UNIVERSIDAD DE EL SALVADOR.

FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE.

DEPARTAMENTO DE MATEMATICAS.

PRACTICA 6. SEMANA 8

CARRERA:

LICENCIATURA EN ESTADISTICA.

ASIGNATURA:

ANALISIS ESTADISTICO CON EL PAQUETE R

DOCENTE:

JAIME ISAAC PEÑA

PRESENTADO POR:

NELSON DE JESUS MAGAÑA GODINEZ

FECHA:

20 SEPTIEMBRE DE 2022

1 CONSTRASTE DE HIPÓTESIS

1.1 Contrastes de hipótesis para la media de una población normal con Varianza conocida

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.
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

$$H_0 := \mu = 55$$

$$H_1 := \mu \neq 55$$

Expresión 5: Contraste de hipótesis 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.

```
getwd()

## [1] "C:/Users/pc 1/Desktop/PAQUETE R/PRACTICAS_S8"

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

Una vez hecho esto en R, se introduce el nivel de significancia que proporciona el enuciado.

```
alpha <-0.05
```

A continuación, calculamos el valor del estadístico de contraste.

```
media <- mean(datos$Flexiones)
mu_0= 55
```

```r
var <- 7.5
n<-nrow(datos)
z <- (media-mu_0)/(sqrt(var)/sqrt(n))
z

## [1] -15.47408
```

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

```r
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 Contrastes de hipótesis para la media de una población normal con Varianza desconocida

Supongamos, de nuevo, una muestra aleatoria **X1,X2,...,Xn**, de tamaño n de valores de la variable aleatoria que sigue una distribución normal de media μ y desviación típica σ, ambas desconocidas. Para resolver el contraste de hipótesis para μ en este caso partimos del estadístico de contraste

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

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.

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

$$H_0 := \mu = 50$$

$$H_1 := \mu \neq 50$$

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

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.

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 Contrastes de hipótesis para el parámetro p de una distribución Binomial

Supongamos que X es una variable aleatoria con distribución de probabilidad binomial con parámetro n y π, $X \to B(n, \pi)$, de la que se extrae una muestra aleatoria **X1,X2,...,Xn** de tamaño n. Sea p la proporción poblacional. Se desea contrastar si el parámetro π puede ser igual a un valor π_0.

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

$$H_0 := \pi_H \geq 0.05$$

$$H_1 := \pi_H < 0.05$$

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, omo 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.

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

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.

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 Contrastes de hipótesis para la diferencias de medias de dos poblaciones normales e independientes.

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

donde σ^2_M representa la varianza del número de flexiones realizadas por los chicos σ^2_M y representa la varianza del número de flexiones realizadas por las chicas.

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

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

```
var.test(Flexiones.chicas, Flexiones.chicos, conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  Flexiones.chicas and Flexiones.chicos
```

```
## F = 1.1428, num df = 31, denom df = 42, p-value = 0.679
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5964636 2.2647730
## sample estimates:
## ratio of variances
##           1.142781
```

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.

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?

$$H_0 := \mu_H = \mu_M$$

$$H_1 := \mu_H \neq \mu_M$$

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

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.

```
#read.table("Flexiones.txt.txt", header=TRUE)
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

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
```

```
t.test(Flexiones.chicos, Flexiones.chicas, alternative = "two.sided", mu = 0, var.equal = TR
```

```
##
##  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

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.

1.5 Contrastes de hipótesis para la diferencia de medias de dos poblaciones normales relacionadas

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.

¿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:

$$H_0 := \mu_a = \mu_d$$

$$H_1 := \mu_a \neq \mu_d$$

donde μ_a y μ_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.

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

```
Antes <- c(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)
```

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 mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.4135884  0.2469217
## sample estimates:
## mean difference
##       -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.

1.6 Contrastes de hipótesis para la diferencia de proporciones

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:

$$H_0 : \pi_H - \pi_M = 0$$
$$H_0 : \pi_H - \pi_M \neq 0$$

donde μ_H y μ_M representan la proporciones de chicos y chicas que practican deporte, respectivamente.

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_deportes<-c(11,19)
vector_sexo<-c(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_deportes,vector_sexo, alternative = "two.sided", conf.level = 0.92)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  vector_deportes 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
```

egún 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.

