

**UNIVERSIDAD DE EL SALVADOR**  
**FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE**  
**DEPARTAMENTO DE MATEMÁTICAS**

**LICENCIATURA EN ESTADÍSTICA**



**PRACTICAS REALIZADAS EN EL SOFTWARE R**

**DOCENTE:**  
**LICENCIADO. JAIME ISAAC PEÑA**

**PRESENTADO POR:**  
**MORIS SALVADOR HENRIQUEZ LIMA**

**Viernes 23 de Septiembre del 2022**



## Índice

<b>1. CONTRASTES DE HIPÓTESIS</b>	<b>2</b>
1.1. Supuesto Práctico 1	2
1.2. Supuesto Práctico 2	5
1.3. Supuesto Práctico 3	7
1.4. Supuesto Práctico 4	8
1.5. Supuesto Práctico 5	9
1.6. Supuesto Práctico 6	10
1.7. Supuesto Práctico 7	12
1.8. Supuesto Práctico 8	13
1.9. Supuesto Práctico 9	14
1.10. Supuesto Práctico 10	15
1.11. Supuesto Práctico 11	16
1.12. Supuesto Práctico 12	17
1.13. Supuesto Práctico 13	18
1.14. Supuesto Práctico 14	19
1.15. Supuesto Práctico 15	21
1.16. Supuesto Práctico 16	21
1.17. Supuesto Práctico 17	22
1.18. Ejercicios Guiados	24
1.18.1. Ejercicio Guiado 1	24
1.18.2. Ejercicio Guiado 2	26
1.18.3. Ejercicio Guiado 3	27
1.18.4. Ejercicio Guiado 4	29
1.18.5. Ejercicio Guiado 5	30
1.19. Ejercicios Propuestos	31
1.19.1. Ejercicio Propuesto 1	31
1.19.2. Ejercicio Propuesto 2	33
1.19.3. Ejercicio Propuesto 3	35
1.19.4. Ejercicio Propuesto 4	36
1.19.5. Ejercicio Propuesto 5	37
1.19.6. Ejercicio Propuesto 6	38

## 1. CONTRASTES DE HIPÓTESIS

### 1.1. Supuesto Práctico 1

Con el fin de estudiar el número medio de flexiones continuadas que pueden realizar sus alumnos, un profesor de educación física somete a 75 de ellos, elegidos aleatoriamente, a una prueba. El número de flexiones realizado por cada alumno, así como su sexo y si realizan o no deporte fuera del horario escolar se muestran en el fichero **Flexiones.txt**.

	Flexiones	Sexo	Deporte
1	60	H	0
2	41	H	0
3	53	M	1
4	53	M	0
5	41	H	0
6	56	H	0
7	50	H	0
8	53	M	1
9	50	M	1
10	48	M	0
11	50	M	1
12	48	M	1
13	56	H	0
14	52	M	1

Se sabe que el número de flexiones se distribuye según una Normal de varianza poblacional 7.5.  
¿Puede asumirse, considerando un nivel de significación del 5 %, que el número medio de flexiones que realizan los alumnos es de 55?

El contraste de hipótesis asociado a este ejercicio es:

$$H_0 \equiv \mu = 55$$

$$H_1 \equiv \mu \neq 55$$

### Solución del Supuesto Práctico 1

En primer lugar debemos importar en **R** los datos que contienen el número de flexiones realizadas por cada alumno. Para ello, utilizamos la orden **read.table**.

El siguiente Proceso `setwd()` se omite debido a que el archivo en el cual se trabajara ya esta en nuestro directorio de trabajo:

Proceso de Lectura de la Base de Datos a utilizar:

```
> datos<- read.table("Flexiones.txt" , header = TRUE)
> datos
```



	Flexiones	Sexo	Deporte
1	60	H	0
2	41	H	0
3	53	M	1
4	53	M	0
5	41	H	0
6	56	H	0
7	50	H	0
8	53	M	1
9	50	M	1
10	48	M	0
11	50	M	1
12	48	M	1
13	56	H	0
14	52	M	1
15	54	M	0
16	50	H	1
17	50	H	0
18	54	H	0
19	52	H	1
20	48	H	0
21	48	H	1
22	35	M	1
23	50	M	1
24	41	M	1
25	56	M	1
26	52	M	1
27	56	M	0
28	54	H	1
29	53	H	0
30	53	M	0
31	53	H	0
32	41	M	1
33	48	M	0
34	50	H	1
35	50	M	1
36	52	H	0
37	53	M	0
38	35	H	0
39	35	H	0
40	54	M	0
41	46	M	1
42	48	H	0
43	50	M	0
44	48	H	0
45	41	M	0
46	48	M	1
47	60	H	1
48	53	M	0
49	54	M	1
50	56	H	1
51	50	H	1



52	41	H	0
53	60	M	1
54	60	M	1
55	54	H	0
56	54	H	0
57	53	H	0
58	35	M	0
59	54	H	0
60	48	M	0
61	50	H	0
62	54	H	0
63	54	H	0
64	53	H	0
65	52	H	0
66	50	H	0
67	52	H	0
68	48	H	1
69	46	H	1
70	53	H	0
71	50	H	0
72	35	H	0
73	50	H	1
74	60	M	1
75	50	H	0

Una vez hecho esto, introducimos en **R** el nivel de significación que proporciona el enunciado.

```
> alpha <- 0.05  
> alpha
```

```
[1] 0.05
```

A continuación, calculamos el valor del estadístico de contraste:

```
> alpha <- 0.05  
> alpha
```

```
[1] 0.05
```

```
> media <- mean(datos$Flexiones)  
> media
```

```
[1] 50.10667
```

```
> mu_0 <- 55  
> mu_0
```

```
[1] 55
```

```
> varianza <- 7.5  
> varianza
```

```
[1] 7.5
```

```
> n <- nrow(datos)  
> n
```



```
[1] 75
```

```
> Z <- (media - mu_0)/(sqrt(varianza)/sqrt(n))  
> Z
```

```
[1] -15.47408
```

Y también el valor crítico, que en este caso coincide con  $z_{1-\frac{\alpha}{2}}$ , el cuantil  $1 - \frac{\alpha}{2}$  de una distribución normal de media 0 y varianza 1.

```
> cuantil<- qnorm(1 - alpha/2)  
> cuantil
```

```
[1] 1.959964
```

Como el valor absoluto del estadístico de contraste (**15.47408**) es mayor que el valor crítico (**1.959964**), en este caso se rechaza la hipótesis nula en favor de la hipótesis alternativa. Es decir, no puede asumirse que el número medio de flexiones que realizan los alumnos es de 55.

## 1.2. Supuesto Práctico 2

Considerando nuevamente el conjunto de datos que se ha presentado en el Supuesto práctico1, relativo al número de flexiones y el sexo de los alumnos. Contrastar a un nivel de significación del 2% la hipótesis de que el número medio de flexiones realizada por los alumnos es de 50. Suponer en este caso que el número de flexiones se distribuye según una normal de varianza desconocida. El fichero es **Flexiones.txt**.

	Flexiones	Sexo	Deporte
1	60	H	0
2	41	H	0
3	53	M	1
4	53	M	0
5	41	H	0
6	56	H	0
7	50	H	0
8	53	M	1
9	50	M	1
10	48	M	0
11	50	M	1
12	48	M	1
13	56	H	0
14	52	M	1



### Solución del Supuesto Práctico 2

En primer lugar, planteamos el contraste de hipótesis asociado a este supuesto:

$$H_0 \equiv \mu = 50$$

$$H_1 \equiv \mu \neq 50$$

Partiendo del conjunto de datos del supuesto 1, el cual se había importado y almacenado en una variable de nombre **datos**.

En primer lugar, como hicimos anteriormente, debemos importar en R los datos que contienen el número de flexiones realizadas por cada alumno. Para ello, utilizamos la orden **read.table**.

```
> getwd()

[1] "C:/Moris_Henriquez/Practicas_R_Sweave_2022"

> datos<- read.table("Flexiones.txt", header = TRUE)
> head(datos)
```

	Flexiones	Sexo	Deporte
1	60	H	0
2	41	H	0
3	53	M	1
4	53	M	0
5	41	H	0
6	56	H	0

Una vez importado los datos, vamos a resolver este contraste. Para ello, basta con realizar la siguiente llamada a la función **t.test**:

```
> t.test(datos$Flexiones, alternative = "two.sided", mu = 50, conf.level = 0.98)
```

One Sample t-test

```
data: datos$Flexiones
t = 0.15451, df = 74, p-value = 0.8776
alternative hypothesis: true mean is not equal to 50
98 percent confidence interval:
 48.46512 51.74822
sample estimates:
mean of x
 50.10667
```

Entre la información que devuelve la función **t.test**, encontramos la relativa al intervalo de confianza, que se estudió en la práctica 5. En esta práctica nos centraremos en la referente al contraste de hipótesis.

**t = 0.15451, df = 74, p-value = 0.8776**  
**alternative hypothesis: true mean is not equal to 50**

En primer lugar, aparece el valor del estadístico de contraste (**0.15451**) junto a los grados de libertad de la distribución **t de Student (74)** que sigue dicho estadístico de contraste. A continuación, encontramos el p-valor, que en este caso es **0.8776**. Por último, el programa nos recuerda que la hipótesis alternativa que se está contrastando es del tipo  $\neq$ .



Teniendo en cuenta que el p-valor (**0.8776**) es superior al nivel de significación (**0.02**) en este ejemplo no podemos rechazar la hipótesis nula, por lo que podemos asumir que el número medio de flexiones que realizan los alumnos es de 50.

### 1.3. Supuesto Práctico 3

Considerando nuevamente el conjunto de datos que se ha presentado en el Supuesto práctico1, relativo al número de flexiones y el sexo de los alumnos. Contrastar a un nivel de confianza del 95 %, si la proporción de alumnos varones es mayor o igual que 0.5 frente a que dicha proporción es menor. El fichero es Flexiones.txt.

El contraste que debemos resolver es:

$$H_0 \equiv \pi_H \geq 50$$

$$H_1 \equiv \pi_H < 50$$

#### Solución del Supuesto Práctico 3

Para realizar la llamada a la función **prop.test** necesitamos conocer el número de alumnos varones y el número total de estudiantes en la muestra. Para ello utilizamos la función de **R** **table**.

