# Análisis de Datos Categóricos

Leticia Gracia Medrano

II

# Índice general

# Capítulo 1

# Introducción

## 1.1. Ejemplos de donde salen datos categóricos

En muchas ocasiones analizamos variables como: nacionalidad, tipo de escuela a la que asiste, marca de su preferencia, partido por el que se votaría, que son **nominales** y otras parecidas como: escolaridad, calificaciones de algún servicio, que son **ordinales**. Estas son las variables que analizaremos en este curso. En el primer caso se pueden intercambiar las categorías de las variables y no se pierde información, mientras que en el segundo caso debemos mantener las etiquetas pues éstas si mantienen un **orden**. También conviene hacer diferencia entre variable *respuesta o dependiente*($Y$)y las variables *explicativas o independientes* ($X$). En este curso el interés está puesto en las variables respuesta categóricas, las explicativas podrán ser continuas o categóricas, según sea el contexto de estudio.

## 1.2. Modelo Poisson

La distribución Poisson ayuda a modelar conteos como: número de accidentes en un tramo carretero, personas que llegan a formarse en una fila. La distribución está dada por

$$P(y) = \frac{\exp^{-\mu}\,\mu^y}{y!}$$

Por ejemplo: Si se tiene que ocurren un promedio de 2 accidentes semanales en cierto tramo carretero entonces la probabilidad de tener 0 accidentes en una semana dada es:

$$P(0) = \frac{\exp^{-2}\,2^0}{0!} = \exp^{-2} = 0{,}135$$

En esta distribución $E(Y) = Var(Y) = \mu$, aquí entonces si la media aumenta, también aumenta la varianza. En la práctica ocurre frecuentemente que los conteos tienen mayor varianza que la esperada, esto se conoce como *sobre-dispersión*. En ocasiones suponer la distribución Poisson resulta muy simplista, pero en otras resulta muy útil.

## 1.3. Modelo Binomial

En el ejemplo anterior el número de accidentes es aleatorio. Pero podría planterase algo como que se clasifican los accidentes hasta que ocurren $N$, con el propósito de estimar la proporción de estos que resultan fatales, entonces el total de accidentes es fijo. Ahora el número de accidentes fatales ya no es Poisson porque tiene un tope máximo de $N$.

Si se tienen que el número de accidentes fatales en $t$ semanas tiene una media de $2t$, y la tasa para accidentes no-fatales es de $8t$. Cuando se junta un total de $N$ accidentes ocurre que el numéro de accidentes fatales se distribuye como binomial con parámetros $N$ y $\pi = \frac{2t}{2t+8t} = ,2$, la probabilidad de cualquier

accidente resulte fatal. La función de distribución binomial recordemos está dada por:

$$P(y) = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y} \text{ con } y = 0, 1, 2, \ldots, N$$

Para el caso en que $N = 10$ y $\pi = ,2$ la probabilidad de que haya $y = 0$ accidentes es

$$P(0) = \frac{10!}{0!(10)!},2^0(,8)^{10} = (,8)^10 = ,107;$$

Para esta distribución $E(Y) = N\pi$ y $Var(Y) = N\pi(1-\pi)$, entonces aquí la varianza siempre es menor que la media.

Cuando los resultados pueden ser más que dos, se tiene una distribución *multinomial*, que se verá en el último tema.

## 1.4. Modelo Multinomial

En este caso el experimento tiene $c$ posibles resultados, sus probabilidades las denotamos por $\pi_1, \pi_2, \ldots, \pi_c$, donde $\sum \pi_j = 1$.

Para $n$ observaciones independientes, la probabilidad multinomial de que $n_1$ caigan en la categoría 1, $n_2$ caigan en la categoría 2,... y $n_c$ caigan en la categoría c, donde $\sum n_j = n$ es

$$P(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1!n_2!\ldots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\ldots\pi_c^{n_c}$$

La distribución multinomial es multivariada. La marginal para cualquier categoría es binomial. Para la categoría $j$ el conteo $n_j$ tiene media $n\pi_j$ y desviación estándar $\sqrt{n\pi_j(1-\pi_j)}$

## 1.5. Inferencia sobre $\pi$

En el modelo binomial el parámetro es $\pi$ que generalmente es desconocido y a través de una muestra trataremos de estimarlo.

La probabilidad de los datos observados, expresada como una función del parámetro, es la función de verosimilitud. Para una $n = 10$ y una $y = 0$ la función de verosimilitud es:

$$P(0) = (10!/0!10!)\pi^0(1-\pi)^{10} = (1-\pi)^{10}$$

$$\ell(\pi) = (1-\pi)^{10}$$

que alcanza el máximo cuando $\pi = 0$, entonces el resultado $y = 0$ ocurre con una mayor probabilidad cuando $\pi = 0$. OJO esta función depende de los valores que tome $\pi$, para cada valor de $y$ se tiene una función de verosimilitud distinta, así para $y = 6$

$$\ell(\pi) = (10!/6!4!)\pi^6(1-\pi)^4 = 210 * \pi^6(1-\pi)^4$$

cuyo máximo lo alcanza en $\pi = ,6$, en este caso $y = 6$ ocurre con una mayor probabilidad cuando $\pi = ,6$

Se tiene que el estimador máximo verosímil para $\pi$ es $\hat{p} = \sum_{i=1}^{n} x_i/n$ donde $x_i$ es 1 o 0 según se observe éxito o fracaso. Entonces $\hat{p}$ es un promedio por lo que se puede utilizar el Teorema Central del Límite. Para una $n$ grande $y = \bar{x}$ se distribuye aproximadademnete como una normal con media $E(\hat{p}) = \sum E(x_i)/n = n\pi/n = \pi$ y varianza $var(\hat{p}) = \frac{\sum var(xi)}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \pi(1-\pi)/n$ .

Se puede usar la estadística

$$z = \frac{\hat{p} - \pi_o}{\sqrt{\pi_o(1-\pi_o)/n}}$$

para probar la hipótesis nula $H_o : \pi = \pi_o$.

Y se puede construir un intervalo de $100(1-\alpha)$ de confianza para $\pi$ con

$$\hat{p} \pm z_{\alpha/2}SE, \text{con } SE = \sqrt{\hat{p}(1-\hat{p})/n}$$

donde $z_{\alpha/2}$ es el percentil que deja a la derecha una cola tamaño $\alpha/2$.

Hay que tener cuidado, esta aproximación es buena cuando $\pi$ está cerca de 0.5 o cuando la $n$ es muy grande. Si no se tiene eso, el nivel de confianza disminuiría. Y es muy mala cuando $\pi$ se acerca a uno o al cero.

Que más se podría hacer?? Si existe una correspondencia entre pruebas de hipótesis e intervalos de confianza, que dice que el intervalo de confianza es aquel donde la hipótesis nula No se rechaza, o sea la región no crítica. Entonces deben hallarse los valores de $\pi_o$ de manera que:

$$\frac{|\hat{p} - \pi_o|}{\sqrt{(}\pi_o(1 - \pi_o/n)} = 1{,}96 = z_{,05/2}$$

Coooomo??? pues elevando al cuadrado ambos lados de la ecuación anterior y de la que resulta una ecuación cuadrática en $\pi_o$ Para el caso de $\hat{p} = 0{,}90$ y $n = 10$ usando esto se encuentra las raíces $\pi_{o1} = {,}596$ y $\pi_{o2} = {,}982$, que darían un intervalo (.596,.982), mientras que usando la ecuación ■tradicional■ se tendría un intervalo al 90 % de confianza (0.714,1.086).

Otra aproximación es la llamada de Agresti-Coull que es muy fácil, se suman 2 a los éxitos y se suman 2 a los fracasos y entonces ... usando el ejemplo anterior

$\hat{p} = (9 + 2)/(10 + 2 + 2) = {,}786$ y $SE = (,786)(,214)/14 = {,}110$ con lo que se obtiene un intervalo de (.57,1.0).

Este método funciona bien aún con muestras pequeñas.

