Práctica 6-7: Test de Hipótesis

1. Test de Hipótesis. Parte I

1.1.

Se toma una muestra aleatoria de tamaño $n = 25$ de una distribución $N(\mu, \sigma)$. Si $\sum_{i=1}^{n} x_i = 100$ y $\sum_{i=1}^{n} x_i^2 = 600$,

1. Contrasta la hipótesis nula $H_0 : \mu = 2.5$ contra la hipótesis alternativa $H_1 : \mu \neq 2.5$ con un nivel de significación $\alpha = 0.05$

1.1. Solución

$X \sim N(\mu, \sigma)$

- $H_0 : \mu = 2.5$ vs. $H_1 : \mu \neq 2.5$
- Condiciones de aplicación: Población Normal, varianza desconocida

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- Cálculo:

```
library(PASWR2)
media <- 100/25
s <- sqrt((600 - 25*media^2)/ (25-1))
tsum.test(mean.x=media, s.x =s, n.x = 25,
alternative = "two.sided", mu = 2.5, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  User input summarized values for x
## t = 2.5981, df = 24, p-value = 0.01577
## alternative hypothesis: true difference in means is not equal to 2.5
## 95 percent confidence interval:
##  2.808408 5.191592
## sample estimates:
## mean of x
##         4
```

- Decisión: Rechazar H_0 porque p-valor=0.01577 $< \alpha$
- Conclusión: Con un nivel de significación del 5% tenemos evidencias que indican que la media de la población es distinta de 2.5.

2. Calcula la Potencia($\mu_1 = 4$) si asumimos que $\sigma = 2.5$.

```
power.t.test(n = 25, delta = 1.5, sd = 2.5, type = "one.sample",
             alternative="two.sided", strict = TRUE)
```

```
##
##      One-sample t test power calculation
##
##               n = 25
##           delta = 1.5
##              sd = 2.5
##       sig.level = 0.05
##           power = 0.8207219
##     alternative = two.sided
```

1.2

Un agricultor quiere comprobar si una nueva marca de fertilizante incrementa la producción de trigo por parcela. Los siguientes datos corresponden a las producciones en fanegas por parcela de 15 parcelas tratadas con el nuevo fertilizante.

2.5 3.0 3.1 4.0 1.2 5.0 4.1 3.9

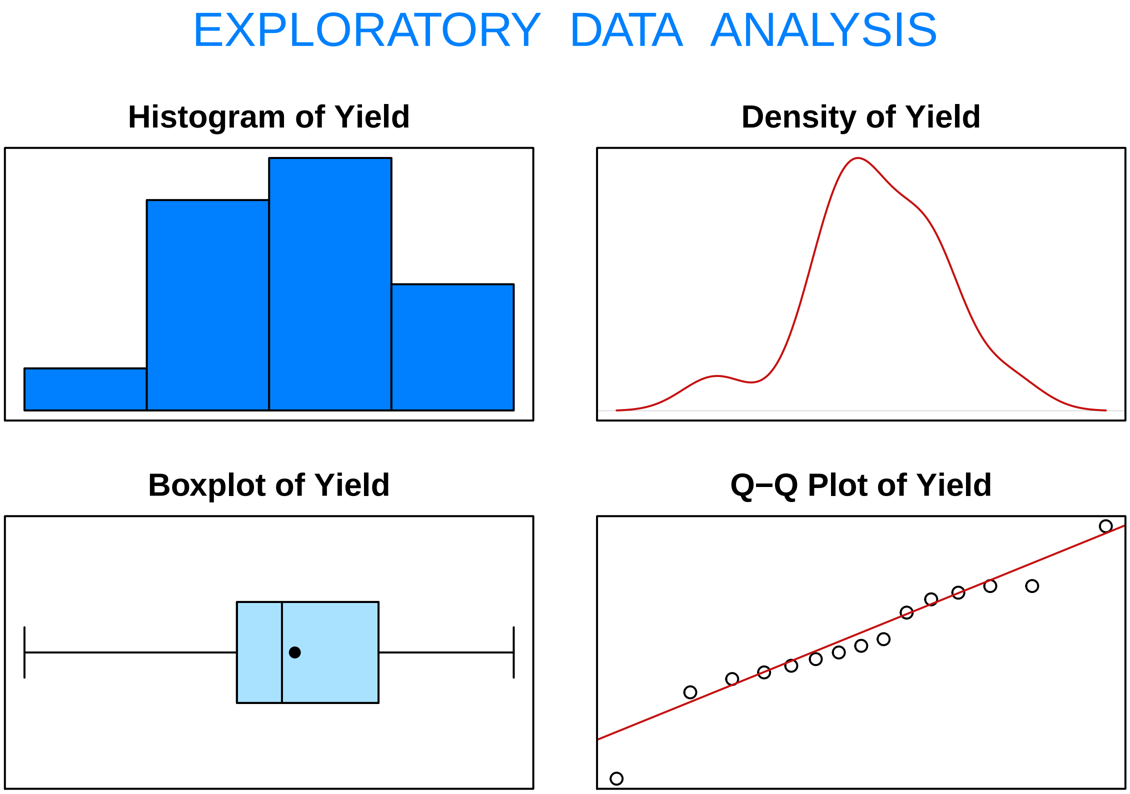
3.2 3.3 2.8 4.1 2.7 2.9 3.7

Verifica que los datos siguen una distribución normal. Si sabe que la producción con el anterior fertilizante es de dos fanegas por parcela, realiza un test de hipótesis para μ con un nivel de significación $\alpha = 0.05$.

1.2 Solución

X: Producción en fanegas de las parcelas

```
Yield <- c(2.5, 3, 3.1, 4, 1.2, 5, 4.1, 3.9,
           3.2, 3.3, 2.8, 4.1, 2.7, 2.9, 3.7)
eda(Yield)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##   15.000    0.000    1.200    2.850    3.300    3.200    3.300    3.950
##      Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    5.000    0.892    0.796    0.230    1.100    3.800    0.122   -0.343
## SW p-val
##    0.652
```

Si nos fijamos en el p-valor del test de Shapiro-Wilk, vemos que es mayor que $\alpha = 0.05$, así que no podemos rechazar la hipótesis nula que establece que los datos provienen de una población normal. Por tanto asumimos normalidad.