En primer lugar, como hicimos anteriormente, debemos importar en **R** los datos que contienen el número de flexiones realizadas por cada alumno. Para ello, utilizamos la orden **read.table**.

```
> getwd()

[1] "C:/Moris_Henriquez/Practicas_R_Sweave_2022"

> datos<- read.table("Flexiones.txt", header = TRUE)
> head(datos)
```

	Flexiones	Sexo	Deporte
1	60	H	0
2	41	H	0
3	53	M	1
4	53	M	0
5	41	H	0
6	56	H	0

Una vez importado los datos, utilizamos la función de R **table** como hemos dicho anteriormente:

```
> table(datos$Sexo)

H  M
43 32
```

De los 75 estudiantes que conforman la muestra, 43 son chicos. Por lo que la llamada a **prop.test** sería la siguiente:

```
> prop.test(43, 75, p = 0.5, alternative = "less", conf.level = 0.95)
```





#### 1-sample proportions test with continuity correction

```
data: 43 out of 75, null probability 0.5
X-squared = 1.3333, df = 1, p-value = 0.8759
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.6693525
sample estimates:
      p
0.5733333
```

De nuevo, los resultados de la función incluyen información sobre el intervalo de confianza y sobre el contraste de hipótesis. Nos centraremos en esta última.

**X-squared = 1.3333, df = 1, p-value = 0.8759**  
**alternativehypothesis: true p islessthan 0.5**

El valor del estadístico de contraste es **1.3333**, con un p-valor de **0.8759**. Como el p-valor es mayor que el nivel de significación, que es 0.05, no rechazamos la hipótesis de que la proporción de alumnos es mayor o igual que 0.5.

### 1.4. Supuesto Práctico 4

Continuando con los datos relativos a las flexiones realizadas por un grupo de estudiantes y asumiendo que las flexiones que realizan los chicos y las que realizan las chicas se distribuyen según sendas distribuciones normales con medias y varianzas desconocidas, contrastar a un nivel de significación del 5% si las varianzas poblacionales de ambas distribuciones pueden asumirse iguales.

El contraste de hipótesis que debemos resolver es:

$$H_0 \equiv \sigma_H^2 = \sigma_M^2$$
$$H_1 \equiv \sigma_H^2 \neq \sigma_M^2$$

Donde  $\sigma_H^2$  representa la varianza del número de flexiones realizadas por los chicos  $\sigma_M^2$  y representa la varianza del número de flexiones realizadas por las chicas.

#### Solución del Supuesto Práctico 4

Lo primero que tenemos que hacer para aplicar la función **var.test** es separar en dos variables los datos relativos a las flexiones realizadas por los chicos y por las chicas.

```
> Flexiones.chicos<- datos$Flexiones[datos$Sexo == "H"]
> Flexiones.chicos

[1] 60 41 41 56 50 56 50 50 54 52 48 48 54 53 53 50 52 35 35 48 48 60 56 50 41
[26] 54 54 53 54 50 54 54 53 52 50 52 48 46 53 50 35 50 50

> Flexiones.chicas<- datos$Flexiones[datos$Sexo == "M"]
> Flexiones.chicas
```



```
[1] 53 53 53 50 48 50 48 52 54 35 50 41 56 52 56 53 41 48 50 53 54 46 50 41 48  
[26] 53 54 60 60 35 48 60
```

A continuación, utilizamos la función `var.test`.

```
> var.test(Flexiones.chicos, Flexiones.chicas, alternative = "two.sided", conf.level = 0.95  
+ )
```

F test to compare two variances

```
data: Flexiones.chicos and Flexiones.chicas  
F = 0.87506, num df = 42, denom df = 31, p-value = 0.679  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.4415454 1.6765483  
sample estimates:  
ratio of variances  
 0.8750585
```

Analizando la información relativa al contraste de hipótesis que se incluye en la salida de `var.test`, vemos que el valor del estadístico de contraste es **0.87506**. La distribución F de Snedecor que sigue el estadístico de contraste tiene 42 grados de libertad en el numerador y 31 en el denominador. El p-valor asociado al contraste es **0.679**. Como este valor es superior al nivel de significación (**que para este ejemplo es 0.05**), no podemos rechazar la hipótesis nula que hemos planteado.

Es decir, se puede considerar que la varianza del número de flexiones realizadas por chicos y la varianza del número de flexiones realizadas por chicas son iguales.

### 1.5. Supuesto Práctico 5

En vista de los resultados obtenidos en el Supuesto Práctico 4, y suponiendo que el número de flexiones que realizan los alumnos y las alumnas se distribuyen de acuerdo a variables normales de medias y varianzas desconocidas, ¿puede suponerse, a un nivel de significación del 5%, que el número medio de flexiones que realizan los chicos y las chicas es igual?

El contraste que debemos resolver en esta ocasión es:

$$\begin{aligned} H_0 &\equiv \mu_H = \mu_M & H_0 &\equiv \mu_H - \mu_M = 0 \\ H_1 &\equiv \mu_H \neq \mu_M & H_1 &\equiv \mu_H - \mu_M \neq 0 \end{aligned}$$

En ambos casos  $\mu_H$ , representa la media poblacional del número de flexiones realizadas por chicos y  $\mu_M$  es la media poblacional del número de flexiones realizadas por las chicas.

#### Solución del Supuesto Práctico 5

Dado que en el Supuesto práctico 4 se concluyó la igualdad de las varianzas del número de flexiones que hacen chicos y chicas, debemos establecer a **TRUE** el valor del parámetro `var.equal` cuando realicemos la llamada a la función `t.test`.

```
> datos<- read.table("Flexiones.txt", header = TRUE)  
> head(datos)
```



```

Flexiones Sexo Deporte
1      60    H      0
2      41    H      0
3      53    M      1
4      53    M      0
5      41    H      0
6      56    H      0

> Flexiones.chicos<- datos$Flexiones[datos$Sexo == "H"]
> Flexiones.chicos

[1] 60 41 41 56 50 56 50 50 54 52 48 48 54 53 53 50 52 35 35 48 48 60 56 50 41
[26] 54 54 53 54 50 54 54 53 52 50 52 48 46 53 50 35 50 50

> Flexiones.chicas<- datos$Flexiones[datos$Sexo == "M"]
> Flexiones.chicas

[1] 53 53 53 50 48 50 48 52 54 35 50 41 56 52 56 53 41 48 50 53 54 46 50 41 48
[26] 53 54 60 60 35 48 60

> t.test(Flexiones.chicos, Flexiones.chicas, alternative = "two.sided", mu = 0, var.equal =
+ TRUE)

```

#### Two Sample t-test

```

data: Flexiones.chicos and Flexiones.chicas
t = -0.06154, df = 73, p-value = 0.9511
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.887271  2.714306
sample estimates:
mean of x mean of y
 50.06977  50.15625

```

Entre la información sobre el contraste de hipótesis que se incluye entre los resultados:

```

data: Flexiones.chicos and Flexiones.chicas
t = -0.06154, df = 73, p-value = 0.9511
alternativehypothesis: true difference in meansisnotequal to 0

```

Se incluye el valor del estadístico de contraste (**-0.06154**), los grados de libertad de la distribución t de Student que sigue el estadístico de contraste (**73**) y el **p-valor (0.9511)**. Como el p-valor es mayor que el nivel de significación fijado (**0.05**), no rechazamos la hipótesis nula del contraste.

## 1.6. Supuesto Práctico 6

Para estudiar los efectos de un programa de control de peso, el profesor de educación física selecciona aleatoriamente a 6 alumnos y se les toma nota de sus pesos antes y después de pasar por el programa.

<i>Antes</i>	72.0	73.5	70.0	71.5	76.0	80.5
<i>Después</i>	73.0	74.5	74.0	74.5	75.0	80.2



¿Puede suponerse, a un nivel de significación del 5 %, que el programa para el control de peso es efectivo? O, dicho de otra forma, ¿el peso medio de los alumnos antes de someterse al programa es igual al peso medio tras el programa?

El contraste de hipótesis que debemos resolver es el siguiente:

$$\begin{aligned}H_0 &\equiv \mu_a = \mu_d \\ H_1 &\equiv \mu_a \neq \mu_d\end{aligned}$$

Donde  $\mu_a$  y  $\mu_d$  hacen referencia al peso medio poblacional antes y después de pasar por el programa de control de peso, respectivamente.

Como puede observarse, los datos vienen por parejas: peso antes y después, dos datos por individuo. Parece lógico que los datos se encuentren relacionados entre sí.

### Solución del Supuesto Práctico 6

En primer lugar, vamos a introducir los datos en **R**.

```
> Antes <- c(72.0, 73.5, 70.0, 71.5, 76.0, 80.5)
> Antes
```

```
[1] 72.0 73.5 70.0 71.5 76.0 80.5
```

```
> Despues<- c(73.0, 74.5, 74.0, 74.5, 75.0, 82.0)
> Despues
```

```
[1] 73.0 74.5 74.0 74.5 75.0 82.0
```

A partir de estos datos, vamos a aplicar la función **t.test**, para resolver el contraste de hipótesis que hemos planteado.

```
> t.test(Antes, Despues, alternative = "two.sided", mu = 0, paired = TRUE)
```

Paired t-test

```
data: Antes and Despues
t = -2.2238, df = 5, p-value = 0.07676
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.4135884  0.2469217
sample estimates:
mean of the differences
 -1.583333
```

Según los datos que se incluyen en la salida de la función, el estadístico de contraste toma un valor de **-2.2238** y sigue una distribución t de Student con 5 grados de libertad. El p-valor asociado al contraste es **0.07676**. Como este **p-valor** es mayor que **0.05**, que es el nivel de significación del contraste, podemos afirmar que la diferencia entre los pesos medios de los alumnos antes y después de seguir el programa de control de peso es nula o, equivalentemente, que ambos pesos medios pueden suponerse iguales.



## 1.7. Supuesto Práctico 7

Retomando el conjunto de datos relativo a las flexiones que realizan un grupo de estudiantes, contrastar, a un nivel de significación del 8 % si la proporción de alumnos y de alumnas que practican deporte pueden considerarse iguales.

El contraste que vamos a resolver es:

$$\begin{aligned}H_0 &\equiv \pi_H - \pi_M = 0 \\H_1 &\equiv \pi_H - \pi_M \neq 0\end{aligned}$$

Donde  $\pi_H$  y  $\pi_M$  representan la proporciones de chicos y chicas que practican deporte, respectivamente.

### Solución del Supuesto Práctico 7

En primer lugar, utilicemos el comando `table` para determinar cuántos chicos y cuantas chicas practican deporte.

```
> table(datos$Sexo, datos$Deporte)

  0  1
H 32 11
M 13 19
```

En total, 11 de los 43 y 19 de las 32 chicas muestreados practican deporte fuera del horario escolar. Vamos a crear dos vectores con esta información: en uno indicaremos el total de chicos y chicas que practican deporte y en el otro el total de chicos y chicas en la muestra.

```
> vector_Deporte<- c(11, 19)
> vector_Deporte

[1] 11 19

> vector_Sexo<- c(43, 32)
> vector_Sexo

[1] 43 32
```

Es muy importante que los valores se introduzcan en el mismo orden en los dos vectores. Ahora ya podemos utilizar la función `prop.test` utilizando estos dos vectores como argumentos.

```
> prop.test(vector_Deporte, vector_Sexo, alternative = "two.sided", conf.level = 0.92)

2-sample test for equality of proportions with continuity correction

data:  vector_Deporte out of vector_Sexo
X-squared = 7.3787, df = 1, p-value = 0.0066
alternative hypothesis: two.sided
92 percent confidence interval:
 -0.5566881 -0.1191840
sample estimates:
 prop 1  prop 2 
0.255814 0.593750
```

Según la salida de la función `prop.test`, el **p-valor** asociado al contraste es **0.0066**, que al ser menor que el nivel de significación (**0.08**), nos lleva a concluir que las proporciones de chicos y chicas que hacen deporte no coinciden.

