Universidad de Alcalá

Estadística

Práctica 10

Contrastes de hipótesis no paramétricos

Universidad de Alcalá

Contenido

Universidad de Alcalá

1. Introducción

Con esta práctica se utiliza R y RStudio para determinar si se puede aceptar una hipótesis, con un nivel de confianza determinado, sobre un parámetro (mediana) de una o dos poblaciones a partir de una o dos muestras respectivamente, cuando los datos no tienen una distribución Normal.

Para ello, lo primero que hay que hacer es comprobar la normalidad de los datos de la muestra o muestras. Si no superan la prueba de normalidad se deben aplicar los contrastes no paramétricos tratados en esta práctica, porque si se supera la prueba de normalidad se pueden aplicar los contrastes paramétricos tratados en la práctica anterior.

Para realizar la prueba de normalidad, se suele usar la función shapiro.test() de R cuando la muestra es pequeña[1] (prueba de Shapiro-Wilk), o la función ks.test() cuando la muestra es grande (prueba de Kolmogorov-Smirnov), con el siguiente formato:

```
> shapiro.test(muestra)

> ks.test(muestra, pnorm, mean(muestra), sd(muestra))
```

Si el resultado obtiene un p-valor<0.05, con un nivel de confianza del 95% no puede afirmarse que los datos tengan una distribución Normal.

Por ejemplo:

```
> x = c(4.4, 3, 0.6, 1.7, 0.4, 1.0, 1.4, 0.9, 4.5, 0.2)
> shapiro.test(x)
	Shapiro-Wilk normality test
data:  x
W = 0.84364, p-value = 0.04878
```

En este caso el p-valor es menor que 0.05, por lo que los datos de la muestra no son Normales y hay que aplicar un contraste de hipótesis no paramétrico.

[1] No hay unanimidad respecto a lo que se considera muestra pequeña o grande. En general, para estas pruebas se suele suponer el límite en 50 datos. Aunque hay autores que defienden que el test de Shapiro-Wilk sólo debería realizarse para una muestra de como mínimo 30 datos.

2. Contraste de hipótesis para la mediana poblacional

Se trata de realizar uno de los siguientes contrastes:

- $H_0 \colon M = m_o$ vs $H_A \colon M \neq m_o$
- $H_0 \colon M = m_o$ vs $H_A \colon M < m_o$
- $H_0 \colon M = m_o$ vs $H_A \colon M > m_o$

Siendo M la mediana poblacional y m_o el posible valor de la mediana frente al que se quiere contrastar.

Se puede utilizar un contraste de signos o un contraste de rangos con signo de Wilcoxon.

Para ello, primero se establece un nivel de significación α para el contraste (normalmente 0.05), después hay que calcular el valor del estadístico de contraste, y finalmente el p-valor, de tal forma que el p-valor es:

- Mayor que α, entonces no se puede rechazar H_0.
- Menor que α, entonces se rechaza H_0.

El cálculo del estadístico de contraste puede realizarse en R como se indica en la siguiente tabla.

Tipo	Estadístico de contraste	Código R

Signos	$S_o = n^{\underline{o}}\ de\ obsevaciones > m_o$	`x=muestra`
		`mo=valor a comprobar en la hipótesis`
		`n=length(x[x!=mo])`
		`so=sum(x>mo)`
		
		`NOTA: con x[x!=mo] se eliminan de la muestra los datos iguales a mo`

Wilcoxon	$W_o = \sum_{i:\, X_i > m} R_i$	`No es necesario calcularlo, lo calcula la función wilcox.test()`

Para el contraste se pueden usar las funciones binom.test y wilcox.test de R.

Tipo	Contraste de hipótesis	Código R
Signos	$H_0: M = m_o$ vs $H_A: M \neq m_o$	`> binom.test(so, n, p=0.5, alternative="two.sided", conf.level=0.95)`
	$H_0: M = m_o$ vs $H_A: M < m_o$	`> binom.test(so, n, p=0.5, alternative="less", conf.level=0.95)`
	$H_0: M = m_o$ vs $H_A: M > m_o$	`> binom.test(so, n, p=0.5, alternative="greater", conf.level=0.95)`
Wilcoxon	$H_0: M = m_o$ vs $H_A: M \neq m_o$	`> wilcox.test(x, mu=mo, alternative="two.sided", conf.level=0.95, exact=FALSE)`
	$H_0: M = m_o$ vs $H_A: M < m_o$	`> wilcox.test(x, mu=mo, alternative="less", conf.level=0.95, exact=FALSE)`
	$H_0: M = m_o$ vs $H_A: M > m_o$	`> wilcox.test(x, mu=mo, alternative="greater", conf.level=0.95, exact=FALSE)`