- $H_0 : \mu = 2$ vs. $H_1 : \mu > 2$
- Condiciones de aplicación: Población Normal, varianza desconocida

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- Cálculo:

```
t.test(x=Yield , alternative = "greater", mu = 2, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  Yield
## t = 5.6443, df = 14, p-value = 3.026e-05
## alternative hypothesis: true mean is greater than 2
```

```
## 95 percent confidence interval:
##  2.894334      Inf
## sample estimates:
## mean of x
##       3.3
```

- Decisión: Rechazar H_0 porque p-valor=$3.026e-05 < \alpha$
- Conclusión: Con un nivel de significación del 5% tenemos evidencias que indican que la producción media se incrementa con el nuevo fertilizante.

1.3

El data frame `STSCHOOL` del paquete `PASWR2` contiene los niveles de satisfacción de graduados de dos facultades, X e Y. Verifica si se puede asumir normalidad e igualdad de varianzas. ¿Hay diferencias significativas entre el nivel medio de satisfacción de los graduados de las dos facultades? Usa un nivel de significación del 5%.

X: nivel de satisfacción de los graduados de la facultad X

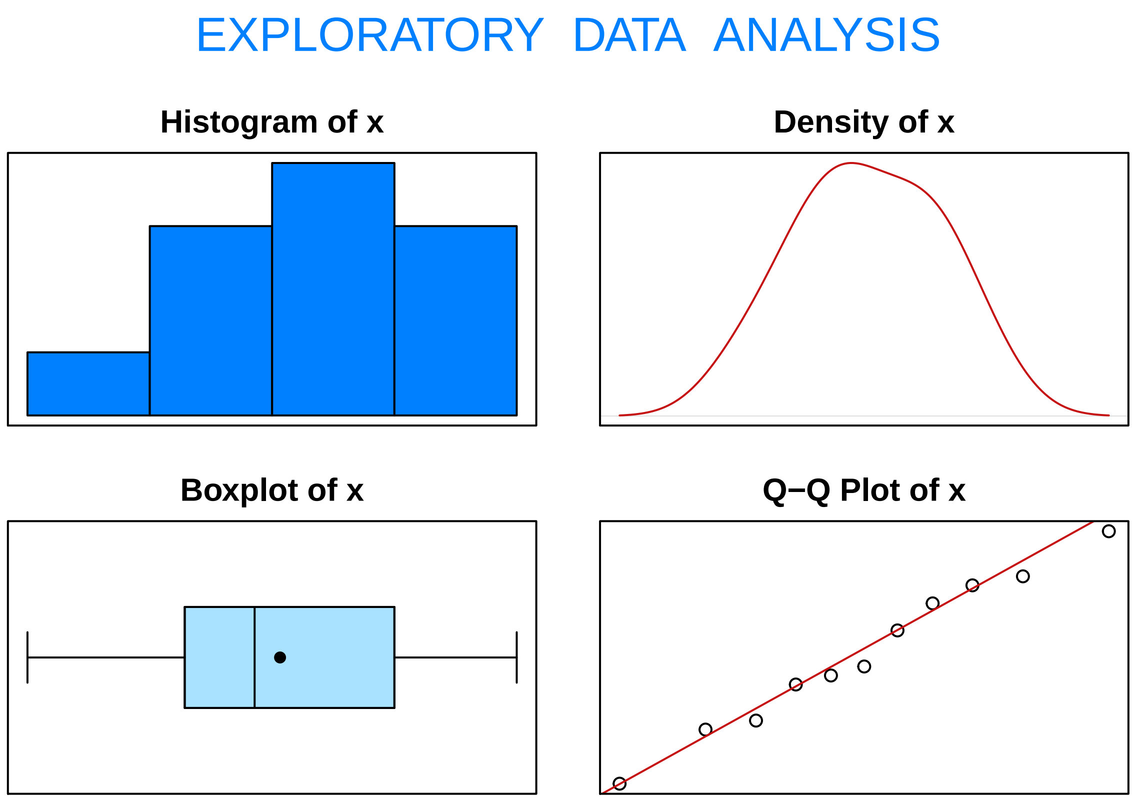
Y: nivel de satisfacción de los graduados de la facultad Y

```r
head(STSCHOOL)
```

```
##    x  y satisfaction school
## 1 69 59           59      y
## 2 75 62           62      y
## 3 76 66           66      y
## 4 80 70           70      y
## 5 81 70           70      y
## 6 82 75           75      y
```