## 1.8. Supuesto Práctico 8

La directora de un hospital quiere comprobar si los ingresos en el hospital se producen en la misma proporción durante todos los días de la semana. Para ello, se anota el número de ingresos durante una semana cualquiera. Los datos se recogen en la siguiente tabla:

<i>Día de la semana</i>	<i>Número de ingresos</i>
<i>Lunes</i>	78
<i>Martes</i>	90
<i>Miércoles</i>	94
<i>Jueves</i>	89
<i>Viernes</i>	110
<i>Sábado</i>	84
<i>Domingo</i>	44

Contrastar, a un nivel de significación del 5 %, si la hipótesis de la directora del hospital puede suponerse cierta. ¿Puede asumirse que las proporciones de ingresos de lunes a domingo son (0.15, 0.15, 0.15, 0.15, 0.20, 0.15, 0.05)?

### Solución del Supuesto Práctico 8

En primer lugar vamos a introducir los datos en R.

```
> frecuencias <- c(78, 90, 94, 89, 110, 84, 44)
> frecuencias
```

```
[1] 78 90 94 89 110 84 44
```

El contraste que se debe resolver es:

$H_0 \equiv$  Los ingresos en el hospital se producen en la misma proporción todos los días de la semana

$H_1 \equiv$  Los ingresos en el hospital no se producen en la misma proporción todos los días de la semana

Para resolver este contraste se usa la función **chisq.test**.

```
> chisq.test(frecuencias)
```

Chi-squared test for given probabilities

```
data: frecuencias
```

```
X-squared = 29.389, df = 6, p-value = 5.135e-05
```

El estadístico de contraste, que sigue una distribución chi-cuadrado, toma el valor **29.389**. Los grados de libertad de la distribución chi-cuadrado para este ejemplo son 6. El p-valor asociado al contraste es menor que 0.05 por lo que, considerando un nivel de significación del 5 %, se rechaza la hipótesis nula. Es decir, se concluye que los ingresos hospitalarios no se producen en la misma proporción todos los días de la semana.

Para comprobar si el vector **(0.15, 0.15, 0.15, 0.15, 0.20, 0.15, 0.05)** puede considerarse como el vector de proporciones de ingresos hospitalarios durante los 7 días de la semana, creamos un vector en R que recoja estos valores:



```
> proporciones <- c(0.15, 0.15, 0.15, 0.15, 0.20, 0.15, 0.05)
> proporciones
```

```
[1] 0.15 0.15 0.15 0.15 0.20 0.15 0.05
```

Volvemos a llamar a la función **chisq.test** incluyendo como argumento el vector que acabamos de definir.

```
> chisq.test(frecuencias, p = proporciones)
```

Chi-squared test for given probabilities

data: frecuencias

X-squared = 9.5286, df = 6, p-value = 0.146

En este caso, el valor del estadístico de contraste es **9.5286**. El p-valor asociado es **0.146** que, al ser superior a 0.05, nos indica que no se puede rechazar la hipótesis nula. Esto equivale a decir que, a un nivel de significación del 5%, puede suponerse que los ingresos hospitalarios se producen según los valores que se recogen en el vector proporciones.

## 1.9. Supuesto Práctico 9

Lanzamos un dado 720 veces y obtenemos los resultados que se muestran en la tabla.

$x_i$	1	2	3	4	5	6
$n_i$	116	120	115	120	125	124

Contrastar la hipótesis de que el dado está bien construido.

### Solución del Supuesto Práctico 9

Comencemos introduciendo en R las frecuencias con las que aparecen los valores del dado.

```
> frecuencias <- c(116, 120, 115, 120, 125, 124)
> frecuencias
```

```
[1] 116 120 115 120 125 124
```

Que el dado esté bien construido equivale a decir que todos sus valores aparecen en la misma proporción. Por tanto, el contraste de hipótesis que se debe resolver es el siguiente:

$H_0 \equiv$  Los valores del dado aparecen en la misma proporción

$H_1 \equiv$  Los valores del dado no aparecen en la misma proporción

Para resolver este contraste de hipótesis se utiliza la función **chisq.test**, que recibe como argumento el vector de frecuencias:

```
> chisq.test(frecuencias)
```

Chi-squared test for given probabilities

data: frecuencias

X-squared = 0.68333, df = 5, p-value = 0.9839



El valor del estadístico de contraste es **0.68333** y el p-valor asociado es igual a **0.984**. Como este p-valor es superior a 0.05 no se puede rechazar la hipótesis nula por lo que, a un nivel de significación del 5 %, concluimos que todos los valores del dado aparecen en la misma proporción. Dicho de otra forma, el dado está bien construido.

### 1.10. Supuesto Práctico 10

a siguiente tabla muestra información sobre el número de ejemplares de 7 especies de peces avistados aguas arriba y aguas abajo en un río.

	<i>Zona</i>	
	<i>Aguas arriba</i>	<i>Aguas abajo</i>
<i>EspecieA</i>	37	19
<i>EspecieB</i>	12	10
<i>EspecieC</i>	10	7
<i>EspecieD</i>	18	20
<i>EspecieE</i>	11	8
<i>EspecieF</i>	16	12
<i>EspecieG</i>	59	24

contrastar, a un nivel de significación del 5 %, si la especie de pez y la zona de avistamiento pueden considerarse variables independientes.

#### Solución del Supuesto Práctico 10

En primer lugar, introduzcamos en R los datos que proporciona el enunciado y construyamos la tabla de contingencia.

```
> frecuencias <- c(37, 19, 12, 10, 10, 7, 18, 20, 11, 8, 16, 12, 59, 24)
> frecuencias

[1] 37 19 12 10 10 7 18 20 11 8 16 12 59 24

> tabla_conting <- matrix (frecuencias, 7, 2, byrow = TRUE, dimnames = list(c("A", "B", "C"
+ ", "D", "E", "F", "G"), c("Aguas_Arriba", "Aguas_abajo")))
> tabla_conting

      Aguas_Arriba Aguas_abajo
A              37          19
B              12          10
C\n           10           7
D              18          20
E              11           8
F              16          12
G              59          24
```

El contraste de hipótesis que se debe resolver es:

$$H_0 \equiv \text{La especie y la zona de avistamiento son independientes}$$

$$H_1 \equiv \text{La especie y la zona de avistamiento no son independientes}$$





A continuación, usaremos la función **chisq.test** (sin aplicar la corrección por continuidad) para resolver el contraste.

```
> chisq.test(tabla_conting, correct = FALSE)
```

Pearson's Chi-squared test

```
data: tabla_conting
```

```
X-squared = 7.7604, df = 6, p-value = 0.2562
```

El estadístico de contraste, que sigue una distribución chi-cuadrado con 6 grados de libertad, toma el valor **7.7604**. El p-valor asociado al contraste es **0.2562**. Como este p-valor es mayor que 0.05, no podemos rechazar la hipótesis nula por lo que concluimos que la especie y la zona de avistamiento son variables independientes. Esto es, para cada especie, se observan el mismo número de peces aguas arriba y aguas abajo en el río.

### 1.11. Supuesto Práctico 11

Se realiza una investigación para determinar si hay alguna asociación entre el peso de un estudiante y un éxito precoz en la escuela. Se selecciona una muestra de 50 estudiantes y se clasifica a cada uno según dos criterios, el peso y el éxito en la escuela. Los datos se muestran en la tabla adjunta

<i>Éxito/Sobrepeso</i>	<i>SI</i>	<i>NO</i>
<i>SI</i>	162	263
<i>NO</i>	38	37

Contrastar, a un nivel de significación del 5 %, si las dos variables estudiadas están relacionadas o si, por el contrario, son independientes.

#### Solución del Supuesto Práctico 11

Introducimos los datos en **R**

```
> frecuencias <- c(162, 263, 38, 37)
```

```
> frecuencias
```

```
[1] 162 263 38 37
```

```
> tabla_conting <- matrix (frecuencias, 2, 2, byrow = TRUE, dimnames = list(c("Exito = Sí",  
+ "Exito = No"), c("Sobrepeso = Sí", "Sobrepeso = No")))
```

```
> tabla_conting
```

```
      Sobrepeso = Sí Sobrepeso = No
Exito = Sí      162      263
Exito = No       38       37
```

El contraste de hipótesis que se debe resolver es:

$$H_0 \equiv \text{El éxito en la escuela y el sobrepeso son independientes}$$

$$H_1 \equiv \text{El éxito en la escuela y el sobrepeso no son independientes}$$



Vamos a resolver el contraste usando la función **chisq.test** (sin aplicar la corrección por continuidad).

```
> chisq.test(tabla_conting, correct = FALSE)
```

Pearson's Chi-squared test

```
data:  tabla_conting  
X-squared = 4.183, df = 1, p-value = 0.04083
```

El p-valor asociado a este contraste es **0.04083**. Como este **p-valor** es menor que **0.05**, se rechaza la hipótesis nula del contraste, por lo que concluimos que el éxito escolar y el sobrepeso son variables dependientes. Esto es, los valores de una dependen de los valores de la otra.

### 1.12. Supuesto Práctico 12

Se quiere comprobar si la proporción de hombres y mujeres en un municipio andaluz es la misma o no. Para ello, se selecciona una muestra aleatoria de habitantes del municipio, de los cuales 258 son hombres y 216 son mujeres. A un nivel de significación del 5 %, ¿puede asumirse cierta la igualdad en el número de hombres y mujeres?

#### Solución del Supuesto Práctico 12

Comencemos planteando las hipótesis del contraste. En este caso, se quiere probar la igualdad de hombres y de mujeres en el municipio. Para ello, es posible plantear el contraste de hipótesis de dos formas distintas. Por un lado, se puede contrastar si la proporción de hombres es de 0.5 (en cuyo caso la proporción de mujeres será también 0.5 y habrá equidad entre ambos géneros) frente a que esta proporción es distinta de 0.5. Pero, alternativamente, se puede contrastar si la proporción de mujeres es de 0.5 (lo que implica que la proporción de hombre será, igualmente, de 0.5 y habrá equidad entre géneros) frente a que esta proporción es distinta de 0.5.

En cualquier caso, el contraste a resolver es:

$$\begin{aligned}H_0 &\equiv p = 0,5 \\ H_1 &\equiv p \neq 0,5\end{aligned}$$

Donde  $p$  representa la proporción de hombres (o de mujeres, dependiendo de la forma de resolver el contraste que se siga) en la población.