Ejemplo:

$Muestra = (4.4, 3, 0.6, 1.7, 0.4, 1.0, 1.4, 0.9, 4.5, 0.2)$
$m_o = 3$

1) Cálculo del estadístico S_o para el contraste de signos

```
> (x=c(4.4, 3, 0.6, 1.7, 0.4, 1.0, 1.4, 0.9, 4.5, 0.2))
 [1] 4.4 3.0 0.6 1.7 0.4 1.0 1.4 0.9 4.5 0.2
> (mo=3)
[1] 3
> (n=length(x[x!=mo]))
[1] 9
> (so=sum(x>mo))
[1] 2
```

2) Contrastes aplicando el método de los signos y el método de los rangos con signo de Wilcoxon

Contraste de hipótesis	Tipo	Código R

$H_0: M = 3 \text{ vs } H_A: M \neq 3$	Signos	`> binom.test(so,n,p=0.5,alternative="two.sided", conf.level=0.95)`
		
		`... p-value = 0.1797`

	Wilcoxon	`> wilcox.test(x, mu=mo, alternative="two.sided", conf.level=0.95, exact=FALSE)`
		
		`... p-value = 0.04401`

$H_0: M = 3 \text{ vs } H_A: M < 3$	Signos	`> binom.test(so,n,p=0.5,alternative="less", conf.level=0.95)`
		
		`... p-value = 0.08984`

	Wilcoxon	`> wilcox.test(x, mu=mo, alternative="less", conf.level=0.95, exact=FALSE)`
		
		`... p-value = 0.02201`

$H_0: M = 3 \text{ vs } H_A: M > 3$	Signos	`> binom.test(so,n,p=0.5,alternative="greater", conf.level=0.95)`
		
		`... p-value = 0.9805`

	Wilcoxon	`> wilcox.test(x, mu=mo, alternative="greater", conf.level=0.95, exact=FALSE)`
		
		`... p-value = 0.9835`

Puede comprobarse que con ambos métodos se rechaza la hipótesis alternativa $H_A: M > 3$ porque en ambos casos el p-valor es mayor que 0.05. Pero el en caso de $H_A: M < 3$ con el contraste de Wilcoxon se acepta mientras que con el contraste de signos se rechazaría. Es más preciso el contraste de rangos con signo Wilcoxon, por lo que es mejor utilizarlo en lugar del contraste de signos.

Universidad de Alcalá

3. Ejemplo: Contraste de hipótesis sobre la mediana de calificaciones de una población de estudiantes

ENUNCIADO

Mediante una encuesta, se saben las calificaciones de acceso a la universidad de 74 estudiantes de un curso de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá, de una población de 108 matriculados. Responder a las siguientes preguntas con contrastes de rangos con signo de Wilcoxon:

a) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de toda la población es diferente a 7.5?**
b) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de toda la población de alumnos matriculados en la asignatura es inferior a 7.5?**
c) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de toda la población de alumnos matriculados en la asignatura es superior a 7.5?**

SOLUCIÓN

Primero hay que leer los datos de las notas, disponibles en el archivo encuesta.csv:

```
> encuesta = read.csv2("encuesta.csv")
> (nota=encuesta$NOTA)
 [1]  8.500  7.100  8.635  8.624  8.200  8.700  7.210  7.630  8.400  8.300
[11]  8.200  9.100  9.789 10.107  8.015  7.310  7.500  8.710  8.336  9.210
[21]  7.800 10.300  7.990  6.900  7.800 10.000  8.590  7.000  8.050 10.799
[31]  7.994  8.550  7.340  6.750  9.560  7.417  6.936  7.210  7.680 10.277
[41]  7.860 10.260  7.270  5.800  7.300  7.140  8.600  7.500  8.000  7.540
[51]  7.292  7.830  6.750  9.806  6.800  6.445  6.650  7.804 10.270  7.600
[61]  7.870  7.000  7.085  7.480  8.070  5.820  6.500  9.900  7.500  6.500
[71]  9.456  8.000  7.800  7.654
```

NOTA: Con estos datos realmente se podría utilizar un contraste paramétrico, porque si se realiza una prueba de normalidad con el test de Kolmogorov-Smirnov (por ser una muestra grande), se obtiene un p-valor mayor que 0.05, por lo que se puede aceptar que son datos Normales:

```
> ks.test(nota,pnorm,mean(nota),sd(nota))

      Asymptotic one-sample Kolmogorov-Smirnov test

data:  nota
D = 0.11842, p-value = 0.2505
alternative hypothesis: two-sided
```

Si se aplicase un contraste paramétrico, al suponer que es una población Normal, donde la media y la mediana son iguales, el contraste sobre la media y la mediana coinciden.

