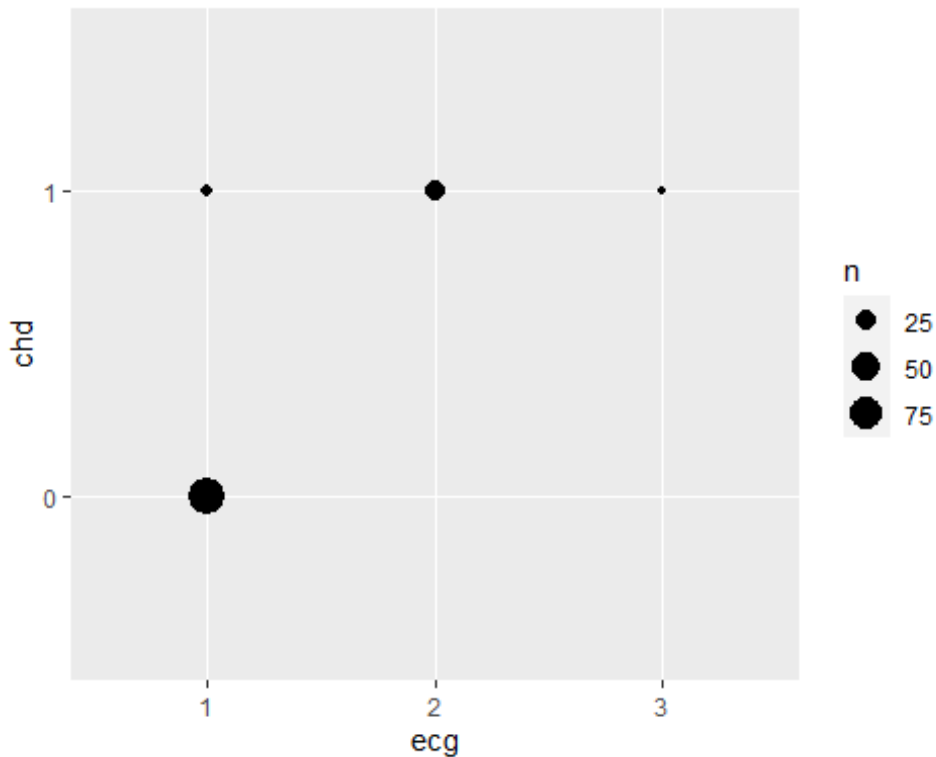
95% family-wise confidence level



d) Puede existir relación entre los antecedentes coronarios y el ECG. ¿ Como establecerías la relación. Muestra e interpreta los resultados.

Las dos variables son categóricas. Queremos estudiar si el hecho de tener antecedentes(0 y 1, no y si, respectivamente) y el resultado del ECG (1,2,3) estan relacionados.

```
chd<- as.character(dades140$CHD)
ecg<- as.character(dades140$ECG)
tabla <- table(dades140$CHD, dades140$ECG, dnn = c("antecedente coron", "ECG"))
ggplot(data = dades140) +
  geom_count(mapping = aes(x = ecg, y = chd))
```