# Capítulo 2

# Tablas de contingencia

## 2.1. Notación

Cuando se tienen dos criterios o variables de clasificación de las observaciones, al hacer el "cruce" de éstas se genera una tabla de frecuencias como sigue:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1J} \\ n_{21} & n_{21} & \cdots & n_{2j} & \cdots & n_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{iJ} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{I1} & n_{I2} & \cdots & n_{Ij} & \cdots & n_{IJ} \end{bmatrix}$$

Donde la variable renglón tiene $I$ categorías y la variables columna tiene $J$ categorías.

Los totales por columna son:

$$\sum_{i=1}^{I} n_{ij} = n_{\cdot j}$$

los totales por renglón:

$$\sum_{j=1}^{J} n_{ij} = n_{i.}$$

el total general :

$$\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} = n$$

y las frecuencias relativas: $r_i = n_{i.}/n$ y $c_j = n_{.j}/n$

Con $i = 1, \cdots, I$ y $j = 1, \cdots, J$

## 2.2. Esquemas de muestreo

En el análisis de tablas de contingencia de dos dimensiones se utilizan tres esquemas de muestreo que ocurren en la práctica: Esquema Poisson, Esquema Multinomial, y Esquema Multinomial-Producto.

### 2.2.1. Esquema Poisson

Supone que se observa cualquier cantidad de datos $\{n_{ij}\}$ durante un intervalo de tiempo. La distribución marginal de las observaciones $n_{ij}$ es $Poisson(m_{ij})$, para $i = 1, \ldots, I,\ j = 1, \ldots, J$. Las probabilidades marginales son

$$P(n_{ij}|m_{ij}) = e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!},$$

donde $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n_{..}$, $E(n_{ij}) = m_{ij}$.

La función de verosimilitud queda expresada como

$$L(m) = L(\{m_{ij}\}) = \prod_{i=1}^{I}\prod_{j=1}^{J} e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!}.$$

### 2.2.2. Esquema Multinomial

Este esquema supone que el número total de observaciones $n_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ es fijo. Se tienen $I \times J$ categorías con probabilidad $Multinomial(n_{..}, \{p_{ij}\})$, en

donde las probabilidades están dadas por

$$P(n_{ij}|m_{ij}) = \binom{n_{..}}{n_{11}n_{12}\cdots n_{IJ}} \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{n_{ij}} = \frac{n_{..}!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{n_{ij}},$$

donde $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n_{..}$, $\prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij} = 1$, $m_{ij} = E(n_{ij}) = n_{..}p_{ij}$, y $p_{ij} = \frac{m_{ij}}{n_{..}}$. Además $m_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} m_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{..}p_{ij} = n_{...}$

La función verosimilitud es

$$\begin{aligned}
L(m) &= L(\{m_{ij}\}) \\
&= \frac{n_{..}!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{n_{ij}} \qquad \text{dado que} \quad p_{ij} = \frac{m_{ij}}{n_{..}} \\
&= \frac{n_{..}!}{\prod_{ij} n_{ij}!} \prod_{ij} \left(\frac{m_{ij}}{n_{..}}\right)^{n_{ij}} \\
&= \frac{n_{..}!}{\prod_{ij} n_{ij}!} \prod_{ij} \left(\frac{m_{ij}}{n_{..}}\right)^{n_{ij}} \frac{e^{-\sum_{ij} m_{ij}}}{e^{-m_{..}}} \qquad \text{dado que} \quad n_{..} = m_{..} \\
&= \frac{n_{..}!}{e^{-n_{..}} \prod_{ij} n_{..}^{n_{ij}}} \prod_{ij} \left(e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!}\right) \\
&\propto \prod_{ij} e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!},
\end{aligned}$$

que indica que el esquema Multinomial es equivalente al esquema Poisson.

### 2.2.3. Esquema Multinomial-Producto

En este esquema se supone que los totales marginales por renglón (o columna) $n_{1\cdot}, n_{2\cdot}, \ldots, n_{I\cdot}$ (o $n_{\cdot 1}, n_{\cdot 2}, \ldots, n_{\cdot J}$) están fijos. Para estos, en cada renglón se tiene una distribución multinomial. Entonces, para el renglón $i$, $i = 1, \ldots, I$, las probabilidades son

$$P(n_{ij}|m_{ij}) = \binom{n_{i\cdot}}{n_{i1}n_{i2}\cdots n_{iJ}} \prod_{j=1}^{J} p_{j|i}^{n_{ij}} = \frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{j|i}^{n_{ij}},$$

donde $\sum_{i=1}^{I} n_{i\cdot} = n_{..}$, $\prod_{j=1}^{J} p_{j|i} = 1$, $m_{ij} = E(n_{ij}) = n_{i\cdot}p_{j|i}$, y $p_{j|i} = \frac{m_{ij}}{n_{i\cdot}}$.

Además. La verosimilitud es

$$\begin{aligned}
L(m) &= L(\{m_{ij}\}) \\
&= \prod_{i=1}^{I}\left[\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!}\prod_{j=1}^{J} p_{j|i}^{n_{ij}}\right] \qquad \text{dado que} \quad p_{j|i} = \frac{m_{ij}}{n_{i\cdot}} \\
&= \frac{\prod_i n_{i\cdot}!}{\prod_{ij} n_{ij}!}\prod_{ij}\left(\frac{m_{ij}}{n_{i\cdot}}\right)^{n_{ij}} \\
&= \frac{\prod_i n_{i\cdot}!}{\prod_{ij} n_{ij}!}\prod_{ij}\left(\frac{m_{ij}}{n_{i\cdot}}\right)^{n_{ij}}\frac{e^{-\sum_{ij} m_{ij}}}{e^{-m_{\cdot\cdot}}} \qquad \text{dado que} \quad n_{\cdot\cdot} = m_{\cdot\cdot} \\
&= \frac{\prod_i n_{i\cdot}!}{e^{-n_{\cdot\cdot}}\prod_{ij} n_{i\cdot}^{n_{ij}}}\prod_{ij}\left(e^{-m_{ij}}\frac{m_{ij}^{n_{ij}}}{n_{ij}!}\right) \\
&\propto \prod_{ij} e^{-m_{ij}}\frac{m_{ij}^{n_{ij}}}{n_{ij}!},
\end{aligned}$$

que indica que el esquema Multinomial-Producto es equivalente al esquema Poisson.

### 2.2.4. Hipótesis de no asociación

La hipótesis nula en la Ji-Cuadrada ($\chi^2$) de no asociación corresponde a "diferentes interpretaciones" para cada esquema de muestreo:

- Esquema Poisson: No Asociación (las variables no están relacionadas).
- Esquema Multinomial: Independencia (la probabilidad conjunta es el producto de las probabilidades marginales).
- Esquema Multinomial-Producto: Homogeneidad (la distribución es la misma en cada renglón).

## 2.3. Prueba de Independencia

La pregunta que surge es si las dos variables son **independientes**, es decir si los datos se acomodan en la tabla de manera proporcional al total de los

renglones y el total de las columnas. Si los datos no se acomodan de manera proporcional diremos que ciertas categorías de las variables están **asociadas**. Es claro que las proporciones no resultan exactas sino que hay variaciones aleatorias, pero si las diferencias son muy grandes con esas proporciones "esperadas", se dirá que las variables no son independientes.

Si en la población de la que se saca la muestra, la probabilidad de que una observación pertenezca a la celda $i, j$ se llama $p_{i,j}$, entonces la frecuencia esperada $F_{i,j}$ de observaciones para esa celda, luego de sacar una muestra tamaño $N$ es $F_{ij} = n \times p_{i,j}$.