Sin embargo, en este ejercicio, a pesar de ser una población con distribución Normal, se puede usar también un contraste no paramétrico, ya que este tipo de contraste se puede aplicar en cualquier caso, tanto si los datos son normales o no; aunque si son normales es recomendable realizar un contraste paramétrico.

a) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población es diferente a 8.6?

Se formula el siguiente contraste de hipótesis:

$$H_0: M = 7.5 \text{ vs } H_A: M \neq 7.5$$

```
> wilcox.test(nota, mu=7.5, alternative="two.sided", conf.level=0.95, exact=FALSE)

	Wilcoxon signed rank test with continuity correction

data:  nota
V = 1867.5, p-value = 0.000738
alternative hypothesis: true location is not equal to 7.5
```

Como el p-valor es menor que alfa (0.05), se rechaza la hipótesis nula y se acepta la alternativa de que la mediana poblacional es diferente a 7.5.

El resultado de la función `wilcox.test()` incluye más información:

- Indica que el valor del estadístico de contraste W_o (llamado V en el resultado de la función) es 1867.5.
- Indica que la hipótesis alternativa es que la mediana fuera diferente a 7.5, como así se indicó al llamar al a la función con el parámetro `alternative="two.sided"`
- Indica que el nivel de confianza es del 95%, como así se indicó al llamar a la función con el parámetro `conf.level=0.95`.

b) ¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de toda la población de alumnos matriculados en la asignatura es inferior a 7.5?

Se formula el siguiente contraste de hipótesis:

$$H_0: M = 7.5 \text{ vs } H_A: M < 7.5$$

```
> wilcox.test(nota, mu=7.5, alternative="less", conf.level=0.95, exact=FALSE)

	Wilcoxon signed rank test with continuity correction

data:  nota
V = 1867.5, p-value = 0.9996
alternative hypothesis: true location is less than 7.5
```

Como el p-valor es mayor que 0.05 no se puede rechazar la hipótesis nula, y por tanto no se puede afirmar que la mediana sea inferior a 7.5.

c) ¿Se puede afirmar con una confianza del 95% que la mediana de notas de acceso de toda la población es superior a 7.5?

Se formula el siguiente contraste de hipótesis:

$$H_0: M = 7.5 \text{ vs } H_A: M > 7.5$$

```
> wilcox.test(nota, mu=7.5, alternative="greater", conf.level=0.95, exact=FALSE)

      Wilcoxon signed rank test with continuity correction

data:  nota
V = 1867.5, p-value = 0.000369
alternative hypothesis: true location is greater than 7.5
```

Como el p-valor es menor que 0.05, entonces se rechaza la hipótesis nula, y se acepta la hipótesis alternativa, por lo que la respuesta es que **SÍ se puede afirmar con una confianza del 95% que la mediana poblacional sea superior a 7.5**.

4. Contraste de hipótesis sobre la diferencia de medianas

Si x e y son dos vectores con las muestras de las poblaciones X e Y independientes, entonces se utilizarán el método de la suma de rangos con signo de Mann-Whitney-Wilcoxon.

Tipo	Contraste de hipótesis	Función R

Wilcoxon
 	
Muestras pareadas del mismo tamaño n
	$H_0: M_{(X-Y)} = 0$ vs $H_A: M_{(X-Y)} \neq 0$	`> wilcox.test(x, y,`
		`paired=TRUE,`
		`alternative="two.sided",`
		`conf.level=0.95, exact=FALSE)`

	$H_0: M_{(X-Y)} = 0$ vs $H_A: M_{(X-Y)} < 0$	`> wilcox.test(x, y,`
		`paired=TRUE,`
		`alternative="less",`
		`conf.level=0.95, exact=FALSE)`

	$H_0: M_{(X-Y)} = 0$ vs $H_A: M_{(X-Y)} > 0$	`> wilcox.test(x, y,`
		`paired=TRUE,`
		`alternative="greater",`
		`conf.level=0.95, exact=FALSE)`