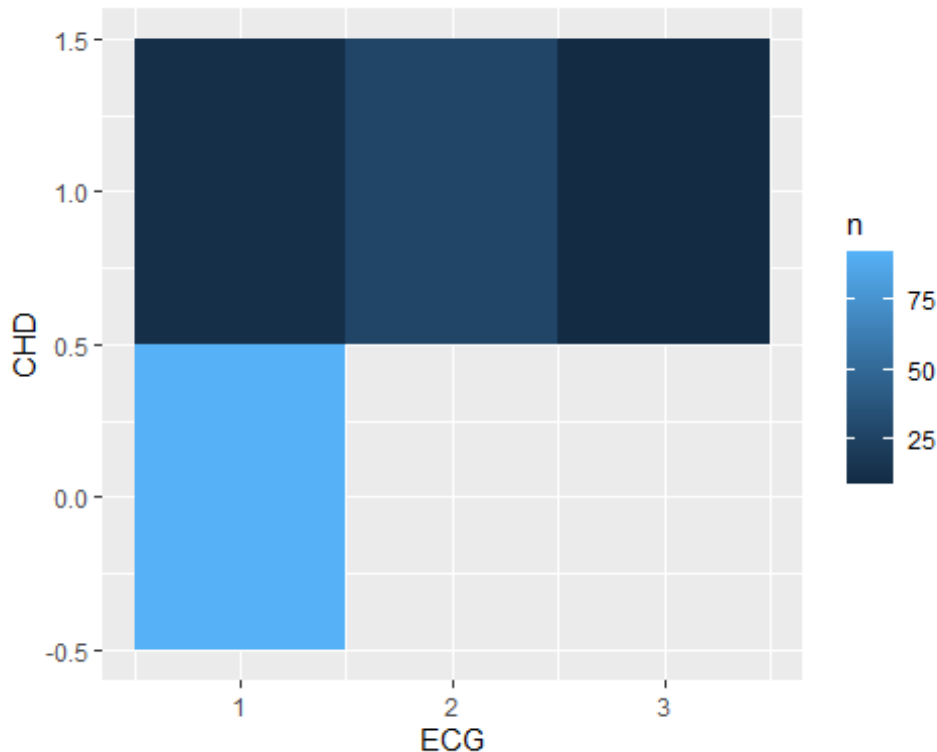
El tamaño de cada círculo en la gráfica muestra cuántas observaciones corresponden a cada combinación de valores.

```
dades140 %>%
  count(ECG, CHD)

## # A tibble: 4 x 3
##       ECG      CHD      n
##   <dbl> <dbl> <dbl>
## 1 1 [Normal]    0 [No]   92
## 2 1 [Normal]    1 [Si]   12
## 3 2 [Frontera]  1 [Si]   27
## 4 3 [Anormal]  1 [Si]    9

dades140 %>%
  count(ECG, CHD) %>%
  ggplot(mapping = aes(x = ECG, y = CHD)) +
  geom_tile(mapping = aes(fill = n))

## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to continuous.
```



Test de Fisher.

El test de Fisher trabaja con frecuencia de eventos, por lo tanto con tablas de contingencia en las que se resume el número de eventos de cada tipo.

tabla

```
##          ECG
## antecedente coron  1  2  3
##          0 92  0  0
##          1 12 27  9
```

La prueba de Fisher es el test exacto utilizado cuando se quiere estudiar si existe asociación entre dos variables cualitativas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable. Es posible utilizarlo con tablas 2xK niveles, como en este caso.

El test de Fisher es más preciso que sus equivalentes aproximados (test chi-square de independencia o G-test de independencia) cuando el número de eventos esperado por nivel es pequeño. Se recomienda utilizarlo siempre que sea posible (tiempo de computación) aunque para observaciones totales >1000 los resultados de los test aproximados son muy parecidos. Condiciones del test: - Independencia: las observaciones de la muestra deben ser independientes unas de otras. - Muestreo aleatorio. - Tamaño de la muestra < 10% población. - Cada observación contribuye únicamente a uno de los niveles. - Las frecuencias marginales de columnas y filas tienen que ser fijas. Si esta condición no se cumple, el test de Fisher deja de ser exacto.

Ho : Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable. Ha: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

```
fisher.test(x = tabla, alternative = "two.sided")  
  
##  
## Fisher's Exact Test for Count Data  
##  
## data:  tabla  
## p-value < 2.2e-16  
## alternative hypothesis: two.sided
```

Rechazamos la hipótesis, no existe relación.

Dado que el test de Fisher contrasta si las variables están relacionadas, al tamaño del efecto se le conoce como fuerza de asociación. Existen múltiples medidas de asociación, entre las que destacan phi o Cramer's V. Los límites empleados para su clasificación son: pequeño: 0.1 mediano: 0.3 grande: 0.5 En R se pueden calcular mediante la función `assocstats()` del paquete `vcd`.

```
assocstats(x = tabla)  
  
##              X^2 df P(> X^2)  
## Likelihood Ratio 105.629  2      0  
## Pearson          92.885  2      0  
##  
## Phi-Coefficient   : NA  
## Contingency Coeff.: 0.632  
## Cramer's V        : 0.815
```

El tamaño de la fuerza de asociación (tamaño de efecto) cuantificado por phi o Cramer's V es grande.