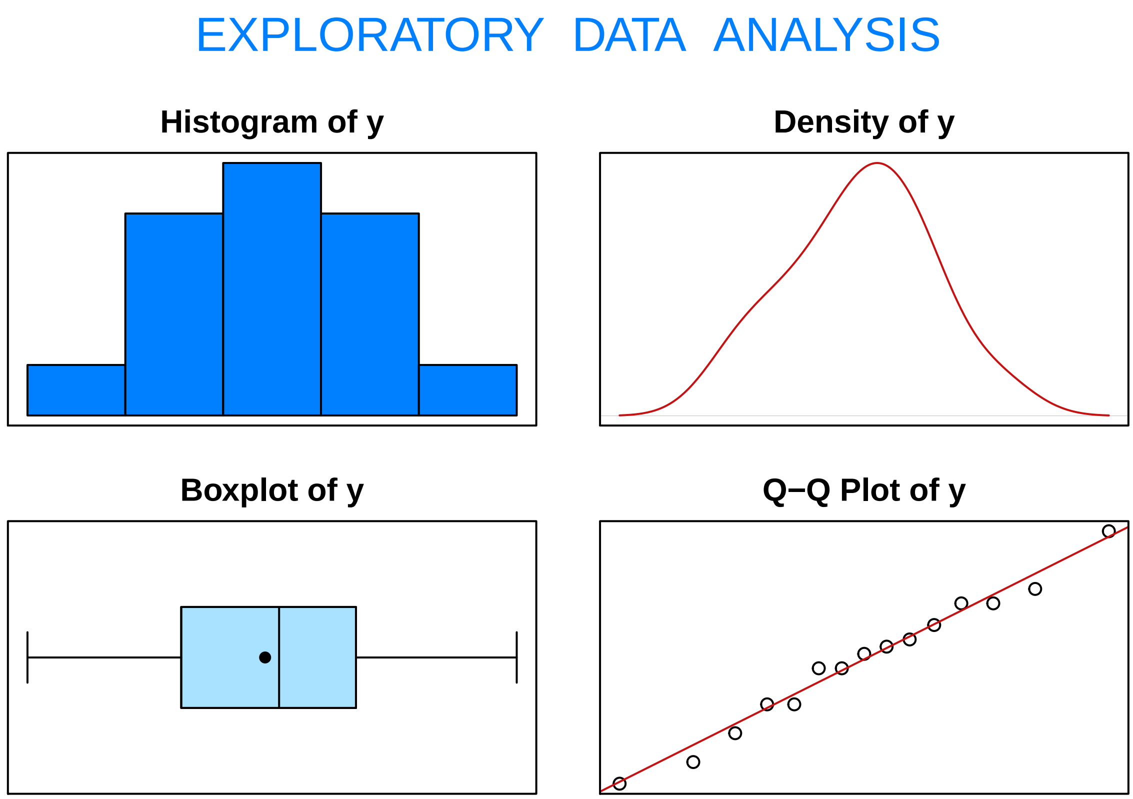
```r
x<- STSCHOOL$satisfaction[STSCHOOL$school=="x"]
y<- STSCHOOL$satisfaction[STSCHOOL$school=="y"]

eda(x)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##    11.000    0.000   69.000   78.000   83.455   82.000   83.455   90.000
##       Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    97.000    8.407   70.673    2.535   12.000   28.000   -1.266   -0.057
## SW p-val
##     0.977
```

```
eda(y)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##    15.000    0.000   59.000   70.000   76.000   77.000   76.000   82.500
##       Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    94.000    9.449   89.286    2.440   12.500   35.000   -0.839   -0.085
## SW p-val
##     0.985
```

En ambos casos, si nos fijamos en el p-valor del test de Shapiro-Wilk, al ser mayor o igual que α, vemos que no podemos rechazar la hipótesis nula que establece que los datos provienen de poblaciones normales. Por tanto asumimos normalidad.

- Condiciones de aplicación para el intervalo de confianza del cociente de varianzas: Poblaciones normales independientes (son independientes porque son alumnos distintos cogidos al azar de las dos facultades).

```r
var.test(x=x, y=y, conf.level = 0.95)$conf
```

```
## [1] 0.2515314 2.8102720
## attr(,"conf.level")
## [1] 0.95
```

Dado que el intervalo de confianza contiene el 1 podemos asumir que las varianzas son iguales.

- $H_0 : \mu_X - \mu_Y = 0$ vs. $H_1 : \mu_X - \mu_Y \neq 0$
- Condiciones de aplicación: Poblaciones normales independientes, varianzas desconocidas y asumimos iguales (ya justificado)
- Cálculos:

```r
t.test(x=x,y=y,var.equal=TRUE,alternative="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 2.0798, df = 24, p-value = 0.0484
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.05691592 14.85217499
## sample estimates:
## mean of x mean of y
##  83.45455  76.00000
```

- Decisión: Rechazar H_0 porque el p-valor = 0.0484 < α=0.05.
- Conclusión: Con un nivel de significación del 5% tenemos evidencias suficientes que indican que el nivel medio de satisfacción de los graduados de las dos escuelas es distinto.

1.4

El data frame `WATER` del paquete `PASWR2` almacena el nivel de contenido en sodio del agua de dos fuentes independientes X e Y. Verifica si se puede asumir normalidad e igualdad de varianzas. Realiza un test para ver si el contenido medio en sodio del agua de la fuente X es menor que el contenido medio en sodio del agua de la fuente Y. Usa un nivel de significación del $\alpha = 0.05$.

1.4 Solución

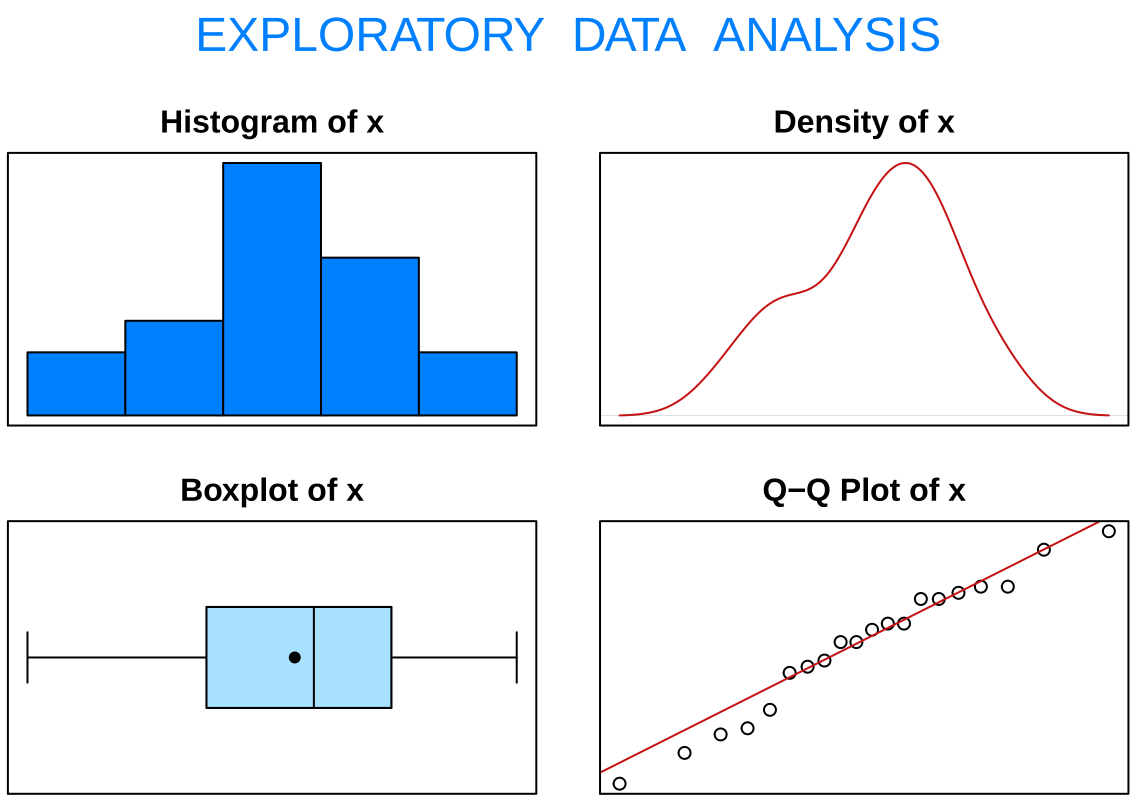
X: Nivel de contenido en sodio del agua de la fuente X