1.7 Contrastes de hipótesis no paramétricos

En la sesión anterior hemos estudiado contrastes de hipótesis acerca de parámetros poblacionales, tales como la media y la varianza, de ahí el nombre de contrastes paramétricos. En estadística paramétrica se trabaja bajo el supuesto de que las poblaciones poseen distribuciones conocidas, donde cada función de distribución teórica depende de uno o más parámetros poblacionales. Sin embargo, en muchas situaciones, es imposible especificar la forma de la distribución poblacional. El proceso de obtener conclusiones directamente de las observaciones muestrales, sin formar los supuestos con respecto a la forma matemática de la distribución poblacional se llama teoría no paramétrica.

En esta sesión vamos a realizar procedimientos que no exigen ningún supuesto, o muy pocos acerca de la familia de distribuciones a la que pertenece la población, y cuyas observaciones pueden ser cualitativas o bien se refieren a alguna característica ordenable. En estos casos, cuando no se dispone de información acerca de qué distribución de probabilidad sigue la variable a nivel poblacional, se pueden utilizar técnicas estadísticas no paramétricas para el planteamiento y resolución de contrastes de hipótesis no paramétricos. .Estas técnicas se basan exclusivamente en la información que se recoge en la muestra para resolver los contrastes.

Así, uno de los objetivos de esta sesión es el estudio de contrates de hipótesis para determinar si una población tiene una distribución teórica específica. La técnica que nos introduce a estudiar esas cuestiones se llama Contraste de la Chi-cuadrado para la Bondad de Ajuste. Una variación de este contraste se emplea para resolver los Contrastes de Independencia. Tales contrastes pueden utilizarse para determinar si dos características (por ejemplo preferencia política e ingresos) están relacionadas o son independientes. Y, por último estudiaremos otra variación del contraste de la bondad de ajuste llamado Contraste de Homogeneidad. Tal contraste se utiliza para estudiar si diferentes poblaciones, son similares (u homogéneas) con respecto a alguna característica. Por ejemplo, queremos saber si las proporciones de votantes que favorecen al candidato A, al candidato B o los que se abstuvieron son las mismas en dos ciudades.

1.8 El procedimiento Prueba de la Chi-cuadrado

Hemos agrupado los procedimientos en los que el denominador común a todos ellos es que su tratamiento estadístico se aborda mediante la distribución Chi-

cuadrado. El procedimiento Prueba de Chi-cuadrado tabula una variable en categorías y calcula un estadístico de Chi-cuadrado. Esta prueba compara las frecuencias observadas y esperadas en cada categoría para contrastar si todas las categorías contienen la misma proporción de valores o si cada categoría contiene una proporción de valores especificada por el usuario.

1.9 Contraste de hipótesis no paramétrico para la independencia de los valores de una variable cualitativa

Supongamos que se dispone de información sobre una variable cualitativa, X, y se quiere comprobar si todas las categorías de la variable aparecen por igual. Es decir, se pretende comprobar si las categorías de la variable son independientes o no. El contraste de hipótesis que se debe resolver es el siguiente:

$$H_0 : \text{Las categorías de la variable X aparecen igual}$$

$$H_1 : \text{Las categorías de la variable X no aparecen igual}$$

Para resolver este contraste en R se utiliza la función chisq.test (que ya se presentó en la práctica 3). Los argumentos de esta función son:

chisq.test(x, p = rep(1/length(x), length(x)))

donde
x es un vector que recoge las frecuencias con las que aparece cada categoría de la variable.
p es un vector, de la misma dimensión que x, que recoge las proporciones que se quieren probar para cada categoría de la variable. Por defecto, se contrasta si todos los valores de la variable aparecen en la misma proporción.

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.

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
En primer lugar vamos a introducir los datos en R.

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

El contraste que se debe resolver es:

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

H_1 : 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.

```r
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:

```r
proporciones <- c(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.

```r
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.

Supuesto Práctico 9

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

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

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

Contrastar la hipótesis de que el dado está bien construido

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 Los valores del dado aparecen en la misma proporción
H_1 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 Contraste de hipótesis no parametricos para la independencia de dos variables cualitativas

Supongamos que se dispone de datos de dos variables cualitativas, X e Y, y se quiere comprobar si los valores que toma una de ellas dependen en cierta medida

de los valores que toma la otra. En tal caso, se dice que las variables X e Y son dependientes

Supuesto Práctico 10

La siguiente tabla muestra información sobre el número de ejemplares de 7 especies de peces avistados aguas arriba y aguas abajo en un río.
Contrastar, a un nivel de significación del 5%, si la especie de pez y la zona de avistamiento pueden considerarse variables independientes.

Solución
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)
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           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 La especie y la zona de avistamiento son independientes

H_1 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.

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

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

Solución
Introducimos los datos en R

```
frecuencias <- c(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 : El exito en la escuela y el sobrepeso son independientes

H_1 El exito 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.

