Ser o no ser: variables aleatorias discretas binarias

En los primeros números de este año hemos estudiado un tipo de variables de respuesta en las que los resultados eran recuentos, no teníamos decimales, siempre números enteros y que estaban expresadas en frecuencias, las variables de Poisson. En los siguientes números vamos a estudiar variables de respuesta aleatorias discretas que pueden ser atributos.

Morillo Alujas[1], Daniel Villalba Mata[2] y Emilio López Cano[3]

[1]Tests and Trials SLU
[2]Universidad de Lleida
[3]Universidad Rey Juan Carlos (Madrid)

Las variables de respuesta aleatorias discretas pueden expresarse por ejemplo como:

- Vivo o muerto,
- Sano o enfermo,
- Macho o hembra,
- Tose o no tose,
- Tiene diarrea o no tiene diarrea,
- Es positivo o negativo frente a una enfermedad.

Todas estas variables son variables aleatorias discretas binarias. Y aprovechamos la ocasión para recordar que una variable aleatoria es una cantidad variable que expresa los resultados que se van obteniendo en un experimento aleatorio, por ejemplo, si vamos detectando la gestación en cerdas, iremos obteniendo los valores "gestante" y "no gestante" a los que podremos asignar posteriormente los valores 1 y 0 si es necesario para su análisis. Y son discretas porque solo pueden tomar unos ciertos valores dentro de un intervalo. Si además los valores que pueden tomar se reducen a dos posibilidades, se habla de variables discretas binarias.

El resumen de las variables binarias se expresa en muchos casos en forma de proporciones. La proporción resume de una forma rápida cuántos objetos experimentales hay en una categoría, por ejemplo, muertos, machos, sanos, etc., pero también, y esto es primordial, cuántos hay en la otra categoría, por ejemplo, vivos, hembras, enfermos, etc.

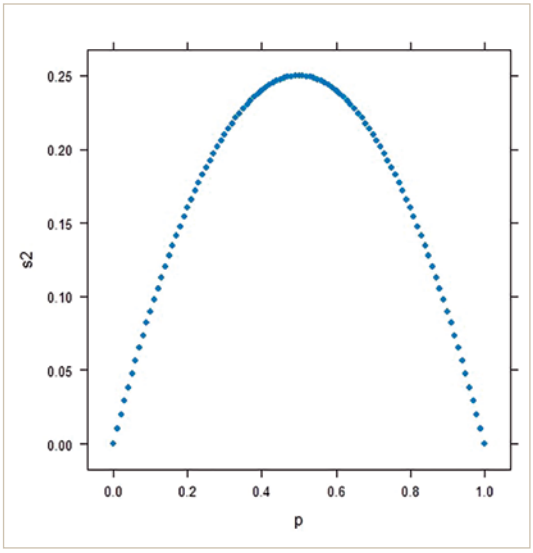

Figura 1. Gráfico de la probabilidad de éxito (p) frente a su varianza (s2).

Tenemos que usar un poco de teoría de la probabilidad antes de comenzar a analizar unos datos. Llamaremos "p" a la probabilidad de que aparezca un suceso de interés o éxito y llamaremos "q" a la probabilidad de que no aparezca. Y como recordamos que la probabilidad máxima es 1, la suma de "p+q = 1", luego si conocemos el número total de todos los elementos de un conjunto (por ejemplo, el total de cerdas a las que pasaremos el ecógrafo) y sabemos cuántas hay "gestantes" ("p", éxito o suceso favorable), restando del total, sabremos cuántas hay "no gestantes" ("q", fracaso o suceso desfavorable). Aunque es común utilizar este lenguaje de éxito/fracaso o suceso favorable/desfavorable, el suceso de interés para el análisis no tiene por qué corresponderse con el éxito empresarial. Por ejemplo, si estuviéramos interesados en estudiar la proporción de abortos, este sería el suceso de interés y nos referiríamos a él como "éxito", aunque obviamente en sentido literal no lo sea.

Así, ya tenemos que, para estudiar nuestra distribución binomial, necesitamos tres parámetros:

- El número total de datos de la muestra, "n".
- El número total de éxitos, que, al dividirlo por el número total de la muestra, tendremos la proporción de éxitos ("p").
- El número total de fracasos, que, al dividirlo por el número total de la muestra, tendremos la proporción de fracasos ("q"), que también podemos obtener mediante la sencilla operación 1-p.

Hemos dicho que necesitábamos tres parámetros, pero en realidad solo necesitamos dos, "n" y "p", ya que restando obtenemos "q".