Ahora, si $p_{i.}$ es la probabilidad de pertenecer al renglón $i$ y $p_{.j}$ es la probabilidad de pertenecer a la columna $j$, cuando las variables son independientes ocurre que $p_{ij} = p_{i.} \times p_{.j}$

Entonces las frecuencias esperadas cuando las variables son independientes son:

$$F_{ij} = N \times p_{i.} \times p_{.j}$$

Estas probabilidades no se conocen, pero pueden ser estimadas con

$$\hat{p}_{i.} = \frac{n_{i.}}{n}$$

y

$$\hat{p}_{.j} = \frac{n_{.j}}{n}$$

y las frecuencias esperadas se estiman con

$$E_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

Hay ocasiones que los totales por renglón son fijos, esto por el diseño del muestreo, si la variables respuesta $Y$ es binaria se tiene un modelo binomial, si tiene más categorías se tiene un esquema multinomial. Y en ese caso nos fijamos en las distribuciones condicionales para cada nivel dela variable $X$. Aquí la independencia entre $X$ y $Y$ puede expresarse también como que las

distribuciones condicionales de $Y$ para cada nivel de la variable $X$ son las mismas.

Otra situación es cuando $n$ es fija y clasificamos a los individuos al ■cruzar■ las dos variables respuesta, en ese caso se tien una distribución multinomial con $I \times J$ categorías.

## 2.4. La $\chi^2$

Para analizar si las variables son independientes se puede utilizar la estadística $\chi^2$, dada por

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Cuando esta estadística toma valores "grandes" se rechaza la hipótesis nula de independencia.

La distribución asintótica de esta estadística puede hallarse suponiendo que las frecuencias observadas siguen una distribución multinomial y que las frecuencias esperadas no son muy pequeñas, y corresponde a una distribución **ji cuadrada** con $(I-1)*(J-1)$ grados de libertad.

Muchos paquetes trabajan con la Ji Cuadrada con corrección de Yates, esto es:

$$\chi^2 = \sum_{ij} [|o_{ij} - e_{ij}| - 0{,}5]^2 / e_{ij}$$

## 2.5. Hago una tabla de 2 por 2 en R

### 2.5.1. Ejemplo 1

Haciendo esto muy rápido nada más doy los datos (pero no se ve bonito).

```
> v1 <- matrix(c(25, 11, 12, 14), ncol = 2)

> chisq.test(v1)

        Pearson s Chi-squared test with Yates continuity correction

data:  v1
X-squared = 2.5041, df = 1, p-value = 0.1135
```

En este caso para un $\alpha = ,10$ la prueba resulta no significativa, es decir no rechazamos $H_0$, es decir no hay evidencia para decir que no son independientes

Si desean hacer una salida más bonita ponen:

```
> FUMA <- c("NO FUMA", "NO FUMA", "SI FUMA", "SI FUMA")
> GENERO <- c("FEM", "MASC", "FEM", "MASC")
> conteos <- c(25, 11, 12, 14)
> TABLA <- data.frame(FUMA, GENERO, conteos)
> xtabs(conteos ~ FUMA + GENERO, data = TABLA)

          GENERO
FUMA       FEM MASC
  NO FUMA  25   11
  SI FUMA  12   14

> chisq.test(xtabs(conteos ~ FUMA + GENERO, data = TABLA))
```

```
        Pearson s Chi-squared test with Yates continuity correction

data:  xtabs(conteos ~ FUMA + GENERO, data = TABLA)
X-squared = 2.5041, df = 1, p-value = 0.1135
```

### 2.5.2. Ejemplo 2

La proporción de niños de bajo peso al nacer es la misma en las mujeres que fuman, que en las que no fuman?

```
> FUMA <- c("NO FUMA", "NO FUMA", "SI FUMA", "SI FUMA")
> BAJOPESO <- c("SI", "NO", "SI", "NO")
> conteos <- c(105, 1645, 43, 207)
> TABLA <- data.frame(FUMA, BAJOPESO, conteos)
> xtabs(conteos ~ FUMA + BAJOPESO, data = TABLA)

          BAJOPESO
FUMA         NO   SI
  NO FUMA 1645  105
  SI FUMA  207   43

> chisq.test(xtabs(conteos ~ FUMA + BAJOPESO, data = TABLA))

        Pearson s Chi-squared test with Yates  continuity correction

data:  xtabs(conteos ~ FUMA + BAJOPESO, data = TABLA)
X-squared = 38.4266, df = 1, p-value = 5.685e-10
```

La prueba es altamente significativa. Decimos que la proporción de niños de bajo peso es estadísticamente diferente en el grupo de las mamás que fuman (43/250=.172) que en las que no fuman (105/1645=0.0638) . .

### 2.5.3. Ejemplo 3

La hipótesis que se desea probar con los datos de la siguiente tabla, es si el tipo de tuberculosis por el que la persona muere es independiente del género

```
> tipotuberculosis <- c("resp", "otra", "resp", "otra")
> genero <- c("m", "m", "f", "f")
> conteos <- c(3534, 270, 1319, 252)
> TABLA <- data.frame(tipotuberculosis, genero, conteos)
> xtabs(conteos ~ tipotuberculosis + genero, data = TABLA)

               genero
tipotuberculosis    f    m
            otra  252  270
            resp 1319 3534

> tabla3 <- xtabs(conteos ~ tipotuberculosis + genero, data = TABLA)
Para calcular los marginales por renglon
> margin.table(tabla3, 1)

tipotuberculosis
otra resp
 522 4853


Para calcular los marginales por columna
> margin.table(tabla3, 2)

genero
   f    m
1571 3804
```

```
Para pedir el resumen de la tabla
> summary(tabla3)

Call: xtabs(formula = conteos ~ tipotuberculosis + genero, data = TABLA)
Number of cases in table: 5375
Number of factors: 2
Test for independence of all factors:
        Chisq = 101.41, df = 1, p-value = 7.483e-24
```

Con estos valores se concluye que las variables no son independientes, es decir que la proporción de hombres que muere por tuberculosis de tipo respiratorio 3534/3804=.929 es significativamente diferente de la proporción 1319/1571=.840 de mujeres que mueren por ese tipo de tuberculosis.

ENCONTRAR ASOCIACION A TRAVES DE LA JI CUADRADA NO IMPLICA NECESARIAMENTE NINGUNA RELACION CAUSAL.

## 2.6. Más de la Ji Cuadrada

### 2.6.1. Sensibible al tamaño de muestra

```
> abortoafavor <- c("si", "si", "no", "no")
> raza <- c("b", "n", "b", "n")
> conteos <- c(49, 51, 51, 49)
> TABLA <- data.frame(abortoafavor, raza, conteos)
> tabla4 <- xtabs(conteos ~ abortoafavor + raza, data = TABLA)
> summary(tabla4)

Call: xtabs(formula = conteos ~ abortoafavor + raza, data = TABLA)
Number of cases in table: 200
Number of factors: 2
```

```
Test for independence of all factors:
        Chisq = 0.08, df = 1, p-value = 0.7773
> abortoafavor <- c("si", "si", "no", "no")
> raza <- c("b", "n", "b", "n")
> conteos <- c(98, 102, 102, 98)
> TABLA <- data.frame(abortoafavor, raza, conteos)
> tabla4 <- xtabs(conteos ~ abortoafavor + raza, data = TABLA)
> summary(tabla4)
Call: xtabs(formula = conteos ~ abortoafavor + raza, data = TABLA)
Number of cases in table: 400
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.16, df = 1, p-value = 0.6892
> abortoafavor <- c("si", "si", "no", "no")
> raza <- c("b", "n", "b", "n")
> conteos <- c(4900, 5100, 5100, 4900)
> TABLA <- data.frame(abortoafavor, raza, conteos)
> tabla4 <- xtabs(conteos ~ abortoafavor + raza, data = TABLA)
> summary(tabla4)
Call: xtabs(formula = conteos ~ abortoafavor + raza, data = TABLA)
Number of cases in table: 20000
Number of factors: 2
Test for independence of all factors:
        Chisq = 8, df = 1, p-value = 0.004678
```

El valor de la ji cuadrada queda multiplicado por la constante que multiplique las entradas de la tabla. Los grados de libertad de la ji cuadrada no se modifican, y por tanto con una $n$ muy grande esta prueba resulta significativa.