Mann-Whitney-Wilcoxon
 	
Muestras Independientes de tamaños n y m
	$H_0: M_X - M_Y = 0$ vs $H_A: M_X - M_Y \neq 0$	`> wilcox.test(x, y,`
		`paired=FALSE,`
		`alternative="two.sided",`
		`conf.level=0.95, exact=FALSE)`

	$H_0: M_X - M_Y = 0$ vs $H_A: M_X - M_Y < 0$	`> wilcox.test(x, y,`
		`paired=FALSE,`
		`alternative="less",`
		`conf.level=0.95, exact=FALSE)`

	$H_0: M_X - M_Y = 0$ vs $H_A: M_X - M_Y > 0$	`> wilcox.test(x, y,`
		`paired=FALSE,`
		`alternative="greater",`
		`conf.level=0.95, exact=FALSE)`

4.1 Ejemplo con muestras pareadas

Las siguientes muestras representan el peso de 10 personas antes (x) y después (y) de una dieta.

x = (90.1, 97.7, 66.1, 65.3, 70.6, 98.4, 65.2, 88.2, 90.4, 90.8)

y = (89.2, 96.1, 68.5, 67.4, 69.6, 98.9, 67.2, 90.1, 89.9, 91.1)

Primero se comprueba si tienen distribución Normal, aplicando el test de normalidad de Shapiro-Wilk (por ser una muestra pequeña).

```
> x=c(90.1, 97.7, 66.1, 65.3, 70.6, 98.4, 65.2, 88.2, 90.4, 90.8)
> shapiro.test(x)
      Shapiro-Wilk normality test
a:  x
W = 0.82241, p-value = 0.0271

> y=c(89.2, 96.1, 68.5, 67.4, 69.6, 98.9, 67.2, 90.1, 89.9, 91.1)
> shapiro.test(y)
      Shapiro-Wilk normality test
data:  y
W = 0.81267, p-value = 0.02066
```

En ambos casos se comprueba que el p-valor es menor que 0.05, por lo que las muestras no tienen distribución Normal, por lo que no se pueden aplicar contrastes paramétricos como z.test o t.test, sino contrastes no paramétricos, como wilcox.test.

Vamos a realizar el contraste:

$$H_0: M_{(X-Y)} = 0 \text{ vs } H_A: M_{(X-Y)} < 0$$

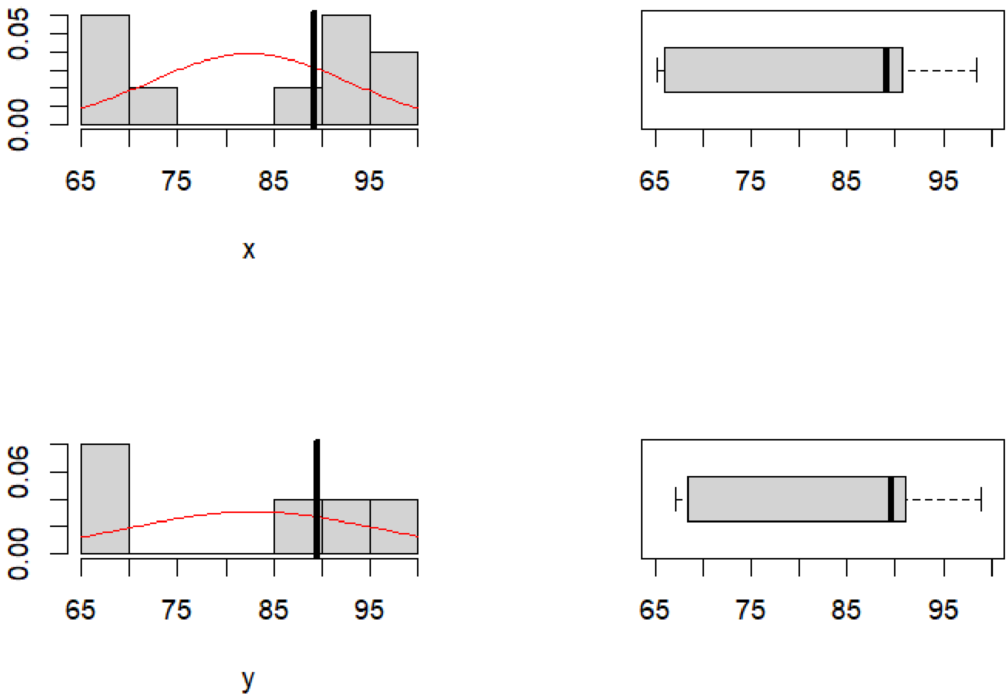
```
> wilcox.test(x, y, paired=TRUE, alternative="two.sided", conf.level=0.95,
exact=FALSE)
      Wilcoxon signed rank test with continuity correction
data:  x and y
V = 17.5, p-value = 0.3326
alternative hypothesis: true location shift is not equal to 0
```

Como el p-valor es menor que 0.05, no se puede rechazar H_0, por lo que se puede afirmar que las medianas son iguales, es decir no ha habido un cambio de peso significativo de las personas después de hacer la dieta.