La media de nuestra variable aleatoria discreta binomial es igual a "np". Su varianza es particular: depende de la media. De hecho, la varianza de una distribución binomial con media "np" es igual a "npq". Si lo vemos en un gráfico, será mejor. La *figura 1* muestra los valores de la varianza en el eje vertical frente a los posibles valores de "p" y un determinado valor de "n" fijo. La varianza es baja cuando "p" es o muy alta, de 0,8 en adelante, o cuando "p" es muy baja, de 0,2 hacia abajo y es máxima cuando "p", y por lo tanto, también "q" es igual a 0,5. Todos los procedimientos estadísticos que conocemos pueden usarse con nuestros datos en proporciones: el análisis de varianza y

todos los tipos de regresiones pueden usarse para modelizar estas variables.

EJEMPLO 1

El primer ejemplo que vamos a usar está extraído del libro "The R Book" de Michael J. Crawley (2013). Nos interesa el archivo "germination.txt" que nosotros ya hemos modificado convenientemente para el ejemplo. Puede descargarse en *www.testsandtrials.com/blog* con el nombre "orobanche.xlsx". Así que lo primero que haremos será activar R en nuestro ordenador. Una vez activado, cargaremos nuestro ayudante RCommander mediante la instrucción *library(Rcmdr)* e importaremos el archivo "orobanche.xlsx", yendo a Datos/Importar datos/desde un archivo Excel y le daremos el nombre "orobanche". Elegiremos nuestro archivo y veremos qué datos contiene, yendo a Estadísticos/Resúmenes/Conjunto de datos activos.

```
> summary(orobanche)
Celo            tot_nul         granja    tratamiento
 Min.   : 0.00   Min.   : 4.00   a73:10    a:10
 1st Qu.: 8.00   1st Qu.:16.00   a75:11    b:11
 Median:17.00    Median:39.00
 Mean   :20.19   Mean   :39.57
 3rd Qu.:26.00   3rd Qu.:51.00
 Max.   :55.00   Max.   :81.00
```

Los datos pueden ser el resultado de un tratamiento con dos protocolos (a y b) en 21 lotes de cerdas nulíparas que están en dos granjas diferentes (a73 y a75) y hemos estudiado su salida a celo tras un tiempo. La variable "celo" nos indica el número de cerdas nulíparas en celo y la variable "tot_nul", el número inicial de nulíparas tratadas. Nuestra hipótesis es que el tratamiento no tiene un resultado diferente en las granjas que hemos usado, es decir, no hay interacción entre tratamiento y granja.
Podemos ver también los primeros registros yendo a la pestaña "Visualizar conjunto de datos" y obtendremos el listado total del estudio (*tabla*).
Para estudiar variables binarias en R debemos indicarle en la variable respuesta el número de éxitos y el número de fracasos, es decir, "p" y "q". Esto lo haremos construyendo una variable respuesta con dos vectores. Usaremos una función llamada "cbind", que nos unirá los éxitos y fracasos en un único objeto. Con este ejemplo lo veremos muy fácilmente.
Si el número de nulíparas en celo es la variable "celo", el número de nulíparas que no salieron en celo es igual al total de tratadas menos las que salieron en celo:

```
tot_nul−celo
```

y crearemos una nueva variable que será nuestra variable de estudio y que denominaremos sencillamente "y". Lo haremos escribiendo la siguiente instrucción en la consola y ejecutándola:

```
y <- cbind(celo, tot_nul- celo)
```

Para ver cómo es esta variable, simplemente escribiremos en la consola "y" y lo ejecutaremos, obteniendo todos los pares de combinaciones de nulíparas en celo y no en celo (se muestran los primeros 6 registros solamente):

```
> y
     celo
[1,]   10 29
[2,]   23 39
[3,]   23 58
[4,]   26 25
[5,]   17 22
[6,]    5  1
```

Para estudiar nuestra hipótesis usaremos un modelo lineal generalizado. Para ello iremos a Esta-

Tabla. Resultados de salida a celo.

celo	tot_nul	granja	tratamiento
10	39	a75	a
23	62	a75	a
23	81	a75	a
26	51	a75	a
17	39	a75	a
5	6	a75	b
53	74	a75	b
55	72	a75	b
32	51	a75	b
46	79	a75	b
10	13	a75	b
8	16	a73	a
10	30	a73	a
8	28	a73	a
23	45	a73	a
0	4	a73	a
3	12	a73	b
22	41	a73	b
15	30	a73	b
32	51	a73	b
3	7	a73	b

dísticos/Ajuste de modelos/Modelo lineal generalizado, en el nombre del modelo, escribiremos "model", y especificaremos que nuestra variable respuesta es "y" en el constructor del modelo y las variables granja y tratamiento como predictoras. Como queremos ver la interacción de ambas, especificaremos que el análisis factorial se realizará mediante el símbolo "*" como puede verse en la *figura 2*.
Al aceptar, veremos una salida completa de este modelo, pero lo que nos interesa es la primera parte:

```
Coefficients:
                              Estimate Std. Error  z value  Pr(>|z|)
(Intercept)                   -0.4122     0.1842   -2.238    0.0252 *
granja[T.a75]                 -0.1459     0.2232   -0.654    0.5132
tratamiento[T.b]               0.5401     0.2498    2.162    0.0306 *
granja[T.a75]:tratamiento[T.b] 0.7781     0.3064    2.539    0.0111 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

Number of Fisher Scoring iterations: 4
```