## 2.7. Comparación de proporciones en tablas de 2 por 2

Notación para una tabla de 2 por 2

| | | |
|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| $n_{21}$ | $n_{21}$ | $n_{2.}$ |
| $n_{,1}$ | $n_{,2}$ | $n_{..}$ |

La estimación de las proporciones está dada por:

$\hat{\pi}_1 = p_1 = n_{11}/n_{1.}$ y $\hat{\pi}_2 = p_2 = n_{21}/n_{2.}$

## 2.8. Diferencia de proporciones

Para los sujetos en la primera fila se tiene que la probabilidadd éxito es $\pi_1$ y para la fila 2 es $\pi_2$, si comparamos $\pi_1 - \pi_2$, las versiones muestrales son $p_1$ y $p_2$ cuando las muestras son de tamaño $N_1$ y $N_2$, del curso propedéutico recordamos que para muestras grandes se tiene que:

$$\hat{\sigma}_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$$

Y el intervalo de $(1-\alpha)\,\%$ de confianza para $\pi_1 - \pi_2$ es

$$(p_1 - p_2) \pm z_{\alpha/2}\hat{\sigma}_{p_1-p_2}$$

## 2.9. Riesgo Relativo

Tal vez ocurre que una diferencia entre proporciones sea más importante cuando se está en los extremos, cerca de 0 o de 1 que cuando se está en el centro. La diferencia entre .010 y .001 es la misma que entre .410 y .401, pero

la primera diferencia es más fuerte pues una es 10 veces la otra, entonces es preferible que consideremos el cociente de proporciones.

En tablas de $2 \times 2$ el *riesgo relativo* es **el cociente de las probabilidades de éxito en los dos grupos**,

$$\frac{\pi_1}{\pi_2}$$

Para el ejemplo anterior se tiene:

$$r.r_{caso1} = \frac{0{,}010}{0{,}001} = 10$$
$$r.r_{caso2} = \frac{{,}410}{{,}401} = 1{,}02$$

OJO hay que definir aquí cual es la variable respuesta para definir la probabilidad de éxito. Es decir qué ponemos como columna y qué como renglón.

Se tiene un riesgo relativo igual a 1 cuando $\pi_1 = \pi_2$, es decir cuando la respuesta es independiente del grupo.

## 2.10. Cociente de Momios

Para la fila 1 el *momio* está dado por: $\text{momio}_1 = \frac{\pi_1}{(1-\pi_1)}$ y para la fila 2 por $\text{momio}_2 = \frac{\pi_2}{(1-\pi_2)}$. Así si $\pi_1 = {,}75$ entonces el momio es .75/.25=3. Entonces si el momio=4, el éxito es cuatro veces más probable que un fracaso. Esperamos ver 4 éxitos por cada fracaso.

Si despejamos la probabilidadse tiene que $\pi = \frac{\text{momio}}{\text{momio}+1}$, si momio=4 entonces $\pi = 4/(4+1) = {,}8$

Ahora el cociente de momios se define como:

$$\theta = \frac{\text{momio}_1}{\text{momio}_2} = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_2}{(1-\pi_2)}}$$

Este NO es un cociente de probabilidades como en el riesgo relativo.

Si $X$ y $Y$ son independientes $\pi_1 = \pi_2$, $\text{momio}_1 = \text{momio}_2$ y también $\theta = \text{momio}_1/\text{momio}_2 = 1$. Cuando $1 < \theta < \infty$ los momios de éxito son mayores

en la fila 1 que en la 2, es decir $\pi_1 > \pi_2$, cuando $0 < \theta < 1$ un éxito es menos probable en la fila 1 que en la 2, es decir $\pi_1 < \pi_2$.

Cuando $\theta$ se aleja del 1, ya sea hacia arriba o hacia abajo, representa mayores niveles de asociación. Una $\theta = 4$ está más lejos de la independencia que una $\theta = 2$, lo mismo una $\theta = ,25$ está más lejos de la independencia que una $\theta = ,50$.

ATENCION!!!!!!!!!!!!

Si se intercambian las filas y se tenía una $\theta = 4$ ahora se tendrá una $\theta = ,25$, lo mismo ocurre si se voltean las columnas.

La $\theta$ no cambia si la tabla se presenta traspuesta, es decir las columnas son los renglones y los renglones son columnas. Como las trata de manera simétrica, no importa cual variable es considerada como respuesta. OJO en riesgo relativo si importa.

El estimador de $\theta$ está dado por

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Ejemplo de Infarto

```
> infarto <- c("isi", "isi", "no", "no")
> grupo <- c("placebo", "taspirin", "placebo", "taspirin")
> conteos <- c(189, 104, 10845, 10933)
> TABLA <- data.frame(infarto, grupo, conteos)
> tabla4 <- xtabs(conteos ~ grupo + infarto, data = TABLA)
> tabla4

          infarto
grupo        isi    no
  placebo    189 10845
  taspirin   104 10933
```

```
> margin.table(tabla4, 1)

grupo
 placebo taspirin
   11034    11037

> margin.table(tabla4, 2)

infarto
  isi    no
  293 21778

> summary(tabla4)

Call: xtabs(formula = conteos ~ grupo + infarto, data = TABLA)
Number of cases in table: 22071
Number of factors: 2
Test for independence of all factors:
        Chisq = 25.014, df = 1, p-value = 5.692e-07
```

Claramente se ve que no son independientes, pero para dónde están jalando las cosas.

Momio para el grupo con placebo:

```
> 189/10845

[1] 0.01742739
```

Momio pra el grupo con aspirina

```
> 104/10933

[1] 0.009512485
```

Cálculo de Cociente de momios:

```
> (189/10845)/(104/10933)
```

```
[1] 1.832054
```

Los momios pues son 83 % más grandes para el grupo placebo.

## 2.11. Residuales

Una forma de ver qué categorías son las que provocan la asociación de las variables es fijarse en aquellas que tengan los más grandes *residuales ajustados* dados por:

$$\frac{(n_{ij} - E_{ij})}{\sqrt{E_{ij} * (1 - p_{i.})(1 - p_{.j})}}$$

## 2.12. Ejemplo

Ejemplo con una tabla de 2x3.

```
> partido <- c("democ", "democ", "independ", "independ", "republic", "republi
> genero <- c("fem", "masc", "fem", "masc", "fem", "masc")
> conteos <- c(279, 165, 73, 47, 225, 191)
> TABLA <- data.frame(partido, genero, conteos)
> TABLA <- xtabs(conteos ~ genero + partido, data = TABLA)
> TABLA
```

```
      partido
genero democ independ republic
  fem    279       73      225
  masc   165       47      191
```

Para estos datos calculamos las marginales por renglon y columna y el total.

```
> margin.table(TABLA, 1)

genero
 fem masc
 577  403

> margin.table(TABLA, 2)

partido
   democ independ republic
     444      120      416

> sum(TABLA)

[1] 980
```

Se calcula la prueba ji cuadrada, y podemos accesar a los valores esperados y los residuales como se muestra

```
> prueba <- chisq.test(xtabs(conteos ~ genero + partido, data = TABLA))
> prueba

        Pearson s Chi-squared test

data:  xtabs(conteos ~ genero + partido, data = TABLA)
X-squared = 7.0095, df = 2, p-value = 0.03005

> prueba$expected

      partido
genero    democ independ republic
  fem  261.4163 70.65306 244.9306
  masc 182.5837 49.34694 171.0694
```

```
> prueba$p.value

[1] 0.03005363

> prueba$residuals

      partido
genero      democ   independ   republic
  fem   1.0875350  0.2792134 -1.2735005
  masc -1.3013036 -0.3340963  1.5238229
```

Para calcular los residuales ajustados debemos hacerle un pequeño ajuste

```
> proprenglon <- (margin.table(TABLA, 1)/sum(TABLA))
> proprenglon

genero
      fem      masc
0.5887755 0.4112245

> propcol <- (margin.table(TABLA, 2)/sum(TABLA))
> propcol

partido
    democ  independ  republic
0.4530612 0.1224490 0.4244898

> auxiliar <- as.matrix(1 - proprenglon) %*% t(1 - as.matrix(propcol))
> resajustados <- matrix(nr = 2, ncol = 3)

> for (i in 1:2) {
+     for (j in 1:3) {
+         resajustados[i, j] <- (prueba$residuals[i, j])/sqrt(auxiliar[i, j]
+     }
+ }
```

```
> resajustados

          [,1]       [,2]     [,3]
[1,]  2.293160  0.4647941 -2.61778
[2,] -2.293160 -0.4647941  2.61778

> p1x <- 577/980
> thetademrepublic <- (279 * 191)/(225 * 165)
> thetademrepublic

[1] 1.435394
```

Los residuales ajustados se muestran grandes en las mujeres demócratas y en los hombres republicanos, esto se muestra también en la $\theta = 1{,}43$, entonces decimos que los momios de identificarse con los demócratas en vez de los republicanos son 44 % más grandes en las mujeres que en los hombres.

