Variables binarias

MCP

25 de febrero de 2018

Dos variables binarias

Ahora consideremos la situación en que el mismo ensayo de Bernoulli es medido en unidades que pueden ser clasificadas en grupos. El caso trivial es cuando una población consiste en dos grupos; mujeres/hombres, agua dulce/salada, empresas mexicanas/extranjeras.

Ejemplo: Lanzamiento de tiros libres de Larry Bird

En un tiro libre s'olo hay posibles resultados: encestar (éxito) o fallar (fracaso). En la NBA los tiros libres son hechos en pares; después de un primer tiro, se ejecuta un segundo intento sin importar el resultado del primero.

El ex-jugador de la NBA Larry Bird (https://es.wikipedia.org/wiki/Larry_Bird) fue uno de los más exitosos durante su carrera, con una proporción del 88.6%. En comparación, el promedio durante la época era cercana al 75% (http://www.basketball.reference.com).

Durante las temporadas 1980 – 1981 y 1981 – 1982 los registros de Bird fueron:

		Segundo		
		Éxito	Fracaso	Total
Primero	Éxito	251	34	285
	Fracaso	48	5	53
	Total	299	39	338

Podría argumentarse que los resultados del segundo tiro libre pueden depender de los resultados del primero; Si el jugador falla el primer tiro, ¿influye la frustración o determinación puede alterar la ejecución del segundo tiro? En tal caso, se esperaría que la probabilidad de éxito en el segundo intento sea diferente dependiendo de si se tuvo éxito o no en el primer intento.

Uno de los objetivos de este tema es investigar si hay o no dependencia entre las variables dadas por las filas y las variables de las columnas.

Ejemplo: Prueba clínica de Salk de una vacuna

Una de las pruebas clínicas más famosas fue hecha en 1954 por Jonas Salk. Su objetivo era determinar la eficacia de la vacuna contra la polio. Los datos obtenidos fueron:

	Polio	Sin Polio	Total
Vacuna	57	200688	200745
Placebo	142	201087	201229
Total	199	401775	401974

Se puede observar que 57 de 200745 niños en el grupo Vacuna desarrollaron polio durante el estudio, en contraste, 142 de los 201229 niños del grupo Placebo desarrollaron la enfermedad. ¿Es efectiva la vacuna contra la polio? En esta sección se tratará de desarrollar medidas de comparación para responder a esta pregunta.

Notación y modelo

Consideraremos dos variables aleatorias de Bernoulli Y_1 y Y_2, una para cada grupo. Las probabilidades de éxito para cada grupo son π_1 y π_2, respectivamente. Se observan n_j ensayos de Y_j, obteniéndose w_j éxitos

$j \in \{1, 2\}$.

$n_+ := n_1 + n_2$ es el número total de ensayos, mientras que $w_+ := w_1 + w_2$ es el número total de éxitos observados.

Podemos resumir la notación en las siguientes tablas:

		Respuesta		
		1	2	
Grupo	1	π_1	$1-\pi_1$	1
	2	π_2	$1-\pi_2$	1

		Respuesta		
		1	2	
Grupo	1	w_1	$n_1 - w_1$	n_1
	2	w_2	$n_2 - w_2$	n_2
		w_+	$n_+ - w_+$	n_+

Denotaremos a la variable que representa el número de éxitos en el grupo j por W_j y escribiremos su función de probabilidad como

$$P(W_j = w_j) = \binom{n_j}{w_j} \pi_j^{w_j} (1-\pi_j)^{n_j - w_j},\ w_j = 0, 1, \cdots, n_j,\ j = 1, 2$$

Asumiremos que Y_1 y Y_2 son independientes, *i.e* los resultados de una no influyen en los de la otra. Por ejemplo, en la prueba clínica de Salk, los niños a los que se les da vacuna no pueden pasar inmunidad o enfermedad a los niños del grupo placebo y *vice versa*. Esta suposición es crítica en el desarrollo posterior; este modelo es conocido como *modelo independiente binomial*. Cuando la independencia no se satisface, entonces otros modelos deben tener en cuenta la dependencia entra las variables aleatorias.

El método para simular este modelo en R nos será de importancia más adelante, cuando usemos estos conteos simulados para evaluar problemas de inferencia, como intervalos de confianza, para determinar si funcionan como es esperado.

Consideremos $\pi_1 = 0.2, \pi_2 = 0.4, n_1 = 10, n_2 = 10$. El siguiente código muestra cómo simular un conjunto de conteos para una tabla de contingencia.

```
pi1<-0.2
pi2<-0.4
n1<-10
n2<-10

set.seed(8191)
w1<-rbinom(n = 1, size = n1, prob = pi1)
w2<-rbinom(n = 1, size = n2, prob = pi2)

c.table<-array(data = c(w1, w2, n1-w1, n2-w2), dim = c(2,2), dimnames = list(Grupo = c(1,2),
             Respuesta = c(1, 2)))
c.table
```

```
##      Respuesta
## Grupo 1 2
##     1 1 9
##     2 3 7
```

```
c.table[1,1] # w1
```

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

```
c.table[1,2]  # n1-w1
```

```
## [1] 9
```

```
c.table[1,]  # w1 y n1-w1
```

```
## 1 2
## 1 9
```

```r
sum(c.table[1,])  # n1
```

```
## [1] 10
```