Si realizamos un Chi-test, la versión no exacta del Fisher-test, R nos da error porque los valores contienen ceros. También se puede solucionar mediante simulación de p-value:

```
chisq.test(tabla, simulate.p.value = TRUE, B = 5000)  
  
##  
## Pearson's Chi-squared test with simulated p-value (based on 5000  
## replicates)  
##  
## data:  tabla  
## X-squared = 92.885, df = NA, p-value = 2e-04
```

3. Algunas cosas más:

a) Indica para que te podrían servir las técnicas de bootstrap en los análisis estadísticos que has realizado.

El bootstrap es una herramienta para construir intervalos de confianza y calcular errores estándar para estadísticas difíciles. Se basa en investigar la distribución de la muestra de un estadístico simulando repeticiones de la distribución observada. Si la muestra original es representativa de la población, la distribución del estadístico calculada a partir de las pseudomuestras (bootstrapping distribution) se asemejará a la distribución muestral que se obtendría si se pudiera acceder a la población para generar nuevas muestras.

Aplicaciones:

- Nos proporciona una manera de construir intervalos cuando no existe normalidad.
- Sirve para calcular significación estadística (p-value) para la diferencia entre dos poblaciones: contrasta la hipótesis de que ambas muestras proceden de la misma población (distribución), pero lo hace mediante el estudio de las diferencias debidas al muestreo aleatorio.
- Calcula intervalos de confianza para la diferencia entre dos poblaciones: para esta finalidad, se considera como hipótesis nula que las observaciones proceden de dos poblaciones distintas. Se emplea el sampling with replacement con la observaciones de cada muestra (sin mezclarlas) para generar dos nuevas pseudomuestras independientes y se calcula la diferencia del estadístico.

El método de bootstrapping se subdivide en paramétrico o no paramétrico dependiendo si se considera que la distribución generada sigue un determinado modelo teórico.

El bootstrapping no asume una asignación aleatoria de los grupos, sino que las muestras han sido obtenidas aleatoriamente de la o las poblaciones. Se aplica por lo tanto en diseños muestrales, no experimentales. Esta es la diferencia clave respecto a los test de permutación.