2 Otros contrastes no paramétricos

2.1 El procedimiento Prueba binomial

El procedimiento Prueba binomial compara las frecuencias observadas de las dos categorías de una variable dicotómica con las frecuencias esperadas en una distribución binomial con un parámetro de probabilidad especificado. Por defecto, el parámetro de probabilidad para ambos grupos es 0.5. Se puede cambiar el parámetro de probabilidad en el primer grupo. Siendo la probabilidad en el segundo grupo igual a uno menos la probabilidad del primer grupo.

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
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

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

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
```

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

$H_0 : p \geq 0.9$ eltratamiento no es efectivo

$H_1 : p < 0.9$ eltratamiento 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.

2.2 Contraste de aleatoriedad. Test de Rachas

El procedimiento Prueba de Rachas contrasta la aleatoriedad de un conjunto de observaciones de una variable continua. Para ello, el test de rachas cuenta las cadenas de valores consecutivos que presenta la variable por encima y por debajo de un determinado punto de corte. Cada uno de estas cadenas recibe el nombre de racha (de ahí el nombre del contraste). Un número muy elevado o muy reducido de rachas apuntarán hacia la no aleatoriedad de los datos que componen la muestra.

Una racha es una secuencia de observaciones similares, una sucesión de símbolos idénticos consecutivos. Ejemplo: $+ + - - - + - - + + + + - - -$ (6 rachas). Una muestra con un número excesivamente grande o excesivamente pequeño de rachas sugiere que la muestra no es aleatoria.

Las hipótesis del contraste son las siguientes:

H_0 : Los datos de la muestra son aleatorios

H_1 : Los datos de la muestra no son aleatorios

Para resolver el contraste con R se utiliza la función runs.test del paquete randtests. De manera que el primer paso es instalar y cargar este paquete.

```
install.packages("randtests")
```

```
## Installing package into 'C:/Users/pc 1/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
## Error in contrib.url(repos, "source"):  trying to use CRAN without
setting a mirror
```

```
ibrary(randtests)
```

```
## Error in ibrary(randtests):  no se pudo encontrar la función "ibrary"
```

A continuación, ya podemos realizar la llamada a la función runs.test. Sus argumentos son los siguientes:

uns.test (x, alternative = "two.sided", threshold, plot)

donde

- **x** es un vector numérico que contiene las observaciones de la variable continua

- **alternative** indica el tipo de la hipótesis alternativa. Puede tomar los valores "two.sided" (hipótesis alternativa bilateral, del tipo $\neq$), que es el valor por defecto; "left.sided" (hipótesis alternativa unilateral, del tipo ¡) o "right.sided" (hipótesis alternativa unilateral, del tipo ¿).
threshold es un valor numérico que indica el punto de corte a partir del cual se transformarán los valores del vector numérico en valores dicotómicos.
plot es un valor lógico que indica si se incluye un gráfico en la salida o no.

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
Formulamos el contraste que debemos resolver.

H_0 : Los datos de la muestra son aleatorios

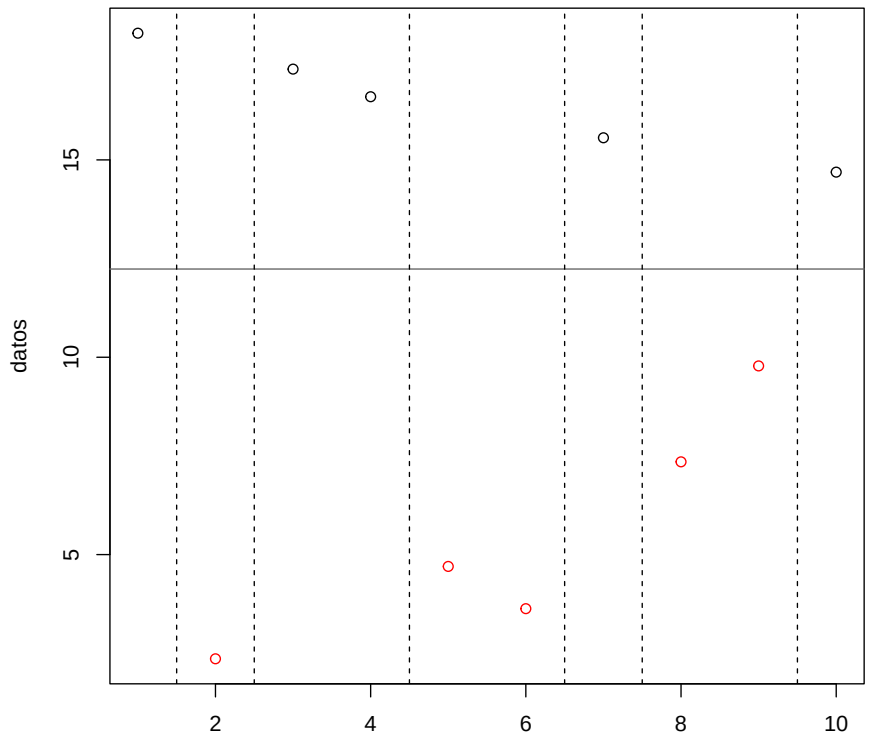
H_1 : Los datos de la muestra no son aleatorios

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)
```

