Problemas de independencia

1. Se seleccionó una muestra de 3000 naranjas de València. Cada naranja se clasificó según su color (claro, medio y oscuro) y se determinó su contenido de azúcar (dulce o no dulce). Los resultados fueron:

Color	Muy Dulce	No Dulce	Totales
Claro	1300	200	1500
Medio	500	500	1000
Oscuro	200	300	500
Totales	2000	1000	3000

   Probar la hipótesis de que la dulzura y el color son independientes.

Solución

En primer lugar definimos la tabla `naranjas` que contine las frecuencias empíricas:

```
(naranjas=matrix(c(1300,500,200,200,500,300),nrow=3))
```

```
##      [,1] [,2]
## [1,] 1300  200
## [2,]  500  500
## [3,]  200  300
```

A continuación realizamos el test χ^2 de independencia:

```
chisq.test(naranjas)
```

```
##
##  Pearson's Chi-squared test
##
## data:  naranjas
## X-squared = 555, df = 2, p-value < 2.2e-16
```

Como el p-valor es muy pequeño, concluimos que tenemos indicios suficientes para afirmar que la dulzura y el color no son independientes.

2. Nos dan las notas de cierta asignatura de 3 grupos de alumnos A, B i C:

A	4.6	5.	5.1	5.6	4.6	5.	5.7	5.4	4.4	8.
B	4.6	3.4	5.3	4.	3.5	4.	5.	4.7	3.6	4.1
C	7.2	7.3	5.7	4.1	5.7	6.1	6.	7.8	7.	3.8

Los clasificamos según 2 criterios: por grupo y por nota teniendo en cuenta que: **Suspenso** significa una nota más pequeña que 5 (nota < 5) y **Aprobado** significa una nota entre 5 y 6 ($5 \leq$ nota ≤ 6), **Notable** significa una nota mayor que 6 (nota > 6).
Hallar a partir del test χ^2 el p-valor para aceptar que los dos criterios son independientes

Solución

El siguiento código crea la tabla de frecuencias por grupo y por nota:

```
grupo.A = c(4.6,5,5.1,5.6,4.6,5,5.7,5.4,4.4,8)
grupo.B = c(4.6,3.4,5.3,4,3.5,4,5,4.7,3.6,4.1)
grupo.C = c(7.2,7.3,5.7,4.1,5.7,6.1,6,7.8,7,3.8)
grupo = rep(c("A","B","C"),each=10)
nota = function(x){aux=ifelse(x < 5,"Suspenso",ifelse(x <=6, "Aprobado","Notable")); return(aux)}
notas.alumnos =c(grupo.A,grupo.B,grupo.C)
tabla.notas = data.frame(sapply(notas.alumnos,nota),grupo)
names(tabla.notas)=c("Notas","Grupo")
table(tabla.notas)
```

```
##           Grupo
## Notas      A B C
##   Aprobado 6 2 3
##   Notable  1 0 5
##   Suspenso 3 8 2
```

El test χ^2 es el siguiente:

```
chisq.test(table(tabla.notas))
```

```
## Warning in chisq.test(table(tabla.notas)): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(tabla.notas)
## X-squared = 14.133, df = 4, p-value = 0.006883
```

R nos avisa que la aproximación puede ser incorrecta. Para resolver este inconveniente, simularemos el valor del p-valor:

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

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 5000
##  replicates)
##
## data:  table(tabla.notas)
## X-squared = 14.133, df = NA, p-value = 0.005799
```

El p-valor no ha variado mucho. Concluimos que como el p-valor es muy pequeño, tenemos evidencias suficientes para afirmar que el grupo del alumno y nota obtenida por el mismo no son independientes.

3. Clasificamos N individuos según dos criterios. Cada criterio tiene dos niveles. La tabla de contingencia es la siguiente:

$C_2 \backslash C_1$	A_1	A_2
B_1	10	5
B_2	5	10

   Hallar el p-valor para poder aceptar que los dos criterios son independientes usando el test χ^2.

Solución

La tabla de frecuencias empíricas es la siguiente:

```
tabla.frec.emp. = matrix(c(10,5,5,10),nrow=2)
```

El test χ^2 es el siguiente:

```
chisq.test(tabla.frec.emp.)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla.frec.emp.
## X-squared = 2.1333, df = 1, p-value = 0.1441
```

Como el p-valor es bastante grande, concluimos que no tenemos evidencias para rechazar que los dos criterios no sean independientes.

4. En un estudio se quiso contrastar si había relación entre el grupo sanguíneo de una persona y el hecho que sea o no portador de un cierto antígeno raro. Para hacerlo, se eligieron 150 portadores del antígeno y 500 no portadores y se les miró el grupo sanguíneo. Los resultados son los de la tabla siguiente:

Grupo	Portadores	No portadores
0	72	230
A	54	192
B	16	63
AB	8	15

   Usar el test χ^2 sobre estos datos para contrastar si las frecuencias de los diferentes tipos sanguíneos son diferentes en los portadores y los no portadores.

Solución

La tabla de frecuencias empíricas es la siguiente:

```
tabla.frec.emp = matrix(c(72,54,16,8,230,192,63,15),nrow=4)
```

El test χ^2 es el siguiente:

```
chisq.test(tabla.frec.emp)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla.frec.emp
## X-squared = 2.4052, df = 3, p-value = 0.4927
```

Como el p-valor es muy grande, concluimos que no tenemos evidencias para rechazar que ser portador del antígeno y el grupo sanguíneo de la persona no sean independientes.

5. En un estudio del 2003 se investigó si el historial familiar del paciente con desorden bipolar tiene influencia en la edad en que se le manifiesta este desorden. Se tomó un grupo de enfermos al azar y se anotó su historial familiar y la edad en que se manifestó la dolencia, clasificándolos en este punto en Precoces (en los cuales el desorden bipolar se manifestó antes de los 18 años) y Tardíos (en los cuales el desorden bipolar se manifestó después de los 18 años). Los resultados son los siguientes:

Historial Familiar	Precoces	Tardíos
Negativo	28	35
Desorden bipolar	19	38
Desorden unipolar	41	44
Desórdenes unipolar y bipolar	53	60

Queremos contrastar si el historial familiar influye en la edad en la cual se manifestó el desorden bipolar en el paciente. Usar el test χ^2 sobre estos datos para contrastar si las frecuencias de los diferentes tipos de desorden difieren según la edad en la cual se manifestó el desorden bipolar en el paciente.

Solución

La tabla de frecuencias empíricas junto con el test χ^2 son:

```
frbip=matrix(c(28,35,19,38,41,44,53,60),nrow=4,byrow=TRUE)
chisq.test(frbip)
```

```
##
##  Pearson's Chi-squared test
##
## data:  frbip
## X-squared = 3.6216, df = 3, p-value = 0.3053
```

Como el p-valor es grande, concluimos que no tenemos evidencias para rechazar que el historial familiar y la edad en la cual se manifestó el desorden bipolar no sean independientes.