## 2.13. Intervalo de confianza para $\log(\theta)$

La distribución muestral del riesgo relativo y cociente de momios es muy asimétrica, debido a esto se usa la función $\log(\theta)$. Esta función resulta simétrica alrededor del cero en el sentido de que si se invierte el orden de las columnas o renglones entonces por ejemplo $\log(2{,}0) = 0{,}7$ (al voltear los renglones) se tendría $\log(0{,}5) = -{,}7$, digamos que representa el mismo nivel de asociación. El doblar el logaritmo de cocientes de momios representa elevar al cuadrado el cociente de momios. La distribución de $\log(\theta)$sigue siendo asimétrica pero muchos más cercana a la normal.

Para una muestra grande tiene media $\log(\theta)$ y una desviación estándar asintótica de:

$$ASE(\log(\hat{\theta})) = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$$

Entonces el intervalo de confianza para $\log(\theta)$ es de la forma:

$$\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log(\hat{\theta}))$$

Si se exponencían los extremos del intervalo se obtiene un intervalo de confianza para $\theta$

Para calcular el intervalo de confianza para $\log(\theta)$ en el ejemplo anterior hacemos:

```
> ase <- sqrt(1/189 + 1/10933 + 1/10845 + 1/104)
> ase

[1] 0.1228416

> log(1.823) - ase * qnorm(0.975)

[1] 0.3597183

> log(1.823) + ase * qnorm(0.975)

[1] 0.8412487

> exp(log(1.823) - ase * qnorm(0.975))

[1] 1.432926

> exp(log(1.823) + ase * qnorm(0.975))

[1] 2.319261
```

Entonces, el intervalo [1.43,2.31] no contien al 1, los momios son diferentes para cada grupo, viendo el límite inferior del intervalo se tiene que LOS MOMIOS DE INFARTO AL MIOCARDIO SON AL MENOS 43 % MAS ALTOS EN EL GRUPO DE PLACEBO QUE EN EL GRUPO DE ASPIRINA.

El intervalo NO es simétrico. Si una celda es cero la $\hat{\theta}$ está indefinida o es cero. Si se usa este otro estimador no se tendría ese problema:

$$\tilde{\theta} = \frac{(n_{11} + ,5)(n_{22} + ,5)}{(n_{12} + ,5)(n_{21} + ,5)}$$

con desviación estándar asintótica de

$$ASE(\tilde{\theta}) = \sqrt{1/(n_{11} + ,5) + 1/(n_{12} + ,5) + 1/(n_{21} + ,5) + 1/(n_{22} + ,5)}$$

En el ejemplo anterior $\tilde{\theta} = 1{,}828$ muy cercano a $\hat{\theta} = 1{,}832$

## 2.14. Intervalo de confianza para $\log(RR)$

También el riesgo relativo tiene una distribución muy asimétrica y de manera análoga el intervalo de confianza para $\log(RR)$ está dado por:

$$log(\hat{RR}) \pm z_{\alpha/2} * \sqrt{\frac{1 - p_1}{n1. * p_1} + \frac{1 - p_2}{n2. * p_2}}$$

### 2.14.1. Ejemplo caso y controles

Los datos se refieren a 262 mujeres de edad intermedia (menores a 69 años) que son admitidas en unidades médicas con infarto agudo al miocardio (MI) en un lapso de 5 años, cada caso es apareado con dos pacientes control recibidos en los mismos hospitales con algún otro padecimiento agudo. Se les clasifica como *si*, sin son fumadoras o exfumadoras y como *no* a aquellas que nunca han fumado. Por el diseño la distribución marginal de MI, está fija, habiendo 2 controles por cada caso. Estos estudios son conocidos como *caso y controles*, este diseño permite tener suficientes casos con la enfermedad (característica) de interés, y despues se buscan ciertas características en su historia clínica, es decir es un estudio retrospectivo.

Se desea comparar fumadoras versus no fumadoras en cuanto a la proporción de personas que sufren infarto al miocardio. Esto se refiere a la distribución marginal de MI dado el estatus de fumador. En esta muestra aproximadamente un tercio de ésta sufrió MI, no tiene sentido usar $1/3$ como estimador de la probabilidad de sufrir MI. ($1/3 = P(MI) = P(MI/F)P(F) + P(MI/NF)P(NF)$). Lo que si se puede calcular es la distribución condicional de ser fumador dado que se sufrió un MI.

```
    infarto
fuma control miocardio
  NO     346        90
  SI     173       172
```

```
        Pearsons Chi-squared test with Yates continuity correction

data:  xtabs(conteos ~ fuma + infarto, data = TABLA)
X-squared = 72.4241, df = 1, p-value < 2.2e-16
```

El cociente momios es:

```
> theta <- (172 * 346)/(90 * 173)
> theta

[1] 3.822222
```

Para las mujeres que sufrieron infarto, la proporción de fumadoras es:

```
> p1 <- 172/262
> p1

[1] 0.6564885
```

Para las mujeres que no sufrieron infarto, la proporción de fumadoras es:

```
> p2 <- 173/519
> p2
```

```
[1] 0.3333333
```

Como vimos que la probabilidad de sufrir infarto es pequeña para ambos grupos, entonces podemos pensar que el $r.r$ es parecido a 3.82. Entonces decimos que: *las mujeres que han fumado alguna vez tienen una probabilidad de sufrir un infarto casi 4 (3.82) veces mayor que las mujeres que no han fumado.*

## 2.15. La prueba exacta de Fisher

La prueba exacta de Fisher para tablas de 2 por 2 cuando se tiene $n's$ pequeñas y no se puede usar la aproximación por $\chi^2$. Para usar esta prueba se requiere que los totales marginales sean fijos La estadística está dada por:

$$P = \frac{(n_{11}+n_{12})!(n_{21}+n_{22})!(n_{11}+n_{21})!(n_{12}+n_{22})!}{n_{11}!n_{12}!n_{21}!n_{22}!N!}$$

La prueba exacta de Fisher calcula la probabilidad bajo la hipótesis de independencia de obtener un arreglo como el observado y la de otros arreglos que mostraran mayor evidencia de asociación, siempre suponiendo que las marginales son fijas. Si la suma de esas probabilidades es menor a $\alpha$ se rechaza la hipótesis de independencia.

Un ejemplo deja ver esto más claramente: La tabla de abjao tiene una entrada más pequeña que 5, que es 2 y esa celda es la más extrema en los datos observados.

```
> tabla2
```

```
           paciente
ideassuicid 1psy 2neur
        1si    2     6
        2no   18    14
```

La la prueba exacta de Fisher es:

$$probde2 = \frac{8!32!20!20!}{2!6!18!14!40!}$$

Las fecuencias más extremas que las observadas (que 2 en tabla2) corresponderían a tabla1 y tabla0:

```
> tabla1
```

```
           paciente
ideassuicid 1psy 2neur
        1si    1     7
        2no   19    13
```

```
> tabla0
```

```
           paciente
ideassuicid 1psy 2neur
        1si    0     8
        2no   20    12
```

Y entonces, para esas tablas se tendrían las probabilidades:

$$probde1 = \frac{8!32!20!20!}{1!7!19!13!40!}$$

y

$$probde0 = \frac{8!32!20!20!}{0!8!20!12!40!}$$

Para calcular entonces la probabilidad total de tener una tabla como la observada o una que sugiera alejarse más de la independencia es la suma de las tres anteriores, obteniedo los siguientes resultados:

```
> probde2
```

```
[1] 0.0957601
```

```
> probde1
```

```
[1] 0.02016002
```

```
> probde0
```

```
[1] 0.001638002
```

```
> probde2 + probde1 + probde0

[1] 0.1175581
```

Como 0,1175581 es mayor al $\alpha = 0{,}05$ entonces los datos **no** dan evidencia de que los psicóticos y los neróticos difieran en sus síntomas. En este caso con sólo calcular *probde*2 hubiese bastado pues esta cantidad es mayor a 0,05. Esta prueba es de una cola, es decir toma una sola dirección, mientras que la prueba de Ji cuadrada va en ambas direcciones. Pero la potencia de la prueba no es muy alta (probabilidad de rechazar la hipótesis nula cuando ésta no es cierta).