Ejemplo aplicado: Calcula el intervalo utilizando Bootstrap de la mediana del tiempo de seguimiento.

```
library(rsample)
library(ggpubr)

## Loading required package: magrittr

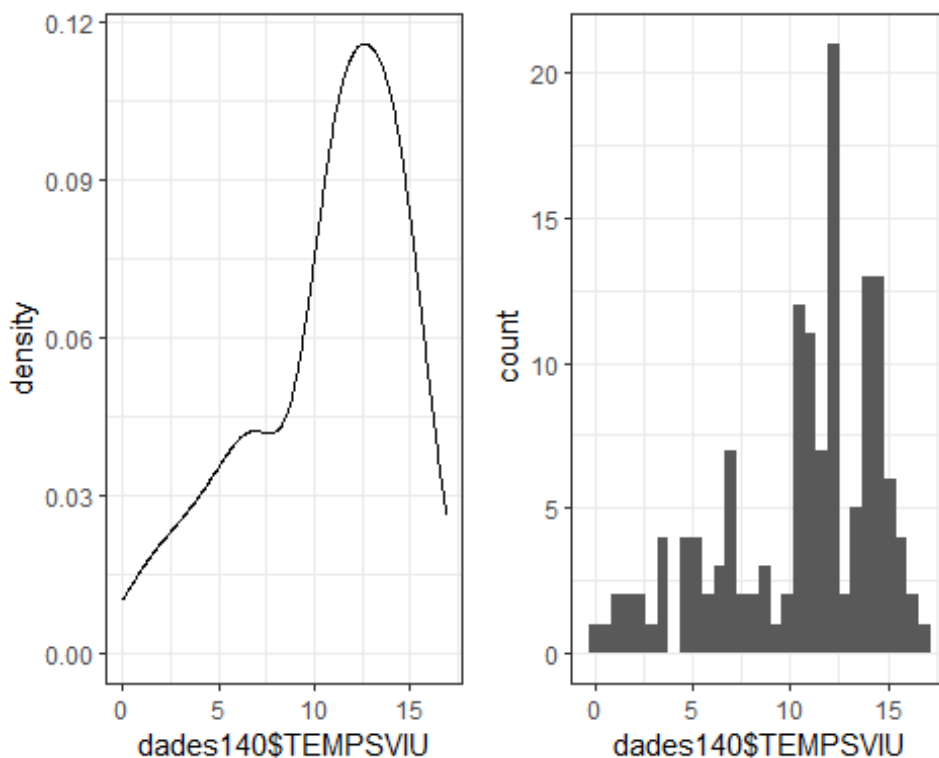
p1 <- ggplot(data = data_frame(dades140$TEMPSVIU), aes(x = dades140$TEMPS
VIU)) +
  geom_density() +
  theme_bw()
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

p2 <- ggplot(data = data_frame(dades140$TEMPSVIU), aes(x = dades140$TEMPSVIU)) +
  geom_histogram() +
  theme_bw()

ggarrange(p1, p2)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Los datos de TEMPSVIU no se distribuyen de forma normal. Esto implica que la aproximación basada en el teorema del límite central para estimar el error estándar $SE = \frac{sd}{\sqrt{n}}$ deja de ser buena y, por lo tanto, también los intervalos paramétricos basados en la estructura $[\text{parámetro estimado} \pm t_{\alpha} SE]$ que se generen con esa estimación del SE. Una alternativa para poder calcular intervalos de confianza es emplear Bootstrapping.

Si la muestra original es representativa de la población, la distribución del estadístico calculada a partir de las pseudomuestras (bootstrapping distribution) se asemejará a la distribución muestral que se obtendría si se pudiera acceder a la población para generar nuevas muestras. De tal forma que:

- La sd de la bootstrapping distribution es un estimador del SE.
- La media de la bootstrapping distribution es un estimador del verdadero parámetro poblacional.

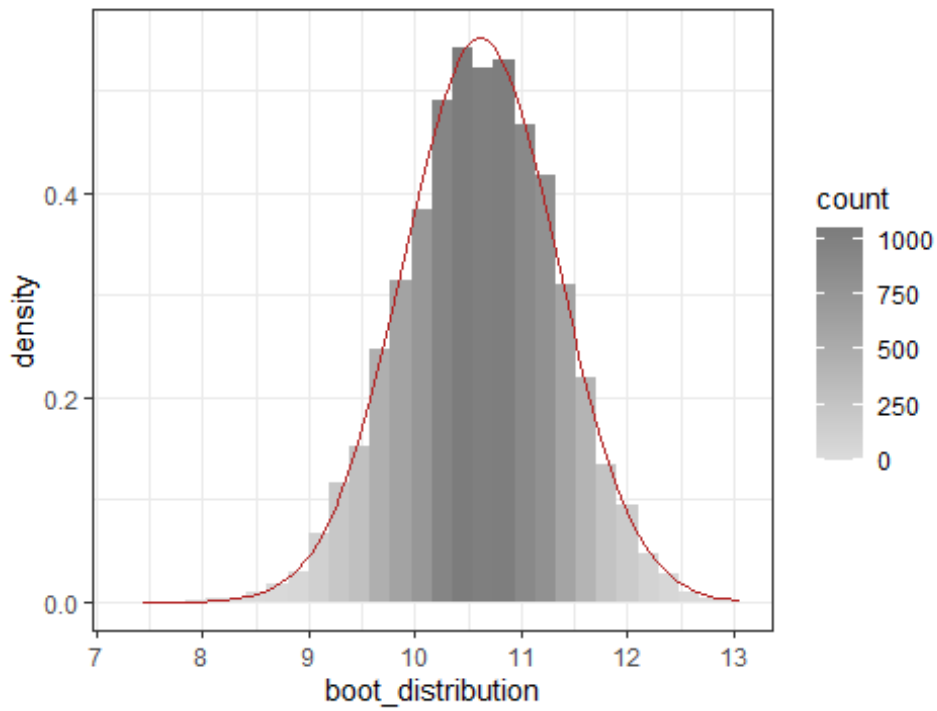
Así pues, a partir de la bootstrapping distribution se pueden obtener los valores necesarios para crear el intervalo de confianza sin recurrir al teorema del límite central.