Utilicemos la función **binom.test** para resolver el contraste.

```
> binom.test(258, n = 474, p = 0.5, alternative = "two.sided", conf.level = 0.95)
```

Exact binomial test

```
data:  258 and 474  
number of successes = 258, number of trials = 474, p-value = 0.05956  
alternative hypothesis: true probability of success is not equal to 0.5  
95 percent confidence interval:  
 0.4982562 0.5897954  
sample estimates:  
probability of success  
 0.5443038
```



En la salida aparecen los datos de entrada que se han usado para resolver el contraste (258 hombres de 474 habitantes muestreados) así como el tipo de la hipótesis alternativa (distinto de) y la proporción que se ha usado como referente para el contraste (0.5).

También aparece un p-valor, que es el que nos ayuda a resolver el contraste. En este caso, el p-valor es **0.05956**. Como es mayor que 0.05, no podemos rechazar la hipótesis nula, por lo que podemos asumir que la proporción de hombres en la población es de 0.5. Consecuentemente, la proporción de mujeres también puede considerarse igual a 0.5 y puede concluirse que el número de hombres y mujeres en el municipio es el mismo.

Por último, en la salida se incluye un intervalo de confianza al nivel de confianza indicado en la llamada a **binom.test** (95 % en nuestro caso), para la proporción de hombres en el municipio. Este intervalo es **(0.4982, 0.5897)**. Como era de esperar, la proporción de referencia pertenece al intervalo calculado.

Si se hubiese optado por considerar p como la proporción de mujeres en el municipio y resolver el contraste a partir de esta proporción se llegaría a la misma conclusión, tal y como se muestra a continuación.

```
> binom.test(216, n = 474, p = 0.5, alternative = "two.sided", conf.level = 0.95)
```

```
Exact binomial test
```

```
data: 216 and 474
number of successes = 216, number of trials = 474, p-value = 0.05956
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4102046 0.5017438
sample estimates:
probability of success
 0.4556962
```

### 1.13. Supuesto Práctico 13

Entre los pacientes con cáncer de pulmón, el 90 % o más muere generalmente en el espacio de tres años. Como resultado de nuevas formas de tratamiento, se cree que esta tasa se ha reducido. En un reciente estudio sobre 150 paciente diagnosticados de cáncer de pulmón, 128 murieron en el espacio de tres años. ¿Se puede afirmar que realmente ha disminuido la tasa de mortalidad?

#### Solución del Supuesto Práctico 13

En primer lugar, vamos a plantear las hipótesis del contraste:

$$H_0 \equiv p \geq 0.9 \text{ el tratamiento no es efectivo}$$
$$H_1 \equiv p < 0.9 \text{ el tratamiento es efectivo}$$

A continuación, utilizaremos la función **binom.test** para resolver el contraste. Teniendo en cuenta el número de pacientes de la muestra que fallecieron (128), el número de pacientes totales en la muestra (150), la proporción que se quiere contrastar (0.9) y la forma de la hipótesis alternativa ("menor que").

```
> binom.test(128, 150, p = 0.9, alternative = "less", conf.level = 0.95)
```

```
Exact binomial test
```

```
data: 128 and 150
number of successes = 128, number of trials = 150, p-value = 0.04396
alternative hypothesis: true probability of success is less than 0.9
```



95 percent confidence interval:

0.0000000 0.8985727

sample estimates:

probability of success

0.8533333

El p-valor asociado al contraste es **0.04396**. De manera que, considerando un nivel de significación del 5 %, rechazamos la hipótesis nula, por lo que se puede concluir que la proporción de pacientes que fallecieron en el espacio de tres años es inferior a 0.9 y, consecuentemente, que el tratamiento es efectivo.

### 1.14. Supuesto Práctico 14

Se realiza un estudio sobre el tiempo en horas de un tipo determinado de escáner antes de la primera avería. Se ha observado una muestra de 10 escáner y se ha anotado el tiempo de funcionamiento en horas: 18.21; 2.36; 17.3; 16.6; 4.70; 3.63; 15.56; 7.35; 9.78; 14.69. A un nivel de significación del 5 %, ¿se puede considerar aleatoriedad en la muestra?

#### Solución del Supuesto Práctico 14

Formulamos el contraste que debemos resolver:

$H_0 \equiv$  Los datos de la muestra son aleatorios

$H_1 \equiv$  Los datos de la muestra no son aleatorios

Comenzamos introduciendo los datos en **R**:

```
> datos <- c(18.21, 2.36, 17.3, 16.6, 4.70, 3.63, 15.56, 7.35, 9.78, 14.69)
> datos
```

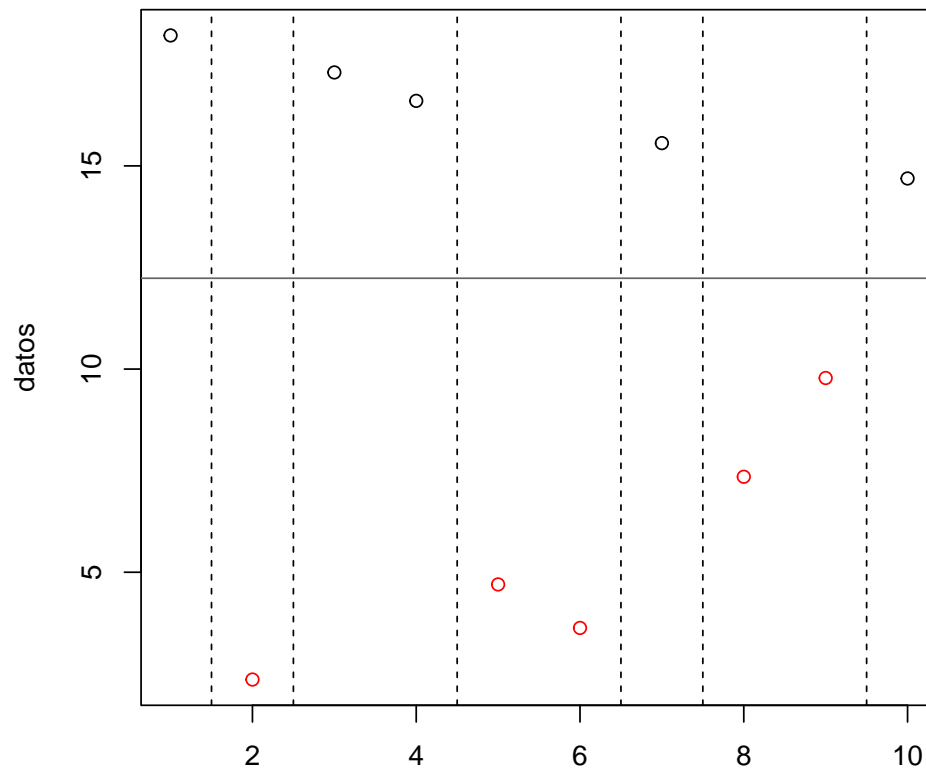
```
[1] 18.21  2.36 17.30 16.60  4.70  3.63 15.56  7.35  9.78 14.69
```

Para resolver el contraste, cargamos el paquete **randtests** y, a continuación, llamamos a la función **runs.test**. Cuando llamamos a esta función, debemos tener en cuenta que la hipótesis alternativa es del tipo “**distinto de**”. Por otra parte, como el enunciado no especifica ningún punto de corte para transformar los valores del vector numérico en valores dicotómicos, este punto de corte vendrá dado por la mediana de los datos (**función median en R**).

```
> library(randtests)
> runs.test (datos, alternative = "two.sided", threshold = median(datos), plot = TRUE)
```

Runs Test

```
data:  datos
statistic = 0.67082, runs = 7, n1 = 5, n2 = 5, n = 10, p-value = 0.5023
alternative hypothesis: nonrandomness
```



Según los resultados del test de rachas, se han encontrado 7 rachas (runs), que vienen separadas por líneas discontinuas verticales. Hay 5 valores por encima de la mediana ( $n_1$ ), marcados en negro, y otros 5 valores por debajo de la mediana ( $n_2$ ), marcados en rojo.

El **p-valor** asociado al contraste es **0.5023** superior a 0.05, por lo que no es posible rechazar la hipótesis nula. Por tanto, podemos concluir que los datos de la muestra son aleatorios.



### 1.15. Supuesto Práctico 15

Las puntuaciones de 10 individuos en una prueba de una oposición han sido las siguientes: 41.81, 40.30, 40.20, 37.14, 39.29, 38.79, 40.73, 39.26, 35.74, 41.65. ¿Puede suponerse, a un nivel de significación del 5 % que dichas puntuaciones se ajustan a una distribución normal de media 40 y desviación típica 3?

#### Solución del Supuesto Práctico 15

El contraste de hipótesis que se plantea es el siguiente:

$$H_0 \equiv \text{Los datos de la muestra proceden de una distribución } N(40,3)$$
$$H_1 \equiv \text{Los datos de la muestra no proceden de una distribución } N(40,3)$$

Comenzamos introduciendo los datos en **R**:

```
> datos <- c(41.81, 40.30, 40.20, 37.14, 39.29, 38.79, 40.73, 39.26, 35.74, 41.65)
> datos

[1] 41.81 40.30 40.20 37.14 39.29 38.79 40.73 39.26 35.74 41.65
```

A continuación, se resuelve el contraste mediante una llamada a la función **ks.test**. Debemos tener en cuenta que la distribución de comparación es la distribución normal (por tanto, el argumento **Supuesto Práctico 16** y tomará el valor **pnorm**) de media igual a 40 y desviación típica igual a 3.

```
> ks.test(datos, y = pnorm, 40, 3, alternative = "two.sided")
```

One-sample Kolmogorov-Smirnov test

```
data: datos
D = 0.27314, p-value = 0.3752
alternative hypothesis: two-sided
```

En este caso, el valor del estadístico de contraste es **0.27314** y el **p-valor** asociado al contraste es **0.3752**. Como el p-valor es superior a 0.05 no podemos rechazar la hipótesis nula, por lo que concluimos que los datos de la muestra proceden de una distribución normal de media 40 y de desviación típica 3.

### 1.16. Supuesto Práctico 16

En unos grandes almacenes se realiza un estudio sobre el rendimiento de ventas de los vendedores. Para ello, se observa durante 10 días el número de ventas de dos vendedores:

Vendedor A: 10 40 60 15 70 90 30 32 22 13  
Vendedor B: 45 60 35 30 30 15 50 20 32 9

Contrastar, considerando un nivel de significación del 5%, si los rendimientos medianos de ambos vendedores pueden asumirse iguales.