# Capítulo 3

# Modelos para datos binarios

## 3.1. Modelos Estadísticos

En un fenómeno de transmisión y recepción de información, se tiene una **seÃśal** que es distorsionada por **ruido**.

La señal puede verse como la parte determinista y el ruido como la parte aleatoria, estas son las dos componentes que forman un modelo estadístico. La componente del mensaje la describiremos como un función matemática que nos da las características principales, en la componente del ruido quedan las características que no alcanza a explicar la componente determinista. A partir de los datos trataremos de sacar la mayor información posible acerca de la señal, lo demás lo atribuiremos al ruido. De la variabilidad una parte se debe a la señal y otra al ruido. Un buen modelo será aquel en el que una gran parte de la variabilidad esté explicada por la función matemática elegida y quede poca variabilidad debida al ruido. Entre más parámetros metamos en la función matemática más explicaremos a los datos, sin embargo se busca también tener un modelo sencillo, fácil de interpretar, por lo que habrá que balancear ambos conceptos (explicar mucho vs sencillez del modelo: principio de parsimonia). Los parámetros son entonces esas cantidades que le dan forma

al modelo (rectas, curvas, superficies) y en función de éstos se plantearán las hipótesis de interés.

En un modelo donde se tiene una variable respuesta binaria $Y$, llamaremos $\pi = P(Y = 1)$ y $1 - \pi = P(Y = 0)$.

Se tiene una muestra $y_1, y_2, \ldots, y_n$ y $\pi_1, \pi_2, \ldots, \pi_n$ son sus probabilidades asociadas, cada una de éstas va cambiando. Se tratará de explicar estos cambios como función de la variable $X$ que toma valores $x_1, x_2, \ldots, x_n$.

Un primer modelo es el modelo lineal

$$\pi = X'\beta$$

que tiene la desventaja de no poder garantizar que los valores ajustados por $X'\hat{\beta}$ caigan dentro de [0,1].

Una forma de obligar a que caigan en el [0,1] es usar un modelo de la forma:

$$\pi = g^{-1}(X'\beta) = \int_{-\infty}^{t} f(s)ds$$

a $f(s)$ se le conoce como **función de tolerancia** y es una función de densidad es decir $f(s) \geq 0$ y $\int_{-\infty}^{\infty} f(s)ds = 1$.

## 3.2. Modelos de dosis-respuesta

En estos modelos la variable respuesta corresponde al proporción de seres muertos expuestos a diferentes dosis de sustancias tóxicas.

1. Cuando la función de tolerancia es constante en un intervalo

$$f(s) = \begin{cases} 1/(c_2 - c_1) & \text{si } c_1 \leq s \leq c_2 \\ 0 & \text{en otro caso} \end{cases}$$

entonces la proporción de muertes es de forma lineal:

$$\pi(x) = \int_{c_1}^{x} f(s)ds = \frac{x - c_1}{c_2 - c_1} = \frac{-c_1}{c_2 - c_1} + \frac{x}{c_2 - c_1} \text{ para} c_1 \leq x \leq c_2$$

Este modelo no es muy usado.

2. Cuando la tolerancia es la función normal

$$\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\frac{s-\mu}{\sigma})\right)^2 ds$$

En este caso

$$\pi(x) = \Phi(\frac{x-\mu}{\sigma})$$

a la dosis $x = \mu$ se le conoce como dosis letal mediana, que es la que se requiere para que mueran la mitad delos individuos. A este modelo se le conoce como modelo PROBIT

$$\Phi^{-1}(\pi) = \beta_o + \beta_1 x$$

con $\beta_o = -\mu/\sigma$ y $\beta_1 = 1/\sigma$.

3. El modelo logístico tiene como función de tolerancia a:

$$f(s) = \frac{\beta_1 \exp(\beta_o + \beta_1 s)}{[1 + exp(\beta_o + \beta_1 s)]^2}$$

y el modelo para $\pi$ queda:

$$\pi(x) = \frac{exp(\beta_o + \beta_1 x)}{1 + exp(\beta_o + \beta_1 x)}$$

si se despeja al componente lineal se obtiene la función **liga** en este caso se conoce como LOGIT.

$$log(\frac{\pi}{1-\pi}) = \beta_o + \beta_1 x$$

4. Cuando $f$ es la distribución de valores extremos:

$$f(s) = \beta_1 \exp((\beta_o + \beta_1 s) - \exp(\beta_o + \beta_1 s))$$

y entonces la proporcion $\pi$ es:

$$\pi = 1 - \exp(-\exp(\beta_o + \beta_1 x))$$

con función liga

$$\log(-\log(1-\pi)) = \beta_o + \beta_1 x$$

conocida como función COMPLEMENTARIA LOGLOG.

Entonces de las funciones que transforman a $\pi$ en $(-\infty, \infty)$, la logit y la probit son simétricas respecto a 0.5 y la cloglog no lo es. Para valores muy pequeños de $\pi$ apenas se distinguen la logit y la cloglog.

Se usa más la logit pues los parámetros pueden interpretarse como logaritmo del cociente de momios.

## 3.3. Ejemplo

En este ejemplo se tienen los conteos de mosquitos muertos luego de 5 horas de exposición a gas carbón bisulfito a diferentes concentraciones. El modelo1 es el logit, el modelo2 es el probit y modelo3 es el cloglog.

```
> dose <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.861, 1.8839)
> numberinsects <- c(59, 60, 62, 56, 63, 59, 62, 60)
> numberkilled <- c(6, 13, 18, 28, 52, 53, 61, 60)
> props <- numberkilled/numberinsects
> beetle.data <- data.frame(dose = dose, numberinsects = numberinsects, numbe
+     props = props)
> beetle.data

    dose numberinsects numberkilled     props
1 1.6907            59            6 0.1016949
2 1.7242            60           13 0.2166667
3 1.7552            62           18 0.2903226
4 1.7842            56           28 0.5000000
5 1.8113            63           52 0.8253968
6 1.8369            59           53 0.8983051
7 1.8610            62           61 0.9838710
8 1.8839            60           60 1.0000000
```