Y: Nivel de contenido en sodio del agua de la fuente Y

```r
head(WATER)
```

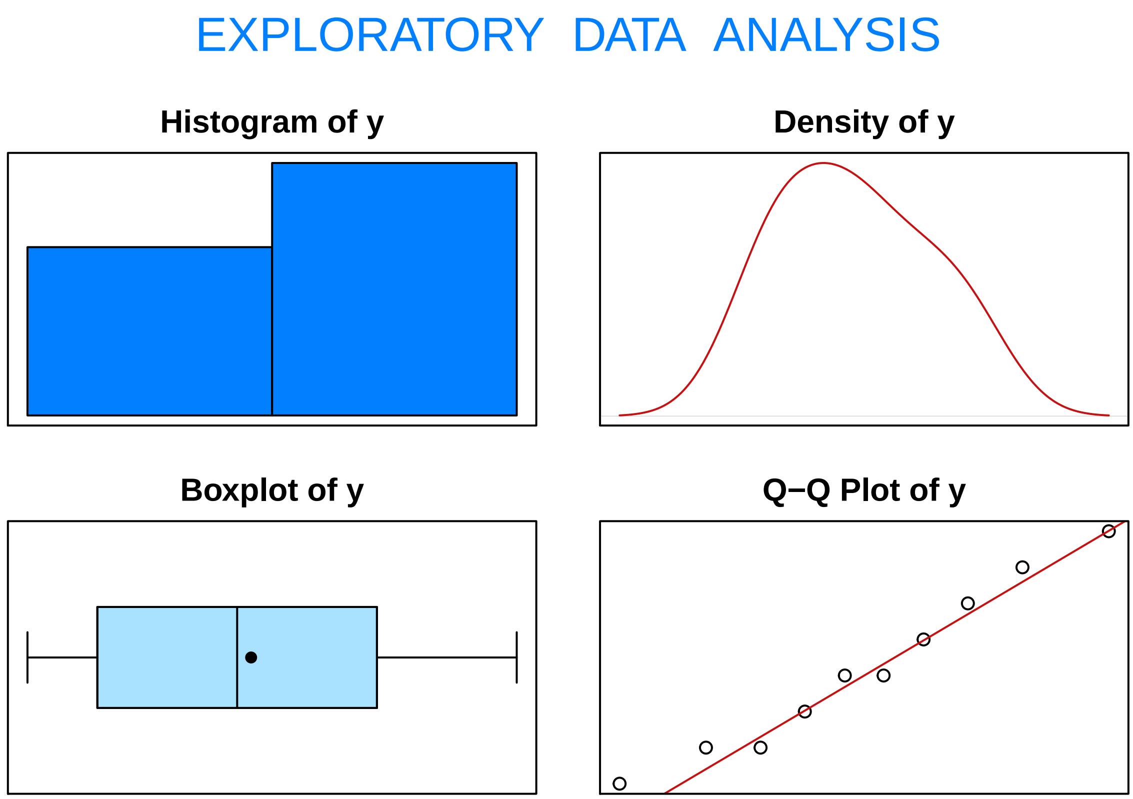
```
##    x  y sodium source
## 1 84 78     84      x
## 2 73 79     73      x
## 3 92 84     92      x
## 4 84 82     84      x
## 5 95 80     95      x
## 6 74 85     74      x
```

```r
x<-WATER$sodium[WATER$source=="x"]
y<-WATER$sodium[WATER$source=="y"]
eda(x)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##    20.000    0.000   54.000   70.500   76.400   78.000   76.611   84.250
##       Max    Stdev      Var  SE Mean  I.Q.R.    Range Kurtosis Skewness
##    95.000   11.081  122.779    2.478   13.750   41.000   -0.879   -0.350
## SW p-val
##     0.658
```

```
eda(y)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##    10.000    0.000   78.000   79.250   81.200   81.000   81.200   82.750
##       Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    85.000    2.300    5.289    0.727    3.500    7.000   -1.446    0.225
## SW p-val
##     0.799
```

En ambos casos, si nos fijamos en el p-valor del test de Shapiro-Wilk, al ser mayor o igual que α, vemos que no podemos rechazar la hipótesis nula que establece que los datos provienen de poblaciones normales. Por tanto asumimos normalidad.

- Condiciones de aplicación para el intervalo de confianza del cociente de varianzas: Poblaciones normales (ya justificado) e independientes (el enunciado dice que las fuentes son independientes).

```r
var.test(x=x, y=y, conf.level = 0.95)$conf
```

```
## [1]  6.302573 66.858988
## attr(,"conf.level")
## [1] 0.95
```

Dado que el intervalo de confianza no contiene el 1 por lo que tenemos evidencias de que las varianzas son distintas.