#### Solución del Supuesto Práctico 16

Comenzamos introduciendo los datos de ventas de los dos vendedores:

```
> datosA <- c(10, 40, 60, 15, 70, 90, 30, 32, 22, 13)
> datosA

[1] 10 40 60 15 70 90 30 32 22 13
```



```
> datosB <- c (45, 60, 35, 30, 30, 15, 50, 20, 32, 9)
> datosB
```

```
[1] 45 60 35 30 30 15 50 20 32 9
```

A continuación, vamos a plantear el contraste que se debe resolver:

$$H_0 \equiv M_{eA} - M_{eB} = 0$$
$$H_1 \equiv M_{eA} - M_{eB} \neq 0$$

O, equivalentemente,

$$H_0 \equiv M_{eA} = M_{eB}$$
$$H_1 \equiv M_{eA} \neq M_{eB}$$

amos a resolver el contraste utilizando la función **wilcox.test**. Para ello, tendremos en cuenta que los datos proceden de muestras independientes, que el valor de la diferencia entre las medianas que se pretende comprobar es 0 y que la hipótesis alternativa del contraste es del tipo “distinto de”. Además, indicaremos que se incluya el intervalo de confianza para la diferencia de las medianas entre las salidas de la función y que no se aplique la corrección por continuidad.

```
> wilcox.test(datosA, y = datosB, alternative = "two.sided", mu = 0, paired = FALSE,
+ correct = FALSE, conf.int = TRUE, conf.level = 0.95)
```

Wilcoxon rank sum test

```
data:  datosA and datosB
W = 52.5, p-value = 0.8497
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -17.00003  25.00003
sample estimates:
difference in location
 0.5611639
```

En este caso, el p-valor asociado al contraste es, aproximadamente, 0.85. Como este p-valor es mayor que 0.05 no se puede rechazar la hipótesis nula, considerando un nivel de significación del 5%. Por tanto, concluimos que las medianas de las ventas de ambos vendedores pueden asumirse iguales. El intervalo de confianza para la diferencia de las medianas incluye, como era de esperar, el valor 0.

### 1.17. Supuesto Práctico 17

En un encinar de Navarra se pretende comprobar si un tratamiento ayuda a disminuir el nivel de húmedas de las hojas de las encinas. Para ello, se realiza un estudio a 10 encinas, en las que se seleccionan aleatoriamente 10 hojas y se registra el nivel de humedad de las hojas antes y después del tratamiento. Los resultados son los siguientes:

<i>Antes</i>	10.5	9.7	13.3	7.5	12.8	15.2	11.2	10.7	5.2	18.9
<i>Después</i>	11.2	7.8	9.2	3.4	8.9	10.8	11.4	8.5	6.2	11.1

Suponiendo un nivel de significación del 5%, ¿Puede suponerse efectivo el tratamiento?



### Solución del Supuesto Práctico 17

En primer lugar, introduzcamos los datos en dos vectores numéricos en **R**.

```
> datosAntes <- c(10.5, 9.7, 13.3, 7.5, 12.8, 15.2, 11.2, 10.7, 5.2, 18.9)
> datosAntes
```

```
[1] 10.5  9.7 13.3  7.5 12.8 15.2 11.2 10.7  5.2 18.9
```

```
> datosDespues <- c(11.2, 7.8, 9.2, 3.4, 8.9, 10.8, 11.4, 8.5, 6.2, 11.1)
> datosDespues
```

```
[1] 11.2  7.8  9.2  3.4  8.9 10.8 11.4  8.5  6.2 11.1
```

El contraste que se debe resolver es el siguiente:

$$H_0 \equiv Me_{Antes} = Me_{Despues}$$
$$H_1 \equiv Me_{Antes} > Me_{Despues}$$

Vamos a resolver el contraste usando la función **wilcox.test**. Hay que recordar que como los datos son relacionados, debemos asignar al parámetro **paired** el valor **TRUE**.

```
> wilcox.test (datosAntes, y = datosDespues, alternative = "greater", mu = 0, paired = TRUE
+ , correct = FALSE)
```

```
Wilcoxon signed rank exact test
```

```
data:  datosAntes and datosDespues
```

```
V = 49, p-value = 0.01367
```

```
alternative hypothesis: true location shift is greater than 0
```

En este ejemplo, el p-valor asociado al contraste es 0.013, inferior a 0.05, por lo que se rechaza la hipótesis nula considerando un nivel de significación del 5 %. Esto quiere decir que el tratamiento utilizado es efectivo para reducir el nivel de humedad de las hojas de las encinas.





## 1.18. Ejercicios Guiados

### 1.18.1. Ejercicio Guiado 1

Un fabricante diseña un experimento para estimar la tensión de ruptura media de una fibra es 20. Para ello, observa las tensiones de ruptura, en libras, de 16 hilos de dicha fibra seleccionados aleatoriamente.

- a) Si la tensión de ruptura se distribuye según una normal de desviación típica
- b) Si la tensión de ruptura se distribuye según una normal de desviación típica desconocida.

Las tensiones son: 20.8, 20.6, 21.0, 20.9, 19.9, 20.2, 19.8, 19.6, 20.9, 21.1, 20.4, 20.6, 19.7, 19.6, 20.3, 20.7.

### Solución del Ejercicio Guiado 1

En ambos casos, el contraste de hipótesis que debemos resolver es:

$$\begin{aligned}H_0 &\equiv \mu = 20 \\H_1 &\equiv \mu \neq 20\end{aligned}$$

En primer lugar, introduciremos en un vector los datos de las 16 tensiones observadas.

```
> tensiones <- c(20.8, 20.6, 21.0, 20.9, 19.9, 20.2, 19.8, 19.6, 20.9, 21.1, 20.4, 20.6,
+ 19.7, 19.6, 20.3, 20.7)
> tensiones

[1] 20.8 20.6 21.0 20.9 19.9 20.2 19.8 19.6 20.9 21.1 20.4 20.6 19.7 19.6 20.3
[16] 20.7
```

También indicamos el nivel de significación,  $\mu_0$  y la desviación típica poblacional de la variable que proporciona el enunciado.

```
> alpha <- 0.02
> alpha
```

```
[1] 0.02
```

```
> mu_0 <- 20
> mu_0
```

```
[1] 20
```

```
> desv_tipica <- 0.45
> desv_tipica
```

```
[1] 0.45
```

- a) Si la tensión de ruptura se distribuye según una normal de desviación típica

En este primer caso, y dado que conocemos la desviación típica poblacional de la distribución de la tensión de la fibra, debemos calcular manualmente los valores del estadístico de contraste y del valor crítico, que serán:

```
> n <- length(tensiones)
> n
```



```
[1] 16
```

```
> media <- mean(tensiones)
> media
```

```
[1] 20.38125
```

```
> Z <- (media - mu_0) / (desv_tipica/sqrt(n))
> Z
```

```
[1] 3.388889
```

El siguiente Proceso:

```
> cuantil <- qnorm(1 - alpha/2)
> cuantil
```

```
[1] 2.326348
```

De este modo, ya tenemos todo lo necesario para la resolución del contraste. Como el valor absoluto del estadístico de contraste **3.3888** es mayor que el cuantil  $Z_{1-\frac{\alpha}{2}}$ , rechazamos la hipótesis nula en favor de la alternativa. Es decir, no puede asumirse que la tensión media de ruptura de la fibra sea de 20 unidades.

**b) Si la tensión de ruptura se distribuye según una normal de desviación típica desconocida.**

Cuando la desviación típica no se conoce, usamos la función `test` para obtener el intervalo de confianza.

```
> t.test(tensiones, alternative = "two.sided", mu = 20, conf.level = 0.98)
```

One Sample t-test

```
data: tensiones
t = 2.9154, df = 15, p-value = 0.01066
alternative hypothesis: true mean is not equal to 20
98 percent confidence interval:
 20.04092 20.72158
sample estimates:
mean of x
 20.38125
```

En este segundo caso, el valor del estadístico de contraste es **2.9154**. El p-valor asociado al contraste es **0.01066**, que al ser menor que **0.02**, el nivel de significación, nos lleva también al rechazo de la hipótesis nula.

En este segundo caso, el intervalo de confianza para la tensión media de la fibra, al 98 % de confianza, es **(20.04092, 20.72158)**.



### 1.18.2. Ejercicio Guiado 2

En una muestra de 40 alumnos, 25 de ellos están conformes con las decisiones que ha tomado el profesor con respecto a las calificaciones. ¿Puede suponerse, con un nivel de significación del 5 %, que la mitad o más de los alumnos están de acuerdo con las calificaciones del profesor?

#### Solución del Ejercicio Guiado 2

En este caso, el contraste que se debe resolver es:

$$\begin{aligned}H_0 &\equiv \pi \geq 0,5 \\ H_1 &\equiv \pi < 0,5\end{aligned}$$

En este caso, debemos utilizar la función **prop.test** para resolver el contraste de hipótesis anterior. Disponemos tanto del número de alumnos que presentan la característica de interés (estar conforme con el profesor) como del número total de alumnos en la muestra, de manera que podemos realizar la llamada a la función tal y como sigue:

```
> prop.test(25, 40, p = 0.5, alternative = "less", conf.level = 0.95)
```

```
1-sample proportions test with continuity correction
```

```
data: 25 out of 40, null probability 0.5
X-squared = 2.025, df = 1, p-value = 0.9226
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.7501004
sample estimates:
      p
0.625
```

El p-valor para este contraste es **0.9226**, el cual es mayor que el nivel de significación, que es 0.05. Por ello, no podemos rechazar la hipótesis nula del contraste y concluiremos diciendo que la mitad o más de los alumnos están de acuerdo con las calificaciones del profesor.



### 1.18.3. Ejercicio Guiado 3

Una agencia estatal vigila la calidad del agua para la cría de peces. Esta agencia desea comparar la cantidad media de cierta sustancia tóxica en dos ríos contaminados por desperdicios industriales. Se seleccionaron 11 muestras en un río y 8 muestras en el otro. Los resultados de los análisis fueron:

Río 1: 10, 10, 12, 13, 9, 8, 12, 12, 10, 14, 8

Río 2: 11, 8, 9, 7, 10, 8, 8, 10

Si las dos poblaciones son normales e independientes, ¿puede suponerse que la cantidad media de sustancia tóxica presente en ambos ríos es la misma? Considerar un nivel de significación del 5 %.

### Solución del Ejercicio Guiado 3

En primer lugar introducimos los datos en R:

```
> Rio1 <- c(10, 10, 12, 13, 9, 8, 12, 12, 10, 14, 8)
> Rio1
[1] 10 10 12 13 9 8 12 12 10 14 8
> Rio2 <- c(11, 8, 9, 7, 10, 8, 8, 10)
> Rio2
[1] 11 8 9 7 10 8 8 10
```

Aunque el enunciado nos pide resolver un contraste de hipótesis para la diferencia de la cantidad media de sustancia tóxica en ambos ríos, primero debemos saber si la variabilidad del nivel de sustancia tóxica en ambos ríos puede considerarse igual. Para ello, resolveremos el siguiente contraste de hipótesis:

$$\begin{aligned}H_0 &\equiv \sigma_1^2 = \sigma_2^2 \\H_1 &\equiv \sigma_1^2 \neq \sigma_2^2\end{aligned}$$

Para resolver este primer contraste, empleamos la función **var.test**.

```
> var.test(Rio1, Rio2, alternative = "two.sided", conf.level = 0.90)

F test to compare two variances

data:  Rio1 and Rio2
F = 2.1846, num df = 10, denom df = 7, p-value = 0.3119
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.6007504 6.8498698
sample estimates:
ratio of variances
 2.184643
```

Según los resultados de **var.test**, el estadístico de contraste toma el valor **2.1846**. El p-valor asociado al contraste es **0.3119**, que es mayor que el nivel de significación (0.10). Por tanto, no podemos rechazar la hipótesis nula o, equivalentemente, podemos asumir que ambas varianzas son iguales en ambos ríos.