Puede parecernos al principio que existe una interacción entre la granja y el tratamiento dado que la significación de esta es muy alta, siendo p = 0,0111. Pero debemos ver que el modelo es válido. Y para ello debemos estudiar si existe sobredispersión, que es una característica que nos indica que la variación es superior a la esperada y que podemos verla dividiendo la desviación residual (*Residual deviance*) con los grados de libertad (*degrees of freedom*). Este resultado es casi 2 (1,957588) cuando debería no ser mayor que 1.

Esto nos dice que los errores no se distribuyen siguiendo la distribución binomial, por lo que usaremos la distribución quasibinomial que añade un parámetro adicional para describir la varianza adicional de los datos que no pueden explicarse mediante la binomial.
Para ello iremos de nuevo a "Estadísticos/Ajuste de modelos/Modelo lineal generalizado", en el nombre del modelo, escribiremos "model2", y especificaremos que nuestra variable respuesta es "y" en el constructor del modelo y las variables granja y tratamiento como predictoras, pero elegiremos la familia quasibinomial como puede verse en la *figura 3*.
Y obtendremos la salida:

```
Coefficients:
                              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                   -0.4122     0.2513   -1.640    0.1193
granja[T.a75]                 -0.1459     0.3045   -0.479    0.6379
tratamiento[T.b]               0.5401     0.3409    1.584    0.1315
granja[T.a75]:tratamiento[T.b] 0.7781     0.4181    1.861    0.0801.
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " "

(Dispersion parameter for quasibinomial family taken to be 1.861832)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Y como podemos observar, la significación de la interacción entre granja y tratamiento ha desaparecido (p = 0,0801), aunque todavía mantenemos sobredispersión que trataremos de corregir, creando un tercer modelo ya sin interacción.
Dejaremos la continuación de este ejemplo para el siguiente número donde veremos también cómo obtener mucha más información en estudios de variables de respuesta binomiales.

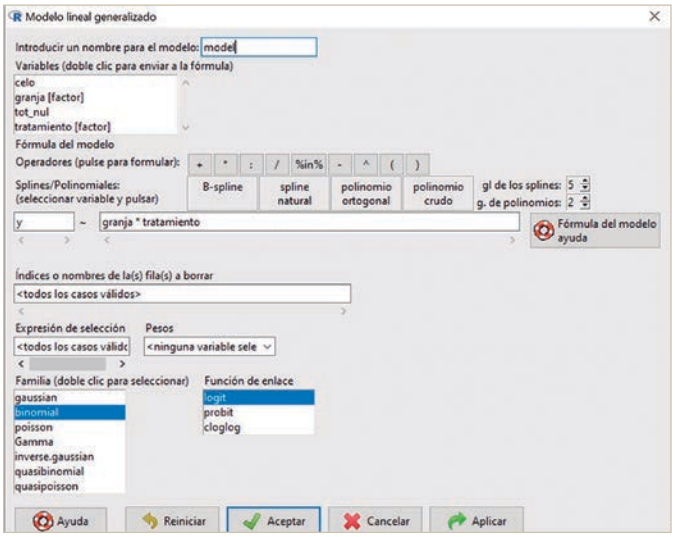

Figura 2. Construcción del modelo.

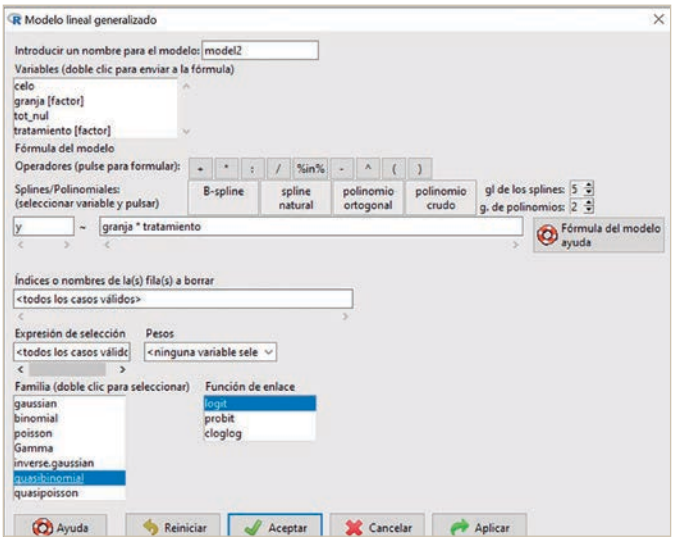

Figura 3. Construcción del modelo quasibinomial.