- $H_0 : \mu_X - \mu_Y = 0$ vs. $H_1 : \mu_X - \mu_Y < 0$
- Condiciones de aplicación: Poblaciones normales independientes, varianzas desconocidas y distintas (justificamos)

- Cálculos:

```
t.test(x,y, var.equal = FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -1.8589, df = 22.069, p-value = 0.03822
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3665724
## sample estimates:
## mean of x mean of y
##      76.4      81.2
```

- Decisión: Rechazar H_0 porque el p-valor=0.03822 $< \alpha$
- Conclusión: Con un nivel de significación del 5% tenemos evidencias suficientes que indican que el contenido medio en sodio del agua de la fuente X es menor que el de la fuente Y.

1.5

Usa el data frame `barley` del paquete `lattice` para contrastar si la producción media de cebada en 1932 en la zona de Morris es mayor que la producción media de cebada en 1932 en la zona de Crookston. Para ello, se cuenta con los datos de producción en 1932 de 10 variedades de cebada cultivadas tanto en Morris como en Crookston. Usa un nivel de significación de $\alpha = 0.05$.

1.5 Solución

X: Producción de cebada en 1932 en Morris

Y: Producción de cebada en 1932 en Crookston

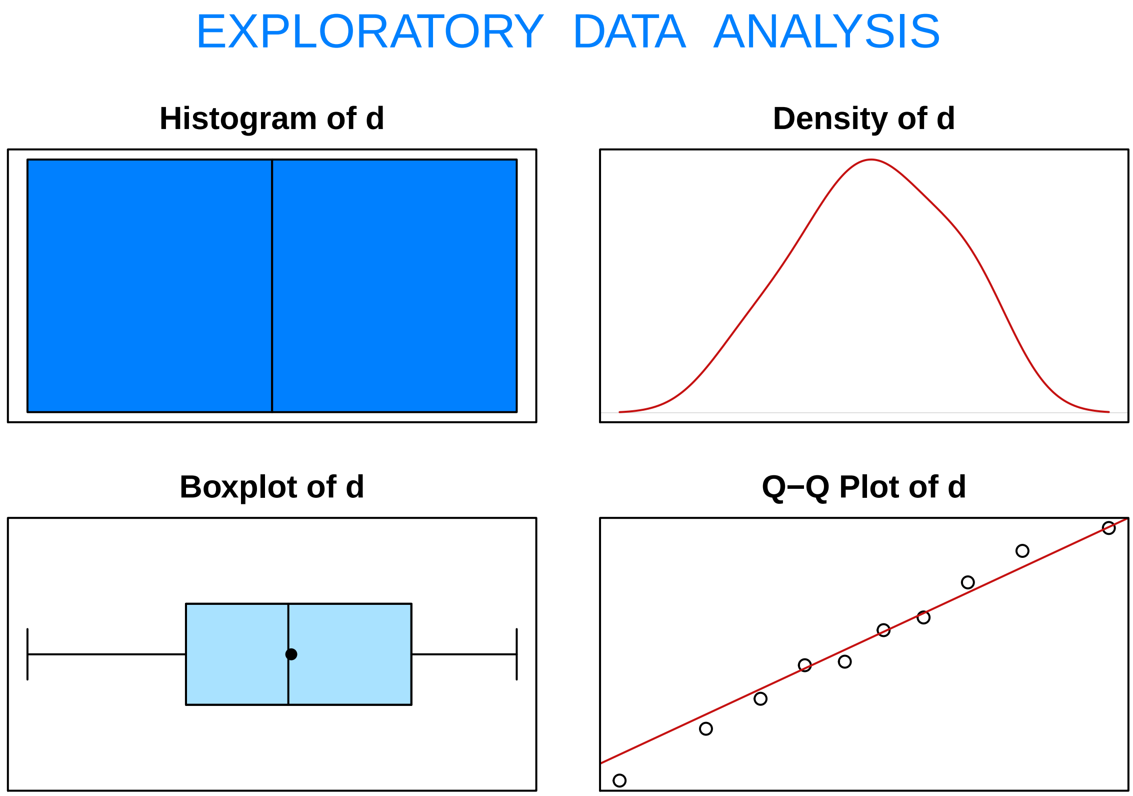
```
library(lattice)
head(barley)
```

```
##      yield   variety year            site
## 1 27.00000 Manchuria 1931 University Farm
## 2 48.86667 Manchuria 1931          Waseca
## 3 27.43334 Manchuria 1931          Morris
## 4 39.93333 Manchuria 1931       Crookston
## 5 32.96667 Manchuria 1931    Grand Rapids
## 6 28.96667 Manchuria 1931          Duluth
```

```
yieldMor32 <- barley$yield[barley$year == "1932" & barley$site =="Morris"]
yieldCro32 <- barley$yield[barley$year == "1932" & barley$site =="Crookston"]
```

Se trata de muestras relacionadas ya que estamos midiendo la producción de cada variedad de cebada en los dos sitios distintos.

```r
d <- yieldMor32 - yieldCro32
eda(d)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##   10.000    0.000    1.400    7.317   10.333   10.233   10.333   13.825
##      Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##   17.967    5.193   26.968    1.642    6.508   16.567   -1.272   -0.130
## SW p-val
##    0.979
```

Si nos fijamos en el p-valor del test de Shapiro-Wilk, al ser mayor o igual que α, vemos que no podemos rechazar la hipótesis nula que establece que los datos provienen de una población normal. Por tanto asumimos normalidad.

- $H_0 : \mu_X - \mu_Y = 0$ vs. $H_1 : \mu_X - \mu_Y > 0$
- Condiciones de aplicación: 2 poblaciones, muestras relacionadas. D=X-Y sigue una distribución normal (ya justificado)
- Cálculos:

```r
t.test(yieldMor32, yieldCro32, paired=TRUE, alternative="greater")
```

```
##
##  Paired t-test
##
## data:  yieldMor32 and yieldCro32
## t = 6.2924, df = 9, p-value = 7.113e-05
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  7.323012      Inf
## sample estimates:
## mean of the differences
##                10.33333
```

```r
# OR

t.test(d, alternative = "greater") # t-test on d
```

```
##
##  One Sample t-test
##
## data:  d
## t = 6.2924, df = 9, p-value = 7.113e-05
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  7.323012      Inf
## sample estimates:
## mean of x
##  10.33333
```

- Decisión: Rechazamos H_0 porque el p-valor=7.113e-05 $< \alpha$
- Conclusión: Con un nivel de significación del 5% tenemos evidencias que indican que la producción media en Morris en 1932 fué mayor que la producción media en Crookston.