Podemos dibujar los histogramas de x e y, y sobre ellos la función de distribución Normal y una línea en la posición de la mediana en cada caso. Y también los diagramas de caja.

```
> par(mfrow=c(2,2))
> hist(x, freq=FALSE, main="", ylab="")
> curve(dnorm(x,mean(x),sd(x)), 65, 100, add=TRUE, col="red")
> abline(v=median(x), lwd=3)
> boxplot(x, horizontal = TRUE, ylim=c(65,100))
> hist(y, freq=FALSE, main="", ylab="")
> curve(dnorm(x,mean(y),sd(y)), 65, 100, add=TRUE, col="red")
```

```
> abline(v=median(y), lwd=3)
> boxplot(y, horizontal = TRUE, ylim=c(65,100))
```



El valor de la mediana de cada muestra es el siguiente:

```
> median(x)
[1] 89.15
> median(y)
[1] 89.55
```

Son valores muy parecidos a nivel de muestras, y según el resultado del contraste, con un 95% de confianza puede afirmarse que también san similares a nivel de población.

4.2 Ejemplo con muestras independientes: contraste de hipótesis sobre la diferencia de medianas de las calificaciones de estudiantes del turno de mañana y del turno de tarde

ENUNCIADO

Mediante una encuesta, se saben las calificaciones de acceso a la universidad de 42 estudiantes del turno de mañana y de 32 estudiantes del turno de tarde en un curso de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá, de una población de 52 matriculados en el turno de mañana y 56 en el turno de tarde. Si se supone que las calificaciones tienen una distribución Normal, responder a las siguientes preguntas:

a) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde?**
b) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde?**
c) **¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde?**

SOLUCIÓN

Podemos leer los datos por turno del archivo encuesta_turno.csv ya utilizado en una práctica anterior.

```
> encuesta.con.turno=read.csv2("encuesta_turno.csv")

> (nota.mañana=encuesta.con.turno[encuesta.con.turno$TURNO=="Mañana",]$NOTA)
 [1]  6.75  6.90  6.94  7.00  7.10  7.21  7.21  7.31  7.34  7.42
[11]  7.50  7.63  7.68  7.80  7.80  7.86  7.99  7.99  8.02  8.05
[21]  8.20  8.20  8.30  8.34  8.40  8.50  8.55  8.59  8.62  8.63
[31]  8.70  8.71  9.10  9.21  9.56  9.79 10.00 10.11 10.26 10.28
[41] 10.30 10.80
> (nota.tarde=encuesta.con.turno[encuesta.con.turno$TURNO=="Tarde",]$NOTA)
 [1]  5.80  5.82  6.44  6.50  6.50  6.65  6.75  6.80  7.00  7.08
[11]  7.14  7.27  7.29  7.30  7.48  7.50  7.50  7.54  7.60  7.65
[21]  7.80  7.80  7.83  7.87  8.00  8.00  8.07  8.60  9.46  9.81
[31]  9.90 10.27
```

Primero se comprueba si tienen distribución Normal, aplicando el test de normalidad de Shapiro-Wilk.

```
> shapiro.test(nota.mañana)

      Shapiro-Wilk normality test

data:  nota.mañana
```

```
W = 0.94009, p-value = 0.02879

> shapiro.test(nota.tarde)

      Shapiro-Wilk normality test

data:  nota.tarde
W = 0.91995, p-value = 0.02072
```

En ambos casos se comprueba que el p-valor es menor que 0.05, por lo que las muestras no tienen distribución Normal, por lo que no se pueden aplicar contrastes paramétricos sino contrastes no paramétricos, en este caso el de Mann-Whitney-Wilcoxon al ser muestras independientes.

a) ¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

$$H_0: M_{Mañana} = M_{Tarde} \text{ vs } H_A: M_{Mañana} \neq M_{Tarde}$$

```
> wilcox.test(nota.mañana, nota.tarde, paired=FALSE, alternative="two.sided", conf.level=0.95, exact=FALSE)
      Wilcoxon rank sum test with continuity correction
data:  nota.mañana and nota.tarde
W = 956, p-value = 0.001977
```