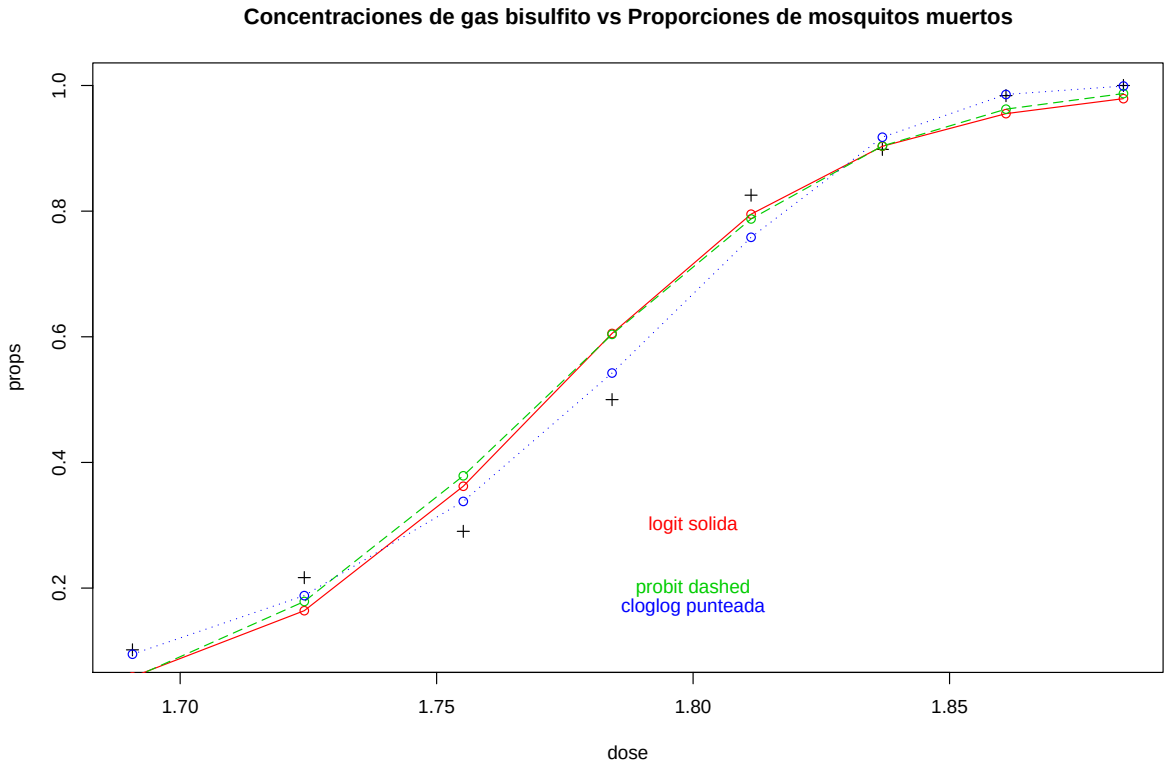
```
> modelo1 <- glm(props ~ dose, family = binomial(link = "logit"), weights = numberin
> summary(modelo1)

Call:
glm(formula = props ~ dose, family = binomial(link = "logit"),
    weights = numberinsects)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4

> modelo2 <- glm(props ~ dose, family = binomial(link = "probit"), weights = numberi
> summary(modelo2)

Call:
glm(formula = props ~ dose, family = binomial(link = "probit"),
```

```
    weights = numberinsects)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
dose          19.728      1.487   13.27   <2e-16 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  10.120  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4

> modelo3 <- glm(props ~ dose, family = binomial(link = "cloglog"), weights =
> summary(modelo3)

Call:
glm(formula = props ~ dose, family = binomial(link = "cloglog"),
    weights = numberinsects)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-0.80329  -0.55135   0.03089   0.38315   1.28883

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -39.572      3.240  -12.21   <2e-16 ***
dose          22.041      1.799   12.25   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   3.4464  on 6  degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4
```

En la gráfica siguiente puede verse que el modelo cloglog es el que mejor se ajusta los datos, luego el probit y luego el logit.

Para el modelo logístico puede verse para que $\pi = 0{,}5$ se requiere que $\log(\pi/(1-\pi)) = 0$ es decir que $\beta_0 + \beta_1 x = 0$ por lo tanto el punto $x = -\beta_0/\beta_1$ nos da el nivel $x$ de carbón bisulfito para el que se tiene la misma probabilidad de que el mosquito muera o que no, en este caso -(-66.717/34.270)=1.77. (verificar esto en la gráfica).


Concentraciones de gas bisulfito vs Proporciones de mosquitos muertos
props
dose
0.2
0.4
0.6
0.8
1.0
1.70
1.75
1.80
1.85
logit solida
probit dashed
cloglog punteada

# Capítulo 4

# Modelo Logístico

Un modelo loístico con $k$ variables explicativas se puede escribir como

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

y al despejar $\pi$ se tiene

$$\pi = \frac{exp(\beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + exp(\beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

Llamando $\eta$ a la componente lineal se tiene que

$$logit(\pi) = \eta$$

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

## 4.1. Estimación de parámetros

Los parámetros en regresión logística se estiman por máxima verosimilitud. Como las observaciones son Bernoullis, la función $L$ de verosimilitud queda:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)}$$

sacando logaritmo

$$\ln(L(\beta) = \sum_{i=1}^{n} y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i))$$

$$= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi(x_i)}{(1 - \pi(x_i))}\right) + \ln(1 - \pi(x_i))$$

Sustituyendo a la función liga y haciendo unas operaciones se obtiene

$$= \sum_{i=1}^{n} y_i \eta_i - \ln\left(1 + \exp(\eta_i)\right)$$

Para maximizar se deriva parcialmente con respecto a las betas, lo haré para el modelo con una sola variable explicativa.

$$\frac{\partial \ln(L(\beta)}{\partial \beta_o} = \sum \left(y_i - (1 + \exp(\beta_o + \beta_1 x_i))^{-1} \exp(\beta_o + \beta_1 x_i)\right)$$

$$= \sum \left(y_i - \frac{\exp(\eta_i)}{(1 + \exp(\eta_i))}\right)$$

$$= \sum \left(y_i - \pi(x_i)\right)$$

De manera análoga para $\beta_1$

$$\frac{\partial \ln(L(\beta)}{\partial \beta_1} = \sum \left(x_i y_i - (1 + \exp(\beta_o + \beta_1 x_i))^{-1} \exp(\beta_o + \beta_1 x_i) x_i\right)$$

$$= \sum x_i \left(y_i - \pi(x_i)\right)$$

Ahora se igualan las parciales a cero y se obtiene un sistema de ecuaciones NO LINEAL, para resolverlo se usan métodos iterativos, el más usado es el de iteración de cuadrados ponderados.

## 4.2. Interpretación de los parámetros

Un modelo muy sencillo es aquel donde se relaciona la probabilidad de ocurrencia de una enfermedad con un solo factor de exposición. Las categorías de

exposición son expuesto y no expuesto. Supongamos que los datos están agrupados y por lo tanto se tienen las proporciones de individuos que desarrollaron la enfermedad en los grupos de expuestos y no expuestos. Así que el modelo se escribe:

$$logit(p_j) = \beta_0 + \gamma_j,\ j = 1, 2.$$

En este modelo tendría sobreparametrización (más parámetros que datos), así que se puede escribir como

$$logit(p_j) = \beta_0 + \beta_1 x_j$$

donde $x_j$ es una variable indicadora que vale cero para los no expuestos y uno para los expuestos.

Despejando el cociente de momios de la ecuación anterior

$$\frac{p_1}{1-p_1} = \exp\beta_0$$
$$\frac{p_2}{1-p_2} = \exp(\beta_0 + \beta_1)$$

Y el cociente de momios es:

$$\phi = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \exp\beta_1$$

De aquí que $\beta_1 = \log(\phi)$

Luego de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$ el estimador del cociente de momios es $\hat{\phi} = \exp(\beta_1)$

Por propiedades de los modelos lineales generalizados se sabe que la desviación estándar de $\hat{\beta}_1$ es la misma que la desviación de $\log(\hat{\phi})$ esto es $\sigma_{\hat{\beta}_1} = \sigma_{log(\phi)}$ y se puede calcular un intervalo de confianza para $\log\phi$ de la manera siguiente:

$$\log(\hat{\phi}) \pm z_{\alpha/2}\sigma_{\log(\hat{\phi})}$$

Y para ontener un intervalo para $\phi$ solo exponenciamos los extremos del intervalo de arriba.

## 4.3. Ejemplo

En un estudio acerca de enfermedades respiratorias en infantes, donde se registra si los niños desarrollan bronquitis o neumonía en su primer anõ de vida, su tipo de alimentación y su género. Los datos aparecen enla librería de **faraway**

```
> library(faraway)
> data(babyfood)
> babyfood

  disease nondisease  sex   food
1      77        381  Boy Bottle
2      19        128  Boy  Suppl
3      47        447  Boy Breast
4      48        336 Girl Bottle
5      16        111 Girl  Suppl
6      31        433 Girl Breast

> xtabs(disease/(disease + nondisease) ~ sex + food, babyfood)

      food
sex         Bottle     Breast      Suppl
  Boy  0.16812227 0.09514170 0.12925170
  Girl 0.12500000 0.06681034 0.12598425

> md1 <- glm(cbind(disease, nondisease) ~ sex + food, family = binomial, babyfood)
> summary(md1)

Call:
glm(formula = cbind(disease, nondisease) ~ sex + food, family = binomial,
    data = babyfood)


Deviance Residuals:
```

```
      1        2        3        4        5        6
 0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4

> exp(-0.669)

[1] 0.5122205
```

Con la alimentación al pecho los momios de enfermarse son 51 % de los momios de los que toman biberón.