Teniendo en cuenta esta información, resolveremos el contraste para la diferencia de medias, que en este caso toma la forma:



$$H_0 \equiv \mu_1 - \mu_2 = 0$$
$$H_1 \equiv \mu_1 - \mu_2 \neq 0$$

Vamos a realizar una llamada a la función **t.test** para resolver este contraste.

```
> t.test(Rio1, Rio2, alternative = "two.sided", mu = 0, var.equal = TRUE, conf.level =  
+ 0.90)
```

Two Sample t-test

```
data: Rio1 and Rio2  
t = 2.2564, df = 17, p-value = 0.0375  
alternative hypothesis: true difference in means is not equal to 0  
90 percent confidence interval:  
 0.424258 3.280287  
sample estimates:  
mean of x mean of y  
 10.72727  8.87500
```

En este caso, el p-valor asociado al contraste es **0.0375**, que es menor que 0.10, el nivel de significación. Por tanto, rechazamos la hipótesis nula y concluimos que la cantidad media de tóxico en ambos ríos no es la misma.



#### 1.18.4. Ejercicio Guiado 4

Una empresa farmacéutica está interesada en la investigación preliminar de un nuevo medicamento que parece tener propiedades reductoras del colesterol en la sangre. A tal fin se toma una muestra al azar de 6 personas, y se determina el contenido en colesterol antes y después del tratamiento. Los resultados han sido los siguientes:

Antes: 217, 252, 229, 200, 209, 213

Después: 209, 241, 230, 208, 206, 211

Comprobar, a un nivel de significación del 4% si la aplicación del medicamento es efectiva. Es decir, comprobar si el nivel medio de colesterol en sangre de los pacientes antes de la aplicación del medicamento es mayor o igual al nivel medio de colesterol en sangre después del tratamiento.

#### Solución del Ejercicio Guiado 4

El contraste que debemos resolver es:

$$\begin{aligned}H_0 &\equiv \mu_a - \mu_d \geq 0 \\H_1 &\equiv \mu_a - \mu_d < 0\end{aligned}$$

Introducimos los datos en R:

```
> Antes <- c(217, 252, 229, 200, 209, 213)
> Antes
[1] 217 252 229 200 209 213
> Despues <- c(209, 241, 230, 208, 206, 211)
> Despues
[1] 209 241 230 208 206 211
```

Estos datos son un claro ejemplo de datos pareados, ya que a los mismos individuos se les ha medido el nivel de colesterol en sangre antes y después del tratamiento.

Ahora sólo nos queda realizar la llamada a la función `t.test`, sin olvidar indicar mediante el parámetro `paired` la relación que existe entre los conjuntos de datos.

```
> t.test(Antes, Despues, alternative = "less", mu = 0, paired = TRUE, conf.level = 0.96)
```

Paired t-test

```
data: Antes and Despues
t = 0.91186, df = 5, p-value = 0.7982
alternative hypothesis: true difference in means is less than 0
96 percent confidence interval:
 -Inf 8.506849
sample estimates:
mean of the differences
      2.5
```

En este caso, el estadístico de contraste toma el valor **0.91186** y el p-valor es **0.7982**. Este p-valor supera el nivel de significación, que recordemos es 0.04. Por tanto, no podemos rechazar la hipótesis nula y concluimos que los niveles medios de colesterol antes y después del tratamiento pueden considerarse iguales, poniendo así en duda la efectividad del mismo.



### 1.18.5. Ejercicio Guiado 5

Una determinada empresa quiere saber si su nuevo producto tendrá más aceptación en la población adulta o entre los jóvenes. Para ello, considera una muestra aleatoria de 400 adultos y 600 jóvenes, observando que sólo a 100 adultos y 300 jóvenes les había gustado su producto. Tomando un nivel de significación del 1 %, ¿puede suponerse que el producto gusta por igual en adultos y jóvenes?

#### Solución del Ejercicio Guiado 5

Para responder a la pregunta que se nos plantea, resolveremos el siguiente contraste de hipótesis:

$$H_0 \equiv \pi_A - \pi_J = 0$$
$$H_1 \equiv \pi_A - \pi_J \neq 0$$

Donde  $\pi_A$  y  $\pi_J$  es la proporción de adultos y jóvenes, respectivamente, a los que gusta el producto.

En este caso debemos utilizar la función `prop.test` para resolver este contraste. Pero primero, debemos crear un vector que indique el número de adultos y jóvenes a los que les gusta el producto así como un segundo vector con el número total de adultos y jóvenes encuestados.

```
> Adul_Jov_Gusta_Producto <- c(100, 300)
> Adul_Jov_Gusta_Producto
```

```
[1] 100 300
```

```
> Adul_Jov_Total <- c(400, 600)
> Adul_Jov_Total
```

```
[1] 400 600
```

Una vez hecho esto, llamamos a la función **prop.test**.

```
> prop.test(Adul_Jov_Gusta_Producto, Adul_Jov_Total, alternative = "two.sided", conf.level
+ = 0.99)
```

2-sample test for equality of proportions with continuity correction

```
data: Adul_Jov_Gusta_Producto out of Adul_Jov_Total
X-squared = 61.463, df = 1, p-value = 4.512e-15
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.3287296 -0.1712704
sample estimates:
prop 1 prop 2
 0.25  0.50
```

Según los resultados que proporciona **prop.test**, el p-valor asociado a este contraste es muy pequeño, concretamente **4.512e-15**. Este p-valor es menor que 0.01, el nivel de significación. Por eso, rechazamos la hipótesis nula en favor de la alternativa y podemos afirmar que el producto no gusta por igual entre adultos y jóvenes.



## 1.19. Ejercicios Propuestos

### 1.19.1. Ejercicio Propuesto 1

Se realiza un experimento para estudiar el nivel (en minutos) que se requiere para que la temperatura del cuerpo de un lagarto del desierto alcance los 45<sup>o</sup> partiendo de la temperatura normal de su cuerpo mientras está en la sombra. Se supone que la varianza es conocida. Se obtuvieron las siguientes observaciones: 10.1 ; 12.5 ; 12.2 ; 10.2 ; 12.8 ; 12.1 ; 11.2 ; 11.4 ; 10.7 ; 14.9 ; 13.9 ; 13.3. Se pide:

- a) Hallar estimaciones puntuales de la media y la varianza
- b) Supóngase que la variable X: “Tiempo en alcanzar los 45<sup>o</sup> sigue una ley Normal
- b1) ¿Puede concluirse que el tiempo medio requerido para alcanzar la dosis letal es de 15 minutos?
- b2) ¿Puede concluirse que el tiempo medio requerido para alcanzar la dosis letal es inferior a 13 minutos?

### Solución del Ejercicio Propuesto 1

Introducimos los datos en R:

```
> Temperatura <- c(10.1 , 12.5 , 12.2 , 10.2 , 12.8 , 12.1 , 11.2 , 11.4 , 10.7 , 14.9 ,  
+ 13.9 , 13.3)  
> Temperatura  
  
[1] 10.1 12.5 12.2 10.2 12.8 12.1 11.2 11.4 10.7 14.9 13.9 13.3
```

#### a) Hallar estimaciones puntuales de la media y la varianza:

Proceso para encontrar la Media:

```
> media <- mean(Temperatura)  
> media  
  
[1] 12.10833
```

Proceso para encontrar la Varianza:

```
> var <- var(Temperatura)  
> var  
  
[1] 2.186288
```

Con los Resultados obtenidos establecemos que para el literal a) la Media es **12.10833** y la Varianza es **2.186288**

#### b1) ¿Puede concluirse que el tiempo medio requerido para alcanzar la dosis letal es de 15 minutos?

Nos piden el siguiente contraste de hipótesis:

$$\begin{aligned}H_0 : \mu &= 15 \\H_1 : \mu &\neq 15\end{aligned}$$





Proceso:

```
> t.test(Temperatura, alternative = "two.sided", mu = 15, conf.level = 0.95)
```

One Sample t-test

```
data: Temperatura
t = -6.7746, df = 11, p-value = 3.055e-05
alternative hypothesis: true mean is not equal to 15
95 percent confidence interval:
 11.16887 13.04780
sample estimates:
mean of x
 12.10833
```

El valor del estadístico de contraste experimental, **-6.7746**, deja a la derecha una área menor que **0.000 < 0.025**. Por lo tanto, se rechaza la hipótesis nula de que el tiempo medio requerido para alcanzar la dosis letal es de 15 minutos.

**b2) ¿Puede concluirse que el tiempo medio requerido para alcanzar la dosis letal es inferior a 13 minutos?**

Nos piden el siguiente contraste de hipótesis:

$$\begin{aligned} H_0 : \mu &\geq 13 \\ H_1 : \mu &< 13 \end{aligned}$$

```
> t.test(Temperatura, alternative = "less", mu = 13, conf.level = 0.95)
```

One Sample t-test

```
data: Temperatura
t = -2.089, df = 11, p-value = 0.03037
alternative hypothesis: true mean is less than 13
95 percent confidence interval:
 -Inf 12.87489
sample estimates:
mean of x
 12.10833
```

El valor del estadístico de contraste experimental, **-2.089**, deja a la derecha una área **0.030 < 0.05**. Por lo tanto se rechaza la hipótesis nula y se concluye que el tiempo medio requerido para alcanzar la dosis letal es inferior a 13 minutos.



### 1.19.2. Ejercicio Propuesto 2

Se quieren comparar dos poblaciones de ranas pipiens aisladas geográficamente. Para ello se toman dos muestras de ambas poblaciones de tamaño 12 y 10 y se les mide la longitud del cuerpo expresado en milímetros.

Población 1: 20,1; 22,5; 22,2 ; 30,2 ; 22,8 ; 22,1 ; 21,2 ; 21,4 ; 20,7 ; 24,9 ; 23,9 ; 23,3

Población 2: 25,3 ; 31,2 ; 22,4 ; 23,1 ; 26,4 ; 28,2 ; 21,3 ; 31,1 ; 26,2 ; 21,4

Contrastar la hipótesis de igualdad de medias a un nivel de significación del 5 %. (Suponiendo que la longitud se distribuya según una Normal).