2. Test de Hipótesis. Parte II

2.1

Este mes, un fabricante ha recibido más del doble de quejas de lo habitual con respecto a la variabilidad del diámetro de las arandelas de 4 cm. El data frame `WASHER` del paquete `PASWR2` guarda los diámetros de 20 de sus arandelas de 4 cm cogidas al azar. Si el pasado mes la varianza del diámetro de las arandelas fue de $\sigma^2 = 0.004cm^2$, contrasta si ha habido un incremento significativo de la variabilidad del diámetro. Usa $\alpha = 0.05$.

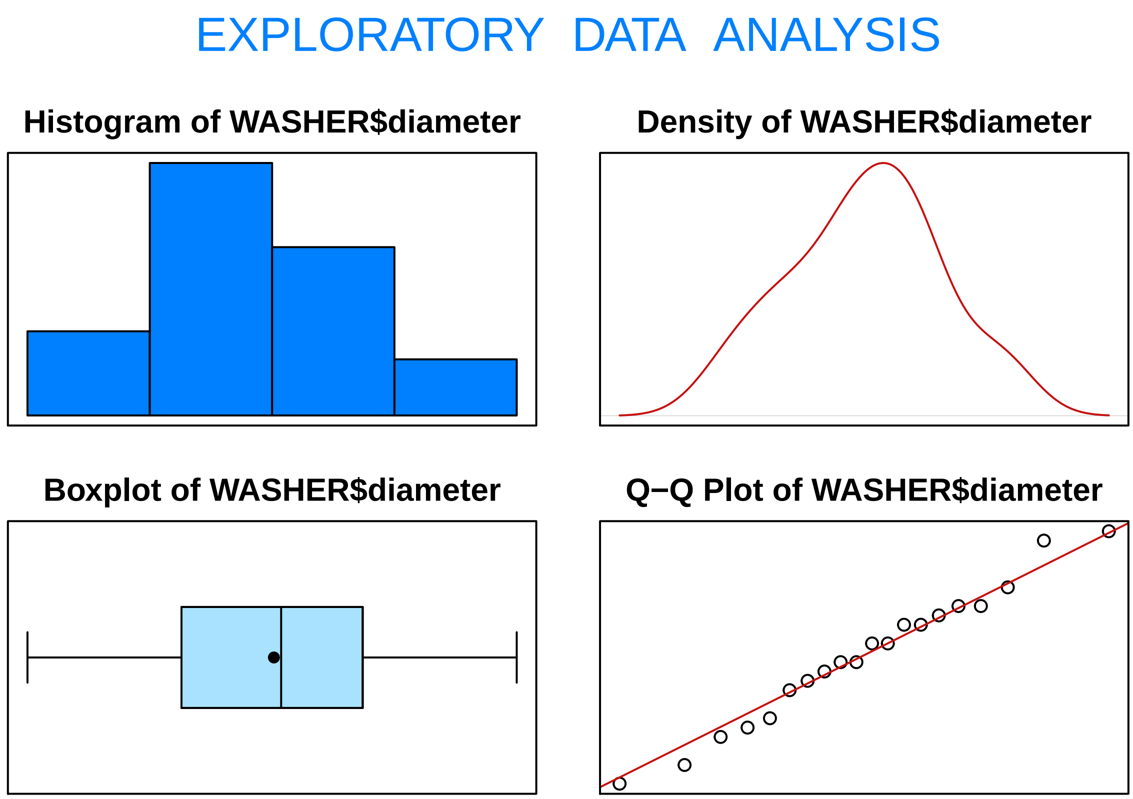
2.1 Solución

X: diámetro de las arandelas de 4cm

```r
WASHER$diameter
```

```
##  [1] 4.06 4.02 4.04 4.04 3.97 3.87 4.03 3.85 3.91 3.98 3.96 3.90 3.95 4.11 4.00
## [16] 4.12 4.00 3.98 3.92 4.02
```

```r
eda(WASHER$diameter)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##   20.000    0.000    3.850    3.943    3.986    3.990    3.987    4.032
##      Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    4.120    0.073    0.005    0.016    0.089    0.270   -0.799   -0.059
## SW p-val
##    0.936
```

Si nos fijamos en el p-valor del test de Shapiro-Wilk, al ser mayor o igual que α, vemos que no podemos rechazar la hipótesis nula que establece que los datos provienen de una población normal. Por tanto asumimos normalidad.

- $H_0 : \sigma^2 = 0.004$ vs. $H_1 : \sigma^2 > 0.004$
- Condiciones de aplicación: 1 población normal (ya justificado)
- Cálculos:

```
cv <- qchisq(0.95,19)             # Critical Value
s2 <- var(WASHER$diameter)
n <- length(WASHER$diameter)
n
```

```
## [1] 20
```

```
Chi2Obs <- (n - 1)*s2 / 0.004      # Standardized Test Statistic's Value
Chi2Obs
```

```
## [1] 25.26375
```

```r
pvalue <- 1-pchisq(Chi2Obs, n-1)
c(CriticalValue = cv, Chi2Obs = Chi2Obs, pvalue = pvalue)
```

```
## CriticalValue       Chi2Obs        pvalue
##    30.1435272    25.2637500     0.1520425
```

- Decisión: No rechazamos H_0 porque el p-valor= $0.1520425 \geq \alpha$
- Conclusión: Con un nivel de significación del 5% no tenemos evidencias suficientes que indiquen que la variabilidad del diámetro de las arandelas haya aumentado.