Si quiséramos repetir el proceso 1000 veces, el código sería

```r
set.seed(8191)
w1<-rbinom(n = 1000, size = n1, prob = pi1)
w2<-rbinom(n = 1000, size = n2, prob = pi2)

# Tabla 1
c.table1<-array(data = c(w1[1], w2[1], n1-w1[1], n2-w2[1]), dim = c(2,2),
                dimnames = list(Grupo = c(1,2), Respuesta = c(1, 2)))
c.table1
```

```
##      Respuesta
## Grupo 1 2
##     1 1 9
##     2 3 7
```

```r
c.table1[1,1]  # w1
```

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

```r
c.table1[1,2]  # n1-w1
```

```
## [1] 9
```

```r
c.table1[1,]  # w1 y n1-w1
```

```
## 1 2
## 1 9
```

```r
sum(c.table1[1,])  # n1
```

```
## [1] 10
```

```r
# Tabla 5
c.table5<-array(data = c(w1[5], w2[5], n1-w1[5], n2-w2[5]), dim = c(2,2),
                dimnames = list(Grupo = c(1,2), Respuesta = c(1, 2)))
c.table5
```

```
##      Respuesta
## Grupo 1 2
##     1 2 8
##     2 5 5
```

```r
pihat1<-w1/n1
mean(pihat1)  # Cercano a pi1
```

```
## [1] 0.2001
```

```r
pihat2<-w2/n2
mean(pihat2)  # Cercano a pi2
```

```
## [1] 0.4053
```

Verosimilitud y estimados

El objetivo principal es estimar las probabilidades π_1 y π_2 y compararlas. Si Y_1 y Y_2 son independientes, también lo son W_1 y W_2 y así la función de verosimilitud es

$$L(\pi_1, \pi_2 | w_1, w_2) = L(\pi_1 | w_1) \cdot L(\pi_2 | \pi_2),$$

que es maximizada por los valores $\hat{\pi_1} = w_1/n_1$ y $\hat{\pi_2} = w_2/n_1$, las proporciones muestrales.

Ejemplo: Lanzamientos de tiro libre de Larry Bird

```
# Creamos la tabla de contingencia
c.table<-array(data = c(251, 48, 34, 5), dim = c(2,2), dimnames =
                 list(Primero = c("éxito", "fracaso"),Segundo = c("éxito", "fracaso")))
list(Primero = c("éxito", "fracaso"), Segundo = c("éxito", "fracaso"))
```

```
## $Primero
## [1] "éxito"   "fracaso"
##
## $Segundo
## [1] "éxito"   "fracaso"
```

```
c.table  # Tabla de contingencia
```

```
##          Segundo
## Primero   éxito fracaso
##   éxito     251      34
##   fracaso    48       5
```

```
c.table[1,1]  # w1
```

```
## [1] 251
```

```
c.table[1,]  # w1 y n1-w1
```

```
##   éxito fracaso
##     251      34
```

```
sum(c.table[1,])  # n1
```

```
## [1] 285
```

```
# Para calcular las proporciones muestrales
rowSums(c.table)  # n1 y n2
```

```
##   éxito fracaso
##     285      53
```

```
pi.hat.table<-c.table/rowSums(c.table)
pi.hat.table
```

```
##          Segundo
## Primero       éxito    fracaso
##   éxito   0.8807018 0.11929825
##   fracaso 0.9056604 0.09433962
```

```
sum(pi.hat.table[1,])
```

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

Pero de manera frencuente los datos son presentados como medidas en cada ensayo en vez de conteos;:

```r
miss.miss<-matrix(rep(c("fracaso", "fracaso"), 5), 5,2, byrow=T)
miss.make<-matrix(rep(c("fracaso", "exito"), 48), 48,2, byrow=T)
make.miss<-matrix(rep(c("exito", "fracaso"), 34), 34,2, byrow=T)
make.make<-matrix(rep(c("exito", "exito"), 251), 251,2, byrow=T)

# Guardamos los datos sin editar en una tabla
all.data<-rbind(miss.miss, miss.make, make.miss, make.make)
all.data2<-data.frame(all.data)


# Asignamos nombres a las columnas
names(all.data2)<-c("primero", "segundo")

# Reordenamos las filas para simular una forma en como los datos pueden ser presentados
set.seed(9212)
all.data2<-all.data2[sample(x = 1:nrow(all.data2), replace = FALSE),]
row.names(all.data2)<-NULL  # Quitamos los números de fila originales
head(all.data2)
```

```
##   primero segundo
## 1   exito   exito
## 2 fracaso   exito
## 3   exito fracaso
## 4 fracaso fracaso
## 5   exito   exito
## 6 fracaso   exito
```

```r
# Primera forma de hacer la tabla de contingencia
bird.table1<-table(all.data2$primero, all.data2$segundo)
bird.table1
```

```
##
##           exito fracaso
##   exito     251      34
##   fracaso    48       5
```

```r
bird.table1[1,1]  # w1
```

```
## [1] 251
```

```r
# Segunda forma de obtener la tabla de contingencia
bird.table2<-xtabs(formula = ~ primero + segundo, data = all.data2)
bird.table2
```

```
##          segundo
## primero   exito fracaso
##   exito     251      34
##   fracaso    48       5
```

```r
bird.table2[1,1]  # w1
```

```
## [1] 251
```

```r
bird.table2/rowSums(bird.table2)
```

```
##          segundo
## primero        exito    fracaso
##   exito   0.88070175 0.11929825
```

```
##   fracaso 0.90566038 0.09433962
```