También podemios decir: La alimentación con pecho reduce los momios de enfermedades respiratorias al 51 % de los momios de los alimentados con botella.

```
> exp(-0.669 - 1.96 * 0.153)
```

```
[1] 0.3795078
```

```
> exp(-0.669 + 1.96 * 0.153)
```

```
[1] 0.6913424
```

Las cantidades de arriba (.37,.69) dan un intervalo de confianza 95 % aproximado para ese cociente de momios.

```
> exp(-0.313)
```

```
[1] 0.7312499
```

En las niñas los momios de enfermarse son un 73 % de los momios en los niños.

## 4.4. Bondad de ajuste modelo logístico

Una manera de ver si el modelo de **estudio** es adecuado, es compararlo contra otro modelo más general, uno que tenga tantos parámetros como datos observados, el modelo **saturado** (con la misma función de distribución y liga que el modelo bajo estudio).

Si se designa por $\beta_{max}$al vector de parámetros del modelo saturado y $b_{max}$ al estimador de máxima verosimilitud de $\beta_{max}$. Sea $L(b_{max}; y)$ la verosimilitud evaluada en $b_{max}$, que será la mayor que puede obtenerse. Y sea $L(b; y)$ la verosimilitud del modelo estudiado, entonces la razón de verosimilitudes

$$\lambda = \frac{L(b; y)}{L(b_{max}; y)}$$

arroja un buena medida de la bondad del modelo, o bien puede utilizarse la logverosimilitud $\log(\lambda) = l(b; y) - l(b_{max}; y)$.

La estadística

$$D = -2 \log \frac{L(b; y)}{L(b_{max}; y)} = -2[l(b; y) - l(b_{max}; y)]$$

se le conoce como *deviance*. Entonces, valores grandes de la *deviance*, sugieren que el modelo en estudio da una pobre descripción de los datos, en relación a la que da el modelo saturado. Para poder decir que tan grande debe ser esa diferencia, se requiere conocer la distribución de $\log(\lambda)$.

Bajo el supuesto de que el modelo estudiado es tan bueno como el modelo saturado, esta $D$ se distribuye asintóticamente como una $\chi^2_{n-p}$ siendo $n$ el número de datos observados y $p$ número de parámetros en el modelo estudiado.

Cuando las $Y_i's$ tienen una distribución $Bin(n_i, \pi_i)$

$$l(\beta, y) = \sum_{i=1}^{n} [y_i \log(\pi_i) - y_i \log(1-\pi_i) + n_i \log(1-\pi_i) + \log(\binom{n_i}{y_i})]$$

Para el modelo saturado, todas las $\pi_i$ son diferentes entre si, así que $\beta = (\pi_1, \ldots, \pi_n)$ y los estimadores máximo verosímiles son $\hat{\pi}_i = y_i/n_i$, evaluando la logverosimilitud en esos estimadores se tiene:

$$l(b_{max}, y) = \sum_{i=1}^{n} [y_i \log(y_i/n_i) - y_i \log(\frac{n_i - y_i}{n_i}) + n_i \log(\frac{n_i - y_i}{n_i}) + \log(\binom{n_i}{y_i})]$$

Para el modelo en estudio $\hat{y}_i = n_i\hat{\pi}_i$ y su verosimilitud es:

$$l(b, y) = \sum_{i=1}^{n} [y_i \log(\hat{y}_i/n_i) - y_i \log(\frac{n_i - \hat{y}_i}{n_i}) + n_i \log(\frac{n_i - \hat{y}_i}{n_i}) + \log(\binom{n_i}{y_i})]$$

y finalmente

$$D = -2 \sum_{i=1}^{n} [y_i \log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i) \log(\frac{n_i - y_i}{n_i - \hat{y}_i})]$$

NOTA IMPORTANTE

Cuando los datos están desagrupados, es decir son Bernoulli, la logverosimilitud es cero, pues $y_i \log(yi)$ y$(1-y_i)\log(1-y_i)$ valen cero, ya que $y_i \in \{0, 1\}$ y entonces la *deviance* vale:

$$D = -2\sum_{i=1}^{n}[\hat{\pi}_i \log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + \log(1-\hat{\pi}_i)]$$

**Esta cantidad NO compara los valores ajustados con los observados!!! Por tanto NO sirve como medida para evaluar la bondad del ajuste.**

El principal problema en el caso de tener datos binarios es que el modelo saturado tiene tantos parámetros como observaciones y si crece el número de observaciones, también crece el número de parámetros y esto repercute en que la distribución asintótica de la *deviance* está lejos de ser una ji cuadrada. Sin embargo esta cantidad **sí** puede ser usada para la selección de variables en el modelo, pues se hace a través de diferencias de *deviances*, que anulan el efecto del modelo saturado.

## 4.5. Estadística Hosmer-Lemeshow

Hosmer y Lemeshow propusieron una cierta manera de agrupar los datos y luego calculan un estadística tipo ji cuadrada.

Hay dos maneras de hacer los grupos: 1) agrupar los datos de acuerdo a los intervalos definidos por los percentiles de las probabilidades estimadas o 2) agrupar de acuerdo a intervalos fijos de las probabilidades estimadas.

Por ejemplo para el caso 1), una vez ordenadas de menor a mayor las probabilidades estimadas, si g=10 y se tienen $n$ observaciones, entonces el primer grupo contiene a las $n'_1 = n/10$ observaciones con las $\hat{p}$ más pequeÃśas, y el último grupo a las $n'_{10} = n/10$ observaciones con las $\hat{p}$ más grandes. Para el caso 2) el primer grupo contendría a las observaciones con $\hat{p} \leq 0{,}1$, el segundo grupo con las observaciones con $\hat{p} \in [0{,}1, 0{,}2]$, hasta el décimo grupo con $0{,}9 \leq \hat{p} \leq 1$

En cualquiera de los casos, los valores **esperados** para las observaciones

$y = 1$ en el grupo $g \in \{1, \ldots, 10\}$ será $\sum_{i=1}^{n_g} \hat{\pi}_i$ y para las observaciones $y = 0$ será $\sum_{i=1}^{n_g}(1 - \hat{\pi}_i)$. La estadística propuesta resulta de hacer una tabla de gx2, con los observados y los esperados y la estadística queda como:

$$\hat{C} = \sum_{k=1}^{g} \left[ \frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right]$$

donde

$$o_{1k} = \sum_{j=1}^{c_k} y_j,$$

es decir el nÃžmero observado de unos en el grupo $k$,

$$o_{0k} = \sum_{j=1}^{c_k} (m_j - y_j),$$

es decir el nÃžmero observado de ceros en el grupo $k$,

$$\hat{e}_{1k} = \sum_{j=1}^{c_k} m_j \hat{\pi}_j,$$

es decir el nÃžmero esperado de unos en el grupo $k$,

$$\hat{e}_{0k} = \sum_{j=1}^{c_k} m_j (1 - \hat{\pi}_j)$$

es decir el nÃžmero esperado de ceros en el grupo $k$,

$c_k$ es el número de patrones de covariables del $k$-ésimo grupo,

esta $\hat{C}$ se puede reescribir como

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_{1k} - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - n'_k \bar{\pi}_k)}$$

donde $\bar{\pi}_k$ es el promedio de las probabilidades estimadas del grupo $k$

$$\bar{\pi}_k = \frac{1}{n'_k} \sum_{j=1}^{c_k} m_j \hat{\pi}_j$$

$\hat{C}$ se distribuye aproximadamente como una ji cuadrada con $g-2$ grados de libertad. Existen varias medidas de bondad de ajuste para datos desagrupados más, pero no las veremos pues casi ningún paquete las tiene actualmente.