2.2

Los números almacenados en el data frame `BAC` (paquete `PASWR2`) son los valores ordenados del nivel de alcohol en sangre obtenidos con alcoholímetros de las compañías X e Y en dos grupos de personas cogidas al azar. Comprueba si los alcoholímetros de la compañía X tiene menor varianza que los alcoholímetros de la compañía Y usando un nivel de significación del 5%.

2.2 Solución

X: nivel de alcohol medido con el alcoholímetro de la marca X

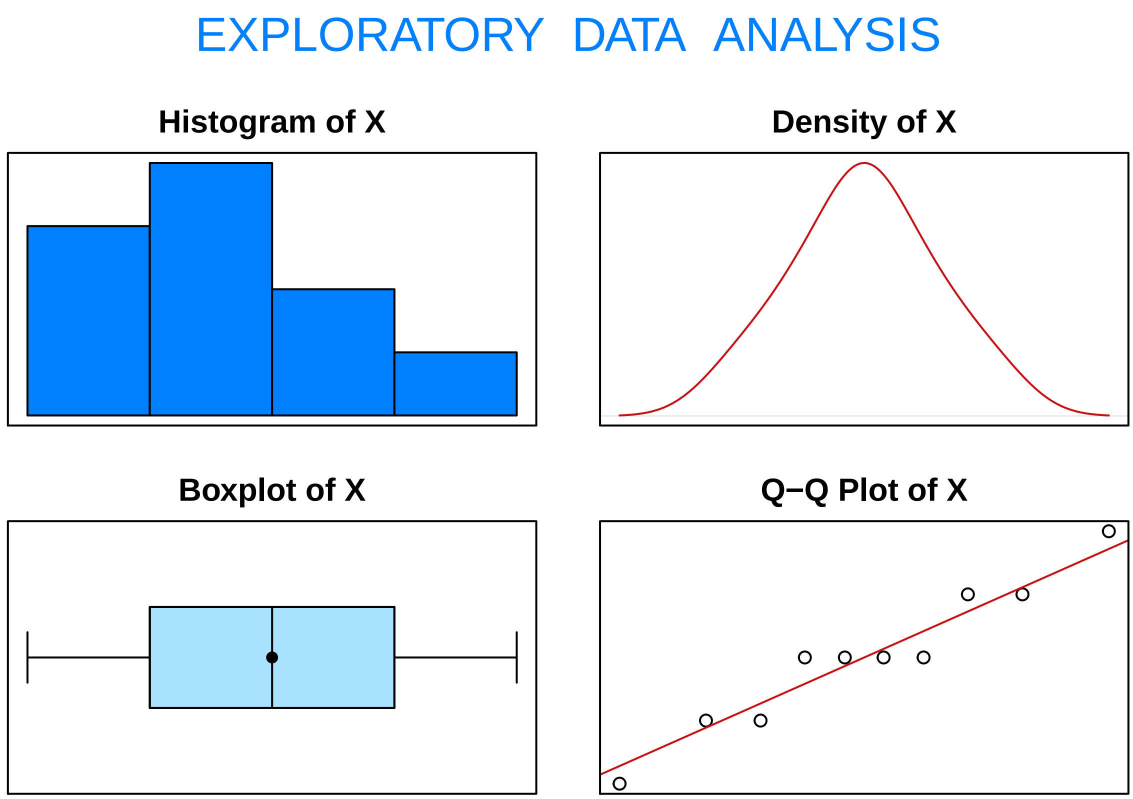
Y: nivel de alcohol medido con el alcoholímetro de la marca Y

```r
BAC
```

```
##       Y    X
## 1  0.00 0.08
## 2  0.03 0.09
## 3  0.04 0.09
## 4  0.04 0.10
## 5  0.05 0.10
## 6  0.05 0.10
## 7  0.06 0.10
## 8  0.07 0.11
## 9  0.08 0.11
## 10 0.08 0.12
```

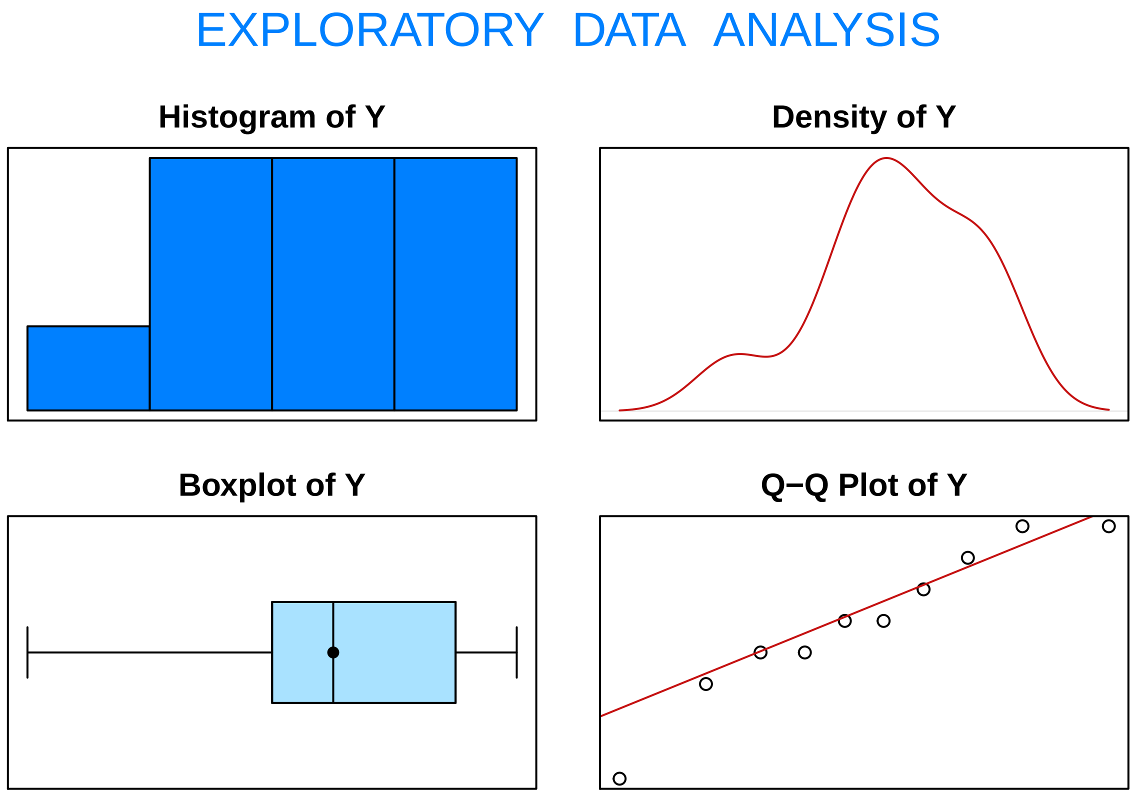
```r
X <- BAC$X
Y <- BAC$Y

eda(X)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##    10.000    0.000    0.080    0.092    0.100    0.100    0.100    0.108
##       Max    Stdev      Var  SE Mean  I.Q.R.    Range Kurtosis Skewness
##     0.120    0.012    0.000    0.004    0.016    0.040   -0.975    0.000
## SW p-val
##     0.703
```

```
eda(Y)
```




```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##   10.000    0.000    0.000    0.040    0.050    0.050    0.050    0.068
##      Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##    0.080    0.024    0.001    0.008    0.028    0.080   -0.717   -0.490
## SW p-val
##    0.549
```