### Solución del Ejercicio Propuesto 2

Introducimos los datos en R:

```
> Poblacion1 <- c(20.1, 22.5, 22.2 , 30.2 , 22.8 , 22.1 , 21.2 , 21.4 , 20.7 , 24.9 , 23.9  
+ , 23.3)  
> Poblacion1
```

```
[1] 20.1 22.5 22.2 30.2 22.8 22.1 21.2 21.4 20.7 24.9 23.9 23.3
```

```
> Poblacion2 <- c(25.3, 31.2 , 22.4 , 23.1 , 26.4 , 28.2 , 21.3 , 31.1 , 26.2 , 21.4)  
> Poblacion2
```

```
[1] 25.3 31.2 22.4 23.1 26.4 28.2 21.3 31.1 26.2 21.4
```

En primer lugar realizamos el contraste de las varianzas:

$$H_0 : \sigma_X^2 = \sigma_Y^2$$
$$H_1 : \sigma_X^2 \neq \sigma_Y^2$$

Proceso para llevar a cabo la Solución:

```
> var.test(Poblacion1, Poblacion2, alternative = "two.sided", conf.level = 0.95)
```

```
F test to compare two variances
```

```
data: Poblacion1 and Poblacion2  
F = 0.51989, num df = 11, denom df = 9, p-value = 0.3043  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.1328934 1.8653086  
sample estimates:  
ratio of variances  
 0.519889
```

Para realizar un contraste de diferencia de medias de dos poblaciones independientes hay que contrastar previamente las varianzas de dichas poblaciones. Esta salida nos muestra el valor experimental del estadístico de contraste (**Fexp = 0.51989**), este valor deja a la derecha un área igual a **0.3043**, por lo tanto no se puede rechazar la hipótesis nula de igualdad de varianzas.

A continuación vamos a realizar el contraste de diferencia de medias:



$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

Proceso para llevar a cabo la Solución:

```
> t.test(Poblacion1, Poblacion2, alternative = "two.sided", mu = 0, var.equal = TRUE)
```

Two Sample t-test

```
data: Poblacion1 and Poblacion2
t = -2.0097, df = 20, p-value = 0.05815
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.5398671  0.1032004
sample estimates:
mean of x mean of y
 22.94167  25.66000
```

A continuación se realiza el contraste para la diferencia de medias suponiendo que las varianzas son iguales. La tabla nos muestra el valor experimental del estadístico de contraste  $t_{exp} = -2,0097$  y el **p-valor = 0.0508**, por lo tanto no se puede rechazar la hipótesis nula de igualdad de medias. También, se puede concluir el contraste observando que el intervalo de confianza para la diferencia de medias (**-5.5398, 0.1032**) contiene al cero.



### 1.19.3. Ejercicio Propuesto 3

Se realiza un estudio, en el que participan 10 individuos, para investigar el efecto del ejercicio físico en el nivel de colesterol en plasma. Antes del ejercicio se tomaron muestras de sangre para determinar el nivel de colesterol de cada individuo. Después, los participantes fueron sometidos a un programa de ejercicios. Al final de los ejercicios se tomaron nuevamente muestras de sangre y se obtuvo una segunda lectura del nivel de colesterol. Los resultados se muestran a continuación.

Nivel previo: 182; 230; 160; 200; 160; 240; 260; 480; 263; 240

Nivel posterior: 190; 220; 166; 150; 140; 220; 156; 312; 240; 250

Se quiere saber si el ejercicio físico ha reducido el nivel de colesterol para un nivel de confianza del 95 %.

### Solución del Ejercicio Propuesto 3

Introducimos los datos en **R**:

```
> Nivelprevio <- c(182, 230, 160, 200, 160, 240, 260, 480, 263, 240)
> Nivelprevio

[1] 182 230 160 200 160 240 260 480 263 240

> Nivelposterior <- c(190, 220, 166, 150, 140, 220, 156, 312, 240, 250)
> Nivelposterior

[1] 190 220 166 150 140 220 156 312 240 250
```

A continuación vamos a realizar el contraste de diferencia de medias de dos poblaciones apareadas:

$$H_0 : \mu_{Pr} - \mu_{Po} \leq 0$$
$$H_1 : \mu_{Pr} - \mu_{Po} > 0$$

Proceso a realizar para llevar a cabo la solución:

```
> t.test(Nivelprevio, Nivelposterior, alternative = "greater", mu = 0, paired = TRUE)
```

Paired t-test

```
data: Nivelprevio and Nivelposterior
t = 2.0525, df = 9, p-value = 0.03516
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.965698      Inf
sample estimates:
mean of the differences
      37.1
```

La salida muestra el valor experimental del estadístico de contraste (**t = 2.0525**) y **Sig. 0.03516**, menor que 0.05, y se debe rechazar la hipótesis nula. Por lo tanto, el nivel medio de colesterol se reducirá con el ejercicio físico.



#### 1.19.4. Ejercicio Propuesto 4

Se ignora la proporción de familias numerosas y con el fin de determinar dicha proporción se toma una muestra de 800 familias siendo la proporción observada de 0.18. Se puede afirmar que la proporción de familias numerosas es 0.20.

#### Solución del Ejercicio Propuesto 4

Se pide realizar el siguiente contraste  $H_0 : P = 0,20$  frente a la alternativa  $H_1 : p \neq 20$ .

Según el enunciado de una muestra de 800 familias la proporción observada de familias numerosas es 0.18. Por lo tanto 144 familias son numerosas y 656 no lo son.

Proceso para llevar a cabo la Solución del Ejercicio:

```
> prop.test(144,800, p = 0.20, alternative = "two.sided", conf.level = 0.95)
```

```
1-sample proportions test with continuity correction
```

```
data: 144 out of 800, null probability 0.2
X-squared = 1.877, df = 1, p-value = 0.1707
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1543404 0.2087896
sample estimates:
      p
0.18
```

El p-valor de la prueba (**Sig. exacta (unilateral)**) es **0.1707** mayor que 0.05. Por lo tanto no se rechaza la Hipótesis nula. Se puede afirmar que la proporción de familias numerosas es 0.20.



### 1.19.5. Ejercicio Propuesto 5

Se sospecha que añadiendo al tratamiento habitual para la curación de una enfermedad un medicamento A, se consigue mayor número de curaciones. Tomamos dos grupos de enfermos de 100 individuos cada uno. A un grupo se le suministra el medicamento A y se curan 60 enfermos y al otro no se le suministra, curándose 55 enfermos. ¿Es efectivo el tratamiento A en la curación de la enfermedad?

### Solución del Ejercicio Propuesto 5

Se pide realizar el siguiente contraste para llevar a cabo la solución de este ejercicio:

$$H_0 : P_x \leq P_y$$
$$H_1 : P_x > P_y$$

En cada grupo hay 100 enfermos a los que se les suministra el medicamento A. De los cuales hay 60 curados en un grupo y 45 curados en el otro grupo.

```
> curados <- c(60, 45)
> curados

[1] 60 45

> enfermos <- c(100, 100)
> enfermos

[1] 100 100

> prop.test(curados, enfermos, alternative = "greater", conf.level = 0.95)

      2-sample test for equality of proportions with continuity correction

data:  curados out of enfermos
X-squared = 3.9298, df = 1, p-value = 0.02372
alternative hypothesis: greater
95 percent confidence interval:
 0.02515435 1.00000000
sample estimates:
prop 1 prop 2
 0.60  0.45
```

El valor del estadístico Chi-Cuadrado es **3.9298** y el p-valor asociado es **0.02372** por lo tanto se debe rechazar la Hipótesis nula. Podemos afirmar que el medicamento A consigue un mayor número de curaciones.

### 1.19.6. Ejercicio Propuesto 6

En 5 zonas de la provincia de Granada (Ladihonda y Fazares, zonas muy secas y Cortijuela, Molinillo y Fardes, zonas húmedas) se hacen una serie de mediciones sobre las hojas de las encinas a lo largo de 3 años consecutivos: 1995, muy seco y 1996 y 1997, muy lluviosos.

El objetivo es medir la simetría fluctuante en dichas hojas como indicador de stress en la planta. Bajo condiciones de stress (sequía, herbivoría, limitación por nutrientes...), la hipótesis es que la asimetría aumente. Contamos con la siguiente información:

- Localización árboles: 5 zonas, dos en zonas muy secas (Hoya Guadix-Baza, Ladihonda y Fazares) y tres en zonas con mayor precipitación (Cortijuela, Molinillo, Fardes). En esta última, Fardes, son árboles situados en la ladera de un río (presumiblemente poco afectados por años más o menos secos).
- Años de climatología diferente: 1995 año muy seco y años 1996 y 1997, años muy lluviosos.
- Situación de la hoja: Canopy (copa de los árboles) y Sprouts (rebrotos, hojas nuevas que salen desde la parte inferior del tronco).

Disponemos de un total de 2101 casos, cedidos por el Departamento de Ecología de la Universidad de Granada (España), de los que hemos seleccionado aleatoriamente una muestra de tamaño 15 que se presenta en la siguiente tabla:

<i>Zona</i>	<i>Parte</i>	<i>Año</i>	<i>Longitud</i>	<i>Asimetría</i>
<i>Cortijuela</i>	<i>Canopy</i>	1995	26.51	0.028
<i>Cortijuela</i>	<i>Canopy</i>	1996	30.17	0.010
<i>Molinillo</i>	<i>Canopy</i>	1995	34.24	0.080
<i>Molinillo</i>	<i>Canopy</i>	1996	31.04	0.340
<i>Molinillo</i>	<i>Canopy</i>	1996	34.99	0.087
<i>Fardes</i>	<i>Canopy</i>	1995	30.48	0.040
<i>Fardes</i>	<i>Canopy</i>	1996	25.07	0.010
<i>Ladihonda</i>	<i>Canopy</i>	1995	25.04	0.021
<i>Ladihonda</i>	<i>Canopy</i>	1996	29.16	0.135
<i>Fazares</i>	<i>Canopy</i>	1995	35.12	0.010
<i>Fazares</i>	<i>Canopy</i>	1996	25.41	0.094
<i>Fazares</i>	<i>Canopy</i>	1996	27.02	0.153
<i>Cortijuela</i>	<i>Sprouts</i>	1995	23.04	0.156
<i>Fazares</i>	<i>Sprouts</i>	1995	27.69	0.172
<i>Fazares</i>	<i>Sprouts</i>	1996	34.71	0.077



1. ¿Se puede admitir que la longitud de las hojas de encina se distribuye normalmente?
2. ¿Se puede admitir que la longitud media de las hojas es igual a 30 cm a un nivel de significación del 5 %? (Suponiendo que la varianza es conocida)
3. Suponiendo que la asimetría de las hojas sigan una distribución Normal; comprobar mediante un contraste de hipótesis si existen diferencias significativas en la asimetría de las hojas teniendo en cuenta la situación de la hoja en el árbol.
4. A un nivel de significación del 5 %, ¿es representativo el ajuste lineal entre la longitud y la asimetría? ¿Cuál sería la expresión del modelo? ¿Cuánto explica el modelo?