```
set.seed(2022)

#Generar la bootstrapping distribution
boot_distribution <- rep(NA, 9999)
for (i in 1:9999) {
  boot_distribution[i] <- mean(sample(x = dades140$TEMPSVIU, size = 30,
replace = TRUE))
}
#Estudio la bootstrapping distribution
ggplot(data = data.frame(boot_distribution), aes(x = boot_distribution))
+
  geom_histogram(aes(y = ..density.. , fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(boot_distribution),
                           sd = sd(boot_distribution))) +
  ggtitle("Histograma + curva normal teórica") +
  theme_bw()

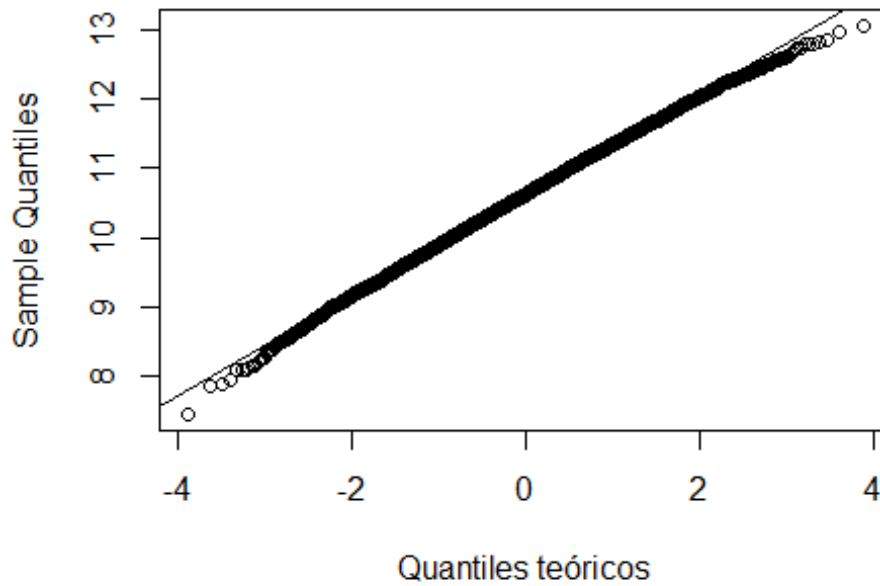
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Histograma + curva normal teórica



```
qqnorm(boot_distribution, xlab = "Quantiles teóricos")  
qqline(boot_distribution)
```

Normal Q-Q Plot



Dependiendo de la distribución que tengan los estadísticos obtenidos por bootstrapping, se emplea un método distinto para estimar los límites del intervalo de confianza. Cuando la distribución obtenida no es de tipo normal, no se pueden emplear intervalos de confianza basados en la t-distribution ni en percentiles. En este caso sí se distribuyen de forma normal. Centro: La media de la bootstrapping distribution debe de ser cercana a la media de la muestra inicial a partir de la cual se está generando el bootstrapping. A esta diferencia se le llama bias.

```
mean(dades140$TEMPSVIU) - mean(boot_distribution)
```

```
## [1] 0.02160599
```