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 (n1), marcados en negro, y otros 5 valores por debajo de la mediana (n2), 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.

2.3 Contraste sobre bondad de ajuste: Procedimiento Prueba de Kolmogorov-Smirnov

Mediante el contraste de bondad de ajuste de Kolmogorv-Smirnov se prueba si los datos de una muestra proceden, o no, de una determinada distribución de probabilidad. Lo que se hace es comparar la función de distribución acumulada que se calcula a partir de los datos de la muestra con la función de distribución acumulada teórica de la distribución con la que se compara.

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
El contraste de hipótesis que se plantea es el siguiente:

H_0 : Los datos de la muestra proceden de una distribución N(40,3)

H_1 : Los datos de la muestra no proceden de 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)
mean(datos)

## [1] 39.491
```

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 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")

##
##  Exact 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.

2.4 Pruebas para dos muestras independientes

El procedimiento Pruebas para dos muestras independientes compara dos grupos de casos existentes en una variable y comprueba si provienen de la misma población (homogeneidad). Estos contrastes, son la alternativa no paramétrica de los tests basados en el t de Student, Al igual que con el test de Student, se tienen dos grupos de observaciones independientes y se compara si proceden de la misma población.

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:

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

Solución
Comenzamos introduciendo los datos de ventas de los dos vendedores:

```
datosA <- c (10, 40, 60, 15, 70, 90, 30, 32, 22, 13)
datosB <- c (45, 60, 35, 30, 30, 15, 50, 20, 32, 9)
```

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

$H_0 : Me_A - Me_B = 0$

$H_1 : Me_A - Me_B \neq 0$

O, equivalentemente,

$H_0 : Me_A = Me_B$

$H_1 : Me_A \neq Me_B$

Vamos 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.

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

```
## Warning in wilcox.test.default(datosA, y = datosB, alternative =
"two.sided", :  cannot compute exact p-value with ties
## Warning in wilcox.test.default(datosA, y = datosB, alternative =
"two.sided", :  cannot compute exact confidence intervals with ties

##
##  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.

2.5 Pruebas para dos muestras relacionadas

Esta prueba es similar a la anterior, con la salvedad de que ahora se supone que los datos de las muestras están relacionados, es decir, no son independientes.

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

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

```r
datosAntes <- c(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)
```

El contraste que se debe resolver es el siguiente:

$H_0 : Me_{Antes} = Me_{Despues}$
$H_1 : Me_{Antes} > Me_{Despues}$

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.

```r
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.

3 Ejercicios

3.1 Ejercicios Guiados

3.1.1 Ejercicio Guiado1

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.

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

$H_0 : \mu = 20$

$H_1 : \mu \neq 20$

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,
mean(tensiones)

## [1] 20.38125

sd(tensiones)

## [1] 0.5230918
```

También indicamos el nivel de significación, μ_0 y la desviación típica poblacional de la variable que proporciona el enunciado.

```
alpha <- 0.02
mu_0 <- 20
desv_tipica <- 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)
z=(media-mu_0)/(desv_tipica/sqrt(n))
z

## [1] 3.388889

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 $Z_{1-\alpha/2}$, 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_0=20, conf.level = 0.98)

##
##  One Sample t-test
##
## data:  tensiones
## t = 155.85, df = 15, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 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).

3.1.2 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:
En este caso, el contraste que se debe resolver es:

$H_0 : \pi \geq 0.5$
$H_1 : \pi < 0.5$
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:

```r
prop.test(25, n=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