### Solución del Ejercicio Propuesto 6

a) ¿Se puede admitir que la longitud de las hojas de encina se distribuye normalmente?

Introducimos los datos en R:

```
> longitud <- c(26.51, 30.17, 34.24, 31.04, 34.99, 30.48, 25.07, 25.04, 29.16, 35.12, 25.41,
+ 27.02, 23.04, 27.69, 34.71)
> longitud

[1] 26.51 30.17 34.24 31.04 34.99 30.48 25.07 25.04 29.16 35.12 25.41 27.02
[13] 23.04 27.69 34.71
```

Calculamos la media y la desviación típica:

```
> mean(longitud)

[1] 29.31267

> sd(longitud)

[1] 4.062451
```

Se contrasta la normalidad mediante el contraste de Kolmogorov-Smirnov.

```
> ks.test(longitud, y = pnorm, 29.31267, 4.062451, alternative = "two.sided")
```

One-sample Kolmogorov-Smirnov test

```
data: longitud
D = 0.15408, p-value = 0.8173
alternative hypothesis: two-sided
```

Mediante la prueba de Kolmogorov-Smirnov obtenemos que el **p-valor es 0.8173**, mayor que el nivel de significación 0.05, por lo tanto no se puede rechazar la hipótesis nula y admitimos que la longitud de las hojas sigue una distribución Normal.

b) ¿Se puede admitir que la longitud media de las hojas es igual a 30 cm a un nivel de significación del 5 %? (Suponiendo que la varianza es conocida)

Nos piden el siguiente contraste de hipótesis:

$$H_0 : \mu = 30$$
$$H_1 : \mu \neq 30$$





Proceso a Realizar:

```
> t.test(longitud, alternative = "two.sided", mu = 30, conf.level = 0.95)
```

One Sample t-test

```
data: longitud
t = -0.65528, df = 14, p-value = 0.5229
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 27.06296 31.56238
sample estimates:
mean of x
 29.31267
```

El valor del nivel crítico o p-valor (Sig. (bilateral)) es **0.5229**, mayor que el nivel de significación 0.05, por lo que no se rechaza la hipótesis nula y admitimos que la longitud media de las hojas de encina es igual a 30 cm.

c) Suponiendo que la asimetría de las hojas sigan una distribución Normal; comprobar mediante un contraste de hipótesis si existen diferencias significativas en la asimetría de las hojas teniendo en cuenta la situación de la hoja en el árbol.

```
> asimetriaCanopy<- c(0.028, 0.010, 0.080, 0.340, 0.087, 0.040, 0.010, 0.021, 0.135, 0.010,
+ 0.094, 0.153)
> asimetriaCanopy
```

```
[1] 0.028 0.010 0.080 0.340 0.087 0.040 0.010 0.021 0.135 0.010 0.094 0.153
```

```
> asimetriaSprouts<- c(0.156, 0.172, 0.077)
> asimetriaSprouts
```

```
[1] 0.156 0.172 0.077
```

Para poder realizar el contraste de igualdad de medias de dos poblaciones independientes es necesario contrastar previamente la igualdad de varianzas.

$$\begin{aligned} H_0 : \sigma_c^2 &= \sigma_s^2 \\ H_1 : \sigma_c^2 &\neq \sigma_s^2 \end{aligned}$$

Donde  $\sigma_c^2$  representa la varianza de la asimetría en el Canopy y  $\sigma_s^2$  representa la varianza de la asimetría en el Sprouts.

```
> var.test(asimetriaCanopy, asimetriaSprouts, alternative = "two.sided", conf.level = 0.95)
```

F test to compare two variances

```
data: asimetriaCanopy and asimetriaSprouts
F = 3.4611, num df = 11, denom df = 2, p-value = 0.4908
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08782899 18.19106196
sample estimates:
ratio of variances
 3.461082
```



Analizando la información relativa al contraste de hipótesis que se incluye en la salida de `var.test`, vemos que el valor del estadístico de contraste es 3.4611. La distribución F de Snedecor que sigue el estadístico de contraste tiene 11 grados de libertad en el numerador y 2 en el denominador. El p-valor asociado al contraste es 0.4908. Como este valor es superior al nivel de significación (que para este ejemplo es 0.05), no podemos rechazar la hipótesis nula que hemos planteado. Es decir, se puede considerar que la varianza de ambas poblaciones son iguales.

Una vez aceptada la igualdad de varianzas, realizamos en contraste de diferencia de medias de dos poblaciones normales independientes.

```
> t.test(asimetriaCanopy, asimetriaSprouts, alternative = "two.sided", mu = 0,
+ var.equal = TRUE)
```

Two Sample t-test

```
data: asimetriaCanopy and asimetriaSprouts
t = -0.88477, df = 13, p-value = 0.3924
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1755285  0.0735285
sample estimates:
mean of x mean of y
  0.084    0.135
```

En la salida se incluye el valor del estadístico de contraste (-0.88477), los grados de libertad de la distribución t de Student que sigue el estadístico de contraste (13) y el p-valor (0.3924). Como el p-valor es mayor que el nivel de significación fijado (0.05), no rechazamos la hipótesis nula y se deduce que las partes de la planta (Canopy y Sprouts) no influyen en la asimetría de las hojas.

**d) A un nivel de significación del 5%, ¿es representativo el ajuste lineal entre la longitud y la asimetría? ¿Cuál sería la expresión del modelo? ¿Cuánto explica el modelo?**

Introducimos los datos de la asimetría en R

```
> longitud <- c(26.51, 30.17, 34.24, 31.04, 34.99, 30.48, 25.07, 25.04, 29.16, 35.12, 25.41,
+ 27.02, 23.04, 27.69, 34.71)
> longitud
```

```
[1] 26.51 30.17 34.24 31.04 34.99 30.48 25.07 25.04 29.16 35.12 25.41 27.02
[13] 23.04 27.69 34.71
```

```
> asimetria <- c(0.028, 0.010, 0.080, 0.340, 0.087, 0.040, 0.010, 0.021, 0.135, 0.010, 0.094
+ , 0.153, 0.156, 0.172, 0.077)
> asimetria
```

```
[1] 0.028 0.010 0.080 0.340 0.087 0.040 0.010 0.021 0.135 0.010 0.094 0.153
[13] 0.156 0.172 0.077
```

```
> reg_lin <- lm(asimetria ~ longitud)
> reg_lin
```

```
Call:
lm(formula = asimetria ~ longitud)
```



Coefficients:

```
(Intercept)    longitud
      0.119847    -0.000875
```

Por defecto, la salida que muestra la función `lm` incluye únicamente las estimaciones para los parámetros, en nuestro caso  $\widehat{\beta}_0$  y  $\widehat{\beta}_1$ . Por tanto, el modelo lineal puede escribirse del siguiente modo:

$$\text{asimetría} = 0.119847 - 0.000875 * \text{longitud}$$

Podemos obtener más información sobre el modelo de regresión que hemos calculado aplicando la función `summary` al objeto que contiene los datos de la regresión, al cual hemos llamado `reg_lin` en este ejemplo.

```
> summary(reg_lin)
```

Call:

```
lm(formula = asimetria ~ longitud)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.087912	-0.072795	-0.009889	0.048489	0.247311

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.119847	0.178754	0.670	0.514
longitud	-0.000875	0.006044	-0.145	0.887

Residual standard error: 0.09187 on 13 degrees of freedom

Multiple R-squared: 0.001609, Adjusted R-squared: -0.07519

F-statistic: 0.02096 on 1 and 13 DF, p-value: 0.8871

Esta salida contiene una información más completa sobre el análisis. Así, por ejemplo, encontramos información sobre los **residuos** (en el apartado `Residuals`), que se definen como la diferencia entre el verdadero valor de la variable dependiente y el valor que pronostica el modelo de regresión. Cuanto más pequeños sean estos residuos mejor será el ajuste del modelo a los datos y más acertadas serán las predicciones que se realicen a partir de dicho modelo.

En la tabla **Coefficients** encontramos los valores de los parámetros que aparecían en la salida por defecto junto a su error estándar. Cada parámetro aparece acompañado del valor de un estadístico *t* de Student y un *p*-valor que sirven para contrastar la **significación del parámetro** en cuestión, es decir, para resolver los siguientes contrastes de hipótesis:

$$\begin{aligned} H_0 : \beta_0 &= 0 \text{ vs } H_1 : \beta_0 \neq 0 \\ H_0 : \beta_1 &= 0 \text{ vs } H_1 : \beta_1 \neq 0 \end{aligned}$$

Lo que se pretende mediante estos contrastes es determinar si los efectos de la constante y de la variable independiente son realmente importantes para explicar la variable dependiente o si, por el contrario, pueden considerarse nulos.



En nuestro ejemplo, los p-valores que nos ayudan a resolver estos contrastes son 0.514 y 0.887, ambos mayores que 0.05. Así, considerando un nivel del significación del 5 %, no rechazamos la hipótesis nula en ambos contrastes, de manera que podemos suponer ambos parámetros no son significativamente distintos de 0. Por lo tanto que concluimos que longitud no es válida para predecir la asimetría según un modelo lineal.

Por último, en la parte final de la salida, encontramos el valor de  $R^2$  (Multiple R-squared) y de  $R^2$  ajustado (Adjusted R-squared), que son indicadores de la bondad del ajuste de nuestro modelo a los datos.  $R^2$  oscila entre 0 y 1, de manera que, valores de  $R^2$  próximos a 1 indican un buen ajuste del modelo lineal a los datos. Por otro lado,  $R^2$  ajustado es similar a  $R^2$ , pero penaliza la introducción en el modelo de variables independientes poco relevantes a la hora de explicar la variable dependiente. Por tanto,  $R^2$  ajustado  $\leq R^2$ . En nuestro ejemplo,  $R^2 = 0.001609$  y  $R^2$  ajustado = -0.07519, por lo que podemos concluir que el modelo lineal no se ajusta a nuestros datos.

El porcentaje de variación de la asimetría de las hojas de la encina explicado por el modelo de regresión lineal es igual al 0.1 %, siendo ésta una cantidad claramente insatisfactoria.

La última línea de la salida incluye un estadístico F de Snedecor y el p-valor correspondiente que se utilizan para resolver el siguiente contraste:

$$\begin{aligned} H_0 : \beta_i &= 0 \quad \forall_i \\ H_1 : \beta_i &\neq 0 \text{ para al menos uno } i \end{aligned}$$

que se conoce habitualmente como **contraste ómnibus**. Mediante este contraste se comprueba si, de forma global, el modelo lineal es apropiado para modelizar los datos. En nuestro ejemplo, el p-valor asociado a este contraste 0.8871 es mayor que 0.05 por lo que, al 5 % de significación no podemos rechazar la hipótesis nula y afirmar que, efectivamente, el modelo lineal no es adecuado para nuestro conjunto de datos.