Como el p-valor es menor que 0.05, se rechaza la Hipótesis H_0 y se acepta H_A, por lo que SÍ se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde.

b) ¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

$$H_0: M_{Mañana} = M_{Tarde} \text{ vs } H_A: M_{Mañana} > M_{Tarde}$$

```
> wilcox.test(nota.mañana, nota.tarde, paired=FALSE, alternative="greater", conf.level=0.95, exact=FALSE)
      Wilcoxon rank sum test with continuity correction
data:  nota.mañana and nota.tarde
W = 956, p-value = 0.0009883
alternative hypothesis: true location shift is greater than 0
```

Como el p-valor es menor que 0.05, se rechaza la Hipótesis H_0 y se acepta H_A, por lo que SÍ se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde.

c) ¿Se puede afirmar con una confianza del 95% que la mediana de las notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

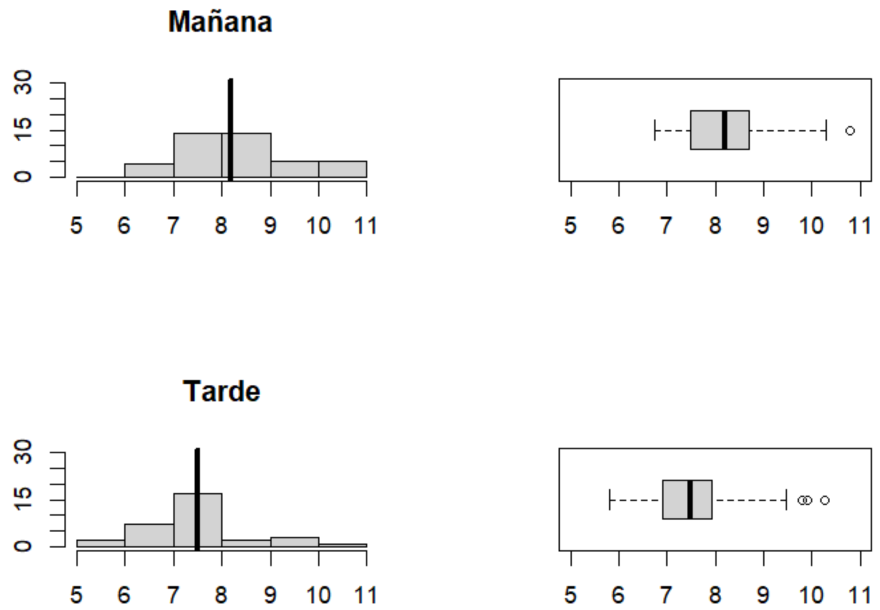
$$H_0: M_{Mañana} = M_{Tarde} \text{ vs } H_A: M_{Mañana} < M_{Tarde}$$

```
>   wilcox.test(nota.mañana,   nota.tarde,   paired=FALSE,   alternative="less", conf.level=0.95, exact=FALSE)
      Wilcoxon rank sum test with continuity correction
data:  nota.mañana and nota.tarde
W = 956, p-value = 0.999
alternative hypothesis: true location shift is less than 0
```

Como el p-valor es mayor que 0.05, no se puede rechazar la Hipótesis H_0 y, por tanto, se rechaza H_A, por lo que NO se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde.

Para finalizar se podrían representar diagramas para hacer una comprobación visual.

```
> par(mfrow=c(2,2))
> L=c(5,6,7,8,9,10,11)
> hist(nota.mañana, breaks=L, ylim=c(0,30), main="Mañana", xlab="", ylab = "")
> abline(v=median(nota.mañana), lwd=3)
> boxplot(nota.mañana,horizontal = TRUE, ylim=c(5,11))
> hist(nota.tarde, breaks=L, ylim=c(0,30), main="Tarde", xlab = "", ylab = "")
> abline(v=median(nota.tarde), lwd=3)
> boxplot(nota.tarde,horizontal = TRUE, ylim=c(5,11))
```



5. Ejercicio propuesto

1) Mediante una encuesta, se sabe el tiempo en minutos del viaje a la Escuela Politécnica de 74 estudiantes de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá, de una población de 108 matriculados. Responder a las siguiente pregunta:
   a. ¿Se puede afirmar con una confianza del 95% que la mediana del tiempo de viaje de toda la población es inferior a 70 minutos?