En ambos casos, si nos fijamos en el p-valor del test de Shapiro-Wilk, al ser mayor o igual que α, vemos que no podemos rechazar la hipótesis nula que establece que los datos provienen de una población normal. Por tanto asumimos normalidad.

- $H_0 : \sigma_X^2/\sigma_Y^2 = 1$ vs. $H_1 : \sigma_X^2/\sigma_Y^2 < 1$
- Condiciones de aplicación: 2 poblaciones normales independientes.
- Cálculo:

```r
var.test(X, Y, alternative = "less")
```

```
##
##  F test to compare two variances
##
## data:  X and Y
## F = 0.22222, num df = 9, denom df = 9, p-value = 0.01764
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0000000 0.7064207
## sample estimates:
## ratio of variances
##          0.2222222
```

- Decisión: Rechazar H_0 porque el p-valor=0.01764 $< \alpha$.
- Conclusión: Con un nivel de significación del 5% tenemos evidencias que indican que la variabilidad del alcoholímetro X es menor que la del Y.

2.3

90 graduados de una muestra aleatoria de 500 obtienen trabajo en su campo de estudio. Si un estudio afirma que el 20% de los graduados encuentra trabajo en su campo de estudio, ¿hay evidencia estadística que desmienta la afirmación del estudio con un nivel de significación $\alpha = 0.05$?

2.3 Solución

X: número de estudiantes que encuentran trabajo en su campo de estudio.

- $H_0 : \pi = 0.2$ vs. $H_1 : \pi \neq 0.2$
- Condiciones de aplicación: muestras grandes ($n \cdot p \geq 10$, $n \cdot (1-p) \geq 10$)
- Cálculos:

```r
prop.test(x = 90, n = 500, p = 0.2, alternative="two.sided", correct = FALSE) ### Sin corrección de con
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  90 out of 500, null probability 0.2
## X-squared = 1.25, df = 1, p-value = 0.2636
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.1488049 0.2160747
## sample estimates:
##    p
## 0.18
```

```r
prop.test(x = 90, n = 500, p = 0.2, alternative="two.sided", correct = TRUE)  ### Con corrección de con
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  90 out of 500, null probability 0.2
## X-squared = 1.1281, df = 1, p-value = 0.2882
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.1478847 0.2171388
## sample estimates:
##    p
## 0.18
```

- Decisión: No rechazar H_0 porque el p-valor=0.2636 $\geq \alpha$.
- Conclusión: Con un nivel de significación del 5% no tenemos evidencias suficientes que indiquen que la afirmación realizada en el estudio sea falsa.

2.4

Usa los datos de la la siguiente tabla para contrastar si los médicos que toman aspirina tienen menor predisposición a sufrir ataques al corazón que los que toman un placebo. Usa un nivel de significación α de 0.05.

Med/attack	Heart attacks	No heart attack	Total
Aspirine	104	10933	11037
Placebo	189	10845	11034
Total	293	21778	22071

####2.4 Solución

X: número de personas que sufren ataque al corazón de las que toman aspirina

Y: número de personas que sufren ataque al corazón de las que toman placebo

- $H_0 : \pi_X - \pi_Y = 0$ vs. $H_1 : \pi_X - \pi_Y < 0$
- Condiciones de aplicación: muestras grandes ($n_X \cdot p_X \geq 10$, $n_X \cdot (1 - p_X) \geq 10$, $n_Y \cdot p_Y \geq 10$, $n_Y \cdot (1 - p_Y) \geq 10$)
- Cálculos:

```r
### Sin corrección de continuidad
prop.test(x = c(104, 189), n = c(11037, 11034), correct = FALSE,
          alternative = "less")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(104, 189) out of c(11037, 11034)
## X-squared = 25.014, df = 1, p-value = 2.846e-07
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.000000000 -0.005173009
## sample estimates:
##     prop 1     prop 2
## 0.00942285 0.01712887
```

```r
### Con corrección de continuidad
prop.test(x = c(104, 189), n = c(11037, 11034), correct = TRUE,
          alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(104, 189) out of c(11037, 11034)
## X-squared = 24.429, df = 1, p-value = 3.855e-07
## alternative hypothesis: less
## 95 percent confidence interval:
```

```
##  -1.000000000 -0.005082393
## sample estimates:
##     prop 1     prop 2
## 0.00942285 0.01712887
```

- Decisión: rechazar H_0 porque el p-valor = 3.855e-07 $< \alpha$.
- Conclusión: Con un nivel de significación del 5% tenemos evidencias que indican que el uso de aspirina reduce la posibilidad de sufrir un ataque al corazón.