En este caso el bias es muy pequeño comparado al valor de la media calculada. Existen varios métodos para generar intervalos de confianza a partir de una bootstrapping distribution. Los intervalos de tipo t y los basados en percentiles emplean la desviación estándar de la bootstrapping distribution como estimación del SE. Son únicamente válidos cuando la bootstrapping distribution se asemeja a una normal y el bias es pequeño.

```
# Intervalo 95% basado en percentiles:
```

```
quantile(x = boot_distribution, probs = c(0.025, 0.975))
```

```
##      2.5%      97.5%
```

```
## 9.183333 11.996833
```

```
library(boot)
```

```
##
```

```
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:survival':
```

```
##
```

```
##      aml
```

```
## The following object is masked from 'package:car':
```

```
##
```

```
##      logit
```

```
set.seed(1010)
```

```
mediana <- function(valores, i) {
```

```
  median(valores[i])
```

```
}
```

```
boot_distribution <- boot(dades140$TEMPSVIU, mediana, R = 99)
```

```
boot_distribution
```

```
##
```

```
## ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
##
```

```
##
```

```
## Call:
```

```
## boot(data = dades140$TEMPSVIU, statistic = mediana, R = 99)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1*      11.6 0.01868687    0.3818982

media_truncada <- function(valores, i) {
  mean(valores[i], trim = 0.25)
}
boot_distribution <- boot(dades140$TEMPSVIU, media_truncada, R = 99)
boot_distribution

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = dades140$TEMPSVIU, statistic = media_truncada, R = 99)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 11.51857 0.03477633    0.3231446
```

b) Indica porqué hay que utilizar técnicas de comparaciones múltiple como la corrección de Bonferroni.

Las comparaciones múltiples surgen cuando un estudio estadístico conlleva la realización de varias comparaciones con el objetivo de identificar aquellos grupos para los que las diferencias son más significativas. Al contrastar m hipótesis nulas,

- la estimación del false positive rate, probabilidad de rechazar la hipótesis nula (test significativo) siendo esta cierta, se obtiene como:
número de falsos positivos (error tipo I)/número de hipótesis nulas verdaderas
- La estimación del false discovery rate, probabilidad de que la hipótesis nula sea cierta a pesar de haber sido rechazada (test significativo), se obtiene como:
número de falsos positivos (error tipo I)/número de hipótesis nulas rechazadas (test significativo), independientemente de que sean ciertas o falsas.

El problema de las comparaciones múltiples se debe al incremento en el error de tipo I como consecuencia del uso repetido de test estadísticos. Si se realizan k comparaciones independientes, la probabilidad de que al menos ocurra un falso positivo, también conocida como family-wise error rate (FWER) viene dada por la ecuación:

$$FWER = 1 - (1 - \alpha)^k$$

Bonferroni es posiblemente el ajuste de significancia más extendido a pesar de que no está recomendado para la mayoría de las situaciones que se dan en el ámbito de la biomedicina. La corrección de Bonferroni consiste en dividir el nivel de significancia α entre el número de comparaciones dos a dos realizadas.

$\alpha_{\text{corregido}} = \alpha / \text{número de grupos}$

Con esta corrección se asegura que la probabilidad de obtener al menos un falso positivo entre todas las comparaciones (family-wise error rate) es $\leq \alpha$. Permite por lo tanto contrastar una hipótesis nula general (la de que todas las hipótesis nulas testadas son verdaderas) de forma simultánea, cosa que raramente es de interés en las investigaciones. Se considera un método excesivamente conservativo sobre todo a medida que se incrementa el número de comparaciones. En R se puede calcular mediante la función `pairwise.t.test()` indicando en los argumentos `p.adj = "bonferroni"`. Ejemplo:

```
pairwise.wilcox.test(x = dadesRIP$TEMPSVIU, g = dadesRIP$ecg, p.adjust.me
thod = "bonferroni", pool.sd = TRUE, paired = FALSE, alternative = "two.si
ded")

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot c
ompute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot c
ompute
## exact p-value with ties

##
## Pairwise comparisons using Wilcoxon rank sum test
##
## data: dadesRIP$TEMPSVIU and dadesRIP$ecg
##
##      1      2
## 2 0.045 -
## 3 0.020 1.000
##
## P value adjustment method: bonferroni
```